Big Data solutions to support Intelligent Systems and Applications

Luciana Lima, Filipe Portela, Manuel Filipe Santos, António Abelha and José Machado.

***Abstract*—** in the last years the number of data available to be used by Intelligent Systems increased significantly. The system have now to have capabilities of storing and processing huge amount of data in real-time. With this new reality arises the Big Data concept. Big Data is much more than a big number of records stored in a database. The data can be in three formats: structured, unstructured and semi-structured. The number of Big Data solutions increase in the market, however it is difficult to understand which type of solutions are able to achieve a set of essential features: Data Integration, Data Visualization, Real-Data Analytics, Interactive Search, Text Analytics, Real-Time and Batch Processing. In order to help the researchers and professionals to have a better comprehension of the vendor's solutions and to make a better choice about what is the better solution to their Intelligent System / applications a comparative study was made. This paper present the study made and the results achieved by comparing a set of Big Data solution.

***Keywords*—**Big Data, Intelligent Systems, Applications, Solutions, Benchmarking.

I. Introduction

THE technological evolution and consequent increased of dependence of society and organizations has led, in recent years, the exorbitant growth in the volume and variety of existing data. The McKinsey Global Institute estimates that the volume of data grow 40% per year and between 2009 and 2020 this growth will be 44 more time [1]. Every two years, the volume in all world doubles and in 2015 it will reach (approximately) 7.9 zeta bytes [1]. At the same time, market evolution requires organizations with the ability to find new ways to improve their products / services; satisfy their customers; prevent some prejudicial situations and finally, avoid the increased costs to achieve these goals.

The Big Data comes in large force, not only by the ability to process high-speed massive amounts and variety of data, but also their ability to provide value to organizations who wants include Big Data in decision-making process.

The relevance of the use of Big Data by organizations of the most diverse sectors and the way how implement a solution Big Data is something that still raises questions and discussion.

Since the appearance of the Big Data buzzword some companies made its own solutions to solve their problem with complexity and volume. The success of these solutions sells Big Data as something very useful and unique. Nowadays companies are invaded by a lot of ideas, tools by all kind of suppliers.

For that reason, this paper tries to make a study of the main solution and vendors in this area. The paper display and categorize this panel of technologies with the purpose of simplify the choice. This papers presents the results achieved after a study made with the goal to find Big Data Solutions and consequently understand their features and capabilities.

Considering the amount of emerging technologies it is provided an overview of the existing technologies divided by Big Vendors (SW/HW, Cloud Deployment options) and Open Source solutions. The solutions offered by each one different and the vendors try to create connected in order to include open-source features. With this work is expected helping the decision-makers to choose the best Big Data solution without any efforts in looking for information and in understanding their main differences and values.

This paper is divided in five section. Behinds a paper introduction it is background where the main concept: Big Data and some of their features and technologies are addressed. Section three make an overview of Big Data solutions have in considerations three vectors (Commercial, Cloud and Open-Source). Then in section four it is made a summary of the results achieved during this benchmarking study. In this section the solutions were compared in two groups: technologies and commercial solutions. Finally a brief conclusion and future work are presented.

II. Background

Big Data is a "Dataset whose size and complexity is beyond the ability of conventional tools of manage, store and analyze data"[2]. Big Data can be seen as a set of technologies capable of store, process and get value from several sources and formats of data in real-time.

Sridhar [3] considered five V's as the core of Big Data. From

This work was FCT - Fundação para a Ciência e Tecnologia within the Project Scope UID/CEC/00319/2013.

Luciana Lima is with Information System Department, University of Minho, Portugal.

Filipe Portela is with Algoritmi Research Centre, University of Minho, Guimarães, Portugal. (Corresponding author to provide phone: +351253510319; fax: +351253510300; e-mail: cfp@dsi.uminho.pt).

Manuel Filipe Santos, is with Algoritmi Research Centre, University of Minho, Guimarães, Portugal. (e-mail: mfs@dsi.uminho.pt).

António Abelha and José Machado are with Algoritmi Research Centre, University of Minho, Braga, Portugal. (e-mail: {Abelha,jmac}@di.uminho.pt).

the five the three with most impact are Volume, Variety, and Velocity. Veracity and Value are also important for Big Data but they are not exclusive for the new way to process data. In fact, to guarantee the Veracity and Value of data for the business is something that already is done in Business Intelligence infrastructures.

Therefore, these three V's can be defined as:

- Velocity - Data is generated, collected and processed very quickly. We fly from Batch to Real-Time and data streaming;
- Variety - Data format and sources are more diverse, much less concerned with schema or rules. Data could be collected from enterprise data warehouses, machines, web pages, etc. Structured, semi-structured or unstructured are processed as well;
- Volume - Refers to large amount s of data generated and collected every day. We jump from TB to PB measure.

As one of the most suitable technologies to storing and management big Data arises Hadoop.

Hadoop is an implementation in Java and Open Source of distributed computing used for the processing and storage of data in large scale by dividing workloads across three, five, or thousands of servers. It is designed to scale up from single servers to thousands of machines, each offering local computation and storage.

Companies seek deeper insights from the massive amount of structured, unstructured, semi-structured, and binary data at their disposal in order to dramatically improve business outcomes. Hadoop can help by:

- Capturing and storing all data for all business functions
- Supporting advanced analytics capabilities
- Sharing customer data quickly and generously with all those who need it
- Continuously accommodating greater data volumes and new data sources

The growth of data volume and complexity leads to the necessity of complement Hadoop with adaptable tools as is for example: Common, Avro, MapReduce, HDFS, Pig, Hive, Hbase, ZooKeeper, Sqoop.

Enterprises traditionally employ Information Technologies (IT) professionals with a package of expertise: implementation, customization, and system integration. However, Big Data deployments needs new specialist competencies – statistical and analytical – in addition to advanced IT/programming skills.

Many IT executives view open source Big Data technologies, such as Hadoop, as immature or unstable and carrying significant security and retooling risks when compared to proprietary tools

Big Data technologies enable detailed tracking and analyses of consumer profiles and behaviors, from non-traditional data sources such as social networking sites, mobile device applications, and sensors. This generates valuable business opportunities for more targeted/personalized services and cross selling. However, enterprises need to be cautious and ensure that this in-depth collection and mining of personal data does not result in privacy and compliance lapses.

III. Big Data Solutions

Since the appearance of the Big Data buzzword some companies made their own solutions to solve their problems with complexity and volume .The success of these solutions sells Big Data as something very useful and unique. Nowadays, companies are invaded by a lot of ideas and tools by all kind of suppliers.

For that reason, this chapter tries to display and categorize this panel of technologies with the purpose of simplifying the choice.

At this point, all the technologic offers are divided into Commercial, Cloud and Open-Source Tools. Some suppliers have their offers in commercial and open-source (sometimes free) distribution that justify their presence on both panels. Moreover, on both deployment options (Commercial, Open Source) a short summary of some technologies is provided. The Commercial section presents Big Data Solutions known as Massively Parallel Processing (MPP) tools.

Finally, in the open source section we can find the summary of Analytic Open Source Tools.

A. Commercial

The commercial or vendor-provided software is a software tool with property rights. The software is designed for sales purposes and it satisfies commercial needs. It is the model where the software developed by a commercial entity is typically licensed for a fee to a customer (either directly or through channels) in object, binary or executable code [4].

To the costumers, this kind of offer is (usually) related as a higher quality, secure and trustworthy software.

The Big Data Technology Panel (Fig. 1) exhibits commercial tools distributed in a Big Data Ecosystem that is composed by four big classes.

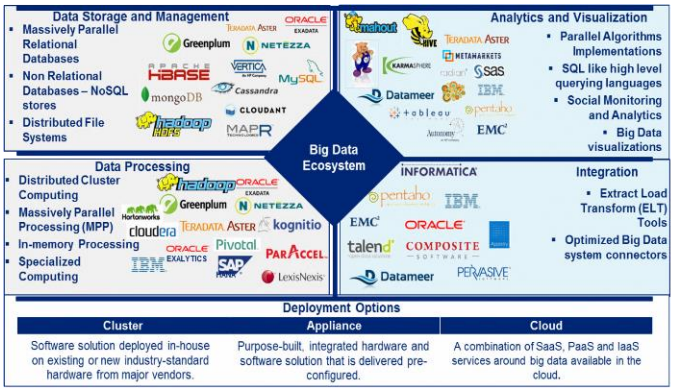

Fig. 1 - Big Data Technology Panel adapted from [5, 6]

Data Storage and Management: Technologies that have the capacity to store large volumes of different types of data. These technologies are also capable of managing and automating their process to maximize and improve the performance of their storage resources. The data storage can be in the Massively Parallel Relational Databases, Non-Relational Databases (NoSQL stores) and Distributed File Systems.

Data Processing: Technologies that are known for collecting and manipulating data to produce meaningful outputs, suitable

ISBN: 978-1-61804-320-7

to be analyzed. The process could be the Distributed Cluster Computing environment, the Massively Parallel Processing (MPP) and the In-memory and Specialized Computing.

Analytics and Visualization: Technologies which are capable of providing to end-users the advanced mechanisms to explore analyze and present data in a pictorial or graphical format. In this class we found tools with Parallel Algorithms Implementations, SQL like high level querying languages, Social Monitoring and Analytics and a Big Data visualizations feature.

Integration: Technologies well suited to combine data from different sources to give an integrated view of valuable data. This process embraces extraction, cleaning, transformation tasks. The existing offer consists of Extract Load Transform (ELT) tools and Optimized Big Data system connectors.

B. Cloud

Cloud Computing it is another theme that is on top of the organizations' minds. As a delivery model for IT services, cloud computing has the potential to enhance business agility and productivity while enabling greater efficiencies and reducing costs [7].

Although Cloud Computing is in an evolution process, it continues to mature and a growing number of enterprises are building efficient and agile cloud environments, as cloud providers continue to expand service offerings.

To manage Big Data challenges like flexibility, scalability and cost to data access, the clouds are already deployed on pools of servers, storage and networking resources and they can scale up or down according to the needs.

Cloud computing offers a cost-effective way to support big data technologies and the advanced analytics applications that can drive business value.

It will take a while for organizations to see the cloud as valid, secure and completely prepared for their needs.

As presented in the cloud vendor map (Fig. 2), there is a wide variety of cloud computing service models to provide storage, management and processing for Big Data. Organizations can leverage SaaS, PaaS or IaaS solutions depending on the needs [5].

	Cloud Solution	Amazon	Microsoft	Google
Storage	Big Data Storage	Amazon S3	HDFS	Cloud Storage (GFS)
	NoSQL Store	DynamoDB	Table Storage API	AppEngine Datastore
	Relational Store	MySQL or Oracle	SQL Azure	Cloud SQL
Processing	Hosting Service	Amazon EC2	Azure Compute	AppEngine
	Map Reduce Service	Elastic MapReduce	Hadoop on Azure	AppEngine Mapper API
	Big Data Analytics	Elastic MapReduce	Hadoop on Azure	BigQuery

Fig. 2 - Prominent Cloud Vendors providing Big Data Solutions retrieved from [5]

C. Open Source

Open Source is a kind of offer defined as source code which must be available for everyone and all modifications made by its user can also be returned to the community.

These technologies are not necessarily free. Anyone who wants to sell an open source program, can make it however, its prices will be low whereas the development to achieve new markets will be fast.

Similar to the earlier sections of this chapter, we provide an open source technologies panel (Fig. 3) that is supplied by "native" open source companies or by renowned Big Vendors who provide some tools in an Open Source Model.

Fig. 3 - Open Source Big Data Technologies Panel retrieved from [8]

Although commercial software is usually related to higher quality and trustworthy software, Open Source Software has attracted substantial attention in the last years [4].

The same happens with cloud solutions, they represent an attractive path when we talk about resource costs but they still have not won the trust of the companies, in terms of the information systems security.

Organizations must evaluate these three options with the purpose of finding the one that best fits their commercial needs.

Be it commercial or open source tools, deployed in a cluster, cloud or appliance, the organizations must take into consideration the costs of acquisition, implementation, maintenance, upgrading and, also, the flexibility of the architecture to integrate other tools with different natures.

Finally, the security that truly depends just as much on how well the software is deployed, configured, updated and maintained, including product vulnerabilities discovered and solved through appropriate and timely updates

IV. Comparative Summary

Once the investment in a Big Data project is approved and finally gathered the information about the market offers, organizations must choose the best solutions that better fits their needs. The large number of options makes it hard for organizations to decide what is best and why. This happens because each Big Data supplier sells its solution as the best one,

the most feasible, robust and scalable, amongst other features, and the organizations cannot base their decisions solely on the kind type of characteristics.

Nowadays, sharing information about it, it is the best "counseling". Tools trial reports, vendor surveys and white papers, benchmark reports from specialists and the opinion of non-professionals (from the social media), these are the best sources to sustain our decisions. More available information simplifies the choice.

A. Analytic Open Source Tools

The table Analytic Open Source Tools (Table 1) describes some of the most used technologies presented in the panel (fig. 3). For now the table focuses only on the analytic tools because they are best known and easily accepted by companies.

This study was made having in consideration the type of features of each one has.

Table 1 - A. Analytic Open Source Tools - Summary

Key Tools	Summary	Features
Knime	Data analytics platform that allows you to perform sophisticated statistics and data mining on your data to analyze trends and predict potential results. Its visual workbench combines data access, data transformation, initial investigation, powerful predictive analytics and visualization The open integration platform offers over 1.000 modules. KNIME[1] is also the open source data analytics platform.	**Data Integration** **Data Mining**
WEKA	WEKA stands for "Waikato Environment for Knowledge Analysis" and it is a collection of machine-learning algorithms in order to solve data mining problems. It is written in Java and thus runs on almost any modern computing platform. It supports different data mining tasks such as clustering, data pre-processing, regression, classification, feature selection as well as visualization [8].	**Data Mining**
Rapid Miner	RapidMiner offers data integration and analysis, analytical ETL and reporting combined in a community edition or enterprise edition. It comes with a graphical user interface for designing analysis processes. The solution offers a metadata transformation, which allows inspecting for errors during design time [8].	**Data Integration** **Data Mining**
Talend	Open source software developed by Talend has developed several big data software solutions, including Talend Open Studio for Big Data, which is a data integration tool supporting Hadoop, HDFS, Hive, Hbase and Pig. The objective is to improve the efficiency of data integration job design through a graphical development environment. Next to open source tools, Talend also sells other commercial products [8].	**Data Integration**
Jaspersoft	• Jaspersoft has developed several open source tools, among others a Reporting and Analytics server, which is a standalone and embeddable reporting server. • The Open Source Java Reporting Library is a reporting engine that can analyze any kind of data and produce reports in any format. • Jaspersoft ETL offers a data integration engine, powered by Talend. They claim it is the world's most used business intelligence software.	**Data Integration**

[1] http://www.knime.org/knime

B. Big Vendor Solutions - Summary

Therefore, Table 2 presents a short summary and an objective comparison of the most popular vendors, in terms of their offers, hardware appliance and Connectors.

Table 2 - Big Vendor Solutions - Summary

Key Vendors	Offers	Hardware Appliance	Connectors
EMC Greenplum	MPP Database Hadoop distribution Chorus – search, explore, visualize, analyze Command Center	Optional- Greenplum Data Computing Appliance (DCA)- single rack expandable in quarter rack increments up to 12 racks) Optional – gNet connector	Connects to most traditional EDWs and BI / analytics applications (SAS, MicroStrategy, Pentaho)
IBM Netezza	IBM Netezza DW Appliance IBM Netezza Analytics	S-Blade servers contain multi-core Intel CPUs and IBM Netezza's unique multi-engineFPGAs. Configuration in single and multiple racks	Data Integration with most IBM and 3rd party solutions. Working with Cloudera to bring in Hadoop connectivity
Oracle Exadata	Oracle Exadata Database Machine InfiniBand High Speed Connectivity Oracle Data Integrator	Exadata Storage Server X2-2 Exadata Database Machine X2-2 Exadata Storage Expansion Racks Exadata X2-2 Memory Expansion	Data Integration with Hadoop, NoSQL and other relational database sources. Special Integration to R. Connectivity to 3rd party solutions. include Cloudera Hadoop distribution and management software
Teradata	Aster Database 5.0 with SQL-MapReduce Teradata Database 14 Hadoop Integrator	Aster MapReduce Appliance Active Enterprise Data Warehouse Extreme Performance Appliance Data Warehouse Appliance Extreme Data Appliance Data Mart Appliance	Data Integration with Hadoop, NoSQL and other relational database sources. Special Integration to SAS. Connectivity to 3rd party solutions.
Cloudera	Hadoop platform distribution (CDH) Data Integrator Automated Cluster Management Search, explore, visualize and analyze engines Web applications that enable you to interact with a CDH cluster(HUE)	CDH (Cloudera's Distribution including Apache Hadoop) Cloudera Express Cloudera Ent	Data Integration with Hadoop, NoSQL and other relational database sources. Connectivity to 3rd party solutions. Connectors for Netezza, Teradata,Tableu, Microstrategy

ISBN: 978-1-61804-320-7

C. Big Data Commercial Evaluation

This evaluation had in consideration most of the important features in Big Data:

Data Integration: Allow a quick and easy integration of multiple data sources (unstructured, semi-structured and structured). At same time it allows organizations to execute data analytics tasks with data virtualization.

Data Visualization: Is the way of the information (knowledge) is displayed and make available. Solutions with this features allow not only the data preparation but also has a data visualization shape in order to present the information stored.

Big Data Analytics: Allow to analyze large data sets containing a variety of data types. Enables organizations to analyze a mix of structured, semi-structured and unstructured data in search of new knowledge and valuable business information.

Interactive Search: Allow an easy way to execute queries by searching particular data / information that the "user" wants (making for example data comparison and if available drill-down and roll-up). Sometimes this type of features has associated information retrieval algorithms.

Text Analytics: Capability to transforms free-form text documents into a chosen intermediate form and deduces patterns on knowledge from the intermediate form.

Real Time: This is a required feature to the Intelligent Systems [9]. It means that a system is able to execute the main tasks in real-time in order to support Data Processing, Data Integration, Data Mining and Data Visualization.

Batch Processing: Also known as intelligent agents [10, 11] it is the execution of a set of tasks (e.g. data processing) on a system without manual intervention.

Table 3 - Big Vendor Solutions - Evaluation

Features	IBM	Oracle	EMC	Teradata	Cloudera
Data Integration	✓	✓	✓	✓	✓
Data Visualization	✓		✓	✓	
Big Data Analytics	✓			✓	
Interactive Search	✓	✓	✓	✓	
Text Analytics	✓				
Real-Time	✓	✓	✓	✓	✓
Batch processing	✓	✓	✓	✓	✓

After analyzing table 3 is easy to understand that there are three essential features for all vendors: Data Integration, Real-Time and Batch processing. From the five vendors evaluated only the IBM has all the features presented in their solution. Following the features list appears Teradata and EMC. The fact of IBM arises in first place does not means that IBM is the best solution (the decision-makers should to choose their solution according to project requirements). Actually it is possible combining solutions in order to achieve a better result. For example if the decision-maker prefers working with Teradata can combine it with a tool presented in table 1 / figure 3 in order to have the same features than IBM or even better.

V. Conclusion

Being real-time and batch processing two required features to deploy an Intelligent System or application all the vendors are able to support this deployment. However the vendor should be chosen in accordance to the project requirements. After choose a Big Data architecture, they can choose an Analytic Open Source Tool in order to support the data analysis process.

Table 2 helps the decision-makers to have a better understanding of which vendor is more suitable for their project. This study allows to have a better understanding about which are the main features of each vendor / solution.

One of the main goals of this paper is to avoid the researcher effort in searching for Big Data solutions to support their system. By reading this paper they can know what are the main open-source technologies and vendors. They also can know what the key vendors offer and what is the hardware appliance and the connectors provided by each.

In the future two studies: Commercial vs Open Source and Appliance vs Cluster will be performed with the goal to display the pros and cons of each solution. Then a Big Data architecture will be designed taken in considerations the studies results.

Acknowledgments

The authors would like to thank Deloitte for some of the resources provided and FCT (Foundation of Science and Technology, Portugal) for the financial support through the contract PTDC/EEI-SII/1302/2012 (INTCare II). This work has been supported by FCT - Fundação para a Ciência e Tecnologia within the Project Scope UID/CEC/00319/2013.

References

[1] J. P. Dijcks, "Oracle: Big data for the enterprise," *Oracle White Paper*, 2012.

[2] M. M. Gobble, "Big data: The next big thing in innovation," *Research-Technology Management*, vol. 56, p. 64, 2013.

[3] P. Sridhar and N. Dharmaji, "A comparative study on how big data is scaling business intelligence and analytics," *Int. J. Enhanced Res. Sci. Technol. Eng*, vol. 2, pp. 87-96, 2013.

[4] G. Hiong, "Open Source and Commercial Software: An In-depth Analysis of the Issues," ed, 2004.

[5] Deloitte, "Title," unpublished|.

[6] P. Pääkkönen and D. Pakkala, "Reference Architecture and Classification of Technologies, Products and Services for Big Data Systems," *Big Data Research*, 2015.

[7] I. I. Center. (2014, April, 24, 2015). *Big Data in the Cloud: Converging Technologies.* Available: http://www.bigdata-startups.com/open-source-tools/,%2012-08-2014/

[8] N. Hague. (2014, April, 24, 2015). *BigData Startups - the big data knowledge platform*. Available: http://www.bigdata-startups.com/open-source-tools/,%2012-08-2014/

[9] F. Portela, M. F. Santos, P. Gago, Á. Silva, F. Rua, A. Abelha, *et al.*, "Enabling real-time intelligent decision support in intensive care," in *25th European Simulation and Modelling Conference- ESM'2011*, Guimarães, Portugal, 2011, p. 446 pages.

[10] L. Cardoso, F. Marins, F. Portela, M. Santos, A. Abelha, and J. Machado, "The Next Generation of Interoperability Agents in Healthcare," *International journal of environmental research and public health*, vol. 11, pp. 5349-5371, 2014.

[11] M. Wooldridge, "Intelligent agents," in *Multiagent systems: a modern approach to distributed artificial intelligence*, ed: MIT Press, 1999, pp. 27-77.