


7-22-2020

Exploring the Factors Influencing Big Data Technology Acceptance

Mohammad Nayemur Rahman
Portland State University

Follow this and additional works at: https://pdxscholar.library.pdx.edu/open_access_etds

 Part of the [Management Sciences and Quantitative Methods Commons](#), and the [Technology and Innovation Commons](#)

Let us know how access to this document benefits you.

Recommended Citation

Rahman, Mohammad Nayemur, "Exploring the Factors Influencing Big Data Technology Acceptance" (2020). *Dissertations and Theses*. Paper 5515.
<https://doi.org/10.15760/etd.7389>

This Dissertation is brought to you for free and open access. It has been accepted for inclusion in Dissertations and Theses by an authorized administrator of PDXScholar. For more information, please contact pdxscholar@pdx.edu.

Exploring the Factors Influencing Big Data Technology Acceptance

by

Mohammad Nayemur Rahman

A dissertation submitted in partial fulfillment of
the requirements for the degree of

Doctor of Philosophy
in
Technology Management

Dissertation Committee:
Tugrul Daim, Chair
Robert Fountain
Nuri Basoglu
Rafaa Khalifa

Portland State University
2020

© 2020 Mohammad Nayemur Rahman

Abstract

The success of new technology depends on user acceptance. Therefore, discovering the antecedents of technology use is pivotal to overcoming the lack of user acceptance in the field of technology adoption. Factors of critical technological capability, in particular, are overlooked and largely neglected in the literature. Accordingly, the body of literature on the field of technology adoption is inconclusive as to which technological capability factors influence technology acceptance.

Big Data has received great attention in academic literature and industry papers. Most of the experiments and studies focused on publishing results of big data technologies development, machine learning algorithms, and data analytics. To the best of our knowledge, there is not yet any comprehensive empirical study in the academic literature on big data technology acceptance. This research makes an attempt to identify factors influencing big data technology acceptance from an industrial-organizational context. With the help of existing technology acceptance theories, literature studies, industry technical papers, and vendor publications on data management technologies ranging from conventional data warehousing to big data storage technologies (e.g., Hadoop Distributed File System), 32 factors have been identified for use in the qualitative study of this research.

By using prominent qualitative research methods including focus groups and one-on-one interviews, this research has identified 12 factors as possible antecedents of

perceived usefulness and intention to use big data technology. These 12 factors include scalability, data storage and processing capabilities, functionality, performance expectancy, security and privacy considerations, reliability, data analytics capability, flexibility, facilitating conditions, output quality, required skills and training, and cost-effectiveness. The qualitative studies were conducted using industry experts with experience in big data technologies as well as traditional data management technologies.

To further validate the factors identified by the qualitative study, a quantitative model is developed. The theoretical foundation of this model is drawn from the Technology Acceptance Model (TAM) developed by Fred Davis (1993). This model allows plugins of external factors to its latent constructs of perceived usefulness (PU) and perceived ease of use (PEOU).

Primary data for the quantitative study were collected from big data (Hadoop User Groups) users in the United States who work in different industries including software and internet services, financial services, healthcare, consulting and professional services, telecommunications, manufacturing, retail, marketing, and logistics. The structural equation modeling (SEM) software, AMOS, was used for empirical verification and validation of our proposed model using 349 survey responses.

The statistical results of this model provide a compelling explanation of the relationships among the antecedent variables and the dependent variables. The analysis of the structural model reveals that the hypothesis tests are significant for eight out of

12 path relationships. This study successfully tests and validates four new variables relating to technological capabilities in adopting new technology: scalability, data storage and processing capability, flexibility, and reliability. The study finds the other four out of the eight variables significant which have also been validated by prior studies: performance expectancy, facilitating conditions, output quality, and required skills and training. Four external variables are found to be non-significant by the proposed model: functionality, security and privacy considerations, data analytics capability, and cost-effectiveness. Our proposed structural model also supports all core constructs of the TAM: perceived usefulness, perceived ease of use, behavioral intention, and actual use.

The model is strongly supported in three important points of measurement which accounts for 80% of the variance in usefulness perceptions, 67% of the variance in usage intentions, and 85% in actual Hadoop usage. These findings make significant contributions to advance theory and provide insights to the foundation for future research to improve our understanding of user acceptance behavior.

Industry big data professionals are the subjects of both qualitative and quantitative studies of this research; therefore, we assert that the industry provides an important input for enhancing the existing TAM model and building information systems (IS) theory. From the practitioners' point of view, this research provides companies with guidance on which technological features and capabilities to look for when buying a

complex form of technology. Limitations of this study are discussed, and several promising new research directions are provided.

Dedication

This dissertation is dedicated to the memory of my father, Md. Lutfur Rahman; my mother, Begum Tahera Khatun; and my elder brother, Md. Aminur Rahman; may Allah have mercy upon them.

About parents Allah says in the Quran (interpretation of the meaning): "We have enjoined on man kindness to his parents; in pain did his mother bear him, and in pain did she give him birth" (Al-Ahqaf; 46:15).

My mother is a source of inspiration for me to be a hard-working person. I remember her hard work and dedication in raising us, the family of eight siblings, brothers and sisters, in a rural area of Bangladesh while my father had to live in a city for job purposes.

My father is a role model for me from morality, justice, and spiritual perspectives.

Lastly, and with gratitude, I remember my elder brother who brought me to live in the city to help me pursue my undergraduate degree. He was instrumental in my academic career development. He returned to Allah while I was pursuing my Ph.D.

Acknowledgments

Paul A. Samuelson (the first American to win the Nobel Memorial Prize in Economic Sciences) wrote in 1946 about the most important macroeconomics book titled *The General Theory of Employment, Interest and Money*, by the British economist John Maynard Keynes:

"it is a badly written book, poorly organized... It is arrogant, bad-tempered, polemical, and not overly generous in its acknowledgements... Flashes of insight and intuition intersperse tedious algebra... When finally mastered [after how many readings?], its analysis is found to be obvious and at the same time new. In short, it is a work of genius."

As regards this dissertation, some readers might not find it easy to read because English is my second language. Nonetheless, I hope it provides some scattered pictures of big data technology acceptance that my readers find valuable.

As I reflect on this journey toward a Ph.D. degree in the College of Engineering and Computer Science, I am amazed by the magnitude of support I have received to make this journey possible.

First and foremost, I would like to express my gratitude to my learned dissertation committee Professor Tugrul Daim, Professor Robert Fountain, Professor Nuri Basoglu, and Dr. Rafea Khalifa for their time, effort and mentoring.

In the words of the Roman statesman and philosopher, Marcus Tullius Cicero:
"The authority of those who teach is often an obstacle to those who want to learn."

I hereby solemnly declare that their authority was no obstacle to my road to learning new things. Their feedback, questions, and efforts to help me complete this dissertation are so appreciated.

I would especially like to thank Dr. Daim, my advisor. He has allowed me much room for creativity in pursuing my research. I am also thankful for Dr. Daim's availability over the years. I had inexhaustible email communications with him, and it took him only a few minutes to a few hours to reply to my emails.

Thanks are due to Dr. Fountain for suggesting improvements, especially in the survey instrument design and statistical analysis sections. We discussed many contemporary issues of our social life as well, and I found our friendly conversation to be of great support.

I greatly appreciate Dr. Basoglu for being detail-oriented in reviewing the first draft of my dissertation and providing constructive feedback.

Many thanks to Dr. Rafea Khalifa for having a bird's-eye view of my dissertation and providing pristine feedback as a reader.

This dissertation was written without any financial support from any institution. I, however, take this opportunity to gratefully acknowledge the financial support I received from my employer to support my graduate work.

I am profoundly grateful to my parents for their tremendous contribution to my upbringing and making me the person I am today.

The person who was the happiest and the person who sacrificed the most is my wife, Shameem Akhter. I can never repay the patience, love, and support over these many years from her. Her constant encouragement, prayers, and sacrifices have earned her the highest honors available.

My sons Rabeeb Rahman, Osman Nayeem, Abrar Nayeem, and daughter Mahjubah Nayeem deserve a big thanks for their understanding, patience, and support. Without their love and support, writing this dissertation would not have been possible. I have promised myself that I will give them more time in the coming days to make up for the lost days.

My appreciation certainly goes to my friends and colleagues Dr. Md. Abu Saleh, Dr. James Gaskin, Dr. Hillol Bala, Dr. Bakhtear Talukdar, Mr. Andy Wong, and Mr. Vipul Kapadia, to name a few, for supporting me in various ways. I appreciate their encouragement during my doctoral journey.

Lastly, I want to thank the industry participants for taking part in the qualitative studies and making the final survey of my research possible. It is tough to conduct research without support from others.

Table of Contents

Abstract.....	i
Dedication.....	v
Acknowledgments.....	vi
List of Tables	xiv
List of Figures	xvi
List of Acronyms.....	xvii
Chapter 1 Research Objectives and Overview.....	1
1.1 Introduction.....	1
1.2 Big Data.....	5
1.3 Characteristics of Big Data	7
1.4 Big Data Technology and Evolution.....	11
1.5 An Overview of Two Hadoop-Based Application Systems	15
1.6 Big Data Market.....	17
1.7 Research Objectives.....	18
1.8 Research Approach	19
1.9 Statement of Problem.....	21
1.10 Research Questions	22
1.11 Significance of Studying Big Data Technology Acceptance	24
Chapter 2 Literature Review	26
2.1 Relevant Theories Used to Study the Adoption and Use of IS.....	26
2.1.1 Theory of Reasoned Action.....	27
2.1.2 Theory of Planned Behavior.....	28
2.1.3 Diffusion of Innovation.....	29
2.1.4 Technology Acceptance Model.....	31
2.1.5 Technology, Organization and Environment	34
2.1.6 Resource Based View.....	35
2.1.7 Unified Theory of Acceptance and Use of Technology.....	36
2.2 Studies Related to Technology Adoption.....	37
2.3 Taxonomy Factors.....	56
2.4 Research Related to Big Data Technology Adoption.....	58

2.5 Research Gaps	60
Chapter 3 Developing Research Model and Research Hypotheses.....	63
3.1 Defining Perceived Usefulness	64
3.2 Brainstorming Session.....	67
3.3 Focus Group Session.....	71
3.4 Individual Interviews	72
3.5 Results of the Qualitative Studies	75
3.6 Developing Research Model	79
3.7 Proposed Research Model.....	84
3.8 Developing Research Hypotheses	86
3.8.1 Hypothesis H1 - Scalability	86
3.8.2 Hypothesis H2 - Data Storage & Processing.....	87
3.8.3 Hypothesis H3 - Cost Effectiveness.....	87
3.8.4 Hypothesis H4 - Performance Expectancy.....	88
3.8.5 Hypothesis H5 - Security and Privacy Considerations.....	89
3.8.6 Hypothesis H6 - Reliability.....	89
3.8.7 Hypothesis H7 - Data Analytics Capability	90
3.8.8 Hypothesis H8 - Training and Required Skills	91
3.8.9 Hypothesis H9 - Flexibility	91
3.8.10 Hypothesis H10 - Output Quality.....	91
3.8.11 Hypothesis H11 - Functionality.....	92
3.8.12 Hypothesis H12 - Facilitation Conditions.....	93
3.8.13 Hypothesis H13 - Perceived Usefulness.....	93
3.8.14 Hypothesis H14 - Perceived Ease of Use	94
3.8.15 Hypothesis H15 - Behavioral Intention	94
Chapter 4 Research Methodology	95
4.1 Research Design	95
4.2 Survey Instrument Development	95
4.3 Instrument Validation Steps.....	97
4.3.1 Instrument Validation Phase One	100
4.3.2 Instrument Validation Phase Two	104

4.3.3 Pilot Test Results	105
4.4 Instrument Reliability	105
4.5 Instrument Administration	106
4.6 Sampling Strategy	107
4.6.1 Sampling Methods	107
4.6.2 Targeted Population	108
4.6.3 Sampling Frame	109
4.6.4 Sample Size	110
4.6.5 Approaches to Increase Sample Size	117
4.6.6 Approaches to Address Concern with Low Responses	119
4.6.7 Survey Administration	121
Chapter 5 Data Screening, Measurement Development and Structural Model Testing	123
5.1 Sample Demographics and Data Screening	123
5.2 Measurement Development	126
5.3 Confirmatory Factor Analysis	130
5.3.1 CFA: Scalability	135
5.3.2 CFA: Data Storage and Processing	136
5.3.3 CFA: Cost-Effectiveness	137
5.3.4 CFA: Performance Expectancy	138
5.3.5 CFA: Security and Privacy	139
5.3.6 CFA: Reliability	140
5.3.7 CFA: Data Analytics Capability	141
5.3.8 CFA: Training and Required Skills	142
5.3.9 CFA: Flexibility	143
5.3.10 CFA: Output Quality	144
5.3.11 CFA: Functionality	144
5.3.12 CFA: Facilitating Conditions	145
5.3.13 CFA: Perceive Usefulness	146
5.3.14 CFA: Perceived Ease of Use	147
5.3.15 CFA: Behavioral Intention	148
5.3.16 CFA: Actual Use	149

5.3.17 Overall Measurement Model Fit.....	150
5.4 SEM Path Analysis – A Hypothesized Model	152
5.5 Discriminant Validity.....	158
Chapter 6 Hypotheses Testing and Discussion	161
6.1 Hypotheses Testing.....	161
6.1.1 Scalability and Perceived Usefulness	163
6.1.2 Data Storage and Processing, and Perceived Usefulness.....	166
6.1.3 Flexibility and Perceived Usefulness	168
6.1.4 Data Analytics Capability and Perceived Usefulness.....	170
6.1.5 Output Quality and Perceived Usefulness.....	171
6.1.6 Performance Expectancy and Perceived Usefulness.....	173
6.1.7 Reliability and Perceived Usefulness.....	174
6.1.8 Security and Privacy, and Perceived Usefulness.....	175
6.1.9 Training and Skills, and Perceived Usefulness	176
6.1.10 Functionality and Perceived Usefulness.....	178
6.1.11 Perceived Ease of Use and Perceived Usefulness.....	179
6.1.12 Perceived Usefulness and Behavioral Intention to Use	181
6.1.13 Perceived Ease of Use and Behavioral Intention to Use.....	183
6.1.14 Facilitating Conditions and Actual Use.....	184
6.1.15 Cost-Effectiveness and Actual Use	186
6.1.16 Behavioral Intention to Use and Actual Use	187
6.2 Controlling Common Method Biases.....	188
6.3 Non-Response Error: Wave Analysis	190
6.4 Summary of the Chapter	194
Chapter 7 Conclusions, Research Contributions, Limitations, Research Direction	196
7.1 Theoretical Contribution	199
7.2 Implications for Practitioners.....	202
7.3 Implications for Researchers	205
7.4 Limitations.....	205
7.5 Future Research Direction	208
References	210

Appendices.....	234
Appendix A: Cover Letter and Survey Questionnaire.....	234
Appendix B: Pilot Test Survey Questionnaire	241
Appendix C: Initial Survey Questionnaire Validation.....	242
Appendix D: Hadoop User Groups in the U.S.	243
Appendix E: Final CFA	244
Appendix F: Cronbach's Alpha	245
Appendix G: EFA – Pattern Matrix	246
Appendix H: Technology Acceptance Factors.....	247

List of Tables

Table 1: Big Data Characteristics – 5 V's.....	8
Table 2: Relevant Theories to Study Adoption of Information Technology	27
Table 3: Summary of TAM Studies (1989-2019).....	32
Table 4: Taxonomy of Factors Based on Literature Review	56
Table 5: Empirical Research on Big Data Technology Adoption.....	59
Table 6: Research Gaps and Research Goals	62
Table 7: Participants in the Brainstorming Session	70
Table 8: Summary of Steps to Develop the Qualitative Study	74
Table 9: Results of Qualitative Study	75
Table 10: Final List of Factors for Use in the Proposed Research Model	78
Table 11: Steps to Validate Survey Instrument	99
Table 12: Example of Measures from Survey Instrument	101
Table 13: Survey Respondents' Job Profiles	124
Table 14: Survey Respondents' Company Profiles	125
Table 15: Survey Questions Ratings.....	131
Table 16: Summary of Initial Findings (CFA): Scalability.....	135
Table 17: Summary of Initial Findings (CFA): Data Storage and Processing.....	137
Table 18: Summary of Initial Findings (CFA): Cost-Effectiveness	138
Table 19: Summary of Initial Findings (CFA): Performance Expectancy.....	138
Table 20: Summary of Initial Findings (CFA): Security and Privacy Considerations	139
Table 21: Summary of Initial Findings (CFA): Reliability	140
Table 22: Summary of Initial Findings (CFA): Data Analytics Capability.....	141
Table 23: Summary of Initial Findings (CFA): Training and Required Skills	142
Table 24: Summary of Initial Findings (CFA): Flexibility.....	143
Table 25: Summary of Initial Findings (CFA): Output Quality.....	144
Table 26: Summary of Initial Findings (CFA): Functionality	145
Table 27: Summary of Initial Findings (CFA): Facilitating Conditions	146
Table 28: Summary of Initial Findings (CFA): Perceive Usefulness.....	147

Table 29: Summary of Initial Findings (CFA): Perceived Ease of Use.....	147
Table 30: Summary of Initial Findings (CFA): Behavioral Intention	148
Table 31: Summary of Initial Findings (CFA): Actual Use.....	149
Table 32: Single Measurement Model – Estimates and Fit Indices.....	150
Table 33: Summary of Overall Measurement Model (CFA).....	151
Table 34: Regression Weights – Path Model: Results of Five Iterations	155
Table 35: CFA Construct Reliability	156
Table 36: Summary of Overall CFA: Fit Indices	156
Table 37: Summary of Overall Path Model.....	158
Table 38: Path Model Standard Regression Weights	158
Table 39: Discriminant Validity Analyses	159
Table 40: Path Model Estimates	162
Table 41: Single Factor Total Variance Explained	189
Table 42: Survey Wave Analysis - Perceived Usefulness	192
Table 43: Survey Wave Analysis - Perceived Ease of Use.....	192
Table 44: Survey Wave Analysis - Behavioral Intention	193
Table 45: Survey Wave Analysis - Actual Use	193

List of Figures

Figure 1: Hadoop and Reporting Application	16
Figure 2: Proposed Research Model	84
Figure 3: Confirmatory Factor Analysis (CFA)	133
Figure 4: Final Research Model – Big Data Technology Acceptance	157
Figure 5: Path Diagram (SEM) of the Final Research Model.....	157

List of Acronyms

Acronym	Definition
AI	Artificial Intelligence
AMOS	Analysis of Moment Structures
ANOVA	Analysis of Variance
AU	Actual Use
AVE	Average Variance Extracted
AWS	Amazon Web Services
BI	Behavioral Intention
CEO	Chief Executive Officer
CFI	Comparative Fit Index
CIO	Chief Information Officer
CMIN	Chi-square Statistics in AMOS
COST	Cost-Effectiveness
CR	Composite Reliability
CTO	Chief Technology Officer
DA	Data Analytics Capability
DB	Database
DF	Degrees of Freedom
DOI	Diffusion of Innovation
DQ	Data Quality
DS	Data Storage and Processing
DV	Dependent Variable
CFA	Confirmatory Factor Analysis

EFA	Exploratory Factor Analysis
EMR	Amazon Elastic MapReduce
ERP	Enterprise Resource Planning
ETL	Extract, Transform, Load
FC	Facilitating Conditions
FL	Flexibility
FN	Functionality
GCS	Google Cloud Storage
HDFS	Hadoop Distributed File System
IFI	Incremental Fit Index
IS	Information Systems
IT	Information Technology
IV	Independent Variable
ML	Machine Learning
RMSEA	Root Mean Square Error of Approximation
OQ	Output Quality
PE	Performance Expectancy
PEOU	Perceived Ease of Use
PSU	Portland State University
PU	Perceived Usefulness
RBV	Resource-Based View
RL	Reliability
SC	Scalability
SEM	Structural Equation Modeling
SOX	Sarbanes-Oxley Act

SP	Security and Privacy
SPSS	Statistical Package for the Social Sciences
SQL	Structural Query Language
S3	Simple Storage Service
TAM	Technology Acceptance Model
TLI	Tucker Lewis Index
TOE	Technology, Organization and Environment
TPB	Theory of Planned Behavior
TR	Training and Skills
TRA	Theory of Reasoned Action
UTAUT	Unified Theory of Acceptance and Use of Technology

Chapter 1 Research Objectives and Overview

1.1 Introduction

Data, data everywhere (The Economist, 2010). Data has hit the big time with ‘big data.’ In the early twenty-first century, the term ‘big data’ has received great attention in computer science, data science, technology management, and information systems (IS) literature (Agarwal & Dhar, 2014; Chen et al., 2012; George et al., 2014; Goes, 2014; Grover et al., 2020; Hilbert, 2016; Jain et al., 2016; Jin et al., 2015; Kambatla et al., 2014; McAfee & Brynjolfsson, 2012; Singh & Reddy, 2015; Tsai et al., 2015). However, references to ‘big data technology acceptance’ are scarce in the practitioner and research papers (Caesarius & Hohenthal, 2018; Kwon et al., 2014; Surbakti et al., 2020). This section explores the concept of technology acceptance. In technology acceptance discipline, technology acceptance is synonymous with user acceptance. The extant literature spells out the concept of acceptance as below (Dillon & Morris, 1996, p. 3).

The “user acceptance is defined as the demonstrable willingness within a user group to employ information technology for the tasks it is designed to support. Thus, the concept is not being applied to situations in which users claim they will employ it without providing evidence of use.”

The stakes are high for technology developers, practitioners, and researchers for getting a technology accepted by its intended users, given that millions of dollars are invested in technology development and procurement. Understanding why potential

users accept technology is important because that helps in designing and developing better methods.

Consistent with the concept of acceptance presented by Dillon and Morris (1996), current research proposes an operational definition of technology acceptance from the technological rigor and complexity that is encountered in an industry setting. Past research (Hess et al., 2014; Lee et al., 2003) synthesize the term technology acceptance from an individual and organization perspective consisting of non-technical constructs and items. One of the existing models, TOE, is defined consisting of technology, organization, and environment (Chau & Tam, 1997). In this model the keyword technology is mentioned but, technical factors have not been identified. Fred Davis (Davis, 1993) develops the technology acceptance model (TAM). As part of the technology acceptance model by Davis (Davis, 1993), the latent constructs like perceived usefulness (PU) and perceived ease of use (PEOU) have been named but Benbasat and others criticize this as having a lack of proper definitions of these two terms (Benbasat & Barki, 2007, Chuttur, 2009). Hence, they consider these two terms as a black box (Lee et al., 2003). The question of concern is, to what (specific) factors make technology useful? This dissertation makes an attempt to look at the PU and PEOU from a technical implication standpoint. This researcher makes an attempt to come up with an operational definition of these terms based on current-day technological aspects and the utility theory of economics (Bentham, 1824; Kapteyn, 1985; Stigler, 1950). Then the

researcher develops big data technology acceptance model based on Davis' TAM (Davis, 1993).

Regarding technology acceptance from industry context, some researchers (Kwon et al., 2014; Russom, 2013) suggest that acceptance by the CEO, CIO, or CTO is reasonable to understand the acceptance of the technology. However, these C-suite executives make decisions based on certain factors that may not comply with the constructs of the TAM and UTAUT technology acceptance models as proposed by Davis (1989) and Venkatesh (2003) respectively. For example, the TAM by Davis (1989, 1993) contains certain external factors as well as internal constructs (PU, PEOU, BI, AU) but CEOs might take decisions by completely bypassing them. This researcher observes based on his industry experience that a CEO might consider purchasing a certain tool or technology which might be inefficient from a usage perspective. But the CEO expects that their own company's products be purchased by that company to reciprocate. A company might have an alliance with another company and hence make a decision to purchase the alliance company's B-class product. These purchase decisions ignore the basics of technology acceptance models.

The present study takes the technology acceptance models from a practical usage perspective. As such, the author asserts that technology acceptance decisions need to come from the real users of a company as opposed to company executives. Company executives are not supposed to know the technical details or features of technology (Wheelock, 2013). Hence, they cannot answer the survey that contains questions on

technical features as well as challenges encountered in using the technology. In order to give acceptance decisions, a person needs to have hands-on experience of the tool or technology. That way, actual users can provide valuable inputs about different features of a technology. This is compliant with the Dillon and Morris (1999) paper which suggests that one needs to be a real user to be an evaluator as well as an adopter of technology. Dillon and Morris (1996) state that Taylor's theory was to get things done by employees, using financial rewards, regardless of whether they like it or not. But, in today's world, it is not that easy to motivate users to get things done with a technology that they do not like.

Silva (1997) observes that in many cases information technology adoption decisions become tools of power and politics in organizations. The author comments in such scenarios that there is a risk of adopting and institutionalizing a "poor" information system. The author laments that in such cases owner satisfaction gets priority over user satisfaction (Silva, 1997).

Davis (1989, 1993) himself has alluded to "physically using the system" to define the user. He relates the construct perceived usefulness to the actual users: "perceived usefulness concerns the expected overall impact of system use on job performance (process and outcome), whereas ease of use pertains only to those performance impacts related to the process of using the system per se" (Davis, 1993, p. 477). He further elaborates on the ease of use: "given that some fraction of a user's total job content is devoted to physically using the system per se, if the user becomes more

productive in that fraction of his her job via greater ease of use, then he or she should become more productive overall” (Davis, 1993, p. 477). Hence, we assert that our plan to use actual Hadoop users of organizations as the subject of this research instead of company chief executive or chief technology officers is consistent with the vision of Davis’ technology acceptance model. Davis reports in his paper (Davis, 1993) that he used 112 professional and managerial employees of a large North American company as subjects of his survey – not CEO’s or CTO’s. Davis’ original model was developed under the assumption that the system is available for voluntary use by employees as opposed to management’s strictures (Davis, 1993).

By taking this into consideration, the author designs his research such that big data technology acceptance decision needs to come from big data technology (e.g., Hadoop) users. The author conducts a survey on Hadoop users. Several Hadoop-user groups have been included in the sampling frame. The conceptual definition of technology acceptance for this study is the extent to which a decision-maker is a hands-on person, that is the actual user of that technology.

1.2 Big Data

Big data is large and complex, and it cannot be stored in conventional data storage/database systems. Caesarius and Hohenthal (2018) posit that the novelty of big data is distinct in terms of its complexity and data structures. Big data has emerged during the last decade. Before the emergence of big data, we used to deal with transactional data that are structured and hence could be stored in conventional relational database

systems (Rahman & Sutton, 2016). The relational database system has been on the market since the early 70s after Dr. Codd gave a model for relational databases based on the mathematical set theory (Codd, 1970). With the advent of new technologies, the internet, advancement in software and hardware engineering, social network tools, and automation, the data volume has increased significantly. For example, as of 2012, Walmart used the technology to create and collect several petabytes of transactional data every hour from its customers (McAfee & Brynjolfsson, 2012).

Most of the internet and social media data are unstructured (Baesens et al., 2016; Das & Kumar, 2013; Rahman & Rutz, 2015). Data has been growing in all sectors. For example, the U.S. government mandated that in healthcare all patient records need to be stored digitally. In healthcare, big data management requirements in terms of personal data, sensitive data, genomic sequencing data, payor records, wearable devices data, complex and heterogeneous data are called out from big data technological capability perspectives (Viceconti et al., 2015). A large volume of healthcare data related to chronic diseases of 140 million patients in the United States require management and processing as well as for analytics (Bardhan et al., 2020). There is also support for open data by government agencies (Jetzek et al., 2019). With the rapid growth of digital publishing data, managing and analyzing the data have become a challenge (Xia et al., 2017). Data storage cost has also been decreasing gradually. As a result, organizations find it worthwhile to store and process big data to

find business opportunities in them. Early users of big data include Google, Yahoo, Facebook, and Amazon to name a few.

1.3 Characteristics of Big Data

Big data has five characteristics compared to conventional data – 5 V's. These include Volume, Velocity, Variety, Veracity, and Value (Baesens et al., 2016; Xia et al., 2017; Marr, 2015). Big data volume is meant for hundreds of terabytes to petabytes of data and when projected data growth at a particular time is much higher than conventional transactional data growth (Abbasi et al., 2016). Associated factor: scalability, big data streaming happens very fast or near real-time for which receiving tools and storage systems need to be very efficient to handle that (Velocity). The speed of data creation is one of the key characteristics of big data (Abbasi et al., 2016). Big data consists of sensor data, mobile phone data, social media data (unstructured), video streaming, and pictures (variety) to name a few. With big data in the picture, organizations are now dealing with structured, semi-structured, and unstructured data. Big data is unstructured and because of that it is challenging to compare data in origin and target (veracity). Since there is a variety of big data sources, credibility, and reliability of this data vary. Hence, dealing with veracity characteristics of big data is a challenge (Abbasi et al., 2016). The existing literature suggests the text analysis using supervised learning is commonly used to assess big data veracity (Lozano et al., 2020).

Big data is a huge volume (low value) and businesses want to find business value (high value) in them by using sophisticated tools and technologies. Big data include both

structured and unstructured data but mostly unstructured (Baesens et al., 2016; Rahman and Aldhaban, 2015). The value characteristic of big data is associated with business value in terms of decisions and actions. Researchers have attempted to view the value creation of big data from a variety of perspectives. Dong et al. (2020) conduct an empirical study on big data analytics which suggests that social media diversity and big data analytics have a positive influence on business value creation and improving the market performance. Lycett (2013) coined the idea of big data value creation and delivery using the concept of datafication in terms of dematerialization (identify information aspect), liquidity (manipulation and dissemination), and density (a combination of resources). Mesgari and Okoli (2019) propose IT materiality, discovery aspects, and action orientation in value creation and the sense-making of new IT. Mikalef et al. (2020) propose tangible (data and technology), intangible (data-driven culture and organizational learning), and human skills (technical and managerial skills) to develop big data analytics capability to maintain competitive performance. Abbasi et al. (2016) suggest assessing the value of big data IT artifacts.

Table 1: Big Data Characteristics – 5 V's

Characteristics	Description	Influencer
Volume	A few terabytes to hundreds of terabytes to petabytes of data need to be captured, processed, stored, and analyzed	Data volume keeps growing in source
Velocity	Given the volume the data need to be captured, processed, and displayed faster for right time business intelligence and decision making	Increase in data sources. Improved computing, processing, BI & Visualization technologies
Variety	Includes a variety of data sources with unstructured, semi-structured, and structured	Sensors, social media sites, digital pictures, video,

	data. More than 90% unstructured (Das & Kumar, 2013)	transaction records, and communication surveillance
Veracity	The quality and provenance of received data. As in most cases data is not structured data consistency is an issue	Data-based decisions require traceability and justification
Value	Provides greater insights generating new business value	Corporate business value

The five V's of big data have some similarity and/or connection with the 12 factors selected as part of the current research model. The 12 factors include scalability, data storage and processing, cost-effectiveness, performance expectancy, security and privacy, reliability, data analytics capability, training and required skills, flexibility, output quality, functionality, and facilitating conditions. Abbasi et al. (2016) emphasize investigating adoption and adaptation of big data techniques and technologies. The scalability factor points to the volume characteristics of big data (Garcia-Gil et al., 2017; Menon & Sarkar, 2016). To handle a large volume of data big data technology Hadoop is considered scalable. The data storage and processing capability factor refers to the volume and velocity characteristics of big data. The flexibility factor relates to velocity characteristics as big data technology is capable to handle small set to large set data, and batch files to streaming data. This factor is also associated with the variety characteristics of big data. Big data technology is capable to handle both structured and unstructured data. The data analytics capability factor is associated with the velocity characteristics (Chardonens et al., 2013). Big data technology is capable to process and display both streaming and static set of data. It has the capability to visualize data in real-time (Berengueres & Efimov, 2014; Garzo et al., 2013; Kranjc et al., 2013). The use

case includes fraud detection (Bologa et al., 2010). The output quality factor refers to veracity characteristics. Big data comes from different external sources and is unstructured, hence data quality of received data is critical (Baesens et al., 2016). This research investigates if data quality provided by big data is a matter of concern in accepting this technology.

The performance expectancy factor is connected with velocity characteristics. Big data technologies are thought to be capable to perform reasonably with a huge volume of data set. The reliability factor relates to big data volume and velocity characteristics. Big data Hadoop is considered to be reliable in retaining data intact, meaning that there is no data loss due to node failure. For example, the HDFS component of Hadoop retains multiple copies of the same data in different nodes. The security and privacy factor relates to the veracity characteristics of big data. There is a concern about the privacy of big data (Richards & King, 2014; Tene & Polonetsky, 2013; Wu et al., 2017). Abbasi et al. (2016) suggest taking privacy and security concerns as a research agenda of big data and behavioral research. The security and privacy factor is a part of this research to understand if this factor has a positive or negative impact on big data technology adoption. The training and skill factor is associated with the variety and other big data characteristics. The unstructured (90%) nature of big data makes it different from conventional transactional data owned by companies (Das & Kumar, 2013).

The distinct, unstructured characteristic of big data causes the use of a new set of big data tools for data receiving, storing, processing, and visualizing. The functionality factor is associated with the volume and velocity of big data characteristics. This refers to Hadoop's capability to receive, store, process, and display data. The facilitating condition factor is not directly associated with big data characteristics, but it speaks for using this technology with some vendor or internal IT infrastructure support (a mediating factor). This study investigates if Hadoop system usage is influenced by this factor. Last but not the least, the cost-effectiveness factor is associated with the value characteristics of big data. This also relates to the initial cost as well as any licensing cost. This particular factor of the model will be assessed to understand this technology from cost perspectives to a business value perspective (Kohli et al., 2012).

1.4 Big Data Technology and Evolution

The extant literature suggests that over the past three decades the information technology field has shown the biggest technological advances (Krugman & Wells, 2017). Big technology Hadoop is one of them. To handle big data, a completely new set of tools and technologies have been emerging since the last decade (Cloudera, 2012; Landset et al., 2015; Rahman et al., 2014). Apache Hadoop is a prominent software framework in the big data world. The evolution of Hadoop is now spanning over 10 years. The seeds of Hadoop were planted back in 2002 by two creative thinkers: Doug Cutting (then-Internet Archive director) and Mike Cafarella (a University of Washington graduate student). Their project name was Nutch which was originally aimed to develop

a state-of-the-art open-source search engine based on Internet archives with the capability to crawl and index millions of pages (Harris, 2013). The project was able to crawl and index hundreds of millions of pages. But to work on billions of pages, a more robust architecture and scalability were needed. And right after their first working version, Google published papers on the Google File System in October 2003 and the MapReduce in December 2004 which helped to build Nutch (Harris, 2013). In a few months, Cutting and Cafarella came up with the underlying file systems and processing framework that eventually became Hadoop (Harris, 2013). In 2006, Cutting went to work with Yahoo to build Hadoop as part of an open-source Apache Software Foundation project by spanning out the storage and processing parts of Nutch along with Google's work on MapReduce (Dolev et al., 2019; Harris, 2013).

Yahoo made a significant contribution to building Hadoop. As of 2011, Yahoo and Hortonworks (spun off from Yahoo) had "contributed more than 80% of the lines of code in Apache Hadoop trunk" (Brockmeier, 2011). There are other contributors to Hadoop in terms of lines of code such as Cloudera, Facebook, LinkedIn, eBay, IBM, Apple, Twitter, and Amazon (Brockmeier, 2011). Cloudera (a Hadoop vendor) was launched in 2008. In 2009, IBM and Greenplum started using Hadoop. In 2010, MapR (another Hadoop vendor acquired by Hewlett Packard Enterprises as of 2019) and Microsoft® Azure started using Hadoop. Hadoop is designated, particularly for large-scale, on-premise deployments.

There are several prominent companies that built platforms and applications on top of the Hadoop Distributed File System (HDFS). Google presented the concept of the big table (for big data); Yahoo contributed to SQL-like infrastructure, Hive; Amazon introduced web services – AWS and Redshift; Microsoft launched big data landscape, Azure; and IBM provided Watson research on big data analytics. Big data potentials include real-time data ingestion, storing, transforming, processing, and new opportunity of business intelligence with big data (Li et al., 2020; Schlesinger & Rahman, 2015). There are some other file systems developed including Lustre and General Parallel File System (GPFS) by IBM. But they do not scale as high as HDFS. GridGrain offers a substitute architecture which is an in-memory based data grid, but it can handle much fewer data compared to HDFS (Monteith et al., 2013).

By the year 2020, a few cloud-based big data platforms (public clouds) have evolved along with their own storage systems as an alternative to HDFS: Microsoft Azure, Google Cloud, and Amazon Elastic MapReduce, to name a few. These are economical, pre-built distributed computing services. The Microsoft Azure related data storage and processing tools include Azure Data Explore, Cosmos DB, Azure Data Lake, Azure HDInsight, and Azure Stream Analytics. Google Cloud Platform has come up with data storage called GCS (Google Cloud Storage), Dataproc, BigQuery, and Cloud SQL. The Amazon Elastic MapReduce (EMR) has its storage system, Amazon S3 (stands for Simple Storage Service) along with other tools and technologies including Apache Spark, Apache Hive, and Apache HBase.

Besides Hadoop's two main components HDFS and MapReduce/Spark, the big data ecosystem consists of a handful of tools and technologies. This section provides a brief overview of some of them. *MapReduce* is one of the two main components of Hadoop. It is a software component that processes data at node-level and provides aggregated data via Map results in terms of the answer to queries. MapReduce suffers from performance. It is good for batch processing. As a substitute for MapReduce, a new software, *Spark*, was developed by UC Berkley which is considered a new generation software and addresses the performance issues. There are several other tools and technologies that are part of the Hadoop platform ecosystem. They include HBase, Hive, Pig, Mahout, MLlib, Flume, and Sqoop. The *HBase* is a non-relational database system that sits on top of the Hadoop file system (HDFS). It allows for quick retrieval of rows based on keys. It also provides the capability to conduct inserts, updates, and deletes. But relational joins cannot be done to pull data from multiple tables the way it is done in traditional database systems.

Hive is a tool that accepts queries (SQL) and converts it to MapReduce or Spark jobs to connect to HDFS and retrieve data in a structured format. This tool is used as an alternative to traditional ETL tasks. *Pig* is a scripting language used to write MapReduce transformations to manipulate data in HDFS. Mahout is a data mining library that runs against HDFS through MapReduce jobs. *MLlib* is a new generation of machine learning libraries based on Spark programs as an alternative to Mahout which uses MapReduce. Flume is a framework used to extract data from external sources and load into Hadoop.

Flume is capable to handle the streaming of data flows and insert into Hadoop. *Sqoop* is another tool that helps to extract data from external sources, mainly relational databases (Teradata, Oracle, SQL Server, etc.) into Hadoop (Rahman, 2016). Companies take advantage of Hadoop by storing huge volumes of historical data (expensive to maintain in relational databases) into Hadoop.

1.5 An Overview of Two Hadoop-Based Application Systems

This section provides an overview of big data Hadoop applications. This researcher was part of the application development team. Figure 1 shows an end-to-end data flow – source (input) and reporting (output). This application was built based on Cloudera Hadoop Distribution and other big data tools (Pig, Sqoop, Hive, and Impala). The goal was to architect a high-performance extract, transform, and load (ETL) platform that supports data visualization and exploration. This application was built for a large company to understand the impact of email on employee productivity. One of the goals was to determine whether the use of alternative collaboration tools would be more effective for teamwork and communication.

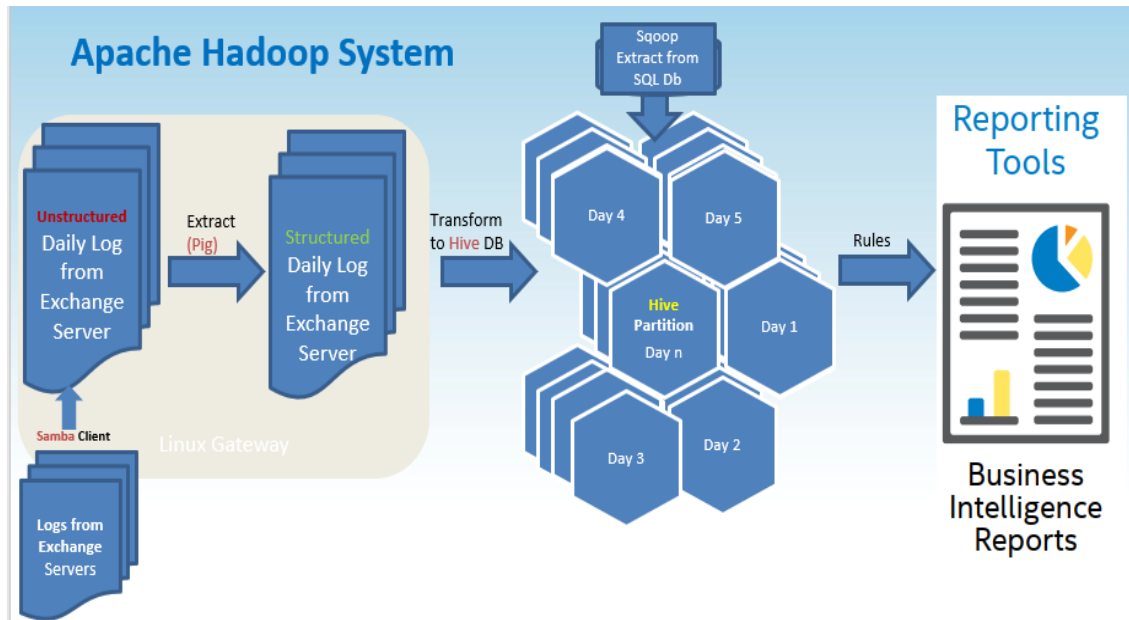


Figure 1: Hadoop and Reporting Application

The left side of Figure 1 (derived from Chowdhury et al., 2015) shows data source, email servers. Unstructured data is pulled using Pig (extract tool) and landed in a staging area of the Hadoop system. Then further processing and transformation are done to prepare data in a structured format. Approximately four billion rows worth four months of data are stored in Hadoop. After required formatting data is stored in Hive table format that resides in the Hadoop Distribution File System (HDFS). There is another source of data that comes from the traditional database system. This data is extracted by using Sqoop and loaded into the Hadoop Storage System. By combining these data, a reporting layer is built into the Hadoop System. A reporting environment is created using Impala which retrieves data from Hadoop and displays via business intelligence reports.

This application achieved several goals: store data in a highly scalable platform (Hadoop). A fault-tolerant tool, Hive was used to store transformed data in Hadoop. A high-performance tool, Impala was used for reporting purposes. Impala is considered Hadoop's high-performance engine which allows for massively parallel processing of queries.

1.6 Big Data Market

The industry research firm, IDC (2019), forecasts that revenues for big data and business analytics are expected to reach \$189.1 billion during the year 2019. The report also forecasts a double-digit per-year growth through 2022. Another research firm, Technavio (2020) provides its latest market research by stating that the big data market is projected to grow by \$142.5 billion during 2020-2024. The report observes that North America had the largest big data market share in 2019. And the report also mentions that the region is expected to offer many growth opportunities to market vendors during the same period of time. It reports that 47% of the market's growth is expected to appear in the North American market during the forecast period Technavio (2020). These latest industry market research reports suggest that the United States is one of the critical markets for big data for the next few years. One of the important sources of economic growth is progress in technology. Technology provides the technological means for other companies to increase the productivity of goods and services (Krugman & Wells, 2017).

1.7 Research Objectives

The purpose of this study is to conduct empirical research to advance knowledge in the field of technology acceptance. We investigate the factors that influence the acceptance of big data technology by companies. This study conducts research among companies in the United States that use big data. Most of the research done in technology acceptance is in the area of personal use (e.g., smartphone). This study consists of technology acceptance by a company through the users of that company. A handful of variables/factors are evaluated by previous research using Davis' (1989) Technology Acceptance Model (TAM).

TAM by Davis (1993) is considered parsimonious and it reportedly has a wealth of empirical supports (Lee et al., 2003). Additionally, TAM posits that technology acceptance is determined by two factors: perceived usefulness (PU) and perceived ease of use (PEOU), which determine behavioral intention (BI), and actual use (AU). Previous research identified PU more effective in technology acceptance. But experts in this field question what makes technology useful (Benbasat & Barki, 2007; Lee et al., 2003). They comment that previous research used PU and PEOU as a black box – that is without giving any specific definition of PU.

"While we do not doubt that Davis et al.'s (1989) original intention was that the influence of system and other characteristics be studied through TAM's constructs, study after study has reiterated the importance of PU, with very little research effort going into investigating what actually makes a system useful. In other words, PU and PEOU have largely been treated as black boxes that very few have tried to pry open." (Benbasat & Barki, 2007, p. 212).

This research makes an effort to define PU in terms of utility theory (Bentham, 1824; Read, 2004; Stigler, 1950) and other relevant information systems (IS) theories. A research model is proposed to determine the factors influencing big data technology acceptance.

1.8 Research Approach

This dissertation consists of several key steps including qualitative and quantitative studies to conduct research on big data technology acceptance. The dissertation provides an overview of big data characteristics (5 Vs) and big data technologies. It provides the importance of studying technology acceptance in general and big data technology acceptance in particular.

This research highlights previous research done on technology acceptance. An overview of extant literature about prominent information systems (IS) theories about technology acceptance was provided. The research provides an update on research done on big data technology and acceptance. It also has taken into consideration the research done on traditional data management software acceptance. The research points out the methodologies used in existing research. In this regard, the research gap in technology acceptance and big data technology acceptance have been identified.

The research model is developed using a methodical approach. First, this study collects most of the variables from existing IT theory (Davis, 1993; Rogers, 2003; Venkatesh et al., 2003), utility theory of economics (Kapteyn, 1985; Stigler, 1950), adoption factors taxonomy based on prior research, industry technical papers, and

other documentation. Through this method, 32 factors have been identified. Later these factors were presented to industry experts who have hands-on experience in both big data technologies (e.g., Hadoop) and traditional data management software including Teradata, Oracle, MS SQL server (Rahman, 2013, 2016). The qualitative studies consisting of the brainstorming sessions, expert panel, focus groups, and interviews were used to get the input in selecting the most important variables of big data technology adoption. Out of 32 factors, the top 12 factors (by voting) are selected to be part of this study. Thus, this research model consists of 12 factors that are used to understand big data technology adoption. More than 60 construct-items are developed using these variables and are finally used in the survey instrument.

Hypotheses have been developed based on 12 factors identified by the qualitative study results. The survey instrument is developed based on the questionnaire used in the existing literature and on new questions added based on big data specific factors. The survey instrument is tested and validated. A web-based survey was developed and sent to big data user groups in the United States. Out of 14 big data user groups (available on the Internet) consisting of 33 thousand subscribers, two Hadoop user groups were sent survey questions. A cluster sampling technique is used by randomly selecting these two user groups. Collected data are analyzed using the statistical software, AMOS. Conclusions are drawn relating to theoretical contribution and practical implications.

1.9 Statement of Problem

Companies have a large volume of enterprise data. There are data (big data) available from external sources (e.g., social media) that could be used by organizations to draw insights, develop products and services, and increase revenue. Both academic and industry papers suggest that organizations are not sure about the prospect of big data projects (Gartner, 2015). An industry survey conducted in 2019, to understand the state of big data and artificial intelligence (AI), indicates that a large majority (73.3%) of organizations identify business adoption of big data and AI initiatives as a challenge (Bean, 2020). The same survey report reveals that 73.2% of the firms have not been able to forge a data culture within the organization. As many as 62.2% of the firms have not been able to create a data-driven organization. As many as 54.9% of the firms are not competing on data and analytics. Half of the firms are not able to identify data as a business asset (Bean, 2020). Researchers suggest that for making organizations data-driven the leadership needs to foster an organization's agility (Holst, 2020).

Industry experts suggest that there are practical obstacles in implementing big data projects (Moktadir et al., 2019; Rahman & Aldhaban, 2015). Chen et al. (2020) report that in healthcare big data management, technology adoption barriers are closely related to skillsets, resource allocation, operational complexity, patient protection laws, and other regulations. The IT leadership, management, knowledge workers, and data architects need to agree on creating a data-driven organization. Since big data uses a completely new set of tools and technologies, an IT department's preparedness,

developers and knowledge workers' required training and skill set is very important (McAfee & Brynjolfsson, 2012). But there is little information as to what factors affect the acceptance of big data technology. Caesarius and Hohenthal (2018) assert that companies might be less inclined to adopt big data technologies particularly if the value in return is unknown. We also know that there is a strong connection between IT capability and firm performance (Chae et al., 2018). There is a need to understand the factors that present significant challenges in adopting big data technologies. Understanding the key factors that affect an organization's use of big data may provide useful information that could allow business executives to implement big data projects and thus increase the business value of big data.

1.10 Research Questions

Based on the background of this study and the research problem, we need to understand the importance of the factors that influence big data technology acceptance. The key research question to understand from this study is:

What factors influence the acceptance of big data technology – Hadoop? What technological capabilities make technology useful?

To get the answer to the above research question this study develops a big data technology acceptance model. Data are collected and model is tested based on survey data from the big data user community in the United States. The findings of this study are expected to help IT managers and company executives to make the decision of adopting big data technologies.

This study is expected to help understand the challenges and/or barriers in adopting big data technology (Moktadir et al., 2019). The study is expected to provide insights as to what actually makes big data product or technology useful to the users. According to TAM, the perceived usefulness (PU) is considered the driving factor. This research attempts to elaborate on a practical definition of PU. We need to understand, what specific features of a complex technology are the determinants of its acceptance. The literature studies reveal that there is little research conducted to explore independent variables from the technological capability standpoint when it comes to IS research related to technology acceptance (e.g., Petter et al., 2013; Surbakti et al., 2020). Our research delves into identifying factors from that perspective. This study is expected to provide insight as to how the user's experience of big data tools and technologies can be improved. This study is also expected to provide information on whether some new factors such as scalability, data storage and processing capability and flexibility have an impact on the perceived usefulness of TAM. The latest studies suggest that the firms that use the highest organizational information technology capability can improve market value by about 45% to 76% (Saunders, 2016). Besides technological factors, this study is expected to provide insight as to how organizational and environmental factors influence big data acceptance, especially in industrial/organizational level acceptance context.

1.11 Significance of Studying Big Data Technology Acceptance

Big data is in its early stage of use by many organizations (Russom, 2013). It is important to investigate the user perception of big data technologies. The extant literature calls for investigating the adoption of big data techniques and technologies (Abbasi et al., 2016). This research is expected to make a contribution to theory and enhancements to existing knowledge. Traditional data management software that holds transactional data, has been in the market for the last 5 decades. With the emergence of the Internet, sensors, social media data is no longer just an organization's transactional data. Big data is mostly non-transactional or unstructured data. Big data has 5 distinct characteristics – volume, velocity, variety, veracity, and value. To handle big data a distinct set of new tools and technologies have emerged. They are different from traditional data management tools and technologies. So, it is important to understand how users perceive these new technologies.

In technology acceptance research, most of the research was done in terms of individual product user's acceptance. Most of the surveys in those studies were conducted on undergraduate and graduate students as subjects. This research investigates technology acceptance by users of organizations. Surveys are conducted on knowledge workers of those organizations as opposed to student groups who are not actual users. Previous research on technology acceptance used TAM which consists of PU and PEOU. Perceived usefulness (PU) needs to be understood based on some clear

definitions guided by IS and economics theories. We hope that will provide new insights on technology acceptance.

Chapter 2 Literature Review

This chapter reviews the existing models, theories, and variables related to technology acceptance used by them. This chapter also provides an account of variables used in different surveys, and experiments conducted, as well as prominent research published in peer-reviewed academic journals and conferences proceedings. It also reviews the industry technical papers, Gartner's papers, software documentation related to big data technology (e.g., Apache Foundation site), and the sites of the Hadoop platform vendors such as Cloudera, Hortonworks, and MapR. The goal was to identify the variables and come up with a list of variables that could be used in a qualitative study. In this process, variables are adopted from existing technology adoption models, theories, survey-based research papers, and industry technical white papers. A list of 32 variables is identified which are presented to qualitative study participants in brainstorming, focus group, and individual interview sessions. The qualitative study provides a selective list of 12 variables that are used as independent variables (IV) in the proposed research model.

2.1 Relevant Theories Used to Study the Adoption and Use of IS

Over the last few decades, scholars have introduced several theoretical models (Table 2) to predict and understand the acceptance of new technology at both the individual level (e.g., smartphone) and the organizational level (e.g., data warehousing technology).

User acceptance is “the demonstrable willingness within users’ group to employ

information technology for the tasks it is designed to support” (Dillon & Morris, 1996, p. 3).

Table 2: Relevant Theories to Study Adoption of Information Technology

Theory/Model	Exogenous Variables	Discipline	Introduced By
Theory of Reasoned Action (TRA)	Attitude Toward Behavior, Subjective Norm	Social Psychology	Fishbein & Ajzen (1975)
Diffusion of Innovation (DOI)	Relative Advantage, Compatibility, Complexity, Trialability, Observability	Communication Studies	Rogers (1983)
Technology Acceptance Model (TAM)	Perceived Usefulness, Perceived Ease of Use	Information Systems	Davis (1989)
Technology, Organization and Environment (TOE) Framework	Technological, Organizational, Environmental	Information Systems	Tornatzky & Fleischer (1990)
Resource-Based View (RBV)	Value, Rareness, Imitability, and Substitutability	Competitive Strategy	Barney (1991)
Theory of Planned Behavior (TPB)	Attitude Toward Behavior, Subjective Norm, Perceived Behavioral Control	Social Psychology	Ajzen (1991)
Unified Theory of Acceptance and Use of Technology (UTAUT)	Performance Expectancy, Effort Expectancy, Social Influence, Facilitating Conditions.	Information Systems	Venkatesh et al. (2003)

2.1.1 Theory of Reasoned Action

The theory of reasoned action (TRA) model was introduced by Fishbein and Ajzen (1975). The TRA consists of two factors, ‘attitude toward behavior’ and ‘subjective norm’ to explain users’ behavioral intention followed by actual behavior to use new technology. This model was widely used in information technology (IT) and other fields (Bagozzi, 1982; Davis et al., 1989; Hartwick & Barki 1994; Mathieson, 1991; Moore & Benbasat, 1996; Sheppard et al., 1988; Venkatesh et al., 2003).

Davis et al. (1989) used this model to predict the adoption of MS Windows and word processing software. Liker and Sindi (1997) employed this model to understand

the adoption of computer-based information systems in the general expert systems in particular. The authors find that intention to use was influenced by subjective norm (i.e., social influence) which encourages to use of the new technology. Karahanna et al. (1999) conduct a cross-sectional comparison between pre-adoption and post-adoption beliefs in technology acceptance. The authors find that pre-adoption behavior is based on perceived usefulness, perceived ease of use, and results-demonstrability while post-adoption is dependent on some instrumental beliefs of usefulness and image perceptions.

Thus, we attempt to investigate the influence of the Hadoop system use by virtue of intention to use. Research suggests that a system might be underutilized or not utilized if the user's psychological reactions are ignored. In this research, the intention to use is taken as one of the constructs of the actual model. The intention is defined as to whether the user will or will not take action to use the system (i.e., Hadoop). Davis' TAM borrowed the construct, 'intention' from TRA. Since this research will use TAM as the primary model, the intention is considered part of the actual model.

2.1.2 Theory of Planned Behavior

Ajzen (1991) has developed the theory of planned behavior (TPB) which has its root in social psychology. The TPB proposes three factors that include 'attitude toward behavior', 'subjective norm', and 'perceived behavioral control'. The TPB model originates from the TRA model and it includes one additional construct, 'perceived behavioral control', to better predict behavioral intention (Cheung et al., 2000; Taylor &

Todd, 1995). Perceived behavioral control speaks for how easy or difficult it is for a person to perform a certain behavior or interest. With that, TPB states that a person's behavioral outcome depends on intention which in turn is influenced by attitude, subjective norm, and perceived behavioral control. On the other hand, the behavior is also determined by perceived behavioral control. Since TPB deals with an individual's behavioral intention it is widely used in social psychology (Rhodes & Courneya, 2003). In IT, TPB's effectiveness toward acceptance of innovation has been investigated by several studies (George, 2004; Mathieson, 1991; Pavlov & Chai, 2002).

2.1.3 Diffusion of Innovation

Rogers (1983) developed and introduced the diffusion of innovation (DOI) model which posits five factors including relative advantage, compatibility, complexity, trialability, and observability. Innovation is deemed to have a relative advantage if it is "technically superior in terms of cost and functionality than the technology it supersedes" (Fichman & Kemerer, 1993, p. 10). Fichman and Kemerer (1993) assert that innovation needs to be compatible "with existing values, skills, and work practices of potential adopters." Regarding complexity, Fichman and Kemerer's (1993, p. 10) general observation is that "innovation is relatively difficult to understand and use." Big data is very large and complex in terms of its characteristics (volume, velocity, variety, veracity, and value). But it is understood given the complexity of big data characteristics. Hence, the users might favor the acceptance of big data technologies. Trialability is related to the risk of no benefit or value. Fichman and Kemerer (1993, p. 9) state that "adopters look

unfavorably on innovations that are difficult to put through a trial period or whose benefits are difficult to see or describe. These characteristics increase the uncertainty about the innovation's true value." In regard to observability, "the results and benefits of the innovation's use can be easily observed and communicated to others" (Fichman & Kemerer, 1993, p. 10).

A large number of past empirical studies have proven this model's effectiveness (Moore & Benbasat, 1996; Teo & Ranganathan, 2004; Wu & Chiu, 2015). Tan and Teo (2000) use relative advantage, compatibility, complexity, and trialability to understand an individual account holder's adoption of online banking. Moore and Benbasat (1996) apply DOI attributes, relative advantage, compatibility, trialability, and observability to understand the adoption of IT by end-users. The DOI is primarily focused on the individual-level rate of adoption as compared to the adoption process from an organizational context (Hameed et al., 2012).

Big data technology capability conforms to technology diffusion attributes such as relative advantage and trialability. In regard to relative advantage, big data technologies are open-source, and technologies are cheaper to store and process complex and large volumes of data. An innovation that has a relative advantage provides economic and organizational political legitimacy in making adoption decisions (Ramamurthy et al., 2008). From a trialability standpoint, big data technologies have positive points. There are quite a few big data tools and technologies (big data ecosystem) that have appeared during the last decade to receive, process, store, and

analyze big data. The most important achievement is that a handful of open source technologies are provided by the Apache Software Foundation that allows any organization to start big data projects (Rahman & Aldhaban, 2015). Thus, big data allows for trialability to understand the benefits of it.

2.1.4 Technology Acceptance Model

Fred Davis (1989) introduced the technology acceptance model (TAM) which is rooted in TRA (Dishaw, 1998). Later, Venkatesh and Davis (2000) developed a revised version called TAM2. Legris et al. (2003) report that overall, the two (TAM and TAM2) can explain about 40% of the system's use. The TAM consists of two constructs, 'perceived usefulness' (PU) and 'perceived ease of use' (PEOU) which are influenced by independent variables that in turn determine the latent variable, 'behavioral intention to use'. The 'intention to use' in TAM overlaps with TRA and TPB. The perceived usefulness and perceived ease of use replace 'attitudes' and 'subjective norms' used in TRA. On the other hand, those two TAM factors (PU & PEOU) replace the effect of attitude, subjective norm, and perceived behavioral control under TPB (Bagozzi, 2007). Davis et al. (1989) and Venkatesh et al. (2003) studies proved that TAM outperforms TRA and TPB in terms of explaining variances. However, in their paper on TAM titled, 'Reexamining perceived ease of use and usefulness', Segars and Grover (1993) comment that "no absolute measures for these constructs exist across varying technological and organizational contexts." The authors observe that task and user characteristics change the nature and importance of perceptions that explain technology use. We assert that

besides task and user characteristics, it is important to independently evaluate technology in terms of its usefulness and core capabilities.

The TAM is considered the most influential and widely used model, especially in the information systems (IS) field (Venkatesh et al., 2007). Bagozzi (2007) identifies parsimony as the main strength of TAM. Several TAM studies in IS research are listed in Table 3. Note, most of these are light technologies and/or applications. This research attempts to extend the TAM to more complex adoption scenarios such as acceptance of the complex platform/ infrastructure, Hadoop by its intended users. One study (Hood-Clark, 2016) has investigated TAM using big data as the application. It finds all core constructs of TAM valid. However, this study has not used big data-related independent variables. What makes big data technology useful? What technological capabilities make big data technology useful? Therefore, in addition to employing TAM's core constructs, antecedents specific to the big data technology and technological capabilities are sought by our study.

Table 3: Summary of TAM Studies (1989-2019)

Authors	Constructs	Applications	Methodology
Davis (1989)	Perceived Usefulness (PU), Perceived Ease of Use (PEOU), Usage (U)	XEDIT	Survey
Davis et al. (1989)	PU, PEOU, Attitude (A), Behavioral Intention (BI), U	Write One	Experiment
Basoglu et al. (2007)	PU, PEOU, U	ERP	Survey
Mathieson (1991)	PU, PEOU, A, BI, U	Spreadsheet	Experiment
Adams et al. (1992)	PU, PEOU, U	E-mail, WordPerfect	Survey
Straub et al. (1995)	PU, PEOU, U	V-mail	Survey
Igbaria et al. (1995)	PU, PEOU, U	Micro-Computer	Survey
Szajna (1996)	PU, PEOU, BI, U	E-mail	Experiment

Hendrickson & Collins (1996)	PU, PEOU, U	1-2-3, WordPerfect	Experiment
Morris & Dillon (1997)	PU, PEOU, A, BI, U	Netscape	Survey
Gefen & Straub (1997)	PU, PEOU, U	E-mail	Survey
Lederer et al. (2000)	PU, PEOU, A, BI, U	World wide web	Survey
Qin et al. (2011)	PU, PEOU, BI	Online Social Networks	Survey
Choi and Ji (2015)	PU, PEOU, BI	Autonomous Vehicle	Survey
Rajan & Baral (2015)	PU, PEOU, BI, U	ERP	Survey
Wang et al. (2012)	PU, PEOU, U	Instant Messaging	Survey
Hood-Clark (2016)	PU, PEOU, A, BI, U	Big Data	Survey

One key aspect of TAM is that it provides a framework to examine the influence of external factors on the usage of a system. Several external factors have been applied to TAM factors. For the construct, perceived usefulness (PU) these external variables have been used: job relevance; result demonstrability; image; complexity; managerial support; social presence; attitude; anxiety; accessibility; perceived enjoyment; facilitating conditions; self-efficacy; end user support (Lee et al., 2003). For the construct, perceived ease of use (PEOU) these external variables have been tested: attitude; anxiety; accessibility; usability; playfulness; perceived enjoyment; facilitating conditions; self-efficacy; social influence (i.e., subjective norm, social pressure) and managerial support (Lee et al., 2003).

Turner et al. (2010) conduct a systematic literature review of 79 empirical studies in 73 articles that published results of empirical studies that used TAM. The authors find that BI is correlated with actual usage. The authors also report that PU and PEOU constructs are not as good at predicting actual technology use as BI.

Scholars of TAM study point out that TAM's two key constructs (perceived usefulness and perceived ease of use) have been used in so many studies including the information technology acceptance field without first defining what makes a system useful (Benbasat & Barki, 2007). Current research makes an attempt to come up with a definition of 'usefulness'. That helps in the qualitative study process in identifying external factors that point to perceived usefulness. Straub and Burton-Jones (2007) observe that only a few studies are conducted on actual system use. Hence, we add this construct to our research model.

Hood-Clark's (2016) research on big data usage using original TAM constructs identify relationship independent variables (perceived usefulness, perceived ease of use, and attitude toward use) and dependent variables (behavioral intention to use, and actual use). This research has not used any big data-specific external variables. That means the author limits its research within TAM core constructs. This type of study attempts to test the validity of the model. Prior literature also conducts such studies (Davis et. al., 1989; Lederer et al., 2000; Mathieson, 1991; Taylor & Todd, 1995) which helps TAM to be one of the mainstream technology acceptance models.

2.1.5 Technology, Organization and Environment

Tornatzky and Fleischer (1990) introduced the technology, organization, and environment (TOE) Framework. This framework has also been widely used (Chau & Tam, 1997; Kuan & Chau, 2001; Malaka & Brown, 2015; Zhu & Kraemer, 2005). This model proposes factors from aspects of technological, organizational, and environmental. It

has been reportedly used to explain organization level technology adoption behavior. Chau and Tam (1997) suggest that innovation adoption needs to be studied from the context of variables that pertain to technological characteristics. In their research, the authors used technology variables such as the complexity of IT infrastructure and formalization on system development and management (Chau and Tam, 1997).

Malaka and Brown (2015) study the organizational adoption of big data by employing TOE. The authors use variables such as data integration, veracity, and performance and scalability from a big data characteristics perspective. This research takes TOE factors into consideration for big data technology acceptance as part of the qualitative study. Possible variables include scalability, data storage, processing capability, data mining capability (technological factors), training and skill of big data users (organizational factor) and facilitating conditions (environmental factor).

2.1.6 Resource Based View

Barney (1991) proposes resource-based view (RBV) of the firm which consists of variables, value, rareness, imitability, and substitutability to achieve competitiveness by a firm. The resource-based view posits that firms should be capable to produce resources (Wernerfelt, 1984). Here “resources mean strengths or assets of the firm that may be tangible (e.g., financial assets, technology) or intangible (e.g., reputation, managerial skills)” (Eisenhardt & Schoonhoven, 1996). We posit that from big data capability standpoint companies can develop three key resources including big data

technology capabilities, technical skillsets associated with big data, and data scientist and analytics expertise (Lee, 2017).

2.1.7 Unified Theory of Acceptance and Use of Technology

Venkatesh et al. (2003) propose a modified and enhanced model called unified theory of acceptance and use of technology (UTAUT). This model consolidates other models including that of TAM. The authors claim this model to be a parsimonious model. The UTAUT is an impressive-sounding name but make no mistake, the pundits of technology acceptance research consider this “parsimonious claim” deceptive (Straub & Burton-Jones, 2007). For example, performance expectancy is defined as one of the five UTAUT constructs. The authors list as many as five underlying constructs including perceived usefulness, extrinsic motivation, job-fit, relative advantage, and outcome expectations. Nonetheless, several empirical studies have tested the effectiveness of this model (Gupta et al., 2008; Im et al., 2011; Venkatesh & Zhang, 2010; Venkatesh et al., 2012;). The UTAUT proposes five predictors, ‘performance expectancy’, ‘effort expectancy’, ‘social influence’, ‘facilitating conditions.’ Since the introduction of this model in 2003 this model has been used extensively mainly in IS research (Venkatesh et al., 2016).

Bagozzi (2007) reports that the knowledge of technology acceptance is increasingly becoming fragmented with little coherent integration. The author cites the example of UTAUT which has five predictors but with as many as “41 underlying independent variables for predicting intentions and at least eight independent variables for predicting behavior” (Bagozzi, 2007, p. 245). The author also observes that with such

a model, technology acceptance is reaching a stage of chaos (Bagozzi, 2007). Bagozzi (2007) brands these five predictors as fundamental, generic, or universal, and uncovering any new predictors by future research might not embody the existing predictors.

The factors of the above theoretical models are taken into consideration in the qualitative study of this research.

2.2 Studies Related to Technology Adoption

This section of the study provides a consolidated list of factors/variables (Appendix H) that is used in the qualitative study of this research (proposed model provided in Chapter 3, Figure 2). As part of the qualitative study using a brainstorming session, focus group session, and individual session a dozen factors are identified out of these 32 factors. As mentioned in a previous section, these factors are derived from papers published in various academic journals, conference proceedings, industry technical papers, Gartner's reports, Hadoop Software documents (e.g., Apache Software Foundation wiki), and Hadoop vendor software documents.

As part of the literature review on big data this study searched the terms 'technology adoption', and "big data technology" in peer-reviewed articles written during the 2011 – 2018 period. The term was searched in digital libraries including ACM, IEEE Xplore, EBSCOHOST, and Google Scholar. It provided more than three hundred papers from dozens of diverse journals including technology management, information systems, computer science, social and business journals, and well conference

proceedings. This study took a cursory look at the titles, abstracts, actual work done, and conclusions of each of the papers and filtered out those papers that did not focus on big data topics. With these criteria, the study came up with a little over one hundred papers. These papers covered different areas of big data. The search and analysis focused on research papers employing scientific research methodologies. These criteria allowed to filter down papers that were industry papers as well as discussion papers. As part of the literature review, an effort was also made to see how data management technologies (data warehousing, database system) prior to big data technologies had been adopted previously. Some factors are selected from those papers as well. Some factors are incorporated from big data-related industry papers, vendor publications, and software documents. These factors are be used for qualitative studies in this research. As part of the qualitative study, the industry big data experts are given shortlist factors which are later used to develop the research model of this research.

1. Performance Expectancy: The performance expectancy factor relates to users' usability of software technology, infrastructure performance in terms of runtime, and computing resources utilization. Venkatesh (2000) has used this factor as one of the independent variables in his model (UTAUT). In IT, knowledge workers have a desire to be successful and attain achievement on the job (Venkatesh & Zhang, 2010; Zhang, 2017). Performance expectancy implies that users realize gains (Mithas et al., 2011) by using technology. This model has been used a lot in recent days. This research includes

the performance expectancy factor for consideration in the qualitative study. Industry experts of the qualitative study will make a decision about whether it could be part of the proposed model of this research.

2. Relative Advantage: This factor originates in the Diffusion of innovation (DOI) theory developed by E.M. Rogers in 1962 (Rogers, 2003). In his seminal book titled, “The Diffusion of Innovation,” Rogers (2003) identifies relative advantage as one of the top five innovation attributes which influence the rate of adoption. Prior research using meta-analysis in technology innovation adoption finds relative advantage as one of the top three innovation attributes (Ramamurthy et al., 2008). Fichman and Kemerer (1993, p. 10) state that “innovation is considered to have a relative advantage if it is technically superior in terms of cost and functionality than the technology it supersedes.” big data technologies are open-source, and technologies are cheaper to store, and process complex and large volumes of data compared to commercial database systems (Rahman and Sutton, 2016). The HDFS is capable to store such data, whereas some other conventional data storage systems are not. An innovation that has a relative advantage provides economic and organizational political legitimacy in making adoption decisions (Ramamurthy et al., 2008; Arts et al., 2011). In their big data adoption framework Sun et al (2018) mention that, this factor might be an influential factor in adopting big data. Hence, this factor is included in the qualitative study part of this research study for further investigation.

3. Scalability: Scalability has been identified as one of the most important capabilities that is needed to run a data warehouse efficiently (Rahman & Rutz, 2015; Sen & Jacob, 1998; Sen & Sinha, 2005). In big data analytics, scalability is identified as one of the important dimensions of efficient data analytics (Anagnostopoulos & Triantafillou, 2020; Menon & Sarkar, 2016; Tsai et al., 2015). Most of the traditional relational databases lack scalability in dealing with hundreds of terabytes of data. The industry papers on big data technology identify scalability as an important driving force behind Hadoop's popularity and adoption (Shvachko, 2011). In big data, new NoSQL technologies emerged to provide performance and scalability (Lourenco et al., 2015; Rahman, 2013). One of the major capabilities of Hadoop distributed file systems is its scale-out storage system (Aye & Thein, 2015). Hadoop's scalability capability is, at least, in three areas: storage, data processing, and machine learnings (García-Gil et al., 2017; Li et al., 2020; Rahman, 2018a). Big data pioneer user companies like Facebook and Google choose Hadoop and HBase for availability, tolerance, and scalability reasons (Borthakur et al., 2011; Olson, 2010). To the best of our knowledge, this factor has not been used as an independent variable of any technology acceptance model. Since the importance of this factor mentioned in both academic and industry papers, we include this factor in the qualitative study of this research.

4. Compatibility: The compatibility factor originates from Rogers' DOI theory (Rogers, 2003). It is one of the five important innovation characteristics. Big data and its tools and technologies are not compatible with conventional data storage systems, transformation tools, and reporting tools. This is because big data is unstructured, in large volume, and in high velocity. Hence, developers also need to acquire new skill sets to use big data tools and technologies (Lee, 2017). Conventional tools, technologies, and skillsets are developed around 'normal data', that is, dealing with transactional data only as opposed to structured data. Fichman and Kemerer (1993, p. 10) assert that innovation needs to be compatible "with existing values, skills, and work practices of potential adopters." Prior research suggests compatibility as an important innovation characteristic to adopt big data (Arts et al., 2011; Chen et al., 2015; Sun et al., 2018). Chen et al. (2015) validate compatibility as a predictor variable of big data analytics use for supply chain value creation. All these research findings beg a reality check with the industry experts about this. Hence, this factor has been included in the qualitative study of this research.

5. Complexity: The complexity factor also originates in the DOI theory (Rogers, 2003). Big data is very large and complex in terms of its characteristics (volume, velocity, variety, veracity, and value). But for big companies who have experts and highly skilled developers, it might not be as complex as needed to implement big data technologies in their organization. Leavitt (2013) observes that big data adds business value, but it is too

complex and expensive for smaller businesses. The analytics, machine learning and different reporting tools need to be run on HDFS using MapReduce and Spark processing engines. Big data velocity requires real-time complex analysis (Chardonens et al., 2013) and extracting complex patterns (Najafabadi et al., 2015). Jin et al. (2015) describe the challenges of big data processing in terms of data complexity, computational complexity, and system complexity. Russom (2013) reports data integration complexity of big data. Amudhavel et al. (2015) state that big data is so large or complex that traditional data processing applications are not capable to handle it. Hence, users may or may not favor the acceptance of big data technologies. For a reality check, we subject this factor to the experts of the qualitative study of this research.

6. Cost effectiveness: Economists claim that new technology causes cost growth, but they say it brings benefits as well (Hodgson, 2011; Kohli et al., 2012). Most of the Hadoop-based big data tools and technologies are open source and are therefore, supposed to be cost-effective. Also, several case studies' results show that big data applications have resulted in organizations' ability to avoid the cost. Bologa et al. (2010) report that big data has made it possible to detect insurance fraud within a reasonable time frame. Villars et al. (2011) state that timeliness of the response using big data helped in eliminating the legal and financial costs associated with fund recovery. Russom (2013) and Hartmann et al. (2014) also report cost containment and cost advantage by using big data technologies. This factor has not been used as an

independent variable in the technology acceptance model. Since big data industry papers suggest this as an import factor, we include it in the qualitative study of this research.

7. Total Cost of Ownership: The capability of a technology that is cost-effective does not incur significant hidden cost during the lifecycle and is easy to dispose of at the end of life. Big data tools are mostly open source. However, if vendor support is needed it would be interesting to see how much total cost of ownership is involved. Hence, we include this factor in the qualitative study of this research.

8. Trialability: The trialability factor has originated in the DOI theory (Rogers, 2003). Innovation needs to be able to be tested on a trial basis with little or no expense (Fichman & Kemerer, 1993). This factor has been validated by prior research (e.g., Arts et al., 2011). Trialability is related to the risk of no benefit or value. Fichman and Kemerer (1993, p. 9) state that “adopters look unfavorably on innovations that are difficult to put through a trial period or whose benefits are difficult to see or describe. These characteristics increase the uncertainty about the innovation’s true value.” Hadoop tools and technologies provided by the Apache Software Foundation are open sources. That means these technologies allow for trialability to understand the benefits of it. Hence, we include this factor in the qualitative study of this research.

9. Security and Privacy Considerations: Data privacy is reported to be one of the concerns of big data adoption (Jain et al., 2016; Raguseo, 2018; Sun et al., 2018; Wessel & Helmer, 2020; Wu et al., 2017). The extant literature suggests that big data technologies must fulfill some specific requirements such as handling sensitive data relating to individuals, firms, and governments (Lee, 2017; Menon & Sarkar, 2016). Richards and King (2014) state that big data technologies need to ensure privacy, confidentiality, and identity as many data originate from users' personal data. Gray (2014) reported that for safe enterprise data deployment, Hadoop lacks security functionality. Martin (2015), and Wessel and Helmer (2020) point out that one of the ethical issues arise from reselling consumers' data to the secondary market for big data. Tang et al. (2019) state that complex big data systems are becoming attack targets by emerging threat agents. The authors present a statistical model for vulnerability disclosures to provide organizations with important insights, so they can become more proactive in the management of cyber risks. We also need to see how all these factors influence big data technology acceptance. Since the data security and privacy concerns get significant attention these days, we take this factor into consideration as part of the qualitative study of this research.

10. Observability: Observability is one of the five innovation characteristics in the DOI theory (Rogers, 2003). This characteristic makes it easy to observe a technology's effectiveness and benefits, and also easy to communicate with others (Fichman &

Kemerer, 1993). Brown-Liburd et al. (2015) observe that big data causes too much information that sometimes goes beyond decision-makers' limited ability to process large amounts of information. However, Leavitt (2013) observes that big data adds business value, but it is too complex and expensive for smaller businesses. Hence, we need to understand big data acceptance in terms of observability attributes. This factor has been used by empirical studies that used DOI as a research model (e.g., Arts et al., 2011).

11. Flexibility: Extant literature suggests flexibility as an important capability of information technology infrastructure (Byrd & Turner, 2000). A system or technology's capability of flexibility allows for having positive results in its use and hence influences its acceptance by the user community (Basoglu et al., 2007; Seneler et al., 2008). Big data tools and technologies provide greater flexibility to collect data from many different sources into one single storage system (Rahman & Rutz, 2015). Abouzeid et al. (2009) emphasize query interface flexibility as it is important for analytical data management as business analysts. These sources include traditional data such as transactional data from enterprise resource planning (ERP), new data such as social media, sensor data, email messages, etc. Hadoop can be used for a wide variety of purposes, such as real-time streaming and processing, log processing, develop recommendation systems, build a data warehousing environment, perform predictive analytics, market campaign analysis, and fraud detection (Li et al., 2020; Nemschoff,

2013). Consolidated data within a single platform provides robust machine learning and data analytics capabilities (Rahman, 2018a; Rahman & Iverson, 2015). Hence, this factor has been subjected to the qualitative study of this research.

12. Fault Tolerance: The fault tolerance factor is derived from big data industry papers. To the best of our knowledge, this factor was not used in any technology acceptance empirical study. Big data technology Hadoop is best known for its fault tolerance capabilities. Hadoop's distributed file system uses commodity hardware to process by providing high throughputs with fault tolerance capabilities (Abouzeid et al., 2009). It maintains multiple copies of the same data into different nodes in the cluster so in the event of failure another copy can be made available for use (Nemschoff, 2013). Hadoop has this particular advantage over conventional database systems. Hence, this factor has been included in the qualitative study.

13. Reliability: Reliability of technology is considered a basic and important characteristic (Barlow, 1984). This factor is identified as one of the important factors of technology adoption taxonomy (Seneler et al., 2008). The Hadoop Distributed File System (HDFS) is destined to store hundreds of terabytes to petabytes of data reliably (Shvachko et al., 2010). Hadoop's distributed file system is fault tolerant. If one node goes down other nodes take over. Data is replicated into three copies into other nodes. Hence, data loss possibility is much less. In data management space, reliability is related

to the volume and velocity of data movement. Data management tools and technologies are expected to withstand the velocity data movement. Hence, we include this factor in the qualitative study of this research. To the best of our knowledge, this factor has not been part of technology acceptance models.

14. Data Storage and Processing Capability: The data storage and processing capability factor has not been used as an independent variable in technology acceptance studies. Big data platform consists of two main components: big data storage and big data processing. Hadoop is known for its high scalability from storage and data processing perspectives (Shvachko et al., 2010). Most of the traditional database systems are not capable of handling hundreds of terabytes of data and also not scalable. Hadoop's storage capacity and data processing capability might be considered an important factor to influence on big data acceptance. Hence, we add this factor to the qualitative study of this research.

15. Output Quality: The output quality factor originates in TAM2 (Venkatesh & Davis, 2000). As part of TAM2, Venkatesh and Davis (2000) present that output quality is a measure in terms of how well a system performs the tasks which it is destined to perform. This factor has been tested and validated by subsequent studies (Wixom et al., 2001). In data management discipline, the output quality is meant for the quality of the data. Côte-Real et al. (2020) conduct an empirical study that reveals data quality in

terms of completeness, and accuracy, and currently can significantly impact firm performance both directly and indirectly. Big data are mostly unstructured. After processing such unstructured data using Hadoop's processing software, the quality of data comes into picture and question. This factor could be considered an important factor of big data technology acceptance. Hence, the output quality factor has been subjected to the qualitative study of this research.

16. Organizational commitment: Organizational commitment is reported as one of the organizational factors for data warehouse success (Ramamurthy et al., 2008). In big data adoption, management support is called out (Russom, 2013). An organization's IT department and data scientist need to take initiative to show the business value of big data to get top management support (Rajpurohit, 2013).

17. Top Management Support: Top management support is identified as one of the organizational dimensions that influence the adoption of data warehouse technology (Hwang et al., 2004). Since big data is a new area of data management, top management support might be crucial for Hadoop adoption. Hence, this factor has been incorporated into the qualitative study of this research.

18. Facilitating Conditions: The facilitating conditions factor originates in the technology acceptance model, UTAUT, developed by Venkatesh et al. (2003). Facilitating conditions

is considered as one of the key factors in data warehouse architecture selection (Ariyachandra & Watson, 2010; Rahman, 2017). Seneler et al. (2008) identify this factor as one of the factors of technology adoption taxonomy. Since big data technologies are complex, we assume that big data technology acceptance is influenced by facilitating conditions. Facilitating conditions might be available in the external environment (e.g., vendor support). Facilitating conditions might need to be available within the organization as well, such as in IT infrastructure support. Hence, we add this factor to the qualitative study of this research.

19. Image: The image factor has been used in TAM2 as an independent variable (Venkatesh & Davis, 2000). Image is the degree to which the use of new technology enhances one's image or status within the organization. Originally, Moore and Benbasat (1991) introduced and validated this factor in the innovation acceptance model. They point out that the users are mindful of whether the use of technology enhances their image, status, prestige, and profile within the organization and outside the organization. Venkatesh and Davis (2000) suggest that job performance by using technology eventually enhances one's image. In big data space, some professionals might believe that their image could be increased if they work in big data. We wonder why things like image, status, or prestige would influence a user's acceptance of the technology. The use of technology should not be influenced by the fact that others also use this

technology considering that use is personal or individual in nature. Nonetheless, we include this factor in the qualitative study of this research.

20. Self-Efficacy: Self-Efficacy is the “belief that one has the capability to perform a particular behavior” (Lee et al., 2003, p. 761). Igarria et al. (1995) introduced this factor in the technology acceptance model to examine the belief in terms of one’s capabilities of using a computer to accomplish certain specific tasks. Sun et al. (2016) posit that the user’s mindful state is also a crucial factor in adopting the technology. The authors assert that mindful adopters will be more likely to perceive technology as useful. Since big data technology is complex and requires certain skillset, we include this factor in the qualitative study to examine this factor’s influence on big data adoption.

21. Subjective Norms/Social Influence: The subjective norms/social influence factor originates in the theory of reasoned action (TRA) developed by Fishbein and Ajzen (1975). Later it was used in the theory of planned behavior (TPB) introduced by Ajzen (1991). Subjective norms/ social influence is meant for a person’s “perception that most people who are important to him think he should or should not perform the behavior in question” (Lee et al., 2003, p. 761; Venkatesh & Davis, 2000). In his original TAM version, Davis has not included subject norm or social influence perhaps due to the fact the subject norm construct is context-driven (Dillon & Morris, 1996). With big data being, a new field, and since learning its new technologies is considered next-generation

tools learning, social influence in terms of peers in the organization or industry might play a pivotal role in using and accepting those new tools and technologies. Hence, it might be worth taking social influence as an important factor.

22. Job Relevance: The job relevance factor originates in TAM2 (Venkatesh & Davis, 2000). "The capabilities of a system to enhance an individual's job performance" (Lee et al., 2003, p. 761). Job relevance is considered to have an influence on perceived usefulness (Venkatesh & Davis, 2000). The technology acceptance is dependent on one's job relevance, which is defined as, whether the user finds it useful or whether the system is capable of supporting the user's daily job performance. Hence, we include this factor for the qualitative study.

23. Results Demonstrability: Results demonstrability is the "degree to which the results of adopting/using the IS innovation are observable and communicable to others" (Karahanna et al., 1999, p. 188; Venkatesh & Davis, 2000). Originally, Moore and Benbasat (1991) came up with the idea that results in demonstrability are meant for the tangibility of the results of using innovation. Later, Venkatesh and Davis (2000) theorized in the TAM2 model that results in demonstrability have a direct influence of perceived usefulness. Agarwal and Prasad (1999) also validated and found a significant correlation. Hence, this factor is added to the qualitative study of this research.

24. Functionality: Functionality is property or features that meet the functional aspects of the technology that a user is looking for. In big data space, functionality is big data tools and technologies' capability or feature that can handle a large volume of data most of which is unstructured and cannot be received or processed using the conventional data storage systems and associated tools and technologies. This factor has not been used in TAM research. We include this factor in the qualitative study.

25. Effort Expectancy: This factor originates in the technology acceptance model, UTAUT, presented by Venkatesh et al. (2003, 2012). Effort expectancy is "related to the degree of ease associated with the use of technology" (Venkatesh et al., 2003, p. 450). Since big data is complex, due to its unstructured nature, it will be interesting to see how easy the big data tools are to use and operate. Hence, we include this factor in the qualitative study of this research.

26. Voluntariness: The voluntariness factor is used as a mediating factor in TAM2, developed by Venkatesh and Davis (2000). Voluntariness is the "degree to which use of the innovation is perceived as being voluntary, or of free will" (Barki & Hartwick, 1994; Lee et al., 2003, p. 761; Venkatesh & Davis, 2000). Originally, Moore and Benbasat (1991) proposed voluntariness as a factor in accepting innovation. The authors attempted to understand whether voluntary use of technology, as opposed to mandatory use, makes any difference in accepting a technology.

27. Data Analytics Capability: Ghasemaghaei (2019) presents that data analytics competency in terms of big data utilization, analytics capability, and tools sophistication mediated by knowledge sharing can improve decision making quality. This factor has not been used in technology acceptance research. We believe this is an important factor in the data management field. Analytical, data mining and reporting tools can run against the Hadoop distributed file system. With Hadoop, there is great prospect of running robust data mining against a complete set of data stored in HDFS (Rahman, 2018a). Zhang et al. (2019) present big data analytics capability air pollution management for sustainability. Wlodarczky and Hacker (2014) provide an account of current trends in predictive analytics of big data. Hadoop has reach machine learning libraries including Mahout (MapReduce) and MLib (Spark) which are developed to perform analytics based on a large and complex set of data that resides in HDFS (Tsai et al., 2015). Wu et al. (2019) report a strong relationship between data analytics capabilities, innovation, and firm productivity. Verma et al. (2018) report that big data analytics might have direct and indirect effects on the acceptance of big data technologies.

28. Enjoyment: Enjoyment is the extent to which the “activity of using a specific system is perceived to be enjoyable in its own right, aside from any performance consequences resulting from system usage” (Chin & Gopal, 1995, p. 47). We are curious if this factor plays any role in Hadoop adoption since Hadoop technology is a bit new, robust, and

complex. This factor has been validated as part of TAM (Wu et al., 2007). Hence, we include this factor in the qualitative study of this research.

29. Absorptive Capacity: Bradford and Saad (2014) state that absorptive capacity is very important for a firm's ability to recognize the value of, and to have resources, human capital, and willingness to exploit external new knowledge and promote that for products and services development. Absorptive capacity is also one of the organizational factors in data warehouse success (Rahman, 2017; Ramamurthy et al., 2008). Big data consists of a large number of tools and technologies. To handle these technologies, adequate skillset and financial resources are also needed. Small and medium-sized business firms might find it challenging to build a comprehensive big data infrastructure and ecosystem. We need to study whether absorptive capacity plays a role in big data acceptance. Hence, this factor has been added to the qualitative study of this research.

30. Organizational Size: Organizational size in terms of the workforce in IT might play a role in adopting and maintaining new technologies (Sun et al., 2018). Since big data tools and technologies are new capabilities in data management, learning those tools and maintaining them requires a workforce and other resources. Ramamurthy et al. (2008) identify organizational size as one of the organizational factors to adopt data warehousing technology. Hence, we include this factor in the qualitative study.

31. Competitive/Industry Pressure: The competitive/industry pressure factor is suggested as one of the environmental factors of technology acceptance (Chen et al., 2015; Hwang et al., 2004). In big data research, it was mentioned that the organizations that adopt big data would be ahead of the competition. Big data is used by organizations to drive business performance. Spiess et al. (2014) report their use of big data helps to improve customers' performance as well as business performance. Barney (1991) defines competitive advantage: "A firm is said to have a competitive advantage when it is implementing a value-creating strategy not simultaneously being implemented by any current or potential competitors" (Barney, 1991, p. 102). We believe that by using big data strategically, organizations can achieve business value and stay ahead of competitors (Hagiu & Wright, 2020). Hence, this factor is included in the qualitative study.

32. Training and Required Skills: In big data, one big challenge is the lack of required skills in analyzing big data (Lee, 2017). It requires the use of a handful of tools and a skillset is needed in programming languages (Davenport & Patil, 2012). In traditional data management, companies have developed skills over a period of time that are useful in dealing with traditional data analysis only (Russom, 2013; Wixom et al., 2001). Big data is a new and different phenomenon for analyzing big data. Brown-Liburd et al. (2015) reported that required training and skills might play an important role in

adopting big data technologies. Hence, the training and required skill factor is include in the qualitative study of this research.

2.3 Taxonomy Factors

A literature review on data management software has provided 32 factors (Section 2.2) that are categorized in a taxonomy into six dimensions (Table 4). These dimensions include environmental, individual, organizational, technological, economic, and legal. Under those six dimensions consisting of 32 factors 12 factors have been selected by an expert panel of big data to use in the proposed research model (see sections 3.5 – 3.6 in Chapter 3). In Chapter 5, we have mentioned that eight of those 12 factors got validated and accepted by statistical analysis using structural equation modeling (SEM) software.

Table 4: Taxonomy of Factors Based on Literature Review

Adoption of Big Data Technology					
Environmental	Individual	Organizational	Technological	Economic	Legal
Facilitating Conditions	Image	Organizational commitment	Performance Expectancy	Cost effectiveness	Security and Privacy
Subjective Norm/Social Influence	Self-Efficacy	Top Management Support	Relative advantage	Total Cost of Ownership	
Competitive/Industry Pressure	Voluntariness	Job Relevance	Scalability		
		Organizational size	Compatibility		
		Training and required skills	Complexity		
		Facilitating Conditions	Observability		
			Flexibility		
			Fault tolerance capability		
			Reliability		
			Data Storage & Processing Capability		
			Output Quality		
			Results Demonstrability		
			Functionality		
			Effort Expectancy		
			Data Analytics Capability		
			Enjoyment		
			Absorptive capacity		
			Triability		

Some of the factors classified as adoption taxonomy have reference to different technology adoption theory factors and some from industry papers. The TAM has reference to perceived usefulness and perceived ease of use. The TAM framework allows for applying external factors identified under six dimensions (Table 4). Past research applied several of these factors using TAM (Benbasat & Barki, 2007; Lee et al., 2003). These factors are task performance, efficiency, innovativeness, management commitment, results from demonstrability, quality, relative advantage, compatibility, complexity, observability, subjective norms, visibility, facilitating conditions and prior experience. Many of these variables belong to factors classified under environmental, organizational, and technological classifications in Table 4. Resource-based view (RBV) theory has reference to environmental and economic dimensions which include business value, rareness, imitability, and substitutability to achieve competitiveness by a firm (Eisenhardt & Schoonhoven, 1996; Jelinek & Bergey, 2013; Wernerfelt, 1984; Teece et al., 1997). Big data capability has implications for important resources such as technological, strategic and economic. Several factors in Table 4 have reference to other technology acceptance models (Fishbain & Ajzen, 1975; Kuan & Chau, 2001; Venkatesh et al., 2003): TRA (subjective norms), TPB (perceived behavioral control), TOE (technological, organizational and environmental) and UTAUT (performance, facilitating conditions) (Venkatesh et al., 2012).

2.4 Research Related to Big Data Technology Adoption

As big data is a new discipline, there are a few studies conducted on big data technology adoption (Chen et al., 2015; Kwon et al., 2015; Malaka and Brown, 2015; Esteves and Curto, 2013). One of the studies (Kwon et al., 2015) examines big data adoption based on two factors, data quality management and data usage experience among South Korean companies using RBV and Isomorphism theorems. The authors point out that their research was an initial study of big data technology adoption (Kwon et al., 2015). The authors first suggest continuing this type of study on the firm's other internal and external conditions of business, and the second, they suggest conducting further study to identify organizational variables and other conditions to understand big data technology adoption. In this comprehensive big data technology, acceptance research model steps were taken to tackle these factors.

The second study was conducted by Malaka and Brown (2015) on a South African telecommunications organization using the TOE model. The scope of this research was very limited. They interviewed seven participants from IT and business. Their findings revealed technology challenges "to the adoption of big data analytics as being data integration, data privacy, return on investment, data quality, cost, data integrity, and performance and scalability." And from an organizational standpoint, "the major challenges were ownership and control, skill shortages, business focus and prioritization, and unclear processes." From the environmental context, market

competition, vendor reliance, and data security and privacy were examined but no major challenges are reported.

The third study was conducted by Esteves and Curto (2013) using a mix of TPB, DOI, and TAM theoretical models. The authors used as many as 15 factors in the research model but did not provide enough information in regard to measures of those 15 factors. Also, the discussion section of the paper was a bit brief which leaves the reader with little or no convincing information. Hence, no valid conclusion could be made about those 15 identified factors used in the empirical model.

Fourth, Verma et al. (2018) conduct an empirical study on big data analytics adoption consisting using latent constructs of TAM: PU, PEOU, Attitude, and Behavioral Intention to use. The authors use big data analytics system quality and information quality along with a mediating factor along with beliefs in the benefits of big data analytics to assess the influence of PU and PEOU. They find that both system quality and information quality influence the core TAM constructs by virtue of user belief in the benefits of big data analytics.

Table 5 provides a summary of four empirical research outcomes on big data technology adoption.

Table 5: Empirical Research on Big Data Technology Adoption

Research Topic	Theory/ Model	Exogenous Variables	Endogenous Variables	Results
Data Quality Management, Data Usage Experience and Acquisition	RVB, Isomorphism	Data Usage Experience, Data Consistency, Data Completeness, and Resource Facilitating Conditions.	Acquisition Intention of big data analytics	Data Usage Experience, Data consistency, Data completeness,

Intention of Big Data Analytics (Kown et al., 2014).				and facilitating conditions – all positive
Challenges of the Organizational Adoption of Big Data Analytics: A Case Study in the South African Telecommunications Industry (Malaka and Brown, 2015)	TOE Framework	Technology: Time and Cost, Data Integration, Veracity, Performance and Scalability; Organization: Ownership and control, skill shortage, communication processes; Environmental/External: Industry/Market competition, vendor reliance, and Data security and privacy	Adoption and Usage	Major challenges experienced were technological and organization but, not with external environment
Influences on the use and behavioral intention to use big data (Hood-Clark, 2016)	TAM	Perceived usefulness, perceived ease of use, and attitude toward use	Behavioral intention to use, and actual use	The main challenge of using and adopting the use of big data is transforming the culture, processes, and people in the organizations
An extension of the technology acceptance model in the big data analytics system implementation environment (Verma et al., 2018).	TAM	System quality, information quality, beliefs of system benefits, perceived usefulness, perceived ease of use, and attitude toward use	Behavioral intention to use	Both system quality and information quality influence the core constructs of TAM through a mediating factor, belief in the benefits of big data analytics

2.5 Research Gaps

Existing literature provides the state of big data technology development (Saheb & Saheb, 2020) and results of case studies, machine learning techniques, predictive modeling, surveys, and experiments (Al-Jarrah et al., 2015; Chardonens et al., 2013; Kambatla et al., 2014; Kiron et al., 2013; LaValle et al., 2011). But this literature did not provide much insight into the overall usage of big data tools and technologies.

Technology acceptance is considered to be the determinant of the success of a product or technology. Studying acceptance from the users' perspective gives new insight about likes and dislikes of different features, the product itself, and the user's attitude toward the product. A systematic study of the review of big data is needed to understand the overall picture of the big data technology acceptance rate.

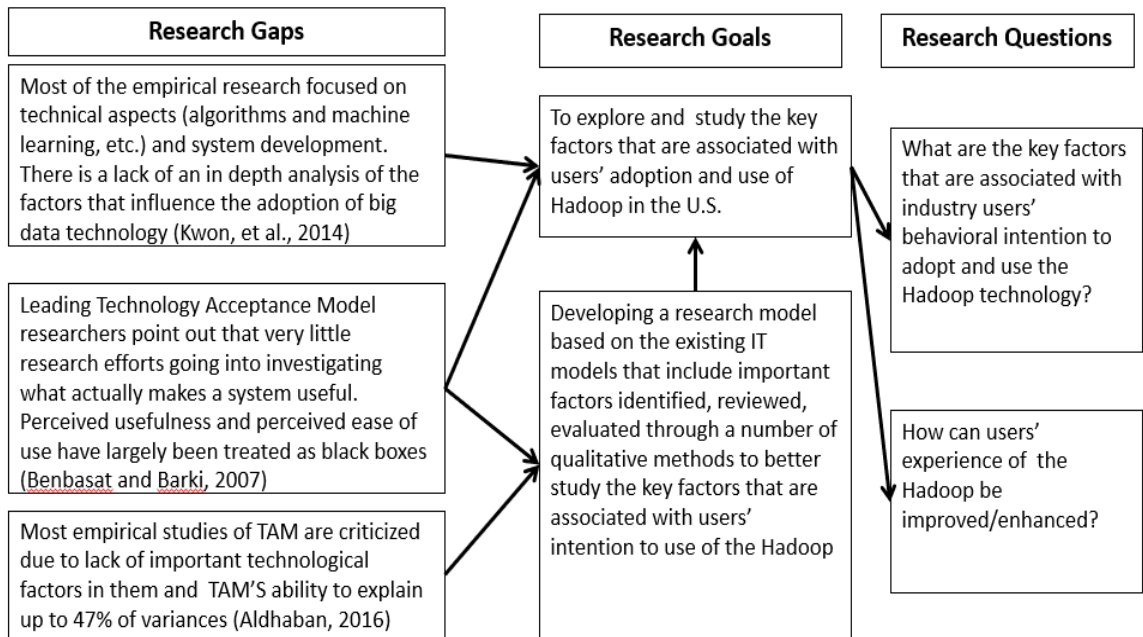
The technology acceptance model (TAM) has been developed by Fred D. Davis (Davis, 1986) as part of his doctoral dissertation at MIT Sloan School of Management to empirically test new end-user information systems. Since then, TAM has been applied frequently for research into the acceptance of new information technology.

This model has gained popularity among practitioners and researchers over the last two decades. The model has been tested and applied in many fields. These include switching cost on accounting software use (Gogus & Ozer, 2014), enterprise resource planning (ERP) software system implementation (Amoako-Gyampah & Salam, 2004; Basoglu et al., 2007; Rajan & Baral, 2015), software evaluation and choice (Szajna, 1994), worldwide web (Lederer et al., 2000), ease of use and usage of information technology (Adams et al., 1992; Davis, 1989), and user acceptance of computer technology (Davis et al., 1989; Davis, 1993), to name a few. In their 2007 paper in the Journal of AIS, Venkatesh, Davis, and Morris put it in the title as to whether TAM is "dead or alive" (Venkatesh et al., 2007). And later, in the conclusion section of the paper, they pronounced the verdict that the research on technology adoption is not

dead! However, they suggest continuing research on TAM by focusing on interesting questions that solve business problems.

To our knowledge, there are a few empirical studies on big data technology (e.g., Hadoop) that used TAM (Hood-Clark, 2016). This makes sense since big data, core big data technologies, and big data ecosystems have emerged during the middle of the last decade. This could be considered a research gap. This study conducts formal research on the user acceptance of big data technology, namely, the Hadoop Distributed File System (HDFS). The research gaps are provided in Table 6.

Table 6: Research Gaps and Research Goals



Chapter 3 Developing Research Model and Research Hypotheses

This dissertation consists of distinct studies: qualitative study and quantitative study. This chapter covers the qualitative studies. Chapter four will discuss quantitative studies. Discovering the antecedents of technology use is viewed as a pivotal factor in the field of technology adoption (Dillon & Morris, 1996). Sekaran and Bougie (2016) suggest that the research model needs to be grounded upon existing theories and previous research. This research took several steps to identify factors affecting big data technology acceptance. First, it reviewed the existing theories of technology acceptance that came from different disciplines including Information Systems (IS), Psychology, Communications, and Economics. Chapter two provided details of existing theories of technology acceptance. The factors used in these models are taken into consideration for this research. Second, this research has done an extensive review of previous research relating to data management software acceptance including database systems, data warehousing, and big data. With the help of extant literature on data management technologies ranging from conventional data warehousing to big data storage technologies (e.g., Hadoop Distributed File System), relevant factors have been taken into consideration. Third, this research also reviews big data white papers, industry technical papers, big data vendor documents, and Gartner reports on big data. Based on these literature reviews, 32 factors (Table 4) have been identified out of which 12 factors have been selected through a qualitative study and used as exogenous variables in a comprehensive big data technology acceptance research model. These 12 factors

fall under five major areas including technology, organizational, environmental, economic, and legal. In the final model, eight factors are accepted by the SEM model (discussed in chapter 5 of this dissertation).

Besides depending on theories of technology acceptance and empirical research on data management software, we made additional steps using a qualitative study to identify possible factors that might affect big data technology acceptance. As part of this qualitative study, we conducted brainstorming sessions consisting of nine experts who work in the industry in the big data fields (section 3.2 in Chapter 3). We conduct a focus group session consisting of 10 experts in big data discipline (section 3.3 in Chapter 3). We also conduct individual interview sessions with 21 professionals who are experts in the big data field (section 3.4 in Chapter 3). The latter is to make sure they could suggest the most important factors as well as new factors relevant to big data and Hadoop that might not be available in previous research since technology changes faster and industrial users' perception of technology use also change.

3.1 Defining Perceived Usefulness

Davis' technology acceptance model includes two key factors, perceived usefulness and perceived ease of use (Davis, 1989). This model has been tested successfully in IS research (Adams et al., 1992; Davis, 1989). This model is reported to explain 47% variance (Dillon & Morris, 1996; Lee et al., 2003). Even though this is a widely used model in IS there is some valid criticism of this model made by scholars of technology acceptance theories. Benbasat and Barki (2007) and a host of other researchers argue

that study after study has been conducted using this model but without making effort to clarify what is meant by 'usefulness'. This research makes an attempt to shed light in regard to the meaning of usefulness.

One definition of usefulness states that "a product, website or application should solve a problem, fill a need or offer something people find useful." (Sauro, 2011).

According to the Merriam-Webster Dictionary, usefulness is "the quality of having utility and especially practical worth or applicability." The Utility Theory of economics states that a product must have the ability to satisfy needs or wants and the consumer of that product has to experience satisfaction. The theory of utility also emphasizes that a rational person will choose the option that provides the highest utility.

Bentham (1824) define utility for the first time: "By 'utility' is meant the property of something whereby it tends to produce benefit, advantage, pleasure, good, or happiness or to prevent the happening of mischief, pain, evil, or unhappiness to the party whose interest is considered." In this definition, the keyword 'property' has implications for the technological capability of Hadoop relating to its five characteristics (5 V's). This research takes the technological capabilities of big data such as scalability, reliability, flexibility, and the robustness of data storage and processing capability into consideration. Previous research applied TAM on light technologies or products: email, spreadsheet, micro-computer, word-perfect, write-one, and so on. Compared to these, the big data technology, Hadoop, is technologically complex and robust as it was designed to deal with hundreds of terabytes of data most of which are unstructured. In

many cases, this data comes from the source very fast. This data needs to be processed faster. The machine learning model needs to run on the Hadoop platform faster. Thus, we need to see the applicability of TAM in explaining big data technology acceptance from that perspective.

Swanson (2019) suggests that technology needs to be associated as a concept with routines as well as patterns of action to allow for providing capabilities. The author suggests four principal modes of change: design concept in creating new tools, execution plan to ensure routines in operating the technology, diffusion of technology and routines to maximize its use, and the mindset of the shift in adapting technology and routines to keep up with best practices.

A look at attitude theory from psychology dictates that a product's design features follow the perception of attitude and then finally end up with usage. Existing literature on big data technology development and application suggests that big data technologies have come into the picture to address certain capabilities issues of data management. Those capabilities are mainly related to five characteristics of big data: volume, velocity, variety, veracity, and value. Big data tools need to be scalable, robust, and efficient due to the magnitude of data that needs to be handled by big data technology and the rate data needs to be received and processed. By taking these into consideration, it is assumed that big data technology acceptance might be dependent on scalability, data storage, processing, flexibility, reliability, and machine learning capability.

3.2 Brainstorming Session

This research is designed to study a small set of factors that are influential and provide insights into big data technology acceptance. In order to narrow down the list of factors (listed in Table 5) the researcher used a qualitative study that consists of an expert panel comprised of experts and knowledgeable persons who have worked in big data projects for three or more years.

One important aspect of a qualitative study is to make an effort to find something which a researcher is not able to see or observe or make sense of due to a different view of the world. In such cases, the qualitative study helps to collect the views of others who might view the world or phenomenon differently than the researcher does.

“There are numerous famous examples where major discoveries were delayed or where observations were ignored because they did not fit prevalent theory and thus inhibiting progress and knowledge generation.” (Atlas.ti, 2017).

This expert panel discussed all of the identified big data factors via one meeting and recommended a shortlist of factors that they think would be important ones. Research suggests that expert panels can 1) provide inputs that is meaningful, rich, and not influenced by the researcher; 2) provide a deeper understanding of the phenomena being studied; and 3) provide researchers the ability to capture deeper information more economically than individual interviews.

The researcher had scheduled a one-hour virtual meeting inviting 13 people with expert knowledge in big data, from the IT department of a local company. The virtual nature of the meeting allowed participants to join the session from multiple locations and sites of the company. All of them have big data project experience of three or more years. They worked in big data projects in various capacities (product manager, project manager, business user representatives, and developers). Nine out of 13 participants attended the meeting. They have diverse backgrounds of Hadoop: backend and frontend users, data scientists, business intelligence architects, solution architects, and managers.

At the start of the meeting, the researcher gave a background of the research. The participants were assured that their personal identity would not be disclosed anywhere in the research report. They had been given an explanation as to what is meant by big data technology and adoption. They were also informed about the specific big data technology the researcher was undertaking for this research. They had been provided information about the literature review efforts on this topic. Also, they were provided with a list of factors that were extracted from academic journals, industry papers, Gartner reports, and vendor documents about big data technology and its adoption. The researcher also briefly went over existing technology adoption models and theories to make them familiar with the factors used by those models. Since the researcher had identified a large number of factors based on theories, models, and academic research, the participants were requested to help in identifying important factors in terms of real-world business implications. They were also asked to propose

any new factor not on the list that they thought it is associated with users' adoption and use of big data. One of the participants commented that the factors to be chosen needs to be relevant to the five V's (characteristics) of big data: volume, velocity, variety, veracity, and value (Marr, 2015). This was a valuable input so the researcher asked the participants to choose the factors that relate to these five characteristics of big data since big data tools and technologies should deal with these five characteristics. The participants were also asked to select factors by taking into consideration as to what (especially technical aspects) make technology useful.

The participants were requested to select the factors by taking three main questions into consideration:

Q1: After the participants were provided with background information about this research the researcher let them take a pause to review the list of 32 factors. They are provided with definition/explanation of each factor. They are requested to provide their thoughts about these factors and also provide any new factors they know would be important but were not on the list provided.

Q2: Next the participants were asked to read the list of factors again including the new factors proposed as part of Q1. They were asked to eliminate any factors that they felt were similar or duplicate in terms of underlying meaning. They were asked to list down only important ones.

Q3: The participants were asked to review the short-listed factors again, reevaluate and validate the factors.

The TAM has two core constructs (dependent variables), perceived usefulness (PU) and perceived ease of use (PEOU) that are connected with external variables (Davis, 1993). We have asked participants to take these two variables into account when selecting external variables (out of 32 factors). The participants discussed the importance of factors among themselves and selected the factors by providing reasons for selecting a particular factor. Sometimes they debated and eventually came to a decision in selecting individual factors. During the selection process, participants were encouraged to select factors from across different categories such as technological, organizational, environmental, legal, and economic. They ended up selecting factors from technology, organizational, environmental, economic, and legal categories (Table 5).

Table 7: Participants in the Brainstorming Session

Participants	Affiliation/ Title	Years of using Hadoop
1.	Anonymous/ Big Data Product Manager	More than three years
2.	Anonymous/ Senior Hadoop Developer	More than three years
3.	Anonymous/ Senior Hadoop Developer	More than three years
4.	Anonymous/ Big Data ETL Developer	More than three years
5.	Anonymous/ Hadoop Developer	More than three years
6.	Anonymous/ Big Data ETL Developer	More than three years
7.	Anonymous/ Project Manager	More than three years
8.	Anonymous/ Big Data Business Analyst/ User Rep.	More than three years

Brainstorming session participants were given the below guidelines:

1. Be familiar with the definition of 'perceived usefulness' and 'perceived ease of use'. Think about the possible technological capabilities of Hadoop.
2. Review the brief description/ definition of each of the 32 factors.
3. Evaluate all 32 factors provided in the spreadsheet file.
4. Add any new factors which you believe might be associated with users' adoption and use of the Hadoop.
5. Select all the important factors.
6. When selecting the factors, please take into consideration what makes technology useful.

NOTE: Only brainstorming participants were asked to add any new factors because the session was conducted first.

3.3 Focus Group Session

A Focus group session is one of the data collection methods used in qualitative studies.

In this research, a focus group session was conducted to evaluate and identify factors of big data technology acceptance out of a list of factors listed based on theory, previous research, and brainstorming sessions described in the previous section. The focus group members were selected based on their in-depth knowledge, experience, and expertise in the big data domain. In this focus group session, 13 professionals were invited out of which 10 persons attended the session. They come from three different companies.

They are Hadoop users: backend and front-end users, architects, managers, and more.

They discussed and evaluated a list of 32 factors and later individually provided their inputs on important factors.

Focus Group session participants were given the below guidelines:

1. Be familiar with the definition of 'perceived usefulness' and 'perceived ease of use'. Think about the possible technological capabilities of Hadoop.
2. Review the brief description/ definition of each of the 32 factors provided.
3. Evaluate all factors provided in the spreadsheet file (includes any new factor proposed by the brainstorming session conducted earlier).
4. Select important factors that are relevant to Hadoop adoption.
5. When selecting the factors, please take into consideration what makes technology useful.

Note: Focus group participants were not asked to add any new factors because brainstorming session participants will not have a chance to vote for any new factors proposed by focus group session participants. The brainstorming session was already conducted.

3.4 Individual Interviews

The personal interview is considered one of the most widely used data collection methods in qualitative research. In this research, individual interviews are conducted to refine the factors of big data technology acceptance achieved, followed by findings based on theory, previous research, brainstorming, and focus group sessions. Here, individuals interviewed were selected based on their in-depth knowledge, experience,

and expertise in the big data domain. They come from 13 different companies and variety of job roles: CEO, data scientists, Hadoop architects, BI Analysts, program manager, product manager, backend, and frontend users. The persons interviewed were provided with a list of 32 factors that were developed using the technology acceptance theories, literature review, brainstorming, and focus group sessions. They were requested to review the list of factors, select, and then validate the most important factors related to users' intention to adopt Hadoop. The individual interview provides the researcher with an opportunity to review factors with a more in-depth perspective. The individual interview is typically conducted through face-to-face, telephone, or emails. The researcher used face-to-face and telephone interview methods. Interviews can be conducted using structured or unstructured methods.

This research used a semi-structured method which means that the individual interviewed were asked certain questions based on a predefined format and the remaining questions as a follow-up. Individuals interviewed were provided with an introduction of research and what is expected out of the personal interviews. They were offered to maintain the confidentiality of personal info as well as the organization at which they were employed. Any concerns of the person interviewed were addressed. An example could be publishing interview results in summarized format and thus personal or organizational information would be kept confidential. In regard to the topic of the interview, the person interviewed was requested to provide deep thoughts about the factors of Hadoop acceptance. Experienced users were chosen, and they were

encouraged to provide thoughts with an open mind. The individual interview results were each recorded to make sure they were authentic. At the end of the interview, each individual participant provided their selected list of factors in a spreadsheet document.

Individual-Interview session participants were given the below guidelines:

1. Be familiar with the definition of ‘perceived usefulness’ and ‘perceived ease of use’.
2. Review the brief description/ definition of each of the 32 factors provided.
3. Evaluate all factors provided in the spreadsheet file (includes any new factor proposed by the brainstorming session conducted earlier).
4. Select important factors that are relevant to Hadoop adoption.
5. When selecting the factors, please take into consideration what makes technology useful.

Note: Individual-interview participants were not asked to add any new factors.

The steps of the qualitative studies are summarized below.

Table 8: Summary of Steps to Develop the Qualitative Study

Research Steps	Description	Target Participants
Literature Review	An extensive literature search related to technology acceptance in general and big data technology acceptance in particular has been conducted.	
Brainstorming	An extensive interactive session to be conducted with nine industry experts via a one-hour session.	Experienced user of big data technology has been invited. They have more than three years of experience.

Focus Group	A one-hour session was conducted with another group of big data users consisting of 10 participants.	The criteria for selecting participants were based on experience as developers, systems analysts, user community.
Interviews	This was a one on one interview with a total of 21 persons. Interviews took 15 to 20 minutes for each participant.	The persons interviewed had hands-on experience with the big data tools and technologies development and use.

3.5 Results of the Qualitative Studies

This qualitative study consisted of three parts: Brainstorming, Focus Group session, and Individual one-on-one sessions. As part of this study, the participants were provided with 32 factors from which they were requested to select the important ones. These participants perform a variety of job roles: CEO, data scientists, Hadoop architects, BI Analysts, program manager, product manager, backend and frontend users. Tables 9 shows the results of this study. The top 15 out of 32 factors are shown in Table 9.

Table 9: Results of Qualitative Study

Rank	Factors/ Variables	No. participants voted for (out of total 40 participants)
1	Scalability	35
2	Data Storage and Processing	32
3	Cost-Effectiveness	32
4	Performance Expectancy	30
5	Security and Privacy Considerations	26
6	Reliability	26
7	Data Analytics Capability	25
8	Training and Required Skills	25
9	Flexibility	24
10	Output Quality	24
11	Functionality	24
12	Total Cost of Ownership (direct & indirect cost)	20

13	Facilitating Conditions (e.g., Vendor/Infrastructure/Customer Support)	18
14	Top Management Support	18
15	Fault Tolerance Capability	18

Defining the conceptual domain of individual constructs has a significant influence on maintaining the distinctiveness of each construct (Petter et al., 2007). A poorly defined construct can cause confusion as to what it does or does not refer to (Mackenzie et al., 2011; Petter et al., 2007). If the definition of a construct is not specified properly, its measures might be deficient, or the definition might overlap with the other constructs that already exist and are validated. Hence, the construct might draw invalid conclusions with other constructs in terms of relationships (Mackenzie et al., 2011).

A variable that is abstract and latent rather than concrete and observable is defined as a construct (Mackenzie et al., 2011; Nunally & Bernstein, 1994). Mackenzie et al. (2011) provide a guideline conceptualizing the constructs that involves examining the constructs used in extant literature in a particular subject, identifying the constructs in terms of entity and properties, specifying the constructs in terms of attributes or characteristics as succinctly as possible, and defining constructs clearly and concisely. We have identified and defined the constructs by following these guidelines. As part of the literature review, we have gathered academic journal papers, industry publications, big data-related software documentations, and vendor documents. As part of specifying the construct-nature, we have identified construct entity type and construct properties.

This helps in developing the construct items. In order to identify the specific conceptual themes, we have provided sufficient thoughts on attributes or characteristics to these constructs. They include common characteristics, unique characteristics, dimensionality, and stability of the constructs. For example, when a construct is meant for multi-dimensionality, it is important to reflect that in the item/measure generation against each dimension of the construct. Based on these characteristics, we have successfully developed the construct-items during the survey instrument development phase. Lastly, we tried to maintain the distinct definition of the constructs and thus avoided any ambiguity. We made sure the constructs are not subject to more than one interpretation. We also made sure the constructs are not overly technical (Mackenzie et al., 2011).

Based on the guidelines proposed by Mackenzie et al. (2011), this research has established a standard definition of 32 factors/constructs. We have presented 32 factors along with definitions to the experts of this qualitative study. The factors have been ranked based on participants voting. Table 9 shows the top 15 factors according to the rank (number of votes for each factor). We have picked up factors/ variables ranked 1 to 13 in table 10. We have decided to merge numbers #3 and #12 as one variable, as was recommended by several participants. They suggest that numbers #3 and #12 are the same finance area factors. Participants suggested to consider them as one factor. Since most of the participants in the qualitative studies voted for cost-effectiveness (Ranked

3) we decided to use this factor for further research (quantitative study) and exclude number #12 (total cost of ownership).

Here is the finalized list of factors identified based on brainstorming, focus group and individual interview methods (Table 10).

Table 10: Final List of Factors for Use in the Proposed Research Model

Factors	Taxonomy of Factors	Comments
Scalability	Technological	Hadoop has a built-in capability to scale-out storage by expanding the number of nodes.
Data storage and processing	Technological	Compared to traditional data storage systems (DBMS, DW) Hadoop can store and process hundreds of terabytes of data.
Cost-effectiveness	Economic	Cost containment by virtue of holding huge data compared to the cost incurred by conventional data storage systems.
Performance Expectancy	Technological	Performance expectancy in terms of data receiving, data storing, and data processing.
Security and Privacy	Legal	Big data consists of unstructured data most of which come from social media, personal data.
Reliability	Technological	Hadoop maintains reliability by keeping the same copy of data in more than one node.
Data Analytics Capability	Technological	Capability to run robust data mining algorithms (Mahout, MLLib libraries) on top of huge data volume. No scalability and performance issues.
Training and Required Skills	Organizational	Big data technologies are complex and new. Training and Skillset is important.
Flexibility	Technological	Hadoop accommodates both structured and unstructured data; it can collect and store data from heterogeneous sources.

Output Quality	Technological	The capability of Hadoop to maintain valid data that can generate business value
Functionality	Technological	Capability to serve the purpose of Hadoop technology.
Facilitating conditions	Environmental, Organizational	Internal big data infrastructure and external support from vendors are crucial.

The factors that are finalized as part of the qualitative study are consistent with the big data literature. Surbakti et al. (2020) conduct a review of big data literature. The authors report that the organizational aspects theme is studied the most, followed by technological aspects including systems, tools, and technologies. Next, the people theme related to leadership, training, and the skillset is discussed in many articles. The data privacy and security issue are widely discussed. The data quality theme is also dominated by big data literature (Surbakti et al., 2020).

3.6 Developing Research Model

This section first provides the descriptions of the top 12 factors selected by experts that participated in the brainstorming sessions, focus group sessions, and individual interviews as part of the qualitative study of this research. The participants provided the justifications listed below for the factors they have selected:

Scalability: The capability of software and hardware is to handle the increase in workload in terms of bandwidth and data volume. A software scalable with it can scale in users and functionality. Hadoop provides a scale-out storage system and can be

expanded by adding nodes and commodity servers as needed. One of the participants suggested that scalability is a big factor in big data adoption. It offers horizontal scaling rather than vertical scaling; hence old hardware does not become obsolete all of a sudden. Another participant pointed out that scalability is the basic advantage provided by a big data system when compared to traditional technology.

Data Storage and Processing Capability: Compared to traditional data storage systems (i.e., conventional databases) the Hadoop can store and process hundreds of terabytes of data using MapReduce/ Spark. One of the participants commented that the ability to ingest anything is an important key feature of any big data Hadoop system. Another participant mentioned that big data technologies are very cost-effective for Big data storage and processing with relative ease. Another participant pointed out that the advantage of a big data system is to provide relatively huge storage.

Cost effectiveness: The capability of a technology that is considered effective and productive compared to its costs. Cost containment and cost advantage are by virtue of open source software and vendor support considerations. One of the participants suggested that most big data technologies are based on open source and thus are very cost-effective to start implementing in Business. Economists suggest that new technology plays a significant role in cost growth but, they observe that it brings benefits as well (Hodgson, 2011). Kohli et al. (2012) suggest IT investments need to be

made based on whether there are contributions to the firm's market value. They also suggest that a firm's market value needs to be measured through accounting measures.

Performance Expectancy/Usability: Performance expectancy is related to the degree a technology is effective in its use. One of the participants of the qualitative study point out that with big data technology, simple queries with the Hive tool and faster results with Impala are a necessity.

Security and Privacy Considerations: The security and privacy considerations are essential to keeping the data with confidentiality, no vulnerability, and no security breaches by hackers (Menon & Sarkar, 2016; Tsai et al., 2015; Wu et al., 2017). One of the participants at a healthcare company mentioned that in the health care setting, security and privacy is a big deal. Another participant mentioned that it is important that big data technology should be able to protect sensitive data.

Reliability: Big data tools and technologies provide greater reliability as the same copy of data stored in more than one node. One of the participants pointed out that being able to maintain data with consistency is important. Wang and Zhang (2018) propose software reliability prediction using a data-driven method, deep learning model. The authors report their proposed model has better prediction performance.

Data Analytics Capability: This category is the ability to discover patterns from a large data set or from incoming streaming data. It involves the prospect of running robust data mining against a complete set of data stored in HDFS with machine learning libraries (e.g., Mahout and MLlib). One of the participants observed that this is where most of the BI/Analytics is going. Another participant pointed out that Hadoop has the ability to apply ML on big data instead of worrying about data size and performance.

Training and Required Skills: This category is the training and skills needed to develop a capability or use technology. Big data is managed through a set of new technologies and hence, training and required skills are important (McAfee and Brynjolfsson, 2012). One of the participants mentioned that the ability to retrain the developer community is a critical aspect for any organization to adopt any new technology. Another participant asks if this is going to be a niche product or is there is enough overlap with existing technology that ramp-up time would be shorter.

Flexibility: Big data tools and technologies provide greater flexibility to extract, process, and load data from many different sources, both structured and unstructured. One of the participants pointed out that big data technologies are open source and developed with flexibility in mind. Due to this feature, it can be adjusted to newer technology, and hence lockdown in any particular technology is not needed. Another participant suggested considering whether technology can be used and/or switched out seamlessly.

Yet another participant pointed out that any new tool needs to interface with the existing ecosystem, hence the flexibility of new technology is key for broader adoption.

Output Quality: The output quality is the competence of the system in maintaining the quality of corporate data. Extant literature suggests that there is a significant relationship between system quality and output quality (Wixom et al., 2001). One of the participants pointed out that output quality is an essential and basic expectation.

Organizations take output quality seriously to make sure they are providing an accurate picture of performance to decision-makers (Lederer et al., 2000). When it comes to financial reporting, accurate numbers are very important, and in some cases, it has implications of external reporting and SOX audit regulations. We can expect firms most likely to adopt the Hadoop technology are those that perceive it ensures output quality.

Functionality: The more a tool provides the capability to perform the job it is intended for, the more likely it will be accepted by users. Some organizations claim that Hadoop meets or exceeds functionality from a data management standpoint, and hence, they will likely use Hadoop for data management and data analytics purposes. Hence, we hypothesize that 'functionality' is positively related to 'perceived usefulness.

Facilitating Conditions: Facilitating conditions are "the control beliefs relating to resource factors such as time and money and IT compatibility issues that may constrain

usage” (Taylor & Todd, 1995). Facilitating conditions include external, organizational, and technical infrastructure support to help undertake big data projects.

3.7 Proposed Research Model

Based on the qualitative studies, we have come up with 12 factors for further study. We also have core constructs of the TAM, PU, PEOU, BI, and AU, by default in our research model.

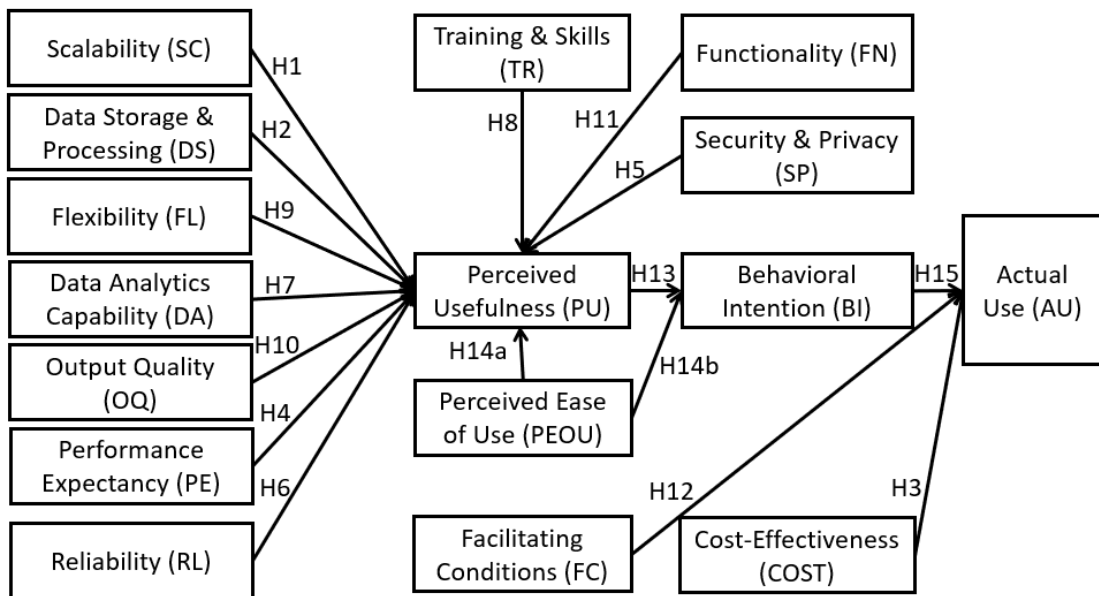


Figure 2: Proposed Research Model

The research model (Figure 2) is primarily based on Davis’ (1989; 1993) technology acceptance model (TAM) which includes factors such as perceived usefulness (PU), perceived ease of use (PEOU), behavioral intention (BI), and actual use (AU). One key aspect of TAM is that it provides a framework to examine the influence of external

factors on the usage of a system (Davis, 1989). The TAM is frequently used to examine the usage behavior of a system from an individual perspective. This research uses this model to examine the usage behavior from an organizational context. In this model, 12 antecedent factors have been selected through an extensive qualitative study (as discussed in sections 3.2 – 3.4 in Chapter 3). Among these factors a few of them were tested in past empirical research: output quality (Venkatesh & Davis, 2000; Wixom et al., 2001), facilitating conditions (Kwon et al., 2014; Ramamurthy et al., 2008; Taylor & Todd, 1995), and performance expectancy (Venkatesh, 2000). The research has incorporated nine new factors including scalability, data storage and processing, flexibility, data analytics capability, reliability, security and privacy, training and skills, functionality, and cost -effectiveness. Successful testing of the influence of these factors on TAM is expected to contribute to the body of knowledge. These factors are related to five characteristics of big data. For example, volume and velocity (data storage and processing), variety (flexibility), veracity (output quality), and value (cost -effectiveness). Big data technology and ecosystem tools have been built based on its five characteristics.

Since this model is built based on 12 factors that are selected out of 32 factors this research would like to validate these factors through survey data. This research uses the structural equal model (SEM) which allows for factor analysis and performance of other statistical analysis to understand which factor and items under each factor will be

influential (Bagozzi & Yi, 1988). This statistical analysis can be used to identify the desired factors. Hence, we develop hypotheses in the next section.

3.8 Developing Research Hypotheses

In order to evaluate the research model, the outcome of hypotheses tests must be informative. The results of a hypothesis tests need to draw correct conclusions about the population. “If the model is truly a good model in terms of its level of fit in the population, we wish to avoid concluding that the model is a bad one. Alternatively, if the model is truly a bad one, we wish to avoid concluding that it is a good one” (MacCallum et al., 1996). Based on the proposed research model we have developed the following hypotheses against each construct. The measures from previous studies are incorporated to reflect the big data context in this study. There are several new constructs and measures developed as well (See Appendix A).

3.8.1 Hypothesis H1 - Scalability

Most of the traditional relational databases lack scalability in dealing with hundreds of terabytes of data. In big data, new NoSQL technologies emerged to provide performance and scalability (Lourenco et al., 2015). Research findings revealed one of the technological challenges to the adoption of big data analytics is performance and scalability (Malaka & Brown, 2015). Big data technologies are scalable in terms of storage, data processing, and building robust machine learning model. Big data pioneer

companies like Facebook choose Hadoop and HBase for availability, tolerance, and scalability reasons (Borthakur et al., 2011). Hence,

***Hypothesis H1:** Scalability in terms of Hadoop scale-out-storage system has a positive effect on perceived usefulness.*

3.8.2 Hypothesis H2 - Data Storage & Processing

Hadoop is considered highly scalable in terms of storage and data processing. “By distributing storage and computation across many servers, the resource can grow with demand while remaining economical at every size” (Shvachko et al., 2010, p. 1).

Traditional databases are not capable to handle hundreds of terabytes of data and are also not scalable. It is worth checking if Hadoop’s storage capacity and data processing capability are related to big data acceptance. Hence,

***Hypothesis H2:** Data storage and processing have a positive effect on perceived usefulness.*

3.8.3 Hypothesis H3 - Cost Effectiveness

Several case studies results show that big data applications have made organizations avoid the cost. Balac et al. (2013) developed a predictive analytics model for real-time energy management using the Time Series approach. Their model is destined to realize tangible improvements in energy efficiency and cost reductions (Balac et al., 2013).

Bologa et al. (2010) report that big data has made it possible to detect insurance fraud within a reasonable time. They point out that in the past, in many cases, insurance fraud

detection was not considered efficient due to the cost and duration of the investigation were very high. The author provides analysis methods for detecting fraud in health insurance. (Bologa et al., 2010). Villars et al. (2011) state that timeliness of the response using big data helped in eliminating the legal and financial costs associated with fund recovery. One of the big data characteristics is that its tools and technology can hold a large volume of data with minimal cost. This allows for analyzing almost all data rather than a small subset or sample (Cao et al., 2015). Srinivasan and Arunasalam (2013) reported that their big data application was able to detect claim anomalies to identify hidden cost overruns of health insurers. Russom (2013) and Hartmann et al. (2014) also report cost containment and cost advantage by using big data technologies.

Roger (1983) asserts that the less expensive the technology, the greater the possibility that it will be adopted. The cost of technology is associated with the benefit achieved. For small companies, the cost might be a major barrier to procure innovation (Premkumar & Potter, 1995). Firms that perceive the cost of big data Hadoop to be high might not adopt it. On the other hand, the medium and large companies might not perceive the cost as a barrier. Hence,

Hypothesis H3: Cost effectiveness is positively related to actual use of Hadoop.

3.8.4 Hypothesis H4 - Performance Expectancy

The performance of the technology is a pivotal factor for technology acceptance.

Successful innovations cannot take place without reasonable performance expectancy.

If technology has the necessary performance capability it would be perceived as useful.

Hence,

Hypothesis H4: Performance Expectancy is positively related to perceived usefulness of Hadoop.

3.8.5 Hypothesis H5 - Security and Privacy Considerations

Big data are mostly unstructured and come from many places including health care.

Security and privacy concerns are getting attention these days (Jain et al., 2016; Tsai et al., 2015). Data breach gets news headlines quite often. User's private information gets into the hands of hackers. Companies are subject to spending millions of dollars to compensate for such data breaches. Hence,

Hypothesis H5: Security and Privacy is positively related to perceived usefulness of Hadoop.

3.8.6 Hypothesis H6 - Reliability

Reliability is the degree to which the new technology is perceived to be dependable by the users. Organizations adopt new technology to overcome the unreliability, deficiencies, or to embark onto new generation tools and technologies to achieve reliability and efficiency. Before accepting any tools or technology users want to be sure that it is reliable and able to show proof that spending money on it is worth it. Hence,

Hypothesis H6: Reliability is positively related to perceived usefulness of Hadoop.

3.8.7 Hypothesis H7 - Data Analytics Capability

One key aspect of the Hadoop-based model is data that is stored in the Hadoop distributed file system (HDFS) with no data movement needed to relational database systems. All analytical, data mining and reporting tools will run against HDFS. With Hadoop distributed files system there is a great prospect of running robust data mining against a complete set of data stored in HDFS. Kranjc et al. (2013) developed a capability to mine real-time streams by transforming batch data processing into a real-time stream mining platform. Tsumoto and Hirano (2013) applied clustering data mining rules to a large dataset consisting of ten years of historical data stored in the hospital information system to discover knowledge from massive healthcare claims data. Wu et al. (2014) published a paper titled, "Data Mining with Big Data" in which they propose a big data processing model, from the data mining capabilities standpoint. Chen et al. (2012) listed areas of emerging research in (big) data analytics, especially using machine learning and data mining. Data analytics capability is the driver of today's business operations. Zhang et al. (2019) and Tsai et al. (2015, 2014) provide a detailed framework for big data analytics. This is worth studying. Hence,

Hypothesis H7: Data analytics capability is positively related to perceived usefulness of Hadoop.

3.8.8 Hypothesis H8 - Training and Required Skills

Training and skillset let company developers and knowledge workers use technology effectively and efficiently. This ensures productivity. Hence, we hypothesize,

***Hypothesis H8:** Training and required skills are positively related to perceived usefulness of Hadoop.*

3.8.9 Hypothesis H9 - Flexibility

Big data tools and technologies providing greater flexibility bring data from different sources and store into a single place (i.e., Hadoop HDFS). These sources include traditional data such as transactional data from enterprise resource planning (ERP), new data such as social media, sensor data, email messages, etc. Hadoop can be used for a wide variety of purposes, such as real-time streaming and processing, log processing, developing recommendation systems, building a data warehousing environment, market campaign analysis, and fraud detection (Nemschoff, 2013). Consolidated data into a single platform provides improved data mining and business intelligence capabilities (Rahman & Iverson, 2015). Hence,

***Hypothesis H9:** Hadoop's flexibility to consolidate data from various sources to single place (HDFS) will have a positive effect on perceived usefulness of Hadoop.*

3.8.10 Hypothesis H10 - Output Quality

Data integrity and quality fall under veracity which is one of the five characteristics of big data. New tools are emerging to map out data lineage (Rahman et al., 2014). This

effort is still at the beginning stage. The empirical study by Kwon et al. (2014) suggests that “a firm’s intention for big data analytics can be positively affected by its competence in maintaining the quality of corporate data.” Lu et al. (2014) assert that if big data cannot provide quality decisions due to data veracity, newly mined knowledge will not be convincing to the analytical community. However, big data is also considered to have the capability to improve quality monitoring clinical trials and decreasing spending from patients to the government level. (Nambiar et al., 2013). Hence,

Hypothesis H10: Output Quality are positively related to the perceived usefulness of Hadoop.

3.8.11 Hypothesis H11 - Functionality

Functionality is the aspects of what technology, a product, or a system can do for users. Functionality includes the features of the product or technology. Functionality is the ability of technology to interact as expected by the users. Hadoop is expected to perform certain functions such as access, and to process data from many sources, tools, and devices. Hadoop provides a distributed file system. Hadoop replicates data sets on commodity servers making the process run in parallel. These functionalities beg validation. Hence,

Hypothesis H11: Functionality is positively related to perceived usefulness of Hadoop.

3.8.12 Hypothesis H12 - Facilitation Conditions

"The degree to which an individual believes that an organizational and technical infrastructure exists to support the use of the system" (Venkatesh et al., 2003, p. 453).

Facilitating conditions is considered as one of the key factors in data warehouse architecture selection (Ariyachandra & Watson, 2010). Even though Hadoop is an open-source system there are vendors like Cloudera, Horton Works, and MapR that have come up with customized versions of the system with features that might help companies in using it easily (Villars et al., 2011). These vendors take care of the newer versions of the software as well as customization (Ceci et al., 2019). Some companies might not want to invest resources to customize and make enhancements to this system. In such cases, those companies might be willing to use the technology. Some companies might have internal platform infrastructure teams to maintain it and provides support in initiating projects. We need to see if big data technology acceptance is influenced by facilitating conditions. Hence,

***Hypothesis H12:** Facilitating Conditions have positive effect on actual use of Hadoop.*

3.8.13 Hypothesis H13 - Perceived Usefulness

This factor is the core construct of TAM. It has been tested and validated by prior empirical research. Therefore, the following hypothesis has been developed:

Hypothesis H13: Perceive Usefulness has positive effect on Behavioral Intention in using Hadoop.

3.8.14 Hypothesis H14 - Perceived Ease of Use

This factor is the core construct of TAM. Two other core constructs, perceived usefulness, and behavioral intention have a dependency on this construct. It has been tested and validated by prior empirical research. Therefore, the following two hypotheses have been developed:

Hypothesis H14a: Perceived Ease of Use (PEOU) has positive effect on Perceive Usefulness (PU) in using Hadoop.

Hypothesis H14b: Perceived Ease of Use (PEOU) has positive effect on Behavioral Intention to using Hadoop.

3.8.15 Hypothesis H15 - Behavioral Intention

This factor is the core construct of TAM. The extant literature reveals that behavioral intention is the strongest influencer of the actual use of a system (Davis, 1993; Dillon & Morris, 1996). It has been tested and validated by prior empirical research. This is one of the two constructs that directly influence the actual use of Hadoop. Therefore, the following hypothesis has been developed:

Hypothesis H15: Behavioral Intention (BI) has positive effect on Actual Use of Hadoop.

Chapter 4 Research Methodology

This dissertation consists of distinct studies: qualitative study and quantitative study.

This chapter covers the quantitative studies. Chapter three discussed qualitative studies.

4.1 Research Design

The previous chapters provide details on relevant theories, review of literature, results of qualitative studies, the proposed model, and hypothesis developed. This chapter concentrates on research design relating to data collection, survey instrument development, instrument validation, and survey administration. This research intends to test hypotheses based on the primary data collection method. Data is collected using survey instruments. Survey designs are distinguished in terms of cross-sectional and longitudinal designs (Pinsonneault & Kraemer, 1993). In a cross-sectional design, the population is described at one point in time as opposed to multiple points in time in a longitudinal design. This research conducts cross-sectional design as big data is a new field and it would not be possible to collect adequate responses at multiple points in time.

4.2 Survey Instrument Development

A survey instrument is used to “gather information about the characteristics, actions, or opinions of a large group of people, referred to as a population” (Tanur, 1982). The study attempts to find relationships between variables that might give insight into users’ adoption of big data. As part of the survey, questions are designed to get answers to the

questions asked in relation to each hypothesis. Survey research questions are developed based on previous empirical studies (Davis, 1989; Kwon et al., 2014; Venkatesh et al., 2003) as well as incorporation of new questions relevant to the topic of research. Some of these questions are borrowed from existing theories (Davis, 198; Venkatesh, 2000) and some others are derived from empirical studies (in big data case: Kwon et al., 2014). In this research, survey questions are inherited from several theories and empirical studies (Davis, 1989; Venkatesh et al., 2003). Survey questions are classified into two broad categories: open-ended and closed-ended. Since this research uses a quantitative method of studies the questions being asked are closed-ended. As part of closed-ended questions, Likert's five-point scale is used (Likert, 1932). Likert scale questions consist of 'strongly disagree', 'disagree', 'neutral', 'agree', and 'strongly agree'.

We have studied two prominent publications on construct item development, measurement, and validation. Morgado et al. (2017) classify "item generations" into two categories: deductive and inductive. The deductive method consists of a literature review and scales used by empirical studies. The inductive method could be considered as gathering information using qualitative studies including focus groups, brainstorming, and individual sessions. The researcher might brainstorm items based on real-life experience. By using these methods, we have developed a sizable list of construct measurements. The extant literature suggests 35.2% of studies used deductive methods, 7.6% used inductive methods, and 56.2% used both deductive and inductive approaches to develop construct items (Morgado et al., 2017). Exclusive use of the

deductive method is reported as a limitation of qualitative research (Morgado et al., 2017). Compared to that, this research used both deductive and inductive approaches to generate construct items. One of the limitations in scale development is that items with ambiguity or difficulty in answering are reported to be the main weakness (Morgado et al., 2017). The ultimate goal of construct-items generation is to develop a set of items that sufficiently captures the essential aspects of a construct (Mackenzie et al., 2011; Petter et al., 2007). But we also need to make sure that an item defined under a construct does not belong to another construct. Additionally, we need to ask ourselves why we ask a particular question (in terms of measure). Asking a question in the survey without sufficient reason would be inefficient or non-beneficial in terms of all types of resource usage.

4.3 Instrument Validation Steps

The next step is to assess content validity which plays a big role in finalizing the survey instrument (Morgado et al., 2017). This validity also requires following some methodical steps including the opinion from the expert panel. As part of further theoretical analysis, 74.2% of empirical studies used expert panels while others used the opinions of a subset of target populations (Morgado et al., 2017). Our study use both expert opinions and surveying the target population using a pilot study. By using a pilot study survey, this research use construct validity using the exploratory factor analysis (EFA). This helped identify and remove weak measures and finalize the constructs. As part of psychometric analysis, 86.6% of the studies use EFA for construct validity (Morgado et al., 2017).

Extant literature suggests that multiple studies found 50% of the items got lost as part of item validation steps (Morgado et al., 2017).

Instrument validity is to measure the accuracy of the instrument as much as possible. Instrument validity ensures that data collection reflects the opinions of the population about the subject being studied (Straub, 1989). Instrument validity is typically conducted in three areas: content validity, criterion-related validity, and construct validity.

- Content validity makes sure that the test question does match the content or subject matter that it is intended to measure. Experts in a given domain typically judge the content. Content validity is conducted through the use of an expert panel. This research relies on an expert panel based on big data experts from the industry that has big data platform along with a lot of big data applications. The expert panel provides valuable opinions on the content of the instrument.
- Criterion-related validity measures the validity of the instrument by comparing the outcome of the test with the performance of another test, usually using correlation. Criterion-related validity is used as predictive of later behavior.
- Construct validity measures the underlying theoretical constructs. For example, in big data acceptance cases, if the measures delve more into an application's validity rather than its usefulness or performance then it diverts from the original intent of the test instrument.

This research uses expert panels based on big data user communities. The expert panel makes the judgement on the survey instrument in terms of content validity and the theoretical nature of construct validity. The initial version of the survey instrument is based on the questionnaire used in previous research. Additional questions are added to the questionnaire based on the intent of the subject matter of this study. Then this enhanced instrument was given to the expert panel to validate. Based on expert panel recommendation the instrument was modified and enhanced as necessary.

To conduct survey instrument validation there are two primary areas taken into consideration. The first one is whether each item represents the factors that are being assessed. The second is whether the questions are easy for participants to answer. Table 12 lists the steps to develop and validate the survey instrument.

Table 11: Steps to Validate Survey Instrument

Steps	Description	Outcome
1. Developing the first version based on previous research survey questions	This was developed base on recent survey question for data management software acceptance	Version One
2. Pre-Validate (Read-aloud)	Using a group of users in Industry improvement areas obtained. Expert panel + Individual interviews with total 12 participants.	Version Two
3. Pilot test 1	Test conducted as part of a web-based survey and email sent to a group of Hadoop users. Total 40 participants.	Version Three

Step one in Table 11 talks about using the questions that were used in similar research in this subject. This gives the validity of the research instrument. This also speaks for consistency with previous research in this field (Venkatesh et al., 2003; Venkatesh et al., 2012).

In step two, the version derived from previous literature is presented to a group of experts to comment on the contents in relation to the study being undertaken. The researcher reads -aloud all the questions along with explanations. Based on that, experts provide their thoughts and opinions. Twelve participants from the industry are invited to this session for about one hour. These experts' thoughts and recommendations are reflected in the survey instrument.

In step three, a pilot test is conducted on the instrument developed and modified in step two above. This test involves 40 participants from among Hadoop users in the industry. This pilot test gives another opportunity to improve the survey instrument. Here it is observed as to whether participants understood the questions and also if they express any concerns about the question format and clarity. The survey instrument is improved based on their response to questions and comments made. Sections 4.3.1 to 4.3.3 provide more details of survey instrument validation.

4.3.1 Instrument Validation Phase One

It is important to make sure that the raters of survey questionnaires have sufficient intellectual ability to rate the survey questions (Mackenzie et al., 2011). It is also important that the raters of the survey questionnaire should represent the main population of interest (Anderson and Garbing, 1991; Mackenzie et al., 2011). The number of questions under each construct needs to be reasonable because the raters of questions can distinguish between items only up to about eight to ten aspects (Mackenzie et al., 2011).

A survey testing tool was used in validating the instrument as an example of the survey instrument validation tool. Below is an example of items for one of the constructs (scalability) of the survey instrument of this research (Table 12).

Table 12: Example of Measures from Survey Instrument

A	B	C	D	E
Construct	Items	Item relevance to the Construct: Put in scale 1 to 5	Ease of answering the question: Put in scale 1 to 5	ANY COMMENTS?
1	1. Construct: Scalability			
3	SC1 - Hadoop is scalable to handle hundreds of terabytes of data compared to relational databases.	5	5	Scale could be in PB scale
4	SC2 - Hadoop is scalable in terms of storage and processing.	5	5	
5	SC3 - Hadoop scalability in terms of higher performance (lower latency) can improve the bottom-line of my company.	3	5	Hadoop latency is generally high. Good for batch and not for real-time.
6	SC4 - With the increase of applications, users, and data volume, Hadoop is able to meet extra load by expanding number of nodes.	5	5	
7	SC5 - Hadoop has built-in capability to scale-out storage compared to our company's traditional data storage systems.	5	5	
8	SC6 - Hadoop's scale-out storage system can store and distribute very large data sets across hundreds of inexpensive servers that operate in parallel.	5	5	
9	SC7 - Hadoop enables my company's businesses to run applications on thousands of nodes involving thousands of terabytes of data.	5	5	
10	SC8 - Hadoop can expand incrementally without having to change the applications and without the users noticing any degradation in performance.	5	5	
11	SC9 - Propose any important item that is missing:			
12	2. Construct: Data Storage and Processing			
13	DS1 - Hadoop system is capable to store and process huge volume of data.	5	5	
14	DS2 - Hadoop is capable to receive and process streaming data real-time.	4	5	
15	DS3 - Hadoop is capable to run analytics on a very large data set.	5	5	
16	DS4 - Hadoop's processing Engine is capable to process both structured and unstructured data.	5	5	
17	DS5 - Hadoop allows for write-once and read many times with its Processing Engine.	5	5	
18	DS6 - Hadoop's processing engine removes the requirement of data summarization which was needed in traditional database systems.	2	2	
19	DS7 - Hadoop's stored data can easily be accessed, used and processed by applications and services.	4	5	In some insatnces, it may not be easy to integrate with some services.
20	DS8 - Hadoop's storage and processing engine can serve all application needs - analytics, processing, machine learning.	5	5	

We have completed Phase 1 of Construct Validation. A total of 32 people had been invited via group meetings or individual invitations. Twelve participants filled out the spreadsheet template with a score for relevance of the construct-items and ease of answering questions. Three participants made comments only and did not score items. A total of 17 people did not accept the invitation. A handful of participants made comments about many of the construct items. We have compiled them and adjusted the questionnaire as appropriate. The participants have not proposed any new item even though they were encouraged to do so.

Based on their comments and suggestions, we were able to remove 18 construct-items from the 12 constructs (which are independent variables). Based on their comments and suggestions, we have also modified several construct items to make them meaningful and easy to understand. One of the participants (who is a professor and an expert in TAM) suggested that we remove “I” and “me” from the item tests and use “my organization” instead since this study is an organizational level study as opposed to an individual -level study. We have made this correction.

After all the fixes, modifications, and adjustments, we still have a total of 79 construct-items with 59 items under the first 12 constructs (Independent variables) and another 20 construct -items (Construct 13 – 16) which are part of the original TAM model. In regard to the first 12 constructs (IV) items, our plan is to bring the number of items down to 4 under each construct via the second round of instrument validation – the pilot test. This is to make sure the instrument is not too long.

Comments from the Respondents as part of instrument validation phase-one:

1. Asked for the meaning of certain keywords
2. Suggested to re-write certain item to make it meaningful
3. Hadoop latency is generally high. Good for batch, not for real-time
4. Hadoop is opensource. But if we depend on vendor then we have to pay
5. Hadoop security is very robust but may not be easy to manage
6. Remove references from the items
7. One question conflicting with other items

8. Rewrite some items to switch from negative (telling disadvantages) to positive contexts.
9. All questions against each variable should start from the same word, use the same tense in questions like past or present or future, don't mix up all. Also, there should be either positive items or negative items, do not mix both and put against each variable.
10. Hadoop, due to the learning curve may not appear cost-effective in the early days of adoption, with the exception of storage cost.
11. Hadoop needs different thinking so training will help with learning curves & change in thinking.
12. The interesting question from a survey perspective is the relevance of these different functions and features of Hadoop to the respondent's bottom line. I'm going to evaluate these questions from that perspective.
13. Can you make this more concrete in order to make it easier to answer?
14. The question seems redundant.
15. A highly technical question that managers won't be able to answer without consulting someone.
16. Seems vague - how much is 'huge'?
17. A complex question to answer.
18. It is not possible to say that, not all apps can use HDFS and MapReduce
19. FN1 - Hadoop system is robust to deal with data" Comment: "not all data"
20. Change from "me" to "my organization"

21. On BI questions.... Not clear, we are already using it for 4 years.

22. some questions seem to be repetitive, is it purposely to verify users' responses each time?

4.3.2 Instrument Validation Phase Two

We conduct a pilot test using Qualtrics survey tool to collect the data. The goal was to collect 15 to 20 responses, but we ended up collecting 40 responses. Many researchers typically use university graduate students to form such an expert panel but since big data discipline is a specialized field, graduate students would most likely not have sufficient knowledge and expertise to be part of the expert panel. To validate the survey instrument for this research we have invited about 70 people who worked in big data domains and have sufficient knowledge and experience in big data tools and technologies, and also on conventional database systems. The criteria suggested in choosing experts are that they have knowledge and experience in the domain and diversity of knowledge in different areas of the subject matter. For example, in big data field, experts could be selected from among developers, systems analysts, application users, platform engineers, project managers, data scientists, and business managers. The meeting type of the expert panel will be an online meeting so participants from different geographical locations can attend the meeting. Based on expert opinions on the survey instrument it has been modified and/or enhanced per recommendation. A pilot test has been conducted among a small group of Hadoop users to test and evaluate the performance of the survey instrument. The pilot test was conducted using a web

survey tool. Based on the outcome of the pilot test, the survey instrument has been modified and enhanced again as appropriate.

4.3.3 Pilot Test Results

We are able to run data using SPSS. The result that the tool generated was not meaningful because a full-length survey instrument (which has 16 constructs including latent variables) with 79 construct items, a large number of survey participants are needed to have statistical packages generate reliable results. We had 40 respondents participate in the pilot survey and out of that, we found 33 responses valid and 7 responses rejected due to incompleteness. The SPSS factor analysis is conducted against the items of each individual construct to identify and remove weak items. By using this process, we are able to identify 4 items for the majority of the constructs and 3 items for the remaining few constructs. With that, we have 62 items under 16 constructs to keep and we removed 17 items as part of this Pilot Test of Survey instruments. The Pilot test survey was conducted via Qualtrics web-based tool (Appendix B).

4.4 Instrument Reliability

Instrument reliability is checked to make sure consistent results are achieved upon repeated applications. Different types of reliability tests are conducted (Research Rundowns, 2018): subject reliability (the ability of the research subject or persons interviewed), observer/ interviewer reliability (abilities of the interviewer), test-retest reliability (consistency of a measure tested over time (in a short time) – measurement

by the same observer/interviewer) (Hendrickson et al., 1993), and internal consistency reliability (consistency of results across items – typically measured using Cronbach's Alpha) (Mackenzie et al., 2011), and instrument reliability (poorly worded questions).

Instrument validity and reliability are inter-related. Instrument validity is a precursor to instrument reliability. A survey instrument needs to be both valid and reliable. A test might be reliable but not valid for the subject of the study. In that case, instrument reliability is not enough. Thus, instrument validity is more important than instrument reliability. In this research, instrument reliability is measured through average variance extracted (AVE), composite reliability (CR), and Cronbach's alpha (Cronbach, 1951).

4.5 Instrument Administration

There are two main types of survey administrations which include structured interviews and self-completion questionnaires. In self-completion questionnaires supervised, postal, email, and web-based online surveys are typically conducted. Web-based surveys are used frequently in IT research because they are easy to communicate, cheaper, and can be sent to a large group of people faster. The barrier to the distant location of participants is not an issue. This dissertation uses a web-based survey method.

In order to facilitate a web-based survey, Portland State University (PSU) has provided a tool and platform called portlandstate.qualtrics.com. For this dissertation, the web-based survey was conducted using Qualtrics (an industry survey tool). Emails

were sent to Hadoop user groups in the United States. with a link to the Qualtrics survey. After initial email invitation reminders, two follow-up emails were sent to the participants.

4.6 Sampling Strategy

In determining a sampling strategy several important considerations need to be made. They include defining a population, establishing the sampling frame, selecting a specific sampling type, and determination of sampling size (probability sampling). There are five steps required to frame sampling strategies which include determining target population, defining a sampling frame, outlining a sampling method, determining the sampling size, and drawing actual sampling (Anderson, 2012).

4.6.1 Sampling Methods

There are four major types of sampling methods found in the literature which include simple random sampling, stratified random sampling, cluster random sampling, and systematic random sampling (Luck and Rubin, 1987). Thus, cluster sampling is considered one of the established sampling methods. In cluster sampling, the population is divided into separate groups. A simple random sample of clusters is selected from different population groups. These groups or clusters need to be homogenous in nature and heterogeneous elements within each group. Each cluster should have distinct subpopulations. The "effective clusters are those that are heterogeneous within and homogenous across" (Lavrakas, 2008).

This research takes advantage of cluster sampling since Hadoop users are already organized in different Hadoop user groups. Hence, the clusters of Hadoop user groups are readily available. There are 21 Hadoop user groups found online, out of which 14 user groups are found active. And out of 14 user groups, two user groups or clusters are randomly selected. This allows sending survey instruments to 10,500 users under two user groups. That means the sample consists of every member of these two Hadoop user groups. Thus, clusters are supposed to reflect the whole population.

In this research, one cluster or Hadoop user group was based in the Bay area which has business importance. The Bay Area is historically an important financial and business center since the late last century. Business activities in this place attract all types of industries. The other cluster or Hadoop user group consists of the users located in the New York area. The New York user group has historical business importance with big companies currently in this area.

4.6.2 Targeted Population

The objective of this dissertation is to study organizations' in the United States that use big data technology, Hadoop. There are no exact statistics as to how many small, medium, and large organizations in the United States use big data. However, the most recent survey suggests that "Big data adoption reached 53% in 2017 for all companies interviewed, up from 17% in 2015, with telecom and financial services leading early adopters" (Columbus, 2017). Since there is no publicly available list of big data user companies this research will use big data user groups available on the Internet to

conduct the survey. Using the user groups as intended users is consistent with the literature that suggests that information technology needs to be accepted by intended users as opposed to “procurers” (Dillon & Morris, 1996). There are 14 active Hadoop user groups in the United States found in the Apache Org Wiki site (HadoopUserGroups, 2019). There are close to 33,000 users belonging to these 14 Hadoop user groups. Selecting all these 33,000 users will be a large number and a poor response might cause a big non-response bias issue. The research will work on two user groups called, ‘Bay Area Hadoop User Group’ and ‘New York group’. These groups consist of 10,500 users.

4.6.3 Sampling Frame

There are 21 Hadoop User Groups found in the Hadoop Wiki site maintained by the Apache Organization (HadoopUserGroups, 2019). Out of 21 sites, only 14 user groups are found to be reachable via the web. Each of these user groups has a few hundred to several thousand members. It is not possible to know what percentage of those users are active in group activities or read user group communication messages. Due to the uncertainty of determining the actual number of active users, we made a decision to limit the sampling frame to members of two user groups or clusters which have been randomly selected. One user group is called ‘Bay Area Hadoop User Group’. This group has 6,440 members. For this user group, there is only one email group. This means that this research has 1 user group’s email group address as opposed to individual email addresses of 6,440 users. The positive side is that no significant time or cost overhead was involved in sending communications to those 6,440 users via 1 user group email

address. We also used a NY-based Hadoop user group with about 4,060 users. These two sites, one on the west coast and the other on the east coast, speaks for two prominent groups. These two places have business significance. These two randomly selected cluster sampling groups with homogeneity among groups and heterogeneity among the elements in each cluster make the sample frame representative of the continental United States Hadoop users.

4.6.4 Sample Size

The sample determination needs to make sure it has adequate power to conduct planned hypothesis tests about model fit. The sample size N needs to have adequate power to detect when hypotheses are false (MacCallum et al., 1996). A sample that is large enough tends to impact time, money, and other resources. A researcher needs to make the trade-off in specifying a sample size. If the sample is too few that might cause the risk of sampling error and hence, not tolerable. On the other hand, if the sample size is too large that could increase the cost of research which might not be affordable but is helpful in reducing the sampling error (Luck & Rubin, 1987).

The tolerable error is the value which is a deviation between the sample estimate and the population parameter that a researcher or decision-maker is willing to accept. The level of confidence in the value that the researcher desires in the sample estimate being within the tolerable error of the population parameter. For example, in social science research the researcher tries to determine the average income of families in a city or community and in that process, the researcher decides that a +/- \$1,500

deviation between the sample mean and true population means is okay and can be accepted with 95% confidence. Determination of Z value (e.g., 1.96) is associated with the desired confidence level specified (in this case 95%). Estimating the standard deviation of the population is based on the standard deviation of the sample being derived using a pilot study or from a previous study comparable to the proposed study.

For determining a sampling size, some general procedures are being followed. They include determining the tolerable error, determining the level of confidence, determining the z value, estimating the standard deviation of the population, using the appropriate statistical formula, and drawing the appropriate sample (Luck & Rubin, 1987).

Formulas are available in selecting an appropriate sample size. The National Education Association has published a formula to determine the sample size for categorical variables (Krejcie & Morgan, 1970):

$$s = \chi^2 \frac{NP(1 - P)}{d^2(N - 1)} + \chi^2 P(1 - P)$$

... where χ^2 is the table value of chi-square for 1 degree of freedom at the desired confidence level ($1.96 * 1.96 = 3.8416$), N = the population size, P = the population proportion (assumed as 50% for maximum sample size), and d = the degree of accuracy expressed as a proportion (typically, .05) (Krejcie & Morgan, 1970).

Another convenient computational formula in determining the sample size n is provided below (Luck & Rubin, 1987):

$$n = \left(\frac{ZS}{e} \right)^2$$

... where e is the tolerable error, Z value is associated with the degree of confidence selected (e.g., 1.645, 1.96, or 2.58 for confidence levels of 90%, 95% or 99% respectively), and s is the sample standard deviation.

So, the tolerable error increase or decrease determines the sample size. The tolerable error selection depends on the sensitivity of the decision outcome. From the above example, a tolerable error of +/- \$1,500 along with the standard deviation of the sample s (\$19,500) will get us a sample size of 649 with a 95% confidence level. But if the researcher or decision -maker is sensitive to the decision outcome and hence wants to stay close to the true population mean by decreasing the tolerable error to +/- \$1,000 in that case sample size would increase to 1460 with a 95% confidence level. On the other hand, if the researcher or decision-maker is a bit less sensitive to the decision outcome and chooses the tolerable error to the range of +/- \$2,000 in that case the sample size needed would decrease drastically to 365 with 95% confidence level.

Now, by leaving both the tolerable error (e = +/- \$1,500) and the sample standard deviation (s = \$19,500) constant if we try sample size determination with different confidence levels, we also get varied sample sizes. With a 90% confidence level the sample size is calculated 457 which means less costly research but with a lowered confidence level. On the other hand, we can get sample sizes of 649 and 1,124 with confidence levels of 95% and 99% respectively. This means that to be more accurate and

confident it requires us to increase the sample size to 1,124. A confidence level of 95% means that there is a 5% risk of true population statistic (mean) to be outside the range of tolerable errors specified.

In sample size determination, the measurement type of variables needs to be taken into consideration. If a categorical variable (e.g., gender, education level) is used as the basis of sample size then sample size needs to be larger compared to a seven-point scale used to measure the continuous variable (Bartlett et al., 2001). In sample size determination two factors need to be taken into consideration: margin of error and alpha level. Cochran (1977) points out that if “the true margin of error exceeds the acceptable margin of error; i.e., the probability that differences revealed by the statistical analyses really do not exist” (Bartlett et al., 2001) then the decision is subject to Type I error (also known as alpha error). In other words, when the statistical analysis reports a difference between the sample estimate and true population parameter exists but actually it does not, in that case it is a Type I error. On the other hand, a Type II error (also known as beta error) occurs when statistical procedures report that a difference between a sample estimate and population parameter does not exist but actually, it does exist (Bartlett et al., 2001).

Sample size calculators are available on the web to determine the sample size. One of them is Raosoft® (Anderson, 2012). Users need to provide input, a margin of error number (e.g., 5%), confidence level (typically, 90%, 95%, or 99%), a population size (if unknown, put 200,000), and response distribution (typically, 50%) (Anderson, 2012).

We use a web survey tool, Qualtrics, as it is available to all PSU students for use (Anderson, 2012). The sample size calculator, Raosoft®, provides an estimate of the required sample size (responses) of 371 for the population size of 10,500 (Anderson, 2012).

Since that we use a web-based survey there is no cost-increase and hence it should not influence our sample size determination. One factor we need to be mindful of is to obtaining data with greater precision of the population statistics with the sample size.

For this research, we use structural equation modeling (SEM) statistical software. The SEM is a statistical modeling technique used to perform confirmatory factor analysis, and regression or path analysis with a graphical interface (Hox & Bechger, 1998). In SEM, the model specification is guided by theories and prior empirical study results (Hox & Bechger, 1998). It is widely used in behavioral science research. There is a dedicated journal titled, 'Structural Equation Modeling: A Multidisciplinary Journal' available that publishes research findings on SEM.

There is a collection of thought, opinions, and conflicting suggestions about sample size determination. This puts new researchers in a tough spot. Several researchers suggested a different sample size for data analysis using SEM (Bentler & Chou, 1987; Hair et al., 2010; Kline, 2015; McQuitty, 2004; Suhr, 2006). McQuitty (2004) suggests that in the SEM program minimum sample size N should never be less than 100. Some other researchers have suggested a thumb rule which consists of a ratio of

20:1 for the number of samples to the number of model parameters (Hair et al., 2010). Suhr (2006) reports that 10:1 might be a realistic target. On the other hand, Bentler and Chou (1987) suggest that, "if the ratio is less than 5:1, the estimates may be unstable." Chin (1998) and Chin and Newsted (1999) suggested having at least 10 responses for each indicator (item) to derive an appropriate sample size.

Boomsma (1982) and Marsh and Bailey (1991) suggest using the ratio (r) of indicators based on P , for indicator variables, and K , for the latent variables. In this case, if $r = 3$ then a minimum sample size of 200 will be required. And when $r = 2$ the sample size needs to be 400 (Ding et al., 1995; Marsh et. al., 1998).

Mulaik et al. (1989) and Pui-Wa et al. (2004) suggested to maintaining at least 200 sample size. Barrett (2007) takes a strong position about sample size for the SEM model by stating that, "SEM analysis based upon samples of less than 200 should simply be rejected outright for publication unless the population from which sample is hypothesized to be drawn is itself small or restricted in size."

This research takes two factors into consideration to come up with reasonable and reliable sample sizes: the use of a sample calculator, and prior research guideline that suggests a reasonable sample needed for a reliable sample for use in structural equation modeling (SEM). First, this research puts the population size of 10, 500 into a sample size calculator (Anderson, 2012). This tool calculated the sample size (required response) of 371 since the members of the online user groups are not active in 100% of the cases. Hence, the sample size calculator's guidelines about sample size cannot be

taken as a rigid sample size. Our survey response size is 349 which is 22 less than the suggested sample size of 371. Hence, the responses of 349 received by this survey could be considered a reasonable size. Second, prior research suggests for data analysis using SEM a minimum sample size of 200 is needed (Barrett, 2007; Mulaik et al., 1989, Pui-Wa et al. 2004). In our case, we have collected and validated a survey response size of 349. Hence, we assume that this is a reasonable sample size. Chapters five and six in this dissertation discuss statistical results based on this sample size.

In quantitative research design, addressing the issue of determining sample size and response bias is essential (Bartlett et al., 2001). A low response rate leads a researcher to a serious problem, which is referred to as a nonresponse error (Luck and Rubin, 1987). The sample might not reflect the population. The concern is that those who have responded might be different from those who did not respond. This is an instance in which the bias from nonresponse emerges. To explain according to the current research, sending survey questions to two Hadoop user groups consisting of 10,500 respondents, and receiving a much lower response might cause nonresponse bias. In the mail survey, nonresponse can result from two sources: cannot locate or reach and refusal to respond. In the case of a web-based survey, the contact email address might have become invalid, the respondent might be busy and hence could not respond, or the respondent is not willing to participate due to lack of time or privacy concerns. To address the non-response bias issue we conduct web analysis, that is,

comparing respondents who participated in the survey after the initial invitation to attend the survey, the first reminder, and the second reminder.

4.6.5 Approaches to Increase Sample Size

Cochran (1977) suggests that one way to attain the target sample size is based on variance estimation. The author proposes taking samples in two steps. By using the results of the first step in terms of variance, a determination could be made as to how many additional responses are needed to achieve the desired sample size. One advantage of this approach is that there is no need to send surveys to a large number of respondents (avoid oversampling). This could help in reducing nonresponse bias which has the most impact in a web-based survey. Bartlett et al. (2001) argue that caution should be used in “raising the sample size above the level indicated by the sample size formula” as it might increase the probability of Type I error.

Besides the oversampling technique, a variety of ways have been proposed to increase the survey response rate. First is an advanced letter informing the respondents that a questionnaire will be on the way very soon and requesting their cooperation. This is reported to have increased the response rate (Luck & Rubin, 1987). Another option is to write a cover letter with the assurance of anonymity or strictly maintaining the confidentiality in dealing with the sensitive issues helps in increasing the response rate (Luck & Rubin, 1987). Also known to be effective is designing the survey with an appropriate survey length. Additionally, it is best to contact participants multiple times, and finally, get the survey pre-tested (Monroe & Adams, 2012). Since low response rates

have continued to be an issue with surveys, as part of sample size increase efforts, this web-based survey research follows these approaches.

Appropriate Length of Survey and Pre-Test: We first design a good survey that is unambiguous, easy to fill out, and be able to be finished in 20 minutes. A well-designed survey that is easy to complete helps in improving response rates and data accuracy. We conduct a pre-test to make sure it is effective. We carefully evaluate pre-test responses and accommodate any reasonable improvement suggestions. This approach was found to be very effective (Dillman et al., 2009; Monroe & Adams, 2012).

Writing Advance Letter: Writing an advance letter to the respondents that a survey to be sent to them very soon. We highlight that the survey response will be used for Ph.D. research purpose only.

Cover Letter and Contacting Participants Multiple Times: We write a strong cover letter by reiterating the importance of this survey and stating that it is intended to be used for Ph.D. work. We hope that participants take it as part of their social responsibility. We assure them that their response will be kept anonymous and contents would be kept strictly confidential. Writing a personalized cover letter has been reported to be helpful in increasing response rates (Atif et al., 2012; Monroe & Adams, 2012).

4.6.6 Approaches to Address Concern with Low Responses

Not getting enough responses per required sample size of a research design is unfortunate for the researcher. Low response rates to a survey cause the sample from which data is collected to be unrepresentative and subject to the existence of bias due to non-response. In such cases, “external validity of the instrument is threatened” (Atif et al., 2012), and making valid conclusions from the data becomes challenging. Extant literature suggests certain measures to address the concern of low response rates.

Late Response Evaluation to Address Non-Response: Armstrong and Overton (1977) report that the most commonly recommended protection against nonresponse bias has been the reduction of nonresponse itself. To address the low response issue, we conduct analysis between different response webs, response to initial invitation, first reminder, and the second reminder. In that case, late respondents could be used as a “proxy for non-respondents in estimating non-response bias” (Atif et al., 2012). These different rounds of response results are compared and checked with the first set of responses to see if the second and third web of responses differs from the first set of responses. This approach checks if late respondents resemble non-respondents. We used this technique in this research. Accordingly, we conduct responses-web analysis using the ANOVA technique in SPSS.

Exclude Unacceptable Measures from the Model: Due to the low response rate if the model fit is found unacceptable measures need to be taken to revise the model when it is meaningful (Suhr, 2006). This research investigates which construct measures are responsible for lack of model fit and whether they could be excluded from the measurement model. We have successfully improved the estimates and model fits by removing poorly performing measures as well as construct. This approach has been practiced by SEM researchers and supported by Anderson and Gerbing (1988).

Commonality Analysis: To address the concern of low response rate all statistical numbers need to be evaluated. MacCallum et al. (1999) assert that the necessary sample size of a given study is dependent on several aspects including the level of commonality of the variables and the level of over-determination factors. An effort could be made to perform commonality analysis which helps to identify the variance of each of the independent variables as to how they are accounted for in a dependent variable. MacCallum et al. (1999) report that as commonalities increase, quality of factor analysis solutions increase and the role of sample size on quality solutions decline. In other words, when commonalities are high (greater than .5) the sample size has little impact on quality solutions. This research evaluates the commonality analysis.

Check the SEM Fit Statistics: The SEM consists of several fit indices out of which the χ^2 is considered the only inferential statistic. Researchers use many descriptive

statistics, hence, in general, rules-of-thumb are applied to assess goodness-of-fit (Iacobucci, 2010). In regard to χ^2 , it is sensitive to sample size (Gerbing & Anderson, 1985) and indicates a poor fit even with modest sample size. Hence, experts in this field suggest, “with some consensus in the psychometric literature, that a model demonstrates reasonable fit if the statistic adjusted by its degrees of freedom does not exceed 3.0: $\chi^2 / df \leq 3$ ” (Kline, 2015; Iacobucci, 2010).

In evaluating the fit statistics Marsh et al. (2004) suggest to not taking the rules-of-thumb very literally. The authors also suggest not to be too much concerned with χ^2 as it simply “will not fit if the sample size is 50 or more.” Further, they suggest seeing if χ^2/df is about 3 or under; to avoid being overly critical if the CFI is not quite .95. On the other hand, Iacobucci (2010) suggests concentrating on asking good theoretical questions as to whether the hypothesized link logically makes sense, and if they are sound, the comprehensive yet parsimonious and a compelling theoretical story exists for the overall model (Iacobucci, 2010).

4.6.7 Survey Administration

After a web-based survey instrument is finalized via Qualtrics an email message with a survey link will be sent to two Hadoop user groups that we have selected. In the cover letter, it will be called out that it would really help in doctoral research if Hadoop user group members respond to our request. We call out that the survey would not be time-consuming as it was designed with the utmost care and has gone through several iterations of exerting review and pilot testing. We also highlight that this is academic

research as opposed to a survey conducted by a marketing firm. The timeline to send out survey email is July 2019 followed by the second round of email as the first reminder – a month later. Depending on response rate the third round of emails as the second reminder was send out about a month later.

Chapter 5 Data Screening, Measurement Development and Structural Model Testing

5.1 Sample Demographics and Data Screening

The data collection for this research is based on two Hadoop user groups including (1) 'Hadoop New York User Group' with 4,060 members, on the east coast, and (2) 'Bay Area Hadoop User Group' with 6,440 members, on the West Coast. This data was collected using a survey instrument via the Qualtrics web-based tool. The survey period spans over a period of three months: July 25, 2019, to September 30, 2019. There are 402 respondents participated in this survey. After data screening 53 responses were found to be incomplete. Hence, we rejected those 53 responses. That means 349 responses are identified as valid.

Examination of Data Entry and Missing Data: The examination of data entry and missing data was done to get significant insight into data characteristics. To make sure data look good we need to validate data over and over – we did a manual check of each row three times. Then we did descriptive statistics including frequency distribution, mean, and standard deviation.

In examining the completeness of the responses, it was found that 53 responses contained missing data for some construct items. These cases were omitted from the preliminary analysis. We used SPSS to test the common method bias in responses. The final sample size consisted of 349 responses.

Table 13: Survey Respondents' Job Profiles

Job Roles	Frequency	Percentage
Hadoop Engineer/Application Developer	135	38.68
Hadoop Administrator	53	15.19
Big Data Architect/Enterprise Architect	45	12.9
No Response	34	9.74
Other	24	6.88
Data Scientist	22	6.3
Data Analyst	19	5.44
Big Data/Information Technology (IT) Manager	10	2.87
Chief Information Officer (CIO) or similar executive	5	1.43
Big Data Program Manager	2	0.57
Total	349	100

Table 13 shows that most of the survey respondents' job role was Hadoop Engineer/Application Developer (39%), Hadoop Administrator (15%), Big Data Architect/Enterprise Architect (13%), Data Scientist (6%), Data Analyst (5%), Big Data/Information Technology (IT) Manager (3%), and Chief Information Officer or similar level experience (1%). About 7% of the respondents identified themselves as having some other job roles, while 10% of the respondents did not answer this question.

Participants consist of different roles because in IT, projects with different roles are involved. Hence, it justifies having opinions from others. Their position signifies the high-profile participation in the survey that adds value to the quality of survey responses.

Table 14: Survey Respondents' Company Profiles

Industries Surveyed	Frequency	Percentage
Software/Internet Services	96	27.51
Financial Services	47	13.47
Healthcare	33	9.46
Consulting/Professional Services	32	9.17
No Response	28	8.02
Telecommunications	26	7.45
Other	26	7.45
Manufacturing	24	6.88
Retail	19	5.44
Insurance	12	3.44
Advertising/Marketing	3	0.86
Transportation/Logistics	3	0.86
Total	349	100

Table 14 shows that survey respondents represent a host of diverse industries. This speaks for the response of many industries as opposed to a single industry. The industries surveyed include Software/Internet Services (28%), Financial Services (14%), Healthcare (10%), Consulting/Professional Services (9%), Telecommunications (7%), Manufacturing (7%), Retail (5%), Insurance (3%), and Advertising/Marketing and Transportation/ Logistics (both less than 1%). About 7% of the respondents identified themselves as belonging to other industries while 8% of the respondents did not answer this question.

5.2 Measurement Development

This dissertation analyzes survey data using structural equation modeling (SEM) software, AMOS. We apply structural equation model techniques in three stages such as single measurement factor model, confirmatory factor analysis (CFA), and a hypothesized structural equation model. "A model is any simplified representation of reality that is used to better understand real-life situations" (Krugman & Wells, 2017). We provide a brief description of these models in several sections of this chapter. Structural equation modeling (SEM) use has been steadily increasing IS research (Chin & Todd, 1995).

To measure the model fits into data there are several statistical techniques used. As part of model-fit steps the reliability test is done via confirmatory factor analyses (CFA) estimates. The reliability tests are done to make sure the internal consistency of the items is maintained. This process allows for determining as to which variables are to be retained and which ones are to be dropped. In this process, an individual model is developed for each construct measure to confirmatory factor analysis.

Structural equation modeling (SEM) is a statistical modeling technique used to draw relationships among variables. The SEM does model specification by linking the variables. SEM is used for quantitative analyses of data through several analytical techniques to specify estimates, to test relationships between observed and unobserved variables, and to check the influence of observed variables on latent variables. The SEM produces a family of statistical analysis including covariance analysis, regression, and

factor analysis. The structural equation model can be considered a model to conduct both factor and multiple regression. The SEM outputs regression weights, variances, and covariance on a set of parameters. It tests both measurement and structural relationships.

To determine the models fit data, several statistical tests are conducted in structural equation modeling (Bagozzi & Yi, 1988). These include the common absolute indices (Chi-Square, RMSEA) and common relative fit indices (IFI, TLI, and CFI).

Absolute fit indices determine how deductive/inferred model fits sample data. With different proposed model variations, the model could be used to see which model fits data much better. This provides the most fundamental information as to how a proposed mode/ theory fits data. Absolute fit indices do not depend on any comparison with a baseline model (Hooper et al., 2008). The tests that fall under absolute indices include the Chi-Squared test and RMSEA (Hooper et al., 2008).

The chi-square is considered a “badness-of-fit” index – smaller values speak for better fit of the model to data. “A chi-square value close to zero indicates little difference between the expected and observed covariance matrices. In addition, the probability level must be greater than 0.05 when chi-square is close to zero” (Suhr, 2006, p. 2).

The Chi-Square has historically been used to measure of overall model fit. It determines the discrepancy between the sample and fitted covariance matrices (Hooper et al., 2008; Hu & Bentler, 1999). One key issue with the chi-square test is that as the

sample size increases its sensitivity also increases. And the consequence is that with such an increase the chi-square test fails. Barrett (2007) explains that it occurs because the sample size is used as a multiplier of the discrepancy function in the model-fit test. Due to this practical limitation, the researcher suggests dividing the chi-square value by the degrees of freedom (chi-square/df). The acceptable ratio range is reported as between 2.0 and 5.0 (Hooper et al., 2008).

The chi-square is reported to be sensitive compared to the sample size and complexity of the model. Kenney and McCoach (2003) report that a more complex model will produce bigger chi-square which more likely ends up rejecting the model. Given the sensitivity of the chi-square, values researchers suggest using a “normed” chi-square in which chi-square is divided by the degrees of freedom (Holmes-Smith et al., 2004). The equation for normed chi-square is derived as $\text{normed chi-square} = \text{chi-square}/\text{df}$. Byrne (2016), Hair et al., 2010, and Holmes-Smith et al. (2004) provide a guideline that a normed chi-square value between 1 and 2 indices should speak for a good model fit.

The root mean square error of approximation (RMSEA) is related to residual in the model (Suhr, 2006). RMSEA values range from 0 to 1. The smaller the value the better the model. A model could be considered fit to data if an RMSEA value of 0.08 or less (Hu & Bentler, 1999; Tabachnick & Fidell, 2012).

The RMSEA is considered the second most important fit indices statistics. The RMSEA is considered to favor parsimony as it chooses the model with relatively less the number of parameters (Hooper et al., 2008). The RMSEA values range from 0 to 1 with a

smaller value indicating a better fit model. Hu and Bentler (1999) reported that an RMSEA value of 0.06 or less speaks for an acceptable model.

In regard to common relative fit indices, the IFI, TLI, and CFI are generally reported by most of the researchers. There are several common relative fit indices and specific rules of thumb applied in regard to the minimum level of the score for a good fit under each fit index (Byrne, 2016). However, Kenny and McCoach (2003) observe that there is no consistent standard or golden rule in choosing the fit indices. The authors generally suggest the indices of CFI and TLI that could be used as common relative fit indices. McQuitty (2004) report that goodness-of-fit statistics are less sensitive to sample size. These include IFI, TLI, CFI (Bentler, 1999; Marsh et al., 1998). So, a few indices are called out by these researchers as prominent fit indices.

The incremental fit index (IFI) is considered close to R-squared. A value with zero means the worst possible model and a value of 1 indicates the highest possible model (Kenny & McCoach, 2003). The TLI (Tucker Lewis Index) is another fit index used in SEM. If the TFI value is greater than one it is set to one. The TLI connected to correlations in the data. If the average correlation between variables is not that high then, the TLI will not be high. A TLI value of ≥ 0.90 is considered acceptable. The Comparative Fit Index (CFI) is equal to the discrepancy function adjusted for sample size (Suhr, 2006). CFI values range from 0 to 1. CFI is considered a “goodness-of-fit” index where larger values mean better fit (Suhr, 2006). Several researchers suggested that an acceptable model fit could be considered when a CFI value is 0.90 or greater (Hu & Bentler, 1999). The

comparative fit index (CFI) takes sample size into account in its calculation. It performs well when the sample size is relatively small (Hooper et al., 2008; Tabachnick & Fidell, 2012). The statistic range for this index is between 0.1 and 1.0. The larger the value the better.

5.3 Confirmatory Factor Analysis

The confirmatory factor analysis (CFA) is conducted to “examine whether or not existing data are consistent with a highly constrained a priori structure that meets conditions of model identification” (Maruyama, 1998). The CFA is also called a “measurement model” in which all factors along with their indicators are connected to one another. The measurement model is destined to represents the theory. The measurement model shows how measured variables come together to represent the theory.

CFA is used to determine if each factor is statistically valid and each factor can be reflected in its indicators. Each factor is linked to its indicators. The factor(s) and measure(s) that are not statistically valid are dropped from the model as part of CFA. In the CFA model, no structural or hypothesized relationship is drawn. Variable are correlated and each variable has its indicators linked to it. The CFA for our research is shown in Figure 4.

As the first step of the CFA, we first evaluate each measure using the single measurement factor model approach that depicts and analyzes data based on a single variable/construct and its measures. As part of a single measurement factor model

standard regression weight results are being evaluated to see if it shows factor loading is good – if it linked to the construct.

As part of single measurement, we conduct instrument reliability test through Cronbach’s alpha test using IBM SPSS. Cronbach’s alpha value greater than or equal to .7 is considered reliable. Table 15 provides results on the quality of survey questions. In most cases the alpha values are greater than .8.

Table 15: Survey Questions Ratings

Construct Name	Number of Items	Cronbach's Alpha	Reliability
Scalability (SC)	4	.901	Reflective
Data Storage & Processing (DS)	4	.776	Reflective
Cost-Effectiveness (COST)	4	.920	Reflective
Performance Expectancy (PE)	4	.869	Reflective
Security & Privacy (SP)	4	.901	Reflective
Reliability (RL)	4	.901	Reflective
Data Analytics Capability (DA)	4	.847	Reflective
Training & Skills (TR)	4	.901	Reflective
Flexibility (FL)	4	.869	Reflective
Output Quality (OQ)	4	.887	Reflective
Functionality (FN)	4	.728	Reflective
Facilitating Conditions (FC)	4	.848	Reflective
Perceived Usefulness (PU)	4	.901	Reflective
Perceived Ease of Use (PEOU)	4	.887	Reflective
Behavioral Intention (BI)	3	.808	Reflective
Actual Use (AU)	3	.787	Reflective

We perform convergent validity of the construct items. Convergent validity is the extent to which an indicative variable aligns or converges on a specific latent construct. The convergent principle state that the measures of constructs that are related to each other should be strongly correlated. The correlations provide evidence that the items all converge on the same construct. Convergence is demonstrated by items having a high

proportion of variance in common having a large commonality. This can be judged from Standardized Regression Estimates in AMOS by looking at output and searching for construct loadings and AVE.

This research uses the AMOS software to perform reliability tests. The reliability tests are conducted to ensure the internal consistency of the items in a measure. This helps to determine whether a construct-item or the construct itself should be retained or dropped from the model. We conduct the calculation of the reliability scores of each construct.

The goal is to see if items under a construct have the homogenous factor loadings. The average variance extract (AVE) needs to be greater than 0.50. The formula for AVE is the following:

$$AVE = \frac{\sum_{i=1}^n \lambda_i^2}{n}$$

... where λ (*Lambda*) represent the standardized factor loading and i is the number of items.

The composite reliability (CR) values need to be greater than 0.70 to be qualified as a good construct for the model. The formula for the CR is the following:

$$CR = \frac{(\sum_{i=1}^n \lambda_i)^2}{(\sum_{i=1}^n \lambda_i)^2 + (\sum_{i=1}^n \delta_i)}$$

... where λ represent the standardized factor loading and i is the number of items. And δ (*Epsilon*) represents error variance terms.

The constructs and their items are evaluated using the individual measurement model and confirmatory factor analysis (CFA). The overall measurement model makes sure that the dimensionality of the constructs is valid, and the measures are valid. Our CFA model is shown in Figure 3. There are 60 construct-items in this model with one item measure dropped from both DS and AU.

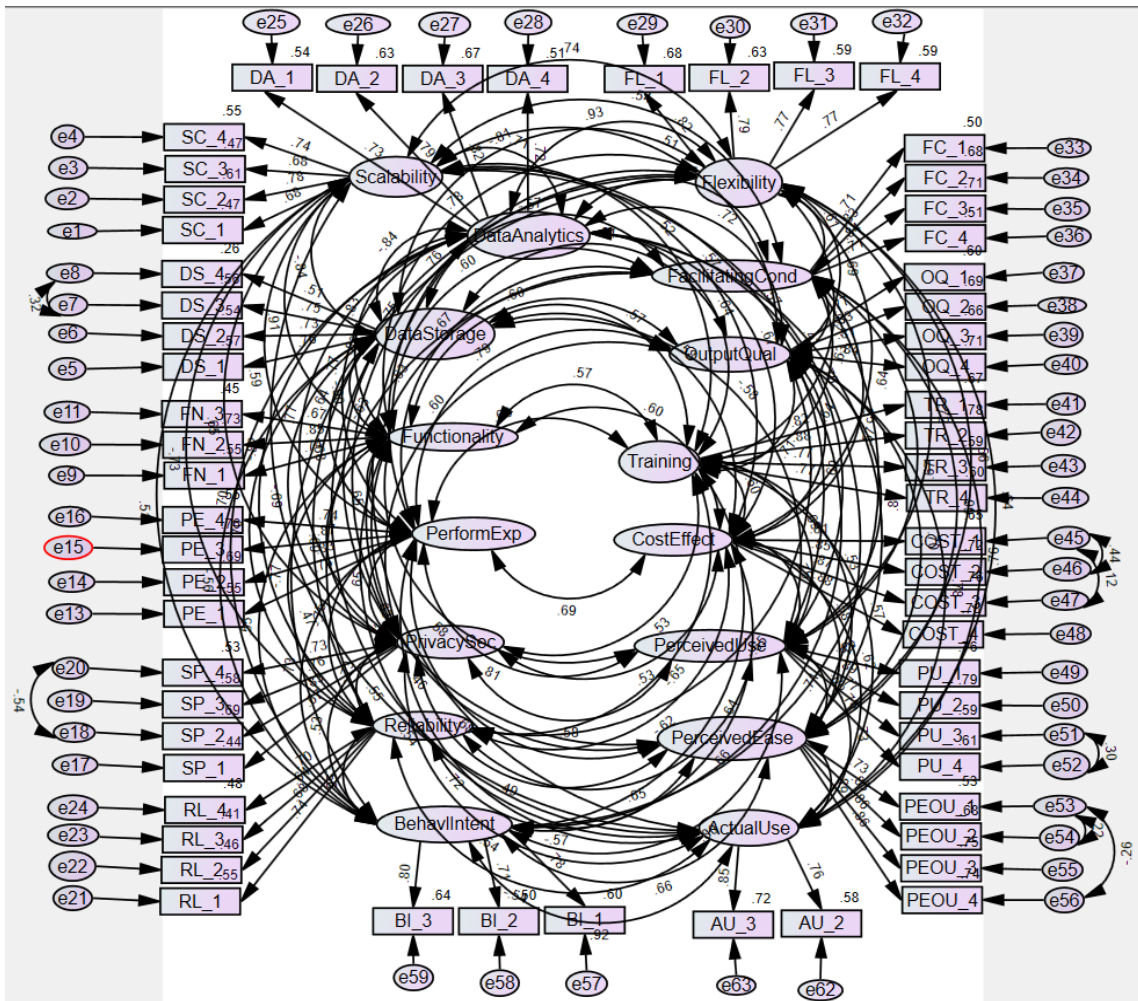


Figure 3: Confirmatory Factor Analysis (CFA)

As part of measurement model (CFA) and path model, we also analyze fit indices using AMOS (sections 5.3.1 to 5.3.17 and 5.4). In section 5.2, we have provided detailed literature findings about recommended threshold numbers of these fit indices. To determine the model's fit data, several statistical tests are conducted in structural equation modeling (Bagozzi & Yi, 1988). These include the common absolute indices (Chi-Square, RMSEA) and common relative fit indices (IFI, TLI, and CFI). Here we provide a few formulas that are used in this research.

The formula for incremental fit index (IFI) is the following (Bollen, 1989).

$$IFI = \Delta_2 = \frac{\hat{C}_b - \hat{C}}{\hat{C}_b - d}$$

... where \hat{C} and d speak for discrepancy and the degrees of freedom for the model being measured, and \hat{C}_b (b as a subscript) and d provide the discrepancy and the degrees of freedom for the baseline model (AMOS, 2020). The AMOS user guide provides details (Amos, 2020).

The Tucker-Lewis Index (TLI) coefficient is shown below (Bentler and Bonett, 1980).

$$TLI = \rho_2 = \frac{\hat{C}_b/d_b - \hat{C}/d}{\hat{C}_b/d_b - 1}$$

... where \hat{C} and d show discrepancy and the degrees of freedom for the model being tested, and \hat{C}_b (b as a subscript) and d_b provide the discrepancy and the degrees of freedom for the baseline model (Amos, 2020). See AMOS user guide for details (Amos, 2020).

The formula for Comparative Fit Index (CFI) is shown below (Bentler, 1990).

$$CFI = 1 - \frac{\max(\hat{C}_B - d, 0)}{\max(\hat{C}_B - d_B, 0)} = 1 - \frac{NCP}{NCP_B}$$

... where \hat{C} , d , and NCP consist of the discrepancy, the degrees of freedom and the non-centrality parameter estimate for the model being assessed, and \hat{C}_B , d_B and NCP_B shed light on the discrepancy, the degrees of freedom and the non-centrality parameter estimate for the baseline model (Amos, 2020). Refer to AMOS user guide for details (Amos, 2020).

5.3.1 CFA: Scalability

Table 16: Summary of Initial Findings (CFA): Scalability

Items	Item Wording	Initial Standardized Loading	Final			
			Standardized Loadings	Variance		
SC1	Hadoop is scalable to handle hundreds of terabytes to petabytes of data compared to relational databases.	0.692	.696	0.521136		
SC2	With the increase of applications, users, and data volume, Hadoop is able to meet extra load by expanding the number of nodes.	0.775	.797	0.399375		
SC3	Hadoop has built-in capability to scale-out storage compared to our organization's traditional data storage systems.	0.774	.673	0.545724		
SC4	Hadoop's scale-out storage system can store and distribute very large data sets across hundreds of inexpensive servers that operate in parallel.	0.674	.723	0.446464		
Average Variance Extracted			0.524			
Composite Construct Reliability			0.814			
Achieved Fit Indices						
	CMIN/DF	p-value	RMSEA	IFI	TLI	CFI
Initial	1.987	.077	0.053	0.989	0.987	0.989
Final	1.712	0	0.045	0.925	0.915	0.924

The average variance extracted (AVE) is 0.52 for this four-item measure. This is above the acceptable level of 0.5 as indicated in the literature (Fornell & Larcker, 1981). This also said as good convergent validity. If AVE is less than .5 then we to remove a poor construct item to improve the AVE value. Rule of thumb is to remove one item at a time. Also, we need to examine the item carefully before deleting it and ensure that there are enough items available.

The Composite Construct Reliability (CR) is 0.81 for the four-item construct, which is well above the acceptable threshold point of .7. Both these reliability indicator values indicate that these four items are reliable and valid for this construct measure.

5.3.2 CFA: Data Storage and Processing

The standardized loadings (regression weights) for DS_1, DS_2, DS_3, and DS_4 are .761, .740, .756, 0.539 respectively. Only DS_4 shows regression weights lower than the weights of the other three items but, the loading is above .5. Hence, all these four items are subjected to Confirmatory Factor Analysis (CFA). The average variance extracted (AVE) is close to .5 (rounded). An AVE value of .5 is acceptable. Hence, these four items passed the convergent validity test. The composite construct reliability (CR) is close to .8 which is above the threshold value of .7. The CR value of .8 also ensures that four items represent this construct well. The CMIN/DF is 7.125 (df = 5 and p-value = 0.000) which is above 2.0. The RMSEA value is .053, which is within the range of 0 to 1. The IFI (.931), TLI (.917) and CFI (.931) values are above the threshold numbers.

Table 17: Summary of Initial Findings (CFA): Data Storage and Processing

Items	Item Wording	Initial Standardized Loading	Final			
			Standardized Loadings	Variance		
DS1	Hadoop is capable to run analytics on hundreds of terabytes to petabytes of data set.		.761	0.420879		
DS2	Hadoop's processing engine is capable to process both structured and unstructured data.		.740	0.4524		
DS3	Hadoop's storage and processing engine can serve many application needs - analytics, processing, machine learning.		.756	0.428464		
DS4	Hadoop is capable to receive and process streaming data real-time.		0.539	0.709479		
Average Variance Extracted			0.497			
Composite Construct Reliability			0.795			
Achieved Fit Indices						
	CMIN/DF	p-value	RMSEA	IFI	TLI	CFI
Initial						
Final	7.125	0.000	0.053	0.931	0.917	0.931

5.3.3 CFA: Cost-Effectiveness

The regression weights for Cost1, Cost2, Cost3, and Cost4 are 0.812, 0.855, 0.857, and 0.883 respectively. All these values show very high standardized loadings. The average variance extracted (AVE) is .73 which above the threshold value of .5. The composite construct reliability (CR) is .91, which also above the threshold value of .7. The CMIN/DF value is 7.125 (df = 4 and p-value = 0.000), which is less than threshold value of 2.0 (Tabachnick & Fidell, 2012). The RMSEA value is .053, which is less than the threshold value of 1.0. The RMSEA values range from 0 to 1 with a smaller value indicating a better fit model. The IFI, TLI, and CFI values are 0.931, 0.917, and 0.931 respectively, all of which are greater than the threshold value of .90.

Table 18: Summary of Initial Findings (CFA): Cost-Effectiveness

Items	Item Wording	Initial Standardized Loading	Final			
			Standardized Loadings	Variance		
Cost1	Hadoop is able to hold hundreds of terabytes to petabytes of data with minimal cost.	0.856	0.812	0.340656		
Cost2	Hadoop offers a cost-effective storage solution for my organization's exploding data sets.	0.896	0.855	0.268975		
Cost3	Hadoop is able to improve the efficiency of business applications and thereby reduce costs.	0.841	0.857	0.265551		
Cost4	Using Hadoop is cost-effective.	0.869	0.883	0.220311		
Average Variance Extracted			0.726			
Composite Construct Reliability			0.914			
Achieved Fit Indices						
	CMIN/DF	p-value	RMSEA	IFI	TLI	CFI
Initial	9.372	0.000	0.155	0.961	0.953	0.961
Final	1.893	0.109	0.051	0.997	0.995	0.997

5.3.4 CFA: Performance Expectancy

The construct, Performance Expectancy, represents PE1, PE2, PE3, and PE4 with standardized values of 0.740, 0.834, 0.866, and 0.743 respectively. These values are higher than .5. The Average Variance Extracted value is 0.64, which is greater than .5 and CR value is 0.87 which is greater than the threshold value of .7. The CMIN/DF (.297), (df = 1 and p-value = 0.586), RMSEA (0.000), IFI (1.001), TLI (1.006), and CFI (1.000) are within the acceptable threshold numbers.

Table 19: Summary of Initial Findings (CFA): Performance Expectancy

Items	Item Wording	Initial Standardized Loading	Final	
			Standardized Loadings	Variance
PE1	The team members of my organization find the Hadoop Platform useful in performing jobs.	0.793	0.740	0.452400
PE2	By using the Hadoop Platform members of my organization are able to accomplish tasks more quickly.	0.818	0.834	0.304444

PE3	The use of the Hadoop Platform increases my organization's productivity.	0.844	0.866	0.250044		
PE4	Hadoop is able to provide a good user experience.	0.739	0.743	0.447951		
Average Variance Extracted			0.636			
Composite Construct Reliability			0.874			
Achieved Fit Indices						
	CMIN/DF	p-value	RMSEA	IFI	TLI	CFI
Initial	6.43	0.000	0.125	0.961	0.953	0.961
Final	0.297	0.586	0.000	1.001	1.006	1.000

5.3.5 CFA: Security and Privacy

This construct consists of four items all of which provide standardized regression weights of 67, 83, 75, and 73. These values are greater than .5 and thus acceptable. The AVE value is .56 and composite construct reliability value is .84. The CMIN/DF (0.399), (df = 1 and p-value = 0.528), RMSEA (.000), IFI (1.001), TLI (1.007), and CFI (1.000) values are also within the threshold points. These four items were subjected to CFA.

Table 20: Summary of Initial Findings (CFA): Security and Privacy Considerations

Items	Item Wording	Initial Standardized Loading	Final			
			Standardized Loadings	Variance		
SP1	Hadoop has data protection capability such as encryption and data masking to prevent sensitive data from being accessed by unauthorized users and applications.	0.668	0.667	0.555111		
SP2	Hadoop has authentication capability such as Kerberos to authenticate Hadoop users.	0.767	.830	0.311100		
SP3	Hadoop provides a capability for providing role-based authorization to both data and metadata stored in HDFS in a Hadoop cluster.	0.762	.759	0.423919		
SP4	Hadoop (HDFS) is able to ensure the confidentiality of stored data in both physical and cyber ways.	0.685	.730	0.467100		
Average Variance Extracted			0.560			
Composite Construct Reliability			0.835			
Achieved Fit Indices						
	CMIN/DF	p-value	RMSEA	IFI	TLI	CFI
Initial	9.456	0.000	0.156	0.912	0.895	0.912

Final	0.399	0.528	0.000	1.001	1.007	1.000
-------	-------	-------	-------	-------	-------	-------

5.3.6 CFA: Reliability

Four construct items, RL1, RL2, RL3, and RL4 have standardized values of .789, .678, .685, and .789 respectively. The CMIN/DF value is 0.433 (df = 1 and p-value = 0.511). The RMSEA value is .000, which is within the threshold value of 0 to 1. The IFI (1.001), TLI (1.009), and CFI (1.000) values are above the threshold value of .9. The AVE value 0.54, which is above the threshold value of .5. However, composite construct reliability is .83, which is greater than the threshold value of 7.

Table 21: Summary of Initial Findings (CFA): Reliability

Items	Item Wording	Initial Standardized Loading	Final			
			Standardized Loadings	Variance		
RL1	Hadoop keeps multiple copies of the same data in different nodes which makes my organization feel comfortable about not losing any critical data.	0.754	.789	0.377479		
RL2	Hadoop is capable to automatically identify data node failing and possible remedy.	0.659	.678	0.540316		
RL3	Hadoop maintains data in raw format which allows data to remain the way it comes from the source, that is, in its original format.	0.631	.685	0.530775		
RL4	Hadoop Platform is able to operate under given conditions, without collapsing.	0.678	.789	0.377479		
Average Variance Extracted			0.544			
Composite Construct Reliability			0.826			
Achieved Fit Indices						
	CMIN/DF	p-value	RMSEA	IFI	TLI	CFI
Initial	5.139	0.000	0.109	0.945	0.934	0.945
Final	0.433	0.511	0.000	1.001	1.009	1.000

5.3.7 CFA: Data Analytics Capability

Four items, DA_1, DA_2, DA_3, and DA_4 have standardized values of .623, .742, .870, and .757 respectively. All these items have loading greater than .5. The CMIN/DF value is .870 (df = 1 and p-value = 0.351) which is below threshold value of 2.0. The RMSEA value is .000 which is below the threshold value of 0.06. The IFI, TLI, and CFI values are 1.000, 1.001, and 1.000 respectively. The AVE value is .59 which is greater than the threshold value of .5 and composite construct reliability value is .85 which is greater than the threshold value of .70.

Table 22: Summary of Initial Findings (CFA): Data Analytics Capability

Items	Item Wording	Initial Standardized Loading	Final			
			Standardized Loadings	Variance		
DA1	Hadoop allows to perform different types of analytics (including Customer, Compliance, Fraud, Operational) to enable making business decisions.	0.745	0.623	0.611871		
DA2	Hadoop's capability to store both historical and current data allows for the discovery of knowledge from massive datasets.	0.819	0.742	0.449436		
DA3	Hadoop's capability to combine data from many sources (external and internal) allows my organization to get 360-degree views of customers and other business entities.	0.789	0.870	0.243100		
DA4	Hadoop provides my organization capability to develop and run machine learning model on a complete set of data (stored in HDFS).	0.709	0.757	0.426951		
Average Variance Extracted			0.589			
Composite Construct Reliability			0.851			
Achieved Fit Indices						
	CMIN/DF	p-value	RMSEA	IFI	TLI	CFI
Initial	5.444	0.000	0.113	0.963	0.955	0.963
Final	0.870	0.351	0.000	1.000	1.001	1.000

5.3.8 CFA: Training and Required Skills

Four items, TR_1, TR_2, TR_3, and TR_4 have standardized values of 0.810, 0.904, 0.775, and 0.749 respectively. All these items have loading greater than .5. The CMIN/DF value is 1.262 (df = 2 and p-value = 0.283) which is below threshold value of 2.0. The RMSEA value is .027 which is below the threshold value of 0.06. The IFI, TLI, and CFI values are .999, .998, and .999 respectively. The AVE value is .66 which is greater than the threshold value of .5 and composite construct reliability value is .88 which is greater than the threshold value of .70.

Table 23: Summary of Initial Findings (CFA): Training and Required Skills

Items	Item Wording	Initial Standardized Loading	Final			
			Standardized Loadings	Variance		
TR1	Having user-support for the Hadoop platform will help users of my organization gain knowledge.	0.838	0.810	0.332511		
TR2	Specialized training will save my organization's users' time on learning how to use the Hadoop platform.	0.852	0.904	0.218544		
TR3	Documentation should be provided for the Hadoop platform for users wanting to learn on their own.	0.805	0.775	0.405559		
TR4	The training gave users of my organization confidence in the Hadoop Platform.	0.754	0.749	0.400924		
Average Variance Extracted			0.661			
Composite Construct Reliability			0.886			
Achieved Fit Indices						
	CMIN/DF	p-value	RMSEA	IFI	TLI	CFI
Initial	4.154	0.001	0.095	0.979	0.975	0.979
Final	1.262	0.283	0.027	0.999	0.998	0.999

5.3.9 CFA: Flexibility

Four items, FL_1, FL_2, FL_3, and FL_4 have standardized values of 0.778, 0.853, 0.780, and 0.817 respectively. All these items have loading greater than .5. The CMIN/DF value is 1.538 (df = 4 and p-value = 0.188) which is below threshold value of 2.0. The RMSEA value is .039 which is below the threshold value of 0.06. The IFI, TLI, and CFI values are .997, .995, and .997 respectively. The AVE value is .65 which is greater than the threshold value of .5 and composite construct reliability value is .88 which is greater than the threshold value of .70.

Table 24: Summary of Initial Findings (CFA): Flexibility

Items	Item Wording	Initial Standardized Loading	Final			
			Standardized Loadings	Variance		
FL1	Hadoop provides greater flexibility to consolidate data from various sources into one single place (i.e., Hadoop HDFS).	0.780	0.778	0.394716		
FL2	Hadoop provides high throughput as well as fault tolerance as data is also replicated to other nodes in the cluster.	0.818	0.853	0.272391		
FL3	Hadoop allows to build programs at a small scale and expand the system as needed.	0.781	0.780	0.391600		
FL4	Hadoop enables businesses to easily access new data sources and tap into different types of data to generate value.	0.779	0.817	0.332511		
Average Variance Extracted			0.652			
Composite Construct Reliability			0.882			
Achieved Fit Indices						
	CMIN/DF	p-value	RMSEA	IFI	TLI	CFI
Initial	4.865	0.000	0.070	0.971	0.966	0.971
Final	1.538	0.188	0.039	0.997	0.995	0.997

5.3.10 CFA: Output Quality

Four items, OQ_1, OQ_2, OQ_3, and OQ_4 have standardized values of 0.799, 0.824, 0.845, and 0.825 respectively. All these items have loading greater than .5. The CMIN/DF value is 1.796 (df = 4 and p-value = 0.127) which is below threshold value of 2.0. The RMSEA value is .048 which is below the threshold value of 0.06. The IFI, TLI, and CFI values are .996, .994, and .996 respectively. The AVE value is .66 which is greater than the threshold value of .5 and composite construct reliability value is .89 which is greater than the threshold value of .70.

Table 25: Summary of Initial Findings (CFA): Output Quality

Items	Item Wording	Initial Standardized Loading	Final			
			Standardized Loadings	Variance		
OQ1	Hadoop Platform's Quality is associated with the satisfaction of my organization's users' work.	0.782	0.799	0.361599		
OQ2	My organization is satisfied with the data consistency in Hadoop Platform.	0.828	0.824	0.321024		
OQ3	My organization is satisfied with the data completeness (no data gaps, missing data) in Hadoop Platform.	0.829	0.845	0.285975		
OQ4	By using the Hadoop, the users of my organization get high quality output.	0.829	0.825	0.319375		
Average Variance Extracted			0.664			
Composite Construct Reliability			0.888			
Achieved Fit Indices						
	CMIN/DF	p-value	RMSEA	IFI	TLI	CFI
Initial	2.818	0.015	0.072	0.988	0.986	0.988
Final	1.796	0.127	0.048	0.996	0.994	0.996

5.3.11 CFA: Functionality

Four items, FN_1, FN_2, and FN_3 have standardized values of 0.743, 0.867, and 0.649 respectively. All these items have loading greater than .5. The CMIN/DF value is 1.471

(p-value = 0.000), which is below threshold value of 2.0. The RMSEA value is .037 which is below the threshold value of 0.06. The IFI, TLI, and CFI values are 0.997, 0.996, 0.997 respectively. The AVE value is .58 which is greater than the threshold value of .5 and composite construct reliability value is .80 which is greater than the threshold value of .70.

Table 26: Summary of Initial Findings (CFA): Functionality

Items	Item Wording	Initial Standardized Loading	Final			
			Standardized Loadings	Variance		
FN1	Hadoop architecture can access and process the data that comes from many sources, tools, and devices.	0.732	0.743	0.447951		
FN2	Hadoop framework provides a distributed file system for big data sets.	0.833	0.867	0.248311		
FN3	The HDFS replicates the data sets on the commodity servers making the process run in parallel.	0.631	0.649	0.578799		
FN4	Hadoop provides rich and robust machine learning libraries (e.g., Mahout).	0.534				
Average Variance Extracted			0.575			
Composite Construct Reliability			0.801			
Achieved Fit Indices						
	CMIN/DF	p-value	RMSEA	IFI	TLI	CFI
Initial	4.475	0.000	0.100	0.953	0.944	0.953
Final	1.471	0.230	0.037	0.997	0.996	0.997

5.3.12 CFA: Facilitating Conditions

Four items, FC_1, FC_2, FC_3, and FC_4 have standardized values of 0.690, 0.837, 0.859, and 0.692 respectively. All these items have loading greater than .5. The CMIN/DF value is 0.458 (df = 4 and p-value = 0.633) which is below threshold value of 2.0. The RMSEA value is .000 which is below the threshold value of 0.06. The IFI, TLI, and CFI values are 1.002, 1.005, and 1.000 respectively. The AVE value is .60 which is greater than the

threshold value of .5 and composite construct reliability value is .86 which is greater than the threshold value of .70.

Table 27: Summary of Initial Findings (CFA): Facilitating Conditions

Items	Item Wording	Initial Standardized Loading	Final			
			Standardized Loadings	Variance		
FC1	My organization takes advantage of new information technologies.	0.767	0.690	0.523900		
FC2	My organization has resources necessary to use the Hadoop Platform.	0.800	0.837	0.299431		
FC3	Given the resources, opportunities, and knowledge it takes to use the Platform, it would be easy for my organization to use the Hadoop Platform.	0.841	0.859	0.262119		
FC4	My organization has internal Hadoop Infrastructure team to support Hadoop Platform users.	0.690	0.692	0.521136		
Average Variance Extracted			0.601			
Composite Construct Reliability			0.857			
Achieved Fit Indices						
	CMIN/DF	p-value	RMSEA	IFI	TLI	CFI
Initial	4.128	0.001	0.095	0.974	0.969	0.974
Final	0.458	0.633	0.000	1.002	1.005	1.000

5.3.13 CFA: Perceive Usefulness

Four items, PU_1, PU_2, PU_3, and PU_4 have standardized values of 0.868, 0.924, 0.738, and 0.741 respectively. All these items have loading greater than .5. The CMIN/DF value is 0.030 (df = 1 and p-value = 0.861) which is below threshold value of 2.0. The RMSEA value is .000 which is below the threshold value of 0.06. The IFI, TLI, and CFI values are 1.001, 1.006, and 1.000 respectively. The AVE value is .69 which is greater than the threshold value of .5 and composite construct reliability value is .90 which is greater than the threshold value of .70.

Table 28: Summary of Initial Findings (CFA): Perceive Usefulness

Items	Item Wording	Initial Standardized Loading	Final			
			Standardized Loadings	Variance		
PU1	Using Hadoop Platform enables my organization to accomplish its tasks more quickly.	0.829	0.868	0.246576		
PU2	Using Hadoop Platform makes it easier for my organization to carry out its tasks.	0.851	0.924	0.146224		
PU3	Hadoop Platform is flexible from varieties of data storage and processing perspectives.	0.831	0.738	0.455356		
PU4	Overall, using Hadoop Platform is advantageous compared to the conventional data management system of my organization.	0.831	0.741	0.450919		
Average Variance Extracted			0.688			
Composite Construct Reliability			0.898			
Achieved Fit Indices						
	CMIN/DF	p-value	RMSEA	IFI	TLI	CFI
Initial	12.252	0.000	0.180	0.938	0.925	0.938
Final	0.030	0.861	0.000	1.001	1.006	1.000

5.3.14 CFA: Perceived Ease of Use

Four items, PEOU_1, PEOU_2, PEOU_3, and PEOU_4 have standardized values of 0.762, 0.882, 0.850, and 0.858 respectively. All these items have loading greater than .5. The CMIN/DF value is 1.433 (df = 2 and p-value = 0.239) which is below threshold value of 2.0. The RMSEA value is .035 which is below the threshold value of 0.06. The IFI, TLI, and CFI values are .998, .998, and .998 respectively. The AVE value is .70 which is greater than the threshold value of .5 and composite construct reliability value is .91 which is greater than the threshold value of .70.

Table 29: Summary of Initial Findings (CFA): Perceived Ease of Use

Items	Item Wording	Initial Standardized Loading	Final	
			Standardized Loadings	Variance

PEOU1	Interacting with Hadoop platform does not require a lot of mental effort.	0.731	0.762	0.419356		
PEOU2	My organization finds Hadoop Platform easy to use when performing its job functions.	0.854	0.882	0.222076		
PEOU3	It is easy for my organization's users to become more skillful and experienced with Hadoop Platform.	0.871	0.850	0.277500		
PEOU4	My organization's interaction with Hadoop Platform is clear and understandable.	0.830	0.858	0.263836		
Average Variance Extracted			0.704			
Composite Construct Reliability			0.905			
Achieved Fit Indices						
	CMIN/DF	p-value	RMSEA	IFI	TLI	CFI
Initial	7.425	0.000	0.136	0.962	0.954	0.962
Final	1.433	0.239	0.035	0.998	0.998	0.998

5.3.15 CFA: Behavioral Intention

Four items, BI_1, BI_2 and BI_3 have standardized values of 0.803, 0.743, and 0.740 respectively. All these items have loading greater than .5. The CMIN/DF value is 1.594 (df = 2 and p-value = 0.203) which is below threshold value of 2.0. The RMSEA value is .041 which is below the threshold value of 0.06. The IFI, TLI, and CFI values are .997, .995, and .997 respectively. The AVE value is .58 which is greater than the threshold value of .5 and composite construct reliability value is .81 which is greater than the threshold value of .70.

Table 30: Summary of Initial Findings (CFA): Behavioral Intention

Items	Item Wording	Initial Standardized Loading	Final	
			Standardized Loadings	Variance
BI1	My organization intends to use Hadoop for its data storage, management, processing, and analytical needs.		0.803	0.355191
BI2	I predict my organization would use Hadoop within the next six months.		0.743	0.447951
BI3	My organization will continue to use Hadoop in the future.		0.740	0.452400

Average Variance Extracted		0.581				
Composite Construct Reliability		0.806				
Achieved Fit Indices						
	CMIN/DF	p-value	RMSEA	IFI	TLI	CFI
Initial						
Final	1.594	0.203	0.041	0.997	0.995	0.997

5.3.16 CFA: Actual Use

Four items, AU_1, AU_2, and BI_3 have standardized values of 0.585, 0.763, and 0.851 respectively. With AU_1 value overall CFA show a bit poor fit. Hence, removed from the CFA (see CFA in Figure 3). All these items have loading greater than .5. The CMIN/DF value is .478 (df = 1 and p-value = 0.489) which is below threshold value of 2.0. The RMSEA value is .000 which is below the threshold value of 0.06. The IFI, TLI, and CFI values are 1.003, 1.008, and 1.000 respectively. The AVE value is .65 which is greater than the threshold value of .5 and composite construct reliability value is .79 which is greater than the threshold value of .70.

Table 31: Summary of Initial Findings (CFA): Actual Use

Items	Item Wording	Initial Standardized Loading	Final			
			Standardized Loadings	Variance		
AU1	My organization uses Hadoop occasionally.	0.585				
AU2	My organization uses Hadoop regularly (daily, weekly, etc.).	0.798	0.763	0.417831		
AU3	My organization is satisfied with using the Hadoop Platform.	0.814	0.851	0.275799		
Average Variance Extracted			0.653			
Composite Construct Reliability			0.790			
Achieved Fit Indices						
	CMIN/DF	p-value	RMSEA	IFI	TLI	CFI
Initial	0.478	0.489	0.000	1.003	1.008	1.000
Final						

5.3.17 Overall Measurement Model Fit

Section 5.3 covered individual measurement model and confirmatory factor analysis (CFA). Statistical estimation and model fit for all independent and dependent variables have been conducted. As part of a single measurement model test, all independent and dependent variables looked good from statistical estimation and model fit indicators perspectives. The fit statistics under the individual measurement model are provided in Table 32.

Table 32: Single Measurement Model – Estimates and Fit Indices

Construct	CMIN/DF	IFI	TLI	CFI	RMSEA	Std. Reg. Wt. 1	Std. Reg. Wt. 2	Std. Reg. Wt. 3	Std. Reg. Wt. 4
Scalability	1.712	0.925	0.915	0.924	0.045	0.696	0.797	0.673	0.723
Data Storage & Processing	7.125	0.931	0.917	0.931	0.053	0.761	0.740	0.756	0.539
Functionality	1.471	0.997	0.996	0.997	0.037	0.743	0.867	0.649	
Performance Expectancy	0.297	1.001	1.006	1.000	0.000	0.740	0.834	0.866	0.743
Security and Privacy	0.399	1.001	1.007	1.000	0.000	0.667	0.830	0.759	0.730
Reliability	0.433	1.001	1.009	1.000	0.000	0.789	0.678	0.685	0.789
Data Analytics Capability	0.870	1.000	1.001	1.000	0.000	0.623	0.742	0.870	0.757
Flexibility	1.538	0.997	0.995	0.997	0.039	0.778	0.853	0.780	0.817
Facilitating Conditions	0.458	1.002	1.005	1.000	0.000	0.690	0.837	0.859	0.692
Output Quality	1.796	0.996	0.994	0.996	0.048	0.799	0.824	0.845	0.825
Training and Required Skills	1.262	0.999	0.998	0.999	0.027	0.810	0.904	0.775	0.749
Cost-Effectiveness	1.893	0.997	0.995	0.997	0.051	0.812	0.855	0.857	0.883
Perceive Usefulness	0.030	1.001	1.006	1.000	0.000	0.868	0.924	0.738	0.741
Perceived Ease of Us	1.433	0.998	0.998	0.998	0.035	0.762	0.882	0.850	
Behavioral Intention	1.594	0.997	0.995	0.997	0.041	0.803	0.743	0.740	

Actual Use	0.478	1.003	1.008	1.000	0.000	0.585	0.798	0.814	
------------	-------	-------	-------	-------	-------	-------	-------	-------	--

The Chi-Square value is evaluated to see if the overall model fits to data. A good model fit should provide CMIN/DF value of less than or equal to 2.0 (Tabachnick & Fidell, 2012). A good model should provide a P-value of ≥ 0.05 . In terms of baseline indicators, three indicators (IFI, TLI, CFI) report how much fit the model is. These values range from 0 to 1 with a larger value indicating a better fit model. Hu and Bentler (1999) reported that IFI, TLI, and CFI value of 0.90 or greater indicate an acceptable fit model. So, a value of greater than or equal to 0.90 should be good and speak for the model fit.

Table 33: Summary of Overall Measurement Model (CFA)

Fit Indices	Overall Measurement Model	
	Initial (62 items) 1	Final (60 items) 2
χ^2 (df)	3096.986 (1709)	2710.611 (1583)
CMIN	1.812	1.712
IFI	.908	.925
TLI	.897	.915
CFI	.907	.924
RMSEA	.048	.045

The initial CFA model examined all 16 constructs (13 independent and three dependent variables) with a total of 62 items. The initial measure model provides the fit indices which are shown under the second column (Initial 1). The TLI value (.897) is less than the threshold value of .900. The other fit indices are above the acceptable threshold numbers. The final measure model consists of 60 items. Two items (FN_4 and AU_3) dropped due to low loadings. We dropped two items and ran it. These results were: 1. Chi-square = 2710.611; 2. Degrees of freedom = 1583; 3. Probability level =

.000; 4. CMIN/DF = 1.712; 5. IFI, TLI, CFI values are .925 .915 and .924 respectively. 6. RMSEA = .045.

Then we have drawn covariance of DS_3 and DS_4, SP_2 and SP_4, COST_1 and COST_2, COST_1 and COST_3, PU_3 and PU_4, PEOU_1_PEOU_2, and PEOU_1 and PEOU_4. This has helped in improving the fit indices shown under the third column (Final 2). All fit indices are above the acceptable threshold numbers. The comparative results between the initial run and final run show that the initial model is weaker than the final model. Therefore, fit statistics justify the deletion of two items from two constructs (Functionality [FN] and Actual Use [AU]). In the final CFA model, chi-square value is reduced by 386.37 (df 126, $p < .001$). The other fit indices also show improved values. This final model suggests a reasonable congruity between data and the CFA model.

5.4 SEM Path Analysis – A Hypothesized Model

Structural mode is meant for representing the theory that shows how constructs are related to other constructs. Scholars comment that SEM has been widely used in business, information systems, and information technology research (Chin & Todd, 1995; McQuity, 2004; Urbach & Ahlemann, 2010) which are mostly empirical studies. Chin and Todd (1995) state that the SEM model plays a key role in addressing IS research problems in assessing IT usage. Research finds the chi-square-test as the most valuable test. Barrett (2007) asserts that the chi-square test should be considered the only significant statistical test for the SEM model to fit the data. Urbach and Ahlemann

(2010) report that during 1994-2008 two top-ranking journals, MIS Quarterly (MISQ) and Information Systems Research (ISR) has published eighty-five research articles that used SEM. One of the critical features of SEM is that it supports latent variables (LVs) (Urbach & Ahlemann, 2010). Straub et al. (2004) provide an exhaustive list of statistical tests and techniques for which SEM is used in Information Systems research. These include discriminant validity, convergent validity, factorial validity, predictive validity, and common method bias as part of construct validity. For reliability testing, internal consistency, split-half, test-retest, inter-rater reliability, unidimensional reliability, the SEM model is used (Straub et al., 2004). Adams et al. (1992) employed the SEM model to evaluate perceived usefulness, ease of use, and usage of information technology in terms of convergent validity of voice and electronic mail data, and discriminant validity of word processing (WordPerfect, Lotus 1-2-3, and Harvard Graphics) data.

A hypothesized model is drawn based on factors (constructs) and associated indicators (measures) in the CFA model. The difference here is that a path model developed with constructs from CFA. Lines with an arrow in one direction are used to show the hypothesized direct relationship between two variables (causal and caused). Lines with an arrow in both directions are used to show the bi-directional relationships (i.e., covariance). Covariance arrows are used among exogenous variables. The hypothesized model for our research is shown in Figure 5 in chapter 5.

In section 5.3, we showed that the CFA model was run successfully with all 16 variables (both dependent and dependent). We have transferred the CFA to the path

model. As part of the first run (Iteration 1) of the path model (SEM), we have included the same number of variables and items that we had in the CFA model.

The results (p-value) of the **Iteration- 1** were shown in Table 34. This iteration shows that p-values are greater than an acceptable limit of 0.05 for most of the factors except PU, BI, and FC. That means the model was not quite right. We reviewed the p-values and decided to remove the variable Cost-Effectiveness factor ($AU \leftarrow COST$) and run the model again.

The result is shown under **Iteration-2** in Table-34. In this iteration the p-value has come within the acceptable limit of the p-value, 0.05 for four additional variables: PE, OQ, TR, and PEOU. The Iteration-2 has improved the model a lot. As part of further refinement security and privacy factor ($PU \leftarrow SP$) was removed from the model since this was showing a high p-value (.783) in Iteration-2 run.

After refinement, the model was run again, and p-values are captured under **Iteration-3** in Table-34. This time the p-value reduced a little bit but did not drop p-value below acceptable threshold point for the additional variable. We have removed one more variable, data analytics capability' ($PU \leftarrow DA$) from the model as it was showing greater p-value in Iteration-3.

The p-value of the refined model is shown under **Iteration-4** in Table-34. This time p-value came down within acceptable limit for several factors: 'scalability' ($PU \leftarrow SC$), 'flexibility' ($PU \leftarrow FL$). But still, the p-value is greater than three more variables.

We refined the model one more time by dropping the variable, 'functionality' (PU ← FN).

The final model was run, and the results were captured under **Iteration-5** in Table-34. This time p-value has dropped below an acceptable limit of 0.05 for two more variables: data Storage and processing (PU ← DS) and 'reliability' (PU ← RL). The results of this final iteration show p-value within acceptable limit for nine independent variables (IV) and three dependent variables (DV). The IV's are scalability, data Storage and processing, flexibility, output quality, performance expectancy, reliability, training and skills, facilitating conditions, and perceived ease of use (PEOU). The dependent variables (DV) include perceived usefulness (PU), behavioral intention to use (BI), and actual use (AU).

Table 34: Regression Weights – Path Model: Results of Five Iterations

Regression Path (Influence of IV on DV)	Iteration-1 p-value	Iteration-2 p-value	Iteration-3 p-value	Iteration-4 p-value	Iteration-5 (FINAL) p-value
SC → PU	.330	.083	.070	.032	.004
DS → PU	.592	.401	.397	.397	.027
FL → PU	.430	.552	.550	.013	.005
RL → PU	.696	.082	.076	.068	.013
PE → PU	.846	***	***	***	***
OQ → PU	.507	***	***	***	.002
TR → PU	.776	.023	.024	.022	.038
SP → PU	.560	.783	Dropped	Dropped	Dropped
DA → PU	.354	.536	.484	Dropped	Dropped
FN → PU	.397	.363	.339	.352	Dropped
PEOU → PU	.350	.017	.016	.020	.010
PU → BI	***	***	***	***	***
PEOU → BI	.003	.002	.002	.002	.002
FC → AU	***	***	***	***	***
COST → AU	.731	Dropped	Dropped	Dropped	Dropped
BI → AU	***	***	***	***	***

Given we he had to drop a few constructs and item we have regenerated the CFA. Based on CFA with 12 constructs and 40 items, the fit statistics under individual measurement models are provided in Table 35.

Table 35: CFA Construct Reliability

Construct	Std. Reg. Wt. 1	Std. Reg. Wt. 2	Std. Reg. Wt. 3	Std. Reg. Wt. 4	AVE	CR
Scalability	0.693	0.839	0.643		0.532	0.77
Data Storage & Processing		0.771	0.831	0.600	0.548	0.78
Performance Expectancy	0.740	0.834	0.866	0.743	0.636	0.87
Reliability	0.789	0.678	0.685	0.789	0.544	0.83
Flexibility	0.805	0.807	0.782	0.768	0.625	0.87
Facilitating Conditions	0.708	0.822	0.844	0.714	0.600	0.86
Output Quality	0.778	0.834	0.811	0.837	0.665	0.89
Training and Required Skills	0.788		0.747	0.800	0.606	0.82
Perceive Usefulness	0.863	0.888	0.770	0.778	0.683	0.89
Perceived Ease of Us	0.764	0.844	0.857	0.858	0.692	0.90
Behavioral Intention	0.766	0.726	0.804		0.586	0.81
Actual Use		0.774	0.831		0.645	0.78

It is clear from the Table 36 that the fit statistics justified the deletion of some specific constructs from the model and some items from different construct measures which resulted in the better model fit in terms for that fit indices presented.

Table 36: Summary of Overall CFA: Fit Indices

Fit Indices	Overall Measurement Model	
	CFA (16 Variables: 60 items)	CFA (12 Variables: 40 items)
χ^2 (df)	2710.611 (1583)	1536.635 (894)
CMIN	1.712	1.719
IFI	.925	.939
TLI	.915	.932
CFI	.924	.938
RMSEA	.045	.045

Here is the final Research model, drawn based on the Path Analysis Results (Figure 4).

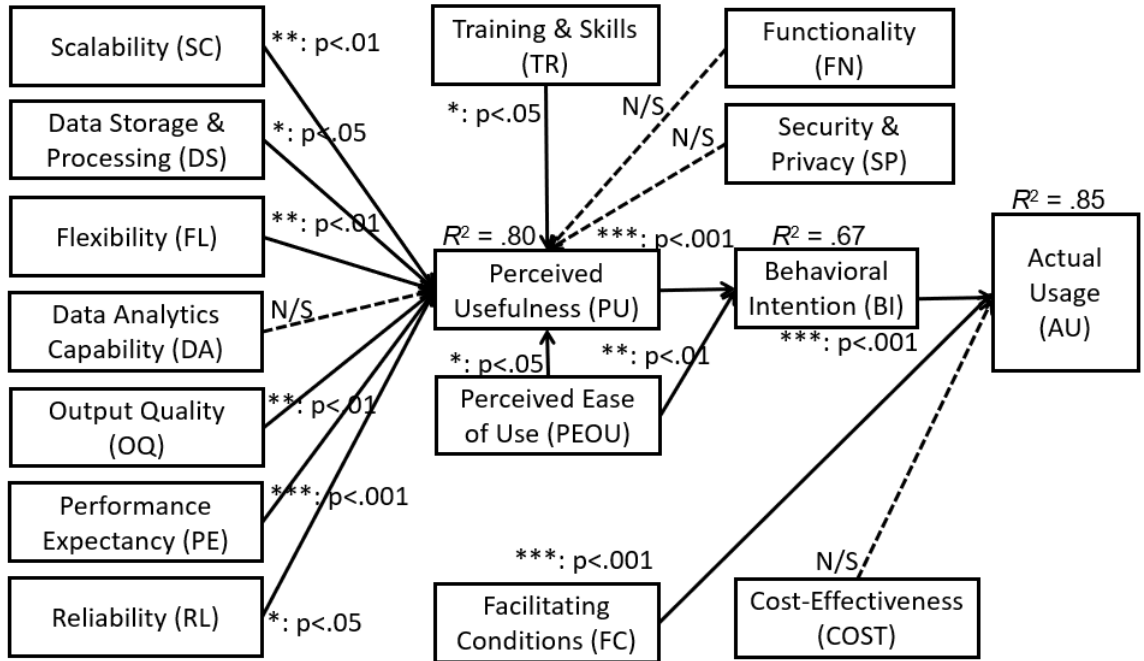


Figure 4: Final Research Model – Big Data Technology Acceptance

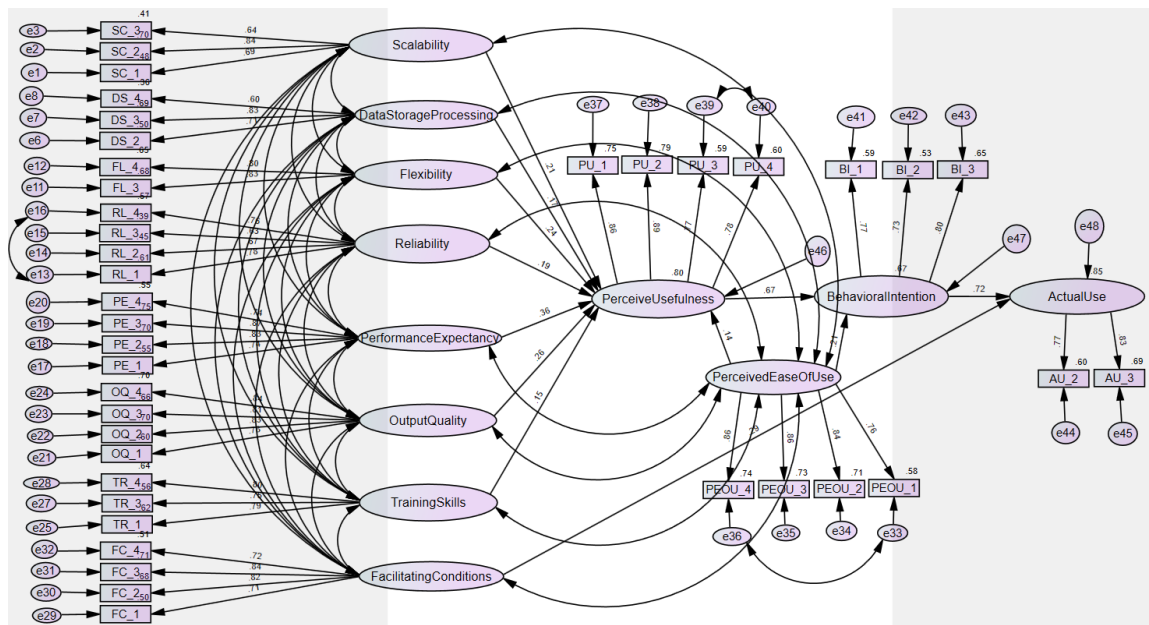


Figure 5: Path Diagram (SEM) of the Final Research Model

Figure 5 shows the R-squared values for PU, BI, and AU are 80, 67, and 85 respectively.

Table 37: Summary of Overall Path Model

Fit Indices	Overall Path Model
	SEM (12 Variables: 40 items)
χ^2 (df)	1228.474 (689)
CMIN	1.783
IFI	.941
TLI	.932
CFI	.940
RMSEA	.047

The path diagram (SEM) of the final research model in Figure 5 show below standard regression weights (Table 38).

Table 38: Path Model Standard Regression Weights

Constructs	Path	Standardized Regression Estimates
Perceived Ease of Use (PEOU)	PEOU → PU	.141
Reliability (RL)	RL → PU	.191
Performance Expectance (PE)	PE → PU	.360
Data Storage & Processing (DS)	DS → PU	.168
Training & Skills (TR)	TR → PU	.149
Scalability (SC)	SC → PU	.208
Output Quality (OQ)	OQ → PU	.261
Flexibility (FL)	FL → PU	.243
Perceived Usefulness (PU)	PU → BI	.667
Perceived Ease of Use (PEOU)	PEOU → BI	.206
Behavioral Intention (BI)	BI → AU	.721
Facilitating Conditions (FC)	FC → AU	.292

5.5 Discriminant Validity

The discriminant validity is one of the most important validities of survey responses in terms of construct values. The discriminant principle state that the measures of different constructs should not correlate highly with each other. The correlations

comparisons should provide evidence that the items on the two constructs discriminate. Discriminant validity measures whether the measure of each construct is distinct and different from the measures of other constructs. In order to demonstrate the discriminant validity of the construct, it is important to show that construct measures are unidimensional (Saleh, 2006). To determine discriminant validity, the literature suggests that squared correlations estimates (i.e., R^2) between each pair of constructs must be less than AVE values of individual constructs. In other words, the square roots of each construct's AVE must be higher than the correlation coefficients of each pair of constructs (Fornell & Larker, 1981). Also, the correlation estimate of each inter-construct must be lower than 0.80 (Bagozzi et al., 1991). Table 39 shows the discriminant validity results. The AVE values are derived from CFA metrics shown in Table 35. The factor correlation estimates consisting of correlations among exogenous variables are derived from SEM model shown in Figure 5 (Correlations – Group Number 1 Default Model).

Table 39: Discriminant Validity Analyses

Correlations	Factor Correlation Estimates	Correlation Squared (r-squared)	AVE1 AVE2 (AVEs should be > r-squared)	AVE1 AVE2 square roots should be > Correlation estimates
SC <--> DS	0.698	0.487	0.524 0.548	0.730 0.740
SC <--> PE	0.602	0.362	0.524 0.636	0.730 0.798
SC <--> RL	0.691	0.477	0.524 0.544	0.730 0.738
SC <--> FL	0.667	0.445	0.524 0.625	0.730 0.791
SC <--> OQ	0.517	0.267	0.524 0.665	0.730 0.815
SC <--> TR	0.516	0.266	0.524 0.606	0.730 0.779
SC <--> PEOU	0.384	0.147	0.524 0.692	0.730 0.832
SC <--> FC	0.533	0.284	0.524 0.600	0.730 0.775
DS <--> PE	0.630	0.397	0.548 0.636	0.720 0.797
DS <--> RL	0.632	0.399	0.548 0.507	0.720 0.738

DS <--> FL	0.721	0.519	0.548 0.625	0.740 0.791
DS <--> OQ	0.560	0.313	0.548 0.665	0.740 0.815
DS <--> TR	0.542	0.294	0.548 0.606	0.740 0.779
DS <--> PEOU	0.420	0.176	0.548 0.692	0.740 0.832
DS <--> FC	0.534	0.285	0.548 0.600	0.740 0.775
PE <--> RL	0.729	0.531	0.636 0.544	0.797 0.712
PE <--> FL	0.711	0.506	0.636 0.625	0.797 0.791
PE <--> OQ	0.786	0.618	0.636 0.665	0.797 0.815
PE <--> TR	0.701	0.491	0.636 0.606	0.797 0.779
PE <--> PEOU	0.675	0.456	0.636 0.692	0.797 0.832
PE <--> FC	0.675	0.456	0.636 0.600	0.797 0.775
RL <--> FL	0.731	0.534	0.544 0.625	0.738 0.791
RL <--> OQ	0.636	0.404	0.544 0.665	0.738 0.815
RL <--> TR	0.636	0.404	0.544 0.606	0.738 0.779
RL <--> PEOU	0.544	0.296	0.544 0.692	0.738 0.832
RL <--> FC	0.606	0.367	0.544 0.600	0.738 0.775
FL <--> OQ	0.658	0.433	0.625 0.665	0.791 0.815
FL <--> TR	0.653	0.426	0.625 0.606	0.791 0.779
FL <--> PEOU	0.532	0.283	0.625 0.692	0.791 0.832
FL <--> FC	0.598	0.358	0.625 0.600	0.791 0.775
OQ <--> TR	0.760	0.578	0.665 0.606	0.815 0.779
OQ <--> PEOU	0.691	0.477	0.665 0.692	0.815 0.832
OQ <--> FC	0.772	0.596	0.665 0.600	0.815 0.775
TR <--> PEOU	0.574	0.329	0.606 0.692	0.779 0.832
TR <--> FC	0.664	0.441	0.606 0.600	0.779 0.775
PEOU <--> FC	0.657	0.432	0.692 0.600	0.813 0.775

Table 39 shows the inter-construct correlation coefficients are lower than the square roots of the corresponding constructs' AVEs. In other words, the squared correlation estimate (i.e., R^2) for each inter-construct is lower than the AVEs of each construct. Inter-construct values of each construct pair also falls below the threshold value of .80. Since we did not violate anything in convergent and discriminant validity, we are going to assume our nomological validity is also good – overall validity.

Chapter 6 Hypotheses Testing and Discussion

This chapter discusses the outputs of the proposed model of this research and the results of hypotheses testing. This research is destined to identify the antecedents of big data technology acceptance. The results of the path model show 10 direct paths and two indirect paths. Eight independent variables have direct path to the dependent variable perceived usefulness (PU). They include scalability (SC), data storage and processing (DS), flexibility (FL), output quality (OQ), performance expectancy (PE), reliability (RL), training, and skills (TR), and perceived ease of use (PEOU). The independent variable, perceived ease of use (PEOU), and one dependent variable, perceived usefulness (PU) point to the dependent variable, behavioral intention to use (BI). Finally, independent variable, facilitating conditions (FC), and behavioral intention to use (BI) point to actual use (AU).

6.1 Hypotheses Testing

In this research, the primary question was what factors influence the big data technology acceptance which was elaborated in chapter one. In chapter three the hypotheses were developed. In this chapter, we discuss the results of the SEM model. The outputs of the model show R-squared values of .80, .67, and .85 for PU, BI, and AU respectively. Here we discuss the hypothesized path results of the final model. These terms are used to identify the independent and dependent variables of this model:

--SC = Scalability (IV)

--DS = Data Storage and Processing (IV)

--FL = Flexibility (IV)

--OQ = Output Quality (IV)

--PE = Performance Expectancy (IV)

--RL = Reliability (IV)

--TR = Training and Skills (IV)

--FC = Facilitating Conditions (IV)

--PEOU = Perceived Ease of Use (IV)

--PU = Perceived Usefulness (DV)

--BI = Behavioral Intention to Use (DV)

--AU = Actual Use (DV)

Table 40: Path Model Estimates

Hypotheses	Paths	SEM Output: Proposed Model				Results*
		Estimate (β)	S.E.	C.R. (t)	p-value	
H1: Scalability in terms of Hadoop scale-out-storage system will have a positive effect on perceived usefulness.	SC → PU	.241	.083	2.907	.004	Supported
H2: Data storage and processing have a positive effect on perceived usefulness.	DS → PU	.198	.089	2.219	.027	Supported
H9: Hadoop's flexibility to consolidate data from various sources to single place (HDFS) have a positive effect on perceived usefulness of Hadoop.	FL → PU	.257	.091	2.827	.005	Supported
H7: Data analytics capability is positively related to perceived usefulness of Hadoop.	DA → PU	.239	.342	.700	.484	Not Supported
H10: Output Quality are positively related to the perceived usefulness of Hadoop.	OQ → PU	.286	.090	3.168	.002	Supported

H4: Performance Expectancy/Usability is positively related to perceived usefulness of Hadoop.	PE → PU	.433	.103	4.185	***	Supported
H6: Reliability is positively related to perceived usefulness of Hadoop.	RL → PU	.249	.100	2.490	.013	Supported
H5: Security and Privacy is positively related to perceived usefulness of Hadoop.	SP → PU	.027	.099	.276	.783	Not Supported
H8: Training and required skills are positively related to perceived usefulness of Hadoop.	TR → PU	.180	.087	2.079	.038	Supported
H11: Functionality is positively related to perceived usefulness of Hadoop.	FN → PU	-.274	.295	-.930	.352	Not Supported
H14a: Perceived Ease of Use (PEOU) have positive effect on Perceived Usefulness (PU).	PEOU → PU	.116	.045	2.561	.010	Supported
H14b: Perceived Ease of Use (PEOU) have positive effect on Behavioral Intention to use Hadoop (BI).	PEOU → BI	.163	.052	3.154	.002	Supported
H13: Perceived Usefulness (PU) have positive effect on Behavioral Intention to use Hadoop (BI).	PU → BI	.645	.070	9.156	***	Supported
H12: Facilitating Conditions have positive effect on attitude toward using Hadoop.	FC → AU	.366	.083	4.411	***	Supported
H3: Cost effectiveness is positively related to adoption of Hadoop.	COST → AU	-.019	.055	-.344	.731	Not Supported
H15: Behavioral Intention (BI) is positively related to Actual Use (AU) of Hadoop.	BI → AU	.748	.080	9.394	***	Supported

*Results Supported as Significance Level: $p \leq .001$, $p \leq .01$, and $p \leq .05$.

The values in the above table reflects the output of Regression Weights: (Group number 1 - Default model) under the Estimates tab.

6.1.1 Scalability and Perceived Usefulness

Scalability is a new factor introduced to this model. This factor was not used in past research. For robust technologies like the one in big data (Hadoop), scalability does

matter when very large volume and complex data are handled (Menon & Sarkar, 2016). Path model results (Table 40) shows Scalability is significantly correlated with Perceived Usefulness, one of the highly correlated independent variables in the model.

The hypothesis test shows 95% confidence ($\beta = .24$, significant at $p \leq .01$). The p-value of 0.004 is smaller than the α of .05 (Table 40). The $p\text{-value} = 0.004 < \alpha = .05$. A $p\text{-value} < \alpha$ (i.e., critical value) is statistically significant. Alpha is usually defined as a 5% level of significance and based on the consensus of the researchers – a 5% probability of incorrectly rejecting the hypothesis is acceptable (based on this data set – to be conservative). If our p-value is lower than alpha, we conclude that there is a statistically significant difference between groups. That is there is less than 5% probability that the null is true. The C.R. value of 2.9 falls outside 2-std (1.96) under a 95% confidence interval. The null hypothesis appears implausible. As a researcher, we really want to reject the null hypothesis, because that is as close as we can get to proving the alternative hypothesis is true. The null hypothesis is rejected here. There is a strong positive correlation between scalability (SC) and perceived usefulness (PU). The experts in the qualitative study of this research have correctly identified it as a significant variable of Hadoop adoption. Industry papers also suggest scalability as an important factor of Hadoop adoption.

The term scalability has been widely used in industry when it comes to buying or using technology. Due to a lack of scalability, we experienced a scalability crisis in large-scale websites, eBay, healthcare.gov (Carr, 2013). Scalability and performance have

received special attention in the software performance review journals as well (Krishnamurthy & Koziolk, 2016). In the data management field, we experience that some database systems cannot expand beyond a certain data size limit. This makes companies switch to another database system. Ariyachandra and Watson (2010) propose that database architecture selection should be based on scalability. Most of the conventional database systems are not built on top of a scalable system except the Teradata database system (Malak and Brown, 2015; Rahman and Sutton, 2013).

In big data space, due to a large volume of data, scalability plays an important role (García-Gil et al., 2017; Lourenco et al., 2015; Menon & Sarkar, 2016). Hadoop is considered a highly scalable storage platform (Nemschoff, 2013). Big data technology and database systems experts of the qualitative study of this research selected scalability as the number one factor for further study as part of this research. Thirty-five of the forty (88%) participants who participated in the qualitative study voted for this factor for study. The performance and scalability challenges are apparent in platform as a Service (PaaS) cloud applications, and network topology (Krishnamurthy & Koziolk, 2016), to name a few. Malaka and Brown (2015) report that scalability is one of the technological challenges that is faced in the data analytics domain. Chen et al. (2015) propose measures of scalability relating to frame theory. Industry papers on big data technologies highlight scalability as one of the important elements of the Hadoop framework (Aye & Thein, 2015; Borthakur, 2007; Lourenco et al., 2015; Nemschoff, 2013).

Scalability has not been part of any IS theory or model. This technological factor has not been tested using any technology acceptance model in general and TAM (Hameed et al., 2012; Hess et al., 2014; Lee et al., 2003) in particular. To the best of our knowledge, this is the first survey-based research that uses scalability as an independent variable under TAM. Our model successfully validates scalability as a predictor variable of the technology acceptance model which exerts influence on perceived usefulness (PU). Future researchers might revalidate this factor as an independent variable of TAM.

6.1.2 Data Storage and Processing, and Perceived Usefulness

This factor is proposed as a new factor in this research. This factor has not been used in past research as part of TAM. Based on the empirical results, this factor emerges as one of the most important factors of Hadoop adoption. The hypothesis test shows a 95% confidence interval ($\beta = .20$, significant at $p \leq .05$). The p-value of 0.027 is smaller than the α of .05 (Table 40). For a significance level of 0.05, the C.R. value of 2.219 exceeds 1.96, which is significant. This ratio speaks for rejecting the null hypothesis. The null hypothesis appears not plausible. Hence, the null hypothesis is rejected. There is a strong positive correlation between 'data storage and processing' (DS) and 'perceived usefulness' (PU). The path model shows that this newly introduced construct has a 17% influence (estimates) on PU.

Organizations have been accumulating large amounts of data for years and years. This data could be internal transactional data of an organization or it could be external data related to an organization's business. With the emergence of online

business, social networking tools, and the advancement of data-generating technologies, organizations are encountering the growth of data volume. These data help in producing insights that revolutionize managerial decision-making (Tambe, 2014). In the past, this data used to be structured data. Now, most of the social media data are unstructured. To store and process, the large volume of data more sophisticated tools are technologies are needed. The exponential data growth necessitates robust data storage and processing of those data efficiently. To address this challenge, emerging big data technologies are thought to play a critical role (Aye & Thein, 2015; Chauhan & Murphy, 2013; Rahman et al., 2014). The Hadoop distributed file system (HDFS) is considered a scalable mass storage system along with MapReduce, its processing engine (Dolev et al., 2019; Shvachko et al., 2010).

This factor has been identified as the number two important factor by the expert-panel of the qualitative study of this research. Thirty-two of the forty (80%) participants who participated in the qualitative study voted for this factor to be included in the research model. The data analysis of the survey responses validates that data storage and processing capability (DS) has a significant influence on the perceived usefulness of the technology acceptance model of this research. This is the first time this factor has been identified as an independent variable of the TAM. Prior research using TAM focused on lightweight technologies. In the data management field, having this factor as a predictor variable for technology acceptance is justified. We hope that the

future researchers in the data management discipline will further study this factor to establish substantial theoretical and empirical support.

6.1.3 Flexibility and Perceived Usefulness

Flexibility is an important term in the software industry. As the software industry is making significant progress and robust systems are being built companies look for flexibility of a system before buying it. Hill (2011) has provide a good definition of flexibility: “When it is used to describe a whole system, flexibility normally refers to the ability for the solution to adapt to possible or future changes in its requirements.” The experts of the qualitative study of this research finds this variable to be an important factor in Hadoop adoption. The extant literature suggests that this factor has not been used in TAM (Lee et al., 2003) or any other IS model before. The hypothesis test shows that the 95% confidence interval for the mean difference ($\beta = .26$, significant at $p \leq .01$). The p -value = $0.005 < \alpha = .05$. The C.R. value of 2.827 is greater than the significance level of 1.96. The null hypothesis appears implausible. The null hypothesis is rejected. There is a strong positive correlation between ‘flexibility’ (FL) and ‘perceived usefulness’ (PU). This construct has a 24% influence (std. reg. estimate) on the perceived usefulness (PU).

Fichman and Kemerer (1993) report that innovation attributes play an important role in adoptions by an organization. The extant literature shows the importance of software flexibility. Scherrer-Rathje and Boyle (2012) have identified five dimensions of enterprise systems flexibility including system connectivity, process integration,

hierarchical integration, user-customizability, and consistency. Gebauer and Lee (2008) emphasize the importance of software flexibility in terms of operational efficiency and long-term effectiveness of an enterprise system. The authors assert that the more an enterprise software system provides flexibility-to-use the more it provides a good fit in relation to characteristics of the business process (Gebauer & Lee, 2008). Byrd and Turner (2000) suggest flexibility as an important capability of information technology infrastructure. The authors report that a flexible IT infrastructure is positively related to the competitive advantage of an organization.

Based on the meta-analysis of 303 studies, Sabherwal and Jeyaraj (2015) observe that firms that take initiative to adopt new technology and make IT alignment find a stronger relationship between IT investment and the business value of information technology. In the data management domain, Hadoop enables us to integrate and access a new source of data, both structured and unstructured, which helps to draw new insights and derive business value. Thus, Hadoop serves a wide variety of purposes including internet and systems log processing, building recommendation systems, building a robust machine learning capability, enabling fraud detection, and conventional data warehousing (Nemschoff, 2013). This factor has not been used in IS theory in general and the technology acceptance model in particular (Hameed et al., 2012; Lee et al., 2003). The expert panel of our qualitative study selected this factor as the number nine factor with 24 (60%) of 40 experts voted for it to be included in the research model. The statistical results of the final survey responses successfully

validated this factor as an independent variable of our model. This factor has a positive influence on perceived usefulness.

6.1.4 Data Analytics Capability and Perceived Usefulness

Data analytics capability in big data space is meant for data analysis of Hadoop's processing engine and machine learning capability using the ML libraries. Hadoop is popular due to its capability to capture and store a very large volume of both structured and unstructured data in its distributed file system (HDFS). Its machine learning libraries are capable to do a robust machine learning model based on a large volume and in many cases a complete set of data. Perhaps that is why the experts in the qualitative study of this research voted for this factor to be part of the current research model. The hypothesis test shows a 95% confidence interval for the mean difference. The p-value of 0.484 is greater than the α of .05 (Table 40). The p-value = 0.354 (initial iteration value) $> \alpha = .05$. The p-value of $> .05$ means not statistically significant. The C.R. value is 0.926 which falls between -1.96 and 1.96, which is not under a 95% confidence interval. We fail to reject the null hypothesis. There is no strong positive correlation between 'data analytics capability' (DA) and 'perceived usefulness' (PU). This factor is non-significant, most probably, Hadoop's main component itself is not a specific tool used for data analytics. However, future researchers might try this variable with a new set of data.

The extant literature has no reference to the use of this factor by any IS theory or model (Hameed et al., 2012; Lee et al., 2003). On the other hand, the latest industry papers on big data suggest the importance of data analytics capability of big data

technology including Hadoop (Abbasi et al., 2016; Akoka et al., 2017; Gandomi & Haider, 2015). The expert panel for our qualitative study also recommends that this factor be included in the research model for further study. However, the statistical analysis of our survey data failed to validate this factor. The single measurement model and CFA results have passed this factor in terms of internal consistency but, the SEM model failed the test. Due to the importance of this factor in the data management field we recommend that this be further tested as part of the technology acceptance model with a new set of sample sizes.

6.1.5 Output Quality and Perceived Usefulness

Output quality should reflect the correct data and be traceable all way back to where it was generated. Output quality also refers to the ease of understanding the information. In the data management space, the output should be reliable and accurate (Baesens et al., 2016). The output quality construct is part of Davis' TAM2 model (Davis, 1989; Holden & Karsh, 2010) as an exogenous variable. The findings of this study results are consistent with theoretical underpinnings as well as findings of several past studies. Path model results suggest the output quality construct has a 26% (std. reg. estimate) influence on PU. The hypothesis test shows a 95% confidence interval for the mean difference ($\beta = .29$, significant at $p \leq .01$). The p-value of .002, means the p-value is less than .01. The p-value of .002 is smaller than the α of .01 (Table 40). The p-value = .002 < $\alpha = .01$. The result of this variable states that with 99% confidence the 'output quality has an influence on 'perceived usefulness.' The null hypothesis is rejected. There is a

strong positive correlation between 'output quality' (OQ) and 'perceived usefulness' (PU).

Davis et al. (1992) used this measure to understand the Extrinsic and Intrinsic Motivation to Use Computers in the Workplace which got published in the journal of applied social psychology. Later, Venkatesh and Davis (2000) proposed this factor as part of TAM2, as a theoretical extension to the model, which appeared in *Management Science*, a leading IS journal. This factor is set to influence perceived usefulness in the model. By output quality, the authors meant to say that how well a system can perform the tasks which match the job goals of users of technology in an organization. The authors also assert that users would be inclined to use a system that is capable to deliver the highest output quality (Venkatesh & Davis, 2000; Wixom et al., 2001). Thus, output quality remains to be a significant determinant of perceived usefulness. Subsequently, this factor along with the TAM2 model was validated by many other researchers (Chismar & Wiley-Patton, 2003; Venkatesh & Bala, 2008). Chismar and Wiley-Patton (2003) successfully validate the TAM2 along with output quality to understand the physicians' intention to use the Internet-based health applications. They report that the output quality and perceived usefulness explain 59% of the variance of usage intentions by pediatricians. Roca et al. (2006) validated the output quality along with TAM2 in their study of e-learning continuance intention. They report that output quality and perceived usefulness are critical to the success of the e-learning system. Our

research model has successfully tested the output quality as a predictor of perceived usefulness. So, this result is consistent with the findings of the extant literature.

6.1.6 Performance Expectancy and Perceived Usefulness

“Performance expectancy is defined as the degree to which an individual believes that using the system will help him or her to attain gains in job performance” (Venkatesh et al., 2003, p. 448). The hypothesis test shows a 95% confidence interval for the mean difference ($\beta = .43$, significant at $p \leq .001$). The p-value of *** (i.e., less than .001) is smaller than the α of .05 (Table 40). The p-value = *** < $\alpha = .05$. The critical ratio of 4.185 is statistically highly significant because of the conventional .05 cutoff level for the statistical significance of 1.96. The C.R. value is, in fact, greater than 2.58, which is a 99.99% confidence interval. So, the null hypothesis is rejected. There is a strong positive correlation between ‘performance expectancy’ (PE) and ‘perceived usefulness’ (PU). The performance expectancy construct has a 36% (std. reg. estimate) influence on PU. This construct was examined and retained by previous research as well. The findings of this study results are consistent with theoretical underpinnings as well as findings of several past studies (Venkatesh et al., 2003).

The performance expectancy construct was introduced by Venkatesh et al. (2003) as part of a “consolidated” technology acceptance model, UTAUT. In this model, the authors theorized that four constructs play a dominant role as determinants of user acceptance and usage behavior: performance expectancy, effort expectancy, social influence, and facilitating conditions. Obviously, performance expectancy construct was

identified as one of the dominant constructs. The authors present that performance expectancy construct is the strongest predictor of intention with item loadings between .88 and .94 (Venkatesh et al., 2003). Subsequently, the construct along with UTAUT was tested by many researchers using a variety of applications including E-government services, clinical decision support system, tablet PC, internet, web-based learning environment, social media and smartphone applications (Aldhaban, 2016; Venkatesh et al., 2012; Venkatesh et al., 2016). Aldhaban (2016) reports that the performance expectancy construct shows the standard regression weight value of 0.339 to determine the intention to use the smartphone. The expert panel of our qualitative study selected this construct as the number 10 independent variable in order of rank to be included in the research model. Our research model shows this construct has a standard regression weight of 0.360. The statistical results of our model show this construct have a positive relationship with the perceived usefulness.

6.1.7 Reliability and Perceived Usefulness

Reliability is the “ability of an apparatus, machine, or system to consistently perform its intended or required function or mission, on-demand and without degradation or failure” (Business Dictionary, 2020). In big data, the reliability factor relates to data volume and velocity characteristics. Reliability is a new construct introduced to this research model. This construct has a 19% (std. reg. estimate) influence on PU. The hypothesis test shows a 95% confidence interval for the mean difference ($\beta = .25$, significant at $p \leq .05$). The p-value of 0.013 is smaller than the α of .05 (Table 40). The

p-value = 0.014 < α = .05. The C.R. value of 2.490 is greater cutoff level for statistical significance of 1.96. The null hypothesis is rejected. There is a strong positive correlation between 'reliability' (RL) and 'perceived usefulness' (PU).

Based on the extant literature (Hameed, 2012; Lee et al., 2003; Zhang and Pham, 2000), this construct has not been tested by IS theories or models in general and technology acceptance models in particular. In the data management field, ensuring the availability of data or no data loss in any circumstance is critical for an organization's sensitive data. Reliability is also critical from a data consistency standpoint. In many cases, data cannot be reproduced. In big data domain, the Hadoop distributed file system (HDFS) keeps multiple copies of the same data in more than one node (Shvachko et al., 2010). This ensures data availability even when one particular node fails. Thus, the Hadoop file system is considered a reliable data management system. The expert panel of the qualitative study of this research has selected this construct as the number six independent variables to be added to the research model for further study. The model has validated this construct with a positive relation to perceived usefulness. This is the first time this construct has been tested as part of the technology acceptance model.

6.1.8 Security and Privacy, and Perceived Usefulness

This construct was not retained in the final model as it failed to pass the confidence interval. The hypothesis test does not show it to falls under a 95% confidence interval for the mean difference. The p-value of 0.783 is greater than the α of .05 (Table 40). The p-value = 0.560 (initial run) > α = .05. The C.R. value of .099 is greater than -1.96 and less

than 1.96 statistical level of significance .05. We fail to reject the null hypothesis. There is no strong positive correlation between 'security and privacy' (SP) and 'perceived usefulness' (PU). It is a bit surprising result that this construct failed the test. Data security and privacy has become important these days. It is worth testing this construct in a future research.

The extant literature shows that this construct is important from the standpoint of data privacy and security (Menon and Sarkar, 2016; Moody et al., 2018; Wu et al., 2017). This concern is more relevant when it comes to big data as this data comes from social media. Personal information needs to be protected (Tsai et al., 2015). In healthcare data, privacy is very important (Viceconti et al., 2015; Wu et al., 2017). In the financial sector, data security is important. This construct has not been used by the technology acceptance model. However, given the data security and privacy has become very important in today's world it is worth testing this construct as part of future research with another set of data.

6.1.9 Training and Skills, and Perceived Usefulness

Education and training are provided to make sure that employees, developers, knowledge workers learn how to use technology, write efficient code, and increase their skillset. In big data space, a new set of tools and technologies are used. Developers and knowledge workers need to increase their skill set as the existing skillset that they used for the conventional data management system is not enough. Using complex technology requires rigorous training (Rajan & Baral, 2015). Therefore, training is an important

factor for the successful implementation of big data technology (McAfee & Brynjolfsson, 2012). In the implementation of other complex technologies, it was found that lack of training was one of the important reasons for the failure of the implementation. Training and education make employees feel comfortable, make them productive, decrease stress, and increase confidence in their ability to use innovative technology. The extant literature suggests that knowledge workers' job performance has a positive relationship with rich use of knowledge management systems, knowledge sharing, and training (Zhang, 2017). This construct has a 15% (std. reg. estimate) influence on PU. The hypothesis test shows a 95% confidence interval for the mean difference ($\beta = .18$, significant at $p \leq .05$). The p-value of 0.038 is smaller than the α of .05 (Table 40). The p-value = 0.038 < $\alpha = .05$. The critical ratio of 2.079 greater than the cutoff level 1.96. The null hypothesis is rejected. There is a strong positive correlation between 'training and skills' (TR) and 'perceived usefulness' (PU).

Recent research on big data highlighted the firm value of big data investments relating to training (Tambe, 2014). There are many tools and technologies related to big data and these are a new set of tools that were not used in the processing and analysis of conventional structured data. Big data technical skill is needed in many areas including data extraction, data processing, machine learning, statistical analysis, learning MapReduce, or Spark programming. Hence, training is important. The developers need the skill set in at least one programming language such as java, python, R or Scala. In TAM research, training is found to be a significant predictor of perceived usefulness (Rajan &

Baral, 2015). Rajan and Baral (2015) report that training has a significant influence on perceived usefulness ($\beta = 0.202, p < 0.001$) in their study of the enterprise resource planning (ERP) tool, SAP. Gupta and George (2016) used a hierarchical model and validated the significance of technical skills ($b = 0.50, p < 0.001$) in achieving big data capability. Extant literature reveals that there is limited research conducted on this construct using TAM. There is non-TAM related research that calls for training needs in big data tools and technologies. Brown-Liburd et al. (2015) report that adequate training and skills play a critical role in adopting big data analytical tools. Malaka and Brown (2015) test the skill shortage in the TOE model related to research on big data analytics. The authors found a shortage of skills as one of the challenges in the adoption of big data analytics. In Hadoop adoption, our research model has found that training and skill construct significantly influence perceived usefulness. Prior to quantitative analysis, we conducted a qualitative study using an expert panel. Most of the expert panel members (63%) selected this factor to be included in our research model.

6.1.10 Functionality and Perceived Usefulness

In information systems (IS), functionality is defined as the aspects of a software or technology that can be provided to users to able to do something useful on the job. The functionality provides users the capability to do on the job tasks by using the software or system. Functionality refers to the features of the software product as well. There are cases in the software industry that high profile software or applications fail to perform its functions due to poor design and functionality. The author of this research is

currently using an industry software that is poorly developed and hence takes more than usual time to develop objects and make workable and have performance issues. We have introduced this construct to a research model based on the qualitative studies of this research. The extant literature suggests that this construct has not been used (Hameed et al., 2012; Lee et al., 2003). The hypothesis test does not show the 95% confidence interval for the mean difference. The p-value of 0.352 is greater than the α of .05 (Table 40). The p-value = 0.397 (initial run) $>$ α = .05. The C.R. value -.930 is greater than cutoff level of -1.96 and less than 1.96. We fail to reject the null hypothesis. There is no strong positive correlation between 'functionality' (FN) and 'perceived usefulness' (PU). I believe this factor was substituted by other capability factors such as scalability, data storage and processing, flexibility.

This construct has not been tested by any IS theory or model in general and TAM in particular (Hess et al., 2014; Lee et al., 2003). However, the expert panel of the qualitative study of this research found it an important factor in big data technology adoption. The individual measurement model and CFA results also validated this construct with strong internal consistency. However, the SEM model failed to validate this construct. Future researchers of TAM might explore this factor further with a different set of data.

6.1.11 Perceived Ease of Use and Perceived Usefulness

The perceived ease of use (PEOU) is a construct of Davis' TAM model. This construct has been repeatedly tested to prove its validity (Davis, 1989; Taylor & Todd, 1995).

Subsequently, much research on technology adoption found this factor influential in technology acceptance. The findings of this study's results are consistent with theoretical underpinnings as well as findings of several past studies (Davis, 1989; Venkatesh & Davis, 2000). In this research model for path analysis, the PEOU shows that it has a 14% influence (std. reg. estimate) on perceived usefulness (PU). The hypothesis test shows 95% confidence ($\beta = .12$, significant at $p \leq .01$). The p-value of 0.010 is smaller than the α of .05 (Table 40). The p-value = 0.010 < $\alpha = .05$. The C.R. value, 2.561 is greater than the 1.96 cutoff level of statistical significance. The null hypothesis is rejected. There is a strong positive correlation between 'perceived ease of use' (PEOU) and 'perceived usefulness' (PU).

This construct was developed by Davis (1993) as part of his original TAM model. It has two flows, with one link to perceived usefulness and the other links to attitude toward using. Davis (1993) reports that perceived ease of use has a very strong influence (0.63) on perceived usefulness compared to attitude toward use (0.13). The author also reports the perceived ease of use has a small direct effect on attitude toward use. This construct exerts its influence on actual system through perceived usefulness: $0.63 * (0.44 + 0.65 * 0.21) = 0.36$ while its influence on actual system use through attitude toward system use is $0.13 * 0.21 = 0.02$ (Davis, 1993). Rajan and Baral (2015) report that perceived ease of use is significantly related to perceived ease of use (beta=0.329, $p < 0.001$). This construct is supported by numerous research findings (Hess et al., 2014). Our model results show that perceived ease of use has a lower

statistical significance ($p < 0.05$) than perceived usefulness ($p < 0.01$). While both these core constructs are statistically significant, our findings indicate that managers and decision-makers consider the usefulness of big data technology, Hadoop is more important than its ease of use. Our model supports this construct along with many other research findings conducted using this construct (e.g., Hess et al., 2014; Lederer et al., 2000).

6.1.12 Perceived Usefulness and Behavioral Intention to Use

Perceived usefulness is an endogenous variable of Davis' original technology acceptance model, TAM (Davis, 1989). This is the core construct of Davis' model and has been used in much research. The path analysis results show that this construct has a 67% influence on behavioral intention to use (BI). The results of this model also show that this factor can explain 80% of the variance. The hypothesis test shows that the 95% confidence interval ($\beta = .65$, significant at $p \leq .001$). The p-value of *** is smaller than the α of .05 (Table 40). The p-value = *** $< \alpha = .05$. The C.R. value, 9.156 is greater than the cutoff value of 1.96, which is statistically highly significant with a 95% confidence interval. The C.R. value is even greater than 2.58, that is, 99.99% confidence interval. The null hypothesis appears not plausible. The null hypothesis is rejected. There is a strong positive correlation between 'perceived usefulness' (PU) and 'behavioral intention to use' (BI).

Perceived usefulness as a significant predictor of behavioral intention to use technology was supported in studies by Davis (1989, 1993), Adams et al. (1992), Igbaria

et al. (1995), Hendrickson et al. (1993), Hess et al. (2014), Brown et al. (2014), and many other researchers (see meta-analysis by Hess et al., 2014; Ma & Liu, 2004; Legris et al., 2003). The extant literature report that perceived usefulness is a major determinant in the U.S. workplace (Igbaria et al., 1995). After the introduction of TAM, Davis (1989) validated the perceived usefulness and perceived ease of use for assessing technology acceptance. The author reported alpha coefficients of .98 and .94 for perceived usefulness and perceived ease of use respectively (Davis, 1989). Subsequently, Adams et al. (1992) retested these two constructs and confirmed the validity and reliability of these scales. Hendrickson et al. (1993) conducted test-retest reliability of perceived usefulness and perceived ease of use scales. The authors report a high degree of test-retest reliability on these two constructs. Hess et al. (2014) conducted a meta-analysis of perceived usefulness, perceived ease of use, and behavioral intentions. As part of an extensive literature search, the authors reviewed 380 articles and reported high-reliability coefficients for perceived usefulness. Perceived usefulness is a core construct of our research model which is set to relate with behavioral intention to use Hadoop. Our test results found an AVE of .68 and composite reliability (CR) value 0.90. Venkatesh and Davis (2000) report 40%–60% of the variance in usefulness perceptions. Compared to that, our model explains 80% variance in usefulness perceptions. Our SEM model successfully tested this construct which is compliant with the findings of prior research.

6.1.13 Perceived Ease of Use and Behavioral Intention to Use

Perceived ease of use (PEOU) is another significant construct of Davis' original TAM model (Davis, 1989). This construct was thought to be an endogenous model but since the extant literature suggests that this construct was less influential compared to PU, this research uses this construct as an exogenous construct that is connected with PU and BI in the path model. This construct has a 21% (std reg. estimate) influence on behavioral intention to use (BI). The hypothesis test shows the 95% confidence interval for the mean difference ($\beta = .16$, significant at $p \leq .01$). The p-value of 0.002 is smaller than the α of .05 (Table 40). The p-value = 0.002 < $\alpha = .05$. The C.R. value, 3.154 is greater than the cutoff level 1.96 statistical significance. The null hypothesis is rejected. There is a strong positive correlation between 'perceived ease of use' (PEOU) and 'behavioral intention to use' (BI). The findings of this study results are consistent with theoretical underpinnings as well as findings of several past studies (Davis, 1989).

Perceived ease of use is a core construct of Davis' original TAM (Davis, 1989). Perceived ease of use as a significant predictor of perceived usefulness and behavioral intention to use technology was supported in studies by Hendrickson et al. (1993), Venkatesh (2000), Gefen and Straub (2000), Ma and Liu (2004), Venkatesh and Bala (2008), and many other researchers (Chin & Todd, 1995; Straub et al., 1995). In measuring system usage: Implications for IS Theory Testing. Perceived ease of use is linked to behavior intention to use both directly (PEOU \rightarrow BI) and indirectly (PEOU \rightarrow PU \rightarrow BI) which has extensive evidence in support of that (Venkatesh & Davis, 2000).

Rajan and Baral (2015) report that perceived ease of use is significantly related (beta=0.266, $p < 0.001$) to behavioral intention to use. Our research model shows that this construct significantly influences (p-value = $0.002 < \alpha = .05$) the behavioral intention to use. However, perceived ease of use has a lower statistical significance ($p < 0.05$) than perceived usefulness ($p < 0.01$) when it comes to influencing the behavioral intention to use. The results of our study are quite consistent with the results reported in recent research.

6.1.14 Facilitating Conditions and Actual Use

Facilitating conditions are meant to provide a wide base of support for the implementation of the technology and system. From big data technology, Hadoop context such supports to include vendor support (software upgrade, custom solutions) and infrastructure support from the internal IT department of a company to facilitate project implementation. The findings of this study's results are consistent with theoretical underpinnings as well as findings of several past studies (Venkatesh et al., 2003). This construct is part of the model, UTAUT introduced by Venkatesh et al. (2003). As part of the UTAUT model, Venkatesh et al. (2003, p. 453) defined this factor as "Facilitating conditions are defined as the degree to which an individual believes that an organizational and technical infrastructure exists to support the use of the system." Path model analysis results show that this construct has a 29% influence (std. reg. estimate) on actual use (AU). The hypothesis test shows the 95% confidence interval ($\beta = .37$, significant at $p \leq .001$). The p-value of *** is smaller than the α of .05 (Table 40). The p-

value = *** < $\alpha = .05$. The C.R. value of 4.441 is greater than the cutoff level .05 statistical significance with a 95% confidence interval. Since the C.R. value is greater than 2.58, that is, 99.99% confidence interval, the null hypothesis is rejected. There is a strong positive correlation between 'facilitating conditions' (FC) and 'actual use' (AU). Moddy et al. (2018) found this construct to be insignificant in their 'unified model of information security policy compliance' model. They commented that it failed the test in their information security model context but, speculated that this factor might pass the test for a more technically challenging action. This research found this construct significant for a complex and challenging technology like Hadoop.

The unified theory of acceptance and use of technology (UTAUT) consists of four key constructs which include performance expectancy, effort expectancy, social influence, and facilitating conditions (Venkatesh et al., 2003). Obviously, the facilitation condition is one of those key factors. A meta-analysis on this model report that this construct was validated and supported by an extensive number of research papers (Venkatesh et al., 2016). Park et al. (2011) hypothesized this construct as organization-level facilitating conditions and found it to explain a larger variance in technology acceptance. Our research model is also developed to Hadoop acceptance from the organizational context. Our model also supports this factor as it was supported by prior research. In this research, we take this construct as something that provides support for Hadoop programmers and analysts. This construct was also validated from an organizational context by Rajan and Baral (2015) to test an ERP system acceptance.

Aldhaban (2016) used this construct to test smartphone acceptance but it was not supported by his construct. The reason might be that smartphone use is very, very common, and does not need any technical support from the vendors. In big data technology adoption facilitating conditions is important since vendor support (e.g., Cloudera, MapR, etc.) is needed by many companies. Companies, especially small and medium-sized, might get customer support and new version upgrade with vendor support (Villars et al., 2011).

6.1.15 Cost-Effectiveness and Actual Use

There is common knowledge and perception that big data tools are cost-effective compared to traditional data management software systems. Typically, cost includes initial investment cost, operational expense, and training cost (Premkumar & Potter, 1995). Based on this understanding the experts of big data systems in the qualitative study of this research voted for this construct to be part of this research model. The hypothesis test does not show that the 95% confidence interval for the mean difference. The p-value of 0.731 is greater than the α of .05 (Table 40). The p-value = 0.731 (initial run) $< \alpha = .05$. The C.R. value of -.344 is greater than the cutoff value of -1.96 and less than 1.96 statistical significance. We failed to reject the null hypothesis. There is no strong positive correlation between 'cost-effectiveness' (COST) and 'actual use' (AU). Organization might not be sensitive to cost given benefits obtained.

This construct was used and successfully validated as part of TAM (Wu & Wang, 2005). This construct was used by researchers using other models as well. Phan and

Daim (2011) successfully validated it for mobile service acceptance. The expert panel of our qualitative study selected it to include it in the research model. Both single measurement models and CFA found this construct valid and reliable. However, the SEM model did not find it a significant influencer of Hadoop adoption. The reason might be that Hadoop is an open-source tool provided by Apache Hadoop. Many companies might find it cheaper compared to conventional data management software. Some companies might not find cost a major barrier. They might use it regardless of costs. They might find the benefits outweigh the cost incurred.

6.1.16 Behavioral Intention to Use and Actual Use

The behavioral intention is the outcome of dyadic behavioral trajectories: perceived usefulness and perceived ease of use. The path model results show that this construct has significant influence (72%) on the actual use of the system. Also, this research model shows that this construct can explain 67% variance. The hypothesis test shows the 95% confidence interval for the mean difference ($\beta = .75$, significant at $p \leq .001$). The p-value of *** is smaller than the α of .05 (Table 40). The p-value = *** < $\alpha = .05$. The C.R. value 9.394 is greater than the cutoff value of 1.96 statistical significance, 95% confidence interval. The C.R. value is even greater than 2.58 statistical significance, that is, 99.99% confidence. The null hypothesis is rejected. There is a strong positive correlation between 'behavioral intention' (BI) and 'actual use' (AU). The findings of this study results are consistent with theoretical underpinnings as well as findings of several past studies (Davis,1989).

This construct is one of the main constructs of TAM developed by Davis (1989). This construct is also used in a later model, UTAUT, developed by Venkatesh et al. (2003). Venkatesh and Bala (2008) incorporated this construct in TAM3 as well. This construct links to the dependent variable, actual use in all these technology acceptance models. Turner et al. (2010) conducted a meta-analysis consisting of 79 empirical studies results published as research articles. Their study shows behavioral intention is likely to be correlated with actual usage. The authors also commented that perceived usefulness and perceived ease of use might not be directly correlated with actual usage (Turner et al., 2010). This means behavioral intention is an important predictor between usefulness and ease of use, and actual usage (Brown et al., 2014). Rajan and Boral validate this construct ($\beta = 0.453$, $p < 0.001$) in their empirical study of ERP system adoption. The author report that the intention to use explained 20.5% of the variance of usage. Venkatesh and Davis (2000) report 34%–52% of the variance in usage intentions. In contrast, our model explains 67% of the variance of usage.

6.2 Controlling Common Method Biases

Both Benbasat and Barki (2007) and Straub and Burton (2007) comment that CMB has never been tested for TAM: "Our view of Benbasat and Barki's characterization of TAM as unassailable is that common methods bias has never been well tested and that TAM linkages may in fact be methodological artifacts" (Straub & Burton, 2007, p. 223).

Burton-Jones (2009) asserts that common method bias can lead to false conclusions.

The author provides a formal definition of that (Burton-Jones, 2009, p. 448):

“Method bias is the difference between the measured score of a trait and the trait score that stems from the rater, instrument, and/or procedure used to obtain score.”

Fuller et al. (2016) observe that researchers take steps to assess common method bias but almost no one reports problematic findings. The authors also comment that a few authors present evidence of bias due to common method bias. Sharma et al. (2009) present a meta-analysis-based technique to estimate the effect of common method variance. The extant literature indicates that compared to other disciplines the empirical studies of IS research have made a rare attempt to assess common method biases (Malhotra et al., 2006). In this research, make an effort to assess such biases. We have followed a few guidelines from the previous research (Podsakoff et al., 2003; Straub et al., 2004) about addressing the common method bias. Both procedural and statistical measures have been taken to control common method bias.

Table 41: Single Factor Total Variance Explained

Factor	Total Variance Explained					
	Total	Initial Eigenvalues		Extraction Sums of Squared Loadings		
		% of Variance	Cumulative %	Total	% of Variance	Cumulative %
1	17.347	43.367	43.367	16.796	41.991	41.991
2	2.396	5.991	49.358			
3	1.597	3.993	53.351			
4	1.371	3.428	56.779			
5	1.278	3.195	59.974			
6	1.117	2.793	62.768			
7	1.072	2.681	65.448			
8	1.000	2.499	67.948			
9	.874	2.186	70.134			

If the variance explained by single factor is less than 50% then no common method bias issue exists. Our test shows this value 41.99%. There are no significant issues of common method bias found in our study (Table 41). Therefore, it passed the test. The extant literature has a strong support of using single factor analysis to check common method bias issue (e.g. Moody et al., 2018).

6.3 Non-Response Error: Wave Analysis

The survey of this study was conducted with the initial invitation to participate in the survey followed by two reminders with intervals. That means we collected 349 responses in three waves with 170 responses as part of the initial invitation, 95 responses as part of the first reminder, and 84 responses as part of the second and last reminder. We have used SPSS ANOVA to perform wave analysis. The level of significance values was measured with a 95% confidence interval. If $p > 0.05$ we say that there was no statistically significant difference between respondents among the three waves of data collection. We define a null hypothesis (H_0) which means no difference between groups being studied. The default, null is correct until we have enough evidence to support rejecting the hypothesis. It is usually kind of a bummer when the null hypothesis is valid because it means we didn't find a difference. In this particular we look for no difference between the waves of survey responses. Hence, we are fine here. The below tables (Table 42 – Table 45) show p values > 0.05 for each construct and each of the items/measures under each construct. We failed to reject the null hypothesis (i.e., mean Initial response = first reminder response = second reminder response).

Prior research suggests that low response rates and non-response are an issue of survey-based research as it threatens the external validity (Armstrong & Overton, 1977; Pinsonneault & Kraemer, 1993; Sivo et al., 2006). The authors propose three post hoc (i.e., after survey, using survey responses) survey strategies to estimate nonresponse error: comparison of demographic and socio-economic difference, comparison of early and late respondents' difference, and weighting adjustments (Sivo et al., 2006).

It is reported that in IS discipline, the comparison between early and later respondents is widely used (Sivo et al., 2006; Aldhaban, 2016). Originally, this strategy was proposed by Armstrong and Overton (1977). Sivo et al. (2006) observe that many researchers do not take initiative to address nonresponse bias issues and then justify the low response rate issues by reporting that other IS researchers also report low response rates. We take this issue more seriously and hence make an attempt to use one of the strategies suggested by Armstrong and Overton (1977) and Sivo et al. (2006). We used a web analysis of different response webs. We used the ANOVA technique using IBM SPSS statistical software. The null hypothesis developed for this purpose was that all the waves of responses are the same. Our ANOVA test failed to reject the null hypothesis for all latent constructs responses. The test shows no significant differences between webs at the 0.05 significant level (Tables 42-45). Hence, we assert that the data collected in the survey three webs responses are the same. And thus, those who did not participate in the survey fall under the category of respondents who participated as part of the last reminders in data collection.

In this research, we have received 349 responses out of 10,500 sample size. This means the response rate is 3.32%. However, even though two Hadoop user groups show the total number of subscribers is 10,500, we strongly believe that in reality, a large number of users are not active members. Hence, we assert that practically our response rate would be much higher.

ANOVA

Table 42: Survey Wave Analysis - Perceived Usefulness

		Sum of Squares	df	Mean Square	F	Sig.
PU_1	Between Groups	.419	2	.210	.309	.734
	Within Groups	234.578	346	.678		
	Total	234.997	348			
PU_2	Between Groups	2.308	2	1.154	1.707	.183
	Within Groups	233.864	346	.676		
	Total	236.172	348			
PU_3	Between Groups	1.775	2	.887	1.523	.220
	Within Groups	201.584	346	.583		
	Total	203.358	348			
PU_4	Between Groups	.434	2	.217	.365	.694
	Within Groups	205.377	346	.594		
	Total	205.811	348			

ANOVA

Table 43: Survey Wave Analysis - Perceived Ease of Use

		Sum of Squares	df	Mean Square	F	Sig.
PEOU_1	Between Groups	1.019	2	.510	.397	.673
	Within Groups	444.224	346	1.284		

	Total	445.244	348			
PEOU_2	Between Groups	1.418	2	.709	.765	.466
	Within Groups	320.880	346	.927		
	Total	322.298	348			
PEOU_3	Between Groups	.611	2	.305	.326	.722
	Within Groups	323.699	346	.936		
	Total	324.309	348			
PEOU_4	Between Groups	.030	2	.015	.017	.983
	Within Groups	302.658	346	.875		
	Total	302.688	348			

ANOVA

Table 44: Survey Wave Analysis - Behavioral Intention

		Sum of Squares	df	Mean Square	F	Sig.
BI_1	Between Groups	.016	2	.008	.010	.990
	Within Groups	278.436	346	.805		
	Total	278.453	348			
BI_2	Between Groups	.969	2	.485	.470	.626
	Within Groups	356.985	346	1.032		
	Total	357.954	348			
BI_3	Between Groups	2.580	2	1.290	1.298	.274
	Within Groups	343.753	346	.994		
	Total	346.332	348			

ANOVA

Table 45: Survey Wave Analysis - Actual Use

		Sum of Squares	df	Mean Square	F	Sig.
--	--	----------------	----	-------------	---	------

AU_1	Between Groups	2.099	2	1.050	.657	.519
	Within Groups	552.818	346	1.598		
	Total	554.917	348			
AU_2	Between Groups	2.018	2	1.009	1.176	.310
	Within Groups	296.790	346	.858		
	Total	298.808	348			
AU_3	Between Groups	.805	2	.403	.488	.614
	Within Groups	285.315	346	.825		
	Total	286.120	348			

6.4 Summary of the Chapter

The hypotheses results show that eight of the 12 independent variables passed the test. These include 'scalability' (SC), 'data storage and processing' (DS), 'flexibility' (FL), 'output quality' (OQ), 'performance expectancy' (PE), 'reliability' (RL), 'training and skills' (TR) and 'facilitating conditions' (FC). Four independent variables could not pass hypothesis test: 'data analytics capability' (DA), 'security and privacy' (SP), 'functionality' (FN), and 'cost -effectiveness' (COST). Among four original TAM variables (that Davis identified), 'perceived ease of use' (PEOU) was used as an independent variable in this research and it passed the hypothesis test. Three other TAM factors include 'perceived usefulness' (PU), 'behavioral intention to use' (BI), and 'actual use' (AU), all of which passed the hypothesis test. The path model results show that actual use (AU) can explain 85% of the variances. Prior studies validated PU and PEOU by showing that the TAM measures can explain 48.7% of the variance in self-reported system use (Dillon & Morris, 1996). Extant literature also reports that the behavioral intention construct in

TAM was able to explain 34%–52% of the variance (Venkatesh & Davis, 2000) and 52% of the variance (Taylor & Todd, 1995) respectively. Straub et al. (1995) report a result of their empirical study of perceived systems use with 49% explained variance. Later, the UTAUT model by Venkatesh et al. (2003) showed that it explained 72% variance. Compared to past research results, our model is able to explain a much higher percentage of variance in usage intention (67%) and 85% in actual use (AU).

It is said that perfection is not always attainable, but we can make our best attempt at excellence. With these high number variances, we believe we have achieved excellence!

Chapter 7 Conclusions, Research Contributions, Limitations, Research Direction

This study explores what factors influence big data technology (Hadoop) adoption. For any organization, the motivation behind adopting new technology is to (a) increase efficiency, (b) reduce cost, and (c) save money (Kohli et al., 2012; Mithas et al., 2011). These motivations are assumed. Having said that, what factors are the organizations looking for in new technology? Perhaps technological capability plays an important role. This has implications for perceived usefulness (PU) of new technology or innovation. During the factor selection process in the qualitative study of this research, the expert panels' participants had been specifically asked as to what makes technology useful. The development and test of our TAM-based model with new factors advance theory and research of the technology acceptance model.

This research examines a host of factors that influence a firm whether to adopt or not adopt the big data technology, Hadoop. Based on a qualitative study this research selected a dozen factors, out of 32, to use them as exogenous variables of the research model. A survey instrument was developed based on construct items from extant literature and also based on several new items relevant to big data technology. An online survey was administered using the survey tool, Qualtrics. Two big data user groups were used which consist of a sample of ten thousand respondents. Those who participated in the survey come from major industries including software/internet services, financial services, healthcare, consulting/professional services, telecommunication, manufacturing, retail, insurance, advertising/marketing, and

transportation/logistics (Table 15). The respondents' profile includes Hadoop engineers/application developers, Hadoop administrators, big data architects/enterprise architects, data scientists, data analysts, big data/information technology (IT) managers, chief information officers, and big data program managers (Table 14).

Four hundred two subjects responded to an email survey about big data technology acceptance out of which 349 responses were found to be complete and sufficient for the statistical analysis. The structural equation modeling (SEM) software, AMOS v26 was used to conduct statistical analysis. The model found eight exogenous variables as significant predictors for the adoption of Hadoop. These factors include scalability, data storage and processing capability, flexibility, reliability, performance expectancy, output quality, training and skills, and facilitating conditions (Figure 4 & 5). The SEM model also found four other exogenous variables to be non-significant. Hence, these factors were rejected: data analytics capability, security and privacy, functionality, and cost-effectiveness. Three of the exogenous variables had been used in past research: output quality, performance expectancy, and facilitating conditions. All these three variables are found to be significant contributors to Hadoop adoption, in this research. This shows consistency between extant literature and the current study results. This research makes a contribution by investigating and testing existing IS theory in a new information technology context. We extended the TAM through the addition of four new external variables. This is a significant contribution to theory and knowledge. There are some counter-intuitive findings as well. Four other new variables are found to

be non-significant in influencing Hadoop adoption: data analytics capability, security and privacy, functionality, and cost-effectiveness. Future research might take these variables into consideration to understand them further.

Lee et al. (2003) list a few limitations in TAM studies based on the meta-analysis of 101 articles published between 1986 and 2003. **First**, the authors report that some researchers use student sample/ university environment to reflect the real working environment. In our research, we have used industry experts who have hands-on experience in using big data technologies. We have used big data professionals in qualitative studies, a pilot study survey, and an actual full-length survey. **Second**, the authors (Lee et al., 2003) report that some researchers use single subject or restricted subjects such as “only one organization, one department, MBA students.” Contrary to that our research uses Hadoop user group members who spread across all prominent industries in the continental United States (see Table 15 for details). And those survey respondents have a few distinct job roles in Hadoop platforms or in the organization (see Table 14 for details). **Third**, another limitation reported was the measurement problems such as the use of single-item scales for a newly developed construct and hence, low validity of the construct and measure. We have introduced a few new independent variables to TAM, but we made sure those variables are represented with at least four items. **Fourth**, some research papers reported low variance scores without explaining the causation of the model (Lee et al., 2003). Our model accurately explains the variances for perceived usefulness, behavioral intention to use, and actual usage of

Hadoop. **Fifth**, some researchers conduct a survey with small sample size such as performing SEM analysis with less than 100 samples. Pundits suggest that SEM analyses need to be performed with at least a sample size of 200. Our research model is developed using SEM and we used 349 samples. However, the data of this survey is as good as the survey responses provided by the subset of the sample of this research.

7.1 Theoretical Contribution

Without theory, there is no knowledge. In the words of W. Edwards Deming:

"Experience teaches nothing. In fact, there is no experience to record without theory...

Without theory there is no learning ..." (Neave, 1990). Thus, our endeavor should be to

try our best to understand things in terms of theory. Our research should be destined

to make a contribution to theory. To that end, our current research has made the best

effort to make a contribution to theory in the technology acceptance field.

A literature review reveals that a few data-storage/DSS-related constructs are applied to TAM (Benbasat & Bakri, 2007; Lee et al., 2003). There is a lack of study that incorporates multiple data-storage/DSS-related constructs to a single study (Kwon et al., 2014). This research makes a contribution to the literature in several ways. **First**, this research has incorporated a few new variables to the model to understand effects and also their relationships to the TAM model (perceived ease of use, perceived usefulness, and behavioral intention). These external variables include scalability, data storage and processing capability, flexibility, and reliability. No other TAM-based research has tested these variables (Lee et al., 2003). We assert that this is a significant contribution to the

body of knowledge since our study successfully tested these new variables to the adoption of a technologically complex system. And this research has proven that these external factors influence the latent variables of TAM, their statistical relationship, and their strength. This research provides insights into how technological characteristics play a role in a large and robust technology like Hadoop. This provides new evidence of taking the technological capabilities into consideration in acquiring new technology. The new factors that are accepted by this research model help us realize the complexity of such robust technologies.

Second, this study applied the technology acceptance theory (TAM) to examine factors of big data technology acceptance. The findings of the study have shown that TAM is valid in a new and technologically complex system implementation (that is, a big data technology context). The technology acceptance model has been mostly applied to understand users' intentions (Holden & Karsh, 2010) from an individual's usage context (e.g., smartphone). This research provides an outcome from industrial/ organizational level users' acceptance context (big data).

Third, it provides an insight into how a complex technology like Hadoop implementation can lead to changes in employees' job characteristics and lead to the urgency of providing more training to the employees. Understanding this important change of work, and the required training and skill is of importance to the theory and practice.

Fourth, it provides us an understanding of the factors (scalability, reliability, flexibility, data storage and processing capability, and performance expectancy) that can influence buying of technologies or platforms like Hadoop and the functioning of employees' job. Many software projects fail due to limitations or inefficient software system. Many organizations switch to another technology due to the bandwidth issue of the existing technology relating to performance, scalability, flexibility, reliability. Thus, we contribute to the IT and data management platform implementation literature by examining the role of these factors.

Fifth, this research presents several new factors that have not been used before. These include scalability, reliability, flexibility, data storage and processing, and training. Prior research tested the TAM using light technologies such as fax machines and word processors. As technologies have proliferated in recent years and in data management space, data volume has increased the new technologies in these areas demanding more capability and performance in terms of scalability, flexibility, and robustness. These new findings are important contributions to our existing knowledge of TAM and IT implementation that was largely overlooked in past research.

Sixth, it contributes to the literature on scalability by identifying a few important measures. This has a great implication for data management platforms. It contributes to the scalability theory (Chen et al., 2015) and systems theory (Paetow et al., 2005).

Finally, perhaps our research would be the first theoretical-based empirical study that examined the effects of certain data management variables in TAM. This is

also expected to provide both academia and practitioners with an understanding of the impact of big data from technological, environmental, and organizational contexts. This study provides findings as to how big data technology overcomes some known limitations of conventional data storage systems (e.g., relational databases).

Our research is based on data collected from actual Hadoop users who have industry job experience in big data field. We developed and validated our model based on industry context (Chiasson & Davidson, 2005). Thus, we evaluate the boundaries of existing IS theory and contribute to enhancing the existing TAM model with new external factors.

7.2 Implications for Practitioners

Prior research suggests that many firms are at the early stage of big data adoption due to a lack of understanding and empirical evidence of the impact of big data technology on organizations (Bean, 2020; Gartner, Inc., 2015). This empirical study provides IT practitioners with insights about whether big data is capable of increasing the data-driven decision performance of organizations.

First, from a managerial perspective, this research provides managers pre- and post-implementation to-dos. This provides companies with insights as to what technology characteristics and capabilities to look for when buying a complex technology. It also provides managers with action plans such as training developers and knowledge workers in order to lessen the negative effects and improve skillsets. Such training will ensure their proper utilization of the newly acquired technology, Hadoop.

Previous research on TAM and UTAUT found that factors like performance expectancy, output quality, and facilitative conditions (Davis, 1993; Venkatesh et al., 2003) are needed to provide seamless access to quality information in an enterprise data management platform. This research introduces new dimensions (e.g., technological) to such data management platforms that are required to handle today's new data (e.g., unstructured data) in an enterprise data management platform.

Second, managers need to be mindful of hiring skilled developers and knowledge workers before planning to implement Hadoop technology in their organization. Existing developers and knowledge workers who work in traditional data management technologies might not have the skills to use Hadoop. They might need the training to brush up their programming language skills. These developers need to be proficient at least in one of the programming languages - Java, Python, Scala, R, etc. (Davenport & Patil, 2012). The managers might expect that the developers and knowledge workers will show low productivity and initially decreases in quality. Some of them who are not confident enough to use this technology might be moved to other job roles. In many cases, new and complex enterprise systems implementation causes major changes in terms of job characteristics and interpersonal relationships in employees' work-life (Bala, 2008).

Third, managers should make sure a facilitating condition exists to support Hadoop developers, knowledge works, data analysts. The Hadoop vendors could be considered to get the latest version software and some custom applications. An internal

IT infrastructure team should exist to facilitate and help in undertaking Hadoop-based project implementations. Facilitating condition refers to the provision of support for users that can influence system utilization.

Finally, big data provides the capability to capture and process a large volume of data. By using Hadoop, organizations might be able to put together internal data (e.g., transactional or dimension data) and external data (e.g., social media and other sources) in HDFS (Rahman, 2018b). That might help business organizations to get a 360-degree view of data and thus improve organizations' decision performance. Given big data is able to consolidate all kinds of data (structured and unstructured) from both internal and external sources the reliability and output quality of those data need to be understood. This is important as data-driven decision making has a dependency on data quality (Baesens et al., 2016). In his seminal paper in Harvard Business Review, David Garvin (Garvin, 1987) pointed out eight dimensions of quality as part of strategic quality management. This research has validated the output quality construct and hence, it speaks for the importance of big data storage systems. The results of this study might be helpful and encouraging for new companies in adopting big data. The new findings of this study are expected to be valuable to big data vendors as well as other stakeholders (e.g., semiconductor manufacturers who supply special server processors for big data processing).

7.3 Implications for Researchers

Previous academic research on big data focused on technical algorithms or system development (Kwon et al., 2014). Since the emergence of big data terminology in the last decade a lot of research was undertaken to develop big data technologies, tools, and techniques (Landset et al., 2015). There are also numerous experiments and use-cases conducted to prove the capability and efficiency of those individual tools and techniques. That indeed made significant research contributions to this new discipline. However, there is very limited research conducted toward understanding the acceptance of big data by business organizations. In this area, one study was conducted by Kwon et al. (2014). That research only investigated the acceptance of big data from data quality and data usage standpoint (internal versus external data usage). This research provides other aspects of big data that are important in understanding the adoption of big data. They include technological variables (e.g., scalability, flexibility, reliability, data storage, and processing capability), organizational variables (e.g., training and skills), and environmental variables (e.g., facilitating condition). With these new variables having been identified by survey results as significantly influential variables, this research is able to contribute to big data adoption research.

7.4 Limitations

This study examined the factors that influence the big data technology adoption. This research was able to identify a few new factors. Despite the potency of these factors, the findings of this study need to be thought about with caution and they warrant

future research attention. This study investigated a limited number of variables out of a pool of three dozen of variables (provided in this dissertation). Future research might consider investigating other variables as well as retesting the ones found influential by this study. In generalizing the findings of this study, the following items need to be verifiably carefully:

First, the findings of this study rely on respondents' self-reported data. Some researchers suggest that self-reported usage does not always reflect actual usage (Burton-Jones, 2009; Szajna, 1996). The concern is that self-reported usage might distort and inflate causal relation between independent and dependent variables (Lee et al., 2003; Podsakoff et al., 2003) and thus cause validity problems. This concern is the strongest when both exploratory variable and dependent variable data is collected from the same person (Podsakoff et al., 2003). Self-reported data is cited as one of the commonly reported limitations (Lee et al., 2003). Self-reported data is also considered as one of the reasons for the common method bias problem. To address this concern, we have conducted the Harman one-factor analysis to check whether variance in the data largely extrapolates to a single factor (Chang et al., 2010). Our study finds no such issue (Table 41). Nonetheless, future researchers might test this model by collecting data for predictor and criterion variables separately (Chang et al., 2010).

The **second** limitation of the study is that it collected data at a single point of time. The IS scholars call out to be careful about the generalization problem of such a single point of time study or collecting data from a homogenous group of subjects (Lee

et al., 2003). The extant literature reveals that in technology acceptance research there is a dominance of cross-sectional study. To avoid the risk of homogenous data collection, we used Hadoop user groups, the members of which belong to all major industries with responses from a variety of stakeholders. Further, to address the issue of cross-sectional study, future research might consider a longitudinal study of these variables. Given the user's perception and intention to change over a period of time, it is worth collecting data at several points of time to perform longitudinal comparisons (Lee et al., 2003).

Third, for the survey of this study, data were collected online from Hadoop User Groups in the United States. There were no individual-level contact numbers. The survey instrument was sent to the Hadoop User groups' address. While online data collection helped in terms of cost, it limits the generalizability of our findings as we do not know exactly what group of respondents did participate in the survey and what groups did not participate. Some populations who do not have internet access got excluded. Hence, future research should test the model with another group of respondents who are directly reachable.

Fourth, the survey responses were collected from many stakeholders (data scientists, data analysts, CTO, application developers, engineers; see Table 14 for details) - the professionals who actually used the tool. This is consistent with the observation that technical persons and consultants are the best people to get input in making the decision to buy a new technology (Wheelock, 2013). Therefore, the study

cannot be generalized as the responses are of the managers and other company executives.

7.5 Future Research Direction

This research has successfully validated the Davis' technology acceptance model along with a few new independent variables. The TAM has not been explored in the data management platform context in terms of independent variables, especially the technological ones. This research provides some insights and directions for future research. As this research has taken on some new challenges using extant as well as new constructs, this opens up avenues for further research.

First, this research has successfully validated a few new independent variables and made them be part of TAM. This is a great contribution to the theory and knowledge. However, it would be tough to make these variables to be part of mainstream TAM research if further research is not conducted. Hence, to give them a widespread validity, further studies on these new variables are warranted.

Second, this study has found four new factors non-significant (functionality, security and privacy, data analytics capability, and cost-effectiveness) even though the expert panel of the qualitative study voted for them and the CFA successfully validated them. These factors failed the SEM validation as part of the path model analysis. We conducted a survey consisting of 62 questions (IV and DV) for which 351 responses were received. The response rate per construct item was 5.63 (349/62). Still, future researchers might run this model with a large number of responses. Some researchers

suggest 10 responses per construct item (Suhr, 2006). Hence, 10 responses per construct item, that is, $62 * 10 = 610$ could be used to see if those four factors get validity. We aspire that this could be the source of new topics for future research.

Third, the survey instrument of this research was destined for the actual users who possess hands-on experience in using the Hadoop. As part of future research, this survey could be conducted using the first-line managers, mid-level managers, and executives of companies as well. This could provide us an insight as to whether collecting data from direct users versus company executives would make any difference. The data were collected from a technology capability and implementation perspective. Future research may investigate whether non-technical questions designed for company executives would make any difference.

Fourth, this study was conducted with data from users in U.S. companies. The results cannot be generalized to organizations outside of the United States. Hence, conducting a comparative analysis of big data technology use or intention to use in similar industries and alternative geographical areas could provide some useful insights.

Finally, big data is here to stay! Given the footprint of data everywhere we do not foresee a paradigm shift in the near future when it comes to big data. Big data technology might change for a good user experience. Research on big data and its technologies is expected to continue from both data-driven and theory-driven research standpoint (Maass et al., 2018).

References

- Abbasi, A., Sarker, S., & Chiang, R.H.L. (2016). Big data research in information systems: Toward an inclusive research agenda. *Journal of the Association for Information Systems*, 17(2). Article 3. DOI: 10.17705/1jais.00423.
- Aboelmaged, M.G. (2014). Predicting e-readiness at firm-level: An analysis of technological, organizational and environmental (TOE) effects on e-maintenance readiness in manufacturing firms. *International Journal of Information Management* 34(2014), 639–651.
- Abouzeid, A. Bajda-Pawlikowski, K., Abadi, D., Silberschatz, A., & Rasin, A. (2009). HadoopDB: An architectural hybrid of MapReduce and DBMS technologies for analytical workloads. *Proceedings of the VLDB '09*, August 24-28, 2009, Lyon, France.
- Adams, D., Nelson, R., & Todd, P. (1992). Perceived usefulness, ease of use and usage of information technology: a replication. *MIS Quarterly*, 16(2), 227-247.
- Agarwal, R., & Dhar, V. (2014). Editorial - Big data, data science, and analytics: The opportunity and challenge for IS research. *Information Systems Research*, 25(3), 443-448.
- Agarwal, R., & Prashad, J. (1999). Are individual differences germane to the acceptance of new information technologies? *Decision Sciences*, 30(2), 361-391.
- Ajzen, I. (1991). The theory of planned behavior. *Organizational Behavior and Human Decision Processes*, 50(2), 179–211.
- Akoka, J., Comyn-Wattiau, I., & Laoufi, N. (2017). Research on big data - A systematic mapping study. *Computer Standards & Interfaces*, 54, 105-115.
- Aldhaban, F. (2016). Exploratory study of the adoption and use of the smartphone technology in emerging regions: Case of Saudi Arabia. *PhD Dissertation in Technology Management*, Portland State University, 2016.
- Al-Jarrah, O.Y., Yoo, P.D., Muhaidat, S., Karagiannidis, G.K., & Taha, K. (2015): Efficient machine learning for big data: A review. *Big Data Research*, 2, 87-93.
- Amoako-Gyampah, K., & Salam, A.F. (2004). An extension of the technology acceptance model in an ERP implementation environment. *Information & Management*, 41, 731-745.
- Amos (2020). IBM® SPSS® Amos(TM) User Guide. Retrieved from: Ibm <http://amosdevelopment.com/webhelp/index.html?ifi2.htm>

- Amudhavel, J., Padmapriya, V., Gowri, V., LakshmiPriya, K., Prem Kumar, K., & Thiyagarajan, B. (2015). Perspectives, Motivations and Implications Of Big Data Analytics. *Proceedings of the 2015 International Conference on Advanced Research in Computer Science Engineering & Technology (ICARCSET 2015)*, March 2015. Article 34, 1–5.
- Anagnostopoulos, C., & Triantafillou, P. (2020). Large-scale predictive modeling and analytics through regression queries in data management systems. *International Journal of Data Science and Analytics*, 9, 17–55.
- Anderson, J.C., & Gerbing, D.W. (1988). Structural equation modeling in practice: A review and recommended two-step approach. *Psychological Bulletin*, 103(3), 411-423.
- Anderson, J.C., & Gerbing, D.W. (1991). Predicting the performance of measures in a confirmatory factor analysis with a pretest assessment of their substantive validities. *Journal of Applied Psychology*, 76(5), 732-740.
- Anderson, T.R. (2012). *Research methods for technology management and other fields*. Portland State University, OR, USA.
- Ariyachandra, T., & Watson, H. (2010). Key organizational factors in data warehouse architecture selection. *Decision Support Systems*, 49(2010), 200–212.
- Armstrong, J.S., & Overton, T.S. (1977). Estimating nonresponse bias in mail surveys. *Journal of Marketing Research*, 14(3), 396-402.
- Arts, J.W.C., Frambach, R.T., & Bijmolt, T.H.A. (2011). Generalizations on consumer innovation adoption: A meta-analysis on drivers of intention and behavior. *International Journal of Research in Marketing*, 28, 134–144.
- Atif, A., Richards, D., & Bilgin, A. (2012). Estimating non-response bias in web-based survey of technology acceptance: A case study of unit guide information systems. *Proceedings of the 23rd Australasian Conference on Information Systems (ACIS)*, December 3 – 5, 2012, Geelong, Australia.
- Atlas.ti (2017). Qualitative research. *Qualitative Data Analysis*. Atlas.ti. Retrieved from <http://atlasti.com/qualitative-research/>
- Aye, K.N., & Thein, T. (2015). A platform for big data analytics on distributed scale-out storage system. *International Journal of Big Data Intelligence*, 2(2), 127-141.
- Baesens, B., Bapna, R., Marsden, J.R., Vanthienen, J., & Zhao, J.L. (2016). Transformational issues of big data and analytics in networked business. *MIS Quarterly*, 40(4), 807-818.
- Bagozzi, R.P. (1982). A field investigation of causal relations among cognitions, affect, intentions, and behavior. *Journal of Marketing Research*, 19, 562-584.

- Bagozzi, R.P. (2007). The legacy of the technology acceptance model and a proposal for a paradigm shift. *Journal of the Association for Information Systems*, 8(4), 244-254.
- Bagozzi, R.P., & Yi, Y. (1988). On the evaluation of structural equation models. *Journal of the Academy of Marketing Science*, 16(1), 74–94.
- Bagozzi, R.P., Yi, Y., & Phillips, L.W. (1991). Assessing construct validity in organizational research. *Administrative Science Quarterly*, 36(3), 421-458.
- Bala, H.K. (2008). Nothing endures but change: Understanding employees' responses to enterprise systems implementation and business process change. *Doctoral Dissertation*, University of Arkansas. USA: ProQuest.
- Balac, N., Sipes, T., Wolter, N., Nunes, K., Sinkovits, B., & Karimabadi, H. (2013). Large scale predictive analytics for real-time energy management. *Proceedings of the 2013 IEEE International Conference on Big Data (BigData 2013)*, 657-664. October 6-9, 2013, Santa Clara, CA, USA.
- Brockmeier, J. (2011). Who wrote Hadoop? It's the community, stupid. Retrieved from <https://readwrite.com/2011/10/05/who-wrote-hadoop-its-the-commu/>
- Bardhan, I., Chen, H., & Karahanna, E. (2020). Connecting systems, data, and people: A multidisciplinary research roadmap for chronic disease management. *MIS Quarterly*, 44(1), 185-200.
- Barki, H., & Hartwick, J. (1994). Measuring user participation, user involvement, and user attitude. *MIS Quarterly*, 18(1), 59-82.
- Barlow, R.E. (1984). Mathematical theory of reliability: A historical perspective. *IEEE Transactions on Reliability*, 33(1), 16-20.
- Barney, J. (1991). Firm resources and sustained competitive advantage, *Journal of Management*, 17(1), 99-120.
- Barrett, P. (2007). Structural equation modeling: adjudging model fit. *Personality and Individual Differences*, 42(2007), 815-824.
- Bartlett, J.E., Kotrlik, J.W., & Higgins, C.C. (2001). Organizational research: Determining appropriate sample size in survey research. *Information Technology, Learning, and Performance Journal*, 19(1), 43 - 50.
- Basoglu, N., Daim, T., & Kerimoglu, O. (2007). Organizational adoption of enterprise resource planning systems: A conceptual framework. *Journal of High Technology Management Research*, 18, 73-97.
- Bean, R. (2020). Firms must overcome human barriers to enable data-driven transformation. *Forbes*. Retrieved from <https://www.forbes.com/sites/ciocentral/2020/01/02/firms-must-overcome-human-barriers-to-enable-data-driven-transformation/#5d8267f42a56>

- Benbasat, I., & Barki, H. (2007). Quo vadis, TAM? *Journal of the Association for Information Systems* 8(4), 211-218.
- Bentham, J. (1824/1987). An introduction to the principles of morals and legislation. In *J. S. Mill and J. Bentham, Utilitarianism and Other Essays*, Harmandsworth: Penguin.
- Bentler, P.M. (1990). Comparative fit indexes in structural models. *Psychological Bulletin*, 107, 238-246.
- Bentler, P.M. & Bonett, D.G. (1980). Significance tests and goodness of fit in the analysis of covariance structures. *Psychological Bulletin*, 88, 588-606.
- Bentler, P. M., & Chou, C. P. (1987) Practical issues in structural modeling. *Sociological Methods & Research*, 16, 78-117.
- Berengueres, J., & Efimov, D. (2014). Airline new customer tier level forecasting for real-time resource allocation of a miles program. *Journal of Big Data*, 1(3), 1-13.
- Bollen, K.A. (1989). A new incremental fit index for general structural equation models. *Sociological Methods and Research*, 17, 303-316.
- Bologa, A., Bologa, R., & Florea, A. (2010). Big data and specific analysis methods for insurance fraud detection. *Database Systems Journal*, 1(1), 30-39.
- Boomsma, A. (1982). The robustness of LISREL against small sample sizes in factor analysis models. *Systems under indirect observation: Causality, structure, prediction*, 1, 148-173.
- Borthakur, D. (2007). The Hadoop distributed file system: Architecture and design. White Paper. *The Apache Software Foundation*, 1-14.
- Bradford, J., & Saad, M., (2014). Towards a method for measuring absorptive capacity in firms. *International Journal of Technology Management and Sustainable Development*, 13(3), 237-249.
- Brown, S.A., Venkatesh, V., & Goyal, S. (2014). Expectation confirmation in information systems research: A test of six competing models. *MIS Quarterly*, 38(3), 729-756.
- Brown-Liburd, H., Issa, H., & Lombardi, D. (2015). Behavioral implications of big data's impact on audit judgment and decision making and future research directions. *Accounting Horizon*, 29(2), 451-468.
- Burton-Jones, A. (2009). Minimizing method bias through programmatic research. *MIS Quarterly*, 33(3), 445-471.
- Byrd, T.A., & Turner, D.E. (2000). Measuring the flexibility of information technology infrastructure: Exploratory analysis of a construct. *Journal of Management Information Systems*, 17(1), 167-208.

- Byrne, B.M. (2016). *Structural equation modeling with AMOS: Basic concepts, applications, and programming*. Routledge. Third Edition. June 24, 2016.
- Caesarius, L.M., & Hohenthal, J. (2018). Searching for big data: How incumbents explore a possible adoption of big data technologies. *Scandinavian Journal of Management, 34*, 129-140.
- Cao, M., Chychyla, R., & Stewart, T. (2015). Big data analytics in financial statement audits. *Accounting Horizon 29*(2), 423-429.
- Carr, D.F. (2013). How We'd Fix HealthCare.gov: Scalability Experts Speak. *InformationWeek*. Retrieved from <https://www.informationweek.com/healthcare/policy-and-regulation/how-wed-fix-healthcaregov-scalability-experts-speak/d/d-id/1112867>
- Ceci, F., Masini, A., & Prencipe, A. (2019). Impact of IT offerings strategies and IT integration capability on IT vendor value creation. *European Journal of Information Systems, 28*(6), 591-611.
- Chae, H.-C., Koh, C.E., & Park, K.O. (2018). Information technology capability and firm performance: Role of industry. *Information & Management, 55*(5), 525-546.
- Chang, S.-J., van Witteloostuijn, A., & Eden, L. (2010). From the editors: Common method variance in international business research. *Journal of International Business Studies, 41*, 178–184.
- Chardonens, T., Cudre-Mauroux, P., & Grund, M. (2013). Big data analytics on high velocity streams: A case study. *Proceedings of the 2013 IEEE International Conference on Big Data (BigData 2013)*, 784-787. October 6-9, 2013, Santa Clara, CA, USA.
- Chauhan, H., & Murphy J. (2013). Harnessing Hadoop: Understanding the big data processing options for optimizing analytical workloads. *Cognizant 20-20 Insights*, 1-8.
- Chau, P.Y.K., & Tam, K.Y. (1997). Factors affecting the adoption of open systems: An exploratory study. *MIS Quarterly, 21*(1), 1-24.
- Chen, D.Q., Preston, D.S., & Swink, M. (2015). How the use of big data analytics affects value creation in supply chain management. *Journal of Management Information Systems, 32*(4), 4-39.
- Chen, H., Chiang, R.H.L., & Storey, V.C. (2012). Business intelligence and analytics: From big data to big impact. *MIS Quarterly, 36*(4), 1165-1188, 2012.
- Chen, P.-T., Lin, C.-L., & Wu, W.-N. (2020). Big data management in healthcare: Adoption challenges and implications. *International Journal of Information Management, (2020)*, 1-11.

- Cheung, W., Chang, M. K., & Lai, V. S. (2000). Prediction of internet and world wide web usage at work: A test of an extended Triandis model. *Decision Support Systems*, 30(1), 83–101.
- Chiasson, M.W., & Davidson, E. (2005). Taking industry seriously in information systems research. *MIS Quarterly*, 29(4), 591-605.
- Chin, W.W. (1998). The partial least squares approach for structural equation modeling. *Modern Methods for business research*, 295-336.
- Chin, W.W., & Gopal, A. (1995). Adoption intention in GSS: relative importance of beliefs. *The DATABASE for Advances in Information Systems*, 26(2-3), 42-64.
- Chin, W.W., & Newsted, P.R. (1999). Structural equation modeling analysis with small samples using partial least squares. *Statistical strategies for small sample research*, 1, 307-341.
- Chin, W.W., & Todd, P.A. (1995). On the use, usefulness, and ease of use of structural equation modeling in MIS Research: A note of caution. *MIS Quarterly*, 19(2), 237-246.
- Chismar, W.G., & Wiley-Patton, S. (2003). Does the extended technology acceptance model apply to physicians. *Proceedings of the 36th Hawaii International Conference on System Sciences (HICSS'03)*, Hawaii, USA.
- Choi, J.K., & Ji, Y.G. (2015). Investigating the importance of trust on adopting an autonomous vehicle. *International Journal of Human-Computer Interaction*, 31, 692-702.
- Chowdhury, T., Ramineni, K., Rahman, N., Krishnan, M., Sharma, S., & Wong, J. (2015). Using big data to understand the impact of email on business. *Intel White Paper*. October 2015. 1-10.
- Chuttur, M. (2009). Overview of the technology acceptance model: Origins, developments and future directions. *All Sprouts Content*. 290.
- Cloudera, Inc. (2012). Ten common Hadoopable problems: Real-world Hadoop use cases. *Cloudera White Paper*. Cloudera.com, Palo Alto, CA, USA.
- Cochran, W.G. (1977). *Sampling techniques* (3rd Ed.). New York: John Wiley & Sons.
- Codd, E.F. (1970). A relational model of data for large shared data banks. *Communications of the ACM*, 13(6), 377-387.
- Columbus, L. (2017). 53% of companies are adopting big data analytics. *Forbes*. Retrieved from <https://www.forbes.com/sites/louiscolombus/2017/12/24/53-of-companies-are-adopting-big-data-analytics/#62fbcefc39a1>

- Côrte-Real, N., Ruivo, P., & Oliveira, T. (2020). Leveraging internet of things and big data analytics initiatives in European and American firms: Is data quality a way to extract business value? *Information & Management*, 57(1).
- Cronbach, L.J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika*, 16(3), 297-334.
- Das, T.K., & Kumar, P.M. (2013). BIG data analytics: A framework for unstructured data Analysis. *International Journal of Engineering and Technology*, 5(1), 153-156.
- Davenport, T.H., & Patil, D.J. (2012). Data scientist: The sexiest job of the 21st Century. *Harvard Business Review*, October 2012. 70-76.
- Davis, F.D. (1986). A technology acceptance model for empirically testing new end-user information systems: Theory and results, *Doctoral Dissertation*, MIT Sloan School of Management, Cambridge, MA, USA.
- Davis, F.D. (1989). Perceived usefulness, perceived ease of use, and user acceptance of information technology, *MIS Quarterly*, 13(3), 319-340.
- Davis, F.D. (1993). User acceptance of computer technology: system characteristics, user perceptions, *International Journal of Man-Machine Studies*, 38(3), 475-487.
- Davis, F.D., Bagozzi, R.P. & Warshaw, P.R. (1989). User acceptance of computer technology: A comparison of two theoretical models, *Management Science*, 35(8), 982-1003.
- Davis, F.D., Bagozzi, R.P., & Warshaw, P.R. (1992). Extrinsic and intrinsic motivation to use computers in the workplace. *Journal of Applied Social Psychology*, 22(14), 1111-1132.
- Davis, F.D., & Venkatesh, V. (1996). A critical assessment of potential measurement biases in the technology acceptance model: three experiments. *International Journal of Human-Computer Studies*, 45(1), 19-45.
- Business Dictionary (2020). Reliability. *Business Dictionary*. Retrieved from <http://www.businessdictionary.com/definition/reliability.html>
- Dillman, D.A., Smyth, J.D., & Christian, L.M. (2009). *Mail and internet surveys: The tailored design method* (3rd Ed.). New York: John Wiley and Sons, USA.
- Dillon, A., & Morris, M.G. (1996). User acceptance of new information technology: Theories and models, *Annual Review of Information Science and Technology*, 31, 3-32.
- Ding, L., Velicer, W.F., & Harlow, L.L. (1995). Effects of estimation methods, number of indicators per factor, and improper solutions on structural equation modeling fit indices. *Structural Equation Modeling: A Multidisciplinary Journal*, 2, 119-143, 1995.

- Dishaw, M.T. (1998). The construction of theory in MIS research. *International Journal of Information Management*, 7(1), 39-52.
- Dolev, S., Florissi, P., Gudes, E., Sharma, S., & Singer, I. (2019). A survey on geographically distributed big-data processing using MapReduce. *IEEE Transactions on Big Data*, 5(1), 60-80.
- Dong, J.Q., & Yang, C.-H. (2020). Business value of big data analytics: A systems-theoretic approach and empirical test. *Information & Management*, 57(1).
- Economist (2010). Data, data everywhere. *The Economist*, Special report, 2010.
- Eisenhardt, K.M., & Schoonhoven, C.B. (1996). Resource-based view of strategic alliance formation: Strategic and social effects in entrepreneurial firms. *Organizational Science*, 7(2), 136-150.
- Esteves, J., & Curto, J. (2013). A risk and benefits behavioral model to assess intentions to adopt big data. *Journal of Intelligence Studies in Business*, 3(2013), 37-46.
- Fichman, R.G., & Kemerer, C.E. (1993). Adoption of software engineering process innovations: The case of object orientation. *Sloan Management Review*, 34(2), 7-22.
- Fishbein, M., & Ajzen, I. (1975). *Beliefs, attitude, intention, and behavior: An Introduction to theory and research*. Reading, MA: Addison-Wesley.
- Fornell, C., & Larcker, D.F. (1981). Evaluating structural equation models with unobservable variables and measurement error. *Journal of Marketing Research*, 18(1), 39-50.
- Fuller, C.M., Simmering, M.J., Atinc, G., Atinc, Y., & Babin, B.J. (2016). Common methods variance detection in business research. *Journal of Business Research*, 69, 3192-3198.
- Gandomi, A., & Haider, M. (2015). Beyond the hype: Big data concepts, methods, and analytics. *International Journal of Information Management*, 35, 137-144.
- García-Gil, D., Ramírez-Gallego, S., García, S., & Herrera, F. (2017). A comparison on scalability for batch big data processing on Apache Spark and Apache Flink. *Big Data Analytics*, 2(1), 1-11.
- Gartner, Inc. (2015). Gartner survey highlights challenges to Hadoop adoption, *Gartner, Inc.*, Stamford, CT, USA (<https://www.gartner.com/en/newsroom/press-releases/2015-05-13-gartner-survey-highlights-challenges-to-hadoop-adoption>).
- Garvin, D.A. (1987). Competing on the eight dimensions of quality. *Harvard Business Review*, 65(6), 100 - 109.

- Garzo, A., Benczur, A.A., Sidlo, C.I., Tahara, D., & Wyatt, E.F. (2013). Real-time streaming mobility analytics. *Proceedings of the 2013 IEEE International Conference on Big Data (BigData 2013)*, 697--702. October 6-9, 2013, Santa Clara, CA, USA.
- Gebauer, J., & Lee, F. (2008). Enterprise system flexibility and implementation strategies: Aligning theory with evidence from a case study. *Information Systems Management*, 25(1), 71-82.
- Gefen, D., & Straub, D.W., (1997). Gender differences in the perception and use of e-mail: An extension to the technology acceptance model. *MIS Quarterly*, 21(4), 389-400.
- Gefen, D., & Straub, D.W. (2000). The relative importance of perceived ease of use in IS adoption: A study of e-commerce adoption. *Journal of the Association for Information Systems*, 1(1), 1-28.
- George, G., Haas, M.R., & Pentland, A. (2014). Big data and management. *Academy of Management Journal*, 57(2), 321-326.
- George, J. F (2004). The theory of planned behavior and internet purchasing. *Internet Research*, 14(3), 198–212.
- Gerbing, D.W., & Anderson, J.C. (1985). The effects of sampling error and model characteristics on parameter estimation for maximum likelihood confirmatory factor analysis. *Multivariate Behavioral Research*, 20, 255–271.
- Ghasemaghaei, M. (2019). Does data analytics use improve firm decision making quality? The role of knowledge sharing and data analytics competency. *Decision Support Systems*, 120, 14-24.
- Goes, P.B. (2014). Editor's comments - Big data and IS research. *MIS Quarterly*, 38(3).
- Gogus, C.G., & Ozer, G. (2014). The roles of technology acceptance model antecedents and switching cost on accounting software use. *Academy of Information and Management Sciences Journal*, 17(1), 1-24.
- Gray, D. (2014). Hadoop as open-source software: pros and cons. Dataconomy. Retrieved from <http://dataconomy.com/hadoop-open-source-software-pros-cons/>
- Grover, V., Lindberg, A., Benbasat, I., & Lyytinen, K. (2020). The perils and promises of big data research in information systems. *Journal of the Association for Information Systems*, 21(2).
- Gupta, B., Dasgupta, S., & Gupta, A. (2008). Adoption of ICT in a government organization in a developing country: *An empirical study*. *Journal of Strategic Information Systems*, 17, 140–154.
- Gupta, M., & George, J.F. (2016). Toward the development of a big data analytics capability. *Information & Management*, 53, 1049-1064.

- HadoopUserGroups (2019). *Hadoop User Groups*. Retrieved from <https://cwiki.apache.org/confluence/display/HADOOP2/HadoopUserGroups>
- Hagiu, A., & Wright, J. (2020). When data creates competitive advantage. *Harvard Business Review*, January–February 2020.
- Hair, J.F., Black, W., Babin, B.J., & Anderson, R.E. (2010). *Multivariate data analysis* (7th ed.): Pearson, 2010.
- Hameed, M.A., Counsell, S., & Swift, S. (2012). A conceptual model for the process of IT innovation adoption in organizations. *Journal of Engineering and Technology Management*, 29, 358-390.
- Harris, D. (2013). The history of Hadoop: From 4 nodes to the future of data. Retrieved from <https://gigaom.com/2013/03/04/the-history-of-hadoop-from-4-nodes-to-the-future-of-data/>
- Hartmann, P.M., Zaki, M., Feldmann, N., & Neely, A. (2014). Big data for big business? A taxonomy of data-driven business models used by start-up firms. *University of Cambridge*, 1-29.
- Hartwick, J., & Barki, H. (1994). Explaining the role of user participation in information system use. *Management Science*, 40(4), 440-465.
- Hendrickson, A.R., Massey, P.D., & Cronan, T.P. (1993). On the test-retest reliability of perceived usefulness and perceived ease of use scales. *MIS Quarterly*, 17(2), 227-230.
- Hess, T.J., McNab, A.L., & Basoglu, K.A. (2014). Reliability generalization of perceived ease of use, perceived usefulness, and behavioral intentions. *MIS Quarterly*, 38(1), 1-28.
- Hilbert, M. (2016). Big Data for development: A review of promises and challenges. *Development Policy Review*, 34(1), 135-174.
- Hill, D. (2011). Software flexibility. *Informative Architecture*. Retrieved from <https://informativearchitecture.wordpress.com/2011/10/04/software-flexibility/>
- Hodgson, A. (2011). *Essays on the evolution of healthcare technology*. Doctoral Dissertation of the Department of Economics, Graduate Division of the University of California, Berkeley, USA. Publisher: ProQuest LLC.
- Holden, R.J., & Karsh, B.-T. (2010). The technology acceptance model: Its past and its future in health care, *Journal of Biomedical Informatics*, 43, 159-172.
- Holmes-Smith, P., Coote, L., & Cunningham, E. (2004). Structural equation modeling: From the fundamentals to advanced topics, ACSPRI-Summer Training Program, Canberra, Australia.

- Holst, A. (2020). Big data adoption barriers among firms in U.S. and worldwide 2019. *Statista*. Retrieved from <https://www.statista.com/statistics/742983/worldwide-survey-corporate-big-data-adoption-barriers/>
- Hood-Clark, S.F. (2016). Influences on the use and behavioral intention to use big data. *Doctoral Dissertation of the School of Business and Technology, Capella University, USA*. Publisher: ProQuest LLC.
- Hooper, D., Coughlan, J., & Mullen, M. R. (2008). Structural equation modelling: Guidelines for determining model fit. *The Electronic Journal of Business Research Methods, 6*(1), 53-60.
- Hox, J.J., & Bechger, T.M. (1998). An introduction to structural equation modeling, *Family Science Review, 11*, 354-373.
- Hu, L., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling, 6*, 1–55.
- Hwang, H.-G., Ku, C.-Y., Yen, D.C., & Cheng, C.-C. (2004). Critical factors influencing the adoption of data warehouse technology: a study of the banking industry in Taiwan. *Decision Support System, 37*(2004), 1-21.
- Iacobucci, D. (2010). Structural equations modeling: Fit Indices, sample size, and advanced topics. *Journal of Consumer Psychology, 20*(2010), 90-98.
- IDC (2019). IDC forecasts revenues for big data and business analytics solutions will reach \$189.1 billion this year with double-digit annual growth through 2022. *IDC Press, FRAMINGHAM, Mass., April 4, 2019*.
- Lavrakas, P.J. (2008). Cluster sample. *Encyclopedia of Survey Research Methods*. SAGE Publishing. DOI: <https://dx.doi.org/10.4135/9781412963947.n67>
- Igbaria, M., Iivari, J., & Maragahh, H. (1995). Why do individuals use computer technology? A Finnish case study. *Information & Management, 29*, 227-238.
- Im, I., Hong, S., & Kang, M.S. (2011). An international comparison of technology of technology adoption: Testing the UTAUT model. *Information & Management, 48*(1), 1-8.
- Jain, P., Gyanchandani, M., & Khare, N. (2016). Big data privacy: A technological perspective and review. *Journal of Big Data, 3*(25), 1-25.
- Jelinek, M., & Bergey, P. (2013). Innovation as the strategic driver of sustainability: Big data knowledge for profit and survival. *IEEE Engineering Management Review, 41*(2), 14-22.

- Jetzek, T., Avital, M., & Bjorn-Andersen, N. (2019). The sustainable value of open government data. *Journal of the Association for Information Systems*, 20(6), Article 6.
- Jin, X., Wah, B.W., Cheng, X., & Wang, Y. (2015). Significance and challenges of big data research. *Big Data Research*, 2, 59-64.
- Kambatla, K., Kollias, G., Kumar, V., & Grama, A. (2014). Trends in big data analytics. *Journal of Parallel Distributed Computing*, 74, 2561-2573.
- Kapteyn, A. (1985). Utility and economics. *De Economist*, 133(1), 1–20.
- Karahanna, E., Straub, D. W., & Chervany, N. L. (1999). Information technology adoption across time: A cross-sectional comparison of pre-adoption and post adoption beliefs. *MIS Quarterly*, 23(2), 183–213.
- Kenny, D.A., & McCoach, D.B. 2003. Effect of the number of variables on measures of fit in structural equation modeling. *Structural Equation Modeling*, 10, 333-351.
- Kiron, D., Ferguson, R.B., & Prentice, P.K. (2013). From value to vision: Reimagining the possible with data analytics. Research Report, *MIT Sloan Management Review*, 1-19. 2013.
- Kline, R.B. (2015). *Principles and practice of structural equation modeling (4th Ed.)*. New York: The Guilford Press.
- Kohli, R., Devaraj, S., & Ow, T.T. (2012). Does information technology investment influence a firm's market value? A case of non-public traded healthcare firms. *MIS Quarterly*, 36(4), 1145-1163.
- Kranjc, J., Podpecan, V., & Lavrac, N. (2013). Real-time data analysis in CloudFlows. *Proceedings of the 2013 IEEE International Conference on Big Data (BigData 2013)*, 15-22. October 6-9, 2013, Santa Clara, CA, USA.
- Krejcie, R.V., & Morgan, D.W. (1970). Determining sample size for research activities. *Educational and Psychological Measurement*, 30, 607 – 610.
- Krishnamurthy, D., & Koziolk, A. (2016). Special issue on challenges in software performance. *Performance Evaluation Review*, 43(4), 1-2.
- Krugman, P., & Wells, R. (2017). *Macroeconomics (5th Ed.)*. New York, NY: Worth Publishers.
- Kuan, K.K.Y., & Chau, P.Y.K. (2001). A perception-based model for EDI adoption in small businesses using a technology-organization-environment framework. *Information & Management*, 38, 507-521.
- Kwon, O., Lee, N., & Shin, B. (2014). Data quality management, data usage experience and acquisition intention of big data analytics. *International Journal of Information Management*, 34(2014) 387–394.

- Landset, S., Khoshgoftaar, T.M., Richter, A.N., & Hasanin, T. (2015). A survey of open source tools for machine learning with big data in the Hadoop ecosystem. *Journal of Big Data*, 2(24), 1-36.
- LaValle, S., Lesser, E., Shockley, R., Hopkins, M.S., & Kruschwitz, N. (2011). Big data, analytics and the path from insights to value. *MIT Sloan Management Review*, 52(2).
- Leavitt, N. (2013). Bringing big analytics to the masses. *Computer*, 46(1), 20-23.
- Lederer, A.L., Maupin, D.J., & Sena, M.P., Zhuang, Y. (2000). The technology acceptance model and the world wide web. *Decision Support Systems*, 29(3), 269-282.
- Lee, I. (2017). Big data: Dimensions, evolution, impacts, and challenges. *Business Horizons*, 60(3), 293-303.
- Lee, Y., Kizar, K.A., & Larsen, K.R.T. (2003). The technology acceptance model: Past, present, and future. *Communications of the Association for Information Systems*, 12, 752-780.
- Legris, P., Ingham, J., & Colletette, P. (2003). Why do people use information technology? A critical review of the technology acceptance model. *Information & Management*, 40, 191-204.
- Li, R., Ruan, S., Bao, J., Li, Y., Wu, Y., Hong, L., & Zheng, Y. (2020). Efficient path query processing over massive trajectories on the cloud. *IEEE Transactions on Big Data*, 6(1), 66-79.
- Liker, J.K., & Sindi, A.A. (1997). User acceptance of expert systems: a test of the theory of reasoned action. *Journal of Engineering and Technology Management*, 14(1997), 147-173.
- Likert, R. (1932). A technique for the measurement of attitudes. *Archives of Psychology*, 22, 5-55.
- Lourenco, J.R., Cabral, B., Carreiro, P., Vieira, M., & Bernardino, J. (2015). Choosing the right NoSQL database for the job: a quality attribute evaluation. *Journal of Big Data*, 2(18), 1-26.
- Lozano, M.G., Brynielsson, J., Franke, U., Rosell, M., & Vlassov, V. (2020). Veracity assessment of online data. *Decision Support Systems*, 129.
- Lu, R., Zhu, H., Liu, X., Liu, J.K., and Shao, J. (2014). Toward Efficient and Privacy-Preserving Computing in Big Data Era. *IEEE Network*, 46-50.
- Luck, D.J., & Rubin, R.S. (1987). *Marketing research* (7th Ed.). Englewood Cliffs, NJ: Prentice-Hall, Inc.
- Lycett, M. (2013). 'Datafication': making sense of (big) data in a complex world. *European Journal of Information Systems*, 22(4), 381-386.

- Maass, W., Parsons, J., Purao, S., Storey, V.C., & Woo, C. (2018). Data-driven meets theory-driven research in the era of big data: Opportunities and challenges for information systems research. *Journal of the Association for Information Systems*, 19(12).
- Ma, Q., & Liu, L. (2004). The technology acceptance model: A meta-analysis of empirical findings. *Journal of Organizational and End User Computing*, 16(1), 59-72.
- MacCallum, R. C., Browne, M. W., & Sugawara, H. M. (1996). Power analysis and determination of sample size for covariance structure modeling. *Psychological Methods*, 1, 130-149.
- MacCallum, R.C., Widaman, K.F., Zhang, S., & Hong, S. (1999). Sample size in factor analysis. *Psychological Methods*, 4(1), 84-99.
- MacKenzie, S.B., Podsakoff, P.M., & Podsakoff, N.P. (2011). Construct measurement and validation procedures in MIS and behavioral research: Integrating new and existing techniques. *MIS Quarterly*, 35(2), 293-334.
- Malaka, I., & Brown, I. (2015). Challenges to the organisational adoption of big data analytics: A case study in the south African telecommunications industry. *Proceedings of the ACM 2015 Annual Conference of the South African Institute of Computer Scientists and Information Technologists (SAICSIT 2015)*. September 28-30, 2015. Stellenbosch, South Africa.
- Malhotra, N.K., Kim, S.S., & Patil, A. (2006). Common method in IS research: A comparison of alternative approaches and a reanalysis of past research. *Management Science*, 52(12), 1865-1883.
- Marr, B. (2015). *Big data: Using SMART big data, analytics and metrics to make better decisions and improve performance* (1st Ed.). USA: Wiley.
- Marsh, H.W., & Bailey, M. (1991). Confirmatory factor analyses of multi trait-multimethod data: A comparison of alternative models. *Applied psychological measurement*, 15, 47-70, 1991.
- Marsh, H.W., Hau, K.-T., Balla, J.R., & Grayson, D. (1998). Is more ever too much? The number of indicators per factor in confirmatory factor analysis. *Multivariate Behavioral Research*, 33, 181-220, 1998.
- Marsh, H. W., Hau, K. -T., & Wen, Z. (2004). In search of golden rules. *Structural Equation Modeling*, 11(3), 320-341.
- Martin, K.E. (2015). Ethical issues in the big data industry. *MIS Quarterly Executive*, 67-85.
- Maruyama, G.M. (1998). Basics of structural equation modeling, USA: Sage Publications.

- Mathieson, K. (1991). Predicting user intention: comparing the technology acceptance model with the theory of planned behavior. *Information Systems Research*, 2(3), 173–191.
- McAfee, A., & Brynjolfsson, E. (2012). Big data: The management revolution. *Harvard Business Review*, 61-68.
- McNeely, C.L., & Hahm, J. (2014). The big (data) bang: Policy, prospects, and challenges. *Review of Policy Research*, 31(4), 304-310.
- McQuitty, S. (2004). Statistical power and structural equation models in business research. *Journal of Business Research*, 57, 175-183.
- Medina-Quintero, J.-M., & Chaparro-Peláez, J. (2007). The impact of the human element in the information systems quality for Decision Making and User Satisfaction. *Journal of Computer Information Systems*, 48(2), 44-52.
- Menon, S., & Sarkar, S. (2016). Privacy and big data: scalable approaches to sanitize large transactional databases for sharing. *MIS Quarterly*, 40(4), 963-981.
- Mesgari, M., & Okoli, C. (2019). Critical review of organisation-technology sensemaking: towards technology materiality, discovery, and action. *European Journal of Information Systems*, 28(2), 205-232.
- Mikalef, P., Krogstie, J., Pappas, I., & Pavlou, P. (2020). Exploring the relationship between big data analytics capability and competitive performance: The mediating roles of dynamic and operational capabilities. *Information & Management*, 57(2), 103169.
- Mithas, S., Ramasubbu, N., & Sambamurthy, V. (2011). How information management capability influence firm performance? *MIS Quarterly*, 35(1), 237-256.
- Moktadir, M.A., Ali, S.M., Paul, S.K., & Shukla, N. (2019). Barriers to big data analytics in manufacturing supply chains: A case study from Bangladesh. *Computers & Industrial Engineering*, 128(2019), 1063-1075.
- Monroe, M.C., & Adams, D.C. (2012). Increasing response rates to web-based surveys. *Journal of Extension*, 50(6). Article 6TOT7.
- Monteith, J.Y., McGregor, J.D., & Ingram, J.E. (2013). Hadoop and its evolving ecosystem. *Proceedings of 5th International Workshop on Software Ecosystems (IWSECO 2013)*. Potsdam, Germany, June 11, 2013.
- Morgado, F.F.R., Meireles, J.F.F., Neves, C.M., Amaral, A.C.S., & Ferreira, M.E.C. (2017). Scale development: ten main limitations and recommendations to improve future research practices. *Psicologia: Reflexão e Crítica*, 30(3), 1-20.
- Mulaik, S.A., James, L.R., Van Alstine, J., Bennett, N., Lind, S., & Stilwell, C.D. (1989). Quantitative methods in psychology. *Psychological Bulletin*, 105, 430-445.

- Moody, G.D., Siponen, M., & Pahnla, S. (2018). Toward a unified model of information security policy compliance. *MIS Quarterly*, 42(1), 285-311.
- Moore, G.C., & Benbasat, I. (1991). Development of an instrument to measure the perceptions of adopting an information technology innovation. *Information Systems Research*, 2, 192-222.
- Moore, G.C., & Benbasat, I. (1996). Integrating diffusion of innovations and theory of reasoned action models to predict utilization of information technology by end-users. *Diffusion and Adoption of Information Technology*, 132-146.
- Morris, M.G., & Dillon, A. (1997). How user perceptions influence software use. *IEEE Software*, 14(4), 58-65.
- Najafabadi, M.M., Villanustre, F., Khoshgoftaar, T.M., Seliya, N., Wald, R., & Muharemagic, E. (2015). Deep learning applications and challenges in big data analytics. *Journal of Big Data*, 2(1), 1-21.
- Nambiar, R., Sethi, A., Bhardwaj, R., & Vargheese, R. (2013). A look at challenges and opportunities of big data analytics in healthcare. *Proceedings of the 2013 IEEE International Conference on Big Data (BigData 2013)*, 17-22. October 6-9, 2013, Santa Clara, CA, USA.
- Nemschoff, M. (2013). Big data: 5 major advantages of Hadoop. ITProPortal. Retrieved from <http://www.itproportal.com/2013/12/20/big-data-5-major-advantages-of-hadoop/#ixzz3tEAqQMHL>.
- Neave, H.R. (1990). *The Deming dimension*. Knoxville, TN: SPC Press.
- Nunnally, J.C., & Bernstein, I.H. (1994). *Psychometric Theory (3rd Ed.)*. New York: McGraw Hill.
- Olson, M. (2010). HADOOP: Scalable, flexible data storage and analysis. *IQT Quarterly*, 1(3), 14-18.
- Paetow K., Schmitt M., & Malsch T. (2005). Scalability, scaling processes, and the management of complexity. A system theoretical approach. In: Fischer K., Florian M., Malsch T. (eds) *Socionics. Lecture Notes in Computer Science*, 3413. Springer, Berlin, Heidelberg.
- Park, S.H., Lee, L., & Yi, M.Y. (2011). Group-level effects of facilitating conditions on individual acceptance of information systems. *Information Technology and Management*, 12(4), 315-334.
- Pavlov, P. A., & Chai, L. (2002). What drives electronic commerce across cultures? A cross-cultural empirical investigation of the theory of planned behavior. *Journal of Electronic Commerce Research*, 3(4), 240–253.

- Petter, S., DeLone, W., & McLean, E. (2013). Information systems success: The quest for the independent variables. *Journal of Management Information Systems*, 29(4), 7-61.
- Petter, S., Straub, D., & Rai, A. (2007). Specifying formative constructs in information systems research. *MIS Quarterly*, 31(4), 623-656.
- Phan, K., & Daim, T. (2011). Exploring technology acceptance for mobile services. *Journal of Industrial Engineering and Management*, 4, 339-360, 2011.
- Pinsonneault, A., & Kraemer, K.L. (1993). Survey research methodology in management information systems: an assessment. *Journal of management information systems*, 10(2), 75-105, 1993.
- Podsakoff, P.M., MacKenzie, S.B., Lee, J.-Y. Y., & Podsakoff, N.P. (2003). Common method biases in behavioral research: A critical review of the literature and recommended remedies. *Journal of Applied Psychology*, 88(5), 879-903.
- Premkumar, G., & Potter, M. (1995). Adoption of computer aided software engineering (CASE) technology: An innovation adoption perspective. *The DATABASE for Advances in Information Systems*, 26(2-3), 105-124.
- Pui-Wa, L., & Dunbar, S.B. (2004). Effects of score discreteness and estimating alternative model parameters on power estimation methods in structural equation modeling. *Structural Equation Modeling*, 11, 20-44.
- Qin, L., Kim, Y., Hsu, J., & Tan, X. (2011). The effects of social influence on user acceptance of online social networks. *International Journal of Human-Computer Interaction*, 27(9), 885-889.
- Raguseo, E. (2018). Big data technologies: An empirical investigation on their adoption, benefits and risks for companies. *International Journal of Information Management*, 38, 187-195.
- Rahman, N. (2018a). A taxonomy of data mining problems. *International Journal of Business Analytics*, 5(2), 73-86.
- Rahman, N. (2018b). Data warehousing and business intelligence with big data. *Proceedings of the 2018 International Annual Conference of the American Society for Engineering Management (ASEM '18)*. The Coeur d' Alene Resort, Idaho, USA. October 17 - 20, 2018.
- Rahman, N. (2017). An empirical study of data warehouse implementation effectiveness. *International Journal of Management Science and Engineering Management*, 12(1), 55-63.
- Rahman, N. (2016). SQL scorecard for improved stability and performance of data warehouses. *International Journal of Software Innovation*, 4(3), 22-37.

- Rahman, N. (2013). SQL optimization in a parallel processing database system. *Proceedings of the IEEE 26th Canadian Conference of Electrical and Computer Engineering (CCECE 2013)*, Regina, Saskatchewan, Canada, May 5 - 8, 2013.
- Rahman, N., & Aldhaban, F. (2015). Assessing the effectiveness of big data initiatives. *Proceedings of the IEEE Portland International Center for Management of Engineering and Technology (PICMET 2015) Conference*, Portland, Oregon, USA, August 2 - 6, 2015.
- Rahman, N., & Iverson, S. (2015). Big data business intelligence in bank risk analysis. *International Journal of Business Intelligence Research*, 6(2), 55-77.
- Rahman, N., & Rutz, D. (2015). Building data warehouses using automation. *International Journal of Intelligent Information Technologies*, 11(2), 1-22.
- Rahman, N., Rutz, D., Akhter, S., & Aldhaban, F. (2014). Emerging technologies in business intelligence and advanced analytics. *ULAB Journal of Science and Engineering*, 5(1), 7-17.
- Rahman, N., & Sutton, L. (2016). Optimizing SQL performance in a parallel processing DBMS architecture. *ULAB Journal of Science and Engineering*, 7(1), 33-44.
- Rajan, C.A., & Baral, R. (2015). Adoption of ERP system: An empirical study of factors influencing the usage of ERP and its impact on end user. *IIMB Management Review*, 27, 105-117.
- Rajpurohit, A. (2013). Big data for business managers - bridging the gap between potential and value. *Proceedings of the 2013 IEEE International Conference on Big Data (BigData 2013)*, 12-14. October 6-9, 2013, Santa Clara, CA, USA.
- Ramamurthy, K., Sen, A., & Sinha, A.P. (2008). An empirical investigation of the key determinants of data warehouse adoption. *Decision Support System*, 44, 817-841.
- Read, D. (2004). Utility theory from Jeremy Bentham to Daniel Kahneman. *Operational Research working papers, LSEOR 04.64*. Department of Operational Research, London School of Economics and Political Science, London, UK. ISBN 0753016899.
- Research Rundowns (2018). Instrument, validity, reliability. *Research Rundowns*. Retrieved from <https://researchrundowns.com/quantitative-methods/instrument-validity-reliability/>
- Rhodes, R. E., & Courneya, K. S. (2003). Investigating multiple components of attitude, subjective norm, and perceived control: An examination of the theory of planned behaviour in the exercise domain. *The British Journal of Social Psychology*, 42, 129-146.
- Richards, N.M., & King, J.H. (2014). Big data ethics. *Wake Forest Law Review*, 49, 393-432.

- Roca, J.C., Chiu, C.-M., & Martinez, F.J. (2006). Understanding e-learning continuance intention: An extension of the technology acceptance model. *International Journal of Human-Computer Studies*, 64, 683–696.
- Rogers, E.M. (1983). *Diffusion of innovations* (3rd Ed.). New York, NY: Free Press.
- Rogers, E.M. (2003). *Diffusion of innovations* (5th Ed.). New York, NY: Free Press.
- Russom, P. (2013). Managing big data. *TDWI Best Practices Report, TDWI Research*, 1-40.
- Sabherwal, R., & Jeyaraj, A. (2015). Information technology impacts on firm performance: An extension of Kohli and Devaraj (2003). *MIS Quarterly*, 39(4), 809-836.
- Saheb, T., & Saheb, T. (2020). Understanding the development trends of big data technologies: an analysis of patents and the cited scholarly works. *Journal of Big Data*, 7(12), 1-26.
- Saleh, M.A. (2006). Antecedents of commitment to an importer supplier. *Doctoral Dissertation. Queensland University of Technology, Brisbane*, 2006.
- Saunders, A. (2016). Valuing information technology related intangible assets. *MIS Quarterly*, 40(1), 83-110.
- Sauro, J. (2011). *Measuring usefulness*. Measuring U. Retrieved from: <https://measuringu.com/usefulness/>
- Scherrer-Rathje, M., & Boyle, T.A. (2012). An end-user taxonomy of enterprise systems flexibility: Evidence from a leading European apparel manufacturer. *Information Systems Management*, 29(2), 86-99.
- Schlesinger, P.A., & Rahman, N. (2015). Self-service business intelligence resulting in disruptive technology. *Journal of Computer Information Systems*, 56(1), 11-21.
- Segars, A.H., & Grover, V. (1993). Re-examining perceived ease of use and usefulness: A confirmatory factor analysis. *MIS Quarterly*, 17(4), 517-525.
- Sekaran, U., & Bougie, R. (2016). *Research methods for business: A skill building approach* (7th Ed.). New York, NY: John Wiley & Sons.
- Sen, A., & Jacob, V.S. (1998). Industrial-strength data warehousing. *Communications of the ACM*, 41(9), 28-31.
- Sen, A., & Sinha, A.P. (2005). A comparison of data warehousing methodologies. *Communications of the ACM*, 48(3), 79-84.
- Seneler, C.O., Basoglu, N., & Daim, T.U. (2008). A taxonomy for technology adoption: A human computer interaction perspective. *IEEE PICMET 2008 Proceedings*, 27-31 July, Cape Town, South Africa.

- Sharma, R., Yetton, P., & Crawford, J. (2009). Estimating the effect of common method variance: The method – Method for pair technique with an illustration from TAM research. *MIS Quarterly*, 33(3), 473-490.
- Sheppard, B.H., Hartwick, J., & Warshaw, P.R. (1988). The theory of reasoned action: A meta-analysis of past research with recommendations for modifications and future research. *Journal of Consumer Research*, 15(3), 325–343.
- Shvachko, K.V. (2011). Apache Hadoop: The scalability update. *login: The Usenix Magazine*, 36(3), 7-13.
- Shvachko, K., Kuang, H., Radia, S., & Chansler, R. (2010). The Hadoop distributed file system. *Proceedings of the 2010 IEEE 26th Symposium on Mass Storage Systems and Technologies (MSST)*, 1-10, IEEE Computer Society, Washington, DC, USA.
- Silva, L. (1997). Power and politics in the adoption of information systems by organisations: the case of a research centre in Latin America. Doctoral Dissertation, Department of Information Systems, *London School of Economics and Political Science*, University of London, UK.
- Singh, D., & Reddy, C.K. (2014). A survey on platforms for big data analytics. *Journal of Big Data*, 1(8).
- Sivo, S.A., Saunders, C., Chang, Q., & Jiang, J.J. (2006). How low should you go? Low response rates and the validity of inferences in IS questionnaire research. *Journal of Management Information Systems*, 7(6), 351-414.
- Spieß, J., T'Joens, Y., Dragnea, R., Spencer, P., & Philippart, L. (2014). Using big data to improve customer experience and business performance. *Bell Labs Technical Journal*, 18(4), 3-17.
- Srinivasan, U., & Arunasalam, B. (2013). Leveraging big data analytics to reduce healthcare costs. *IT Pro*, 21-29.
- Stigler, G.J. (1950). The development of utility theory. *Journal of Political Economy*, 58(4), 307-327.
- Straub, D.W. (1989). Validating instruments in MIS research. *MIS Quarterly*, 13(2), 147-169.
- Straub, D., Boudreau, M.C., & Gefen, D. (2004). Validation guidelines for IS positivist research. *Communications of the Association for Information Systems*, 13(2004), 380-427.
- Straub, D., & Burton-Jones, A. (2007). Veni, vidi, vici: Breaking the TAM logjam. *Journal of the Association for Information Systems*, 8(4), 223-229.
- Straub, D., Limayem, M., & Karahanna-Evaristo, K. (1995). Measuring system usage: Implications for IS theory testing. *Management Science*, 41(8), 1328-1342.

- Suhr, D. (2006). *The basics of structural equation modeling*. Retrieved from <http://www.lexjansen.com/wuss/2006/tutorials/TUT-Suhr.pdf>.
- Sun, H., Fang, Y., & Zou, H. (2016). Choosing a fit technology: Understanding mindfulness in technology adoption and continuance. *Journal of the Association for Information Systems*, 17(6), 377-412.
- Sun, S., Cegielski, C.G., Jia, L., & Hall, D.J. (2018): Understanding the factors affecting the organizational adoption of big data. *Journal of Computer Information Systems*, 58(3), 193-203.
- Surbakti, F.P.S., Wang, W., Marta Indulska, M., & Sadiq, S. (2020). Factors influencing effective use of big data: A research framework. *Information & Management*, 57(1), 103146.
- Swanson, E.B. (2019). Technology as Routine Capability. *MIS Quarterly*, 43(3), 1007-1024.
- Szajna, B. (1994). Software evaluation and choice: predictive evaluation of the technology acceptance instrument, *MIS Quarterly*, 18(3), 319-324.
- Szajna, B. (1996). Empirical evaluation of the revised technology acceptance model. *Management Science*, 42(1), 85-92.
- Tabachnick, B.G., & Fidell, L.S. (2012): *Using multivariate statistics* (6th Ed.), New York, NY: Pearson.
- Tambe, P. (2014). Big data investment, skills, and firm value. *Management Science*, 60(6), 1452-1469.
- Tan, M., & Teo, T.S.H. (2000). Factors influencing the adoption of internet banking. *Journal of the Association for Information Systems*, 1(1), 1-41.
- Tang, M., Alazab, M., & Luo, Y. (2019). Big data for cybersecurity: Vulnerability disclosure trends and dependencies. *IEEE Transactions on Big Data*, 5(3) 317-329.
- Tanur, J.M. (1982). Advances in methods for large-scale surveys and experiments. In R. Mcadams, N.J. Smelser, & D.J. Treiman (eds.), *Behavioral and Social Science Research: A National Resource, part II*. Washington, DC: National Academy Press, 1982.
- Taylor, S., & Todd, P.A. (1995). Understanding information technology usage: A test of competing models. *Information Systems Research*, 6(2), 144-176.
- Technavio (2020). *Big data market 2020-2024: Growing investment in smart city initiatives to boost growth*. Technavio. Link: <https://www.businesswire.com/news/home/20200309005078/en/Big-Data-Market-2020-2024-Growing-Investment-Smart>
- Teece, D.J., Pisano, G., & Shuen, A. (1997). Dynamic capabilities and strategic management. *Strategic Management Journal*, 18(7), 509-533.

- Tene, O., & Polonetsky, J. (2013). Big data for all: Privacy and user control in the age of analytics. *Northwestern Journal of Technology and Intellectual Property* 11(5), 239-273.
- Teo, T. S. H., & Ranganathan, C. (2004). Adopters and non-adopters of business-to-business electronic commerce in Singapore. *Information and Management*, 42, 89–102.
- Tornatzky, L., & Fleischer, M. (1990). *The process of technology innovation*. Lexington, M.A: Lexington Books.
- Tsai, C.-W., Lai, C.-F., Chao, H.-C., & Vasilakos, A.V. (2015): Big data analytics: a survey. *Journal of Big Data*, 2(21), 1-32.
- Tsai, C.-W., Yang, Y.-L., Chiang, M.-C., & Yang, C.-S. (2014). Intelligent big data analysis: a review. *International Journal of Big Data Intelligence*, 1(4).
- Tsumoto, S., & Hirano, S. (2013). Granularity-based temporal data mining in hospital information system. *Proceedings of the 2013 IEEE International Conference on Big Data (BigData 2013)*, 32-40. October 6-9, 2013, Santa Clara, CA, USA.
- Turner, M., Kitchenham, B., Brereton, P., Charters, S, and Budgen, D. (2010). Does the technology acceptance model predict actual use? A systematic literature review. *Information and Software Technology*, 52, 463-479.
- Urbach, N., & Ahlemann, F. (2010). Structural equation modeling in information systems research using partial least squares. *Journal of Information Technology Theory and Application*, 11(2), 5-40.
- Venkatesh, V. (2000). Determinants of perceived ease of use: Integrating control, intrinsic motivation, and emotion into the technology acceptance model. *Information Systems Research*, 11(4), 342–365.
- Venkatesh, V., & Bala, H. (2008). Technology Acceptance Model 3 and a Research Agenda on Interventions. *Decision Sciences*, 39(2), 273-315.
- Venkatesh, V., & Davis, F. D. (2000). A theoretical extension of the technology acceptance model: four longitudinal field studies. *Management Science*, 46(2), 186–204.
- Venkatesh, V., Davis, F., & Morris, M.G. (2007). Dead or alive? The development, trajectory and future of technology adoption research. *Journal of the Association for Information Systems*, 8(4), 267-286.
- Venkatesh, V., Morris, M. G., Davis, G. B., & Davis, F. D. (2003). User acceptance of information technology: Toward a unified view. *MIS Quarterly*, 27(3), 425–478.
- Venkatesh, V., Thong, J.Y.L., & Xu, X. (2012). Consumer acceptance and use of information technology: extending the unified theory of acceptance and use of technology. *MIS Quarterly*, 36(1), 157-178.

- Venkatesh, V., Thong, J.Y.L., & Xu, X. (2016). Unified theory of acceptance and use of technology: A synthesis and the road ahead. *Journal of the Association for Information Systems*, 17(5), 328-376.
- Venkatesh, V., & Zhang, X. (2010). Unified theory of acceptance and use of technology: U.S. vs. China. *Journal of Global Information Technology Management*, 13(1), 5–27.
- Verma, S., Bhattacharyya, S.S., & Kumar, S. (2018). An extension of the technology acceptance model in the big data analytics system implementation environment. *Information Processing and Management*, 54, 791-806.
- Viceconti, M., Hunter, P., & Hose, R. (2015). Big data, big knowledge: Big data for personalized healthcare. *IEEE Journal of Biomedical and Health Informatics*, 19(4), 1209-1215.
- Villars, R.L., Olofson, C.W., & Eastwood, M. (2011). Big data: What it is and why you should care, *IDC White Paper*, Paper ID: 228827, 1-14.
- Wang, J., & Zhang, C. (2018). Software reliability prediction using a deep learning model based on the RNN encoder–decoder. *Reliability Engineering and System Safety*, 170, 73–82.
- Wang, W., Hsieh, J.J.P.-A., & Song, B. (2012). Understanding user satisfaction with instant messaging: An empirical survey study. *International Journal of Human-Computer Interaction*, 28, 153-162.
- Wernerfelt, B. (1984). A resource-based view of the firm. *Strategic Management Journal*, 5, 171-180.
- Wessel, M., & Helmer, N. (2020). A crisis of ethics in technology innovation. MIT Sloan Management Review. Spring 2020.
- Wheelock, M.D. (2013). Factors influencing the adoption of cloud storage by information technology decision makers. *Ph.D. Dissertation, School of Business and Technology, Capella University. ProQuest LLC. ISBN: 978-1-3036-2041-6.*
- Wixom, B.H., & Watson, H.J. (2001). An empirical investigation of the factors affecting data warehousing success. *MIS Quarterly*, 25(1), 17-32.
- Wlodarczky, T.W., & Hacker, T.J. (2014). Current trends in predictive analytics of big data. *International Journal of Big data Intelligence*, 1(3), 172.
- Wu, I.-L., & Chiu, M.-L. (2015). Organizational applications of IT innovation and firm's competitive performance: A resource-based view and the innovation diffusion approach. *Journal of Engineering and Technology Management* 35, 25–44.
- Wu, J.-H., Chen, Y.-C., & Lin, L.-M. (2007). Empirical evaluation of the revised end user computing acceptance model. *Computers in Human Behavior*, 23(2007), 162–174.

- Wu, J., Li, H., Liu, L., & Zheng, H. (2017). Adoption of big data analytics in mobile healthcare market: An economic perspective. *Electronic Commerce Research and Application*, 22, 24-41.
- Wu, J.-H., & Wang, S.-C. (2005). What drives mobile commerce? An empirical evaluation of the revised technology acceptance model. *Information & Management*, 42(2005), 719–729.
- Wu, L., Hitt, L., & Lou, B. (2019). Data analytics, innovation, and firm productivity. *Management Science*, 66(5).
- Wu, X., Zhu, X., Wu, G.-Q., & Ding, W. (2014). Data mining with big data. *IEEE Transactions on Knowledge and Data Engineering*, 26(1), 97-107.
- Xia, F., Wang, W., Bekele, T.M., & Liu, H. (2017). Big scholarly data: A survey. *IEEE Transactions on Big Data*, 3(1) 18-35.
- Zhang, D., Pan, S.L., Yu, J., & Liu, W. (2019). Orchestrating big data analytics capability for sustainability: A study of air pollution management in China. *Information & Management*, 103231. <https://doi.org/10.1016/j.im.2019.103231>
- Zhang, X. (2017). Knowledge management system use and job performance: A multilevel contingency model. *MIS Quarterly*, 41(3), 811-840.
- Zhang, X., & Pham, H. (2000). An analysis of factors affecting software reliability. *The Journal of Systems and Software*, 50, 43-56.
- Zhu, K., & Kraemer, K. L. (2005). Post-adoption variations in usage and value of e-business by organizations: cross-country evidence from the retail industry. *Information Systems Research*, 16(1), 61–84.

Appendices

Appendix A: Cover Letter and Survey Questionnaire

Dear Participant,

This survey is part of an academic research project undertaken by Nayem Rahman (Ph.D. candidate) and Dr. Tugrul U. Daim (Ph.D. advisor) from the Department of Engineering and Technology Management, Portland State University, Oregon, USA.

You are being invited to participate in this survey because of your expertise and experience in the field. Your name will not be used in any published reports about this study.

Your participation in this survey is completely voluntary. You have the right to choose not to participate or to withdraw your participation at any point in this study.

The survey is being undertaken to explore the factors influencing big data technology (Hadoop) adoption.

The survey is expected to provide an outcome from industry/organization-level users' acceptance context.

The final results of the survey will provide the basis for a dissertation towards my Ph.D. degree at the Maseeh College of Engineering and Computer Science, Portland State University, Oregon, USA.

If you consent to participate in this survey, please click on the RIGHT-ARROW below to continue in this Survey.

This survey uses 5-point Likert-scale with the scale being Strongly Disagree (1) to Extremely Agree (5).

Thank you very much for volunteering and taking the time to help me by responding to this survey.

Thanks & Regards,
Nayem Rahman
Ph.D. Candidate,
Department of Engineering & Technology Management
Portland State University
Portland, OR 97201, USA
E-mail: nayem.rahman@yahoo.com (primary); rahmanm@pdx.edu (alternative).

SURVEY QUESTIONNAIRE

5-point Likert scales (Strongly Disagree, Disagree, Neutral, Agree, Strongly Agree) were used for all constructs except 2 demographic questions at the end.

1. Scalability (SC)

New items

SC1 - Hadoop is scalable to handle hundreds of terabytes to petabytes of data compared to relational databases.

SC2 - With the increase of applications, users, and data volume, Hadoop is able to meet extra load by expanding the number of nodes.

SC3 - Hadoop has built-in capability to scale-out storage compared to our organization's traditional data storage systems.

SC4 - Hadoop's scale-out storage system can store and distribute very large data sets across hundreds of inexpensive servers that operate in parallel.

2. Data Storage and Processing (DS)

New items

DS1 - Hadoop is capable to run analytics on hundreds of terabytes to petabytes of data set.

DS2 - Hadoop's processing engine is capable to process both structured and unstructured data.

DS3 - Hadoop's storage and processing engine can serve many application needs - analytics, processing, machine learning.

DS4 - Hadoop is capable to receive and process streaming data real-time.

3. Cost-Effectiveness

New items

Cost1 - Hadoop is able to hold hundreds of terabytes to petabytes of data with minimal cost.

Cost2 - Hadoop offers a cost-effective storage solution for my organization's exploding data sets.

Cost3 - Hadoop is able to improve the efficiency of business applications and thereby reduce costs.

Cost4 - Using Hadoop is cost-effective.

4. Performance Expectancy

Partially Adapted from Venkatesh, Morris, Davis (2003)

PE1 - The team members of my organization find the Hadoop Platform useful in performing jobs.

PE2 - By using the Hadoop Platform members of my organization are able to accomplish tasks more quickly.

PE3 - The use of the Hadoop Platform increases my organization's productivity.

PE4 - Hadoop is able to provide a good user experience.

5. Security and Privacy Considerations

New items

SP1 - Hadoop has data protection capability such as encryption and data masking to prevent sensitive data from being accessed by unauthorized users and applications.

SP2 - Hadoop has authentication capability such as Kerberos to authenticate Hadoop users.

SP3 – Hadoop provides a capability for providing role-based authorization to both data and metadata stored in HDFS in a Hadoop cluster.

SP4 - Hadoop (HDFS) is able to ensure the confidentiality of stored data in both physical and cyber ways.

6. Reliability

New items

RL1 - Hadoop keeps multiple copies of the same data in different nodes which makes my organization feel comfortable about not losing any critical data.

RL2 - Hadoop is capable to automatically identify data node failing and possible remedy.

RL3 - Hadoop maintains data in raw format which allows data to remain the way it comes from the source, that is, in its original format.

RL4 - Hadoop Platform is able to operate under given conditions, without collapsing.

7. Data Analytics Capability

New items

DA1 - Hadoop allows to perform different types of analytics (including Customer, Compliance, Fraud, Operational) to enable making business decisions.

DA2 - Hadoop's capability to store both historical and current data allows for the discovery of knowledge from massive datasets.

DA3 - Hadoop's capability to combine data from many sources (external and internal) allows my organization to get 360-degree views of customers and other business entities.

DA4 - Hadoop provides my organization capability to develop and run machine learning model on a complete set of data (stored in HDFS).

8. Training and Required Skills

Partially adapted from Amoaky-Gyampah & Salam (2004); Rajan & Baral (2015)

TR1 - Having user-support for the Hadoop platform will help users of my organization gain knowledge.

TR2 - Specialized training will save my organization's users' time on learning how to use the Hadoop platform.

TR3 - Documentation should be provided for the Hadoop platform for users wanting to learn on their own.

TR4 - The training gave users of my organization confidence in the Hadoop Platform.

9. Flexibility

New items

FL1 - Hadoop provides greater flexibility to consolidate data from various sources into one single place (i.e., Hadoop HDFS).

FL2 - Hadoop provides high throughput as well as fault tolerance as data is also replicated to other nodes in the cluster.

FL3 - Hadoop allows to build programs at a small scale and expand the system as needed.

FL4 - Hadoop enables businesses to easily access new data sources and tap into different types of data to generate value.

10. Output Quality

Partially adapted from Medina-Quintero & Chaparro-Peláez (2007); Venkatesh & Davis, 2000

OQ1 - Hadoop Platform's Quality is associated with the satisfaction of my organization's users' work.

OQ2 - My organization is satisfied with the data consistency in Hadoop Platform.

OQ3 - My organization is satisfied with the data completeness (no data gaps, missing data) in Hadoop Platform.

OQ4 - By using the Hadoop the users of my organization get high quality output.

11. Functionality

New items

FN1 - Hadoop architecture can access and process the data that comes from many sources, tools, and devices.

FN2 - Hadoop framework provides a distributed file system for big data sets.

FN3 - The HDFS replicates the data sets on the commodity servers making the process run in parallel.

FN4 - Hadoop provides rich and robust machine learning libraries (e.g., Mahout).

12. Facilitating Conditions

Adapted from Kwon et al. (2014); Venkatesh (2000)

FC1 - My organization takes advantage of new information technologies.

FC2 - My organization has resources necessary to use the Hadoop Platform.

FC3 - Given the resources, opportunities, and knowledge it takes to use the Platform, it would be easy for my organization to use the Hadoop Platform.

FC4 – My organization has internal Hadoop Infrastructure team to support Hadoop Platform users.

13. Perceive Usefulness (PU)

Adapted from Davis (1993)

PU1 - Using Hadoop Platform enables my organization to accomplish its tasks more quickly.

PU2 - Using Hadoop Platform makes it easier for my organization to carry out its tasks.

PU3 - Hadoop Platform is flexible from varieties of data storage and processing perspectives.

PU4 - Overall, using Hadoop Platform is advantageous compared to the conventional data management system of my organization.

14. Perceived Ease of Use (PEOU)

Adapted from Davis (1993); Venkatesh & Davis (2000)

PEOU1 - Interacting with Hadoop platform does not require a lot of mental effort.

PEOU2 - My organization finds Hadoop Platform easy to use when performing its job functions.

PEOU3 - It is easy for my organization's users to become more skillful and experienced with Hadoop Platform.

PEOU4 - My organization's interaction with Hadoop Platform is clear and understandable.

15. Behavioral Intention (BI) to Use the System

Adapted from Venkatesh et al. (2003)

BI1 - My organization intends to use Hadoop for its data storage, management, processing, and analytical needs.

BI2 - I predict my organization would use Hadoop within the next six months.

BI3 - My organization will continue to use Hadoop in the future.

16. Actual Use (AU)

Adapted from Davis (1993); Davis & Venkatesh (1996); Rajan & Baral (2015)

AU1: My organization uses Hadoop occasionally.

AU2: My organization uses Hadoop regularly (daily, weekly, etc.).

AU3: My organization is satisfied with using the Hadoop Platform.

Note:

5-point Likert scales (Strongly Disagree, Disagree, Neutral, Agree, Strongly Agree) were used for all constructs except below 2 demographic questions

Demographic Questions

Your core business falls under which of the following organizations? Choose One:

Adapted from Russom (2013)

Manufacturing

Financial Services

Consulting/Professional Services

Software/Internet Services

Healthcare

Insurance

Retail

Telecommunications

Government

Transportation/Logistics

Advertising/Marketing

Other

What is your job role in your organization? Choose One:

Adapted from Russom (2013)

Hadoop Engineer/Application Developer

Big Data Architect/Enterprise Architect

Hadoop Administrator

Data Scientist

Data Analyst

Big Data/Information Technology (IT) Manager

Big Data Program Manager

Chief Information Officer (CIO) or similar executive

Other.


Appendix B: Pilot Test Survey Questionnaire

Survey Instrument created as part of Pilot Test (partial picture shown here) Big Data (Hadoop) Tech Adoption - Survey Instrument Validation

▼ INTRODUCTION

*****Big Data (Hadoop) Tech Adoption - Survey Instrument Validation*****



INTRO Dear Participant-

 You are being asked to participate in this Pilot Test of Survey based on the input of which a final Survey Instrument will be developed for the actual survey of a large population. This Pilot Test is being conducted by Nayem Rahman (Ph.D. candidate) and Dr. Tugrul U. Daim (Ph.D. advisor), from the Department of Engineering and Technology Management, at Portland State University in Portland, Oregon.

This research will develop a technology acceptance model for big data technologies.
Your name will not be used in any published reports about this study.
If you consent to participate in this survey, please click on the right arrow below to continue in this Pilot Test Survey.
Thanks & Regards
Nayem Rahman

▼ Survey Questionnaire Block Options ▼

Q1 1. Scalability

	Strongly Agree	Agree	Neutral	Disagree	Strongly Disagree
SC1 - Hadoop is scalable to handle hundreds of terabytes to petabytes of data compared to relational databases. (Processing, storing, etc)	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
SC2 - With the increase of applications, users, and data volume, Hadoop is able to meet extra load by expanding the number of nodes.	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
SC3 - Hadoop has built-in capability to scale-out storage compared to our organization's traditional data storage systems.	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
SC4 - Hadoop's scale-out storage system can store and distribute very large data sets across hundreds of inexpensive servers that operate in parallel.	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
SC5 - Hadoop can expand incrementally without having to change the applications and without the users noticing any degradation in performance.	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>

Q2 2. Data Storage and Processing

Appendix C: Initial Survey Questionnaire Validation

This was conducted before Pilot Test and Final Survey Data Collection (Partial List)

A	B	C	D	E
		Item relevance to the Construct: Put in scale 1 to 5	Ease of Answering the question: Put in scale 1 to 5	ANY COMMENTS?
1	Construct Items			
2	1. Construct: Scalability			
3	SC1 - Hadoop is scalable to handle hundreds of terabytes of data compared to relational databases.	5	5	Scale could be in PB scale
4	SC2 - Hadoop is scalable in terms of storage and processing.	5	5	
5	SC3 - Hadoop scalability in terms of higher performance (lower latency) can improve the bottom-line of my company.	3	5	Hadoop latency is generally high. Good for batch and not for real-time.
6	SC4 - With the increase of applications, users, and data volume, Hadoop is able to meet extra load by expanding number of nodes.	5	5	
7	SC5 - Hadoop has built-in capability to scale-out storage compared to our company's traditional data storage systems.	5	5	
8	SC6 - Hadoop's scale-out storage system can store and distribute very large data sets across hundreds of inexpensive servers that operate in parallel.	5	5	
9	SC7 - Hadoop enables my company's businesses to run applications on thousands of nodes involving thousands of terabytes of data.	5	5	
10	SC8 - Hadoop can expand incrementally without having to change the applications and without the users noticing any degradation in performance.	5	5	
11	SC9: Propose any important item that is missing:			
12	2. Construct: Data Storage and Processing			
13	DS1 - Hadoop system is capable to store and process huge volume of data.	5	5	
14	DS2 - Hadoop is capable to receive and process streaming data real-time.	4	5	
15	DS3 - Hadoop is capable to run analytics on a very large data set.	5	5	
16	DS4 - Hadoop's processing Engine is capable to process both structured and unstructured data.	5	5	
17	DS5 - Hadoop allows for write-once and read many times with its Processing Engine.	5	5	
18	DS6 - Hadoop's processing engine removes the requirement of data summarization which was needed in traditional database systems.	2	2	
19	DS7 - Hadoop's stored data can easily be accessed, used and processed by applications and services.	4	5	In some insatnces, it may not be easy to integrate with some services.
20	DS8 - Hadoop's storage and processing engine can serve all application needs - analytics, processing, machine learning.	5	5	

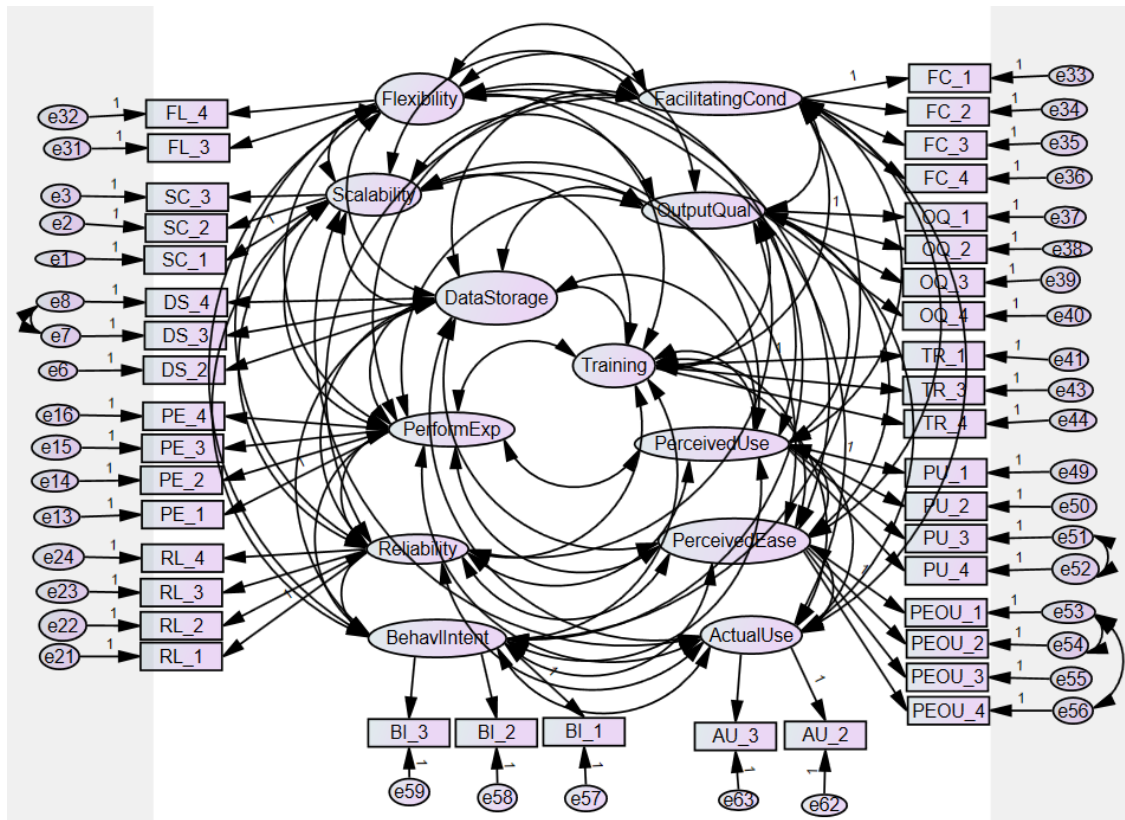
Appendix D: Hadoop User Groups in the U.S.

Hadoop User Groups – U.S. [Search on Google: As of July 25, 2019] pp pp pp pp pp pp pp

Hadoop User Group Name	Location	Link
1. Atlanta Hadoop Users Group (HUG) [Members: 2,737]	Atlanta, GA	https://www.meetup.com/Atlanta-Hadoop-Users-Group/?_cookie-check=M9Oyj8wv5UK4CIYj
2. Bay Area Hadoop User Group [Members: 6,440]	San Francisco, CA	https://www.meetup.com/hadoop/
3. Phoenix Hadoop User Group [Members: 1,568]	Boston, MA	https://www.meetup.com/Phoenix-Hadoop-User-Group/
4. Chicago area Hadoop User Group [Members: 2,951]	Chicago, IL	https://www.meetup.com/Chicago-area-Hadoop-User-Group-CHUG/
5. Cleveland Hadoop User Group [Members: 3,337]	Cleveland, OH	https://www.meetup.com/Cleveland-Hadoop/
6. DFW Bigdata Meetup Group [Members: 3,220]	Dallas, TX	https://www.meetup.com/DFW-BigData/
7. Florida HUG [Members: 163]	Saint Augustine, FL	https://www.meetup.com/HUGNOFA/
8. New Jersey HUG [Members: 1,368]	Flemington, NJ	https://www.meetup.com/nj-dapp/
9. Hadoop-NYC [Members: 4,060]	New York, NY	https://www.meetup.com/Hadoop-NYC/
10. Pittsburgh HUG [Members: 730]	Pittsburgh, PA	https://www.meetup.com/HUG-Pittsburgh/
11. Los Angeles HUG [Members: 2,049]	Los Angeles, CA	https://www.meetup.com/LA-HUG/
12. St. Louis HUG [Members: 1,395]	Saint Louis, MO	https://www.meetup.com/St-Louis-Hadoop-Users-Group/
13. Big Data (native Hadoop) Ingest & Transform, Washington DC [Members: 1,084]	Washington, DC	https://www.meetup.com/Big-Data-Ingest-Washington-DC/members/
14. Charlotte HUG [Members: 891]	Charlotte, NC	https://www.meetup.com/CharlotteHUG/

Appendix E: Final CFA

Total 12 Constructs along with 40 Items



Appendix F: Cronbach's Alpha

Construct Name	Number of Items	Cronbach' Alpha	Reliability
Scalability (SC)	4	.901	Reflective
Data Storage & Processing (DS)	4	.776	Reflective
Cost-Effectiveness (COST)	4	.920	Reflective
Performance Expectancy (PE)	4	.869	Reflective
Security & Privacy (SP)	4	.901	Reflective
Reliability (RL)	4	.901	Reflective
Data Analytics Capability (DA)	4	.847	Reflective
Training & Skills (TR)	4	.901	Reflective
Flexibility (FL)	4	.869	Reflective
Output Quality (OQ)	4	.887	Reflective
Functionality (FN)	4	.728	Reflective
Facilitating Conditions (FC)	4	.848	Reflective
Perceived Usefulness (PU)	4	.901	Reflective
Perceived Ease of Use (PEOU)	4	.887	Reflective
Behavioral Intention (BI)	3	.808	Reflective
Actual Use (AU)	3	.787	Reflective

Appendix G: EFA – Pattern Matrix

Pattern Matrix^a

	Factor							
	1	2	3	4	5	6	7	8
SC_1					.719			
SC_2					.797			
SC_3					.727			
DS_2							-.360	
DS_3							-.991	
DS_4							-.552	
RL_1				-.436				
RL_2				-.723				
RL_3				-.682				
RL_4				-.555				
FL_3								.691
FL_4								.747
TR_1						.749		
TR_3						.866		
TR_4						.676		
PE_1		.634						
PE_2		.946						
PE_3		.833						
PE_4		.565						
OQ_1			.407					
OQ_2			.636					
OQ_3			1.066					
OQ_4			.691					
FC_1	.588							
FC_2	.867							
FC_3	.928							
FC_4	.676							

Extraction Method: Maximum Likelihood.

Rotation Method: Promax with Kaiser Normalization.

a. Rotation converged in 8 iterations.

Appendix H: Technology Acceptance Factors

Technology Acceptance factors identified based on literature review

Sl.	External Variables	Variable Description	Authors	Theory/ Model
1	Performance Expectancy	User experience focused (Object Usability)	Venkatesh, 2000.	UTAUT
2	Relative advantage	The degree to which an innovation is perceived as being better than its precursor (Lee et al., 2003).	Arts et al., 2011; Chin & Gopal, 1995; Fichman & Kemerer, 1993; Moore & Benbasat, 1991; Premkumar & Potter, 1995; Ramamurthy et al., 2008; Wu & Chiu, 2015; Moore & Benbasat, 1996; Tan & Teo, 2000; Taylor & Todd, 1995.	DOI
3	Scalability	Capability of software and hardware to handle increase of workload in terms of bandwidth and data volume.	Aye & Thein, 2015; Borthakur et al., 2011; Lourenco et al., 2015; Malaka & Brown, 2015; Rahman & Rutz, 2015; Sen & Jacob, 1998; Sen & Sinha, 2005;	TOE
4	Compatibility	The degree to which an innovation is perceived as being consistent with the existing values, needs, and past experiences of potential adopters (Lee et al., 2003).	Arts et al., 2011; Chin & Gopal, 1995; Fichman & Kemerer, 1993; Luo et al., 2010; Moore & Benbasat, 1991; Premkumar & Potter, 1995; Wu & Chiu, 2015; Rajan & Baral, 2015; Moore & Benbasat, 1996; Taylor & Todd, 1995; Wu & Wang, 2005.	DOI, TAM
5	Complexity	The degree to which an innovation is perceived as being difficult to use (Lee et al., 2003).	Arts et al., 2011; Chau & Tam, 1997; Fichman & Kemerer, 1993; Premkumar & Potter, 1995; Ramamurthy et al., 2008; Wu & Chiu, 2015; Rajan & Baral,	DOI, TAM

			2015; Tan & Teo, 2000; Taylor & Todd, 1995.	
6	Cost effectiveness	Capability of a technology that is effective and productive enough in relation to its costs.	Balac et al., 2013; Bologa et al., 2010; Cao et al., 2015; Hartmann et al., 2014; Russom, 2013; Villars et al., 2011; Phan & Daim, 2011; Premkumar & Potter, 1995; Wu & Wang, 2005.	None
7	Total Cost of Ownership	Capability of a technology that is cost effective, does not incur significant hidden cost during the lifecycle, and easy to dispose of at the end of life.	Malaka & Brown, 2015; Kohli et al., 2012.	None
8	Trialability	The degree to which an innovation may be experimented with before adoption (Lee et al., 2003).	Fichman & Kemerer, 1993; Moore & Benbasat, 1991; Tan & Teo, 2000; Karahanna et al., 1999; Lee et al., 2003	DOI, TAM
9	Security and Privacy Considerations	Security and privacy against intangible harm that something can cause.	Gray, 2014; McNeely & Hahm, 2014; Martin, 2015; Richards & King, 2014; Tene & Polonetsky, 2013; Viceconti et al., 2015; Wu et al., 2017.	TOE
10	Observability	The degree to which the results of an innovation are observable to others (Lee et al., 2003).	Arts et al., 2011; Fichman & Kemerer, 1993; Moore & Benbasat, 1991; Moore & Benbasat, 1996; Karahanna et al., 1999; Lee et al., 2003	DOI
11	Flexibility	"Technology characteristic that allows or enables adjustments and other changes to the business process" (Nelson & Nelson, 1997).	Basoglu et al. 2007; Nelson & Nelson, 1997; Nemschoff, 2013; Abouzeid et al. 2009.	None

12	Fault tolerance capability	"Software fault tolerance is a set of software facilities to detect and recover from faults that cause an application process to crash or hang and that are not handled by the underlying hardware or operating system" (Huang & Kintala, 1993).	Abouzeid et al., 2009; Nemschoff, 2013; Huang & Kintala, 1993	None
13	Reliability	Capability of software and hardware to work smoothly according to specifications.	Barlow, 1984; Shvachko et al., 2010; Zhang and Pham, 2000.	None
14	Data storage and processing capability	Capability of technology to store very large volume of data and process them to derive meaningful information.	Aye & Thein, 2015; Shvachko et al., 2010; Li et al., 2020.	None
15	Output Quality	Validity of data/ system to use for business purposes.	Kwon et al., 2014; Venkatesh & Davis, 2000	TAM2
16	Organizational commitment	"Organizational commitment is the individual's psychological attachment to an organization" (The Oxford Review).	Rajpurohit, 2013; Ramamurthy et al., 2008	
17	Top Management Support	Refers to executives of an organization who support is needed to implement a project, tool or technology.	Hwang et al., 2004; Karahanna et al., 1999; Premkumar & Potter, 1995.	TRA, TAM, TOE
18	Facilitating conditions	The control beliefs relating to resource factors such as time and money and IT compatibility issues that may constrain usage (Lee et al., 2003).	Ariyachandra & Watson, 2010; Im et al., 2011; Kwon et al., 2014; Tan & Teo, 2000; Taylor & Todd, 1995; Venkatesh et al., 2003.	TPB, TAM2, UTAUT, TOE

19	Image	The degree to which use of an innovation is perceived to enhance one's image or status in one's social system.	Lee et al., 2003; Moore & Benbasat, 1991; Venkatesh & Davis, 2000	TRA, TAM
20	Self-Efficacy	The belief that one has the capability to perform a particular behavior.	Lee et al., 2003; Igbaria et al., 1995; Rajan & Baral, 2015; Venkatesh, 2000; Tan & Teo, 2000; Taylor & Todd, 1995.	TPB, TAM
21	Subjective Norm/Social Influence	Person's perception that most people who are important to him/her think he/she should or should not perform the behavior in question.	Lee et al., 2003; Choi & Chung, 2013; Venkatesh & Davis, 2000; Im et al., 2011; Tan & Teo, 2000; Liker & Sindi, 1997.	TPB, UTAUT
22	Job Relevance	The capabilities of a system to enhance and individual's job performance.	Lee et al., 2003; Venkatesh & Davis, 2000	TAM
23	Results Demonstrability	The degree to which the results of adopting/using the IS innovation are observable and communicable to others.	Lee et al., 2003; Moore & Benbasat, 1991; Venkatesh & Davis, 2000; Karahanna et al., 1999.	TRA, TAM
24	Functionality	Meets or exceeds functionality		None
25	Effort Expectancy	Effort expectancy is related to the degree of ease associated with the use of a technology.	Im et al., 2011; Venkatesh et al., 2003	UTAUT
26	Voluntariness	The degree to which use of the innovation is perceived as being voluntary, or free will.	Moore & Benbasat, 1991; Venkatesh & Davis, 2000; Lee et al., 2003	TAM
27	Data Analytics Capability	Ability to discover patterns from a large data set or from incoming streaming data.	Zhang et al., 2019.	None
28	Perceived Enjoyment	The extent to which the activity of using a specific	Davis et al., 1992; Chin & Gopal, 1995; Teo et al.	TAM

		system is perceived to be enjoyable in its own right, aside from any performance consequences resulting from system usage.	1999; Lee et al., 2003; Venkatesh et al., 2000	
29	Absorptive capacity	Capability of a firm to assimilate new knowledge about something (e.g., tools or technologies) by an organization.	Bradford & Saad, 2014; Ramamurthy et al., 2008.	None
30	Organizational size	Capability of an organization for executive succession.	Aboelmaged, 2014; Hwang et al., 2004; Ramamurthy et al. 2008.	TOE
31	Competitive/Industry Pressure	Competitive pressure from Industry. The state of business organization that can develop a competitive strategy.	Aboelmaged, 2014; Kuan & Chau, 2001; Malaka & Brown, 2015; Hagi & Wright, 2020	TOE
32	Training and required skills	Training and skills needed to develop a capability or use a technology	Brown-Liburd et al., 2015; Malaka & Brown, 2015; Rajan & Baral, 2015; Russom, 2013; Wixom, & Watson, 2001	RBV