Chapter 11
Integrating Heterogeneous Data for Big Data Analysis

Richard Millham
Durban University of Technology, South Africa

ABSTRACT

Data is an integral part of most business-critical applications. As business data increases in volume and in variety due to technological, business, and other factors, managing this diverse volume of data becomes more difficult. A new paradigm, data virtualization, is used for data management. Although a lot of research has been conducted on developing techniques to accurately store huge amounts of data and to process this data with optimal resource utilization, research remains on how to handle divergent data from multiple data sources. In this chapter, the authors first look at the emerging problem of "big data" with a brief introduction to the emergence of data virtualization and at an existing system that implements data virtualization. Because data virtualization requires techniques to integrate data, the authors look at the problems of divergent data in terms of value, syntax, semantic, and structural differences. Some proposed methods to help resolve these differences are examined in order to enable the mapping of this divergent data into a homogeneous global schema that can more easily be used for big data analysis. Finally, some tools and industrial examples are given in order to demonstrate different approaches of heterogeneous data integration.

INTRODUCTION

In today's business world, the strategic and tactical decisions of many business departments, such as marketing or inventory, is based on information derived from its large data stores. The rise in companies using "Big data"(Agrawal, 2010) is growing by 40% annually. The amount spent on collecting, storing, retrieving, and analyzing big data has been predicted to grow from $3.2 billion (US) in 2010 to $17.2 billion in 2015. (Leavitt, 2013) The growth in big data can be attributed to several factors: inexpensive storage, more sensor and data capture technologies used within a firm,

DOI: 10.4018/978-1-4666-5864-6.ch011

increasing access to information through the use of the cloud and virtualized storage infrastructures, and new analysis tools. The type of data collected has also increased in both size and variety; social media information, telephone conversations, and video surveillance are being increasingly included in a firm's data store (Gantz, 2011). Along with the rise of big data is an increasing business need to perform business analytics on this data to enable more accurate forecasting, more comprehensive reports, et al. One major aspect of data analytics is having a uniform, global view of data, regardless of the actual underlying data structures, (which is provided by data virtualisation) to enable uniform analysis of this data throughout the organization.

In this chapter, we look at the emerging problem of the greatly increasing volume and variety of data within many businesses and how it can bottleneck Big data analytics. As a solution, we look at data virtualisation which provides an abstract, global view with data mechanisms underneath this view to handle the diverse and distributed nature of its data. In this chapter, we briefly examine data virtualisation, with possible underlying data management mechanisms, and techniques to overcome the value, syntactical, semantic, and structural differences in data in order to map this data into an integrated global schema.

BACKGROUND: BUSINESS ANALYTICS WITH THEIR CHALLENGES

Over 90% of Fortune 500 companies have a Big data initiative this year. An IBM study has discovered that companies which use Big data analytics perform better than those who do not. (Leavitt, 2013) However, until legislative changes occur, certain industries, such as Finance and Healthcare, are currently required to keep all of their data in-house (Leavitt, 2013).

Analytics provides up-to-the-minute business insights, which have been derived from business data, which helps manage business risks and reduce compliance penalties. (Composite Customer Value Framework, 2012) However, the growing volume and complexity of business data increases business risks and reduces business agility in responding to new threats and opportunities. (Data Virtualization Platform Maturity Model, 2012) Notably, there is a rise in semi-structured data from Web services and non-relational data stores which must be integrated and analyzed for business insights. (Turbo Charge Analytics with Data Virtualization, 2013) Data access and integration pose the biggest bottleneck for analytics. (Data Virtualization Applied, 2012) An example, when a business is analyzing a typical marketing campaign, they must integrate and analyze diverse data from multiple sources: Website click statistics for their marketing Web site, email responses for leads, revenue feeds from Web services, et al.

The data are diverse:

- Third-party/desktop data
- Semi-structured data
- Unstructured, from multiple platforms.

However, all this sale and marketing data must be integrated in order for the business to understand the true impact of their marketing campaign. With this integrated and analyzed data, a broader analysis of the whole marketing campaign is made possible which can reveal which marketing components are more effective than others. Consequently, the most effective components can be enhanced to provide a large marketing impact. With real-time sale data available to sales agents, they can be more responsive to their customer's needs which results in higher sales revenues. With the easier and quicker integration of sales and marketing data, through data virtualisation, and a more thorough and faster analysis of this integrated data, faster marketing campaigns can be produced with a quicker response time (Turbo Charge Analytics with Data Virtualization, 2013).