A Comparative Study of Hadoop-based Big Data Architectures

Allae Erraissi, Abdessamad Belangour, Abderrahim Tragha
Department of Mathematics and Computer Science
Faculty of Sciences Ben M'Sik, Casablanca, Morocco
erraissi.allae@gmail.com
belangour@gmail.com
atragha@yahoo.fr

ABSTRACT: *Big Data is a concept popularized in recent years to reflect the fact that organizations are confronted with large volumes of data to be processed and this, of course, presents a strong commercial and marketing challenge. This trend around the analysis and collection of Big Data has given rise to new solutions that combine traditional data warehouse technologies with Big Data systems in a logical architecture. It is important to note here that several distributions ready to use for managing a system Big Data are available in the market, namely HortonWorks, Cloudera, MapR, IBM Infosphere BigInsights, Pivotal HD, Microsoft HD Insight, etc. The different distributions have an approach and different positioning in relation to the vision of a platform Hadoop. In this article, we shall first explain the architectures and components of the five distributions of Hadoop solutions for Big Data. Then we shall present our comparative study in which we shall use 34 relevant criteria to define the strengths and weaknesses of the main Hadoop distribution providers.*

Keywords: Big Data, Architecture Distribution Hadoop, Comparison

Received: 13 June 2017, Revised 18 July 2017, Accepted 25 July 2017

1. Introduction

The spectacular development of social networks, the internet, the connected objects, and mobile technology is causing an exponential growth of data which all companies are confronted with. These technologies widely produce amounts of data which has to be collected, categorized, deployed, stored, analyzed, and so on. Hence, it appears an urgent need for a robust system capable of doing all the treatments within organizations. Consequently, the technology of big data began to flourish and several vendors offer ready-to-use distributions to deal with the Big Data, namely HortonWorks [1], Cloudera [2], MapR [3], IBM Infosphere BigInsights [4], Pivotal HD [5], Microsoft HD Insight [6], etc. In fact, each distribution has its own approach for a Big Data system, and the choice will be based on one or the other solution depending on several requirements. For example, if the solution is open source, the maturity of the solution, and so on. These solutions are Apache projects and therefore available. However, the interest of a complete package lies in the simplicity of installation, the compatibility between the components, the support and so on. In this article, we shall present our second comparative study on the five main Hadoop distribution providers in order to distinguish the strengths and weaknesses of each Hadoop distribution.

2. Hadoop Distributions of the Big Data

There are several distributions that permit to manipulate a Big Data system and to manage its main components: HortonWorks, Cloudera, MapR, IBM Infosphere BigInsights, Pivotal, Microsoft HDInsight, etc. In this part, we shall talk about the four best known and used distributions in the world of Big Data. These are: HortonWorks, Cloudera, Pivotal HD, and IBM Infosphere BigInsights.

2.1 HortonWorks Distribution

In 2011, members of the Yahoo team in charge of the Hadoop project formed HortonWorks. All components of this distribution are open source and licensed from Apache so as to facilitate the adoption of the Apache Hadoop platform. HortonWorks 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. This distribution is most consistent with Apache's Hadoop platform.

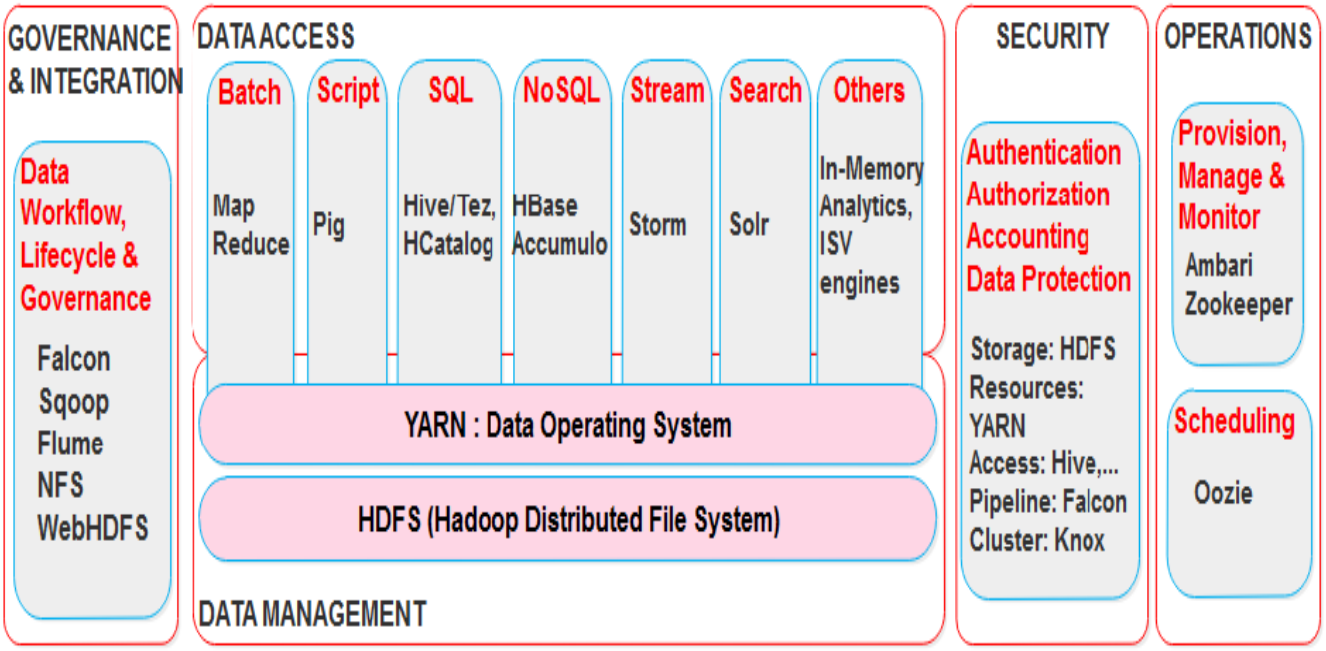

Figure 1. HortonWorks Hadoop Platform (HDP) [1]

The following elements make up the HortonWorks platform [1]:

- Heart Hadoop (HDFS/MapReduce) [10]
- Querying (Apache Hive [11])
- Integration services (HCatalog APIs [13], WebHDFS, Talend Open Studio for Big Data [14], Apache Sqoop [12])
- NoSQL (Apache HBase [15])
- Planning (Apache Oozie [16])
- Distributed Log Management (Apache Flume [17])
- Metadata (Apache HCatalog [13])
- Coordination (Apache Zookeeper [18])
- Learning (Apache Mahout [19])
- Script Platform (Apache Pig [20])
- Management and supervision (Apache Ambari [21])

2.2 Cloudera Distribution

Cloudera was firstly founded by Hadoop experts from Facebook, Google, Oracle and Yahoo. This distribution is largely based on the components of Apache Hadoop and it is complemented by essentially house components for cluster management. The aim of Cloudera's business model is not only to sell Licenses but to sell support and training as well. Cloudera offers a fully open source version of their platform (Apache 2.0 license) [2].

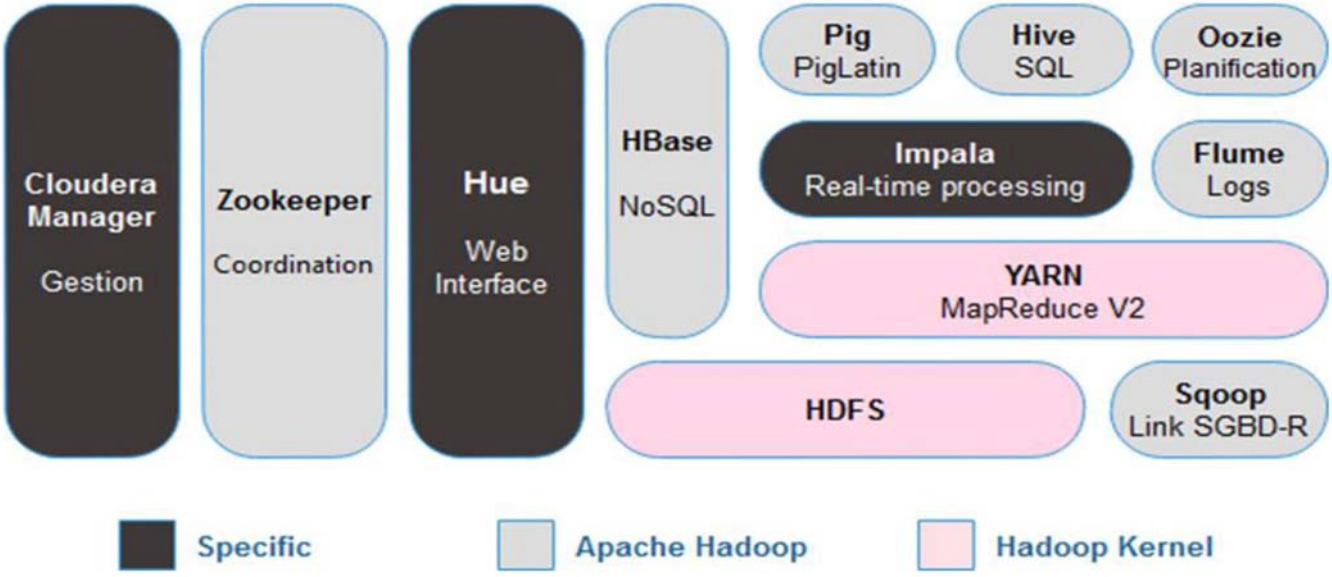

Figure 2. Cloudera Distribution for Hadoop Platform (CDH) [2]

2.3 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 console. In June 2013, IBM launched the Infosphere BigInsights Quick Start Edition. This new edition provided massive data volume analysis capabilities on a business-centric platform [4]. It both combines Apache Hadoop's Open Source solution with enterprise functionality and hence, provides a large-scale analysis, characterized by its resilience and fault tolerance. In short, this distribution supports structured, unstructured and semi-structured data and offers maximum flexibility.

2.4 Pivotal HD Distribution

Pivotal Software, Inc. (Pivotal) is a software and services company based in San Francisco and Palo Alto, California, with several other offices. The divisions include Pivotal Labs for consulting services, the Pivotal Cloud Foundry development group, and a product development group for the Big Data market. In March 2013, an Apache Hadoop distribution called Pivotal HD was announced, including a version of Greenplum software [24] called Hawq. Pivotal HD Enterprise is a commercially supported distribution of Apache Hadoop [5]. The figure below shows how each Apache and Pivotal component integrates into the overall architecture of Pivotal HD Enterprise:

3. Comparison between Distributions

In effect, we earlier carried out a comparative study of the Hadoop distributions architecture of Big Data for the purpose of making an evaluation between the distributions. Our main goal was to identify the strengths and weaknesses of the five major Hadoop distribution providers: Cloudera, HortonWorks, IBM InfoSphere BigInsights, MapR and Pivotal HD. Indeed, this work is an advanced analysis of the first comparative study we made before [27]. We based our work on three principal studies. The first one is the evaluation made by Forrester Wave [7] of the same Hadoop distributions, in which they used 35 evaluation criteria grouped into three high-level buckets: Current Offering, Strategy, and Market presence. The two other comparative studies are those proposed by Robert D. Schneider [8] and V.Starostenkov [9] on the three HortonWorks, Cloudera, and MapR distributions.

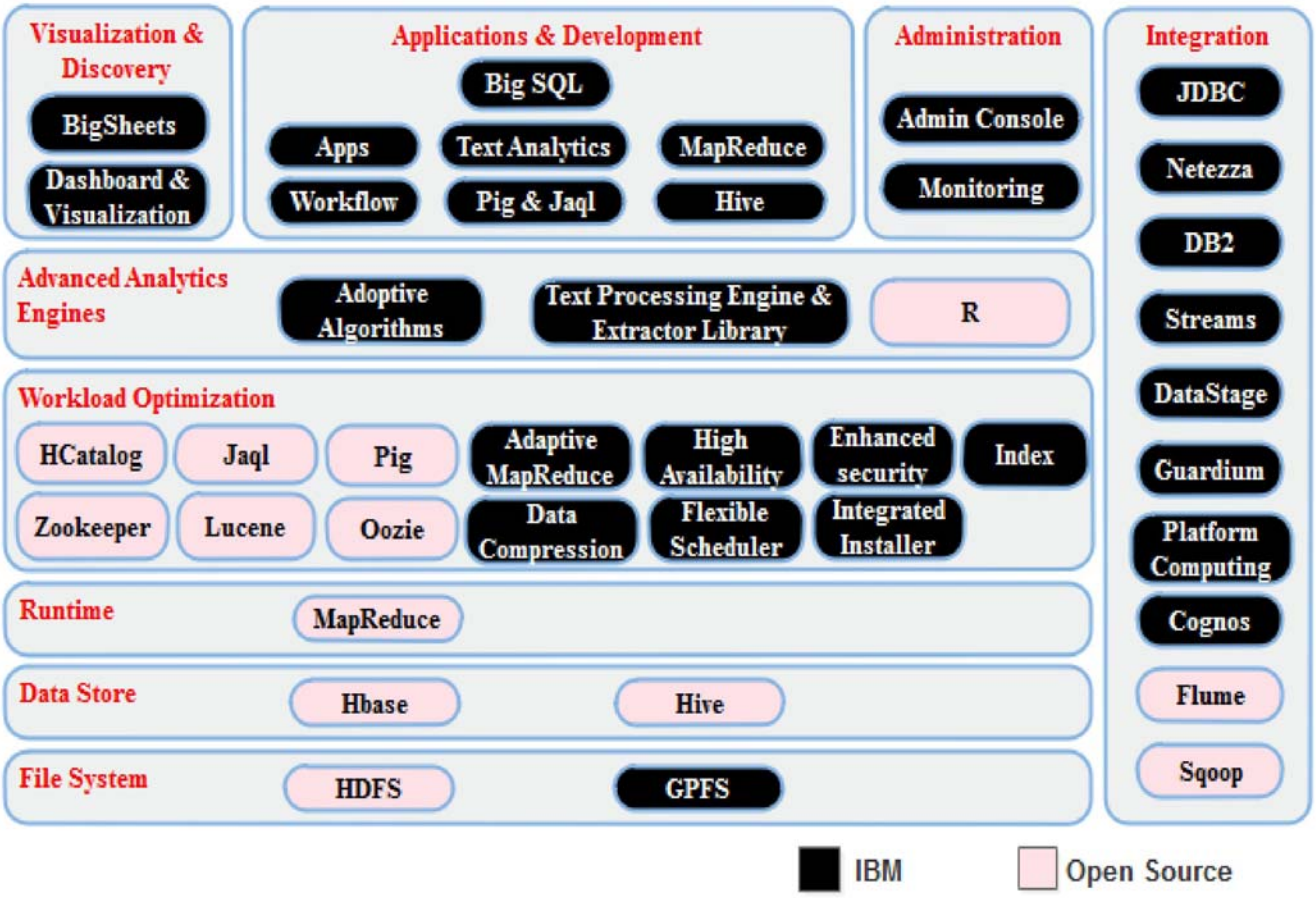

Figure 3. IBM InfoSphere BigInsights Entreprise Edition [4]

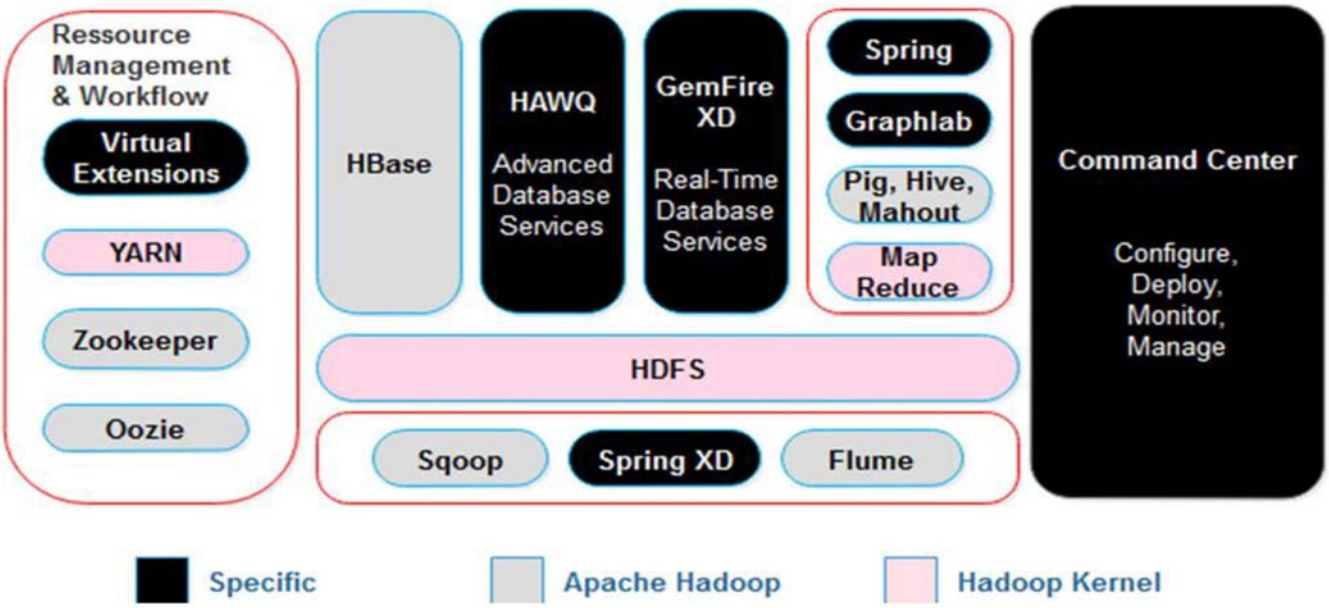

Figure 4. Pivotal HD Entreprise [5]

In this analysis, we shall propose 34 relevant criteria to try to distinguish and differentiate the different architectures available for the five distributions of Big Data solutions.

3.1 Criteria for Comparison

To compare the five distributions, we shall use the following criteria:

• **Disaster Recovery:** It can help prevent data loss in the event of a computer center failure. It has the possibility to rebuild the infrastructure and to restart applications that support the activity and survival of a company. Therefore, Disaster Recovery must be able to take care of the computer needs necessary for the survival of the organization in case of a Big Data system disaster.

• **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.

• **Management Tools:** These are management consoles used by different Hadoop solution providers to manage a Hadoop distribution. Thanks to these tools, you can effortlessly deploy, configure, automate, report, track, troubleshoot, and maintain a Big Data system.

• **Data and Job Placement Control:** It allows to control the placement of data and jobs on a Hadoop cluster and, hence, permets to choose nodes to execute jobs presented by different users and groups.

• **Heatmaps, Alarms, Alerts:** These are tools and notifications that allow to keep a global view of our Big Data system. The term Heatmap means a graphical representation of the data as a color, and a color block represents each host in the cluster.

• **DFS:** c'est le système de fichiers distribué pour le stockage.

• **Security ACLs:** It is a list of permissions attached to an object. An ACL specifies processes or system users which are allowed to access objects, and define clearly permissible operations on the object.

• **Data Ingestion:** Data Ingestion 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 ingested in batches. When it is ingested in real time, it is imported as it is transmitted by the source. However, when data is ingested in batches, it is imported in discrete blocks at periodic time intervals.

• **MetaData Architecture:** here we shall talk about two types of architectures used by the Hadoop distributions at MetaData level. The first one is a centralized architecture where everyone depends on the same authority. The second one is a decentralized architecture that has no center, no more and no less. This means that every entity can be a part of a network that has no main authority and that these authorities can talk to each other.

• **MapReduce:** It is the dedicated programming model for making parallel and distributed computations of potentially very large data.

• **Apache Hadoop YARN** [25] (Yet Another Resource Negotiator) is a technology for managing clusters and making Hadoop more suitable for operational applications that cannot wait for the completion of batch processing. YARN is among the key features of Hadoop 2, the second generation of the distributed processing infrastructure of Apache Software Foundation.

• **Non-Relational Data Base:** These are databases that do not include the key/table model that relational database management systems (RDBMS) use. These databases put into action data manipulation techniques and dedicated processes so as to provide solutions to the major data issues that great organizations are confronted with. The most popular emerging non-relational database is NoSQL (Not Only SQL).

• **Meta Data Services:** It is an object-oriented reference technology that can be integrated into enterprise information systems or into applications that use the MetaData process.

• **Scripting Platforms:** These are dedicated platforms for programming languages that use high-level constructs to interpret and

execute one command at a time.

• **Data Access and Query:** queries are utilized by users to express their information needs and to access data.

• **Workflow Scheduler:** It is a representation of a sequence of operations or tasks carried out by a person, a group of people or an organization. It refers to the passage of information from one episode to another.

• **Coordination Cluster:** In order to avoid gaps and overlaps in Cluster work, coordination aims to ensure a coherent and complementary approach by identifying ways to work together for the purpose of achieving better collective outcomes.

• **Bulk Data Transfer between RDB and Hadoop:** These are tools designed to efficiently transfer data between Apache Hadoop and structured data stores such as relational databases.

• **Machine Learning:** It concerns the analysis, design, development, and implementation of methods that allow a machine to evolve through a systematic process, and consequently, to solve problems by more conventional algorithmic means.

Data Analysis: It allows to process a large number of data and to identify the most interesting aspects of the structure of these data. Besides, it eventually provides graphical representations which can reveal relations that are difficult to grasp by the direct data analysis.

• **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. Cloud services are characterized by their great flexibility.

• **Parallel Query Execution Engine:** These are parallel query execution engines, intending to optimize the execution of queries and indexes.

• **Full-text Search:** It is a search technique in a document or database on all the words. This technique tries to match the words to those provided by the users.

• **Data Warehousing:** Means a database used to collect, order, log and store information from operational databases. It also provides a basis for business decision support.

• **Extract, Transform and Load (ETL):** It is a computer technology known as ETL that allows massive synchronization of information from one data source to another.

• **Authentication:** It is a process allowing access to the resources of an information system by an entity. It permits the system to validate the legitimacy of the access of the entity. After this, the system assigns this entity the identity data for that session.

• **Authorization:** It determines whether the authenticated subject can place the desired action on the specified object.

• **Accountability:** It is an obligation to report and explain, with an idea of transparency and traceability, by identifying and documenting the measures implemented mainly for the purpose of complying with the requirements of the IT and freedoms regulations.

• **Data Protection:** It urges the data controller in the Big Data distributions to adopt internal rules and to implement appropriate measures to guarantee and demonstrate that the processing of personal data is carried out in compliance with the IT regulations and freedoms.

• **Provise, Manage & Monitor:** These are tools for configuration management, management and monitoring of the computer system. Provisioning allows you to remotely install and configure software, allocate disk space, power, or memory. Monitoring is the permanent monitoring of the computer system for a preventive purpose. It allows it to be alerted in case of abnormal operations detections.

• **Scheduler:** It to define the links between the processes and the way to launch them. The notion of processing can be quite

general since it is any executable command on one or more computing machines.

3.2 Comparison

This table clusters the comparative study carried out as well as the results for each Hadoop distribution.

Criteria \ Distributions		Horton Works	Clou-dera	Map-R	IBM BigInsights	Pivotal HD
Disaster recovery		-	+	+	+	+
Replication Data		+	+	+	+	+
Replication MetaData		-	-	+	+	+
Management tools		+	+	+	+	+
Data and Job Placement Control		-	-	+	+	-
Heatmaps, Alarms, Alerts		+	+	+	+	+
DFS		+	+	+	+	+
Security ACLs		+	+	+	+	+
Data Ingestion	Batch	+	+	+	+	+
	Streaming	-	-	+	+	+
MetaData Architecture	Centralized	+	+	-	-	-
	Distributed	-	-	+	+	+
Map Reduce		+	+	+	+	+
Non-Map Reduce Tasks (YARN)		+	+	+	+	+
Non-Relational Data Base		+	+	+	+	+
Meta Data Services		+	+	+	+	+
Scripting platform		+	+	+	+	+
Data Access and Query		+	+	+	+	+
Workflow scheduler		+	+	+	+	+
Cluster coordination		+	+	+	+	+
Bulk Data transfer between RDB and Hadoop		+	+	+	+	+
Distributed Log Management services		+	+	+	+	+
Machine learning		+	+	+	+	+
Data Analysis		+	+	+	+	+
Cloud services		+	+	+	+	+
Parallel Query Execution Engine		+	+	+	+	+
Full-Text search		+	+	+	+	+
Data warehousing		+	+	+	+	+
Extract, Transform and Load (ETL)		+	+	+	+	+
Data Interaction and Analysis		+	+	+	+	+
Authentication		+	+	+	+	+
Authorization		+	+	+	+	+
Accountability		+	+	+	+	+
Data Protection		+	+	+	+	+
Provise, Manage & Monitore		+	+	+	+	+
Scheduling		+	+	+	+	+

Table 1. Comparison between the five distributions Hadoop for Big Data

4. Discussion

Prior to starting our discussion, it is of utmost importance to point out briefly that several large organizations have contributed to the establishment of several distributions only for the purpose of managing large Big Data and of drawing valuable information drowned in the mass.

Accordingly, we based our comparative study on these distributions, mainly on the architectures of the different Hadoop distribution providers in the Big Data. We rely on 34 relevant criteria that must have any solution to manage and administer clusters, as well as to collect, sort, categorize, move, analyze, store, and process Big Data. At this point we eventually draw some conclusions. Firstly, we found out that the majority of providers offer distributions based on Apache Hadoop and associated open source projects, and that they give a software solution that organizations can install on their own infrastructure on-site in private cloud and/or public cloud. We also found out that most of the five different distributions are based on the majority of the criteria we have proposed and in this context, we deduce that there is not really an absolute winner in the market since each supplier focuses on main features dedicated to Big Data systems such as integration, security, scale, performance critical to business adoption and governance.

5. Conclusion

In short, The Big Data refers to the explosion of the volume of data in companies and to the technological means proposed by the publishers to answer them. It also includes all the technologies for storing, analyzing and processing heterogeneous data and content, in order to bring out added value and wealth. This trend around the collection and analysis of Big Data has given rise to new distributions to manage a Big Data system. Seeing that there are several distributions that facilitate the adoption of Apache's Hadoop platform and manage clusters namely Cloudera, HortonWorks, MapR, IBM Infosphere BigInsights, Microsoft HD Insight, Pivotal HD, and so on. The work related to comparative studies will help us to detect the common features and the specificities of the main Hadoop distributions of Big Data in order to try to standardize the concepts of Big Data in our next works.

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

[8] Schneider, R. D. (2014). HADOOP BUYER'S GUIDE.

[9] Starostenkov, V., Senior, R., Developer, D. (2013). Hadoop Distributions: Evaluating Cloudera, Hortonworks, and MapR in Micro-benchmarks and Real-world Applications.

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

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

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

[13] Wall, L. (2015). About the Tutorial Copyright & Disclaimer, p. 2.

[14] Barton, Daniel, Rick. (2013). Talend Open Studio Cookbook. Birmingham, UK : Packt Publishing, 2013.A

[15] Vohra, Deepak. (2016). Apache HBase Primer. 1st ed. edition. New York, NY : Apress.

[16] Islam, Kamrul, Mohammad., Srinivasan, Aravind. (2015). Apache Oozie: The Workflow Scheduler for Hadoop. 1 edition. Sebastopol: O'Reilly Media.

[17] Hoffman, Steve. (2015). Apache Flume: Distributed Log Collection for Hadoop - Second Edition. 2nd edition. Birmingham, England; Mumbai, India: Packt Publishing - ebooks Account.

[18] Bagai, Chandan. (2016). Characterizing & Improving the General Performance of Apache Zookeeper: Sub-Project of Apache Hadoop. LAP LAMBERT Academic Publishing.

[19] Lyubimov, Dmitriy, Palumbo, Andrew. (2016). Apache Mahout: Beyond MapReduce. 1 edition. CreateSpace Independent Publishing Platform.

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

[21] Eadline, Douglas. (2015). Hadoop 2 Quick-Start Guide: Learn the Essentials of Big Data Computing in the Apache Hadoop 2 Ecosystem. 1 edition. New York: Addison-Wesley Professional.

[22] Covert, Michael, Loewengart, Victoria. (2015). Learning Cascading. Birmingham: Packt Publishing - ebooks Account.

[23] Dunning, Friedman, Ellen., Tomer Shiran Ted. (2016). Apache Drill: The SQL Query Engine for Hadoop and NoSQL. 1 edition. O'Reilly Media.

[24] Gollapudi, Sunila. (2013). Getting Started with Greenplum for Big Data Analytics. Birmingham, UK: Packt Publishing.

[25] Alapati, Sam, R. (2016). Expert Hadoop Administration: Managing, Tuning, and Securing Spark, YARN, and HDFS. 1 edition. Boston, MA: Addison-Wesley Professional.

[26] Russell, John. (2014). Getting Started with Impala: Interactive SQL for Apache Hadoop. 1 edition. Sebastopol, CA: O'Reilly Media.

[27] Erraissi, Allae., Belangour, Abdessamad., Tragha, Abderrahim. (2017). A Big Data Hadoop Building Blocks Comparative Study." *International Journal of Computer Trends and Technology.* Accessed June 18, 2017. http://www.ijcttjournal.org/archives/ijctt-v48p109.