A Big Data Hadoop building blocks comparative study

Allae Erraissi[1], Abdessamad Belangour[2], Abderrahim Tragha[3]

Laboratory of Information Technology and Modeling LTIM
University Hassan II, Faculty of Sciences Ben M'sik, Casablanca, Morocco

Abstract *— These last years, the new technologies produce each day large quantities of 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 whole issue, for companies and administrations, is not to pass by valuable information drowned in the mass. 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 a system Big Data namely HortonWorks [1], Cloudera [2], MapR [3], IBM Infosphere BigInsights [4], pivotal HD [5], Microsoft HD Insight [6], etc. The different distributions have an approach and a different positioning in relation to the vision of 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 article, we shall discuss the world of big data by defining these characteristics and its architecture. Then we shall talk about some distributions Hadoop, and finally, we shall conclude by a comparative study on the top five suppliers of Hadoop distributions of Big Data.*

Keywords — *Big Data, 5 V's, Distribution Hadoop, comparison*

I. INTRODUCTION

In recent years, new technologies have daily produced large amounts of data that need to be collected, sorted, categorized, moved, analyzed, stored, and so on. As a result of this, we enter the Big Data era in which several publishers offer ready-to-use distributions to manage a Big Data system, namely HortonWorks [1], Cloudera [2], MapR [3], IBM Infosphere BigInsights [4], Pivotal HD [6], etc. Indeed, different distributions have a different approach and positioning with regard to the vision of a Hadoop platform. The choice will be made on one or on the other solution according to several requirements. For example, if the solution is open source, Maturity of the solution, etc. Some editions have been supplemented with additional bricks, which make it possible to simplify the operation of the platforms that remains complex because of the number of components required. Thus, our work is to make a comparative study on the main Hadoop distribution providers to define the strengths and weaknesses of each distribution.

II. THE BIG DATA

A. Definition

There is no perfect definition of Big Data. The term is used by many companies and it is becoming more and more popular [7]. Here we are talking about the Big Data which are the mega data or sometimes called mass data. Every day, users create much content such as blog posts, tweets, photos and videos on social networks. More than that, servers continuously record messages in Log files to keep track of Events, and companies also record information about sales, suppliers, operations, customers, etc. A study shows that each year the amount of data will increase by 40 percent by 2020 [8].

Faced with this exponential growth in data volumes, companies are confronted with certain problems, such as how to collect, store, analyze and exploit these large volumes of data in order to create added value. The challenge for companies and administrations is not to miss out on valuable information drowned in the masses. This is where the "Big Data" technology comes in, based on a very fine analysis of masses of data.

B. Big Data Features

In order to talk about the Big Data, we shall talk about the 3Vs. A Big Data solution must process the " 3 Vs " of the large data [9]:

Volume: This is the weight of the data to be collected. The emergence of social networks has accentuated this production of data. It should be noted that 90% of the existing data in the world at present have been created during the last two years [10]. Before we talked about gigabytes, now we talk about terabytes, petabytes, exabytes and even zettabytes.

Velocity: By velocity, we mean the rate at which data arrives and how fast it must be treated. In the age of the Internet and of the almost instantaneous information, a decision-making must be rapid so that the company can not be overtaken by its competitors. The velocity varies from batch to real time. Therefore, immediate data processing would be the key element of a large data model.

Variety: Since the data comes from multiple sources such as messages published on social networking sites, digital photos and videos, sensors used to collect climate information, purchase transaction records, GPS signals from mobile phones, etc. It is not just traditional relational data that is structured, but also unstructured and even semi-structured data as well.

Some researchers and analysts add other components to these "3Vs":

Veracity: Veracity is meant to build confidence in the data collected and presented since most of the decision-makers do not trust the data on which they base their decisions.

Value: It is about being able to focus on data having real value and being operable. It represents the value that can be derived from these data and the uses they produce.

III. BIG DATA HADOOP DISTRIBUTIONS

There are several distributions which allow to manipulate a Big Data system and to manage its main components, namely HortonWorks [1], Cloudera [2], MapR [3], IBM Infosphere BigInsights [4], Pivotal [5], Microsoft HDInsight [6], etc. In this part, we shall talk about the three best known and used components in the world of Big Data. These are HortonWorks, Cloudera, and MapR.

A. HortonWorks

HortonWorks was formed in June 2011 by members of the Yahoo team in charge of the Hadoop project. Their goal is to facilitate the adoption of the Apache Hadoop platform that is why all components are open source and licensed Apache [1]. The purpose of the economic model of HortonWorks is not to sell the license but to sell support and training. Thus, this distribution is the most consistent with Apache's Hadoop platform and HortonWorks is a big Hadoop contributor.

B. Cloudera

Cloudera wants to be like the commercial company Hadoop which was founded by Hadoop experts from Facebook, Google, Oracle, and Yahoo. If their platform is largely based on Apache's Hadoop, It is complemented with home components primarily used for cluster management [2]. The purpose of Cloudera's economical model is the sale of licenses as well as support and training. Hence, Cloudera offers a fully open source version of their platform (Apache 2.0 license).

C. MapR

MapR was founded in 2009 by former Google members. Although its approach is commercial, MapR contributes to Apache Hadoop projects like HBase, Pig, Hive, ZooKeeper and especially Drill [3]. MapR differs mainly from the version of Apache Hadoop by taking distance with the core of the platform. They propose their own distributed file system as well as their own version of MapReduce: MapR FS and MapR MR [3].

IV. BIG DATA ARCHITECTURE

Before starting with Big Data, you have to make sure that all the essential components of the architecture for analyzing all aspects of a large amount of data are in place. Without this correct configuration, you will find it difficult to deal with such a mass of data.

An architecture of a Big Data system should be able to consume myriad data sources in a fast and inexpensive way. It should also have the following layers: Data sources, Ingestion Layer, Visualization Layer, Hadoop Platform management Layer, Hadoop Storage Layer, Hadoop Infrastructure Layer, Security Layer, and Monitoring Layer [11].

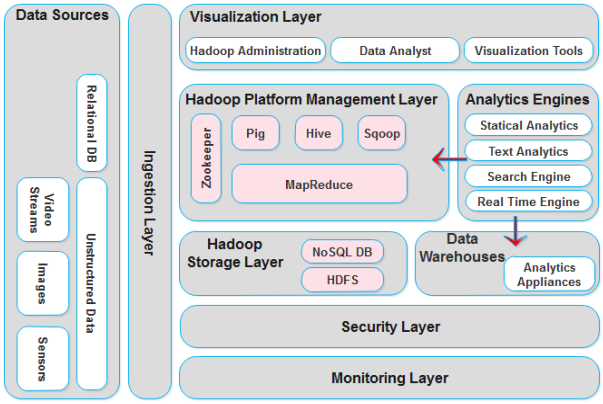

Fig.1. The Big Data architecture [11]

This figure describes the necessary components of the architecture that should be part of a Big Data system. It is necessary to choose open source or licensed frameworks to take full advantage of all the features of the different components of a Big Data system.

The architecture of a Big Data system consists of the following layers:

Visualization: This layer is useful for data analysts and scientists to understand better and faster, and hence to be increasingly able to examine different aspects of the data in different visual modes.

Hadoop Platform Management: This layer provides the necessary tools for processing the MapReduce as well as the query languages in order to access the NoSQL databases by using the distributed HDFS storage file system (PIG, HIVE, Sqoop, etc.)

Hadoop Storage: This layer is dedicated for storing data by using massively distributed storage and processing which constitute a change in the way a company manages the data. Hadoop uses HDFS, which is a distributed file system designed to store a very large volume of information (terabytes or petabytes) through a large number of machines in a cluster. It stores data reliably, runs on basic hardware, uses blocks to store a file or part of a file, and so on.

Security: This layer is designed for data protection since the security of this data becomes a

major concern. Client purchasing habits, patient medical history, genetic disease demographics, all these and many other types and uses of data need to be protected, both to meet compliance requirements and to protect The private life of the individual. These security requirements must be part of any Big Data system from the beginning.

Monitoring: With so many distributed data storage clusters and multiple data source ingestion points, it is important to get a complete picture of the Big Data system thanks to monitoring systems. Therefore, this layer defines the concepts used by these monitoring systems to increase the performance of Hadoop.

Ingestion: This layer allows to separate the noise from the relevant information. It must be able to handle the huge volume, high speed, and variety of data. It should also have the ability to validate, clean, transform, reduce and integrate the data so that the Hadoop ecosystem can use them later.

Data Sources: This layer defines the different types of data sources inside and outside the company that needs to be analyzed in a Big Data solution. The data in the Big Data are characterized by an enormous volume, Variety, velocity, and value. Therefore, it's a complex data flow that needs to be perfectly treated in the ingestion layer.

V. BIG DATA DISTRIBUTION ARCHITECTURES

During our study, we based on architectures of the different distributions Hadoop. Here there is the example of the three architectures: Cloudera distribution for Hadoop Platform, pivotal HD business, and HortonWorks data platform.

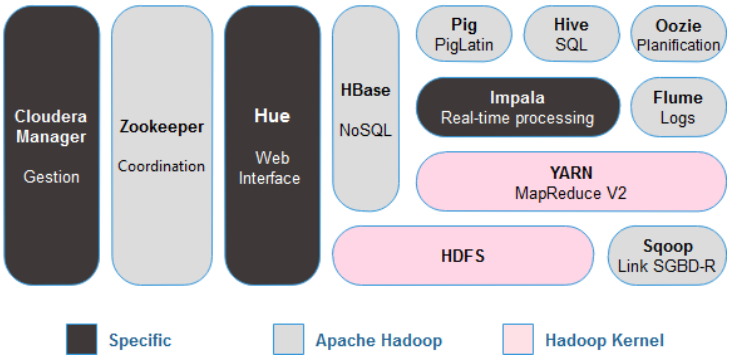

Fig.2. Cloudera Distribution for Hadoop Platform (CDH) [12]

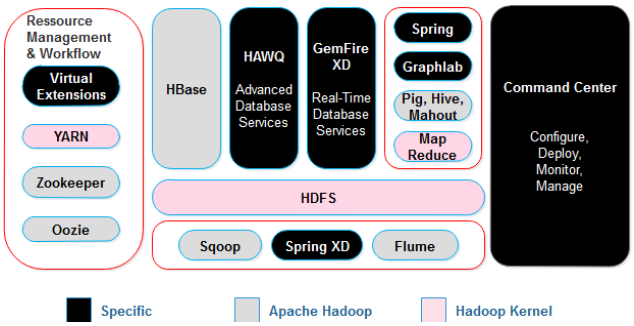

Fig.3. Pivotal HD Enterprise [5]

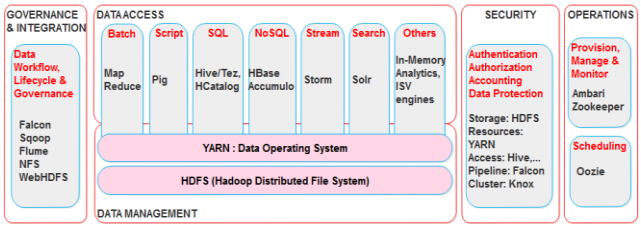

Fig.4. HortonWorks Data Platform [1]

According to the architectural study, the majority of Hadoop distributions are based on Hadoop's core components, HDFS, MapReduce and YARN [17]. In addition to that, there is Apache Hadoop products as well as specific products for each solution.

VI. COMPARISON BETWEEN DISTRIBUTIONS

In order to evaluate distributions, we conducted a comparative study to identify the strengths and weaknesses of the five major Hadoop distribution providers: Cloudera, HortonWorks, IBM InfoSphere BigInsights, MapR, and Pivotal.

We focus on an evaluation by Forrester Wave [13] on the same Hadoop distributions on which they used 35 evaluation criteria grouped into three high-level buckets: Current offering, Strategy, and Market presence.

Forrester's evaluation of Hadoop's distributions of Big Data revealed that there are four leaders and a Strong Performer, who is Pivotal Software. Cloudera, MapR Technologies, IBM, and HortonWorks are the leaders [13].

A. Criteria for comparison

To compare the five distributions, we used the following criteria:

- **Editor**: This is the company that designed, developed and commercialized the Hadoop solution.
- **Available edition**: It is for knowing the different distributions marketed by the publishers.
- **Base**: This is the basis on which publishers have based themselves to improve their solutions. We note that all publishers have concentrated on the HDFS and the second version of MapReduce which is YARN.

Table 1. Comparison distributions Big Data

Solution	Editor	Available edition	Base	Administ-ration console	Components of the solution
MapR	MapR	▪ M3 (free) ▪ M5 ▪ M7	HDFS, YARN	MapR control system	Tez, Spark, Cascading, Pig, MapReduce, GraphX, MLLib, Mahout, Drill, Shark, Impala, Hive, Accumulo, Solr, HBase, Storm, Hue, Flume, Sqoop, Knox, Sentry, Falcon, Oozie, Whirr, Zookeeper.
Cloudera	Cloudera	▪ Cloudera Express ▪ Cloudera Entreprise (Basic Edition, Flex Edition, Data Hub Edition)	HDFS, YARN	Cloudera manager	Hive, Pig, HBase, Hue, Avro, Whirr, Flume, Yarn, Mahout, Cloudera Impala, Cloudera Manager, Sqoop, Oozie, Zookeeper, Apache Sentry
Horton Works Data Platform	Horton Works	HortonWorks Data platform 2.5	HDFS, YARN	Ambari	Hive, Pig, HBase, Hue, Tez, Yam, Zookeeper, Mahout, Flume, Sqoop, Oozie, Whirr, Storm, Apache Ganglia, Apache Falcon, WebHDFS, NFS, Spark, Accumulo, Knox
Pivotal Data Suite	Pivotal	Pivotal HD Entreprise	HDFS, YARN	Command center	Yarn, Zookeeper, Oozie, HBase, HDFS, Sqoop, Spring XD, HAWQ, GemFire XD, Spring, Graphlab, Pig, Hive, Mahout, MapReduce, Flume,
IBM InfoSphere BigInsights	IBM	▪ Quick Start Edition ▪ Standard Edition ▪ Entreprise Edition	HDFS, YARN	Web Console	BigSheets, Dashboard & Visualization, Text Analytics, MapReduce, Workflow, Pig, Jaql, Hive, Text processing Engine & Extractor Library, R, Integrated Installer, Zookeeper, Oozie, Jaql, Lucene, Pig, Hive, MapReduce, Symphony, HBase, HDFS, Sqoop, Flume, Netezza, etc.

- **Administration consoles**: These are the tools used to manage Hadoop. Thanks to these management consoles, we can deploy, configure, automate, track, report, troubleshoot robustly and effortlessly and maintain.
- **Components of the solution**: The elements that make up each solution.

B. Comparison

In this part, we apply the five comparison criteria that we have proposed on the five Hadoop distributions and we group the result for each solution in Table 1.

Since we are in the era of Big Data, several large organizations have contributed to solutions in order to manage this large mass of data and to draw relevant information.

After the comparative study that was made on the five main suppliers of Hadoop distributions in order to manipulate the big data, the thing to note is that the majority of the suppliers offer distributions based on Apache Hadoop and projects open sources associated. They also provide a software solution that organizations can install on their own infrastructure on place in private cloud and/or public cloud.

Even if most of the components that form Hadoop are open source, it is interesting to note that there are several gains to pay a supplier in order to subscribe to a business offer of the platform. This subscription fee will give us for example access to technical support, to functions that are not available in the community version as well as to the training.

Currently, there is no absolute winner on the market because each of the suppliers focuses on the main features such as security, integration, the scale, the essential performance to the adoption of the business and governance.

VII. 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 namely HortonWorks, Cloudera, MapR,

IBM Infosphere BigInsights, pivotal HD, Microsoft HD Insight, etc. The work related to the comparative studies will help us to detect the common points and the specificities of the different distributions Hadoop of Big Data.

References

[1] HortonWorks Data Platform HortonWorks Data Platform: New Book. (2015).

[2] Menon, R. (2014). Cloudera Administration Handbook

[3] Dunning, T., & Friedman, E. (2015). Real-World Hadoop

[4] Quintero, D. (n.d.). Front cover implementing an IBM InfoSphere BigInsights Cluster using Linux on Power.

[5] Pivotal Software, I. (2014). Pivotal HD Enterprise Installation and Administrator Guide.

[6] Sarkar, D. (2014). Pro Microsoft HDInsight. Berkeley, CA: Apress.

[7] Thibaud Chardonnens, "Big Data analytics on high velocity streams: specific use cases with Storm", Software Engineering Group, Department of Informatics, University of Fribourg, Switzerland, 2013.

[8] McKinsey Global Institute. Big data: The next frontier for innovation, competition, and productivity. Paper, June 2011. 7, 9, 10, 11

[9] Nauman Sheikh, "Big Data, Hadoop, and Cloud Computing, Implementing Analytics", Morgan Kaufmann, 2013.

[10] C. Dobrea, and F. Xhafa b, "Intelligent services for Big Data science", Future Generation Computer Systems, Volume 37, 2014, pp. 267-281.

[11] Sawant, N., & Shah, H. (Software engineer). (2013). Big data application architecture & A a problem-solution approach. Apress.

[12] Lenovo, I. (2015). Lenovo Big Data Reference Architecture for Cloudera Distribution for Hadoop, (August).

[13] Read, W., Report, T., & Takeaways, K. (2016). The Forrester WaveTM: Big Data Hadoop Distributions, Q1 2016.

[14] Gates, Alan, and Daniel Dai. Programming Pig: Dataflow Scripting with Hadoop. 2 edition. O'Reilly Media, 2016.

[15] Capriolo, Edward, Dean Wampler, and Jason Rutherglen. Programming Hive: Data Warehouse and Query Language for Hadoop. 1 edition. Sebastopol, CA: O'Reilly Media, 2012.

[16] Ting, Kathleen, and Jarek Jarcec Cecho. Apache Sqoop Cookbook: Unlocking Hadoop for Your Relational Database. 1 edition. Sebastopol, CA: O'Reilly Media, 2013.

[17] Murthy, Arun, Vinod Vavilapalli, Douglas Eadline, Joseph Niemiec, and Jeff Markham. Apache Hadoop YARN: Moving beyond MapReduce and Batch Processing with Apache Hadoop 2. 1 edition. Upper Saddle River, NJ: Addison-Wesley Professional, 2014.