Chapter 1
Big Data Overview

Yushi Shen
Microsoft Corporation, USA

Ling Wu
EMC² Corporation, USA

Yale Li
Microsoft Corporation, USA

Shaofeng Liu
Microsoft Corporation, USA

Qian Wen
Endronic Corp, USA

ABSTRACT

This chapter provides an overview of big data and its environment and opportunities. It starts with a definition of big data and describes the unique characteristics, structure, and value of big data, and the business drivers for big data analytics. It defines the role of the data scientist and describes the new ecosystem for big data processing and analysis.

INTRODUCTION

Today we have heard a lot about Big Data. What is Big Data? (Press, 2013) Is there a definite size over which data becomes Big Data? Is it the number of rows or the number of columns? Is a spreadsheet that contains a million rows Big Data? Is a database that has a billion records Big Data? How big is Big Data?

Wikipedia defines big data as "a collection of data sets so large and complex that it becomes difficult to process using the available database management tools. The challenges include how to capture, curate, store, search, share, analyze and visualize big data. The trend to larger data sets is due to the additional information derivable from the analysis of a single large set of related data, as compared to separate smaller sets with the same total amount of data, allowing such correlations to be found to spot business trends, determine the quality of research, prevent diseases, link legal citations, combat crimes and determine real-time roadway traffic conditions. " (Wikipedia, 2012)

In today's environment, we have access to more types of data. These data sources include online transactions, social networking activities, mobile device services, internet gaming and etc. While the public open data is growing, increasingly powerful ERPs also bring corporate data to a new level. Big data is changing the world. Data sources are expanding, data from Facebook, twitter, YouTube, Google and etc., are to grow 50X in the next 10 years.

DOI: 10.4018/978-1-4666-9840-6.ch001

An IDC study shows that in 2010, there have been 1.2 zettabytes (1,200,000,000,000,000,000,000) of information, a trillion billion bytes of information to be managed and analyzed. It is estimated that by 2020, there is going to be 35 zeta bytes of information. Data deluge is to grow 44X in this decade. About 90% of this information being created is unstructured, like website clicks, mobile phone calls, Facebook posts, call center conversations, tweets, videos and emails. (Gens, 2013) Where are all these big data going? It is going to be run in the cloud. When we talk about cloud computing, we cannot miss big data.

BIG DATA DEFINITION

Big data is defined in Wikipedia that as the "data sets that are too large for storage, management, processing and analysis, it present challenges beyond traditional IT techniques." (Wikipedia, 2012) BIG is a term that is relative to the size of the data, and the scope of the IT infrastructure that is in place. Transforming big data could benefit scientific discovery, environmental and biomedical research, and national security. In order to do that, big data requires the use of new technical architectures and analytics tools, to generate business value from the huge volume of data, in order to create insights.

Big Data comes in all kinds of forms: from highly structured ERP (Enterprise Resource Planning) data, or CRM (Customer Relation Management) data, to multi-million rows of text file, to video files and machine generated sensor data. The common feature is the high data volume and data complexity. Most of big data is unstructured or semi-structured, and require new techniques and tools to analyze.

Big Data examples are everywhere in our lives. With the popularity of mobile computing and the self-expression tolls, everyone has the ability to share their thoughts and ideas worldwide. Smart phones carry sensors like GPS, accelerometer, microphone, camera and Bluetooth which can collect huge amounts of data, and allow research on behavioral and social science, with the large scale mobile data to characterize and understand real-life phenomena. In 2011, there have been 6 billion mobile phone subscribers, growing 45 percent annually for the past four years. (ITU, 2011) A quarter of them use smartphones. By 2014, mobile internet use should overtake desktop internet use.

There are more than 845 million active Facebook users by the end of 2011, 50 percent log onto Facebook every day; 30 billion pieces of content are shared every month. Every 60 seconds, there are 510,000 posted comments, 293,000 status updates and 135,000 uploaded photos. 20 million Facebook applications are installed per day. In just 20 minutes, over 1 million links are shared. (Protalinski, 2012)

Twitter has 100 million active users around the world; more than half of them log in to twitter each day to follow their interest. The average user has 115 followers. An average of 190,000,000 tweets are sent per day. Tweeter handles 1.6 billion queries per day. 34% of marketers have generated leads using Twitter and 20% have closed deals. (Twitter, 2011)

YouTube has over 800 million unique visitors per month, which generates 92 billion page views. 72 hours of videos are uploaded every minute and over 4 billion hours of video are watched each month. More videos are uploaded to YouTube in 60 days, than the three major US TV network programs created in 60 years. (YouTube.com Statistics)

Google has 4.7 billion searches per day in 2011. Google Plus reaches out to 10 million users in 16 days, it is been reported that it has 400 million users in just one month. (Statistic Brain, 2012)

LinkedIn has 135,000,000 users; Wikipedia hosts over 17 million articles; Foursquare sees 2,000,000 check-ins a week; Instagram reaches out to 13 million users in 13 months after its launch, and have 150,000,000 photos uploaded; Flickr hosts over 5 billion images. (Decision Stats, 2011) (Pring, 2012)