Big Data in Cloud: A data architecture

Jorge Oliveira e Sá[1], César Martins[2,] and Paulo Simões[3]

[1]University of Minho, Portugal, jos@dsi.uminho.pt
[2]University of Minho, Portugal, pg21441@alunos.uminho.pt
[3]ISEG – School of Economics and Management, Portugal, paulo.simões@iseg.utl.pt

Abstract. Nowadays, organizations have at their disposal a large volume of data with a wide variety of types. Technology-driven organizations want to capture process and analyze this data at a fast velocity, in order to better understand and manage their customers, their operations and their business processes. As much as data volume and variety increases and as faster analytic results are needed, more demanding is for a data architecture. This data architecture should enable collecting, storing, and analyzing Big Data in Cloud Environment. Cloud Computing, ensures timeliness, ubiquity and easy access by users. This paper proposes to develop a data architecture to support Big Data in Cloud and, finally, validate the architecture with a proof of concept.

Keywords: Big Data, Cloud Computing, Cloud Computing Systems Architecture, Big Data Analytics, and Technologies for Big Data.

1 Introduction

Nowadays, organizations want to get meaning from large volume of different types of data and they face many different challenges. Organizations face difficulties in processing data, i.e., collecting, storing and making data available for analytical processes, losing opportunities to better support technical or tactical decisions. Organizations are confronted with data from several sources (Kimball & Ross, 2013).

Thus, arise the need of solutions that enable collecting and integrating data (of different types) from various sources, but in addition, allowing analysis of data anywhere (space) and any time (time) because the data access is crucial to decision making (Russom, 2013).

These data have different characteristics, i.e., can be structured, semi-structured or unstructured. Due to this diversity the term Big Data arises and covers different types of data with different origins. Big Data covers several concepts, creating much confusion and misunderstanding.

It is commonly accepted that Big Data concept was first formulated by Doug Laney in February 2001. Laney proposed three important dimensions: Volume (large amount of data); Velocity (capture, store and analyze data quickly to help better support to operational activities); and Variety (process data with distinct characteristics) (Laney, 2001).

If data to be analyzed are only typical structured data, e.g., numeric or alphanumeric attributes in entity relationship model (ER), the relational databases supported by massively parallel processing technology would be sufficient to meet the challenges. However, variety (the needed to process and analyze unstructured data) introduces a disruptive dimension that needs to be addressed with the current computer resources (Kimball & Ross, 2013).

These three initial dimensions (now being referred as "three V's"), defined Big Data concept and associated with other characteristics contributed to the proliferation and understanding of this concept, (Stonebraker, 2012).

Thus, the need for an architecture which allows joining all these data dimensions arises. This architecture should allow storage of data from different sources, this requires a thorough knowledge of data sources, as well as data extraction, validation and transformation process in order to help the process of decision-making (Russom, 2013). To do this, data will be stored in a repository that is created with specific technologies for this type of solutions (Lohr, 2012). Finally, data should be available for analysis, where users (decision makers, analysts, etc.) can explore it and this exploration will be as dynamic as possible, i.e., users can define which variables and values want to explore. Data analytics can be performed via dashboards, reporting, online analytical processing (OLAP), etc. Architecture needs to allow data analytics in any place and time. This can be achieved by using Cloud Computing environment (Agrawal et. al., 2010).

The goal of this work is to develop a conceptual and technological architecture to support Big Data and Cloud Computing, especially in terms of integration, storage and availability of data.

This work aims to answer the following question: What are the architecture characteristics to collect, store and provide data for further analysis in a Cloud environment?

This paper begins by presenting a theoretical framework of the concepts used, i.e., Big Data and Cloud Computing. In the next section a conceptual architecture will be presented and will be instantiated and implemented with some specific technology solutions available in the market which allows present a technological architecture. Finally, a test case will be used and to serve as proof of concept.

2 Theoretical Framework

This section provides definitions of Big Data in section 2.1 and Cloud Computing in section 2.2

2.1 Big Data

Big Data term is increasingly used, mainly because organizations are realizing the value and the gains that can obtain through the huge amounts of data they possess (Stonebraker, 2012; Oswaldo et al., 2010). The growth, proliferation and influence of social networks in society were the main influencing factors for the importance of Big Data term (Manyika et. al, 2011).

Handle Big Data cannot be achieved with a single technology, but rather as a combination of various technologies, some more ancient and some more recent. Big Data seeks to help organizations to achieve greater efficiency in managing large amount of data, as well, to solve problems associated with its storage. Data can be provided by internal and/or external sources, and the data may have different types of data structures namely (Halper & Krishnan, 2013; Guoliang et. al, 2008):

- **Structured Data**: all data are organized into semantic blocks (entities), entities are grouped together through associations and classes, an entity of a particular group can have the same descriptions and attributes that can be performed for all entities of a group, may thus contain the same shape, size and in the same order. All these data are stored in Database Management Systems (DBMS) with a rigid structure that was previously projected through an ER model. Structured data can be provided, e.g., from management applications, like ERP, CRM, SCM, etc.
- **Semi-structured data**: these data cannot be stored into a DBMS and show a high degree of heterogeneity, which means that these data cannot be modeled into a rigid structure. Examples of semi-structured data are: Extensible Markup Language (XML), Resource Description Framework (RDF), Web Ontology Language (OWL), etc.
- **Unstructured data**: these data does not necessarily have a format or sequence, do not follow rules and are not predictable. Unstructured data is currently getting a lot of attention mainly due the proliferation of a variety of mobile data devices. However there are other data sources such as sensor machines, smart devices, collaborative technologies and social networking. These data are not related but diverse data, some examples of this type of data are: text, video, images, etc.

The main characteristics which allow differentiate the various types of data are shown in Table 1.

Big Data has also provided an increase in data complexity and the database management systems (based on relational model) show difficulties to storing them (Manyika et. al, 2011).

Big Data can have several interpretations, but it is important to emphasize that is based on the principle of three V's: Volume, Variety and Velocity. However, scientific and academic communities propose other two V’s: Variability and Value (Stonebraker, 2012). Next the five V's are described:

- **Volume -** represents a large amount of data to be collected and analyzed in order to use Structured Query Language (SQL), analytics (count, sum, max, min, average and group by), regressions, machine learning and large complex data analytics.
- **Variety** - corresponds to manage several data types with different structures.
- **Velocity** – big volumes of data are generated quickly and incrementally, which decreases the time window available for decision-making. The challenge here is to collect and store large volumes of data in a timely manner, seeking to use historical and real-time data to support operational, on the fly, decisions.
- **Variability**- allows the classification of various data sources (structured, semi-structured and unstructured) according to their quality, according to aspects such as accuracy and timeliness of the data provided.
- **Value** - is the value that users obtain by using Big Data to decision-making.

Table 1 - Differences between Structured, Semi-Structured and Unstructured Data, adapted from (Guoliang et. al, 2008)

Structured Data	Semi-Structured Data	Unstructured Data
Predefined structure	There is not always a schema	There is no schema
Regular Structure	Irregular Structure	Irregular Structure
Separate data Structure	Embedded in the data structure	The structure is dependent on the source of data
Reduced Structure	Extensive structure (in particular each data set may have its own organization)	Extensive structure depends largely on the type of data
Little evolutionary and quite rigid.	Evolutionary far, the structured could change very often.	Evolutionary far, the structured changes quite often
Has closed schemas and integrity constraints	A schema there is no associated data	A schema there is no associated data
Clear distinction of data structure	There is not clear distinction between the data structure	Unable to distinguish between data structure

2.2 Cloud Computing

The term Cloud Computing is used when referring to technology that provides flexible resources and IT services based on Internet (Böhm et al., 2011). Cloud Computing can also be considered as a set of concepts associated with various areas of expertise such as Service-Oriented Architecture (SOA), distributed computing, grid computing (computational model that divides the tasks in several machines) and virtualization (Youseff et al, 2008).

Beyond the technological vision that is associated with Cloud Computing, the concept can also be understood as an innovation, especially in the provision of IT services (Böhm et al., 2011). Many believe that this is a potential to be exploited, mainly in the way of development and deployment of computational resources and applications, looking for new business models especially for companies providing Software (Youseff et al, 2008; Stuckenberg et al., 2011).

According to National Institute of Standards and Technology (NIST), there are many benefits of adopting Cloud Computing and the most relevant are (Olivier et al, 2012):

- Enables economies from client side, i.e., because the service provider has the responsibility to ensure services and infrastructure, this allows the client to focus on what is goal of the business and thus improve their productivity, on the other hand, ensures that the client has a constant adaptation of the service to their needs, often by reducing the cost of services and infrastructure.

- Enables good services from provider side, i.e., the provider of Cloud services seeks to ensure continuous improvement of its services through better matching of hardware and software needs required by their customers. It is also required that the service provider makes available a set of security policies that defend the relevant information of each of its customers.

Cloud also has some models that make possible its implementation as commercial solutions. Currently existing models are:

- **Private Cloud**: The user of these solutions is a specific organization or an organizational unit, which may be internal to the organization or contracted to a Cloud service provider. The Advantages of Cloud cannot be fully exploited by this model due to the high degree of customization.
- **Community Cloud:** The service is used by multiple members of a group, being offered by various vendors, internal or external to the community.
- **Public Cloud:** Services are available to general public; the service is offered by a single vendor and in this model the stability and resources has pooling can be fully exploited.
- **Hybrid Cloud:** A hybrid Cloud offers a varied combination of models, e.g., some data can reside in a private Cloud while other data and applications can reside in a public Cloud.

3 Data Architecture

The data architecture will be explained using two steps:

1. **Conceptual Architecture**: describes the levels that constitute the architecture and the explanation of the activities that will be conducted in each of the levels, see section 3.1;
2. **Technological Architecture**: describes a technological solution. It is emphasized that there may be other distinct technological solutions presented, see section 3.2.

Data sources represented in conceptual and technological architecture represents examples of different sources and data types can be used.

3.1 Conceptual Architecture

Conceptual architecture, see figure 1, is divided in two levels: Data Staging Area and Data Analysis and Visualization. Each level supports a set of activities which will be described next.

Data Staging Area

This level is responsible for two important processes, one is called ETL (Extract, Transform and Load) process, which involves the extraction of data from multiple sources, transformation and cleaning of the same, to ensure that the processed data are loaded for later a storage area. The activities that are performed by ETL process are cleaning data, detecting and correcting errors, extracting data for further analysis. The second process consists in storing data that were previously processed. Data will be stored in a Big Data database that allows storing data from several sources into a single data repository.

Data Analysis and Visualization

This level is responsible for defining business indicators to be analyzed. It is possible to define a set of metrics that allow decision-making. The results of Data Analysis are presented in a Cloud environment (e.g., results available via web). The relevance of this Data Visualization component is related to the need for rapid access to data from any location at any time. So, a user with a device (as computers, laptops, tablets, smartphones, etc.) with internet connection can perform Data Analysis.

3.2 Technological Architecture

Technological architecture, see figure 2 shows an example of some technologies that can be used at both levels of architecture. The choice of technologies had taken into account the market value of vendors.

The architecture was implemented using the Azure a Microsoft online platform that enables the provision of resources such as an operating system, memory, disk and processor that can be accessed from anywhere provided you have an internet connection. Using this solution Windows Server 2008 R2 is installed, within which was then installed a virtual machine (Hortonworks Sandbox 2.1), and also the tools ETL and data integration Talend OpenStudio, Big Data, Data Quality and Data Integration and as tools for the analysis of Microsoft BI Power and Tableau data. The purpose of the installation of all this technology is the ability of architecture validation, in order to subsequently adopt the same in a real scenario.

However it is also important to note that the use of the Azure platform allowed all the work could be run in the Cloud environment, and the use of this environment is an asset especially because it allows the work to be performed at any time regardless of location and without limitation of time and space. In addition there is a physical border with processing capacity limit of a machine because with this type of environments you can add features that respond to the needs. One of the benefits of using Azure has to do with the fact that there is a large initial investment in the acquisition of machinery and subsequent hardware upgrade needs to meet the needs required for the use of a specific technological tool.

The use of Cloud resources also requires costs and although these are less than the costs associated with the acquisition of hardware equipment, the availability of resources in the Azure platform is instant and if an upgrade or increase storage capacities necessary, Random Access Memory (RAM) and processor can be made and released quickly and very easily.

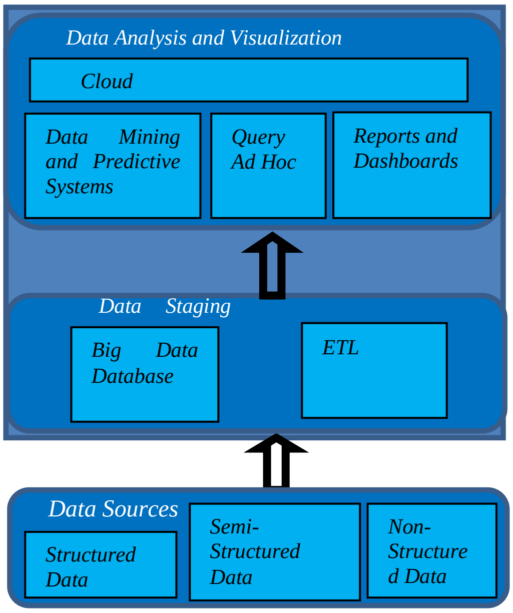

Figure 1 - Conceptual Architecture

4 Test Case

Test case was conducted by selecting a dataset of electronic commerce. This dataset was obtained from the American company Omniture - this company was acquired by Adobe in 2009, and provides data sets for analysis in various projects.

The first step was collecting and validating data by eliminating nulls and duplicate values. Next step was to select data to feed business indicators previously defined and send this data to a repository.

Two questions were defined to be answered by the solution:

1) What is the behavior of online users regarding purchases made in USA?

2) What age group has used more online shopping and what products they bought?

To take full advantage of customers information when they visit a sales website it was necessary cross-checking information of customers' logs and records of purchases and aggregate this information in a single repository of data.

Dashboards and reports were built to allow analysis and visualization of information.

This solution was implemented in a Cloud environment with implementing some levels of security to access and analyze data. The repository enables data encryption which increases user security confidence level to use a Cloud environment.

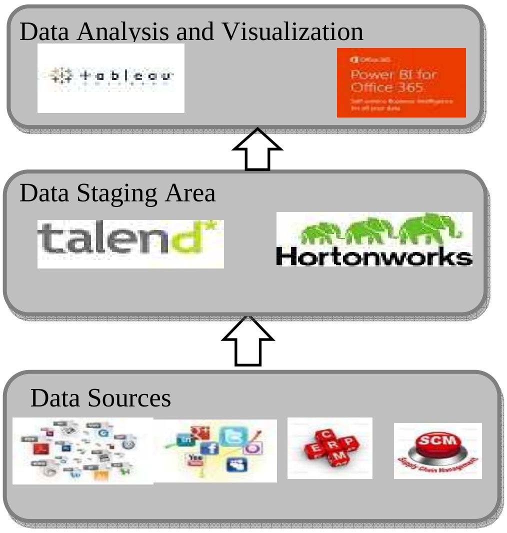

Figure 2 - Technological Architecture

4.1 Analysis Performed

To answer to the first question "What are the states with the highest volume of online shopping?" a dashboard was created and shows the states of USA where large volumes of purchases are made and it is possible to see the most bought item was clothing. It is also possible to check that, in most states, clothing is more representative. To answer to the second question "What are the most popular products by age group?" a second dashboard was created, see figure 3

4.2 Outcome Assessment

Results are a reference to volume purchases that are made through online shopping. It is not easy to relate such data, use of unstructured data forces to realize the value that data can provide and how much it can contribute to a better organizational decision making process. It is found that unstructured data work as a complement to structured data because it causes decisions to be made with better confidence.

Finally, should be noted that solution performance depends on many factors, but it is important to highlight two of these factors:

1. results rely heavily on data. If data are not appropriate the solution is not the best;
2. importance of data transforming, cleaning and data modeling, ensures integrity of data contributing to achievement of best results in data analytics.

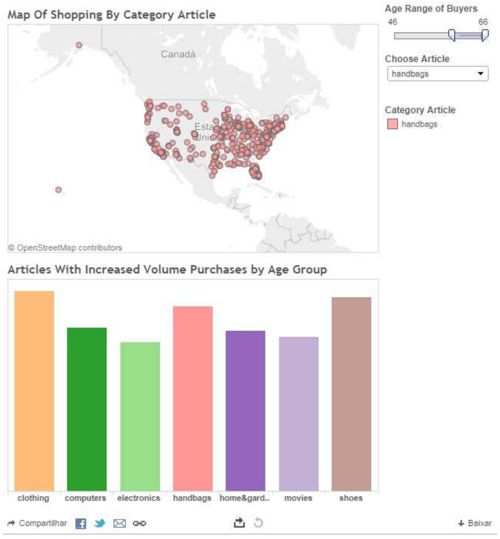

Figure 3 - Online Shopping Analysis

5 Conclusions

Big data architecture in Cloud should allow storage of various data types as structured, semi-structured and unstructured data, there must be mechanisms that allow interaction with these data, including the use of SQL and NoSQL. Organization and distribution of data must be done using storage systems with the capability to store various types and sources of data, systems such as HDFS and Hive, however these systems must have some features like: fast access to data, ensuring consistency and data integrity through security mechanisms that allow data access only to users who have permission to manipulate that data.

Correct implementation of access management policies, allows users rely on use of Cloud solutions, especially solutions for data analysis, mainly due to confidentiality issues as importance of data on organizational processes. An important feature that databases should provide is data encryption and multiple levels of security, this feature can prevent, e.g., Database Administrator (DBA) have access to most sensitive organization's data without authorization. With these measures, users gain confidence in Cloud solutions.

Use of Big Data in Cloud environment, requires rethinking of data architectures, new architectures must take into account characteristics such as robustness, availability and easy access of data for analysis, allowing data structured, semi-structured and unstructured, while preserving crucial aspects such as security and data integrity.

In order to validate the architecture it will be necessary to implement in a real project. One aspect very important is the incorporation of text mining analysis and data mining.

Acknowledgments. This work was financed by FCT - Fundação para a Ciência e Tecnologia for the project: PEst-OE/EEI/UI0319/2014.

References

[1] Agrawal, D., Das, S., & Abbadi, A. E. (2010). Big Data and Cloud Computing: New wine or just new bottles. *PVLDB, 3(2)*, pp. 1647–1648.

[2] Böhm, M., Leimeister, S., Riedl, C., & Krcmar, H. (2011). Cloud Computing–Outsourcing 2.0 or a new Business Model for IT Provisioning?. Application Management. *Gabler Verlag Springer Fachmedien Wiesbaden GmbH, Wiesbaden*, pp. 31-56.

[3] Guoliang, L., Beng, C. O., Jianhua, F., Jianyoung, W., & Lizhu, Z. (2008). EASE: an effective 3-in-1 keyword search method for unstructured, semi-structured and structured data. *SIGMOD '08 Proceedings of the 2008 ACM SIGMOD international conference on Management of data*, pp. 903-914

[4] Halper, F., & Krishnan, K. (2013). TDWI Big Data Maturity Model Guide Interpreting Your Assessment Score. *TDWI Benchmark Guide 2013–2014*.

[5] Kimball, R., & Ross, M. (2013). The Data Warehouse Toolkit: The Definitive Guide to Dimensional Modeling, Third Edition. *Wiley: The Data Warehouse Toolkit: The Definitive Guide to Dimensional Modeling, Third Edition*.

[6] Laney, D. (2001). 3D data management: Controlling data volume, velocity and variety. META Group Research Note, 6.

[7] Lohr, S. (2012). The Age of Big Data. *The New York Times.*

[8] Manyika, J., Chui, M., Brown, B., Bughin, J., Dobbs, R., Roxburgh, C., & Byers, A. H. (2011). Big data: The next frontier for innovation, competition, and productivity.

[9] Olivier, B., Thomas, B., Heinz , D., Hanspeter , C., Babak , F., Markus, F., Grivas, S., et al. (2012). Cloud Computing. *Swiss Academy of Engineering Sciences.*

[10] Oswaldo, T., Pjotr, P., Marc, S., & Ritsert, C. J. (2010). Big data, but are we ready. pp. 647-657.

[11] Russom, P. (2013). Managing Big Data. *TDWI Best Practices Report Fourth Quarter 2013*.

[12] Stonebraker, M. (2012). What Does 'Big Data' Mean? *Communications of the ACM*.

[13] Stuckenberg, S., Fielt, E., & Loser, T. (2011). The Impact Of Software-As-A-Service On Business Models Of Leading Software Vendors. *Experiences From Three Exploratory Case Studies. Proceedings of the 15th Pacific Asia Conference on Information Systems (PACIS 2011)*.

[14] Youseff, L., Butrico, M., & Da Silva, D. (2008). Toward a Unified Ontology of Cloud Computing. Grid. *GCE'08*.