CS 580K Advanced Topics in Cloud Computing – Assignment 3

Ketan Deshpande
B00816854
kdeshpa5@binghamton.edu

Q.1. Can you figure out what are the main steps do we need to run a Hadoop mapreduce task (i.e., wordcount here)?

➔ To set up a Hadoop cluster, you will need to install JAVA and Hadoop binary (the mapreduce programming framework and HDFS file system), and configure the environment. But for the sake of simplicity we will use docker image and its git repository.

1. Pull docker image
2. Clone git hub repository
3. Create Hadoop network to connect Hadoop connectors
4. Start Hadoop container by going into hadoop-cluster-docker repository
5. Start Hadoop clusters
6. Then directly run the run-wordcount.sh program which then runs the map and reduce function and calculate words in 2 files file1.txt and file2.txt

Q.2 What does this command mean — "hdfs dfs -put ./input/* input"?

➔ As per the Hadoop documentation the put command is used to copy single src, or multiple srcs from local file system to the destination file system. It also reads input from stdin and writes to destination file system if the source is set to "-". Basically, we're transferring the contents of the input folder to Hadoop' distributed file system.

Q.3 How many mappers and reducers are launched for executing the above wordcount program?

➔ Number of mappers = 2
Number of reducers = 1

Q.4 How much time do mappers and reducers spend for the above tasks, separately?

➔ Total time spent by all map tasks (ms)=10970
Total time spent by all reduce tasks (ms)=5038

Q.5 After execution, what are the files in the output folder in HDFS, and what content do they contain?

➔ The output folder contains following files
_SUCCESS – file is empty
part-r-00000 – contains the output of the program i.e.
Docker 1
Hadoop 1
Hello 2

Q.6 How many master and slave containers do you launch separately this time?

➔ 1 master and 4 slave containers

Q.7 Please figure out what a master container/node and a slave container/node are used for.

➔ Master containers -

- HDFS cluster contains a single master node, which is also called Name node.
- It manages the file system namespace and regulates access to file system by clients
- It executes the file system operations like opening, closing, and renaming files and directories
- It also determines the mapping of blocks to data nodes

Slave containers -

- Slave containers store the actual data blocks
- These are responsible for serving read and write requests from the file system's clients
- They also perform block creation, deletion, and replication upon instruction from the Master

Q.8 How many mappers and reducers are launched for executing the above wordcount program?

➔ Number of mappers = 3
Number of reducers = 1

Q.9 How much time do mappers and reducers spend for the above tasks, separately?

➔ Total time spent by all map tasks (ms)=33204
Total time spent by all reduce tasks (ms)=5934

Q.10 What are the two most frequently occurring words, and how many times do they occur?

➔ The – 42
Of – 27

Q.11 Please describe the basic steps in the map function of WordCount.java.

➔ In the map function, the input text is tokenized using the object of StringTokenizer.

- We will repeat the loop unless there are no more tokens can be generated i.e. end of the input
- In each loop, we will create a key value pair, each token i.e. word will be a new key with value 1
- We will write these key value pairs

Q.12 Please describe the basic steps in the reduce function of WordCount.java.

➔ The basic steps are follows:

- We will repeat the loop on the IntWritable values object
- The reducer will be called on each different key, and it will sum the number of entries in the values.
- Then the reducer will store these sums for each unique key.

Q.13 How many mappers and reducers are launched for executing the above wordcount program?

➔ Number of mappers = 2
Number of reducers = 1

Q.14 How much time do mappers and reducers spend for the above tasks, separately?

➔ Total time spent by all map tasks (ms)=12252
Total time spent by all reduce tasks (ms)=8724