Big Data: Review Paper

Abdulbaset Salem Albaour, Yousof Abdulrahman Aburawe

[1]*Misurata University*
[2]*Misurata University*

Abstract

Big data is a new technological paradigm for data that is generated at high speed, high volume, and with great diversity. Big data is seen as a revolution that has the potential to revolutionize the way companies operate in many industries. This paper introduces big data and the dimensions of data quality where the challenges of the quality factors of big data quality are discussed, and the lifecycle of big data analytics is discussed.

Key Words: *Big data, Big Data Quality, Big Data Quality Dimensions, Big Data Analysis*

INTRODUCTION

Big data refers to the concept of very large data sets involving three major dimensions or properties named (3Vs). First is a volume according to the amount of data located in the storage medium. Second is Variety which refers to the various heterogeneous and complex types of data. Data can be structured, unstructured, or semi-structured generated either by humans or machines. The third is velocity which indicates the speed of data processing required to handle that large amount of data.

Most definitions of big data focus on the size of data in storage. Size matters, but there are other important attributes of big data, namely data variety and data velocity. The three Vs of big data (volume, variety, and velocity) constitute a comprehensive definition, and they bust the myth that big data is only about data volume. In addition, each of the three Vs has its own ramifications for analytics.[1] (See Figure 1.)

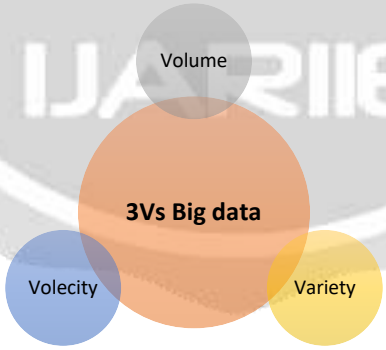

Figure 1: Three Vs Big data

"Big data is a term that describes large volumes of high velocity, complex and variable data that require advanced techniques and technologies to enable the capture, storage, distribution, management, and analysis of the information." [2]

A large data volumes are daily generated at unprece-dented rate from heterogeneous sources (e.g., health, government, social networks, marketing, financial). This is due to many techno- logical trends, including the Internet of Things, the proliferation of the Cloud Computing as well as the spread of smart devices. [3].

Some of the difficulties related to big data include capture, storage, search, sharing, analytics, and visualizing. Today, enterprises are exploring large volumes of highly detailed data so as to discover facts they didn't know before.[4]

Big data is data whose scale, distribution, diversity, and/or timeliness require the use of new technical architectures, analytics, and tools in order to enable insights that unlock new sources of business value. Three main features characterize big data: volume, variety, and velocity, or the three V's. The volume of the data is its size, and how enormous it is. Velocity refers to the rate with which data is changing, or how often it is created. Finally, variety includes the different formats and types of data, as well as the different kinds of uses and ways of analyzing the data.[5]

Big Data Quality & Big Data Quality Dimensions

Huge quantities of data does not automatically guarantee quality. And with larger volumes it becomes more important to focus on the quality in order to derive some meaningful insights out of the available data. In most contexts the worth of the data is determined by its 'fitness for use', this criteria of data is still the central part to determine the necessity or mandate for organizations to invest in big data.[6]

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.[7]

Figure 2: Big Data Quality Dimensions

Accessibility and availability are related to the ability of the user to access data from his or her culture, physical status/functions, and technologies available [8]

Consistency, cohesion and coherence refer to the capability of data to comply without contradictions to all properties of the reality of interest, as specified in terms of integrity constraints, data edits, business rules and other formalisms [8]

Completeness: it measures the degree with which a dataset is complete. It is evaluated by assessing the ratio between the amount of values currently available in the dataset and the expected amount of values. The expected amount of values considers both null values in available registrations and missing registrations. Note that, as regards data streams, missing registrations are easy to detect if data are sensed with a specific frequency. If data are not collected at a regular pace it is possible to rely on historical data to estimate the sampling frequency that often varies over time [9]

Consistency: [10] it refers to the violation of semantic rules defined over a 7 set of data items. Therefore, this dimension can be calculated only if there is the availability of a set of rules that represent dependencies between

attributes. We have developed a module that detects functional dependencies and checks that values in the dataset respect them.

Timeliness: refers to the time expectation for accessibility and availability of information. Timeliness can be measured as the time between when information is expected and when it is readily available for use, this concept is of particular interest, because synchronization of data updates to application data with the centralized resource supports the concept of the common, shared, unique representation. The success of business applications relying on master data depends on consistent and timely information. Therefore, service levels specifying how quickly the data must be propagated through the centralized repository should be defined so that compliance with those timeliness constraints can be measured.[11]

Validity: Is the data presented in the correct and pre-defined format, type or range so as to be applicable to the given analytical task

The data set may be complete but does it tell the user what it purports to and is it valid for the current business context. Data quality not only depends on the completeness but also on the business environment and the business purpose it is supposed to serve. Only the data that conform to the business requirements can be considered valid. In health care, huge volume of sensor-generated data is generated by remote monitoring or continuous monitoring of patients. If the data generated are not valid, the analysis will lead to 28 erroneous results and making decisions based on those results would put the patient's life at risk. Storing and analyzing Big Data is a complex procedure, as Big Data storage requires special hardware, and analysis requires Big Data analytics tools. To perform Big Data analysis, a company requires financial investment. IoT is widely used for environmental research like air quality monitoring, water quality monitoring, weather prediction, and natural disaster prediction. If the data used for analysis is biased, the results obtained will not be valid and making decisions or judgments based on the invalid result might lead to disaster.[12]

In Big Data, other dimensions have to be considered. For example, the large number of sources makes trust and credibility important. The trustworthiness of a data item is the probability that its value is correct and depends on data provenance [9]

The Challenges of Big Data Quality

Big Data can bring cost saving, risk control, improvement of management efficiency, and increment of value into enterprise. In the meanwhile, Big Data brings some challenges:

- **Unevenness of Data Quality**

  The large amount of data. Though the first step of processing data is to gather data, if the gather all data in spite of quality, it is possible to make wrong predictions and decisions. according to view of this condition, after gathering data, it is necessary to select relative data and clean conflicting data.[13]

- **Lack of skills**

  Big data application requires enterprise to design new data analysis models. That's because traditional models are fit to process structured data not big data including multi-type data. Thus, it needs some data science to apply to enterprise data management. The enterprise is short of talents who can design new data analysis models. The talents who not only can design new data analysis models but also know the financial management are fewer. Lack of talents is a severe and long-term issue. Big Data is a sword with two blades. Through affecting the idea, function, mode, and method of financial management, it can bring cost saving, risk control, improvement of management efficiency, and increment of value into enterprise. In the meanwhile, it brings a lot of challenges. Only through fostering strengths and circumvent weaknesses, can an enterprise remain invincible in Big Data era.[13]

- **Big data Features Volume**

  Refers to the tremendous volume of the data. We usually use TB or above magnitudes to measure this data volume. Velocity means that data are being formed at an unprecedented speed and must be dealt with in a timely manner. Variety indicates that big data has all kinds of data types, and this diversity divides the data into structured data and unstructured data. These multi typed data need higher data processing capabilities.[14]

- The diversity of data sources brings abundant data types and complex data structures and increases the difficulty of data integration

  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 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.
- 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.

  In order to guarantee the product quality and improve benefits to enterprises, in 1987 the International Organization for Standardization (ISO) published ISO 9000 standards. Nowadays, 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. By contrast, the study of data quality standards began in the 1990s.[14]

Big Data Analysis

Big data analytics is differences from traditional analytics Because of the big increase in the volume of data and that led to Many researchers have suggested commercial DBMS and this not suitable with size of data. This type of data is impossible to handle using traditional relational database management systems. New innovative technologies were needed and Google found the solution by using a processing model called MapReduce. There are more solutions to handle Big Data, but the most widely-used one is Hadoop, an open source project based on Google's MapReduce and Google File System. Hadoop was founded by the Apache Software Foundation. The main contributors of the project are Yahoo, Facebook, Citrix, Google, Microsoft, IBM, HP, Cloudera and many others. Hadoop is a distributed batch processing infrastructure which consists of the Hadoop kernel, Hadoop Distributed File System (HDFS), MapReduce and several related projects.[15]

Big Data Analytics refers to the process of collecting, organizing, analyzing large data sets to discover different patterns and other useful information. Big data analytics is a set of technologies and techniques that require new forms of integration to disclose large hidden values from large datasets that are different from the usual ones, more complex, and of a large enormous scale. It mainly focuses on solving new problems or old problems in better and effective ways

Types of Big Data Analytics

a) Descriptive Analytics

It is a preliminary stage of data methods organize data and help uncover patterns that offer insight. Descriptive analytics provides future probabilities and trends and gives an idea about what might happen in the future.[16]

b) predictive Analytics

Is a method through which we can extract information from existing data sets to predict future outcomes and trends and also determine patterns? It does not tell us what will happen in future. It forecasts what might happen in future with acceptable level of reliability. It also includes what if-then-else scenarios and risk assessment. Applications areas of Predictive Analytics are.[17]

Big Data Analytics Lifecycle Overview

The Big Data Analytics Lifecycle defines analytics process best practices spanning discovery to project completion. The lifecycle draws from established methods in the realm of data analytics and decision science.

This synthesis was developed after gathering input from data scientists and consulting established approaches that provided input on pieces of the process.

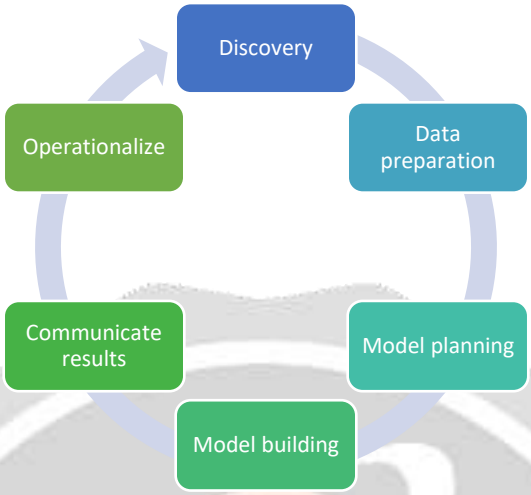

Figure 3: Big data analysis stages

The Data Analytics Lifecycle is designed specifically for Big Data problems and data science projects. The lifecycle has six Stages, and project work can occur in several Stages at once. For most Stages in the lifecycle, the movement can be either forward or backward. This iterative depiction of the lifecycle is intended to more closely portray a real project, in which aspects of the project move forward and may return to earlier stages as new information is uncovered and team members learn more about various stages of the project. This enables participants to move iteratively through the process and drive toward operationalizing the project work. [18]

Here is a brief overview of the main Stages of the Data Analytics Lifecycle:

- Stage 1: Discovery: In Stage 1, the team learns the business domain, including relevant history such as whether the organization or business unit has attempted similar projects in the past from which they can learn.
- Stage 2: Data preparation: Stage 2 requires the presence of an analytic sandbox, in which the team can work with data and perform analytics for the duration of the project.
- Stage 3: Model planning: Stage 3 is model planning, where the team determines the methods, techniques, and workflow it intends to follow for the subsequent model building Stage.
- Stage 4: Model building: In Stage 4, the team develops datasets for testing, training, and production purposes. In addition, in this Stage the team builds and executes models based on the work done in the model planning Stage.
- Stage 5: Communicate results: In Stage 5, the team, in collaboration with major stakeholders, determines if the results of the project are a success or a failure based on the criteria developed in Stage 1. The team should identify key findings, quantify the business value, and develop a narrative to summarize and convey findings to stakeholders.
- Stage 6: Operationalize: In Stage 6, the team delivers final reports, briefings, code, and technical documents. In addition, the team may run a pilot project to implement the models in a production environment.

CONCLUSION

Big data refers to the set of numerical data produced by the use of new technologies for personal or professional Purposes. In this paper, we have studied Big Data characteristics and discussed the challenges of big data quality

might affect raised by Big Data. Also, Big Data analytics is the process of examining these data in order to uncover hidden patters. Big Data Analytics is a fast growing technology. But difficult degree of analysis of these data in the framework of the Big Data is a process that depended on kind of process which required.

REFERENCES

[1] P. Russom, "Big data analytics - TDWI Best Practices Report," TDWI Best Pract. Report, Fourth Quart., no. August, p. 38, 2011.

[2] M. Agarwal, S. Ajemian, G. Tim, B. Microstrategy, R. Campbell, and S. Coggeshall, "Demystifying Big Data: A Practical Guide To Transforming The Business of Government Listing of Leadership and Commissioners," pp. 1–40, 2013.

[3] A. Oussous, F. Z. Benjelloun, A. Ait Lahcen, and S. Belfkih, "Big Data technologies: A survey," *J. King Saud Univ. - Comput. Inf. Sci.*, vol. 30, no. 4, pp. 431–448, 2018, doi: 10.1016/j.jksuci.2017.06.001.

[4] P. Russom, *Big Data Analytics. In: TDWI Best Practices Report*. 2011, pp. 1–40.

[5] N. Elgendy and A. Elragal, "Big Data Analytics: A Literature Review Paper," in *Advances in Data Mining. Applications and Theoretical Aspects*, 2014, pp. 214–227.

[6] A. Ramasamy and S. Chowdhury, "Big Data Quality Dimensions: a Systematic Literature Review," *J. Inf. Syst. Technol. Manag.*, vol. 17, no. 0, 2020, doi: 10.4301/s1807-1775202017003.

[7] 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, Aug. 2014, doi: 10.1016/j.ijpe.2014.04.018.

[8] 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, Jan. 2015, doi: 10.4018/JDM.2015010103.

[9] D. Ardagna, C. Cappiello, W. Samá, and M. Vitali, "Context-aware data quality assessment for big data," *Futur. Gener. Comput. Syst.*, vol. 89, pp. 548–562, 2018, doi: 10.1016/j.future.2018.07.014.

[10] C. Batini and M. Scannapieco, *Data and Information Quality*. Cham: Springer International Publishing, 2016.

[11] D. Loshin, "Data Quality and MDM," *Master Data Manag.*, pp. 87–103, 2009, doi: 10.1016/b978-0-12-374225-4.00005-9.

[12] J. Farmasi and S. Dan, "No 主観的健康感を中心とした在宅高齢者における 健康関連指標に関する共分散構造分析Title," vol. 14, no. 1, pp. 55–64, 2017.

[13] M. Ke and Y. Shi, "Big Data , Big Change : In the Financial Management," no. October, pp. 77–82, 2014.

[14] L. Cai and Y. Zhu, "The Challenges of Data Quality and Data Quality Assessment in the Big Data Era," pp. 1–10, 2020.

[15] M. Adam and I. Fakharaldien, "Big Data Analysis and Storage," no. March, 2017.

[16] A. Salina, "A Study on Tools of Big Data Analytics," no. October 2016, 2018, doi: 10.15680/IJIRCCE.2016.

[17] A. P. F. O. R. T. Ext, A. Udio, and V. Ideo, "B IG D ATA A NALYTICS : C HALLENGES AND," vol. 5, no. 1, pp. 41–51, 2016.

[18] B. Y. David Dietrich, Barry Heller, *Data Science and Big Data Analytics: Discovering, Analyzing, Visualizing and Presenting Data*. John Wiley & Sons, Inc. 10475 Crosspoint Boulevard Indianapolis, IN 46256.