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)

# Configuration Steps

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. And other parameters can be changed according to the application’s requirement, the details of each parameter is explained in “Application\_Programmer\_Manual”.
4. Copy the “Example/WordCounter” or “Example/Twitter” to the root directory on the MapReduce Master. The root directory is specified in the “MapReduce/config.txt”. (preparation for running new tasks)

# How to start/stop/monitor 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”.
2. When the MapReduce master start to launch, 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. Switch to another terminal, type “make run”, and select the role of each machine to “s sid”(slave). Here, “sid”(slave id) must be set to a number according to the order of the IP address written in the config.txt. For example, “s 1”, or “s 2” or “s 3”.
5. When the MapReduce slaves are connected to the MapReduce mster 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. On the Master’s terminal, a new task can be started by “new taskName”.
8. When the task is launched, we can use the following commands to monitor the info of all the Slaves, the status of tasks and result file locations.
9. When a certain Slave is failed, Master will reschedule the jobs on that Slave to accomplish the overall tasks.
10. All the facility can be stopped by input “exit” in the Master’s terminal.

# How to run a new task

Assume a MapReduce Master and several MapReduce Slaves have been started and connected (Refer to former section to know how to start and connect)

1. Input “New XXX” on the Master’s Terminal (XXX is the taskName, e.g. WordCounter, Twitter)
2. “Spliting file for task XXX” will be displayed to show file split phase
3. “Sending jobs of XXX out to slaves” will be displayed to show job assignment phase.
4. “Job assignment of XXX finished” will be displayed to show job assignment finished.