Quality and Security in Big Data: Challenges as opportunity to make a Powerful Conclude Clarification

Mr. Ashutosh Kumar

Department of Computer Science and Engineering, Institute of Technology, Gopeshwar,India

Abstract

Quality and Security are two critical issues in Big Data that present various challenges. High volume, heterogeneity and fast of data age and handling are, among others, ordinary challenges that must be would in general before setting up any information quality organization structure or information security system. This collection gives an summarize of data quality and data security in a Big Data setting and features the contentions that may exist during the usage of these frameworks. Such a contention causes the setting to up of such frameworks significantly progressively mind-boggling and the reflection into new arrangements turns into a significant essential. In this paper, we consider these challenges to introduce a global respond for assessing the quality of data without affecting data security and without it becoming a obstruction.

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

INTRODUCTION

Big Data, as shown in Fig. 1, is frequently characterized by the term 5V: **Volume** refers to the large 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 amounts of data.

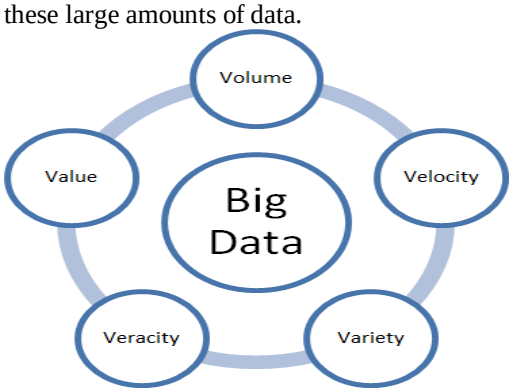

Fig 1. Big Data characteristics

Several works think about Complexity as the sixth trait of Big Data. Complexity gauges the level of interconnection and association of data structures so a little change or a blend of little changes, in one or In a Big Data condition, data frameworks are engaged with complex data trades and frequently work from heterogeneous and unstructured data gathered from outer sources. As an outcome, the general quality of the data that streams across data frameworks can quickly debase after some time if the quality of the two procedures and data inputs isn't controlled. Besides, the nonstop assortment of a lot of data, the decent variety of data sources, the preparing of data "progressing", and so on all assume a job in making security vulnerabilities.

We conclude that the imperatives forced by the Big Data setting present numerous difficulties for both the quality and security of data. The vast majority of these issues have been tended to in writing, from examines that attention on data quality to those that manage data security. Nonetheless, not very many articles have revealed that data security can be an obstruction to data quality or the other way around. At the point when we consider the contentions that may exist between the two frameworks, the complexity turns out to be much more prominent and, subsequently, we should consider new adjusted arrangements. These difficulties are in this manner a decent chance to think about new research topics wherein the data quality administration and data security frameworks are commonly fortifying.

I. Big Data Quality

Many research and industry reports are clear in demonstrating the extreme harm brought about by the nearness of poor data quality in various settings and at a wide range of levels. Data of low quality can prompt wrong discoveries and can essentially undermine a scope of choices and approach making forms. The expenses of poor data quality can make loss of chance, loss of income, re-execution of procedures because of data blunders, quality improvement costs, etc. There might be various explanations behind poor data quality, for example, data passage mistakes, utilizing broken strategies to

gather data, inability to refresh data that may change after some time, twisting business rules, copy records, absent or off base data values, and so on. Data quality is a fundamental subject for organizations, giving precise data so as to settle on right choices appropriately. Crosby characterizes quality by consistence with prerequisites. This definition has been taken over by ISO/IEC 25012 norms. which joins data quality to how much a lot of data attributes meets the prerequisites. In this segment, we will introduce some key ideas and thoughts identified with data quality all in all and afterward follow the primary difficulties that we experience in a big data setting.

II. Data Quality Management System

Improving data quality spotlights on assessing direct and indirect expenses of data quality just as recognizing methodologies and strategies to accomplish the ideal quality goals. In,the creators distinguish two primary kinds of quality improvement systems. The principal sort of technique straightforwardly includes changing the data values. For instance, erasing copies, refreshing out of date data values, rectifying mistaken values, and so on. A few methodologies can be ordered in this kind of system, for example, obtaining of new data, normalization of values, record linkage, combination of data and blueprints, data cleaning, and so forth. The second sort of procedure comprises in rethinking forms that make or change data. For instance, including a data position registration before putting away, including an approval step for data source unwavering quality, and so on. Data Quality is portrayed by a lot of attributes called "data quality measurements". Contingent upon the setting of each venture, the quality of data can be broke down from at least one measurements. Wang and Strong characterize a "data quality measurement" as a lot of data quality traits that speak to a solitary viewpoint or build of data quality. It's a quantifiable property that speaks to certain data qualities. In writing, regardless of whether there is no broad concurrence on all the properties characterizing the quality of data or the specific significance of every property, numerous quality measurements are concentrated, for example, accuracy, fulfillment, consistency and practicality. The standard ISO/IEC 25012 gives a definition to the most examined measurements and proposes a Data Quality Model from three perspectives:

1. Internal Data Quality: it alludes to the capacity of a lot of static data credits to fulfill expressed and inferred needs when the data are utilized under indicated conditions. These attributes allude to data itself and give the standards to guarantee and check the quality of data values, data type and length, data definitions including metadata, data rules and connections between data. A case of an inner Data Quality trademark is Consistency which alludes to data autonomously of equipment and programming angles.

***2. External Data Quality*:** it alludes to the capacity of data to fulfill expressed and inferred needs when data are utilized under indicated conditions inside a PC framework. Outer Data Quality attributes are "acquired" by data from PC frameworks' abilities that can be executed on such data relying upon client prerequisites. A case of an External Data Quality trademark is Security since it relies upon equipment and programming capacities.

3. Data Quality being used: it alludes to the ability of the data to empower explicit clients to accomplish explicit objectives with viability, profitability, security and fulfillment in explicit use settings. Data Quality in utilization settings means to characterize the data quality attributes that express the client's abstract perspective about data they are working with as far as how such data fulfills his/her data needs. A case of an External Data Quality trademark is Credibility which speaks to the degree to which data fulfill clients' needs and is viewed as obvious by them. To gauge a measurement, one or a few measurements can be related with it. A quality measurement characterizes how to assess a measurement. It tends to be objective, when it depends on quantitative measures (for example the aftereffect of a condition, a scientific condition, a total equation, and so forth.) or emotional, when it depends on subjective assessments, for example, observations, needs and encounters of partners (for example an input poll, client reviews, and so on.). As indicated by the kind of data utilized as a quality marker, Bizer in orders quality assessment measurements into three classes. A measurement can be:

- Content-based: the data itself is utilized as a quality marker.
- Context-based: metadata are utilized as the quality marker of conditions in which the data was made or utilized.
- Rating-based: the measurement depends on express appraisals for the data itself, data sources or data suppliers.

III. Big Data Quality Challenges

In spite of the fact that there are various models for assessing data quality in a customary setting, none of these models are appropriate for Big Data environments. We have recognized four significant difficulties for any data quality administration framework:

High Volume: The volume of data is huge and it's hard to assess and improve the data quality

inside a sensible time. What's more, the steady increment in the volume of data requires the execution of an adaptable arrangement. Versatility mirrors the capacity of data quality strategies to process, in a pertinent way, progressively bigger and complex datasets.

Heterogeneity: The data created these days are much of the time, semi-organized ounstructured. This kind of data is more mind boggling to process than organized data. Understanding the semantics and connections between's unstructured data is a troublesome errand. In conclusion, despite the fact that the relocation of organized data from a social database to a non-social database is conceivable, for instance, the transformation of semi-organized data into organized data is troublesome or inconceivable.

Data change: Nowadays, data changes rapidly and can quickly get out of date. In the event that associations can't gather the necessary data, state-of-the-art and continuously, they may create pointless or deceiving ends, conceivably prompting choice blunders.

Data security: On the one hand, a data quality administration framework includes read access to all data to play out the data quality assessment process and compose access to all data to manufacture the way toward improving their quality. Then again, a data security framework expects to shield data from unauthorized read and write access.

A) BIG DATA SECURITY

Despite the fact that there are various models for assessing data quality in a customary Traditionally, Security is centered around Confidentiality, Integrity, and Availability. Privacy is proposed to shield data from unapproved get to. Respectability is tied in with shielding data from unapproved changes. Accessibility manages making data open to approved elements and clients. ISO/IEC 27001 considers different properties that might be engaged with data security to be specific realness, obligation, non-denial and unwavering quality. Moreover, in the Big Data field, a few examinations are keen on Privacy so as to ensure individual and delicate data and assign it as one of the principle security goals. Security can be considered as a disposition of classification that considers extra components, for example, client assent the board in regards to their own data, consistence with administrative and lawful commitments, and so forth. In this segment, we will introduce a rundown of dangers and dangers affecting data security as indicated by the Big Data process. We \

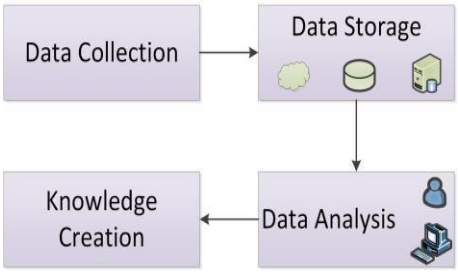

Fig. 2. Big Data Process

will at that point diagram the fundamental difficulties that a security framework faces.

B. Security risks in Big Data

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

From a security perspective, every one of these stages can introduce dangers and security dangers:

- **Data collection stage:** it's basic to gather data from dependable sources and to ensure that they are secure and secured against spillage (phishing, spamming, ridiculing). Different measures can be applied to data assortment, for example, get to control and encryption of touchy data.
- **Data storage stage:** data gathered in the past stage must be put away and secured so as to guarantee a protected situation for examination. During this stage, data stockpiling plates might be assaulted (replicating, taking or harming) or unapproved access may happen to target and investigate data bearers so as to extricate helpful information. Data anonymization, parceling and swapping are valuable procedures for ensuring put away data.
- **Data analysis stage:** data investigation is performed to remove significant data by applying Data Mining strategies. It is important to introduce a protected preparing condition to keep the data from being available to unapproved substances that can investigate and investigate it so as to separate significant or potentially close to home data. Various procedures make it conceivable to conclude individual data, for example, re-recognizable proof and connection.
- **Knowledge creation stage:** the data produced from the past stage is touchy and their insurance is obligatory. The security dangers during this stage for the most part identify with data spillage (phishing and satirizing) and the risk to the protection of people. Receiving a compelling access

control procedure and encoding the pertinent outcomes seem, by all accounts, to be acceptable approaches to improve security.

a) Data Security Challenges

Big Data, with its decent variety of data types between organized, semi-organized and unstructured, has carried numerous difficulties to the security and protection of people. The security challenges as far as cryptography, log and occasion examination, interruption discovery and anticipation, and access control have taken another measurement. In this area, we will take a gander at a lot of difficulties that can compromise data security and protection in a Big Data setting:

• **Confidentiality:** privacy includes setting up a lot of rules and limitations to constrain access to private data. It is commonly treated with get to control and cryptographic components. The regions of research to improve the secrecy of data in Big Data are worried about issues, for example, consolidating and incorporating of access control strategies, programmed the executives of these arrangements, programmed organization of approvals, use of access control on Big Data stages, and so on.

• **Integrity:** with regards to Security, uprightness implies safeguarding data against unapproved changes. In Big Data, preparing is generally dispersed more than a few hubs. Uprightness infers the consistency of data between various duplicates. It likewise infers that the data isn't changed or adjusted by unapproved parties during advances between hubs. Trustworthiness is commonly affected by equipment and programming mistakes, human blunders and interruptions. To keep up data trustworthiness, the difficulties to be tended to incorporate, notwithstanding the administration of access approvals, the confirmation of the dependability of data and their sources, the foundation of instruments for recognition and counteraction of data misfortune, duplication of data without affecting accessibility, and so on.

• **Availability:** accessibility implies that data must stay open to approved clients and elements. This alludes to the anticipation and recuperation of equipment and programming blunders, human mistakes, and vindictive access that can make data inaccessible. For the most part, the accessibility of data is fulfilled by applying various data replications. In any case, replication can prompt data honesty gives that depend on duplication to guarantee consistency across data. Also, the accessibility of data can hurt protection by improving the blend and the investigation of data and the conclusion of delicate data on people.

• **Privacy:** classification concerns any sort of data, with regards to individual data of people, it is called Privacy. It's tied in with controlling the sharing of by and by recognizable data (PII).

Likewise, the sharing of individual data is regularly controlled by security laws. The assurance of individual data in Big Data requires specific procedures, for example, straightforwardness which comprise of including concerned clients, getting the assent of clients who may deny it whenever, recognition and counteraction of re-recognizable proof procedures, safeguarding of diagnostic outcomes, consistence with administrative and lawful imperatives, and so on.
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 network as depicted in Fig. 3.

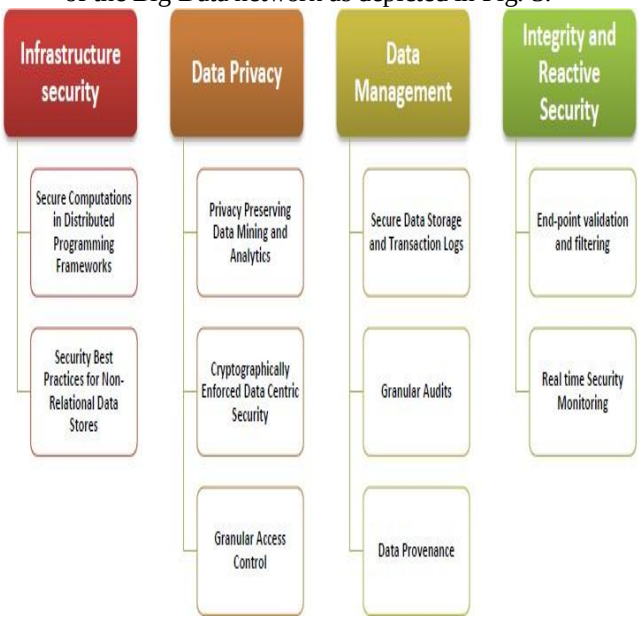

Fig. 3. Classification of the Top 10 security Challenges in the Big Data network

To meet these challenges, CSA presented, three years later in June 2016, a report detailing the 100 best practices to consider.

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 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. 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.

CONFLICT BETWEEN DATA QUALITY AND DATA SECURITY

Data Quality and Data Security are two primary subjects that contradict various difficulties in Big Data, for example, high volumes and heterogeneity of data, believability of data and their sources, the speed at which data is gathered and prepared, and so on. Regarding Data Security, three primary security properties are distinguished and are unmistakably characterized: Confidentiality, Integrity and Availability (CIA). Then again, as far as Data Quality, there is no broad concession to all properties characterizing the quality of data or the specific importance of every property. A confining for all attributes characterizing both quality and security of data is then essential. Besides, a few properties are regular to data quality and data security, yet their implications are unique. For instance, in data security, Integrity alludes to how much data is ensured against unapproved get to, though in data quality there's no reasonable definition in regards to data Integrity. However, a few investigations have connected Integrity of data quality to three properties; in particular Accuracy, Completeness and Consistency. Accuracy speaks to how much a data value fits in with its genuine or indicated value. Culmination is characterized as the degree to which every single vital value have been relegated and put away in the PC framework. Consistency alludes to the nonappearance of obvious logical inconsistencies in the data. We subsequently need to unite meanings everything being equal and decide purposes of combination and divergence. Moreover, the guideline of data security, particularly classification and trustworthiness, is to ensure data against unapproved get to. Be that as it may, executing a data quality administration framework requires adaptable peruse and compose access to all data. This prerequisite can make numerous security issues on the grounds that the data quality framework can trade data with different frameworks or be controlled by various individuals of various profiles who don't really have a similar access rights. Consequently, data security can be a boundary to data quality, and conversely, setting up a quality administration framework may require a degree of resilience that can make security vulnerabilities. This contention between these two frameworks makes their execution progressively unpredictable and requires considering new access control arrangements adjusted to the Big Data setting and empowering quality procedures to get to the necessary data without trading off their security. Such a strategy can be accomplished by executing or broadening a fine-grained get to control model, for example, 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), and so on.

What's more, a few components and procedures of Security and Quality are incongruent. For instance, we can specify the deduplication and the encryption of data. The motivation behind data deduplication, notwithstanding liberating extra room and the transmission capacity of the system is to expel copies to guarantee data consistency. The reason for encryption is to secure data against unapproved get to. Be that as it may, as referenced in [28], customary encryption procedures are contradictory with data deduplication. In particular, customary encryption requires various clients to encode their data with their own keys. Thus, identical data duplicates from various clients will prompt distinctive figure writings, making deduplication unimaginable. To beat this contention, we should set up very much adjusted systems. For our case, we should for instance consider bringing together the mystery keys inside a committed element that will permit the deduplication procedure to unscramble the data appropriately, inferring that the data is scrambled simply in the wake of confirming their uniqueness, and so forth.

Most Big Data examine centers around data quality or data security independently. In any case, the two subjects cause issues of combination. In such conditions, the reinforcing of data security instruments to the detriment of data quality procedures or the reception of certain security resiliences to improve data quality are two methodologies that require careful discretion. In Table 1, we list a progression of purposes of intermingling and dissimilarity between data quality and data security, just as a lot of activities to determine clashes.

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

IV. SECURE DATA ACCURACY ASSESSMENT SOLUTION

Our strategy to address the difficulty is to deal with the quality of the data through one or numerous properties obviously characterized. Every one of these properties must be dissected from a security perspective so as to recognize the potential dangers of contention. The goals of these contentions is to permit access to the data by the quality administration framework without trading off their security. In the accompanying, we propose a model for evaluating data quality through the "Accuracy" measurement.

A few examinations have recognized Accuracy as the key component of Data Quality. Accuracy is characterized in writing as a proportion of the vicinity of a data value v to some other value v' that is viewed as right .This definition appears to be pertinent to organized and semi-organized data, as it tends to be contrasted and reference data speaking to this present reality. In any case, for unstructured data, this definition doesn't appear to be sufficient to all circumstances. Unquestionably, unstructured data may contain data that can be contrasted and this present reality, for example, a target portrayal of an item, an image of something, a reality previously, a scientific condition, and so forth. The issue lies in circumstances where data can't be contrasted and reality. For instance, the data that just identifies with the individuals who gave it, for example, individual impressions, conclusions regarding a matter, expectations, plans to do later on, individual investigation around a subject, and so on. This sort of data needs validity and can't be contrasted and as far as anyone knows right data. Unstructured data is in this way increasingly complex to assess and can't be assessed similarly as organized or semi-organized data. Surveying the accuracy of unstructured data will be increasingly pertinent on the off chance that it centers around metadata identified with the data as opposed to the data itself.

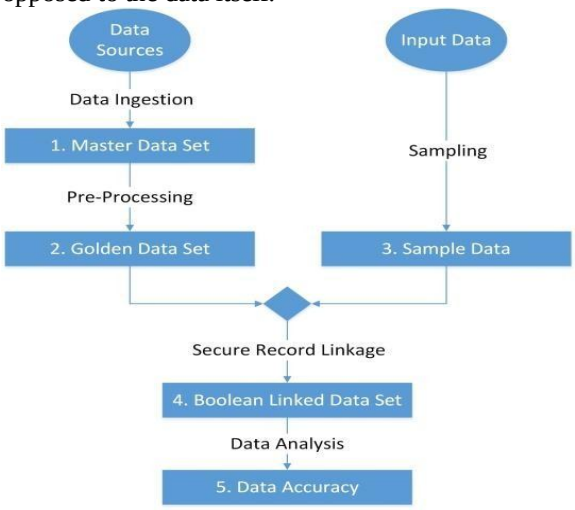

Fig. 4. Secure Data Accuracy Assessment Model

In Fig. 4, we propose our model to build up an incredible answer for survey the accuracy of data in a Big Data setting without struggle with security framework

Our solution consists in 5 steps:

1. Master Data Set: we consider that the Data Lake in its crude state is our Master Data Set that contains all data gathered from various data sources

2. Golden Data Set: this data set is an upgrade of the Master Data Set to which various procedure are applied, for example, data cleaning, copy erasure, refreshing of out of date data, adjustment of wrong values, and so on. The objective is to improve the quality from the Master Data Set to get right data in this Golden Data Set.

3. Sample Data: these data are gotten by sampling the data to survey. Sampling is a method generally talked about in writing to deal with enormous volumes of data and appears to be successful for our concern. When endeavoring to examine a data set to evaluate its quality, we can be happy with the investigation of a delegate test of the whole data set. For certain kinds of issues, sampling gives results in the same class as playing out a similar investigation utilizing all the data , however for specific cases, particularly the examination of huge volumes of data, sampling is by all accounts the most suitable arrangement.

To make an example of a dataset, various methods exist, for example, Simple Random Sampling, Stratified Sampling, Cluster Sampling, Multistage Sampling, Systematic Sampling, and so forth. A few methods can be consolidated to make a successful example. Whatever the method utilized, all data units ought to have a similar possibility of being chosen in the example. To know the size of the example, it will be important to know ahead of time the size of the data to be tested which isn't anything but difficult to acquire in a Big Data venture. To take care of this issue, there is a compelling methodology called "Store Sampling" at first presented by Vitter in 1985. Repository Sampling is a group of randomized calculations for irregular choice of an example of k components in an enormous data set of size n or in a data stream of size n, where n is obscure or hard to know. The proficiency of this calculation lies in its enhancement of making an example from a huge volume of data, without the need to know from the earlier its size. All while guaranteeing a similar opportunity to all the data units of the set to test.

1. **Boolean Linked Data Set:** this is the key advance of our answer. It includes playing out a record linkage between the example of data to evaluate and the Master Data Set. The record linkage process probably read access to all Master Data Sets at the same time, not at all like a conventional record linkage process and to safeguard data security, the aftereffects of our procedure will be as a table containing for each field of each record a Boolean value:
    - **None:** the record isn't coupled; for each field in the record, the procedure brings none back.
    - **False:** the record is combined with another from the Master Data Set, yet the value of the field is not quite the same as that of its reporter.
    - **True:** the record is combined with another from the Master Data Set and the value of the field is like that of its reporter. Closeness doesn't infer careful equality between values, yet from a specific comparability limit, we can consider that the values are equivalent.

    Along these lines, the way toward assessing data quality isn't hindered by security locks since it is necessitated that the record linkage process has free access to all data however without uncovering the data since it restores a Boolean value dependent on a similitude computation calculation.
2. **Data Accuracy:** the last advance is tied in with investigating the Boolean connected table from the past advance to conclude the general accuracy.
3. Our answer addresses the issue of this article by permitting read access to all reference data. This benefit is adequate to assess the quality of the data. Security, then, is safeguarded since we have presented two layers of data insurance:

- Comparing the data to be assessed with the reference data restores a Boolean table containing three sorts of values: True (if the datum is right), False (if the datum isn't right), or None (if the data to be assessed doesn't coordinate the reference data).
- The return of the outcomes will be as a normal value speaking to the general accuracy of the considerable number of data to be assessed.

The security framework can be additionally reinforced by confining the arrival of results to elements with explicit rights.

CONCLUSION

In this article we concentrated on the issues of quality and security of data on a similar level. The difficulties forced by the setting of Big Data mess up them two. We expect that the gave answers for take care of the issues of high volume, heterogeneity and validity of data won't just be utilized to set up quality administration frameworks yet in addition to create security ones. We have emphasized the conflicts that may exist, making the implementation of these systems more complex and requiring reflection of new solutions. At long last, we introduced an answer for assess data accuracy without affecting data security. Our answer at this stage stays hypothetical and, to legitimize its achievability and unwavering quality, it requires the execution of a lot of procedures, for example, the foundation of a Data Lake facilitating right data and setting-up of a record linkage process with read access to all the data in the Data Lake without trading off their security, and so forth.

REFERENCES

[1] WANG RY, STRONG DM. BEYOND *"ACCURACY: WHAT DATA QUALITY MEANS TO DATA CONSUMERS"*,JOURNAL OF MANAGEMENT INFORMATION SYSTEMS 1996; 12:5-33.

[2] Redman TC. *"Data's Credibility Problem"*, Harvard Business Review 2013.

[3] 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.

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

[5] Merino J, Caballero I, Rivas B, Serrano M, Piattini M*,"A Data Quality in Use model for Big Data"*. Future Generation Computer Systems 2015.

[6] Cai L, Zhu Y. *"The Challenges of Data Quality and Data Quality Assessment in the Big Data Era"*. Data Science Journal 2015; 1-10.

[7] Chen M, Mao S, Liu Y.*" Big Data: A Survey"*. Springer Science+Business Media 2014.

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

[9] 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.

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

[11] Hakuta K, Sato H. *"Cryptographic Technology for Benefiting from Big Data"*. Springer, The Impact of Applications on Mathematics 2014;85-95.

[12] Sudarsam SD,Jetley RP, Ramaswamy S.*"Security and Privacy of Big Data. Big Data: A Primer* "2015;121-136.

[13] Xu L, Shi W. "*Security Theories and Practices for Big Data*", Springer International Publishing Switzerland 2016;157- 192.

[14] 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.

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

[16] 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.

[17] 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.

[18] Immonen A, Pääkkönen P, Ovaska E,"Evaluating the Quality of Social Media Data in Big Data Architectur", IEEE Access 2015;3.

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

[20] Dr.E.Kesavulu Reddy" The Analytics of Clouds and Big Data Computing"SSRG International Journal of Computer Science and Engineering, volume 3 Issue 11–November 2016.