Big Data Techniques and Applications

B

Gamze Özel
Hacettepe University, Turkey

INTRODUCTION

Big data is a new term but not a wholly new area since many of the research-oriented agencies such as NASA, the National Institutes of Health and Energy Department laboratories have been engaged with aspects of big data for years, though they probably never called it that. It's the recent explosion of digital data in the past few years. Every day the world creates 2.5 quintillion bytes of data (IBM, 2012). As more and more organizations are stepping out of the traditional boundaries of the enterprise to understand the impact of their environment on their business, big data keeps growing bigger.

Big data is a collection of data sets so large and complex that it becomes difficult to process using on-hand database management tools or traditional data processing applications (White, 2009). It turns imperfect, complex, often unstructured data into actionable information and comes in many different formats that go beyond the traditional transactional data formats (UN Global, 2012). Big comes from just about everywhere: posts to social media sites, on-line purchase transactions, weather sensors, digital pictures and videos posted, online sensors used to gather climate information, and from cell phone GPS signals to name a few. Growing volumes of structured, semi-structured and unstructured data are streaming into organizations at a faster and faster rate.

Big data has increased the demand of information management specialists: the usual suspects, for example, Software AG, Oracle Corporation, IBM, Microsoft, SAP, HP and EMC have spent more than $15 billion acquiring software firms only specializing in data management and analytics. On its own, this piece of the industry worth more than $100 billion and growing at almost 10 percent a year which is about twice as fast as the software business as a whole. IBM® has developed a comprehensive, integrated and industrial strength big data platform that allows you to address the full spectrum of big data business challenges. The four core capabilities of the platform include Hadoop, stream computing, data warehousing, and information integration and governance.

Big data requires exceptional techniques to efficiently process large quantities of data within tolerable elapsed times. Companies are applying a range of technologies to deal with big data such as massively parallel processing (MPP) databases, distributed databases and file systems, cloud computing and scalable storage systems among others. Manyika et al. (2011) suggests suitable technologies include A/B testing, association rule learning, classification, cluster analysis, crowd sourcing, data fusion and integration, ensemble learning, genetic algorithms, machine learning, natural language processing, neural networks, pattern recognition, anomaly detection, predictive modeling, regression, sentiment analysis, signal processing, supervised and unsupervised learning, simulation, time series analysis and visualization.

In this chapter, past and current research on big data techniques and its applications are investigated. Some categories of big data techniques applicable across a range of industries is provided. Then, illustrative examples of big data are explored for understanding very large-scale data and complex analyses in order to make better decisions.

DOI: 10.4018/978-1-4666-5202-6.ch032

BACKGROUND

Big data is a popular term used to describe the exponential growth, availability and use of information, both structured and unstructured. Big data is data that exceeds the processing capacity of conventional database systems (SAS, 2012). The data is too big, moves too fast, or doesn't fit the structures of your database architectures. As seen in Figure 1, big data is currently defined using three data characteristics: volume, variety and velocity. It means that some point in time, when the volume, variety and velocity of the data are increased, the current techniques and technologies may not be able to handle storage and processing of the data. At that point the data is defined as Big Data. Therefore, big data are sometimes called the "3Vs": more volume, more variety and higher rates of velocity (Douglas, 2001). The following three dimensions of big data define the expansion of a data set along various fronts to where it merits to be called big data.

Volume: There is more data than ever before, its size continues increasing, but not the percent of data that our tools can process. Many factors contribute to the increase in data volume such as text data constantly streaming in from social media, increasing amounts of sensor data being collected, etc. Excessive data volume created a storage issue in the past. On the other hand, other issues emerge including how to determine relevance amidst the large volumes of data and how to create value from data that is relevant with today's decreasing storage costs.

Velocity: There are many different types of data, as text, sensor data, audio, video, graph, and more. Velocity means that both how fast data is being produced and how fast the data must be processed to meet demand (Douglas, 2013). It is a combined data infrastructure and data management process that addresses different concerns that are visible after the creation and addition of big data objects. It covers factors like website or application response, transaction execution time, data analysis and automatic and quick updates across all data stores.

Variety: From excel tables and databases, data structure has changed to loose its structure and to add hundreds of formats. Pure text, photo, audio, video, Web, GPS data, sensor data, relational data bases, documents, SMS, pdf, flash, etc. Structure can no longer be imposed like in the past in order to keep control over the analysis. As new applications are introduced new data formats come to life.

As seen in Figure 1, the 3Vs together describe a set of data and a set of analysis conditions that clearly define the concept of big data. Real world big data applications commonly address one or two of the "V"s. However, there are many organizations with big data projects that do indeed incorporate all three; these usually involve high volumes of streaming data from a variety of sources (Gartner, 2011). For this reason, Gartner updated its definitions in 2012 as follows: "Big data are high volume, high velocity, and/or high variety information assets that require new forms of processing to enable decision making, insight discovery and process optimization."

Laney (2001) was the first one in talking about 3V's in Big Data management but nowadays, there are two more V's:

Variability: There are changes in the structure of the data and how users want to interpret that data.

Figure 1. The 3 V's model is the most commonly used description of big data
(Source: http://www.thinkinc.com/blog/the-future-of-big-data-and-the-data-scientist-in-2013/)