Chapter 7
Hadoop MapReduce Programming

ABSTRACT

The second major component of Hadoop is MapReduce. It is the software framework for Hadoop environment. It consists of a single resource manager, one node manager per node, and one application manager per application. These managers are responsible for allocating necessary resources and executing the jobs submitted by clients. The entire process of executing a job is narrated in this chapter. The architecture of MapReduce framework is explained. The execution is implemented through two major operations: map and reduce. The map and reduce operations are demonstrated with an example. The syntax of different user interfaces available is shown. The coding to be done for MapReduce programming is shown using Java. The entire cycle of job execution is shown. After reading this chapter, the reader will be able to write MapReduce programs and execute them. At the end of the chapter, some research issues in the MapReduce programming is outlined.

INTRODUCTION

The major components of Hadoop systems are Hadoop Distributed File System (HDFS) and MapReduce. The MapReduce is a software framework for easily writing applications. It is mainly used for processing larger amounts of data in-parallel on large clusters (thousands of nodes) in a reliable and fault-tolerant manner. This chapter gives an overview of MapReduce programming.

DOI: 10.4018/978-1-5225-3790-8.ch007

Also it explains clearly the different APIs available for programming. The programming can be done in different languages like C, C++, C#, Java, Perl, PHP, Python and Ruby. But this chapter focuses on programming with Java only.

BACKGROUND

The growth of data in the recent applications in the Internet is highly alarming. Analyzing such kind of large data (data analytics) is the demand in the business process requirements. Though many algorithms and techniques have been developed to mine such kind of large data and have been invested in the analytics, the turnaround time is not satisfactory. The huge storage requirements and computing requirements have dictated the distributed computing environment. The Hadoop based Distributed File System has enabled this. The principle involves dividing the jobs into small independent pieces (in many cases split manually) and mapping to various computing system and combining the solution back in a synchronized manner.

WORKING OF MAPREDUCE

The MapReduce framework consists of a single master ResourceManager, one slave NodeManager per cluster-node, and AppMaster per application. The user/application can submit the work to be executed as a job. The input and output of the job are stored in file system. The framework takes care of splitting the job into number of smaller tasks, scheduling the tasks across different nodes and monitoring them. If the task fails, the framework re-executes the job automatically without user intervention. The tasks are normally scheduled in the nodes where data is already present and hence the network bandwidth is properly utilized.

The applications can specify the input/output locations and other job parameters in "job configuration". Then the client submits the jar/executable file of the job along with its configuration to the ResourceManager. The ResourceManager then:

- Distributes software/configuration to the slaves
- Schedules the tasks