Big Data Analytics: A Preliminary Study of Open Source Platforms

Jorge Nereu[1], Ana Almeida[1] and Jorge Bernardino[2]
[1]*Computer Engineering Department (DEI), ISEP, Polytechnic of Porto, Porto, Portugal*
[2]*ISEC-CISUC, Polytechnic of Coimbra, Coimbra, Portugal*

Keywords: Big Data Analytics, BI, Open Source Big Data Platforms.

Abstract: Nowadays organizations look for Big Data as an opportunity to manage and explore their data with the objective to support decisions within its different operational areas. Therefore, it is necessary to analyse several concepts about Big Data Analytics, including definitions, features, advantages and disadvantages. By investigating today's big data platforms, current industrial practices and related trends in the research world, it is possible to understand the impact of Big Data Analytics on smaller organizations. This paper analyses the following five open source platforms for Big Data Analytics: Apache Hadoop, Cloudera, Spark, Hortonworks, and HPCC.

1 INTRODUCTION

Nowadays we observe huge volumes of data in constant growth, due to the evolution of technology together with the massive exchange of information. Therefore it is essential to make use of sophisticated platforms to deal with this massive quantity of data.

There are two types of platforms available for handling Big Data - Open Source and Proprietary Software - which are used by organizations to manage their information. However, many of the organizations do not know the benefits, advantages, and disadvantages that these platforms offer in cost, operation, and information management.

In recent times all types of organizations are present on the Internet, and this channel has a great impact on their business, taking care of what customers want and also serving as a guide for new products and what is offered. This process also highlights the huge deal of information on what has to do with products and services for sale.

This is the main reason why this research work is carried out to analyse in particular the Open Source platforms for analytics that best fit in Small and Medium-sized Enterprises (SMEs) and Non-governmental organizations (NGO).

Currently, organizations and companies have opted for the adoption of open source and proprietary software platforms oriented to Big Data to solve problems of handling, management, storage, and analysis of information.

In order to justify this work, an analysis will be carried out between the open source platforms that can be adopted by SMEs and that cannot or do not wish to acquire proprietary platforms.The objective is to discover what kind of platforms and tools would be most suitable for their environment.

This paper analyses the following open source platforms for Big Data Analytics: Apache Hadoop, Cloudera, Spark, Hortonworks, and HPCC.

The rest of this paper is structured as follows. Section 2 presents the related work, and section 3 describes Big Data and Analytics. In section 4 we describe the analysed platforms for Big Data Analytics. Section 5 presents a comparision of the main features of the analysed platforms. Finally, conclusions and future work are summarized in Section 6.

2 RELATED WORK

Multiple research works have been done to compare and evaluate existing Big Data platforms with some research focused on a specific capability, technology or purpose (Lapa et al., 2014), (Bernardino, 2011/ 2015), (Neves and Bernardino, 2015).

Almeida and Bernardino (2015) focus on the capability of mining data, and in a mix of technical parameters and features that are suitable for Small and Medium Enterprise environments.

Nereu, J., Almeida, A. and Bernardino, J.
Big Data Analytics: A Preliminary Study of Open Source Platforms .
DOI: 10.5220/0006470104350440
In *Proceedings of the 12th International Conference on Software Technologies (ICSOFT 2017)*, pages 435-440
ISBN: 978-989-758-262-2

On the other hand, Morshed et al. (2016) focused their work on platforms addressing distributed real-time data analytics and concluded that the platforms analysed do not cover all the features that are required for distributed computation in real-time.

Miller et al. (2016) works on platforms written in SCALA programming language that supports both the object-oriented and functional programming paradigms built on top of JAVA.

Landset et al. (2015) presented a comprehensive survey of open source tools for machine learning with big data in the Hadoop ecosystem to researchers or professionals in machine learning but is inexperienced with big data.

(Sagiroglu and Sinanc (2013) provides an overview of big data such as samples, methods, advantages and challenges. They compare Hadoop and HPCC by their architectures, primary languages, and indexes in a Distributed File System, data warehouse abilities and performance tests where HPCC shows the best results.

Another recent paper describes an experiment with 40-node using Hadoop Platforms (Hortonworks, Cloudera or Apache), Spark for streaming data processing, HBase and OpenTSDB to store time series sensor data. The authors present the characteristics, requirements, and configurations of Hadoop platforms (Liu et al., 2016).

Consequently, there exist few works which do an evaluation based on specific capability, technology or purpose. Our work contributes to the identification of the Big Data platforms for analytics that may be suitable for SMEs in their operations.

3 BIG DATA AND ANALYTICS

Organizations find it difficult to perform a detailed analysis and provide new advantages and opportunities to their stakeholders. Some collected data which ranges from customers' names, addresses, available products, purchases as well as the employees recruited, has become very important for daily operations ("Ventana Research," 2014).

With this data, it is even more evident that technology is imperative in data storage and its recovery. Technological developments contribute to an increase in capabilities to store more data as well as more methods of collecting this data. Additionally, huge amounts of data have been made easily accessible (Inoubli et al., 2016).

Presently, organizations explore large data volumes that are highly detailed to discover the facts that they were not aware of initially.

Big Data provides government and business organizations new ways to combine miscellaneous digital data sets and after that, use statistics and other data mining techniques to extract from them both occult information and astonishing correlations (Rubinstein, 2012). In short, Big Data is described as an enormous volume of structured, semi-structured and unstructured data that is so big that it is difficult or impossible to process using traditional database systems and software techniques.

3.1 Big Data Analytics

Big Data Analytics is becoming a trending practice that many companies are adopting to build valuable information (Sivarajah et al., 2017). The main objective of Big Data Analytics is to become an asset for making business decisions as well as for data scientists and other analytics professionals to analyse enormous volumes of transaction data.

Platforms oriented to Big Data Analytics are the greatest promoters of the paradigm shift of Big Data. These platforms manage large volumes of data and also work as an application of various analytical techniques for large volumes of data (Miller et al., 2016). To extract useful information from large data volume tools, it is appropriate to collect, store and process data from various analytical perspectives (Prasad and Agarwal, 2016).

3.2 Big Data Ecosystems

The ecosystem of big data includes several aspects such as data, the lifecycle models of big data, and finally the infrastructure that is used for support (Murthy and Bowman, 2014).The maturity of big data and predictive analysis leads to more open source contributors to the technologies used to empower the solutions. Presently, all types and sizes of vendors are making use of open sources for big data processing and the predictive analytics process (Pääkkönen and Pakkala, 2015). In some cases, the cloud, as well as open sources for storage and computing, are the technological catapults that enable start-ups and the emergence of small companies to compete with the more established ones (Sen et al., 2016). Big Data open source platforms are divided into several categories, which are data storage and access, development tools, and platforms for analytics and reporting (Miller et al., 2016).

In the next section, we will analyse five of the most popular open source big data platforms.

4 BIG DATA PLATFORMS

A Big Data platform should be a solution that is specifically designed to meet the needs of one organization (Chandrasekhar et al., 2013).

The next section describes the characteristics of five most popular platforms for Big Data (Landset et al., 2015): Apache Hadoop, Cloudera, Spark, Hortonworks, and HPCC.

4.1 Apache Hadoop

The Apache Hadoop is a free software project of the Apache foundation that implements the MapReduce paradigm and the Hadoop Distributed File System (HDFS). This open source platform allows distributed processing of large data sets across clusters of servers using simple programming models, where one cluster is designated as the master node and other as a slave node (Prasad and Agarwal, 2016). This platform has been projected to scale from one server to thousands of servers where each has its own local processing and storage ("Apache™ Hadoop®," 2016).

The two most important functions that characterize the platform are MapReduce and HDFS, where MapReduce supports analysis of data and HDFS supports storage of data (Saraladevi et al., 2015). HDFS is at the base of the architecture as shown in Figure 1.

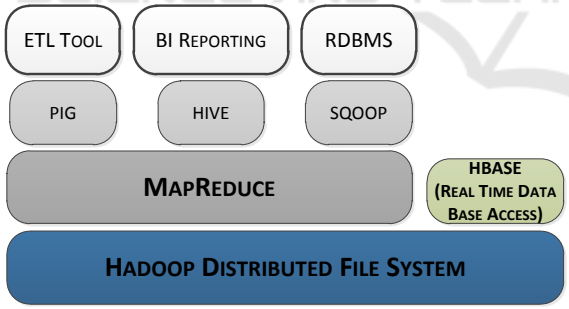

Figure 1: Hadoop Architecture (Saraladevi et al., 2015).

MapReduce main advantage is the accomplishment of parallelization and failover by splitting the work into multiple units (Chandrasekhar et al., 2013; Miller et al., 2016). Another significant advantage of Hadoop MapReduce pointed by authors is that it permits non-expert users an easy way to run analytical jobs over Big Data.

The platform uses Hadoop Distributed File System (HDFS), which is based on the distributed Google File System – GFS. It supports a scalable distributed file system that stores huge files in various and distributed machines in a reliable and efficient way (Inoubli et al., 2016).

The HDFS automatically replicates data across various nodes for fault tolerance (Inukollu et al., 2014). There are two types of nodes in a cluster. The first is the name-node (master) and the second is the data-node (slave). The name-node manages files, blocks, and mapping in a formation of the data-nodes, the data-node is responsible for storing data from a block unit into a number of locations separately. HDFS files are also replicated in multiples in order to provide parallel processing of large amounts of data (Khan et al., 2014).

4.2 Cloudera

Cloudera is the most well-known platform based on Apache Hadoop, which offers an effective platform that empowers organizations to gain insights from all their data (structured or unstructured) (Chandrasekhar et al., 2013). Cloudera is on the front line of the data management. Furthermore, Cloudera is the most innovative and contributes most for the open source Apache Hadoop platform (Sabapathi and Yadav, 2016). Cloudera is the leader in Hadoop-based platforms (Chandrasekhar et al., 2013) and has the same methods, functions, and main properties present in Hadoop, but it includes other efficient tools for social media (Murthy and Bowman, 2014). Cloudera maximizes the capabilities of Hadoop in storage, retrieval, and analysis (Murthy and Bowman, 2014) and enables enterprises to take advantage of its features of SQL tools to achieve real-time analytics (Prasad and Agarwal, 2016).

Where this platform stands out from the original Hadoop system is that it offers big data processing at faster speeds (Prasad and Agarwal, 2016), and with its user-friendly interface with many features and useful tools like Cloudera Impala. We can see the Cloudera Impala status in the Hadoop Stack in Figure 2.

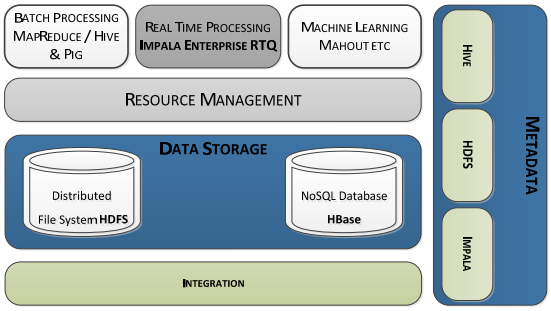

Figure 2: Cloudera Impala Status in Hadoop Stack analytics (Prasad and Agarwal, 2016).

Impala is a real-time, parallelized processing engine with an SQL-based interface that queries the storage (HDFS and HBASE). Impala is seen as the fastest querying engine present in the Hadoop-based platforms. Moreover, is not just the Impala that stands out from the other platforms; the Cloudera Manager is more stable and complete in features than the Ambari (HDP) and resource manager (Hadoop) (Azarmi, 2015).

4.3 Spark

Spark is an open source framework that was originally developed at UC Berkley in 2009 (Inoubli et al., 2016). This platform stands out for running programs faster than Hadoop MapReduce on disk or memory. Spark API supports Java, Scala, Python and R to develop applications quickly, and can be integrated to work with other platforms or standalone ("Apache Spark™," 2016).

Apache Spark is particularly appropriate and efficient for the analytics of heterogeneous data (Inoubli et al., 2016) and for stateful computations when precisely a delivery is useful indifferent whether it takes too long or not. Spark supports real-time distributed features, and integrates a complete SQL interface (Spark-SQL). It uses Hive for standard query languages, and also Domain Specific Language – DSL for query structured data (Morshed et al., 2016). It is similar to Impala in features and performance (Azarmi, 2015).

Spark uses a resilient distributed dataset (RDD) as a basic abstraction for a distributed dataset. The core operations (map, reduce and groupByKey) can be accomplished on the elements of the RDD and any one of those operations is evaluated lazily (transformations) or eagerly (actions). The distinct property of RDD is that they are unchangeable; operations on the RDDs create new RDDs (Miller et al., 2016).

Apache Spark is best suitable for near real-time data processing, and not for real-time processing because Spark uses mini batches that are not suitable for event level processing. The attractive feature of Spark is the capability to manage Machine Learning (ML) efficiently, due to its memory caching capacity that is impressive. Almost all of the popular streaming data sources can be easily integrated into the Spark API (Morshed et al., 2016).

4.4 Hortonworks

Hortonworks Data Platform (HDP) is based on Apache Hadoop. It offers its free and open source version of Hadoop along with services and training (Dinsmore, 2016). HDP agglutinates the stable components instead of distributing the latest version of the Hadoop project (Azarmi, 2015). Contrasting with Cloudera, HDP is 100% open source and totally free. It is an excellent choice for organizations that need the capability and cost-effectiveness of Apache Hadoop, with ready business tools (Chandrasekhar et al., 2013; "HDP," 2016).

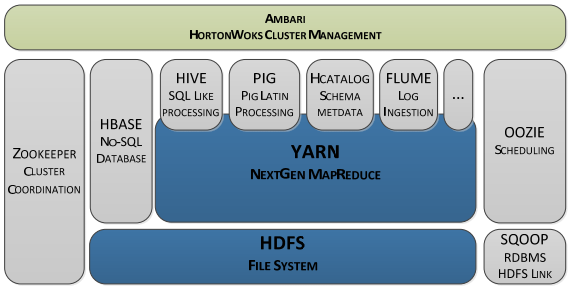

Figure 3: Hortonworks distribution (Azarmi, 2015).

As seen in Figure 3, HDP contains an integrated solution comprised of open source solutions such as Hadoop, Pig, Hive, Yarn, etc. (Khalifa et al., 2016). The components of Hadoop core stack are represented in blue, the components of the Hadoop Ecosystem project are in grey, and the specific component from HDP is represented in green (Azarmi, 2015). To deal with the performance issues, the HDP promotes Apache Tez as a performance optimizer (Dinsmore, 2016). This platform does not view the Hadoop as an alternative to traditional data management platforms and thus focuses on offering integration components for traditional data management platforms ("HDP," 2016). HDP looks for Hadoop as a tool to complement the existing data platforms, a similar vision with the Proprietary Software vendors.

4.5 HPPC System

The High-Performance Computing Cluster (HPCC) Systems Big Data is an open source framework that is used for manipulating, querying, transforming, as well as data warehousing. This framework is typically used as a choice instead of the Hadoop-based platforms, and there are two versions of the platform, one paid and one free (Chandrasekhar et al., 2013).

The HPCC uses the Linux operating system to support the layers of custom-built middleware components, thus providing an environment for running and supporting the distributed file system for data-intensive computing. It makes use of Thor

data refinery that is identical to the Hadoop-MapReduce combination, with its functions and capabilities, however, with similar configurations, it offers a much better performance (Furht and Villanustre, 2016). The HPPC data delivery engine Rapid Online XML Inquiry Engine (Roxie) as the name suggests is an online high performance structured query and analysis tool that supports parallel data access processing requests per node per second with sub-seconds response times (Furht and Villanustre, 2016) and the ECL – Enterprise Control Language. This Easy-to-learn and consistent programming language (ECL) was designed specifically for big data processing. There is another version called the community edition, which is a free HPCC version and is also supported by active developers and enthusiasts' community through online forums of discussion. The HPCC Systems platform has the same core technology that LexisNexis has used for years to analyse enormous data sets for its customers in industry, law enforcement, government, and science ("HPCC Systems Platform," 2016).

Due to the high-performance and cost-effectiveness of its implementation, the HPCC has been adopted by several government agencies, companies and research laboratories (Furht and Villanustre, 2016).

5 PLATFORMS COMPARISON

This work aimed at analysing five of the most popular open source big data platforms describing some of the more significant qualities, characteristics, capabilities, and functionalities of each platform. Table 1 shows a succinct description and the key features, contributing to the identification of the Big Data platforms for analytics that may be suitable for SMEs in their day-to-day business operations.

6 CONCLUSIONS AND FUTURE WORK

Big Data and Big Data Analytics have a direct relationship with the generation of knowledge since it is a fundamental and necessary element for decision-making within an organization, where information has been acquired.

In the open source platforms analysed Hadoop is the most used and serves as base for some other platforms. We suggest that the Cloudera is better suited for all contexts, particularly when users intend to deal and interact with large data sets in real-time. However, for integration with existing traditional data management systems we propose Hortonworks Data Platform because it has its own data integration modules that allows better support for other systems in an approach in terms of processes, analysis, and manipulation of various data sources.

As future work we propose to test in more detail the platforms characteristics, capabilities and functionalities in Big Data Analytics. We intend to experiment and explore the platforms in a real business environment.

Table 1: Big Data Platforms – comparative table.

	Description	Strong Points
Apache Hadoop	The most popular platform that implements the MapReduce paradigm and uses the HDFS.	-Largest community -Popularity -Forefront
Cloudera	The most well-known Hadoop-based platform. Same methods, functions, main properties as Hadoop, but more efficient in storage, retrieval, and analysis.	-Innovative -Efficient tools for social media -SQL tools for real-time analytics -User-friendly interface -Stability -Training & Support
Apache Spark	This platform runs programs faster than MapReduce on disk or memory and can be integrated to work with others platforms.	-Supports several programming languages -Integration with other platforms -Efficient analytics -Memory caching capacity -Complete SQL interface
Hortonworks	This platform is also Hadoop-based but only uses the stable components. Promotes the Apache Tez to deal with performance issues and the Apache Ambari as the cluster manager.	-Training & Support -Stability -Ready business tools -Low complexity for integration into an IT infrastructure -Windows support
HPCC	Typically chosen as alternative to Hadoop-based platforms, uses Thor data refinery as a distributed file system and for processing data across several nodes.	-High-performance -Consistent programming language (ECL) -Experienced -Robust solution

REFERENCES

Almeida, P.D.C. d, Bernardino, J., 2015. Big Data Open Source Platforms, in: 2015 IEEE International Congress on Big Data, pp. 268–275.

Apache Spark™ [WWW Document], 2016. Apache Spark™ - Light.-Fast Clust. Comput. URL http://spark.apache.org/ (accessed 11.16.16).

Apache™ Hadoop® [WWW Document], 2016. URL http://hadoop.apache.org/ (accessed 11.15.16).

Azarmi, B., 2015. Scalable Big Data Architecture: A practitioners guide to choosing relevant Big Data architecture. Apress.

Bernardino, J., 2011. Open source business intelligence platforms for engineering education. WEE2011 - Proc. of the 1st World Engineering Education Flash Week.

Bernardino, J. 2015. Open Business Intelligence for Better Decision-Making. In I. Management Association (Ed.), Economics: Concepts, Methodologies, Tools, and Applications, IGI Global (pp. 611-628).

Chandrasekhar, U., Reddy, A., Rath, R., 2013. A comparative study of enterprise and open source big data analytical tools, in: 2013 IEEE Conference on Information Communication Technologies. Presented at the 2013 IEEE Conference on Information Communication Technologies, pp. 372–377.

Dinsmore, T.W., 2016. Disruptive Analytics: Charting Your Strategy for Next-Generation Business Analytics, 1st ed. edition. ed. Apress, New York, NY.

Furht, B., Villanustre, F., 2016. Big data technologies and applications. Springer, Cham.

HDP [WWW Document], 2016. . Hortonworks Data Platf. HDP. URL http://hortonworks.com/products/data-center/hdp/ (accessed 2.4.17).

HPCC Systems Platform [WWW Document], 2016. . HPCC Syst. Platf. HPCC Syst. URL https://hpccsystems.com/download/hpcc-platform (accessed 11.15.16).

Inoubli, W., Aridhi, S., Mezni, H., Jung, A., 2016. Big Data Frameworks: A Comparative Study. ArXiv161009962 Cs.

Inukollu, V.N., Arsi, S., Ravuri, S.R., 2014. HIGH LEVEL VIEW OF CLOUD SECURITY: ISSUES AND SOLUTIONS. Conf. Comput. Sci. Eng. Appl. 4.

Khalifa, S., Elshater, Y., Sundaravarathan, K., Bhat, A., Martin, P., Imam, F., Rope, D., Mcroberts, M., Statchuk, C., 2016. The Six Pillars for Building Big Data Analytics Ecosystems. ACM Comput Surv 49, 33:1–33:36.

Khan, N., Yaqoob, I., Hashem, I.A.T., Inayat, Z., Mahmoud Ali, W.K., Alam, M., Shiraz, M., Gani, A., 2014. Big Data: Survey, Technologies, Opportunities, and Challenges. Sci. World J. 2014, e712826.

Landset, S., Khoshgoftaar, T.M., Richter, A.N., Hasanin, T., 2015. A survey of open source tools for machine learning with big data in the Hadoop ecosystem. J. Big Data 2, 24.

Lapa, J., Bernardino, J., Figueiredo, A., 2014. A Comparative Analysis of Open Source Business Intelligence Platforms, in: Proc. of the Int. Conf. on Information Systems and Design of Communication, ISDOC '14. ACM, New York, NY, USA, pp. 86–92.

Liu, F.C., Shen, F., Chau, D.H., Bright, N., Belgin, M., 2016. Building a research data science platform from industrial machines, in: 2016 IEEE International Conference on Big Data (Big Data)., pp. 2270–2275.

Miller, J.A., Bowman, C., Harish, V.G., Quinn, S., 2016. Open Source Big Data Analytics Frameworks Written in Scala, in: 2016 IEEE International Congress on Big Data (BigData Congress), pp. 389–393.

Morshed, S.J., Rana, J., Milrad, M., 2016. Open Source Initiatives and Frameworks Addressing Distributed Real-Time Data Analytics, in: 2016 IEEE International Parallel and Distributed Processing Symposium Workshops (IPDPSW). Presented at the 2016 IEEE International Parallel and Distributed Processing Symposium Workshops (IPDPSW), pp. 1481–1484.

Murthy, D., Bowman, S.A., 2014. Big Data solutions on a small scale: Evaluating accessible high-performance computing for social research. Big Data Soc. 1, 2053951714559105.

Neves, P., Bernardino, J., 2015. Big Data Issues, in: Proceedings of the 19th Int. Database Engineering & Applications Symposium. ACM, pp. 200–201.

Pääkkönen, P., Pakkala, D., 2015. Reference Architecture and Classification of Technologies, Products and Services for Big Data Systems. Big Data Res. 2, 166–186.

Prasad, B.R., Agarwal, S., 2016. Comparative Study of Big Data Computing and Storage Tools : A Review. Int. J. Database Theory Appl. 9, 45–66.

Rubinstein, I., 2012. Big Data: The End of Privacy or a New Beginning? (SSRN Scholarly Paper No. ID 2157659). Social Science Research Network, Rochester, NY.

Sabapathi, R., Yadav, S., 2016. Big Data:Technical Challenges towards the Future and its Emerging Trends. AADYA-Natl. J. Manag. Techno. 6, 130–137.

Sagiroglu, S., Sinanc, D., 2013. Big data: A review, in: 2013 International Conference on Collaboration Technologies and Systems (CTS). Presented at the 2013 International Conference on Collaboration Technologies and Systems (CTS), pp. 42–47.

Saraladevi, B., Pazhaniraja, N., Paul, P.V., Basha, M.S.S., Dhavachelvan, P., 2015. Big Data and Hadoop-a Study in Security Perspective. Procedia Comput. Sci. 50, 596–601.

Sen, D., Ozturk, M., Vayvay, O., 2016. An Overview of Big Data for Growth in SMEs. Procedia - Soc. Behav. Sci., 12th International Strategic Management Conference, ISMC 2016, 28-30 October 2016, Antalya, Turkey 235, 159–167.

Sivarajah, U., Kamal, M.M., Irani, Z., Weerakkody, V., 2017. Critical analysis of Big Data challenges and analytical methods. J. Bus. Res. 70, 263–286.

Ventana Research: Big Data Analytics [WWW Document], 2014. Pentaho. URL http://www.pentaho.com/resource/ventana-research-big-data-analytics (accessed 2.3.17).