Chapter 50
Evaluating NoSQL Databases for Big Data Processing within the Brazilian Ministry of Planning, Budget, and Management

Ruben C. Huacarpuma
University of Brasília, Brazil

Rafael T. de Sousa Júnior
University of Brasília, Brazil

Daniel da C. Rodrigues
University of Brasília, Brazil

Lizane Leite
University of Brasilia, Brazil

Antonio M. Rubio Serrano
University of Brasília, Brazil

Edward Ribeiro
University of Brasilia, Brazil

João Paulo C. Lustosa da Costa
University of Brasília, Brazil

Maristela Holanda
University of Brasilia, Brazil

Aleteia P. F. Araujo
University of Brasilia, Brazil

ABSTRACT

The Brazilian Ministry of Planning, Budget, and Management (MP) manages enormous amounts of data that is generated on a daily basis. Processing all of this data more efficiently can reduce operating costs, thereby making better use of public resources. In this chapter, the authors construct a Big Data framework to deal with data loading and querying problems in distributed data processing. They evaluate the proposed Big Data processes by comparing them with the current centralized process used by MP in its Integrated System for Human Resources Management (in Portuguese: Sistema Integrado de Administração de Pessoal – SIAPE). This study focuses primarily on a NoSQL solution using HBase and Cassandra, which is compared to the relational PostgreSQL implementation used as a baseline. The inclusion of Big Data technologies in the proposed solution noticeably increases the performance of loading and querying time.

DOI: 10.4018/978-1-4666-9840-6.ch050

INTRODUCTION

Over the past years, Big Data storage and management have become challenging tasks. According to Russom (2011), when the volume of data started to grow exponentially in the early 2000s, storage and processing technologies were overwhelmed managing hundreds of terabytes of data. In addition, the heterogeneous nature of the data presents challenges that must be taken into consideration. Such characteristics can be observed in different domains such as social network operations, gene sequencing or cellular protein concentration measurement (Andrew, 2012). Moreover, improved Internet connections and new technologies, such as smartphones or tablets, require faster data storage and querying. For these reasons, organizations and enterprises are becoming more and more interested in Big Data technologies (Collet, 2011).

Along with well-known IT companies such as Google and Facebook, governments are also interested in Big Data technologies in order to process information related to education, health, energy, urban planning, financial risks and security. Efficient processing of all this data reduces operating costs, thereby economizing the investment of public resources in a more rational way (Office of Science and Technology Policy of The United States, 2012). In the same way as other governments, Brazil is starting to employ Big Data technologies in its IT systems.

In this chapter, we propose the use of Big Data technology to solve the limitations observed on the SIAPE database processing. Notably, the SIAPE system controls payroll information regarding all federal public sector employees in Brazil. Given its growth rate of 16GB per month, the SIAPE database can be characterized as a relevant Big Data case. Specifically we focus on the Extract, Transform and Load (ETL) (Jun, 2009) modules in this system, which in our proposal are modified in order to operate with NoSQL data storage, using HBase and Cassandra. Thus, the SIAPE database is used as a case study in this chapter in order to validate our proposal.

The remainder of this chapter is structured as follows: Section 2 presents basic concepts related to Big Data; in Section 3, we describe the use case including our proposed solution; Section 4 discusses our implementation results. Finally, Section 5 presents our conclusions.

BIG DATA

The current data we manage is very diverse and complex. This is a consequence of social network interactions, blog posts, tweets, photos and other shared content. Devices continuously send messages about what they or their users are doing. Scientists are generating detailed measurements of the world around us with sensors installed within devices such as mobile telephones, tablets, watches, cars, computers, etc. and finally the internet is the ultimate source of data with colossal dimensions (Marz, 2013).

Big data is exceeding the conventional database systems capacities. The data is too big, moves too fast or does not fit into existing database architectures (Dumbill, 2012). Although the literature usually defines Big Data based on the size of the data, in this work point of view, Big Data is not only defined by the size, but also according to Russom (2011), we take into account the so called by 3Vs factors, i.e. Volume, Variety and Velocity.