# Big Data Analytics:
## Trends and Case Studies

B

**Hoda Ahmed Abdelhafez**
*Suez Canal University, Egypt*

## INTRODUCTION

Big Data is a platform for transforming all of this data into actionable items for business decision making (Smith, 2012). Big data refers to the explosion in the quantity and quality of available data, largely the result of recent and unprecedented advancements in data recording and storage technology (McBurney, 2011).

Big Data is the data sets whose volume, variety, velocity and complexity make it impossible for current relational databases and architectures to store and manage it. Big Data is not about the content that is created, nor is it about consumption but it is more about the analysis of the data and how that needs to be done. Also the value the organizations can derive from this big data and the resulting require these organizations to rethink their information strategies to extract the value (Carter, 2011)

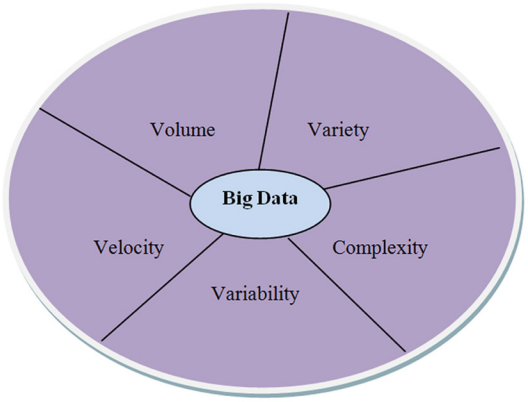
Big data analytics is the application of advanced analytic techniques to large big data sets. It is a collection of different tools and techniques such as NoSQL database and data visualization (Russom, 2012). New technologies for big data have emerged and are being used in many big data analytics environments. These technologies like NoSQL databases, Hadoop and MapReduce form the core of an open source software framework that supports the processing of large data sets across clustered systems (Rouse, 2012).

## VOLUME, VARIETY, AND VELOCITY OF BIG DATA

The three attributes or components of big data are volume, variety and velocity (Russom, 2011), in addition to complexity and variability (SAS, 2012) as shown in Figure 1.

*Data volume* is the primary attribute of big data. It can be quantified by terabytes or petabytes and it can also be quantified by counting records, tables, transactions and files (Kaisler et al., 2013; Russom, 2011). For instance, the total book stack in Library of Congress measures 15Terabytes and Google processes more than 1Petabytes every hour; also Bank of America Merrill Lynch manages petabytes of data for advanced analytics and new regulatory requirements (Forsyth Communications, 2012). Organizations that are facing huge amount of data cannot manage and analysis this data with traditional IT structures but they need scalable storage and distributed approach to process this data (Dumbill, 2012).

*Figure 1. The components of big data*




DOI: 10.4018/978-1-4666-5202-6.ch029

*Data Variety* means the diverse source of data because of explosion of sensors, smart devices and social collaboration technologies (Kaisler et al., 2013; Zikopoulos, et al., 2012). Variety including structured, semi structured (XML, RSS feeds) and unstructured (text and human language) as well as data, data comes from audio, video, and other devices. In addition to multidimensional data which can be drawn from a data warehouse (Russom, 2011). About 80 percent of a company's data is unstructured including office productivity documents, e-mails, Web content, in addition to social media (Forsyth Communications, 2012). Text, video and other forms of media will need a completely different architecture and technologies to perform the required analysis. For example, many marketing departments want to find ways to do sentiment and brand analysis based on what is being posted on Facebook, Twitter and YouTube. This dynamic becomes challenge in Asia with local social media sites such as Nate in Korea and RenRen in China (Carter, 2011).

*Data velocity* means the speed of the data generation and the speed of the data delivery. For example, the data stream coming from social media or any kind of device or sensor such as robotic manufacturing machines (Kaisler et al., 2013, Russom, 2011; Zikopoulos, et al., 2012). The Internet and mobile era for instance allow online retailers to compile large histories of customers' every click and interaction not just the final sales. In addition to the Smartphone increases the rate of data inflow, as a streaming source of audio data and geo-located imagery (Dumbill, 2012a). Organizations need to use new technologies to analyze big data streams near real time in order to gain more insights.

*Complexity* involves things for moving operational data into big data platforms and the Difficulties in managing the data across multiple sites and geographies. Complexity focuses on (1) the increased data with the expanding universe data sources, (2) the need to link, match, and transform data across business entities & systems and (3) understanding the relationships among all data such as complex hierarchies and data linkages (SAS, 2012).

*Variability* represents contexts associated with a given piece of data. The dataflow can be highly variable – with daily, seasonal and event-triggered peak loads that can be challenging to manage (SAS, 2012).

## DOES RELATIONAL DATABASE SOLVE BIG DATA PROBLEM?

Relational database and schema manage the storage and retrieval of the structured data. Relational database do not able to handle streaming data because they must slowly import data into a native representation before they can be queried. The streaming technologies don't integrate well into the relational database engines themselves and these engines don't parallelize effectively to massive amounts of data (Madden, 2012).

NoSQL databases are sometimes called key-value stores, separate data management and data storage, whereas relational databases attempt to satisfy both concerns with databases (Kavis, 2010; Cattell, 2011). NoSQL databases could solve big data problem instead of traditional database especially when the organizations need to analysis massive amount of unstructured and semi-structured data and handling queries in a simple manner. If most of the data are in legacy systems, RDBMSs are needed in this case. With this, both tools RDBMS and NoSQL can be used to integrate relational model with different approach to storage in order to help the organizations gain competitive advantage. (see Table 1)

## BIG DATA ANALYTICS

Big data analytics is the process of examining large volume of data of a variety of types to discover useful information. Such information can provide competitive advantages over rival organizations and result in business benefits. The main goal of