Chapter 9
Hadoop Tools

ABSTRACT

As the name indicates, this chapter explains the various additional tools provided by Hadoop. The additional tools provided by Hadoop distribution are Hadoop Streaming, Hadoop Archives, DistCp, Rumen, GridMix, and Scheduler Load Simulator. Hadoop Streaming is a utility that allows the user to have any executable or script for both mapper and reducer. Hadoop Archives is used for archiving old files and directories. DistCp is used for copying files within the cluster and also across different clusters. Rumen is the tool for extracting meaningful data from JobHistory files and analyzes it. It is used for statistical analysis. GridMix is benchmark for Hadoop. It takes a trace of job and creates a synthetic job with the same pattern as that of trace. The trace can be generated by Rumen tool. Scheduler Load Simulator is a tool for simulating different loads and scheduling methods like FIFO, Fair Scheduler, etc. This chapter explains all the tools and gives the syntax of various commands for each tool. After reading this chapter, the reader will be able to use all these tools effectively.

INTRODUCTION

The core part of the Hadoop is MapReduce which supports distributed processing of huge amount of data in a cluster of commodity machines. Another major part is Hadoop Distributed File System (HDFS) which maintains the needed file system. Although the HDFS and MapReduce support the major operations, additional tools are provided to support the users. This chapter describes all the additional tools available and how each one can be used.

DOI: 10.4018/978-1-5225-3790-8.ch009

BACKGROUND

Even though Hadoop framework supports all necessary functions for distributed processing, some more tools are needed. The archiving of data, running executable files and more data analysis are needed for efficient processing. Also benchmarking has to be done. These additional functionalities are provided with the help of some tools. Those tools are described in this chapter.

TOOLS

The additional tools provided by the Hadoop distribution are:

- Hadoop Streaming
- Hadoop Archives
- DistCp
- Rumen
- Gridmix
- Scheduler Load Simulator
- Benchmarking

Let us see all these tools one by one in detail.

HADOOP STREAMING

Hadoop streaming is a utility that is used to create and run Map/Reduce jobs with any executable or script as the mapper and/or the reducer.

Working of Hadoop Streaming

Both mapper and reducer can be executables. These executables read the input line by line from stdin and gives the output to stdout. When the mapper is initialized, each mapper task will launch the executable as a separate process. The mapper tasks covert its input into lines and feed them to stdin. The mapper collects lines from stdout of the process and coverts each line to key, value pair. This key, value pair is output of the mapper.