Chapter 4
A Holistic View of Big Data

Won Kim
Gachon University, South Korea

Ok-Ran Jeong
Gachon University, South Korea

Chulyun Kim
Gachon University, South Korea

ABSTRACT

Today there is much hype about big data. The discussions seem to revolve around data mining technology, social Web data, and the open source platform of NoSQL and Hadoop. However, database, data warehouse and OLAP technologies are also integral parts of big data. Big data involves data from all sources, not just social Web data. Further, big data requires not only technology, but also a painstaking process for identifying, collecting, and preparing sufficient amounts of relevant data. This paper provides a holistic view of big data.

INTRODUCTION

Big data is one of the current IT buzzwords. It refers roughly to extraction of actionable intelligence from a large amount of data, including social Web data, and applying it to some important needs of an organization. The data may be stored in the proprietary databases of an organization or purchased from third-party data providers or may be gathered from the Internet.

Although there is much hype about big data today, big data has been around for at least three decades or even longer, depending on how it is defined. From the 1970s, database systems, report generators and decision support systems were the technologies used for managing and analyzing large amounts of data. In the 1990s data warehousing and data migration technologies made organization-wide decision making easier over data across various data sources. At about the same time, data mining technology emerged to allow for semi-automatic extraction of grouping and classification of data. In all these, relational database systems and file systems have been used for storing data. Recently, the Hadoop open-source platform has become popular for storing and processing big data.

DOI: 10.4018/978-1-4666-9840-6.ch004

(For expositional simplicity, henceforth we will use the term “big data” to mean not just “a huge amount of data”, but also “storing, managing, and analyzing big data”. Big data certainly requires technologies.) The current hype about big data in the trade press appears to make big data seem like it is all about technologies, is a fully automated magic, and is a requirement for the survival of every organization. In reality, big data is not all about technologies, it requires considerable expert human efforts, and it can give competitive advantages to an organization only if used properly. In fact, big data requires the following three critical elements, besides technologies.

1. Big data requires data. This may sound even dumb. However, the point is that the data must be the right kinds, must be sufficient in quantity, and must be clean. If relevant data is not available, no actionable intelligence can be discovered. If the amount of data is not sufficient, there may be no statistical significance in the results of big data. Even if there is a huge amount of data, when much of it is dirty, the data is not usable.
2. Big data involves a painstaking process. If this process is not properly followed, efforts to extract actionable intelligence from big data are not likely to succeed. The starting point of the process is to identify important objective for big data, and exploring feasibility of successfully meeting the objective. The process ends after the actionable intelligence discovered is applied to the business needs. In between, the data must be analyzed for suitability for analysis and be cleansed, transformed and encoded for analysis. The suitability analysis and preparation for analysis require substantial human efforts.
3. Big data requires people who understand how to use the technologies and how to execute each step of the big data process. Data mining technology is based on approximate computations that group data based on some measures of “mathematical similarity”, without understanding the meanings of the data. There are many mining tasks, including grouping (clustering) similar objects, classifying new objects into one of the existing groups, detecting anomalous data (outliers), etc. These tasks must be performed on various types of data, including numeric data, words, text, Web pages, multimedia data, sequence data, etc. Further, these tasks must support the special requirements and characteristics of numerous types of applications. To make the matter even worse, for any given task, there are many algorithms with different tradeoffs. There has been much progress in the usability of the data mining software that embody the algorithms. However, there is still a long way to go.

In other words, big data is difficult to do. In this paper, we provide a holistic view of big data, including technologies and non-technology elements, so that the readers may have a more complete perspective of big data, rather than get sidetracked by the current hype.

The remainder of this article is organized as follows. In the Process section, we will discuss the big data process, along with the technologies relevant to big data. In the Data Mining Technologies section, we will review data mining technologies. In the Database Platform section, we will discuss the big data platform issues. In the Conclusion section, we will outline R&D directions and conclude the article.