Seema Rawat *et al*, International Journal of Computer Science and Mobile Applications,
Vol.4 Issue. 11, November- 2016, pg. 13-19 **ISSN: 2321-8363**
Impact Factor: 5.515

An Analytical Study on Research Challenges and Issues in Big Data Analysis

Seema Rawat
Research Scholar, Swami Vivekanand University, Sagar, M.P
seemarawatcse@gmail.com

Shankar Ramamoorthy
Professor, Swami Vivekanand University Sagar, M.P
shankar.Imc@yahoo.com

Abstract:- The Big Data is the new technology in the field of research in recent years and is not only big in Amount , but also produced at speed and variety, which endeavors the research upsurge in multidisciplinary fields like Decision making , Healthcare industry and business analysis. Due to the basic features (Volume, speed, Velocity, and Variety) of Data when it is difficult to store and analyse with existing tools and techniques. It explains distinctive tasks in scalability, storage, computational complexity, analytical, security issues. Hence here the salient features of big data and how this shakes the storage tools and processing technique. It shows the taxonomy of Big Data application areas and explanation of data characteristics, privacy and security issues. Furthermore, It also discover research issues and challenges in big data storage, privacy and security of data and data processing.
Keywords: Big data, Analytics, Data Science, Hadoop, map reduce.

1. INTRODUCTION:

Big data is the technology that maintains the huge amount of data. Millions of data is been generated in a single day through various technology with high requirement of proper backup. As most of the data is in continuous form which is used for different-different analysis and other purposes like business trends, legal citation linking, etc,. According to the scenario [5] now the user's requirement are touching the peak. The data storage which previously takes its limits under Gigabyte or Terabyte that now needs more then Zettabyte or Yottabyte.

Fig.1 Big Data

Big Data was developed and came to use as the time of spread sheets are over. It is well-known that the big data is the most popular term used to express exponential growth of data and to collect, store, manage, process, analyze, utilize, or to visualize the data in the data in such a way that it produce a effective results.

The following are the characteristics of big data which are also known as 4V's of big data:

Volume: Variety of data is been collected by organizations from various sources such as social media, business world of transaction, and information from machine to machine data or sensors. Looking in the past storing the data was one of the big issue which was resolved via. Hadoop.

Velocity: Streaming of data is in miraculous speed with proper timely manner. It mainly consists the rate at which data is getting produced. In this, motion data also been store continuously within respond time in some milliseconds such as live streaming of any matches, news, etc. Sensors,[2] smart metering, and RFID tags proves better in demand to accord with cascade of data s specially in near-real- time scenario.

Variety: Different forms of data such as structured data which is properly managed like MySql, semi-Structural data in which some of the data is maintained well and some are not like json & xml, and unstructured data which is not properly maintained like text, audios, videos, etc. are available in this huge database named big data.

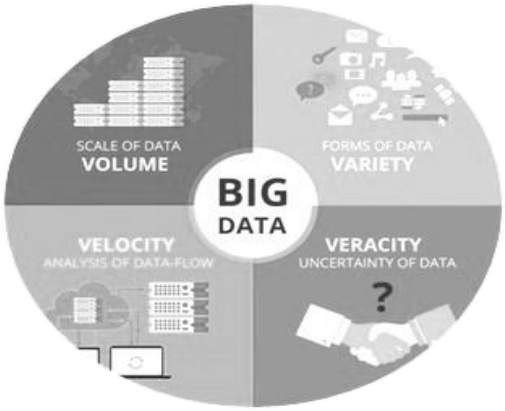

Fig. 2 Four V's of Big Data

Veracity: The expansion of data velocity and variety causes data to be highly inconsistent, ambiguous, incomplete, deception, latency, approximate, or uncertain. Due to which management of data become more challenging and even more in case of unstructured data.

Looking onto the difference between relational database management system and Big Data. The relational database management system i.e. RDBMS consists structural data which is in form of rows and columns. for example excel files, sql data, xml files, oracle data and many more. Whereas, in big data the data could be in any form[6] i.e. structural, semi structural or unstructured data such as images, logs, text, files, videos, etc. using format of hdfs (hadoop distributed file system). It include some software help such as if the data lies in Gigabytes than SQL/Oracle/RDBMS could come in help but if the data is in Terrabyte than Teradata[9] is been used and if it goes more than Petabyte then hadoop technology is much more preferable and to process business logic it seeks help of MapReduce.

Hadoop:

Hadoop is build up by apache. It is widely used as an open source framework in today's era and a data management tool which provide distributed processing of huge datasets on the cluster of merchandise hardware with the best advantage of scale out of storage. For creating a Hadoop cluster one has to know how much data it has to analyze in coming 6 or 8 months. It is nothing but just a group of system in which Hadoop is installed. It is freely available.

Hadoop Architecture- In Hadoop architecture, HDFS is used for storage/reads-writes and MapReduce or YARN is used for processing.

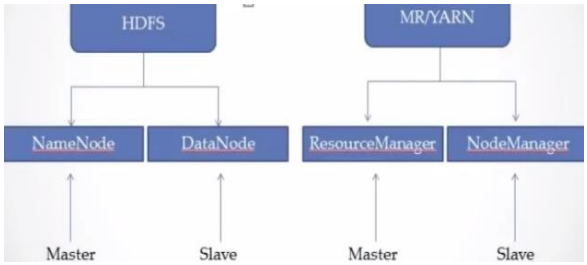

Fig.3 Hadoop Architecture

Storage part is handled by NameNode and DataNode via HDFS. Where NameNode is master and DataNode is slave. Similarly, MP/YARN is handled by ResourceManager and NodeManager. Where ResourceManager is master and NodeManager is slave. The system runs on master system that are known as master daemons and the system runs on slave system that are known as slave daemons. DataNode and ResourceManager will run on master daemon and DataNode and NodeManager will run on slave daemons.

Secondary NameNode is run on master daemons. Which will only take hourly backup and stores it. It will never take place in case NameNode is crashed. But, it can restart the crashed Hadoop[8] cluster. It is very important daemon for Hadoop1 however, In Hadoop 2 it is not much important.

Hadoop 2: YARN

HDFS federation brings important measures of scalability and reliability to Hadoop. YARN, the other major advance in Hadoop 2, brings significant performance improvements for some applications, supports additional processing models, and implements a more flexible execution engine.

YARN is a resource manager that was created by separating the processing engine and resource management capabilities of MapReduce as it was implemented in Hadoop 1. YARN is often called the operating system of Hadoop because it is responsible for managing and monitoring workloads[10], maintaining a multi-tenant environment, implementing security controls, and managing high availability features of Hadoop.

Like an operating system on a server, YARN is designed to allow multiple, diverse user applications to run on a multi-tenant platform. In Hadoop 1, users had the option of writing MapReduce programs in Java, in Python, Ruby or other scripting languages using streaming, or using Pig, a data transformation language. Regardless of which method was used, all fundamentally relied on the MapReduce processing model to run.

YARN supports multiple processing models in addition to MapReduce. One of the most significant benefits of this is that we are no longer limited to working the often I/O intensive,[11] high latency MapReduce framework. This advance means Hadoop users should be familiar with the pros and cons of the new processing models and understand when to apply them to particular use cases.

Hadoop Tools

The most salient advantages of Hadoop 2 over Hadoop 1 is the improved reliability and speed improvements that come with the improvements to HDFS federation and the introduction of YARN, which separates processing management from resource management. HDFS federation reduces the chance of cluster-wide disruptions to the failure of a single Namenode. Different types of jobs can now run on a Hadoop cluster at the same time. Developers are no longer limited to writing multi-pass MapReduce programs when a better option can be modeled using a directed acyclic graph approach.

It should be noted that although YARN is a mature top-level Apache project, other tools in the Hadoop ecosystem are still in incubator [7] status. These tools can be used but they may require more effort to configure, manage and maintain than top-level projects. You are also less likely to find large volumes of support messages and threads on

community sites such as Stackoverflow.com. For example, a search for Hadoop Hive returns over 3,800 hits in StackOverflow while a search for Hadoop Spark returns about 100. Fortunately, for enterprises moving to Hadoop 2 there are commercial support options from Hortonworks, MapR and Cloudera.

Early adopters of Hadoop were limited to MapReduce-based processing models. Hadoop 2 has introduced a new processing model that lends itself to common big data use cases including interactive SQL over big data, machine learning at scale, and the ability to analyze big data scale graphs. For developers who were waiting for more support for their use cases, now is the time for another close look at Hadoop. HDFS[12] (hadoop distributed file system) which is also used for storage and YARN/MRv2 (yet another resource negotiator/ MapReduce version 2) which is used for cluster management. Whereas, in this there is one master which is active an another is on standby so in case active Namenode crashed then standby master will take place. Whereas, in this there is one master which is active an another is on standby so in case active Namenode crashed then standby master will take place.

MapReduce:

MapReduce is a kind of model which is based on distributed computing and also a processing technique. This algorithm consists two important tasks, Map and Reduce. One can take its advantage in unravel surprising trends in real world data. It is programmed in core java to write business logic for better processing.

In MapReduce, the input and output both are key value pair. Mapper first need to[14] get data i.e. input file by default it get data in text form which can be configurable. For processing data one has to mention business logic in it and finally it write the output. Mapper's output is further forward to Reducer. Any file defined in Hadoop is divided in blocks.

After coding and removing the jar file and put that file on execution at the same time processing will start wherever the data is and mapper start running. The number of mapper runs is the same number of blocks is available. Reducer get data from Mapper and one has to write business logic after its process completion it take out the final summarized output. It's algorithm has two task to be done i.e. Summarization operation and by default one reducer. Once the data received by reducer then its work is to summarize[3] the data and by default one reducer means all the mappers will run but there will be only one reducers which is fully configurable on the basis of number of output required by programmer.

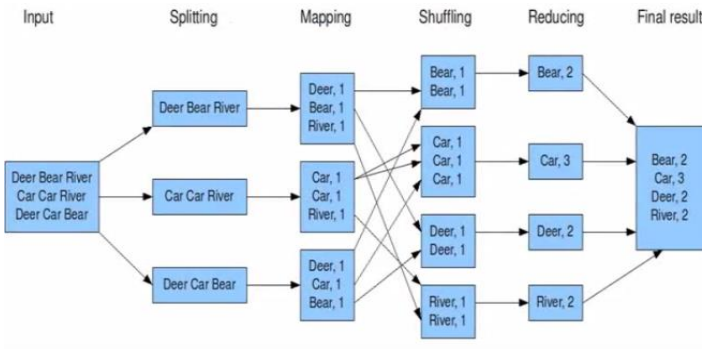

Fig. 4 Working criteria/example of MapReduce

Here, Mapper input is unorganized which is further divided into different blocks after execution of map function it gave the output. Then shuffling process starts it merge similar input afterwards reducer summarized it and provide the final output. Looking[16] onto the major benefit comparative with traditional way so, Traditionally big data was split into various machines[13] and each system need to code accordingly and then processing starts. Afterwards in the end all the results again need to aggregated for final result. And if any system is crashed then it forms the data loss. But in this, it make most of its work by itself such as dividing the file, making replication, converting it into file blocks. Secondly, program is[15] written at only once which makes process faster. If any system is crashed so, data can be achieved from another system. Here, competition is taken closer to data which make summarization faster and easy and it is done using clustering technique.

2. Research challenges in Big Data

The Big Data Processing leads to many open challenges that can be mainly categorized to three areas,
1) Data accessing and mining platform
2) Semantics and Application domain knowledge and 3) Big Data Analytics.

2.1 **Data accessing and processing:** Big Data deals with huge amount of data which is generated at high rate by variety, its big challenge is to where to process the data and how to access it. The data storage organizes the collected [15] data in a convenient format for value extraction and analysis. To accommodate data persistently along with the reliability, also it should provide a scalable interface to access, query and analyze. It's a challenge to Storage Infrastructure and processing tools.

i) Storage Infrastructure: The collected data is physically stored on storage devices like RAM, Magnetic Disks, Disk Arrays and Storage class Memory etc. These devices have their own performance metrics which leads to building high performance and scalable storage systems. Storage infrastructure is among the areas which provide various research challenges.

ii) Data storage Tools: Data storage tools play a major role in organizing data over Storage Infrastructure so that data can be accessed in convenient manner for processing. Even before the origin of Big Data such tools were actively researched. The basis of data storage is given by file system, which attracts the attention of academic and industry. The Google File System (GFS)[3] is the first file system for Big Data. The main disadvantage of GFS is it provides poor performance for small size file and suffers with single point failure The characteristics of Big Data demands parallel distributed programming models to provide analytical results at high rate. The famous programming models like Spark, Mapreduce, Dryad, Pregel, GraphLab, Storm, are widely used in Big Data analytics. Even though there are many file systems, Database technologies and Programming models available to support Big Data characteristics

2.2 Different Type of Application area

The most significant challenges and issues are
i) Data Privacy and security
ii) Knowledge of Application type.
i) Data Privacy and security: The main aim of digital world is to share the data and it has got such growth due to data distribution. The motivation for sharing data over multiple systems is clear, but true concern is emphasized on the sensitive data, involving banking transaction and medical records processed by Big Data analytics. For such applications simple data exchange polices cannot resolve privacy concern. Two common approaches to protect privacy are:

1) Restrict data access: The security can be added to data or access control can be executed on data so that sensitive data is accessed only. In this approach challenges are to address the designing of security approach, so that sensitive information can't be accessed by unauthorized users.

2) Data Field Anonymization: It is a process where individual record in sensitive data cannot be pinpointed. The main objective of anonymization is to inject randomness to data to ensure variety of privacy goals.

ii) Knowledge of Application Type: The vital condition for the Big Data analytics is knowledge of application domain, which plays a vital role to decide and design Big Data mining algorithms and framework and also facilitates veracious features for modeling the underlying data.

2.3 Big Data Analytics

"The process of identifying the hidden patterns and unknown correlations by using analytical algorithms running on powerful machines is termed as Big Data analytics". In batch processing store and analyse technique is used, where it stores the data first then analysis will be done. In batch processing[6] a chain of jobs are executed without human intervention. Batch processing is the most popular scenario which processes huge data in a single run. The popular open source tool MapReduce supports batch processing which is a part of Hadoop framework. Hadoop is scalable, fault tolerant, flexible and easy-to-code. MapReduce is distributed parallel programming model which works on Single Program Multiple Data concept.

3. Research challenges and issues in Big Data

Many researchers are focused on addressing the issues in Big Data storage, access, security and analysis. But there are several exiting issues and challenges where researchers can look into. The challenges are as follows.

• **Heterogeneity and Incompleteness**: Large amount of data is being generated from sources. Data is in the form of text, data logs, videos, images, audios, structured, semi structured, unstructured data from sources like sensors, airplanes, social networks, retail industry, mobiles etc.,. Also uncertainty is created by incomplete data during analysis which should be managed correctly.

•**Scale and complexity:** Managing rapidly increasing huge volume of data is a challenging issue. The conventional mechanisms are not suitable for managing analyzing and retrieving of Big Data, which is an open challenge due to its complexity and scalability.

• **Timeliness:** The time required to analyze the data will increase due to rapidly increasing its generation. Hence there are some situations where misuse of data need to be addressed.

•**Security and Privacy:** Huge amount of data is being generated, processed and analyzed[9]. During this process, the users and organizations are worried about data privacy and security related issued.

•**Fault tolerance:** Fault tolerance is most important issue that needs to be addressed in big data. When a process started by involving many network nodes in the entire computation process, becomes cumbersome.

• **Data analysis:** Deciding an appropriate technique to analyse a huge data is crucial phase of data analytics. To extract desired patterns from huge data there must be analytical algorithms to generate results. However the desired patterns purely depend on the timely requirements so the existing algorithms are not suitable for Big Data analysis.

•**Knowledge Discovery:** The prime issue of big data is Knowledge discovery and representation of data which includes types of sub domains: preserving, archiving, and retrieving of information, authentication and data management. There are many existing tools and techniques to address these issues. But most of these techniques are specific to some problems.

Conclusions:

There is a potential for making faster advancements in scientific discipline for analyzing the large amount of data .The technical challenges are most common across the large variety of application domains, therefore new cost effective and faster methods must be implemented to analyse the big data. The volume of data being generated is high and continues to increase timely. There is an expansion in data variety and the velocity of data being generated and is high due to automation, social media, smart phones, sensor connected devices and internet. This endeavors the research upsurge in multidisciplinary fields as well as Government, Healthcare and business performance applications. Hence we explore the salient features of Big Data and present and discuss the classification of datasets based on their behavior and respective available tools to address it. We also addressed the research challenges in the field of big data domain. The big data is the current trend and next era is ruled by big data and there is huge scope for open research challenges. It has been expect that, the research on big data storage, privacy of data and analytical techniques will continue to grow in forthcoming days.

References

[1]. Sivarajah, U., Kamal, , Z., Weerakkody, V.: Critical analysis of Big Data challenges and analytical methods. J. Bus. Res. 70, 263–286 (2017).
[2]. Du, Li, A., Zhang, L.: Survey on the Applications of Big Data in Chinese Real Estate Enterprise. Procedia–Procedia Comput. Sci. 30, 24–33 (2014).
[3]. Mauroandrea, . De, Greco, M., Grimaldim, M., Table, V.: What is Big Data ? A Consensual Definition and a Review of Key Research Topics. 97, (2015).
[4]. Gandomi, A., Haider, M.: Beyond the hype: Big data concepts, methods, and analytics. Int. J. Inf. Manage. 35, 137–144 (2015).
[5]. Özköse, H., P.L.Q., Gencer, C.: Yesterday , Today and Tomorrow of Big Data. 195, 1042–1050 (2015).
[6]. Abaker, I., Hashem, I., Badrul, N., Mokhtar, S., Gani, A., Ullah, S.: The rise of " big data " on cloud computing : review and open research issues. Inf. Syst. 47, 98–115 (2015).
[7]. Cao, L.: Data Science: A Comprehensive Overview. 50, (2017).
[8]. Tan, M.B., Saleh, I., Dustdar, S.: Social-Network-Sourced Big Data Analytics. IEEE Internet Comput. 17, 62–69 (2013).
[9]. Williams, G.J., Office, A.T.: Big Data Opportunities and Challenges: Discussions from Data Analytics Persoectives. Comput. Intell. Mag. IEEE. 9,62–74 (2014).
[10]. Kumar, Praveen, Bhawna Dhruv and Vijay S. Rathore. "Present and future access methodologies of big data." Int J Adv Res Sci Eng 8354, no. 3 (2014): 541-547.
[11]. Hu, H. , Chua, T.-S., Li, X.: Toward Scalable Systems for Big Data Analytics: A Technology Tutorial. IEEE Access. 2, 652–687 (2014).
[12]. Calheiros, R.N., Bianchi, S., Netto, M.A.S., Buyya, R.: Big Data computing and clouds: Trends and future directions. J. Parallel Distrib. Comput. 79–80, 3–15 (2015).
[13]. Chen, M., Mao, Y.: Big Data : A Survey. 171–209 (2014). 17. Singh, D., Reddy, C.K.: A survey on platforms for big data analytics. J. Big Data. 2, 8 (2014).
[14]. Colombo, E.: Privacy Aware Access Control for Big Data: A Research Roadmap. Big Data Res. 2, 145–154 (2015).
[15]. Saxena, Piyush, Satyajit Padhy, and Praveen Kumar. "Use of storage as a service for online operating system in Cloud Computing." International Conferenc on Telecom and Networks (2013).
[16]. Rumbold, J.M.M., Pierscionek, B.K.: What Are Data? A Categorization of the Data Sensitivity Spectrum. Big Data Res. (2017).
[17]. Kumar, Praveen, and Vijay Singh Rathore. "Improvising and optimizing resource utilization in big data processing." In Proceedings of Fifth International Conference on Soft Computing for Problem Solving, pp. 345-353. Springer, Singapore, 2016.

