ISSN(Online): 2320-9801
ISSN (Print) : 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Website: www.ijircce.com

Vol. 6, Issue 8, August 2018

Big Data Analytics using Hadoop Technologies: A Study based on CDH for Big Data

Prof. Daivashala R. Deshmukh[1], Shubham Deepak Jadhav[2], Pranav Dattatray Joshi[3], Dnyaneshwari Pawar[4]

Assistant Professor, Department of Computer Science & Engineering, Maharashtra Institute of Technology, Aurangabad (MS), India[1]

B.Tech Student, Department of Computer Science & Engineering, Maharashtra Institute of Technology, Aurangabad (MS), India[2]

B.Tech Student, Department of Computer Science & Engineering, Maharashtra Institute of Technology, Aurangabad (MS), India[3]

B.Tech Student, Department of Computer Science & Engineering, Maharashtra Institute of Technology, Aurangabad (MS), India[4]

ABSTRACT: Big data refers to huge amount of data increasing rapidly by various data sources and different data format. Many organizations are facing with certain problems of collecting, storing, analysing and exploiting these large volumes of data and different formats in order to create the added value. It is here where the technology of the "Big Data" intervenes. This technology is based on an analysis of very fine masses of data. It is interesting to note that there are several publishers who offer distributions ready to use for managing Big Data namely Hortonworks, Cloudera [1], MapR, etc. These different distributions have an approach and a different positioning in relation to the vision towards a platform Hadoop. These solutions are the Apache Projects and therefore available. Yet, the interest of a complete package resides in the compatibility between the components, the simplicity of installation, support, etc. In this paper, we shall discuss the world 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: Hadoop, Big Data Analysis, HDFS, MapReduce

I. INTRDUCTION

Nowadays data is being generated with very high rate from different areas like social networking sites, business, emails, blogs, etc. To analyse and process this large amount of data and to extract information for usersthere is a need of deploying data intensive application and storage clusters.
Data sets are so large and complex that it becomes difficult or impossible to process those using traditional database management applications.

A. Major challenges with Big Data:

The first challenge is storing the BIGGER amount of data- Storing huge data in a traditional database system is not possible, because of the storage limitation and tremendous increase rate of data. The second challenge is storing heterogeneous data - The data is not only huge, but it is also present in various formats i.e. unstructured, semi-structured and structured. So, you need to make sure that you have a system to store different types of data that is generated from various sources.The third problem, which is the processing speed - Now the time taken to process this

ISSN(Online): 2320-9801
ISSN (Print) : 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Website: www.ijircce.com

Vol. 6, Issue 8, August 2018

huge amount of data is quite high as the data to be processed is too large.

B. Big Data: BIG Impact

a. Definition

Big data is a term for data sets that are so large or complex that traditional data processing application software is inadequate to deal with them. Big data challenges include capturing data, data storage, data analysis, search, sharing, transfer, visualization, querying, and updating and information privacy. Lately, the term "big data" tends to refer to the use of predictive analytics, user behaviour analytics, or certain other advanced data analytics methods that extract value from data. There is little doubt that the quantities of data now available are indeed large, but that's not the most relevant characteristic of this new data ecosystem.

b. Characteristics of Big Data (5 V's. Volume velocity variety veracity value)

- Volume - Volume describes the amount of data generated by organizations or individuals.
- Velocity - Velocity describes the frequency at which data is generated, captured and shared.
- Variety -. It means unstructured text, video, audio those have important impact.
- Veracity - the quality and uncertainty of the data
- Value - business value to be derived.

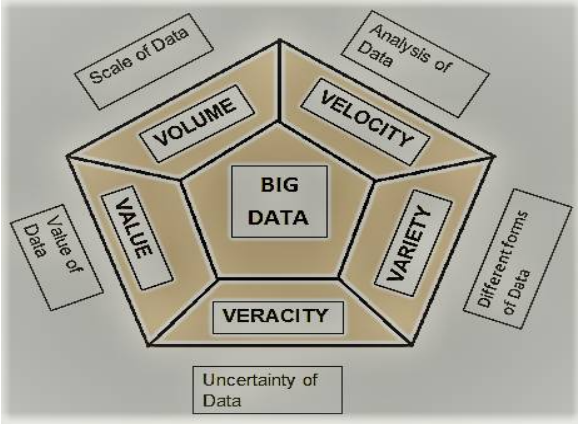

Figure 1:- 5 V's of Big Data

c. Examples of Big Data:

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

- Social media interactive platforms like Twitter, Facebook, YouTube, Instagram are the source of large amount of images, videos and audios.
- Remote sensors and machine-generated data from IoT also contribute to valuable source of big data
- E-Commerce sites like Amazon, Flipkart, eBay, etc. generate data from orders, products, visits, baskets which is useful for increasing their sales.

II. BIG DATA HADOOP DISTRIBUTIONS

A. Cloudera

Cloudera Inc. was founded by big data geniuses from Facebook, Google, Oracle and Yahoo in 2008. It was the first company to develop and distribute Apache Hadoop-based software and still has the largest user base with most number

ISSN(Online): 2320-9801
ISSN (Print) : 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Website: www.ijircce.com

Vol. 6, Issue 8, August 2018

of clients. Although the core of the distribution is based on Apache Hadoop, it also provides a proprietary Cloudera Management Suite to automate the installation process and provide other services to enhance convenience of users which include reducing deployment time, displaying real time nodes count, etc.

B. Hortornworks

Hortonworks, founded in 2011, has rapidly emerged as one of the top and leading vendors of Hadoop. The distribution provides open source platform based on Apache Hadoop for analysing, storing and managing big data. Hortonworks is the only commercial vendor to distribute complete open source Apache Hadoop without additional proprietary software. Hortonworks' distribution HDP2.6.5 can be directly downloaded from their website free of cost and is easy to install. The engineers of Hortonworks are behind most of Hadoop's recent innovations including YARN, which is better than MapReduce in the sense that it will enable inclusion of more data processing frameworks.

C. MapR

MapR is privately held company at California that contributes to Apache Hadoop projects like HBase, Pig, Hive and ZooKeeper. Their products include MapR FS file system, MapR-DB NoSQL database and MapR Streams. It develops technology for both, commodity hardware and cloud computing services.In its standard, open source edition, Apache Hadoop software comes with a number of restrictions. Vendor distributions are aimed at overcoming the issues that the users typically encounter in the standard editions. Under the free Apache license, all the three distributions provide the users with the updates on core Hadoop software. But when it comes to handpicking any one of them, one should look at the additional value it is providing to the customers in terms of improving the reliability of the system (detecting and fixing bugs etc), providing technical assistance and expanding functionalities.All three top Hadoop distributions, Cloudera, MapR and Hortonworks offer consulting, training, and technical assistance. But unlike its two rivals, Hortonworks' distribution is claimed to be 100 precent open source. Cloudera incorporates an array of proprietary elements in its Enterprise 4.0 version, adding layers of administrative and management capabilities to the core Hadoop software.Going a step further, MapR replaces HDFS component and instead uses its own proprietary file system, called MapRFS. MapRFS helps incorporate enterprise-grade features into Hadoop, enabling more efficient management of data, reliability and most importantly, ease of use. In other worlds, it is more production ready than its other two competitors.Up to its M3 edition, MapR is free, but the free version lacks some of its proprietary features namely, JobTracker HA, NameNode HA, NFS-HA, Mirroring, Snapshot and few more.

III. HADOOP CORE COMPONENTS

A. 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, which receives the output from a map as an input and merge those data tuples into a smaller set of tuples. 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. Dividing a data processing technique into mappers and reducers 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.

B. HDFS

HDFS connects together the file system on many local nodes to make them into one large file system. HDFS handles failure of nodes, 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 clusters of servers. Once the quantity and quality of enterprise data is accessible in HDFS, and YARN permit multiple data access applications to process it.HDFS is a scalable, fault-tolerant, distributed storage system that works closely with a huge variety of serialize data access applications, arranged by YARN and it remains economical at every amount

ISSN(Online): 2320-9801
ISSN (Print) : 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Website: www.ijircce.com

Vol. 6, Issue 8, August 2018

of storage.The file content is divided into large blocks in gigabytes, and each block of the file is independently replicated at multiple Data Nodes. The blocks are stored on the local file system on the Data Nodes. The NameNode actively monitors the number of replicas of a block. When a replica of a block is lost due to a Data Node failure or disk failure, the NameNode creates another replica of the block. The NameNode maintains the namespace tree and the mapping of blocks to Data Nodes, holding the entire namespace image in RAM. The NameNode does not directly send requests to Data Nodes. It sends instructions to the Data Nodes by replying to heartbeats sent by those Data Nodes.

IV. BIG DATA DISTRIBUTION ARCHITECTURE OR CDH

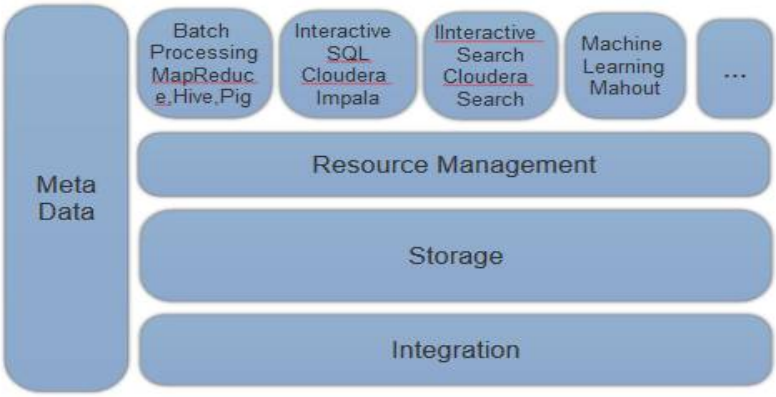

Figure 2: Cloudera Distribution for Hadoop Platform (CDH)

A. Sqoop:-"SQL to Hadoop and Hadoop to SQL"
Sqoop is a hadoop tool designed to transmit data to Hadoop and relational database servers and vice versa. It is used to import data from relational databases such as Oracle, MySQL, MariaDB to Hadoop HDFS and export from Hadoop distributed file system to relational databases. It is provided by the Apache Cloudera Software Foundation.

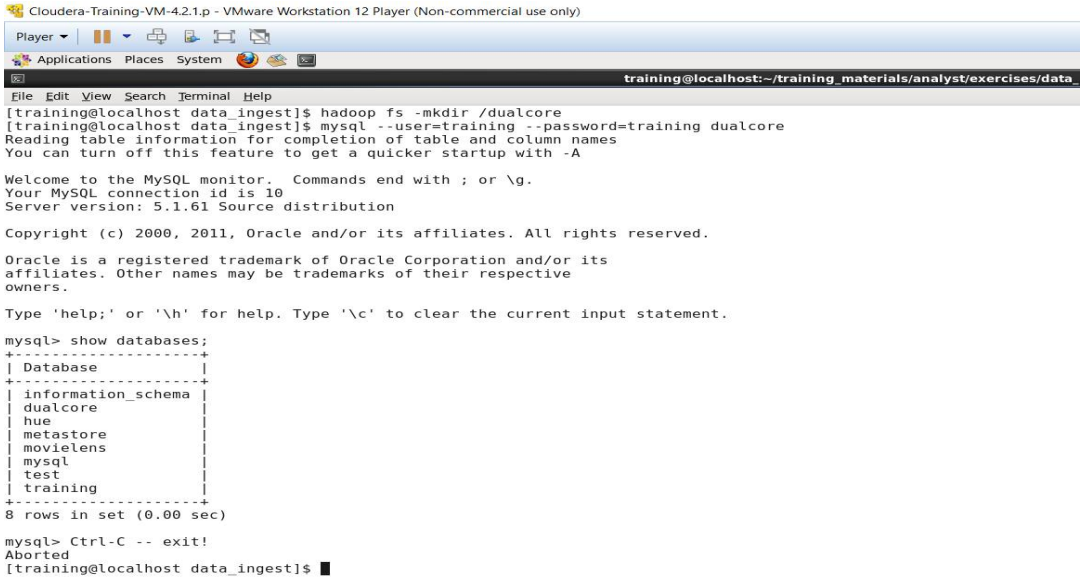

Figure 3: Apache Sqoop

ISSN(Online): 2320-9801
ISSN (Print) : 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Website: www.ijircce.com

Vol. 6, Issue 8, August 2018

B. Pig

Pig is a high level scripting language that is used with Apache Hadoop framework. Pig tool enables data workers to write complex data transfers without knowing Java program. Pig's simple SQL-like scripting language is called Pig Latin.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 JRuby, Jython 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.
Pig works with data information from many sources, including structured and unstructured data, and loads the results into the Hadoop Data distribution File System.
Pig scripts are transformed into a sequence of MapReduce tasks that runs on the Apache Hadoop cluster.

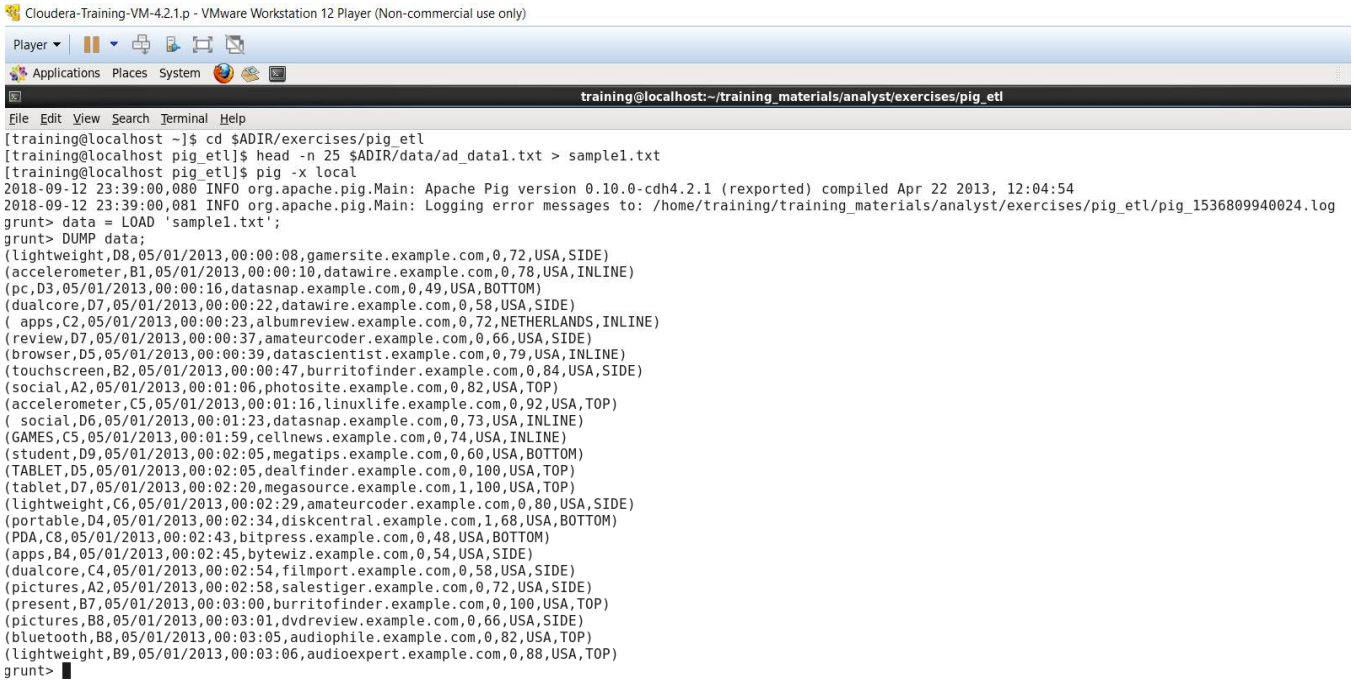

Figure 4: Apache Pig

C. Hive

Hive makes use of equivalent query servers with intelligent in-memory caching to avoid Hadoop's batch-oriented latency and provide as fast as sub-second query response times against smaller data volumes, while Hive on Tez continues to provide excellent batch query performance against petabyte-scale data sets.
The tables in Hive are similar to tables in a relational database, and data units are organized in a taxonomy from larger to more granular units. Databases are comprised of tables, which are made up of partitions. Data can be accessed via a simple query language and Hive supports overwriting or appending data.
Within a particular database, data in the tables is serialized and each table has a corresponding Hadoop Distributed File System (HDFS) directory. Each table can be sub-divided into partitions that determine how data is distributed within sub-directories of the table directory. Data within partitions can be further broken down into buckets.
Hive supports all the common primitive data formats such as BIGINT, BINARY, BOOLEAN, CHAR, DECIMAL, DOUBLE, FLOAT, INT, SMALLINT, STRING, TIMESTAMP, and TINYINT. In addition, analysts can combine primitive data types to form complex data types, such as structs, maps and arrays.

ISSN(Online): 2320-9801
ISSN (Print) : 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Website: www.ijircce.com

Vol. 6, Issue 8, August 2018

D. Impala

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 integrated with Hadoop to use the same file and data formats, metadata, security and resource management frameworks used by Map Reduce, Apache Hive, Apache Pig and other Hadoop software.Impala is promoted for analysts and data scientists to perform analytics on data stored in Hadoop via SQL or business intelligence tools. The result is that large-scale data processing (via MapReduce) and interactive queries can be done on the same system using the same data and metadata – removing the need to migrate data sets into specialized systems and/or proprietary formats simply to perform analysis.Features include:Supports HDFS and Apache HBase storage,Reads Hadoop file formats, including text, SequenceFile, 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 ApacheHive.

D. Hue:

Hue is a query based interactive editor for hadoop stack, like hive and impala.

Hue, the web-based interface that makes Apache Hadoop easier to use, helps you do that through a GUI in your browser — instead of logging into a Hadoop gateway host with a terminal program and using the command line.

Applications for HUE are usually implemented in DJango, a popular MVC web framework that understands the application namespaces. On top of that, the SDK lets the application bundle and start helper daemons which might, for example, talk to various interfaces in Hadoop, HDFS, or one of the numerous other applications that ship with CDH.

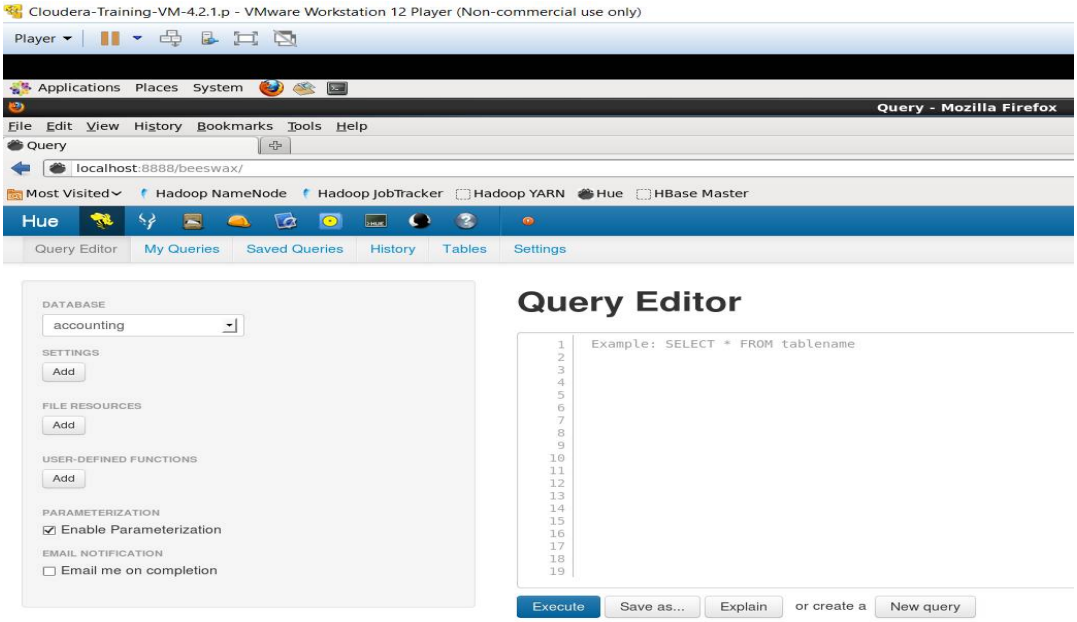

Figure 5: Hue Query Editor with Hive and Impala

F. Oozi:

Apache Oozie Workflow Scheduler for Hadoop is a workflow and coordination service for managing Apache Hadoop jobs:

- Oozie Workflow jobs are Directed Acyclic Graphs (DAGs) of *actions*; *actions* are typically Hadoop jobs (MapReduce, Streaming, Pipes, Pig, Hive, Sqoop, etc).
- Oozie Coordinator jobs trigger recurrent Workflow jobs based on time (frequency) and data availability.
- Oozie Bundle jobs are sets of Coordinator jobs managed as a single job.

Oozie is an extensible, scalable and data-aware service that you can use to orchestrate dependencies among jobs running on Hadoop.Oozie v3 is a server based *Bundle Engine* that provides a higher-level oozie abstraction that will

ISSN(Online): 2320-9801
ISSN (Print) : 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Website: www.ijircce.com

Vol. 6, Issue 8, August 2018

batch a set of coordinator applications. The user will be able to start/stop/suspend/resume/rerun a set coordinator jobs in the bundle level resulting a better and easy operational control.Oozie v2 is a server based *Coordinator Engine* specialized in running workflows based on time and data triggers. It can continuously run workflows based on time (e.g. run it every hour), and data availability (e.g. wait for my input data to exist before running my workflow).Oozie v1 is a server based *Workflow Engine* specialized in running workflow jobs with actions that execute Hadoop Map/Reduce and Pig jobs.

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,

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 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/

BIOGRAPHY

1. 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 is Red hat(RHEL 7.0) RHCSA and RHCE certified and also completed training of Cloudera's and Hortornwork's Big Data Course. Her current research area of interest is Big Data.

2. Mr. Shubham Jadhav is pursuing his bachelor's degree from Maharashtra Institute of Technology. He is a student of final year, computer science and engineering department and completed training on Cloudera's Big Data Analytics Course from MIT's Big Data Academy and also completed training on Red hat RHCSA.

3. Mr. Pranav Joshi is pursuing his bachelor's degree from Maharashtra Institute of Technology. He is a student of final year, computer science and engineering department and completed training on Cloudera's Big Data Analytics Course from MIT's Big Data Academy.

4. Ms. Dnyaneshwari Pawaris pursuing her bachelor's degree from Maharashtra Institute of Technology. She is a student of final year, computer science and engineering department and completed training on Cloudera's Big Data Analytics Course from MIT's Big Data Academy.