International Journal of Advance Research in Science and Engineering
Vol. No. 5, Issue No. 07 , July 2016
www.ijarse.com
IJARSE
ISSN 2319 - 8354

# BIG DATA AND CLOUD COMPUTING

Sunita Sharma

*BCIIT, New Delhi (India)*

**ABSTRACT**

*The Term Big data burst upon scene in first decade on 21st century. Big data is a technique that is used to handle large volume of data. This Technique is used store, manage, and analyze very high velocity of data and this data can be in any form structured or unstructured form. It is difficult to process large volume of data using data base technique like RDBMS. Main challenge is data storage, Transfer, querying, visualization. Analyze and information privacy. There are so many methods to tackle big data.*

***Key words: Hadoop, HDFS***

## I INTRODUCTION

Big data means really a big data; it is a collection of large datasets that cannot be processed using traditional computing techniques. Big data is not merely a data; rather it has become a complete subject, which involves various tools, techniques and frameworks. 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. 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.

## II BIG DATA CHALLENGE

There are so many major challenges with this term. Analysis of data, Capturing data and data, storing the data, transfer, sharing and presentation of data

## III TOOLS AND TECHNOLOGY

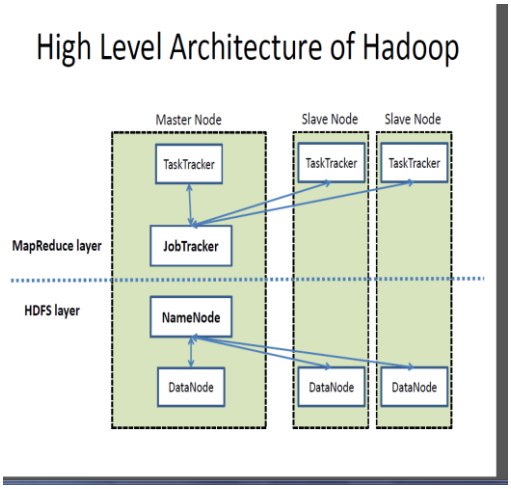
Hadoop is a free, Java-based programming framework that supports the processing of large data sets in a distributed computing environment. It is part of the Apache project There are so many major challenges with this term. Analysis of data, Capturing data and data, storing the data, transfer, sharing and presentation of data. sponsored by the Apache Software Foundation. This approach is used for to scale up from single source or can say single server to multiple machines. Foe each there will be a local area for computation and storage. It does not depend on hardware to deliver high availability services.

### 3.1 HDFS Architecture

Hadoop is a fault-tolerant system and it is called the Hadoop Distributed File System, or HDFS. HDFS is able to store huge amounts of information, scale up incrementally and survive the failure of significant parts of the storage infrastructure without losing data. Hadoop is based on Master /Slave architecture. And this architecture is used for computational and distributed storage. In distributed storage name node work as master and data node play role of slave. In same way Job seeker and task seeker is master slave correspondingly in distributed computing. This can work with any distributed file system directly. Hadoop Distributed File System (HDFS) is most

Common file used by Hadoop.

But the main problem with this technology is that it requires a lot of bandwidth for computations.




### 3.2 Cloud Computing

Cloud computing is a platform that enables on-demand access to computing and storage of all data and resources on internet. Customer no needs to take overhead. All responsibility of Third party service provider. companies offering cloud services Clients can easily outsource large amounts of data and computation to remote locations, as well as run applications directly from the cloud resources who are unwilling or unable to procure and maintain their own computing infrastructure. The ever increasing need for computing power and storage.

## IV MODEL BUILDING AND SCORING

The data storage and Data as a Service (DaaS) capabilities provided by Clouds are important, but for analytics, it is equally relevant to use the data to build models that can be utilised for forecasts and prescriptions. Moreover, as models are built based on the available data, they need to be tested against new data in order to evaluate their ability to forecast future behaviour. Termed here as model building and scoring – to Cloud

providers and ways to parallelise certain machine learning algorithms in the Cloud and exposed via Web Services interfaces. Users can access the models with Web browser technologies to compose their data mining solutions.

data analysis and model building that can run either on a customer's premises or be allocated as SaaS using Infrastructure as a Service (IaaS) provided by solutions such as Amazon EC2 and IBM Smart Cloud Enterprise Google Prediction API allows users to create machine learning models to predict numeric values for a new item based on values of previously submitted training data or predict a category that best describes an item. The prediction API allows users to submit training data as comma separated files following certain conventions, create models, share their models or use models that others shared. With the Google Prediction API, users can develop applications to perform analytics tasks such as sentiment analysis purchase prediction, provide recommendations, analyse churn, and detect spam. The Apache Mahout project ] aims to provide tools to build scalable machine learning libraries on top of Hadoop using the MapReduce paradigm. The provided libraries can be deployed on a Cloud and be explored to build solutions that require clustering, recommendation mining, document categorisation, among others.

## V OTHER DIFFERENT COMPONENT

**HBase:** It is written in Java and modelled after Google's Big Table. HBase is an example of a No SQL data store. This method use distributed cluster to store and process big data.

**Hive**: Its name is HIVE QL. Basically it is used for SQL access with the help of Map reduces.

**Cascading: When** size of data exceeds and further complexity is also increased in data. To solve that problem Cascading technique is used.

**Avro:** Main purpose of use of Avro for data serialization. Avro can store data definition and data simultaneously into a one message which helps to dynamically understand the big data.

**Big Top**: It is used for packaging and testing the Hadoop ecosystem.

**Oozie**: Two basic jobs of oozie are oozie workflow and oozie coordinator. This tool provides grater control over jobs.

## VI CONCLUSIONS AND FUTURE WORK

Big Data is very large data set that has highly volume, diversity, and complexity requires new architecture, techniques, algorithms, and analytics to manage it and extract value and hidden knowledge from it. In today time, Data can be s generated from various different sources. It is a very big issue to process these large amounts of data today. We discussed Hadoop tool for big data in this paper. Hadoop is the core platform for structuring Big Data, and solves the problem of making it useful for analytics purposes. In this paper we discussed some hadoop components. Components are helped out to support the processing of large data sets in distributed

computing environments. In future for better result we can use some others clustering techniques and enhance the performance.

## REFERENCES

[1] Yingyi Bu _ Bill Howe _ Magdalena Balazinska _ Michael D. Ernst **"The HaLoop Approach to Large-Scale Iterative Data Analysis"** VLDB 2010 paper "HaLoop: Efficient Iterative Data Processing on Large Clusters.

[2] S.Vikram Phaneendra & E.Madhusudhan Reddy **"Big Data- solutions for RDBMS problems-A Survey"** In 12th IEEE/IFIP Network Operations & Management Symposium (NOMS 2010) (Osaka, Japan, Apr 19{23 2013).

[3] Harshawardhan S. Bhosale1 , Prof. Devendra P. Gadekar2 "A Review Paper on Big Data and Hadoop" International Journal of Scientific and Research Publications, Volume 4, Issue 10, October 2014 1 ISSN 2250-3153 www.ijsrp.org A

[4] Rotsnarani Sethy , Mrutyunjaya Panda "Big Data Analysis using Hadoop: A Survey" nternational Journal of Advanced Research in Computer Science and Software Engineering Volume 5, Issue 7, July 2015

[5] Albert Bifet **"Mining Big Data In Real Time"**Informatica 37 (2013) 15–20 DEC 2012.

[6] R. Saraswathy, P. Priyadharshini, P. Sandeepa "HBase Cloud Research Architecture for Large Scale Image Processing" International Journal of Advanced Research in Computer Science and Software Engineering Volume 4, Issue 12, December 2014

[7] Albert Bifet **"Mining Big Data In Real Time"**Informatica 37 (2013) 15–20 DEC 2012.

[8] K.Arun, Dr.L.Jabasheela "Big Data: Review, Classification and Analysis Survey" International Journal of Innovative Research in Information Security (IJIRIS) ISSN: 2349-7017(O) Volume 1 Issue 3 (September 2014) ISSN: 2349-7009(P)

[9] https://hadoop.apache.org/

[10] https://hadooptutorial.wikispaces.com/Hadoop+architecture

[11] Bernice Purcell **"The emergence of "big data" technology and analytics"** Journal of Technology Research 2013. 1994 2/13/04