# Big Data challenges and Hadoop as one of the solution of big data with its Modules

Tapan P. Gondaliya, Dr. Hiren D. Joshi

Abstract— this is an era of the tools, trend and technology. And these tools, trend and technology era is mainly depending upon the information. Or this information is called the data and that information is stored in database just like a My SQL, Oracle, Access and other many more database available in different companies. But problem is in our daily life we used tools and devices for the entertainment, websites, Scientific instrument, and many more different devices that generate a large amount of data and that is in zeta bytes or in Petabytes that is called big data and that data create a big problem in our day to day life. Very tough to organize or manage this kind of big data but through Apache Hadoop it's trouble-free. Basically in this paper we describe the main purpose of Apache Hadoop and how can Hadoop organize or manage the different kinds of data and what are the main techniques are to be used behind that.

**Index Terms**— Big Data, Hadoop, Map Reduce, Hadoop Distributed File System, Big Data Challanges

## 1 INTRODUCTION

This is the era of digitalization and we all live in digital world. In this era we need more and more facility for our life is being easier. So we used different kind of new things for entertaining our life using audio video and digital images, through social stuff like a Facebook Google+ Tweeter, using mobile as well as sensor devices, we used company or government sector that all this increases the digitalization and create a large amount of structured, semi structured and unstructured data. Face book and Google+ or other social website activity daily billions of data are being shared at the same time. Through sensor network, scientific instrument and mobile devices also generates large amount of data in couple of hours. In paper of the big data spectrum they said one interesting thing is IDC Terms this as the digital universe and IDC also predict that in year 2015 this digital universe is generates an unbelievable 8 Zeta bytes of big data. [2][4]

Big Data is just simply like a smaller data but difference is it's a large volume of data is in Zeta bytes that's why it's called big data. Large data create a problem to the company as well as government organization for storage, capture, search, transfer, sharing, analysis, and visualization. To over comes this problem apache introduced one free open source tool that is called the Hadoop.[9] Since 80 percent of data is in "unstructured" form, it must be formatted in a way that makes it suitable for subsequent analysis and data mining. Apache Hadoop is core platform for structuring Big Data, and cracks the problem of making it useful for analytics purpose


Fig.1. Sources whose generated massive amount of data [24] [25] [26] [27]

## 2 Hadoop

First of all Google was introduced the GFS and MapReduce technique and they published that paper was in 2004 for announcing the new technology that is the successful technology

- *Tapan P. Gondaliya, PhD Scholar, School of Computer Science, R.K. University, Rajkot, Gujarat, India, E-mail: tapan.gondaliya@rku.ac.in*
- *Dr. Hiren D. Joshi, Associated Professor cum I/C, School of Computer Science, Dr. Babasaheb Ambedkar Open University, Ahemadabad, Gujarat, India, E-mail: hiren.joshi@baou.edu.in*

but after that doug cutting got an opportunities and develop an open source version of MapReduce and that is called the apache Hadoop. Now Hadoop is a core part in different big companies or websites like a Facebook, Yahoo, LinkedIn, and Tweeter etc. Apache Hadoop is an open source project governed by an ASF (Apache Software Foundation) that allows you to gain insight from large amounts of semi structured, unstructured, structured data quickly and without significant investment. [16]

Hadoop gives a distributed file system and a framework for the analysis and transformation of very large data sets using the MapReduce techniques [1]. Basically Hadoop is the product of apache and also an open source distributed software platform for processing and storing. Hadoop code is mainly written in java language, it runs on a cluster of industry standard servers configured with direct attached storage. Purpose behind using apache Hadoop is we can store Petabytes or Exabyte of data reliably on tens of thousands of servers while scaling performance cost-effectively by simply adding inexpensive nodes to the cluster. Or we can also say that Hadoop is designed to run on hardware and can scale up or scale down without any interruption or fault. It contains three main functionality storage, processing and resource management. [8] Hadoop architecture is described as under.

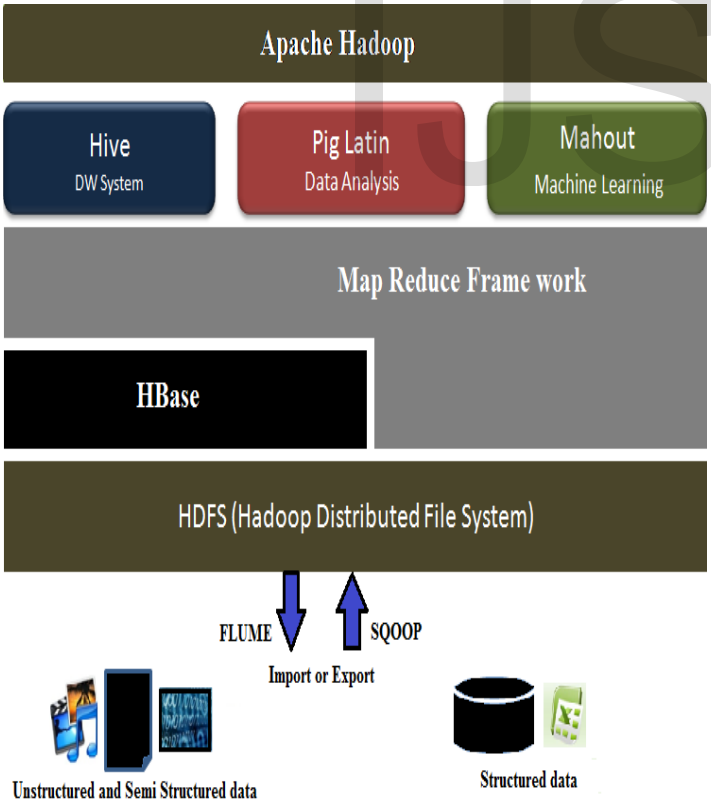


Fig.2 Ecosystem of Hadoop [13]

In above figure here we describe the Apache Hadoop is the kind of tools and all other is the components of the Hadoop ecosystem and what are the main use of that component is explaid as under one by one.

### 2.1 Apache HIVE

An Apache Hive is one kind of the (DW) data warehouse infrastructure built on top most level in Apache Hadoop. Apache Hive providing a functionality of ad-hoc query, analysis of large datasets and data summarization. And it is also provides a technique to project structure in to the data in Hadoop and query that data using a SQL-like language that is called HiveQL. Hive tool also permits to the user for explore and structure the data, for analyse it, and then turn it into business approach. [16] Hives ease integration between Hadoop and tools for data visualization and business approach. Hive database table are same as tables in a relational database, in hive data units are managed in taxonomy from larger to more granular units. Database is comprised of different tables and which are basically made up of partitions. Data can be access through a simple query language that is called HiveQL, which is same like a Structure Query Language (SQL). Hive support appending data or overwriting, but not deleting and updating. Apache Hive supports primitive type of data formats like a string, Boolean, float, binary, decimal, integer, double, smallint, and bigint. Hive also combines both kind of data types primitive data types and complex data types just like maps, array and structs. [14]

### 2.2 Apache Pig

Apache Pig permits to user to write compound MapReduce conversions using a simple scripting language. Basically Pig translates the Pig Latin script into MapReduce so that it can be executed within Hadoop. The language Pig Latin defines a bunch of transformations on a dataset such as join, aggregate and sort. Pig Latin is sometimes extended using User Defined Functions (UDFs), and that function developers can write in Java or a scripting language and then call that function directly from the Pig Latin. Pig Latin was designed for performing a long series of data operations, making it model for three categories of Big Data operations: data pipelines, research on raw data, standard extract-transform-load (ETL) and iterative processing of data. [16]

### 2.3 Apache Mahout

Mahout is also a project of apache. Apache Mahout is a collection of scalable machine-learning algorithms; apply on the top of level in Apache Hadoop and using the MapReduce paradigm. Some algorithm as define here Distributed Item-based Collaborative Filtering, Canopy Clustering, Hierarchical Clustering, K-Means Clustering, Spectral Clustering. Basically Machine learning is a field of artificial intelligence hub on enabling machines to learn without being explicitly programmed, and it is mainly used to improve future performance based on previous outcomes. When once big data is stored in the Apache Hadoop Distributed File System (HDFS), Mahout giving the best data science tools to manually find meaningful

patterns in the big datasets. The Mahout Project main aims to make it faster and easier to turn big data into big information. [16] Apache Mahout gives an applying of various machine learning algorithms, including local mode and distributed mode. Some algorithm as define here Distributed Item-based Collaborative Filtering, Canopy Clustering, Hierarchical Clustering, K-Means Clustering, Spectral Clustering.[16]

### 2.4 Apache HBase

Apache HBase is the Hadoop database, a distributed, and scalable, big data store tools. Apache HBAS project's main goal is hosting a very large table that contains billions of rows and millions of columns a top cluster of commodity hardware. Apache HBase Used when you need random as well as real-time read or write access to your Big Data. HBase provide a good features to the user Strictly steady reads and writes, Easy to use Java API for client access, Linear and modular scalability, Automatic and configurable sharing of tables. Apache HBase is a column-oriented database management system that runs on top of HDFS. It is well suited for sparse data sets, which are common in many big data use cases. [17]

### 2.5 Map Reduce Framework

MapReduce Framework is the heart of apache Hadoop. It is a programming model that permits to the user for massive scalability across large number of servers in a Hadoop cluster. The MapReduce concept is very easy to understand for those who well known about the scale-out data processing solutions in clustered. MapReduce is an actually refers to two combine tasks that is basically Hadoop programs perform. The first is the map job, and second is reduce job. In map task set of data and converts it into another set of data, where individual elements are broken down into key/value pairs. Reduce job takes the output from a map as input and combines those data tuples into a smaller set of tuples. [18]

### 2.6 Hadoop Distibuted File System

HDFS is the distributed file system. The Hadoop Distributed File System (HDFS) is mainly used to store very large data set simply and reliably, and to streaming those dataset at high bandwidth to user applications. In a large cluster, thousands of servers both host directly attached storage and execute user application tasks. By distributing storage and computation across many servers, the resource can grow with demand while remaining economical at every size.

### 2.7 Apache Flume

Flume is a reliable, distributed and available service for efficiently collecting, moving and aggregating big amounts of log data. Flume has a simplest and flexible type of architecture based on streaming data flows. Flume is a robust and fault tolerant technique with changeable reliability mechanisms and many failover and recovery mechanisms. Flume uses a simple extensible data model that allows for online analytic application. [16]

### 2.8 Apache Sqoop

Apache Sqoop is a one kind of application run through command-line interface for transferring purpose of data between Hadoop and relational database. [8] Apache Sqoop mainly supports a free form SQL query as well as saved jobs which can be run multiple times to import updates made to a database since the last import or incremental loads of a single table. Hive or HBase also provide a facility of imports and it is used to populate tables in. And exports can be used to put data from Hadoop into a relational database. In year March 2012 Sqoop became a top level Apache project. [16]

## 3 MODULE OF HADOOP

Hadoop Apache Hadoop provide a so many module to the user and that provide a different functionality to the user but main four module of Hadoop is Hadoop Common, HDFS, MapReduce and YARN. First one is Hadoop Common is a kind of common utility that provide functionality for support the other Hadoop modules. Second module of the Hadoop is Hadoop distributed File System for storing a variety of data, and Third module is Hadoop MapReduce for parallel processing engine. Last but not least the Hadoop YARN is providing a framework for job scheduling and cluster resource management. These all module is used for the mange or organized big data.

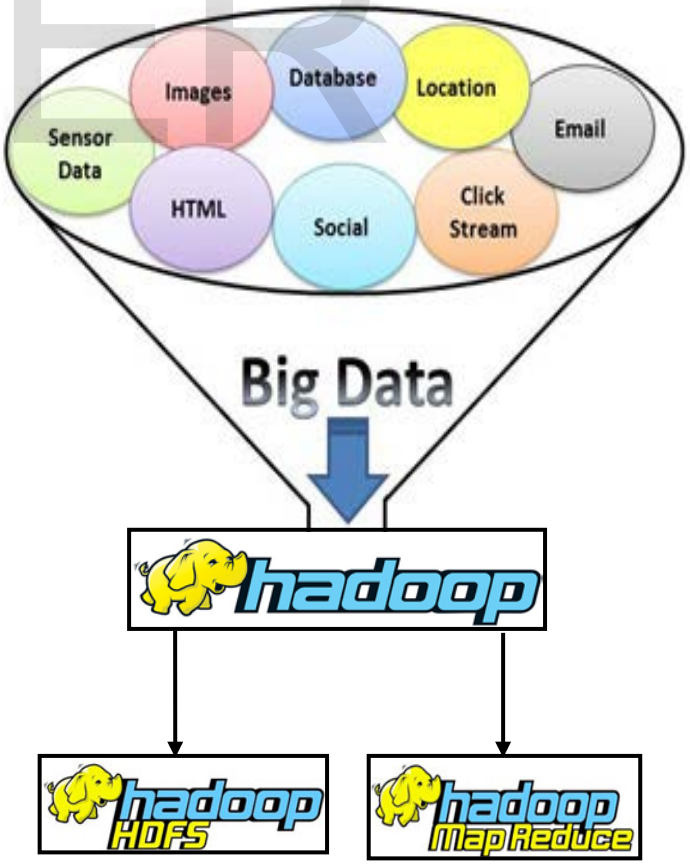


Fig. 3 Big data is the problem and Hadoop is the solution. (With two main techniques) [28][29][30][31]

### 3.1 Hadoop Distributed File System

The Apache platform also includes the Hadoop Distributed File System. Basically HDFS which is also used for fault tolerance and scalability. Hadoop distributed file system stores large files by dividing them into small part usually 64 or 128 MB and replicate that parts on three or more servers. In other word we can simply say that Hadoop Distributed File System will split data files into chunks which are managed by different nodes in cluster. [8]

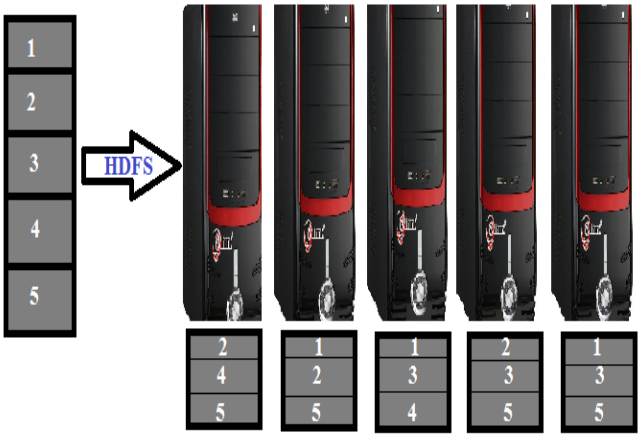


Fig. 4 Apache Hadoop distributed file system

In above figure HDFS breaks incoming files into blocks and store them redundantly in a cluster. Hadoop Distributed File System can also provide an APIs for Map Reduce applications to read and write data in parallel. Performance and capacity can be scaled by adding data nodes, and a single name node technique manages data monitors and data placement server availability. [8] Distributed File System of Hadoop is clusters in production use today reliably hold Petabytes of data on thousands of nodes. Hadoop distributed file system gives a facility to the user for scalable, fault-tolerant storage at low cost of a data. The HDFS software detects and compensate for system or hardware oriented issues, including server failure and disk problems or other physical problem. [5]

### 3.2 Apache Map Reduce

Central to the scalable of Hadoop is the distributed processing framework is called MapReduce. Basically MapReduce help to programmers to resolve data parallel problems for which the data set can be further divided into smaller parts and processed separately. MapReduce it allows ordinary developers, not just those skilled in high-performance computing, to use parallel programming constructs without worrying about the complex details of intra-cluster communication, task monitoring, and failure handling. MapReduce simplifies all that. Mainly the system splits into the input data-set and multiple chunks, all assigned a map task that can process the data in parallel.[8] Each map task reads the input as a set of (value & key) pairs and produces a transformed set of (value & key) pairs as the output. The framework shuffles and sorts outputs of the map tasks, sending the intermediate (value & Key) pairs to the reduce tasks, which group them into final results. MapReduce uses Task Tracker and Job Tracker mechanisms to schedule tasks, monitor them, and restart any that fail. [8]

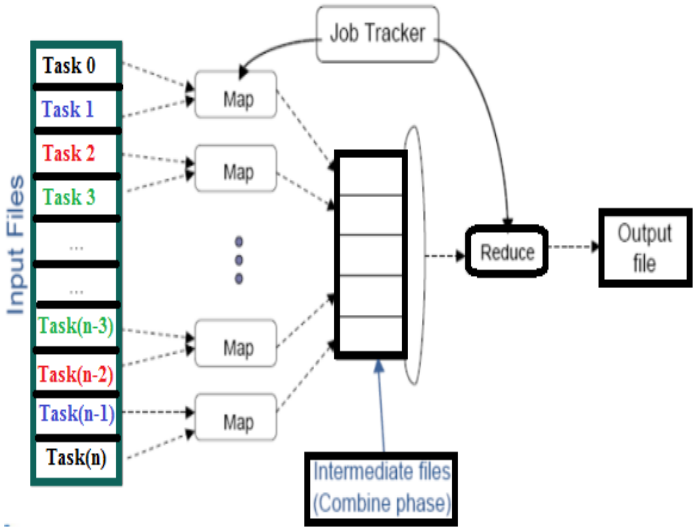


Fig. 5 Apache MapReduce Technique [23]

Mainly the system splits into the input data-set and multiple chunks, all assigned a map task that can process the data in parallel. Each map task reads the input as a set of (value & key) pairs and produces a transformed set of (value & key) pairs as the output. The framework shuffles and sorts outputs of the map tasks, sending the intermediate (value & Key) pairs to the reduce tasks, which group them into final results. MapReduce uses Task Tracker and Job Tracker mechanisms to schedule tasks, monitor them, and restart any that fail.[8] [6]

## 4 Live Challanges and Solution Through Hadoop In Big Companies

In lots of companies are worried about the total amount of data is becoming so large that it is tough to find out the most important type of information. Recently, Companies and Organization have been limited to using subsets of their data, or they were controlled to simplistic analyses because the amount of data over whelmed their processing platforms. There are lots of challenges are against the big data include storage, search, capture, transfer, sharing, analysis, and visualization. [21]

Here we describe the one live example of the Oracle product that used the Hadoop for the part of big data solution and that product is the Oracle MoviePlex. Oracle MoviePlex is an online movie streaming company just like many others online video or movies stores YouTube and many more websites as well they needed the cost effective approach to the big data challenges. Oracle MoviePlex implemented the big data Hadoop platform for better manage their business, identify key opportunities and enhance customer satisfaction as well. [20]

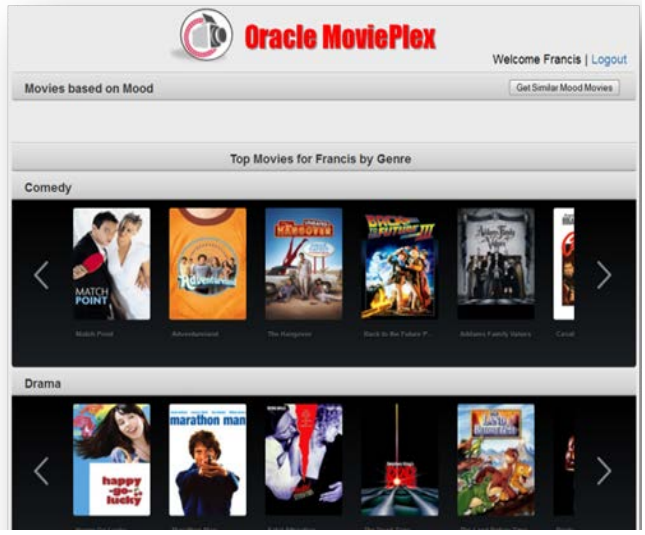


Fig.6 Oracle MoviePlex Application [20]

Oracle MoviePlex applications are generating massive amount of unstructured data that describe user behaviour and application performance. And this huge amount of the data creates a different kind of the issues like a derived data at the right time; better understand the viewing trends of various customer segments?; Bandwidth and streaming problems. That's why the Oracle MoviePlex Used the Oracle Big data platform that is basically used the Apache Hadoop and its different kind of module like a HDFS, Map Reduce, Pig and Hive. [20]

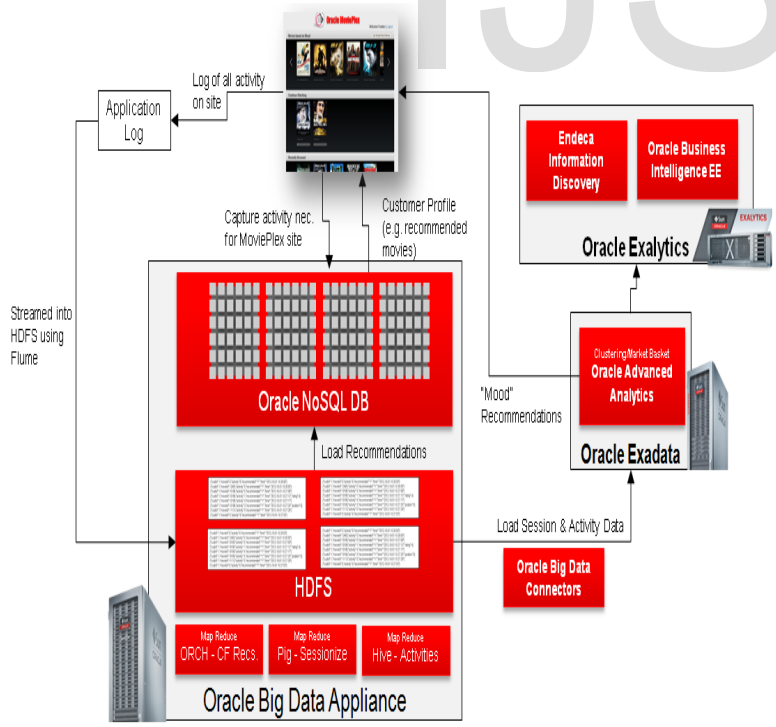


Fig.7 Oracle Big Data Platform [20]

## 5 Hadoop Used By Different Companies

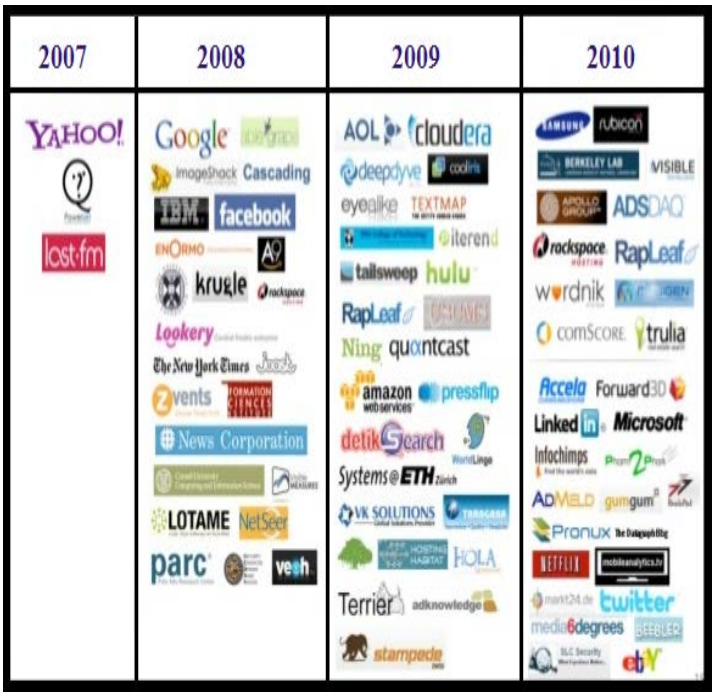


Fig.8 Companies Use Apache Hadoop [22]

## 6 Conclusion

In this paper we well explained the problem as well as the best solution of those problems. Now a day's data is generate in a large amount in different sector and that is called the big data and that create a problems in many Companies, government and private organization, and Science and Technology. The Problem is how can managed or organized this large amount of data and it's a tough task to using data warehouse or database. Best solution for that problem is Apache Hadoop. First of all in this paper we describe the introduction of big data and what are the main problems. Second thing here we describe is Apache Hadoop is the best solution of big data problem. Using some of the modules and we also describe the whole Hadoop Ecosystem and its different parts with live example using Oracle application.

## 5 Acknowledgement

First of all I would sincerely thank to my revered guide, Dr. Hiren Joshi (Associated Professor cum I/C) in School of Computer Science at Dr. Ambedkar Open University Ahmadabad for his valuable guidance and helpful suggestions in this paper. I would also thank to Dr. Tushar Deshai, Head of Doctoral Study at R.K.University Rajkot for motivate me. I wish to express heartiest thanks to my parents, my friends and colleagues for their support, love and inspiration.

## 6 REFERENCES

[1] Konstantin Shvachko, Hairong Kuang, Sanjay Radia, Robert Chansler, "The Hadoop Distributed File System", Shv, Hairong, SRadia, Chansler@Yahoo-Inc.com, IEEE 2010

[2] John Gantz , David Reinsel, "*Extracting Value from Chaos*", IDC IVIEW, June 2011

[3] S. Vikram Phaneendra, E. Madhusudhana Reddy, "*Big Data - Solutions for RDBMS Problems – A Survey*", International Journal of Advanced Research in Computer and Communication Engineering, Vol. 2, Issue 9, ISSN: 2278-1021, SEP 2013

[4] Contributing Authors, "*Big Data Spectrum*", Infosys Limited Bangalore India, 2012

[5] Papineni Rajesh, Y. Madhavi Latha, "*HADOOP the Ultimate Solution for BIG DATA Problems*", IJCTT, Vol-4 Issue-4,April 2013

[6] Azza Abouzeid, Kamil BajdaPawlikowski, Daniel Abadi, Avi Silberschatz, Alexander Rasin (August 2009), "*HadoopDB: An Architectural Hybrid of MapReduce and DBMS Technologies for Analytical Workloads*" VLDB '09, Lyon, France.

[7] Jeffrey Dean, Sanjay Ghemawat, "*MapReduce: Simplified Data Processing on Large Clusters*", Google, Inc. , OSDI 2004

[8] White Paper Big Data Analytics, " *Extract,Transform and Load Big Data with Apache Hadoop*", Intel, 2013

[9] Umesh V. Nikam, Anup W. Burange, Abhishek A. Gulhane, "*Big Data and HADOOP: A Big Game Changer*", International Journal of Advance Research in Computer Science and Management Studies, Volume 1, Issue 7, ISSN: 2321-7782, DEC 2013

[10] Jaya Singh, Ajay Rana, "Exploring The Big Data Spectrum", IJETAE, Vol-3 Issue-4,April 2013

[11] Shilpa, Manjit Kaur, "*BIG Data and Methodology-A review*", International Journal of Advanced Research in Computer Science and Software Engineering(Ijarcsse), Volume 3, Issue 10, ISSN: 2277 128X ,October 2013

[12] Payal Malik, Lipika Bose, "Study and Comparison of Big Data with Relational Approach", International Journal of Advanced Research in Computer Science and Software Engineering, Volume 3, Issue 8, ISSN: 2277 128X, August 2013

[13] http://en.wikipedia.org/wiki/Apache_Hadoop

[14] Tom White," Hadoop: The Definitive Guide", O'Reilly Media, Inc,2009

[15] http://www.guruzon.com/6/introduction/map-reduce

[16] http://hortonworks.com/hadoop

[17] http://hbase.apache.org

[18] http://www.01.ibm.com/software/data/infosphere/hadoop

[19] http://blog.enablecloud.com/2012/06/what-lies-at-core-of-hadoop.html

[20] https://oraclecn.activeevents.com/connect/fileDownload/session/4617511313E404EE4C4B45AF6F1DFEC2/CON1476_Gubar.ppt.pptx

[21] http://en.wikipedia.org/wiki/Big_data

[22] http://www.snia.org/sites/default/education/tutorials/2013/spring/big/RobPeglar_Introduction_to_Analytics_and_Big_Data_Hadoop.pdf

[23] http://institute.lanl.gov/isti/irhpit/projects/

[24] http://evbdn.eventbrite.com/s3-s3/eventlogos/114011/mediaonmobile.jpg

[25] http://www.techywood.com/wp-content/uploads/2013/12/wpid-social-media-logos_15773.png

[26] http://uwaterloo.ca/institute-nanotechnology/sites/ca.institute-nanotechnology/files/resize/uploads/images/WatLAB-500x332.jpg

[27] http://dsn.east.isi.edu/images/soldier_sm.jpg

[28] http://www.scaledb.com/images/big_data.jpg

[29] http://readwrite.com/files/_hadoopelephant_rgb1.png

[30] http://blogs.mulesoft.org/wp-content/uploads/2013/05/hdfs-logo-square.jpg

[31] http://www.rtgx.com/images/sub/solutions/bigdata/mapreduce-logo.jpg

[32] http://www.cloudera.com/content/cloudera/en/resources/library/aboutcloudera/beyond-batch-the-evolution-of-the-hadoop-ecosystem-doug-cutting-video Video explain doug-cutting Hadoop ecosystem