The Opportunities, Challenges and Quality Assessment of Data in Big Data Age

Punyaban Patel[1]
Professor
Department of Computer Science & Engineering
Malla Reddy Institute of Technology
Maisamaguda, Dulapally, Secunderabad,
India
punyaban@gmail.com

Bibekananda Jena[2]
Asst. Professor
Department of Electronics & Comm. Engineering
Anil Neerukonda Institute of Technology and Sciences
Visakhapatnam
India

Abstract: The issue of data quality (DQ) is of growing importance in Remote Sensing (RS) data, biodiversity and commerce databases due to the widespread use of digital services that exploit and depends on data. With the advent of Internet of Things (IoT) and web technologies, there has been a remarkable growth in the amount of data generated. Big data is closely linked to the new, old question of data quality. Whoever pursues a new research perspective such as big data and wants to zero out irrelevant data is confronted with questions of data quality. Therefore, the European General Data Protection Regulation (GDPR) requires data processors to meet data quality standards; in case of non-compliance, severe penalties can be imposed. But what does data quality actually mean? And how does the quality requirement fit into the dogmatic systems of civil and data protection law? This paper emphasizes on the need for big data, technological advancements, quality assessment, tools and techniques being used to process big data are discussed. Technological improvements and limitations of existing storage techniques are also presented.

Key Words: Data, Big Data, Data Quality, Quality Assessment, Veracity, Velocity, Data Quality Metrics

1. Introduction

Data is rapidly growing and to quantify the data units also increased day by day. The improved availability of online satellite imagery, biodiversity and commerce databases have been greatly expanded the use of spatial data in science, technology, and to conserve it [1]. The reliable and safe health, economic & social care depends on access to, and the use of, high quality data. High quality information is an important resource for service providers in planning, managing, delivering and monitoring high quality safe care for the society. The information is data that has been processed or analysed to yield something useful and can be used intelligently in practice and real life situation. Data are numbers, symbols, words, images and graphics that have yet to be organised or analysed. Data quality refers to data that are accurate, valid, reliable, relevant, legible, complete and available in a timely manner to decision-makers for planning purposes. Data can be said to be of good quality when it meets the requirements of people who need to access data and information to support service delivery, quality improvement, and performance reporting and the planning[3,4]. For example, healthcare professionals need access to good quality data on patients such as known allergies or previous adverse reactions to drugs prior to prescribing or administering medications. As good quality health information is dependent on good quality data the most appropriate starting point for this work will focus on efforts to improve the information on which decisions are based to ensuring that data is collected, processed and analysed appropriately. National standards for data quality ascertain the structures and processes organisations should have in place to create an environment that enables data quality. Standards for data quality outline a framework to enable the collection, analysis and use of good quality data to support the delivery of health care, social care, commerce and to report on performance.

Big Data [6] is something like a set of huge data sets which are multifaceted and requires tedious jobs to capture, store, process and analyse them. The definition of "BIG DATA" may vary from organization to organization, institute to institute, person to person depending upon their use cases and their value generation from their data and data characteristics such as data size, capacity, competence of human resource, techniques used for analysis and, etc. For example, some organization, handling few GB of data may be a burdensome job where as for others it may be some terabytes. Big Data may be referred to data which is being generated in a very large quantity / volume at a high velocity/ rate in many different formats of data. According to traditional practice only some of the sample data taken for analysis instead of taking complete data due to many technical challenges to handle complete data set. Data was sampled and analysis was done on the sample data for decision making. However, with the help of Big Data and associated technologies and frameworks such as Hadoop, we would be able to process and analyse complete data set. So, we can achieve very accurate results from the complete data set as it is not biased for decision making.

This paper has been organizes as the section 1: Introduction, section 2: Data quality and Big Data, section 3: Features of Big data, section 4: Data quality assessment, section 5: Improving the Quality of Data: Challenges and Opportunities , and the section 6 concluded the paper.

2. Data quality and Big Data

Data Quality has been defined as "the totality of features and characteristics of a data set, that bear on its ability to satisfy the needs that result from the intended use of the data". Data quality therefore refers to data that is fit for purpose or "fit for use". This generally accepted view recognises that the quality of data is determined by the consumer – the person who will use it and who will ultimately decide if it is fit for whatever purpose it is intended

Data quality issues are as old as databases themselves. But Big Data has added a new dimension to the area; with many more Big Data applications coming online, the problem of bad data quality problems can be far more extensive and catastrophic.

3. Features of Big Data

If Big Data is to be used, organisations need to make sure that this information collection sticks to a high standard. To understand the problems, we need to look at them in terms of the important aspects of Big Data itself [3,5,6]. These are:

- **Velocity**

The speed at which data is generated can make it difficult to measure data quality given the finite amount of time and resources. By the time a quality assessment is concluded, the output could be obsolete and useless.
One way to overcome this is through sampling, but this is at the expense of bias as samples rarely give a truthful picture of the entire dataset.

- **Variety**

Data comes in all shapes and sizes in Big Data and this affects data quality. One data metric may not suit all the data collected. Multiple metrics are needed as evaluating and improving data quality of unstructured data is vastly more complex than for structured data. Data from different sources can have different semantics and this can impact things. Fields with identical names, but from different parts of the business, may have different meanings. To make sense of this data, reliable metadata is needed (e.g. sales data should come with time stamps, items bought, etc.). Such metadata can be hard to obtain if data is from external sources.

- **Volume**

The massive size and scale of Big Data projects makes it nigh-on impossible to undertake a wide-ranging data quality assessment. At best, data quality measurements are imprecise (these are not absolute values, more probabilities).

- **Value**

The value of data is all about how useful it is in its end purpose. Organisations use Big Data for many business goals and these drive how data quality is expressed, calculated and enhanced.
Data quality is dependent on what your business plans to do with the data; it's all relative. Incomplete of inconsistent data may not impact how useful the data is in achieving a business goal. The data quality may good enough to ignore improving it.
This also has a bearing on the cost vs benefit of improving data quality; is it worth doing and what issues need to take priority.

- **Veracity**

Veracity is directly tied to quality issues in data. It relates to the imprecision of data along with its biases, consistency, trustworthiness and noise. All of these effect data accountability and integrity.
In different organisations and even different parts of the business, data users have diverse objectives and working processes. This leads to different ideas about what constitutes data quality.
Data quality metrics have to be redefined based on particular attributes of the Big Data project, in order that metrics have a clear meaning, which can be measured and used for evaluating the alternative strategies for data quality improvement.

4. Data Quality Assessment

The hierarchical big data quality assessment is shown in [5,9,10] Table 1;

Dimensions	Elements	Indicators
1) Availability	1) Accessibility	■ Whether a data access interface is provided ■ Data can be easily made public or easy to purchase
	2) Timeliness	■ Within a given time, whether the data arrive on time ■ Whether data are regularly updated ■ Whether the time interval from data collection and processing to release meets requirements
2) Usability	1) Credibility	■ Data come from specialized organizations of a country, field, or industry ■ Experts or specialists regularly audit and check the correctness of the data content ■ Data exist in the range of known or acceptable values
3) Reliability	1) Accuracy	■ Data provided are accurate ■ Data representation (or value) well reflects the true state of the source information ■ Information (data) representation will not cause ambiguity

	2) Consistency	■ After data have been processed, their concepts, value domains, and formats still match as before processing ■ During a certain time, data remain consistent and verifiable ■ Data and the data from other data sources are consistent or verifiable
	3) Integrity	■ Data format is clear and meets the criteria ■ Data are consistent with structural integrity ■ Data are consistent with content integrity
	4) Completeness	■ Whether the deficiency of a component will impact use of the data for data with multi-components ■ Whether the deficiency of a component will impact data accuracy and integrity
4) Relevance	1) Fitness	■ The data collected do not completely match the theme, but they expound one aspect ■ Most datasets retrieved are within the retrieval theme users need ■ Information theme provides matches with users' retrieval theme
5) Presentation Quality	1) Readability	■ Data (content, format, etc.) are clear and understandable ■ It is easy to judge that the data provided meet needs ■ Data description, classification, and coding content satisfy specification and are easy to understand

Table 1: Hierarchy of Big Data quality assessment [10]

Descriptions of the data quality elements are given below[6,7,10].

- ***Accessibility***
  Accessibility refers to the difficulty level for users to obtain data. Accessibility is closely linked with data openness, the higher the data openness degree, the more data types obtained, and the higher the degree of accessibility.
- ***Timeliness***
  Timeliness is defined as the time delay from data generation and acquisition to utilization. Data should be available within this delay to allow for meaningful analysis. In the age of big data, data content changes quickly so timeliness is very important.
- ***Authorization***
  Authorization refers to whether an individual or organization has the right to use the data.
- ***Credibility***
  Credibility is used to evaluate non-numeric data. It refers to the objective and subjective components of the believability of a source or message. Credibility of data has three

key factors: reliability of data sources, data normalization, and the time when the data are produced.

- ***Definition/Documentation***
  Definition/document consists of data specification, which includes data name, definition, ranges of valid values, standard formats, business rules, etc. Normative data definition improves the degree of data usage.
- ***MetaData***
  With the increase of data sources and data types, because data consumers distort the meaning of common terminology and concepts of data, using data may bring risks. Therefore, data producers need to provide metadata describing different aspects of the datasets to reduce the problems caused by misunderstanding or inconsistencies.
- ***Accuracy***
  To ascertain the accuracy of a given data value, it is compared to a known reference value. In some situations, accuracy can be easily measured, such as gender, which has only two definite values: male and female. But in other cases, there is no known reference value, making it difficult to measure accuracy. Because accuracy is correlated with context to some extent, data accuracy should be decided by the application situation.
- ***Consistency***
  Data consistency refers to whether the logical relationship between correlated data is correct and complete. In the field of databases, it usually means that the same data that are located in different storage areas should be considered to be equivalent. Equivalency means that the data have equal value and the same meaning or are essentially the same. Data synchronization is the process of making data equal.
- ***Integrity***
  The term data integrity is broad in scope and may have widely different meanings depending on the specific context. In a database, data with “integrity” are said to have a complete structure [8]. Data values are standardized according to a data model and/or data type. All characteristics of the data must be correct – including business rules, relations, dates, definitions, etc. In information security, data integrity means maintaining and assuring the accuracy and consistency of data over its entire life-cycle. This means that data cannot be modified in an unauthorized or undetected manner.
- ***Completeness***
  If a datum has multiple components, we can describe [7] the quality with completeness. Completeness means that the values of all components of a single datum are valid. For example, for image colour, RGB can be used to describe red, green, and blue, and RGB represents all parts of the colour data. If the colour value of a certain component is missing, the image cannot show the real colour and its completeness is destroyed.
- ***Auditability***
  From the perspective of audit application, the data life cycle includes three phases: data generation, data collection, and data use. But here auditability means that auditors can fairly evaluate data accuracy and integrity within rational time and manpower limits during the data use phase.
- ***Fitness***
  Fitness has two-level requirements: 1) the amount of accessed data used by users and 2) the degree to which the data produced matches users’ needs in the aspects of indicator definition, elements, classification, etc.
- ***Readability***
  Readability is defined as the ability of data content to be correctly explained according

to known or well defined terms, attributes, units, codes, abbreviations, or other information.

- ***Structure***
  More than 80% of all data is unstructured, therefore, structure refers to the level of difficulty in transforming semi-structured or unstructured data to structured data through technology.

5. Improving the Quality of Data: Challenges and Opportunities

We must consider the extent to which the quality of data in open data portals can be improved given the variety of such data (in terms of domains covered and intended use by "anyone ... for any purpose") and the typically subjective, use oriented view of quality [2,6]. We believe this highlights an opportunity for data quality researchers to consult with the open data community to document the quality issues they experience, with the aim of identifying two categories of quality metrics: those generally relevant to all open datasets; and those relevant to the various types of data that are routinely published. Examples of the former include completeness of the dataset, representational consistency (including how data and missing data are described), accessibility, conformance to file formats, and metrics developed by [7] to assess the quality of metadata in repositories as part of improving discoverability;

The big data quality faces the following challenges [10]:

- ***The diversity of data sources brings abundant data types and complex data structures and increases the difficulty of data integration.***
  In the past, enterprises only used the data generated from their own business systems, such as sales and inventory data. But now, data collected and analyzed by enterprises have surpassed this scope. Big data sources are very wide, including: 1) data sets from the internet and mobile internet; 2) data from the Internet of Things; 3) data collected by various industries; 4) scientific experimental and observational data, such as high-energy physics experimental data, biological data, and space observation data. These sources produce rich data types. One data type is unstructured data, for example, documents, video, audio, etc. The second type is semi-structured data, including: software packages/modules, spreadsheets, and financial reports. The third is structured data. The quantity of unstructured data occupies more than 80% of the total amount of data in existence.

- ***Data volume is tremendous, and it is difficult to judge data quality within a reasonable amount of time***
  It is difficult to collect, clean, integrate, and finally obtain the necessary high-quality data within a reasonable time frame. Because the proportion of unstructured data in big data is very high, it will take a lot of time to transform unstructured types into structured types and further process the data. This is a great challenge to the existing techniques of data processing quality.

- **Data change very fast and the "timeliness" of data is very short, which necessitates higher requirements for processing technology**

Due to the rapid changes in big data, the "timeliness" of some data is very short. If companies can't collect the required data in real time or deal with the data needs over a very long time, then they may obtain outdated and invalid information. Processing and analysis based on these data will produce useless or misleading conclusions, eventually leading to decision-making mistakes by governments or enterprises. At present, real-time processing and analysis software for big data is still in development or improvement phases; really effective commercial products are few.

- **No unified and approved data quality standards have been formed in China and abroad, and research on the data quality of big data has just begun**.

Now-a-days, there are more than 100 countries and regions all over the world actively carrying out these standards. This implementation promotes mutual understanding among enterprises in domestic and international trade and brings the benefit of eliminating trade barriers.

6. Conclusion

Data quality is one component of a suite of information governance standards and supporting guidance that are being developed by the Authority. This international review will be used to inform the Authority of international practice in relation to data quality. The next step will be to develop the standards and guidance for data quality in Ireland to support the delivery of better safer care using the international review as a resource. These standards will be developed in conjunction with key stakeholders. These standards will be developed in conjunction with key stakeholders. Consequently the big data quality assessment will be done properly and the challenges can be face.

References:

[1] English L. 10 years of Information Quality Advances: What Next? Information Management Magazine . 2001.

[2] R.Y. Wang and D.M. Strong. 1996. Beyond Accuracy: What Data Quality Means to Data Consumers. Journal of Management Information Systems 12, 4 (1996), 5–33.

[3] Kerr KA, Norris T, & Stockdale R.," The strategic management of data quality in healthcare" , Health Informatics Journal, 14 p.259, 2008.

[4] Deming WE. ,"Out of the Crisis", The MIT Press; 2000.

[5] Wang RY, Ziad M, & Lee Y, "Data Quality", . Kluwer Academic Publishers; 2001.

[6] Batini C, Scannapieca M.," Data Quality: Concepts, Methodologies and Techniques", Springer; 2006.

[7] S. Neumaier, J. Umbrich, and A. Polleres. 2016. Automated Quality Assessment of Metadata Across Open Data Portals. Journal of Data and Information Quality 8, 1, Article 2 (Oct. 2016), 29 pages. Open Knowledge. 2012. The Open Data Handbook. Retrieved Feb 1, 2016 from http://opendatahandbook.org/.

[8] Mohanty S," The four essential V's for a big data analytics platform". Dataconomy-Online, http://dataconomy.com/the-four-essentials-vs-for-a-big-data-analytics-platform/. Accessed 4 Apr, 2017.

[9] Sotto LJ & Simpson AP, " United States. In: Roberton G (ed) Data protection & privacy", Law Business Research Ltd, London, pp 208–214, 2015.

[10] 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-2015-002