CSE 532 Theory of Database Systems, Spring 2020, Assignment 4: Readme

Tianao Wang 112819772

April 27, 2020

1 Environment

VB + Cloudera Quickstart VM + jdk1.8 + hadoop2.6 + spark2.4.5

2 Task 1

2.1 Task 1.1

Command:
javac -cp ‘hadoop classpath‘ Covid19_1.java -d build -Xlint

jar -cvf Covid19_1.jar -C build/ .

hdfs dfs -rm -r /cse532/output

hadoop jar Covid19_1.jar Covid19_1 /cse532/input/covid19_full_data.csv true /cse532/output/

hdfs dfs -cat /cse532/output/*

```
        HDFS: Number of write operations=2
Job Counters
        Launched map tasks=1
        Launched reduce tasks=1
        Data-local map tasks=1
        Total time spent by all maps in occupied slots (ms)=8910
        Total time spent by all reduces in occupied slots (ms)=8932
        Total time spent by all map tasks (ms)=8910
        Total time spent by all reduce tasks (ms)=8932
        Total vcore-milliseconds taken by all map tasks=8910
        Total vcore-milliseconds taken by all reduce tasks=8932
        Total megabyte-milliseconds taken by all map tasks=9123840
        Total megabyte-milliseconds taken by all reduce tasks=9146368
```

2.2 Task 1.2

Command:
javac -cp ‘hadoop classpath‘ Covid19_2.java -d build -Xlint

jar -cvf Covid19_2.jar -C build/ .

hdfs dfs -rm -r /cse532/output

hadoop jar Covid19_2.jar Covid19_2 /cse532/input/covid19_full_data.csv 2020-01-01 2020-03-31 /cse532/output/

hdfs dfs -cat /cse532/output/*

```
Job Counters
        Launched map tasks=1
        Launched reduce tasks=1
        Data-local map tasks=1
        Total time spent by all maps in occupied slots (ms)=33118
        Total time spent by all reduces in occupied slots (ms)=20796
        Total time spent by all map tasks (ms)=33118
        Total time spent by all reduce tasks (ms)=20796
        Total vcore-milliseconds taken by all map tasks=33118
        Total vcore-milliseconds taken by all reduce tasks=20796
        Total megabyte-milliseconds taken by all map tasks=33912832
        Total megabyte-milliseconds taken by all reduce tasks=21295104
```

2.3 Task 1.3

Command:
javac -cp ‘hadoop classpath‘ Covid19_3.java -d build -Xlint

jar -cvf Covid19_3.jar -C build/ .

hdfs dfs -rm -r /cse532/output

hadoop jar Covid19_3.jar Covid19_3 /cse532/input/covid19_full_data.csv
hdfs://quickstart.cloudera:8020/cse532/cache/populations.csv /cse532/output/

hdfs dfs -cat /cse532/output/*

```
        HDFS: Number of write operations=2
Job Counters
        Launched map tasks=1
        Launched reduce tasks=1
        Data-local map tasks=1
        Total time spent by all maps in occupied slots (ms)=17927
        Total time spent by all reduces in occupied slots (ms)=28178
        Total time spent by all map tasks (ms)=17927
        Total time spent by all reduce tasks (ms)=28178
        Total vcore-milliseconds taken by all map tasks=17927
        Total vcore-milliseconds taken by all reduce tasks=28178
        Total megabyte-milliseconds taken by all map tasks=18357248
        Total megabyte-milliseconds taken by all reduce tasks=28854272
```

3 Task2

3.1 Task 2.1

Command:
javac -cp ”/home/cloudera/Desktop/spark/spark-2.4.5-bin-hadoop2.7/jars/*” SparkCovid19_1.java

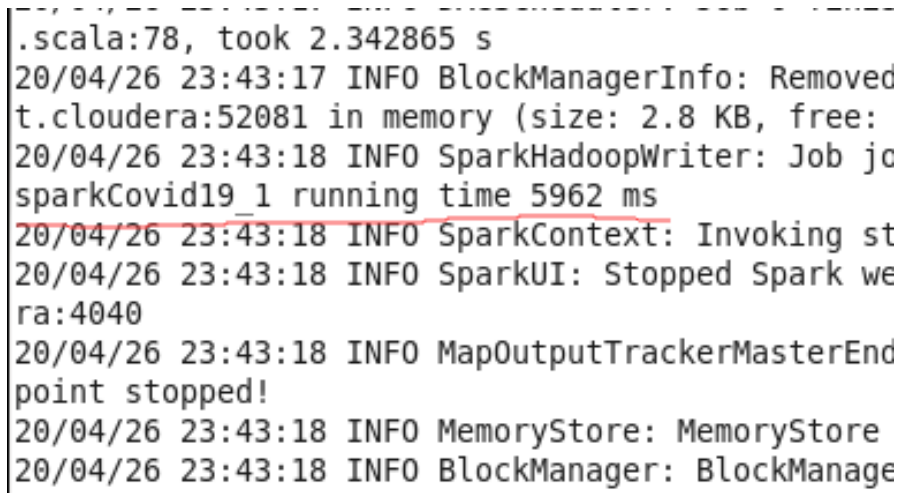
-d build -Xlint

jar -cvf SparkCovid19_1.jar -C build/ .

hdfs dfs -rm -r /cse532/output

spark-submit –class SparkCovid19_1 SparkCovid19_1.jar
hdfs://quickstart.cloudera:8020/cse532/input/covid19_full_data.csv 2020-01-01
2020-03-31 hdfs://quickstart.cloudera:8020/cse532/output/

hdfs dfs -cat /cse532/output/*



3.2 Task 2.2

Command:
javac -cp ”/home/cloudera/Desktop/spark/spark-2.4.5-bin-hadoop2.7/jars/*” SparkCovid19_2.java
-d build -Xlint

jar -cvf SparkCovid19_2.jar -C build/ .

hdfs dfs -rm -r /cse532/output

spark-submit –class SparkCovid19_2 SparkCovid19_2.jar
hdfs://quickstart.cloudera:8020/cse532/input/covid19_full_data.csv
hdfs://quickstart.cloudera:8020/cse532/cache/populations.csv
hdfs://quickstart.cloudera:8020/cse532/output/

hdfs dfs -cat /cse532/output/*

```
.scala:78, took 9.017890 s
20/04/26 23:33:28 INFO SparkHadoopWriter: Job job_2020042623
sparkCovid19_2 running time 25406 ms
20/04/26 23:33:29 INFO SparkContext: Invoking stop() from sh
20/04/26 23:33:29 INFO SparkUI: Stopped Spark web UI at http
ra:4042
```