International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395 -0056
Volume: 03 Issue: 06 | June-2016 www.irjet.net p-ISSN: 2395-0072

Traditional Data Storage Methods and the Big Data Concepts

Can Razbonyalı[1], Erdal GÜVENOĞLU[2]

[1]Ph.D. Student, Okan University, Istanbul, Turkey
[2]Assistant Professor, Computer Engineering Department, Maltepe University, Istanbul, Turkey

---

Abstract - *The current century we live in is defined as the information age. Because of this, information/data has become a valuable resource for all stakeholders. All organizations are now intensely using information/data in their decision making processes. Growing business potentials and evolving relationships caused information/data to increase very rapidly. Thus, collecting, storing and accessing information/data has achieved a major importance in organizations' development processes. The innovations in computer technology and the demand for information/data have shown a continuous improvement from the aspect of storing and processing information/data. This study aims to investigate the continuously improving information/data storage and access techniques and the effects of new approaches on these techniques. In the study, firstly, traditional data structure techniques are mentioned. Then the study goes on to explain the concepts of traditional database and data mining. In the final section, Big Data and its effect on traditional methods have been explained including the application of a typical example.*

Key Words: Data, information, memory, storage, access, database, data warehouse, data mining, big data

1. INTRODUCTION

A computer system has four components in general (Fig-1). These are; hardware (CPU, memory, G/Ç units), processing systems, application programs (compilers, assemblers, loaders, database systems, etc.) and users (people, other computers).

Basic computer sources are provided by the hardware. The use of these sources is realized by application programs that are prepared for the solution of user problems. The processing unit, on the other hand, provides the communication between the application programs and the hardware [1, 2].

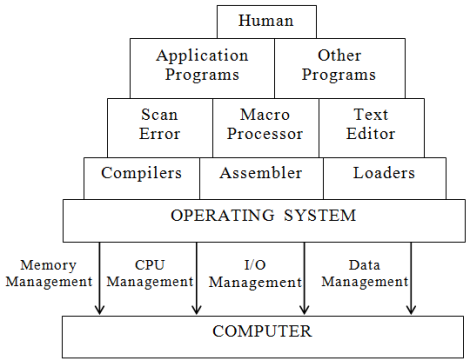

Fig-1: The Components of the Computer System

One of the most important units in the computer system is the main memory unit. Using such an important unit as the data storage medium in the most proper manner is the main purpose and necessity. This necessity made us examine the structure of the data, and investigate the relations between data structures and the memory. The term "in the most proper manner" defines the properties like access speed, small memory place usage and the facility of the method used. Methods have been developed in order to solve the problems that appear in the organization of the data structure in the memory, in structural relations as well as in the usage of the memory of the computer in the most proper manner. The methods showing the design or the form of the data being held in the memory are defined as the data structure.

To examine and define a data structure, following the stages below will ensure hypothetical clarification and safe program writing [3]

- the definition of the data structure,
- the notation of the data structure,
- access to the data structure audit group,
- access to the data element,
- usage and important algorithms.

The definition of the data structure is the abstract form of the data structure seen by the user algorithms. The compulsory initial values of the data structure and the validity indicators are defined in this stage.

www.manaraa.com

The notation of the data structure is the form of the placement of the data structure in the computer memory. A proper placement in the memory is realized by considering the size of the memory words of the computer used. Three basic placement designs, which are row first-column first and hierarchial series are used.

Access to the data structure audit group is the access to the origin of the data structure notation placed on any part of the computer memory. In other words, it is the matching of the symbolic name of the data structure and the placement address in the memory. Access to the data element is providing the validity while passing through the data structure audit group. In upper-class programming languages, when the compiler analyzes the word directory analysis of the definition of the data structure defined for the compiler, it is actually establishing the audit group of the data structure by creating the specified parameters. Once the audit group is established, it is ensured that the access, which has passed from the reliability audit of the data structure audit group, has the right to reach the right address in the memory. Various methods have been developed parallel to the development process of the computer technology for the purpose of placing the data to the memory in an efficient way, and for the purpose of accessing and processing these data [4].

2. TRADITIONAL DATA STORAGE AND PROCESSING METHODS

When traditional data storage and data processing methods are mentioned, the first methods that come to the mind are; simple and not simple data structures, database, and data mining. The data structures show the storage style of the data in the memory and the design of this storage. The data structures appear in the form of simple and not simple data structures. The information that is defined as simple data structure consists of simple numbers and letters [5]. This information is represented in one byte, and constitutes the smallest unit of the computer memory that can be addressed. The simple data structures that are defined in this way are classified as numerical simple data structures and character simple data structures. In addition to this classification, there are also logical and pointer data structures. The simple data structures consist of decimal numbers, double whole numbers, floating-point numbers, character values, logical simple data structures (TRUE/FALSE) and pointers.

The non-linear data structures, on the other hand, consist of linear lists, tree-like structures and graph structures. The linear lists has the image of single-dimensional series consisting of elements that are put in order side by side, and created by sequence of the connections between the elements. In other words, the addresses of the elements in a linear list are in consequent order. There are also lists whose addresses are not consequent, and these are called as "connector linear lists" [6]. The linear lists are separated into two categories according to their memory positions as ordinary and connective memory positions. Ordinary memory positions are separated into three sub-categories according to the form of the processes (addition, extraction, access, etc.) perfromed over the information in the lists as array, stack and queue.

When data are stored in a series, there appear problems in memory use and adding/extracting data. In order to eliminate these problems, the data structures that are called as "connective lists" are created by storing the information showing its order as well as the information that has to be kept. Although it is considered as if more memory is used in connective list usage, it requires less data place when compared with the ordinary list because a used cell is returned to the memory again [7]. Adding/extracting data in connector lists is performed more easily. Although access to a random element is easier in order assignment, the access time changes according to the distance between the searched element and the first element; because, all of the elements will be scanned in connector assignment. Connector lists are suitable for ordinary processes, and joining two or more lists is made more easily. In addition, connector lists are more beneficial because they allow the representation of complex structures. Aside from these characteristics, another benefit of connector lists is the ability of making the information according to any elements in a series without changing their places, and only by changing the connection information and initial list information. Connector lists have two different structures as circular and bilateral connector lists. In circular connector lists, the address of the first element is placed in the connection area of the last element. For this reason, concepts like "the front side of the list" or "the back side of the list" are not valid for this type. In bilateral connector lists, on the other hand, the connector information is bilateral, forward and backward. Tree-like structures: The data structures that are designed in the form of a tree and in accordance with the concepts like root, branch, leaves, etc. are called as tree-like structures. Tree-like structures are recursive, and the distribution form of the upper branches is not so different from the lower branches [8]. A tree is a cluster formed of loops in finite number, and consists of a special loop defined as the root, and sub-clusters that do not have common elements. The number of the sub-trees of a loop is called "the degree of the loop". The loops whose degrees are zero are called as "leaves". The level of the special loop that is defined as the root in a tree is taken as the first level, and the other loops are given numbers according to this special loop (the root).

Graph structures, on the other hand, are the data structures that are created by joining the data of the same cluster (Fig-2). The loops show the joining point, and the edges show the connection relation between the loops [5]. The whole of the data or a part of them may be placed in the loop or in the edge information section. Bilateral relation may be observed

www.manaraa.com

in graph-like structures. In graph-like data structures, there are no orderly situations, which is the case in tree-like structures.

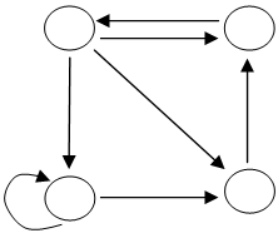

Fig-2: Oriented graph notation

Graph-like structures have an important place in computer software, and they are used in the solution of many problems. For example, the optimization of traffic or water carriage infrastructures is suitable for graph-like structures.

The complex data and file structures, too many interfile relations, and the access to them brought the problem of inadequacy. In order to solve this problem, new software technologies have been suggested in data storage and access to data, and the Database Management Systems (DBMS) approach has been proposed [9]. In this approach, the data entry and data storage are the main issues, and this process is independent from the access to the relevant data, and the smallest change in the registry and file structures causes the change of the application programs and leads to re-collection of them. Database systems are a component of computer systems, and consist of data and programs related with each other. This collection of data is called as the database. The database is the medium where the information is kept, and the database systems are the management of this medium with various software. The database includes any types of data that are necessary at the moment or that will be needed in the future [10].

In the developing technological process, the increase in the use of computers also gave rise to the increase in the amount of data/information. This increase led to the exceeding of the data analysis and the abilities of the data storage media. This inadequacy led to the development of new analysis tools aside from data structures and database concepts. Data mining is defined as obtaining the useful data by applying the rules and relations in an intense data medium [11]. The development in hardware and software technologies led to the development of proper media in developing the decision-support systems, and led to the emergence of "data warehouse" concept. The data warehouse means ensuring that all the data are used by creating the technological infrastructure of decision-support systems. The data warehouse is important in inter-relating the modular applications. It ensures the information infrastructure that is needed for the analytical processes in the time dimension (Fig-3). The data warehouse is the collection of the data prepared to support the managers in their decision-making processes. The data have a time dimension and have the properties of being intended for the subject matter. They are also integrated and are read-only [12].

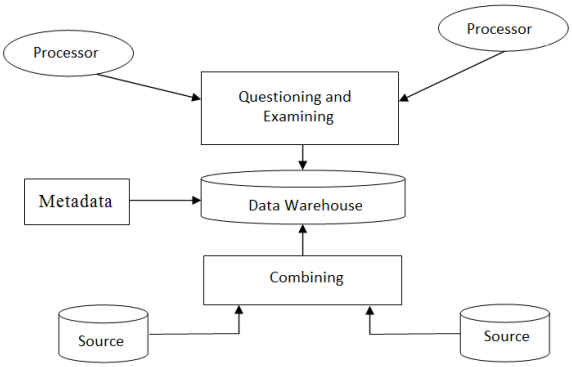

Fig-3: The architecture of data warehouse

In today's institutions and organizations, the infrastructure needed for the creation of decision support systems at strategic level is provided by the data warehouse. For this reason, the data warehouse ensures that the data are ready for query whether inside or outside the institutions and organizations [13].

Two basic structures, which are predictive and descriptive, are used in data mining. In predictive models, a model is developed by using firstly the data whose results are already known. Then, by making use of these models, the prediction of the results of the data clusters whose results are not known is performed. In descriptive models, on the other hand, it is ensured that the structures in the data that will pioneer in decision-making process are defined. These processes are shown in Fig-4.

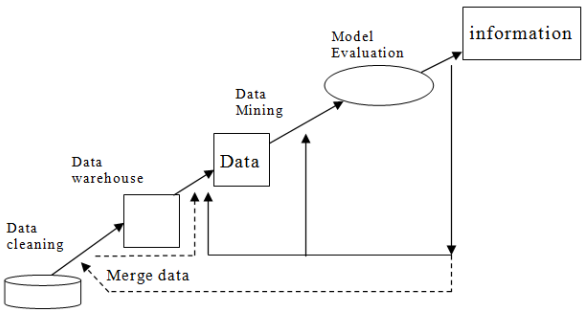

Fig-4: The Process of Obtaining Data

As the information age in which we are living requires, the importance and intensity of information have increased at a great deal. Smart phones, computers and the management of information technologies, which are the elements of the information society, have entered to every aspect of our lives. For this reason, an intensified data amount has begun to be collected in a qualified and meaningful manner. As a result of this, the speed of the access to the data has also increased parallel to the increase in the amount of data. The changes in the data/information in terms of quantity brought with it the changes in terms of quality as well.

www.manaraa.com

The intense collection of data/information in a way that created a meaningful whole was first performed in astronomy and genetics sciences. Today, this phenomenon has shown itself in every aspect of our lives. Big data are defined as the intense and complex data clusters that cannot be processed by existing data systems. In other words, the data clusters that exceed the collection, storing and analyzing ability of the known database management systems and software elements are called as big data. Today, this size has increased from terabyte to petabyte (10^{15} bytes). The data collected from various sources like social media sharing, blogs, photographs, videos, log files etc. have reached a size that cannot be stored in traditional structures. These intense data must be converted into meaningful and processable form. For this reason, big data consist of the logs of the web providers, internet statistics, social media publishing, blogs, micro blogs, climate sensors and similar other sensors and the call registers of the GSM operators[14].

Today's traditional structures are not adequate for storing these data. Since the basis is the integrity of the data in relational databases, it is slower when compared with big data analyses. In addition, while the processes are at gigabyte level in relational databases; petabyte level and batch processing are mentioned for the analyses of big data. Since the big data approach works according to the distributed file system, it is not possible to talk about the integrity of the data. In other words, there are no charts and relation tables, which is the case in relational databases. Since big data have no possibility of ensuring all the rules like consistency, availability and partition tolerance; the loss of some data or some data being incorrect are not very important when the size of the data is considered. For this reason, solutions have been created to join the big data with distributed file systems of simple hardware (like MapReduce, Hadoop, Storm, Hana and NoSQL). The MapReduce has been developed by Google in order to process the problems by dividing them into different units. Facebook, which is one of today's social networks, has an extremely big Hadoop cluster. Another social network, Twitter, has developed Strom, which allows the processing of real-time data. Hana, which is developed by SAP, enables faster processing of the data in the main memory unit instead of keeping them in disk medium. NoSQL (Not only SQL) and Hadoop are the most frequently used ones in today's world.

In this study, Apache Hadoop has been used for the purpose of keeping and query of the data in big amounts. Apache Hadoop ensures that big data are analyzed in different computers simultaneously. The data to be analyzed are kept on HDFS (Hadoop File System), and Hadoop processes on the clusters created by the other computers [15]. This structure ensures that both the data and the jobs are distributed. Apache Pig and Apache Hive ensure that SQL-type queries are performed and converted into Hadoop jobs, and speeds up the development stage. This phenomenon, which is an open source, has abstracted the Map-Reduce algorithm, and facilitated the learning threshold. Apache Oozie, on the other hand, has been developed to ensure that the Hadoop jobs that are defined in a flow are processed with an order and with certain intervals.

In this study, the performance measurements of data in big amounts over Hadoop and over traditional database management systems have been examined. Different queries have been operated on two different datasets, which are 4 GB and 6 GB in size. The same data have been transferred onto a relational database. Then, the performance measurements of the queries have been performed on Hadoop and relational database management system (MSSQL) on two computers with the same configurations.

3. APPLICATION and PERFORMANCE ANALYSIS

This study has been tested on datasets, which are 4 GB and 6 GB in size. A database, which is provided by Minnesota University as free-of-charge, has been made use of in order to test the results of the study. The database has been prepared as two sets, which are 4 GB and 6 GB in size. Firstly, these datasets are transferred to a traditional database management system. For this reason, the queries given in Table 1 have been operated separately in computers with the same configurations as the test medium, and the time measurements have been performed in this way.

Table-1: Query operated on Hadoop and traditional database

Query Number	Query
S1	Select * from ratings
S2	Select movieId from ratings
S3	Select * from ratings order by userId
S4	Select count(*) from ratings
S5	Select count(*) from ratings group by userId
S6	Select * from ratings where ratings like '5'
S7	Select movieId from ratings where ratings like '5'
S8	Select count(*) from ratings where ratings like '5'
S9	Select max(ratings) from ratings
S10	Select min(ratings) from ratings
S11	Select average(rating) from ratings
S12	Select average(rating) from ratings group by movieId
S13	Select * from ratings order by userId asc
S14	Select * from ratings order by userId desc
S15	Select * from ratings order by rating
S16	Select count(*) as nbratings, rating from ratings group by rating order by rating desc
S17	Select count(userid) as nbusers, avgrating from (select round(avg(rating),1) as avgrating, userId from ratings group by userId) as avgratingbyusers group by avgrating order by avgrating desc
S18	Select useId, col1, movieId, count(*) as sumRows From (Selectuseld, col1, movieId, row_number() over (Partition Byuseld) as row from ratings) rs Where row <= 10000 Group Byuserld, col1, movieId
S19	Select * From(Select userId from ratings Union All Select movieId from ratings) rn
S20	Select * From (Select userId from ratings Union All Select movieId from ratings) rn Order by userId

www.manaraa.com

The results obtained by operating the queries in computers with the same configuration are given in Table 2. The queries have been operated on the datasets in different sizes.

Table-2: Query Times operated on Hadoop and traditional database

The Size of the Dataset	Query Number	Traditional DBMS (sec)	Hadoop (sec)
4 GB	S1	8	7.1
	S2	8.2	7.5
	S3	8.5	7.8
	S4	8.3	7.6
	S5	11.54	7.65
	S6	8.9	8
	S7	15.4	8.7
	S8	8	6.2
	S9	10	8.4
	S10	10	8.4
	S11	11.2	10.4
	S12	13.1	12.1
	S13	8.1	6.4
	S14	8.2	6.2
	S15	7.9	7
	S16	14.2	12.6
	S17	15	13.7
	S18	16	15.2
	S19	13.2	10.7
	S20	14.7	13.5
8 GB	S1	10	8.4
	S2	11.1	8.8
	S3	10.5	9.2
	S4	12	8.8
	S5	15.2	8.6
	S6	12.6	11.3
	S7	18.2	11.7
	S8	9.14	6.88
	S9	12	10.2
	S10	12	10.2
	S11	13.8	12.4
	S12	16	14.3
	S13	10	9
	S14	11	8.4
	S15	9.6	8.1
	S16	16.7	15.1
	S17	18.2	15.5
	S18	22.5	20.3
	S19	23.4	20.7
	S20	24.5	22.9

The result graphics obtained from the queries operated on data whose data sizes are bigger than 4 GB are given in Chart-1; and the result graphics obtained from the queries operated on data whose data sizes are bigger than 6 GB are given in Chart-2.

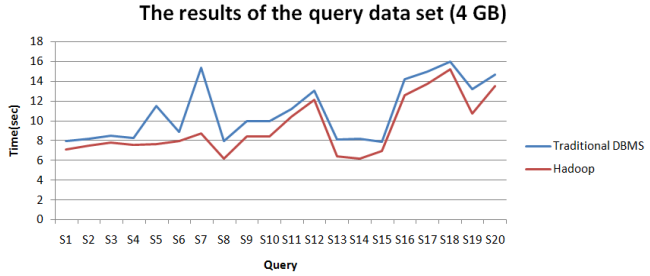

Chart-1: 4 GB Dataset query results

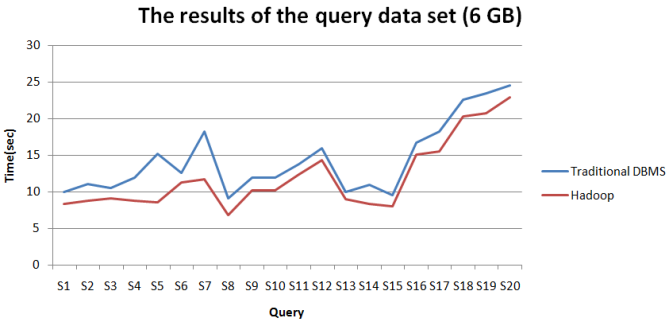

Chart-2: 6 GB Dataset query results

The complexity of the query and the size of the dataset used are important factors influencing the query times. The complexity of the query and the size of the dataset used are important factors influencing the query times. When the result graphics are considered, it is observed that the query times in both datasets are extremely short in queries whose query times are short. On the other hand, it has been observed that the query times are longer in queries whose query times are bigger. In addition, it has also been observed that the resulting duration of the same queries is relatively longer in traditional database management systems. It has been determined that the queries in datasets like Hadoop result in relatively shorter durations no matter how complex the query is or how big the dataset is. This situation ensures that the query, analysis etc. processes are performed in a more easily way.

4. CONCLUSIONS

The big data has begun to occupy an important place among daily activities of many institutions. In addition to this, the big data technology will be the new generation technology that will be applied after a short while by almost all companies. Traditional database management systems are incapable of covering the growing data needs with their inadequacy in incapability in collimating and dividing more than one. Hadoop is an open source code used commonly and accepted widely in order to calculate the big data analytics in an easily scalable medium. In addition, Hadoop has the characteristics of storing and analyzing unstructured data and supports reliable, low-cost, distributed parallel programming. As a result of this, it is preferred by Google, Yahoo and Facebook, the pioneers of the sector. The previous versions of Hadoop did not have real-time data analysis component; however, it introduced the Apache Spark for real-time big data analysis in recent times. Spark is based on data that are distributed in a flexible manner, and it is claimed that it provides the results in a time no more than half a second. As a future study, establishing a real-time big data analytic engine will be interesting.

www.manaraa.com

REFERENCES

[1] M. Andrews, "Computer Organization", Computer Science Press, 1987.

[2] T. Bartee, "Digital Computer Fundamentals", McGraw-Hill Book Co., 1985.

[3] E. Horowitz and S. Sahni, "Fundamentals of Data Structures", Computer Science Press Inc., 1982.

[4] E. Horowitz and S. Sahni, "Fundamentals of Computer Algorithms", Pitman Publishing Ltd., 1976.

[5] R. Çölkesen, "Veri Yapıları ve Algoritmalar", Papatya Yayıncılık, 2004, İstanbul, Türkiye.

[6] J. Tremblay and P. Sorenson, "An Introduction to Data Structures with Applications", McGraw-Hill Book Co., 1984, New York, USA.

[7] S. Kurnaz, "Veri Yapıları ve Algoritma Temelleri", Papatya Yayıncılık, 2004, İstanbul, Türkiye.

[8] C.A.R. Hoare, "Recursive Data Structures", International Journal of Computer and Information Sciences, vol. 4, 1975, pp. 105-32.

[9] F. Keskinel, "Veri Tabanı Kavramı", Enka Yayınları, 1985, İstanbul, Türkiye.

[10] Y. Özkan,"Veri Tabanı Sistemleri", Alfa Basım Yayım Dağıtım Ltd. Şti., 2009.

[11] H. Kaya and K. Köymen,"Veri Madenciliği Kavramı ve Uygulama Alanları", Doğu Anadolu Bölgesi Araştırmaları, 2008, pp. 159-164.

[12] Y. Özkan, "Veri Ambarı", Türkiye Bilişim Ansiklopedisi, 2006, pp. 879-883.

[13] A. Baykal, "Veri Madenciliği Uygulama Alanları", D.Ü.Ziya Gökalp Eğitim Fakültesi Dergisi, Vol. 7, 2006, pp.95-107.

[14] B. Demirtaş and M. Argan,"Büyük Veri ve Pazarlamadaki Dönüşüm: Kuramsal Bir Yaklaşım", Pazarlama ve Pazarlama Araştırmaları Dergisi, Vol. 15, 2015, pp. 1-21.

[15] İ. Karaca, "Büyük Veri Analizlerinin Kurumsal Faaliyetlerde Kullanım Alanları", Ankara Üniversitesi, PhD Thesis, 2015.

BIOGRAPHIES

Can Razbonyalı graduated from Maltepe University, Turkey, in 2007. He took M.Sc. degree from Trakya University, Turkey, in 2011, all in Computer Engineering. He took Ph.D. degree from Computer Engineering, Okan University, Turkey, in 2016. His interest areas include big data, data mining, machine learning and computer programming.

Erdal Güvenoğlu graduated from Trakya University, Turkey, in 2001. He took M.Sc. degree from Trakya University, Turkey, in 2005, all in Computer Engineering. He took Ph.D. degree from Computer Engineering, Trakya University, Turkey, in 2012. His interest areas include image processing, image encryption, steganography and computer programming.

المنارة للاستشارات
www.manaraa.com