Future Generation Computer Systems 63 (2016) 100–107

Contents lists available at ScienceDirect

Future Generation Computer Systems

journal homepage: www.elsevier.com/locate/fgcs

From Business Intelligence to semantic data stream management

Marie-Aude Aufaure [a], Raja Chiky [b,*], Olivier Curé [c], Houda Khrouf [d], Gabriel Kepeklian [d]

[a] MAS Lab Ecole Centrale Paris, France
[b] ISEP - LISITE, Paris, France
[c] LIP6 (UMR 7606/CNRS), Université Pierre et Marie Curie (UPMC), Paris, France
[d] Atos Integration, Paris, France

HIGHLIGHTS

- Evolution of Business Intelligence with emergence of Big Data technologies.
- New technologies and approaches the 3Vs (Volume, Velocity and Variety) of Big data.
- Stream reasoning over Big Data.
- Summarizing data streams (semantic and classic data).
- Semantic data matching in stream context.

ARTICLE INFO

Article history:
Received 21 May 2015
Received in revised form
4 November 2015
Accepted 10 November 2015
Available online 2 December 2015

Keywords:
Data stream
Linked Data
Business Intelligence
Stream reasoning

ABSTRACT

The Semantic Web technologies are being increasingly used for exploiting relations between data. In addition, new tendencies of real-time systems, such as social networks, sensors, cameras or weather information, are continuously generating data. This implies that data and links between them are becoming extremely vast. Such huge quantity of data needs to be analyzed, processed, as well as stored if necessary. In this position paper, we will introduce recent work on Real-Time Business Intelligence combined with semantic data stream management. We will present underlying approaches such as continuous queries, data summarization and matching, and stream reasoning.

1. Introduction

The main objective of Business Intelligence is to transform data into knowledge for a better decision-making process. The constant growth of data and information, coming from heterogeneous data sources has led to new ways of interaction and the integration of new models and tools to cope with this heterogeneity. We manipulate more and more unstructured data documents, emails, social networks, contacts that need to be integrated with classical structured data like CRM, data stored in relational databases. We also need more and more interactivity, flexibility, dynamicity and expect the system to be proactive and reactive. Users expect immediate feedback, and want to find information rather than merely look for it. Moreover, the company tends to be organized in a collaborative way, called enterprise 2.0 [1]. All these evolutions induce challenging research topics for Business Intelligence, such as providing efficient mechanisms for a unified access and model to both structured and unstructured data. Semantic technologies are a perfect fit for integrating and matching data. Business Intelligence integrates collaborative and social software, by combining BI with elements from both Web 2.0 and the Semantic Web. Extracting value from all these data, a crucial advantage for companies, requires business analytics. In order to synthesize information and derive insights from massive, dynamic, ambiguous data, the use of data visualization techniques and visual analytics becomes critical. Business Intelligence is also impacted by big data, and need to account for the volume of data sources as well as the need of response in real-time for extracting value from trusted data.

This position paper addresses the integration of real-time analytics with semantic technologies. Many research work has been done separately in these two fields, but, to the best of our

* Corresponding author.
E-mail addresses: marie-aude.aufaure@ecp.fr (M.-A. Aufaure), raja.chiky@isep.fr (R. Chiky), olivier.cure@lip6.fr (O. Curé), houda.khrouf@atos.net (H. Khrouf), gabriel.kepeklian@atos.net (G. Kepeklian).

http://dx.doi.org/10.1016/j.future.2015.11.015
0167-739X/

knowledge, only a few ones provide an integrated view. This is mainly due to scalability issues for semantic reasoning.

The rest of this paper is organized as follows. Section 2 describes the new needs in Business Intelligence and presents a generic architecture for semantic data stream management platform. Section 3 focuses on related work in the area of semantic data streaming. Section 4 describes data matching in an RDF stream context. Section 5 gives an overview of reasoning in the context of RDF stream processing. Finally, Section 6 concludes this paper and gives an outlook upon future research for managing large-scale semantic streaming data.

2. From BI to semantic data stream management

Business Intelligence (BI) refers to a set of tools and methods dedicated to collecting, representing and analyzing data to support decision-making in enterprises. BI is defined as the ability of an organization to take all input data and convert them into knowledge, ultimately, providing the right information to the right people at the right time via the right channel. During the last two decades, numerous tools have been designed to make available a huge amount of corporate data for non-expert users. Business Intelligence is a mature technology, widely adapted, but faces new challenges for incorporating new data such as unstructured data or data coming from sensors or social networks into analytics. A key issue is the ability to analyze in real-time these constantly growing amounts of data, taking their meaning into account. The complexity of BI tools and their interface is a barrier for their adoption. Thus, personalized systems and user modeling [2] have emerged to help provide more relevant information and services to the user. Information visualization and dynamic interaction techniques are key for enhancing the user experience in using such tools [3].

Traditional BI systems offer tools for structuring and storing data in a data ware-house, in which data are modeled with a multidimensional model representing the analysis axis. Key performance indicators can be computed from this model and restituted to the user in a static dashboard.

These systems can be extended with semantic technologies to capture the meaning of data and new ways of interacting with data, intuitive and dynamic. Semantic technologies [4,5] focus on the meaning of data and are capable of dealing with both unstructured and structured data. Having the meaning of data and a reasoning mechanism may assist a user during his analysis task. The vision of the FP7 CUBIST[1] project was to extend the ETL process to both structured and unstructured data, to semantically store data in a triple store and to provide user-friendly visual analytics capabilities leading to dynamic dashboards. Then, the information provided to the user is not composed of only quantitative values like key performance indicators, but can also integrate qualitative values represented by a graph or a lattice extracted from formal concepts (a formal concept is a set of objects sharing properties; the formal concepts are then organized in a lattice linked together by a relation of inclusion). The user can then navigate into these semantic data through a visual analytic tool [6].

More recently, business intelligence has been impacted by big data and, in particular, need to take into account the velocity i.e. the ability to provide information or alerts in real-time from streams. With the exponential growth of sensor networks, web logs, social networks and interconnected application components, large collections of data are continuously generated with high speed. These data are called "data streams": there is no limit on the total volume of data and there is no control over the order in which data arrive. The analysis methods (data mining, machine learning) should self-adapt to these data and process them on the fly in one pass and in the order of their arrival. These heterogeneous data streams [7] are produced in real time and consequently, should be processed on the fly. Then, they are maintained, interpreted and aggregated in the purpose of reusing their semantics and recommending relevant alerts to the targeted stakeholders in order to react to interesting phenomena occurring in the input streams. A precious decision-making value can be enhanced through the semantic analysis of data streams, especially while crossing them with other information sources.

Coming back to semantic technologies, numerous techniques can be used to extract some meaning or knowledge from data sources. Among them, we can cite natural language processing techniques, data mining, machine learning and ontology engineering. These techniques are used to extract patterns or models, to structure data and to transform any information in actionable knowledge. Semantic Web technologies can be used for linking, publishing or searching for data on the web, but also for large-scale structuring and enriching data with the RDF semantic model.

Semantic-based approaches are useful to simplify the integration of heterogeneous data sources by the mean of ontologies and for offering a unified metadata layer. Semantics can also be used for discovering and enriching information, and finally, to provide a unified data access mechanism. Semantics addresses the variety from the 3 V of Big Data (Volume, Variety and Velocity) to generate value from heterogeneous data. The value of data also increases when they can be linked to other data (Linked Data). Semantic technologies can then be seen as a great opportunity to reduce the cost and complexity of data integration.

Fig. 1 represents a generic architecture of what could be a Real-Time BI platform in which structured and unstructured data streams are processed on the fly. In a Real-Time BI platform, multiple heterogeneous data sources can be connected, and data can be static or dynamic. The static data comes from standard databases or from open data, and does not change or in a minor way. Dynamic data comes as a stream, in a semantic format (RDF for example) or not (raw data). To process raw data, they need first to be converted into semantic format using ontologies. The idea is to maintain into the system an homogeneous format and a meaningful model that can be processed by machines. This architecture is composed of different components for processing streams: semantic filtering and continuous queries, data summarization, matching and reasoning. Semantic filtering is used for processing a large volume of streams on the fly. Existing systems are mainly based on RDF and SPARQL. To manage infinite real-time data stream, the platform has to provide the ability to create persistent continuous queries, which allow users to receive new results when they become available. Moreover, in the context of Big Data with a huge volume of data coming in high velocity, the platform provides some summarizing and load shedding techniques [8] that randomly drop data from the streams when the load of the platform increases beyond what it can handle. Data matching is used to enrich data with additional knowledge and to combine data streams coming from distributed data sources. Discovering relations between data is a key factor to add contextual information which may enhance decision making. This task is particularly challenging in a stream context where time-efficient techniques are needed to ensure scalability over high stream rates. Finally, reasoning is a key component in an RDF stream context and still considered as an open problem, mainly for reasoning using parallelized computation and expressive ontology language. These components are not fully integrated in existing architectures for enabling semantic stream processing. We introduce in the following sections a survey of research work related to semantic data stream management, summarizing techniques, data matching and reasoning.

[1] CUBIST EU FP7 project: http://www.cubist-project.eu/index.php?id=378.