Chapter 2
Real-Time Big Data Warehousing

Francisca Vale Lima
University of Minho, Portugal

Carlos Costa
University of Minho, Portugal

Maribel Yasmina Santos
University of Minho, Portugal

ABSTRACT

The large volume of data that is constantly being generated leads to the need of extracting useful patterns, trends, or insights from this data, raising the interest in business intelligence and big data analytics. The volume, velocity, and variety of data highlight the need for concepts like real-time big data warehouses (RTBDWs). The lack of guidelines or methodological approaches for implementing these systems requires further research in this recent topic. This chapter presents the proposal of a RTBDW architecture that includes the main components and data flows needed to collect, process, store, and analyze the available data, integrating streaming with batch data and enabling real-time decision making. Using Twitter data, several technologies were evaluated to understand their performance. The obtained results were satisfactory and allowed the identification of a methodological approach that can be followed for the implementation of this type of system.

INTRODUCTION

The technological evolution of the last years has called the attention of organizations for the analysis of data, increasing the interest in Business Intelligence (BI). BI allows the understanding of the business needs and opportunities, and represents a competitive advantage (H. Chen, Chiang, & Storey, 2012). The technologies available for data analysis have been increasingly requested, with a special focus on ways of extracting information from large volumes of data and identifying patterns and trends that support decision-making (Di Tria, Lefons, & Tangorra, 2014b).

DOI: 10.4018/978-1-5225-5516-2.ch002
The volume, velocity and variety of data have imposed considerable challenges to traditional data storage and processing technologies, being almost impossible using them to extract useful information from data (Cuzzocrea, Song, & Davis, 2011). Traditional technologies fail to respond to requests on time and, therefore, solutions based on Big Data concepts were introduced (M. Chen, Mao, & Liu, 2014; Goss & Veeramuthu, 2013; Zikopoulos, Eaton, DeRoos, Deutsch, & Lapis, 2011), substituting traditional data storage and processing technologies with much more efficient ones (H. Chen et al., 2012).

Data Warehouses (DWs) in Big Data contexts, i.e., Big Data Warehouses (BDWs), allow the analysis of large volumes of data, extracting relevant information from them, in order to fulfill organizational analytical needs (Di Tria, Lefons, & Tangorra, 2014a). The use of BDWs increases the ability to question data faster than usual, also enhancing the access to real-time data. Having more updated data prepared for analysis is nowadays of upmost importance, creating the need for Real-Time Big Data Warehouses (RTBDWs), as traditional tools are not able to process large volumes of data in real-time.

This is even more relevant in a context where the advent of real-time technology (e.g., distributed message queueing systems and stream processing) makes faster data changes, requiring an up-to-date analysis of large amounts of data. The challenge lies in real-time data access with no processing delays and a reduced latency of processing operations (Li & Mao, 2015). Therefore, it is relevant to understand the real-time requirements for modeling and implementing BDWs, in order to obtain updated information in a faster way, enhancing business’s competitive advantage.

In order to materialize the real-time requirements for BDWs, this work proposes a BDW architecture for real-time processing. Although some research contributions can be found in the literature (as can be seen in the next section), this work proposes an innovative approach that allows timely collection, processing, storage and analysis of real-time data.

This work explores and evaluates the role of each technology included in the proposed architecture, and summarizes a set of considerations for the implementation of BDWs in real-time. By analyzing the performance of the technologies in different scenarios, it is possible to understand and identify best practices regarding streaming data collection, processing and storage. For that, both real-time data repository and a historical data repository are used, allowing data to flow faster from the data source to the data analysis component. The logical architecture proposed in this work can be adopted by any organization, making available an analytical environment able to integrate historical data with streaming data.

The following sections, in outline, include: i) an overview of related work and the scientific contributions of this work; ii) the BDW architecture for real-time contexts; iii) the demonstration case describing data collection, processing and storage; iv) the results obtained from the several benchmarks, discussing the main findings; v) some remarks about the undertaken work and its future developments.

**Big Data Warehousing**

Nowadays, organizations are increasingly aware of BI and Business Analytics. However, as the volume of available data has been increasing at ever growing rates, its integration for analysis purposes becomes increasingly complex (M. Chen et al., 2014; Goss & Veeramuthu, 2013). The DW is the essential component of a BI infrastructure, integrating strategic indicators that bring analytical value to the decision-making process (Vaisman & Zimányi, 2012), which highlight relevant hidden patterns or insights in data (Di Tria et al., 2014a). Although a DW allows the collection, storage and management of data to support the decision-making process (Krishnan, 2013), it is constrained by the volume, velocity and variety of