Project title : YAH – Yet Another Hadoop

Team Member Details:
KS Abhisheka - PES1UG19CS202
Lithesh Shetty - PES1UG19CS245
Vinay P - PES1UG19CS567

Design details:

We started the project by creating a namenode heartbeat,datanode heartbeat functions and established communication(heartbeat transfer) between them using UDP client-server model.Then we created the datanodes and namenodes based on the configuration given in the configuration file and dynamically made the heartbeat function to run as soon as the datanode is setup. Meanwhile we implemented the logging of information to the metadata file, log file for the specified location for each operation that is done to the DFS.We also made a secondary namenode and implemented such a way that if there is any fault in the primary namenode,the secondary namenode will receive the information and takes care of further actions to be taken.Secondary namenode also makes a backup of the primary namenodes content,checkpoint for every sync period given.Then we implemented the splitting of the datafile based on the size of each block and put it into a temporary folder till other operations are carried out like renaming it to required type, hashing.We then created the Command Line Interface for commands specified in the specification with corresponding operation for them.Then we implemented the map-reduce job and did the final touch-up of the complete project of what was left out.

Surface level implementation:

- Heartbeat-UDP client/server
- Splitting of file-split command
- We used shutil library for removing the directory before creation and also for copying the mapper and reducer for datanodes for map-reduce operation .

- Used socket, time, json libraries for the udp client/server communication, keeping track of time and for parsing the json file respectively.
- We have divided the metafile into 2 parts one is metadata of input files which keeps track of which file got uploaded to the hdfs and tracks its split location which will be helpful while reading the file in constant time.Second part of the metadata file is to keep track of datanodes basically it keeps track of which datanode has how many free blocks and how many occupied blocks

Reason behind design decision:

We implemented the heartbeat functions as UDP client/servers as the heartbeats should be continuously sent between Primary namenode, Secondary namenode and datanodes and also as udp doesn’t require to establish a connection. It looked logical to implement this way. And also for the fault tolerance of primary namenode, instead of making the secondary namenode converting it to primary namenode and another secondary namenode to be created, we created another primary namenode so as to decrease any extra time and resources for creation and switch over. Rest everything is same as the design implementation of hadoop

Takeaway from the project:

We got the hands-on experience of creating a Distributed File System. Almost all the topics taught in class helped us throughout the project and this project was very helpful in making us clearly understand depth of the topics i.e each and every details of the hadoop DFS. Starting with zero knowledge of where and how to start with the project, we spent ample amount of time discussing with the team various times and came up with some logical solutions for implementation of various features. Overall we learnt in depth the architecture of Hadoop, its working and running of map-reduce job. Implementing Hadoop from scratch improved our understanding of the subject by a lot.