A Comparative Study of Hadoop-Based Big Data Architectures

Vanika[1], Aman Kumar Sharma[2]
[1]Dept of Computer Science
[2]Professor, Dept of Computer Science
[1, 2] Himachal Pradesh University, Shimla, HP, INDIA.

Abstract- *Big Data is a popularized concept which has presented a strong commercial and marketing challenge in the organizations. Market has given rise to new solutions regarding collection of Big Data combining the traditional technologies with the new ones. HortonWorks, Cloudera, MapR, IBM Infosphere BigInsights and Pivotal HD are few distributions used for managing Big Data that are available in the market. The different distributions have an individual approach with respect to Hadoop. In this paper, the architectures and components of the five distributions of Hadoop solutions for Big Data are explained and a comparative study is made to define the strengths and weaknesses of the various Hadoop distributions.*

Keywords- Big Data, Hadoop Distribution, HortonWorks, Cloudera, MapR, IBM Infosphere BigInsights, Pivotal HD.

I. INTRODUCTION

Big data is the upcoming technology which brings huge benefits to the business organizations. The huge development in the information technology has been causing growth of data. This huge amount of data has to be collected, categorized, deployed, stored and analyzed. Hence, it appears an urgent need for a robust system capable of doing all the analysis within organizations. Several distributions used for managing Big Data in Hadoop architecture are available in the market, namely HortonWorks [1], Cloudera [2], MapR [3], IBM Infosphere BigInsights [4] and Pivotal HD [5], etc. Each distribution has its own approach for a Big Data system and the choice of selecting the distribution will be based on various parameters depending on several requirements. All distributions are easy to install and are compatible supporting each other.

In this paper, Section I contains the introduction of Big data and various Hadoop distributions. Section II discusses the various Hadoop distributions namely Hortonworks, Cloudera, MapR, IBM Infosphere BigInsights and Pivotal HD. In Section III, a comparative study on the five distributions is putforth in order to distinguish the strengths and weaknesses of each of these Hadoop distributions. In the last section, conclusion is given.

II. BIG DATA HADOOP DISTRIBUTIONS

There are several distributions that permit to manipulate a Big Data system along with managing its main components: HortonWorks, Cloudera, MapR, IBM Infosphere BigInsights and Pivotal HD. In this paper, five most commonly used distributions of Big Data [6] are discussed namely HortonWorks, Cloudera, MapR, IBM Infosphere BigInsights and Pivotal HD.

2.1 HORTONWORKS DISTRIBUTION

In 2011, members of the Yahoo team in charge of the Hadoop project formed HortonWorks. It is a big Hadoop contributor and its economic model is not for the purpose of selling a license but of selling exclusively support and training [1]. This distribution is most consistent with Apache's Hadoop platform that develops, supports and provides expertise on a set of open source software designed to manage data and processing for things such as Internet of Things (IOT) and machine learning. Hortonworks has three interoperable product lines: Hortonworks Data Platform (HDP), Hadoop Distributed File (HDF) and Data Plane Services [7]. There are five pillars to Hadoop that make it enterprise ready [8]:

Data Management –It supports storing and processing of data linearly.

- **Apache Hadoop YARN (Yet Another Resource Negotiator)** – It is a next-generation framework for Hadoop data processing which extends MapReduce capabilities.
- **Hadoop Distributed File System (HDFS)** –It is a Java-based file system that provides scalable and reliable data storage for large clusters.

Data Access – It is the interaction with data from batch to real-time.

- **MapReduce** – It is a framework for writing applications that process large amounts of structured and unstructured data, in a reliable and fault-tolerant manner.

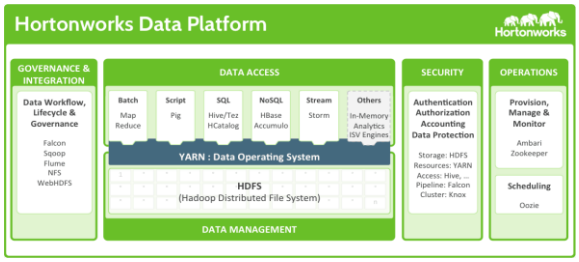

Fig. 1 HortonWorks Data Platform (HDP) [1]

- **Apache Hive** – It is a data warehouse that enables easy data summarization and ad-hoc queries via an SQL-like interface for large datasets stored in HDFS.
- **Apache Tez** – It generalizes the MapReduce paradigm to a framework for executing a complex Directed Acyclic Graph (DAG).
- **Apache HCatalog** – It is a table and metadata management service that provides a centralized way for data processing systems.
- **Apache Storm** – It is a distributed real-time computation system for processing large streams of data.
- **Apache Solr** – It is the open source platform for searches of data stored in Hadoop.
- **Apache Mahout** – It provides scalable machine learning algorithms that data science for clustering and classification.
- **Apache HBase** – A column-oriented NoSQL data storage system that provides read/write access for user applications.
- **Apache Accumulo** – It is a high performance data storage and retrieval system with cell-level access control.

Data Governance and Integration – It manages data according to policy.

- **Workflow Management** – Workflow Manager allows to easily create and schedule workflows and monitor workflow jobs.
- **Apache Flume** – Flume allows efficiently collecting and moving large amounts of log data from different sources to Hadoop.
- **Apache Sqoop** – Sqoop is a tool that speeds and eases movement of data in and out of Hadoop.
- **Apache Pig** – A platform for processing and analyzing large data sets.

Security –It deals with authentication, authorization, accounting and data protection.

- **Apache Knox** – The Knox Gateway provides a single point of authentication and access for Apache Hadoop services in a cluster.

Operations – Provision, managing and monitoring are the main operations.

- **Apache Ambari** – An open source installation lifecycle management, administration and monitoring system for Apache Hadoop clusters.
- **Apache Oozie** – It is a Java based web application used to schedule Apache Hadoop jobs. Oozie combines multiple jobs sequentially into one logical unit of work.
- **Apache ZooKeeper** – A highly available system for coordinating distributed processes.

2.2 CLOUDERA DISTRIBUTION

Cloudera was first founded by Hadoop experts from Facebook, Google, Oracle and Yahoo in 2008. Cloudera started Cloudera Distribution Including Apache Hadoop (CDH) that targeted enterprise-class deployments of that technology [2]. This distribution is largely based on the components of Apache Hadoop along with components for cluster management. Cloudera provides a scalable, flexible, integrated platform that makes it easy to manage rapidly increasing data. Cloudera provides the following products and tools [9]:

- **CDH** — The Cloudera distribution of Apache Hadoop provides security and integration with various hardware and software solutions.
- **Apache Impala** — It is a massively parallel processing SQL engine for interactive analytics and business intelligence. It queries Hadoop data files from different sources like MapReduce jobs or Hive tables.
- **Cloudera Search** — It provides real-time access to data stored in Hadoop and HBase.
- **Cloudera Manager** — It is an application used to deploy, manage, monitor and diagnose issues with CDH deployments. It includes the Cloudera Manager API, which is used to obtain cluster health

information and metrics, as well as to configure Cloudera Manager [9].

- **Cloudera Navigator** — It is a data management and security tool for the CDH platform. It enables administrators, data managers and analysts to explore the data in Hadoop and simplifies the storage and management of encryption keys.
- **Hue** - It is a web-based interactive query editor in the Hadoop stack that helps in visualizing and sharing of data.

2.3 MAPR DISTRIBUTION

MapR is a software company based in Santa Clara, California [10]. MapR provides access to different data sources from a single computer cluster, including big data workloads such as Apache Hadoop and Apache Spark, a distributed file system, a multi-model database management system and event stream processing, combining analytics in real-time with operational applications [11]. It runs on both commodity hardware and public cloud computing services. The MapR Data Platform provides a unified data solution for structured data (tables) and unstructured data (files) [12].

MapR-FS - The MapR File System is a read-write distributed file system that eliminates the Namenode associated with cluster failure in other Hadoop distributions. Its main features are :

- **Storage pools** – It is a group of disks in which MapR-FS writes data.
- **Containers** – It is an abstract entity that stores files and directories in MapR-FS.
- **CLDB (Container Location Database)** – It is a service that tracks the location of every container in MapR-FS.
- **Volumes** – It is a management entity that stores and organizes containers in MapR-FS.
- **Snapshots** – It is the read-only image of a volume at a particular time which is used to access deleted data.
- **Direct Access NFS** – It helps applications to read/write data directly into the cluster.

MapR-DB - The MapR Database provides distributed data replication for structured and unstructured data. The table manages structured data, while the file manages unstructured data. The MapR-DB tables store structured data as a nested sequence of key/value pairs.

2.4 IBM INFOSPHERE BIGINSIGHTS DISTRIBUTION

InfoSphere BigInsights for Hadoop was firstly introduced in 2011 in two versions: the Enterprise Edition and the basic version, which was a free download of Apache Hadoop bundled with a web management support [13]. Later, IBM launched the InfoSphere BigInsights Quick Start Edition in 2013. This new edition provided massive data volume analysis capabilities on a business centric platform [4]. InfoSphere BigInsights supports structured, unstructured and semi-structured data and offers maximum flexibility. It also features Hadoop and its related technologies as a core component [13].

- **File systems** - InfoSphere BigInsights can be installed with either the Hadoop Distributed File System (HDFS) or the General Parallel File System (GPFS).
- **MapReduce frameworks** – It is the core of Apache Hadoop which provides programming model to servers in a Hadoop cluster.
- **Text Analytics** – It extracts information from unstructured and semi-structured data and produce documents that provide valuable insights into original data.
- **IBM Big SQL** – It is a data warehouse system that is used to summarize, query and analyze data.
- **InfoSphere BigInsights Console** – It is an integrated console that is used to view cluster, manage files, cluster instances and schedule jobs and tasks from a single location.

2.5 PIVOTAL HD DISTRIBUTION

Pivotal Software, Inc. (Pivotal) is a software and services company based in San Francisco and Palo Alto, California. The divisions include Pivotal Labs for consulting services, the Pivotal Cloud Foundry development group, and a product development group for the Big Data market [5]. In March 2013, an Apache Hadoop distribution called Pivotal HD was announced, including a version of Greenplum software [14] called HAWQ. Pivotal HD allows enterprises to simplify development, expand Hadoop's capabilities, increase productivity and cut costs. Pivotal HD Enterprise is a commercially supported distribution of Apache Hadoop [15].

- **Command Center** – It is a command line web-based tool for installing, managing and monitoring Pivotal HD cluster.
- **Pivotal Data Loader** – It is a high-speed data ingest tool for Pivotal HD cluster.

- **Unified Storage System (USS)** – It is a framework that provides HDFS protocol layer on top of external file systems.
- **Spring Data** – It provides support for writing Apache Hadoop applications.
- **HAWQ** – It is a parallel SQL query engine that combines the merits of the Greenplum Database Massively Parallel Processing (MPP) relational database engine and the Hadoop parallel processing framework [14].
- **Pivotal Extension Framework (PXF)** – It is used to provide support for external data formats such as HBase and Hive.

III. COMPARISON AND ANALYSIS

A comparative study of the Hadoop distributions architecture of Big Data is presented for an evaluation between the distributions namely HortonWorks, Cloudera, MapR, IBM InfoSphere BigInsights and Pivotal HD so as to identify their strengths and weaknesses. In this analysis, relevant criteria are used to distinguish and differentiate the different architectures for the five distributions of Big Data solutions.

3.1 CRITERIA FOR COMPARISON

The following criteria are used to distinguish between different distributions.

1. **MapReduce:** It is the programming model for making parallel and distributed computations of data.
2. **Apache Hadoop YARN:** It is a technology for managing clusters and making Hadoop more suitable for operational applications that cannot wait for the completion of batch processing [11].
3. **Non-Relational Data Base:** This database considers data manipulation techniques and dedicated processes to provide solutions to different data related issues.
4. **Disaster Recovery:** It helps in preventing data loss in case of a computer centre failure. It is also required to take care of the computer needs of the organization in case of a Big Data system disaster.
5. **Replication:** In order to improve reliability, fault tolerance and availability, the different Big Data Hadoop distributions use a process of information sharing to ensure data consistency across multiple redundant data sources. Data replication is called if the data is duplicated on multiple storage locations.
6. **Management Tools:** It helps in deploy, configure, automate, report, track, troubleshoot and maintain a Big Data system.
7. **Data and Job Placement Control:** It allows controlling the placement of data and a job on a Hadoop cluster and hence permits to choose nodes to execute jobs presented by different users and groups.
8. **DFS (Distributed File System):** DFS is a file system in which the data stored on a server is accessed and processed in the same way as if it was stored on the local client machine [16]. The server allows the client users to share files and store data locally.
9. **Data Ingestion:** It is the process of importing and obtaining data for immediate use or storage in a database or HDFS. Thus, Data can be broadcast in real-time or in batches.
10. **MetaData Architecture:** Two types of architectures are used - a centralized architecture where everyone depends on the same authority and a decentralized architecture that has no central authority.
11. **Data Access and Query:** Queries are utilized by users to express their information needs and to access data.
12. **Cluster Coordination:** It aims to ensure a coherent and complementary approach to avoid gaps in cluster work, by identifying ways to work together for the purpose of achieving better collective outcomes.
13. **Machine Learning:** It mainly concerns with the analysis, design, development and implementation of methods to solve problems.
14. **Cloud Services :** These are the dedicated services to exploit the computing and storage power of remote computer servers via a network that is the internet [17].
15. **Scheduler:** These are used to define the links between the processes and the way to launch them.

3.2 COMPARISON

Based on the literature study it was determined regarding each of the distributions about the criteria's being satisfied by the distributions. Also the distributions were executed to confirm the presence of these criterions.

Table 1 - Comparison between the five distributions [18][19]

CRITERIA/DISTRIBUTIONS	HORTONWORKS	CLOUDERA	MAPR	IBM BIGINSIGHTS	PIVOTAL HD
MapReduce	Yes	Yes	Yes	Yes	Yes
Apache YARN	Yes	Yes	Yes	Yes	Yes
Non-Relational Database	Yes	Yes	Yes	Yes	Yes
Disaster Recovery	No	Yes	Yes	Yes	Yes
Replication MetaData	No	No	Yes	Yes	Yes
Management Tools	Yes	Yes	Yes	Yes	Yes
Data and Job Placement Control	No	No	Yes	Yes	No
DFS	Yes	Yes	Yes	Yes	Yes
Data Ingestion/ Batch	Yes	Yes	Yes	Yes	Yes
Streaming	No	No	Yes	Yes	Yes
MetaData Architecture/ Centralized	Yes	Yes	No	No	No
Distributed	No	No	Yes	Yes	Yes
Data Access and Query	Yes	Yes	Yes	Yes	Yes
Cluster Coordination	Yes	Yes	Yes	Yes	Yes
Machine Learning	Yes	Yes	Yes	Yes	Yes
Cloud Services	Yes	Yes	Yes	Yes	Yes
Scheduler	Yes	Yes	Yes	Yes	Yes

Table – 1 shows the comparative study of the different Hadoop distribution providers in the Big Data. Various criteria are used to manage clusters, as well as to collect, sort, categorize, move, analyse, store and process Big Data. Hortonworks, Cloudera, MapR, IBM InfoSphere BigInsights and Pivotal HD contains Mapreduce as a programming model, Apache Yarn for managing data present in clusters and NoSql for manipulation of data using distributed file system. But HortonWorks and Cloudera fail to provide replication of data. All the five distributions contain schedulers for processing and querying jobs and have cluster coordination to gap in the cluster. In Table- 1, a 'Yes' denotes presence of criteria whereas a 'No' represents not satisfying the criteria.

After the analysis, some conclusions are drawn. Firstly, distributions based on Apache Hadoop gives a software solution to organizations so that they can install on their own infrastructure in private cloud and/or public cloud. Also, the five distributions are satisfying almost the same criteria laid down in the study that was proposed and each distribution focuses on core features dedicated to Big Data systems such as integration, security, scale performance and governance.

IV. CONCLUSIONS AND FUTURE SCOPE

Big Data refers to the release of the voluminous data from companies. Big data analytics is a process of collecting, organizing and analyzing large sets of data to discover patterns and useful information. This trend of collection and analysis of Big Data has given rise to new distributions to manage a Big Data system. Several distributions are there to manage clusters namely HortonWorks, Cloudera, MapR, IBM Infosphere BigInsights and Pivotal HD. The comparative study shows that the core of all five distributions is based on open source Hadoop architecture.

Although all of them have same platform but there is slight differences in terms of included projects and variants. These Hadoop distributions simplify the process; however these implementations still require a lot of advancements for mapreduce jobs in integrating data sources into Hadoop.

Future work may include optimization of the technology and approaches used in order to face the increasing multi-streams and Big Data challenges which further helps to improve performance, scalability and results accuracy.

REFERENCES

[1] Giles Stephen, Khan Umair, "Pro Hortonworks Data Platform: Harness the Power and Promise of Big Data with HDP", Apress, ISBN 978-1-4842-0668-3, 2015.
[2] Menon Rohit, "Cloudera Administration Handbook", Packt Publishing, ISBN 978-1-7835-5896-4, 2014.
[3] Dunning Ted, Friedman Ellen, "Real-World Hadoop", O'Reilly Media Inc., ISBN 978-1-4919-2266-8, 2015.
[4] Quintero Dino, "Implementing an IBM InfoSphere BigInsights Cluster using Linux on Power", IBM RedBooks Publication, ISBN 978-0-7384-4074-3, 2015.
[5] Pivotal Software Inc., "Pivotal HD Enterprise Installation and Administrator Guide", EMC Corporation, 2013.
[6] Starostenkov V, Senior R, Developer D, "Hadoop Distributions: Evaluating Cloudera, Hortonworks, and MapR in Micro-benchmarks and Real-world Applications", Altoros Systems Inc., 2013.
[7] HortonWorks homepage, https://en.wikipedia.org/wiki/Hortonworks accessed on 25/6/2018 at 1100 hrs.
[8] Hadoop Tutorial – Getting started with HDP, https://hortonworks.com/tutorial/hadoop-tutorial-getting-started-with-hdp/section/1/ accessed on 27/6/2018 at 1300 hrs.
[9] Cloudera, https://www.cloudera.com/documentation/enterprise/5-8-x/topics/introduction.html accessed on 17/7/2018 at 2300 hrs.
[10] Sharma Meenakshi, Chauhan Vaishali, Kishore Keshav, "A Review: MapReduce and Spark for Big Data Analytics", International Journal of Advanced Technology in Engineering and Science, Vol. 04, Issue 06, pp. 42-50, 2016.
[11] Alapati Sam R, "Expert Hadoop Administration: Managing, Tuning and Securing Spark, YARN and

HDFS", Addison-Wesley Professional, ISBN 978-0-13-459719-5, 2016.

[12] MapR Data Platform, http://doc.mapr.com/display/MapR/MapR+Data+Platform accessed on 20/7/2018 at 1800 hrs.

[13] IBM Knowledge Center, https://www.ibm.com/support/knowledgecenter.html accessed on 22/7/2018 at 1400 hrs.

[14] Gollapudi Sunila, "Getting Started with Greenplum for Big Data Analytics", Packt Publishing, ISBN 978-1-4493-6204-1, 2013.

[15] Overview of Apache Stack and pivotal Components, http://pivotalhd-210.docs.pivotal.io/doc/1110/OverviewofApacheStackandPivotalComponents.html accessed on 25/7/2018 at 1900 hrs.

[16] Alam B. M., "A New HDFS Structure Model to Evaluate the Performance of Word Count Application on Different File Size", International Journal of Computer Applications, Vol. - 111, No. 3, pp. 1–4, 2015.

[17] Trifan Mircea, Ionescu Dan, Ionescu Bogdan, "An Architecture and Methods for Big Data Analysis", Conference Paper, DOI 10.1007/978-3-319-18296-4_39, pp. 1-25, 2014.

[18] Erraissi Allae, Belangour Abdessamad, Tragha Abderrahim, "A Big Data Hadoop Building Blocks Comparative Study", International Journal of Computer Trends and Technology, Vol. 48 No. 1, pp. 109, 2017.

[19] Oussous Ahmed, Benjelloun Fatima-Zahra, Lahcen Ait Ayoub, Belfkih Samir, "Big Data technologies: A survey", Journal of King Saud University – Computer and Information Sciences, pp. 1-18, 2017.