Task I:

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

1. [Hadoop Installation](https://www.dezyre.com/hadoop-tutorial/big-data-hadoop-tutorial)
2. install JAVA
3. ssh must be installed and sshd must be running to use the Hadoop scripts that manage remote Hadoop daemons.
4. Start hadoop container clusters : Single node hadoop cluster must be configured,
   1. start 3 containers with 1 master and 2 slaves
   2. which will lead inside hadoop-master container
5. Start hadoop clusters command

./start-hadoop.sh

1. Starting namenode and yarn daemons and Resource Manager daemon
2. Create 2 inputfiles
3. Create input directory on HDFS
4. Put input files to HDFS

*Q.2 What does this command mean — “hdfs dfs -put ./input/\* input”? (Hint, HDFS is Hadoop’s distributed ﬁle system*

Hadoop will copy the input and Put input files in HDFS

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

Launched map tasks=2

Launched reduce tasks=1

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

Total time spent by all map tasks (ms)=10431

Total time spent by all reduce tasks (ms)=4805

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

on completion file generated are “\_SUCCESS “ and “part-r-00000” where (r=reduce m=mapper)

wordcount output:

Docker 1

Hadoop 1

Hello 2

This contains the word count for every word in input files

Task II:

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

1 master and 4 slave

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

HDFS has a master/slave architecture. An HDFS cluster consists of a single NameNode, a master server that manages the file system namespace and regulates access to files by clients. In addition, there are a number of DataNodes, usually one per node in the cluster, which manage storage attached to the nodes that they run on. HDFS exposes a file system namespace and allows user data to be stored in files. Internally, a file is split into one or more blocks and these blocks are stored in a set of DataNodes. The NameNode executes file system namespace operations like opening, closing, and renaming files and directories. It also determines the mapping of blocks to DataNodes. The DataNodes are responsible for serving read and write requests from the file system’s clients. The DataNodes also perform block creation, deletion, and replication upon instruction from the NameNode.

Master: NameNode

Slave: {Datanode}…..{Datanode}

-     The Master (NameNode) manages the file system namespace operations like opening, closing, and renaming files and directories and determines the mapping of blocks to DataNodes along with regulating access to files by clients

-     Slaves (DataNodes) are responsible for serving read and write requests from the file system’s clients along with perform block creation, deletion, and replication upon instruction from the Master (NameNode).

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

Launched map tasks=3

Launched reduce tasks=1

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

Total time spent by all map tasks (ms)=30196

Total time spent by all reduce tasks (ms)=5498

Task III:

*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*

1. Extend Mapper interface
2. Create IntWritable one to store value 1 to be assigned to word
3. Text word to store key in string format
4. Write map function that receives line by line input from file
5. Split the line by space using StringTokenizer itr
6. In while loop set the word with the key word
7. In context which outputs the value add word and one

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

1. In the reduce function, Iterable<IntWritable> values contains the list from map
2. Iterate the list and sum the value
3. Add the value with the key word to output context

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

Launched map tasks=3

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)=26401

Total time spent by all reduce tasks (ms)=6388

Task IV:

Code added to display number of “and”

*String value1 = key.toString();*

*if(value1.equals("and"))*

Task V:

***In this problem statement, we will find the days on which each basement has more trips.***

**Mapper Class:**

From theMapper, we will take the combination of the basement and the day of the week as key and the number of trips as value.

First, we will parse the date, which is in string format into date format using **SimpleDateFormat**class in Java. Now, to take out the day of the date, we will use the **getDay()**which willreturn an integer value with the day of the week’s number. So, we have **created an array** which consists of all the days from Sunday to Monday and have passed the value returned by **getDay()**into the array in order to get the day of the week.

Now, after this operation, we have returned the combination of **Basement\_number+Day of the week**as keyand the **number of trips** as value.

**Reducer Class:**

In the reducer, we will calculate the sum of trips for each basement and for each particular day, by using the below lines of code.