Harnessing the Power of Big Data Analytics

Billie Anderson
Bryant University, USA

J. Michael Hardin
University of Alabama, USA

INTRODUCTION

Every two days we create as much data as we did up to 2003. -Eric Schmidt, CEO of Google

The age of big data is upon us. Data is being collected by businesses at a rate never encountered before, through Web sources, cellular phones and social media. The growth of Internet businesses has led to a whole new scale of data processing challenges. Companies such as Google, Facebook, Yahoo, Twitter, and Amazon now routinely collect and process hundreds to thousands of terabytes of data on a daily basis. This represents a significant change in the volume of data which can be processed, a major reduction in processing time required, and reduction of the cost required to store data.

Organizations have been collecting data for years, but the digital age has brought with it a substantial increase in the amount of data that is available to the modern-day business. For example, the genealogy site Ancestry.com stores about 2.5 petabytes of data (White, 2012). Twitter collects 7 terabytes of new data each day (Soffer & Heid, 2012). This data size growth rate can be attributed to several factors. The first is a more prominent presence in the online community. Many of the major retailers such as Apple, Wal-Mart, Target, Macy's, Best Buy, Kohl's, and Walgreens have much more of an online presence than they did 10 years ago. This online retail presence increases the amount of data each company has access to and can collect. From the financial services and healthcare sectors more data is being produced from a business protection standpoint. That is, more backup, recovery, and monitoring of customer or patient records.

In 2011, researchers from MIT Sloan Management Review and IBM asked 3,000 executives, managers and analysts how they obtain value from their massive amounts of data. The study found that organizations that used business information and analytics outperformed organizations who did not. Specifically, these researchers found that top-performing businesses were twice as likely to use analytics to guide future strategies and guide day-to-day operations as their lower-performing counterparts (LaValle, Lesser, Shockley, Hopkins, & Kruschwitz, 2011).

In order to extract value from big data, companies need to be able to easily work with terabytes and petabytes of data constantly generated by employees, customers, competitors, and Websites. It is not only the size of the data sets that distinguishes the big data movement, but also the differing types of data that must be handled. The scope of data collected by organizations is more diverse than ever. Data comes in a variety of different forms, such as structured and unstructured, spread across internal and external sources. Data is also more dynamic in the age of big data. Data constantly changes and evolves in real-time, making the window for taking action considerably shorter than in the past (McAfee & Brynjolfsson, 2012).

This chapter will define big data and big data analytics. Emerging data architectures that can handle vast amounts of data such as Hadoop will be examined. Hive, the new programming language developed by Facebook that makes Hadoop more accessible, will be explained. A survey of how

DOI: 10.4018/978-1-4666-5202-6.ch101

software and hardware companies are creating new businesses and technology from the big data architectures will be provided. The chapter will conclude with the future work of big data that is on the horizon.

BACKGROUND

Typical business practice for large-scale data analysis has traditionally focused on Enterprise Data Warehouses (EDWs). EDWs dominated academic research and industrial development throughout the 1990's. A data warehouse is a large repository of historical and current transaction data of an organization. An EDW is a centralized data warehouse that is accessible to the entire organization. EDWs are considered to be the cornerstone of good information technology (IT) (Cohen, Hellerstein, Dolan, Welton, & Dunlap, 2009). EDWs play a pivotal role in organizations that are very information-centered in industries such as retail and telecommunications. The EDW serves as the central meeting location for data integration within a large organization. The EDW has traditionally been an advantage for computing enterprise wide analytics since it has the ability to gather and organize data information from all elements of the organization. The main focus of an EDW is to compute data intensive reports for high levels of decision-making management.

With the rise of big data in the last 10 years, storage capacities of traditional EDWs have become overwhelmed, causing a scalability issue for many IT departments (Russom, 2011). Today, many organizations are looking for an alternative to the traditional EDW paradigm, for several reasons. First, the sizes of the databases themselves have grown remarkably in the past 10 years. Second, the value of advanced data analysis is at the forefront of Business Intelligence (BI). BI was previously used to perform basic data aggregation functions such as counts or summaries of the data at a high level. BI is evolving into digging deeper into the data and analyzing highly detailed data using more advanced mathematical and statistical techniques. Finally, storage has become inexpensive. Until recently, costs were too high to store data anywhere other than disk. But the price of memory has plunged in the past several years. The price of one megabyte of memory is currently less than one cent (Elliott, 2011).

Another type of IT system that is vital to a business that collects and analyzes large-scale data is On-line Analytical Processing (OLAP). OLAP is a business data warehouse that is primarily used for analytics, data mining, and decision making. An OLAP database can contain aggregated or historical data. For example, an OLAP database may contain years of data about flight reservations. An analyst can use this data to gain meaningful insight into the data such as flight trends, types of customers who are traveling in various classes such as first class, business or coach. Queries are often complex and involve aggregations. OLAP Queries have significant importance in strategic decision making. The top level of management typically uses the analysis from an OLAP query to aid in decision making.

In this changed climate of widespread large-scale data collection, companies have altered their traditional methods of organizing and analyzing data. A primary storage and analysis system for big data is Hadoop. Hadoop is a programming framework that supports the storage, processing, and analysis of big data in a distributed computing environment: a distributed computing environment involves setting up and managing computing and data exchange in a system of distributed computers. Typically, distributed computing is used in a larger network of computing systems where data is distributed across computers in an effort to save computation time for big data. Hadoop was invented by Google in order to usefully index all the rich textural and structural information they were collecting, and then present meaningful and actionable results to users (Ghemawat, Gobioff, & Leung, 2003; Dean & Ghemawat, 2008). There was no application on the market that would collect unorganized data from many different