Cluster Based Data Storage Management System Using Hadoop Distributed File System

Myat Kyaw
University of Computer Studies, Yangon
mk.closeyoureyes@gmail.com

Abstract

Data storage is one of the important resources in cloud computing. There is a need to manage the data storage. This system is a cluster based data storage file system using Hadoop Distributed File System with the low cost devices. Hadoop Distributed File System (HDFS) is the primary distributed storage used by Hadoop applications. A hadoop cluster primarily consists of a NameNode that manages the file system metadata and DataNodes that store the actual data [1]. HDFS has master-slave architecture. HDFS is well suited for distributed storage and distributed processing using commodity hardware. It is fault tolerant, scalable, and extremely simple to expand MapReduce programming model, well known for its simplicity and applicability for large set of distributed applications, is an integral part of Hadoop. Furthermore, the map-reduce processing this system will implement the data deduplication service to be more efficient on data storage. In this system, map, reduce and deduplication algorithm are implemented to be efficient on the storage size with low cost devices.