International Journal of Scientific Research in Science and Technology
Print ISSN: 2395-6011 | Online ISSN: 2395-602X (www.ijsrst.com)
doi : https://doi.org/10.32628/IJSRST

Big Data: Research Issues and Challenges

Qamar Rayees Khan

Department of Computer Science, Baba Ghulam Shah Badshah University, Rajouri, Jammu & Kashmir, India

ARTICLE INFO

Article History:

Accepted: 05 Oct 2017
Published: 16 Oct 2017

Publication Issue
Volume 3, Issue 7
September-October-2017

Page Number
1485-1493

ABSTRACT

In the present digital and computing world, it is believed that over two billion people are connected to the internet, and over five billion people own mobile phones. By 2025, there is going to be a TSUNAMI of 50 billion devices to be connected with Internet. The prediction of data will be 45 times more than in couple of years and is growing at an exponential rate. These datasets generated by various sources are not only huge in volume but also high in velocity and variety, thus making it difficult to handle using traditional tools and techniques. There is a need to gain valuable insights in order to handle and extract knowledge from these data sets. Big data analytics help to provide that value and knowledge by discovering the unexpected patterns in the ocean of data. The aim of this paper is to provide overview of Big Data Analytics, characteristics of Big data, technologies, issues and challenges related with Big Data.

Keywords—Big Data, Analytics, Hadoop, HDFS, Map Reduce,

I. INTRODUCTION

In the current international population of 7.2 billion, more than 2 billion people are connected to the Internet [1]. According to McKinsey (2013), 5 billion individuals use various mobile devices [2]. This technological revolution leads to the generation of tremendous amount of data through the gigantic use of such devices. Data is generated at an exponential rate. According to the famous book “The Human Face of Big Data”, the amount of data generated by humans in 24 hours is equal to 70 times the information held in the Library of congress. [2]. Among various sources of data, sensors produce different variety of data that is either Structured or unstructured. This data is called as Big Data. [3]. Diebold et.al; [4] states that the term “big data” was coined by John Mashey in lunch term conversation at Slicon graphics (SGC) in 1990. The data holds the prime importance in today’s world. Organisations have realised that if that want to be competitive, they can’t ignore the big data at any cost. Following quotes from various data Scientist’s indicate the importance of Big Data.

“Information is the oil of the 21st century, and analytics is the combustion engine.” (Peter Sondergaard, Gartner Group)

“Data is the new science. Big Data holds the answers.” (Pat Geisinger, EMC)

“Data are becoming the new raw material of business.” (The Economist 2010)

A. Definitions

When the term "Big Data" is talked about, Size come into mind, however, there are other characteristics of big data that have emerged recently.

Laney et al; [5] suggested about the three dimensions of challenges in data management as Volume, Variety and Velocity (the three V's). **Gartner** also defines the Big data in similar fashion as, "Big data 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" ("Gartner IT Glossary.")

Apache Hadoop defines big data as; "Data sets which could not be captured, managed, and processed by general computers within an acceptable scope" [6].

McKinsey & Company defines Big Data as; "Big Data as the next frontier for innovation, competition, and productivity. Big data shall mean such datasets which could not be acquired, stored, and managed by classic database software" [6].

NIST defined Big data as; "Big data shall mean the data of which the data volume, acquisition speed, or data representation limits the capacity of using traditional relational methods to conduct effective analysis of the data which may be effectively processed with important horizontal zoom technologies" [6].

Tec America foundation defines big data as follows;
"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. [7]"

B. Description of Various V's in Big Data

- **Volume:** The word "volume" itself indicates about the big data. In recent days Enormous amount of data in terabytes and petabytes is produced that is very difficult for traditional databases to handle [8][9]. IBM conducted a survey in mid 2012 that revealed that more than half of the 1144 individuals considered datasets over one terabyte as big data. One tera byte means that the data that can be store. 16 million Facebook photographs or which can be stored on 1500 CDs or 220 DVDs [10]. Volume of big data is relative and depends upon time and type of data. As storage capacities increases day by day allowing bigger datasets to be stored, the data that may seem bigdata today may not be bigdata in future. Different datasets of same size require different technologies for management based upon their type, e.g., tabular, audio and video data. Thus, it is impossible to define a threshold for big data volume.
- **Variety:** Variety refers to the structural diversity in a dataset. Technological advancement help enterprise to use data in various types of data. The data can be in any format such as structured, unstructured, semi structured or mixture of these three. Structured data refers to the data that has a defined length and format as in spreadsheets or relational databases. 5% of the existing data is structured [11]. Unstructured data refers to the data that lack the structural organisation required by machines for analysis e.g., Audio, Video, pdf files, images and text. Semi structured data does not have a fixed Schema but has a simple label/value pairs. E.g. XML, EDI and SWIFT.
- **Velocity:** The Velocity refers to the rate of generation of data and the speed in which the data is analysed. The increasing of digital services such as sensors and smartphones has led to an imaginary rate of data creation. Various big retailers are generating data in higher rates. For example, Wal-Mart processes more than one million transactions in a single hour [11]. The generation of data from mobile devices produce violent flow of information that is used to generate real time offers for everyday customers. This data provides information about customers such as past buying patterns, geospatial location, demographics that can be analysed in real- time to create real customer value

- **Veracity:** This fourth dimension of big data was coined by IBM. It refers to the abnormality, biases and noise in data. Customer sentiments in social media is an example of veracity as it is uncertain and depends upon the human judgement, yet it contains valuable information. Various tools and techniques are developed for management and mining of such uncertain data.
- **Variability:** SAS introduced two additional dimensions of big data as Variability and Complexity. Variability refers to the variation in the rate of flow of dat. Big data velocity is not consistent and has periodic peaks and troughs. Complexity refers to the fact that big data are generated through a number of sources.
- **Value:** Oracle introduced Value as another attribute of big data. According to Oracle, big data are characterized by relatively "low value density". That is, the data received in the original form usually has a low value relative to its volume. Thus, high value of data can be obtained by analysing large volumes of such data.

The important fact about the dimensions discussed is that they are not independent of each other. When one-dimension changes, there is like hood of changing the other dimension. There exists a 'three-V tipping point' for every firm beyond which traditional data management and analysis technologies become inadequate for deriving intelligence. This tipping point is the threshold beyond which firms start dealing with big data.

C. History of Big Data

Megabyte to gigabyte: Historical business data introduced "Big data" challenge in moving from the size of Megabyte to Gigabyte. The pressing need at the time was to have the data housed and run relational queries for business analysis and reporting. Various efforts were done by researchers to give birth to database machine that resulted in integrated hardware and software in resolving the problems. The aim was that such integration would provide better performance at lower cost. After some time, it became evident that hardware-specialized database machines can't cope up with the pace at which general computers progressed. Thus, today's database systems are software systems that impose constraints on hardware but can run on general purpose computers.

Gigabyte to Terabyte: In 1980's, the advancement of digital technology caused data volumes to increase to several gigabytes, even to terabytes that was beyond the storage and processing capabilities of large computer system. Data Parallelization was proposed to increase the and to improve the performance by distributing the data and related tasks into incongruous hardware. Several parallel databases were built including shared-memory databases, shared disk databases, and shared- nothing databases.

Terabyte to Petabyte: During late 1990's, swift development of web 1.0 led the whole world into the internet era having semi structured and unstructured web-pages holding terabytes of data.

D. Different phases in big data

Different phases in Big data are data generation, data acquisition, data storage, and data analysis [6]. These phases are also considered to be the challenges of Big data.

i) Data generation

Big data generation is the first phase. The common sources of data generation are enterprises, IOT, Medical, Sensors, Mobile phones etc. Huge data generator sources are Sensors, search entries, weather forecasting, social networking sites. Enterprise data includes online trading data, production data, inventory data, sales data, financial data etc; which is needed to record information and data-driven activities in enterprises [6].

ii) Data Acquisition/Collection

Data acquisition is the second phase that includes the collection of data, transmission and its pre- processing.

Efficient transmission is used to send the collected dataset to an appropriate storage management system. As redundant and useless data is also present, collected dataset may require large storage. [6] [12] Data compression is applied to the collected dataset for efficient storage.

iii) Data Storage

This phase of big data is concerned with storage and management of large volume of data so that data is available whenever and wherever needed. On one hand, Reliability is the major concern in various Storage systems like massive storage systems, distributed storage system and big data storage system; on the other hand, it must allow in accessing the data using various query for analysis. Reliability should also be achieved in data processing [6]

iv) Data Analysis

Data Analysis is the final phase which involve extracting, finding, resolving and refining the data. The role of Data analysis is to plan user demands and needs and prediction of future trends. Traditional Analysis is used for structured data in which stastical methods are used. Other type of data analysis is called Big data analysis which is used for Structured, Unstructured and Semi Structured data. [6]

II. BIG DATA ANALYTIC PROCESSING

Big data sizes are increasing from a few dozen terabytes (TB) to many petabytes (PB) in a single dataset. Some of the challenges related to big data include capturing, storing, searching, sharing, analytics and visualizing. Today enterprise, are discovering facts that they didn't know before by exploring large volumes of highly detailed data. [18]. Big data analytics are the advanced analytic techniques that are applied on big data sets to reveal and leverage business change [18]. Nowadays, having piles of information is no longer enough to make better decisions. The evolution of technology and the increased amount of data flowing in and out of the organisation has led to the faster and more efficient data analysis. As big data is impossible to be analysed by the traditional tools and techniques, therefore, there arises a need for new tools and methods to be used for big data analytics. The new architecture is also needed for storing and managing such data. Thus, big data has an effect on everything on data from its collection, to processing till final extracted decisions.

A. HADOOP (Highly Archived Distributed Object-Oriented Programming)

Hadoop (Highly Archived Distributed Object-Oriented Programming) is an open source software that facilitates scalability and provides the distributed computing on the group of economical servers [13]. Input to Hadoop may be different amount of data from different sources like images,

videos, Audios, Sensor records, Files, Folders, e mail conversations. Thus, data may be of any format as Structured data, Un structured data and Semi structured data [14]. Hadoop also offers Documentation, location awareness, source code and work scheduling. The two main components of Hadoop are Hadoop Distributed File System (HDFS) and Map Reduce [15]. It also contains other components like Flume, HBase, Hive, Lucene, Pig, Oozie, Zookeeper, Avro, Chukwa. Hadoop works on Master and Slave architecture. Master node consists of Data node, Name node, Job tracker and Task tracker. Whereas Slave node consists of Task tracker and Data node. Slave Node is also called as Worker Node and is responsible for computation of data. Job Tracker performs the Scheduling of various Jobs.

a) Hadoop Distributed File System (HDFS)

HDFS is a special file system that works on distributed architecture. Actually, it is a framework written in java. HDFS works on Master Slave Architecture.

Name Node of HDFS acts as a Master and stores all the information whereas Data Node acta as a slave. The Name Node includes various type of information like Meta Data, Attributes, File Location, Job Tracker, Task Tracker, Active Node and Passive Node, Free Space and data they store, replication of data etc. It also keeps the record of location of files, metadata, attributes and data block of the Data Node [16]. When the client wants to read a file in HDFS, the name Node is contacted to collect the position of Data Blocks related to required files. The Master node (Name Node) always stores the doppelganger and periodic logs of a system. It also keeps track of replicas of files and free blocks in the system. The slave node (Data Node) saves the data and comprises of a task tracker that helps to track the active work of a Data Node and Job impending of Name Node [20]. Data Node keeps track of every block and sends its block activity log after every hour to the Name Node. The information about block replica is also provided to Name Node thus remaining up to date. Data Node also sends its heartbeat to Name Node after every ten seconds to specify its availability and operating accuracy. If the Name Node doesn't receive heart beat within ten seconds, it assumes the data node to be dead and produce replicas of file on the other nodes. [16]

b) Map Reduce (MP)

MapReduce was proposed by Google. It is a programming model that has been inspired by "Map" and "Reduce" of functional languages. It is a software program that is used to process distributed processing on huge datasets of the clusters in the computers. MapReduce is the core of Hadoop that performs the main task of data processing and analytics [21]. MapReduce paradigm is based on scaling out rather than scaling up i, e adding more resources or computers rather than increasing the power or storage capacity of a single computer [22]. The main idea here is to break down a task into stages and stages and execute the stages in parallel so as to reduce the time needed to complete the task [21].

In the first phase, MapReduce maps input values to a set of key/value pairs as output. The role of "Map" function is to divide large computational tasks into smaller tasks and allocate them to appropriate key/value pairs [21]. For instance, unstructured data such as text is mapped to a structured key/value pair. The key may be the word in the text and the value may the number of occurrences of word. The output from this process is then input to the "Reduce" function [22]. Reduce performs the task of collection and combination by combining all those values which share the same key value, thus providing the final result of computational task [21].

The MapReduce function depends upon Job tracker and the Task tracker. The Job tracker nodes distributes the mapper and reducer functions to the available task trackers and also monitor the results [22]. The MapReduce job starts by the Job Tracker in such a way that it assigns a portion of input file on the HDFS to a map task running on a node [23]. The Task Tracker is actually responsible for running the jobs and communicating the results back to the Job Tracker. Inter node communication is minimized as communication between nodes is often through files and directories [22].

III.ISSUES AND CHALLENGES IN BIG DATA

In this new technological era, data is everywhere whether it comes from social networking sites, banking and Education Many crucial challenges need to be focussed on handling the Big data and analytical process [24]. Table 1. Summarises the Big data processes, Challenges and Solutions thereof.

Table 1 Big Data Processes, Challenges and Solutions.

Process	Challenges	Solutions	Key references
Data Access and Collection	• Easy access to data offered in standardized formats. No practical limit to the size of these data offering unlimited scalability • Efficiently obtain detailed data for a large number of agents • Protocols on security, privacy, and data rights	• Sensors • Web scraping • Web traffic and communications monitoring	[27][28]
Data Storage	• Tools for data storage, matching and integration of different big datasets • Data reliability • Warehousing	• SQL, NoSQL, Apache Hadoop • Save essential information only and update in real time	[29][30]
Data Processing	• Use non-numeric data for quantitative analyses	• Text mining tools to transform text into numbers • Emotion recognition	[31][32]
Data Analysis	• Large number of variables • Causality • Find latent topics and attach meaning • Data too large to process	• Ridge, lasso, principal compo-nents regression, partial least squares, regression trees • Topic modeling, latent Dirichlet allocation, entropy- based measures, and deep learning • Cross-validation and holdout samples • Field experiments • Parallelization, bags of little boot-strap, sequential analysis	[33][34][35]
Reporting and Visualization	• Facilitate interpretation, representation with external partners and knowledge users • Difficult to understand complex patterns	• Describe data sources • Describe methods and specifications • Bayesian analysis • Visualization and graphic interpretations	[36][37]

Some of the challenges on which researchers are mainly focussing on are briefly discussed in subheadings, a-k

a) Storage:

There are many sources of data production whether it is Internet or other sources. Today, data produced by internet is in Exabytes that has led to the data explosion. Data will get much bigger in future. Since the traditional database management tools will not be able to store or process such kind of data, it need's the different mechanism to handle such data.

b) Representation of Data:

There is a Heterogeneity of data that exists in structure, type, semantics, granularity and accessibility. The representation of data is more important for data analytics and user analysis. Improper representation of data reduces its originality and disturbs the data analysis process.

c) Life Cycle management of data:

As the existing data storage system can't support huge amount of data, the role of data life cycle management is to decide which data should be stored and which data should be discarded during analytical process.

d) Analysis:

Data is generated from broad range of resource that may vary in volume and structure. Therefore, Huge time and resources are required in analysing the data. To solve this issue, special scaled out Architectures are needed to process this data in a distributed manner. Data is divided into different fragments and handled in a number of computers in a distributed fashion. Then the processed data is combined.

e) Processing issue:

To process the large amount of big data using traditional tools is impossible. While collecting the data, Processing time is considerably reduced using an Index. E.g. Assuming the process of an exabyte of data. Processor expands 200 instructions in one block at 5 GB, the time required for processing is 20 seconds. The generation of new analytics algorithm need to be produced in order to provide timely and actionable information [26]

f) Reporting:

Reporting involves displaying the data in the form of values. For the huge data, it is challenging and difficult to understand the results. In that case, results should be displayed in a manner that it becomes easy to understand the results.

g) Reducing redundancy & Data Compression:

The cost of system is decreased by using the mechanism of Redundancy reduction and Data compression. e.g. Sensors produce highly redundant data that can be filtered to reduce the Redundancy.

h) Energy Management:

The power consumption control and management mechanism are to be implemented for big data as data growth rate, analytical process, transmission management and storage management will consume more electric energy. Thus, Energy management ensures less consumption of energy.

i) Security and privacy:

The Security and Privacy is an important challenge for every domain. As Big data is generated from various sources where it contains private data such as credit card data, personal ID and other sensitive assets.

j) Confidentiality of Data:

The owners and Service providers of the big data can't analyse the huge datasets in a proper manner so Data confidentiality becomes another challenge for big data. They are dependent on professionals or third-party tools to analyse such data. Thus, there is a potential risk to involving such third-party tools. Hence, is an Important issue for the researchers.

k) Management and transport issue:

Management issue deals with building a database which deals with huge data. Data are distributed geographically managed by multiple entities. The data sources can vary both spatially and temporally [25]. Transport issue is also one of the challenge in Big Data. If the data is processed using current technologies, it will overwhelm the network communication [25].

IV. CONCLUSION

The above study provides the basic information about Big data, characteristics of Big Data, and Big Data analytics. History of Big data and Issues & Challenges are discussed in a detailed manner. The paper focusses about the working of HADOOP. The detailed explanation about components of HADOOP like

HDFS (Hadoop distributed file system) and MapReduce are provided.

V. REFERENCES

[1]. Worldometers, “Real time world statistics,” 2014, http://www.worldometers.info/world-population/.

[2]. http://www.digitalpromise.org/blog/entry/realizing- theopportunity-for-big-data-in-education

[3]. D. Che, M. Safran, and Z. Peng, “From Big Data to Big Data Mining: challenges, issues, and opportunities,” in Database Systems for Advanced Applications, pp. 1–15, Springer, Berlin,Germany, 2013.

[4]. Diebold, F. X. (2012). A personal perspective on the origin(s) and development of “big data”: The phenomenon, the term, and the discipline (ScholarlyPaper No. ID 2202843). Social Science Research Network

[5]. Laney, D. (2001, February 6). 3-D data management: Controlling data volume, velocity and variety. Application Delivery Strategies by META Group Inc

[6]. Min Chen, Shiwen Mao, and Yunhao Liu, “Big Data: A Survey”, 2014 Springer.

[7]. Amir Gandomi and MurtazaHaider, “Beyond the hype:Big data concepts, methods, and analytics”, 2015 ELSEVIER.

[8]. Avita Katal, Mohammad Wazid, and R H Goudar, “BigData: Issues. Challenges, Tools and Good Practices”, 2013 IEEE.

[9]. Uzma Shafaque, Parag D. Thakare, Mangesh M. Ghonge, and Milindkumar V. Sarode, “Algorithm and Approaches to Handle Big Data”, National Level Technical Conference “XPLORE 14, 2014 IJCA.

[10].Schroeck, M., Shockley, R., Smart, J., Romero-Morales, D., & Tufano, P.(2012). Analytics: The real-world use of big data.How innovative enterprises extract value from uncertain data. IBM Institute for Business Value.

[11].Cukier K., The Economist, Data, data everywhere: A special report on managing information, 2010

[12].“Challenges and Opportunities with Big Data”, Acommunity white paper developed by leading researchers across the United States, 2012.

[13].Dhole Poonam B, GunjalBaisa L, “Survey Paper on Traditional Hadoop and Pipelined Map Reduce”,International Journal of Computational Engineering Research Vol 03, Issue12

[14].“Leveraging Massively Parallel Processing in an Oracle Environment for Big Data”, An Oracle White Paper, November 2010.

[15].Jeffrey Dean and Sanjay Ghemawat, “Map Reduce: Simplified Data Processing on Large Clusters”, Google, Inc.

[16].Vibhavari chavan, Rajesh N. Phursule, “Survey paper on Big Data, IJCSIT , Vol. 5, 2014, PP. 7932- 7939”. Technologies, Vol. 5 (6), 2014,7932-7939”.

[17].Kubick, W.R.: Big Data, Information and Meaning. In: Clinical Trail insights, Pp. 26-28, 2012

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

[19].Elgendy, N.: Big Data Analytics in Support of the Decision Making Process. MSc Thesis, German University in Cairo, p. 164 (2013)

[20].Suman Arora, Dr. Madhu Goel, “Survey Paper on Scheduling in Hadoop”, International Journal of Advance Research in Computer Science and Software Engineering Volume 4, Issue 5, May 2014

[21].Cuzzocrea, A., Song, I., Davis, K.C.: Analytics over Large-Scale Multidimensional Data:The Big Data Revolution! In: Proceedings of the ACM International Workshop on Data Warehousing and OLAP, pp. 101–104 (2011)

[22].EMC: Data Science and Big Data Analytics. In: EMC Education Services, pp. 1–508 (2012)

[23].Lee, R., Luo, T., Huai, Y., Wang, F., He, Y., Zhang, X.: Ysmart: Yet Another SQL-to-MapReduce Translator. In: IEEE International Conference on Distributed Computing Systems (ICDCS), pp. 25–36 (2011)

[24].Chen H, Chiang RHL, Storey VC. Business intelligence and analytics: from big data to big impact. MIS Quart.2012; 36(4):1165–88

[25].Stephen Kaisler, Frank Armour, J. Alberto Espinosa, andWilliam Money, "Big data: Issues and challenges Moving forward", 2012 IEEE.

[26].Avita Katal, Mohammad Wazid, and R H Goudar, "BigData: Issues. Challenges, Tools and Good Practices", 2013 IEEE.

[27].Chaffin, D., Heidl, R., Hollenbeck, J. R., Howe, M., Yu, A., Voorhees, C., & Calantone, R. The promise and perils of wearable sensors in organizational research. Organizational Research Methods. Published online ahead of print. doi: 10.1177/1094428115617004, 2015.

[28].Sismeiro, C., & Bucklin, R. E. Modeling purchase behavior at an e-commerce web site: A task-completion approach. Journal of Marketing Research, 41: 306–323, 2004.

[29].Varian, H. R. Big data: New tricks for econometrics.The Journal of Economic Perspectives, 28: 3–27, 2014.

[30].Prajapati, V. Big data analytics with R and Hadoop.Birmingham, England: Packt Publishing,2013.

[31].Manning, C. D., Raghavan, P., & Schu¨ tze, H, Intro- duction to information retrieval. Cambridge, England: Cambridge University Press, 2009.

[32].Teixeira, T., Wedel, M., & Pieters, R. Emotion-induced engagement in internet video advertisements. Journal of Marketing Research, 49: 144–159, 2012

[33].Hastie, T., Tibshirani, R., & Friedman, J. 2009. The el- ements of statistical learning: Data mining, in- ference and prediction (2nd ed.). Berlin, Germany: Springer.

[34].George, E. I., & McCulloch, R. E,. Variable selection via Gibbs sampling. Journal of the American Statis- tical Association, 88: 881–889, 1993

[35].Archak, N., Ghose, A., & Ipeirotis, P. G. . Deriving the pricing power of product features by mining consumer reviews. Management Science, 57: 1485– 1509,2011.

[36].Loughran, T., & McDonald, B. When is a liability not a liability? Textual analysis, dictionaries, and 10-Ks. The Journal of Finance, 66: 35–65,2011.

[37].Simonsohn, U., Simmons, J. P., & Nelson, L. D, Specification curve: Descriptive and inferential statistics on all reasonable specifications. Available online at SSRN. doi: 10.2139/ssrn.2694998, 2015

Cite This Article :

Qamar Rayees Khan , " Big Data: Research Issues and Challenges", International Journal of Scientific Research in Science and Technology(IJSRST), Print ISSN : 2395-6011, Online ISSN : 2395-602X, Volume 3, Issue 7, pp.1485-1493, September-October-2017.

Journal URL : https://ijsrst.com/IJSRST523102114