Term Project Progress Report

Jinesh Dhruv
jad6566

Nisha Bhanushali
nnb7791

Title: Map Reduce Implementation

Goal: Parallel processing of a file in a distributed system based on the size of the file using map reduce.

Challenges:

1. Determining the number of servers required to process the file based on the size of the file (i.e. performance and scalability).
2. Splitting the file data into independent chunks (key-value pair) which are processed by the map tasks in a completely parallel manner over a group of servers.
3. Reducing all the independent chunks and assembling those chunks based on the same key and producing the final result.

Solution to the above challenges:

1. We won't assign random number of servers to process a particular file. Rather, we will be predefining the range for various sizes of the files that will help in determining the required number of servers needed to process those files. This will help us in achieving scalability (For instance, if the file size in between 1-2 MB use 2 servers, 2-3 MB use 3 servers and so on). We will not divide the file if the size of the file is small, as computational time will be high if we process this file using map reduce compared to process that file without map reduce. So, we will use map reduce to process only the large file whose size is greater than the predefined threshold. Thus, we will improve the performance for processing the different size files.
2. We will use a master that will divide the file into independent chunks and map each chunk to the respective server where the mapper function will generate the key value pair for the assigned chunk.
3. We will use a reducer function that will assemble all the values of the same key from all the servers that are performing mapping functions and do further computation on it to get the result.

Progress made so far:

1. Understood the working and structure of map-reduce.
2. Defined the scope of the project.
3. Created the sample data set of different sizes for processing.
4. Designed the basic structure for map-reduce implementation.
5. We have selected the threshold value for determining the number of servers needed to process the file.
6. Developed and tested the sorting algorithm.
7. Designed the master that will divide file into independent chunks.
8. Designed the user interface.

What to show in the demo:

1. User will be given an option to choose a file from the list of files.
2. Based on the file selection, we will display the information of file size and the number of servers needed for processing.
3. Then, display the size of newly created independent chunks at each server.
4. Now we will display the time required to process each chunk at the respective server.
5. Then we will display notification message when all the intermediate results of chunks are combined at one server.
6. Later, output the total time taken to process the file selected in step 1 using map reduce.
7. Show input file and the generated output file using map reduce.