International Journal of Scientific & Engineering Research, Volume 8, Issue 4, April-2017
ISSN 2229-5518

A Study of Analysis the Data Using Big Data Tools and Technologies

P. Sivagami and Dr. A. Priya

Abstract-Data concerns massive amount of data. The data set that include different types such as structured, unstructured and semi-structured data. This data can be generated from different sources like social media, audios, images, log files, sensor data, transactional applications, web etc. To process or analyze this huge amount of data or extracting meaningful information is a challenging task now a days. Big Data are very large data sets which are difficult to manage with ordinary data handling techniques. It exceeds the processing capability of traditional database to capture, manage, and process the voluminous amount of data. Big Data analysis is concerned with finding patterns, trends and associations hidden within .The three main characteristics of Big Data are volume, variety and velocity. In this paper first introduce the general background of big data and then focus on Hadoop platform using map reduce algorithm. Hadoop plays a major role in the IT market. It is a framework for managing the massive amount of heterogeneous data. Many leading giants such as IBM, Microsoft, Yahoo, Amazon are working with this technology. Almost 100% of the other big giants would move on to Hadoop in the upcoming years. This paper is a study of Hadoop - tools and techniques.

Keywords — Big Data, Characteristics of Big Data, Sources of Big Data, Technologies of Big Data, Hadoop Tool, Map Reduce.

1. INTRODUCTION

Every day, we create 2.5 quintillion bytes of data so much that 90% of the world is data was generated last few years. Due to new technologies, devices, communication like social network. The amount of produced by the beginning 2003 year billion of gigabytes. The same amount of data generated every two days in 2011 year. Every ten minutes same amount of data generated in 2013 year. Big Data is as a collection of large dataset. The size of data is Multi terabyte, Peta bytes, Exabyte's & quintillion bytes. Big Data is Capturing data, Curation , Storage, Searching, Sharing, Transfer, Analysis, Presentation. Big Data that cannot be processed using traditional computing techniques. Big Data is not merely a data rather it has become a complete subject which involve various tools, techniques and framework.

2. SOURCES OF BIG DATA

Big data involves the data produced by different devices and applications. Given below are some of the fields that come under the umbrella of Big Data.

- **Black Box Data** : It is a component of helicopter, airplanes, and jets, etc. It captures voices of the flight crew, recordings of microphones and earphones, and the performance information of the aircraft.
- **Social Media Data** : Social media such as Facebook and Twitter hold information and the views posted by millions of people across the globe.
- **Stock Exchange Data** : The stock exchange data holds information about the 'buy' and 'sell' decisions made on a share of different companies made by the customers.
- **Power Grid Data** : The power grid data holds information consumed by a particular node with respect to a base station.
- **Transport Data** : Transport data includes model, capacity, distance and availability of a vehicle.
- **Search Engine Data** : Search engines retrieve lots of data from different databases.

P.sivagami, Research scholar, Asst.Professor,D.K.M College for Women,Vellore.
Dr. A. Priya, Assistant Professor, Department of ComputerScience, Thiruvalluvar University college (Autonomous) Gajalnaikenpatti,Tirupattur.

Figure 1: Sources of Big Data

3. CHALLENGES OF BIG DATA

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 also 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 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 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 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.

3.1 TYPES OF BIG DATA

Thus Big Data includes huge volume, high velocity, and extensible variety of data. The data in it will be of three types.

- **Structured data** : Relational data.
- **Semi Structured data** : XML data.
- **Unstructured data** : Word, PDF, Text, Media Logs.

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.

4. BIG DATA TECHNOLOGIES

Big data technologies are important in providing more accurate analysis, which may lead to more concrete decision-making resulting in greater operational efficiencies, cost reductions, and reduced risks for the business.

To harness the power of big data, you would require an infrastructure that can manage and process huge volumes of structured and unstructured data in real time and can protect data privacy and security.

There are various technologies in the market from different vendors including Amazon, IBM, Microsoft, etc., to handle big data. While looking into the technologies that handle big data, we examine the following two classes of technology:

4.1 OPERATIONAL BIG DATA

NoSQL Big Data systems are designed to take advantage of new cloud computing architectures that have emerged over the past decade to allow massive computations to be run inexpensively and efficiently. This makes operational big data workloads much easier to manage, cheaper, and faster to implement. Some NoSQL systems can provide insights into patterns and trends based on real-time data with minimal coding and without the need for data scientists and additional infrastructure.

4.2 ANAYTICAL BIG DATA

This includes systems like Massively Parallel Processing (MPP) database systems and MapReduce that provide analytical capabilities for retrospective and complex analysis that may touch most or all of the data. MapReduce provides a new method of analyzing data that is complementary to the capabilities provided by SQL, and a system based on MapReduce that can be scaled up from single servers to thousands of high and low end machines.

4.3 BENEFITS OF BIG DATA

The main importance of Big Data consists in the potential to improve efficiency in the context of use a large volume of data, of different type. If Big Data is defined properly and used accordingly, organizations can get a better view on their business therefore leading to efficiency in different areas like sales, improving the manufactured product and so forth.

Using the information kept in the social network like Facebook, the marketing agencies are learning about the response for their campaigns, promotions, and other advertising mediums.

Using the information in the social media like preferences and product perception of their consumers, product companies and retail organizations are planning their production.

Using the data regarding the previous medical history of patients, hospitals are providing better and quick service.

5. TECHNOLOGIES AND METHODS

Big data is a new concept for handling massive data therefore the architectural description of this technology is very new. There are the different technologies which use almost same approach i.e. to distribute the data among various local agents and reduce the load of the main server so that traffic can be avoided. There are endless articles, books and periodicals that describe Big Data from a technology perspective so we will instead focus our efforts here on setting out some basic principles and the minimum technology foundation to help relate Big Data to the broader IM domain.

5.1 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, ptimized for high throughput, large block sizes, tolerant of software and hardware failure.

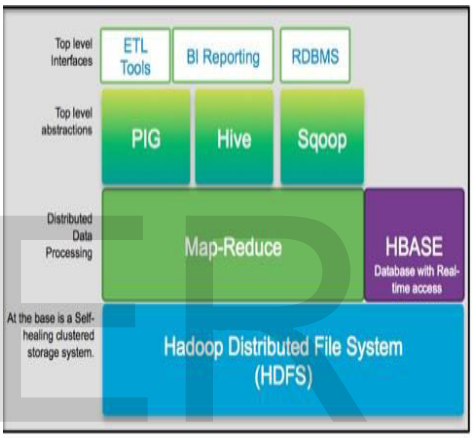

Fig. 2 : Architecture of Hadoop

5.1.1 COMPONENTS OF HADOOP

HBase : It is open source, distributed and Non-relational database system implemented in Java. It runs above the layer of HDFS. It can serve the input and output for the Map

Reduce in well mannered structure.

Oozie : Oozie is a web - application that runs in ajava servlet. Oozie use the database to gather the information of Workflow which is a collection of actions. 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 or vice versa.

Avro : It is a system that provides functionality ofdata serialization and service of data exchange. It is basically used in Apache Hadoop. These services can be used together as well as independently according the data records.

Chukwa : Chukwa is a framework that is used fordata collection and analysis to process and analyze 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 MapReduce framework is created which is used with Hadoop platform. It is a high level data processing system where the data records are analyzed that occurs in high level language.

Zookeeper : It is a centralization based service thatprovides distributed synchronization and provides group services along with maintenance of the configuration information and records.

Hive : It is application developed for datawarehouse that provides the SQL interface as well as relational model. Hive infrastructure is built on the top layer of Hadoop that help in providing conclusion, and analysis for respective queries.

Hadoop was created by Doug Cutting and Mike Cafarella in 2005. Doug Cutting, who was working at Yahoo! at the time, named it after his son's toy elephant. It was originally developed to support distribution for the Nutch search engine project. Hadoop is open - source software that enables reliable, scalable, distributed computing on clusters of inexpensive servers.

Hadoop is:

Reliable : The software is fault tolerant, it expects and handles hardware and software failures

Scalable : Designed for massive scale of processors, memory, and local attached storage.

Distributed : Handles replication. Offers massively parallel programming model, Map Reduce.

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 click stream. 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 Map Reduce, 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.

5.1.2 HDFS

The Hadoop Distributed File System (HDFS) is the file system component of the Hadoop framework. HDFS is designed and optimized to store data over a large amount of low - cost hardware in a distributed fashion.

Name Node : Name node is a type of the master node, which is having the information that means meta data about the all data node there is address(use to talk), free space, data they store, active data node , passive data node, task tracker, job tracker and many other configuration such as replication of data.

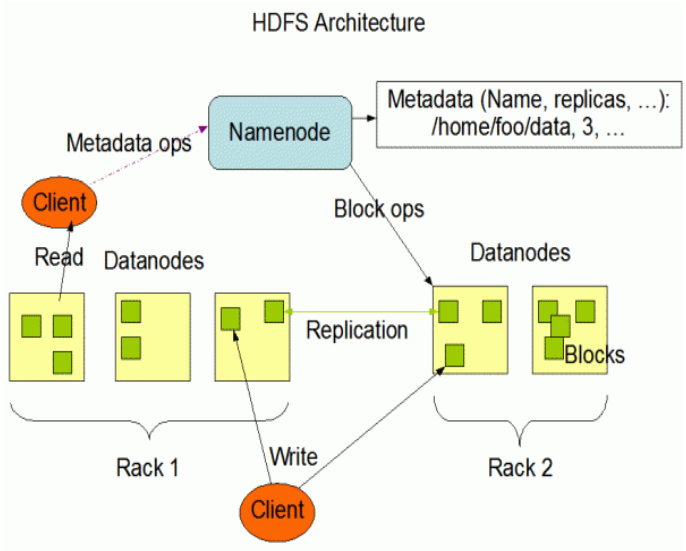

Figure 3 : HDFS Architecture

The Name Node records all of the metadata, attributes, and locations of files and data blocks in to the Data Nodes. The attributes it records are the things like file permissions, file modification and access times, and namespace, which is a hierarchy of files and directories. The Name Node maps the namespace tree to file blocks in Data Nodes. When a client node wants to read a file in the HDFS it first contacts the Name node to receive the location of the data blocks associated with that file.

A Name Node stores information about the overall system because it is the master of the HDFS with the Data Nodes being the slaves. It stores the image and journal logs of the system. The Name Node must always store the most up to date image and journal. Basically, the Name Node always knows where the data blocks and replicates are for each file and it also knows where the free blocks are in the system so it keeps track of where future files can be written.

Data Node: Data node is a type of slave node in the hadoop, which is used to save the data and there is task tracker in data node which is use to track on the ongoing job on the data node and the jobs which coming from name node.
The Data Nodes store the blocks and block replicas of the file system. During startup each Data Node connects and performs a handshake with the Name Node. The Data Node checks for the accurate namespace ID, and if not found then the Data Node automatically shuts down. New Data Nodes can join the cluster by simply registering with the Name Node and receiving the namespace ID. Each Data Node keeps track of a block report for the blocks in its node. Each Data Node sends its block report to the Name Node every hour so that the Name Node always has an up to date view of where block replicas are located in the cluster. During the normal operation of the HDFS, each Data Node also sends a heartbeat to the Name Node every ten minutes so that the Name Node knows which Data Nodes are operating correctly and are available.

The base Apache Hadoop framework is composed of the following modules:
Hadoop Common : contains libraries and utilities needed by other Hadoop modules.
Hadoop Distributed File System(HDFS) : a distributed file system that stores data on commodity machines, providing very high aggregate bandwidth across the cluster.
Hadoop MapReduce : a programming model for large scale data processing.

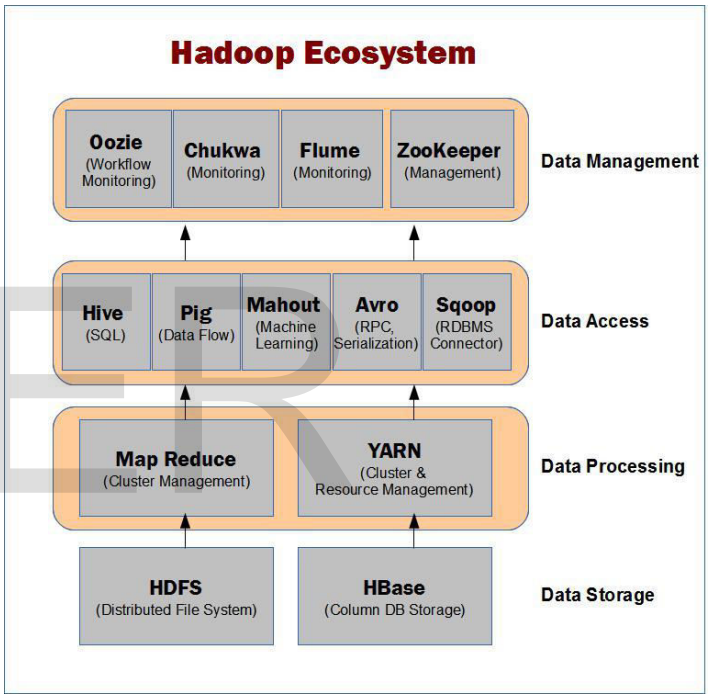

Figure 4 : Hadoop Ecosystem

5.2 MAPREDUCE

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(in_key,in_value) ---
>list(out_key,intermediate_value)Reduce(out_key,list(inter mediate_value))---
>list(out_value)

The parameters of map () and reduce () function is as follows:

map (k1,v1) ! list (k2,v2) and reduce (k2,list(v2)) ! list (v2)

A Map Reduce framework is based on a master-slave architecture where one master node handles a number of slave nodes . Map Reduce works by first dividing the input data set into even-sized data blocks for equal load distribution. Each data block is then assigned to one slave node and is processed by a map task and result is generated. The slave node interrupts the master node when it is idle. The scheduler then assigns new tasks to the slave node. The scheduler takes data locality and resources into consideration when it disseminates data blocks.

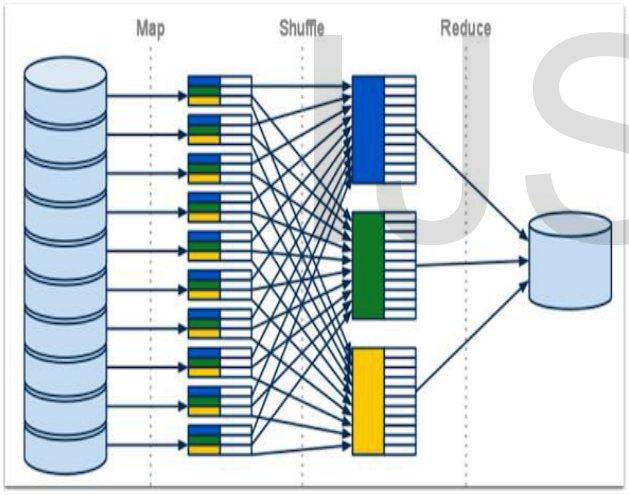

Figure 5 : Architecture of Map Reduce

Figure 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 gather 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. To address the volume aspect, new techniques have been proposed to enable parallel processing using Map Reduce framework. 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.

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

5.2.1 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.

Map Reduce Framework Map Reduce is a software framework for distributed processing of large data sets on computer clusters. It is first developed by Google .Map Reduce is intended to facilitate and simplify the processing of vast amounts of data

in parallel on large clusters of commodity hardware in a reliable, fault - tolerant manner.

MapReduce is the key algorithm that the Hadoop MapReduce engine uses to distribute work around a cluster. Typical Hadoop cluster integrates MapReduce and HFDS layer.

In MapReduce layer job tracker assigns tasks to the task tracker.

Master node job tracker also assigns tasks to the slave node task tracker figure 6

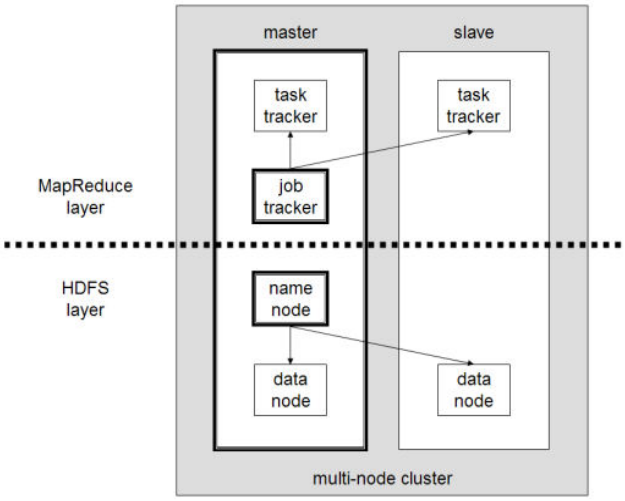

Figure 6 Map Reduce based on Master - Slave Architecture

Master Node contains :
Job tracker node (Map Reduce layer)Task tracker node (Map Reduce layer) Name node (HFDS layer).
Data Node(HFDS layer).
Multiple slave Nodes contain :
Task tracker Node (Map Reduce layer) Data node (HFDS layer)
Map Reduce layer has job and task tracker nodes HFDS layer has name and data nodes .

6. CONCLUSION

The data is generated and proliferating worldwide both from machines and human beings at different speeds and in different formats due to tweeters, face book, stock trading sites, news sources and so on. As I entered the era of Big Data processing large Volumes of data have been greater. Big Data is becoming the new area of research. Big data analysis helps business people to make better decisions and researchers to identify new opportunities. Technical challenges must be addressed for efficient and fast processing of Big Data.

This paper presents fundamental concepts of Big data like characteristics, sources, frameworks and technologies to handle big data. Through better Big data analysis tools like Map Reduce over Hadoop and HDFS, guarantees faster advances in many scientific disciplines and improving the profitability and success of many enterprises. MapReduce has received a lot of attentions in many fields, including data mining, information retrieval, image retrieval, machine learning, and pattern recognition. However, as the amount of data that need to be processed grows, many data processing methods have become not suitable or limited. This paper exploits the MapReduce framework for efficient analysis of big data and for Solving challenging data processing problems on large scale datasets in different domains. MapReduce provides a simple way to scale your application. It Effortlessly scale from a single machine to thousands, providing Fault tolerant & High performance.

7. FUTURE ENHANCEMENT

To provide security threats for big data. Over the next years spending on cloud - based Big Data and analytics (BDA) solutions will grow three times faster than spending for on-premise solutions. Hybrid on/off premise deployments will become a requirement. unified data platform architecture will become the foundation of BDA strategy. The unification will occur across information management, analysis, and search technology. Rich media (video, audio, image) analytics will at least triple in 2015 and emerge as the key driver for BDA technology investment.

8. REFERENCES

1. https://en.wikipedia.org/wiki/Big_data accessed ondated 13/03/2016.
2. Bernice Purcell, The emergence of "big data" technology and analytics, Journal of Technology Research, Holy Family university.
3. Dr. Nitin V. Wnkhade 1, Anusha B. Dhakite 2, Big Data: Overview International Journal of Advanced Research in Computer science and Software Engineering, Volume 6, Issue 6, June 2016,(ISSN 2277 – 128X(Online).
4. Salil Jagtap 1, Shraddha Malviya 2, Big Data: The New Era of Storing Data International Journal of Advanced Research in Computer science and Software Engineering, Volume 6, Issue 7, June 2016,(ISSN 2277 – 128X(Online).
5. G. Aloisioa,b, S. Fiorea,b, Ian Fosterc, D. Williamsd, Scientific big data analytics challenges at large scale.
6. Apache Hive. Available at http://hive.apache.org.
7. http://www.forbes.com/sites/gartnergroup /2013/03/27/gartners big data definition consists of three parts not to be confused with three vs/
8. Apache Hadoop ! (hadoop.apache.org)
9. Hadoop on Wikipedia(http://en.wikipedia.org/wiki/Hadoop)
10. www.guruzon.com/6/introduction/Hadoop
11. Cloudera - Apache Hadoop for the Enterprise(http://www.cloudera.com)
12. Varsha B.babade, Big Data and Hadoop: International research Journal of Engineering, Volume 3, Issue 1, Jan 2016,(e-ISSN 2395 – 0056(Online).