IOSR Journal of Engineering (IOSRJEN) **www.iosrjen.org**
ISSN (e): 2250-3021, ISSN (p): 2278-8719
PP 10-15

A Review on Big Data

[1]R.Anusuya, [2]Dr.S.Krishnaveni
Research Scholar Pioneer college of arts and science
Research Supervisor

Abstract: *The term big data arose under the volatile increase of global data as a technology that is able to store and process big and varied volumes of data, providing both enterprises and science with deep insights over its clients/experiments. Although big data solves much of our current problems it still presents some gaps and issues that raise concern and need improvement. Big data is the term for any collection of datasets so large and complex that it becomes difficult to process using traditional data processing applications. The challenges include analysis, capture, search, sharing, storage, transfer, visualization, and privacy violations. 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. Data that is so large in volume, so diverse in variety or moving with such velocity is called Big data.This paper provides an overview of big data mining and discusses the related challenges and the new opportunities.*

Keywords— *Big Data, Hadoop, Map Reduce,*

I. INTRODUCTION

The concept of big data became a major strength of innovation across both academics and corporations. The pattern is viewed as an effort to understand and get proper insights from big datasets (big data analytics), providing summarized information over huge data loads. Big data is a largest drone phrases in domain of IT, new technologies of personal communication driving the big data new trend and internet population grew day by day but it never reach by 100%. The need of big data generated from the large companies like facebook, yahoo, Google, YouTube etc for the purpose of analysis of enormous amount of data which is in unstructured form or even in structured form. Google contains the large amount of information. So there is the need of Big Data tools that is the processing of the complex and massive datasets This data is different from structured data in terms of five parameters –variety, volume, value, veracity and velocity (5V's). The five V's (volume,variety, velocity, value, veracity) are the challenges of big data management are:

1. Volume:

Data is ever-growing day by day of all tpes ever MB, PB, YB, ZB, KB, TB of information. The data results into large files. Excessive volume of data is main issue of storage. This main issue is resolved by reducing storage cost. Data volumes are expected to grow 50 times by 2020.

2. Variety:

Data sources are extremely heterogeneous. The files comes in various formats and of any type, it may be structured or unstructured such as text, audio, videos, log files and more. The varieties are endless, and the data enters the network without having been quantified or qualified in any way.

3. Velocity:

The data comes at high speed. Sometimes 1 minute is too late so big data is time sensitive. Some organizations data velocity is main challenge. The social media messages and credit card transactions done in millisecond and data generated by this putting in to databases.

4. Value:

It is a most important v in big data. Value is main buzz for big data because it is important for businesses, IT infrastructure system to store large amount of values in database.

5. Veracity:

The increase in the range of values typical of a large data set. When we dealing with high volume, velocity and variety of data, the all of data are not going 100% correct, there will be dirty data. Big data and analytics technologies work with these types of data.

Huge volume of data (both structured and unstructured) is management by organization, administration and governance. Unstructured data is a data that is not present in a database. Unstructured data may be text, verbal data or in another form. Textual unstructured data is like power point presentation, email messages, word documents, and instant massages. Data in another format can be.jpg images,.png images and audio files.

II. THE HISTORY OF BIG DATA IN STAGES:

Megabyte to Gigabyte: In the 1970s and 1980s, historical business data introduced the earliest ``big data'' challenge in moving from megabyte to gigabyte sizes. The urgent need at that time was to house that data and run relational queries for business analyses and reporting. Research efforts were made to give birth to the ``database machine'' that featured integrated hardware and software to solve problems. The underlying philosophy was that such integration would provide better performance at lower cost. After a period of time, it became clear that hardware- specialized database machines could not keep pace with the progress of general-purpose computers. Thus, the descendant database systems are soft-ware systems that impose few constraints on hardware and can run on general- purpose computers.

Gigabyte to Terabyte: In the late 1980s, the popularization of digital technology caused data volumes to expand to several gigabytes or even a terabyte, which is beyond the storage and/or processing capabilities of a single large computer system. Data parallelization was proposed to extend storage capabilities and to improve performance by distributing data and related tasks, such as building indexes and evaluating queries, into disparate hardware. Based on this idea, several types of parallel databases were built, including shared-memory databases, shared-disk databases, and shared-nothing databases, all as induced by the underlying hardware architecture.

Petabyte to Exabyte: Under current development trends,data stored and analyzed by big companies will undoubt- edly reach the PB to exabyte magnitude soon. However, current technology still handles terabyte to PB data; there has been no revolutionary technology developed to cope with larger datasets.

III. BIG DATA APPLICATIONS

Smart Grid case*:* It is crucial to manage in real time the national electronic power consumption and to monitor Smart grids operations. This is achieved through multiple connections among smart meters, sensors, control centers and other infrastructures. Big Data analytics helps to identify at-risk transformers and to detect abnormal behaviors of the connected devices. Grid Utilities can thus choose the best treatment or action. The real-time analysis of the generated Big Data allow to model incident scenarios. This enables to establish strategic preventive plans in order to decrease the corrective costs. In addition, Energy-forecasting analytics help to better manage power demand load, to plan resources, and hence to maximize prots.

E-health: Connected health platforms are already used to personalize health services (e.g., CISCO solution. Big Data is generated from different heterogeneous sources (e.g., laboratory and clinical data, patients symptoms uploaded from distant sensors, hospitals operations, pharmaceutical data). The advanced analysis of medical data sets has many beneficial applications. It enables to personalize health services (e.g., doctors can monitor online patients symptoms in order to adjust prescription); to adapt public health plans according to population symptoms, disease evolution and other parameters. It is also useful to optimize hospital operations and to decrease health cost expenditure.

Internet of Things (IoT): IoT represents one of the main markets of big data applications. Because of the high variety of objects, the applications of IoT are continuously evolving. Nowadays, there are various Big Data applications supporting for logistic enterprises. In fact, it is possible to track vehicles positions with sensors, wireless adapters, and GPS. Thus, such data driven applications enable companies not only to supervise and manage employees but also to optimize delivery routes. This is by exploiting and combining various information including past driving experience. Smart city is also a hot research area based on the application of IoT data.

Public utilities*:* Utilities such as water supply organizations are placing sensors in the pipelines to monitor flow of water in the complex water supply networks. It is reported in the Press that Bangalore Water Supply and Sewage Board is implementing a real-time monitoring system to detect leakages, illegal connections and remotely control valves to ensure equitable supply of water to different areas of the city. It helps tp reduce the need for valve operators and to timely identifying and fixing water pipes that are leaking.

Transportation and logistics: Many public road transport companies are using RFID (Radiofrequency Identification) and GPS to track buses and explore interesting data to improve there services… For instance, data collected about the number of passengers using the buses in different routes are used to optimize bus routes and the frequency of trips. various real-time system has been implemented not only to provide passengers with recommendations but also to offer valuable information on when to expect the next bus

which will take him to the desired destination. Mining Big Data helps also to improve travelling business by predicting demand about public or private networks. For instance, in India that has one of the largest railway networks in the world, the total number of reserved seats issued every day is around 250,000 and reservation can be made 60 days in advance. Making predictions from such data is a complicated issue because it depends on several factors such as weekends, festivals, night train, starting or intermediate station. By using the machine learning algorithms, it is possible to mine and apply advanced analytics on past and new big data collection. In fact advanced analytics can ensure high accuracy of results regarding many issues.

Political services and government monitoring: Many governments such as India and United States are mining data to monitor political trends and analyze population sentiments. There are many applications that combine many data sources: social network communications, personal interviews, and voter compositions. Such systems enable also to detect local issues in addition to national issues. Furthermore, governments may use Big Data systems to optimize the use of valuable resources and utilities. For instance, sensors can be placed in the pipelines of water supply chains to monitor water flow in large networks. So it is possible for many countries to rely on real-time monitoring system to detect leakages, illegal connections and remotely control valves to ensure equitable supply of water to different areas of the city.

IV. BIG DATA CHALLENGES

The mining of Big Data offers many attractive opportunities. However, researchers and professionals are facing several challenges when exploring Big Data sets and when extracting value and knowledge from such mines of information. The difficulties lye at different levels including: data capture, storage, searching, sharing, analysis, management and visualization. Furthermore, there are security and privacy issues especially in distributed data driven applications. Often, the torrent of information and distributed streams surpass our capability to harness. In fact, while the size of Big Data keeps increasing exponentially, the current technological capacity to handle and explore Big Data sets, is only in the relatively lower levels of petabytes, exabytes and zettabytes of data.

1. Dealing with data growth

The most obvious challenge associated with big data is simply storing and analyzing all that information. In its Digital Universe report, IDC estimates that the amount of information stored in the world's IT systems is doubling about every two years. By 2020, the total amount will be enough to fill a stack of tablets that reaches from the earth to the moon 6.6 times. And enterprises have responsibility or liability for about 85 percent of that information.

Much of that data is unstructured, meaning that it doesn't reside in a database. Documents, photos, audio, videos and other unstructured data can be difficult to search and analyze.

In order to deal with data growth, organizations are turning to a number of different technologies. When it comes to storage, converged and hyper converged infrastructure and software-defined storage can make it easier for companies to scale their hardware. And technologies like compression, deduplication and tiering can reduce the amount of space and the costs associated with big data storage.

2. Generating insights in a timely manner

Of course, organizations don't just want to store their big data — they want to use that big data to achieve business goals. According to the NewVantage Partners survey, the most common goals associated with big data projects included the following:

1. Decreasing expenses through operational cost efficiencies
2. Establishing a data-driven culture
3. Creating new avenues for innovation and disruption
4. Accelerating the speed with which new capabilities and services are deployed
5. Launching new product and service offerings

All of those goals can help organizations become more competitive — but only if they can extract insights from their big data and then act on those insights quickly. PwC's Global Data and Analytics Survey 2016 found, "Everyone wants decision-making to be faster, especially in banking, insurance, and healthcare."

4. Integrating disparate data sources

The variety associated with big data leads to challenges in data integration. Big data comes from a lot of different places — enterprise applications, social media streams, email systems, employee-created documents, etc. Combining all that data and reconciling it so that it can be used to create reports can be incredibly difficult. Vendors offer a variety of ETL and data integration tools designed to make the process easier, but many enterprises say that they have not solved the data integration problem yet.

In response, many enterprises are turning to new technology solutions. In the IDG report, 89 percent of those surveyed said that their companies planned to invest in new big data tools in the next 12 to 18 months. When asked which kind of tools they were planning to purchase, integration technology was second on the list, behind data analytics software.

5. Validating data

Closely related to the idea of data integration is the idea of data validation. Often organizations are getting similar pieces of data from different systems, and the data in those different systems doesn't always agree. For example, the ecommerce system may show daily sales at a certain level while the enterprise resource planning (ERP) system has a slightly different number. Or a hospital's electronic health record (EHR) system may have one address for a patient, while a partner pharmacy has a different address on record.

Solving data governance challenges is very complex and is usually requires a combination of policy changes and technology. Organizations often set up a group of people to oversee data governance and write a set of policies and procedures. They may also invest in data management solutions designed to simplify data governance and help ensure the accuracy of big data stores — and the insights derived from them.

6. Securing big data

Security is also a big concern for organizations with big data stores. After all, some big data stores can be attractive targets for hackers or advanced persistent threats (APTs).However, most organizations seem to believe that their existing data security methods are sufficient for their big data needs as well.

V. BIGDATA TECHNOLOGIES

Hadoop

Hadoop is a framework that can run applications on systems with thousands of nodes and terabytes. It distributes the file among the nodes and allows to system continue work in case of a node failure. This approach reduces the risk of catastrophic system failure.

In which application is broken into smaller parts (fragments or blocks).Apache Hadoop consists of the Hadoop kernel, Hadoop distributed file system (HDFS), map reduce and related projects are zookeeper, Hbase, Apache Hive. Hadoop Distributed File System consists of three Components: the Name Node, Secondary Name Node and Data Node. The multilevel secure (MLS) environmental problems of Hadoop by using security enhanced Linux (SE Linux) protocol. In which multiple sources of Hadoop applications run at different levels.

This protocol is an extension of Hadoop distributed file system. Hadoop is commonly used for distributed batch index building; it is desirable to optimize the index capability in near real time. Hadoop provides components for storage and analysis for large scale processing. Now a day's Hadoop used by hundreds of companies.The advantage of Hadoop is Distributed storage & Computational capabilities, extremely scalable, optimized for high throughput, large block sizes, tolerant of software and hardware failure.

Hadoop is an Open Source implementation of a large-scale batch processing system. That use the Map-Reduce framework introduced by Google by leveraging the concept of map and reduce functions that well known used in Functional Programming. Although the Hadoop framework is written in Java, it allows developers to deploy custom- written programs coded in Java or any other language to process data in a parallel fashion across hundreds or thousands of commodity servers. It is optimized for contiguous read requests(streaming reads), where processing includes of scanning all the data. Depending on the complexity of the process and the volume of data, response time can vary from minutes to hours. While Hadoop can processes data fast, so its key advantage is its massive scalability.

Hadoop is currently being used for index web searches, email spam detection, recommendation engines, prediction in financial services, genome manipulation in life sciences, and for analysis of unstructured data such as log, text, and clickstream. While many of these applications could in fact be implemented in a relational database(RDBMS), the main core of the Hadoop framework is functionally different from an RDBMS. The following discusses some of these differences Hadoop is particularly useful when:

Complex information processing is needed:

- Unstructured data needs to be turned into structured data.
- Queries can't be reasonably expressed using SQL Heavily recursive algorithms.
- Complex but parallelizable algorithms needed, such as geo-spatial analysis or genome sequencing.

Machine learning:

- Data sets are too large to fit into database RAM, discs, or require too many cores (10's of TB up to PB) .
- Data value does not justify expense of constant real-time availability, such as archives or special interest

info, which can be moved to Hadoop and remain available at lower cost.
- Results are not needed in real time Fault tolerance is critical.

Significant custom coding would be required to:
- Handle job scheduling.

Hadoop was inspired by Google's MapReduce, a software framework in which an application is broken down into numerous small parts. Any of these parts (also called fragments or blocks) can be run on any node in the cluster. Doug Cutting, Hadoop's creator, named the framework after his child's stuffed toy elephant. The current Apache Hadoop ecosystem consists of the Hadoop kernel, MapReduce, the Hadoop distributed file system (HDFS) and a number of related projects such as Apache Hive, HBase and Zookeeper. The Hadoop framework is used by major players including Google, Yahoo and IBM, largely for applications involving search engines and advertising. The preferred operating systems are Windows and Linux but Hadoop can also work with BSD and OS X.

Map Reduce

Map-Reduce was introduced by Google in order to process and store large datasets on commodity hardware. Map Reduce is a model for processing large-scale data records in clusters. The Map Reduce programming model is based on two functions which are map() function and reduce() function. Users can simulate their own processing logics having well defined map() and reduce() functions. Map function performs the task as the master node takes the input, divide into smaller sub modules and distribute into slave nodes. A slave node further divides the sub modules again that lead to the hierarchical tree structure. The slave node processes the base problem and passes the result back to the master Node. The Map Reduce system arrange together all intermediate pairs based on the intermediate keys and refer them to reduce() function for producing the final output. Reduce function works as the master node collects the results from all the sub problems and combines them together to form the output.

Map Reduce Components:

1. **Name Node**: manages HDFS metadata, doesn’tdeal with files directly.
2. **Data Node**: stores blocks of HDFS—default replication level for each block: 3.
3. **Job Tracker**: schedules, allocates and monitorsjob execution on slaves—Task Trackers.
4. **Task Tracker**: runs Map Reduce operations.

VI. BIG-DATA SYSTEM ARCHITECTURE

A big-data system is complex, providing functions to deal with different phases in the digital data life cycle, ranging from its birth to its destruction. At the same time, the system usually involves multiple distinct phases for different applications.The details for each phase are explained as follows.

Data generation concerns how data are generated. In this case, the term ``big data" is designated to mean large, diverse, and complex datasets that are generated from various longitudinal and/or distributed data sources, including sensors, video, click streams, and other available digital sources.

Normally, these datasets are associated with different levels of domain specific values. In this paper, we focus on

datasets from three prominent domains, business, Internet, and scientific research, for which values are relatively easy to understand. However, there are overwhelming technical challenges in collecting, processing, and analyzing these datasets that demand new solutions to embrace the latest advances in the information and communications technology (ICT) domain.

Data acquisition refers to the process of obtaining information and is subdivided into data collection, data transmission, and data pre-processing. First, because data may come from a diverse set of sources, websites that host formatted text, images and/or videos - data collection refers to dedicated data collection technology that acquires raw data from a specific data production environment. Second, after collecting raw data, we need a high-speed transmission mechanism to transmit the data into the proper storage sustaining system for various types of analytical applications.

Data storage concerns persistently storing and managing large-scale datasets. A data storage system can be divided into two parts: hardware infrastructure and data management. Hardware infrastructure consists of a pool of shared ICT resources organized in an elastic way for various tasks in response to their instantaneous demand. The hardware infrastructure should be able to scale up and out and be able to be dynamically recognized to address different types of application environments. Data management software is deployed on top of the hardware infrastructure to maintain large-scale datasets. Additionally, to analyze or

interact with the stored data, storage systems must provide several interface functions, fast querying and other programming models.

Data analysis leverages analytical methods or tools to inspect, transform, and model data to extract value. Many application elds leverage opportunities presented by abundant data and domain specifi analytical methods to derive the intended impact. Although various elds pose different application requirements and data characteristics, a few of these elds may leverage similar underlying technologies. Emerging analytics research can be classified into six critical technical areas: structured data analytics, text analytics, multimedia analytics, web analytics, net-work analytics, and mobile analytics.

VII. CONCLUSIONS

In this paper we have surveyed various technologies to handle the big data and there architectures. In this paper we have also discussed the challenges of Big data (volume, variety, velocity, value, veracity) and various advantages and a disadvantage of these technologies. This paper discussed an architecture using Hadoop and MapReduce distributed data processing. The main goal of our paper was to make a survey of various big data handling techniques those handle a massive amount of data from different sources and improves overall performance of systems.

REFERENCES

[1]. Yuri Demchenko ―The Big Data Architecture Framework (BDAF)‖ Outcome of the Brainstorming Session at the University of Amsterdam 17 July 2013.

[2]. Amogh Pramod Kulkarni, Mahesh Khandewal, ―Survey on Hadoop and Introduction to YARN‖, International Journal of Emerging Technology and Advanced Engineering Website: www.ijetae.com (ISSN 2250-2459, ISO 9001:2008 Certified Journal, Volume 4, Issue 5, May 2014).

[3]. Sagiroglu, S.Sinanc, D.,‖Big Data: A Review‖,2013, 20-24.

[4]. Ms. Vibhavari Chavan, Prof. Rajesh. N. Phursule, ―Survey Paper On Big Data‖ International Journal of Computer Science and Information Technologies, Vol. 5 (6), 2014.

[5]. Margaret Rouse, April 2010―unstructured data‖.

[6]. Kyuseok Shim, MapReduce Algorithms for Big Data Analysis, DNIS 2013, LNCS 7813, pp. 44–48, 2013.

[7]. Dong, X.L.; Srivastava, D. Data Engineering (ICDE),‖ Big data integration‖ IEEE International Conference on , 29(2013) 1245–1248.

[8]. V. Gudivada, D. Rao, and V. Raghavan, "Big Data–Driven Natural Language–Processing Research and Applications," *Big Data Analytics*, V. Govindaraju, V. Raghavan, and C.R. Rao, eds., Elsevier, 2015 (in press).

[9]. A. Halevy, P.Norvig, and F.Pereira,The Unreasonable Effectiveness of Data," *IEEE Intelligent Systems*, vol. 24, no. 2, 2009, pp. 8–12.

[10]. V. Gudivada, D. Rao, and V. Raghavan, "Renaissance in Data Management Systems: SQL, NoSQL, and NewSQL," *Computer* (in press).

[11]. R. Baeza-Yates. "Big Data or Right Data?" *Proc. 7th Alberto Mendelzon Int'l Workshop on Foundations of Data Management* (AMW 13), 2013, vol. 1087, paper 14; http://ceur-ws.org/Vol-1087/paper14.pdf.

[12]. Tekiner F. and Keane J.A., Systems, Man and Cybernetics (SMC), ―Big Data Framework‖ 2013 IEEE International Conference on 13–16 Oct. 2013, 1494–1499.

[13]. Mrigank Mridul, Akashdeep Khajuria, Snehasish Dutta, Kumar N ―Analysis of Big Data using Apache Hadoop and Map Reduce‖ Volume 4, Issue 5, May 2014‖.

[14]. Suman Arora, Dr.Madhu Goel, ―Survey Paper on Scheduling in Hadoop‖ International Journal of Advanced Research in Computer Science and Software Engineering, Volume 4, Issue 5, May 2014.

[15]. Aditya B. Patel, Manashvi Birla and Ushma Nair, "Addressing Big Data Problem Using Hadoop and Map Reduce," in Proc. 2012 Nirma University International Conference On Engineering.

[16]. Jimmy Lin ―Map Reduce Is Good Enough?‖ The control project, IEEE Computer 32 (2013).