Chapter 1
Big Data Overview

ABSTRACT

Big data is now a reality. Data is created constantly. Data from mobile phones, social media, GIS, imaging technologies for medical diagnosis, etc., all these must be stored for some purpose. Also, this data needs to be stored and processed in real time. The challenging task here is to store this vast amount of data and to manage it with meaningful patterns and traditional data structures for processing. Data sources are expanding to grow into 50X in the next 10 years. An International Data Corporation (IDC) forecast sees that big data technology and services market at a compound annual growth rate (CAGR) of 23.1% over 2014-19 period with annual spending may reach $48.6 billion in 2019. The digital universe is expected to double the data size in next two years and by 2020 we may reach 44 zettabytes (10^{21}) or 44 trillion gigabytes. The zettabyte is a multiple of the unit byte for digital information. There is a need to design new data architecture with new analytical sandboxes and methods with an integration of multiple skills for a data scientist to operate on such large data.

INTRODUCTION

Andrew Brust (2012) stated, "We can safely say that Big Data is about the technologies and practice of handling data sets so large that conventional database management systems cannot handle them efficiently, and sometimes cannot handle them at all". The attributes to be dealt for such big data stands out to be:

DOI: 10.4018/978-1-5225-3790-8.ch001

- Huge volume
- Complexity of types and structures
- Speed of new data growth and its processing

How a big data is differentiated from other data? It is not only voluminous. There are 3 'V's to characterize this data (Viceconti, Hunter, & Hose, 2015). They are:

- **Volume:** It refers to the size of big data, which is definitely huge. Most organizations are struggling to manage the size of their databases and it has become overwhelming. From 2010 to 2020, data increases from 1.2 ZetaBytes (ZB) to 35.2 ZB.
- **Velocity:** It includes the speed of data input and output i.e. given based on the aspects of throughput of data and latency. Here the machine generated data explodes even in milliseconds. For example, Communication Service Provider CSP that generates GPS data, data streaming from websites etc. The challenging fact is to embed analytics for data-in-motion with reduced latency (www.turn.com conducts analytics for online advertisement in 10 milliseconds).
- **Variety:** It refers to various types of data which cannot be easily managed by traditional database. The data in warehouse was compiled from variety of sources and transformed using ELT (Extract, Load and Transform) but this is restricted for structured content. Here, this includes data expanded across horizons which comprises of textual, geo-spatial, mobile, video, weblogs, social media data, and transaction data etc...

Nowadays two more V's are added for big data: 1) veracity; 2) value.

- **Veracity**: It refers to the trustworthiness of data. Some data may have ambiguity. For example, all the posts in Twitter cannot be trusted. The volume may be responsible for lack of accuracy.
- **Value:** It refers to the value/quality usage of data. It is nothing but how far the data is useful to that particular organization.