TuEngr Group

International Transaction Journal of Engineering, Management, & Applied Sciences & Technologies

http://TuEngr.com

PAPER ID: 11A9M

TECHNOLOGICAL TOOLS FOR DATA SECURITY IN THE TREATMENT OF DATA RELIABILITY IN BIG DATA ENVIRONMENTS

Luay Abdulwahid Shihab [1*]

[1] *Department of Basic Science, College of Nursing, University of Basrah, Basrah, IRAQ.*

ARTICLEINFO

Article history:
Received 03 January 2020
Received in revised form 05 March 2020
Accepted 16 March 2020
Available online 28 March 2020

Keywords:
Information security; Data authenticity; Technological security tools; Reliability of data; Information security mechanism; Big Data technologies; Data legality; Network security measures; Veracity in Big Data; Data privacy.

ABSTRACT

The set of new technological solutions that allow organizations to better manage their information, commonly known as "Big Data", have a growing role in all types of public and private organizations. As a Big Data problem, how data grows in volume, speed and variety can be contemplated. This is due to the great advance and use of information technologies, and the daily use that people make of them. Within the state of the art are found from various definitions of the term Big Data to existing technologies to start a project in an institution of any productive, commercial, or educational branch. This article gives an overview of the data security technology processes, defining those that lead to rising data veracity in Big Data environments. As a result of this analysis, a series of criteria was established relevant to the authenticity of the data and the use of network security measures were suggested for each of these criteria. The article also seeks to lead to further work on information security within Information Science, as it would provide a perspective on the methods available for approaching information security, leading to increasing the reliability of knowledge obtained from contexts containing significant volumes of knowledge. This work proposes adding two criteria for veracity, highlighted as a contribution of this work, in addition to the previous criteria. These are legality and privacy.

Disciplinary: Computer Sciences and Information Technology.

1. INTRODUCTION

Increasing the amount of data generated by numerous applications and day-to-day operations in society has forged the need to alter, refine and produce data management and treatment approaches and models that account for database and computer system limitations. Big Data emerges in reaction to this, a concept that incorporates multiple techniques involved with handling vast volumes of data from diverse sources and is easily generated. (Li et al 2015).

*Corresponding author (L.A.Shihab). Email: Luaay.abdulwahid@uobasrah.edu.iq

The word Big Data is generally synonymous with exorbitant volumes of data. This idea must be put aside because Big Data is not only targeted at a massive scale, it rather includes both the quantity and range of data and the speed and efficiency of entry. The process has now been moved to contact, to get the most value from the information produced minute by minute (Mohanty et al., 2015).

With the rise of Big Data, a new concept, Data Science has also been accommodated, in a generic way to refer to the series of techniques necessary for the treatment and manipulation of massive information from an approach of statistical and computer science. Also, the emergence of a new professional profile, the "Data Scientist" (Aalst, 2014), the people trained in this profile should know about the business, computational tools, and statistical analysis and interpretation.

Among the objectives of Information Science is to provide a means for making relevant information available to individuals, groups, and organizations involved with science and technology (Douglas, 2006). Bhadani *et al.* (2016) conceptualize information security as a set of measures that aims to protect and preserve information and information systems, that any information must be correct, accurate, and available to be stored, retrieved, processed, and made available safely and reliably. Oussous (2017) conceptualizes Big Data's environments that have large volumes of data, and that if the processed data is not authentic, the information generated will not be reliable. In addition to the volume, other characteristics are speed, variety, veracity, value, variability, and visualization (Zikopoulos et al., 2011).

To contribute to increasing the reliability of the data and information generated in the Big Data environments, this study proposes to carry out an analysis focusing on the characteristic veracity in Big Data, presenting the relations of this aspect with the technological mechanisms of information security.

1.1 INFORMATION SECURITY AND ITS MECHANISMS

Information security transcends computational controls, thus allowing different approaches to be used in its conceptual defines information security as a process to protect information from misuse, whether intentional or not, carried out by people inside or outside the organization (DeCandia et al., 2007). Information security makes it possible to use the resources that support the necessary information for strategic, tactical, and operational activities in a reliable manner in an organization (Bahga and Madisetti, 2012). Feng (2014) explores data security as an expertise field devoted to securing data assets from unlawful entry, improper modification, or unavailability of information assets. Implementing an adequate range of safeguards, including rules, protocols, practices, operational structure, and device and hardware functions ensure information security. In its origins and for historical reasons, information security is generally associated with the principle of confidentiality/secrecy where security controls in a more rudimentary version were adopted to guarantee the confidentiality of critical information. Confidentiality is an important principle within information security, but others must be considered to ensure the effective protection of data and information. (Marchal et al.,2014) states that three basic principles guide the implementation of information security, including confidentiality, integrity, and availability. For critical information must be protected to prevent its destruction, temporary unavailability, tampering, or unauthorized disclosure. Manogaran (2016) also describes that the degree of availability, integrity, and confidentiality will protect information for the organization to operationalize its business and meet its objectives.

Hashizume et al. (2013) present two additional principles for the protection of information to be achieved, authenticity and non-repudiation - also found in the literature as irreversible. Already (Sabahi, 2011) also introduces the principle of timeliness that references to electronic documents, so that they have the same degree of reliability in the existing paper documents such as authenticity, integrity, and timeliness. For the development of this study, the six principles of information security presented by these authors are considered:

- Confidentiality is the principle that ensures that information is accessed only by authorized entities.
- Integrity ensures that the information is complete and faithful, that is, that it does not change by entities not authorized by its owner.
- Availability deals with the principle that ensures that information is available whenever necessary.
- Authenticity guarantees that the entities involved in a process containing digital information are authentic, that is, if they are true, ensuring that the information comes from the advertised source.
- Irretractability (also called non-repudiation) refers to the guarantee that an entity does not deny the authorship of something done by it.
- Timeliness is the principle that guarantees the validity of digital information over time.

For each of the principles, specific technological mechanisms are adopted, which are the tools, technologies, and protocols used in the information security process. The technological mechanisms of information security to be used in the analysis are proposed. Note that digital certificates and time stamps are technologies that implement security mechanisms, however, they are addressed in this work because they are technologies consolidated in the legal scope.

1.2 TECHNOLOGICAL MECHANISMS FOR INFORMATION SECURITY

- **Cryptography:** It is of fundamental importance for information security since it is the basis for several technologies and protocols used to guarantee confidentiality, integrity, authentication, and non-retractability of information. This mechanism transforms readable data into unreadable data using a code in such a way that only authorized entities and owners of it can decrypt and interpret them (LiYi et al., 2017) and the sleek and encryption systems, electronic high security because of the strong and difficult to break and therefore the confidentiality of this system depends on the secret key Encryption converts the text flowing to the explicit text of the encrypted beta - beta at the same time. (Luaay Abdulwahed Shihab et al, 2012)
- **Hashing:** They are mathematical calculations used in algorithms that produce the history of the information, making it possible to identify whether it has been altered. Calculation algorithms hashing are used to ensure the integrity and identify whether there were unforeseen changes. (Harfoushi, Osama & Obiedat, Ruba., 2018)
- **Digital signature**: It is the combination of hashing and encryption mechanisms, used to guarantee the authenticity, integrity, and irritability of information. (Ye et al., 2015)
- **Access control**: It deals with the limitation of access to information and must be implemented considering the “need to know” and the “need for access”. The standard recommends that access permissions be approved by the information officer. Also, the “profile” feature can be adopted to authorize not only access but also individual actions or those of a group of users (Hu, Vincent & Grance et al 2014).
- **Backup**: They are backup copies that guarantee the recovery of information in case of loss or unavailability of them in their original bases (Hu, Vincent & Grance et al 2014).
- **Digital Certificates**: They materialize the use of the digital signature and enable the use of cryptography, being issued by certifying authorities that certify that the information used in its generation is true and valid for a certain time. With the use of mathematical functions, it

is possible to obtain a guarantee of authenticity, irreversibility, integrity, and confidentiality (Tanwar et al., 2014).

- **Timestamp**: It guarantees the validity of digitally signed information over time. It is a seal that certifies the date and time that a document was digitally signed, ensuring that it was not tampered with in the time interval between signing and consulting the document. This mechanism adds a temporal anchor to the electronic document so that some characteristics present in physical documents, such as identification of authorship and imperceptible alteration in the document, are also present in electronic documents to avoid possible legal challenges (Jain et al, 2016)

1.3 BIG DATA

The term *Big Data* refers to large volumes of data that have different characteristics, are heterogeneous and that originate from different sources. In a reality in which organizations are generating huge amounts of data, which requires a specific management process to guarantee its quality, *Big Data* solutions and practices are necessary when traditional technologies and techniques are not sufficient for execution of activities related to the management of large *datasets* (Tian, 2017). Another issue to be considered is the number of actors involved in the generation of data and information, since, with the advent of the internet, social networks and mobile devices, this number has increased considerably. Currently, this scenario has expanded even more with the progressive implementation of the Internet of Things, a proposal that assumes the interconnection of all "things", generating information about events and transactions carried out and captured by them. (Ahmed et al., 2018) reinforce this observation by citing that the massive adoption of cell phones by society and a large number of existing computers and information systems, generates an avalanche of information and complements that the speed with which this data is generated and accumulated data is an important factor in defining *Big Data*. Kashyap et al. (2018) states that while the volume factor receives all the attention, one of the most challenging aspects of Big Data is related to the lack of structure to deal with all this data. Environments with single servers, databases structured in rows and columns, and static repositories need to be adequate to store large amounts of data, whether structured or not, being of different types and formats, thus generating an intense and continuous flow.

According to Bhogal et al. (2015), there are different points of view in the literature regarding the characteristics that make up *Big Data*. What seems to be a consensus is that, for a data set to be considered as such, it must have at least one of these particularities? Three of them were initially identified by Doug (Matturdi et al., 2014). Some years later, other characteristics were added to the set of aspects related to B*ig Data* and they are: veracity, value, variability and visualization, where:

- **Veracity**: it is related to the quality and fidelity of the data, that is, to the degree of precision and reliability that the data has;
- **Value**: refers to the usefulness of the data and its importance within a given context.
- **Variability**: it is the change of meaning that the data undergoes over time;
- **Visualization**: refers to the effectiveness of the way data is presented.

1.4 VERACITY IN *BIG DATA*

The characteristic veracity in *Big Data* environments refers to the degree of credibility of the data, and they must have significant reliability to provide value and utility to the results generated from them. In Inukollu et al. (2014) conception, data need to be evaluated for veracity, objectivity, and credibility to guarantee the production of true, objective, and credible information.

Zhang et al. (2016) states that it is necessary to rethink traditional repository architectures so that they are prepared to receive and process large volumes of structured and unstructured data. Unstructured data by nature brings a significant amount of inaccuracy and uncertainty, similar to data originated from social media, which are inherently inaccurate. Levels of data inaccuracy and uncertainty can vary from one database to another and decisions need to be made on the data that has the highest level of truth.

Concerning the requirements to verify the veracity of the data, Claverie-Berge et al. (2012) proposes the inconsistency that refers to the divergent data on the same fact; the incompleteness that is related to the lack of essential data to reach a certain objective; the ambiguity regarding the possibility of different and mistaken interpretations of the data; the latency that is related to the time the data is collected until the moment when it generates some kind of result; and finally, the approximation models (algorithms) that consider the correlation between the data to treat them.

From Rubin (2013) point of view, the theme veracity in Big Data is related to the management of uncertainties. The authors present a proposal for the reduction of uncertainties regarding the content of textual data, associating computational linguistics tools that can be used to measure three dimensions of veracity: the objectivity that is related to the particular form of writing of those who generate the information; the veracity that refers to the degree of truth in the information; and the credibility that concerns how much of the information is credible. It is possible to observe a scarcity of studies on the requirements of the veracity of the data, with different approaches to deal with the subject. Some with similar understandings and others complimentary, however, all with the same objective - obtaining greater accuracy and credibility of the data.

2. PROCEDURE

The methodological procedures has a qualitative approach since it is not concerned with numerical representativeness, but with deepening the understanding of a social group, an organization, etc. As for nature, it is basic research because it aims to generate new knowledge, useful for the advancement of science. As for the objectives, it is exploratory research because it aims at a better understanding of the problems, having the purpose of promoting greater familiarity with the themes to make them more explicit and also to help in the construction of new ones.

This study is classified as bibliographic research, elaborated from surveys of theoretical references already analyzed and published in written and electronic media, such as books, scientific articles, and websites. To verify the originality and possible contribution of this study to the area of Information Science, this study considers main databases of *WoS*, IEEE, Scopus, For the terms "*Big Data*" *and Veracity* more than 26 publications were found, for the terms "*Big Data* " *and* "*Information Security*", more than 5 publications and finally for the terms that contemplated the three subjects treated in this study, "*Big Data* " *and* "*Veracity*" *and* " *Information Security*", the result was no publication. Therefore, given the figures presented, it is noted that studies in the *Big Data* area are still incipient from the perspective of veracity related to information security.

3. BIG DATA TREATMENT

Currently, the techniques and technologies for the capture, rubbing, analysis, and visualization of Big Data are classified into three categories of processing tools: i) batch; ii) by flows (streaming);

and iii) interactive analysis (Chen et al., 2014). These are based on the Apache Hadoop infrastructure, such as Mahout and Dryad. For streaming processing tools Storm or S4 platforms are usually used, while for interactive analysis the platforms are usually Google's Dremel and Apache Drill. Data analytics refers to any mathematical or scientific method that allows obtaining a new vision of the data or discovering patterns of that data (Dietrich, 2014). Big data and data analytics have been used to describe the relationship between data sets and data analysis techniques for applications that require terabyte or exabyte data handling and where the technology to store, manage, and analyze data is essential. The Big data analytics architecture, first, the data is acquired from various sources, internal to the same organization or external to other organizations; They are then curated (prepared) in the format of Excel, XML, CSV files, and other unstructured formats such as text files, emails, web pages, twits. Subsequently, a series of technological tools are applied to manipulate this data, and finally, data analytics is applied, in the form of queries, reports, data mining techniques, statistical techniques. The results of the analysis are presented graphically using different types of diagrams, trees, pie diagrams, lines, 3D, etc.

3.1 BIG DATA TECHNOLOGIES

The listed used techniques are inexhaustive, considering growing developments. There is a wide variety of techniques and technologies to add, manipulate, analyze and visualize Big Data developed from different fields e.g. statistics, computer science, applied mathematics and economics, from flexible and multidisciplinary approachs. EU Commission (2013) produced the list of techniques, technologies, and tools, Table 1. Figure 1 shows Big Data technologies to be reviewed in this article.

Table 1: Different types of Techniques, Technologies, and Big Data Visualization Tools. (after EU Commission (2013)).

Techniques	Technologies	Visualization
• A/B Testing • Association rule learning • Data Mining • Genetic algorithms • Neural Networks	• Cassandra • Cloud computing • Extract, transform and load • Hadoop • SQL • NoSQL	• Tag clouds • Conversation clouds • Cluster graphics • Historical flows • Spatial information flows

Figure 1: Big Data Technologies

4. RESULTS

Techniques: The techniques listed here are some of those used, so this list is not intended to be exhaustive, bearing in mind that developments in this area are constantly growing.

- **A/B testing:** a control group is compared to a variety of test groups to determine what is the best procedure to achieve a particular objective. Big Data allows several tests to be carried out, ensuring large enough groups.
- **Association rule learning**: set of techniques to discover relationships between variables in large databases
- **Data mining**: Techniques for identifying patterns from large datasets, incorporating computational and machine learning approaches. These approaches include studying the law of association, the study of the groups, grouping, and regression.
- **Genetic algorithms**: Technique of optimization inspired by the mechanisms of evolution (the survival of the fittest). These are commonly called "evolutionary algorithms."
- **Machine Learning:** Related to the design and creation of algorithms that allow computers to learn empirical data-driven behaviors. NLP (Natural Language Processing), is a machine learning example.
- **Neural Networks:** Computational models inspired by the framework of biological neural networks (e.g. cells and brain connections) to scan and refine data set patterns

Technologies

- **Cassandra:** open source database administrator for the treatment of large amounts of data in a distributed system. It is an Apache Software project.
- **Cloud computing:** (or computing cloud) Computing model in which highly scalable computing services are delivered as a service over a network, often implemented as a distributed system.
- **Extract, transform, and load (ETL):** Computer methods used to retrieve data from external databases, convert it and load it into a database to fulfill organizational needs.
- **Hadoop:** Open source program framework for the management of large distributed datasets. Its architecture was inspired by Google's MapReduce and Google File System, originally built on Yahoo!, and currently operated as a project of the Apache Software Foundation.
- **HBase:** Open, distributed and incompatible table, using Google's Big Table as a platform. It was originally developed by Powerset and is now operated as part of Hadoop by the Apache Software Foundation.
- **MapReduce:** Google-inspired programming model for processing large databases on distributed systems. It was also implemented by Hadoop. In Fig. 3.2 the outline of a MapReduce process is presented and then, a description of each phase involved is made.

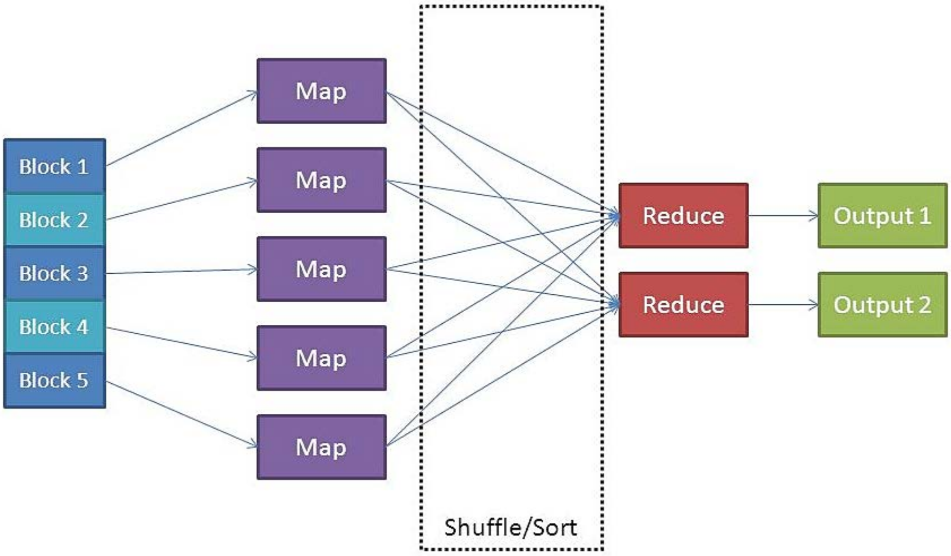

Figure 2: General diagram of the MapReduce processes

- **SQL:** (structured query language), a computer language designed to manage relational databases, allowing to specify different types of operations

Display

The visualization tools allow to communicate, understand and improve the analysis of the results obtained from Big Data, efficiently.

- **Tag Cloud:** or weighted text cloud, where the most frequently used words appear larger than the least frequent.
- **Clustergram:** visualization technique used to show how members of a database are assigned to a cluster as the number of clusters increases.
- **History Flow:** simulation methodology that maps a document's history as it is written by many contributing authors.
- **Spatial information flow:** demonstrates how data flows spatially, that is, from one area to another, country, place, etc.

5. DISCUSSION

To select the requirements related to the veracity of the data to be part of this study, an analysis of the requirements presented in the session that deals with the veracity in Big Data was carried out according to Arnaboldi (2017) and Gill (2015). The analysis considered the complementarity, similarities and equivalences existing between them. From this analysis, the requirements were compiled and selected to form the set of requirements to be worked on in this article. Table 2 presents the selection of the requirements obtained from the analysis from the literature.

Table 2: Selection of veracity requirements presented in the literature.

Researcher	Selected Requirements	Excluded Requirements
Arnaboldi (2017)	inconsistency, incompleteness, ambiguity, latency and approximation models	-
Gill (2015)	Reliability	-
Wamba et al (2015)	Veracity	objectivity and credibility

This study use the full requirements proposed by Claverie-Berge (2012). The selection of the veracity requirement proposed by Rubin and Lukoianova (2014) is justified because it is understood in the presented context, it is complementary to the reliability requirement Sordi (2008).

Thus, seven requirements related to the veracity of the data are presented. Based on the analyzes carried out in this work, in addition to these, it was decided to include two other requirements understood as important to meet the objectives of this study, they are:

Legality: Legislation applied in in the country is important in the context of *Big Data*, both for access, as for the use and availability of data. This requirement was included because it is understood that the legislation relevant to each situation, applicable to the locations to which the data is being accessed, used and made available, needs to be met in order to transmit trust to *Big Data* users also from a legal perspective.

Privacy: refers to the protection of private data, referring to its use, storage and availability. It is important to consider what data can be obtained and maintained so that adequate privacy for each type is offered and thus provide users of these environments with the confidence that private data will be treated as such, from its acquisition to its availability. It is also important to consider that data published in different media can be viewed publicly and be used to generate erroneous conclusions, as they are isolated and out of context data. However, the privacy requirement from this point of view is not the scope of this study.

Therefore, the requirements regarding the veracity of data in *Big Data* environments to be analyzed in this work totaled nine items, as defined: inconsistency, incompleteness, ambiguity, latency, approximation models, reliability, veracity, legality and privacy.

The results of the analysis on each of the requirements will be presented below, the security mechanisms that can contribute to their treatment and the justifications for their proposals.

The proposition of security mechanisms for the inconsistency requirement presented in Table 3, considers two aspects, the inconsistency generated in the data source (in the case of unreliable sources), where the proposed mechanism is the use of a digital certificate issued by a Public Key Infrastructure (PKI) of international accreditation in the data sources, for that its authenticity is proven. This mechanism attests that the basis is not false, as it has been certified by an authorized and recognized digital certificate regulator and the inconsistency in the storage base of the collected data, where the proposition is the adoption of mechanisms for validating the input data in the relevant fields of the system. This proposal considers the possibility of a security breach that allows inconsistent data (which does not correspond to reality) to be recorded in the systems.

Table 3: The proposition of security mechanisms (Source: author).

Security mechanisms	Requirement	Mechanisms
Information security mechanisms for the inconsistency requirement	Inconsistency	-Digital certificate in the data sources
		-Validation mechanisms of input data in the relevant fields, in the storage system
	Incompleteness	- Use of hashing in data traffic.
		- Use of hashing in data storage.
		- Authentication for access to the collected data base.
		- Adoption of the data access profile feature.
		- Adoption of data replication systems.
Information security mechanisms for the ambiguity requirement	Ambiguity	- Address the ambiguity of the data, no technological mechanisms of information security were directly identified.
Information security mechanisms for the latency requirement	Latency	- Digital signature.
		- Timestamp.
Information security mechanisms for the approach models requirement	Approach models	- Deal with the approximation models, no directly associated information security mechanisms were identified.
Information security mechanisms for the reliability requirement	Reliability	- Digital certificate in data sources.
		- The digital signature of data, at the source.
Information security mechanisms for the veracity requirement	Veracity	- Digital certificate in data sources.
		- The digital signature of data, at the source.
Information security mechanisms for the legality requirement	Legality	- Authentication for access to the database.
		- Adoption of the data access profile feature.
Information security mechanisms for the privacy requirement	Privacy	- Traffic encryption and data storage.
		- Authentication for access to the database.

Table 3, the security mechanisms for the incompleteness requirement refer to the adoption of a traffic *hashing* mechanism that allows it to be verified whether the data received are the same extracted from the source, that is, it checks whether the data received is faithful to the originals (source). The completeness of the data must also be maintained for as long as the data remains stored. In this way, it is possible to identify the existence of changes in the data and the accesses are segmented according to the needs - with the options "read-only" of the data or "total access". These mechanisms reduce the risk of data deletions and changes by unauthorized entities and, if they occur, that are identified.

The replication of the collected data can also be adapted to promote the "backup " of the data and guarantee its completeness, in case of storage servers unavailability. Besides, replication systems also use *hashing* mechanisms to ensure that data remains complete after replication and integration. It is worth mentioning that no security mechanisms were identified to deal with incompleteness when the data is generated incomplete and if there are flaws in the data collection mechanism. The mechanisms presented here aim to ensure completeness during data traffic and storage.

Ambiguity refers to the imprecision of data in a way that generates different and misinterpretations. Even though these are not security mechanisms, it is suggested that consistent metadata, semantic techniques be adopted, such as the use of ontologies and data dictionaries to minimize the chances of resulting in false interpretations.

The mechanisms of digital signature and time stamp can be used to contribute to the latency requirement since together they allow to verify that even the data being used after some time after its collection, they were authentic and valid when they were digitally signed in its generation (in this context generation is understood as the moment when the data was collected and stored for later analysis). It is worth mentioning that the mechanisms proposed for this requirement do not attest to the obsolescence of the data, but confirm its validity in the generation, after a while.

Approximation models are applied in algorithms to treat the data and identify the correlation between them. In this case, the recommendation is that the algorithms adopt, as far as possible, the relevant security mechanisms, presented for the other requirements.

The data needs to be authentic and thus be perceived by its users to be reliable. As presented in Table 3 it is proposed that the security mechanisms for the reliability requirement are, the adoption of data sources that have a digital certificate issued by PKI with international accreditation, since this mechanism certifies that the source is not false, and the use of digital signatures to guarantee authenticity and legitimacy of data at source.

For the veracity requirement, the proposition is the use of a digital certificate and the digital signature in the source data. If in the reliability of the data, the perception is generated by the trust in the source that generated the data, the veracity also considers this factor, that data generated by suitable sources have a high degree of truth in its content. Therefore, the proposed mechanisms corroborate this, as they provide the authenticity and legitimacy of the data at the source.

Table 3 shows the proposition of the security mechanisms for the legality requirement, where the protection of the data in its availability after the analysis was considered. These should be made available only to authorized entities considering the current legislation and according to the type of data. The suggested security mechanism is authentication with the use of a profile to access the databases, limiting only those entities authorized by law. It is worth mentioning that as important as the context presented, it is to observe the legality of the data under other aspects, such as legality in the data collection, so that it is only from open and/or authorized sources, the use of data protected by intellectual property and other controls such as digital curation to obtain a more complete treatment for this requirement, however for these cases.

For data to have their privacy guaranteed, the proposal is the adoption of encryption, both in traffic and in its storage, in this way it is possible to ensure that only authorized entities that have the key to decrypt the data, have access to them. Also, the adoption of authentication to access the database. However, it is worth mentioning that the mechanisms presented do not guarantee that privacy will be respected in its entirety, as they do not ensure compliance with privacy in the collection, use and availability of this data. The item must be complemented with the adoption of ethical procedures and processes for using the data by the law.

Given the presented, it is understood that to ensure the veracity of data in *Big Data* environments, they must be consistent and complete, ambiguity, latency and approximation models need to be addressed and the data must be reliable, and factors related to their legality and privacy must be respected and maintained.

6. CONCLUSION

The term *Big Data* is approached in the literature from different perspectives and there is no single concept in relation to it, however, the main idea can be defined as environments that involve the use of large amounts of data to make decisions more precisely. In general, seven factors characterize *Big Data*, however, the expectation for the use of this data is that it is based on accurate and reliable data. This condition makes the characteristic veracity in *Big Data* an indispensable factor to obtain value and the results are in line with expectations.

The requirements for veracity and information security mechanisms were presented in this study in a broad and general way. However, *Big Data* environments are used for different purposes. Therefore, the applicability of these mechanisms, to a lesser or greater degree, must be assessed considering the specificities and needs of each scenario.

In this study, whose objective was to contemplate an analysis of the technological mechanisms of information security that contribute to the veracity of data in *Big Data* environments, the factors that characterize an environment such as *Big Data* were presented. Emphasizing the veracity factor and studying the technological mechanisms of information security that could be related to this factor. The main requirements related to the veracity of the data were also identified and those that would be part of the analysis and that would be linked to the information security mechanisms were selected. As a contribution of this work, the suggestion of including two veracity requirements proposed by the authors of this article, in addition to the requirements identified based on the literature, is highlighted: they are legality and privacy.

The mechanisms: encryption, access control, *hashing*, *backup*, data replication, digital certificate, digital signature, and time stamp can contribute to the requirements of data veracity in *Big Data* environments. For the ambiguity requirements and approximation models, no directly associated information security mechanisms were identified. It is recommended to address ambiguity, actions in the scope of people and processes and for approximation models, when possible and applicable, to implement the security mechanisms presented for the other veracity requirements. It is also concluded that the technological mechanisms presented in this work do not address the requirements of veracity in their entirety but can contribute to this end. For a complete solution, non-technological security mechanisms must also be considered, that is, those related to people and processes.

This work does not intend to exhaust the issues related to information security mechanisms and the requirements for the veracity of data in *Big Data* environments. Mechanisms such as authorizations for data changes, evaluation of the methods used for data hygiene, methods used in data collection, the *expertise* of the data analyst and confidence in the supplier that stores the data are some of the points to be considered for a more complete treatment of the data.

7. DATA AND MATERIAL AVAILABILITY

Data can be provided by contacting the corresponding author.

8. REFERENCES

Aalst, W. M. P. (2014). Data Scientist: The Engineer of the Future. In Enterprise Interoperability VI(7), K. Mertins, F. Bénaben, R. Poler, and J. P. Bourrières, Eds. Springer, 13-26.

Ahmed, E. et al. (2018). Recent Advances and Challenges in Mobile Big Data. IEEE Communications. 56. DOI: 10.1109/MCOM.2018.1700294.

Arnaboldi, M.; Busco, C.; Cuganesan, S. (2017). Accounting, accountability, social media and big data: Revolution or hype. Account. Audit. Account. J. 30, 762–776

Bahga, A., & Madisetti, V. K. (2011). Analyzing massive machine maintenance data in a computing cloud. IEEE Transactions on Parallel and Distributed Systems, 23(10), 1831-1843.

Bhadani, A., Jothimani, D. (2016). Big data: Challenges, opportunities and realities, In Singh, M.K., & Kumar, D.G. (Eds.), Effective Big Data Management and Opportunities for Implementation (1-24).

Bhogal, N. & Jain, S. (2017). A review on big data security and handling. 6.

DeCandia, G., et al. (2007). Dynamo: amazon’s highly available key-value store. SOSP, 7, 205–220.

Deka, G. C. (2013). A survey of cloud database systems. IT Professional, 16(2), 50-57.

EU Commission (2013). “Big Data. Analytics & Decision Making”. Business Innovation Observatory.

Feng, D. G., Zhang, M., & Li, H. (2014). Trusted Computing and Information Assurance Laboratory, Institute of Software, Chinese Academy of Sciences. Big Data Security and Privacy Protection. 37(1), 246-258.

Gill, J., & Singh, S. (2015). Enormous Possibilities in Big Data: Trends and Applications. Asian Journal of Computer Science and Technology, 4(2), 23-26.

Harfoushi, O., Obiedat, R. (2018). Security in Cloud Computing Using Hash Algorithm: A Neural Cloud Data Security Model. Modern Applied Science. DOI: 10.5539/mas.v12n6p143.

Hashizume K. Rosado D. G. Fernández-Medina E. Fernandez E. B. (2013). An analysis of security issues for cloud computing. Journal of Internet Services and Applications, 4(1), 1–13. 10.1186/1869-0238-4-5

Hu, V. et al. (2014). An Access Control Scheme for Big Data Processing. DOI: 10.4108/icst.collaboratecom.2014.257649.

Inukollu, V. N., Arsi, S., & Ravuri, S. R. (2014). Security issues associated with big data in cloud computing. International Journal of Network Security & Its Applications, 6(3), 45-56.

Jain, P. et al. (2016). Big data privacy: a technological perspective and review. Journal of Big Data. 3. DOI: 10.1186/s40537-016-0059-y.

Ji, Y., Y. Tian, F. Shen, and J. Tran. (2016). Experimental Evaluations of MapReduce in Biomedical Text Mining. Information Technology: New Generations, Springer. 665-675

K.C. Li, H. Jiang, L. T. Yang, and A. Cuzzocrea, Big Data: Algorithms, Analytics, and Applications, Chapman &. CRC Press, 2015.

Kashyap, Ramgopal & Piersson, A.. (2018). Impact of Big Data on Security (Chapter 15). 10.4018/978-1-5225-4100-4.ch015.

Klein, M., et al. (2017). Biospark: scalable analysis of large numerical datasets from biological simulations and experiments using Hadoop and Spark. Bioinformatics, 33(2), 303-305.

L. Douglas, "3D data management: Controlling data volume, velocity and variety," Gartner, Retrieved 6 (2001).

LiYi,Keke Gai,Longfei Qiu, Meikang Qiu, ZhaoHuid, Intelligent cryptography approach for secure distributed big data storage in cloud computing, Information Sciences, Volume 387, May 2017, Pages 103-115.

Luaay Abdulwahed Shihab. (2012). Wireless LAN Security and Management, International Journal of Engineering and Advanced Technology (IJEAT), 2(1).

Manogaran, G., Thota, C., & Kumar, M. V. (2016). MetaCloudDataStorage architecture for big data security in cloud computing. Procedia Computer Science, 87, 128-133.

Marchal, S., Jiang, X., State, R., & Engel, T. (2014). A Big Data Architecture for Large Scale Security Monitoring. In IEEE Big Data Congress, 56-63. DOI: IEEE. 10.1109/BigData.Congress.2014.18

Matturdi, Bardi & Zhou, Xianwei & Li, Shuai & Lin, Fuhong. (2014). Big Data security and privacy: A review. China Communications. 11. DOI: 135-145. 10.1109/CC.2014.7085614.

Mohanty, H., Bhuyan, P., & Chenthati, D. (Eds.). (2015). Big data: A primer. 11. Springer.

Oussous, A., et al. Big Data technologies: A survey. Journal of King Saud University – Computer and Information Sciences (2017), http://dx.doi.org/10.1016/j.jksuci.2017.06.001

Rubin, V., & Lukoianova, T. (2014). Veracity roadmap: Is big data objective, truthful and credible? Advances in Classification Research Online, 24(1), 4-15.

Singh, S. and N. Ahuja, “Article recommendation system based on keyword using map-reduce,”in 2015 Third International Conference on Image Information Processing (ICIIP), 2015, pp. 548-550.

Sabahi, F. (2011). Virtualization-level security in cloud computing. In Communication Software and Networks, IEEE 3rd International Conference. DOI: IEEE. 10.1109/ICCSN.2011.6014716

Singh, S., & Ahuja, N. (2015). Article recommendation system based on keyword using map-reduce. In IEEE Third International Conference on Image Information Processing (ICIIP) (548-550)..

Tanwar, S., Prema, V. (2014). Role of Public Key Infrastructure in Big Data Security. CSI Communications, 45-48

Tian, Y. (2017). Towards the Development of Best Data Security for Big Data. Communications and Network. 09. 291-301. DOI: 10.4236/cn.2017.94020.

Wamba, S.F., Akter, S., Edwards, A., Chopin, G., Gnanzou, D. (2016). How ‘big data’ can make big impact: Findings from a systematic review and a longitudinal case study. Int. J. Prod. Econ.165, 234-246

Ye, F., Y. Qian, R.Q. Hu. (2015). An Identity-Based Security Scheme for a Big Data Driven Cloud Computing Framework in Smart Grid. 2015 IEEE Global Communications Conference (GLOBECOM), San Diego, CA, 1-6.

Zhang, Yingfeng & Ren, Shan & Liu, Yang & Si, Shubin. (2016). A big data analytics architecture for cleaner manufacturing and maintenance processes of complex products. Journal of Cleaner Production. DOI: 10.1016/j.jclepro.2016.07.123.

Zikopoulos, P., Eaton, C., et al., 2011. Understanding Big Data: Analytics for Enterprise Class Hadoop and Streaming Data. McGraw-Hill.

Luay Abdulwahid Shihab is a Lecturer at Department of Basic Medical Science, Nursing College, University of Basra, Iraq. He is a PhD student at UTM, Malaysia. He got an MSc in Computer Science (Network) from College of Skobad, Agra University, India. His research is focused on Wireless Networks, Image Processing and Communication Security.

Trademarks Disclaimer: All product names including trademarks™ or registered® trademarks mentioned in this article are the property of their respective owners, using for identification and educational purposes only. The use of them does not imply any endorsement or affiliation.