Big Data and Quality: A Literature Review

Guma Abdulkhader Lakshen, Sanja Vraneš, *Member, IEEE,* and Valentina Janev, *Member, IEEE,*

***Abstract* — Big Data refers to data volumes in the range of Exabyte (10^{18}) and beyond. Such volumes exceed the capacity of current on-line storage and processing systems. With characteristics like volume, velocity and variety big data throws challenges to the traditional IT establishments. Computer assisted innovation, real time data analytics, customer-centric business intelligence, industry wide decision making and transparency are possible advantages, to mention few, of Big Data. There are many issues with Big Data that warrant quality assessment methods. The issues are pertaining to storage and transport, management, and processing. This paper throws light into the present state of quality issues related to Big Data. It provides valuable insights that can be used to leverage Big Data science activities.**

***Keywords* —Big Data, Quality assessment, stream processing, survey, Big Data frameworks.**

I. Introduction

THE term "Big Data" originally meant the volume of data that could not be processed (efficiently) by traditional database methods and tools [1]. It is a term applied to a new generation of software, applications, system, storage and architecture, all designed to derive business value from unstructured data. Big data has become an indispensable area of research as it is expected to add big value to enterprises in the real world. In this paper we report on a study related to "Big Data" challenges, where "quality" is assumed to be one of them [2].

A simple Web search of the terms "Data quality" through any search engine returns over twelve millions pages, which clearly indicates the importance of data quality and its issues. Data quality issues evolved from traditional structured data managed relational databases to Big Data. Advanced tools, software, and systems are required to capture, store, manage, and analyze the data sets, all in a timeframe that preserves the intrinsic value of the data. Big data refers to environment in which data sets have grown too large to be handled, managed, stored and retrieved in an acceptable timeframe.

In literature, there is no unified definition of "Big Data", it has been defined differently in technological, industrial, research or academic perspectives [5]. For instance, "Big Data" was defined as "data quality is consistently meeting knowledge worker and end-customer expectations" [3], but also as "data that fit for their intended uses in operations, decision making, and planning" [4].

Data quality dimensions and issues are not yet well explored and understood for Big Data, and may be very different from the quality dimensions for traditional sized data sets. This paper primarily refers to the challenge of quality in Big data and the functionalities needed for Big data processing.

Corresponding Guma Abdulkhader Lakshen is a PhD student at the School of Electrical Engineering, University of Belgrade, Bul. Kralja Aleksandra 73, 11120 Belgrade, Serbia.

Sanja Vraneš is with the Mihajlo Pupin Institute, Volgina 15, 11060 Belgrade, Serbia.

Valentina Janev is with the Mihajlo Pupin Institute, Volgina 15, 11060 Belgrade, Serbia.

II. Literature Review on Big Data

A. Characteristics of Big Data

Generally, Big data is considered as structured, semi structured, and unstructured datasets with massive data volumes (Terabytes and above) that cannot be easily captured, stored, manipulated, analyzed, managed and presented by traditional hardware, software and database management technologies.

Big data is usually described by its characteristics. Laney [6] was the first to propose three dimensions that characterize the challenges and opportunities of increasing large data volumes: Volume, Velocity and Variety, known as the (3V's) of data. Additional V's of data have been added over the years. It is apparent that defining big data and its characteristics will be an ongoing endeavour, but it nevertheless will not have negative impact on big data handling and processing. While 3V's have been continuously used to describe big data, the additional dimensions of Veracity and value have been added to describe data integrity and quality to become what is known as 5 V's of big data [7]. A brief description of the characteristics of the V's of big data is given in Table 1.

Suthaharan [8] even argued that the first three V's (*volume, velocity, and variety*) cannot support early detection of big data characteristics for its classification and proposed 3Cs: *cardinality*, *continuity*, and *complexity*.

B. Importance of Big Data

The world has recognized the importance of Big Data. In August 2010, President Barrack Obama announced the *"Transparency and Open Government"* in the "Memorandum for the Heads of Executive Departments and Agencies" proclaiming that "Big Data is a national challenge and priority along with healthcare and national security" [9]. The National Science Foundation, the National Institutes of Health, the U.S. Geological Survey, the Departments of Defense and Energy, and the Defense Advanced Research Projects Agency announced a joint R&D initiative in March 2012 that will invest more than $200 million to develop new big data tools and

978-1-5090-4086-5/16/$31.00 ©2016 IEEE

techniques. Its goal is to advance our "…understanding of the technologies needed to manipulate and mine massive amounts of information; apply that knowledge to other scientific fields "as well as address the national goals in the areas of health energy defense, education and researcher" [10]. Value arises from the ability to analyze the data to develop actionable information. From the literature survey, we have identified five generic ways that are illustrated in Fig. 1.

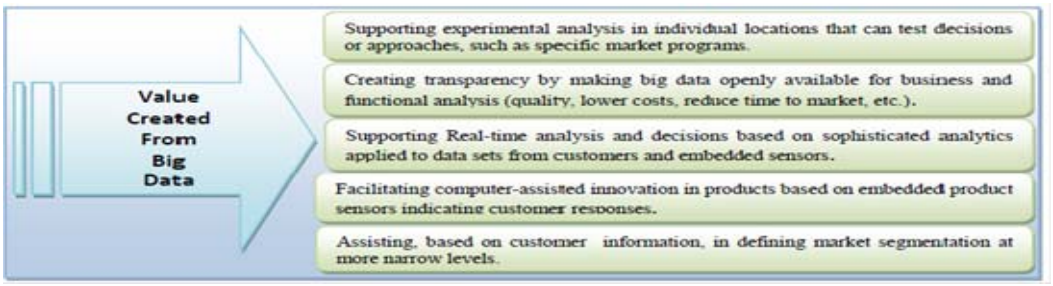

Fig. 1. Created value by the use of Big data

TABLE 1: CHARACTERISTICS OF BIG DATA.

Volume	Measures the **amount** of data available to an organization, which does not necessarily have to own all of it as long as it can access it.
Velocity	Measures the **speed** of data creation, streaming and aggregation (see Panasas® Parallel Storage for Big Data Applications, November 2012)
Variety	Measures the **richness** of the data representation – text, images, video, audio, etc.
Veracity	Measures the **understandability** of the data – biases, noise, abnormality, etc.
Value	Measures the **usefulness** of data in making decisions.

TABLE 2: RESEARCH ISSUES AND APPLICATION DOMAINS.

Project	**Research**	**Applications / Case Studies (CS)**
https://www.big-data-europe.eu	(storage (Hive, Cassandra), message passing (Kafka, Flume), multi-purpose data processing and analysis (Apache Hadoop, Apache Spark, Apache Flink), and publishing (Geotriples))	• Test generic infrastructures is found in the **Health domain**; Drug discovery; System monitoring in wind energy production unit; Viticulture; Crowd-sourcing in transport; aggregation platform in the transport sector; Climate pilot, **Secure Societies**.
http://byte-project.eu/	(Elements of social impact; Evaluating and addressing positive and negative externalities; Foresight analysis; Road-mapping; The big data community; Stakeholder engagement; Dissemination; Project management.)	• Environment (Earth and space observation portals and associated initiatives; Utilities / smart cities); Cultural Heritage; **Health CS**: A Genetic Research Initiative ; Transport CS: Shipping industry stakeholder; **Crisis informatics** CS: A Research Institute for Crisis Computing .
http://optique-project.eu/	(Real time stream processing; Query evaluation with Elastic Cloud; End-user oriented Query interface.)	• **Health care**, Energy.

II. LITERATURE REVIEW ON DATA QUALITY

Due to the relevance of data quality, its nature, and the variety of data types and information systems, achieving data quality is a complex, multidisciplinary area of investigation [14].

A. Main issues

Data quality involves several research topics and real-life application areas. Table 2 shows the research issues and application domains discussed in data quality literature and EU research projects.

The most common research issues discussed are *models*, *techniques*, *tools*, *frameworks* and *methodologies* in addition to *dimensions* which are briefly described in Section III.B:

Models are mainly used in databases to represent data and data schemas, as well as in information systems to represent business processes of organizations [11].

Techniques refer to algorithms, heuristics, knowledge-based procedures and learning processes that help to identify and solve a DQ related problem.

Methodologies provide guidelines to choose appropriate techniques or tools as the effective way of DQ measurement and improvement processes.

Tools and frameworks: A *tool* is a software procedure designed, automated and provided with an interface to evaluate the DQ activities.

A framework consists of a suite of coherent tools for a domain or task field. Furthermore, data quality is getting more attention in diverse application domains in recent years: E-government, Life Sciences, Web data, and Health care, etc. However here we discuss the DQ in web data contexts.

B. Data quality assessment dimensions

Data quality dimension is a term used by data management professionals to describe a feature of data that can be measured or assessed against defined standards in order to determine the quality of data [12]. It is also used to describe the measure of the quality of data. The key data quality dimensions are not universally agreed yet [13], however, the core data quality dimensions are shown in Fig. 2. The dimensions are defined as follows:

Completeness: The proportion of stored data against the potential of "100% complete"

Uniqueness: Nothing will be recorded more than once based upon how that thing is identified.

Timeliness: The degree to which data represent reality from the required point in time.

Fig. 2. Data quality dimensions.

Validity: Data are valid if it conforms to the syntax (format, type, range) of its definition.

Accuracy: The degree to which data correctly describes the "real world" object or event being described.

Consistency: The absence of difference, when comparing two or more representations of a thing against a definition.

There are additional factors which can have an impact on the effective use of data in addition to the above six dimensions such as *Usability, Flexibility, Confidentiality, and Value Timing issues of the data* [15,16].

There are several conditions that contributed to problems of data quality such as lack of validation routines [17], data valid, but not correct [18], mismatched syntax, formats, and structures [19], unexpected changes in source system, spiderweb of interfaces, lack of referential integrity checks, poor system design and data conversion errors. According to TDWI's Data Quality Survey [20], 40% of the surveyed companies have suffered losses, problems, or costs due to poor quality data and 43% not yet studied the issue. The most common problems caused by poor quality data, reported by of 286 respondents, are 'extra time required to reconcile data' 87%, 'loss of credibility in the system or application' 81%, 'extra costs' 72%, 'customer dissatisfaction' 67% and delay in deploying a new system 64%.

On the other hand, companies invested in improving management of data quality gained tangible and intangible benefits, usually opposes problems mentioned above [1]. Some of benefits from high quality data are 'single version of the truth' 19%, 'increased customer satisfaction' and 'greater confidence in analytical systems' 17%.

According to a survey of top managers at 600 firms, the report found that about 60% had cut their processing costs, more than 40% boosted sales through better analysis of customer data, and more than 30% had won a significant contract through better analysis of data. According to the *PriceWaterhouseCoopers* Global Data Management Survey 2001, companies that manage their data as a strategic resource and invest in its quality are already pulling ahead in terms of reputation and profitability from those that fail to do so.

III. The Challenges of Data Quality

Enterprises use big data usually confronted with many challenges such as heterogeneity and incompleteness**,** diversity of data sources, huge data volume, short data timeline, non-existing and approved data quality standards, lack of structure, error-handling, privacy, timeliness, provenance, and visualization. These technical challenges are common across a large variety of application domains, and therefore not cost-effective to address in the context of one domain alone. Currently, comprehensive analysis and research of quality standards and quality assessment methods for big data are lacking [2].

Based on our analysis of case studies elaborated in EU research projects (see Table 2), we could select example scenarios such as *Natural disaster affecting several states in Europe* (S1), *Aircraft accidents during the landing phase that cause closing the runway* (S2), and *Data/ Document/ Information exchange in public sector administration of European member states e.g. Health Information Exchange* (S3) and link them to quality assessment challenges and functionalities required for building Big Data applications (see Table 3).

IV. Analysis of Big Data Frameworks

There is a number of diverse big data frameworks available and also used in existing Big Data applications, therefore we'll concentrate on the most popular streaming solutions. Five frameworks were selected for analysis, the summary of which is presented in the following Table 4 (Source: P. Zapletal, Comparison of Apache Stream Processing Frameworks, www.cakesolutions.net).

Streaming applications and stream processors are very diverse. There are two main runtime designs, Dataflow based (Storm, Samza, Flink, Apex) and Micro-batch based (Spark).

Storm is a pioneer of real-time analytics, distributed dataflow abstraction with low-level control, time windowing, and state introduced recently, whereas *Spark* has a unified batch and stream processing over a batch runtime, good integration with batch programs, but lags behind recent. *Flink* is a leader of open source streaming innovation and is a unified batch and stream processing over dataflow engine which makes it highly flexible, robust stateful, and windowing computations. *Samza* builds heavily on *Kafka*'s log based philosophy and has a pluggable components, but runs best with *Kafka*, and it's time windowing is basic. Streaming advancements but evolving quickly as with *Spark 2.0* comes with new streaming engine. Lastly, *Apex* is a native streaming engine built natively on *YARN* and has advanced partitioning support with locality optimizations.

Social platforms such as Facebook, Twitter, LinkedIn are cloud service providers for person-to-person communication. Designing these platforms can be performed by SQL, NoSQL, Cache Augmented SQL, graph databases and many more tools. Each platform has it's own style of representation such as tabular, while others use different approach, which leads to different merits of each. Comparing these platforms and their merits requires **using benchmarks** such as BG.

BG is a benchmark (see http://bgbenchmark.org/BG/) to evaluate the performance of a data store for interactive social networking actions and sessions which can be either read or update a very small amount of the entire set. BG can be used to compute a social Action Rating (SoAR) or a Socialites rating of a data store. The ratings compute the number of concurrent actions performed by a system.

An RDF benchmark for RDF engines consists of datasets (including data generators in the case of synthetic benchmarks), query workloads, performance metrics and rules that should be followed when executing a benchmark. Benchmarks are distinguished between those that use real datasets and those that produce synthetic datasets using special purpose data generators, as well as benchmark generators (see e.g. EU project https://project-hobbit.eu/).

V. Conclusion

In this paper, a review was conducted on the innovative topic of big data, which has gained a lot of attention and interest recently. There are lots of issues in big data technologies today and in this paper we addressed the challenge of quality and the functionalities needed for both processing and storage. This paper has covered also the latest tools and technologies that deal with big data with the aim to make the data processing more efficient and meaningful.

However, many technical challenges described in this paper must be addressed first before this potential can be realized fully. The challenges include not just the obvious issues of scale, but also heterogeneity, lack of structure, error-handling, privacy, timeliness, provenance, and visualization, at all stages of the analysis pipeline from data acquisition to result interpretation. Furthermore, these challenges will require transformative solutions, and will not be addressed naturally by the next generation of industrial products. Therefore, in order to achieve the promised benefits of Big Data, fundamental research towards addressing the identified technical challenges should be supported and encouraged.

Acknowledgement

The research presented in this paper is partly financed by the Ministry of Science and Technological Development of the Republic of Serbia (SOFIA project, Pr. No: TR-32010).

References

[1] S. Kaisler, J. A. Espinosa, F. Armour, W. Money, Advanced Analytics for Big Data, Encyclopedia of Information Science and Technology (3rd Edition), IGI-Global.

[2] Cai, L. & Zhu, Y., (2015). The Challenges of Data Quality and Data Quality Assessment in the Big Data Era. Data Science Journal. 14, p.2. DOI: http://doi.org/10.5334/dsj-2015-00.

[3] English L. P. (1999). Improving Data Warehouse and Business Information Quality, Wiley Computer Publishing, New York, USA.

[4] Wang R. Y., Strong D. M., Guarascio L. M. (1994). Data Consumers' Perspective on Data Quality, Beyond Accuracy: What Data Quality Means to Data Consumers, TDQM-94-01, MIT.

[5] Chen, M., Mao, S., & Liu, Y. (2014). Big Data: A Survey. Mobile Networks and Applications, 19(2), 171- 209.

[6] L. Douglas."3D Data Management: Controlling Data Volume, Velocity and Variety" (PDF). Gartner. Retrieved 6 February 20.

[7] P. Russom. "Big Data Analytics." TDWI Best Practices Report, Fourth Quarter 2011. TDWI Research. 2011.

[8] Suthaharan, S. (2014). Big Data Classification: Problems and Challenges in Network Intrusion Prediction with Machine Learning. Performance Evaluation Review, 41(4), 70-73.

[9] American Institute of Physics (AIP). 2010. College Park, MD, (http://www.aip.org/fyi/2010/)

[10] Mervis, J. 2012. "Agencies Rally to Tackle Big Data", Science, 336(4):22, June 6, 2012.

[11] P. Hitzler and K. Janowicz, "Linked data, big data, and the 4th paradigm," Semantic Web, vol. 4, no. 3, pp. 233–235, 2013.

[12] Batini, M. Scannapieco (2006): Data quality – Concepts, Methodologies and Techniques, Springer publications

[13] R. Jugulum, Competing with High Quality Data: Concepts, Tools, and Techniques for Building a Successful Approach to Data Quality. John Wiley & Sons, 2014.

[14] State Smart Transportation Initiative SSTI http://www.ssti.us/wp/wp-content/uploads/2014/10/Big-Data-ES.pdf

[15] D. Boyd and K. Crawford, "Critical questions for big data: Provocations for a cultural, technological, and scholarly phenomenon," Inf. Commun. Soc., vol. 15, no. 5, pp. 662–679, 2012.

[16] L. Sebastian-Coleman, Measuring Data Quality for Ongoing Improvement: A Data Quality Assessment Framework. Newnes, 2012.

[17] R. Y. Wang and D. M. Strong, "Beyond accuracy: What data quality means to data consumers," J. Manag. Inf. Syst., pp. 5–33, 1996.

[18] D. McGilvray, Executing Data Quality Projects: Ten Steps to Quality Data and Trusted InformationTM. Morgan Kaufmann, 2010.

[19] Y. W. Lee, "Crafting rules: Context-reflective data quality problem solving," J. Manag. Inf. Syst., vol. 20, no. 3, pp. 93–119, 2003.

[20] P. Russom. "Big Data Analytics" TDWI Best Practices Report, Fourth Quarter 2011. TDWI Research. 2011.

TABLE 3: CHALLENGES OF QUALITY IN THE BIG DATA ERA

Challenge	Quality assessment functionalities needed
S1: Data volume is remendous, the proportion of unstructured data in big data is very high and it is difficult to judge data quality within a reasonable amount of time.	• accuracy and consistency checking • provenance and trustfulness of sources • dealing with uncertainty e.g. changing weather
S2: Data change very fast and the "timeliness" of data is very short	• high-precision algorithms (or rules) needed to support the decision-making process
S3: The diversity of data sources brings abundant data types and complex data structures and increases the difficulty of data integration.	• validity checking to ensure conformity to agreed exchange standards • consistency checking to achieve single version of the truth

TABLE 4: OVERVIEW OF STREAMING SOLUTIONS

	STORM	**SPARK**	**SAMZA**	**FLINK**	**APEX**
Streaming Model	Native	Micro-batching	Native	Native	Native
API	Compositional	Declarative	Compositional	Declarative	Compositional
Fault tolerance	Record ACKs	RDD-based	Log-based	Check-points	Check-points
Guarantee	At-least-once	Exactly-once	At-least-once	Exactly-once	Exactly-once
State	Only in Trident	State as DStrem	Stateful operator	Stateful operator	Stateful operator
Windowing	Not built-in	Time based	Not built-in	Flexible	Policy based
Latency	Very-Low	Medium	Low	Low	Very-Low
Throughput	Medium	Very-High	High	Very-High	Very-High