Journal of Ubiquitous Systems & Pervasive Networks
Volume 12, No. 1 (2020) pp. 09-15

Quality and Security in Big Data: Challenges as opportunities to build a powerful wrap-up solution

Mohamed TALHA*, Nabil ELMARZOUQI, Anas ABOU EL KALAM

ENSA UCA, Marrakesh, Morocco, 40000

Abstract

Quality and Security are two major issues in Big Data that pose many challenges. High volume, heterogeneity and high speed of data generation and processing are, amongst others, common challenges that must be addressed before setting up any data quality management system or data security system. This document provides an overview of data quality and data security in a Big Data context and highlights the conflicts that may exist during the implementation of these systems. Such a conflict makes the setting up of such systems even more complex and the reflection into new solutions becomes a major prerequisite. In this paper, we consider these challenges to present a global solution to evaluate the quality of data without impacting data security and without it becoming a barrier.

Keywords: *Big Data, Data Quality, Data Security, Accuracy Assessment, Record Linkage, Big Data Sampling.*

1. Introduction

Big Data, as shown in Fig. 1, is often characterized by the term 5V: **Volume** refers to the huge amounts of data generated every second; **Velocity** refers to the speed at which data is produced and processed; **Variety** refers to the heterogeneity of data and their sources; **Veracity** refers to the consistency and reliability of the data; and finally, **Value** refers to the profits that can be made from these large volumes of data.

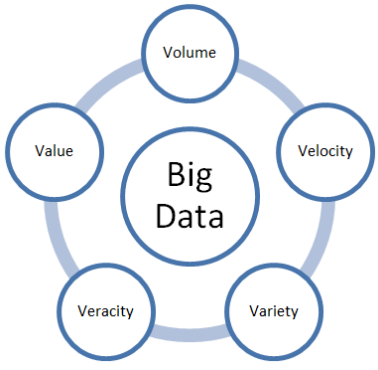

Fig. 1. Big Data characteristics

Several works consider **Complexity** as the 6th characteristic of Big Data [1], [2], [3]. Complexity measures the degree of interconnection and interdependence of data structures so that a small change or a combination of small changes, in one or more elements can produce very large changes that reverberate through the system and affect considerably its behavior.

In a Big Data environment, information systems are involved in complex information exchanges and often operate from heterogeneous and unstructured data collected from external sources. As a consequence, the overall quality of the data that flows across information systems can rapidly degrade over time if the quality of both processes and information inputs is not controlled [4]. Moreover, the continuous collection of large amounts of data, the diversity of data sources, the processing of data "on the move", etc. all play a role in creating security vulnerabilities. We deduce that the constraints imposed by the Big Data context pose many challenges for both the quality and security of data. Most of these issues have been addressed in literature, from studies that focus on data quality to those that deal with data security. However, very few articles have reported that data security can be a barrier to data quality or vice versa. When we consider the conflicts that may exist between the two systems, the complexity becomes even greater and, consequently, we must think of new adapted solutions. These challenges are therefore a good opportunity to reflect on new research themes in which the data quality management and data security systems are mutually reinforcing.

The rest of this paper is organized as follows: Section 2 describes data quality as well as the process of implementing a Data Quality Management System. Section 3 is dedicated to Data Security and the challenges imposed by the Big Data context. Section 4 highlights the conflict between Data Quality and Data Security Systems. In Section 5 we present a solution for assessing data quality without impacting security. Finally, Section 6 presents our conclusion and plan for future work.

* Corresponding author. Tel.: +33613670360
E-mail: mohamed.talha@icloud.com

DOI: 10.5383/JUSPN.12.01.002

2. Big Data Quality

Many research and industry reports are clear in indicating the severe damage caused by the presence of poor data quality in diverse contexts and at many different levels [5]. Data of poor quality can lead to incorrect findings and can significantly undermine a range of decisions and policy-making processes [6]. The costs of poor data quality can create loss of opportunity, loss of revenue, re-execution of processes due to data errors, quality improvement costs, and so on. There may be different reasons for poor data quality such as data entry errors, using faulty methods to collect data, failure to update data that may change over time, misapplying business rules, duplicate records, missing or incorrect data values, etc. Data quality is an essential topic for businesses, providing accurate information in order to make correct decisions accordingly. Crosby [7] defines quality by compliance with requirements. This definition has been taken over by ISO/IEC 25012 standard [8] which links data quality to the degree to which a set of data characteristics meets the requirements. In this section, we will present some fundamental concepts and notions related to data quality in general and then trace the main challenges that we encounter in a big data context.

2.1. Data Quality Management System

Improving data quality focuses on evaluating direct and indirect costs of data quality as well as identifying strategies and techniques to achieve the desired quality objectives. In [4], the authors identify two main types of quality improvement strategies. The first type of strategy directly involves changing the data values. For example, deleting duplicates, updating obsolete data values, correcting incorrect values, etc. Several approaches can be classified in this type of strategy such as acquisition of new data, standardization of values, record linkage, integration of data and schemas, data cleaning, etc. The second type of strategy consists in redefining processes that create or modify data. For example, adding a data format check-up before storing, adding a validation step for data source reliability, etc.

Data Quality is characterized by a set of characteristics called "data quality dimensions". Depending on the context of each project, the quality of data can be analyzed from one or more dimensions. Wang and Strong [9] define a "data quality dimension" as a set of data quality attributes that represent a single aspect or construct of data quality. It's a measurable property that represents some data characteristics. In literature, even if there is no general agreement on all the properties defining the quality of data or the exact meaning of each property [4], many quality dimensions are studied such as accuracy, completeness, consistency and timeliness. The standard ISO / IEC 25012 [8] provides a definition for the most discussed dimensions and proposes a Data Quality Model from three points of view:

- **Internal Data Quality**: it refers to the capability of a set of static data attributes to satisfy stated and implied needs when the data are used under specified conditions. These characteristics refer to data itself and provide the criteria to ensure and verify the quality of data values, data type and length, data definitions including metadata, data rules and relationships between data. An example of an internal Data Quality characteristic is Consistency which refers to data independently of hardware and software aspects.
- **External Data Quality**: it refers to the capability of data to satisfy stated and implied needs when data are used under specified conditions within a computer system. External Data Quality characteristics are "inherited" by data from computer systems' capabilities that can be implemented on such data depending on user requirements. An example of an External Data Quality characteristic is Security since it depends on hardware and software capabilities.
- **Data Quality in Use**: it refers to the capability of the data to enable specific users to achieve specific goals with effectiveness, productivity, safety and satisfaction in specific usage contexts. Data Quality in usage contexts aims to define the data quality characteristics that express the user's subjective point of view about data they are working with in terms of how such data satisfies his/her information needs. An example of an External Data Quality characteristic is Credibility which represents the extent to which data satisfy users' needs and is regarded as true by them.

To measure a dimension, one or several metrics can be associated with it. A quality metric defines how to evaluate a dimension. It can be objective, when it's based on quantitative measures (e.g. the result of a condition, a mathematical equation, an aggregation formula, etc.) or subjective, when it's based on qualitative evaluations such as perceptions, needs and experiences of stakeholders (e.g. a feedback questionnaire, user surveys, etc.) [4], [10], [11]. According to the type of information used as a quality indicator, Bizer in [12] categorizes quality evaluation metrics into three categories. A metric can be:

- **Content-based**: the information itself is used as a quality indicator.
- **Context-based**: metadata are used as the quality indicator of circumstances in which the information was created or used.
- **Rating-based**: the metric relies on explicit ratings for the information itself, information sources or information providers.

2.2. Big Data Quality Challenges

Although there are different models for evaluating data quality in a traditional context, none of these models are suitable for Big Data environments [13]. We have identified four major challenges for any data quality management system:

- **High Volume**: The volume of data is enormous and it's difficult to evaluate and improve the data quality within a reasonable time [14]. In addition, the constant increase in the volume of data requires the implementation of a scalable solution. Scalability reflects the ability of data quality techniques to process, in a relevant way, increasingly larger and complex datasets [15].
- **Heterogeneity**: The data produced nowadays are, in most cases, semi-structured or unstructured. This type of data is more complex to process than structured data. Understanding the semantics and correlations between unstructured data is a difficult task [16]. Lastly, although the migration of structured data from a relational database to a non-relational database is possible, [17] as an example, the conversion of semi-structured data into structured data is difficult or impossible [1].
- **Data change**: Nowadays, data changes very quickly and can rapidly become obsolete. If organizations cannot collect the required data, up-to-date and in real-time, they may produce unnecessary or misleading conclusions, potentially leading to decision errors [14].
- **Data security**: On the one hand, a data quality management system involves read access to all data to

perform the data quality assessment process and write access to all data to build the process of improving their quality. On the other hand, a data security system aims to protect data from unauthorized read and write access. Thus, a potential conflict may arise where the data security system can make the quality management process slower and more complex. As confirmed by Strong et al. [18], Privacy and Confidentiality mechanisms can be barriers to data accessibility.

3. Big Data Security

Traditionally, Security is focused on **Confidentiality**, **Integrity**, and **Availability**. Confidentiality is intended to protect data from unauthorized access. Integrity is about protecting data from unauthorized changes. Availability deals with making data accessible to authorized entities and users. ISO/IEC 27001 [19] considers other properties that may be involved in data security namely authenticity, responsibility, non-repudiation and reliability. Furthermore, in the Big Data field, several studies are interested in **Privacy** in order to protect personal and sensitive information and designate it as one of the main security objectives. Privacy can be considered as a particularity of confidentiality that takes into account additional elements such as user consent management regarding their personal data, compliance with regulatory and legal obligations, etc. In this section, we will present a list of threats and risks impacting data security according to the Big Data process. We will then outline the main challenges that a security system faces.

3.1. Security risks in Big Data

A Big Data Process, as shown in Fig. 2, consists of collecting and storing large amounts and wide varieties of data sets, and extracting valuable information and/or knowledge by analyzing the data sets [20].

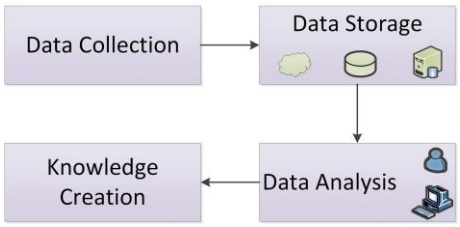

Fig. 2. Big Data Process

From a security point of view, each of these phases can present threats and security risks:

- **Data collection phase**: it's essential to collect data from reliable sources and to make sure that they are secure and protected against leakage (phishing, spamming, spoofing). Various measures can be applied to data collection, such as access control and encryption of sensitive data.
- **Data storage phase**: data collected in the previous phase must be stored and protected in order to ensure a secure environment for analysis. During this phase, data storage disks may be attacked (copying, stealing or damaging) or unauthorized access may occur to target and explore data carriers in order to extract useful knowledge. Data anonymization, partitioning and swapping are very useful techniques for protecting stored data.
- **Data analysis phase**: data analysis is performed to extract important information by applying Data Mining techniques. It is necessary to present a secure processing environment to prevent the data from being accessible to unauthorized entities that can analyze and explore it in order to extract relevant and/or personal information. Different techniques make it possible to deduce personal information such as re-identification and correlation.
- **Knowledge creation phase**: the information generated from the previous phase is very sensitive and their protection is mandatory. The security risks during this phase mainly relate to data leakage (phishing and spoofing) and the threat to the privacy of individuals. Adopting an effective access control strategy and encrypting the relevant results appear to be good ways to improve security.

3.2. Data Security Challenges

Big Data, with its diversity of data types between structured, semi-structured and unstructured, has brought many challenges to the security and privacy of individuals. The security challenges in terms of cryptography, log and event analysis, intrusion detection and prevention, and access control have taken a new dimension [21]. In this section, we will look at a set of challenges that can threaten data security and privacy in a Big Data context:

- **Confidentiality**: confidentiality involves setting up a set of rules and restrictions to limit access to confidential data. It is generally treated with access control and cryptographic mechanisms [21], [22]. The areas of research to improve the confidentiality of data in Big Data are concerned with issues such as merging and integrating of access control policies, automatic management of these policies, automatic administration of authorizations, application of access control on Big Data platforms, etc. [22].
- **Integrity**: when it comes to Security, integrity means preserving data against unauthorized changes. In Big Data, processing is usually distributed over several nodes. Integrity implies the consistency of data between different copies. It also implies that the data isn't altered or modified by unauthorized parties during transitions between nodes [23]. Integrity is generally impacted by hardware and software errors, human errors and intrusions [21]. To maintain data integrity, the challenges to be addressed include, in addition to the management of access authorizations, the assurance of the reliability of data and their sources, the establishment of mechanisms for detection and prevention of data loss, deduplication of data without impacting availability, etc.
- **Availability**: availability means that data must remain accessible to authorized users and entities. This refers to the prevention and recovery of hardware and software errors, human errors, and malicious access that can make data unavailable. Generally, the availability of data is satisfied by applying multiple data replications. However, replication can lead to data integrity issues [24] that rely on deduplication to ensure consistency across data. In addition, the availability of data can harm privacy by simplifying the combination and the analysis of information and the deduction of sensitive information on individuals [22].
- **Privacy**: confidentiality concerns any type of data, when it comes to personal information of individuals, it is called Privacy. It's about controlling the sharing of personally identifiable information (PII). This personal data cannot be shared without the informed consent of their owner. In

addition, the sharing of personal data is often regulated by privacy laws. The protection of personal data in Big Data requires particular processes such as transparency which consist of involving concerned users, obtaining the consent of users who may revoke it at any time, detection and prevention of re-identification processes, preservation of analytical results, compliance with regulatory and legal constraints, etc.

Cloud Security Alliance (CSA), a nonprofit organization that promotes best practices for improving cloud security, published in April 2013 a report in which it ranked the top ten security challenges according to four aspects of the Big Data ecosystem, as depicted in Fig. 3.

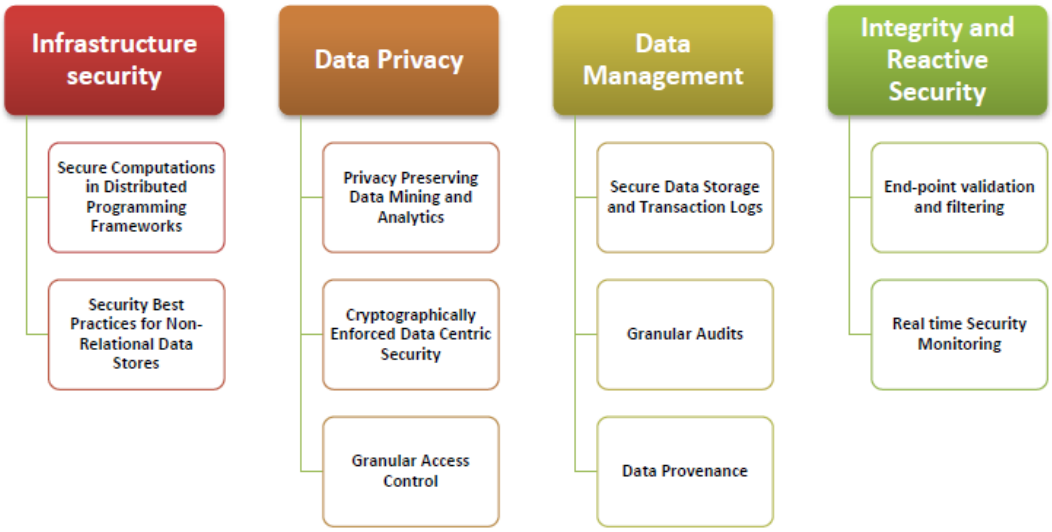

Fig. 3. Classification of the Top 10 security Challenges in the Big Data ecosystem [25]

To meet these challenges, CSA presented, three years later in June 2016, a report detailing the 100 best practices to consider [26]. Among these actions, we can quote:

- For companies working with distributed programming frameworks, like Hadoop, CSA recommends to use Kerberos authentication, or an equivalent, to create a trusted environment.
- To preserve privacy, all personally identifiable information such as name, address, gender, date of birth, contact information, etc. must be hidden or deleted.
- Companies dealing with large data sets may benefit by migrating from traditional relational databases to NoSQL databases which accommodate and process huge volumes of static and streaming data for predictive analytics or historical analysis. The authors in [17] propose an approach to migrate a relational database to a NoSQL database.
- NoSQL databases are deprived of advanced security features. In particular, the NoSQL DBMS do not offer the equivalent of the GRANT / REVOKE commands present in any relational DBMS making it possible to define access control policies [27]. In such a situation, CSA advocates the use of powerful encryption solutions such as Advanced Encryption (AES), RSA encryption, or Secure Hash Algorithm 2 (SHA-256).
- CSA also recommends using different storage spaces for code and encryption keys, as well as for data or repository. The encryption keys must be saved in an offline secure space.

4. Conflict between Data Quality and Data Security

Data Quality and Data Security are two main topics that oppose different challenges in Big Data such as high volumes and heterogeneity of data, credibility of data and their sources, the speed at which data is collected and processed, etc. In terms of Data Security, three main security properties are identified and are clearly defined: Confidentiality, Integrity and Availability (CIA). On the other hand, in terms of Data Quality, there is no general agreement on all properties defining the quality of data or the exact meaning of each property [4]. A framing for all characteristics defining both quality and security of data is then necessary. Furthermore, several properties are common to data quality and data security, but their meanings are different. For example, in data security, Integrity refers to the degree to which data is protected against unauthorized access, whereas in data quality there's no clear definition regarding data Integrity. Yet, several studies have attached Integrity of data quality to three properties; namely Accuracy, Completeness and Consistency [9]. Accuracy represents the degree to which a data value conforms to its real or specified value. Completeness is defined as the extent to which all necessary values have been assigned and stored in the computer system. Consistency refers to the absence of apparent contradictions in the data [8]. We therefore need to bring together definitions of all dimensions and determine points of convergence and divergence.

Moreover, the principle of data security, especially confidentiality and integrity, is to protect data against unauthorized access. However, implementing a data quality management system requires flexible read and write access to all data. This requirement can create many security problems because the data quality system can exchange data with other systems or be manipulated by different people of different profiles who do not necessarily have the same access rights. Thus, data security can be a barrier to data quality, and inversely, setting up a quality management system may require a level of tolerance that can create security vulnerabilities. This conflict between these two systems makes their implementation more complex and requires thinking about new access control policies adapted to the Big Data context and

enabling quality processes to access the required data without compromising their security. Such a policy can be achieved by implementing or extending a fine-grained access control model such as TBAC (Task Based Access Control), RBAC (Role Based Access Control), ABAC (Attribute Based Access Control), OrBAC (Organization Based Access Control), PuRBAC (Purpose-Aware Role-Based Access Control), etc.

In addition, several mechanisms and techniques of Security and Quality are incompatible. As an example, we can mention the deduplication and the encryption of data. The purpose of data deduplication, in addition to freeing storage space and the bandwidth of the network, is to remove duplicates to ensure data consistency. The purpose of encryption is to protect data against unauthorized access. However, as mentioned in [28], traditional encryption techniques are incompatible with data deduplication. Specifically, traditional encryption requires different users to encrypt their data with their own keys. Thus, identical data copies from different users will lead to different cipher-texts, making deduplication impossible. To overcome this conflict, we must establish well-adapted strategies. For our case, we must for example think of centralizing the secret keys within a dedicated entity that will allow the deduplication process to decrypt the data properly, implying that the data is encrypted only after verifying their uniqueness, etc.

Most Big Data research focuses on data quality or data security separately. However, the two subjects cause problems of convergence. In such circumstances, the strengthening of data security mechanisms at the expense of data quality processes or the adoption of certain security tolerances to improve data quality are two strategies that require vigilant arbitration. In Table 1, we list a series of points of convergence and divergence between data quality and data security, as well as a set of actions to resolve conflicts.

Table 1. Data Security vs Data Quality

	Data Security	Data Quality	Required actions
Big Data Challenges	Volume, Variety, Velocity, Veracity, etc.		Expose Big Data challenges to solve
Properties	- CIA - Clear definitions	No general agreement on all properties and their exact meanings.	A framing for all characteristics defining quality and security
Purpose conflicts	The purpose of data security is to protect data against unauthorized access.	Implementing a data quality management system requires flexible read and write access to all data.	Specific Access Control Model
Technical conflicts	Data encryption: protect data against unauthorized access	Data deduplication: remove duplicates to ensure data consistency	- Centralize secret keys - Data is encrypted only after verifying their uniqueness

5. Secure Data Accuracy Assessment Solution

Our methodology to meet the challenge is to manage the quality of the data through one or many properties clearly defined. Each of these properties must be analyzed from a security point of view in order to identify the potential risks of conflict. The resolution of these conflicts is to allow access to the data by the quality management system without compromising their security. In the following, we propose a model for assessing data quality through the "Accuracy" dimension.

Several studies have identified Accuracy as the key dimension of Data Quality [10], [29], [30], [31]. Accuracy is defined in literature as a measure of the proximity of a data value v to some other value v' that is considered correct [10], [29], [32], [33]. This definition seems applicable to structured and semi-structured data, as it can be compared with reference data representing the real world. However, for unstructured data, this definition does not seem adequate to all situations. Certainly, unstructured data may contain information that can be compared with the real world, such as an objective description of an object, a picture of something, a fact in the past, a mathematical equation, etc. The problem lies in situations where data cannot be compared with reality. For example, the information that only relates to the people who provided it, such as personal impressions, opinions on a subject, intentions, plans to do in the future, personal analysis around a topic, etc. This kind of information lacks credibility and cannot be compared with supposedly correct data. Unstructured data is therefore more complex to evaluate and cannot be evaluated in the same way as structured or semi-structured data. Assessing the accuracy of unstructured data will be more relevant if it focuses on metadata related to the data rather than the data itself. In [34], confirming our hypothesis, the authors present a solution to evaluate the quality of data collected from social networks by integrating a metadata management system in the Big Data life cycle.

In Fig. 4, we propose our model to develop a powerful solution to assess the accuracy of data in a Big Data context without conflict with security system.

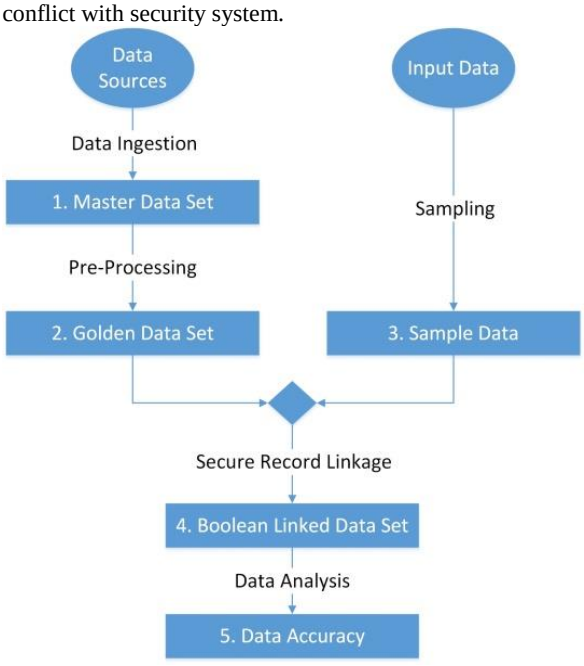

Fig. 4. Secure Data Accuracy Assessment Model

Our solution consists in 5 steps:

1. **Master Data Set**: we consider that the Data Lake in its raw state is our Master Data Set that contains all data collected from different data sources.
2. **Golden Data Set**: this data set is an enhancement of the Master Data Set to which a number of process are applied such as data cleaning, duplicate deletion, updating of

obsolete data, correction of incorrect values, etc. The goal is to improve the quality from the Master Data Set to get correct data in this Golden Data Set.

3. **Sample Data**: these data are obtained by sampling the data to assess. Sampling is a technique widely discussed in literature to handle large volumes of data and seems effective for our problem. When attempting to analyze a data set to assess its quality, we can be satisfied with the analysis of a representative sample of the entire data set. For some types of problems, sampling gives results as good as performing the same analysis using all the data [35], but for particular cases, especially the analysis of large volumes of data, sampling seems to be the most appropriate solution [30], [36], [37].
   To create a sample of a dataset, different techniques exist such as Simple Random Sampling, Stratified Sampling, Cluster Sampling, Multistage Sampling, Systematic Sampling, etc. Several techniques can be combined to create an effective sample. Whatever the technique used, all data units should have the same chance of being selected in the sample. To know the size of the sample, it will be necessary to know in advance the size of the data to be sampled which is not easy to obtain in a Big Data project. To solve this problem, there is an effective approach called "Reservoir Sampling" initially introduced by Vitter [38] in 1985. Reservoir Sampling is a family of randomized algorithms for random selection of a sample of k elements in a large data set of size n or in a data stream of size n, where n is unknown or difficult to know. The efficiency of this algorithm lies in its optimization of creating a sample from a large volume of data, without the need to know a priori its size. All while ensuring the same chance to all the data units of the set to sample.
4. **Boolean Linked Data Set**: this is the key step of our solution. It involves performing a record linkage between the sample of data to assess and the Master Data Set. The record linkage process must have read access to all Master Data Sets but, unlike a traditional record linkage process and to preserve data security, the results of our process will be in the form of a table containing for each field of each record a Boolean value:
   - **None**: the record is not coupled; for each field in the record, the process returns none.
   - **False**: the record is coupled with another one from the Master Data Set, but the value of the field is different from that of its correspondent.
   - **True**: the record is coupled with another one from the Master Data Set and the value of the field is similar to that of its correspondent. Similarity does not imply exact equality between values, but from a certain similarity threshold, we can consider that the values are equal.

   In this way, the process of evaluating data quality is not blocked by security locks since it is required that the record linkage process has free access to all data but without revealing the information since it returns a Boolean value based on a similarity calculation algorithm.
5. **Data Accuracy**: the last step is about analyzing the Boolean linked table from the previous step to deduce the overall accuracy.

Our solution meets the need of this article by allowing read access to all reference data. This privilege is sufficient to evaluate the quality of the data. Security, meanwhile, is preserved since we have introduced two layers of data protection:

- Comparing the data to be evaluated with the reference data returns a Boolean table containing three types of values: True (if the datum is correct), False (if the datum is not correct), or None (if the data to be evaluated does not match the reference data).
- The return of the results will be in the form of an average value representing the overall accuracy of all the data to be evaluated.

The security system can be further strengthened by restricting the return of results to entities with specific rights.

6. Conclusion

In this article we focused on the issues of quality and security of data on the same level. The challenges imposed by the context of Big Data cause problems for both of them. We assume that the provided solutions to solve the problems of high volume, heterogeneity and credibility of data will not only be used to set up quality management systems but also to develop security ones. We have emphasized the conflicts that may exist, making the implementation of these systems more complex and requiring reflection of new solutions. Finally, we presented a solution to evaluate data accuracy without impacting data security. Our solution at this stage remains theoretical and, to justify its feasibility and reliability, it requires the implementation of a set of processes such as the establishment of a Data Lake hosting correct data and setting-up of a record linkage process with read access to all the data in the Data Lake without compromising their security, etc.

References

[1] Katal A, Wazid M, Goudar RH. Big Data: Issues, Challenges, Tools and Good. IEEE, the 6th International Conference on Contemporary Computing (IC3) 2013 https://doi.org/10.1109/IC3.2013.6612229

[2] KAISLER S, ARMOUR F, ESPINOSA JA. Big Data: Issues and Challenges Moving Forward. IEEE, the 46th Hawaii International Conference on System Sciences 2013. https://doi.org/10.1109/HICSS.2013.645

[3] Gani A, Siddiqa A, Shamshirband S, Hanum F. A survey on indexing techniques for big data: taxonomy and performance evaluation. Springer-Verlag London 2015. https://doi.org/10.1007/s10115-015-0830-y

[4] Batini C, Cappiello C, Francalanci C, Maurino A. Methodologies for Data Quality Assessment and Improvement. ACM Computing Surveys 2009; 41 https://doi.org/10.1145/1541880.1541883

[5] Laranjeiro N, Soydemir SN, Bernardino J. A Survey on Data Quality: Classifying Poor Data. IEEE 21st Pacific Rim International Symposium on Dependable Computing 2015; 179-188 https://doi.org/10.1109/PRDC.2015.41

[6] Bertot JC, Choi H. Big Data and e-Government: Issues, Policies, and Recommendations. The Proceedings of the 14th Annual International Conference on Digital Government Research 2014 https://doi.org/10.1145/2479724.2479730

[7] Crosby PB. Quality is free. New York:McGraw-Hill 1979

[8] ISO, ISO/IEC 25012:2008—Software engineering. Software product quality requirements and evaluation

(SQuaRE). Data quality model, Report, International Organization for Standarization 2009

[9] Wang RY, Strong DM. Beyond Accuracy: What Data Quality Means to Data Consumers. Journal of Management Information Systems 1996; 12:5-33 https://doi.org/10.1080/07421222.1996.11518099

[10] Redman TC. Data's Credibility Problem. Harvard Business Review 2013

[11] Cappiello C, Francalanci C, Pernici B. Data quality assessment from the user's perspective. The 2004 international workshop on Information quality in information systems 2004; 68-73 https://doi.org/10.1145/1012453.1012465

[12] Bizer C. Quality-driven information filtering in the context of web-based information systems. Ph.D. Thesis. Freie Universit, Berlin 2007

[13] Merino J, Caballero I, Rivas B, Serrano M, Piattini M. A Data Quality in Use model for Big Data. Future Generation Computer Systems 2015 https://doi.org/10.1016/j.future.2015.11.024

[14] Cai L, Zhu Y. The Challenges of Data Quality and Data Quality Assessment in the Big Data Era. Data Science Journal 2015; 1-10 https://doi.org/10.5334/dsj-2015-002

[15] Chen M, Mao S, Liu Y. Big Data: A Survey. Springer Science+Business Media 2014 https://doi.org/10.1007/s11036-013-0489-0

[16] Barna S, Divesh S. Data Quality: The other Face of Big Data. IEEE, the 30th International Conference on Data Engineering 2014

[17] Sayeb Y, Ayari R, Naceur S, Ben Ghezala H. From Relational Database to Big Data: Converting Relational to graph database, MOOC database as example. Journal of Ubiquitous Systems & Pervasive Networks 2017; 8:15-20.

[18] Strong DM, Yang WL, Wang RY. Data uamity in context. Communications of the ACM 1997; 41:103-110 https://doi.org/10.1145/253769.253804

[19] ISO, ISO/IEC 27001:2013-Information technology -- Security techniques -- Information security management systems -- Requirements, International Organization for Standarization 2013

[20] Hakuta K, Sato H. Cryptographic Technology for Benefiting from Big Data. Springer, The Impact of Applications on Mathematics 2014;85-95 https://doi.org/10.1007/978-4-431-54907-9_6

[21] Sudarsam SD, Jetley RP, Ramaswamy S. Security and Privacy of Big Data. Big Data: A Primer 2015;121-136 https://doi.org/10.1007/978-81-322-2494-5_5

[22] Bertino E. Big Data – Security and Privacy. IEEE International Congress on Big Data 2015 https://doi.org/10.1109/BigDataCongress.2015.126

[23] Xu L, Shi W. Security Theories and Practices for Big Data. Springer International Publishing Switzerland 2016;157-192 https://doi.org/10.1007/978-3-319-27763-9_4

[24] Terzi DS, Terzi R, Sagiroglu S. A Survey on Security and Privacy Issues in Big Data. IEEE, The 10th International Conference for Internet Technology and Secured Transactions 2015 https://doi.org/10.1109/ICITST.2015.7412089

[25] Big Data Working Group. Expanded Top Ten Big Data Security and Privacy Challenges. CLOUD SECURITY ALLIANCE 2013.

[26] Big Data Working Group. Big Data Security and Privacy Handbook - 100 Best Practices in Big Data Security and Privacy. CLOUD SECURITY ALLIANCE 2016.

[27] GABILLON A. Contrôler les accès aux données numériques. Revue de l'Electricité et de l'Electronique 2013.

[28] Li J, Chen X, Li M, Li J, Lee PPC, Lou W. Secure Deduplication with Efficient and Reliable Convergent Key Management. IEEE TRANSACTIONS ON PARALLEL AND DISTRIBUTED SYSTEMS 2014;25 https://doi.org/10.1109/TPDS.2013.284

[29] Wand Y, Wang RY. Anchoring data quality dimensions in ontological foundations. ACM 1996; 39. https://doi.org/10.1145/240455.240479

[30] Motro A, Rakov I. Estimating the Quality of Databases. Springer-Verlag Berlin Heidelberg 1998. https://doi.org/10.1007/BFb0056011

[31] Redman TC. Measuring Data Accuracy: A Framework and Review. In: Information Quality. London and New York: Taylor & francis group 2005, (0-7656-1133-3); 21-36.

[32] Redman TC. Data Quality for the Information Age. Artech House 1996, (9780890068830).

[33] Scannapieco M, Missier P, Batini C. Data Quality at a Glance. Datenbank-Spektrum 2005.

[34] Immonen A, Pääkkönen P, Ovaska E. Evaluating the Quality of Social Media Data in Big Data Architecture. IEEE Access 2015; 3. https://doi.org/10.1109/ACCESS.2015.2490723

[35] Dean J. Big Data, Data Mining, and Machine Learning. Canada: John Wiley & Sons 2014. https://doi.org/10.1002/9781118691786

[36] DASU T, JOHNSON T. Exploratory Data Mining and Data Cleaning. Canada: John Wiley & Sons 2003. https://doi.org/10.1002/0471448354

[37] Prajapati V. Big Data Analytics with R and Hadoop. Birmingham: Packt Publishing 2013, (978-1-78216-328-2).

[38] VITTER JS. Random Sampling with a Reservoir. ACM Transactions on Mathematical Software 1985; 11:37-57. https://doi.org/10.1145/3147.3165