Chapter 4
Hadoop Framework for Handling Big Data Needs

Rupali Ahuja
University of Delhi, India

ABSTRACT

The data generated today has outgrown the storage as well as computing capabilities of traditional software frameworks. Large volumes of data if aggregated and analyzed properly may provide useful insights to predict human behavior, to increase revenues, get or retain customers, improve operations, combat crime, cure diseases, etc. In conclusion, the results of effective Big Data analysis can be used to provide actionable intelligence for humans, as well as for machine consumption. New tools, techniques, technologies and methods are being developed to store, retrieve, manage, aggregate, correlate and analyze Big Data. Hadoop is a popular software framework for handling Big Data needs. Hadoop provides a distributed framework for processing and storage of large datasets. This chapter discusses in detail the Hadoop framework, its features, applications and popular distributions, and its Storage and Visualization tools.

INTRODUCTION

The exponential rise in usage of internet and mobile devices along with adoption of technologies like Cloud Computing, Mobile Computing, Internet of Things, sensor based networks have led to the explosion of data generated each second. It's not only the size but the pace at which data is being generated and the wide variety of data formats that has become a point of concern in academia, research and industry. This new generation of data produced today is being termed "Big Data" because of its voluminous nature, high speed and wide variety of structured and unstructured data formats that it supports. Big Data has outgrown the Storage as well as computing capabilities of conventional software frameworks. Many new tools, techniques, technologies and methods are being developed to store, retrieve, manage, aggregate, correlate and analyze Big Data as each industry now wants to gain insights from this huge pile of data.

DOI: 10.4018/978-1-5225-3142-5.ch004

Large volumes of complex data can hide important insights. This data, when captured, formatted, stored, aggregated and analyzed can help an organization to gain useful insight to increase revenues, increase number of customers, reduce costs, increase productivity, increase efficiency and improve operations. A lot of business organizations are using Big Data Analytics to measure customer experience. Analysis of data sets can find new correlations, to spot business trends, prevent diseases, reduce crime rate, predict human behavioral patterns, weather forecasts, reveal astronomical information, etc. (The Economist, 2010)

Hadoop is the most popular and powerful open source distributed framework for handling Big Data needs. Hadoop is capable of storing large amounts of data across clusters of possible cluster of nodes comprising of commodity hardware. It can efficiently perform computationally intensive analysis on the data stored. It is a highly reliable, scalable and fault tolerant software model. It replicates every data block so that it is always available. It performs job monitoring so that every job is completed on time and if any node fails, it automatically shifts work to a different machine on the cluster.

Hadoop framework is based on master slave architecture. There is one master and several worker nodes in a cluster. The main components of Hadoop are MapReduce and Hadoop Distributed File System (HDFS). MapReduce facilitates processing of large data sets and HDFS is the distributed file system for storing large amounts of data. Both these components are based on master slave architecture. MapReduce's master component is called *JobTracker* which divides a job among various tasks and distributes them among several worker nodes called *TaskTrackers*. In HDFS, the master component is called *NameNode* that hosts and manages the file system. The data across files is distributed on *DataNodes*. *JobTrackers* are aware of locations where data resides and assigns map or reduce tasks to *TaskTrackers* which are nearest to the location of data they require.

The chapter discusses Hadoop framework in detail. It describes the features and applications of Hadoop along with a list of its popular distributions. It briefly describes the Hadoop Ecosystem and mainly focuses on the Storage and Visualization components of this system i.e. HDFS, HBase, Hive, Sqoop, Ambari, QlikView, Jaspersoft and Tableau, etc.

BACKGROUND

The foundation of Hadoop was laid by Doug Cutting and Mike Cafarella at Yahoo Incorporated (Wikipedia, 2015b) in 2005. Hadoop was originally developed to support web indexing in Nutch search engine project. The main components of Hadoop i.e. MapReduce and HDFS are inspired from Google papers. MapReduce is a user-defined function developed by Google in early 2000 for indexing the Web. HDFS has been derived from Google File System (GFS). The database component of Hadoop, HBase is inspired by the Google Bigtable. Currently, Hadoop is an open source project of the Apache Software Foundation (Apache Hadoop, 2015) and is being continuously improved and enhanced by thousands of contributors worldwide. Top IT giants like Yahoo, Facebook, Google, Microsoft, eBay, EMC, etc. are using Hadoop to handle Big Data needs. Hadoop is a Java based framework and requires Java Runtime Environment for its execution. Yahoo has more than 100,000 CPUs in over 40,000 servers running Hadoop, with its biggest Hadoop cluster running 4,500 nodes (Assay, 2014). Figure 1 depicts the Hadoop timeline.

In its report on Hadoop predictions for 2015 (Gualtieri, Kisker, Leaver, Hopkins & Murphy, 2014), tech analyst firm Forrester calls Hadoop a "rising star" in data analytics and claims "Hadooponomics,"