Big Data Quality: A Survey

Ikbal taleb
CIISE
Concordia University
Montreal, QC, Canada
i_taleb@live.concordia.ca

Mohamed Adel Serhani
College of Information Technology
UAE University
Al Ain, UAE
serhanim@uaeu.ac.ae

Rachida Dssouli
CIISE
Concordia University
Montreal, QC, Canada
rachida.dssouli@concordia.ca

***Abstract*—With the advances in communication technologies and the high amount of data generated, collected, and stored, it becomes crucial to manage the quality of this data deluge in an efficient and cost-effective way. The storage, processing, privacy and analytics are the main keys challenging aspects of Big Data that require quality evaluation and monitoring. Quality has been recognized by the Big Data community as an essential facet of its maturity. Yet, it is a crucial practice that should be implemented at the earlier stages of its lifecycle and progressively applied across the other key processes. The earlier we incorporate quality the full benefit we can get from insights. In this paper, we first identify the key challenges that necessitates quality evaluation. We then survey, classify and discuss the most recent work on Big Data management. Consequently, we propose an across-the-board quality management framework describing the key quality evaluation practices to be conducted through the different Big Data stages. The framework can be used to leverage the quality management and to provide a roadmap for Data scientists to better understand quality practices and highlight the importance of managing the quality. We finally, conclude the paper and point to some future research directions on quality of Big Data.**

Keywords—Big Data, Data Quality, Quality Management framework, Quality of Big Data.

I. INTRODUCTION

Big Data (BD) has become a very attractive area of research and development for both academia and industries. With the spread of Broadband Internet everywhere and the large number of services that emerged recently (VOD, Cloud storage and services, data clusters), a huge amount of data is generated every day strengthening the Big Data Era. Many IT professionals, researchers, scientists, and companies are working heavily to define, describe, and analyze the new challenges and the possible technologies and approaches that might be used to address these challenges. Exploring existing technologies and platforms, data scientists are processing, and analyzing this huge amount of data to produce relevant insights that might have a big impact on society and human wellbeing. For instance, predicting market growth, tracking and isolating infection diseases, managing road traffic, and predicting meteorology. However, traditional tools, techniques, and algorithms used for traditional datasets are not anymore suitable since Big Data is dynamic, continuous in nature, takes various format, unstructured, and of big size. Therefore, it is important to adapt, rewrites, redesign, from scratch these tools and algorithms to respond the new Data characteristics and related challenges.

In Big Data, data originally comes in different aspects, from multiples sources that must be cleaned, filtered, processed, integrated, merged, partitioned, transported, sketched, and stored. All these steps are executed in real-time, in batch or in parallel and preferably on the cloud. While it is well-known that, in theory, more high quality data leads to better predictive power and overall insight, this raw data must be channeled through a quality assessment in the pre-processing phase in which activities such as data cleansing, de-duplication, compression, filtering, and format conversion take place. This mandatory step is essential to refine, valuate the data and ensure its quality.

In order to keep track of data value and relevance as well as the severity of the impact of the aforementioned pre-processing, and processing transformations, a concept of data quality is paramount importance. Moreover, the nature of targeted data, such as those generated from social networks and which are unstructured with no quality references, suggests that data must be profiled and provided with certain quality information's at the inception phase. This also means that data attributes quality must be assessed, improved and controlled all along its lifecycle as it directly impacts the results of the analysis phase.

Data quality is a well-known concept within database community and have been an active area of research for many years [1], [2]. However, a direct application of these quality concepts to Big Data faces severe challenges in terms of time and costs of data pre-processing. The problem is exacerbated by the fact that these techniques were developed for well-structured data. Big data reveals new characteristics that make its quality assessment very challenging. The variety of Big data brings complex data structure which increases the difficulty of its quality evaluation. Also, Big data high volume involves time and resources for processing which hardly influence the process of its quality evaluation. In addition, variability, velocity, and volatility features introduce new challenges in managing, and assessing the quality of Big data given the speed in which it is generated and fluctuated. To the best of our knowledge there is no standard quality management framework for Big data that has emerged yet. Most of existing works on Big data quality management are still under investigations and have not reached a good level of maturity. Past work in the database community cannot be fully adopted as it is because of the above mentioned Big data new challenges. However, some quality assessment practices can be readapted to cope with these new issues.

In this context, the data quality model should be developed to follow some Big Data key concepts as the origin, domain, nature, format, and type of data it is applied on. A proper management of these quality schemes is essential when dealing with large datasets. In addition, existing Big Data architectures do not support quality management processes. However, some

978-1-5386-7232-7/18/$31.00 ©2018 IEEE
DOI 10.1109/BigDataCongress.2018.00029

initiatives are still limited to specific application domain and scope. Moreover, the evaluation and estimation of Quality must be handled in all the lifecycle phases from data inception to its analytics. This evaluation is crucial to provision value-added services and achieve the Big Data vision. Quality measurement, assessment, enforcement, monitoring, and adaptation are key quality processes that will illustrate what Big Data quality management means.

The rest of paper is organized as follow: next section introduces Big Data and data quality foundations, definition, characteristics, and lifecycle. Section 3 introduces a holistic quality management model for Big Data value chain. Section 4 surveys and classify the most important research works on Big Data quality evaluation and management. Section 5 identifies the main challenges and the open research directions in Big Data quality management in general and along the quality management sub-processes. Finally, the last section concludes the paper with ongoing and challenging directions.

II. BIG DATA AND DATA QUALITY FOUNDATIONS

According to IBM [3], Gartner [4], [5], McKinsey [6],and [7]–[9] every day huge amounts of data are generated; this data represents 2.5 quintillion bytes (Exabyte (EB) = 1018 bytes). In year 2000, 800,000 Petabyte (1 PB= 10^{15} bytes) of data were stored. Twenty years later, in 2020 this number will reach 35 Zettabytes (1 ZB= 10^{21} bytes) [10]. This exponential increase of data storage is originated from Web search companies as Google, Yahoo; who had to query very large distributed aggregations of loosely structured data sources. Moreover, application domains including Facebook, Amazon, Twitter, YouTube, Internet of things sensors, mobile smart phones are the main actors and data generators. The amount of data they generate daily is from 5 to 10 Terabytes (1 TB= 10^{12} bytes).

A. Big Data

If we need to define Big Data, we must introduce its evolution through the years while linking it to its characteristics. As the name implies, it was somehow about the large size of data files that cannot be handled by traditional databases [1]. Then extended to cover the difficulty to analyze these data using the traditional software algorithms. Big Data means the whole value chain that includes several stages: data generation, collection, acquisition, transportation, storage, preprocessing, and processing, analytics, and visualization. The insights that we can extract from this chain are from the continuous data growth using new techniques and new architectures.

1) *Definition*: there is no clear and final definition of Big Data according to many references such as: [8], [9], [11], [12]. It is high-volume, high-velocity and high-variety information assets that demand cost-effective, innovative forms of information processing for enhanced insight and decision-making. "Big Data" is used to describe a massive volume of both structured and unstructured data; therefore, it's difficult to process it using traditional database and software techniques. It refers also to the technologies and storage facilities that an organization requires to handle and manage the large amounts of data that derives from multiples sources.

2) *Origins*: the data originates from everywhere: sensors used to gather climate information, posts to social media sites, digital pictures and videos uploaded to media portals, purchase transaction records, and cell phone GPS signals to name a few. The gigantic volume of data did not mean it's the only characteristics to consider.

B. Charachteristics

In 2011, early days of Big Data, McKinney Global Institute report [6] identifies three main original dimensions that characterize it from any other data concepts. The Volume, Velocity and Variety, also called the 3 V's as illustrated in figure 2. It is essential that these characteristics are not limited in number. Because Big Data is more than simply a matter of size; it is an opportunity to find insights in new and emerging types of data and content, to make useful decisions. Lately the number of dimensions increased to 4, 7 and even to 10 V's [13]–[15].

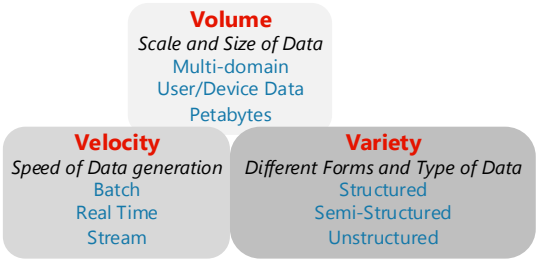

Figure 2. Big Data Original 3 V's

C. Big Data Lifecycle

As mentioned early and showed in depicted in figure 3, the Big Data ecosystem is organized as a value chain lifecycle from data inception to visualization. In the following, a brief description of all the main stages of Big Data lifecycle.

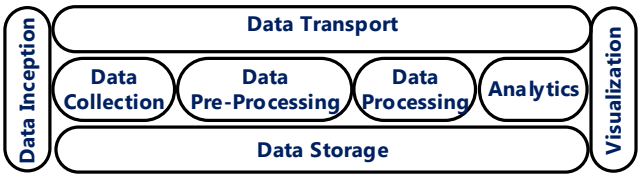

Figure 3 Big Data Lifecycle

1) *Data Generation/Inception:* is the phase where data is created, many data sources are responsible for these data: electrophysiology signals, sensors used to gather climate information, surveillance devices, posts to social media sites, videos and still images, transaction records, to name a few.

2) *Data Acquisition:* consists of data collection, data transmission, and data pre-processing [12][16] .

   - *Data Collection:* the data is gathered in specific data formats from different sources: real world data measurements using sensors and RFID, or data from any sources using a specific designed script to crawl the web.
   - *Data Transport:* to transfer the collected data into storage data centers using interconnected networks.
   - *Data Pre-Processing:* it consists of the typical pre-processing activities like *Data Integration, Enrichment, Transformation, Reduction, and Cleansing.*

3) *Data Storage:* the infrastructure data center where the data is stored and distributed among several clusters, data centers spread geographically. The storage systems ensure several fault tolerance levels to achieve reliability and efficiency.

4) *Data Processing & Analytics:* application of Data Mining algorithms, Machine Learning, Artificial Intelligence and Deep Learning to process the data and extract useful insight for better decision making. Data scientists are the most expected users of this phase since they have the expertise to apply what needed on what must be analyzed.

5) *Data Visualization:* the best way of assessing the value of processed data is to examine it visually and taking decision accordingly. Application of visualization methods in Big Data is of an importance as it closes the loop value chain.

D. Data Quality

Most studies in the area of Data Quality (DQ) are from database management research communities [1], [2]. According to [17], data quality is not easy to define, its definitions are data domain aware. In general, there is a consensus that data quality is always dependent on the quality of the data source [18].

1) *Definition:* it is recognized that DQ has many definitions that are related to the context, domain, area or the fields in which it is used [19], [20]. DQ is differently understood in academia than in industry. In [21], the authors summarized data quality from the most known and used definitions from ISO 25012 Standard. In the literature, data quality is "fitness for use". In [19], data quality is defined as the appropriateness for use or meeting user needs.

2) *Data Quality Dimensions(DQD's):* according to [19], [22], [23], a DQD offers a way to measure and manage data quality. There are several quality dimensions each of which is associated with a specific metrics. DQD's usually fall into four categories illustrated in Figure 4: intrinsic, contextual, and representational and accessibility [23]–[27]. For instance, the contextual dimensions are related to the information, and intrinsic refers to objective and native data attributes. Examples of intrinsic DQD's include:

- Accuracy: measures whether data was recorded correctly and reflect realistic values.
- Timeliness: measures whether data is up to date as data currency and volatility [28].
- Consistency: measures whether data agrees with its format and structure. Mostly the respect of data constraints.
- Completeness: describes whether all relevant data are recorded. It measures missing values for an attribute.

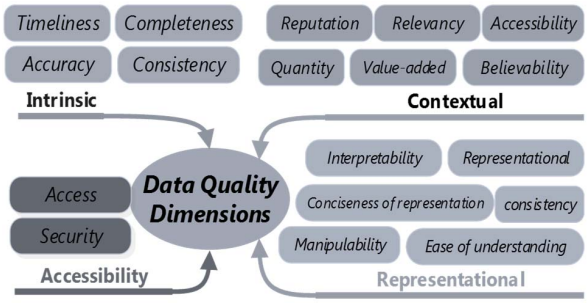

Figure 4. Data Quality Dimensions (Characteristics)

3) *Data Quality Metrics:* each DQD need to be quantified and measured. DQD metrics represents the steps to evaluate these dimensions. From simple formulas to more complex multivariate expressions, the metrics gives the measurability property to DQD's. For example the computation of missing values for an attribute is considered as the measure to rate the DQD completeness [29]–[31].

4) *Data Quality Evaluation:* following a data driven strategy requires handling quality evaluation on the already generated data. Hence, it is mandatory to measure and quantify the DQD. For structured or semi-structured data, the data is available as a set of attributes represented in columns or rows and respectively their values are recorded. Any data quality metric should specify whether the values of data respect the quality attributes (Dimensions). The author in [32], quoted that data quality measurement metrics tend to evaluate a binary results correct or incorrect (0 and 100% respectively), and use universal formulas to compute these attributes applied to quality dimensions (e.g. accuracy). The measurements generate DQD's scores using their related defined metrics (e.g. the accuracy score is the number of correct instances values divided by the total number of instances of the measured attributes).

E. Big Data Quality Evaluation

The importance of data quality in the Big Data lifecycle redefines the way data supervision is handled. Managing the data quality involves adding more functionalities in each stage with an ongoing quality control and monitoring to avoid quality failure during all phases of the lifecycle. Big Data quality evaluation is concerned about properties such as the performance, the value and the cost. In the next section, we illustrate more details on how quality must be managed and assessed in Big Data lifecycle.

F. Big Data Quality Issues

Data quality issues take place when quality requirements are not met on data values [33]. These issues are due to several factors or processes happened in different levels: 1) the data sources: unreliability, trust, data copying, inconsistency, multi-sources, and data domain, 2) the generation level: human data entry, sensors devices readings, social media, unstructured data, and missing values, and 3) the process and/or application level (acquisition: collection, transmission). The data pre-processing improves data quality by executing many tasks and activities like data transformation, integration, fusion, and normalization.

The authors in [21], [34] enumerate many causes of poor data that affect the data quality and came up with a list of elements that affects the data quality and its related dimensions. In [19], [34], The authors addressed a compilation of poor data causes classified by DQD, granularity level, and data source type while highlighting the causality mapping between these.

III. A HOLISTIC QUALITY MANAGEMENT MODEL FOR BIG DATA

In Big Data, the Data management is impartial to its quality management. Therefore, we need to identify the quality issues and requirements at each stage of the lifecycle. To ensure a high-quality value chain, an improvement procedure is inevitable and should be built-in within each process of the lifecycle. The goal is that Quality management activities should be undertaken without adding extra communication, processing and cost overhead on the different Big Data ecosystem layers.

To initiate any quality project for Big Data, a set of parameters and concepts must be enumerated to identify the processes

involved and to characterize the data and workflow type used as crucial inputs for its management. Two strategies are being perpetrated in data quality: *data driven* mentioned early in the paper and process driven strategies. The data driven approach act on the data itself, assess its quality with the objective of augmenting its quality. The *process driven* is a predictive approach that focus on the quality of the process used to generate or manipulate data.

In figure 5, we propose a holistic quality management model that captures important quality aspects and explore how to deal with Big Data quality throughout its lifecycle. We identified the processes that must handles and address data quality problems and provide a quality assessment schemes to ensure its effective management. The most important stages in this respecting order are: (1) data creation, (2) data source (3) data collection, (4) data transport, (5) data storage, (6) data preprocessing, (7) processing and analytics, and (8) visualization. Moreover, Big Data lifecycle stages and their related quality information's and processes must be addresses to achieve an end-to-end Big Data quality management driven lifecycle.

origins, first, we need to examine the existing data that is accumulated in the past and build a knowledge base from this data to better predict a redesign its creation for Big Data. In this case, the data design must be an iterative process that compromise data inspection and data quality evaluation (usually follows a data driven approach when using existing data) to enhance processes that create a high-quality data ready to be used by processes of the Big Data lifecycle.

A process driven strategy is adopted in the data creation phase, where quality constraints are set to eliminate bad data before it is created. Moreover, after data is created, a collection process is initiated to gather data into more structured and organized format, thus becomes useful data. The set of techniques used to reorganize data themselves ensure and guarantee the quality of data. The data collection processes are considered as data management processes that can positively impacts the data quality if it use the proper data science methods and techniques [35].

The frameworks and tools used to design data varies from

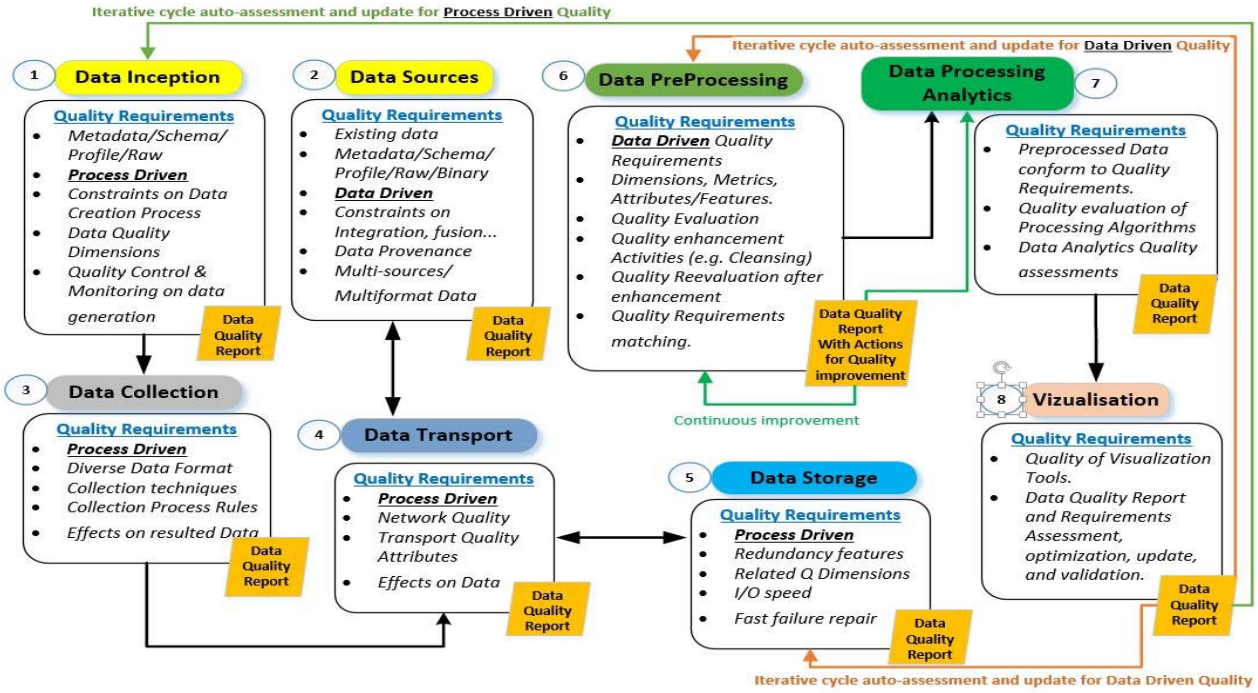

Figure 5. A Holistic Quality Management Model for Big Data Value Chain

In the following, we highlight the importance of quality requirements considerations, quality implementation and enforcement across different processes of the Big Data value chain. We also describe quality propagation across different processes of the value chain and the level of coordination between these processes for the sake of supporting quality management. Finally, we illustrate continuous improvement of quality through loopbacks and inter-processes interactivity.

A. Quality enabled data acquisition

The data acquisition is handled in two stages that are: data inception, and data collection. However, data sources (2 in Figure 5) is related to already existing data generally stored in many formats. Before addressing the Big Data quality at its

traditional database systems to more Big Data related systems like NoSQL databases (MongoDb, Cassandra, Hadoop HBase and other associated Hadoop ecosystem tools). A database design based on NoSQL will discourages the complicated tabular structure queries and lead to use connector tools.

B. Quality-aware data transport and storage

For data transport phase, it is very important to support QoS considering the requirements that should be met during service provisioning. Ensuring data quality is more specifically related to the underlying networks used to transmit data, and the security measures that guarantee the transmission of data between multiple points without data losses or corruptions. The quality of data networks is based on client's QoS requirements,

provider QoS offering and the real-time QoS measurements. The primary goal of QoS is to provide priority, including dedicated bandwidth, controlled jitter and latency (e.g. Big Data real-time stream processing), and improved loss characteristics. Moreover, SLA, proactive and reactive provisioning policy should be considered by providers to deliver the required QoS.

For data storage it is handled for Big Data quality through data distribution and replication. For example, storage using Hadoop ecosystem relies on several nodes that duplicate data to avoid any catastrophic data loss and ensure continuity when failure happens. Moreover, data storage quality relies on the storage medium I/O performances for several types of Big Data. Reading and writing ratios must follow the I/O requirements for each data type, e.g., HD Video data, and stream processing.

C. Quality driven data preprocessing

Preprocessing represents the last remedy to quality management in Big Data. It is the process of cleansing, integrating, normalizing, transforming, integrating data to improve it quality before being consumed by the processing stage. It follows a data driven approach that relies heavily on data values. For example, data cleansing relies on completeness, consistency. The quality settings (e.g. requirements, baseline/model, and quality report) are essential for the process to achieve data quality improvements. It identifies what quality is expected from the data, dimensions and scores. The model or baseline represents the basic quality requirements bounds and can be adjusted based on the quality evaluation process and the matching of quality with its requirements. Finally, the quality report similarly to data provenance, records the data path since its inception to its final analytics stage. The Quality report contents is iteratively updated and augmented with new parameters to serves as a Big Data quality repository.

Many data pre-processing tools has emerged, in addition others existing tools used in the database domain have been updated to handle Big Data. Mostly, Hadoop and spark based preprocessing tools and framework are widely used (e.g. Talend Open Studio for Big Data [36], and Open Refine [37]).

D. Quality enabled data processing and analytics

Processing and analytics take advantages from the quality assessments undertaken in the previous quality assessment activities. In the processing stage, quality management consists of validating that preprocessed data is conform to processing quality requirements, in addition, to the quality evaluation of processing algorithms (e.g. Hadoop based processing). However, analytics consist of assessing different analytics schemes and algorithms (e.g. deep learning) to ensure the validity of quality evaluation before visualization of insights retrieved from analytics.

As for Big data preprocessing, the processing frameworks are Hadoop and spark related. In [38], many frameworks and methods for the preprocessing and processing of have been detailed and compared.

E. Quality enabled data visualization

It consists of aggregating the visualization tools quality requirements, the quality report, and the quality of diverse visualization tools. It also assesses the visualized data that leads to accurate decision. This stage is very important as it is the last process in closing the quality assessment loop in the Big Data value chain and triggers some improvements recommendations for continuous quality monitoring. Many visualization tools for Big Data are used like Google Chart, Tableau, DataWrapper.

F. Quality propagation and continous quality improvement

The quality management model proposed in figure 5 illustrates how quality management activities propagate across the different processes of the Big Data value chain. At each stage, a quality requirement is taken into considerations, quality measurements are reflected, and a quality report is generated. This quality report is forwarded to the next stage of the value chain to track, validate, enrich, and enhance the quality of both the data and the process. A continuous quality improvement is possible through loopback between processes to monitor, revise, and enhance quality whenever required. This is a very important feature that ensure accuracy and improvement of quality assessment across the Big Data value chain.

IV. Big Data Quality: Research Classification

As shown in Figure 6, we have selected numerous papers that mainly address, and debated the data quality in Big Data. Some of them went deeply into specific quality properties, however, others addressed Big Data quality issues, and proposed some solutions based on quality assessment, improvement while applying diverse techniques. The designated literature reveals that quality of data has been addressed from diverse perceptions (e.g. the data, processes, applications, and its management) with focus on quality properties assessment, quality evaluation processes, and quality models related to Big Data.

		2017 [42]	2016 [43]	2016 [50]	2016 [44]	2015 [45]	2015 [52]	2016 [61]	2016 [46]	2016 [59]	2016 [51]	2016 [39]	2016 [35]	2017 [60]	2016 [47]	2017 [40]	2015 [53]	2016 [54]	2015 [55]	2016 [48]	2015 [49]	2014 [25]	2017 [56]	2015 [57]	2017 [58]	2016 [41]	2015 [31]
Big Data Storage	I											X	X			X									X		
Big Data Pre-Processing	II	X	X										X		X	X					X						
Big Data Processing & Analytics	III	X	X		X	X			X			X	X		X	X				X							
Big Data Characteristics: V's	IV						X		X	X	X	X	X		X	X	X	X	X		X			X	X		X
Big Data Management	V			X	X				X	X	X		X		X	X			X	X	X		X		X	X	
Big Data Problems, Data Quality Issues	VI		X			X	X		X	X	X			X			X	X	X	X	X	X			X	X	X
Data Quality Dimensions and Metrics	VII	X	X				X	X	X	X	X						X	X	X	X	X	X	X		X	X	X
Data Quality Assessment	VIII									X	X	X	X						X						X	X	X
Data Quality Management	IX	X		X	X	X	X	X	X	X	X	X					X		X					X			X
Big Data Applications and Quality Improvement	X	X	X		X	X				X					X				X	X	X	X					

Figure 6. Big Data Quality Classification

The goal of this classification is to extract major research trends related to Big Data quality, identify what have been addressed so far in its quality management, and what need further explorations to reach its full potential. Also, we projected these findings to draw the path towards future research directions. We come-up with 10 classes (from *I* to *XI*) that identify the key research trends in Big Data Quality. These categories are also grouped into 6 main clusters (*A* to *F*) that define the most important areas of interest tied with data quality and Big Data.

A. Big Data Value Chain (I-II-III)

In [35], [39]–[49], the authors emphasized many data quality issues and their effects in Big data lifecycle stages; as preprocessing, processing, and storage. They mostly proposed a management process flow as taxonomy to control quality of pre-processing and processing for verification of data quality, data sources and data formats before analytics. Others proposed a combined quality model to detect preprocessing defects and act upon to correct tasks flaws.

For data processing, some studies leveraged analytics techniques, machine learning and classification practices and evaluate their suitability for Big data. Then, they analyzed the effects of data size to evaluate the accuracy on the applications of these methodologies. Others proposed to adapt these techniques to handle the Big Data characteristics and tackles different processing quality issues.

As storage is crucial for Big Data, authors addressed the multi-storage providers and its impact on the performance and efficiency of pre-processing, processing when high data distribution on multiple cloud providers is adopted. Most of the previous works, addressed quality in Big data in an ad hoc manner while not following a comprehensive model that considers the quality characteristics, the processes, and the underlying infrastructure. Such model will assure an end-to-end quality management in the value chain. Most of the works addressed separately Big data stages without leveraging their effects through its whole lifecycle.

B. Big Data Management & Characteristics (IV-V)

In this category, two important aspects of Big Data were targeted in [31],[35],[39]–[41],[44],[46]–[59] that are its characteristics (V's) and management. Most of the authors agreed that the V's represent a significant aspect of Big Dat and have a high impact on data quality when it is not well managed in an efficient way. They typically linked the V's with DQD's in order to find correlations and impact on each other. Others emphasized that the scalability, integrity and resource optimization is highly proportional to Big Data V's as they represents the key elements for its Mangement solutions.
In the Big Data management, authors surveyed and proposed management models tackling storage, pre-processing, and processing. Moreover, an up-to-date review of techniques and methods for each process involved in the management processes were carefully studied. Managing Big Data requires a mapping of its value chain stages with its related processes and sub-processes. The importance of the Quality in Big Data Management was not generally addressed. Such a Framework with end-to-end Quality management and enforcement is very challenging.

C. Big Data Problems & Data Quality Issues (VI)

Congregating Big Data problems with Data quality issues is justified by the strong relationship between these two concepts. Any data quality issues will be reflected in the analytics. The authors of selected literature [25], [31], [41], [43], [45], [46], [48], [49], [51]–[56], [58]–[60], have stressed that it is very important to discover quality issues and map them with Big data problems in the lifecycle as early as possible. This will help isolate and adapt the processes that must handle both concerns. Most of the data quality issues have been addresses heavily in the research community and yet it is still not adapted in Big Data. Further discussions related the DQDs and V's were considered proportional.

D. Data Quality (VII-VIII-IX):

Data quality has been fully investigated in the following work [25], [31], [39], [41]–[46], [48]–[59], [61] which confirm its importance for Big Data. Most of authors, consider the DQD's as a crucial model to use. They developed a Quality model for Big Data based on DQD's mapped with V's to address scalability and reliability issues. Others addressed the problem of choosing DQD's and metrics for unusual data as images, binary data, and unstructured data. They extracted features and combined many DQD's to measure a quality dimension score.
There are many challenges that need to be tackled to come up with a quality assessment scheme for Big Data. The authors listed many functionalities that must be considered: e.g. accuracy, consistency, provenance, uncertainty.
Accordingly, there is no complete reference model for Data Quality and its Management in Big Data. DQM must ensure data conformity and follows all the steps since its inception to quality assessment. In other words, new technological and decision-making challenges are making DQM applications more complicated.

E. Big Data Applications and Quality Improvements (X)

It is a value chain-based application that follows stages from data creation to visualization. Some of the authors focus on how to manage the quality within these applications by evaluating metrics for resource management like storage and processing. Others, proposed solutions to enhance quality of the data while applying cleansing tasks and activities that are parts of pre-processing (e.g. BigDansing, and Nadeef) [51]– [53], [56], [68]. Data quality consists of assessing quality dimensions, along with combining all the above to aggregate one quality score that reflects the quality of Big Data Lifecycle applications.

V. DISCUSSION AND FUTURE DIRECTIONS

Ensuring Quality is recognized as one of the most challenging issues in Big Data era. Current approaches and solutions emerged both from academia and industry that tackled quality have not reached yet a convincing level of maturity. Evaluate the importance of assessing quality of Big Data versus the value it generates for its users (e.g. governments, businesses) is of paramount importance. In addition, following well studied data quality management plan, using the right assessment scheme, adopting the appropriate quality measurement approaches, and utilizing the suitable tools and platforms to conduct different quality evaluation activities all together will help achieving

high quality assessment results. Furthermore, addressing quality across the Big Data value chain enforces an end-to-end quality evaluation and management and leads to better quality improvements. Finally, evaluating the overhead of quality assessment guarantees a cost-effective quality management processes.

Based on the above, future research directions in Big Data quality should be geared towards the development of solutions that consider the following:

a) Assessment of quality as earlier as possible and it end-to-end integration across its.
b) Implementation of continuous quality improvement and enforcement mechanisms in Big Data quality management.
c) Specification of Big Data Quality metrics that should cope with the data dynamic nature and its unconventional characteristics.
d) Development of new quality dimensions with specific measurement attributes for unstructured, and schema less data.
e) Enforcement of quality requirements, generation of quality reports and feedbacks to support assessment activities.
f) Development of more Online automated real-time dashboards for Big data quality monitoring.
g) Application of higher degree of statistical proof in different Big data quality evaluation processes including sampling, regression, correlation, and matching.
h) Development of effective quality outcomes prediction.
i) Evaluation of quality of a representative set of data samples then generate a quality model to apply on the whole Data. This will get a glimpse of the data quality and proceed with the equality results applied on all the data [45]– [47].

Finally, it is worth mentioning that research work and solutions on Big Data quality are still in its preliminary phase, and there is much to do in its development and standardization. It is a multidisciplinary, complex, and multi-variant domain where new assessment, processing techniques, and analytics algorithms, storage technologies and processing platforms will play a great role in the development and the maturation of this active research area. We anticipate that researchers from academia will contribute to the development of new Big data quality approaches, algorithms, optimizations techniques that go beyond the traditional ones used in databases and data warehouses. However, industries will lead development initiatives of new platforms, solutions, and technologies that support end-to-end quality management within the Big Data lifecycle.

VI. Conclusion

Big Data has emerged as new paradigm for handling huge, continuous, varying, and complex data. Its quality is the key for its acceptance and usefulness. A poor data quality might lead to severe consequences. This will lose the benefit of analyzing and exploring large-scale data sets in an appropriate way. Using conventional techniques to manage Big Data is not any more appropriate. Therefore, the design and application of efficient approaches to manage the quality is highly demanded. In this paper, we identified the key research challenges in Big Data quality and we highlighted their importance. We then surveyed classified and discussed the most comprehensive research initiatives. Afterwards, we proposed a holistic view of Big Data quality management model that emphasized the key quality assessment activities to be conducted across the value chain. Finlay, we discussed the main tendencies in Big Data quality assessment and we point to some future research directions. We are planning to further extend the scope of this work and deeply describe how quality assessments activities can be implemented in a context of a real Big Data project and where quality matters.

References

[1] P. Z. Yeh and C. A. Puri, "An Efficient and Robust Approach for Discovering Data Quality Rules," in *2010 22nd IEEE International Conference on Tools with Artificial Intelligence (ICTAI)*, 2010, vol. 1, pp. 248–255.

[2] F. Chiang and R. J. Miller, "Discovering data quality rules," *Proc. VLDB Endow.*, vol. 1, no. 1, pp. 1166–1177, 2008.

[3] "IBM - What is big data?" [Online]. Available: http://www-01.ibm.com/software/data/bigdata/what-is-big-data.html. [Accessed: 30-May-2016].

[4] "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-May-2016].

[5] D. Laney, "The importance of'Big Data': A definition," *Gart. Retrieved*, vol. 21, pp. 2014–2018, 2012.

[6] J. Manyika *et al.*, "Big data: The next frontier for innovation, competition, and productivity," *McKinsey Glob. Inst.*, pp. 1–137, 2011.

[7] A. De Mauro, M. Greco, and M. Grimaldi, "What is big data? A consensual definition and a review of key research topics," in *AIP conference proceedings*, 2015, vol. 1644, pp. 97–104.

[8] I. Emmanuel and C. Stanier, "Defining Big Data," in *Proceedings of the International Conference on Big Data and Advanced Wireless Technologies*, New York, NY, USA, 2016, p. 5:1–5:6.

[9] J. S. Ward and A. Barker, "Undefined by data: a survey of big data definitions," *ArXiv Prepr. ArXiv13095821*, 2013.

[10] P. Zikopoulos and C. Eaton, "Understanding Big Data: Analytics for Enterprise Class Hadoop and Streaming Data," 2011.

[11] 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: 29-Nov-2017].

[12] H. Hu, Y. Wen, T.-S. Chua, and X. Li, "Toward Scalable Systems for Big Data Analytics: A Technology Tutorial," *IEEE Access*, vol. 2, pp. 652–687, 2014.

[13] P. Géczy, "Big data characteristics," *Macrotheme Rev.*, vol. 3, no. 6, pp. 94–104, 2014.

[14] M. Ali-ud-din Khan, M. F. Uddin, and N. Gupta, "Seven V's of Big Data understanding Big Data to extract value," in *American Society for Engineering Education (ASEE Zone 1), 2014 Zone 1 Conference of the*, 2014, pp. 1–5.

[15] "Big Data Technology with 8 V′s," *M-Brain*. .

[16] M. Chen, S. Mao, and Y. Liu, "Big Data: A Survey," *Mob. Netw. Appl.*, vol. 19, no. 2, pp. 171–209, 2014.

[17] P. Oliveira, F. Rodrigues, and P. R. Henriques, "A Formal Definition of Data Quality Problems.," in *IQ*, 2005.

[18] M. Maier, A. Serebrenik, and I. T. P. Vanderfeesten, *Towards a Big Data Reference Architecture*. University of Eindhoven, 2013.

[19] F. Sidi, P. H. Shariat Panahy, L. S. Affendey, M. A. Jabar, H. Ibrahim, and A. Mustapha, "Data quality: A survey of data quality dimensions," in *2012 International Conference on Information Retrieval Knowledge Management (CAMP)*, 2012, pp. 300–304.

[20] I. Caballero and M. Piattini, "CALDEA: a data quality model based on maturity levels," in *Third International Conference on Quality Software, 2003. Proceedings*, 2003, pp. 380–387.

[21] M. Chen, M. Song, J. Han, and E. Haihong, "Survey on data quality," in *2012 World Congress on Information and Communication Technologies (WICT)*, 2012, pp. 1009–1013.

[22] P. Glowalla, P. Balazy, D. Basten, and A. Sunyaev, "Process-Driven Data Quality Management – An Application of the Combined Conceptual Life Cycle Model," in *2014 47th Hawaii International Conference on System Sciences (HICSS)*, 2014, pp. 4700–4709.

[23] C. Batini, C. Cappiello, C. Francalanci, and A. Maurino, "Methodologies for data quality assessment and improvement," *ACM Comput. Surv.*, vol. 41, no. 3, pp. 1–52, Jul. 2009.

[24] B. T. Hazen, C. A. Boone, J. D. Ezell, and L. A. Jones-Farmer, "Data quality for data science, predictive analytics, and big data in supply chain management: An introduction to the problem and suggestions for research and applications," *Int. J. Prod. Econ.*, vol. 154, pp. 72–80, 2014.

[25] B. Saha and D. Srivastava, "Data quality: The other face of Big Data," in *2014 IEEE 30th International Conference on Data Engineering (ICDE)*, 2014, pp. 1294–1297.

[26] C. Cappiello, A. Caro, A. Rodriguez, and I. Caballero, "An Approach To Design Business Processes Addressing Data Quality Issues," 2013.

[27] Y. Wand and R. Y. Wang, "Anchoring data quality dimensions in ontological foundations," *Commun. ACM*, vol. 39, no. 11, pp. 86–95, 1996.

[28] W. Fan, F. Geerts, and J. Wijsen, "Determining the currency of data," *ACM Trans. Database Syst. TODS*, vol. 37, no. 4, p. 25, 2012.

[29] V. Goasdoué, S. Nugier, D. Duquennoy, and B. Laboisse, "An Evaluation Framework For Data Quality Tools.," in *ICIQ*, 2007, pp. 280–294.

[30] M. A. Serhani, H. T. E. Kassabi, I. Taleb, and A. Nujum, "An Hybrid Approach to Quality Evaluation across Big Data Value Chain," in *2016 IEEE International Congress on Big Data (BigData Congress)*, 2016, pp. 418–425.

[31] D. Firmani, M. Mecella, M. Scannapieco, and C. Batini, "On the Meaningfulness of 'Big Data Quality' (Invited Paper)," in *Data Science and Engineering*, Springer Berlin Heidelberg, 2015, pp. 1–15.

[32] H. M. Sneed and K. Erdoes, "Testing big data (Assuring the quality of large databases)," in *2015 IEEE Eighth International Conference on Software Testing, Verification and Validation Workshops (ICSTW)*, 2015, pp. 1–6.

[33] C. Fürber and M. Hepp, "Towards a Vocabulary for Data Quality Management in Semantic Web Architectures," in *Proceedings of the 1st International Workshop on Linked Web Data Management*, New York, NY, USA, 2011, pp. 1–8.

[34] N. Laranjeiro, S. N. Soydemir, and J. Bernardino, "A Survey on Data Quality: Classifying Poor Data," in *2015 IEEE 21st Pacific Rim International Symposium on Dependable Computing (PRDC)*, 2015, pp. 179–188.

[35] A. Siddiqa *et al.*, "A survey of big data management: Taxonomy and state-of-the-art," *J. Netw. Comput. Appl.*, vol. 71, pp. 151–166, Aug. 2016.

[36] "Big Data: Talend Big Data Integration Products & Services." [Online]. Available: https://www.talend.com/products/big-data/. [Accessed: 30-Jan-2018].

[37] "OpenRefine." [Online]. Available: http://openrefine.org/. [Accessed: 30-Jan-2018].

[38] S. García, S. Ramírez-Gallego, J. Luengo, J. M. Benítez, and F. Herrera, "Big data preprocessing: methods and prospects," *Big Data Anal.*, vol. 1, no. 1, Dec. 2016.

[39] S. Li *et al.*, "Geospatial big data handling theory and methods: A review and research challenges," *ISPRS J. Photogramm. Remote Sens.*, vol. 115, pp. 119–133, May 2016.

[40] A. Oussous, F.-Z. Benjelloun, A. Ait Lahcen, and S. Belfkih, "Big Data technologies: A survey," *J. King Saud Univ. - Comput. Inf. Sci.*, Jun. 2017.

[41] M. Scannapieco and L. Berti, "Quality of Web Data and Quality of Big Data: Open Problems," in *Data and Information Quality*, Springer, Cham, 2016, pp. 421–449.

[42] S.-T. Lai and F.-Y. Leu, "Data Preprocessing Quality Management Procedure for Improving Big Data Applications Efficiency and Practicality," in *Advances on Broad-Band Wireless Computing, Communication and Applications*, vol. 2, L. Barolli, F. Xhafa, and K. Yim, Eds. Cham: Springer International Publishing, 2017, pp. 731–738.

[43] K. Sharma and others, "Quality issues with big data analytics," in *Computing for Sustainable Global Development (INDIACom), 2016 3rd International Conference on*, 2016, pp. 3589–3591.

[44] J. Ding, D. Zhang, and X. H. Hu, "A Framework for Ensuring the Quality of a Big Data Service," in *2016 IEEE International Conference on Services Computing (SCC)*, 2016, pp. 82–89.

[45] D. Becker, B. McMullen, and T. D. King, "Big data, big data quality problem," in *2015 IEEE International Conference on Big Data (Big Data)*, 2015, pp. 2644–2653.

[46] A. F. Haryadi, J. Hulstijn, A. Wahyudi, H. Van Der Voort, and M. Janssen, "Antecedents of big data quality: An empirical examination in financial service organizations," in *Big Data (Big Data), 2016 IEEE International Conference on*, 2016, pp. 116–121.

[47] M. H. ur Rehman, V. Chang, A. Batool, and T. Y. Wah, "Big data reduction framework for value creation in sustainable enterprises," *Int. J. Inf. Manag.*, vol. 36, no. 6, pp. 917–928, Dec. 2016.

[48] J. Gao, C. Xie, and C. Tao, "Big Data Validation and Quality Assurance – Issuses, Challenges, and Needs," in *2016 IEEE Symposium on Service-Oriented System Engineering (SOSE)*, 2016, pp. 433–441.

[49] Z. Khayyat *et al.*, "Bigdansing: A system for big data cleansing," in *Proceedings of the 2015 ACM SIGMOD International Conference on Management of Data*, 2015, pp. 1215–1230.

[50] J. Merino, I. Caballero, B. Rivas, M. Serrano, and M. Piattini, "A Data Quality in Use model for Big Data," *Future Gener. Comput. Syst.*, vol. 63, pp. 123–130, Oct. 2016.

[51] G. A. Lakshen, S. Vraneš, and V. Janev, "Big data and quality: A literature review," in *2016 24th Telecommunications Forum (TELFOR)*, 2016, pp. 1–4.

[52] N. Abdullah, S. A. Ismail, S. Sophiayati, and S. M. Sam, "Data quality in big data: a review," *Int. J. Adv. Soft Comput. Its Appl.*, vol. 7, no. 3, 2015.

[53] C. Batini, A. Rula, M. Scannapieco, and G. Viscusi, "From data quality to big data quality," *J. Database Manag.*, vol. 26, no. 1, pp. 60–82, 2015.

[54] J. Liu, J. Li, W. Li, and J. Wu, "Rethinking big data: A review on the data quality and usage issues," *ISPRS J. Photogramm. Remote Sens.*, vol. 115, pp. 134–142, May 2016.

[55] A. Immonen, P. Paakkonen, and E. Ovaska, "Evaluating the Quality of Social Media Data in Big Data Architecture," *IEEE Access*, vol. 3, pp. 2028–2043, 2015.

[56] D. Brown, "Encyclopedia of Big Data," in *Encyclopedia of Big Data*, L. A. Schintler and C. L. McNeely, Eds. Cham: Springer International Publishing, 2017, pp. 1–3.

[57] "A Suggested Framework for the Quality of Big Data." [Online]. Available: https://statswiki.unece.org/display/bigdata/2014+Project. [Accessed: 11-Nov-2017].

[58] P. Pääkkönen and J. Jokitulppo, "Quality management architecture for social media data," *J. Big Data*, vol. 4, no. 1, p. 6, Dec. 2017.

[59] M. Kläs, W. Putz, and T. Lutz, "Quality Evaluation for Big Data: A Scalable Assessment Approach and First Evaluation Results," in *2016 Joint Conference of the International Workshop on Software Measurement and the International Conference on Software Process and Product Measurement (IWSM-MENSURA)*, 2016, pp. 115–124.

[60] M. Janssen, H. van der Voort, and A. Wahyudi, "Factors influencing big data decision-making quality," *J. Bus. Res.*, vol. 70, pp. 338–345, Jan. 2017.

[61] P. Ciancarini, F. Poggi, and D. Russo, "Big Data Quality: A Roadmap for Open Data," in *2016 IEEE Second International Conference on Big Data Computing Service and Applications (BigDataService)*, 2016, pp. 210–215.

[62] I. Taleb, H. T. E. Kassabi, M. A. Serhani, R. Dssouli, and C. 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)*, 2016, pp. 759–765.

[63] I. Taleb and M. A. Serhani, "Big Data Pre-Processing: Closing the Data Quality Enforcement Loop," in *2017 IEEE International Congress on Big Data (BigData Congress)*, 2017, pp. 498–501.

[64] I. Taleb, R. Dssouli, and M. A. Serhani, "Big Data Pre-processing: A Quality Framework," in *2015 IEEE International Congress on Big Data (BigData Congress)*, 2015, pp. 191–198.