Digging into Hadoop-based Big Data Architectures

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

[1,2,3] Laboratory of Information Technology and Modeling LTIM,
Hassan II University, Faculty of Sciences Ben M'sik, Casablanca, Morocco

Abstract

During the last decade, the notion of big data invades the field of information technology. This reflects the common reality that organizations have to deal with huge masses of information that need to be treated and processed, which represents a strong commercial and marketing challenge. The analysis and collection of Big Data have brought about solutions that combine traditional data warehouse technologies with the systems of Big Data in a logical and coherent structure. Thus, many vendors offer their own Hadoop distributions such as HortonWorks, Cloudera, MapR, IBM Infosphere BigInsights, Pivotal HD, Microsoft HD Insight, and so on. Their main purpose was to supply companies with a complete, stable and secure Hadoop solution for Big Data. They even compete with each other's to find efficient and complete solutions to satisfy their customers need and, hence, make benefit from this fast-growing market. In this article, we shall present a comparative study in which we shall use 34 relevant criteria to determine the advantages and drawbacks of the most outstanding Hadoop distribution providers.

Keywords: *Big Data, Big Data distributions, Hadoop Architectures, comparison, Big Data solutions comparison.*

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 have to handle. These technologies widely produce amounts of data, which Data analysts have to collect, categorize, deploy, store, analyze 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 build ready-to-use distributions to deal with the Big Data, namely HortonWorks [1], MapR [3], Cloudera [2], 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 customer will choose between the different solutions relying on several requirements. For example, he will consider if the solution is open source, or if it is a mature one, and so on. These solutions are Apache projects and therefore available. However, the success of any distribution lies in the suppleness of the installation, the compatibility between the constituents, the support and so on. This work is an advanced analysis of the first comparative study we made before [27]. In this paper, we shall present our second comparative study on the five main Hadoop distribution providers in order to explore the benefits and challenges of each Hadoop distribution and consider them.

2. Hadoop Distributions of the Big Data

Given the challenge that lies ahead with the endless growth of data, many distributions emerge data processing field to handle the rising capacity demands of a Big Data system. The most outstanding of these solutions are HortonWorks, Cloudera, MapR, IBM Infosphere BigInsights, Pivotal and Microsoft HD Insight. Indeed, our work in this part will focus on the five best known and used distributions, which are HortonWorks, Cloudera, MapR, Pivotal HD, and IBM Infosphere BigInsights.

2.1 HortonWorks distribution

In 2011, members of the Yahoo team firstly found HortonWorks. All the components of this distribution are open source and licensed from Apache. Yet, the defined objectives of this distribution are to simplify the adoption of the Apache Hadoop platform. HortonWorks is a big Hadoop contributor and Hadoop vendors provide a commercial model where they sell licenses with technical support and training services. Apache's Hadoop platform and HortonWorks solution are compatibles with each other's.

https://doi.org/10.20943/01201706.5259

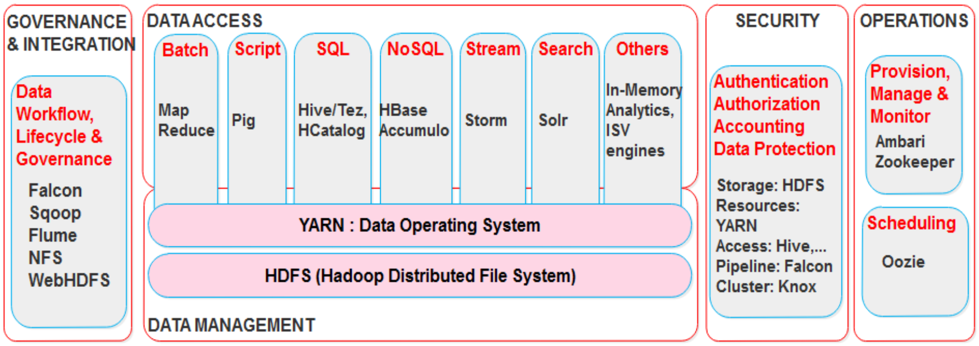

Figure 1: HortonWorks Hadoop Platform (HDP) [1]

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

Hadoop experts from Facebook, Google, Oracle and Yahoo succeed in finding Cloudera. This distribution includes the components of Apache Hadoop and it succeeds to develop effectively house components for cluster management. The aim of Cloudera's business model is not only to sell customers Licenses but also to sell them training and support services as well. Cloudera provides a fully open source version of their platform (Apache 2.0 license) [2]. The following elements make up the Cloudera platform [2]:

Apache Components:

- HDFS: Hadoop Distributed File System.
- MapReduce: Parallelized processing framework.
- HBase: NoSQL database.
- Hive: SQL query type.
- Pig: Scripting and Hadoop query.
- Oozie: Workflow and planning of Hadoop jobs.
- Sqoop: transfer of data between Apache Hadoop and the relational databases.
- Flume: Exploiting files (log) in Hadoop.
- Zookeeper: Coordination service for distributed applications.
- Mahout: Hadoop learning and data mining framework.

Original components Cloudera:

- Hadoop Common: A set of common utilities and libraries that support other Hadoop modules.
- Hue: SDK to develop user interfaces for Hadoop applications.
- Whirr: Libraries and scripts for running Hadoop and related services in the Cloud.

Components not Apache Hadoop:

- Impala Cloudera: “is Cloudera’s open source massively parallel processing (MPP) SQL query engine for data stored in a computer cluster running Apache Hadoop” [26].
- Cloudera Manager: Deployment and management of Hadoop components.

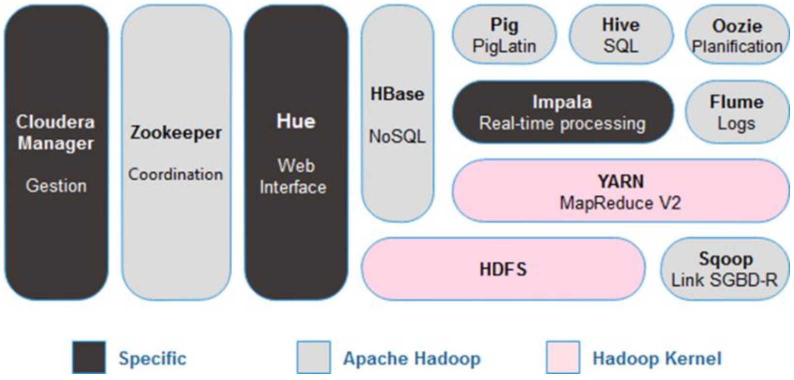

Figure 2: Cloudera Distribution of Hadoop Platform (CDH) [2]

2.3 MapR distribution

In 2009, former genius members of Google develop MapR distribution. It mainly contributes to Apache Hadoop projects like HBase, Hive, Pig, Zookeeper and especially Drill [3]. They successfully propose their own version of MapReduce and distributed file system: MapR FS and MapR MR [3]. Thus, three versions of their solution are available:

- M3: Open source version.
- M5: Version that adds high-availability features and support.
- M7: Includes an optimized HBase environment.

The MapR M3 distribution is an open source version that consists of [3]:

Apache Components:

HBase, Pig, Hive, Mahout, Cascading [22], Sqoop, Flume.

Other components:

- MapR FS: MapR proposes its own file system by replacing the HDFS.
- MapR Control System (MCS): MCS allows management and supervision of the Hadoop cluster. It is a web-based tool for managing cluster resources (CPU, Ram, Disk, etc.) as well as services and jobs.
- Apache Cascading: Java framework dedicated to Hadoop. It allows a Java developer to find his brands (JUnit, Spring, etc.) and to manipulate the concepts of Hadoop with a high-level language without knowing the API.
- Apache Vaidya: Hadoop Vaidya is a performance analysis tool for MapReduce jobs.
- Apache Drill [23]: Drill completes MapReduce and is an API for faster query creation based on the SQL model.

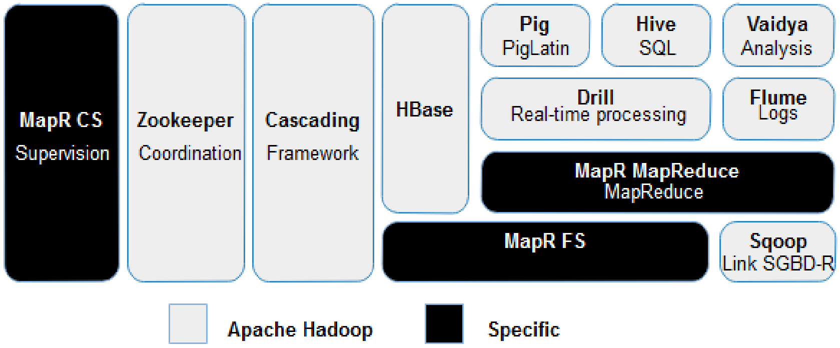

Figure 3: MapR (M3) [3]

2.4 IBM InfoSphere BigInsights distribution

In 2011, IBM professionals develop InfoSphere BigInsights for Hadoop 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 provides massive data analysis capabilities on a business-centric platform [4]. It both combines Apache Hadoop's Open Source solution with company performance and hence, give way to a large-scale analysis, marked by fault tolerance and resilience. In short, this distribution supports structured, unstructured and semi-structured data and offers maximum flexibility.

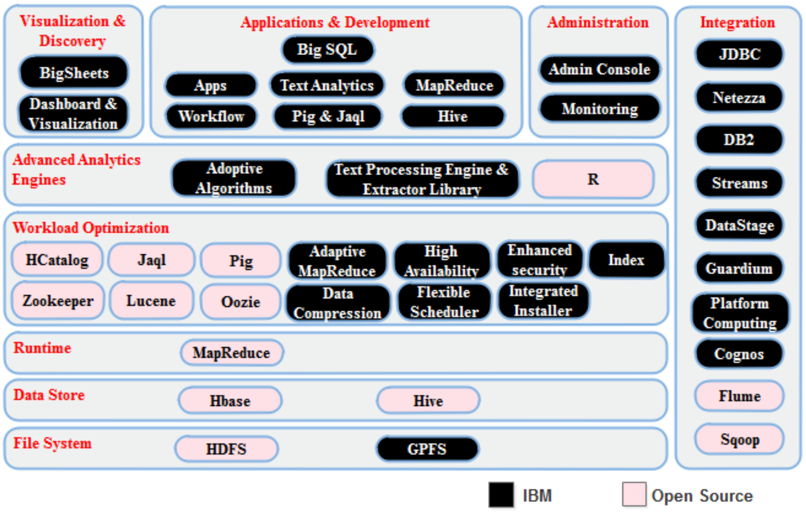

Figure 4: IBM InfoSphere BigInsights Enterprise Edition [4]

2.5 Pivotal HD distribution

Pivotal Software Inc is a Software company, which is headquartered in San Francisco, California. Its main offices involve Pivotal Lab, the Pivotal Cloud Foundry development group, and a product development group for the Big Data market. It was not until 2013 that developers have built the Apache Hadoop distribution called Pivotal HD. This distribution includes a version of Greenplum software [24], also called Hawq. In short, 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:

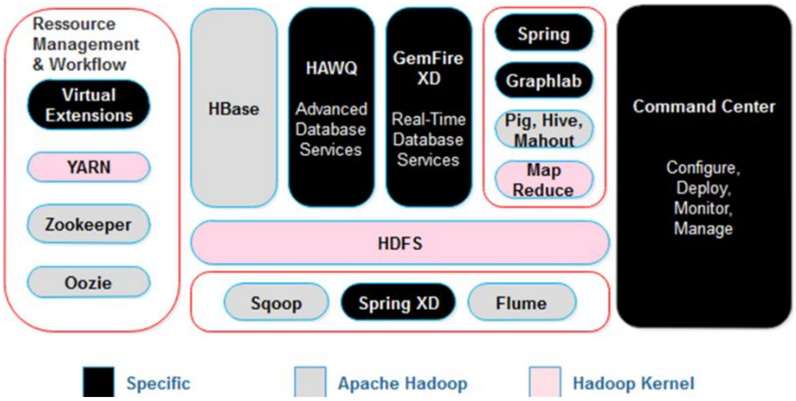

Figure 5: Pivotal HD Entreprise [5]

3. Comparison between distributions

In effect, we earlier carried out a comparative study of the Hadoop distributions architecture of Big Data. Our main goal was to make an evaluation between the distributions and thus provide the advantages and the drawbacks 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 base our subject on three principal studies. The first one is the

evaluation made by Forrester Wave [7] of the five Hadoop distributions. They make use of 35 evaluation criteria, which they divide into three high-level buckets: Current Offering, Strategy, and Market presence. The two other studies are those proposed by Robert D. Schneider [8] and V.Starostenkov [9] on the three HortonWorks, Cloudera and MapR distributions.
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 prevent data loss in the event of a computer center failure. It has the ability to rebuild the infrastructure and to restart applications that support the activity 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**: The different Big Data Hadoop distributions use a process of information sharing to improve reliability, fault tolerance and availability. They also aim 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 controlling the placement of data and jobs on a Hadoop cluster and, hence, permits to choose nodes to execute jobs presented by different users and groups.
- **Heat maps, Alarms, Alerts**: These tools and notifications allow keeping a global view of our Big Data system. The term Heat map means a graphical representation of the data as a color, and a color block represents each host in the cluster.
- **DFS**: Distributed file system for storage.
- **Security ACLs**: It is a list of permissions attached to an object. An ACL specifies and processes system users. These system users may access objects, and define clearly permissible operations on that 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 databases do not include the key/table model that relational database management systems (RDBMS) use. They practice data manipulation techniques and dedicated processes. Their main goal is to provide solutions to the major data issues that confront great organizations. The most popular emerging non-relational database is NoSQL (Not Only SQL).
- **Meta Data Services:** is an object-oriented reference technology. Suppliers integrate this technology into information systems of enterprises or into applications that use the Metadata process.
- **Scripting platforms:** They are dedicated platforms for programming languages that use high-level constructs to interpret and execute one command at a time.

- **Data access and query:** Users utilize queries 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**: It attempts to avoid gaps and overlaps in Cluster work and aims to ensure a coherent and complementary approach to identify ways to work together. Its primary goal is to achieve better collective outcomes.
- **Bulk Data Transfer between RDB and Hadoop**: These are tools designed to transfer efficiently data between Apache Hadoop and structured data stores such as relational databases.
- **Distributed Log Management services**: These services aim to handle large volumes of log messages in a distributed manner. They typically cover connection collection, distributed log aggregation, long-term log storage, log analysis, Log search and reporting, and so forth.
- **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**: Cloud services, also called dedicated services, exploit the computing and storage power of remote computer servers via a network that is the internet. They 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. It 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 the processing of personal data, which 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.

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

Criteria \ Distributions		Horton Works	Cloudera	MapR	IBM BigInsights	Pivotal HD
Disaster recovery		-	+	+	+	+
Replication Data		+	+	+	+	+
Replication Meta Data		-	-	+	+	+
Management tools		+	+	+	+	+
Data and Job Placement Control		-	-	+	+	-
Heatmaps, Alarms, Alerts		+	+	+	+	+
DFS		+	+	+	+	+
Security ACLs		+	+	+	+	+
Data Ingestion	Batch	+	+	+	+	+
	Streaming	-	-	+	+	+
Meta Data 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		+	+	+	+	+

4. Discussion

Prior to starting our discussion, it is of utmost importance to point out briefly that many companies play a leading role in the establishment of several distributions. The main object of these distributions is to manage large Big Data, draw valuable information from the mass and provide digital transformation technology and services.

Accordingly, we based our comparative study on these distributions, mainly on the structure of the various 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, most providers have created their own distributions relying on Apache Hadoop and associated open source projects. They equally give a software solution that numerous organizations can benefit from by installing it on their own infrastructure on-site in private cloud and/or public cloud. We eventually found out that most of the five different distributions are based on the majority of the criteria we have proposed. In this context, we deduce that there is not really an absolute winner in the market since each supplier focuses on major 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 benefit and wealth. This tendency of Big Data collection and processing has given rise to new distributions designed to manage a Big Data system. The aim of these distributions is to pave the way for the adoption of Apache's Hadoop platform and to manage clusters primarily Cloudera, HortonWorks, MapR, IBM Infosphere BigInsights, Microsoft HD Insight, Pivotal HD, and so forth. The work related to our comparative studies leads us to examine and detect the common features and characteristics of the main Hadoop distributions of Big Data in order to seek 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] R. D. Schneider, "HADOOP BUYER'S GUIDE," 2014.

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

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

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

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

[13] L. Wall et al., "About the Tutorial Copyright & Disclaimer," p. 2, 2015.

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

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

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

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

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

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

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

[21] Eadline, Douglas. 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, 2015.

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

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

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

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

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

[27] Allae Erraissi, Abdessamad Belangour, Abderrahim Tragha. "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.

Allae Erraissi is a Ph.D. student of computer science at the Faculty of Sciences at the Hassan II University, Casablanca, Morocco. He won his Master Degree in Information Sciences and Engineering from the same University in 2016 and is currently working as Mathematics teacher in a High school in Casablanca, Morocco. His main interests are the new technologies namely Model-driven engineering, Cloud Computing and Big Data.

Abdessamad Belangour is an Associate Professor at the Faculty of Sciences at the Hassan II University, Casablanca, Morocco. He is mainly working on Model Driven Engineering approaches and their applications on new emerging technologies such as Big Data, Business Intelligence, Cloud Computing, Internet of Things, Real-time embedded systems etc.

Abderrahim Tragha is a Full Professor at the Faculty of Sciences at the Hassan II University, Casablanca, Morocco. He is specialized in cryptography and is recently interested in Automatic Language Processing applied on The Arabic language, in Model-driven engineering and in Big Data.