ASSIGNMENT-2

Part A : Assignments based on the Hadoop

Aim:

Design and develop a distributed application to find the coolest/hottest year from the available weather data. Use weather data from the Internet and process it using MapReduce.

Introduction

MapReduce is a framework using which we can write applications to process huge amounts of data, in parallel, on large clusters of commodity hardware in a reliable manner. MapReduce is a processing technique and a program model for distributed computing based on java.

The MapReduce algorithm contains two important tasks, namely Map and Reduce.Map takes a set of data and converts it into another set of data, where individual elements are broken down into tuples (key/value pairs).

Secondly, reduce task, which takes the output from a map as an input and combines those data tuples into a smaller set of tuples. As the sequence of the name MapReduce implies, the reduce task is always performed after the map job.

The major advantage of MapReduce is that it is easy to scale data processing over multiple computing nodes.

Under the MapReduce model, the data processing primitives are called mappers and reducers.
Decomposing a data processing application into mappers and reducers is sometimes nontrivial. But, once we write an application in the MapReduce form, scaling the application to run over hundreds, thousands, or even tens of thousands of machines in a cluster is merely a configuration change. This simple scalability is what has attracted many programmers to use the MapReduce model.

The Algorithm

MapReduce program executes in three stages, namely map stage, shuffle stage, and reduce stage.
Mapstage : The map or mapper's job is to process the input data. Generally the input data is in the form of file or directory and is stored in the Hadoop file system (HDFS). The input file is
passed to the mapper function line by line. The mapper processes the data and creates several small chunks of data.
Reduce stage : This stage is the combination of the Shuffle stage and the Reduce stage. The Reducer's job is to process the data that comes from the mapper. After processing, it
produces a new set of output, which will be stored in the HDFS.

Inserting Data into HDFS:

•The MapReduce framework operates on <key, value> pairs, that is, the framework views the input to the job as a set of <key, value> pairs and produces a set of <key, value> pairs as the output of the job, conceivably of different types.
• The key and the value classes should be in serialized manner by the framework and hence, need to implement the Writable interface. Additionally, the key classes have to implement the Writable-Comparable interface to facilitate sorting by the framework.
•Input and Output types of a MapReduce job: (Input) <k1,v1> -> map -> <k2, v2>-> reduce -> <k3, v3> (Output).

The input for our program is weather data files for each year This weather data is collected by National Climatic Data Center – NCDC from weather sensors at all over the world. You can find weather data for each year from ftp://ftp.ncdc.noaa.gov/pub/data/noaa/.All files are zipped by year and the weather station. For each year, there are multiple files for different weather stations .

Steps for Compilation & Execution of Program:

```
#sudo mkdir analyzelogs
   ls
 #sudo chmod -R 777 analyzelogs/
          cd
         ls
         cd ..
         pwd
         ls
         cd
         pwd
#sudo chown -R hduser analyzelogs/
         cd
         ls
 #cd analyzelogs/
         ls
         cd ..
```

Copy the Files (Mapper.java,Reduce.java,Driver.java to Analyzelogs Folder)
#sudo cp /home/mde/Desktop/count_logged_users/* -/analyzelogs/

Start HADOOP
#start-dfs.sh
#start-yarn.sh
#jps

```
        cd
        cd analyzelogs
        ls
        pwd
        ls
#ls -ltr
#ls -al
 #sudo chmod +r *.*
        pwd
#export CLASSPATH="$HADOOP_HOME/share/hadoop/mapreduce/hadoop-mapreduce-client-core-
2.9.0.jar:$HADOOP_HOME/share/hadoop/mapreduce/hadoop-mapreduce-client-common-
2.9.0.jar:$HADOOP_HOME/share/hadoop/common/hadoop-common-
2.9.0.jar:~/analyzelogs/SalesCountry/*:$HADOOP_HOME/lib/*"
```

Compile Java Files

```
# javac -d . SalesMapper.java SalesCountryReducer.java
   SalesCountryDriver.java ls
 #cd SalesCountry/
   ls
  cd ..
  #sudo gedit Manifest.txt
  #jar -cfm analyzelogs.jar Manifest.txt
    SalesCountry/*.class ls
   cd
   jps
#cd analyzelogs/
```

Create Directory on Hadoop

```
#sudo mkdir ~/input2000
    ls
   pwd
 #sudo cp access_log_short.csv ~/input2000/
# $HADOOP_HOME/bin/hdfs dfs -put ~/input2000 /
 # $HADOOP_HOME/bin/hadoop jar analyzelogs.jar /input2000 /output2000

# $HADOOP_HOME/bin/hdfs dfs -cat /output2000/part-00000

# stop-all.sh
  # jps
```

Name: Rohan S Kadu PRN no:71901492B Roll no: T1851004

Output:

```
rohan@rohan-HP-205-G1-AiO-Business-PC:~$ su hduser
Password:
hduser@rohan-HP-205-G1-AiO-Business-PC:/home/rohan$  cd
hduser@rohan-HP-205-G1-AiO-Business-PC:~$ pwd
/home/hduser
hduser@rohan-HP-205-G1-AiO-Business-PC:~$ sudo mkdir Temperature
[sudo] password for hduser:
hduser@rohan-HP-205-G1-AiO-Business-PC:~$ ls
Temperature
hduser@rohan-HP-205-G1-AiO-Business-PC:~$ sudo chmod -R 777 Temperature/
hduser@rohan-HP-205-G1-AiO-Business-PC:~$ sudo chown -R hduser Temperature/
hduser@rohan-HP-205-G1-AiO-Business-PC:~$ cd Temperature
hduser@rohan-HP-205-G1-AiO-Business-PC:~/Temperature$ pwd
/home/hduser/Temperature
hduser@rohan-HP-205-G1-AiO-Business-PC:~/Temperature$ sudo cp -R
/home/rohan/Desktop/Assignment2/* ./
hduser@rohan-HP-205-G1-AiO-Business-PC:~/Temperature$ ls
hottestncoolest.txt  input_dataset  MaxTemperatureDriver.java  MaxTemperatureMapper.java
MaxTemperatureReducer.java
hduser@rohan-HP-205-G1-AiO-Business-PC:~/Temperature$ ls -ltr
total 20
-rw-r--r-- 1 root root 1870 May 19 14:40 hottestncoolest.txt
drwxr-xr-x 2 root root 4096 May 19 14:40 input_dataset
-rw-r--r-- 1 root root 1431 May 19 14:40 MaxTemperatureDriver.java
-rw-r--r-- 1 root root  561 May 19 14:40 MaxTemperatureReducer.java
-rw-r--r-- 1 root root  942 May 19 14:40 MaxTemperatureMapper.java
hduser@rohan-HP-205-G1-AiO-Business-PC:~/Temperature$ sudo chmod +r *.*
hduser@rohan-HP-205-G1-AiO-Business-PC:~/Temperature$ export
CLASSPATH="$HADOOP_HOME/share/hadoop/mapreduce/hadoop-mapreduce-client-core-
2.10.1.jar:$HADOOP_HOME/share/hadoop/mapreduce/hadoop-mapreduce-client-common-
2.10.1.jar:$HADOOP_HOME/share/hadoop/common/hadoop-common-
2.10.1.jar:~/Temperature/MaxMinTemp/*:$HADOOP_HOME/lib/*"
hduser@rohan-HP-205-G1-AiO-Business-PC:~/Temperature$ javac -d . MaxTemperatureMapper.java
MaxTemperatureReducer.java MaxTemperatureDriver.java
Note: MaxTemperatureDriver.java uses or overrides a deprecated API.
Note: Recompile with -Xlint:deprecation for details.
hduser@rohan-HP-205-G1-AiO-Business-PC:~/Temperature$ ls
hottestncoolest.txt  input_dataset  MaxMinTemp  MaxTemperatureDriver.java
MaxTemperatureMapper.java  MaxTemperatureReducer.java
hduser@rohan-HP-205-G1-AiO-Business-PC:~/Temperature$ cd MaxMinTemp
hduser@rohan-HP-205-G1-AiO-Business-PC:~/Temperature/MaxMinTemp$ ls
MaxTemperatureDriver.class  MaxTemperatureMapper.class  MaxTemperatureReducer.class
hduser@rohan-HP-205-G1-AiO-Business-PC:~/Temperature/MaxMinTemp$ cd ..
hduser@rohan-HP-205-G1-AiO-Business-PC:~/Temperature$ sudo gedit Manifest.txt

(gedit:2101): Tepl-WARNING **: 14:46:15.855: GVfs metadata is not supported. Fallback to
TeplMetadataManager. Either GVfs is not correctly installed or GVfs metadata are not supported on this
platform. In the latter case, you should configure Tepl with --disable-gvfs-metadata.
hduser@rohan-HP-205-G1-AiO-Business-PC:~/Temperature$ jar -cfm temperature.jar Manifest.txt
MaxMinTemp/*.class
hduser@rohan-HP-205-G1-AiO-Business-PC:~/Temperature$ ls
hottestncoolest.txt  Manifest.txt  MaxTemperatureDriver.java  MaxTemperatureReducer.java
input_dataset        MaxMinTemp    MaxTemperatureMapper.java  temperature.jar
```

Name: Rohan S Kadu PRN no:71901492B Roll no: T1851004

```
hduser@rohan-HP-205-G1-AiO-Business-PC:~/Temperature$ ls
hottestncoolest.txt  Manifest.txt  MaxTemperatureDriver.java  MaxTemperatureReducer.java
input_dataset        MaxMinTemp    MaxTemperatureMapper.java  temperature.jar
hduser@rohan-HP-205-G1-AiO-Business-PC:~/Temperature$ cd
hduser@rohan-HP-205-G1-AiO-Business-PC:~$ start-dfs.sh
21/05/19 14:53:26 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your
platform... using builtin-java classes where applicable
Starting namenodes on [localhost]
Enter passphrase for key '/home/hduser/.ssh/id_rsa':
hduser@localhost's password:
localhost: starting namenode, logging to /usr/local/hadoop/logs/hadoop-hduser-namenode-rohan-HP-205-
G1-AiO-Business-PC.out
Enter passphrase for key '/home/hduser/.ssh/id_rsa':
hduser@localhost's password:
localhost: starting datanode, logging to /usr/local/hadoop/logs/hadoop-hduser-datanode-rohan-HP-205-G1-
AiO-Business-PC.out
Starting secondary namenodes [0.0.0.0]
Enter passphrase for key '/home/hduser/.ssh/id_rsa':
hduser@0.0.0.0's password:
0.0.0.0: starting secondarynamenode, logging to /usr/local/hadoop/logs/hadoop-hduser-
secondarynamenode-rohan-HP-205-G1-AiO-Business-PC.out
21/05/19 14:54:23 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your
platform... using builtin-java classes where applicable
hduser@rohan-HP-205-G1-AiO-Business-PC:~$ start-yarn.sh
starting yarn daemons
starting resourcemanager, logging to /usr/local/hadoop/logs/yarn-hduser-resourcemanager-rohan-HP-205-
G1-AiO-Business-PC.out
Enter passphrase for key '/home/hduser/.ssh/id_rsa':
hduser@localhost's password:
localhost: starting nodemanager, logging to /usr/local/hadoop/logs/yarn-hduser-nodemanager-rohan-HP-
205-G1-AiO-Business-PC.out
hduser@rohan-HP-205-G1-AiO-Business-PC:~$ jps
2880 SecondaryNameNode
3013 ResourceManager
2645 DataNode
3147 NodeManager
3261 Jps
2510 NameNode
hduser@rohan-HP-205-G1-AiO-Business-PC:~$ cd Temperature
hduser@rohan-HP-205-G1-AiO-Business-PC:~/Temperature$ sudo chmod -R 777 input-dataset/
chmod: cannot access 'input-dataset/': No such file or directory
hduser@rohan-HP-205-G1-AiO-Business-PC:~/Temperature$ cd
hduser@rohan-HP-205-G1-AiO-Business-PC:~$ sudo chmod -R 777 input-dataset/
chmod: cannot access 'input-dataset/': No such file or directory
hduser@rohan-HP-205-G1-AiO-Business-PC:~$ cd Temperature
hduser@rohan-HP-205-G1-AiO-Business-PC:~/Temperature$ ls
hottestncoolest.txt  Manifest.txt  MaxTemperatureDriver.java  MaxTemperatureReducer.java
input_dataset        MaxMinTemp    MaxTemperatureMapper.java  temperature.jar
hduser@rohan-HP-205-G1-AiO-Business-PC:~/Temperature$ sudo chmod -R 777 input_dataset/
hduser@rohan-HP-205-G1-AiO-Business-PC:~/Temperature$ cd
hduser@rohan-HP-205-G1-AiO-Business-PC:~$ HADOOP_HOME/bin/hdfs dfs -put
/home/hduser/Temperature/input_dataset /
bash: HADOOP_HOME/bin/hdfs: No such file or directory
```

Name: Rohan S Kadu PRN no:71901492B Roll no: T1851004

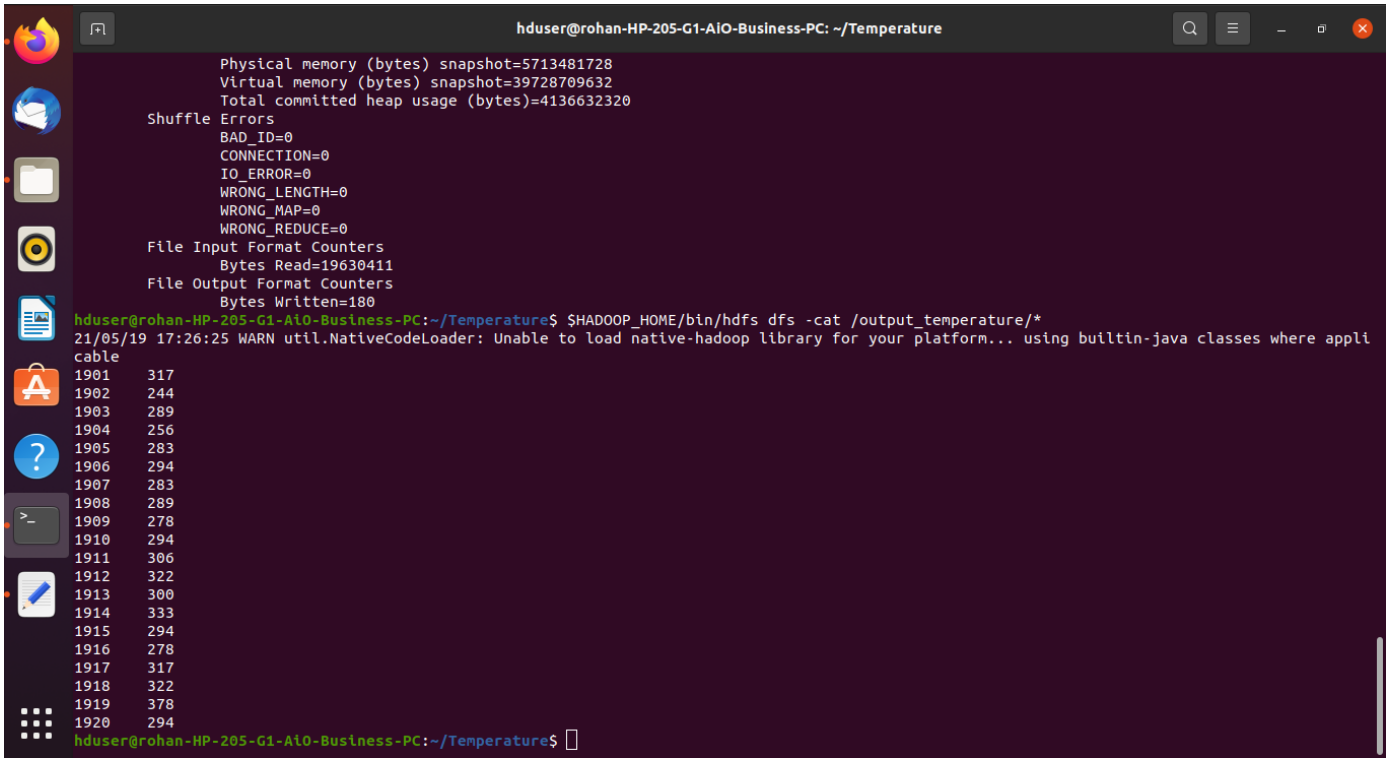
```
hduser@rohan-HP-205-G1-AiO-Business-PC:~$ $HADOOP_HOME/bin/hdfs dfs -put
/home/hduser/Temperature/input_dataset /
21/05/19 15:02:26 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your
platform... using builtin-java classes where applicable
hduser@rohan-HP-205-G1-AiO-Business-PC:~$ cd Temperature
hduser@rohan-HP-205-G1-AiO-Business-PC:~/Temperature$ $HADOOP_HOME/bin/hadoop jar temp.jar
/input_dataset /output_temperature
JAR does not exist or is not a normal file: /home/hduser/Temperature/temp.jar
hduser@rohan-HP-205-G1-AiO-Business-PC:~/Temperature$ $HADOOP_HOME/bin/hadoop jar
temperature.jar /input_dataset /output_temperature
21/05/19 15:11:55 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your
platform... using builtin-java classes where applicable
21/05/19 15:11:58 INFO client.RMProxy: Connecting to ResourceManager at /0.0.0.0:8032
21/05/19 15:12:00 WARN mapreduce.JobResourceUploader: Hadoop command-line option parsing not
performed. Implement the Tool interface and execute your application with ToolRunner to remedy this.
21/05/19 15:12:02 INFO input.FileInputFormat: Total input files to process : 20
21/05/19 15:12:02 INFO mapreduce.JobSubmitter: number of splits:20
21/05/19 15:12:03 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_1621416303401_0001
21/05/19 15:12:04 INFO conf.Configuration: resource-types.xml not found
21/05/19 15:12:04 INFO resource.ResourceUtils: Unable to find 'resource-types.xml'.
21/05/19 15:12:04 INFO resource.ResourceUtils: Adding resource type - name = memory-mb, units = Mi,
type = COUNTABLE
21/05/19 15:12:04 INFO resource.ResourceUtils: Adding resource type - name = vcores, units = , type =
COUNTABLE
21/05/19 15:12:06 INFO impl.YarnClientImpl: Submitted application application_1621416303401_0001
21/05/19 15:12:06 INFO mapreduce.Job: The url to track the job: http://rohan-HP-205-G1-AiO-Business-
PC:8088/proxy/application_1621416303401_0001/
21/05/19 15:12:06 INFO mapreduce.Job: Running job: job_1621416303401_0001
21/05/19 15:12:33 INFO mapreduce.Job: Job job_1621416303401_0001 running in uber mode : false
21/05/19 15:12:33 INFO mapreduce.Job:  map 0% reduce 0%
21/05/19 15:14:11 INFO mapreduce.Job:  map 15% reduce 0%
21/05/19 15:14:14 INFO mapreduce.Job:  map 30% reduce 0%
21/05/19 15:15:42 INFO mapreduce.Job:  map 30% reduce 10%
21/05/19 15:15:43 INFO mapreduce.Job:  map 35% reduce 10%
21/05/19 15:15:44 INFO mapreduce.Job:  map 45% reduce 10%
21/05/19 15:15:45 INFO mapreduce.Job:  map 55% reduce 10%
21/05/19 15:15:49 INFO mapreduce.Job:  map 55% reduce 18%
21/05/19 15:16:45 INFO mapreduce.Job:  map 65% reduce 18%
21/05/19 15:16:46 INFO mapreduce.Job:  map 80% reduce 23%
21/05/19 15:16:53 INFO mapreduce.Job:  map 80% reduce 27%
21/05/19 15:17:37 INFO mapreduce.Job:  map 85% reduce 27%
21/05/19 15:17:38 INFO mapreduce.Job:  map 95% reduce 27%
21/05/19 15:17:39 INFO mapreduce.Job:  map 100% reduce 27%
21/05/19 15:17:43 INFO mapreduce.Job:  map 100% reduce 100%
21/05/19 15:17:48 INFO mapreduce.Job: Job job_1621416303401_0001 completed successfully
21/05/19 15:17:49 INFO mapreduce.Job: Counters: 50
        File System Counters
                FILE: Number of bytes read=1567044
                FILE: Number of bytes written=7509885
                FILE: Number of read operations=0
                FILE: Number of large read operations=0
                FILE: Number of write operations=0
                HDFS: Number of bytes read=19632511
                HDFS: Number of bytes written=180
```

HDFS: Number of read operations=63
HDFS: Number of large read operations=0
HDFS: Number of write operations=2
Job Counters
Killed map tasks=3
Launched map tasks=21
Launched reduce tasks=1
Data-local map tasks=21
Total time spent by all maps in occupied slots (ms)=1453825
Total time spent by all reduces in occupied slots (ms)=189085
Total time spent by all map tasks (ms)=1453825
Total time spent by all reduce tasks (ms)=189085
Total vcore-milliseconds taken by all map tasks=1453825
Total vcore-milliseconds taken by all reduce tasks=189085
Total megabyte-milliseconds taken by all map tasks=1488716800
Total megabyte-milliseconds taken by all reduce tasks=193623040
Map-Reduce Framework
Map input records=142622
Map output records=142458
Map output bytes=1282122
Map output materialized bytes=1567158
Input split bytes=2100
Combine input records=0
Combine output records=0
Reduce input groups=20
Reduce shuffle bytes=1567158
Reduce input records=142458
Reduce output records=20
Spilled Records=284916
Shuffled Maps =20
Failed Shuffles=0
Merged Map outputs=20
GC time elapsed (ms)=28255
CPU time spent (ms)=78250
Physical memory (bytes) snapshot=5713481728
Virtual memory (bytes) snapshot=39728709632
Total committed heap usage (bytes)=4136632320
Shuffle Errors
BAD_ID=0
CONNECTION=0
IO_ERROR=0
WRONG_LENGTH=0
WRONG_MAP=0
WRONG_REDUCE=0
File Input Format Counters
Bytes Read=19630411
File Output Format Counters
Bytes Written=180

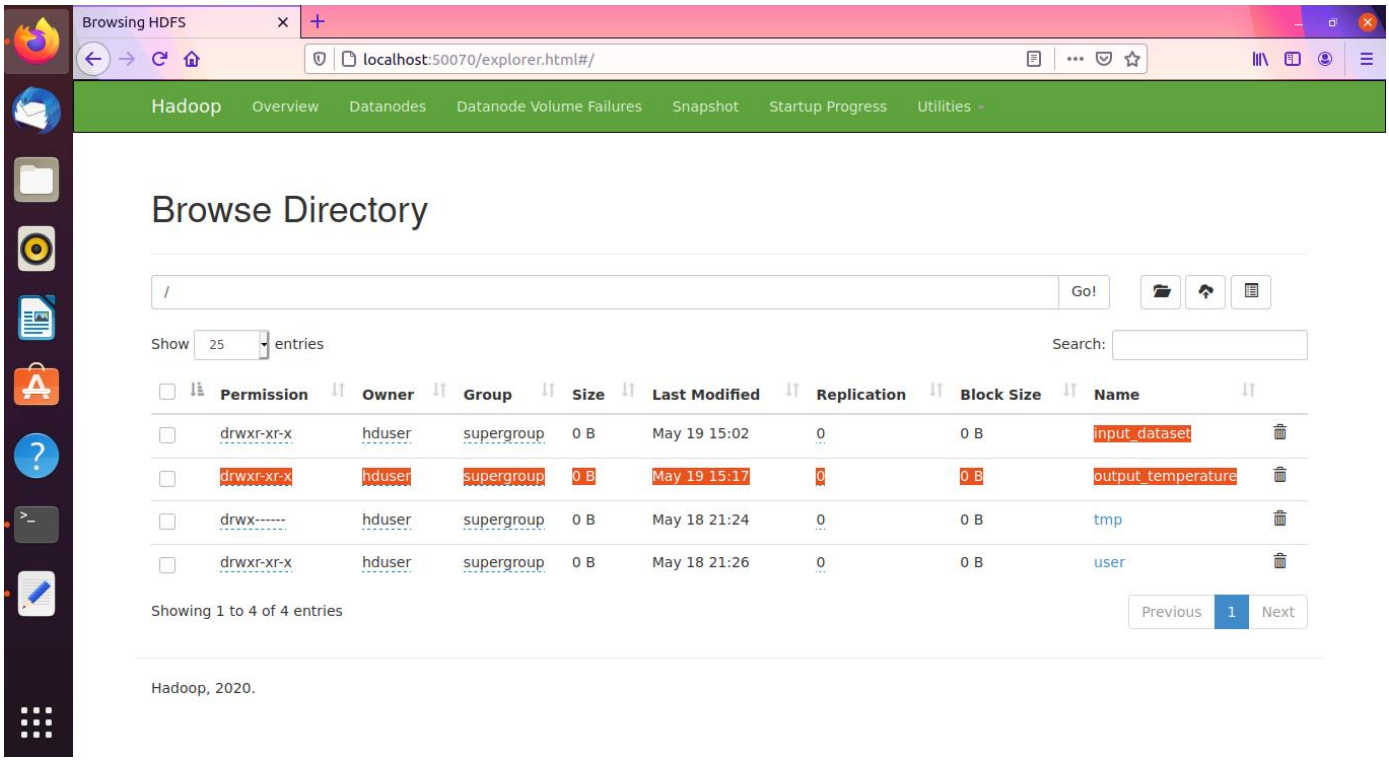
```
hduser@rohan-HP-205-G1-AiO-Business-PC:~/Temperature$ $HADOOP_HOME/bin/hdfs dfs -cat /output_temperature/*
21/05/19 17:26:25 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applicable
1901    317
1902    244
```

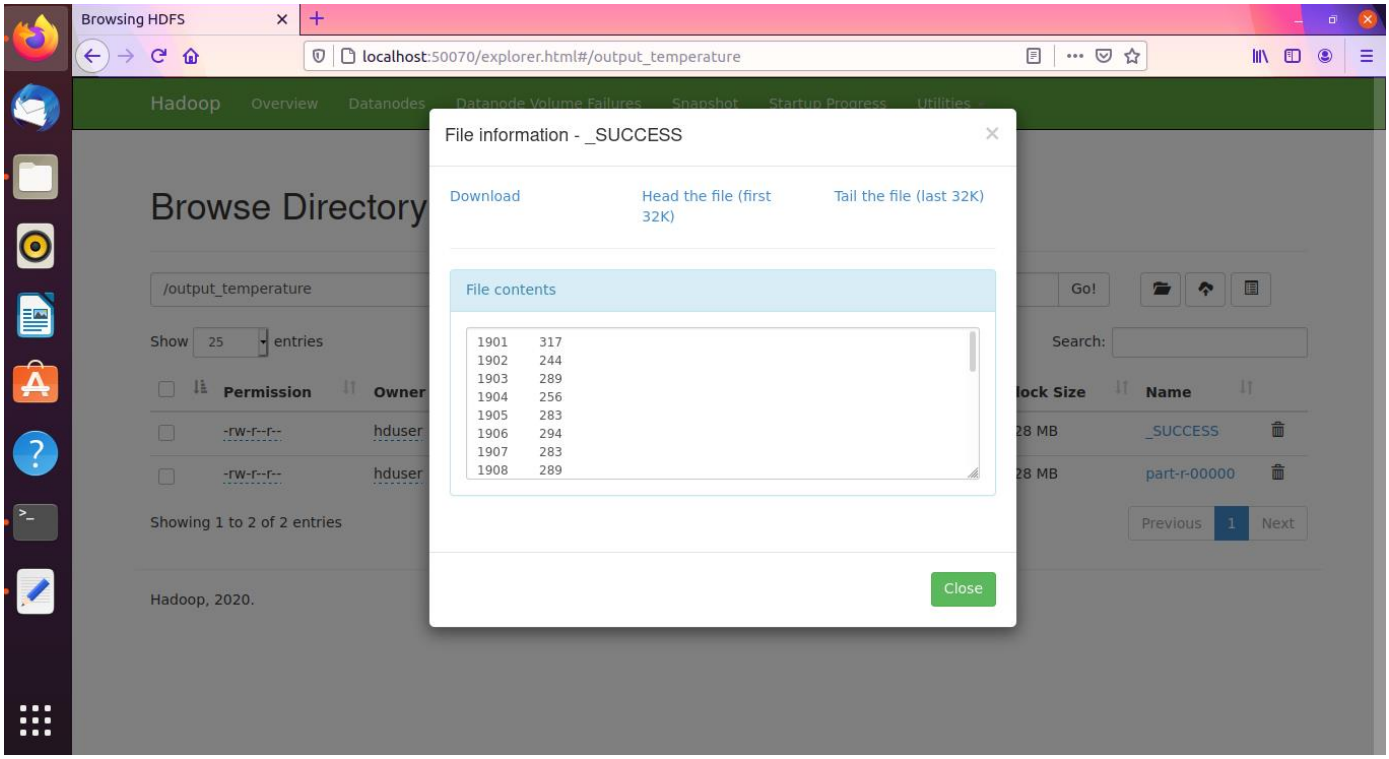
Name: Rohan S Kadu PRN no:71901492B Roll no: T1851004

1903 289
1904 256
1905 283
1906 294
1907 283
1908 289
1909 278
1910 294
1911 306
1912 322
1913 300
1914 333
1915 294
1916 278
1917 317
1918 322
1919 378
1920 294



Name: Rohan S Kadu PRN no:71901492B Roll no: T1851004





Conclusion: Thus we have learnt how to design a distributed application using MapReduce and process a log file of a system.