Study report on Issues and Challenges on Big Data

Kakkerla Shivakumar
Assistant Professor, Department of CSE,
Geethanjali College of Engineering & Technology, Hyderabad, Telangana, India.

***Abstract*—**In the last few years, the growth in hardware technology has made it for many companies to store huge amount of data. The data streams are infinite and it can be found in many applications. Data will arrive continuously as a stream in a dynamic fashion, processing and querying of these streams are challenging tasks. Big data is a set of techniques and technologies that require new forms of integration to uncover large hidden values from large datasets that are diverse, complex, and of a massive scale. Big data environment is used to acquire, organize and analyze the various types of data. There is a requirement of efficient datamining techniques which can handle the streaming data. In this paper we explore various issues and challenges on Big Data.

***Keywords*—** *Data Mining, Data Stream, Big Data, Hadoop*

I. INTRODUCTION

Data mining [1] is a process of extracting knowledge from large databases. With the very high voluminous structured or unstructured continuous data being generated from various applications and devices, the concept of data is no more static but is turning out to be dynamic. This brings a lot of challenges in analyzing the data.

The data streams are infinite and it can be found in many applications such as telecommunication systems, where data will arrive continuously as a stream in a dynamic fashion, processing and querying of these streams are challenging tasks. Satellite systems, online transactions, video surveillance, sensor networks and web applications data is not simply load the arriving into traditional databases and perform operations on it.

Traditional data mining algorithms are not suitable for handling data streams because the algorithms designed perform multiple scans over the data which is not possible when handling the data streams [2]. This brings actual challenge before the data mining researchers working in the area of data streams. Further, many of the existing data mining algorithms available for clustering [13], classification and finding frequent pattern are suitable for only static data sets and are no more practically suitable for handling data streams or for mining the stream data.

Data from data streams will arrive continuously as chunks of data with high rate. The data sets are endlessly grown over the years. The data streams [16] are infinite and it can be found in many applications such as telecommunication systems, where data will arrive continuously as a stream in a dynamic fashion, processing and querying of these streams are challenging task.

The present century is the century of data. We are collecting and processing data of all kinds on scales unimaginable earlier. Examples of such data are internet traffic, financial tick-by-tick data and DNA Microarrays which feed data in large streams into scientific and business bases worldwide.

Big Data is the word used to describe massive volumes of structured and unstructured data that are so large that it is very difficult to process this data using traditional databases and software technologies. Big Data is continuously generated on a massive scale. It is generated by online interactions among people, by transactions between people and systems and by sensor enabled instrumentation. The problems of such data include collection, storage, search, sharing, transfer, visualization and analysis. An important advantage of analysis of Big Data is the additional information that can be obtained from a single large set as opposed to separate smaller sets. Big Data allows correlations to be found, for instance, to spot business trends.

Big data is particularly a problem in business analytics because standard tools and procedures are not designed to search and analyze massive datasets.

The rest of this paper is organized as follows. In Section II,we present the definition of big data, its features, characteristics and data storage management. In Section III, we introduce Hadoop and its architecture. Then, in Section IV, we introduce the big data issues and challenges. A brief conclusion with recommendations for future studies is presented in section V.

II. BIG DATA

A. What is Big Data

Big data is the term for a collection of data sets which are large and complex, it contain structured and unstructured both type of data. Data comes from everywhere, for example sensors used to gather climate information, posts to social media sites, digital pictures and videos etc. This data is known as big data. Useful data can be extracted from this big data with the help of data mining [5].
There are two types of big data:

Structured data: Data can be numbers and words that can be easily categorized and analyzed. These data are generated by things like smart phones; sensors embedded n electronic devices and Global Positioning System (GPS) devices. Structured data also include account balances, and transaction data.

***Unstructured data*:** Data include more complex information, such as customer reviews from commercial websites, photos and comments and posts on social networking sites. These data cannot easily be separated into categories or analyzed numerically.

Due to explosive growth of the Internet in recent years means that the variety and amount of big data continue to grow. Much of that growth comes from unstructured data.

***Example of Big Data*:** An example of big data might be petabytes (1,024 terabytes) or exabytes (1,024 petabytes) of data consisting of billions or trillions of records coming from different sources (e.g. Web, sales, customer contact center, social media, mobile data and so on). The data is typically loosely structured data that is often incomplete and inaccessible.

B. Features of Big Data

- The data from data streams arrive continuously.
- Nature of the data streams are dynamic
- The size of the data stream is unlimited.

- Frequent accessing of data from data streams is costly.

Because of the above features of data streams, mining the data streams poses a confrontation to many researchers extracting knowledge from streamed data [12].

C. *Charecteristics of Big Data*

The Big data [15] can be characterized by wellknown 4Vs as shown in following fig1.

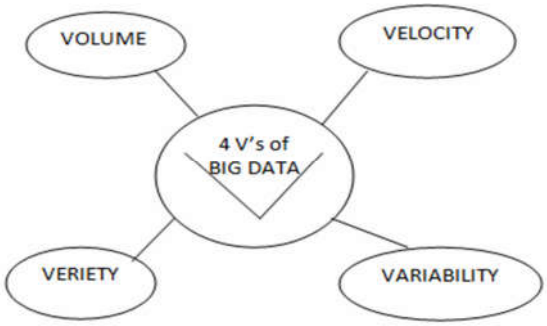

Fig1: Characteristics of Big Data

- ***Volume*:**Data volume measures the amount of data available to an organization, which does not necessarily have to own all of it as long as it can access it. As data volume increases, the value of different data records will decrease in proportion to age, type, richness, and quantity among other factors.
- ***Velocity***: Data velocity measures the speed of data creation, streaming, and aggregation. eCommerce has rapidly increased the speed and richness of data used for different business transactions. Data velocity management is much more than a bandwidth issue.
- ***Variety*:** Data today comes in all types of formats. Data variety is a measure of the richness of the data representation – text, images video, audio, etc. From an analytic perspective, it is probably the biggest obstacle to effectively using large volumes of data. Incompatible data formats, non-aligned data structures, and inconsistent data semantics represents significant challenges that can lead to analytic sprawl.
- ***Variability*:** The inconsistency of data can show at times- which can hamper the process of handling and managing the data effectively.

Another characteristic of big data is:

- Complexity:Today's data comes from multiple sources. And it is still an undertaking to link, match, clean and transform data across systems. However, it is necessary to connect and correlate relationships, hierarchies and multiple data linkages or your data can quickly spiral out of control.

D. *Big Data Storage*

The rapid growth of Data highly requires efficient storage management system and also requires some efficient data storage device for managing a large amount of data. The functionality of big Data Storage includes the management of large scale and unstructured data sets with the ease of reliability and availability of the data accessing.

The big data storage paradigm is designed for providing data storage service with reliability and authenticity. On the other hand, the storage amount should also provide the flexibility for data accessing, query and analysis of large amount of data [7].

As an auxiliary equipment of server data storage devices are used for managing and analyzing the unstructured datasets with structured RDBMS. But many companies own their big cloud storage capacity but which obtains cost effectiveness and it is competitive in nature thus there is a compelling need of research direction towards data storage [8].

III. HADOOP

When making an attempt to know about big data, the word such as Hadoop cannot be avoided. Hadoop is open-source software that enables reliable, scalable, distributed architecture for data storage and processing [6].

Hadoop doesn't do data mining. Hadoop manages data storage (via HDFS, a very primitive kind of distributed database) and it schedules computation tasks, allowing you to run the computation on the same machines that store the data. It does not do any complex analysis.

The Hadoop Distributed File System (HDFS) is designed and optimized to store data over a large amount of low-cost hardware in a distributed fashion. Table 1 shows the difference between RDBMS [18] and Hadoop.

TABLE1: Difference between RDBMS and Hadoop

RDBMS	Hadoop
Traditional row column databases used for transactional systems, reporting and archiving.	Hadoop manages data storage via HDFS, which is designed and optimized to store large amount of data in distributed fashion.
It supports structured type of data only	It supports structured, semi structured and unstructured type of data
Maximum size of data in Terabytes	Maximum size of data exceeds hundreds of pita bytes
Queries processed on finite data sets from disk.	Queries processed on continuous and time varying data streams
It can process 1000 queries/second	It can process millions of queries per second
It can relatively works with low data rate	It can extremely works with high data rate.

Apache Hadoop [14] is an open-source software framework that supports massive data storage and processing. Instead of relying on expensive, proprietary hardware to store and process data, Hadoop enables distributed processing of large amounts of data on large clusters of commodity servers.

Hadoop has many advantages, and the following features make Hadoop particularly suitable for big data management [17] and analysis:

- *Scalability*: Hadoop allows hardware infrastructure to be scaled up and down with no need to change data formats. The system will automatically redistribute data and computation jobs to accommodate hardware changes.

- *Cost Efficiency*: Hadoop brings massively parallel computation to commodity servers, leading to a size able decrease in cost per terabyte of storage, which makes massively parallel computation affordable for the ever growing volume of big data.

- *Flexibility*: Hadoop is free of schema and able to absorb any type of data from any number of sources. Moreover, different types of data from multiple sources can be aggregated in Hadoop for further analysis. Thus, many challenges of big data can be addressed and solved.

- *Fault tolerance*: Missing data and computation failures are common in big data analytics. Hadoop can recover the data and computation failures caused by node breakdown or network congestion.

Hadoop Architecture:

The Apache Hadoop architecture [17] consisting of several modules, including HDFS, MapReduce, HBase and Chukwa. These modules fulfill the functions of a big data. The layered architecture of the core library is shown in Fig 2. We will introduce different modules from the bottom-up in examining the structure of the big data.

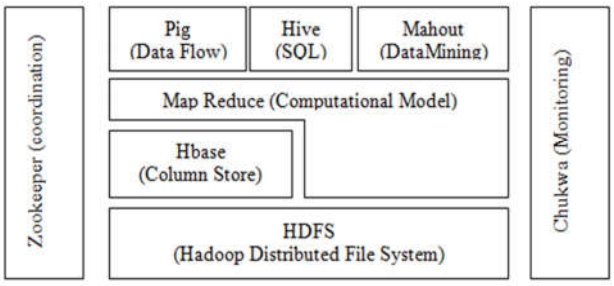

Fig 2: Architecture of Hadoop

***HDFS*:**

HDFS is a distributed file system developed to run on commodity hardware that references the GFS design. HDFS is the primary data storage of Hadoop applications. An HDFS cluster consists of a single Name Node that manages the file system metadata, and collections of Data Nodes that store the actual data. A file is split into one or more blocks, and these blocks are stored in a set of Data Nodes. Each block has several replications distributed in different Data Nodes to prevent missing data.

***HBase*:**

Apache HBase is a column-oriented database management system on top of HDFS. It is well suited for sparse data sets, which are common in many big data use cases. Unlike relational database

systems, HBase does not support a structured query language like SQL; Because HBase is not a relational data store at all. HBase can serve both as the input and the output for MapReduce jobs run in Hadoop and may be accessed through Java API, REST, Avor and Thrift APIs.

***MapReduce*:**

Hadoop MapReduce [10] is the computation core for massive data analysis and is also modeled after Google's MapReduce. The MapReduce framework consists of a single master JobTracker and one slave TaskTracker per cluster node. The master is responsible for scheduling jobs for the slaves, monitoring them and re-executing the failed tasks. The slaves execute the tasks as directed by the master. The MapReduce framework and HDFS run on the same set of nodes, which allows tasks to be scheduled on the nodes in which data are already present.

***Pig & Hive*:**

Pig and Hive are two SQL-like high-level declarative languages that express large data set analysis tasks in MapReduce programs. Pig is suitable for data flow tasks and can produce sequences of MapReduce programs, whereas Hive facilitates easy data summarization and ad hoc queries.

***Mahout*:**

Mahout is a data mining library implemented on top of Hadoop that uses the MapReduce paradigm. Mahout contains many core algorithms for clustering, classification, and batch-based collaborative filtering.

***Zookeeper & Chukwa*:**

Zookeeper and Chukwa are used to manage and monitor distributed applications that run on Hadoop. Specifically, Zookeeper is a centralized service for maintaining configuration, naming, providing distributed synchronization, and providing group services. Chukwa is responsible for monitoring the system status and can display, monitor, and analyze the data collected.

Table 2 presents a quick summary of the function classification of Hadoop modules. Under this classification, HDFS and HBase are responsible for data storage, MapReduce, Pig, Hive and Mahout perform data processing and query functions, and Zookeeper and Chukwa coordinate different modules being run in the big data platform.

TABLE 2: Hadoop Module Summarization

Function	Module	Description
Data Storage	HDFS HBase	Distributed file system Column based data store
Computation	Map Reduce	Group aggregation computation framework
Query & Analysis	Pig Hive Mahout	SQL like language for data flow tasks SQL like language for data query Data Mining library
Management	Zookeeper Chukwa	Service Configuration, Synchronization, etc. System Monitoring

IV. RESEARCH ISSUES AND CHALLENGES IN BIG DATA

Issues of Big data are Storage issues:

Current disk technology limits are about 4 terabytes per disk. So, 1 exabyte would require 25,000 disks. Even if an exabyte of data could be processed on a single computer system, it would be unable to directly attach the requisite number of disks. Access to that data would overload current communication networks. Thus, transferring an exabyte would take longer time to transmit the data from a storage point to a processing point [3].

Processing Issues:

Effective processing of exabytes of data will require new analytics algorithms in order to provide timely and actionable information.

Management issues:

Data management will be the most difficult problem to address with big data. The sources of data are varied - both temporally and spatially, by format, and by method of collection. Individuals contribute digital data in mediums comfortable to them: documents, drawings, pictures, sound and video recordings, models,etc – with or without adequate metadata describing what, when, where, who, why and how it was collected and its provenance.

There are lots of open research issues in the field of data stream mining[4]. Some of the issues are given in the following.

- Preprocessing of data from data streams.
- Reliability of incoming data.
- Satisfying the user requirements
- Scalability of data stream mining systems.
- Accuracy of results while dealing with continuous flow of data.

Research Challenges [9] in Handling Big Data

- A major challenge for a very large heterogeneous data set is to figure out what data one has and how to analyze it.
- An emerging challenge for big data users is accessing of more data (quantity), they often want even more (quality).
- Changing of asymmetrical arrival of data rate.
- Providing quality of results using data mining methods.
- Small amount of memory and huge volume of data streams.
- Resources are limited for storage and computing.
- Processing big data is a major challenge than the storage or management problem.

V. CONCLUSION

Now a days extracting knowledge from Big Data is a high touch business. Because of increase in the amount of data in the fields of telecommunication, media, education, meteorology, biology, environmental research, it becomes difficult to handle the large data sets.

Technologies today not only support the collection of large amounts of data but also help in utilizing such data effectively. Due to rapid growth of internet, everyday people are sending posts, comments, videos, etc. As a result, big data simply requires a new way of thinking about how to store and analyze data to accommodate these new realities and turn insights into actionable decisions.

In this paper we concentrated and presented survey on Big data issues and challenges in concise manner.The world is now ready and able to work in 'n' dimensions. Finally, there is a need to have better data models which would handle high dimensional data.

REFERENCES

[1] J. Han and M. Kamber, Data Mining: Concepts and Techniques, J. Kacprzyk and L. C. Jain, Eds. Morgan Kaufmann, 2006, vol. 54, no. Second Edition

[2] M.S.B PhridviRaj, C.V. GuruRao. Data Mining-past,present and future- a typical survey on data steams. Proceedings of 7thInternational Conference Interdisciplinarity in Engineering(INTER-ENG 2013), published by Elsevier, 2014: 255-63.

[3] Stephen Kaisler et.al. Big Data: Issues and Challenges Moving Forward, Proceeedings of 46th Hawaii International Conference on System Sciences, 2013.

[4] Vinay Kumar K, Srinivasan R, Elijah Blessing Raj Sigh. Data Stram Model-Issues, Challenges and Clustering Techniques, International Journal of Recent Development in Engineering and Technology, Volume 1, Special Issue 1, Oct 2013,pages 18-24.

[5] Bharti Thakur et al. Data Mining for Big Data: A Review. International Journal of Advanced Research in Computer Science and Software Engineering, Volume 4, Issue 5, 2014, pp. 469-473.

[6] Vibhavari Chavan et.al.Survey on Big Data, International Journal of Computer science and Information Technologies, Volume 5,Issue 6, 2014, 7932-7939.

[7] Rajeswari.D. State of the Art of Big Data Analytics: A Survey,International Journal of ComputerApplications(0975-8887),Volume 120-No.22, 2015, 39-46.

[8] Cevher, V., Becker, S., Schmidt, M.2014.Convex Optimization for Big Data: Scalable, randomized, and parallel algorithms for big data analytics. Signal Processing Magazine, IEEE, Vol.31, No.5, pp.32-43

[9] Min Chen.Shiwen Mao.Yunhao Liu. Big Data: A Survey. Published by Springer Science + Business Media New York, 2014: 171-209.

[10] Dilpreet Singh and Chandan K Reddy. A survey on platforms for big data analytics, Journal of Big Data, a Springer Open Journal, 2014.

[11] Puneet Singh Duggal and Sanchita Paul. Big Data Analysis: Challenges and Solutions, International Conference on Cloud, Big Data and Trust 2013, Nov 13-15, RGPV.

[12] Fabian M. Suchanek et al. Knowledge Bases in the Age of Big Data Analytics. Proceedings of 40th International Conference on Very Large Data Bases, Volume 7, No. 13, 2014.

[13] S.Guha, N.Mishra, R.Motwani, and L.O'challaghan clustering data streams.In Proceedings of the Annual Symposium on Foundations of Computer Science.IEEE,November 2000.

[14] Wang, F. et al. Hadoop High Availability through Metadata Replication. ACM(2009)

[15] Venkat Narasimha Inkollu et al. Security Issues Associated with Big data In Cloud Computing. International Journal of Network security and its Applications(IJNSA), Volume 6, No.3, 2014.

[16] Yu.Bao.Liu et.al. Clustering Text data streams, Journal of computer science and technology, volume 23, issue 1, pages 112-128, 2008.

[17] Hu, H., Wen, Y., Chua, T-S., Li, "Toward Scalable Systems for Big Data Analytics: A Technology Tutorial," Access, IEEE , Vol.2, No., pp.652-687, 2014.

[18] Bansal et. Al. “Transitioning from Relational Databases to Big Data”,International Journal of Advanced Research in Computer Science and Software Engineering, Volume 4, pp. 626-630, 2014

BIOGRAPHY

KAKKERLA SHIVAKUMAR has completed B.Tech in MVSR College of Engineering &technology, Hyderabad, Telangana, INDIA. And M.Tech in NOVA College of Engineering & technology, Hyderabad, Telangana, INDIA. Present working in Geethanjali College of Engineering & technology, Hyderabad, Telangana, INDIA as a assistant Professor and total teaching experience is 2.5 years and his interest area are Cloud Computing, Data Mining and bigdata

.