OPEN ACCESS
International Journal of Cooperative Information Systems
Vol. 29, No. 4 (2020) 2030001 (27 pages)

DOI: 10.1142/S0218843020300016

World Scientific
www.worldscientific.com

A Cross-Domain Comparative Study of Big Data Architectures

Martin Macak*, Mouzhi Ge and Barbora Buhnova
Institute of Computer Science, Masaryk University
Sumavska 15, 602 00 Brno, Czech Republic
Faculty of Informatics, Masaryk University
Botanicka 68a, 602 00 Brno, Czech Republic
**macak@mail.muni.cz*

Received 1 October 2019
Revised 28 April 2020
Accepted 19 August 2020
Published 28 October 2020

Nowadays, a variety of Big Data architectures are emerging to organize the Big Data life cycle. While some of these architectures are proposed for general usage, many of them are proposed in a specific application domain such as smart cities, transportation, healthcare, and agriculture. There is, however, a lack of understanding of how and why Big Data architectures vary in different domains and how the Big Data architecture strategy in one domain may possibly advance other domains. Therefore, this paper surveys and compares the Big Data architectures in different application domains. It also chooses a representative architecture of each researched application domain to indicate which Big Data architecture from a given domain the researchers and practitioners may possibly start from. Next, a pairwise cross-domain comparison among the Big Data architectures is presented to outline the similarities and differences between the domain-specific architectures. Finally, the paper provides a set of practical guidelines for Big Data researchers and practitioners to build and improve Big Data architectures based on the knowledge gathered in this study.

Keywords: Big Data; Big Data architecture; cross-domain comparison; domain-specific architectures; architectural variety.

1. Introduction

Software system architectures are important to organize and understand complex systems when designing and managing system components with long-term rationality by a set of standards, in which its modularity and integration can influence important system characteristics.[1] With the development of Big Data research, a typical Big Data life cycle usually combines the processes of data collection, extraction, cleaning, pre-processing, analysis, visualization, aggregation, as well as

*Corresponding author.

storing.[2] In each process, Big Data tools and technologies such as Apache Hadoop or Apache Spark can be used in the complex ecosystem, and those data-driven technologies are emerging as the backbone infrastructure in different application domains.[3,4] In order to manage the Big Data complexity, different Big Data architectures are proposed to organize the Big Data tools, components, and processes. However, the diversity in Big Data architectures makes it more and more difficult for organizations to build or choose an effective and suitable Big Data architecture.

In order to tackle the architectural diversity and complexity, some general Big Data reference architectures have been proposed over the last decade. For example, Paakkonen and Pakkala[5] have proposed a reference architecture for Big Data systems, which is based on an extensive analysis of published implementation architectures of Big Data use cases. A slightly different approach was taken by Gkalp *et al.*,[6] who performed a systematical review of open-source Big Data tools and building-blocks of those tools (to store, manage and analyze Big Data) and used these findings to design a reference architecture for open-source Big Data analytics. The components of these general architectures quite well resemble each other, although some approaches lay special emphasis on the aspects that they intend to highlight. Nadal *et al.*,[7] for instance, extend these views with a new semantic layer, which consists of semantic annotations of the Big Data architecture, to govern the data life cycle. One of the characteristics that these general architectures have in common is that while they claim that the components of the architecture are optional and depend on the application domain, they do not specify the differences and particularity of domains, nor they give examples that would help to understand how the architecture can be adapted to an application domain.

This all makes it very difficult for Big Data architects to design an architecture for a specific domain. When they want to understand the Big Data architecture best fitting their domain, they are left to gather a collection of the domain-specific Big Data architectures on their own, which is a very time-consuming process. Besides, it is difficult for them to understand the reasons behind the differences among the identified architectures in the domain and to select a representative Big Data architecture that they can follow. This is especially problematic in the application domains that contain a large number of proposed architectures. For example, in the smart cities domain, there are more than 20 Big Data architectures to investigate. Furthermore, as the number of domain-specific architectures still raises,[8] this problem grows more significant as time passes.

In different application domains, the maturity of Big Data architecture varies. Thus, it is valuable to learn from the experience and practice of Big Data architecture in different domains and transfer the knowledge from one domain to another. There is, however, a barrier to learn the Big Data architecture across domains because each domain devotes its own particularity. For example, in agriculture, there is a significant emphasis on supporting multiple user roles in the system, while in the energy management domain, supporting multiple user roles is seldom mentioned. This means that we cannot easily reuse the architecture from the energy

management domain in the agriculture domain because it might not be able to support the desired functionality. On the other hand, the Big Data architectures in different domains share a certain level of commonalities. For example, data processing frameworks such as Apache Spark and NoSQL database are typically used across many domains. It is therefore highly desirable to understand the architectures of each domain and be aware of the domain specifics, similarities, and differences, in order to design and implement an effective Big Data architecture.

In this paper, we survey an extensive collection of Big Data architectures in different domains to understand the state-of-the-art in domain-specific Big Data architectures, as well as their similarities and differences across different domains, such as healthcare, energy management, transportation, smart buildings, smart cities, manufacturing, military, aviation, agriculture, education, and environmental monitoring. Within our analysis, we have, for instance, observed that while some domains rely on service-oriented architectures (e.g. smart cities), others prefer a process-oriented view (e.g. manufacturing). We have also observed specific popularity of modern architecture approaches, such as fog computing, in some domains (e.g. in healthcare). The key contributions of this work are as follows.

- We have surveyed the Big Data architecture papers in different application domains.
- We have selected the most representative Big Data architecture for each domain and identified the typical features in the domain-specific architectures.
- We have conducted a pairwise comparison among the domains and reported the similarities and differences among the Big Data architectures in different domains.
- We have proposed practical guidance on how to build and implement Big Data architectures.

The remainder of the paper is structured as follows. Section 2 defines the scope and methodology of the paper in terms of how the domains and architectural papers were selected. Next, Sec. 3 discusses the Big Data architectures used in each identified domain. Based on the discussion, Sec. 4 presents the cross-domain findings, such as the similarities and differences among domain-specific architectures. Finally, Sec. 5 concludes the paper and outlines future research.

2. Scope and Methodology

Big Data technologies have emerged in different application domains in daily life, with different levels of emphasis on Big Data architecture. To understand the domains where architecture is explicitly studied, we have performed a search of academic databases, including ScienceDirect, Google Scholar, ACM Digital Library, IEEE Xplore Digital Library, Springer, in combination with general Google search, using *Big Data architecture* as the search term. We limited the search to the up-to-date papers over the last six years, from 2013 to 2019. One screening condition is that detailed descriptions of Big Data architecture should exist in the papers. We paid special attention to the survey papers on Big Data research.

On the search result, we have performed an extraction of domain keywords discussed in the titles and abstracts of the papers, which were explicitly *transportation, traffic, farming, agriculture, smart farming, energy management, smart grid, education, adaptive learning, learning, military, war, defense, national security, healthcare, medicine, e-health, biology, chemistry, aviation, aeronautics, environment, environmental monitoring, smart buildings, smart home, smart cities, manufacturing, industry, industrial control system, social media, astronomy, banking, finance, security surveillance, sport, science, tourism, government.* These keywords have been clustered into the following domains used in this paper: *Transportation* (transportation, traffic), *agriculture* (farming, agriculture, smart farming), *energy management* (energy management, smart grid), *education* (education, adaptive learning, learning), *military* (military, war, defense), *healthcare* (healthcare, medicine, e-health), *aviation* (aviation, aeronautics), *environmental monitoring* (environment, environmental monitoring), *smart buildings* (smart buildings, smart home), *smart cities* (smart cities), *manufacturing* (manufacturing, industrial control system), *banking* (banking, finance), and then also *biology*, *sport*, *tourism*, *e-science*, *security surveillance*, *network monitoring*, *social media*, and *astronomy.*

Note that there might be architectures relevant to multiple Big Data domains, e.g. there can be an aviation Big Data architecture, which is at the same time relevant to the military, or an emergency transportation Big Data architecture relevant to the healthcare domain. Each work has therefore been associated with its primary domain to which it contributes more strongly (which can more generally benefit from it) and been considered part of this domain during the analysis.

Overall, a total of 70 papers have been identified and assigned to 20 domains. In nine Big Data domains, however, we have found only one proposed architecture. Those are biology, sport, tourism, e-science, security surveillance, network monitoring, banking, social media, and astronomy. As these could add noise to the results of the study, our analysis keeps its focus on the domains with more than one identified architecture, therefore covering 61 papers in 11 domains. The papers and their related domains are listed in Table 1 (the architectures in domains that have one architecture only are listed as others).

Table 1. List of the found architectures.

Domain	Architecture papers	Representative
Healthcare	Refs. 9–15	Ref. 13
Energy management	Refs. 16–20	Ref. 18
Transportation	Refs. 21–24	Ref. 21
Smart buildings	Refs. 25–29	Ref. 27
Smart cities	Refs. 30–52	Ref. 30
Manufacturing	Refs. 53–56	Ref. 53
Military	Refs. 57–59	Ref. 58
Aviation	Refs. 60–62	Ref. 62
Agriculture	Refs. 63–65	Ref. 65
Education	Refs. 66 and 67	Ref. 66
Environmental monitoring	Refs. 68 and 69	Ref. 68
Others	Refs. 70–78	—

To facilitate the discussion, we have selected a representative Big Data architecture in each domain based on three criteria: (1) the majority of architectures from this domain can be mapped to it, (2) it illustrates the majority of obtained specific properties, and (3) it is concise and easily readable. If none or multiple architectures satisfied the first condition, we applied the second condition. If after this application also several architectures remained, we used the third condition and picked the most concise and readable architecture as a representative.

3. Big Data Architectures Across Domains

Many studies have proposed architectures for Big Data systems, in which it can be general or domain-specific. As far as we know, the most widely accepted representative of a general Big Data architecture is presented in Fig. 1, which can map most of the existing Big Data architectures. This architecture is, therefore, valuable for us to derive the domain-specific representatives.

This general architecture consists of nine components. It contains functionalities (rectangles), data stores (ellipses), and data flows (arrows). The data processing process is shown as a pipeline, in which data go mostly from left to right. *Data sources* component refers to, for example, streaming sources of data. These data can be structured, unstructured, and semi-structured. *Data extraction* component moves data into the system. They may be temporarily stored or transferred and

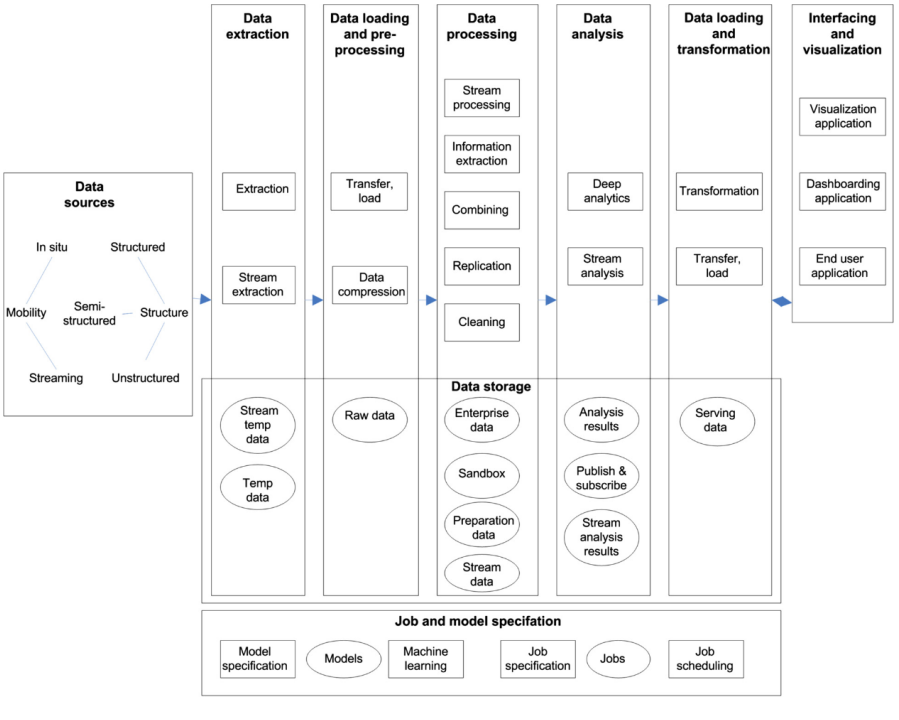

Fig. 1. Example of general Big Data architecture.[5]

loaded into a raw data store in *Data loading and pre-processing* component. From this raw data store, data can be processed by the *Data processing* component. For example, they can be cleaned, replicated, or some information can be extracted from them. Then this data can be stored for further analysis for *Data analysis* component. Analyzed data may be transformed in *Data loading and transformation* into a data serving store that can be used in the *Interfacing and visualization* component. The *Data storage* component is used across multiple components, such as data extracting and data processing. Furthermore, jobs, models, or algorithms can be specified in the *Job and model specification* component. In Ref. 5, authors mapped this architecture to several real use cases, like data analytics infrastructures in Facebook, LinkedIn, Twitter, Netflix, and a few others.

3.1. *Healthcare*

Big Data analysis helps the healthcare domain by improving treatment methods, improving the detection of diseases, and monitoring the hospital quality. In this domain, many heterogeneous data can be found, such as clinical data and logs of health monitoring devices.

Security and data protection are critical, which have been explicitly stated in almost all the identified architectures. Another crucial property is latency. That is why some architectures use fog computing in combination with cloud computing to reduce the volume of data that needs to be moved to the cloud.[11,13] On the other hand, some architectures use lambda architecture to ensure low latency.[12] The typical data sources in this domain are several types of sensors, multimedia, and social network. From the tool perspective, usually, Hadoop, with a combination of a NoSQL database, is used.

Seven Big Data architectures have been collected in this domain. In Fig. 2, we present the chosen representative, which multiple architectures in this domain resemble, and to which they can be mapped. It contains a fog layer, which appears to be a typical aspect of the healthcare domain. It is a three-layered architecture that consists of (1) a device layer, containing sensors providing health data; (2) fog layer, which helps with the latency problem; and (3) cloud layer that integrates data and provides the interface for users.

3.2. *Energy management*

In the energy management domain, all the found Big Data architectures are designed for smart grids to improve their efficiency. The data that are used in the analysis are from sensors containing information about energy production, energy consumption, and other properties like quality and reliability.[18,19] Also, other data can be included in the analysis, for example, weather data.[18] In this domain, cloud computing is a common component in multiple architectures. Spark is usually used because it helps with the providing of low latency. In addition, Hadoop and NoSQL databases are also common in this domain.

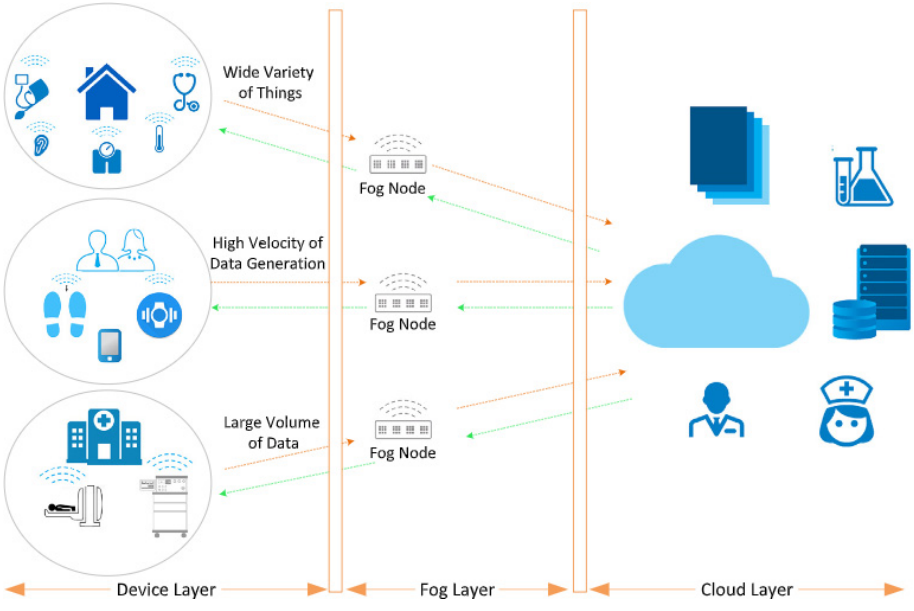

Fig. 2. Representative of the healthcare domain.[13]

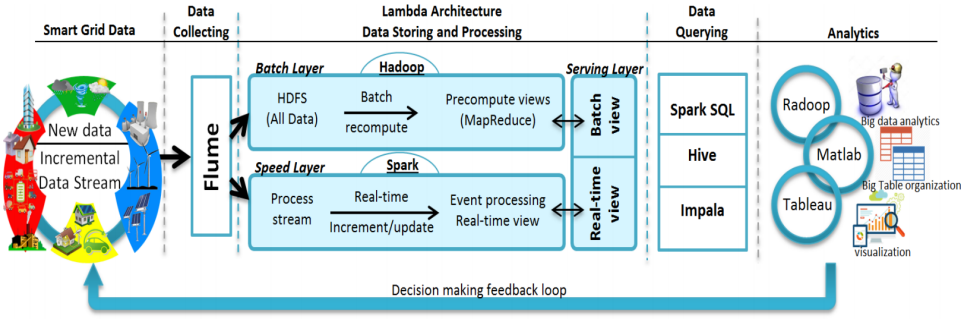

Fig. 3. Representative of the energy management domain.[18]

We have identified five Big Data architectures in the energy management domain. Figure 3 depicts a typical example, whose layered style can also be seen in other architectures, and hence the majority of other architectures from this domain can be mapped to it. Although others are not explicit about the usage of the lambda architecture, the structure follows the same pattern: Collect the data, store and analyze them, extract useful information, and visualize it.

3.3. *Transportation*

Big Data analysis in the transportation domain is mostly about the interpretation of data that is generated by the vehicles. However, in some use cases, the data

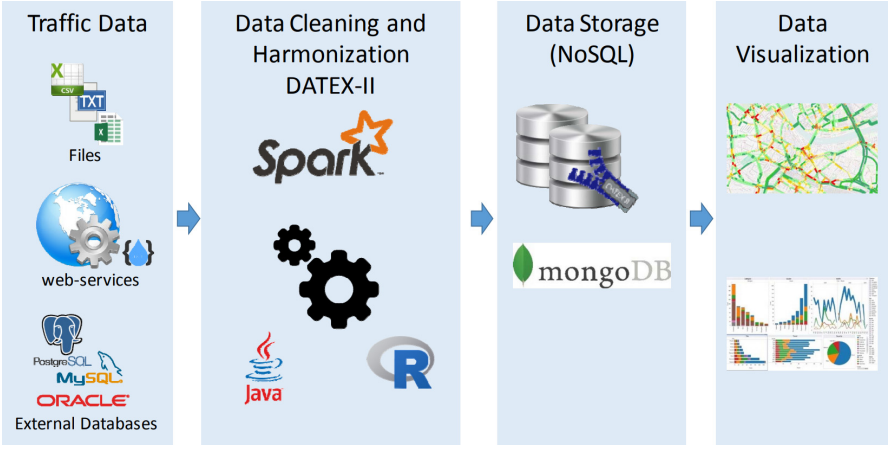

Fig. 4. Representative of the transportation domain.[21]

from mobile phones and social media are included in the data analysis. We have found architectures that consider use cases like traffic prediction that can result in dynamic toll charging for highways,[21] or improving road safety and navigation,[23] accident detection, travel time prediction, bus arrival time prediction,[24] or real-time traffic control.[22]

Big Data in the transportation domain is usually generated from various sources, so data variety is a major issue here. The data analysis uses a statistical approach, but also data mining and machine learning.

Four Big Data architectures have been found in this domain. Each appears to be different because each follows a different use case. The chosen representative architecture of this domain is presented in Fig. 4. It is the architecture for traffic prediction via dynamic toll charging. We chose it because it is the most concise and readable among all four architectures.

3.4. *Smart buildings*

In the smart buildings domain, real-time sensor data from the buildings are usually analyzed. We have found architectures for improving energy efficiency in buildings, like detection of anomalies[28] or energy monitoring,[26] and architectures for smart building maintenance.[25,27] Actually, the system maintenance was the most dominant in this domain. The architectures usually use cloud in their design.

We have found four Big Data architectures. Figure 5 presents a representative architecture. We consider that this architecture clearly illustrates the wide variety of data sources for smart buildings, mostly sensor data, and thus can be considered as a typical architecture in this domain.

3.5. *Smart cities*

The smart cities domain is widely used and may be included in several other domains mentioned before. For example, smart buildings can be a specific part

Fig. 5. Representative of the smart buildings domain.[27]

of smart cities. Therefore, similar to the smart building domain, the architectures in smart cities typically use sensors as data sources. After data extraction from the sources, the architectures focus on data processing that brings specific value for the smart city and improves the quality of life in the city. The main issue in this domain donates in the infrastructure that can handle, for example, the desired workflows, availability, scalability, and security. Because of these issues, those architectures are usually service-oriented, and the systems are hosted in a cloud.

Smart cities, being a multi-domain federation, consists of a large number of Big Data architectures. Specifically, we have found 23 architectures in this domain. In Fig. 6, we present one of them that best describes the general usage of these architectures. Several other smart cities architectures can be mapped to it. It also shows multiple typical properties in these architectures, which is, for example, the usage of sensors, cloud computing, and service orientation. Specifically, service-oriented architectures were mostly used in this domain.

3.6. *Manufacturing*

In this domain, the Big Data architectures are used for cleaner production,[53] planning,[55] managing linked systems,[54] and business process analysis.[56] As a shared characteristic, they rely on sensors as the common data source in this domain. In some cases, the data also comes from specialized internal systems. The architectures mostly need to deal with Big Data volume and variety issues, using mostly Hadoop and NoSQL databases. The prevalent use case in this domain is business process

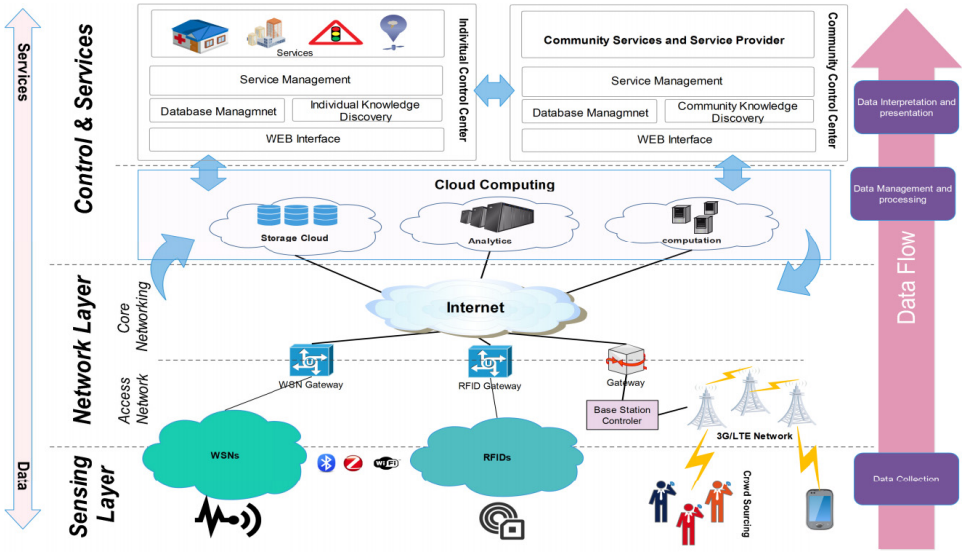

Fig. 6. Representative of the smart cities domain.[30]

analysis, i.e. the analysis aimed at the increase of the efficiency and effectiveness of processes.

We discovered four Big Data architectures in this domain. Figure 7 presents the chosen four-layered architecture, which is a typical example of an architecture in this domain because the majority of architectures can be mapped to it.

3.7. *Military*

Use cases considered in the Big Data architectures for the military domain are usually simulations, crowd-sourcing, and the analysis of sensor data from multiple sources, like surveillance cameras, radars, human intelligence, and web crawlers.[57,58] The challenge in the tactical aspect of this domain is heterogeneous data sources with incomplete and noisy data and a strong emphasis on security.[58] Also, the networking bandwidth is limited, while there are real-time requirements. On the other hand, in the simulation part of this domain, there is no problem with the data quality because the data for the analysis can be carefully chosen.[57] However, there is a big stress on the trustworthiness of the output data, so the models that the simulations are based on should not have any flaws. For the applications that have critical time constraints, the system also needs to process the data as fast as possible.

We have identified three Big Data architectures in this domain. As the found use cases are completely different, the chosen architecture, which is in Fig. 8, was picked because it contains the majority of specific properties, and it is easily readable.

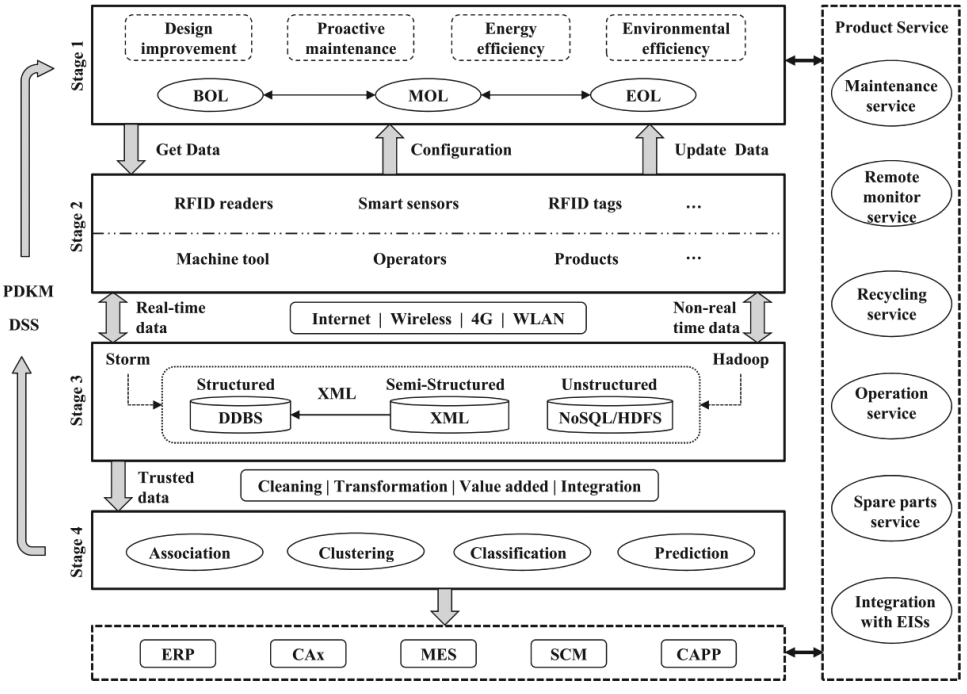

Fig. 7. Representative of the manufacturing domain.[53]

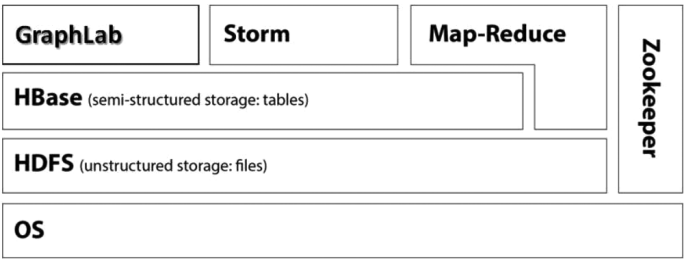

Fig. 8. Representative of the military domain.[58]

3.8. *Aviation*

In this domain, the data are heterogeneous and featured by, for instance, the aircraft, airline and airport data, market information, flight tracking data, passenger information, weather conditions, and air safety reports.[62] The data is typically used for predictions,[60] aircraft tracking,[61] and safety measures.[62] The typical challenge in this domain is to dealing with scalability and multiple data formats. Safety is the property that was mentioned most in this domain.

Three Big Data architectures are found to be available. Because of the particularity of this domain, we infer that there might be more Big Data architectures existing but not published. Figure 9 depicts an exemplary Big Data architecture

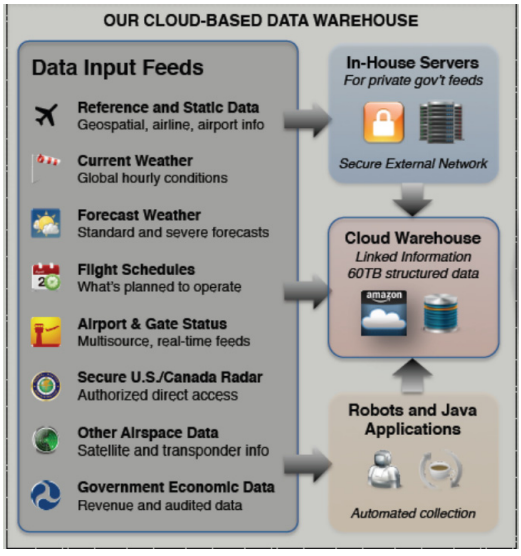

Fig. 9. Representative of the aviation domain.[62]

for aviation, which has been chosen for its simplicity. It shows the architecture of the cloud-based data warehouse that can be queried.

3.9. *Agriculture*

In the agriculture domain, the main goal of Big Data analysis is to improve farming productivity by recommendation, prediction, or problem detection.[63–65] Thus, the information about the plants, soil, water, fertilizers, pesticides, and weather is essential. The data are collected from sensors, weather stations, satellites, and public bodies.[64] The architectures consider a recommendation for multiple types of users to improve their management.

Big Data in the agriculture domain has to deal with all typical data challenges, like volume, velocity, and variety. In the case of IoT devices, it also needs to handle the poor-quality data. For example, the solar radiation data collected shortly after rain should not be used in assessing crop performance.[65] The analysis in this domain is performed in a batch, stream, as well as real-time processing mode. The main focus of the architectures is the scalability, extensibility, visualization, and handling of multiple types of users. In addition, we found that there is the highest number of papers in agriculture that consider recommendation as a use case.

We have found three Big Data architectures in this domain. It is possible because the availability of Big Data in this domain is quite a new phenomenon. In Fig. 10, we present one of them, which, through the SmartFarmNet gateway, manages the communication connected IoT devices. It contains most of the specific properties in this domain. The collected data are stored, processed, and provided to users through different APIs.

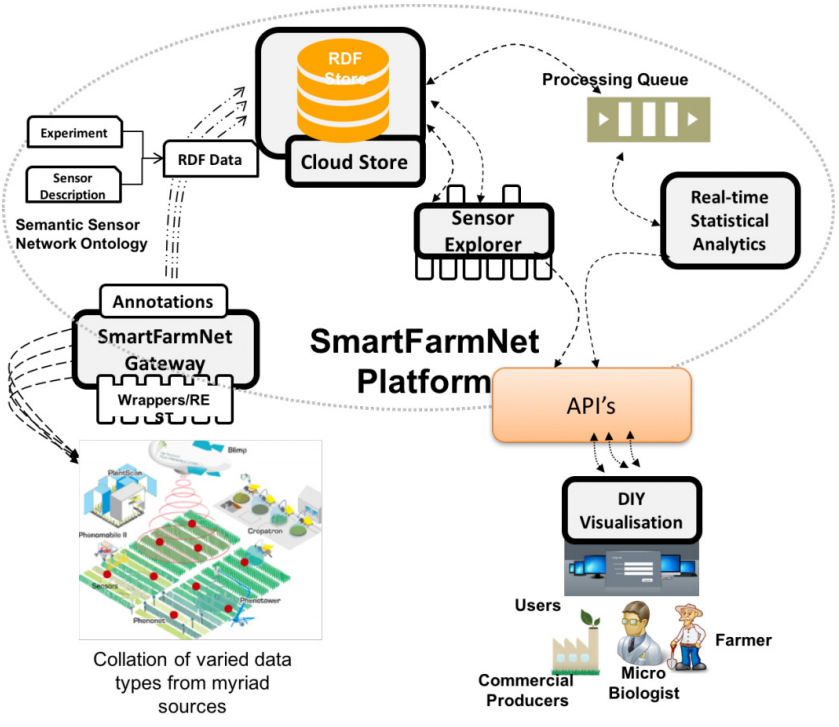

Fig. 10. Representative of the agriculture domain.[65]

3.10. *Education*

Big Data analysis in the education domain can be used for educational statistics, prediction of course enrollments, or identification of the students' status.[67] The data can be collected from information systems, documents, logs from university servers, public portals, or social networks.[66] A typical characteristic of this data is its great variety. Typically, data mining are performed on them.

We have identified only two Big Data architectures in this domain, which might be caused by the fact that the data analyzed in universities is typically not very big. Hence, the Big Data architecture does not need to be studied explicitly. One of the identified architectures is depicted in Fig. 11 to illustrate an example of this domain.

3.11. *Environmental monitoring*

In this domain, data can come from many different sources, for example, sensors, geographical information systems, or global positioning systems. They are then used to model and manage environmental processes.[68] Fazio *et al.*[69] are, for instance, focusing only on the storage of Big Data, combining the document and object storage, while Fang *et al.*[68] use a more typical four-layer architecture. The analysis

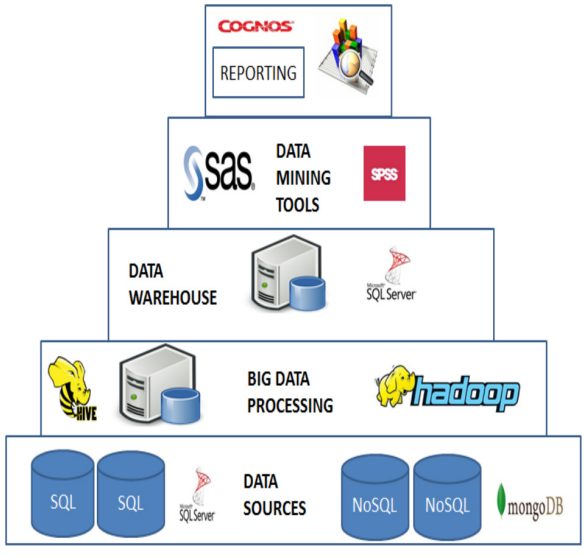

Fig. 11. Representative of the education domain.[66]

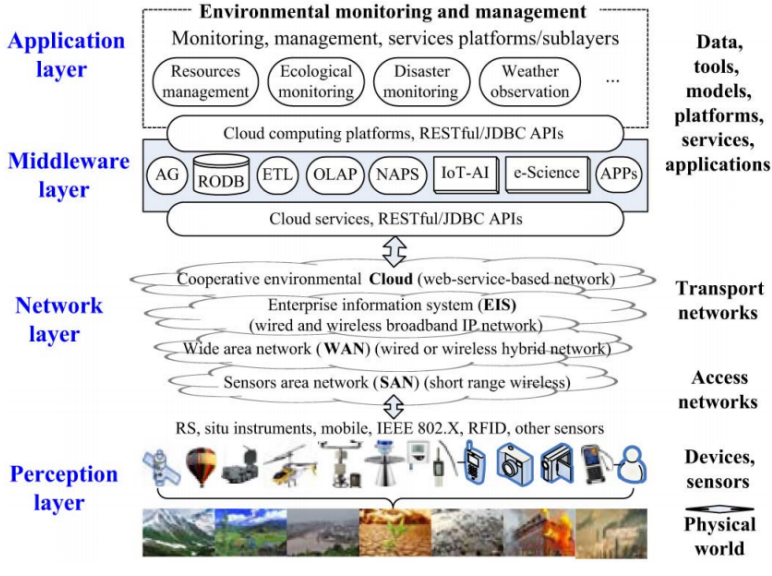

Fig. 12. Representative of the environmental monitoring domain.[68]

is done either on historical data or real-time data. In both cases, cloud technology is used, which helps to ensure the scalability and reliability of these systems. We have found two Big Data architectures in this domain. In Fig. 12, we present a more concise one.

4. Cross-Domain Findings

We have found numerous implications in terms of similarities as well as differences among the application domains and architectures designed for supporting their Big Data analysis. This section discusses the key findings from our study, also presented in Table 2, which shows the common characteristics of the pairwise comparison of the domains.

To construct Table 2, we have first extracted Big Data architecture properties from the analysis in Sec. 3. Specifically, we have linked the properties discussed within Sec. 3 to all the 61 individual architectures, which resulted in Table A.1 in Appendix A. The data from this table was then used to construct the list of specific properties for each domain by taking the properties that were present at least in 50% of the architectures in that domain. When the domain contained only two architectures, we took only the properties that were included in both of them. We also found the properties that were specific only for one domain. Those domain-specific properties were included in the table if their presence distinguishes the domain from others, i.e. the percentage of the included property is higher than in other domains.

This typical properties were then used to construct Table 2, which illustrates the properties of the pairwise comparison of the domains. The information on the diagonal corresponds to the properties that are considered typical for that particular domain. If we did not identify any specific properties for the domain or any specific common properties for the pair of domains, the corresponding cell is left empty.

4.1. *Similarities of Big Data architectures across domains*

This section summarizes the similarities across the Big Data architectures in the application domains, as outlined in Table 2. Moreover, some properties are similar across all the domains. They have not been included in the table for readability. These properties are *data variety*, *batch and real-time processing*, *architectural scalability*, and *data visualization*.

Big Data Sources. The usual data source in most of the domains is a sensor. Sensors are typically used for system monitoring and anomaly detection. In the smart building domain, they can, for example, detect a fire in the building.[29] Another popular data source is data from social networks. For example, in healthcare, social-network data are used to track the mental health of some patients[14] or to predict health trends in regions.[12] Also, in the transportation domain, social network data are used for traffic predictions.[21] Besides these types of data, multimedia are used in healthcare, transportation, military, agriculture, and environmental monitoring for the image, video, or audio analysis. Another common data source is GPS, which is used in transportation, agriculture, and environmental monitoring for analyzing the position of a particular object or person.

Table 2. Cross-domain findings.

	Healthcare	Energy management	Transportation	Smart buildings	Smart cities	Manufacturing	Military	Aviation	Agriculture	Education	Environmental monitoring
Healthcare	**fog computing**										
Energy management	activity monitoring, cloud, data mining, event prediction, Hadoop, latency, NoSQL database, sensors, useful info detection										
Transportation	data mining, event prediction, Hadoop, latency, machine learning, multimedia, NoSQL database, sensors, social network, useful info detection	data mining, Hadoop, latency, NoSQL database, sensors, Spark, statistical analysis, useful info detection									
Smart buildings	activity monitoring, cloud, security, sensors, useful info detection	activity monitoring, cloud, sensors, useful info detection	sensors, useful info detection	**maintenance**							
Smart cities	activity monitoring, cloud, event prediction, security, sensors, social network	activity monitoring, cloud, event prediction, sensors	event prediction, sensors, social network	activity monitoring, cloud, security, sensors	**service orientation**						
Manufacturing	activity monitoring, data mining, event prediction, Hadoop, NoSQL database, sensors, useful info detection	activity monitoring, data mining, Hadoop, NoSQL database, sensors, statistical analysis, useful info detection	data mining, Hadoop, NoSQL database, sensors, statistical analysis, useful info detection	activity monitoring, sensors	activity monitoring, event prediction, sensors	**business process analysis**					
Military	cloud, data mining, event prediction, Hadoop, machine learning, multimedia, NoSQL database, security, sensors, social network, useful info detection	cloud, data mining, Hadoop, NoSQL database, sensors, statistical analysis, useful info detection	data mining, data quality issue, event prediction, Hadoop, machine learning, multimedia, NoSQL database, sensors, social network, statistical analysis, useful info detection	cloud, security, sensors, useful info detection	cloud, event prediction, security, sensors, social network	data mining, Hadoop, sensors, statistical analysis, useful info detection	**simulations**				
Aviation	activity monitoring, data mining, event prediction, security	activity monitoring, data mining, statistical analysis	data mining, event prediction, statistical analysis	activity monitoring, security	activity monitoring, event prediction, security	activity monitoring, data mining, statistical analysis	data mining, event prediction, security, statistical analysis, tracking	**safety**			
Agriculture	activity monitoring, cloud, event prediction, multimedia, NoSQL database, sensors, useful info detection	activity monitoring, cloud, NoSQL database, sensors, statistical analysis, useful info detection	data quality issue, event prediction, GPS, multimedia, NoSQL database, sensors, statistical analysis, useful info detection	activity monitoring, cloud, sensors, useful info detection	activity monitoring, cloud, event prediction, sensors	activity monitoring, NoSQL database, sensors, statistical analysis, useful info detection	cloud, data quality issue, event prediction, multimedia, NoSQL database, sensors, statistical analysis, useful info detection	activity monitoring, event prediction	**multiple types of users, recommendation**		
Education	data mining, event prediction, NoSQL database, social network	data mining, NoSQL database, statistical analysis	data mining, event prediction, NoSQL database, social network, statistical analysis		event prediction, social network	data mining, NoSQL database, statistical analysis	data mining, event prediction, NoSQL database, social network, statistical analysis	data mining, event prediction, statistical analysis	event prediction, NoSQL database, statistical analysis		
Environmental monitoring	activity monitoring, cloud, multimedia, sensors	activity monitoring, cloud, sensors	GPS, multimedia, sensors	activity monitoring, cloud, sensors	activity monitoring, cloud, sensors	activity monitoring, sensors	cloud, multimedia, sensors	activity monitoring	activity monitoring, cloud, GPS, multimedia, sensors		

Big Data Technologies. From the Big Data tool perspective, the majority of Big Data domains uses Apache Hadoop for distributed computing and as the distributed file system. Another popular computational framework, specifically in the transportation and energy management domain, is Apache Spark. Furthermore, many domains indicate the usage of NoSQL databases, such as Apache HBase, MongoDB, or Redis, as the most typical examples.

Big Data Analysis. Data mining is the most common type of analysis in multiple domains, including healthcare, energy management, transportation, manufacturing, military, aviation, and education. Its techniques are used for prediction, detection, and recommendation. Many domains also use statistical analysis. For instance, in the energy management domain, statistical analysis is used to understand the consumption and usage of energy,[20] and in the education domain, statistical analysis is used to understand data about schools.[67] On the other hand, machine learning was not discussed very often (only in healthcare, transportation, and military), which might be caused by the fact that our study only focuses on papers with explicit Big Data architecture, while machine learning approaches rather focus on the algorithmic, than architecture, side of the solution.

Big Data Applications. One of the most popular scenarios discussed together with the Big Data architectures is the monitoring (with the exception of transportation and military domains, where the monitoring was not discussed explicitly). For example, the growth of crops is monitored in the agriculture domain,[64] or the moving vehicles are monitored in the transportation domain.[22] On the other hand, the military domain, instead of monitoring, focuses on tracking specified targets.[58] Similarly, tracking is also relevant in the aviation domain. However, military and aviation are the only domains that mention tracking in connection with Big Data architectures. Event prediction is another common application, which can be found in most of the domains, with the exception of energy management, smart buildings, manufacturing, and environmental monitoring that do not discuss it explicitly in connection to Big Data architectures. For example, in the education domain, the number of students who will enroll in courses can be predicted.[66] Further, in the transportation domain, the traffic[21] and travel time[24] is the focus of the prediction. The third most common use case is the information detection. For example, in healthcare, frauds,[10] diseases,[9] and anomalies in the behavior of patients[15] are detected. Moreover, in the manufacturing domain, deviations from the process models are detected, so that prompt reaction to problems is possible, and the efficiency of the production can be improved.[56]

Big Data Challenges and Architectural Features. Some of the domains highlight the specific focus of the architectures, such as the latency of computing, which is emphasized in healthcare, energy management, and transportation. Moreover, security is receiving specific attention in the Big Data architectures in healthcare, smart buildings, smart cities, military, and aviation. There is also an emphasis on

data quality, specifically in agriculture, due to relying on erroneous sensors and military due to the criticality of the impact of the analysis. In addition, data quality is considered in the architectures in the military and transportation domain. For several architectures, the typical similarity is the utilization of cloud technologies. For example, in smart cities, the cloud is used for the storage and analysis of data from several data sources.[30]

4.2. *Differences of Big Data architectures across domains*

Specific properties. We have found several properties that are unique for a specific domain, i.e. the percentage of the property is higher than in other domains. For instance, in healthcare architectures, several approaches are using fog computing to reduce the volume of data that needs to be sent to the cloud. In other domains, we did not see such a big emphasis on this technique. We have only observed it in some of the smart cities architectures.

In the smart buildings domain, the maintenance of the components connected to the system is a very typical use case. The architecture moreover, needs to manage the communication between the components well. We expect that this use case is highly relevant also in the manufacturing domain, although we have not identified any Big Data architecture that would be explicit about this.

In smart cities architectures, service orientation is very common. As the system needs to operate many services, it is natural to use a service-oriented architecture.

The exclusive type of analysis in the manufacturing domain is business process analysis. Systems analyze the processes that are executed in production. Their aim is to increase the effectiveness of the process by, for example, discovering unusual behavior or bottlenecks in them.

In the military, the simulation use case had a slightly higher percentage of usage, comparing to other domains. This use case was also mentioned in several smart cities papers.

In the aviation domain, the most commonly emphasized property, also reflected by Big Data architectures, is safety. Surprisingly, safety is not much mentioned in other domains, although we expect it to be important also elsewhere than in aviation.

In agriculture, the architectures typically pay special attention to reflecting multiple types of users, such as farmers, data scientists, companies, public administrators, and domain experts, each using the system in a very different way. That is why this property is emphasized in the architectures. In addition, the recommendation use case has a higher usage here than in the other domains. No other domain pays much attention to the user roles in the system.

In energy management, transportation, education, and environmental monitoring, we did not find any unique property, specific only for one domain.

4.3. *Implications for building and improving Big Data architectures*

By reviewing the similarity and difference of domain-specific Big Data architectures, we have derived the following implications and guidelines on how to build, design, and implement a Big Data architecture.

Smart cities and healthcare are the most indicative domains for studying Big Data architectures. Both of the domains feature a high number of architectures, indicating that the Big Data architectures in smart cities and healthcare are well developed, and they may be more mature than the architectures in other domains. We have summarized the majority of the Big Data architectures in the two domains to provide a foundation for practitioners to build or adopt Big Data architectures.

Due to the large variety of application domains, Big Data architectures are not yet proposed in a number of domains. For instance, in agriculture, education, or environment, only a few proposed Big Data architectures can be found. Further, in some domains like biology, astronomy, marketing, or sport, we have found only one or no architecture. Some possible reasons can be that in those domains, researchers and practitioners are not sharing their architectures, one of Big Data architectures has been widely accepted in that domain, or those domains need significant effort in creating a Big Data architectures. On the other hand, it can also be considered as a research gap in those domains. That is, Big Data architectures can be proposed or adopted in domains such as agriculture, education, or environment.

Table 2 can be used as a knowledge base for building a domain-specific Big Data architecture. Similarities of Big Data architectures in different domains have been derived. Thus, a Big Data architect can look into the architectural features from similar domains and reuse features in their own context. We have found many similarities, for example, in healthcare, smart cities, manufacturing, and agriculture.

The knowledge of Big Data architecture in one domain can be brought to other domains. As several domains contain specific property that was not emphasized in other architectures, we can see that, for example, when implementing a fog computing architecture in the transportation domain, one can borrow the experience from healthcare architectures and see how they manage to do it. Furthermore, some domains can be grouped and relatively isolated from other domains. That means the Big Data architectures are scoped in only several domains. Thus, an Big Data architect does not need to go through all the architectures in every domain rather focus on certain domains. For example, the manufacturing and education domains can be considered as a cohort.

There is a popular toolset in Big Data architectures. The majority of architectures consider Apache Hadoop for data storage and processing. Also, NoSQL databases, such as Apache HBase, MongoDB, or Redis, are very commonly specified for data storage. As a data processing alternative for Hadoop, several papers contain Apache Spark in their architectures.[18,20,21,23] Therefore, we can infer that these tools are mature enough to be utilized in the newly created Big Data architectures.

As these Big Data tools and software have been used in different application domains, those aforementioned tools are considered to be not domain-specific and can be used independently from the domains.

There is no "one fits all" Big Data architecture for every domain. When designing a Big Data solution, in most cases, it is better to look for a domain-specific architecture rather than to look for a general architecture, like the one we present in Fig. 1. Although this general architecture is well designed, it does not emphasize domain-specific properties, like the discussed service-orientation, safety, or need for multiple types of users, which should be understood when designing a Big Data solution for a specific domain.

4.4. *Exemplary indications for building a new Big Data architecture*

This section demonstrates an exemplary scenario, in which we intend to design a new Big Data architecture for a specific data-intensive system. Consider a hospital system for predicting medical diagnosis. For the purpose of this illustration, we can set the functional requirements aside and move forward to the point when the analysis of the non-functional requirements of the system indicates that the Big Data architecture of our medical system shall integrate Hadoop and NoSQL as well as apply data mining. For the software architect, this is still a very open assignment where guidance in terms of an existing Big Data architecture template is missing and the guideline is needed to prevent architectural flaws that would later be very hard to correct. To obtain such guidance, the following steps can be taken based on our results.

First, the match to the required properties can be looked up in Table A.1, where we see three architectures in the healthcare domain that match the required properties. As these matching architectures differ in their remaining properties, they guide the architect to consider multiple clarifying questions and decide among the three options. The decision depends on the system design priority. For example, if the new system needs to ensure a rigorous patient privacy, we can decide to follow the architecture in Ref. 10. If we prefer the importance of latency in our system, we can look for the inspiration in Ref. 13 or Ref. 12. Further, if we plan to integrate Fog computing within our solution, we can follow Ref. 13 to design the Big Data architecture.

Furthermore, the architect might benefit from the inspiration from other domains. In Table A.1, we can see that papers in the energy management and transportation domain also share the same required properties. Thus, the two architectural alternatives[18,20] from the energy management domain can be considered to in designing the new Big Data architecture. Besides, there are also good indications in the transportation domain, which are described in Refs. 21 and 23. Similarly, we can filter these architectures based on clarifying questions matching the differences in the properties of the architectures. For example, if we need high-level security, we

can follow the architecture in Ref. 20. If the system needs to integrate multimedia analysis, we can refer to Ref. 18 or Ref. 21.

5. Conclusions

In this paper, we have conducted a comprehensive survey on Big Data architectures in a total of 11 application domains across from traditional domains such as agriculture and education, to emerging domains such as smart building and smart cities. In each of the application domains, we have searched and reviewed a set of Big Data architectures that are proposed in it and selected one representative Big Data architecture to describe how a typical architecture is used in that domain.

Based on the survey of Big Data architectures in all the surveyed domains, we have derived the similarity of the Big Data architectures from different domains, where the common features across different domain-specific Big Data architectures and pairwise comparisons have been presented. Further, we have also found differences and distinctions among the Big Data architectures due to their specific domain property. The similarity and difference of Big Data architectures enable us to interlink the domains and indicate how to complement and improve different Big Data architectures as well as to understand why certain component in a Big Data architecture is critical to a specific domain. Based on the results and analyses, we have proposed a set of practical guidance on how to build and improve Big Data architectures.

As future work, we plan to refine the granularity of the survey further and investigate how to configure a Big Data architecture. Also, since we believe that due to domain features, there is no "one fits all" architecture for every domain, detailed guidance on how to build or improve Big Data architectures in each domain can be constructed. Therefore, in the future, we also plan to derive guidelines or best practice of Big Data architecture for every application domain.

Acknowledgment

The work was supported from by European Regional Development Fund Project *CERIT Scientific Cloud* (No. CZ.02.1.01/0.0/0.0/16_013/0001802).

Table A.1. Properties of the reviewed architectures.

		sensors	social networks	multimedia	GPS	Hadoop	Spark	NoSQL database	data mining	statistical analysis	machine learning	activity monitoring	tracking	event prediction	useful info detection	recommendation	business process analysis	simulations	maintenance	security	latency	safety	multiple types of users	data quality issue	cloud	fog computing	service orientation
Healthcare	[4]	✓		✓		✓		✓			✓	✓		✓	✓					✓							
	[70]	✓	✓	✓		✓		✓	✓	✓	✓	✓		✓	✓	✓				✓		✓		✓	✓		
	[40]	✓				✓		✓				✓		✓						✓	✓				✓	✓	
	[65]	✓	✓	✓		✓		✓	✓			✓		✓	✓						✓						
	[18]	✓		✓		✓		✓	✓	✓	✓	✓		✓	✓					✓	✓				✓	✓	
	[77]	✓	✓	✓					✓	✓	✓	✓		✓	✓	✓		✓		✓				✓	✓		
	[64]	✓	✓									✓			✓					✓	✓				✓	✓	✓
Energy management	[43]	✓				✓		✓				✓		✓											✓		
	[27]								✓	✓		✓		✓	✓												
	[47]	✓		✓		✓	✓	✓	✓	✓		✓		✓	✓						✓				✓		
	[14]	✓				✓	✓	✓			✓	✓		✓	✓				✓	✓	✓				✓		✓
	[39]	✓				✓	✓	✓	✓	✓	✓			✓	✓	✓				✓	✓			✓	✓		
Transportation	[26]	✓	✓	✓	✓	✓	✓	✓	✓		✓			✓							✓			✓			
	[3]	✓	✓	✓	✓	✓	✓	✓		✓	✓			✓	✓	✓				✓	✓	✓					
	[55]	✓	✓			✓	✓	✓	✓	✓	✓			✓	✓						✓			✓			
	[75]		✓		✓	✓		✓	✓		✓			✓	✓										✓		
Smart buildings	[68]	✓						✓			✓	✓			✓				✓	✓	✓						
	[2]	✓							✓			✓		✓					✓	✓	✓				✓		
	[56]	✓										✓							✓	✓					✓		
	[54]	✓							✓	✓		✓		✓	✓									✓			
	[29]	✓		✓		✓				✓		✓			✓	✓			✓			✓		✓	✓		
Smart cities	[30]	✓	✓	✓					✓	✓		✓			✓					✓		✓	✓		✓		✓
	[66]	✓			✓						✓	✓		✓	✓					✓	✓		✓		✓	✓	✓
	[23]	✓				✓	✓	✓		✓		✓	✓							✓		✓					✓
	[12]	✓	✓			✓	✓	✓	✓			✓				✓			✓	✓	✓						✓
	[8]	✓	✓			✓	✓	✓	✓	✓		✓			✓				✓	✓	✓				✓		
	[33]	✓	✓			✓	✓		✓	✓	✓	✓		✓				✓		✓			✓		✓		✓
	[57]	✓	✓	✓	✓	✓	✓			✓		✓	✓	✓	✓					✓			✓				✓
	[74]	✓	✓		✓				✓	✓		✓	✓	✓	✓	✓			✓	✓		✓	✓		✓		✓
	[51]	✓	✓		✓	✓						✓	✓	✓				✓		✓							✓
	[50]	✓	✓									✓						✓		✓	✓				✓	✓	✓
	[6]	✓	✓		✓							✓		✓					✓	✓	✓	✓	✓				✓
	[60]	✓	✓		✓	✓	✓	✓	✓	✓	✓	✓		✓	✓				✓	✓		✓	✓	✓	✓		✓
	[16]								✓			✓			✓			✓		✓	✓	✓			✓			✓
	[10]	✓	✓								✓	✓						✓	✓	✓			✓		✓	✓	✓
	[71]	✓	✓			✓						✓		✓						✓		✓			✓		✓
	[49]	✓	✓		✓				✓	✓		✓			✓					✓					✓		✓
	[28]	✓				✓		✓	✓		✓	✓		✓				✓				✓			✓		✓
	[31]	✓										✓								✓					✓		✓
	[9]	✓	✓		✓	✓				✓		✓	✓	✓	✓					✓		✓			✓		✓
	[1]	✓	✓								✓	✓						✓	✓	✓	✓	✓	✓		✓	✓	✓
	[11]	✓	✓		✓							✓	✓	✓								✓					✓
	[52]	✓	✓			✓	✓	✓	✓	✓	✓	✓	✓	✓						✓	✓				✓	✓	✓
	[67]	✓							✓	✓		✓		✓	✓								✓				
Manufacturing	[76]	✓				✓		✓	✓	✓		✓	✓	✓			✓		✓								
	[38]	✓										✓		✓													
	[35]	✓		✓					✓	✓		✓		✓	✓	✓	✓										
	[73]	✓				✓		✓	✓						✓		✓										
Military	[63]	✓	✓					✓	✓	✓	✓		✓	✓	✓	✓			✓				✓		✓		✓
	[61]	✓	✓	✓		✓		✓	✓	✓			✓	✓	✓					✓				✓	✓		
	[34]	✓	✓	✓		✓	✓	✓			✓	✓	✓	✓	✓					✓	✓			✓			
Aviation	[5]								✓	✓			✓	✓					✓								
	[7]				✓	✓		✓				✓	✓							✓		✓					
	[37]								✓	✓		✓	✓	✓						✓		✓			✓		
Agriculture	[62]	✓	✓	✓	✓		✓	✓		✓				✓	✓	✓							✓	✓	✓		
	[36]	✓		✓	✓	✓		✓			✓	✓		✓	✓								✓		✓		✓
	[32]	✓		✓				✓		✓		✓		✓		✓			✓	✓	✓			✓	✓		
Education	[44]		✓	✓		✓		✓	✓	✓				✓									✓				
	[42]	✓	✓					✓	✓	✓				✓	✓												
Environmental monitoring	[17]	✓		✓	✓							✓		✓								✓			✓		
	[19]	✓		✓	✓			✓				✓													✓		

References

1. E. Crawley *et al.*, The influence of architecture, in *Engineering Systems Monograph* (MIT, 2004).
2. M. Ge, H. Bangui and B. Buhnova, Big data for internet of things: A survey, *Future Gener. Comput. Syst.* **87** (2018) 601–614.
3. B. Rossi *et al.*, Anomaly detection in smart grid data: An experience report, *2016 IEEE Int. Conf. Systems, Man, and Cybernetics (SMC)* (IEEE, 2016), pp. 002313–002318.
4. D. Gesvindr and J. Michalkova and B. Buhnova, System for collection and processing of smart home sensor data, in *IEEE Int. Conf. Software Architecture Workshops* (*ICSAW*) (Gothenburg, 2017), pp. 247–250, doi=10.1109/ICSAW.2017.23.
5. P. Paakkonen and D. Pakkala, Reference architecture and classification of technologies, products and services for big data systems, *Big Data Res.* **2**(4) (2015) 166–186, doi:https://doi.org/10.1016/j.bdr.2015.01.001.
6. M. O. Gokalp *et al.*, Big-data analytics architecture for businesses: A comprehensive review on new open-source big-data tools (Cambridge Service Alliance: Cambridge, UK, 2017).
7. S. Nadal *et al.*, A software reference architecture for semantic-aware big data systems, *Inf. Softw. Technol.* **90** (2017) 75–92, doi:https://doi.org/10.1016/j.infsof.2017.06.001.
8. M. Goudarzi, Heterogeneous architectures for big data batch processing in mapreduce paradigm, *IEEE Trans. Big Data* **5**(1) (2019) 18–33.
9. J. Archenaa and E. M. Anita, A survey of big data analytics in healthcare and government, *Proc. Comput. Sci.* **50** (2015) 408–413, doi:https://doi.org/10.1016/j.procs.2015.04.021.
10. Y. Wang, L. Kung and T. A. Byrd, Big data analytics: Understanding its capabilities and potential benefits for healthcare organizations, *Technol. Forecast. Soc. Change* **126** (2018) 3–13, doi:https://doi.org/10.1016/j.techfore.2015.12.019.
11. G. Manogaran *et al.*, A new architecture of internet of things and big data ecosystem for secured smart healthcare monitoring and alerting system, *Future Gener. Comput. Syst.* **82** (2018) 375–387, doi:https://doi.org/10.1016/j.future.2017.10.045.
12. V.-D. Ta, C.-M. Liu and G. W. Nkabinde, Big data stream computing in healthcare real-time analytics, *2016 IEEE Int. Conf. Cloud Computing and Big Data Analysis (ICCCBDA)* (Chengdu, Chin, 2016), pp. 37–42. doi:10.1109/ICCCBDA.2016.7529531.
13. B. Farahani *et al.*, Towards fog-driven iot ehealth: Promises and challenges of iot in medicine and healthcare, *Future Gener. Comput. Syst.* **78** (2018) 659–676, doi:https://doi.org/10.1016/j.future.2017.04.036.
14. Y. Zhang *et al.*, Health-cps: Healthcare cyber-physical system assisted by cloud and big data, *IEEE Syst. J.* **11**(1) (2017) 88–95, doi:10.1109/JSYST.2015.2460747.
15. G. Suciu *et al.*, Big data, internet of things and cloud convergence – an architecture for secure e-health applications, *J. Med. Syst.* **39**(11) (2015) 141, doi:10.1007/s10916-015-0327-y.
16. M. Mayilvaganan and M. Sabitha, A cloud-based architecture for big-data analytics in smart grid: A proposal, *2013 IEEE Int. Conf. on Computational Intelligence and Computing Research* (Enathi, India, 2013), pp. 1–4. doi:10.1109/ICCIC.2013.6724168.
17. X. He *et al.*, A big data architecture design for smart grids based on random matrix theory, *IEEE Trans. Smart Grid* **8**(2) (2017) 674–686, doi:10.1109/TSG.2015.2445828.
18. A. A. Munshi and Y. A. I. Mohamed, Data lake lambda architecture for smart grids big data analytics, *IEEE Access* **6** (2018) 40463–40471, doi:10.1109/ACCESS.2018.2858256.

19. H. Daki *et al.*, Big data management in smart grid: concepts, requirements and implementation, *J. Big Data* **4** (2017) 13, doi:10.1186/s40537-017-0070-y.
20. X. Liu and P. S. Nielsen, Streamlining smart meter data analytics, in *Proc. 10th Conf. Sustainable Development of Energy, Water and Environment Systems* (International Centre for Sustainable Development of Energy, Water and Environment Systems, 2015).
21. G. Guerreiro *et al.*, An architecture for big data processing on intelligent transportation systems. an application scenario on highway traffic flows, *2016 IEEE 8th Int. Conf. Intelligent Systems (IS)* (Sofia, Bulgaria, 2016), pp. 65–72, doi:10.1109/IS.2016.7737393.
22. S. Amini, I. Gerostathopoulos and C. Prehofer, Big data analytics architecture for real-time traffic control, *2017 5th IEEE Int. Conf. Models and Technologies for Intelligent Transportation Systems, Naples Ital* (2017), pp. 710–715, doi:10.1109/MTITS.2017.8005605.
23. Y. Petalas *et al.*, A big data architecture for traffic forecasting using multi-source information, *Int. Workshop of Algorithm Aspects of cloud Computing Conf.* (Cham, 2017), pp. 65–83, doi:10.1007/978-3-319-57045-7_5.
24. J. Yu, F. Jiang and T. Zhu, Rtic-c: A big data system for massive traffic information mining, *2013 Int. Conf. Cloud Computing and Big Data* (Fuzhou, China, 2013), pp. 395–402, doi:10.1109/CLOUDCOM-ASIA.2013.91.
25. M. Villari *et al.*, Alljoyn lambda: An architecture for the management of smart environments in iot, *2014 Int. Conf. Smart Computing Workshops* (Hong Kong, China, 2014), pp. 9–14, doi:10.1109/SMARTCOMP-W.2014.7046676.
26. A. R. Al-Ali *et al.*, A smart home energy management system using iot and big data analytics approach, *IEEE Trans. Consum. Electron.* **63**(4) (2017) 426–434, doi:org/10.1109/TCE.2017.015014.
27. A. P. Plageras *et al.*, Efficient iot-based sensor big data collection-processing and analysis in smart buildings, *Future Gener. Comput. Syst.* **82** (2018) 349–357, doi:https://doi.org/10.1016/j.future.2017.09.082.
28. M. Pena *et al.*, Rule-based system to detect energy efficiency anomalies in smart buildings, a data mining approach, *Expert Syst. Appli.* **56** (2016) 242–255, doi:https://doi.org/10.1016/j.eswa.2016.03.002.
29. M. S. Hossain, M. A. Rahman and G. Muhammad, Cyber-physical cloud-oriented multi-sensory smart home framework for elderly people: An energy efficiency perspective, *J. Parallel Distrib. Comput.* **103** (2017) 11–21, special Issue on Scalable Cyber-Physical Systems, doi:https://doi.org/10.1016/j.jpdc.2016.10.005.
30. R. Jalali, K. El-khatib and C. McGregor, Smart city architecture for community level services through the internet of things, in *18th Int. Conf. Intelligence in Next Generation Networks* (Pairs, France, 2015), pp. 108–113, doi:10.1109/ICIN.2015.7073815.
31. B. Tang *et al.*, A hierarchical distributed fog computing architecture for big data analysis in smart cities, in *Proc. ASE BigData & SocialInformatics 2015, ASE BD&SI '15*, (ACM, Kaohsiung, Taiwan, 2015), pp. 28:1–28:6 doi:10.1145/2818869.2818898.
32. M. Gohar *et al.*, A big data analytics architecture for the internet of small things, *IEEE Commun. Mag.* **56**(2) (2018) 128–133, doi:10.1109/MCOM.2018.1700273.
33. C. Costa and M. Y. Santos, Basis: A big data architecture for smart cities, *2016 SAI Computing Conf. (SAI)* (London, UIC, 2016), pp. 1247–1256, doi:10.1109/SAI.2016.7556139.
34. B. Cheng *et al.*, Building a big data platform for smart cities: Experience and lessons from santander, *2015 IEEE Int. Congress on Big Data* (New York, USA, 2015), pp. 592–599, doi:10.1109/BigDataCongress.2015.91.

35. Z. Khan *et al.*, Towards cloud based big data analytics for smart future cities, *J. Cloud Comput.* **4**(1) (2015) 2, doi:10.1186/s13677-015-0026-8.
36. M. M. Rathore *et al.*, Urban planning and building smart cities based on the internet of things using big data analytics, *Comput. Networks* **101** (2016) 63–80, industrial Technologies and Applications for the Internet of Things, doi:https://doi.org/10.1016/j.comnet.2015.12.023.
37. C. Yin *et al.*, A literature survey on smart cities, *Sci. China Inf. Sci.* **58**(10) (2015) 1–18, doi:10.1007/s11432-015-5397-4.
38. B. Nathali Silva, M. Khan and K. Han, Big data analytics embedded smart city architecture for performance enhancement through real-time data processing and decision-making, *Wireless Commun. Mobile Comput.* **2017** (2017) 1–12, doi:10.1155/2017/9429676.
39. P. G. V. Naranjo *et al.*, Fo can: A fog-supported smart city network architecture for management of applications in the internet of everything environments, *J. Parallel Distribut. Comput.* **132** (2018) 274–283, doi:https://doi.org/10.1016/j.jpdc.2018.07.003.
40. N. Bawany and J. Shamsi, Smart city architecture: Vision and challenges, *Int. J. Adv. Comput. Sci. Appl.* **6** (2015) 246–255, doi:10.14569/IJACSA.2015.061132.
41. E. F. Z. Santana *et al.*, Software platforms for smart cities: Concepts, requirements, challenges, and a unified reference architecture, *ACM Comput. Surv.* **50**(6) (2017) 78:1–78:37. doi:10.1145/3124391.
42. A. Enayet *et al.*, A mobility-aware optimal resource allocation architecture for big data task execution on mobile cloud in smart cities, *IEEE Commun. Magazine* **56**(2) (2018) 110–117, doi:10.1109/MCOM.2018.1700293.
43. S. J. Clement, D. W. McKee and J. Xu, Service-oriented reference architecture for smart cities, in *Proc. 2017 IEEE Novel Big Data Architecture in Support of Symp. Service-Oriented System Engineering (SOSE)* (IEEE South San Francisco, California, 2017), pp. 81–85, doi:10.1109/SOSE.2017.29.
44. Z. Xiong, Y. Zheng and C. Li, Data vitalization's perspective towards smart city: A reference model for data service oriented architecture, *2014 14th IEEE/ACM Int. Symp. Cluster, Cloud and Grid Computing* (Chicago, UK, USA, 2014), pp. 865–874, doi:10.1109/CCGrid.2014.74.
45. S. V. Nandury and B. A. Begum, Smart wsn-based ubiquitous architecture for smart cities, *2015 Int. Conf. Advances in Computing, Communications and Informatics (ICACCI)* (Kochi, 2015), pp. 2366–2373, doi:10.1109/ICACCI.2015.7275972.
46. X. He *et al.*, Qoe-driven big data architecture for smart city, *IEEE Commun. Mag.* **56**(2) (2018) 88–93, doi:10.1109/MCOM.2018.1700231.
47. M. S. Jamil *et al.*, Smart environment monitoring system by employing wireless sensor networks on vehicles for pollution free smart cities, *Proc. Eng.* **107** (2015) 480–484, https://doi.org/10.1016/j.proeng.2015.06.106.
48. C. Chilipirea *et al.*, An integrated architecture for future studies in data processing for smart cities, *Microprocess. Microsyst.* **52** (2017) 335–342, doi:https://doi.org/10.1016/j.micpro.2017.03.004.
49. D. P. Abreu *et al.*, A resilient internet of things architecture for smart cities, *Ann. Telecommun.* **72**(1) (2017) 19–30, doi:org/10.1007/s12243-016-0530-y.
50. J. Conway-Beaulieu, A. Athaide, R. Jalali and K. El-Khatib, Smartphone-based architecture for smart cities, in *Proc. 5th ACM Symp. Development and Analysis of Intelligent Vehicular Networks and Applications*, (ACM, Cancun Maxico, 2015), pp. 79–83, doi:10.1145/2815347.2826698.

51. A. M. S. Osman, A novel big data analytics framework for smart cities, *Future Genera. Comput. Syst.* **91** (2019) 620–633, doi:https://doi.org/10.1016/j.future.2018.06.046.
52. C. Tao *et al.*, Architecture for monitoring urban infrastructure and analysis method for a smart-safe city, *2014 Sixth Int. Conf. Measuring Technology and Mechatronics Automation* (Zhangjiajie, China, 2014), pp. 151–154, doi:10.1109/ICMTMA.2014.40.
53. Y. Zhang *et al.*, A big data analytics architecture for cleaner manufacturing and maintenance processes of complex products, *J. Clean. Prod.* **142** (2017) 626–641, doi:https://doi.org/10.1016/j.jclepro.2016.07.123.
54. J. Lee, B. Bagheri and H.-A. Kao, A cyber-physical systems architecture for industry 4.0-based manufacturing systems, *Manufact. Lett.* **3** (2015) 18–23, doi:https://doi.org/10.1016/j.mfglet.2014.12.001.
55. J. Krumeich *et al.*, Big data analytics for predictive manufacturing control — a case study from process industry, *2014 IEEE Int. Congress on Big Data* (Anchorage, AK, 2014), pp. 530–537, doi:10.1109/BigData.Congress.2014.83.
56. H. Yang *et al.*, A system architecture for manufacturing process analysis based on big data and process mining techniques, *2014 IEEE Int. Conf. Big Data (Big Data)* (Washington, DC, USA, 2014), pp. 1024–1029, doi:10.1109/BigData.2014.7004336.
57. X. Song *et al.*, Military simulation big data: Background, state of the art, and challenges, *Math. Probl. Eng.* **2015** (2015) 298356, doi:10.1155/2015/298356.
58. O. Savas *et al.*, Tactical big data analytics: Challenges, use cases, and solutions, *SIGMETRICS Perform. Eval. Rev.* **41**(4) (2014) 86–89, doi:10.1145/2627534.2627561.
59. J. Klein *et al.*, A reference architecture for big data systems in the national security domain, in *Proc. 2nd Int. Workshop on BIG Data Software Engineering, BIGDSE '16* (ACM, Austin Texas, 2016), pp. 51–57, doi:10.1145/2896825.2896834.
60. S. Ayhan *et al.*, Predictive analytics with aviation big data, in *2013 Integrated Communications, Navigation and Surveillance Conf. (ICNS)* (Herndon, UA, USA, 2013), pp. 1–13, doi:10.1109/ICNSurv.2013.6548556.
61. E. Boci and S. Thistlethwaite, A novel big data architecture in support of ads-b data analytic, in *Proc. 2015 Integrated Communication, Navigation and Surveillance Conf. (ICNS)* (IEEE, 2015), pp. C1–1.
62. T. Larsen, Cross-platform aviation analytics using big-data methods, *2013 Integrated Communications, Navigation and Surveillance Conf. (ICNS)* (Herndon, VA, USA, 2013), pp. 1–9, doi:10.1109/ICNSurv.2013.6548579.
63. P. Shah, D. Hiremath and S. Chaudhary, Big data analytics architecture for agro advisory system, *2016 IEEE 23rd Int. Conf. High Performance Computing Workshops (HiPCW)* (Hyderabad, 2016), pp. 43–49, doi:10.1109/HiPCW.2016.015.
64. S. Lamrhari *et al.*, A profile-based big data architecture for agricultural context, *2016 Int. Conf. Electrical and Information Technologies (ICEIT)* (Tangiers, 2016), pp. 22–27, doi:10.1109/EITech.2016.7519585.
65. P. P. Jayaraman *et al.*, Internet of things platform for smart farming: Experiences and lessons learnt, *Sensors* **16**(11) (2016) 1–17, doi:10.3390/s16111884.
66. P. Michalik, J. Stofa and I. Zolotova, Concept definition for big data architecture in the education system, *2014 IEEE 12th Int. Symp. Appl. Mach. Int. Inf. (SAMI)* (Herl'pany, Slorakia, 2014), pp. 331–334, doi:10.1109/SAMI.2014.6822433.
67. F. Matsebula and E. Mnkandla, A big data architecture for learning analytics in higher education, *2017 IEEE AFRICON* (Cape Town, 2017), pp. 951–956, doi:10.1109/AFRCON.2017.8095610.
68. S. Fang *et al.*, An integrated system for regional environmental monitoring and management based on internet of things, *IEEE Trans. Ind. Inf.* **10**(2) (2014) 1596–1605, doi:10.1109/TII.2014.2302638.

69. M. Fazio *et al.*, Big data storage in the cloud for smart environment monitoring, *Proc. Comput. Sci.* **52** (2015) 500–506, *the 6th Int. Conf. Ambient Systems, Networks and Technologies (ANT-2015)*, pp. 500–506, doi:https://doi.org/10.1016/j.procs.2015.05.023.
70. C. Wang *et al.*, Heterogeneous cloud framework for big data genome sequencing, *IEEE/ACM Trans. Comput. Biol. Bioinformat.* **12**(1) (2015) 166–178, doi:10.1109/TCBB.2014.2351800.
71. R. Rein and D. Memmert, Big data and tactical analysis in elite soccer: future challenges and opportunities for sports science, *SpringerPlus* **5**(1) (2016) 1410.
72. M. Fuchs, W. Hopken and M. Lexhagen, Big data analytics for knowledge generation in tourism destinations - a case from sweden, *J. Dest. Mark. Manage.* **3**(4) (2014) 198–209, doi:https://doi.org/10.1016/j.jdmm.2014.08.002.
73. Y. Demchenko *et al.*, Addressing big data issues in scientific data infrastructure, in *2013 Int. Conf. on Collaboration Technologies and Systems (CTS)* (San, Diego, CA, USA), 2013, pp. 48–55, doi:10.1109/CTS.2013.6567203.
74. Z. Xu *et al.*, The big data analytics and applications of the surveillance system using video structured description technology, *Cluster Comput.* **19** (2016) 1283–1292, doi:10.1007/s10586-016-0581-x.
75. S. Marchal *et al.*, A big data architecture for large scale security monitoring, *2014 IEEE Int. Congress on Big Data* (Anchorage, Ak, USA, 2014), pp. 56–63, doi:10.1109/BigData.Congress.2014.18.
76. A. Munar, E. Chiner and I. Sales, A big data financial information management architecture for global banking, *2014 Int. Conf. Future Internet of Things and Cloud* (Barcelona, Span, 2014), pp. 385–388, doi:10.1109/FiCloud.2014.68.
77. G. Miranda and T. Delgado, Big data architecture for social media sentiment analysis supporting context aware recommendation systems, *5th International Workshop on Knowledge Discovery, Knowledge Management and Decision Support* (Mexico City, Mexico, 2015).
78. Z. Zhang *et al.*, Scientific computing meets big data technology: An astronomy use case, *2015 IEEE Int. Conf. Big Data (Big Data)* (SantaClara, CA, USA, 2015), pp. 918–927, doi:10.1109/BigData.2015.7363840.