Chapter 4
Hadoop Setup

ABSTRACT

Apache Hadoop is an open source framework for storage and processing massive amounts of data. The skeleton of Hadoop can be viewed as distributed computing across a cluster of computers. This chapter deals with the single node, multinode setup of Hadoop environment along with the Hadoop user commands and administration commands. Hadoop processes the data on a cluster of machines with commodity hardware. It has two components, Hadoop Distributed File System for storage and Map Reduce/YARN for processing. Single node processing can be done through standalone or pseudo-distributed mode whereas multinode is through cluster mode. The execution procedure for each environment is briefly stated. Then the chapter explores the Hadoop user commands for operations like copying to and from files in distributed file systems, running jar, creating archive, setting version, classpath, etc. Further, Hadoop administration manages the configuration including functions like cluster balance, running the dfs, MapReduce admin, namenode, secondary namenode, etc.

INTRODUCTION

Apache Hadoop being an open source framework provides the utility for storage and large scale processing of data on cluster of machines with commodity hardware. This also uses a simple programming model. The framework includes a distributed storage with cluster of computers for computation. Some modules of this hadoop include an environment as hadoop common with

DOI: 10.4018/978-1-5225-3790-8.ch004

mapreduce and hadoop distributed file system (Jain, 2017). Hadoop common has the java libraries with filesystem and OS supportive level abstractions that contain necessary files to build hadoop. Hadoop Distributed File System has a high throughput for storing the application data. Hadoop mapreduce further named as YARN provides a structure for job scheduling and cluster resource management to work in parallel. This as an environment includes various architecture components and models to work with. The node setup for the hadoop architecture is as follows in the chapter

BACKGROUND

As already discussed in previous chapter, primary components at the core of Apache Hadoop high level architecture includes HDFS and Map reduce layer (Lublinsky et al., 2015). The HDFS is a portable file system written in Java, uses TCP/IP layer for communication. As stated, cluster of datanodes handle the block of data over the network. It is highly reliable since the data stored is replicated thrice with two on same rack and one in other. YARN being a framework builds with jobtracker and task tracker that handle the processing. Job tracker schedules the nodes to task tracker for performing the map or reduce task (Moorthy, 2014). Further the working environment of hadoop can be stated as follows: A user can submit a job by specifying the location of input and output file in the distributed system. Java classes in jar file implements the mapreduce processing. For the above given process, job configuration can be in a single node or cluster node by modifying the different parameters to be specified in the chapter below.

SINGLE NODE SETUP

The steps for setting up a single node hadoop should be backed up by HDFS and YARN running on a Linux environment (White, 2015).

The basic requirements behind the installation of hadoop include Java. Check in command prompt to verify if Java is already installed using:

```
$ java - version
```