A Survey on Big Data Technologies

Santhoshi A,
III-B.Tech (IT), SNS College of Technology,
Coimbatore, Tamil Nadu – 641035

***Abstract*— Big data is a collection of massive and complex data sets that include the huge quantities of data, social media analytics, data management capabilities, and real-time data. Big data came into existence when the traditional relational database systems were not able to handle the unstructured data (weblogs, videos, photos, social updates, human behaviour) generated today by organisation, social media, or from any other data generating source. Data that is so large in volume, so diverse in variety or moving with such velocity is called big data. Big data analytics is the process of examining large amounts of data. Analysing Big Data is a challenging task as it involves large distributed file systems which should be fault tolerant, flexible and scalable. The technologies used by big data application to handle the massive data are Hadoop, Map Reduce, Apache Hive, No SQL and HPCC. These technologies handle massive amount of data. In this research paper various technologies for handling big data along with the advantages and disadvantages of each technology for catering the problems in hand to deal the massive data has discussed.**

Keywords: Big Data, Hadoop, Map Reduce, Apache Hive, No SQL

I. INTRODUCTION

A. Definition

With the growth of technologies and services, the large amount of data is produced that can be structured and unstructured from the different sources. Such type of data is very difficult to process that contains the billions records of millions people information that includes the web sales, social media, audios, images and so on. The need of big data comes from the Big Companies like yahoo, Google, Facebook etc., for the purpose of analysis of big amount of data which is in unstructured form. Google contains the large amount of information. So; there is the need of Big Data Analytics that is the processing of the complex and massive datasets. Big data analytics analyse the large amount of information used to uncover the hidden patterns and the other information which is useful and important information for the use.

B. BIG Data Parameters

Big data is characterized by its 7 V's. The challenging 7 V's of big data are: **V**olume, **V**ariety, **V**elocity, **V**eracity, **V**alue, **V**isualization and **V**ariability (**7 V**).

1. **Volume**: Data is ever-growing day by day of all types 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 (even in the same field or in distinct) are extremely heterogeneous. The files come 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 organisations 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.

6. **Visualization**: Visualisations can contain dozens of variables and parameters- a far cry from the x and y variables of your standard bar chart- and finding a way to present this information that makes the findings clear is one of its challenges.

7. **Variability**: Variability refers to data whose meaning is dynamic. This is particularly the case when gathering data relies on language processing.

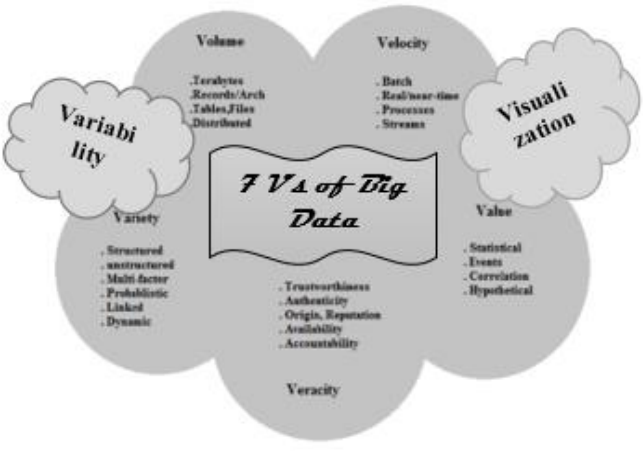

Fig 1. Big Data Parameters

II. LITERATURE REVIEW

The process of the research into complex data basically concerned with the revealing of hidden patterns.

1. **Aditya B. (2012)** in **"Addressing Big Data Problem Using Hadoop and Map Reduce"** defines big data Problem using Hadoop and Map Reduce" reports the experimental research on the Big data problems in various domains. It describes the optimal and efficient solutions using Hadoop cluster, Hadoop Distributed File System (HDFS) for storage data and Map Reduce framework for parallel processing to process massive data sets and records.
2. **Sagiroglu, S., (2013) in**" Big **Data: A Review"** offered the big data content, its scope, functionality, data samples, advantages and disadvantages along with challenges of big data. The critical issue in relation to the Big data is the privacy and protection. Big data samples describe the review about the environment, science and research in biological area. By this paper, we can conclude that any association in any domain having big data can take the benefit from its careful investigation for the problem solving principle. Using Knowledge Discovery from the Big data convenient to get the information from the complicated data records. The overall appraisal describe that the data is mounting day by day and becoming complex. The challenge is not only to gather and handle the data but also how to extract the useful information from that collected data records. In accordance to the Intel IT Centre, there are several challenges related to Big Data which are rapid data growth, data infrastructure, and variety of data,
visualization and data velocity.
3. **John A. Keane** in **2013** proposed a framework in which big data applications can be developed. The framework consist of three stages (multiple data sources, data analysis and modelling, data organization and interpretation) and seven layers (visualisation/presentation layer, service/query/access layer, modelling/ statistical layer, processing layer, and system layer, data layer/multi model) to divide big data application into blocks. The main motive of this paper is to manage and architect a massive amount of big data applications. The advantage of this paper is big data handles heterogeneous data and data sources in timely to get high performance and Framework Bridge the gap with business needs and technical realities. The disadvantage of this paper is too difficult to integrate existing data and systems.

4. **Real Time Literature Review about the Big data** According to **2013**, Facebook has 1.11 billion people active accounts from which 751 million using Facebook from a mobile. Another example is flicker having feature of Unlimited photo uploads (50MB per photo), Unlimited video uploads (90 seconds max, 500MB per video), the ability to show HD Video, Unlimited storage, Unlimited bandwidth. Flickr had a total of 87 million registered members and more than 3.5 million new images uploaded daily.

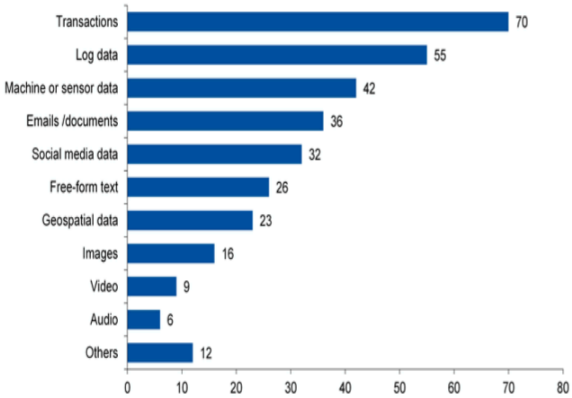

Fig 2. Data sources

III. PROBLEM STATEMENT

Big Data has come up as we are living in society that uses the intensive use of increasing data technology. As the Big data is the latest technology that can be beneficial for the business organizations, so it is necessary that various issues and challenges associated with this technology should bring out into light. The two main issues are the storage capacity and the processing of the data.

IV. BIG DATA TECHNOLOGIES

There are the different technologies which distribute the data among various local agents and reduce the load of the main server so that traffic can be avoided.

A. *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 (HDFS) 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. This protocol is an extension of Hadoop distributed file system (HDFS). Hadoop is commonly used for distributed batch index building. Hadoop provides components for storage and analysis for large scale processing.

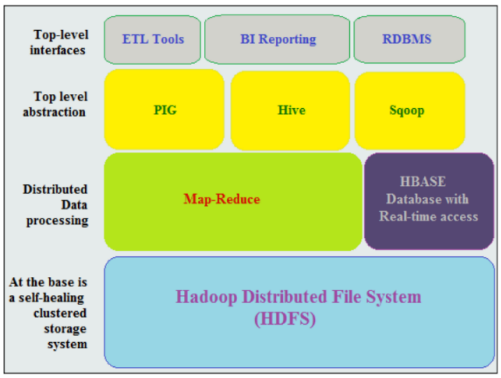

Fig 3: Architecture of Hadoop

Components of Hadoop:

- Base: It is open source, distributed and Non-relational database system implemented in Java. It runs above the layer of HDFS.
- Oozie: Oozie is a web-application that runs in a java servlet. Oozie use the database to gather the information of Workflow. It manages the Hadoop jobs in a mannered way.
- Sqoop: Sqoop is a command-line interface application that provides platform which is used for converting data from relational databases and Hadoop.
- Avro: It is a system that provides functionality of data serialization and service of data exchange.
- Chukwa: Chukwa framework is used for data collection and analysis to process and analyse the massive amount of logs. It is built on the upper layer of the HDFS and Map Reduce framework.
- Pig: Pig is high-level platform where the Map Reduce framework is created which is used with Hadoop platform. It is a high level data processing system where the data records are analysed that occurs in HLL.
- Zookeeper: It is a centralization based service that provides distributed synchronization and provides group services along with maintenance of the configuration information and records.
- Hive: It is application developed for data warehouse that provides the SQL interface as well as relational model.

The advantage of Hadoop is Distributed storage & Computational capabilities, extremely scalable, optimized for high throughput, large block sizes, tolerant of software and hardware failure. The disadvantage of Hadoop is that it is master processes are single points of failure. Hadoop does not offer storage or network level encryption, inefficient for handling small files.

B. Map Reduce

Map-Reduce process and store large datasets on commodity hardware. It 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.

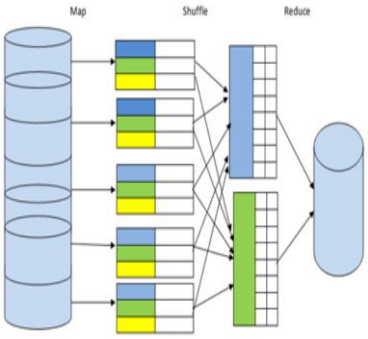

Fig 4. Architecture of Map-reduce

Figure 4 shows the Map Reduce Architecture and Working. It always manages to allocate a local data block to a slave node. If the effort fails, the scheduler will assign a rack-local or random data block to the slave node instead of local data block. When map () function complete its task, the runtime system gathers all intermediate pairs and launches a set of condense tasks to produce the final output. Large scale data processing is a difficult task, managing hundreds or thousands of processors and managing parallelization and distributed environments makes is more difficult. Map Reduce provides solution to the mentioned issues, as is supports distributed and parallel I/O scheduling, it is fault tolerant and supports scalability and it has inbuilt processes for status and monitoring of heterogeneous and large datasets as in Big Data. It is way of approaching and solving a given problem. Using Map Reduce framework the efficiency and the time to retrieve the data is quite manageable. Data aware caching (Dache) framework that made slight change to the original map reduce programming model and framework to enhance processing for big data applications using the map reduce model. The advantage of map reduce is a large variety of problems are easily expressible as Map reduce computations and cluster of machines handle thousands of nodes and fault-tolerance.

Components of Map-Reduce

1. Name Node: manages HDFS metadata, doesn't deal with files directly.
2. Data Node: stores blocks of HDFS—default replication level for each block: 3.
3. Job Tracker: schedules, allocates and monitors job execution on slaves— Task Trackers.
4. Task Tracker: runs Map Reduce operations.

The disadvantage of map reduce is Real-time processing, not always very easy to implement, shuffling of data, batch processing.

C. Hive

Hive is a distributed agent platform, a decentralized system for building applications by networking local system resources. Apache Hive data warehousing component, an element of cloud-based Hadoop ecosystem which offers a query language called HiveQL that translates SQL-like queries into Map Reduce jobs automatically. Applications of apache hive are SQL, oracle, IBM DB2. Architecture is divided into Map-Reduce-oriented execution, Meta data information for data storage, and an execution part that receives a query from user or applications for execution. The advantage of hive is more secure and implementations are good and well-tuned. The disadvantage of hive is only for ad hoc queries and performance is less as compared to pig.

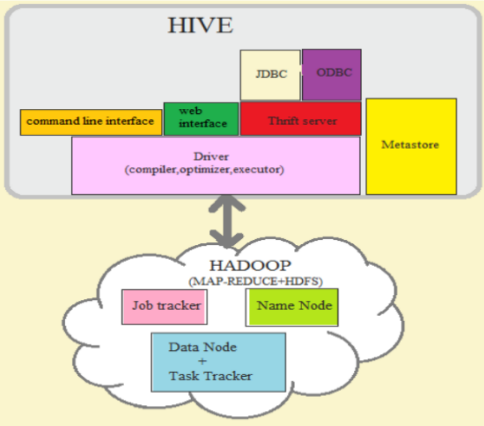

Fig 5: Architecture of HIVE

D. No-SQL

No-SQL database is an approach to data management and data design that's useful for very large sets of distributed data. These databases are in general part of the real-time events that are detected in process deployed to inbound channels but can also be seen as an enabling technology following analytical capabilities. These are only made feasible because of the elastic nature of the NoSQL model where the dimensionality of a query is evolved from the data in scope and domain rather than being fixed by the developer in advance. It is useful when enterprise need to access huge amount of unstructured data. The most popular NoSQL database is Apache Cassandra. The advantage of No-SQL is open source, Horizontal scalability, Easy to use, store complex data types, very fast for adding new data and for simple operations/queries.

The disadvantage of No-SQL is Immaturity, No indexing support, No ACID, Complex consistency models, Absence of standardization.

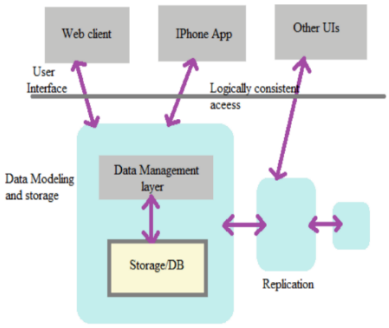

Fig 6: Architecture of No-SQL

E. HPCC

HPCC is an open source platform used for computing and that provides the service for handling of massive big data workflow. HPCC data model is defined by the user end according to the requirements. HPCC system is proposed and then further designed to manage the most complex and data-intensive analytical related problems. HPCC system is a single platform having a single architecture and a single programming language used for the data simulation. HPCC system was designed to analyze the gigantic amount of data for the purpose of solving complex problem of big data. HPCC system is based on enterprise control language which has the declarative and on-procedural nature programming language the main components of HPCC are:

- HPCC Data Refinery: Use parallel ETL engine mostly.
- HPCC Data Delivery: It is massively based on structured query engine used.
- Enterprise Control Language distributes the workload between the nodes in appropriate even load.

V. FUTURE SCOPE

There is nothing concealed that big data significantly influencing IT companies and through development new technologies only we can handle it in a managerial way. Big data totally change the way of organizations, government and academic institution by using number of tools to make the management of big data. In future Hadoop and NoSQL database will be highly in demand moving forward. The amount of data produced by organizations in next five years will be larger than last 5,000 years. In the upcoming years cloud will play the important role for private sectors and organisations to handle the big data efficiently.

VI. CONCLUSION

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 (7 V's) and various advantages and a disadvantage of these technologies. This paper discussed an architecture using Hadoop HDFS distributed data storage, real-time NoSQL databases, and MapReduce distributed data processing over a cluster of

commodity servers. 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] Mukherjee, A.; Datta, J.; Jorapur, R.; Singhvi, R.; Haloi, S.; Akram, W., (18-22 Dec.,2012) , "Shared disk big data analytics with Apache Hadoop"
[2] Aditya B. Patel, Manashvi Birla, Ushma Nair ,(6-8 Dec. 2012),"Addressing Big Data Problem Using Hadoop and Map Reduce"
[3] Sagiroglu, S.; Sinanc, D. ,(20-24 May 2013),"Big Data: A Review" A Real Time Approach with Big Data-A review
[4] Yuri Demchenko "The Big Data Architecture Framework(BDAF)" Outcome of the Brainstorming Session at theUniversity of Amsterdam 17 July 2013.
[5] 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.
[6] Margaret Rouse, April 2010 "unstructured data".
[7] http://searchcloudcomputing.techtarget.com/definition/Hadoop
[8] http://dashburst.com/infographic/big-data-parameters
[9] http://www-01.ibm.com/software/in/data/bigdata/
[10] how-much-data-is-on-the-internet-and-generated-online-every-minute/
[11] Addressing big data problem using Hadoop and Map Reduce Bakshi, K.,(2012)," Considerations for big data: Architecture and approach"