Report of MapReduce Facility
Project 3 of 15-640
Kailiang Chen(kailianc) & Yang Pan(yangpan)

System Requirements

Log in 4 unix.andrew.cmu.edu using Terminals (we assume a User Case that there are 1 MapReduce master + 3 MapReduce slave , the master also acts as HDFS NameNode and 3 slaves also act as 3 DataNodes)

We do not need the AFS to run our system. But

Get the system run

Configure the config.txt

1. Open four Terminals and use “ifconfig” to check the IP addresses of each machine.
2. Choose one machine as MapReduce Master and the other three as MapReduce Slaves
3. Change the “MapReduce/config.txt” content, and set MasterHost, SlaveHosts, and DataMasterHost to the right IP addresses (MasterHost and Data MasterHost should be the same host). Note that the first slave host has the slave id (sid) 1 and so on (more on this later).
4. Change the “MasterRootDir ” and “SlaveRootDir” to the root directory to store the related files in every machine. Note that every directory in “SlaveRootDir” is corresponding to the host in “SlaveHosts” in order.
5. Change the “SlaveCapacity” to specify the capacity of every slave, which is the maximum number of jobs can run concurrently at one time. Also, the number in “SlaveCapacity” is corresponding to the host in “SlaveHosts” in order.
6. Change “NumberOfReducer” and “FileChunkSizeB” (of bytes) accordingly. Note that the numbers are fixed for every task once it’s configured.

7. Other options can be left unchanged.
8. OK. You can now run the system!

Run the system

1. Type “make” and then “make run” in the four Terminals, and then select the role of each machine. “m” means “master”, “s” means “slave”. Be sure to run master before run slave.
2. After the MapReduce master is started, the console will display “master”, which means having started successfully.
3. Then KPFS Data Master will be launched, when the DataMaster displays “KPFS master ready”, Data Master in distributed system has been launched.
4. In the slave console, type “s sid”(slave) to start it as a slave. Here, “sid”(slave id) is the number we talked in step 3 of “Configure the config.txt”.
5. When the MapReduce slaves are connected to the MapReduce master successfully. “Slave 1 connected”, “Slave 2 connected”, “Slave 3 connected” will display on the MasterNode’s console.
6. On the Slave’s terminal, “Connected to master” and “KPFS ready” will be displayed.
7. Now you are ready to start a work on our map-reduce system!

Start a work

Prepare the file

1. For every task, create a new directory with the task name under the root directory of data master specified in config.txt. Remember, task name is the ID of a task.
2. Create a directory named “UserFiles” (or the name specified under “UserDirName” option in config.txt) under the task directory.
3. Put the jar file (.jar) and input data file (.txt) into the “UserFiles” directory. Note that the name of jar and input data file should be the same as task name. Say the task name is “WordCounter”, and then the two files should have the names

"WordCounter.jar" and "WordCounter.txt".

4. After all these steps, you can now start the work!

Start/stop/monitor the work and system

1. On the Master's terminal, you can see an instruction showing all the commands you can type. Start a new task by typing "new [TaskName]".
2. You can know the status of a specific task by typing "show [TaskName]" or know all the tasks by "alltasks".
3. To know the alive slaves, you can type "aliveslaves". Type "slavestatus" to know the working status of every slaves (including dead ones for debug reason).
4. If you want to know the location of every file in KPFS, type "allfiles".
5. If you want to terminate a task, type "end [TaskName]".
6. After a task is finished, you can know from the master's console where the output files are located. Due to the limited time, we do not gather all files to master. So you have to grab the output files in different slaves (if you specific the number of reducers to be more than 1).