future internet

MDPI

Article

Main Issues in Big Data Security

Julio Moreno *, Manuel A. Serrano and Eduardo Fernández-Medina

Alarcos Research Group, University of Castilla-La Mancha, 13005 Ciudad Real, Spain; manuel.serrano@uclm.es (M.A.S.); eduardo.fdezmedina@uclm.es (E.F.-M.)

* Correspondence: julio.moreno@uclm.es; Tel.: +34-926-29-53-00

Academic Editor: Dino Giuli

Received: 24 June 2016; Accepted: 29 August 2016; Published: 1 September 2016

Abstract: Data is currently one of the most important assets for companies in every field. The continuous growth in the importance and volume of data has created a new problem: it cannot be handled by traditional analysis techniques. This problem was, therefore, solved through the creation of a new paradigm: Big Data. However, Big Data originated new issues related not only to the volume or the variety of the data, but also to data security and privacy. In order to obtain a full perspective of the problem, we decided to carry out an investigation with the objective of highlighting the main issues regarding Big Data security, and also the solutions proposed by the scientific community to solve them. In this paper, we explain the results obtained after applying a systematic mapping study to security in the Big Data ecosystem. It is almost impossible to carry out detailed research into the entire topic of security, and the outcome of this research is, therefore, a big picture of the main problems related to security in a Big Data system, along with the principal solutions to them proposed by the research community.

Keywords: Big Data; security; systematic mapping study

1. Introduction

Over the last few years, data has become one of the most important assets for companies in almost every field. Not only are they important for companies related to the computer science industry, but also for organisations, such as countries' governments, healthcare, education, or the engineering sector. Data are essential with respect to carrying out their daily activities, and also helping the businesses' management to achieve their goals and make the best decisions on the basis of the information extracted from them [1]. It is estimated that of all the data in recorded human history, 90 percent has been created in the last few years. In 2003, five exabytes of data were created by humans, and this amount of information is, at present, created within two days [2].

This tendency towards increasing the volume and detail of the data that is collected by companies will not change in the near future, as the rise of social networks, multimedia, and the Internet of Things (IoT) is producing an overwhelming flow of data [3]. We are living in the era of Big Data. Furthermore, this data is mostly unstructured, signifying that traditional systems are not capable of analysing it. Organisations are willing to extract more beneficial information from this high volume and variety of data [4]. A new analysis paradigm with which to analyse and better understand this data, therefore, emerged in order to obtain not only private, but also public, benefits, and this was Big Data [5].

Each new disruptive technology brings new issues with it. In the case of Big Data, these issues are related not only to the volume or the variety of data, but also to data quality, data privacy, and data security. This paper will focus on the subjects of Big Data privacy and security. Big Data not only increases the scale of the challenges related to privacy and security as they are addressed in traditional security management, but also create new ones that need to be approached in a new way [6]. As more data is stored and analysed by organisations or governments, more regulations

are needed to address these concerns. Achieving security in Big Data has, therefore, become one of the most important barriers that could slow down the spread of technology; without adequate security guarantees, Big Data will not achieve the required level of trust [7]. Big Data brings big responsibility [8].

According to the Big Data Working Group at the Cloud Security Alliance organisation there are, principally, four different aspects of Big Data security: infrastructure security, data privacy, data management, and integrity and reactive security [9]. This division of Big Data security into four principal topics has also been used by the International Organisation for Standardisation in order to create a security standard for security in Big Data. Figure 1 contains a scheme showing the main topics related to security in Big Data.

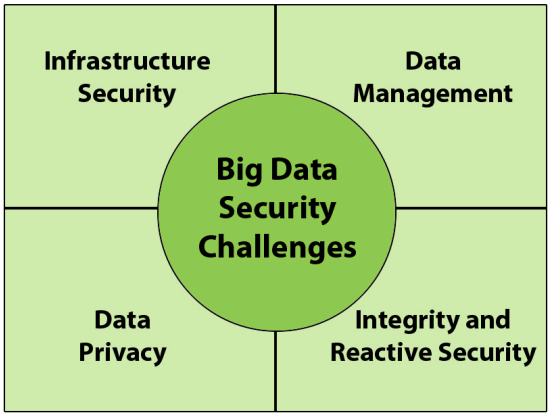

Figure 1. Main challenges as regards security in Big Data security, based on [9].

The purposes of this paper are to highlight the main security challenges that may affect Big Data, along with the solutions that researchers have proposed in order to deal with them. This big picture of the security problem may help other researchers to better understand the security changes produced by the inherent characteristics of the Big Data framework and, consequently, find new research lines so as to carry out more in-depth investigations. This goal has been accomplished by carrying out an empirical investigation by means of the systematic mapping study method with the aim of obtaining a complete background to the security problem as regards Big Data and the proposed solutions.

This paper is consequently structured as follows: first we provide a brief introduction to the subject of Big Data, after which we describe the systematic mapping study process. We then go on to analyse the results obtained, and additionally discuss them. Finally, we present a section concerning our conclusions.

2. Big Data Basis

The term Big Data refers to a framework that allows the analysis and management of a larger amount of data than the traditional data processing technologies [10]. Big Data supposes a change from the traditional techniques in three different ways: the amount of data (volume), the rate of data generation and transmission (velocity), and the types of structured and unstructured data (variety) [11]. These properties are known as the three basic V's of Big Data. Many authors have added new characteristics to the initial group, such as variability, veracity, or value [12].

One of the most important parts of the Big Data world is the use of brand new technologies in order to extract valuable information from data and the ability to combine data from different sources and different formats. Big Data have also changed the way in which organisations store data [13], and have allowed them to develop a more thorough and in-depth understanding of their business, which implies a great benefit [14]. Data was traditionally stored in a structured format, such as a relational database, in order to make it easier to process and to understand that data. There is

now, however, a new tendency: storing the current data volumes in unstructured or semi-structured data [15].

The current maturity of technologies such as Cloud Computing or ubiquitous network connectivity provide a platform on which to easily collect the data, store it, or process it [16]. This set of characteristics has allowed the rapid spread of Big Data techniques. Furthermore, not only can big organisations afford Big Data, but small companies can also obtain benefits from the use of this Big Data ecosystem [17]. A scheme of the typical Big Data ecosystem is shown in Figure 2.

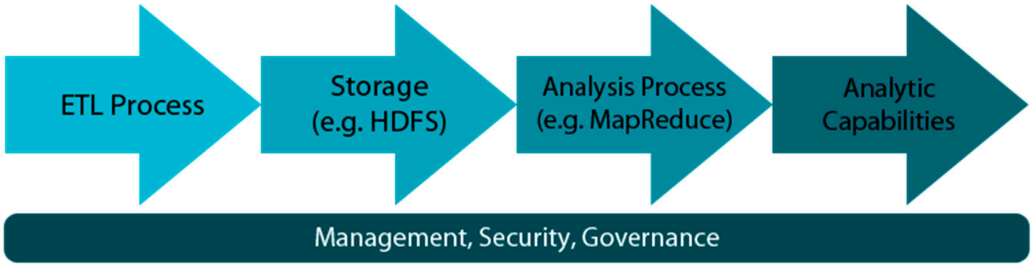

Figure 2. Typical Big Data architecture, based on [18].

This success of the use of Big Data technology can be explained by the release of a type of software: Apache Hadoop. Hadoop is a framework developed by Apache that allows the distributed processing of large dataets across clusters of computers using programming models. It is designed to be scalable from a single server to thousands of them, each of which offers computation and local storage [19]. Hadoop can be considered as a de facto standard [20]. The input for the Hadoop Framework is the data that feed the Big Data system. As explained previously, these data usually originate from very different sources and formats. Hadoop has its own distributed file system (HDFS) which stores the data in different servers with different functions, such as NameNode, which is used to store the metadata, or the DataNodes, which store the application data [21]. The principal characteristic of Hadoop is, however, that of being an open-source implementation of MapReduce [22].

MapReduce is a programming model that is particularly focused on processing and generating large datasets. The MapReduce paradigm accomplishes this goal by describing two different functions [23]:

- The map function, which processes the key/value pair needed to create a set of intermediate key/value pairs.
- The reduce function, which processes the intermediate values generated and merges them to produce a solution.

Apart from the MapReduce framework, there are several different projects that comprise the Hadoop ecosystem, such as Hive, Pig, Sqoop, Mahout, Zookeeper, Spark, or HBase. Each of these tools, along with hundreds of others, provides a range of possibilities that allow organisations to obtain value from their data [3].

Big Data environments do not traditionally prioritise security [24], and it was for this reason that we decided to carry out research in order to discover the main security challenges with respect to Big Data, along with the solutions, methods, or techniques proposed by researchers so as to achieve security in Big Data systems.

3. Systematic Mapping Study

In order to obtain a big picture of the security problem in the Big Data field, we decided to carry out an empirical investigation based on previous literature. We, therefore, resolved to adapt the systematic mapping study method. Mapping studies basically use the same methodology as Systematic lecture reviews, but their main objective is to identify and classify all the research related to

a broad software engineering topic, rather than answering a more specific question. This method has four stages: the research questions, the research method, the case selection and case study roles and procedures and, finally, data analysis and interpretation [25].

3.1. Research Questions

In our case, the questions include the investigation of the main challenges and problems that can be found with respect to the topic of Big Data security, along with another question whose objective is to discover the main security dimensions on which researchers are focusing their efforts. Finally, we wished to discover which different techniques, methodologies or models have already been developed in order to deal with these problems. Table 1 shows a definition of the research questions followed and the motivation behind them.

Table 1. Research questions and their motivations.

Research Questions	Motivation
RQ1. What are the main challenges and problems with respect to Big Data security?	To elicit the main problems and challenges related to Big Data security.
RQ2. What are the main security dimensions on which researchers are focusing their efforts?	To discover what the main focus is for those researching Big Data security.
RQ3. What techniques, methodologies, and models with which to achieve security in Big Data exist?	To explore the different techniques, methodologies, or models used to make Big Data systems secure.

3.2. Research Method

The research method employed was that of carrying out an automatic search of various online libraries: ACM, SCOPUS, and the IEEE Digital Library. The reason for selecting these libraries rather than others was that they contain a great amount of literature related to our main objective and that they would facilitate a specific search. In order to achieve a proper outcome, we created a research string:

$$\begin{gathered}\text{(“Big Data” OR BigData OR Hadoop) AND} \\ \text{(Secur* OR Confidentiality OR Integrity OR Availability OR Privacy)}\end{gathered} \quad (1)$$

The first part of the research string (1) is related to Big Data technology and we, therefore, included the main implementation of Big Data: Hadoop. Hadoop can be considered as a de facto standard. The second part of the string is related to the traditional security dimensions: confidentiality, integrity, and availability. We also decided to include the privacy dimension, even though it can be considered as personal confidentiality. This decision was related to the larger impact that the privacy dimension seems to have with respect to the subject of Big Data in comparison with confidentiality. This perception is also confirmed by the classification made by the CSA [9].

3.3. Previous Works

Before starting our research, we decided to carry out a small search for studies in order to discover whether there were any papers that had already dealt with the subject of our investigation. This goal was accomplished by searching the same online libraries for papers reviewing the security in Big Data between the years 2004 and 2015. We found some reviews focused on more specific topics, such as privacy in social networks [26], but were unable to find any whose objective was to obtain a high level picture of the security in Big Data.

3.4. Case Selection and Case Study Roles and Procedures

The aforementioned research string was then run in the selected online libraries in order search for the words it contained in the titles, keywords, abstracts, and whole texts of papers. This resulted in about 2300 papers. Once we had obtained the papers that conformed to our research question,

it was necessary to make a case selection so as to attain only those that best fitted with our main research aim. In order for us to maintain those that were genuinely related to our research, the selected papers had to fulfil a number of criteria, such as having been published between 2004 (the MapReduce programming paradigm is released by Google [23]) and 2015. We additionally considered only those papers published in journals, conferences, congresses, or important workshops.

We first selected those cases that were truly related to our purpose and made a classification of them. This classification was carried out by focusing our research on the title and abstract of each paper, although it was, in some cases, necessary to read the full paper. We then carried out a new search of the selected literature in order to avoid the inclusion of any duplicates and improve the quality of the research. It is important to highlight that it is possible for a case to belong to different categories, i.e., privacy and integrity. We eventually obtained over 500 papers that adjusted to our parameters and would be useful for our research.

3.5. Main Topics Found

In order to show the main topics found during the research, we have decided to divide them into the four principal aspects employed in the division created by the Big Data Working Group at the Cloud Security Alliance organisation: infrastructure security, data privacy, data management, and integrity and reactive security [9]. This classification has also been used by the NIST Big Data security group for the creation of a security standard for Big Data.

3.5.1. Infrastructure Security

When discussing infrastructure security, it is necessary to highlight the main technologies and frameworks found as regards securing the architecture of a Big Data system, and particularly those based on the Hadoop technology, since it is that most frequently used. In this section we shall also discuss certain other topics, such as communication security in Big Data, or how to achieve high-availability. Figure 3 contains a graphic that shows the main topics found and the quantity of papers dealing with each specific topic.

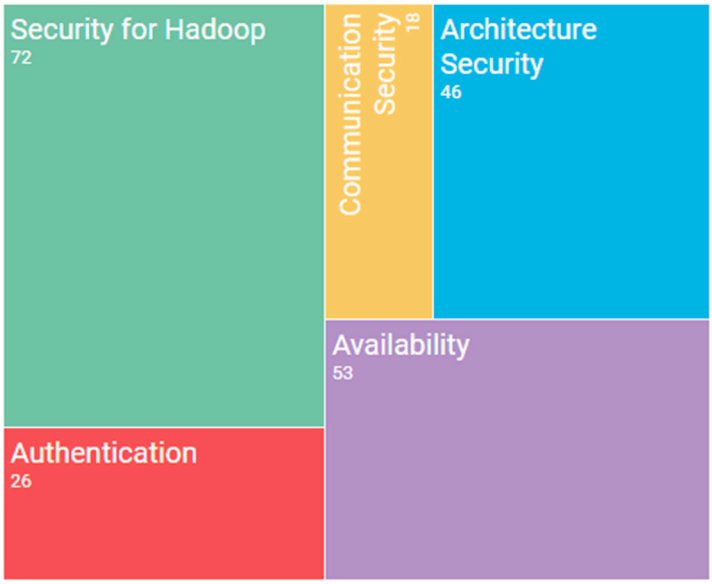

Figure 3. Main topics regarding infrastructure security.

Security for Hadoop

The graphic shows that the main topic dealt with by those researching infrastructure security is security for Hadoop. As explained in previous sections, Hadoop can be considered as a de facto

standard for implementing a Big Data environment in a company. The security problems related to this technology have, therefore, been widely discussed by researchers, who have also proposed various methods with which to improve the security of the Hadoop system. This category is probably the most transverse since, in order to protect it, the solutions use different security mechanisms such as authenticity or cryptography.

For example, there is a proposal for a security model for G-Hadoop (an extension of the MapReduce framework to run on multiple clusters) that simplifies users' authentication and some security mechanisms in order to protect the system from traditional attacks [27]. A few papers focus on protecting the data that is stored in the HDFS by proposing a new schema [24], a secure access system [28], or even the creation of an encryption scheme [29].

Availability

Researchers have also dealt with the subject of availability in Big Data systems. One of the main characteristics of Big Data environments, and by extension of a Hadoop implementation, is the availability attained by the use of hundreds of computers in which the data are not only stored, but are also replicated along the cluster. Finding an architecture that will ensure the full availability of the system is, therefore, a priority.

For instance, in [30] the authors propose a solution with which to achieve high availability by having multiple active NameNodes at the same time. Other solutions are based on creating a new infrastructure of the storage system so as to improve availability and fault tolerance [31,32].

Architecture Security

Another different approach is that of describing a new Big Data architecture, or modifying the typical one, in order to improve the security of the environment. The authors of [33] propose a new architecture based on the Hadoop file system which, when combined with network coding and multi-node reading, makes it possible to improve the security of the system. Another solution focuses on secure group communications in large-scale networks managed by Big Data systems, and this is achieved by creating certain protocols and changing the infrastructure of the nodes [34].

Authentication

The value of the data obtained after executing a Big Data process can, to a great extent, be determined by its authenticity. A few papers deal with this problem by proposing solutions related to authentication. In [35], the authors suggest solving the problem of authentication by creating an identity-based *signcryption* scheme for Big Data.

Communication Security

The security as regards communications between different parts of the Big Data ecosystem is a topic that often is ignored, and only a small number of papers therefore deal with this problem. One paper approaches the topic by explaining the regular data life cycle in a Big Data system, following the different network protocols and applications that the data pass through. The authors also enumerate the main data transfer security techniques [36].

Summary

With regard to the topic of infrastructure security, the main problem dealt with by researchers would appear to be security for Hadoop systems. This is not surprising since, as stated previously, Hadoop can be considered as a de facto standard in industry. The remaining problems addressed in this topic are usually solved by modifying the usual scheme of a Big Data system through the addition of new security layers.

3.5.2. Data Privacy

Data privacy is probably the topic about which ordinary people are most concerned, but it should also be one of the greatest concerns for the organisations that use Big Data techniques. A Big Data system usually contains an enormous amount of personal information that organisations use in order to obtain a benefit from that data. However, we should ask ourselves where the limit regarding the use of that information is.

Organisations should not have total freedom to use that information without our knowledge, although they also need to gain some benefit from the use of that data. Several techniques and mechanisms with which to protect the privacy of the data, and also allow companies to still make a profit from it have therefore been developed, and attempt to solve this problem in various different ways. Figure 4 contains a graphic that shows the main ways in which this problem is dealt with, and the quantity of papers found for each specific topic.

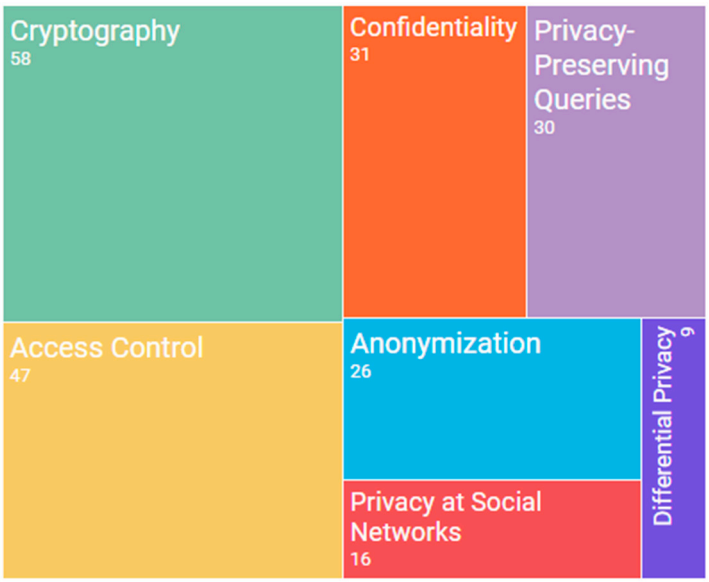

Figure 4. Main topics on data privacy.

Cryptography

The most frequently employed solution as regards securing data privacy in a Big Data system is cryptography. Cryptography has been used to protect data for a considerable amount of time. This tendency continues in the case of Big Data, but it has a few inherent characteristics that make the direct application of traditional cryptography techniques impossible.

One example of the use of cryptography can be found in [37], in which the authors propose a bitmap encryption scheme that guarantees users' privacy. Other authors' research is focused on how to process data that is already encrypted. One paper, for example, explains a technique with which to analyse and programme transformations with *PigLatin* in the case of encrypted data [38].

Access Control

Access control is one of the basic traditional techniques used to achieve the security of a system. Its main objective is to restrict non-desirable users' access to the system. In the case of Big Data, the access control problem is related to the fact that there are only basic forms of access control. In order to solve this problem, some authors propose a framework that supports the integration of access control features [39]. Other researchers focus their attention on the MapReduce process itself, and suggest a framework with which to enforce the security policies at the key-value level [40].

Confidentiality

Although privacy is traditionally treated as a part of confidentiality, we decided to change the order owing to the tremendous impact that privacy has on the general public's perception of Big Data technology.

The authors that approach this problem often propose new techniques such as computing on masked data (CMD), which improves data confidentiality and integrity by allowing direct computations to be made on masked data [41], or new schemes, such as Trusted Scheme for Hadoop Cluster (TSHC) which creates a new architecture framework for Hadoop in order to improve the confidentiality and security of the data [42].

Privacy-Preserving Queries

The main purpose of a Big Data system is to analyse the data in order to obtain valuable information. However, while we manipulate that data we should not forget its privacy. A few papers pay attention to the problem of how to make queries whilst simultaneously not violating the privacy of the data.

One way in which to achieve this protection is by encrypting the data, as discussed previously, but this adds a new problem: how do we analyse the encrypted data? Some authors propose that this problem can be solved by means of a secure keyword search mechanism over that encrypted data [43].

Anonymisation

One of the most extended ways in which to protect the privacy of data is by anonymising it. This consists of applying some kind of technique or mechanism to the data in order to remove the sensitive information from it or to hide it. Big Data usually implies a large amount of data, and this problem, therefore, increases in Big Data environments.

The authors of [44] propose a hybrid method that combines the two most frequently used anonymisation schemes: top-down specialisation (TDS) and bottom-up generalisation (BUG).

Privacy in Social Networks

Social networks are all around us. The popularity of Social networks is currently huge, and almost everybody with access to the Internet has at least one account with them. People share a lot of personal information in these networks without actually worrying about what the organisation behind them will do with their data. This data, along with the strong analysis capability of Big Data, is a huge threat to our personal privacy.

Addressing this problem is not an easy task, and some authors suggest new legislation with which to increase the protection of data privacy [45]. Another paper, meanwhile, proposes a technique that can be used to increase the control that users have over their own data in social networks [46].

Differential Privacy

The objective of differential privacy is to provide a method with which to maximise the value of analysis of a set of data while minimising the chances of identifying users' identities. A few papers focus on achieving privacy in Big Data by applying differential privacy techniques. For example, in [47] the authors attempt to distort the data by adding noise.

Summary

The topic most frequently dealt with by researchers would appear to be privacy. There are a lot of different perspectives as regards ensuring privacy. Authors usually propose different means of encryption, based on traditional techniques but with a few changes in order to adapt these techniques to the inherit characteristics of a Big Data environment. Owing to the large amount of papers found

on this topic in comparison to the others, we believe that it is advisable to split this category up into, on the one hand, data privacy itself, and on the other, cryptography and access control techniques.

3.5.3. Data Management

This section focuses on what to do once the data is contained in the Big Data environment. It not only shows how to secure the data that is stored in the Big Data system, but also how to share that data. We shall also discuss the different policies and legislation that authors suggest in order to use Big Data techniques safely. Figure 5 contains a graphic that shows the topics that will be discussed in this section, along with the quantity of papers found for each specific topic.

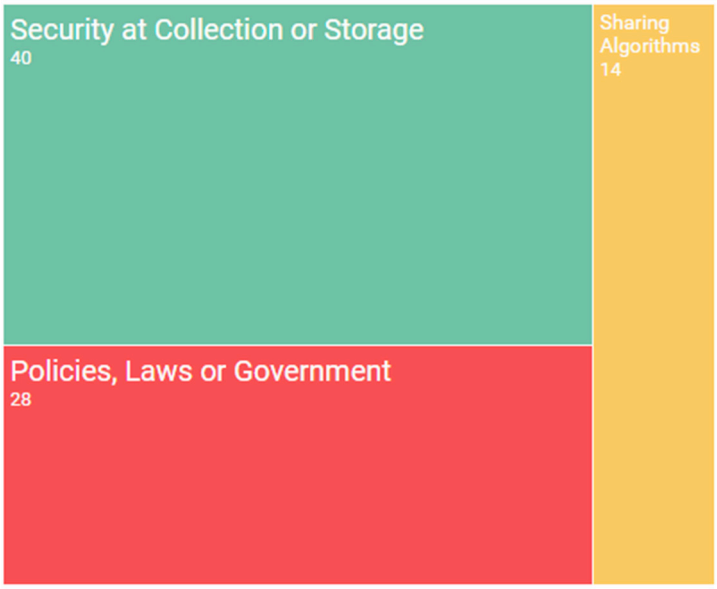

Figure 5. Main topics on data management.

Security at Collection or Storage

As mentioned previously, Big Data usually implies a huge amount of data. It is, therefore, important not only to find a means to protect data when it is stored in a Big Data environment, but also to know how to initially collect that data.

In order to solve these problems, some authors propose a mechanism with which to protect data owners' privacy by creating a parameter to measure the acceptable level of privacy [48]. Another approach found in [49], suggests that security storage can be protected by dividing the data stored in the Big Data system into sequenced parts and storing them in different cloud storage service providers.

Policies, Laws, or Government

Every disruptive technology brings new problems with it, and Big Data is no exception. The problems related to Big Data are mostly related to the increase in the use of this technique to obtain value from a large amount of data by using its powerful analysis characteristics. This could imply a threat to people's privacy. In order to reduce that risk, many authors propose the creation of new legislation and laws that will allow these new problems to be confronted in an effective manner. For example, in [50] the authors propose a legal framework in order to protect students' privacy.

The purpose of this subtopic was also to find some frameworks or initiatives that attempt to establish a robust government of the security of the data in a Big Data system, but were unable to find any papers that cover this problem with respect to the full life cycle of a Big Data environment.

Sharing Algorithms

In order to obtain the maximum possible value from data, it is necessary to share that data among the cluster in which Big Data is running or to share those results for collaboration. However, again, we have the problem of how to guarantee security and privacy when that sharing process is taking place.

Some authors approach this problem by increasing the surveillance of the user taking part in data sharing [51], while others propose securing the transmission itself by creating a new technique based on nested sparse sampling and coprime sampling [52].

Summary

This section includes almost the entire lifecycle of the data used in a Big Data system, from its collection to its sharing, and also includes how to properly govern the security of that data. With regard to collection and sharing, authors propose the creation of new schemas, frameworks, and protocols with which to secure data. Other authors also suggest toughening up the legislation concerning the privacy of the data used by companies. Furthermore, we have found a lack of papers dealing with the need to create a framework that covers security data governance in a Big Data system in its entire lifecycle.

3.5.4. Integrity and Reactive Security

One of the bases on which Big Data is supported is the capacity to receive streams of data from many different origins and with distinct formats: either structural data or non-structural data. This increases the importance of checking that the data's integrity is good so that it can be used properly. This topic also covers the use case of applying Big Data in order to monitor security so as to detect whether a system is being attacked. Figure 6 contains a bar chart that shows the main subtopics found during the systematic mapping study, and the quantity of papers for each specific topic.

Figure 6. Main topics as regards integrity and reactive security.

Integrity

Integrity has traditionally been defined as the maintenance of the consistency, accuracy, and trustworthiness of data. It protects data from unauthorised alterations during its lifecycle. Integrity is considered to be one of the three basic dimensions of security (along with confidentiality

and availability). Ensuring integrity is critical in a Big Data environment, and authors agree as to the difficulty of achieving the proper integrity of data when attempting to manage this problem.

For example, they propose an external integrity verification of the data [53] or a framework to ensure it during a MapReduce process [54].

Attack Detection

As occurs with all systems, Big Data may be attacked by malicious users. Some authors, therefore, take advantage of the inherit characteristics of Big Data and suggest certain indicators that may be a sign that the Big Data environment is under attack.

For instance, in [55] the authors develop a computational system that captures the provenance data related to a MapReduce process. There are also researchers who propose an intrusion detection system especially intended for the specific characteristics of a Big Data environment [56].

Recovery

The main purpose of this topic is to create particular policies or controls in order to ensure that the system recovers as soon as possible when a disaster occurs. Many organisations currently store their data in Big Data systems, signifying that if a disaster occurs the entire company could be in danger.

We have found only a few papers that cover this problem. For example, in [57] there are some recommendations regarding what can be done to recover from a desperate situation.

Summary

In this section, the main topic discussed by researchers would appear to be the integrity of data. In order to secure that integrity, they propose various kinds of verification to ensure that the data has not been modified. This section also covers the possibility of detecting the attacks that a Big Data system may undergo. There is a lack of papers dealing with the possibility of disaster occurring in a Big Data system, and how to recover from it. This is probably a consequence of the high availability that a Big Data system usually achieves, but this topic should not be overlooked.

4. Analysis of Results

Analysing such a large amount of results is not an easy task, and in order to meticulously interpret them we shall, therefore, answer the research questions in the specified order.

The first question was "What are the main challenges and problems as regards security in Big Data?" This was easily answered, because during the first phase of research we found that a few documents have been produced by the Cloud Security Alliance and by the National Institute of Standards and Technology that approach the topic of security in Big Data and highlight the main problems and challenges that concern this technology. These results allowed us to guide the remainder of our research.

Figure 7 shows a pie chart with the amount of papers grouped by the different categories. However, according to the results of our research, and owing to the quantity of papers found that are related to this problem, we believed that it might be useful to split the privacy category into two different categories: privacy, itself, and access control and cryptography.

The second and third questions were "What are the main security dimensions on which researchers are focusing their efforts?" and "What techniques, methodologies, and models with which to achieve security in Big Data exist?" These questions form the main body of our research, and, in order to simplify the visualisation of the results, we have, therefore, created Table 2, which connects the main topics found with the typical security dimensions. The last column shows those papers that are not clearly related to any of these dimensions or deal with security, in general, without specifying more.

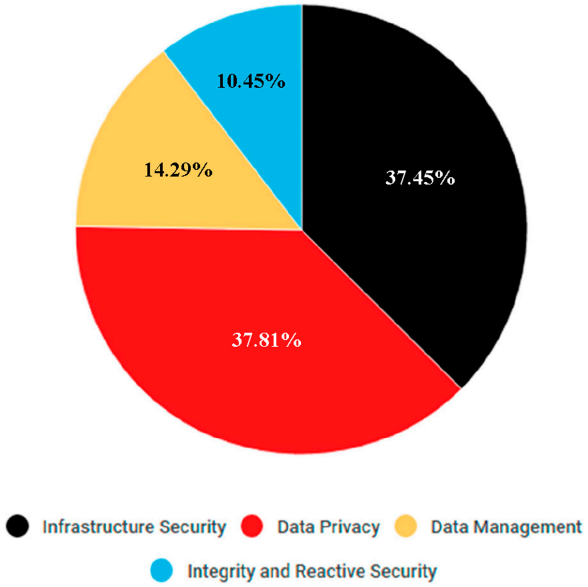

Figure 7. Papers grouped by main categories.

Table 2. Combined results.

	Availability		Confidentiality		Integrity		Privacy		Other	
	Prob. [1]	Sol. [2]	Prob. [1]	Sol. [2]	Prob. [1]	Sol. [2]	Prob. [1]	Sol. [2]	Prob. [1]	Sol. [2]
Infrastructure Security	16	74	7	15	11	33	11	50	27	56
Security for Hadoop	1	16	1	7	1	11	0	14	11	28
Availability	13	46	5	3	6	6	4	6	0	0
Architecture Security	0	8	0	3	1	9	3	17	10	8
Authentication	1	1	1	2	2	5	2	7	4	12
Communication Security	1	3	0	0	1	2	2	6	2	8
Data Privacy	5	7	13	45	6	17	36	108	10	43
Cryptography	0	2	2	11	1	5	4	25	4	18
Access Control	0	2	1	4	0	1	6	17	5	21
Confidentiality	5	3	10	28	5	9	2	9	0	0
Privacy-preserving queries	0	0	0	1	0	2	7	18	0	4
Anonymization	0	0	0	1	0	0	2	25	0	0
Privacy at Social Networks	0	0	0	0	0	0	13	5	1	0
Differential Privacy	0	0	0	0	0	0	2	9	0	0
Data Management	5	7	4	8	3	5	29	34	9	16
Security at Collection or Storage	3	5	2	5	2	4	8	15	5	8
Policies, Laws or Government	0	0	0	1	0	0	19	11	2	3
Sharing Algorithms	2	2	2	3	1	1	2	8	2	5
Integrity and Reactive Security	6	8	5	9	13	39	9	8	2	8
Integrity	6	6	5	9	13	37	6	5	0	0
Attack Detection	0	0	0	0	0	2	3	2	2	7
Recovery	0	2	0	0	0	0	0	1	0	1

[1] Problems, issues, or challenges found; [2] Solutions, frameworks, or mechanisms to solve the problem.

The purpose of this table is to make it easier to visualise the main problems and their relation to the security dimensions. Each column shows the number of papers found, and there are two approaches: on the one hand, those papers that explain the problem with respect to one specific topic, and on the other, those that deal with the problem and propose a solution. Many papers are, of course, located in both columns. That is to say, we have found forty-three papers related to integrity (thirteen express the problem and thirty-seven propose a solution, but there are a few that do both things at the same time).

The table shows that the main problems are principally related to infrastructure problems and privacy issues. In the first case, researchers focus their attention on creating new Big Data architectures that deal with the problems of availability and privacy. Privacy challenges are, meanwhile, by far the most frequently discussed topic. Many authors deal with this topic by explaining the new privacy problems that have arisen as the result of the use of the Big Data technology, while others attempt to

solve the issue by applying variations of traditional techniques that have been adjusted to the inherent characteristics of Big Data.

The other two main topics found are researched to a far lesser extent than those mentioned above. While integrity is well covered, there are only a few papers dealing with the problem of recovering the system in the case of a total failure. Furthermore, we discovered that the topic of data management is not dealt with as frequently as it should be. For example, we have not found any security government frameworks that would make it possible to manage the security of a Big Data system throughout its entire life cycle. We believe that this is crucial if the correct deployment of Big Data technology is to be achieved.

Related with the quantity of studies found per year, we have detected how the number of papers written by the researchers has been continuously increasing until the year 2015. This can be a consequence of the maturity reached by the Big Data technology. Figure 8 contains a graph with the evolution in the quantity of papers found during the considered period.

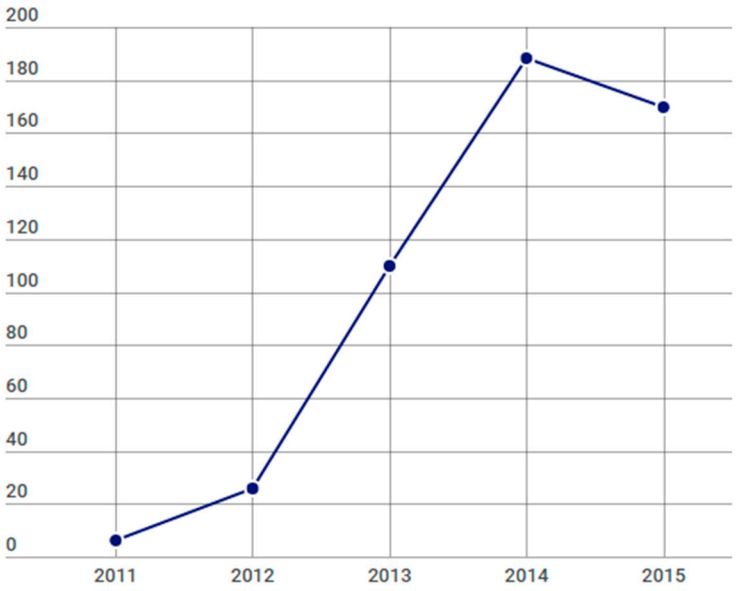

Figure 8. Quantity of papers per year.

5. Conclusions

This paper provides an explanation of the research carried out in order to discover the main problems and challenges related to security in Big Data, and how researchers are dealing with these problems. This objective was achieved by following the systematic mapping study methodology, which allowed us to find the papers related to our main goal.

Having done so, we discovered that the principal problems are related to the inherent characteristics of a Big Data system, and also to the fact that security issues were not contemplated when Big Data was initially conceived. Many authors, therefore, focus their research on creating means to protect data, particularly with respect to privacy, but privacy it is not the only security problem that can be found in a Big Data system; the traditional architecture itself and how to protect a Hadoop system is also a huge concern for the researchers.

We have, however, also detected a lack of investigations in the field of data management, especially with respect to government. We are of the considered opinion that this is not acceptable, since having a government security framework will allow the rapid spread of Big Data technology.

In conclusion, the Big Data technology seems to be reaching a mature stage, and that is the reason why there have been a number of studies created the last year. However, that does not mean that it is no longer necessary to study this paradigm, in fact, the studies created from now should focus on more

specific problems. Furthermore, Big Data can be useful as a base for the development of the future technologies that will change the world as we see it, like the Internet of Things (IoT), or on-demand services, and that is the reason why Big Data is, after all, the future.

Acknowledgments: This work has been funded by the SEQUOIA project (Ministerio de Economía y Competitividad and the Fondo Europeo de Desarrollo Regional FEDER, TIN2015-63502-C3-1-R), and by the SERENIDAD project (Consejería de Educación, Ciencia y Cultura de la Junta de Comunidades de Castilla-La Mancha, y Fondo Europeo de Desarrollo Regional FEDER, PEII-2014-045-P).

Author Contributions: E.F.-M. had the idea of carrying out a systematic mapping review; J.M. created the protocol to follow during the process; M.A.S. validated the protocol; J.M. carried out the research and analyzed the results; E.F.-M. and M.A.S. supervised and helped during the analysis process; J.M. wrote the paper.

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

References

1. Mayer-Schönberger, V.; Cukier, K. *Big Data: A Revolution that Will Transform How We Live, Work, and Think*; Houghton Mifflin Harcourt: Boston, MA, USA, 2013.
2. Sagiroglu, S.; Sinanc, D. Big data: A review. In Proceedings of the 2013 International Conference on Collaboration Technologies and Systems (CTS), San Diego, CA, USA, 20–24 May 2013; pp. 42–47.
3. Hashem, I.A.T.; Yaqoob, I.; Anuar, N.B.; Mokhtar, S.; Gani, A.; Ullah Khan, S. The rise of "big data" on cloud computing: Review and open research issues. *Inf. Syst.* **2015**, *47*, 98–115. [CrossRef]
4. Sharma, S. Rise of Big Data and related issues. In Proceedings of the 2015 Annual IEEE India Conference (INDICON), New Delhi, India, 17–20 December 2015; pp. 1–6.
5. Eynon, R. The rise of Big Data: What does it mean for education, technology, and media research? *Learn. Media Technol.* **2013**, *38*, 237–240. [CrossRef]
6. Wang, H.; Jiang, X.; Kambourakis, G. Special issue on Security, Privacy and Trust in network-based Big Data. *Inf. Sci. Int. J.* **2015**, *318*, 48–50. [CrossRef]
7. Thuraisingham, B. Big data security and privacy. In Proceedings of the 5th ACM Conference on Data and Application Security and Privacy, San Antonio, TX, USA, 2–4 March 2015; pp. 279–280.
8. Rijmenam, V. *Think Bigger: Developing a Successful Big Data Strategy for Your Business*; Amacom: New York, NY, USA, 2014.
9. Big Data Working Group; Cloud Security Alliance (CSA). Expanded Top Ten Big Data Security and Privacy. April 2013. Available online: https://downloads.cloudsecurityalliance.org/initiatives/bdwg/Expanded_Top_Ten_Big_Data_Security_and_Privacy_Challenges.pdf (accessed on 9 December 2015).
10. Meng, X.; Ci, X. Big data management: Concepts, techniques and challenges. *Comput. Res. Dev.* **2013**, *50*, 146–169.
11. Chen, M.; Mao, S.; Liu, Y. Big data: A survey. *Mob. Netw. Appl.* **2014**, *19*, 171–209. [CrossRef]
12. Khan, M.A.-U.-D.; Uddin, M.F.; Gupta, N. Seven V's of Big Data understanding Big Data to extract value. In Proceedings of the 2014 Zone 1 Conference of the American Society for Engineering (ASEE Zone 1 2014), Bridgeport, CT, USA, 3–5 April 2014.
13. Cumbley, R.; Church, P. Is "Big Data" creepy? *Comput. Law Secur. Rev.* **2013**, *29*, 601–609. [CrossRef]
14. Dijcks, J.P. Oracle: Big data for the enterprise. In *Oracle White Paper*; Oracle Corporation: Redwood City, CA, USA, 2012.
15. Minelli, M.; Chambers, M.; Dhiraj, A. *Big Data, Big Analytics: Emerging Business Intelligence and Analytic Trends for Today's Businesses*; John Wiley & Sons: New York, NY, USA, 2013.
16. Demchenko, Y.; De Laat, C.; Membrey, P. Defining architecture components of the Big Data Ecosystem. In Proceedings of the 2014 International Conference on Collaboration Technologies and Systems (CTS 2014), Minneapolis, MN, USA, 19–23 May 2014; pp. 104–112.
17. Kumaresan, A. Framework for building a big data platform for publishing industry. In *Knowledge Management in Organizations*; Springer International Publishing: Cham, Switzerland, 2015; pp. 377–388.
18. Helen, S.; Peter, H. *Oracle Information Architecture: An Architect's Guide to Big Data*; Oracle Corporation: Redwood City, CA, USA, 2012.
19. Apache Hadoop. Available online: http://hadoop.apache.org/ (accessed on 14 March 2016).

20. Cackett, D. Information Management and Big Data A Reference Architecture. Oracle: Redwood City, CA, USA, 2013.
21. Shvachko, K.; Kuang, H.; Radia, S.; Chansler, R. The Hadoop distributed file system. In Proceedings of the 2010 IEEE 26th Symposium on Mass Storage Systems and Technologies (MSST2010), Incline Village, NV, USA, 3–7 May 2010.
22. Jiang, D.; Ooi, B.C.; Shi, L.; Wu, S. The performance of mapreduce: An indepth study. *Proc. VLDB Endow.* **2010**, *3*, 472–483. [CrossRef]
23. Dean, J.; Ghemawat, S. MapReduce: Simplified Data Processing on Large Clusters. *Commun. ACM* **2004**, *51*, 107–113. [CrossRef]
24. Jeong, Y.-S.; Kim, Y.-T. A token-based authentication security scheme for Hadoop distributed file system using elliptic curve cryptography. *J. Comput. Virol. Hacking Tech.* **2015**, *11*, 137–142. [CrossRef]
25. Kitchenham, B.A.; Budgen, D.; Pearl Brereton, O. Using mapping studies as the basis for further research—A participant-observer case study. *Inf. Softw. Technol.* **2011**, *53*, 638–651. [CrossRef]
26. Rekha, H.S.; Prakash, C.; Kavitha, G. Understanding Trust and Privacy of Big Data in Social Networks—A Brief Review. In Proceedings of the 2014 3rd International Conference on Eco-Friendly Computing and Communication Systems (ICECCS 2014), Bangalore, India, 18–21 December 2014; pp. 138–143.
27. Zhao, J.; Wang, L.; Tao, J.; Chen, J.; Sun, W.; Ranjan, R.; Kołodziej, J.; Streit, A.; Georgakopoulos, D. A security framework in G-Hadoop for big data computing across distributed Cloud data centres. *J. Comput. Syst. Sci.* **2014**, *80*, 994–1007. [CrossRef]
28. Yang, C.-T.; Shih, W.-C.; Chen, L.-T.; Kuo, C.-T.; Jiang, F.-C.; Leu, F.-Y. Accessing medical image file with co-allocation HDFS in cloud. *Future Gener. Comput. Syst.* **2015**, *43–44*, 61–73. [CrossRef]
29. Cohen, J.C.; Acharya, S. Towards a trusted HDFS storage platform: Mitigating threats to Hadoop infrastructures using hardware-accelerated encryption with TPM-rooted key protection. *J. Inf. Secur. Appl.* **2014**, *19*, 224–244. [CrossRef]
30. Wang, Z.; Wang, D. NCluster: Using Multiple Active Name Nodes to Achieve High Availability for HDFS. In Proceedings of the 2013 IEEE 10th International Conference on High Performance Computing and Communications & 2013 IEEE International Conference on Embedded and Ubiquitous Computing (HPCC_EUC), Zhangjiajie, China, 13–15 November 2013; pp. 2291–2297.
31. Meye, P.; Raipin, P.; Tronel, F.; Anceaume, E. Mistore: A distributed storage system leveraging the DSL infrastructure of an ISP. In Proceedings of the 2014 International Conference on High Performance Computing & Simulation (HPCS), Bologna, Italy, 21–25 July 2014; pp. 260–267.
32. Azeem, M.A.; Sharfuddin, M.; Ragunathan, T. Support-based replication algorithm for cloud storage systems. In Proceedings of the 7th ACM India Computing Conference, Nagpur, India, 9–11 October 2014; pp. 1–9.
33. Ma, Y.; Zhou, Y.; Yu, Y.; Peng, C.; Wang, Z.; Du, S. A Novel Approach for Improving Security and Storage Efficiency on HDFS. *Procedia Comput. Sci.* **2015**, *52*, 631–635. [CrossRef]
34. He, S.; Wu, Q.; Qin, B.; Liu, J.; Li, Y. Efficient group key management for secure big data in predictable large-scale networks. *Concurr. Comput.* **2016**, *28*, 1174–1192. [CrossRef]
35. Wei, G.; Shao, J.; Xiang, Y.; Zhu, P.; Lu, R. Obtain confidentiality or/and authenticity in Big Data by ID-based generalized signcryption. *Inf. Sci.* **2015**, *318*, 111–122. [CrossRef]
36. Frank, J.B.; Feltus, A. The Widening Gulf between Genomics Data Generation and Consumption: A Practical Guide to Big Data Transfer Technology. *Bioinf. Biol. Insights* **2015**, *9* (Suppl. 1), 9–19.
37. Yoon, M.; Cho, A.; Jang, M.; Chang, J.W. A data encryption scheme and GPU-based query processing algorithm for spatial data outsourcing. In Proceedings of the 2015 International Conference on Big Data and Smart Computing (BIGCOMP), Jeju, Korea, 9–12 February 2015; pp. 202–209.
38. Stephen, J.J.; Savvides, S.; Seidel, R.; Eugster, P. Program analysis for secure big data processing. In Proceedings of the 29th ACM/IEEE international conference on Automated software engineering, Vasteras, Sweden, 15–19 September 2014; pp. 277–288.
39. Colombo, P.; Ferrari, E. Privacy Aware Access Control for Big Data: A Research Roadmap. *Big Data Res.* **2015**, *2*, 145–154. [CrossRef]
40. Ulusoy, H.; Colombo, P.; Ferrari, E.; Kantarcioglu, M.; Pattuk, E. GuardMR: Fine-grained Security Policy Enforcement for MapReduce Systems. In Proceedings of the 10th ACM Symposium on Information, Computer and Communications Security, Singapore, 14–17 April 2015; pp. 285–296.

41. Kepner, J.; Gadepally, V.; Michaleas, P.; Schear, N.; Varia, M.; Yerukhimovich, A.; Cunningham, R.K. Computing on masked data: A high performance method for improving big data veracity. In Proceedings of the 2014 IEEE High Performance Extreme Computing Conference (HPEC), Waltham, MA, USA, 9–11 September 2014; pp. 1–6.
42. Quan, Z.; Xiao, D.; Wu, D.; Tang, C.; Rong, C. TSHC: Trusted Scheme for Hadoop Cluster. In Proceedings of the 2013 Fourth International Conference on Emerging Intelligent Data and Web Technologies (EIDWT), Xi'an, China, 9–11 September 2013; pp. 344–349.
43. Kuzu, M.; Islam, M.S.; Kantarcioglu, M. Distributed Search over Encrypted Big Data. In Proceedings of the 5th ACM Conference on Data and Application Security and Privacy, San Antonio, TX, USA, 2–4 March 2015; pp. 271–278.
44. Irudayasamy, A.; Arockiam, L. Scalable multidimensional anonymization algorithm over big data using map reduce on public cloud. *J. Theor. Appl. Inf. Technol.* **2015**, *74*, 221–231.
45. Mantelero, A.; Vaciago, G. Social media and Big Data. In *Cyber Crime and Cyber Terrorism Investigator's Handbook*; Syngress: Boston, MA, USA, 2014; pp. 175–195.
46. Estivill-Castro, V.; Hough, P.; Islam, M.Z. Empowering users of social networks to assess their privacy risks. In Proceedings of the 2014 IEEE International Conference on Big Data, Washington, DC, USA, 27–30 October 2014; pp. 644–649.
47. Ren, H.; Wang, S.; Li, H. Differential privacy data Aggregation Optimizing Method and application to data visualization. In Proceedings of the 2014 IEEE Workshop on Electronics, Computer and Applications (IWECA 2014), Ottawa, ON, Canada, 8–9 May 2014; pp. 54–58.
48. Xu, L.; Jiang, C.; Chen, Y.; Ren, Y.; Liu, K.J.R. Privacy or Utility in Data Collection? A Contract Theoretic Approach. *IEEE J. Sel. Top. Signal Proc.* **2015**, *9*, 1256–1269.
49. Cheng, H.; Rong, C.; Hwang, K.; Wang, W.; Li, Y. Secure big data storage and sharing scheme for cloud tenants. *China Commun.* **2015**, *12*, 106–115. [CrossRef]
50. Weber, A.S. Suggested legal framework for student data privacy in the age of big data and smart devices. In *Smart Digital Futures*; IOS Press: Washington, DC, USA, 2014; Volume 262.
51. Thilakanathan, D.; Calvo, R.; Chen, S.; Nepal, S. Secure and controlled sharing of data in distributed computing. In Proceedings of the 16th IEEE International Conference on Computational Science and Engineering (CSE 2013), Sydney, Australia, 3–5 December 2013; pp. 825–832.
52. Chen, J.; Liang, Q.; Wang, J. Secure transmission for big data based on nested sampling and coprime sampling with spectrum efficiency. *Secur. Commun. Netw.* **2015**, *8*, 2447–2456. [CrossRef]
53. Liu, C.; Yang, C.; Zhang, X.; Chen, J. External integrity verification for outsourced big data in cloud and IoT. *Future Gener. Comput. Syst.* **2015**, *49*, 58–67. [CrossRef]
54. Wang, Y.; Wei, J.; Srivatsa, M.; Duan, Y.; Du, W. IntegrityMR: Integrity assurance framework for big data analytics and management applications. In Proceedings of the 2013 IEEE International Conference on Big Data, Silicon Valley, CA, USA, 6–9 October 2013; pp. 33–40.
55. Liao, C.; Squicciarini, A. Towards provenance-based anomaly detection in MapReduce. In Proceedings of the 2015 15th IEEE/ACM International Symposium on Cluster, Cloud and Grid Computing (CCGrid), Shenzhen, China, 4–7 May 2015; pp. 647–656.
56. Tan, Z.; Nagar, U.T.; He, X.; Nanda, P.; Liu, R.P.; Wang, S.; Hu, J. Enhancing big data security with collaborative intrusion detection. *IEEE Cloud Comput.* **2014**, *1*, 27–33. [CrossRef]
57. Chang, V. Towards a Big Data system disaster recovery in a Private Cloud. *Ad Hoc Netw.* **2015**, *35*, 65–82. [CrossRef]