Scientific Journal of Impact Factor (SJIF): 4.72

e-ISSN (O): 2348-4470
p-ISSN (P): 2348-6406

International Journal of Advance Engineering and Research Development

Volume 4, Issue 8, August -2017

"A study based on Cloudera's distribution of Hadoop technologies for big data"

Ms. Shalini Khokad[1]
B.Tech CSE Student, Department of Computer Science & Engineering,
Dr. BAMU, Maharashtra Institute of Technology, Aurangabad (MS), India

Ms. Gauri Bhalerao[2]
B.Tech CSE Student, Department of Computer Science & Engineering,
Dr. BAMU, Maharashtra Institute of Technology, Aurangabad (MS), India

Prof. Daivashala Deshmukh[3]
Assistant Professor, Department of Computer Science & Engineering,
Dr. BAMU, Maharashtra Institute of Technology, Aurangabad (MS), India

Abstract :-

These last years, the new technology are producing a large quantities of data i.e Big data. Companies are faced with certain problems of collecting, storing, analyzing and exploiting these large volumes of data in order to create the added value. The complete issue, for organizations and administrations, is not to pass by valuable information drowned in the mass. It is here where the technology named as the "Big Data" intervenes. This technology is based on an analysis of very fine large number of data. It is interesting to note that there are several Organizations who offer distributions ready to use for managing a system Big Data namely Hortonworks , Cloudera, MapR, etc. The different distributions have an approach and a different positioning in relation to the vision of a platform Hadoop. These solutions are the termed as Apache Projects and therefore available in now-a-days. Yet, the interest of a complete package lies in the compatibility between these components, the simplicity of installation as well as support. In this article, we shall focus the era of big data by defining these characteristics and its architecture. Then we shall talk about Cloudera Distribution for Hadoop Platform, and finally, we shall conclude by a study on the tools of Hadoop distributions of Big Data provided by Cloudera.

Keywords- *Big Data, 5 V's, Distribution framework, Hadoop, Analysis.*

I. INTRODUCTION

Nowadays data is being generated with very high rate from different areas like business, specific data, emails, blogs, etc. To analyze and process this large amount of data and to extract information for users there is need of deploying data intensive application and storage clusters.

Data sets so large and complex that it becomes difficult or impossible to process them using traditional database management applications .Relational database such as Data Ware House to Business Intelligence now we are believing and thinking one more level above because we are experiencing unexpected growth in structured and unstructured data is very high. Principally big data exceeds the processing capacity of traditional database systems whether data is too big, moves too fast, or, doesn't fit in the present structure of your database architectures. The most popular choice for Big Data software stack is Hadoop.

II. BIG DATA: BIG IMPACT

A. Definition

Concept of Big data is huge for data sets that traditional data processing application software is incapable to handle with them. Some challenges related to big data includes data capturing, storage, analysis on data, search, sharing, transfer, visualization, querying, updating and maintaining information privacy content.

The term "big data" refer to the use of predictive analytics, user behaviour analytics, or certain other advanced data analytics methods that extract value from data, and sometimes to a particular size of data set. "There is

little doubt that the amount of data now available is indeed large, but that's not the most relevant characteristic of this new data ecosystem"

B. **Characteristics of Big Data (5 Vs. Volume velocity variety veracity value)**

a. Volume - Volume describes the amount of data generated by organizations or individuals.
b. Velocity - Velocity describes the frequency at which data is generated, captured and shared.
c. Variety - Big data means much more than rows and columns. It means unstructured text, video, audio that have important impact.
d. Veracity - the quality and understandability of the data
e. Value – Business value to be derived.

C. Examples of Big Data:

An example of big data may be petabytes or exabytes of data collected from different sources (e.g. Web, sales, customer contact center, social media, mobile data etc.) consisting of billions to trillions of records of millions of people. The data is typically loosely structured data that is often incomplete and inaccessible.

1. eBay uses two data warehouses at 7.5 petabytes and 40PB Hadoop cluster for search, consumer recommendations, and merchandising operations.

2. Walmart handles more than 1 million customer transactions every hour, which are imported into databases estimated to contain more than 2.5 petabytes (2560 terabytes) of data – the equivalent of 167 times the information contained in all the books in the US Library of Congress.

III. BIG DATA HADOOP DISTRIBUTION

a. Cloudera

Big data geniuses from Facebook, Google, Oracle and Yahoo in 2008 established Cloudera .Inc. It was the first leading company to develop and distribute Apache Hadoop-based application software and still has the biggest user support with greater number of clients. Although the core of the distribution is based on Apache Hadoop, providing a proprietary Cloudera Management Suite for automating the installation process as well as providing other services to enhance convenience of users. Making convenient to user reduces deployment time, displays real time nodes count, etc.

b. Hortonworks

The leading vendors of Hadoop, which has quicky emerged known as Hortonworks, founded in 2011. The Cloudera distribution provides open source platform based on Apache Hadoop for performing various tasks such as analysing, storing and managing big data.The only commercial vendor to distribute complete open source Apache Hadoop without any additional proprietary software is Hortanworks. HDP2.0 , Hortonworks' distribution ,can be directly downloaded from their website free of cost and it is simple and easy to install. The engineers of Hortonworks has contributed to most of Hadoop's recent major innovations including YARN, which is better than MapReduce programming paradigm in the sense that it enables involvement of more data processing frameworks.

c. MapR

All three top Hadoop distributions, Cloudera, MapR and Hortonworks offer various services like consulting, training, and technical assistance. But unlike its two opponents, Hortonworks' distribution is claimed to be total 100 percent open source. Cloudera incorporates a collection of proprietary softwares in its Enterprise 4.0 version, with additional layers of administration and management capabilities to the core Hadoop software framework. Moving ahead, MapR replaces HDFS component and instead of using its own proprietary file system, called MapRFS. MapRFS helps to incorporate enterprise-grade features into Hadoop, enabling more efficient management of data, reliability and ease of use. In other worlds, it is more production ready than its other two competitor distributors.

IV. BIG DATA DISTRIBUTION ARCHITECTURE OR CDH

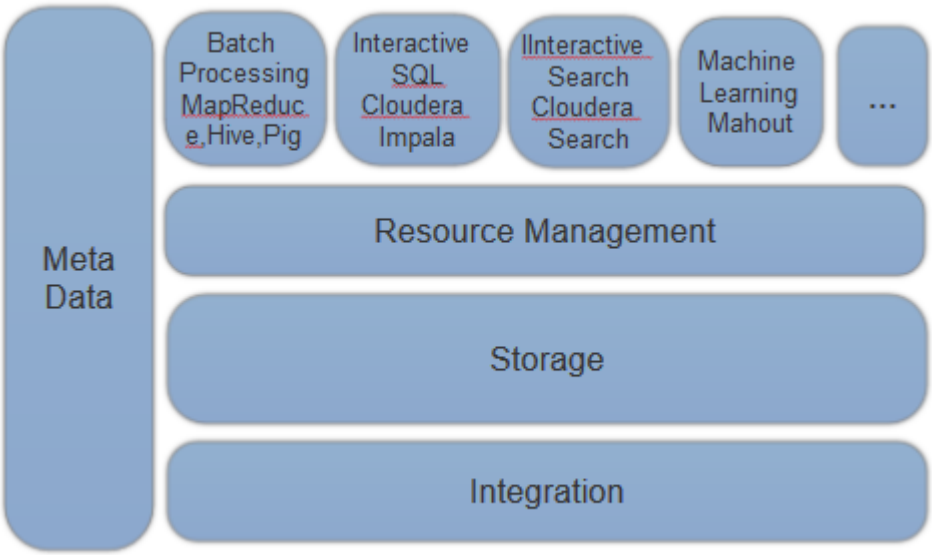

Figure 1. Cloudera Distribution for Hadoop Platform (CDH)

MapReduce

Map Reduce is a distributed computing program model used for processing technique based on java. The Map Reduce algorithm contains two essential tasks, called as Map and Reduce. Map task takes a set of data and transform it into other set of data, where respective elements are segment into key or value pairs called as tuples. Second, reduce task, that takes an output from a map as an input information and merge those data tuples into a smaller set of elements. As the sequence of the name Map and Reduce implies, the reduce job is must performed after the completion of map task.

The main advantage of Map Reduce is easy to range data processing over various different computing nodes. In the Map Reduce technique, the data processing fundamental components are known as mappers and reducers. Segmenting a data processing technique or method into mappers and reducers term as it is somewhat nontrivial. But, once we perform an application in the Map Reduce form, scaling the application to run over thousands of machines in a cluster is hardly a configuration change.

HDFS

A file system that collect all the nodes in a Hadoop cluster for data storage. It connects together the file system on many local nodes to make them into one large file system. HDFS handles nodes will fail, so it achieves reliability by duplicating data across multiple nodes. HDFS is a Hadoop distributed file system based on java that provides not only scalable but also reliable data storage, and it was designed to extent big cluster of servers. Once the quantity and quality of enterprise data or processed information is accessible in Hadoop Distributed File System, and YARN gives permission to multiple data access applications to process it.

As it is a scalable, fault-tolerant, distributed storage file system that co-ordinates closely with a huge variety of serialize data access applications, arranged by YARN. By disseminating storage and computation across many servers, the combined storage resource can grow linearly with demand it remains economical at every amount of storage. The file contains information is distributed into large blocks in gigabytes, and each block of the file i.e. data is independently copied at multiple DataNodes. The blocks are loaded as on the local file system on the DataNodes. The Namenode actively monitors the number of copies of a block containing information. When a replica of a block is lost due to a DataNode failure or disk failure, the NameNode creates another replica of the block. The NameNode maintaining elements as the namespace tree and the mapping of blocks to DataNodes, holding the whole namespace image in

Random access memory for processing. The NameNode does not directly send requests to DataNodes. It provides guidelines for the DataNodes by sending to heartbeats received by those DataNodes.

a) Sqoop (Import and Export)

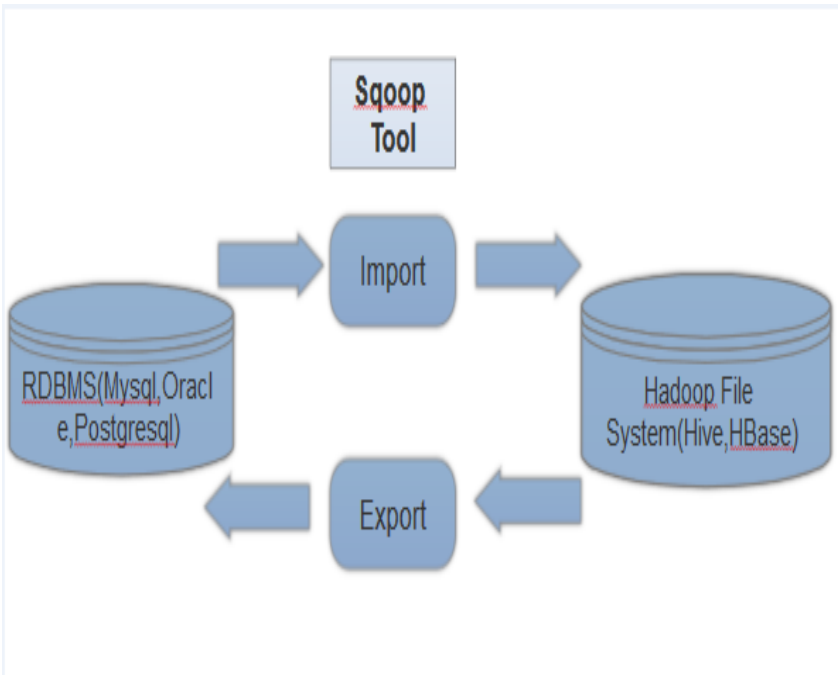

Figure 2. Sqoop Working

"SQL to Hadoop and Hadoop to SQL"

Sqoop is a hadoop tool designed to transmit data to Hadoop and relational database servers. It is used to import data from relational databases such Oracle to Hadoop HDFS, MYSQL and export from Hadoop distributed file system to relational databases. It is provided by the Apache Cloudera Software Foundation.

From Hadoop for analytics and data processing technique requires storing data into clusters and processing it in similar with other data that often resides in relational databases across the enterprise.

Sqoop tool used as the easy importation and exportation of data or information from database which is in the structured form such as relational databases, enterprise data warehouses, data intelligence, data mining and NoSQL systems. From the use of Sqoop tool, we can arrange the data from external system on to HDFS, and tables in database Hive and HBase. Sqoop accommodate with Oozie tool as the cloudera distribution, allowing the schedule and automate import and export jobs. Sqoop tool uses a connection based architecture with the JDBC/ODBC which supports plugins with the device or systems that allows connectivity to new external systems.

In this command, the above choice given are as follows:

import: This is the sub-command that instructs Sqoop to initiate an import.

–connect <connect string>, –username <user name>, –password <password>: These are connection components that are given as to establish connection with the database. This is no different from the connection components that uses when connecting to the database through a JDBC connection.

–table <table name>: This component gives the table information which will be imported.

b) Pig (Express Data Flow)

Pig is a high-level scripting language that is used with Apache Hadoop framework. Pig tool facilitate data workers to write complicated data transforms without knowing the Java program as base. Pig's simple SQL-like scripting language is called Pig Latin, and that uses application to developers already knowing the scripting languages and SQL Structured Query Language.

Pig tool is complete, so we can do all required data manipulations in Apache Hadoop framework with Pig. Through the User Defined Functions(UDF) efficiency in Pig, Pig can request code in many more languages like Ruby, Python and Java. We can enclose Pig script languages or files in other languages. The result is that we can use Pig tool as

a component to build huge and major complex applications that implements real business problems or tasks.P is works with data information from many sources, including structured and unstructured data, and loads the results into the Hadoop Data Distributor File System. Pig scripts are transformed into a sequence of MapReduce tasks that are run on the Apache Hadoop cluster.

c) Hive (Batch Processing)

Hive on LLAP (Live Long and Process) uses equivalent multiple query servers with intelligent in-memory caching to avoid Hadoop's batch-oriented latency and provide as fast as sub-second query response times across smaller data volumes, while Hive on Tez application framework continues to provide excellent batch query performance on petabyte-scale data sets.

The Hive's table are equivalent to tables in a relational database system. Data unit are arranged in a hierarchical form from larger to more granular units. Tables present in databases are made up of partitions. Simple query language is used to access stored data. Hive supports operations on data such as overwriting or appending.

In a database, tables comprising of data are serialized and each table has a corresponding directory known as Hadoop Distributed File System (HDFS). Each table can be divide into sub-partitions by determining how data is distributed within multiple sub-directories of the table directory. Partitioned data can be further broken down in form of buckets.

All the common fundamental data formats type such as BIGINT, BINARY, BOOLEAN, CHAR, DECIMAL, DOUBLE, FLOAT, INT, SMALLINT, STRING, TIMESTAMP, and TINYINT are supported by Hive apache Hadoop tool. Additionally, the fact is that, by the analyst's complex data types are formed, that are structs, maps and arrays by combining all other fundamental data types.

d) Impala (SQL syntax + compute Framework)

Cloudera Impala is an open source Massively Parallel Processing (MPP) query engine that runs natively on Apache Hadoop. Built for performance, Impala uses in memory data transfers with its native query engine allowing users to issue SQL queries against HDFS and receive results in seconds. Impala is combined with Hadoop to use the same file and data formats, metadata, security and resource management frameworks which are used by MapReduce, Apache Hive, Apache Pig and other Hadoop software. Impala promotes data analysts and data scientists to perform analytics on data stored in Hadoop via SQL queries or business intelligence tools. Its output is large-scale data processing using MapReduce technique and interactive SQL database queries can be done on the same system using the methods as for same data and metadata – removing the applications to transmit data sets into specialized concurrent systems and proprietary formats used to perform analysis of the information. Features include: Supports HDFS and Apache HBase storage, Reads Hadoop file formats, including text, Sequence File, Avro, RCFile, and Parquet, Supports Hadoop security (Kerberos authentication), Fine-grained, role-based authorization with Apache Sentry, Uses metadata, ODBC driver, and SQL syntax from Apache Hive.

e) Hue (Hadoop user interface)

Hue is the open source analytics tool workbench arranged for the processed data discovery, new updated query operations, as well as consistent collaboration. It bridges the gap between IT organization and the business for authenticated self-service analytics.

It is web-based interface making Hadoop easier to use, helping you through by providing graphical user interface in internet browser -instead of logging into a Hadoop gateway host with a terminal program and using the command line prompt.

Applications for HUE are usually implemented in Django, a popular MVC web framework that understands the application namespaces. At top level, the Software Development Kit lets the application bundle start and helps daemons which might talk to various interfaces in Hadoop, HDFS, or one of the numerous other applications dispatched with CDH.

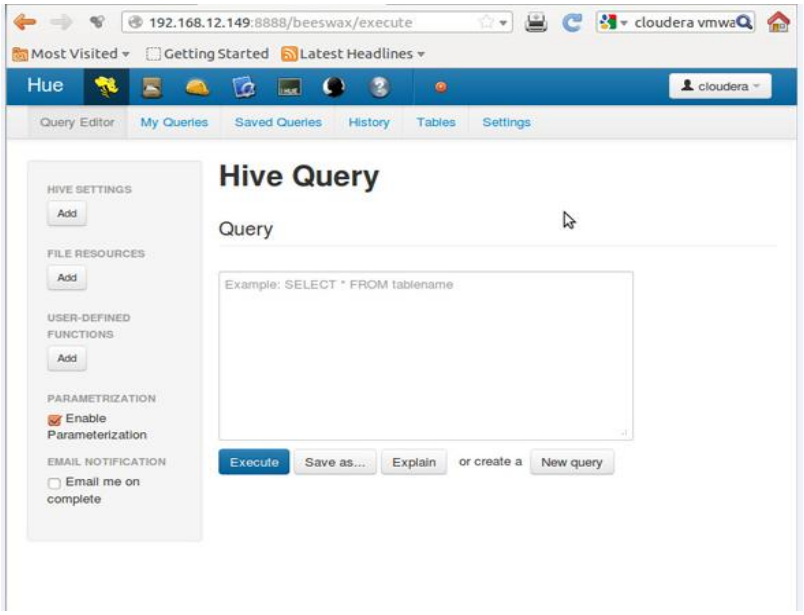

Figure 3. Hive query editor

f) Oozie (Workflow Scheduler):

Apache Oozie is a workflow and coordination service for managing Apache Hadoop tasks:

- Oozie Workflow jobs are Directed Acyclic Graphs (DAGs) of *actions*; which are typically tasks performed by Hadoop such as MapReduce, Streaming, Pig, Hive, Sqoop, etc.
- Oozie Coordinator triggers concurrent Workflow jobs based on time (frequency) and data availability.
- Oozie Bundle jobs are collection of Coordinator task that is managed as a single job.

- Oozie is an extensible, scalable and data-aware service which is used to manage dependencies among jobs that are running on Hadoop.
- Oozie v3: It is a server based Bundle Engine which provides a higher-level abstraction of oozie that will batch a collection of coordinator applications. The user will be able to perform various operations such as start, stop, suspend, resume, rerun a set coordinator jobs in the bundle level resulting into a better and easy operational control.
- Oozie v2: It is a server based Coordinator Engine which is specialized in workflows running based on time as well as data triggers. It can continuously run workflows based on time (e.g. run it every hour or day or week or yearly), and data availability (e.g. waiting for input data to exist before running any workflow).
- Oozie v1: It is a server based Workflow Engine which is specialized in running workflow jobs with actions that execute MapReduce and Pig jobs in Hadoop.

V. CONCLUSION

The Big Data is a concept popularized in recent years to translate the fact that companies are faced with large volumes of data to handle gradually and considerably while presenting a high-stake at the commercial level and marketing. This trend around the collection and analysis of Big Data has given birth to new solutions which combine classic technologies of data warehouse to systems Big Data in a logical architecture. Besides, as there are several distributions that can help to facilitate the adoption of the Platform Hadoop of Apache and manage clusters Cloudera.

HDFS and Map Reduce is scalable and fault tolerant model that hides all complexities for Big Data analytics. The model is built to work efficiently on thousands of machines and massive data sets using commodity hardware.

VI. REFERENCES

[1]. https://www.cloudera.com/products/open-source/apache-hadoop/apache-hive.html
[2]. https://www.experfy.com/blog/cloudera-vs-hortonworks
[3]. Sawant, N., & Shah, H. (Software engineer). (2013). Big data application architecture & A a problem-solution approach. Apress
[4]. Big Data: An Introduction, ARD-IJEET, ISSN- 2320-8821, Volume 5, Issue 1
[5]. Lenovo, I. (2015). Lenovo Big Data Reference Architecture for Cloudera Distribution for Hadoop.
[6]. https://en.wikipedia.org/wiki/Apache_Hive
[7] https://en.wikipedia.org/wiki/Pig_(programming_tool)
[8]. https://impala.apache.org/

VII. AUTHORS PROFILE

1. Ms. Shalini Khokad is pursuing her bachelor's degree from Maharashtra Institute of Technology. She is a student of final year, computer science and engineering department Recently she completed training on Cloudera's Big Data Course from MIT's Big Data Academy.
2. Ms. Gauri Bhalerao is pursuing her bachelor's degree from Maharashtra Institute of Technology. She is a student of final year, computer science and engineering department Recently she completed training on Cloudera's Big Data Course from MIT's Big Data Academy.
3. Prof. Daivashala Deshmukh is an assistant professor in Maharashtra Institute of Technology. Other than academics she is a coordinator and instructor for Big Data Academy in Computer science and Engineering department, Maharashtra Institute of Technology. She completed training of Cloudera's and HortornWork's Big Data Course. Her current research area of interest is Big Data.

