International Journal of Modern Trends in Engineering and Research
www.ijmter.com
e-ISSN No.:2349-9745, Date: 2-4 July, 2015

A Review Approach for Big Data and Hadoop Technology

Prof. Ghanshyam Dhomse[1], Ms. Katkade Komal[2], Ms. Lunawat Manali[3], Ms. Abad Latika[4]
[1]*Department of computer Engineering, SNJB's KBJ COE, Chandwad*
[2]*Department of computer Engineering, SNJB's KBJ COE, Chandwad*
[3]*Department of computer Engineering, SNJB's KBJ COE, Chandwad*
[4]*Department of computer Engineering, SNJB's KBJ COE, Chandwad*

Abstract— This paper is the review paper on Big Data and Hadoop Technology which will give us the summary of various components of Hadoop framework. Facebook, twitter and many other applications randomly generate larger amount of data. So there is need of larger framework to handle such kind of data sets. Therefore Big Data and Hadoop Technology came into picture towards research direction. Hadoop and Big Data are very important not only to increase productivity and growth of industries but also used in various sectors like reveal patterns, trends and association , especially related to human behavior and interaction. Big data refers to the different types of data which ranges in Zetabyte and beyond. Therefore in this paper, we provide a brief survey of various Hadoop components, Big data analytics research, while highlighting the specific concerns in Big data world. We present a taxonomy based on the key issues in this area, and discuss the different methods to tackle these issues. Based on this survey study many midmarket organizations report a need for tools ranging from real-time processing to predictive analytics, data cleansing, and data visualization.

Keywords: Big Data, Parameters, Evolution, Hadoop, HDFS, MapReduce, Hive.

I. INTRODUCTION

One of the biggest new ideas related to computing in 21st century in "BIG DATA". Every day we create 2.5 quintillions of data and 90% of this data has been created from last two years alone. This is Big Data.

Big Data can be defined as large data sets that may be analyzed computationally to reveal patterns, trends and associations, especially relating to human behavior and interactions. Big Data is voluminous amount of structured, semi-structured and unstructured data used to store a very huge quantity of data that is often measured in Exabyte and Petabyte.

Big Data consist of three V's:

1) Volume (amount of data)
2) Variety(range of data type and sources)
3) Velocity(speed of data in and out)

This was firstly defined by Doug Laney of Gartner twelve years ago. They are extremely difficult to maintain and manipulate using common database management tools.

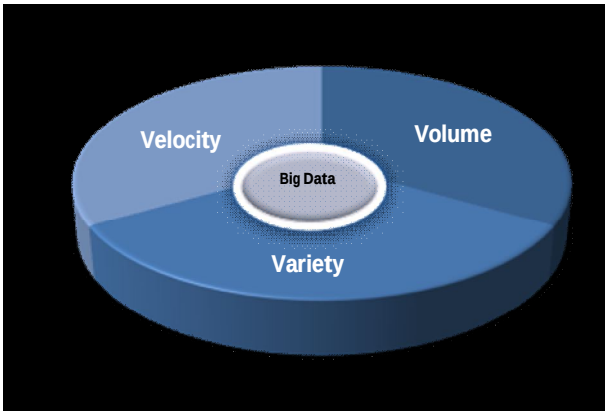

Fig 1: Three V's factor of Big Data

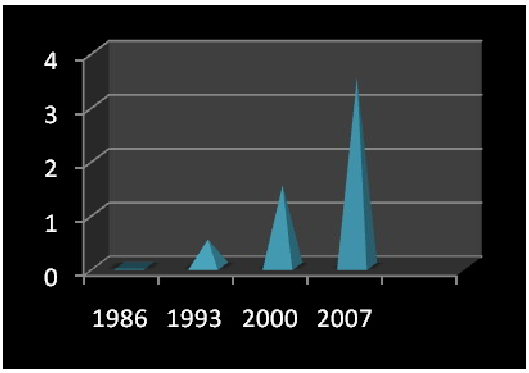

Fig 2: Growth of Big Data

II. HADOOP

Apache Hadoop is an open source software framework written in java. It is an open source project governed by the Apache Software Foundation (ASF). Hadoop was created by Doug Cutting and Mike Cafarella in 2005. It is used for distributed storage and distributed processing of large data sets. Hadoop's code is mostly written in Java and some of its native code is written in C with command utilities as shell scripts.

The modules of base Apache Hadoop framework are as follows:

1) Hadoop Common- contains libraries and utilities which are needed by other Hadoop modules.
2) Hadoop Distributed File System (HDFS) –_a distributed_file system that stores data on commodity machines and also provides high aggregate bandwidth across the cluster.
3) Hadoop YARN-_It is a resource management platform responsible for managing compute resources in clusters and using them from scheduling of users' applications.
4) Hadoop MapReduce-_a programming model for large scale data processing.

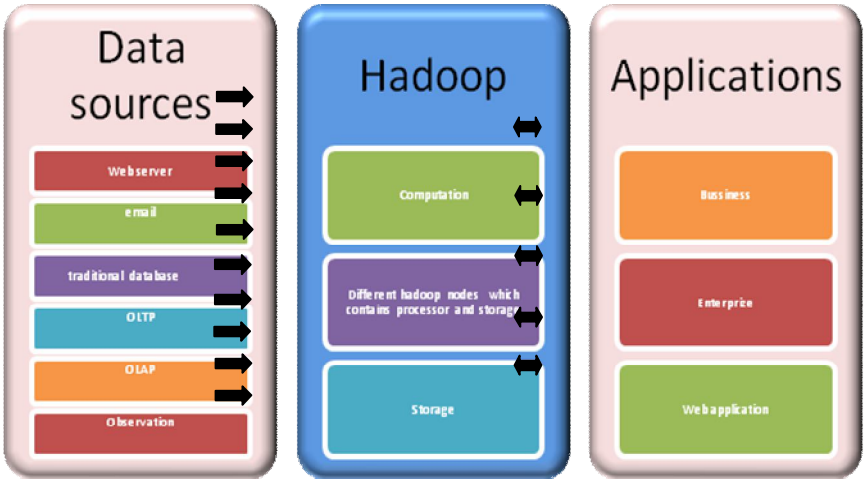

Fig 3: Hadoop Framework

2.1 HDFS

HDFS is nothing but Hadoop Distributed File System. HDFS is designed to handle large data sets and also for applications where large bandwidth is required. It is java based. It is open source. It provides scalable and reliable data storage. HDFS has demonstrated scalability of up to 200 PB of storage and a single cluster of 4500 servers, supporting close to a billion files and blocks. It follows master-slave architecture. It is fault tolerant. It distributes storage and computation across many servers. This combined storage resource can grow with demand .The Hadoop clusters are available all the time and are functional. It stores files and directories in hierarchical manner.

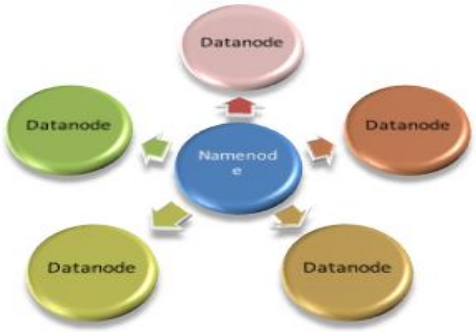

Fig 4: HDFS

HDFS cluster is comprised of NameNode. HDFS consist of single NameNode and multiple DataNodes. NameNode acts as a master and DataNodes acts as slave nodes. DataNodes are connected to NameNode. At the start, NameNode performs handshake operation with DataNodes to establish a connection. Each DataNode has a unique storage ID. Files and directories are represented on a NameNode. Each DataNode is assigned to one cluster. NameNode execute file related operations like opening and closing file. Following are some of the features of HDFS:

1) Rack Awareness
2) Minimal Data Motion
3) Utilities
4) Rollback

2.2 MapReduce

It is a popular open source implementation. MapReduce is the heart of Hadoop. It is a programming model. It is used for processing and generating large datasets. This is done with the help of parallel and distributed algorithm on a cluster. A MapReduce program is composed of map () procedure and reduce () procedure.

map (): Performs filtering and sorting.

reduce (): Performs summary operations

The map jobs take a data and convert it into another set of data. The reduce job takes input from the output of map job and convert it into smaller set of tuples.

It operates on a <key, value> pair.

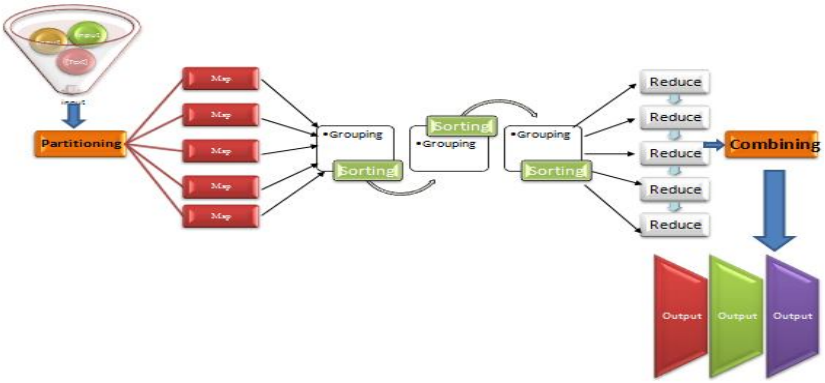

Fig 5: Mapping and Reducing

III.HIVE

Apache Hive is a data warehouse infrastructure .It is built on the top of Hadoop. It provides query optimization, query and analysis. It was initially developed by facebook. Later it was developed by Apache software Foundation. By default Hive stores all its metadata in an embedded Apache Derby Database. It provides SQL like language called as HiveQL with schema on read. It converts queries to map/reduce. HiveQL supports ACID functionality. Features of HiveQL:

- Equijoins between tables.
- Facilitates another table to store result of query.
- Managing tables and partitions.

Hive makes querying and analyzing easy. It can also be called as a platform to develop SQL type scripts to do MapReduce operations. Hive is not a relational database. For storing data, Hive data models are:

- Database
- Table
- Partition
- Buckets

Database is a collection of tables. Tables are like relational database tables. All the contents of tables are store in HDFS. Table shows what data is stored. Tables can have a multiple Partitions. Partition shows how data is stored.

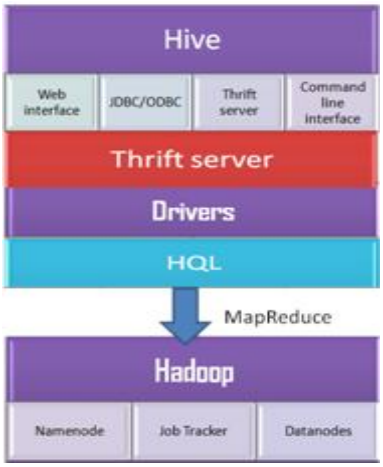

Fig 6: Hive

Data in Hive is stored in different file formats.

Following are some of the features of Hive:

1) It stores schemas in database.
2) It stores processed data into HDFS.
3) It is designed for OLAP.
4) It is fast and scalable.
5) It is extensible.

CONCLUSIONS

As Big Data is emerging on a wide scale in corporate, industries, universities, etc, it is the need to store, process and maintain the security and integrity of data. This huge amount of data cannot be processed efficiently with help of traditional database system.

BIBLIOGRAPHY

Ms. Lunawat Manali: Student of third year Computer Engineering, Shri Neminath Jain Bhramhacharya's Kantabai Bhavarlalji Jain College of Engineering, Chandwad.

Ms. Abad Latika: Student of third year Computer Engineering, Shri Neminath Jain Bhramhacharya's Kantabai Bhavarlalji Jain College of Engineering, Chandwad.

Ms. Katkade Komal: Student of third year Computer Engineering, Shri Neminath Jain Bhramhacharya's Kantabai Bhavarlalji Jain College of Engineering, Chandwad (Pune University)

REFERENCES

[1.] http://www.slideshare.net/
[2.] http://hadoop.apache.org/
[3.] http://en.wikipedia.org/wiki/Hortonworks
[4.] Big Data Analytics with R and Hadoop-Vignesh Prajapati
[5] By Michael Schroeck, Rebecca Shockley, Dr. Janet Smart, Professor Dolores Romero-Morales and Professor Peter Tufano –Real use of Big Data

