**Mini Project 3**

**Task I: Word Count — Hadoop’s “Hello World’**

* **Q.1. Can you figure out what are the main steps do we need to run a Hadoop map reduce task (i.e., wordcount here)? (Ask TA for Clarification)**
  + We start docker master and slave containers using sudo ./start-container.sh starting hadoop-master and hadoop-slave-1 and hadoop-slave-2 container and we get automatically shelled inside the hadoop-master container.
  + We have to launch the hdfs and YARN from Hadoop master container root directory.
  + Create input files containing random strings
  + Move these files into HDFS
  + We run the wordcount program.
* **Q.2 What does this command mean — “hdfs dfs -put. /input/\* input”? (Hint, HDFS is Hadoop’s distributed file system. Please refer to [4].)** 
  + HDFS allows user data to be organized in the form of files and directories. It provides a command line interface called FS shell that lets a user interact with the data in HDFS.
  + So hdfs stands for Hadoop Distributed file System while dfs refers to distributed file system.
  + In the command **“hdfs dfs -put. /input/\* input”** we are transferring the contents of the input folder into HDFS. Here the content will be broken down into chunks and stored on the disk.
* **Q.3 How many mappers and reducers are launched for executing the above wordcount program?**

Total Number of Mappers – 2

Total Number of Reducers – 1

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

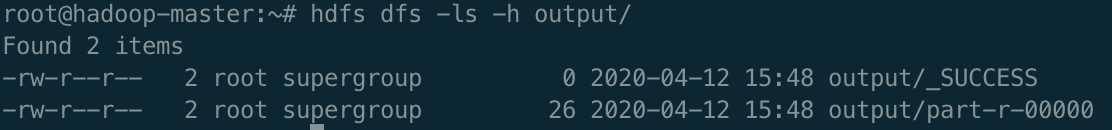
Total time spent by mappers(ms) = 19803

Total time spent by reducers (ms)= 5382

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

The output folder in HDFS contains 2 files –

* + 1. \_SUCCESS
    2. part-r-00000

****

The contents of part-r-00000 are

Docker 1

Hadoop 1

Hello 2

**A picture containing meter

Description automatically generated**

The file \_SUCCESS is empty

****

**Task II: Resizing the Hadoop Cluster**

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

Master Containers – 1

Slave containers - 4

**A close up of a logo

Description automatically generated**

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

Master containers

* The Name Node is the Hadoop master node
* Name Node consults with Data Nodes in the cluster when copying data or running mapReduce operations. Manages HDFS storage. To ensure high availability, you have both an active Name Node and a standby Name Node. Each run on its own, dedicated master node.

Slave Containers

* Data Node is the Hadoop slave node.
* Files copied to Hadoop mountpoint exist as blocks on different Data Nodes in the cluster. The collection of Data Nodes is what we call the HDFS

**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)=43821

Total time spent by all reduce tasks (ms)=6849

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

Two most frequently occurring words -

the – 42 occurrences

of – 27 occurrences

**Task III: Read the Source Code of the WordCount Program**

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

* We create StringTokenizer named Itr. StringTokenizer helps to break string into tokens. If Itr has more tokens the we set the next token into the word object.
* The loop is repeated until the end of input for generation of tokens.

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

* The created keys are arranged, keeping the unique keys intact and combining the replica keys.
* Add the number of entries for a value.
* Store the sum of each unique key.

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

Launched map tasks=2

Launched reduce tasks=1

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

Total time spent by all map tasks (ms)=18242

Total time spent by all reduce tasks (ms)=6030

**Task IV: Write Your Own Simple Hadoop Program**

In this task, we are going to write a variant of the WordCount program. Instead of counting the number of the occurrences of each word from the input files, we only focus on a specific word. Particularly, we would like to only count the occurrences of word “and”. Please write your Hadoop map and reduce functions to realize this. For example, the output will look like (if we use the “text.txt” file as the input): $wordcount output:

$and 18

A picture containing table

Description automatically generated