Challenges in Big Data Application: A Review

Satanand Mishra
Scientist
CSIR-AMPRI,
Bhopal

Vijay Dhote
PG Scholar
VNS Group,
RGPV
Bhopal

G. S. Prajapati
Asstt. Professor
VNS Group,
RGPV
Bhopal

J.P. Shukla
Principal Scientist
CSIR-AMPRI,
Bhopal

ABSTRACT

New invention of advanced technology, enhanced capacity of storage media, maturity of information technology and popularity of social media, business intelligence and Scientific invention, produces huge amount of data which made ample set of information that is responsible for birth of new concept well known as big data. Big data analytics is the process of examining large amounts of data. The analysis is done on huge amount of data which is structure, semi structure and unstructured. In big data, data is generated at exponentially for reason of increase use of social media, email, document and sensor data. The growth of data has affected all fields, whether it is business sector or the world of science. In this paper, the process of system is reviewed for managing "Big Data" and today's activities on big data tools and techniques.

Keywords

Big data, big data challenges and management, Hadoop, HDFS, Hadoop component.

1. INTRODUCTION

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. The primary sources for big data are from business applications, public web, social media, and sensor data and so on. Big data is currently a major topic of discussion across a number of fields, including management and marketing, scientific research, national security, government transparency and open data. Both public and private sectors are making increasing use of big data analytics. The need to process and analyze such massive datasets has introduced a new form of data analytics called Big Data Analytics. Big Data analytics involves analyzing large amounts of data of a variety of types to uncover hidden patterns, unknown correlations and other useful information. Five years ago, only big business could afford to profit from big data: Wal-Mart and Google, specialized financial traders**.** Due to its specific nature of Big Data, it is stored in distributed file system architectures. Today, we are using an open source project called Hadoop, commodity Linux hardware and cloud computing, this power is in reach for everyone [1] [13] [15]. The Big Data is the combination of structured, semi-structured, unstructured, homogeneous and heterogeneous data. With the insertion of big data and massive scientific data sets that are measured in tens of peta bytes, a user has the relevant to transfer even larger amounts of data [35].

1.1 Big data development-

In big data development three basic pillars are used volume, velocity, variety.

Data volume**:** Volume is the first and most flaming feature. In the year 2000, 800,000 petabytes of data were stored in the world. This number is expected to reach 35 zeta bytes by 2020.Social media generate around 10TB data every day. Big data are used to organize the large amount of data. The growth of data volume imposes numerous requirements on I/O performance, which results in I/O bottlenecks in current HEC (High End Computing) machines [14].

Data velocity**:** Velocity is refers to how quickly the data is generated and store, and its suitable rate of data processing and retrieval. Now a day data is more complex, data size in terms of hundreds of terabytes, petabytes, Exabyte's etc. Complex data handling is difficult or impossible for traditional system. So many organizations must choose better analytical tool to effectively deal with big data.

Data variety**:** Variety means different types of data. Today the increase use of smart devices, as well as social technology, sensor, data has became a large and complex because it does not only include traditional data but including some other data like structure data, semi structure, and unstructured data from different sources such as search engine, social media, web pages, sensor data, document etc[2] [3].

The characteristics of Big Data can be broadly divided into five Vs i.e. Volume, Velocity, Varity and Variability; Veracity. Volume refers to the size of the data It is the size of the data which determines the value and potential of the data under consideration and whether it can actually be considered Big Data or not. While Velocity tells about the speed at which data is generated; Varity and Variability tells us about the complexity and structure of data and different ways of interpreting it Variability is a factor which can be a problem for those who analyse the data., Veracity refers accuracy of analysis depends on the veracity of the source data. Fig 1 shows the development pillar of big data and characteristics of big data.

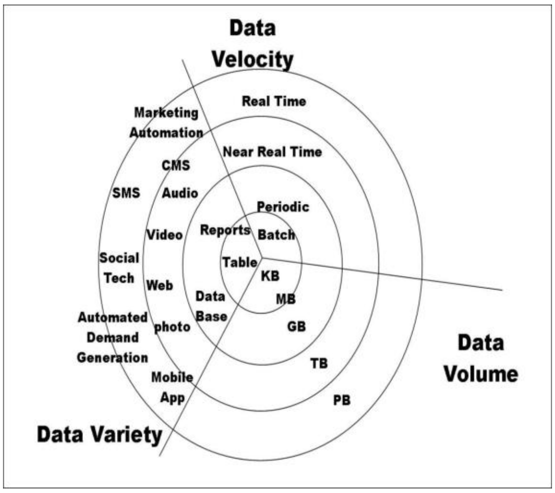

Fig 1: Development pillar of big data and its characteristics

1.2 Big data marketing survey-

Big data has increased the use and demand day by day in information management specialists and many software companies like IBM, Microsoft, HP and Dell have spent more than $15 billion on software firms specializing in data management and analytics. There are 4.6 billion mobile-phone subscriptions worldwide and between 1 billion and 2 billion people accessing the internet [4].

Between 1990 and 2005, more than 1 billion people worldwide entered the middle class which means more and more people who gain money will become more literate which in turn leads to information growth. The world's effective capacity to exchange information through telecommunication networks was 281 petabytes in 1986, 471 petabytes in 1993, 2.2exabytes in 2000, 65 Exabyte's in 2007.Traffic flowing over the internet will reach 667 Exabyte's annually by 2014 and till 2015 it is expected to increase three times[5].

2. CHALLENGES IN BIG DATA

Big data challenges are usually the real implementation interrupt which require immediate attention, if any implementation without handling these challenges may lead to failure of technology and some unfavourable result [20].

***Privacy and security*:** It is most important challenge of big data it is sensitive and technical as well as legal significance. The personal information of any person combine with the large data set, that define to the presuming new facts about that person and its possible that this kind of facts about the person is secretive and the person might not want the data owner to know or any person to know about them.

***Analytical challenge*:** Big data brings along with it some huge analytical challenges. Big data analytics is the process of examining big data to uncover hidden patterns, unknown correlations and other useful information that can be used to make better decisions. The types of analysis to be done on this huge amount of data which is unstructured, semi structured and structured they need large technical skill.

2.1 Technical challenges:

Fault tolerance – New incoming technologies like cloud computing and big data it is always that whenever the failure occurs the damage is done should be within acceptable threshold rather than beginning the whole task from the scratch. Fault tolerant computing is extremely hard, involving difficult algorithms. Thus the main task is to reduce the probability of failure to an acceptable level.

Quality of data- Storage and collection of huge amount of data are more costly. More data if used for decision making or for predictive analysis in business will definitely lead to better results. Big data basically focuses on quality of data storage rather than having very large irreverent data so that better result and conclusion can be drawn.

Heterogeneous data- In big data unstructured data represent almost every kind of data being produced by social media interaction, recorded meeting, handle PDF document, fax transfer, to email and more. Converting unstructured data in to structured data one is also not feasible. Structured data is organized but unstructured data is completely raw and unorganized [6] [19].

2.2 Forecast to the future in big data:

There are many future challenges in big data analytics and management. Some of the challenges that researchers and practitioners will have in future

Distributed mining- *Many* data mining techniques are not trivial to paralyze. To have distributed versions of some methods, a lot of research is needed with practical and theoretical analysis to provide new methods.

Analytics Architecture- It is not clear yet how an optimal architecture of analytics systems should be to deal with historic data and with real-time data at the same time. An interesting proposal is the Lambda architecture of Nathan Marz [31].

Compression- Dealing with Big Data, the quantity of space needed to store it is very relevant. Using compression, we may take more time and less space, so we can consider it as a transformation from time to space. Using sampling, we are losing information, but the gains in space may be in orders of magnitude. For example Feldman et al. [32] use core sets to reduce the complexity of Big Data problems.

Visualization- A main task of Big Data analysis is how to visualize the results of any data. Because of data is so big, it is very difficult to find user-friendly visualizations.

Hidden Big Data- Large quantities of useful data are getting lost since new data is largely untagged file based and unstructured data. [33]

3. DATA MANAGEMENT

In data management most important point is how do we store and process such huge amount of data, most of data which is raw, some semi structure and unstructured. Big data platforms categorized on how to store and process data in scalable, fault tolerant, and efficient manner.

For handling of big data two important information management tool are used such as Relational DBMS (RDBMS) used for systematic workload and non-relational techniques (NOSQL) for handling raw data, semi structure and unstructured data [7]. We conducted extensive experiments where we compared the most representative algorithm from each of the categories using a large number of real (big) data sets. The effectiveness of the candidate clustering algorithms is measured through a number of internal and external validity metrics, stability, runtime, and scalability tests [18].

The goal of big data management is to ensure a high level of data quality and accessibility for business intelligence and big data analytics applications. Effective big data management helps companies locate valuable information in large sets of unstructured data and semi-structured data from a variety of sources, including call detail records, system logs and social media sites [8].

Today big data management problem immediate solving by hadoop, the open source platforms for parallel processing of huge dataset across cluster of commodity servers. In this way you can quickly build big data management loading data in to hadoop distributed file system (HDFS), Hive, HBase, pig based transformations on data stored in hadoop [9].Micro-blogs and Twitter data follow the data stream model. Twitter data arrive at high speed, and algorithms that process them must do so under very strict constraints of space and time. The main Twitter data stream that provides all messages from every user in real time is called Fire hose and was made available to developers in 2010 [29].

Google's technical response to the challenges of web-scale data management and analysis was simple, by database standard, but modern big data to handle the challenges of web- scale storage the Google File System (GFS) is created [10]. To handle the challenge of processing the data in such a large files, Google discover its map Reduce programming model and platform. Map Reduce model basically works on two functions map and reduce [27].

4. TECHNOLOGY AND TECHNIQUES

The purpose of processing a large amount of data, big data requires additional technologies. There are many techniques and technologies have been introduced for manipulating, aggregate visualizing and analysis of big data [11]. These techniques and technologies draw from several fields including statistics, computer science, applied mathematics, and economics. Big data requires exceptional technologies to efficiently process large quantities of data within tolerable elapsed times [34].

There are many solutions for big data handling, but Hadoop is one of the most important technologies for handling big data.

4.1 Hadoop:

Hadoop is a distributed file system and data processing engine that is designed to handle extremely high volumes of data in any structure. Hadoop is a programming framework to support the processing of large data set in distributed computing environment. Hadoop is open source software hosted by Apache Software Foundation (ASF).It consists of many small sub project which is belongs to distributed computing [27] [28].

Hadoop created to inspire by BigTable which is Google's data storage system, Google File System and MapReduce [39]. Hadoop is Java based framework and heterogeneous open source platform. It is not a replacement for database, warehouse or ETL (Extract, Transform, Load) strategy. Hadoop includes a distributed file system, analytics and data storage platforms and a layer that manages parallel computation, workflow and configuration administration [40]. It is not designed for real time complex event processing like streams. HDFS (Hadoop Distributed File System) runs across the nodes in a Hadoop cluster and connects together the file systems on many input and output data nodes to make them into one big file system

Hadoop has two important components -

- The Hadoop distributed file system (HDFS), which supports data in structured relational form, in unstructured form, and in any form in between.
- The Map Reduce programming paradigm for managing applications on multiple distributed servers.

The processing pillar in the Hadoop system is the Map Reduce framework. Map Reduce is a heart of Hadoop system. Map Reduce is programming framework for distributed computing which is developed by Google, in which breaking the large complex data in to small data used divide and conquer method. Map Reduce have two function map and reduce [21] [23]. And one another new function is Global Reduce We use the term "Global Reduce" to distinguish it from the "local" Reduce which happens right after Map execution, but conceptually as well as syntactically, a Global Reduce is just another conventional Reduce function [24].

Map: It distributes out work to different nodes in the distributed clusters.

Reduce: It collects the work and resolves the results in to a single value [16]. Our use of a functional model with user specified map and reduce operations allows us to parallelize large computations easily and to use re-execution as the primary mechanism for fault tolerance [25].Using Map Reduce techniques many advantages in big data, Map Reduce is Simple and easy to use, Independent of the storage, Flexible, High scalable[20]. Typical, implementation of the mapReduce paradigm requires networked attached storage and parallel processing. Hadoop and HDFS by apache are widely used for storing and managing big data [22].

4.2 HDFS (Hadoop Distributed File System):

Hadoop are used a fault-tolerant storage system is called a Hadoop Distributed File System (HDFS).HDFS is a client-server based architecture which comprise of Name Node and many Data Nodes. The name node stores the Meta data for the Name Node. Name Node basically used for keeps track of the Data Node. And another responsibility for Name Node is for file system operation. If any time Name Node is fail then hadoop doesn't support automatic recovery, but the configuration of another node is possible [12] [26] [28]. We have developed a file system connector module for SF-CFS (Cluster File System) to make it work inside Apache Hadoop platform as the backend file system replacing HDFS altogether and also have taken advantage of SFCFS's potential by implementing the native interfaces from this module. Our shared disk Big Data analytics solution doesn't need any change in the Map Reduce applications. Just by setting a few parameters in the configuration of Apache Hadoop, the whole Big Data analytics platform can be made up and running very quickly [30].In fig 2 shows the how to split a large data input into different nodes(small size data).

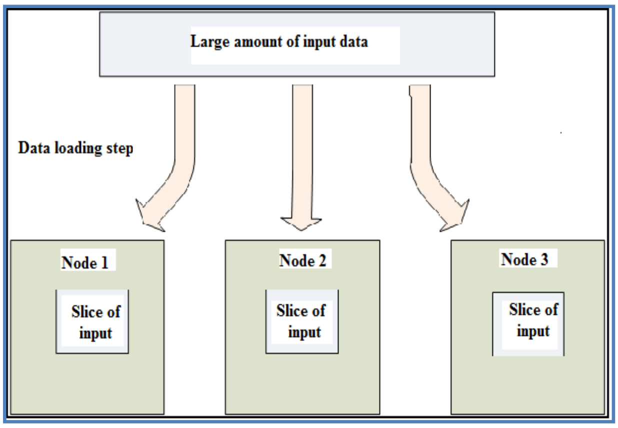

Fig 2: Nodes Distribution of Large data

HDFS stores terabytes of unstructured data in clustered environment and retrieves very fast whenever asked to present the data .Since HDFS uses clustered environment with many processing and data backup nodes, hence it increases the query performance to return the data very fast and also recovery of data from backup nodes in case of fallout of primary storage nodes [36]. Since past few years growth of search engines, social and electronic media, e-publishing and content management houses has produced huge amount of unstructured data which become a challenge to store and manage them in traditional SQL databases. As we know now the need is the mother of invention. Now we need such a data management system which can store and manage thousands of petabytes of unstructured data efficiently and the invention is HDFS. HDFS is one of the core components of Apache Hadoop. Several vendors uses Hadoop framework, enhance it, and make it available as free or commercial software [37]. HDFS is highly available and scalable due to flexible architecture. It implements master/slave architecture. Every cluster of HDFS comprises of NameNode and DataNode. NameNode manages the namespace of file systemas it serves as the Master Node and keeps all meta data information about the physical file. NameNode takes responsibility to serve the HDFS client for all operations been performed on HDFS [36]-[38]. All client requests first come to NameNode where Map and Reduces processes are running. DataNodes are storage node. All data in Hadoop is stored on data nodes. There number of nodes in a HDFS cluster with single NameNode as primary or master node. When any user attempts to store a large file in Hadoop ,then that file is breaks into various blocks and are stored in a set of DataNodes.

Hadoop consist of many other components-

- ***Apache Avro*:** designed for communication between Hadoop nodes through data serialization. These services can be used together as well as independently.
- ***Cassandra and Hbase*:** a non-relational database designed for use with Hadoop .It is run on the top of HDFS it serve as input and output for the mapreduce.
- ***Hive*:** a query language similar to SQL (HiveQL) but compatible with Hadoop.It is data warehousing application that provides SQL interface and relational model.
- ***Mahout*:** an AI tool designed for machine learning; that is, to assist with filtering data for analysis and exploration.
- ***Pig Latin*:** A data-flow language and execution framework for parallel computation.
- ***ZooKeeper*:** Keeps all the parts coordinated and working together. It is provide centralized service for distributed synchronization.
- ***Sqoop*:** Sqoop is a command-line interface platform that is used for transferring data between relational database and Hadoop.
- ***Oozie*:** Oozie is a java based web-application that runs in java servlet. Its manage the Hadoop job.
- ***Flume*:** It is high level architecture which focused on streaming of data from multiple sources [11] [17].

5. CONCLUSION

Big Data is becoming the new Final Frontier for scientific data research and for business applications. Big data is making a new path for the knowledge extraction from the large amount of data. There are various challenges and issues regarding big data. In order to explore Big Data, it is to be analyzed several challenges at the data, model, and system levels. There must be support and encourage in fundamental research towards these issues if we want to achieve the benefits of big data. The critical issue of privacy and security of the big data is the big issue will be discussed more in future. For most organizations, big data analysis is a challenge. Sheer volume of data and the different formats of the data (both structured and unstructured data) that is collected across the entire organization and the many different ways different types of data can be combined, contrasted and analyzed to find patterns and other useful business information.

6. REFERENCES

[1] Gartner IT glossary Big data. http://www.gartner.com/it-glossary/big-data Accessed 25 June 2014.

[2] Zikipoulos, P., Deutsch, T., Deroos, D., Parasuraman, K., Corrigan, D. and Giles J. Y. 2012. Harness the Power of Big Data.

[3] Gartner, Big Data Definition, http://www.gartner.com/it-glossary/big-data/

[4] "Data,data every where"The Economist. 25 February 2010. Retrieved9 December 2012

[5] Hilbert, M. and Lopez, P. 2011. "The World's Technological Capacity to Store, Communicate, and Compute Information"

[6] http://www.slideshare.net/HarshMishra3/harsh-big-data-seminar-report

[7] http://hive.apache.org/.

[8] http://searchdatamanagement.techtarget.com/definition/big-data management.

[9] https://www.talend.com/resource/big-data-management.html

[10] Ghemawat, S., Gobioff, H. and Leung, S. T. 2003. The google file system. In Proc. 19th ACM Symp. On Operating Systems Principles, SOSP '03, New York, NY, USA, ACM.

[11] Addressing big data problem using Hadoop and Map Reduce

[12] Garlasu, D., Sandulescu, V., Halcu, I. and Neculoi G. 2013. "A Big Data implementation based on Grid Computing", Grid Computing.

[13] Pitre, R., and Kolekar, V. 2014. Describe the big data is a new term used to identify the data sets due to their large size and complexity.

[14] Zou, H., Yu, Y., Tang, W. and Chen, H. W. M. 2014. Carried out a flexible data analytics framework for big data application with I/O performance improvement.

[15] Hudda, M. M. G. and Ramannavar, M. M. 2013. Propose an increase in use of social media forums, email, document and sensor data etc.

[16] Kaur, M., and Shilpa, 2013. Describe the big data and methodology.

[17] Sagiroglu, S. and Sinanc, D. 2013. Introduce the review of big data.

[18] Fahad, A., Alshatri, N., Tari, Z., Alamri, A., Khalil, I., Zomaya, A. Y., Foufou, S. and Bouras, A. 2014. Describe the clustering algorithm for big data.

[19] Armour, F., Kaisler, S. and Espinosa, J. A., Money W. 2013. Illustrated the issues and challenges in big data.

[20] Lee, K. H., Choi, T. W., Ganguly, A., Wolinsky, D. I., Boykin, P. O. and Figueired, R. 2011. Presents the parallel data processing with map Reduce.

[21] Grolinger, K.,Michael, Hayes, Higashino, W., Heureux ,A. L., Allison, D. S. and Capretz ,M. A. M. 2014. Discuss the challenges for mapreduce in big data.

[22] Dr. Siddaraju, Sowmya, C. L., Rashmi, K. and Rahul M. 2014. Carried out the analysis of big data using mapreduce framework.

[23] Ekanayak, J., Pallickara, S. and Fox, G. 2008. Proposes the MapReduce for data intensive for scientific analysis.

[24] Luo, Y. and Plale, B. 2012. Carried out the hierarchical mapreduce programming model and scheduling algorithms.

[25] Dean, J. and Ghemawat, S. 2004. Illustrated the Simplifed Data Processing on Large Clusters.

[26] Natarajan, S. and Sehar, S. 2013. Presents the algorithm for distributed data mining in HDFS.

[27] Manikandan, S. G. and Ravi, S. 2014. Discuss Big Data Analysis using Apache Hadoop.

[28] Kala Karun, A. and Chitharanjan, K. 2013. Discuss A Review on Hadoop HDFS Infrastructure Extensions.

[29] Bifet, A. 2012. Present a mining big data in real time and discuss the data streams.

[30] Mukherjee, A., Datta, J., Jorapur, R., Singhvi, R., Haloi, S. and Akram, W. 2012. Shared disk big data analytics with Apache Hadoop.

[31] Marz, N. and Warren, J. 2013. Big Data: Principles and best practices of scalable realtime data systems. Manning Publications.

[32] Feldman, D., Schmidt, M. and Sohler, C. 2013. Turning big data into tiny data: Constant-size coresets for k-means, pca and projective clustering. In SODA.

[33] Fan, W. and Bifet, A., Discribe the big data mining current status and forecast to the future.

[34] Patel, A. B., Birla,M. and Nair, U. 2012. Presents Addressing Big Data Problem Using Hadoop and Map Reduce.

[35] Redid, K. K. and Indira, D. 2013. Introduce the knowledge that big data is combination of structured, semi-structured, unstructured homogeneous and heterogeneous data.

[36] Ousterhout, J.and Agrawal P. 2011 “The case for RAMCloud”,Communications of the ACM, Vol. 54(7).

[37] Petrovic,J. 2008 "Using Memory cached for Data Distribution in IndustrialEnvironment", In Third International Conference on Systems(ICONS08),Cancun,

[38] Ousterhout, J.and Agrawal P. 2010 “The case for RAMCloud” ScalableHigh-Performance Storage Entirely in DRAM”, ACM SIGOPSOperating Systems Review, Vol.43 (4).

[39] Begoli,E. and Horey,J. 2012 "Design Principles for Effective Knowledge Discovery from Big Data", Software Architecture (WICSA) and European Conference on Software Architecture (ECSA) Joint Working IEEE/IFIP Conference on, Helsinki.

[40] Intel IT Center, "Planning Guide: Getting Started with Hadoop", Steps IT Managers Can Take to Move Forward with Big Data Analytics. http://www.intel.com/content/dam/www/public/us/en/documents/guides/getting-started-with-hadoop-planning-guide.pdf