# Hadoop Ecosystem: A Future Analytical Hub

[1]Subita Kumari
[1]Research Scholar,
[1]CSE Deptt., UIET
[1]MDU, Rohtak, India

*Abstract:* Hadoop, the most popular open source implementation of Map-Reduce, stores and processes large datasets in a distributed manner. It is better suitable for offline batch processing. The popularity of Hadoop has grown in the last few years because it carries out analysis of data in a variety of formats, from unstructured data, to semi-structured data, such as logs, to structured data with a fixed schema. The Hadoop Ecosystem includes other tools to address particular needs like Hive and Pig. Hive is a SQL dialect and Pig is a dataflow language for that hide the tedium of creating Map-Reduce jobs behind higher-level abstractions.

***IndexTerms*** **- Hadoop Ecosystem, Map-Reduce, Hive, Pig**

## I. INTRODUCTION

Hadoop is simply the name of a stuffed toy elephant that belonged to the son of its creator "DOUG CUTTING". Hadoop Ecosystem is an open-source integrated framework which stores and processes large datasets in a distributed manner. It is data warehouse solution which stores data in petabytes and carries out advanced analytics on the datasets. It is better suitable for offline batch processing. Its two major modules are Map-Reduce (for processing) paradigm and Hadoop Distributed File System (for storage). The Hadoop Ecosystem is a combination of Hadoop along with different tools such as Pig, Sqoop, and Hive that are used to help Hadoop modules.

## II. HADOOP DISTRIBUTED FILE SYSTEM (HDFS)

HDFS is core component of Hadoop which is designed for storing very large files with streaming data access patterns. Streaming access means write once and read more. It is highly fault tolerant system and provide high throughput. It can be built out of commodity hardware. HDFS is not suitable in situations where lots of small files are there, low latency data access is required and multiple writes operations are required on files. HDFS contains mainly two types of nodes i.e. Name Node and Data Node as shown in Fig 1. Name node is the master of the system. It does not store the data. It is present on very expensive hardware like RAM for fast access. It maintains and manages the blocks which are present on the data nodes. Job tracker daemon runs on the name node. Data nodes are the slaves which are deployed on the commodity machines and provide the actual storage. They are responsible for serving read and write requests for the clients. Task tracker daemons run on the data nodes. HDFS client is an application used to interact with name node (job tracker) and data node (task tracker). Task tracker carries out the actual map-reduce operations. Hive, mySQL, other SQL, Pig and local file systems use HDFS for temporarily processing of their data. Figure 2 shows this concept.

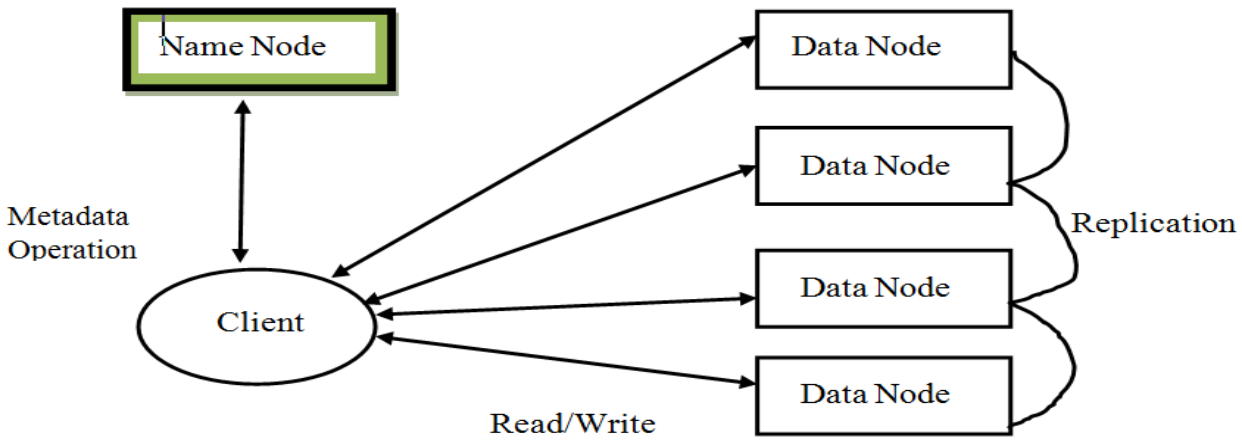


Figure 1: Architecture of Hadoop Distributed File System

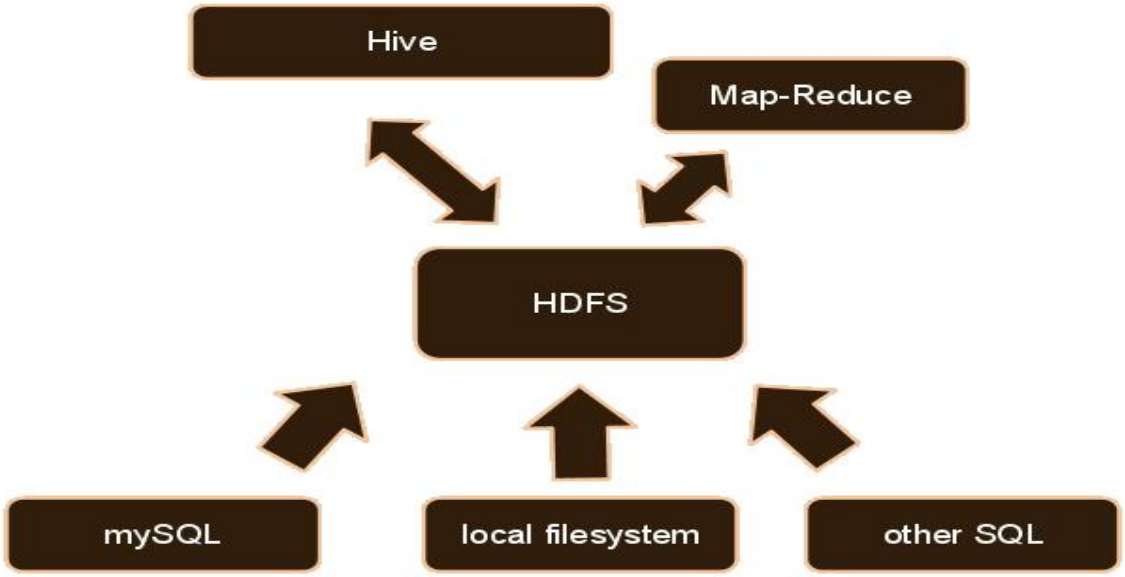


Figure 2: HDFS as Central File System

## III. MAP-REDUCE

Map-Reduce is a parallel and distributive programming paradigm for processing bulk amount of heterogeneous and unstructured data on clusters of inexpensive and easily available hardware. Hadoop sends the map-reduce program to datasets stored on commodity hardware. Map-Reduce operations of Hadoop can be executed using various mechanisms. The conventional method uses Java map-reduce program for semi-structured, structured and unstructured data. The scripting method uses map-reduce for semi-structured and structured data using Pig. To process structured data using Hive map-reduce operations are implemented with the help of Hive QL (Hive Query Language). Figure 3 shows the functioning of map-reduce in Hadoop. Job client submits jobs to Hadoop. Job tracker co-ordinates jobs and task tracker executes job tasks. Step by step functioning of Map-Reduce in Hadoop is as below -

- Client submits jobs to Job Tracker.
- Job Tracker works as a master and talks to Name Node & creates exec plan of job.
- Job tracker submits work to Task Trackers.
- Task Trackers work as slaves and break jobs into map and reduce tasks.
- Task Trackers report progress via heartbeats.
- Job Tracker manages phases and updates status.

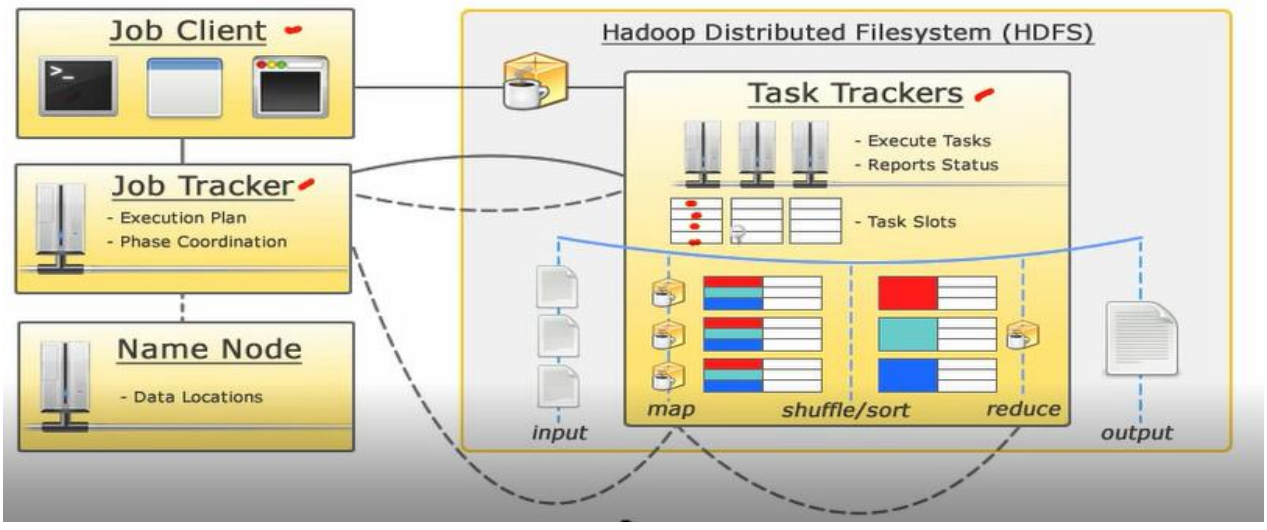


Figure 3: Functioning of Map-Reduce in Hadoop

Figure 4 shows the internal functioning of map-reduce task at data nodes. In map-reduce computing model, two functions are used i.e. map function and reduce function. Both these functions are written by users according to their requirement.

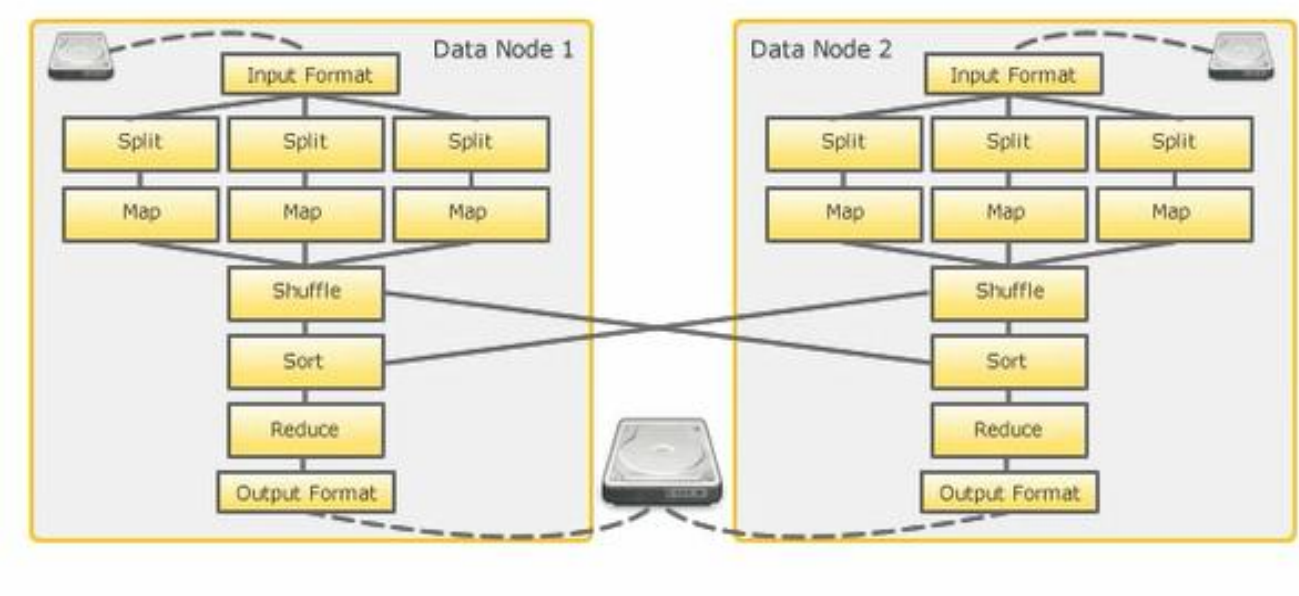


Figure 4: Internal Functioning of Map-Reduce at Data Node

The map function is used to process key-value pair and generates intermediate results and send them to reduce function. Reduce function further condense the values into a smaller set. Programmers find these functions easy to use because they do not need to know the parallel and distributed system. Google's cluster uses more than thousand map reduce jobs per day. The description of each phase is given in table1.

Table 1: Description of Phases of Map-Reduce

| Sr. No. | Name of Phase | Description of The Phase |
|---|---|---|
| 1 | InputFormat | InputFormat determines how the files are parsed into the MapReduce pipeline. |
| 2 | Split Phase | In this phase, the input data is divided into input splits based on the InputFormat. Input splits equate to a map task which all run in parallel. |
| 3 | Map Phase (Mappers) | This phase transforms the input splits into key/value pairs based on user-defined code. |
| 4 | Shuffle & Sort | This phase moves map outputs to the reducers and sort them by the key. |
| 5 | Reduce Phase (Reducers) | This phase aggregates key/value pairs based on user-defined code. |
| 6 | OutputFormat | OutputFormat determines how the results are written to the output directory. |

## IV. HIVE

Facebook is the originator of Hive. But Apache Software Foundation developed it further as open source software under the name Apache Hive. Apache Hive first brought SQL to Hadoop. It is an application developed for a data warehouse that provides the SQL like interface as well as a relational model. Hive infrastructure is built on the top of Hadoop that help in providing analysis for respective queries. It provides a mechanism to project structure onto the data and query the data using SQL like language called HiveQL. HiveQL is pretty much like SQL. Besides common SQL features (WHERE, HAVING, JOIN, GROUP BY, SORT BY etc.), HiveQL also have extensions such as TABLESAMPLE, LATERAL VIEW, OVER etc. Hive translates its own dialect of SQL (HiveQL) queries to a directed acyclic graph of map reduce jobs. Hive uses map reduce and HDFS for processing and storage/retrieval of data. Hive requires users to provide schema, storage format (optional) and serializer/deserializer (SerDe) while creating a table. All this information is saved in the metadata repository and is used whenever the table is referenced. Like any database management system (DBMS), one can run Hive queries from a command line interface (known as the Hive shell), from a Java Database Connectivity (JDBC), Open Database Connectivity (ODBC) or from Web UI (User Interface) as shown in Fig. 5.

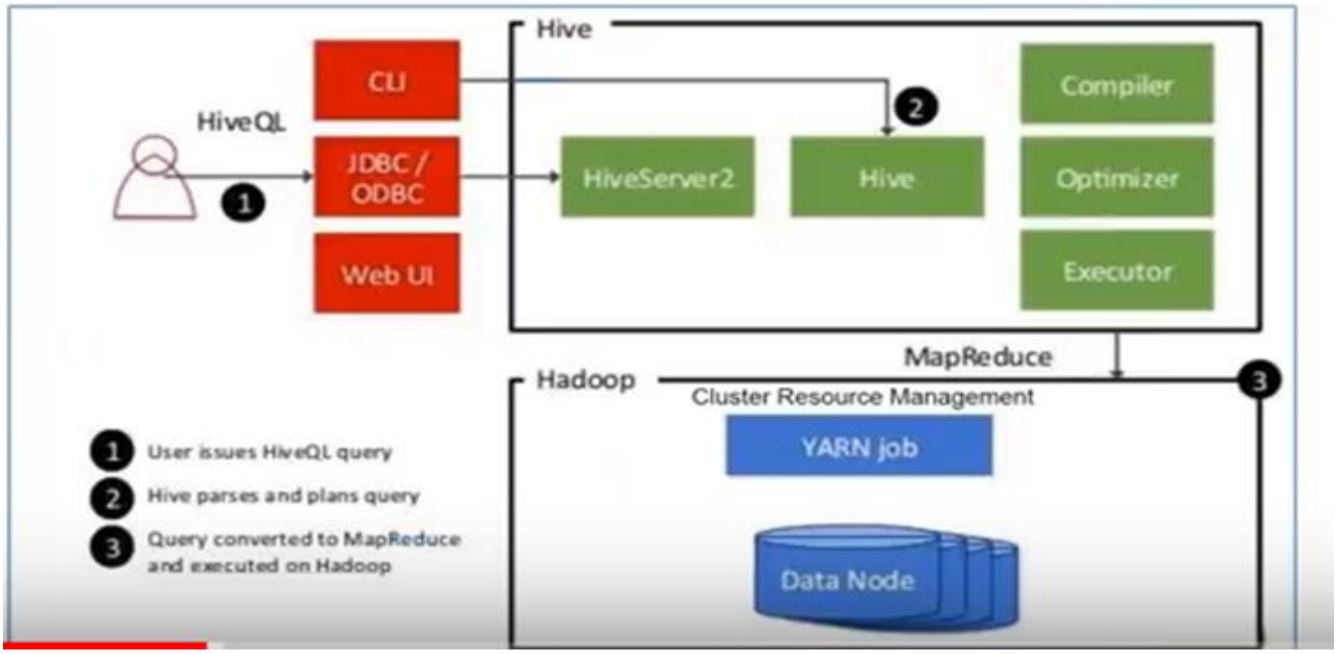


Figure 5: Hive Architecture

## V. PIG

Pig is a platform for designing data flows for ETL processing and analysis of large datasets by generating Map-Reduce jobs. Pig Latin is the programming language for Pig. It gives data manipulation operations, like joining, grouping and filtering. This high-level language allows developers to use HDFS data without writing complex Map-Reduce tasks. The Pig Latin scripting language is not only a higher-level data flow language but also has operators similar to SQL that are translated into a series of map and reduce functions. Pig Latin is designed to fill the gap between the declarative style of SQL and the low-level procedural style of Map-Reduce. Figure 6 shows the architecture of Pig. Table 2 shows the description of components of Apache Pig.

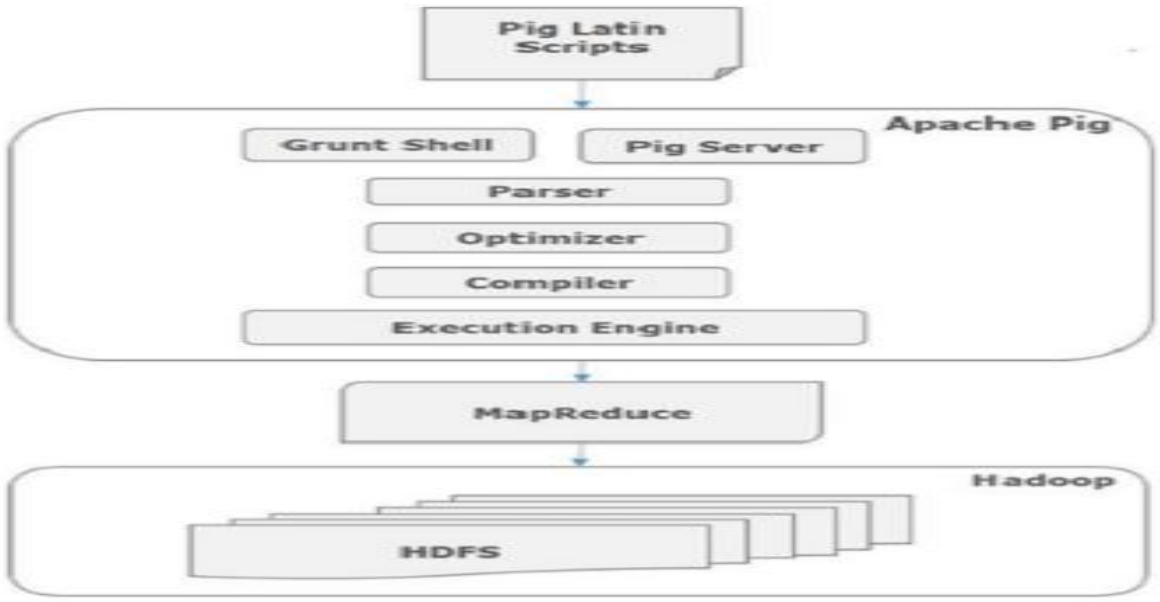


Figure 6: Pig Architecture

Table 2: Description of Components of Apache Pig

| Sr. No. | Name of Component | Function of The Component |
|---|---|---|
| 1 | Parser | Initially the Pig Scripts are handled by the Parser. It checks the syntax of the script, does type checking, and other miscellaneous checks. The output of the parser will be a DAG (directed acyclic graph). |
| 2 | Optimizer | DAG is passed to the logical optimizer, which carries out the logical optimizations such as projection and pushdown. |
| 3 | Compiler | The compiler compiles the optimized logical plan into a series of Map-Reduce jobs. |
| 4 | Execution Engine | Map-Reduce jobs are submitted to Hadoop in a sorted order. Finally, these Map-Reduce jobs are executed on Hadoop producing the desired results. |

## IV. CONCLUSION

Hadoop Ecosystem contains Hadoop, HDFS, Map-Reduce, Pig, Hive and some other high level language tools. Hadoop is data warehouse solution which stores data in petabytes and carries out advanced analytics on the datasets. It is better suitable for offline batch processing. Its two major modules Map-Reduce (for processing) paradigm and Hadoop Distributed File System (for storage) have been explained in this paper. Hive and Pig are high level language tools which use HDFS for data processing. They provide custom code for implementing Map-Reduce thus making user job pretty easier. Their architecture has been explained in this paper. Hadoop Ecosystem seems to be the most useful tool to cater the needs of future analytical requirement.

## REFERENCES

[1] Wenliang Huang, Zhen Chen, Wenyu Dong, Hang Li, Bin Cao, and Junwei Cao, “Mobile Internet Big Data Platform in China Unicom”, TSINGHUA SCIENCE AND TECHNOLOGY, ISSN 1007-0214 10/10 pp95-101 Volume 19, Number 1, (2014) .

[2] Min Chen, Shiwen Mao and Yunhao Liu, “Big Data: A Survey”, Springer, School of Computer Science and Technology, Huazhong University of Science and Technology, 1037 Luoyu Road, Wuhan, 430074, China. (2014).

[3] Raghupathi and Raghupathi, “Big data analytics in healthcare: promise and potential”, Health Information Science and Systems, http://www.hissjournal.com/content/2/1/3 (2014).

[4] Manyika J, McKinsey Global Institute, Chui M, Brown B, Bughin J, Dobbs R, Roxburgh C, Byers AH, “Big data: the next frontier for innovation, competition, and productivity”, McKinsey Global Institute (2011).

[5] “Apache Hive,” http://hortonworks.com / hadoop /hive/,2011-2014.

[6] Sanjeev Dhawan, Sanjay Rathee," Big Data Analytics using Hadoop Components like Pig and Hive ", American International Journal of Research in Science, Technology, Engineering & Mathematics ,pp. 88-93,March-May, 2013.

[7] Ashish Thusoo, Joydeep Sen Sarma, Namit Jain, Zheng Shao, Prasad Chakka, Ning Zhang, Suresh Antony, Hao Liu, Raghotham Murthy ," Hive – A Petabyte Scale Data Warehouse Using Hadoop", VLDB Endowment, ACM, August 24-28, 2009, Lyon, France.

[8] Subita Kumari , Pankaj Gupta," Document Store NoSQL Databases", International Journal of Artificial Intelligence and Knowledge Discovery Vol.5, Issue 3, July, 2015.