Chapter 1
Big Data:
The Data Deluge

Jayshree Ghorpade-Aher
University of Pune, Pune

Reena Pagare
University of Pune, Pune

Anita Thengade
University of Pune, Pune

Santaji Ghorpade
IBM India Pvt. Ltd., India

Manik Kadam
Allana Institute, India

ABSTRACT

Today is the Computer Era, where the data is increasing exponentially. Managing such a huge data is a challenging job. Under the explosive increase of global data, the term of big data is mainly used to describe enormous datasets. The state-of-the-art of big data is discussed here. The discussions aim to provide a comprehensive overview and big-picture to readers of this existing research area. This chapter discusses the different models and technologies for Big Data; It also introduces Big data Storage. Big data has been a potential topic in various research fields and areas like healthcare, public sector, retail, manufacturing personal data, etc.

I. INTRODUCTION

Since last two decades, there is a huge amount of data collection and we need something to handle that, so as the development process we are moving from file to database, data-ware house, datacenters, and now the big term Big Data. Big data is data that exceeds the processing capacity of conventional database systems. The data is too big, moves too fast, or doesn’t fit the structures of your database architectures. One must choose an appropriate way to gain value from this huge data. According to a report from International Data Corporation (IDC), in 2011; the overall created and copied data volume in the world was 1.8ZB (≈10^21B), which increased by nearly nine times within 5 years (Min, Shiwen, Yunhao, 2014). The various source for collection of such data are generated from online transactions, emails, videos, audios, images, click streams, logs, posts, search queries, health records, social networking interactions, science data, sensors and mobile phones and their applications. According to the recent survey, a personal computer holds about 500 terabytes. Various other sources like Facebook (log data of over 10 PB per
Big Data

month), Google (processes 100 PB of data) electricity board, telecom industries, organizations, twitter (12 TB), railway, airlines (black box), stock exchange market etc. also add to the huge amount of data (≈5 EB of data was created by humans until 2003). Today this amount of information is created in just a day. Figure 1 specifies the various components of Big data. Today working with conventional databases and other software tools have not proven efficient while handling the large datasets.

**Definition:** Big data is a term for massive data sets having large, more varied and complex structure with the difficulties of storing, analyzing and visualizing for further processes or results.

In 2010, Apache Hadoop defined big data as datasets which could not be captured, managed, and processed by general computers within an acceptable scope.

May 2011, McKinsey & Company, a global consulting agency announced Big Data as the next frontier for innovation, competition, and productivity.

Doug Laney, an analyst of META (presently Gartner) defined challenges and opportunities brought about by increased data with a 3Vs model as shown in figure 1 [Seref & Duygu, 2013], that is the increase of Volume (size of data increasing day by day), Velocity (processing of data), and Variety (Structured data inserts a data warehouse already tagged and easily sorted but, Unstructured data is random and difficult to analyze, Semi-structured data does not conform to fixed fields but contains tags to separate data elements).

*Figure 1. The 3 V’s of Big Data*
Related Content

Gridifying Neuroscientific Pipelines: A SOA Recipe and Experience from the neuGRID Project
www.igi-global.com/chapter/gridifying-neuroscientific-pipelines/73547?camid=4v1a

A Graph-Based Biomedical Literature Clustering Approach Utilizing Term's Global and Local Importance Information
www.igi-global.com/article/graph-based-biomedical-literature-clustering/1819?camid=4v1a

Gene Expression Mining Guided by Background Knowledge
Jirí Kléma, Filip Železný, Igor Trajkovski, Filip Karel and Bruno Crémilleux (2009). *Data Mining and Medical Knowledge Management: Cases and Applications* (pp. 268-292).
www.igi-global.com/chapter/gene-expression-mining-guided-background/7537?camid=4v1a

Ethical and Legal Data Mining: Paradigms of Organizational Development Practitioners and Organizational Psychologists
Ben Tran (2013). *Ethical Data Mining Applications for Socio-Economic Development* (pp. 201-229).
www.igi-global.com/chapter/ethical-legal-data-mining/76263?camid=4v1a