BIG DATA PROJECT ASSIGNMENT

YET ANOTHER HADOOP

TEAM DETAILS:

NAME	SRN
SAI JATIN	PES2UG19CS353
SAI SHRI KRISHNA	PES2UG19CS356
SAMYUKTHA	PES2UG19CS361
SATHVIK M	PES2UG19CS369

DESIGN DETAILS:

1. Setting up a distributed file system.
2. Creation of data nodes.
3. Creation of name nodes (primary , secondary nodes).
4. Persistence Storage
5. Manipulating distributed file system.
6. Loading the distributed file system.
7. Accessing the distributed file system.
8. Stimulate the Running hadoop jobs.

IMPLEMENTATION DETAILS:

1. The setup process reads the config file and creates the DFS based on the configuration provided. If no such configuration is provided, it will resort to using a default configuration (using heartbeat method).After the setup has been completed, we create a file which

stores information about the DFS in `dfs_setup_config`. This file store the configuration settings to load the DFS for later use.

2. Data Nodes are created which store the data in blocks .Each file is split into the specified block size, and each block is replicated `replication_factor` number of times. Each replicated block is now stored in a Data Node. All information about the position of each replicated block in a data node is stored on the name node.

3.The primary Name Node keeps track of all changes being made in `path_to_fs`.

4.we implemented -put , -cat , -ls, -rm, -mkdir, -rmdir as CLI. CLI never crashes unless the process is terminated. It will run all the hadoop jobs and the output of reducer is stored in the given output directory.

REASON BEHIND THE DESIGN:

Reason behind the design is for easy storage of file system and easy manipulation of files.

TAKEAWAYS:

This project has exposed us to various fundamentals of hadoop file system. Which can be used to access and manipulate the files. The commands used for this are -put , -cat . which makes the viewing and accessing files easy.