International Journal of Science and Research (IJSR)
ISSN: 2319-7064
ResearchGate Impact Factor (2018): 0.28 | SJIF (2018): 7.426

Process Challenges of Big Data - A Comprehensive Study

G. Sandra

Assistant Professor, Sree Narayana Guru College of Advanced Studies, Cherthala, Alappuzha, Kerala, India. Pin. 688 523

Abstract: ***In the present digital world, we are living in an era of big data. There is an exponential growth of data every day. This massive growth of data has many scientific and economic advantages. Through processing this raw data, the users are able to enjoy many insights and discover new knowledge from it. These digital data can be used profitably for business prosperity and making policy decisions. The massive growth of data gives us not only opportunities but also brings forth many challenges. Scholars have broadly classified the big data challenges into three categories: Data challenges; Process challenges and Management challenges. This paper overviews the process challenges of big data.***

Keywords: Big Data, Big data challenges, Process challenges

1. Introduction

In modern digital world, huge amount of data is generated every moment, popularly called Big Data. They are generated by machine and human efforts. Scholars have calculated its tempo differently [1]. Today the machine and human generated data are growing at a ten times faster than the conventional data growth rate. Among them, the machine data is having an exponential growth rate of fifty times faster. Studies have shown how the yearly data generation of world is increased from 1-2 exabytes back in 2000 to 2700 exabytes after 12 years with an expectation of 40,000 exabytes in 2020. This type of data explosion has many advantages and disadvantages to the organizations that are utilizing this data. It is the treasure house for users to getting insights through its analysis, which can be used for their business prosperities and framing policies. At the same time these organizations are facing problems also in their attempt to take benefits from this data [2]. Prominent among are the data center power, space restrictions, problem of growing cluster, data storage, movement of data and management difficulties. Scholars have nomenclature these problems as big data challenges.

In this paper, an attempt is made to overview the group of big data challenges and a comprehensive survey particularly related to challenges of processing the big data. It is organized in the following way. Section II deals with big data challenges. Section III narrates various prominent process challenges encountered by organization and section IV concludes the study with suggestions for future works.

2. Big data Challenges

In the modern times, there is an astounding growth of digital data. They are generated from multiple sources, such as manufacturing, science, business, information technology and from personal lives. This tremendous growth of data volume out-speed the development of our computing infrastructure. As the size of datasets increasing, the conventional data processing technologies are inefficient to process the big data. From this feature of big data, scholars have tried to define big data as the one which should include datasets beyond the ability of conventional software tools to capture, manage and process data [3].

The concept of big data is problematic. Though it is the source of invaluable insights and knowledge discovery, it is an embodiment of challenges that must be address properly to elicit the benefits from it. Majority of these challenges are the function of the characteristics of big data [4]. Mainly they are of 3Vs; Volume, Velocity and Variety, though researchers have proposed more than 15 characteristics of Big data. The challenges generating from high volume of big data are too critical. Since the size of the data generating is extraordinarily high, data scientists find it difficult to quantify them, because of its complexity. Like this, the high velocity challenge of big data is posing many problems to data management system. Though different technologies have been developed to solve the challenges of high velocity, but the capacity of the technologies of data streaming to solve these problems are very limited. Similarly the big data implementation needs the processing of data from sources, where data can be in different formats. This brings forth the challenges of data variety. Hence the big data challenges include the problem of how to deal with data growth, developing technologies for eliciting insights from data growth in timely manner, integrating disparate data sources, validating data, securing big data, organizational resistances, decision making of what are generated and collected, issues of privacy, building solutions for multifaceted data and implementation of new approaches for processing and analysis of data [5]. All these big data challenges can be broadly grouped into three main categories as: Data Challenges, Process Challenges and Management Challenges [6].

Data Challenges are the difficulties faced by the organizations from the characteristics of big data such as Volume, Variety, Velocity, Veracity, Variability, Visualization and Value. The Process Challenges are the group of challenges encountered when processing big data. The main among them are the problems associated with data acquisition, warehousing, data mining, cleansing, data aggregation, data integration, modeling, analysis of data and data interpretation. When assessing, managing and governing the data, the organizations are faced with challenges. They are classified as management challenges, which include problems of security, privacy, data governance, operational expenditures, and information sharing and data ownership. In this study an attempt is made

to do a comprehensive survey on process challenges of big data by keeping the other two categories of challenges beyond the scope of this paper.

3. The Process Challenges

The big datasets, which is very complex in nature, is always a critical challenge to organizations. The ability and efficiency to process this complex data is a real peril to all business firms irrespective of their size. The datasets come before business firms for processing are mainly unstructured and semi-structured. The processing of these non-relational datasets really poses a significant problem to data scientists. The following section attempts to narrate these challenges under the heads; data acquisition and warehousing, data mining, data cleansing, data aggregation, data integration, data analysis and modeling and data interpretation.

3.1. Data acquisition challenges

In big data, the main challenge in acquiring data is to find out the real framework and tools for acquiring and processing data. The data can be acquired from multiple sources, such as web mining, logs, sensors, social networks, etc. It has different formats, like structured, unstructured and semi-structured. Data scientists face great challenges in the processing of unstructured and semi-structured data compared to structured data, which is generating at a high speed and has a definite format and rigid.

The data acquisition is the process of gathering, filtering and cleaning data before depositing them in the storage. The very success in these efforts is depending on the 4Vs of Big Data, viz; Volume, Velocity, Variety and Value [7]. In data acquisition scenarios, these characteristics of big data are assumed as high volume, high velocity, high variety and low value. In this context it has to be remembered that in the formulation of algorithms for the processing of data acquisition steps is governed by the high value fragments of data, which remains a major challenge. The other big data acquisition challenges are emanating from data types, data structure, data scalability, data content, data storage, data integration, data efficiency and big data upgrades [8]. Big data is multi-structured. So it requires knowledge of efficient management tools for accessing different types of data from various origins and contextualize it for data analysis and knowledge extraction. Scalability is another great concern of big data processing because of the necessity of big data system regarding the ability to quickly address and analyze data on demand without being affected by the scale and pace of acquisition of data and querying. Yet another concern is the need of system reliability, which denotes the ability to provide similar performance always and integrate newer data sources and upgrades the existing data without affecting the functionality and performance.

To integrate the process of acquisition of data, we have different big data architectures of IBM, Vivisimo and Oracle. They are used to retrieve the content of data sources, offer scalable storage solutions like NoSQL database, Hadoop Distributed File System (HDFS), etc. Then the stored data is processed with the help of data analytics software and finally analyzing them with appropriate data analytics algorithms.

Different technologies are adopted for acquiring big data with the help of paradigms like message queuing, subscribe paradigm [9] and event processing paradigm [10]. We have different technologies for acquiring data. The most commonly used open protocols are Advanced Message Queuing Protocol (AMQP) and Java Message Services (JMS). The AMQP has the advantages of ubiquity, safety, fidelity, applicability and manageability. AMQP is compatible with JMS. It offers a common way for Java programs to create, send, receive and read an organization's messaging system's messages.

Along with the enterprise-specific open protocols for data acquisitions, many software tools are also in use for data acquisition. The most widely used among them are: Storm; which is an open-source framework used for the computation of data streams. It supports different ranges of programming languages and using in data collecting scenarios. S4 is a scalable streaming system used for processing stream of data [11]. It is inspired by Map Reduce application. Kafka is widely used software for unifying offline and online processing of data. For efficiently collecting log data, Flume is widely used, which is robust and fault tolerant. A popular and open-source framework widely used for scalable and distributed computing on big data is Hadoop. It is written in Java and derived from Google's Map Reduce and the Google File System (GFS).

The above mentioned tools are only a few widely used techniques for data acquisition and processing. The major challenges that the data scientists confronting in the use of the tools for the acquisition of data are; the ability to deal with wide range of tools; to provide systems to connect the data acquisition with the data pre-and post- throughput and storage; to frame models for data analysis of structured, semi-structured and unstructured data and to provide open-source tools for processing the acquired data [12].

3.2. Data warehousing challenges

Data warehousing is referred as the electronic storage of large amount of information by an organization. It helps the users to understand their organization and its performance. It acts as a central repository where information comes from different data sources. The data warehouse merging all the information in one comprehensive database, so that the users can access the processed data through the business intelligence tools, SQL clients and spreadsheets. The data warehousing makes the data mining possible. It is beneficial to decision makers, who use customized and complex process to get information and to discover hidden pattern of data flows. For this different data warehousing tools are employed, such as Oracle, Marklogic, Amazon Redshift, etc.

The data warehousing poses great challenges to big data analysis. It is not an ideal option for unstructured data processing. It is too complex for average users. For organizations also, it is too difficult to build and run data warehouse systems, since data are continuously grow in size.

3.3. Data mining challenges

Data mining techniques offer useful information to big data analytics. Data mining is the process of exploration and analysis of huge quantum of data to find out the pattern of big data analysis [13]. It aims at either classification or prediction. For this, different algorithms are in use. The widely used tools are Logistic regression, mainly employed to predict the probability of occurrences; Classification Trees, which is a prominent data mining techniques, widely employed to classify the dependent variables based on the measurements of predictable variables; Clustering techniques like K-nearest neighbors for identifying group of similar records and Neural Networks, where data is given to the input node and by trial and error system, the algorithm adjusts the weights till it satisfies certain stopping criteria.

Data mining process is also great challenges to big data analytics. At every moment, umpteen quantities of data are generated. Due to the multiplicity of data, generally the software tools employed for its analysis have become inefficient to manage it. Hence the data mining in big data is facing with the problems of data security, privacy issues, redundant data problems, poor data quality, higher cost of data mining, problem to process unstructured data, etc. These issues lead to many data mining challenges in big data. Some of them are: formulation of appropriate big data mining platforms; big data semantics, developing efficient big data mining algorithms for mining complex and dynamic data; mining from sparse, uncertain and incomplete data, model fusion for multiple information sources, etc. As a solution to above challenges, scholars have proposed different open- source software such as Hadoop, Cloudera, MonoDb, Map Reduce, etc., But above proposed tools have their own limitations.

3.4 Data Cleansing challenges

In big data, the data cleansing is the procedure of correcting the inaccurate and corrupt data. Big data may have many inaccurate data, which may drive the business to wrong decision. Hence the data cleansing is essential for big data management. Due to high volume, high velocity and high variety of big data, removing the corrupt data from it is a tedious effort. For the better use of data, filtering of the data is required. Data cleansing eradicates errors from data types and transform metrics and log data into suitable formats. To materialize this objective, high performing computing paradigm is required [14]. Since big data is rather elusive, there is the challenge to develop suitable extraction methods for mining the required information and articulate them in a structured form for making it more legible [15].

3.5 Data Aggregation and Integration challenges

Data aggregation is referred as the process of gathering and expressing the raw data in a summary for statistical analysis. It helps to eliminates redundancy and aims to gain awareness about the resources groups. They are of time aggregation and spatial aggregation. The time aggregation denotes the data point for a single resource over a specified time period and spatial aggregation is the data points for a group of resources over a specified period of time. The data is collected and presented in different time intervals, such as reporting period like daily, weekly, monthly, quarterly and monthly; granularity, which refers to period over which the data points for a given resources are collected for aggregation and polling period; where the time limit that determines how resources are sampled for data. To combine data from multiple sources into one place, different aggregation tools are used for getting new insights and discover new relationship and patterns. The prominent tools that have gained wider popularity are Flume, Sqoop, etc. [16]

Big data analytics confronts challenges not only from data aggregation but from data integration also. It is the process of combining data generated from different sources such as traditional, machine, social media, web data, IoT, etc., into a single framework of analysis to get new insights. It helps the users to have a unified view about the accumulated data. But the process of integration of big data is quite complicated. It comprises of critical challenges like uncertainty of data, accessing data coming from different sources, data synchronization, choosing data management tools, selection of data experts and selection of right strategy. The popular integration techniques used in this domain are the Schema mapping, Record linkage and the Data fusion [17].

3.6 Data Modeling and Analysis challenges

Big data modeling is the logical design of a system, which implies the physical implementation of data base. It plays a crucial role in the big data analytics that helps the data scientists to build set of relationship between data items. It is the process of sorting and storing of data. Good data models can reduce unnecessary data redundancy, reuse computing results and reduce the storage and computing costs for big data system. It is a set of tools and techniques used for understanding and analyzing how an organization should collect, update and store data. Two types of data models techniques are proposed [18]. They are Entity Relationship Models (E-R models) and Unified Modeling Language (UML). Many models have been suggested for hierarchical, network, relational and non-relational databases [19]. Since 1980's the non- relational models have been gained greater prominence with the implementation of Oracle databases, MySQL and Microsoft SQL server.

A relational database is a collection of data items arranged in formally described tables from which data can be accessed in different ways. Generally the relational database use Structured Quary Language (SQL), to access and modify the data stored in database. But the challenge is that the relational database models do not support high scalability. In big data processing, the database has to be partitioned across multiple servers [20]. To solve this challenge, scholars have recommended Non- Relational databases [21], which are classified on the basis of organizing data as key value store structure, document store, graphic store and column-oriented database (COD).

Not only modeling, the process of data analysis is also a big challenge to data scientists. It denotes the process of collecting, organizing and analyzing large sets of data with the objectives of getting insights, pattern and knowledge. At

present different software are used for the above process. Most of them are open-source tools and others are paid. A comparative study of popularly used open source big data software is given below.

Table 1

S. No	Big data Analysis Tools	Category	Main features
1.	Apache Hadoop	Open - source	Written in Java and provides cross-platform support.
2.	Cloudera Distribution for Hadoop (CDH)	Open - source and free	It encompasses Apache Hadoop, Apache Spark, Apache Impala.
3.	Apache Cassandra	Open - source	Used to manage huge volume of data. It employs Cassandra Structured Quary Language.
4.	KNIME (Konstanze Information Miner)	Open - source	It supports Linux, OSX and Windows OS. It is used for enterprise reporting, integration, research, data mining and business intelligence.
5.	Data Wrapper	Open-source	Used for data visualization helps the users to generate simple, precise and embeddable charts very quickly.
6.	MonogoDB	Free and Open-source	It is a NoSQL documents-oriented database written in C, C++ and Java script. It supports multiple operating systems such as Windows Vista, OSX, Linux, Solaris and Free BSD.
7.	Lumify	Free and open - source	It is designed for big data fusion, integration analytics and visualization.
8.	HPCC	Open-source	It stands for high-performance computing cluster. It is a complete big data solution over a highly scalable supercomputing platform.
9.	Apache Storm	Free and open-source	It is a cross-platform, distributed stream processing and a fault-tolerant real- time computational framework.
10.	Apache SAMOA	Open-source	It stands for Scalable Advanced Massive Online Analysis.
11.	Talend	Free and open-source	Its components and connectors are Hadoop and NoSQL. It provides community support only.
12.	Rapidminer	Open-source	A cross-platform tool which offers an integrated environment for data science, machine learning and predictive analysis.
13.	Qubole	Open-source	It is an independent and all inclusive big data platform.
14.	Tableau	Open-source	It is an effective tool for data visualization and exploration.
15.	R	Free and open-source	It is a comprehensive statistical analysis package.

From the above description, it is made clear that we have ample tools available to support big data operations. But the real challenge is the ability to choose the right big data tools wisely as per the requirement.

3.7 Data Interpretation challenges

Data interpretation is very crucial in big data processing for developing and arriving at sound conclusions. It is the process of making sense of numerical data that has been collected, analyzed and presented. The big data is collected from multiple sources. While analyzing it, the nature and interpretation may vary from business to business. When interpreting data, the data scientists must discern the difference between correlations, causation, and coincidence and consider all the factors that may influence the result of data analysis.

Different methods are followed for the interpretation of data, which is broadly classified into quantitative and qualitative methods. The qualitative data analysis is categorical and not described through numerical values. The techniques used for this effort are observations, documentation and interviews. In case of quantitative data interpretation, it is numerical. It refers to a set of process by which numerical data is analyzed. For this the most common statistical methods followed are mean, standard deviation and frequency distribution. Other significant interpretation process of quantitative methods includes regression analysis, cohort analysis and predictive and prescriptive analysis. Data interpretation in big data is used for data identification and explanation, comparing and contrasting data, identification of data outliers and future prediction.

Data interpretation confronts many challenges in big data processing. The astounding growth of big data, the multiplicity of unstructured data, the high variety and velocity of big data have intensely affected the big data processing and derivation of new knowledge from these raw data. Hence the data scientists have to find out critical solutions for the development and selection of right technological solutions for accessing, aggregating, analyzing and interpreting big data for which the critical management challenge is the task of finding out the skilled and experienced persons having sufficient analytical vision [22].

4. Conclusion

To understand, process and utilize the knowledge and new insights hidden in big data are the critical challenges to data scientists. The multiple growths of big data and challenges emanating from it promote a lot of new technology and innovation. By understanding the real process challenges of big data will help researchers to get greater insights into utilization of big data in the right way for making effective business decisions and policy making. Therefore research is highly needed in this direction.

References

[1] International Data Corporation (IDC) survey, World Data Protection as a service forecast, 2019-20.
[2] P. Russom, Managing Big Data, https://tdwi.org/articles/2013/10/01.
[3] A.Labrinidis, H Jagadish, Challenges and opportunities with Big Data, Proceedings of the VLDB, 2012, 5 (12): 2013-2033.
[4] X Jin, B. W. WahCheng X, Wang Y, Significance and Challenges of Big Data Research, Big Data Research, 2(2), 2015, pp.59-64.
[5] T Wang, V Wiebe, Big Data Analytics on the characteristics equilibrium of collective options in social

networks, International Journal of Cognitive Informatics and Natural Intelligence (IJCINI), 8 (3), 2014, pp. 29-44.
[6] R Akerkar, (Ed), Big Data Computing, CRS Press, Taylor & Francis Group, Florida, USA, 2014, pp.103-128.
[7] Klaus Lyko, Marcus Nitzschke, Axel-Cyrille and Ngonga Ngomo, Big Data Acquisition, New Horizons for Data-Driven Economy, Springer Link, 2017, pp. 39-61.
[8] Quasim Maqbool and Ahmed Habib, Resolve Five Big Data, data Acquisition challenges, Control Engineering, March 10, 2019.
[9] A Carzaniga,D.S. Rosenblum, and A.L. Wolf Achieving Expressiveness and Scalability in an Internet-Scale Event Notification Services, Nineteenth ACM Symposium on Principles of Distributed Computing (PODC 2000), Portland, Oregon, July 2000.
[10] Gianpaolo Cugoda and Alessardro Margara, Orocessing flows of information: From data stream to complex event processing, ACM Computing Surveys (CSUR), vol. 44, Issue 3, June 2012.
[11] Leonardo Neumeyer, Bruce Robbins, Anish Nair, Anand Kesari, S4: Distributed steam computing platform, International Conference, IEEE, 2010.
[12] F. Zhang, Liu M, Gui F, Shen W, Shami A, Ma Y, A distributed frequent item set mining algorithm using spark for Big Data analytics, Cluster Computing, 18 (4), 2015, pp. 1493-1501.
[13] Xingquan Zhu, Iam Davidson, Knowledge discovery and Data Mining: Challenges and Relations, ISBN 978-1-59904-252, Hershey, Newyork, 2007.
[14] Nan Tang, Big Data cleansing, Asia Pacific Web Conference, Springer, 2014.
[15] S Lakshmi, An overview on study Data Cleansing, its types and its methods for data mining, International Journal of Pure and Applied Mathematics, vol. 119, No. 12, 2018, 16837-16848.
[16] Pathak Anand Prakashbhai and Hari Mohan Pandey, Influence pattern from Big Data Aggrigation filtering and tagging- A survey, 2014, International Conference, The Next Generation information Technology Submit, Nov. 2014.
[17] X.L. Dong and D Srivastava ,Big Data Integration, Published Data, 2017-08-26.
[18] Misbachul Huda M, Dian Rahma Latifa Hayun and Zhin Martun, A Study on big Data ModelingTechniques, ULTIMA InfoSys, vol.VI, No.1, June 2015.
[19] Uma Bhat and Shraddha Jadhav, Moving Towards Non-Relational Database, International Journal of Computer Application, Vol.1, No.13, 2010.
[20] N Jacana, S Pure, Abuja M Kathuria and Gosain D, A survey and comparison of relational and non-relational database, International Journal of Engineering Research of technology, August 2012.
[21] Han J Haihong E, Guan Le and Jian Du, Survey on NoSQL Database Pervasive Computing Applications (ICPCA), 6th International Conference , Oct 2011, pp. 363-366.
[22] A Bhimani and L Willcocks, Digitisation, Big Data and transformation of accounting information, Accounting and Business Research, 44(4), 2014, pp.469-490.