Big Data Project Report

Project Title : Yet Another Hadoop

Team Details
A Narendiran PES1UG19CS001
Abdul Rahman Shigihalli PES1UG19CS009
Abhishek D PES1UG19CS020
Advit Gandhi PES1UG19CS035

3.1 Project Title Chosen

The project title we chose is Yet Another Hadoop. Yet Another Hadoop is a mini set-up of Hadoop on our system which is a replication of HDFS with the implementation of functionalities of Datanodes, Namenodes and replication of data across nodes. YAH breaks down complex tasks into smaller distributed tasks.

3.2 Design Details

We have datanodes created according to the config file and the details about how the data is stored in the datanode is specified in the namenode. We have implemented basic CLI commands like ls, put and cat.

3.3 Surface level implementation details about each unit

We are creating multiple file splits using the module fsplit.filesplit(), this divides the file and stores it in the datanode.
Namenode - is basically a json file which stores the directory, location of the file and which file chunk belongs to which datanodes.
We have implemented CLI using argparse and created a pip package to run our hadoop. MapReduce runs sequentially in our project.
Multiple virtual directories can be created, giving the user the illusion that a virtual directory exists and that the file chunks are stored in datanodes.

3.4 Reason behind design decisions

Simplicity is everything so, we tried to make the architecture as easy as possible which makes it easy for the third person to understand how hadoop works. We made namenodes as a json object, also made

use of command line arguments which makes it easier for the user to interact with our hadoop system.

3.5 Takeaway from the project

During this course, we used Hadoop for all our assignments and we had a little idea of the architecture behind it and how it works. Working on this project, we learned most of the architectural design of Hadoop by creating a mini Hdfs system, running map-reduce programs with it and working on Hadoop in depth. We understood the concept of virtual directory, fault tolerance, namenode, datanodes and how they work, how chunks of files are stored in datanodes and how they are retrieved from the same.