**Part 2:**

**Approach:**

1. Created an input directory in Hadoop for Covid and Olympics keyword data as part of part 1.
2. Moved the files in the covid\_olympics\_dt directories to Hadoop new input directory to reduce the number of sub-directories.
3. Using MapReduce, I wrote a Java program that counts the number of rows in the downloaded data directory.  The Mapper function in the program takes tweet JSON data as a key, value as 1, and the reducer reduces the (tweet,1) and counts the number of unique lines in the downloaded data. The entire number of lines or row count would be the program's final output. Placed the Java code file (Tharuni\_Samineni\_Program\_2.java) file in csx.
4. Compile the java program (Tharuni\_Samineni\_Program\_2.java).
5. Create a jar file.
6. Used a command to run the MapReduce program.
7. Once the program is compiled and run successfully, the final output of the program can be viewed by writing a command.

**Instructions to run the code:**

2,1) Created an input directory using the following command:

- hdfs dfs -mkdir /user/ssamine/covid\_olympics\_dt

Text

Description automatically generated

Figure 1: Shows the creation of an input directory

2,2) To move the files from directory(covid\_olympics\_dt) to a new hadoop input directory(inputdata) to reduce the number of sub-directories path use the following command:

- hdfs dfs -cp /user/ssamine/covid\_olympics\_dt/ /user/ssamine/covid\_olympics\_ipt



Figure 2: Shows the command of moving downloaded data from covid\_olympics\_dt to covid\_olympics\_ip

2,3) Used another directory as the first directory was using each of the month date and year as sub-directories, so created a new directory with proper command

- hdfs dfs -cp /user/ssamine/covid\_olympics\_ipt/2022/03/06/23/\* /user/ssamine/inpudata

A screenshot of a computer

Description automatically generated with medium confidence

Figure 3: Shows the command to move data from covid\_olympics\_ipt directory to inputdata directory

2,4) To view the data present in the directory use the following command:

- hdfs dfs -ls /user/ssamine/inputdata

A screenshot of a computer

Description automatically generated with medium confidence

A screenshot of a computer

Description automatically generated with medium confidence

Figures 4: Shows the data present in the directory

2,5) To compile the (Tharuni\_Samineni\_Program\_2.java)mapreduce program for line count use the following command:

- $HADOOP\_HOME/bin/hadoop com.sun.tools.javac.Main Tharuni\_Samineni\_Program\_2.java

A screenshot of a computer

Description automatically generated with medium confidence

Figure 5: Shows the command to compile the MapReduce program (Tharuni\_Samineni\_Program\_2)

2,6) To create a jar file use the following command:

- jar cf rc.jar Tharuni\_Samineni\_Program\_2\*.class

A screenshot of a computer

Description automatically generated with medium confidence

Figure 6: Shows the command for creation of jar file

2,7) To run the mapreduce program use the following command:

- hadoop jar rc.jar Tharuni\_Samineni\_Program\_2 /user/ssamine/inputdata /user/ssamine/testop (Use different directories for output, otherwise an error will be shown that directory already exist)

A screenshot of a computer

Description automatically generated with medium confidence

Figure 7: Shows the command to run the mapreduce code

2,8) To check the output of the program use the command:

- hdfs dfs -cat /user/ssamine/testop/part-r-00000

Text

Description automatically generated

Figure 8: Shows the command to print the output of the program

**Expected Final Output**- Total Row count: 30320