Chapter 10
Driving Big Data with Hadoop Technologies

Siddesh G. M.
M. S. Ramaiah Institute of Technology, India

Srinidhi Hiriyannaiah
M. S. Ramaiah Institute of Technology, India

K. G. Srinivasa
M. S. Ramaiah Institute of Technology, India

ABSTRACT

The world of Internet has driven the computing world from a few gigabytes of information to terabytes, petabytes of information turning into a huge volume of information. These volumes of information come from a variety of sources that span over from structured to unstructured data formats. The information needs to update in a quick span of time and be available on demand with the cheaper infrastructures. The information or the data that spans over three Vs, namely Volume, Variety, and Velocity, is called Big Data. The challenge is to store and process this Big Data, running analytics on the stored Big Data, making critical decisions on the results of processing, and obtaining the best outcomes. In this chapter, the authors discuss the capabilities of Big Data, its uses, and processing of Big Data using Hadoop technologies and tools by Apache foundation.

1. INTRODUCTION

The data driven computing world changes the way we perceive and interact with the world based on the interpretation of the large volumes of data produced by the computing world. Conventional relational database systems are used as the primary means of storage of data. These data have to be processed and analyzed effectively that is critical for ensuring decisions made on the introduction of new products, generating the quarterly reports, maintaining the relationships with the customers, manage their finances and thus understand about the world. Since, the internet has reached all round the globe, data is being generated from various sources such as blogs, social networking sites,

DOI: 10.4018/978-1-4666-5864-6.ch010

videos, transactions of various businesses, sensors of traffic flow, GPS information from satellites and so on the list continues with different characteristics termed as BIG DATA. Hadoop is one of the platforms that help in storing and accessing the Big Data across clusters of systems. In this article we discuss how to use Hadoop technologies to process the Big Data which is discussed in brief as follows; Hadoop - It is an Apache open source project that enables distributed processing of large sets of data that spans over different clusters. It consists of two things namely Hadoop Distributed File System (HDFS) and Map Reduce. HDFS - It is a file system supported by Hadoop and fashioned around Master-Slave architecture. It mainly consists of Namenode that acts as the master and Datanodes that acts as the slaves. Any data such as CSV files, sequence files, images, videos etc., can be loaded into HDFS just like other file systems. Map Reduce - It is a programming model that enables processing of data loaded into HDFS across different Clusters. It mainly consists of Map phase and Reduce phase. The Map phase takes care of getting the data from HDFS and Reduce phase takes care of presenting the result to the Client. HBase - It is a distributed, column-oriented database that is on top of HDFS. The data model of HBase allows scalability of data beyond the traditional relational database systems by grouping the columns of data into Columnfamilies. Hive - It is a data warehouse that allows querying on large datasets stored in HDFS using SQL like language interface called HiveQL. Hive is used for ad-hoc queries, data-summarization analysis of large data sets stored in HDFS. Sqoop - It is a command line interface tool that allows transfer of data between the structural databases and Hadoop platforms which might be either of HDFS or Hive or HBase. It also allows exporting data back to the relational databases. Pig- It is a platform that allows analyzing large data sets present in HDFS with its language called Pig Latin. It is built on top of Hadoop that provides an abstraction layer to Map Reduce programming model. It supports the data operations that help in aggregating and separating the data.

2. WHAT IS BIG DATA?

The world of computing is driven by data and can change the way we perceive and interact with the world. The data generated by the computing devices are generally stored in the conventional databases. The data have to be processed and analyzed effectively that is critical for ensuring decisions made on the introduction of new products, generating the quarterly reports, maintaining the relationships with the customers, manage their finances and thus understand about the world (LaValle, Hopkins, Lesser, Shockley, & Kruschwitz, 2010). In the telecom industry, call data records need to be analyzed for ensuring quality of service with the customers (Schroeck, Shockley, Smart, Romero-Morales, & Tufano, 2012 ; Banerjee, 2011). Another example is the online retail industry that keeps track of each click of browsing by the customers for ensuring smarter shipping and inventory decisions (Schroeck et. al., 2012). The Banking sector needs to keep track of both customer and financial details to ensure how money is managed and transferred (Hickins, 2013). Since, the internet has reached all round the globe, data is being generated from various sources such as blogs, social networking sites, videos, transactions of various businesses, sensors of traffic flow, GPS information from satellites and so on the list continues with different characteristics termed as *big data* (Schroeck et. al., 2012). Big data spans over three basic characteristics namely volume, variety and velocity, commonly called as 3 V's that provide a better view of different aspects of Big Data and the platforms available to exploit them.

- **Volume**: It refers to the large collection of data being generated. More data moves across the internet generating terabytes to petabytes of data. For example an air-