Informatics Bulletin, Faculty of Computers and Artificial Intelligence, Helwan University
Published Online ***Vol 3 Issue 2, July 2021***
(https://fcihib.journals.ekb.eg)

Dimensionality for Big Data Quality: A Review

Marwa Gaber Abd El-Wahab
Computer Science Department,
Faculty of Computers and
Artificial Intelligence,
Helwan university,
Cairo, Egypt

Wessam M. H. El-Behaidy
Computer Science Department,
Faculty of Computers and
Artificial Intelligence,
Helwan university,
Cairo, Egypt

Amal Elsayed Aboutabl
Computer Science Department,
Faculty of Computers and
Artificial Intelligence,
Helwan university,
Cairo, Egypt

Abstract **- The analysis and use of big data (BD) in today's enterprise context is dominated by high data quality (DQ). Decisions and opinions derived from poor quality data have negative and unexpected consequences for organizations. Currently, in the data quality field, researchers urgently need to consider improving the big data quality (BDQ), especially because of the lack of detailed research in this area. Considering the big data quality will allow organizations to take sensible, accurate assessment-based decisions. This article examines and summarizes current research into big data quality by discovering the fundamental qualities of big data and its quality. In big data, the main issues of data quality are also examined. The suggestions of some researchers on data quality evaluation are outlined. Examining this subject can help to improve conceptual measurements of high data quality and provide a strong basis for future researches by developing integrated analysis and evaluation of data quality models using appropriate algorithms.**

Index Terms**—Big Data; dimensions of data quality; data quality evaluation; dimensions of Big data quality.**

I. INTRODUCTION

Big Data (BD) is a compelling field of scientific and industrial research and development. Many researchers, IT professionals, scientists and organizations try to identify and analyze new challenges and potential technologies and methods to solve these challenges [1]. Data scientists examine current technologies and platforms, and process and analyze this huge number of data to create relevant insights that can have an important influence on culture and human well-being. For example, forecasting market growth, detecting and isolating infectious diseases, controlling road traffic, and forecasting the weather. However, traditional data sets algorithms are no longer suitable as Big Data is dynamic, constant, unstructured and wide. Thus, these tools and proposed techniques must be updated, rewritten and scrapped to respond to new data features and challenges.

Data is the most critical factor in all life cycle phases. These phases include data processing and analysis. However, without ready data, these phases will not be widespread. However, if the data is not appropriate, clean, and not ready for processing, the data processing can be very sensitive. Inaccurate data can cause biased analysis, primarily due to factors such as poor preparation, format, source, type and other data properties. Data properties should first be defined to define their quality. Data quality, also known as knowledge quality [48], has been described in a variety of ways in the literature [49]. Good quality data may indicate whether data meets user expectations [50], whether they are human user systems, or whether they are defined as "data suitable for use by data consumers" [51]. “The degree of satisfaction of a set of data properties” is defined by ISO/IEC 25012[52].

For example, Coherence and precision [33], in queries expressing the needs and limitations of the problem [53]. Examples include the need for a solution. The definitions mentioned above describe inadequate data quality to comply with data collection requirements.

Without meeting all the requirements, it is not suitable for use by the data consumer and thus affects the components involved (e.g., company, customer). The effect can be of three types: operation (through dissatisfaction and increase costs between customers and employees); tactical (through decision-making and disbelief) and strategic (i.e., affecting the overall strategy of the organization) [54]. After all, any system or company that relies on information is vulnerable to difficulties if the data processing has no expected quality properties.

Define the quality of data is a complex, multifaceted, persistent process. This generally refers to aspects from quality of service and software to data quality. Quality is also (1) the area of interest, (2) the characteristics are defined, and (3) the measurement methods and assessment methods are dependent on. In other words, deep domain knowledge, well-defined data attributes, and quality goals are essential requirements for quality assessment. Therefore, data quality

can be measured using a variety of measurement and evaluation tools in different areas. For big data [9], the data themselves and their qualities are the key issues. Big data has many features that influence data quality directly (DQ). Data diversity is one of the features of BD that describes the various data sources and formats. The variety of information gives an intuitive insight into data quality. For example, a data storage facility is structured, while data from social media is unstructured and schema-free. Data speed is another quality characteristic that quickly generates large quantities of data. For example, high-quality parameters should be considered for quality assessments.

This has a direct effect on data quality in all of these parameters. A precondition for building data confidence and ensuring its quality is therefore necessary.

This paper studies the recent work in the context of BD quality and its dimensions. The remaining article is structured as follows: In the next section, the main definition of data quality, its features and principles of the life cycle will be presented. Section 3 lists and analyses the most significant data quality research studies. Section 4 deals with the critical data quality as important issues and research guidelines. Section 5 concludes, finally, with a continuous and difficult direction.

II. BIG DATA DEFINITION AND CHARACTERISTICS

According to [2]-[8], every day a large amount of information is generated, which is 2.5 bytes (Exabyte (EB) = 1018 bytes). Data were stored with a capacity of 800 000 petabytes (1 PB = 1015 bytes) in 2000. Every second 1.7 megabyte of data on this planet was generated by 2020 [10]. The increased data storage is the result of web searches such as Google and Yahoo that many loosely organized data sources have been invited to consult. Besides, the main players and data manufacturers include application domain such as Social media, Amazon, YouTube, IoT Sensors, handheld smartphones.

To evaluate big data, it should be introduced and its characteristics linked to its evolution over the years.

A. Big Data

At the beginning and as the term "big data" suggests, it was somehow a matter of the data in a large number of files that could not be handled by traditional databases [11]. The definition was then extended to include the difficulty of analyzing these data using traditional computer algorithms. Big data is the entire value chain, covering many different stages: production, accumulation, acquisition, transportation, storage, processing as well as visualization and interpretation. The insights that can be learned from this chain are from the continuous development of data using existing methods and emerging architectures.

Big data is broad and spectrum-speed, varied and cost-efficient, innovative information treatment to enhance decision-making and understanding in [7]-[8], [12]-[13].

"Big data" describes an enormous number of semi-structured or structured information that prevents the use of traditional databases and computer technology. It also refers to technology and storage devices used by an organization to store and conserve massive amounts of data from various sources.

B. Big data features in the DQ context

Big data is regarded as structured, semi-organized and unstructured databases with a large amount of information, as previously mentioned. It is also a nightmare task for traditional computers, software and database management systems to be stored, presented, analyzed, maintained and controlled. Laney [14] is the first to add three main characteristics representing big data: Velocity, Variety and Volume, known as (3V's). Although these key elements are widely used to represent Big data, two additional aspects for data integrity have been developed: Value and Veracity. BD's principal characteristics are summarized in Table 1.

TABLE 1
MAIN BIG DATA CHARACTERISTICS

Characteristics	Explanation
Value	Importance of Data; It clearly illustrates the business value that will be extracted from big data [17]-[19].
Volume	Data size; the amount of collected and maintained data. The data may range from TB to PB [15]-[17].
Veracity	Data quality; if the information collected is not accurate, the accuracy is nearly useless [19].
Variety	Data type; The end of the receipt [15] – [17] is typically in videos, images, audio, and so on.
Velocity	Data speed; the speed of production of information [16]-[18].

III. DATA QUALITY AND DQD

Data quality (DQ) does not require that data should be error-free only. Buggy data is just one piece of data quality state. Many of the experts hold a wider perspective. There are many ways in which data and its quality are used, which has been described in terms of dimensions [27][29].

A. Data Quality

Larry English [40] says that data quality means "reliably meeting the needs of knowledge workers and end-users.". Other [22][34][37] state that the goodness or suitability of data to meet market needs is data quality.

It is known that DQ has multiple definitions that depend on the context, domain where it is used [23]-[24]. DQ is understood differently in academia than in business. In business [21], the authors sum up DQ from well-known and accepted ISO 25012 interpretations. In literature, "fitness for use." is DQ.

In [25] the data quality is described as acceptable for use or to meet the requirements of users.

Big data indicates that data quality can present many problems due to the high number of data sources and to the different types of data sources which can be processed in a rather short time. [28].

The analysis of BD is different depending on the context of their application. In some cases, some data which can irremediably be called 'dirty,' which is not of sufficient quality for application use, may be found very valuable in other contexts. The crucial issue here is that DQ for large datasets depends on their intent to use [29].

The data is distributed via BD storage systems to allow for the unrestricted handling of large amounts of information through distributed computing. Further, there is a high availability of error tolerance, geo-distribution and replication of data. However, this will result in data quality concerns such as consistency across several data centres.

BDQ requires well-defined and lightweight measurement processes which can run in parallel with each phase [30]. These processes provide data quality management, control and keep track of any changes that might increase or decrease the data quality.

B. Data Quality Dimensions (DQD)

DQD is used by data management experts to define data quality characteristics that can be measured or graded according to data quality standards [26]. Some of the common DQDs which are widely cited in the literature are listed below:

1) *Completeness:* To be completed, the values of all components of the data element should be valid. It decides whether or not in the data resource all the data needed to meet current and future company needs are available.
2) *Timeliness:* It calculates the time between acquisition and analysis, including the use of data with the desired results. Failure to collect data within a specified period may affect its usefulness in decision making.
3) *Accuracy:* refer to whether the data represents the dataset.
4) *Consistency:* The lack of distinction when comparing two or more examples of the same thing with their meanings.

In big data, several ranges of quality dimensions have been considered by different researchers in current literary works. In [31], the fact of each dimension is designed to cover specific areas that might fall under the overall idea of data quality, was explained.

Table 2 summarizes BDQD introduced in [31][32].

TABLE 2
KEY DIMENSIONS THAT ARE RELATED TO BDQ [31][32].

BDQ Dimension	Explanation
Uniqueness	The information can be saved once
Validity	If the rules have complied with the data is correct (format, type and range).
Correctness	It often is known as syntactical precision, which means a domain's closeness to a data value.
Currency	Concerns on how prompt data are
Accessibility	The willingness of customers to access data from their context, physical status and available technologies.
Believability	Whether a source provides information that can be considered accurate, genuine and trustworthy
Reputation	Considers how secure the source of knowledge is
Relevancy	The specifics of the actual task are applicable
Ease of understanding	How much data is simple, without confusion and easy to understand.
Volume of data	To what extent the correct data size is used to complete the task.
Ease of manipulation	How easy it is for different tasks to process and submit data.
Free-of-error	Degree the data is truthful, accurate, and reliable.
Interpretability	It is clear to what extent the data are insufficient for units, symbols, languages and importance.
Objectivity	The extent of explicit and unsatisfactory languages or syntax free descriptions of data.
Security	The level of access protection offered to maintain security.
Cohesion	Consistency, cohesion and coherence apply to the capacity of data to conform without interfering with all the properties of the fact of concern, as defined in terms of credibility restrictions, data editing, market rules and other formalities.

Redundancy	Redundancy, minimality, compactness and conciseness apply to the capacity to reflect the fact of importance with limited use of insightful tools.
Value-added	The data seems to be important.

Table 3 summarizes some dimensions that have been reviewed along with BDQ and traditional data quality [31][36].

TABLE 3
SOME QUALITY DIMENSIONS THAT ARE RELATED TO TRADITIONAL AND BIG DATA

Dimension	Traditional data quality [35]	Big Data Quality Research Work [32]	Big Data Quality Research Work [31]	Big Data Quality Research Work [34]	Big Data Quality Research Work [36]	Big Data Quality Research Work [33]
Completeness	✓	✓	✓	✓	✓	✓
Timeliness	-	✓	-	✓	✓	✓
Validity	-	✓	✓	-	-	✓
Uniqueness	✓	-	-	-	✓	-
Accuracy	✓	-	✓	✓	✓	✓
Consistency	✓	-	✓	✓	✓	✓
Correctness	✓	-	✓	-	-	✓
Currency	✓	-	-	-	✓	✓
Accessibility	✓	✓	✓	✓	-	✓
Believability	✓	✓	✓	-	-	✓
Reputation	✓	✓	✓	-	-	✓
Objectivity	✓	✓	-	-	-	✓
Relevancy	✓	✓	-	✓	-	✓
Ease of understanding	✓	✓	✓	-	-	✓
Volume of data	✓	✓	-	-	-	-
Ease of manipulation	✓	✓	-	-	-	-
Free-of-error/ Credibility	✓	✓	-	✓	✓	-
Interpretability	✓	✓	-	-	✓	-
Security	✓	✓	-	-	-	-
Value-added	✓	✓	-	-	-	✓

The mere analysis of the dimensions listed above clearly indicates a substantial degree of consistency and similarity between the various ways in which the data can be analyzed for quality purposes. However, the question that should be appeared here, how much these dimensions can

indeed extend in the light of BD? What are the dimensions that are perceived to be more relevant for data quality in the sense of BD use?

C. The main dimensions of BDQ and its relation to Big Data characteristics

According to [28] it could not be as important to increase DQ for large data as the quantity of inaccurate data. Because this amount of incorrect data is considered negligible to affect the final result after the analysis of data. Therefore, which of those two completely opposing schools of thinking is relevant appears to rely on the impact and amount of the incorrect or 'dirty' data as part of a big data collection. This raises the importance of understanding which dimensions are more relevant for big data.

Caballero et al [29] say that the main dimension of data quality to be tackled for BD is the consistency that they define as the ability of information systems to ensure the integrity of datasets as data is distributed across networks and systems. Their major idea is that a dataset can only be calculated for business value concerning its application. They further subdivide the consistency into three parts (3 C's). Even so, many of the conventional dimensions of data quality were linked to the three consistency sub-domains:

1) *Contextual consistency* [38] means the degree of use in the same field of application of large data sets regardless of the form of data generation, size and speed. As a result, validity, authenticity, ease of understanding, precision and confidentiality are perceived to be essential to contextual continuity development.
2) *Temporal consistency* presents the fact that data has to be understood in a consistent time slot, such that the same data cannot be equivalent if it may not come from the same time slot. Temporal consistency is assumed to include time concurrence, availability, and currency.
3) *Operational consistency* relates to the operational effectiveness of technologies on data creation and use. The principal connected dimensions are available, portable, consistent, complete and traceable.

However, many dimensions of typical data quality have been connected to the consistency sub-bands as follows (Table 4).

TABLE 4
MAPPED 3V'S OF BD TO THE 3C'S OF DATA QUALITY [29]

	Velocity	Volume	Variety
Contextual consistency	Consistency, Credibility, Confidentiality	Completeness, Credibility	Accuracy, Consistency, Understandability
Temporal consistency	Consistency, Credibility, Currentness, Availability	Availability	Consistency, Currentness, Compliance
Operational consistency	Completeness, Accessibility, Efficiency, traceability, Availability, Recoverability	Completeness, Accessibility, Efficiency, Availability, Recoverability	Accuracy, Accessibility, Compliance, Efficiency, Precision, Traceability, Availability

The approach to the specific properties of large data in which DQ dimensions were given was, however, simply not an experimental or research assumption. Research in this sector is immediately required to identify the dimensions that are most relevant for big data through research on the BDQ.

D. Poor Data, DQ Issues and Problems

Many issues, such as easy access, flexible, reliable and timely data utilization can affect data efficiency [28]-[29]. There are various explanations for poor quality data, including lack of authentication protocols [39]. Often the correct data, although errors in syntactic or formatting have been identified[26]. Poor data can also result from inadequate changes and configuration monitoring systems within the organization, and weak system development processes (system design, data conversion errors, validation processes, etc.). In [41], data quality dimensions were linked to the causes of incorrect data, concluding that DQ issues may appear to be inaccurate and contradictory values, missing data, uniqueness constraints, and violations of functional dependence. The Transforming Intelligence Information (TDWI) data quality study [20]

indicated that there was a lack of credibility, additional costs and time to refinish data, frustration with consumers and continuing delays in the provision of new system products as the usual problems induced by faulty data. All these problems are shown in Fig. 1, all these issues are presented.

E. DQ problems for Big Data characteristics

precision, veracity and reliability are also associated with the accuracy of the DQD in the BD context [43]. In [26],[29],[34], an attempt is made to map these characteristics, data and DQ. In another research, the authors of [44] addressed the DQD "Accuracy" versus the "Volume" feature of BD. They assume that the rise in data size has a high effect on the improvement of the DQ.

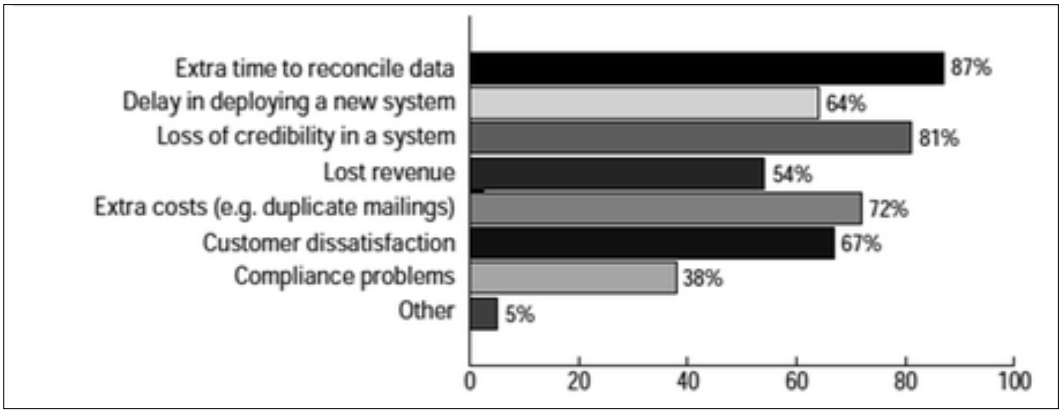

Fig. 1. Some of the reasons that lead to data quality problems [20]

IV. QUALITY ASSESSMENT

DQ can be measured and the metric of data quality should determine whether the attributes agree with the previously defined DQD or not.

A. DQ Evaluation, Metrics, and Measurement

In [9], their work is based on a set of columns, rows and attribute(s), and its values on structured data. Measurements of data quality must determine whether or not the value of the data meets the quality attribute. The author states in [39] that measuring data quality metrics evaluates good or bad binary results, or values from 0 to 100 (100% is optimal) and uses default methods to determine them. This applies to many quality dimensions, such as accuracy, completeness and consistency.

For example, the following is defined by a metric calculating the accuracy of a data attribute:

1) The data type of the attribute and its value.
2) For numeric attributes, a range or sets of valid values (also text) can be defined. All other values are incorrect.
3) The accuracy of an attribute is calculated according to the number of valid values divided into the number of rows or remarks. Table 4 lists the measures of many DQD values scores
4) For other data, types/formats, such as images, video, audio files, a different type of metric should be defined to assess the accuracy or other quality parameter.

The authors of [26] define the usefulness of images as an aspect of its data quality. For this type of data, the extraction features of the data are described and extracted for each data object. These features have drawbacks that define the goodness or badness of the data values.

Some quality metrics functions are designed based on extracted characteristics such as usability, accuracy, integrity (based on multiple attributes), and any other data quality dimension that is a candidate for such data types (e.g., video, image or audio) be judged by experts in the domain. their work is based on structured data represented by a set of attributes, columns and rows and their values. DQ Metrics shall determine whether or not the data value meets the attribute of quality.

TABLE 4
METRIC FUNCTIONS OF DATA QUALITY DIMENSION

DQ Dimension	Metric function
Consistency	Cons= (N_{vrc} / N)
Completeness	Comp= (N_{mv} / N)
Accuracy	Acc = (N_{cv} / N)

N: in the sample dataset, the total number of rows
N_{vrc}: Total value that complies with the limitations
N_{mv}: Amount of missing values
N_{cv}: Amount of the correct values

B. Evaluation of BDQ

In [45], the authors present their proposed quality pipeline mapping from big data processing pipelines [46]. First, a high-quality pipeline was implemented and then three key qualities were concentrated: consistency, accuracy and

confidentiality. Fig. 2 demonstrates their findings.

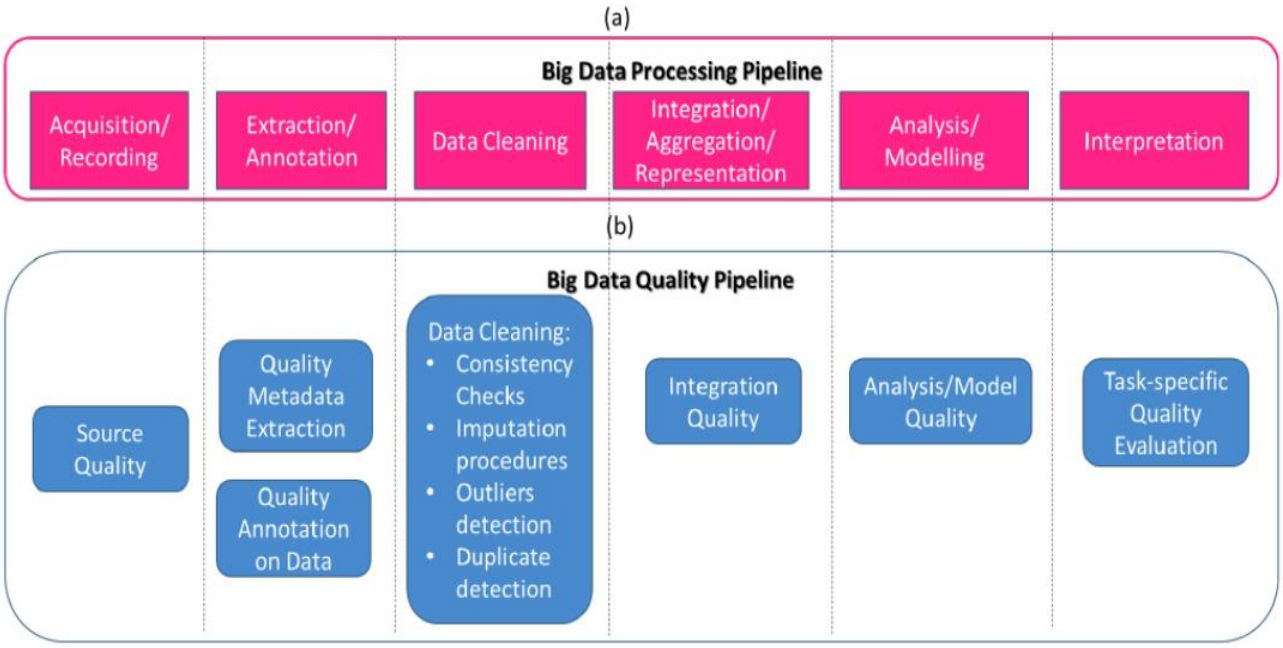

Fig. 2. Part (a) - Pipeline for big data processing; Part (b) - Pipeline of BDQ [45]

In [34], a dynamic feedback system has been suggested with a key data quality evaluation approach that focuses on the characteristics of BD, as seen in the following Fig. 3.

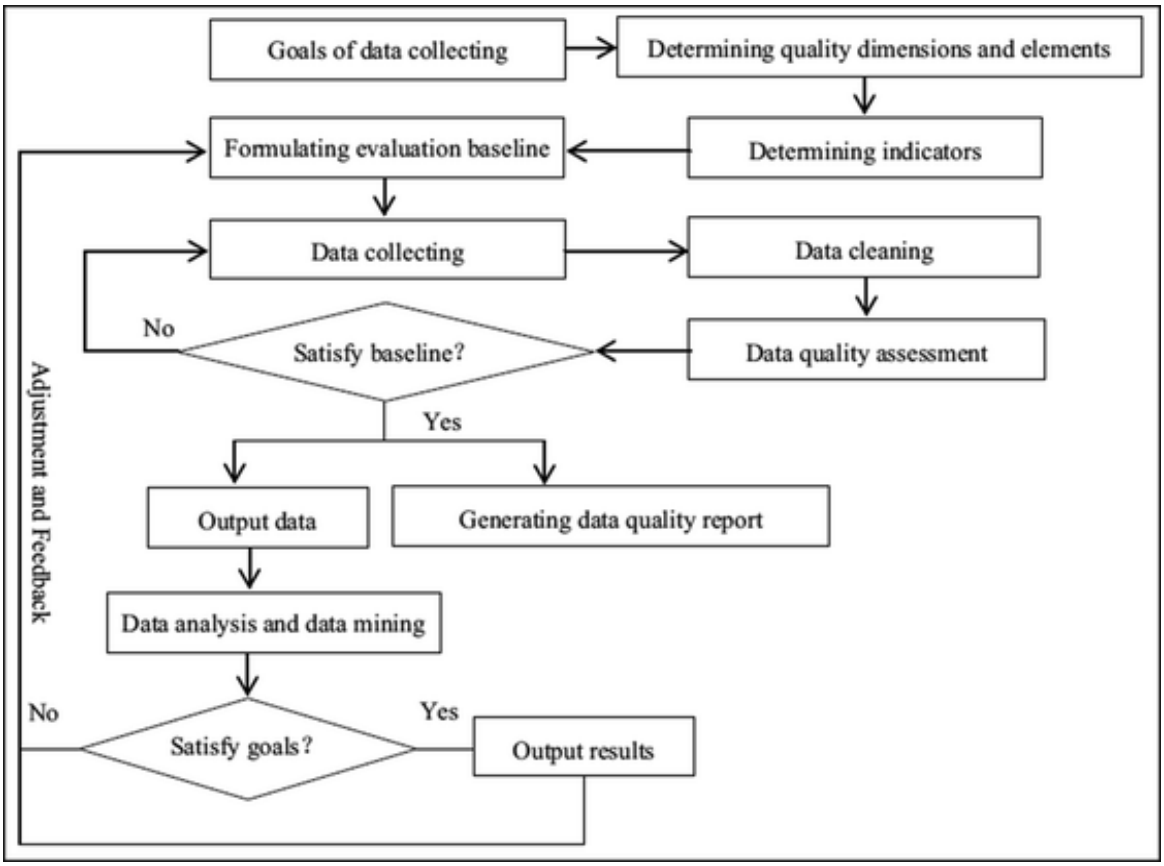

Fig. 3. The BDQ evaluation process [34]

In [47], they propose a quality evaluation model that selects the quality dimension for each type of data and evaluates its extracted characteristics. Because of Unstructured data which has no column values, they use a quantitative approach to data quality based on the content of the data. Figure 4 illustrated this proposed approach.

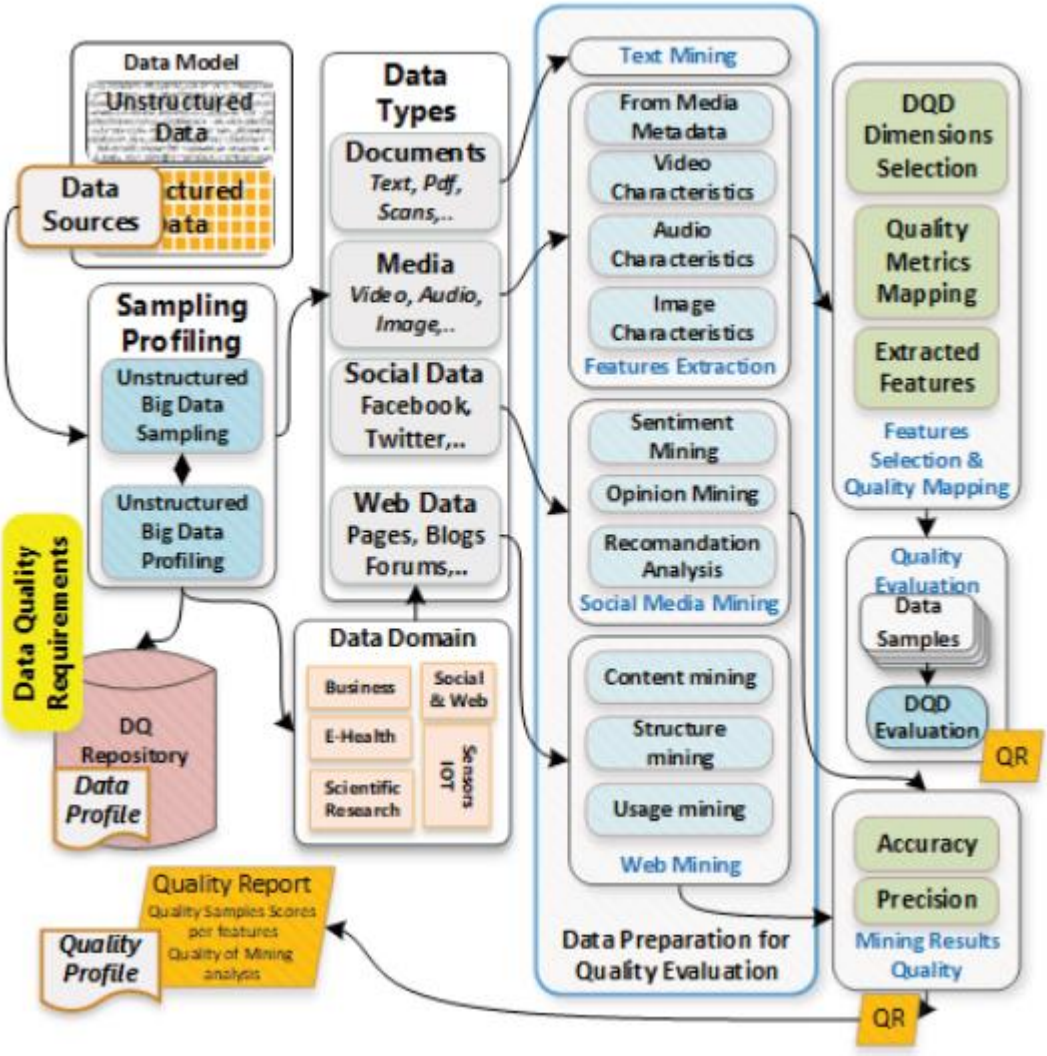

Fig. 4. Model for the evaluation of unstructured BD [47]

V. CONCLUSION

The big data era gives rise to concerns about ensuring BDQ and the analysis of information and expertise. Poor data quality could lead to poor data efficiency and major decision-making errors. This paper reviews current research on BDQ problems, highlights proposed mechanisms and measurement methodologies for the enforcement and improvement of BDQ. This paper also proposed the use of a metric big data model as a guide in respect of data quality. As a future work, according to all the above work, more scientific assessments and the testing of the proposed solutions are recommended. The measuring results for the qualitative dimension should therefore also be based on automatic optimization and quality testing. Besides, a metric model should be used to construct large data in the form of measurements of data quality.

VI. REFERENCES

[1] Taleb, Ikbal, Mohamed Adel Serhani, and Rachida Dssouli. "Big data quality: A survey." In 2018 IEEE International Congress on Big Data (BigData Congress), pp. 166-173., 2018.

[2] Chiang, Fei, and Renée J. Miller. "Discovering data quality rules." Proceedings of the VLDB Endowment 1, no. 1 (2008): 1166-1177.

[3] What Is Big Data? - Gartner IT Glossary - Big Data," Gartner IT Glossary, 25-May-2012. [Online]. Available: http://www.gartner.com/it-glossary/big-data/. [Accessed: 30-Januray-2021].

[4] M. Beyer, D. Laney, The Importance of "Big Data": A Definition-Gartner 2012, G00235055. [Accessed: 30-Januray-2021].

[5] Manyika, James, Michael Chui, Brad Brown, Jacques Bughin, Richard Dobbs, Charles Roxburgh, and Angela Hung Byers. Big data: The next frontier for innovation, competition, and productivity. McKinsey Global Institute, 2011.

[6] De Mauro, Andrea, Marco Greco, and Michele Grimaldi. "What is big data? A consensual definition and a review of key research topics." In AIP conference proceedings, vol. 1644, no. 1, pp. 97-104. American Institute of Physics, 2015.

[7] Emmanuel, Isitor, and Clare Stanier. "Defining big data." In Proceedings of the International Conference on Big Data and Advanced Wireless Technologies, pp. 1-6. 2016.

[8] Ward, Jonathan Stuart, and Adam Barker. "Undefined by data: a survey of big data definitions." arXiv preprint arXiv:1309.5821 (2013).

[9] Taleb, Ikbal, Hadeel T. El Kassabi, Mohamed Adel Serhani, Rachida Dssouli, and Chafik Bouhaddioui. "Big data quality: A quality dimensions evaluation." In 2016 Intl IEEE Conferences on Ubiquitous Intelligence & Computing, Advanced and Trusted Computing, Scalable Computing and Communications, Cloud and Big Data Computing, Internet of People, and Smart World Congress (UIC/ATC/ScalCom/CBDCom/IoP/SmartWorld), pp. 759-765., 2016.

[10] Labrinidis, Alexandros, and Hosagrahar V. Jagadish. "Challenges and opportunities with big data." Proceedings of the VLDB Endowment 5, no. 12 (2012): 2032-2033.

[11] Yeh, Peter Z., and Colin A. Puri. "An efficient and robust approach for discovering data quality rules." In 2010 22nd IEEE International Conference on Tools with Artificial Intelligence, vol. 1, pp. 248-255., 2010.

[12] G. Press, "12 Big Data Definitions: What's Yours?" Forbes. [Online]. Available: https://www.forbes.com/sites/gilpress/2014/09/03/12-big-data-definitions-whats-yours/.
[Accessed: 30-Januray-2021].

[13] Hu, Han, Yonggang Wen, Tat-Seng Chua, and Xuelong Li. "Toward scalable systems for big data analytics: A technology tutorial." IEEE Access 2 (2014): 652-687.

[14] Laney, Doug. "3D data management: Controlling data volume, velocity and variety." META Group research note 6, no. 70 (2001): 1.

[15] Oguntimilehin, Abiodun, and E. O. Ademola. "A review of big data management, benefits and challenges." A Review of Big Data Management, Benefits and Challenges 5, no. 6 (2014): 1-7.

[16] Kaisler, Stephen, Frank Armour, J. Alberto Espinosa, and William Money. "Big data: Issues and challenges moving forward." In 2013 46th Hawaii international conference on system sciences, pp. 995-1004. IEEE, 2013.

[17] Mark Troester (2013), ―Big Data Meets Big Data Analytics, [online], [Accessed: 30-Januray-2021].

[18] Oracle (2013), ―Information Management and Big Data: A Reference Architecture, [online], [Accessed: 30-Januray-2021].

[19] Owais, Suhail Sami, and Nada Sael Hussein. "Extract five categories CPIVW from the 9V's characteristics of the big data." International Journal of Advanced Computer Science and Applications 7, no. 3 (2016): 254-258.

[20] Eckerson, W. (2002) Data Quality and the Bottom Line: Achieving Business Success through a Commitment to High-Quality Data: The Data Warehouse Institute: 1-33

[21] Kaisler, Stephen and J. Alberto Espinosa, Frank Armour, and William Money. "Advanced Analytics for Big Data." In Encyclopedia of Information Science and Technology, Third Edition. edited by Mehdi Khosrow-Pour, D.B.A., 7584-7593. Hershey, PA: IGI Global, 2015. http://doi:10.4018/978-1-4666-5888-2.ch747

[22] Tayi, Giri Kumar, and Donald P. Ballou. "Examining data quality." Communications of the ACM 41, no. 2 (1998): 54-57.

[23] Sidi, Fatimah, Payam Hassany Shariat Panahy, Lilly Suriani Affendey, Marzanah A. Jabar, Hamidah Ibrahim, and Aida Mustapha. "Data quality: A survey of data quality dimensions." In 2012 International Conference on Information Retrieval & Knowledge Management, pp. 300-304. IEEE, 2012.

[24] Caballero, Ismael, and Mario Piattini. "CALDEA: a data quality model based on maturity levels." In Third International Conference on Quality Software, 2003. Proceedings., pp. 380-387. IEEE, 2003.

[25] Batini, Carlo, Cinzia Cappiello, Chiara Francalanci, and Andrea Maurino. "Methodologies for data quality assessment and improvement." ACM computing surveys (CSUR) 41, no. 3 (2009): 1-52.

[26] Firmani, Donatella, Massimo Mecella, Monica Scannapieco, and Carlo Batini. "On the meaningfulness of "big data quality"." Data Science and Engineering 1, no. 1 (2016): 6-20.

[27] Carey, M. J., S. Ceri, P. Bernstein, U. Dayal, C. Faloutsos, J. C. Freytag, G. Gardarin et al. "Data-Centric Systems and Applications." Italy: Springer (2006).

[28] Soares, S. "Big Data Quality, Big Data Governance: An Emerging Imperative." 101-112.

[29] Merino, Jorge, Ismael Caballero, Bibiano Rivas, Manuel Serrano, and Mario Piattini. "A data quality in use model for big data." Future Generation Computer Systems 63 (2016): 123-130.

[30] Cheah, You-Wei, Richard Canon, Beth Plale, and Lavanya Ramakrishnan. "Milieu: Lightweight and configurable big data provenance for science." In 2013 IEEE International Congress on Big Data, pp. 46-53. IEEE, 2013.

[31] Batini, Carlo, Anisa Rula, Monica Scannapieco, and Gianluigi Viscusi. "From data quality to big data quality." In Big Data: Concepts, Methodologies, Tools, and Applications, pp. 1934-1956. IGI Global, 2016.

[32] Pipino, Leo L., Yang W. Lee, and Richard Y. Wang. "Data quality assessment." Communications of the ACM 45, no. 4 (2002): 211-218.

[33] Scannapieco, Monica. Data Quality: Concepts, Methodologies and Techniques. Data-Centric Systems and Applications. Springer, 2006.

[34] Cai, Li, and Yangyong Zhu. "The challenges of data quality and data quality assessment in the big data era." Data science journal 14,2015.

[35] Sidi, Fatimah, Payam Hassany Shariat Panahy, Lilly Suriani Affendey, Marzanah A. Jabar, Hamidah Ibrahim, and Aida Mustapha. "Data quality: A survey of data quality dimensions." In 2012 International Conference on Information Retrieval & Knowledge Management, pp. 300-304. IEEE, 2012.

[36] El Alaoui, Imane, Youssef Gahi, and Rochdi Messoussi. "Big data quality metrics for sentiment analysis approach." In Proceedings of the 2019 International Conference on Big Data Engineering, pp. 36-43. 2019.

[37] Saeed, Faisal, Nadhmi Gazem, Fathey Mohammed, and Abdelsalam Busalim, eds. Recent Trends in Data Science and Soft Computing: Proceedings of the 3rd International Conference of Reliable Information and Communication Technology (IRICT 2018). Vol. 843. Springer, 2018.

[38] CARLO. SCANNAPIECO BATINI (MONICA.). DATA AND INFORMATION QUALITY: Dimensions, Principles and Techniques. SPRINGER, 2018.

[39] Juddoo, Suraj. "Overview of data quality challenges in the context of Big Data." In 2015 International Conference on Computing, Communication and Security (ICCCS), pp. 1-9. IEEE, 2015.

[40] English, Larry P. Improving data warehouse and business information quality: methods for reducing costs and increasing profits. John Wiley & Sons, Inc., 1999.

[41] Laranjeiro, Nuno, Seyma Nur Soydemir, and Jorge Bernardino. "A survey on data quality: classifying poor data." In 2015 IEEE 21st Pacific rim international symposium on dependable Computing (PRDC), pp. 179-188. IEEE, 2015.

[42] Uddin, Muhammad Fahim, and Navarun Gupta. "Seven V's of Big Data understanding Big Data to extract value." In Proceedings of the 2014 zone 1 conference of the American Society for Engineering Education, pp. 1-5. IEEE, 2014.

[43] Goasdoué, Virginie, Sylvaine Nugier, Dominique Duquennoy, and Brigitte Laboisse. "An Evaluation Framework for Data Quality Tools." In ICIQ, pp. 280-294. 2007.

[44] Woodall, Philip, Alexander Borek, Jing Gao, Martin A. Oberhofer, and Andy Koronios. "An Investigation of How Data

Quality is Affected by Dataset Size in the Context of Big Data Analytics." In ICIQ. 2014.

[45] Catarci, Tiziana, Monica Scannapieco, Marco Console, and Camil Demetrescu. "My (fair) big data." In 2017 IEEE International Conference on Big Data (Big Data), pp. 2974-2979. IEEE, 2017.

[46] Bertino, Elisa. "Big Data--Opportunities and Challenges Panel Position Paper." In 2013 IEEE 37th Annual Computer Software and Applications Conference, pp. 479-480. IEEE Computer Society, 2013.

[47] Taleb, Ikbal, Mohamed Adel Serhani, and Rachida Dssouli. "Big data quality assessment model for unstructured data." In 2018 International Conference on Innovations in Information Technology (IIT), pp. 69-74. IEEE, 2018.

[48] Ge, Mouzhi, and Markus Helfert. "A review of information quality research—develop a research agenda." In Paper presented at the International Conference on Information Quality 2007. 2007.

[49] Batini, Carlo, Matteo Palmonari, and Gianluigi Viscusi. "Opening the closed world: A survey of information quality research in the wild." In the Philosophy of Information Quality, pp. 43-73. Springer, Cham, 2014.

[50] Sebastian-Coleman, Laura. Measuring data quality for ongoing improvement: a data quality assessment framework. Newnes, 2012.

[51] Strong, Diane M., Yang W. Lee, and Richard Y. Wang. "Data quality in context." Communications of the ACM 40, no. 5 (1997): 103-110.

[52] International Organization for Standardization /International Electrotechnical Commission. "Software Engineering-Software product Quality Requirements and Evaluation (SQuaRe) Data quality model." ISO/IEC 25012 (2008): 1-13.

[53] "Swebok v3 guide IEEE computer society." [Online]. Available: http://www.computer.org/web/swebok/v3- guide. [Accessed: 30-Januray-2021].

[54] Redman, Thomas C. "The impact of poor data quality on the typical enterprise." Communications of the ACM 41, no. 2 (1998): 79-82.