CS580K Mini_project 2 :

Name : Shubham Patwa

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

For running a hadoop map reduce task these are the following main steps:

Firstly ,

We need to pull the docker image , create Hadoop network, and start the Hadoop container clusters.

Here, we have started the container with 1 master and 2 slaves.

We need to have more than 1 input files in the input folder and then we need to push those files onto HDFS using put command which is taken care of in run-wordcount.sh file

Then, we need to run the run-wordcount.sh file to count the word occurrences.

This file then starts the process which then performs mapping, splitting by the two data nodes.

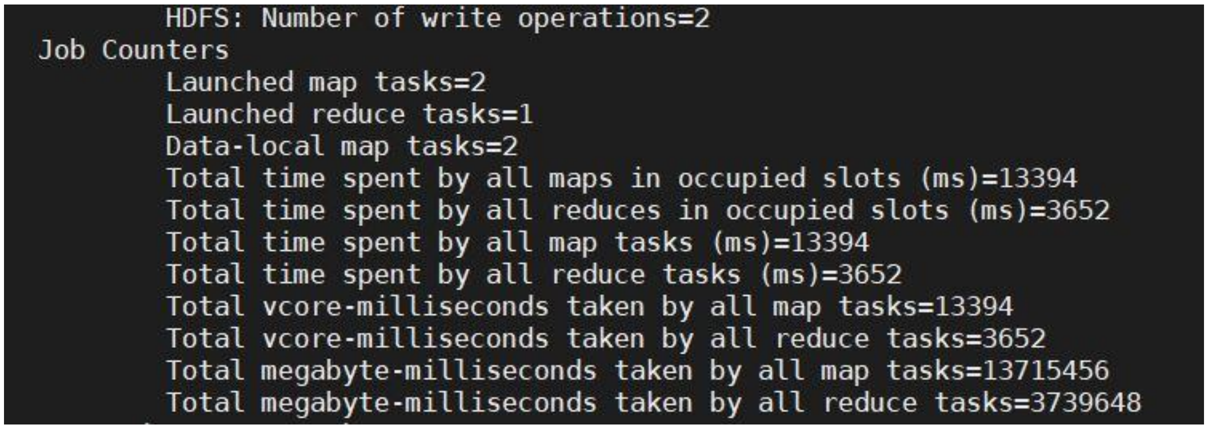
The master node gets reported the data performed by the data nodes which then performs the shuffling and reducing process.

The words and its count are then displayed.

• Q.2 What does this command mean — "hdfs dfs -put ./input/* input"? (Hint, HDFS is Hadoop's distributed file system.)

The input files present in the 'input' folder are put onto the HDFS(Hadoop distributed file system) using the put command.

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



```
		HDFS: Number of write operations=2
	Job Counters
		Launched map tasks=2
		Launched reduce tasks=1
		Data-local map tasks=2
		Total time spent by all maps in occupied slots (ms)=13394
		Total time spent by all reduces in occupied slots (ms)=3652
		Total time spent by all map tasks (ms)=13394
		Total time spent by all reduce tasks (ms)=3652
		Total vcore-milliseconds taken by all map tasks=13394
		Total vcore-milliseconds taken by all reduce tasks=3652
		Total megabyte-milliseconds taken by all map tasks=13715456
		Total megabyte-milliseconds taken by all reduce tasks=3739648
```

Mappers :2

Reducers : 1

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

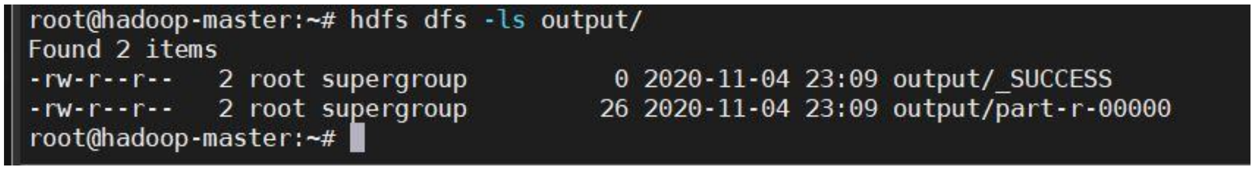
As given in the image above,

Time for mappers : 13394 ms

Time for reducers : 3652 ms

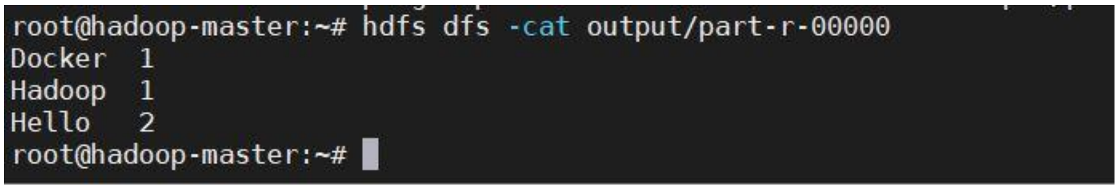
• Q.5 After execution, what are the files in the output folder in HDFS, and what content do they contain?

The output folder contains the following files:



```
root@hadoop-master:~# hdfs dfs -ls output/
Found 2 items
-rw-r--r--   2 root supergroup          0 2020-11-04 23:09 output/_SUCCESS
-rw-r--r--   2 root supergroup         26 2020-11-04 23:09 output/part-r-00000
root@hadoop-master:~#
```

Content is:

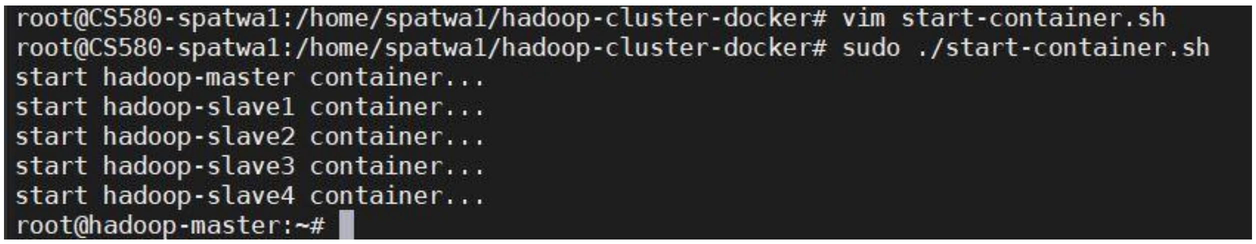


```
root@hadoop-master:~# hdfs dfs -cat output/part-r-00000
Docker  1
Hadoop  1
Hello   2
root@hadoop-master:~#
```

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

Cluster size has been resized to 5.

1 master and 4 slaves were launched.



```
root@CS580-spatwa1:/home/spatwa1/hadoop-cluster-docker# vim start-container.sh
root@CS580-spatwa1:/home/spatwa1/hadoop-cluster-docker# sudo ./start-container.sh
start hadoop-master container...
start hadoop-slave1 container...
start hadoop-slave2 container...
start hadoop-slave3 container...
start hadoop-slave4 container...
root@hadoop-master:~#
```

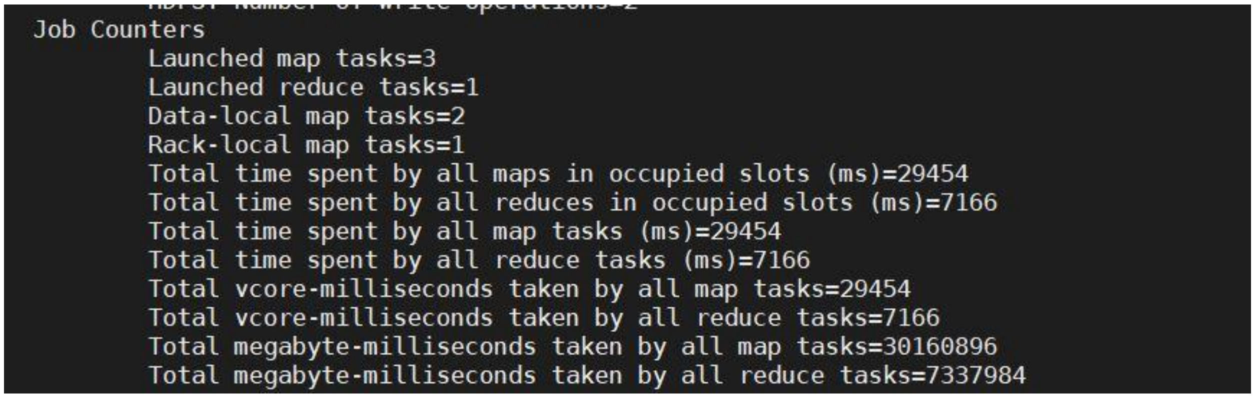
Q.7 Please figure out what a master container/node and a slave container/node are used for.

MasterNode : Master node in a hadoop cluster is responsible for storing data in HDFS and executing parallel computation the stored data using MapReduce. Master Node has 3 nodes – NameNode, Secondary NameNode and JobTracker. JobTracker monitors the parallel processing of data using MapReduce while the NameNode handles the data storage function with HDFS. NameNode keeps a track of all the information on files (i.e. the metadata on files) such as the access time of the file, which user is accessing a file on current time and which file is saved in which hadoop cluster. The secondary NameNode keeps a backup of the NameNode data.

Slave/Worker Node- This component in a hadoop cluster is responsible for storing the data and performing computations. Every slave/worker node runs both a TaskTracker and a DataNode service to communicate with the Master node in the cluster. The DataNode service is secondary to the NameNode and the TaskTracker service is secondary to the JobTracker.

Reference : https://www.dezyre.com/article/hadoop-cluster-overview-what-it-is-and-how-to-setup-one/356#:~:text=Master%20Node%20has%203%20nodes%20%E2%80%93%20NameNode%2C%20Secondary%20NameNode%20and%20JobTracker.&text=Slave%2FWorker%20Node%2D%20This%20component,Master%20node%20in%20the%20cluster.

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



```
		HDFS: Number of write operations=2
	Job Counters
		Launched map tasks=3
		Launched reduce tasks=1
		Data-local map tasks=2
		Rack-local map tasks=1
		Total time spent by all maps in occupied slots (ms)=29454
		Total time spent by all reduces in occupied slots (ms)=7166
		Total time spent by all map tasks (ms)=29454
		Total time spent by all reduce tasks (ms)=7166
		Total vcore-milliseconds taken by all map tasks=29454
		Total vcore-milliseconds taken by all reduce tasks=7166
		Total megabyte-milliseconds taken by all map tasks=30160896
		Total megabyte-milliseconds taken by all reduce tasks=7337984
```

Mappers : 3

Reducers : 1

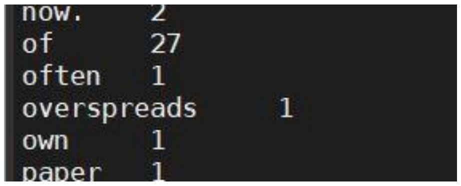
Q.9 How much time do mappers and reducers spend for the above tasks, separately?

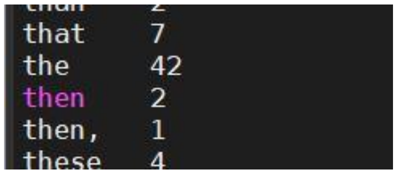
Referring to the above figure, time taken for :

Mappers: 29454ms

Reducers: 7166ms

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


```
now.      2
of        27
often     1
overspreads      1
own       1
paper     1
```


```
that      7
the       42
then      2
then,     1
these     4
```

The most frequent words are :

‘of’ – occurring 27 times and

‘the ‘ – occurring 42 times

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

Reference : https://hadoop.apache.org/docs/r1.2.1/mapred_tutorial.html

Basic Steps:

The mapper implementation via the map method processes one line at a time. It then splits the line into tokens separated by white spaces via the StringTokenizer which then emits a key-value pair of ((word,1)) , that is it converts the words into keys and occurrences as values.

The text is converted to string using the toString function and the while checks for more tokens if available. In the loop, (itr.nextToken) will be stored in word using word.set and then it will be given to context.write which will in turn be given as input to reducer.

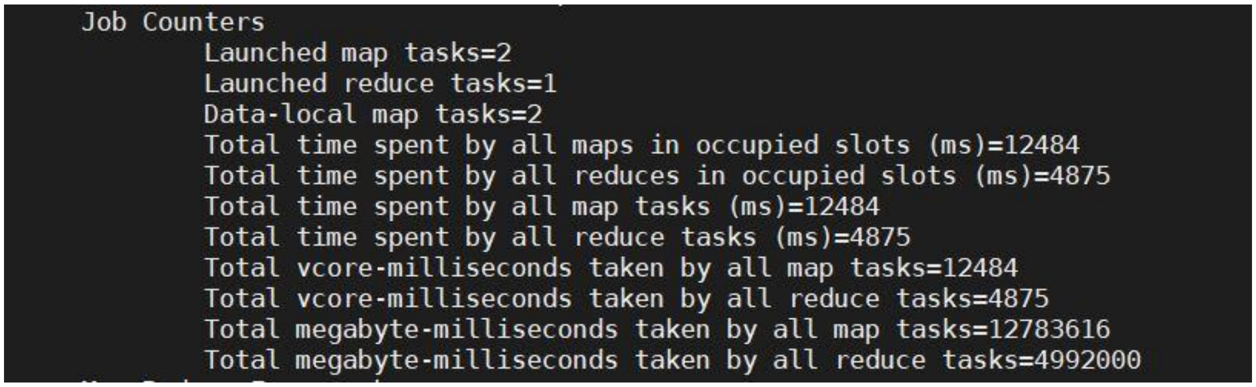
Q.12 Please describe the basic steps in the reduce function of WordCount.java

Reference : https://hadoop.apache.org/docs/r1.2.1/mapred_tutorial.html

The reduced function takes the words as key and occurrences as values and just sums up the values for each key. The values are in the form of IntWritable.

The sum is set using result and is written using context.write with word as key and result as value.

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



```
     Job Counters
             Launched map tasks=2
             Launched reduce tasks=1
             Data-local map tasks=2
             Total time spent by all maps in occupied slots (ms)=12484
             Total time spent by all reduces in occupied slots (ms)=4875
             Total time spent by all map tasks (ms)=12484
             Total time spent by all reduce tasks (ms)=4875
             Total vcore-milliseconds taken by all map tasks=12484
             Total vcore-milliseconds taken by all reduce tasks=4875
             Total megabyte-milliseconds taken by all map tasks=12783616
             Total megabyte-milliseconds taken by all reduce tasks=4992000
```

Mappers : 2

Reducers : 1

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

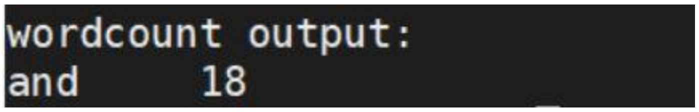
Using the figure above, time taken for:

Mappers: 12484ms

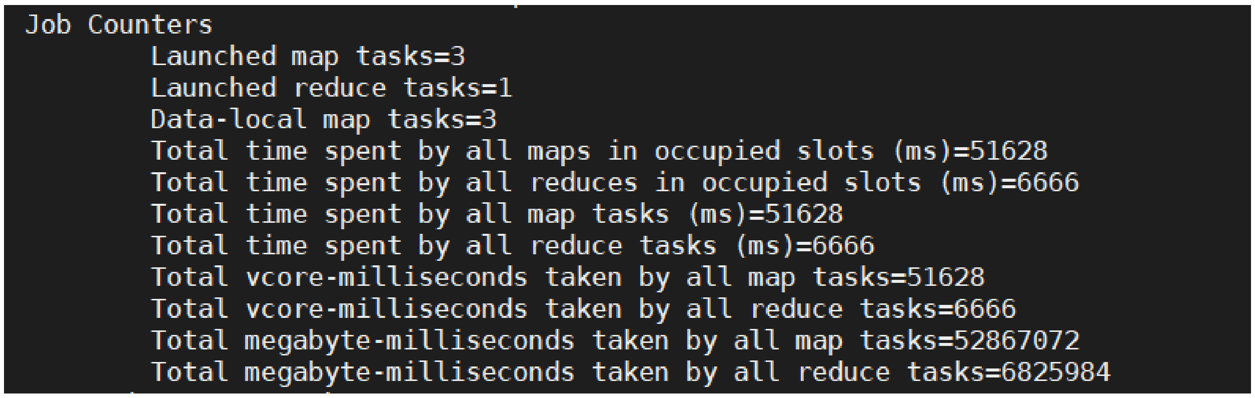
Reducers : 4875ms

Task IV: Write Your Own Simple Hadoop Program :

After modifying some code , the output that I got is:



```
wordcount output:
and      18
```



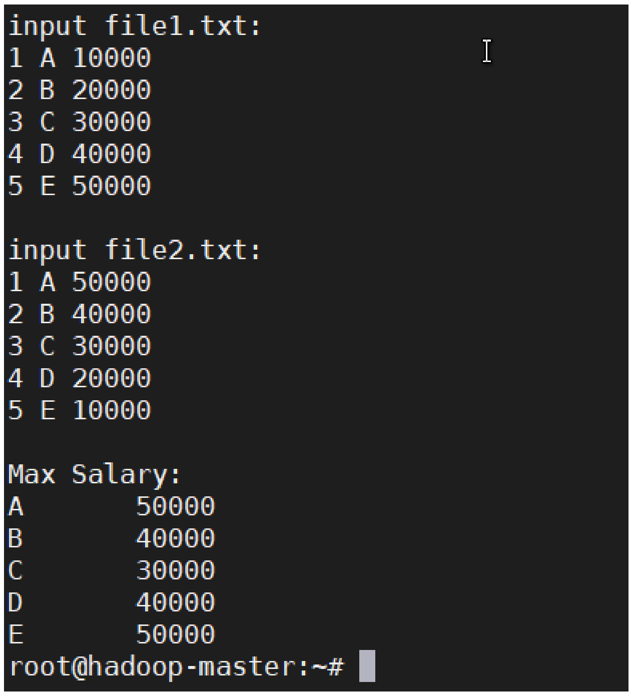
```
 Job Counters
         Launched map tasks=3
         Launched reduce tasks=1
         Data-local map tasks=3
         Total time spent by all maps in occupied slots (ms)=51628
         Total time spent by all reduces in occupied slots (ms)=6666
         Total time spent by all map tasks (ms)=51628
         Total time spent by all reduce tasks (ms)=6666
         Total vcore-milliseconds taken by all map tasks=51628
         Total vcore-milliseconds taken by all reduce tasks=6666
         Total megabyte-milliseconds taken by all map tasks=52867072
         Total megabyte-milliseconds taken by all reduce tasks=6825984
```

Task V: Other Hadoop Use Cases:

I have two text files with serial number, employee name and salary in each of the files.

The program has the function to find max salary of the same person from different files.

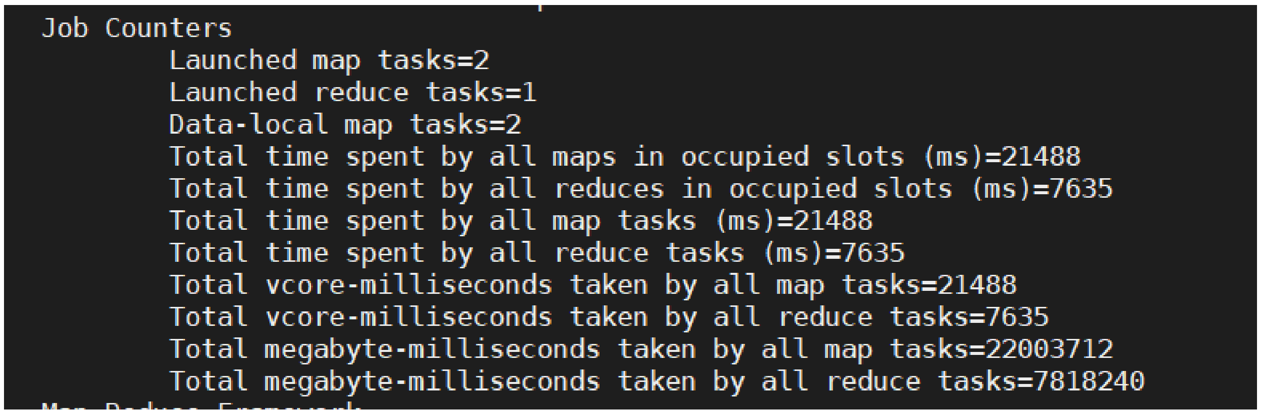
Thus the output we get is as follows: (Modified the given program):



```
input file1.txt:
1 A 10000
2 B 20000
3 C 30000
4 D 40000
5 E 50000

input file2.txt:
1 A 50000
2 B 40000
3 C 30000
4 D 20000
5 E 10000

Max Salary:
A       50000
B       40000
C       30000
D       40000
E       50000
root@hadoop-master:~#
```



```
  Job Counters
          Launched map tasks=2
          Launched reduce tasks=1
          Data-local map tasks=2
          Total time spent by all maps in occupied slots (ms)=21488
          Total time spent by all reduces in occupied slots (ms)=7635
          Total time spent by all map tasks (ms)=21488
          Total time spent by all reduce tasks (ms)=7635
          Total vcore-milliseconds taken by all map tasks=21488
          Total vcore-milliseconds taken by all reduce tasks=7635
          Total megabyte-milliseconds taken by all map tasks=22003712
          Total megabyte-milliseconds taken by all reduce tasks=7818240
```