A Proposed Architecture to Support the Processing and Analysis of Structured and Unstructured Massive Data Sets in the Brazilian Army

Marcio de Carvalho Victorino
Computer Science Department, University of Brasília
Brasília, Brazil
e-mail: mcvictorino@cic.unb.br

Danilo Rodrigues Azeredo Silva
Computer Science Department, University of Brasília
Brasília, Brazil
e-mail: danilo@draconsultoria.com

Marçal de Lima Hokama
Science and Technology Department, Brazilian Army
Brasília, Brazil
e-mail: lima@cds.eb.mil.br

***Abstract*— In recent years, the demand for processing a great volume of data of different formats has increased considerably. In order to address such a demand, a number of new data models have been proposed, which are different from relational models. The databases that implement such models are commonly known as Non-Structured Query Language (NoSQL). NoSQL databases have become excellent persistence devices for the Big Data environment because they provide structural flexibility, horizontal scalability, unstructured data support and distributed processing. However, these databases are not suitable to support ad-hoc analysis, very common to the conventional Online Analytical Processing (OLAP) architecture. On the other hand, the OLAP architecture has a multidimensional data repository, called Data Warehouse (DW), not able to support the new data storage demand. This work presents a preliminary study within the Brazilian Army to establish a new architecture that integrates the OLAP and Big Data environments to provide structured and unstructured data usage for decision support.**

Keywords-NoSQL; OLAP; Big Data.

I. INTRODUCTION

The Relational Database Management System (RDBMS) has become the most widely used Database Management System (DBMS) since the 1980s, due to its simplicity of representation and the data independence it provides.

However, the significant increase in the quantity and complexity of services offered on the web in the last decade has generated massive volumes of data in varying formats, demanding greater processing distribution and high flexibility of these data storage structures.

To support these new demands, a new database generation without rigid schemes arose, based on non-relational models, called NoSQL databases.

Even though the term "NoSQL" was used for the first time in 1990 by Carlos Strozzi [1], it was only in 2011 that this term began representing the database family that may be categorized in the following models [2][3]: graphs, documents, column family and key-value pair.

Based on these application domains, in which there are massive volumes of data, in a high variety of formats, which require processing in an appropriated velocity (3V's), the term "Big Data" was coined [4].

In 2012, a project that accounts for this context of the 3V's (volume, variety and velocity) was implemented in the Brazilian Army – the Integrated System of Border Monitoring (in Portuguese, Sistema Integrado de Monitoramento de Fronteiras, SISFRON) [19]. SISFRON uses sensoring, communication and information technology resources to cover 16.886 Kilometers of the Brazilian border, monitoring nearly 27% of the Brazilian national territory, and generating a huge data volume in various formats. These data will become an important source for decision support. Its cost is estimated at approximately US$ 4 billion (R$ 12 billion) to be invested over a 10 year period.

Notably, from the great volume and variety of data analysis, this project may also create opportunities that have a positive impact not only for the Brazilian Army, but also for the greater society. However, in this initiative, it is important that the organization chooses appropriate technologies to build its Big Data ecosystems [5].

The NoSQL DBMSs are commonly used in Big Data applications. However, these databases are not appropriate for supporting ad-hoc analysis, very common in Decision Support Systems (DSS) [6].

On the other hand, the OLAP architecture, proposed in the 90s [7], provides broad support for ad-hoc analysis; however, it cannot support all the demands of a Big Data environment [8].

Subsequently, with regard to decision support, the Brazilian Army has already deployed a system that uses OLAP architecture. However, it is clear that there is a need to extend this architecture to provide the processing of new data sources, from new systems such as SISFRON, in differing formats.

In this paper, we present the analysis of possible architectures to be adopted by the Brazilian Army, that integrate the OLAP and Big Data characteristics. The aim is to provide the support for the processing and analysis of

 ISBN: 978-1-61208-408-4

massive structured and unstructured data in the Brazilian Army.

The paper is organized as follows: Section II presents the concept of Big Data. Section III presents the Decision Support Environment in the Brazilian Army. Section IV presents proposals of architecture for Big Data Analysis environments. Finally, conclusions are included in Section V.

II. BIG DATA

There are various definitions and understandings for the term "Big Data". One of the most widely accepted is the '3Vs' definition presented by Doug Laney [9] in 2001 and ratified by Gartner [10] in 2012: "Big data is high-volume, high-velocity and high-variety information assets that demand cost-effective, innovative forms of information processing for enhanced insight and decision making."

For tackling these new challenges, a new generation of technologies has been proposed, such as Hadoop [11]. Created by Doug Cutting, Hadoop was inspired by BigTable, which is Google's data storage system; by the Google File System; and by MapReduce [11]. Hadoop is a framework based on Java and a heterogeneous platform with open source license. This framework includes a distributed file system; a data storage platform; and a management parallel computation layer, a work flow and configuration administration. The Hadoop Distributed File System (HDFS) is executed in all the nodes of a Hadoop cluster and connects the file systems in many input and output data nodes, to turn them into a unique large file system.

MapReduce [12], presented by Google [13], has a processing approach that divides the Big Data complex problems into small work units and processes them in parallel.

The term "Hadoop" is also used to designate a family of related projects that are under the infrastructure of distributed computation and data processing in large scale, such as Common, Avro, MapReduce, HDFS, Pig, Hive, HBase, ZooKeeper, Sqoop, among others [11].

III. DECISION SUPPORT IN BRAZILIAN ARMY

The Brazilian Army has many computerized systems in various administrative and operational areas, supporting their respective processes.

Such systems generate and store a large quantity of data related to the processes. However, these data are not related to each other. The data, produced and stored by the Brazilian Army, are assets to those sectors that produce them, but this raw data cannot be used as a strategic resource because it is not integrated.

Aiming to minimize this problem, in 2008, the Brazilian Army began the DSS initiative, called Management Integrated System (MIS). This system is organized according to the OLAP architecture proposed by Kimball [7], comprised of basically four layers: the layer created by relational data sources; the extraction, transformation and load (ETL) layer; the storage layer; and the presentation layer.

In this architecture, the data that come from the source systems go through the ETL processing to be loaded in multidimensional repositories called, Data Warehouse (DW). The goal is to integrate the data that come from various sources and transform them into information.

The MIS is comprised of one Data Mart (DM) in each activity area, initiated by: the Human Resources in 2009, Military Service in 2010, Logistics in 2011, Personal Assessment in 2012 and Finances in 2013. The group formed by these five DM was consolidated in a Brazilian Army DW. However, with new data sources from SISFRON this conventional OLAP architecture has become unable to support the huge data volume and unstructured data from this system.

The SISFRON systemic architecture is composed of the following subsystems: sensoring, decision support, communication and information technology, information security, simulation and logistic. The sensoring subsystem is composed of terrestrial and aerial surveillance radars, weather stations and radars, and electromagnetic and optical sensors. Its main function is to obtain data to provide surveillance and event detection. The decision support subsystem has the aim of treating the data collected by the sensors for following visualization.

For supporting SISFRON data source, the next step of the MIS project will consist in providing unstructured data processing and analysis in the Brazilian Army. However, Kimball and Ross [8] claim that the DW based on RDBMS is not able to store the large data variety available in big organizations, making it necessary to create a Big Data environment.

IV. ARCHITECTURE PROPOSAL

The NoSQL databases have become excellent persistence devices in Big Data environments, for providing structural flexibility, high horizontal scalability, unstructured data support and distributed processing [3].

These databases were created to support operational processes that manipulate a large volume of data in various formats in an appropriate time period, but not to provide more elaborated analysis for decision support [3].

In this context, Big Data Analysis is understood as a procedure set executed on a large scale, on large data repositories, in which the main goal is knowledge extraction [14].

Another important aspect to be observed in data analysis in critical environments is the fact that there are few query tools able to access data repositories available to Hadoop, which is the main framework used in Big Data environments [3]. Examples of these tools are Hive and Pig. Besides the small number of tools, the few that exist require deep programming knowledge to generate analytical queries about the NoSQL databases or other databases.

The OLAP report tools, on the other hand, provide complete support to ad-hoc analysis in which the final user may access the dimensionally organized data in a DW to generate the sheets or graphics without creating any code line.

 ISBN: 978-1-61208-408-4

There are many issues when trying to enable the use of OLAP architecture with DW to support decision-making in a Big Data environment. The main limiting issues for the joint use of these two technologies are the high volume and data complexity available in Big Data environments [15].

The great challenge is to provide architecture that makes it possible to use the structured and unstructured data together in Big Data environments to support decision-making processes.

Davenport [16] presents a proposal to merge the two technologies OLAP and Big Data in an integrated environment. Figure 1 presents conventional OLAP architecture [7]. In the figure, we see that the data sources, mainly those that are structured, are submitted to an ETL process to be stored in DW and, after this, presented to the final users by the OLAP report tools.

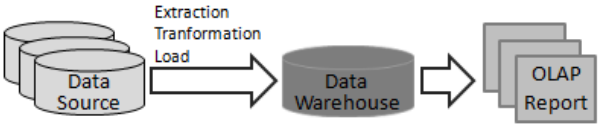

Figure 1. Conventional OLAP architecture [7].

Figure 2 presents an extension of this conventional architecture according to Davenport [16] to provide its use in Big Data environments.

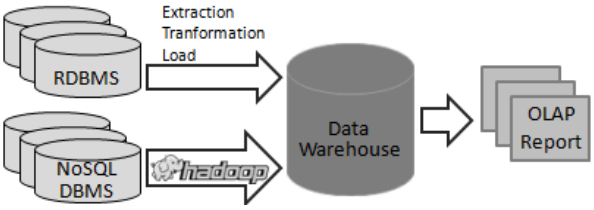

Figure 2. Extended OLAP architecture [16].

In the architecture presented in Figure 2, the data sources are organized in various formats such as relational tables, graphs, documents, column family and key-value pair. In this architecture, the Hadoop framework performs two roles. It may work as a data presentation tool for the operational queries, and may perform the ETL tool role for the stored data in NoSQL databases, since it transports the data from their sources to the DW and executes the transformation process. In this architecture, all the analytic data are stored in a DW verifying an integrated analysis from the information crossing from many areas, generating the result through OLAP reports.

Thus, the various DBMS that compose the SISFRON, such as MySQL and Postgres (relational), Riak (key-value), MongoDB (documents), Cassandra (column family) and Neo4J (graphs) will be used as resources to support decisions.

However, this architecture is limited, since the current OLAP environment is not able to support the high data volume and variety in a Big Data environment [8].

Gartner [17] presents some proposals for architectures in Big Data Analytics environments. Figure 3 presents an adaptation of one of these proposals.

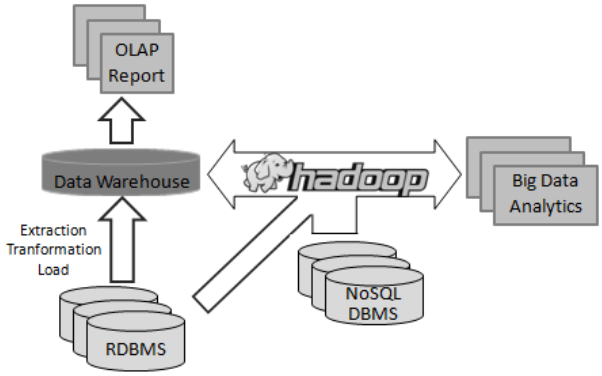

Figure 3. Gartner [17] adapated Big Data Analytics Architecture.

The architecture presented in Figure 3 is broader than the one presented in Figure 2. It provides the generation of analytical reports from the DW through the traditional OLAP tools, or directly from the data repositories that comprise the Big Data (RDBMS and NoSQL DBMS) through the analytical query tools that are part of the Hadoop framework.

In this architecture, the DW stores the data from the RDBMS and part of the data that comes from NoSQL DBMS extracted through the Hadoop, since the DW has storage limitations. On the other hand, as the Hadoop doesn't have limitations related to distributed data source access, it becomes possible to analyze high data volume and variety compared to the OLAP architecture. However, it is worthy to note that the analysis tools available for the Hadoop are not as mature as the OLAP tools.

Hence, as the Brazilian Army already has the OLAP architecture presented in Figure 1, after a thorough study of many architectural options presented in [7] [16] [17], it was concluded that the Brazilian Army will focus efforts, initially, on the architecture presented in Figure 2. However, due to the fact that this architecture has limitations, it is intended, in a second phase, to migrate to a broader architecture presented in Figure 3, bearing in mind that the architecture presented in Figure 3 is an extension of the one presented in Figure 2. No work will be wasted during the transition.

Due to the fact that the architecture initially slated for adoption by the Brazilian Army cannot support all the volume and variety of the existing data in the institution, there is a clear need to consider the following aspects as subsidies to the process of data extraction of the Big Data, stored in NoSQL DBMS to load in DW:

A. Selectivity

Since the DW of the OLAP architecture is not able to store all the data volume of the Big Data, it is necessary to prioritize the activity areas of the observed domain. Subsequently, the data extraction of the Big Data for the greater load in the DW must start with the data of the most important areas.

 ISBN: 978-1-61208-408-4

B. Summarization

Inmon [18] advises storing data in the DW, or, more precisely, in the fact table – the fact in the lowest level of granularity. However, this procedure is not viable in the proposed architecture presented in Figure 2, because, as already mentioned, the DW is not able to store all the data of the Big Data environment. Hence, it is necessary to summarize the data to create important aggregates for the observed domain, decreasing the space needed for data storage in DW. The main drawback of this line of action is to limit the drill down operation, since the level of detailing of the fact will depend on the lowest level of granularity of the aggregates generated and stored in the DW. Even so, this procedure is necessary.

C. Iterative ETL

The iterative ETL refers to the act of executing the ETL processes in an iterative manner, i.e., the employing of one finite sequence of ETL processing, in which the purpose of each one of the steps is to add data to the result of the previous step. Hence, the step sequence will be limited to the DW storage capacity.

Right after the consolidation of the architecture presented in Figure 2, the efforts for the implementation of the architecture presented in Figure 3 will be initiated. In this step, the critical issues for this environment quoted by Cuzzocrea et al. [15] will be considered, among them are: the data volume and complexity; new project methodologies; hardware; query performances; usability and data quality.

A thorough study about Big Data Analytics tools available for the Hadoop framework will carried out. As soon as the tools are mature enough, the architecture of Figure 3 will substitute the one presented in the Figure 2.

V. CONCLUSION

The growing sophistication of the available applications on the web has generated massive volumes of data in various formats that require processing at an appropriate velocity. This environment is called “Big Data”.

The RDBMS have become unable to attend to all of the Big Data environment demands. In this context, users began employing NoSQL DBMS as data persistence devices.

It has become a great challenge to apply advanced analytical techniques to Big Data data sets, Big Data Analytics, since the main framework for Big Data environments, the Hadoop, does not have mature ad-hoc query analytical tools.

On the other hand, OLAP architecture has many mature ad-hoc query analytical tools, but its repository, the DW, is not able to store all the data from Big Data.

Thus, we have presented an initial Brazilian Army study to adopt an architecture capable of providing data extraction from NoSQL data sources for further load in a DW, providing the availability of a Big Data Analytics environment with OLAP architecture in this institution.

From our research results and discussion here, it is possible to conclude that the Brazilian Army plans initially to extend its OLAP architecture to deal with unstructured data for subsequently migrating to a broader architecture.

REFERENCES

[1] NoSQL: a non-SQL RDBMS. [retrieved: 2014, 11]. [Online]. Available: http://www.strozzi.it/cgi-bin/CSA/tw7/I/en_US/NoSQL/Home%20Page

[2] P. J. Sadalage and M. Fowler, “NoSQL Distilled: A Brief Guide to the Emerging World of Polyglot Persistence”, New Jersey, IN: Pearson Education, Inc., 2012.

[3] M. Indrawan-Santiago. “Database Research: Are We at a Crossroad? Reflection on NoSQL’. Fifteenth International Conference on Network-Based Information Systems, 2012, pp. 45–51.

[4] S. Singh and N. Singh, "Big Data Analytics", 2012 International Conference on Communication, Information & Computing Technology Mumbai India, IEEE, October 2011.

[5] N. Heudecker and H. Lehong. Applying the Big Data Ecosystem. [retrieved: 2014, 07]. [Online]. Available: http://www.gartner.com/document/code/252014

[6] M. Indrawan-Santiago. “Database Research: Are We at a Crossroad? Reflection on NoSQL’. Fifteenth International Conference on Network-Based Information Systems, 2012, pp. 45–5.

[7] R. Kimball, L. Reeves, M. Ross and W. Thornthwaite. “The Date Warehouse Lifecycle Toolkit”, New York, IN: John Wiley & Sons, 1998.

[8] R. Kimball and M. Ross, “The Data Warehouse Toolkit: The Definitive Guide to Dimensional Modeling”, 3nd ed., Indiana, IN: John Wiley & Sons, 2013.

[9] D. Laney, "Application Delivery Strategies. Meta Group, Retrieved." [retrieved: 2014, 10]. [Online]. Available: http://blogs.gartner.com/doug- laney/files/2012/01/ad949-3D-Data-Management-Controlling- Data-Volume-Velocity-and-Variety.pdf

[10] M. A. Beyer and D. Laney, “The Importance of ‘Big Data’: A Definition” [retrieved: 2014, 08]. [Online] Available: https://www.gartner.com/doc/2057415/importance-big-data-definition

[11] T. White, "Hadoop: The Definitive Guide", 3nd. ed, IN: O'Reilly Media, 2012.

[12] J. Dean and S. Ghemawat, “MapReduce: Simplified Data Processingon Large Clusters” Google, Inc OSDI 2004.

[13] S. Sagiroglu and D. Sinanc, “Big Data: A review, Collaboration Technologies and Systems (CTS)”, 2013 International Conference, 2013, pp. 42 - 47, IEEE Conference Publications.

[14] A. Cuzzocrea, “Analytics over Big Data: Exploring the Convergence of DataWarehousing, OLAP and Data-Intensive Cloud Infrastructures”, Computer Software and Applications Conference (COMPSAC), 2013 IEEE 37th Annual, 2013, pp. 481 - 483, IEEE Conference Publications.

[15] A. Cuzzocrea, L. Bellatreche, and Il-Yeol Song, “Data warehousing and OLAP over Big Data: current challenges and future research directions”, In Proceedings of the sixteenth international workshop on Data warehousing and OLAP (DOLAP '13). ACM, New York, NY, USA, 2013, pp. 67-70.

[16] T. H. Davenport, “Big Data at Work: Dispelling the Myths”, Harvard Business Press, 2014.

[17] S. Sicular, "Hadoop Architecture: Multiple Choices — Which One Is Right?". Lecture Notes in Gartner Catalyst Conference, Brasília, Brazil, 2014.

[18] W. H. Inmon, “Building the Data Warehouse”, 4nd ed., Indiana, IN: Wiley Publishing, Inc., 2005.

[19] Brazilian Army, "O Sistema Integrado de Monitoramento de Fronteiras (SISFRON)" [retrieved: 2014, 05]. [Online]. Available: http://www.ccomgex.eb.mil.br/index.php/pt_br/sisfron/arquitetura