sustainability

MDPI

Article

Design of a Data Management Reference Architecture for Sustainable Agriculture

Görkem Giray [1] and Cagatay Catal [2,*]

[1] Independent Researcher, Izmir 35320, Turkey; gorkemgiray@gmail.com
[2] Department of Computer Science and Engineering, Qatar University, Doha 2713, Qatar
* Correspondence: ccatal@qu.edu.qa; Tel.: +974-4403-4261

Abstract: Effective and efficient data management is crucial for smart farming and precision agriculture. To realize operational efficiency, full automation, and high productivity in agricultural systems, different kinds of data are collected from operational systems using different sensors, stored in different systems, and processed using advanced techniques, such as machine learning and deep learning. Due to the complexity of data management operations, a data management reference architecture is required. While there are different initiatives to design data management reference architectures, a data management reference architecture for sustainable agriculture is missing. In this study, we follow domain scoping, domain modeling, and reference architecture design stages to design the reference architecture for sustainable agriculture. Four case studies were performed to demonstrate the applicability of the reference architecture. This study shows that the proposed data management reference architecture is practical and effective for sustainable agriculture.

Keywords: sustainability; agriculture; sustainable agriculture; data management; reference architecture; design science research

Citation: Giray, G.; Catal, C. Design of a Data Management Reference Architecture for Sustainable Agriculture. *Sustainability* **2021**, *13*, 7309. https://doi.org/10.3390/su13137309

Academic Editor: José Manuel Mirás-Avalos

Received: 16 May 2021
Accepted: 25 June 2021
Published: 30 June 2021

1. Introduction

The increasing food demand and its large ecological footprint call for action in agricultural production [1]. Inputs and assets should be optimized; long-term ecological impacts should be assessed for sustainable agriculture. Decision-making processes on optimization and assessment need data on several inputs, outputs, and external factors. To this end, various systems have been developed for data acquisition and management to enable precision agriculture [1]. Precision agriculture refers to the application of technologies and principles for improving crop performance and environmental sustainability [2]. Smart farming extends precision agriculture and enhances decision-making capabilities by using recent technologies for smart sensing, monitoring, analysis, planning, and control [1]. Data to be acquired are enhanced by context, situation, and location awareness [1]. Real time sensors are utilized to collect various data and real time actuators are used to fine-tune production parameters instantly.

Murakami et al. [3] and Steinberger et al. [4] pointed out a need for data storage and a processing platform for agricultural production. They utilized web services to send and receive data from a central web application. This web application receives, stores, and processes data and provides the required outputs to its users or any other system. Similarly, Sørensen et al. [5] listed several data processing use cases to assist farmers' decision-making processes. Recent technologies, such as the Internet of Things (IoT), make digital data acquisition and hence, smart farming possible [6]. In recent years, many studies have been performed in smart farming and precision agriculture [7–13]. Currently, Industry 4.0 acts as a transformative force on smart farming processes. Industry 4.0-related technologies, namely the IoT, big data, edge computing, 3D printing, augmented reality, collaborative robotics, data science, cloud computing, cyber-physical systems, digital twins,

cybersecurity, and real-time optimization are integrated into different parts of agricultural systems [14].

To realize operational efficiency, full automation, and high productivity in these systems, different types of data are collected from operational systems using different sensors, stored in big data systems, and processed using machine learning and deep learning approaches. Traditional data management techniques and systems are not sufficient to deal with this scale of data, therefore, big data infrastructures and systems have been designed and implemented. To manage the complexity of this big data, many different aspects of data must be considered during the design of these systems. Different data management reference architectures have been designed to date [15–17]. To the best of our knowledge, none of these studies have focused on sustainable agriculture. There exist several practices for sustainable agriculture that can protect the environment, improve soil fertility, and increase natural resources. It is known that agriculture can affect soil erosion, water quality, human health, and pollination services [18]. As such, sustainable agriculture is crucial to minimize the negative effects of agricultural production. Sustainable agriculture requires an iterative process because each actor in this system has a different responsibility, and the success of this process is highly dependent on the success of each actor.

The goal of this study is to present a data management reference architecture for supporting smart farming and sustainable agriculture. The study builds on the recent developments in data management and processing, i.e., big data, machine learning, and data lake. We designed a data management reference architecture for sustainable agriculture and evaluated it using several case studies. Domain scoping, domain modeling, and reference architecture design stages were followed to create the reference architecture. Based on the reference architecture, we can design different application architectures. During the validation stage of this study, we have shown the applicability of our reference architecture.

The contributions of this study are presented as follows:

- A data management reference architecture design approach is presented, which can be used for different application domains;
- By using this design approach, a novel data management reference architecture for sustainable agriculture was designed for the first time in literature;
- The reference architecture is validated using different case studies obtained from the literature.

The structure of the paper follows the outline proposed by Gregor & Hevner [19] for design science research. Section 2 summarizes the research method adopted in this study. Section 3 defines and structures the problem by analyzing the existing literature. Section 4 starts with the related reference architecture studies and then explains the solution design process and the reference architecture obtained. Section 5 presents the evaluation of the reference architecture by deriving application architectures from it based on some requirements from the sustainable agriculture domain. Section 6 discusses the results and Section 7 concludes the paper.

2. Research Method

The design science research (DSR) method proposed by Hevner et al. [20] was followed in this study. DSR is a problem-solving paradigm and seeks to create artifacts through which information systems can be effectively and efficiently engineered [20]. These artifacts are designed to interact with a problem context to improve something in that context [21].

The activities and the artifacts span two significant dimensions, i.e., problem-solution and theory-practice dimensions [22]. Figure 1 shows the research method used in this study. The first step was the identification of some problem instances occurring in practice and sharing similar aspects. These problem instances were analyzed, and a problem statement was formed using theoretical concepts from the literature. A conceptual solution, i.e., an artifact or artifacts, was designed by following a systematic approach. Domain analysis was used to derive and represent domain knowledge to be used for solution design. Domain analysis involved domain scoping and domain modeling activities [23]. Domain scoping

refers to the identification of relevant knowledge sources to derive the key concepts of the solution [24]. To this end, several searches were conducted on the Scopus database using different search strings. Domain modeling aims at unifying and representing the domain knowledge obtained from relevant sources. The feature model was used to represent the output of domain modeling [25]. A reference architecture was designed as a conceptual solution.

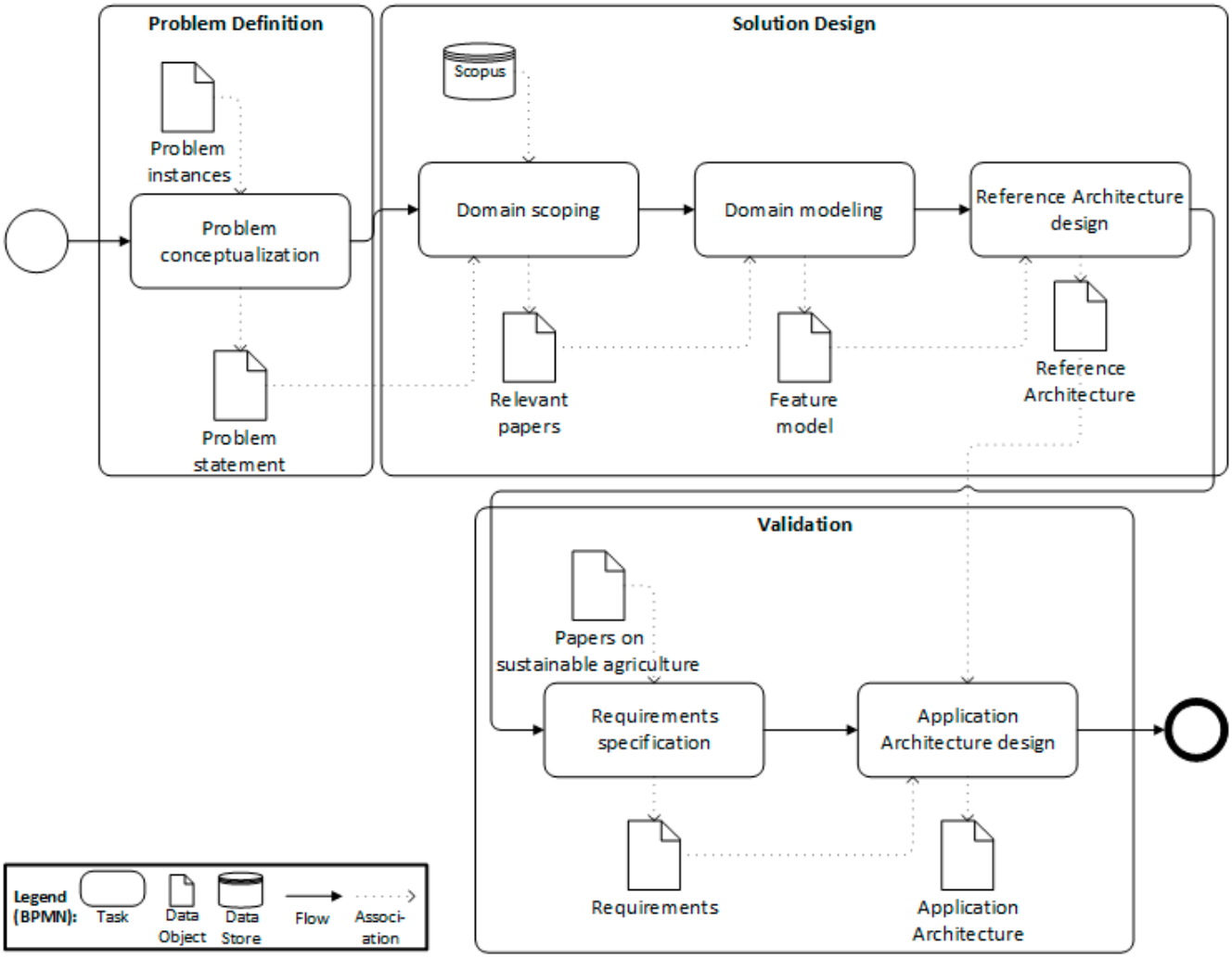

Figure 1. The research method used in this study, which involves the main steps of DSR, i.e., problem definition, solution design, and validation. The iterative nature of the research method was neglected for the sake of simplicity.

To evaluate the reference architecture, requirements were specified using the recent literature on sustainable agriculture [26–30]. Based on these requirements, a concrete application architecture was derived using the reference architecture.

In accordance with the research method, the following sections present problem definition, design of a solution, and an evaluation of the solution.

3. Problem Definition

This study was motivated by different cases involving different data management requirements to support sustainable agriculture. The following three cases were used for understanding and conceptualizing the problem.

- Case 1: Satellite images (e.g., Sentinel-2 data) can be obtained from a data provider. These images can be processed to derive plant parameters such as Leaf Area Index

(LAI), biomass, and chlorophyll content during the growing season [31]. Afterward, the current growth status and development of cultivated crops at each location in the field can be deduced [32]. This information can be used for site-specific plant protection and fertilization measures [33] and hence, support sustainable agriculture.

- Case 2: Harvested crop volume can be quantified and recorded in real time using numerous sensors [34]. Various parameters such as quantity per hectare and flow can be calculated and crop productivity maps can be built [34]. Farmers can use these maps to optimize inputs such as fertilizers, pesticides, and seeding rates, and increase yields [35].
- Case 3: Machinery process data such as speed, angle, pressure, and flow rate can be obtained through sensors in tractors and equipment [4]. Machine, worker, field, and time slot data can be stored, and basic statistics like minimum, maximum, standard deviation can be computed [4]. As a result, automated documentation of the production process and site-specific work can be attained [4].

Table 1 summarizes the above-mentioned cases from a data management perspective. Similar to many cases in various domains, at a high level, digital data are produced and fed to a software system to be processed and stored. Such a system can be named as a data management platform and produce outputs that can lead to better business outcomes. As per the first case, satellite images can be processed via computer vision algorithms to drive plant parameters such as Leaf Area Index (LAI), biomass, and chlorophyll content, which can be used to track the current growth status of cultivated crops and support decision-making activities.

Table 1. A summary of the cases presented above from a data management perspective.

Data Input	Data Processing	Data Output	Outcome
satellite images	derive plant parameters via computer vision algorithms	plant parameters such as Leaf Area Index (LAI), biomass, and chlorophyll content	track current growth status and development of cultivated crop at each location
harvested crops volume via sensors	build crop productivity maps	various parameters such as quantity per hectare and flow on the map	use such maps to optimize inputs such as fertilizers, pesticides, and seeding rates, and increase yields
machinery process data via sensors	compute statistics	machine, worker, field, and time slot data along with basic statistics such as minimum, maximum, and standard deviation	attain automated documentation of production process and site-specific work

The next section describes the solution design and the reference architecture.

4. Solution Design and Artifact Description

This section starts with a summary of related reference architecture studies and then presents the three steps of the solution design phase, namely, domain scoping, domain modeling, and reference architecture design.

4.1. Related Reference Architecture Studies

Before presenting our reference architecture, we discuss the available reference architectures in the literature. While Nikkilä et al. [36] and Kaloxylos et al. [37] presented architectural aspects of Farm Management Information Systems (FMISs), they did not propose a reference architecture.

Tummers et al. [17] designed a reference architecture for FMISs. They first identified the stakeholders and their concerns. Afterward, a feature model for FMISs was created. The reference architecture was designed and represented via context and decomposition views. Three case studies were performed to show the applicability of the proposed reference architecture.

Köksal & Tekinerdogan [6] proposed a reference architecture for IoT-based FMISs. They proposed an architecture design method and showed that the approach is practical and effective. Their architecture includes the following main features: data acquisition, data processing, data visualization, system management, and external services. Each main feature consists of several sub-features. For instance, the data processing feature involves the following sub-features: image/video processing, data mining, decision support, and data logging. They used decomposition, layered, and deployment views to document the reference architecture. For deriving a concrete FMIS architecture, their reference architecture can be used.

Kruize et al. [38] proposed a reference architecture for farm software ecosystems. Farm software ecosystems aim to fulfill the needs of several actors in the smart farming domain. In that respect, their scope is much wider compared to FMISs. The farm software ecosystem reference architecture mainly focuses on the problem of bringing various software and hardware components together to form a platform for multiple actors.

To the best of our knowledge, there is no other study that presents a data management reference architecture for sustainable agriculture. Although some of the previous studies mention several data-related components, a complete architectural view of managing data for sustainable agriculture was missing. As such, our reference architecture study aims to fill this research gap.

4.2. Domain Scoping

The Scopus database was used as the knowledge source for domain scoping. To identify the search keywords, it is crucial to understand the recent factors driving reference architectures for data management. The big data concept emerged to highlight challenges of data management, including volume, velocity, and variety [39]. Machine learning is another hot research topic, which tries to acquire knowledge by extracting patterns from raw data [40,41] and solve problems using this knowledge. Data lake is another recent concept to emerge that addresses the shortcomings of data warehouses. A data lake can be defined as a data management platform that allows the storing of both structured and unstructured data, unlike data warehouses that handle only structured data. This type of platform is designed to enable big data processing, real time analytics, and machine learning. Based on these recent trends, four search strings were used to form the initial paper pool. The phrase "reference architecture" was combined with four phrases representing the recent trends in data processing for sustainable agriculture (i.e., data management, big data, machine learning, and data lake). The search keywords were kept general to have a high recall and relatively low precision. Although this required more effort from the authors, obtaining a broader initial set of papers decreased the possibility of missing relevant studies.

The database search on Scopus was conducted in February 2021. No criterion was set for the publication date. A total number of 270 papers was obtained for the pool of candidate papers. All the results were combined in an Excel sheet, which included useful information about the papers such as title, abstract, keywords, and publication date, which are used in further steps.

To identify the relevant papers for designing a reference architecture, authors applied the exclusion criteria to the papers obtained from Scopus. The papers, which were duplicated, not written in English, or without a full text available, were filtered out. The papers involving a reference architecture to process data in any business domain were included. Seven papers included a reference architecture for data processing, along with the essential components.

The data extraction phase followed the selection of relevant papers. The components of the data processing architecture listed in the papers were extracted and recorded in an Excel sheet. These components were unified by reading the definitions presented in the papers. Table 2 shows the unified list of the components and the source papers where each component was identified.

Table 2. The unified list of components identified in the literature.

	[15]	[42]	[43]	[44]	[45]	[46]	[47]
Ingestion	✓	✓	✓	✓	✓	✓	✓
Information Extraction							✓
Data Quality Management			✓	✓	✓	✓	✓
Integration		✓		✓		✓	✓
Analysis	✓	✓	✓	✓	✓	✓	✓
Storage	✓	✓	✓	✓	✓	✓	✓
Security/Privacy				✓		✓	
Metadata Management	✓						
Replication/Archiving					✓	✓	✓

All the papers address three main components that deal with: collecting data from various sources (i.e., acquisition), processing data to provide some value to data consumers (i.e., analysis), and persist data (i.e., storage). Pääkkönen and Pakkala [47] identified information extraction as a component to extract structured data from unstructured and semi-structured data, such as email or images. Data quality management is another vital component to handle data quality problems. Data received from various sources should be integrated for further analysis. Security and privacy components are used to protect data from unauthorized data and for proper handling of personal data. Metadata management refers to the creation and storage of metadata to document the meaning of data. The replication component ensures redundant storage of data sources to provide better data availability in case of technical problems. The archiving component is responsible for storing cold data for future possible needs.

4.3. Domain Modeling

Feature modeling is one of the approaches to represent domain knowledge in a reusable format [24,25]. Figure 2 shows the feature model, which is derived from the unified list of components presented in Table 2. A feature diagram can include mandatory and optional features. Three components identified by all papers are treated as mandatory features. The remaining features are optional and can be used depending on business requirements.

Data are needed to be onboarded to a data platform for further processing and storage, a process referred to as the ETL (extract, transform, and load) process in traditional data warehousing architectures [48]. Such architectures possess a pre-defined data schema and data are loaded based on this schema. Data ingestion refers to the process of transferring data from providers to a platform for further processing [49]. The umbrella term for such processes is data acquisition. Data can be acquired in batches at regular intervals or in real time (or in near real time) as streams. A component acquiring data streams should be able to handle data with high velocity [50].

Information extraction is intended for obtaining useful information from unstructured and semi-structured data [51]. Unstructured and semi-structured data may include natural language text, image, audio, and video. Several tasks performed under information extraction include classification, named entity recognition, relationship extraction, and structure extraction [16,48,52]. Named entity recognition (i.e., named entity identification) aims at identifying and classifying named entities in unstructured or semi-structured texts into predefined categories such as a person, organization, or location. For instance, Gangadharan and Gupta [53] extracted names of crops, soil types, crop diseases, pathogen names, and fertilizers from documents on agriculture. Relation extraction is the task of detecting and classifying predefined types of associations among recognized entities [52]. For example, relationships among crop diseases and locations can be extracted. The sub-features for information extraction can be expanded depending on domain-specific requirements.

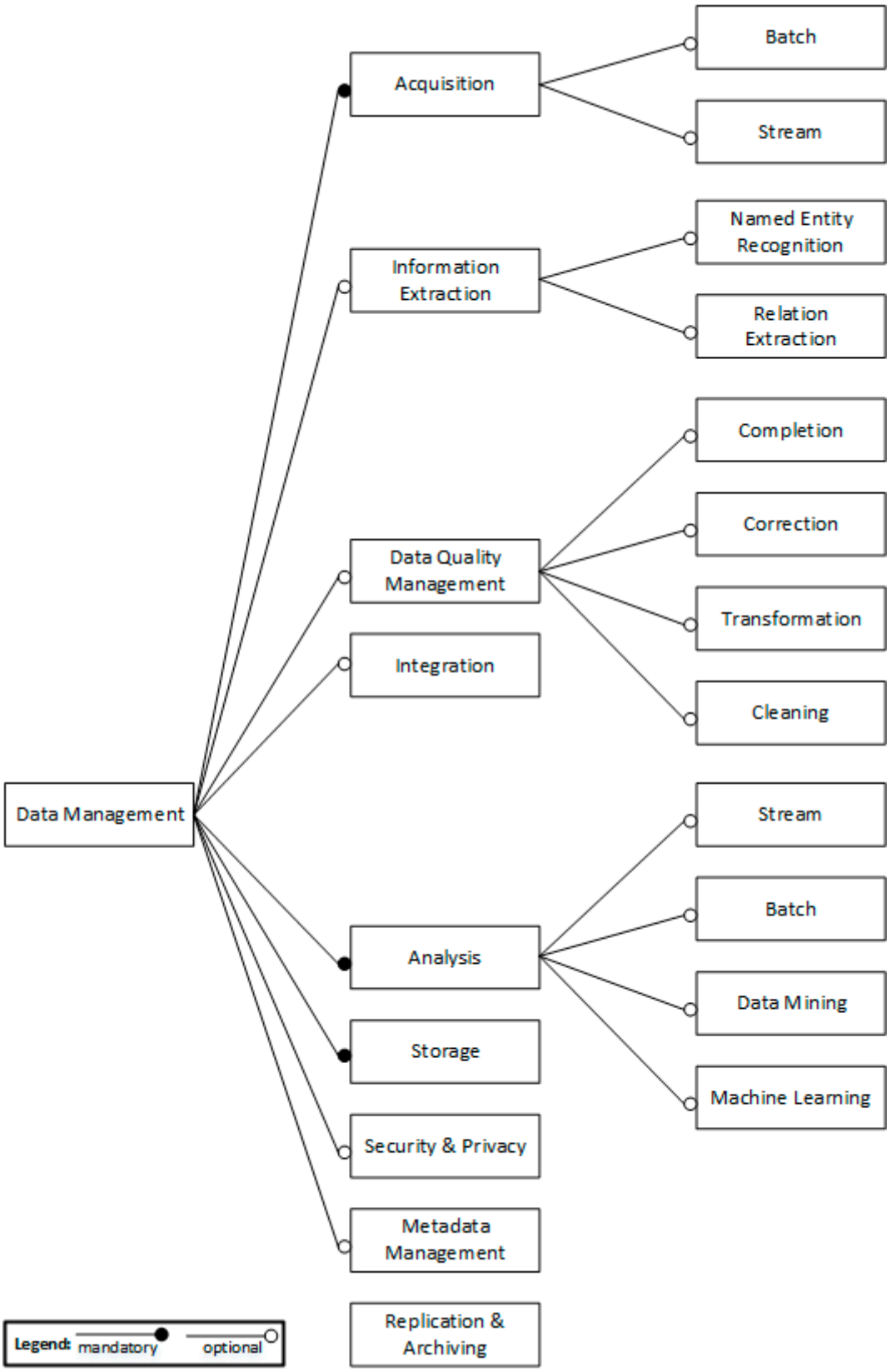

Figure 2. The feature model for data management in the sustainable agriculture domain.

Data quality management refers to handling data quality problems that may arise due to several reasons [48]. There may be missing, incorrect, unusable, or redundant data [54].

To address these main data quality problems, missing data can be completed, incorrect data can be corrected, unusable data can be transformed, and redundant data can be cleaned.

In general, data management platforms acquire data from multiple sources that usually involve differences in data models, schemas, and data semantics [55]. Data integration aims at combining heterogeneous data and providing a unified view of these data [56]. One technique for data integration is schema mapping, which refers to conveying the data schemas of multiple data sources into one global common schema [57].

Data are analyzed to obtain some value from them. Results of analysis may provide insights to users and constitute some intermediate output for further processing [48]. Stream analysis encompasses the timely processing of flowing data and generates required outputs. For instance, an environmental monitoring system can process raw data coming from sensor networks to identify critical cases [58]. On the contrary, batch analysis is conducted on static datasets [59]. Data mining and machine learning, including deep learning algorithms, may be utilized to produce deeper analyses [60,61].

Storage is a feature supporting other features and refers to the temporary and persistent storage of data. To manage the increased volume, velocity, and variety of data [39], different types of data stores are released. Therefore, the storage feature involves various database management systems (e.g., Microsoft SQL Server, Oracle, PostgreSQL, or MongoDB) implementing different data models such as relational or nonrelational (or NoSQL).

The security and privacy feature addresses authentication and authorization, access tracking, and data anonymization [48]. Several standards, guidelines, and mechanisms can affect the realization of this feature, such as data encryption standards and mechanisms, access guidelines, and remote access standards [62].

Metadata management is related to planning, implementation, and control activities to enable access to metadata [62]. This feature mainly involves capabilities related to collecting and integrating metadata from diverse sources and providing a standard way to access these metadata [62].

The replication feature manages the storage of the same data on multiple storage devices [62]. While having replicated instances of data support high-availability, data consistency may become an issue to deal with. The archiving feature addresses the movement of infrequently used data onto media with a lower retrieval performance [62].

4.4. Reference Architecture Design

Based on the abstraction derived from the cases presented in Section 3, Figure 3 shows the context diagram of a data management system. The context diagram shows the overall purpose of the system and its interfaces with the external environment [63]. At a very high level, some data providers send data to a data management system. These may be humans entering data through a graphical user interface or external systems providing input data to be processed. Data obtained from data providers are stored, processed, and served to data consumers based on their requirements.

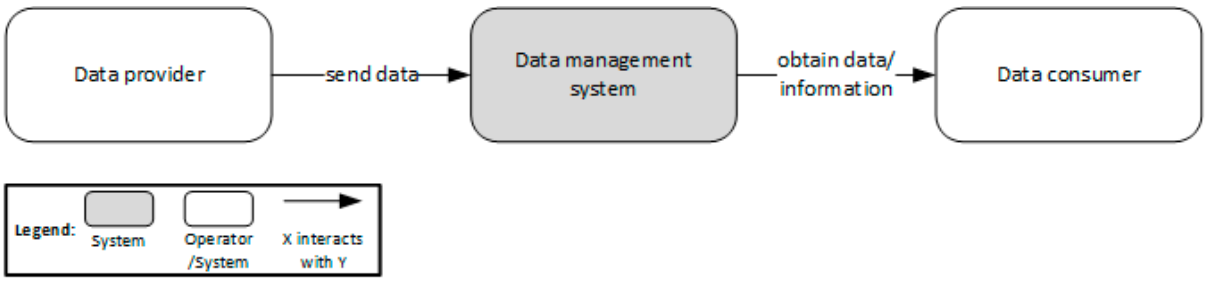

Figure 3. The context diagram for a data management system to conceptualize the cases identified in the sustainable agriculture domain.

Figure 4 shows the decomposition view of the data management reference architecture proposed for data processing.

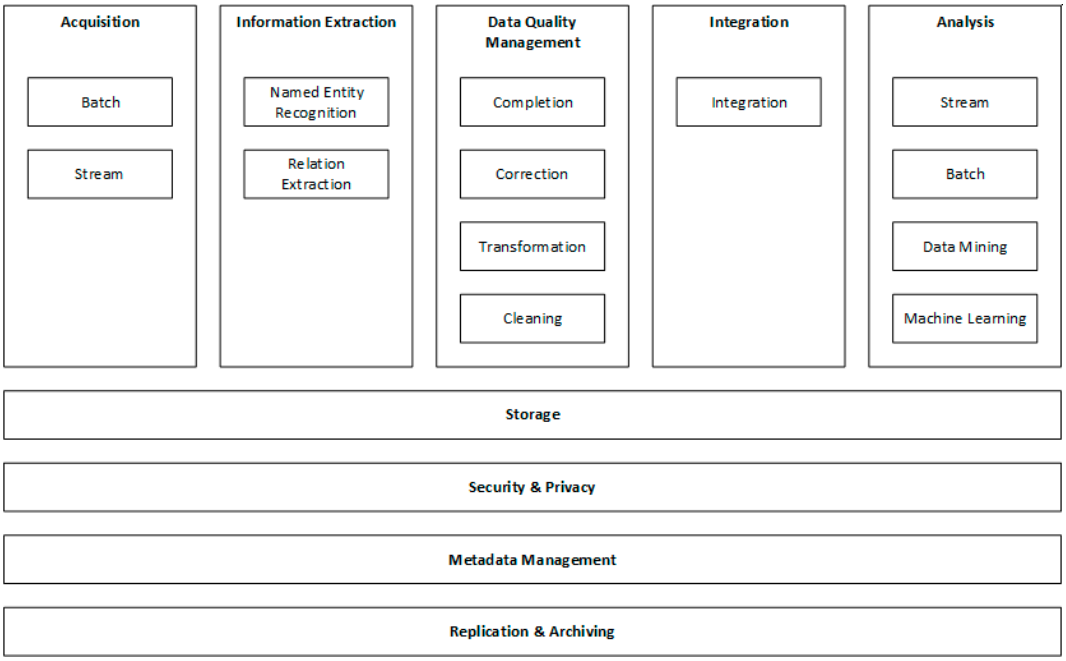

Figure 4. Decomposition view of the data management reference architecture.

Acquisition components are responsible for onboarding data to the data management platform for further processing. Useful information such as named entities and relations can be extracted from the acquired data. Quality problems can be resolved by completing, correcting, transforming, and cleaning the acquired data to obtain more accurate results from analyses. The data obtained from various sources need to be integrated to end up with better and richer insights. Various components can be used to analyze data to support decision-making. The storage component handles different modes of storing data. The security and privacy component is needed to protect data from unauthorized access. Data on the description of acquired data, are handled by the metadata management component. Replication may be needed for high availability. An archive component is usually required to manage the process of archiving unused historical data.

The next section presents how an application architecture can be derived from the reference architecture based on the requirements in sustainable agriculture.

5. Validation

To evaluate the reference architecture, a set of requirements was extracted from recent papers on sustainable agriculture [26–30]. The high-level requirements address issues on crop yield prediction [27], irrigation management [28], real time variable-rate fertilization [30], and exotic animal infectious diseases monitoring [26]. These diverse high-level requirements address various data management aspects in sustainable architecture.

Sustainability is a long-term, high-level goal involving several aspects and sub-goals. Generally, there is a gap between software requirements and sustainability goals. While software requirements tend to be more tangible, sustainability goals tend to be more intangible. Therefore, a tangible decomposition of sustainability and the ability to map it to concrete software requirements are required to monitor the achievement of sustainability goals [64].

Penzenstadler and Femmer proposed a reference model to show the dimensions of sustainability (i.e., individual, social, economic, environmental, and technical) and map them to high-level software requirements [64,65]. Figure 5 shows the high-level software

requirements obtained from the literature and how they are mapped in the sustainability model proposed by Penzenstadler & Femmer [64,65].

The goal of sustainability has several dimensions such as economic and environmental. There are moral and natural goods that are perceived as an expression of a specific dimension [64]. Long-term profit and healthy environment are two examples of values contributing to economic and environmental dimensions, respectively. Indicators are qualitative or quantitative metrics that express a specific degree or score regarding a value [64]. Consumption amounts of resources, water, agricultural pesticide, and weedicide are examples of indicators. Activities are measures taken to contribute to values [64] and can have different levels of granularities, which are associated with each other. Lower level activities such as crop phenotypic monitoring, shown in Figure 5, can contribute to a higher-level activity such as variable-rate fertilization management. Lower level leaf activities, such as predicting crop yield, shown in Figure 5, can be treated as the high-level software requirements (e.g., high-level use cases or epics) against which one or more system features are developed. The components of the data management reference architecture used to realize each high-level software requirement are described as follows.

Crop yield prediction is an essential task for growers and farmers to decide on what and when to grow [66]. However, it is extremely challenging due to numerous complex factors [67]. To overcome this challenge, researchers started to use machine learning (ML) algorithms to predict crop yield based on various input variables [27]. Mohsen et al. [27] suggest using weather, soil, plant population, and planting process data. These historical data are extracted from various sources such as surveys [68,69] and stored (*acquisition: batch*). Unusable data such as rows with missing values are removed (*cleaning*) and some of the values are normalized (*transformation*). The data obtained from different sources are combined (*integration*). Several ML models are built through experimentation (*machine learning: model development*). One of the obtained ML models that exhibits a satisfactory performance is deployed (*machine learning: model deployment*). The performance of the deployed model should be monitored for a possible performance degradation [70,71]. When the performance does not meet expectations, a new ML model should be trained and deployed.

Wireless sensor networks using IoT technologies can be utilized for irrigation management [28]. Sensors can measure real time environmental data such as soil moisture in predefined periods and send these data to the data storage over the Internet (*acquisition: stream*) [72]. A weather forecast can be obtained from a data provider (*acquisition: batch*) to manage irrigation by considering the conditions that affect the irrigation process, such as rain or strong winds [72]. The measurement and forecast data are combined (*integration*) by considering time dimension, i.e., data obtained from different sources must fit into the same temporal window [73]. Based on the combined data, irrigation decisions can be drawn based on predefined rules [28] or a prediction model (*machine learning*) [73]. The data on this decision can be sent to an actuator to control irrigation.

Crop phenotypic information can be used to enable real time variable-rate fertilization [74,75]. The predictors of phenotypic information involve crop three-dimensional size, biomass, and vegetation index as well as other indicators [30]. These data can be acquired using aviation-based [76,77] and ground-based [78] approaches. Data obtained through sensors mounted to UAVs [79] or ground-based phenotypic platforms with a series of sensors and a GPS [80,81] can be ingested into a data management platform (*acquisition: stream*). To obtain accurate predictions of crop phenotypic information, it is necessary to combine multi-source sensor data such as color, depth, and spectral data with environmental and crop physiology data [30] (*integration*) and develop ML models (*machine learning: model development*). ML models are needed to be deployed (*machine learning: model deployment*) and used for real time variable-rate fertilization. As a result, improvements in the level of fertilizer utilization efficiency enable environmental and economic sustainability benefits by maximizing crop output and minimizing fertilizer input [30].

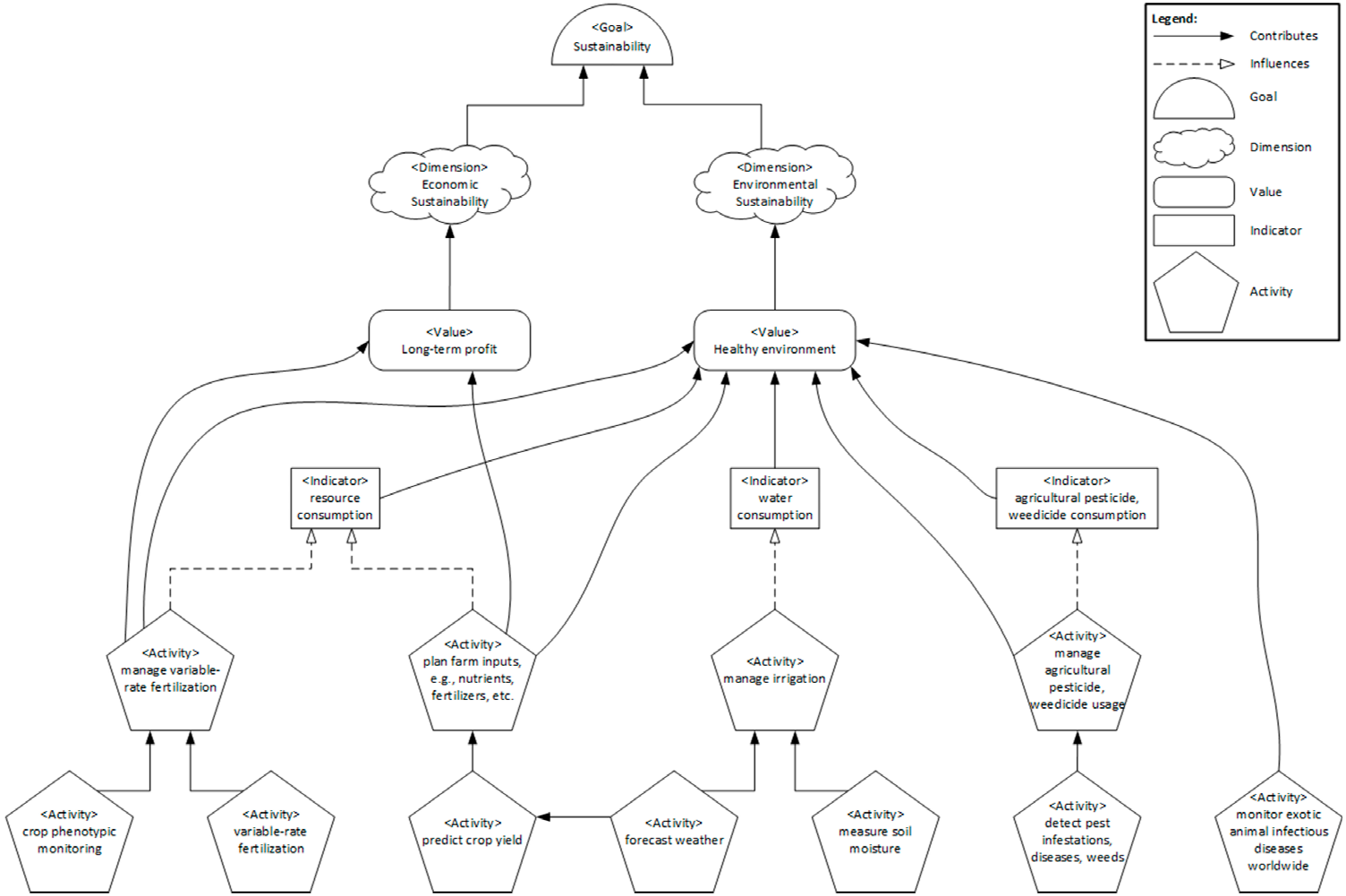

Figure 5. The requirements obtained from the literature on sustainable agriculture. The requirements are mapped to sustainability dimensions using the sustainability model proposed by Penzenstadler & Femmer [64,65].

Exotic animal infectious diseases are considerable threats to global health security and economic stability [26,82]. One of the vital sources for detecting signals of disease is online news platforms. Data can be collected from platforms such as Google News (*acquisition: batch*) and stored for further analysis (*storage*). The named entities such as location, date, disease, hosts, and number of cases can be extracted using Natural Language Processing (NLP) techniques [26] (*information extraction: named entity recognition*). In addition, predefined types of relationships among the recognized entities can be detected (*information extraction: relation extraction*). As an example, the sentence "12 pigs have been infected by African swine fever in Poland" can provide the following entities [26]: number of cases = 12; host = pig; disease = African swine fever; location = Poland.

Table 3 shows the functionalities extracted from the three problem cases (PC) and the four validation cases (VC). Figure 6 presents how these functionalities are mapped to the components of the reference architecture proposed in this study.

Table 3. The functionalities in the problem cases (PC) and validation cases (VC).

Case	#	Functionality
PC1	PC1.1.	obtain and store satellite images
	PC1.2.	process images to derive plant parameters, such as leaf area index (LAI), biomass, and chlorophyll content
	PC1.3.	deduce the current growth status and development of cultivated crops at each location in the field
PC2	PC2.1.	obtain and store harvested crop volume in real time using sensors
	PC2.2.	calculate parameters such as quantity per hectare and flow and build crop productivity maps
PC3	PC3.1.	obtain and store machinery process data; such as speed, angle, pressure, and flow rate through sensors in tractors
	PC3.2.	compute basic figures; such as minimum, maximum, standard deviation and produce documentation of the production process
VC1	VC1.1.	acquire and store historical data on weather, soil, plant population, and planting process
	VC1.2.	remove unusable data and normalize the remaining data
	VC1.3.	combine data on weather, soil, plant population, and planting process
	VC1.4.	build ML models and deploy the one that best satisfies requirements
VC2	VC2.1.	measure and store real time environmental data
	VC2.2.	obtain and store weather forecast
	VC2.3.	combine measurement and forecast data
	VC2.4.	decide on irrigation based on predefined rules or a prediction model
VC3	VC3.1.	obtain and store data, such as crop three-dimensional size, biomass, and vegetation index
	VC3.2.	combine multisource sensor data, such as color, depth, and spectral data, with environmental and crop physiology data
	VC3.3.	build ML models and deploy the one that best satisfies requirements for real time variable-rate fertilization
VC4	VC4.1.	obtain and store news from online news platforms
	VC4.2.	extract named entities, such as location, date, disease, hosts, and number of cases
	VC4.3.	detect predefined types of relationships among recognized entities

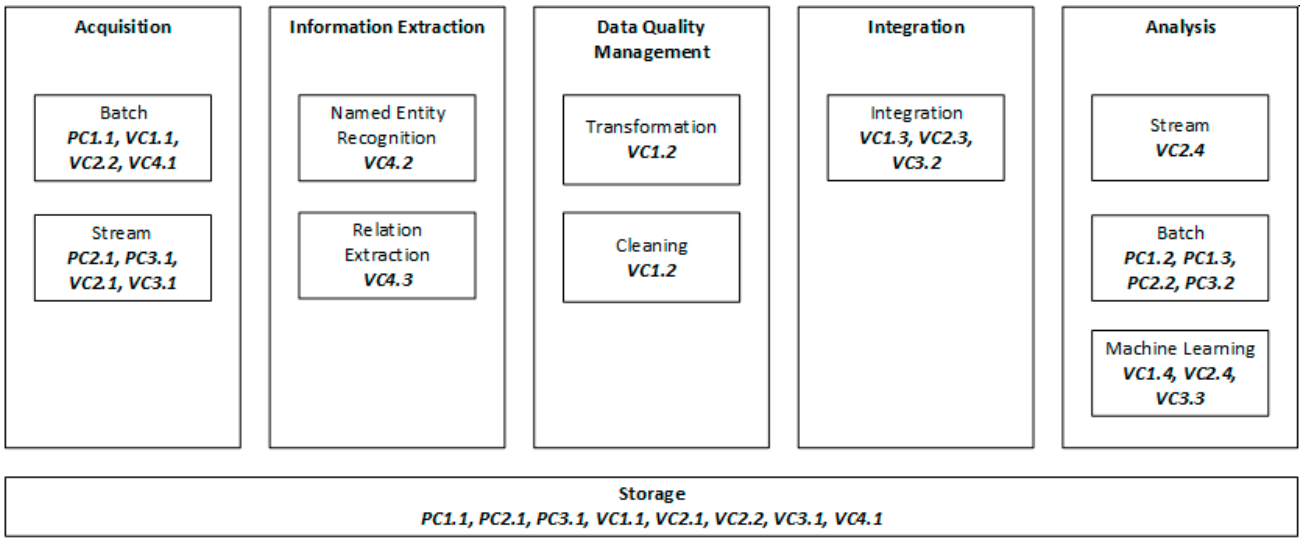

Figure 6. Mapping functionalities to the reference architecture components.

6. Discussion

This research presents a novel data management reference architecture for sustainable agriculture using well-established architecture modeling techniques. As such, this study can pave the way for similar studies on data management reference architectures. The reference architecture was designed based on domain analysis. Other data management application architectures can be developed based on this reference architecture using variant features specified in this study.

The features shown in this study were obtained from peer-reviewed papers in the Scopus database. The inclusion of other databases may reveal unidentified features and the presented reference architecture may be extended accordingly. Since precision agriculture and smart farming are still evolving, new features can be integrated in the future, and data management reference architecture can be adapted. Different tools, techniques, and systems are currently developed by practitioners and researchers in smart farming and therefore, we expect to see new papers in databases that might bring new functionalities and features to the presented reference architecture. However, the presented methodology and the overall reference architecture can be easily changed to reflect the recent changes in smart farming.

The reference architecture is used to derive application architectures based on a multi-case study approach. For the multi-case study, there is a threat of misinterpretation of applied concepts. Although authors discussed the concepts carefully and iteratively, there is a possibility that several concepts may have been interpreted differently compared to the concepts presented in the selected studies. The generalization of the presented findings must be taken with caution because different case studies may require new functionalities. While we can identify many different application scenarios in smart farming, the scenarios share some features from a data management perspective. Therefore, four case studies were shown to demonstrate the applicability of the proposed reference architecture. Other researchers can also evaluate the applicability of this architecture using different case studies in smart farming and create an application architecture for their uses.

It was shown that reference architecture design is useful for the agri-food domain. This study focused on sustainability; however, it can be extended to a larger context by covering other critical aspects of agriculture. For sustainable agriculture, the presented features are beneficial when designing new systems for agricultural production. Further research is needed to evaluate the applicability of the data management reference architecture for different application domains. We expect that increasingly more researchers will focus on sustainability in agriculture in the near future and develop novel models to fully

address sustainability from several aspects. The advancement in machine learning and particularly deep learning techniques can also contribute to the development of novel models addressing sustainability.

7. Conclusions and Future Work

In this study, a data management reference architecture for sustainable agriculture was proposed and evaluated using different case studies. To the best of our knowledge, this is the first study that focuses on sustainability within the context of data management reference architecture. The design science research (DSR) method was applied while designing the reference architecture. For solution design, domain scoping, domain modeling, and reference architecture design stages were followed. Domain scoping was performed based on relevant papers, the domain model was represented as a feature model, and the reference architecture was built at the end of the reference architecture design stage. Three case studies was investigated from different perspectives and the applicability of the data management reference architecture was evaluated. We consider that this research can improve the research in sustainable agriculture with respect to data management and pave the research for designing smart systems for smart farming and precision agriculture.

As future work, we plan to extend this study with new case studies and evaluate the applicability of the presented reference architecture for different scenarios. Another planned study involves the mapping of the reference architecture to the components of the farm management information systems and platforms used in the industry. This mapping can help us identify the possible missing components in the reference architecture. In addition, we can identify enhancement opportunities for the systems and platforms used in the industry.

Author Contributions: Conceptualization, G.G. and C.C.; methodology, G.G.; investigation, G.G.; data curation, G.G.; validation, G.G. and C.C.; writing—original draft preparation, G.G.; writing—review and editing, G.G. and C.C. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Wolfert, S.; Goense, D.; Sørensen, C.A.G. A future internet collaboration platform for safe and healthy food from farm to fork. In Proceedings of the 2014 Annual SRII Global Conference, San Jose, CA, USA, 23–25 April 2014; pp. 266–273.
2. Pierce, F.J.; Nowak, P. Aspects of precision agriculture. *Adv. Agron.* **1999**, *67*, 1–85.
3. Murakami, E.; Saraiva, A.M.; Junior, L.C.R.; Cugnasca, C.E.; Hirakawa, A.R.; Correa, P.L. An infrastructure for the development of distributed service-oriented information systems for precision agriculture. *Comput. Electron. Agric.* **2007**, *58*, 37–48. [CrossRef]
4. Steinberger, G.; Rothmund, M.; Auernhammer, H. Mobile farm equipment as a data source in an agricultural service architecture. *Comput. Electron. Agric.* **2009**, *65*, 238–246. [CrossRef]
5. Sørensen, C.; Bildsøe, P.; Fountas, S.; Pesonen Pedersen, S.; Basso, B.; Nash, E. *Integration of Farm Management Information Systems to Support Real-Time Management Decisions and Compliance of Management Standards*; Center for Research & Technology: Thessaly, Greece, 2009. Available online: http://www.futurefarm.eu (accessed on 14 April 2021).
6. Köksal, Ö.; Tekinerdogan, B. Architecture design approach for IoT-based farm management information systems. *Precis. Agric.* **2019**, *20*, 926–958. [CrossRef]
7. Groeneveld, D.; Tekinerdogan, B.; Garousi, V.; Catal, C. A domain-specific language framework for farm management information systems in precision agriculture. *Precis. Agric.* **2020**, *22*, 1067–1106. [CrossRef]
8. Jin, X.B.; Yu, X.H.; Wang, X.Y.; Bai, Y.T.; Su, T.L.; Kong, J.L. Deep learning predictor for sustainable precision agriculture based on internet of things system. *Sustainability* **2020**, *12*, 1433. [CrossRef]
9. Kaya, A.; Keceli, A.S.; Catal, C.; Yalic, H.Y.; Temucin, H.; Tekinerdogan, B. Analysis of transfer learning for deep neural network based plant classification models. *Comput. Electron. Agric.* **2019**, *158*, 20–29. [CrossRef]
10. Loures, L.; Chamizo, A.; Ferreira, P.; Loures, A.; Castanho, R.; Panagopoulos, T. Assessing the Effectiveness of Precision Agriculture Management Systems in Mediterranean Small Farms. *Sustainability* **2020**, *12*, 3765. [CrossRef]
11. Podlasek, A.; Koda, E.; Vaverková, M.D. The Variability of Nitrogen Forms in Soils Due to Traditional and Precision Agriculture: Case Studies in Poland. *Int. J. Environ. Res. Public Health* **2021**, *18*, 465. [CrossRef] [PubMed]
12. Verdouw, C.; Tekinerdogan, B.; Beulens, A.; Wolfert, S. Digital twins in smart farming. *Agric. Syst.* **2021**, *189*, 103046. [CrossRef]

13. Keceli, A.S.; Catal, C.; Kaya, A.; Tekinerdogan, B. Development of a recurrent neural networks-based calving prediction model using activity and behavioral data. *Comput. Electron. Agric.* **2021**, *170*, 105285. [CrossRef]
14. Catal, C.; Tekinerdogan, B. Aligning education for the life sciences domain to support digitalization and industry 4.0. *Procedia Comput. Sci.* **2019**, *158*, 99–106. [CrossRef]
15. Nadal, S.; Herrero, V.; Romero, O.; Abelló, A.; Franch, X.; Vansummeren, S.; Valerio, D. A software reference architecture for semantic-aware Big Data systems. *Inf. Softw. Technol.* **2017**, *90*, 75–92. [CrossRef]
16. Salma, C.A.; Tekinerdogan, B.; Athanasiadis, I.N. Domain-driven design of big data systems based on a reference architecture. In *Software Architecture for Big Data and the Cloud*; Morgan Kaufmann: Burlington, MA, USA, 2017; pp. 49–68.
17. Tummers, J.; Kassahun, A.; Tekinerdogan, B. Reference architecture design for farm management information systems: A multi-case study approach. *Precis. Agric.* **2020**, *22*, 1–29. [CrossRef]
18. DeLonge, M.S.; Miles, A.; Carlisle, L. Investing in the transition to sustainable agriculture. *Environ. Sci. Policy* **2016**, *55*, 266–273. [CrossRef]
19. Gregor, S.; Hevner, A.R. Positioning and presenting design science research for maximum impact. *MIS Q.* **2013**, *37*, 337–355. [CrossRef]
20. Hevner, A.R.; March, S.T.; Park, J.; Ram, S. Design science in information systems research. *MIS Q.* **2004**, *28*, 75–105. [CrossRef]
21. Wieringa, R.J. *Design Science Methodology for Information Systems and Software Engineering*; Springer: Berlin, Germany, 2014.
22. Runeson, P.; Engström, E.; Storey, M.A. The design science paradigm as a frame for empirical software engineering. In *Contemporary Empirical Methods in Software Engineering*; Springer: Cham, Germany, 2020; pp. 127–147.
23. Köksal, Ö.; Tekinerdogan, B. Feature-driven domain analysis of session layer protocols of internet of things. In Proceedings of the 2017 IEEE International Congress on Internet of Things (ICIOT), Honolulu, HI, USA, 25–30 June 2017; pp. 105–112.
24. van Geest, M.; Tekinerdogan, B.; Catal, C. Design of a reference architecture for developing smart warehouses in industry 4.0. *Comput. Ind.* **2021**, *124*, 103343. [CrossRef]
25. Tekinerdogan, B.; Öztürk, K. Feature-driven design of SaaS architectures. In *Software Engineering Frameworks for the Cloud Computing Paradigm*; Springer: London, UK, 2013; pp. 189–212.
26. Arsevska, E.; Valentin, S.; Rabatel, J.; De Goër de Hervé, J.; Falala, S.; Lancelot, R.; Roche, M. Web monitoring of emerging animal infectious diseases integrated in the French Animal Health Epidemic Intelligence System. *PLoS ONE* **2019**, *13*, e0199960. [CrossRef]
27. Mohsen, S.; Guiping, H.; Huber, I.; Archontoulis, S.V. Coupling machine learning and crop modeling improves crop yield prediction in the US Corn Belt. Scientific Reports (Nature Publisher Group). *arXiv* **2021**, arXiv:2008.04060.
28. Sanjeevi, P.; Prasanna, S.; Siva Kumar, B.; Gunasekaran, G.; Alagiri, I.; Vijay Anand, R. Precision agriculture and farming using Internet of Things based on wireless sensor network. *Trans. Emerg. Telecommun. Technol.* **2020**, *31*, e3978. [CrossRef]
29. Sharma, R.; Kamble, S.S.; Gunasekaran, A.; Kumar, V.; Kumar, A. A systematic literature review on machine learning applications for sustainable agriculture supply chain performance. *Comput. Oper. Res.* **2020**, *119*, 104926. [CrossRef]
30. Shi, Y.; Zhu, Y.; Wang, X.; Sun, X.; Ding, Y.; Cao, W.; Hu, Z. Progress and development on biological information of crop phenotype research applied to real-time variable-rate fertilization. *Plant Methods* **2020**, *16*, 11. [CrossRef]
31. Clevers, J.G.; Kooistra, L.; Van den Brande, M.M. Using Sentinel-2 data for retrieving LAI and leaf and canopy chlorophyll content of a potato crop. *Remote Sens.* **2017**, *9*, 405. [CrossRef]
32. Bach, H.; Migdall, S.; Mauser, W.; Angermair, W.; Sephton, A.J.; Martin-de-Mercado, G. An integrative approach of using satellite-based information for Precision farming: TalkingFields. In Proceedings of the 61st International Astronautical Congress, Prague, Czech Republic, 27 September–1 October 2010.
33. Bach, H.; Mauser, W. Sustainable agriculture and smart farming. In *Earth Observation Open Science and Innovation*; Springer: Cham, Germany, 2018; pp. 261–269.
34. Burlacu, G.; Costa, R.; Sarraipa, J.; Jardim-Golcalves, R.; Popescu, D. A conceptual model of farm management information system for decision support. In *Doctoral Conference on Computing, Electrical and Industrial Systems*; Springer: Berlin/Heidelberg, Germany, 2014; pp. 47–54.
35. Srivastava, S. Space Inputs for Precision Agriculture: Scope for Prototype Experiments in the Diverse Indian Agro-Ecosystems. In Proceedings of the Map Asia 2002, Bangkok, Thailand, 7–9 August 2002; pp. 1–4.
36. Nikkilä, R.; Seilonen, I.; Koskinen, K. Software architecture for farm management information systems in precision agriculture. *Comput. Electron. Agric.* **2010**, *70*, 328–336. [CrossRef]
37. Kaloxylos, A.; Groumas, A.; Sarris, V.; Katsikas, L.; Magdalinos, P.; Antoniou, E.; Politopoulou, Z.; Wolfert, S.; Brewster, C.; Eigenmann, R.; et al. A cloud-based Farm Management System: Architecture and implementation. *Comput. Electron. Agric.* **2014**, *100*, 168–179. [CrossRef]
38. Kruize, J.W.; Wolfert, J.; Scholten, H.; Verdouw, C.N.; Kassahun, A.; Beulens, A.J. A reference architecture for Farm Software Ecosystems. *Comput. Electron. Agric.* **2016**, *125*, 12–28. [CrossRef]
39. McAfee, A. & Brynjolfsson, E. Big data: The management revolution. *Harv. Bus. Rev.* **2012**, *90*, 60–68.
40. Goodfellow, I.; Bengio, Y.; Courville, A. *Deep Learning*; The MIT Press: Cambridge, MA, USA, 2016.
41. Ashmore, R.; Calinescu, R.; Paterson, C. Assuring the Machine Learning Lifecycle: Desiderata, Methods, and Challenges. *arXiv* **2019**, arXiv:1905.04223.
42. Dayal, U.; Akatsu, M.; Gupta, C.; Vennelakanti, R.; Lenardi, M. Expanding global big data solutions with innovative analytics. *Hitachi Rev.* **2014**, *63*, 333–339.
43. Suriarachchi, I.; Plale, B. Crossing analytics systems: A case for integrated provenance in data lakes. In Proceedings of the 2016 IEEE 12th International Conference on e-Science (e-Science), Baltimore, MD, USA, 23–27 October 2016; pp. 349–354.

44. Rao, W.; Jiang, J.; Yang, M.; Peng, W.; Zhou, A. Research on Energy Interconnection Oriented Big Data Sharing Platform Reference Architecture. In *Advanced Computational Methods in Energy, Power, Electric Vehicles, and Their Integration*; Springer: Singapore; pp. 217–225.
45. Sang, G.M.; Xu, L.; De Vrieze, P. Simplifying big data analytics systems with a reference architecture. In *Working Conference on Virtual Enterprises*; Springer: Cham, Germany, 2017; pp. 242–249.
46. Arass, M.E.; Ouazzani-Touhami, K.; Souissi, N. Data Life Cycle: Towards a Reference Architecture. *Int. J.* **2020**, *9*. [CrossRef]
47. Pääkkönen, P.; Pakkala, D. Extending reference architecture of big data systems towards machine learning in edge computing environments. *J. Big Data* **2020**, *7*, 1–29. [CrossRef]
48. Maier, M. Towards a Big Data Reference Architecture. Master's Thesis, University of Eindhoven, Eindhoven, The Netherlands, 2013.
49. Meehan, J.; Aslantas, C.; Zdonik, S.; Tatbul, N.; Du, J. Data Ingestion for the Connected World. In Proceedings of the 8th Biennial Conference on Innovative Data Systems Research (CIDR'17), Chaminade, CA, USA, 8–11 January 2017.
50. Stonebraker, M.; Madden, S.; Dubey, P. Intel "big data" science and technology center vision and execution plan. *ACM SIGMOD Rec.* **2013**, *42*, 44–49. [CrossRef]
51. Adnan, K.; Akbar, R. An analytical study of information extraction from unstructured and multidimensional big data. *J. Big Data* **2019**, *6*, 1–38. [CrossRef]
52. Singh, S. Natural language processing for information extraction. *arXiv* **2018**, arXiv:1807.02383.
53. Gangadharan, V.; Gupta, D. Recognizing Named Entities in Agriculture Documents using LDA based Topic Modelling Techniques. *Procedia Comput. Sci.* **2020**, *171*, 1337–1345. [CrossRef]
54. Oliveira, P.; Rodrigues, F.; Henriques, P.R. *A formal Definition of Data Quality Problems*; ICIQ: Tarragona, Spain, 2005.
55. Ziegler, P.; Dittrich, K.R. Data integration—problems, approaches, and perspectives. In *Conceptual Modelling in Information Systems Engineering*; Springer: Berlin/Heidelberg, Germany, 2005; pp. 39–58.
56. Lenzerini, M. Data integration: A theoretical perspective. In Proceedings of the twenty-first ACM SIGMOD-SIGACT-SIGART symposium on Principles of Database Systems, Madison, WI, USA, 3–5 June 2002; pp. 233–246.
57. Batini, C.; Lenzerini, M.; Navathe, S.B. A comparative analysis of methodologies for database schema integration. *ACM Comput. Surv. (CSUR)* **1986**, *18*, 323–364. [CrossRef]
58. Cugola, G.; Margara, A. Processing flows of information: From data stream to complex event processing. *ACM Comput. Surv. (CSUR)* **2012**, *44*, 1–62. [CrossRef]
59. Carbone, P.; Katsifodimos, A.; Ewen, S.; Markl, V.; Haridi, S.; Tzoumas, K. Apache flink: Stream and batch processing in a single engine. *Bull. IEEE Comput. Soc. Tech. Comm. Data Eng.* **2015**, *36*, 28–38.
60. Begoli, E. A short survey on the state of the art in architectures and platforms for large scale data analysis and knowledge discovery from data. In Proceedings of the WICSA/ECSA 2012 Companion Volume, Helsinki, Finland, 20–24 August 2012; pp. 177–183.
61. Chen, J.; Chen, Y.; Du, X.; Li, C.; Lu, J.; Zhao, S.; Zhou, X. Big data challenge: A data management perspective. *Front. Comput. Sci.* **2013**, *7*, 157–164. [CrossRef]
62. DAMA International. *DAMA-DMBOK: Data Management Body of Knowledge*, 2nd ed.; Technics Publications, LLC: Basking Ridge, NJ, USA, 2017.
63. Kim, C.H.; Weston, R.H.; Hodgson, A.; Lee, K.H. The complementary use of IDEF and UML modelling approaches. *Comput. Ind.* **2003**, *50*, 35–56. [CrossRef]
64. Penzenstadler, B.; Femmer, H. *A Generic Model for Sustainability*; Technical Report; TUM: Munich, Germany, 2012.
65. Penzenstadler, B.; Femmer, H. A generic model for sustainability with process-and product-specific instances. In Proceedings of the 2013 Workshop on Green in/by Software Engineering, Fukuoka, Japan, 26 March 2013; pp. 3–8.
66. van Klompenburg, T.; Kassahun, A.; Catal, C. Crop yield prediction using machine learning: A systematic literature review. *Comput. Electron. Agric.* **2020**, *177*, 105709. [CrossRef]
67. Khaki, S.; Wang, L. Crop yield prediction using deep neural networks. *Front. Plant Sci.* **2019**, *10*, 621. [CrossRef] [PubMed]
68. Natural Resources Conservation Service, United States Department of Agriculture, Web Soil Survey. 2019. Available online: https://data.nal.usda.gov/dataset/natural-resources-conservation-service-web-soil-survey (accessed on 23 April 2021).
69. USDA NASS. *Surveys*; National Agricultural Statistics Service, U.S. Department of Agriculture: District of Columbia, WA, USA, 2019.
70. Wan, Z.; Xia, X.; Lo, D.; Murphy, G.C. How does Machine Learning Change Software Development Practices? *IEEE Trans. Softw. Eng.* **2019**. [CrossRef]
71. Yokoyama, H. Machine learning system architectural pattern for improving operational stability. In Proceedings of the 2019 IEEE International Conference on Software Architecture Companion (ICSA-C), Hamburg, Germany, 25–26 March 2019; pp. 267–274.
72. Glória, A.; Dionisio, C.; Simões, G.; Cardoso, J.; Sebastião, P. Water management for sustainable irrigation systems using internet-of-things. *Sensors* **2020**, *20*, 1402. [CrossRef]
73. Campos, N.G.S.; Rocha, A.R.; Gondim, R.; Coelho da Silva, T.L.; Gomes, D.G. Smart & green: An internet-of-things framework for smart irrigation. *Sensors* **2020**, *20*, 190.
74. Wang, X.C.; Chen, M.; Sun, G.X.; Zhang, Y.; Zhang, Y.N. Design and test of control system on variable fertilizer applicator for winter wheat. *Trans. CSAE* **2015**, *31*, 88–92.
75. Yinyan, S.; Man, C.; Xiaochan, W.; Odhiambo, M.O.; Weimin, D. Numerical simulation of spreading performance and distribution pattern of centrifugal variable-rate fertilizer applicator based on DEM software. *Comput. Electron. Agric.* **2018**, *144*, 249–259. [CrossRef]

76. Boegh, E.; Soegaard, H.; Broge, N.; Hasager, C.B.; Jensen, N.O.; Schelde, K.; Thomsen, A. Airborne multispectral data for quantifying leaf area index, nitrogen concentration, and photosynthetic efficiency in agriculture. *Remote Sens. Environ.* **2002**, *81*, 179–193. [CrossRef]
77. Deery, D.M.; Rebetzke, G.J.; Jimenez-Berni, J.A.; James, R.A.; Condon, A.G.; Bovill, W.D.; Hutchinson, P.; Scarrow, J.; Davy, R.; Furbank, R.T. Methodology for high-throughput field phenotyping of canopy temperature using airborne thermography. *Front. Plant Sci.* **2016**, *7*, 1808. [CrossRef] [PubMed]
78. Guo, Q.; Yang, W.; Wu, F.; Pang, S.; Jin, S.; Chen, F.; Wang, X. High-throughput crop phenotyping: Accelerators for development of breeding and precision agriculture. *Bull. Chin. Acad. Sci.* **2018**, *33*, 940–946.
79. Tian, M.; Ban, S.; Chang, Q.; You, M.; Luo, D.; Wang, L.; Wang, S. Use of hyperspectral images from UAV-based imaging spectroradiometer to estimate cotton leaf area index. *Trans. Chin. Soc. Agric. Eng.* **2016**, *32*, 102–108.
80. Busemeyer, L.; Mentrup, D.; Möller, K.; Wunder, E.; Alheit, K.; Hahn, V.; Maurer, H.P.; Reif, J.C.; Würschum, T.; Müller, J.; et al. BreedVision—A multi-sensor platform for non-destructive field-based phenotyping in plant breeding. *Sensors* **2013**, *13*, 2830–2847. [CrossRef] [PubMed]
81. Sharma, L.K.; Bu, H.; Franzen, D.W.; Denton, A. Use of corn height measured with an acoustic sensor improves yield estimation with ground based active optical sensors. *Comput. Electron. Agric.* **2016**, *124*, 254–262. [CrossRef]
82. Arsevska, E.; Roche, M.; Hendrikx, P.; Chavernac, D.; Falala, S.; Lancelot, R.; Dufour, B. Identification of terms for detecting early signals of emerging infectious disease outbreaks on the web. *Comput. Electron. Agric.* **2016**, *123*, 104–115. [CrossRef]