IOSR Journal of Computer Engineering (IOSR-JCE)
e-ISSN: 2278-0661,p-ISSN: 2278-8727, Volume 22, Issue 1, Ser. II (Jan - Feb 2020), PP 01-05
www.iosrjournals.org

A Brief about Big Data, It's Technology and Challenges

Devesh Malik, Pawan Kumar Goel
Assistant ProfessorDepartment of CSE,Shri Ram Group of Colleges,Muzaffar Nagar (U.P.) India
Associate Professor& Head Department of CSE,Shri Ram Group of Colleges,Muzaffar Nagar (U.P.) India

Abstract:*Big Data is the leading topic in the current e youth, the data they generate while communicating with each other. E youth is not the only reason of generating Big Data. The rapid growth of Internet uses, Sensors, IOT, and computing led to the extensive growth of data in various industries and field. This paper provides a brief about Big Data. It also covers the thought of various and characteristics described by various author about Big Data. The technology used in Big data. This paper also describes about various challenges associated to Big Data.*

Keywords:*Big Data, NOSQL, 5 Vs, Pig, HBase, Cloudera, RC File, Hadoop etc.*

Date of Submission: 15-01-2020 Date of Acceptance: 03-02-2020

I. Introduction

Today generating data is a very simple task for people. There is huge exponential amount of data generated by people and system which overflowing the web. The amount of data generated on the web isnear about 1 exabytes (1018) and 1 zetta bytes (1021). Big data has been converted into a hotspot that attracts great attention from industry, governments and academy around the world [1-3]. This growth of data due to digital sensors, communications between friends, computation done by automated system and storage. By 2025, it is expected that the Internet will exceed the brain capacity of everyone living in the whole world [5]. The term Big Data has been first introduced by "ROGER MAGOULAS" [6]. In same direction , M. Pospiech and C. Felden [7] have presented his views on various aspects of Big Data and have divided them in four domains: first is data provisioning which include data acquisition, data storage and data processing, Second is data utilization which include computation and time complexity, third one Functional data provisioning which include information life cycle management,information managementetc.) and last one is Functional data utilization which include place where big data can be used.In a deepersense big data can be regarded as a bond that connects and integrates the physical world through internet, the Internet of Things and other information technologies, while human society generates its big data-based mapping human-computer interface, brain machine interfaces and by using mobile[8-10]. So big data canbe classified into two categories, first is , data from the physical world, which is obtained through sensors, scientific experiments done by human, equipment created by human and observation or analysed data, neural data, astronomical data and remote sensing data and data from the human society like social network, Internet, Health, Finance, Economics and Transportation. In this paper we are going to talk about the what is big data in section 1, its management in section 2, technology in section 3, and challenges in section 3.

WHAT IS BIG DATA?

Various author define Big Data in different ways like Manyika define Big Data as "Datasets whose size in beyond the ability of traditional database management software to retrieve, store, manage, and analyse data"
Davis and Patterson say "Big data is data too big to be handled and analysed by traditional data base such as SQL"
Both authorsfocus only one aspect of data that is size of data. There are many other aspects.
EddDumbill in explicitly conveys the multi dimension of big data when adding that "the data is too big, move too fast or doesn't fit the strictures of your database architectures".
Many authors use three Vs (Volume, Variety and Velocity) and some of them use 5 Vs to characterize Big Data.
Volume is related to large amount of data.
Variety is related to many forms;
Velocity is related to the various speed data is coming.
Compared to traditional data the feature of big data can be characterize by 5 v namely Volume, Variety, Velocity, Veracityand Value
Variety Is related to diversified data type
Velocity is related to timely and speed
Veracity means uncertainty in data

Value means importance of data
Due to this diversification in data type, the data can be divided in three type

- Structure data like table
- Semi structure like email
- unstructured like text, image and video, and voice.

Big data has great value for information industry, big data is strong ingredients to the next generation of IT industry.

II. Big Data Management

Data management is related to storage, retrieval and modification of data. Thus, key challenges in big data are related to storage, transportation and processing of high throughput and speed data. It is completely different from Big Data challenges to which is full ambiguity, uncertainty and variety.

According to Karmasphere [6] Big Data analysis is divided into four steps: Acquisition or Access, Assembly or Organization, Analyse and Action or Decision. Thus, these steps are also known as the "4 A's" of data management.

Acquisition or Access: it is the step where acquire of high-speed data from a various type of sources like web, DBMS(OLTP), NoSQL, HDFS take place and it has to deal with various access protocols. It is the place where a filter could be established to store data which could be useful, helpful or "raw" data with a lower degree of uncertainty.

Organization: this is the point where data have to be clean and need to be put in a computable mode, integrated, structured or semi-structured form and stored in the right location like existing data warehouse,Operational Data Store, data marts, Complex Event Processing engine, NoSQL database etc [14]. Thus, in this phase extract, transform, load had to be done. Cleaning Big Data is not entirely guaranteed in this phase, in fact "the volume, velocity, variety, and variability of Big Data may preclude us from taking the time to cleanse it all thoroughly".

Analyse: Here we have to run queries, modelling, and building algorithms to find new information. Here Mining of data take place which requires integrated, cleaned, trustworthy data.

Decision: In this phase the user is able to take valuable decisions which means interpret results from analysis. Consequently, it is very important step for the user to "understand and verify" the expected outputs [14].

III. Big Data Technologies

There are various tools which can be used in Big Data management from first stage that is data acquisition to last stage data analysis. Most of these tools are influence by Apache projects and are constructed like Hadoop which written in Java and created by "Doug Cutting", Hadoop brings the cheapest mean process large amounts of data, without considering its structure.
Hadoop is made up of two sub domains: Hadoop Distributed File System (HDFS) and MapReduce.

HDFS

Distributed file system is a file system which is basically used to access file form remote place and for preparing disk less system. Where asHDFS is a distributed file system designed to run on large clusters of based on Google File System (GFS).
"Shvachko "[17, page 1] says "HDFS is designed to store very large datasets reliably and easily, and to provide access of those datasets at high bandwidth to various user". By large, its mean near about from 10 to 100 GB and above form it.
HDFS is not for interactive use rather it is dedicated to batch processing [16,13].
In HDFS, files are written one time and accessed more than one times [16,18]; data coherency is ensured by it and data are accessed with high throughput [16].
HDFS file system metadata that data about data is stored in a dedicated server and the Name Node and the application data is on the other servers known as Data Nodes.
HDFS is also responsible for handling failure at application level. This can be achieved by creating replication that is a copy of data set, they are replicated on a number of data node after division of file into block. all the data nodes containing a replica of a block are not located in the same rack.

MapReduce

MapReduce is a programming model and associated implementation for processing and generating large datasets [19].
MapReduce framework is application specific [20] and is most appropriate for semi structured or unstructured data. The output of MapReduce is a set of <key, value> pairs.

The name "MapReduce" expresses the fact that there is use of an algorithm with two kernel functions: "Map" and "Reduce". The Map function is applied on the input data and output is a list of intermediate <key, value> pairs; and the Reduce function merges all intermediate values associated with the same intermediate key [19] [20].

In a Hadoop cluster, MapReduce program [12] is executed by subsequently breaking it down into pieces called tasks. When a node in Hadoop cluster receives a job, it is able to divide it, and run it in parallel over other nodes [13]. Here the data location is tracked by the Job Tracker which communicates with the Name Node to help data nodes to send tasks to near-data data nodes.

Storage and management Capability:

Cloudera Manager, RCFile (Record Columnar File) [21], It is an efficient storage structure which allows fast data loading and query processing.

Database Capability:

These are the name of technology for data base Capability Oracle NoSQL, Apache HBase, Apache Cassandra, Apache Hive can be seen as a distributed data warehouse [15].Apache ZooKeeper is "an open-source software, in-memory, distributed NoSQL database [4, page 69] that is used for coordination and naming services for managing distributed applications [4,13,12,15].

Processing Capability:

Pig

It is used to analysing large datasets and thus spend less time having to write mapper and reducer programs [12,13]

Chukwa, Oozie

It is an open-source tool for handling complex pipelines of data processing [13,3,12].Using Oozie

Data Integration Capability

Apache Sqoop, Flume is a distributed, reliable, and available service for efficiently collecting, aggregating, and moving large amounts of log data. It is designed to import streaming data flows [12,22].

Visualization techniques

the ultimate goal of Big Data analysis is Making valuable decisions and the achievement of this goal requires good visualization of content.

Real-time updates require large amount processing and not stored in a relational way [4].

The name of techniques for Big Data visualization.9 is as follow

Tag Cloud

Tagging is similar to face book tagging. It is one of the methodsfor linkingand visualizing concepts. Here concepts are written using properties such as font size, Font style, colour.

Clustergram

M. Schonlau [23] defines cluster gram as a visualization technique used for cluster analysis

History Flow

F.B. Viégas, M. Wattenberg and K. Dave [24] present history flow as a visualization technique designed to show the evolution of a document efficiently

IV. Data Analytics

Big Data Analytics is not a simple task it can be defined as the use of very advanced analytic techniques [25].

There are some need for the development of big data Analytics:

- Tools and storage capabilities is in such a way that it can handle big data.
- So, it become easy to provides large statistical samples and enhanced results of experiments.
- Companies and Governments have knowledge of benefits to develop the economics of big data.

there are many variety techniques used for analytics on big data [26].

Association rule learning

It is mainly used to find relationships among entities

Machine learning

It is mainly used to bring computer to learn complex patterns and make intelligent decisions based on it [11].

Data mining

It can be seen as a combination of statistics and machine learning [11].

Cluster analysis

It aims to divide data into smaller clusters having the same set of characteristics not known in advance.

V. Challenges Of Big Data

There is sequence of challenges attached to the big data. Some of them are related Data Integration, complexity of data, complexity of data processing.

As it is known Big Data is big and complex, thus challenges can be classified into two categories

- Tasks related to managing data at a very large scale
- Finding and combining information that is relevant as per needs [27]

The meaningful data integration or combination challenge can be viewed as five-step challenge: (1) define the problem.
(2) identify relevant pieces of data in Big Data
(3) Extract Transform and Load (ETL) it into appropriate formats and store it for further processing
(4) remove it ambiguate
(5) solve the problem.
All these steps itself is a challenge of big data.
2. Billion Triple Challenge which aims to process largescale Record Description Format (RDF) to provide a detailed description of entire entity of the triple in a simple vocabulary and to link each entity to the corresponding sources.
3. Linked Open Data (LOD) Ripper for providing good use cases is another challenge.
Next leading Challenge is Data Complexity.
The typical characteristics of big data are

- Various type and patterns
- Complex computation
- Complicated interrelationship
- Varied data quality

Due to large size of data, traditional data analysis and sentiment analysis become a difficult task. Thus, there is need of new traditional data analysis tool and become a challenge for big data.

The traditional computation method like machine learning, information retrieval and data mining not enough to work with this huge amount and variable velocity of data. Thus, there is need of new novel and highly efficient computation method and be a challenge of big data.

VI. Conclusion

This can be concluded here that Big Data is the one of the leading Topic and leading Issue of coming generation. Due to huge size, volume, velocity, value and veracity there need to change the technology with advancement. There are lot of other challenges associated to the big data include traditional data analytic tool, traditional computation method, problem finding, problem solving, gathering data, organizing Big Data all comes under It.

References

[1]. V. Mayer-Schonberger, K. Cukier, Big Data: A Revolution That Will Transform How We Live, Work, and Think, Houghton Mifflin Harcourt, 2013.
[2]. R. Thomson, C. Lebiere, S. Bennati, Human, model and machine: a complemen-tary approach to big data, in: Proceedings of the 2014 Workshop on HumanCentered Big Data Research, HCBDR '14, 2014.
[3]. A. Cuzzocrea, Privacy and security of big data: current challenges and future research perspectives, in: Proceedings of the First International Workshop onPrivacy and Security of Big Data, PSBD '14, 2014.
[4]. K. Krishnan, Data warehousing in the age of big data, in: The Morgan Kaufmann Series on Business Intelligence, Elsevier Science, 2013.
[5]. A. Reeve, Managing Data in Motion: Data Integration Best Practice Techniques and Technologies, Morgan Kaufmann, 2013.
[6]. D. Agrawal, S. Das, A. El Abbadi, Big data and cloud computing: current state and future opportunities, in: Proceedings of the 14th International EDBT, EDBT/ICDT '11, ACM, New York, NY, USA, 2011, pp. 530–533.10
[7]. M. Pospiech, C. Felden, Big data—a state-of-the-art, in: AMCIS, Association for Information Systems, 2012.
[8]. G. Li, X. Cheng, Research status and scientific thinking of big data, Bull. Chin. Acad. Sci. 27 (6) (2012) 647–657.
[9]. Y. Wang, X. JinXueqi, Network big data: present and future, Chinese J. Comput. 36 (6) (2013) 1125–1138.
[10]. X.-Q. Cheng, X. Jin, Y. Wang, J. Guo, T. Zhang, G. Li, Survey on big data system and analytic technology, J. Softw. 25 (9) (2014) 1889–1908.
[11]. J. Manyika, M. Chui, B. Brown, J. Bughin, R. Dobbs, C. Roxburgh, A.H. Byers, Big Data: The Next Frontier for Innovation, Competition, and Productivity, McKinsey Global Institute, 2011.
[12]. P.Zikopoulos,C.Eaton,UnderstandingBigData:Analyticsfor Enterprise Class Hadoop and Streaming Data, McGraw-Hill Education, 2011.
[13]. I. O'Reilly Media, Big Data Now: 2014 Edition, O'Reilly Media, 2014.
[14]. H.V. Jagadish, D. Agrawal, P. Bernstein, E. e. a. Bertino, ChallengesandOpportunitieswithBigData,TheCommunity Research Association, 2015.
[15]. T. White, Hadoop: The Definitive Guide, first ed., O'Reilly Media, Inc., 2009.
[16]. D. Borthakur, The hadoop distributed file system: Architecture and design, The Apache Software Foundation. (2007) 1–14.
[17]. K. Shvachko, H. Kuang, S. Radia, R. Chansler, The hadoop distributed file system, in: Proceedings of the 2010 IEEE 26th Symposium on Mass Storage Systems and Technologies (MSST), MSST '10, IEEE Computer Society, Washington, DC, USA, 2010, pp. 1–10.
[18]. G. Turkington, Hadoop Beginners Guide, Packt Publishing, Limited, 2013.
[19]. J. Dean, S. Ghemawat, Mapreduce: simplified data processing on large clusters, Commun. ACM 51 (1) (2008) 107–113.

[20]. C. Ranger, R. Raghuraman, A. Penmetsa, G. Bradski, C. Kozyrakis, Evaluating mapreduce for multi-core and multiprocessorsystems,in:Proceedingsofthe2007IEEE13th International Symposium on High Performance Computer Architecture, HPCA '07, IEEE Computer Society, Washington, DC, USA, 2007, pp. 13–24.
[21]. Y. He, R. Lee, Y. Huai, Z. Shao, N. Jain, X. Zhang, Z. Xu, Rcfile: A fast and space-efficient data placement structure in mapreduce-based warehouse systems, in: Proceedings of the 2011IEEE27thInternationalConferenceonDataEngineering, ICDE '11, IEEE Computer Society, Washington, DC, USA, 2011, pp. 1199–1208.
[22]. C. Wang, I.A. Rayan, K. Schwan, Faster, larger, easier: reining real-time big data processing in cloud, in: Proceedings of the PostersandDemoTrack,Middleware'12,ACM,NewYork,NY, USA, 2012, pp. 4:1–4:2.
[23]. M. Schonlau, The clustergram: A graph for visualizing hierarchical and nonhierarchical cluster analyses, Stata J. 2 (4) (2002) 391–402. 12.
[24]. F.B.Viégas,M.Wattenberg,K.Dave,Studyingcooperationand conflict between authors with history flow visualizations, in: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, CHI '04, ACM, New York, NY, USA, 2004, pp. 575–582.
[25]. P. Russom, et al. Big data analytics, TDWI Best Practices Report, Fourth Quarter.
[26]. D. Maltby, Big data analytics, in: 74th Annual Meeting of the Association for Information Science and Technology (ASIST), 2011, pp. 1–6.
[27]. C. Bizer, P. Boncz, M.L. Brodie, O. Erling, The meaningful use of big data: four perspectives – four challenges, SIGMOD Rec. 40 (4) (2012) 56–60.

Devesh Malik, Pawan K. Goel et.al, A Brief about Big Data, It's Technology and Challenges." *IOSR Journal of Computer Engineering (IOSR-JCE)*, 22.1 (2020), pp. 01-05.