Principles Of Big Data Management(CSE 417L)

LAB RECORD

NAME: M.NAGARJUNA

REGISTER NO: AP20110010026

BRANCH: CSE A

ASSIGNMENT-1

PROGRAM:

Implementation of Word Count program using Map Reduce without combiner logic

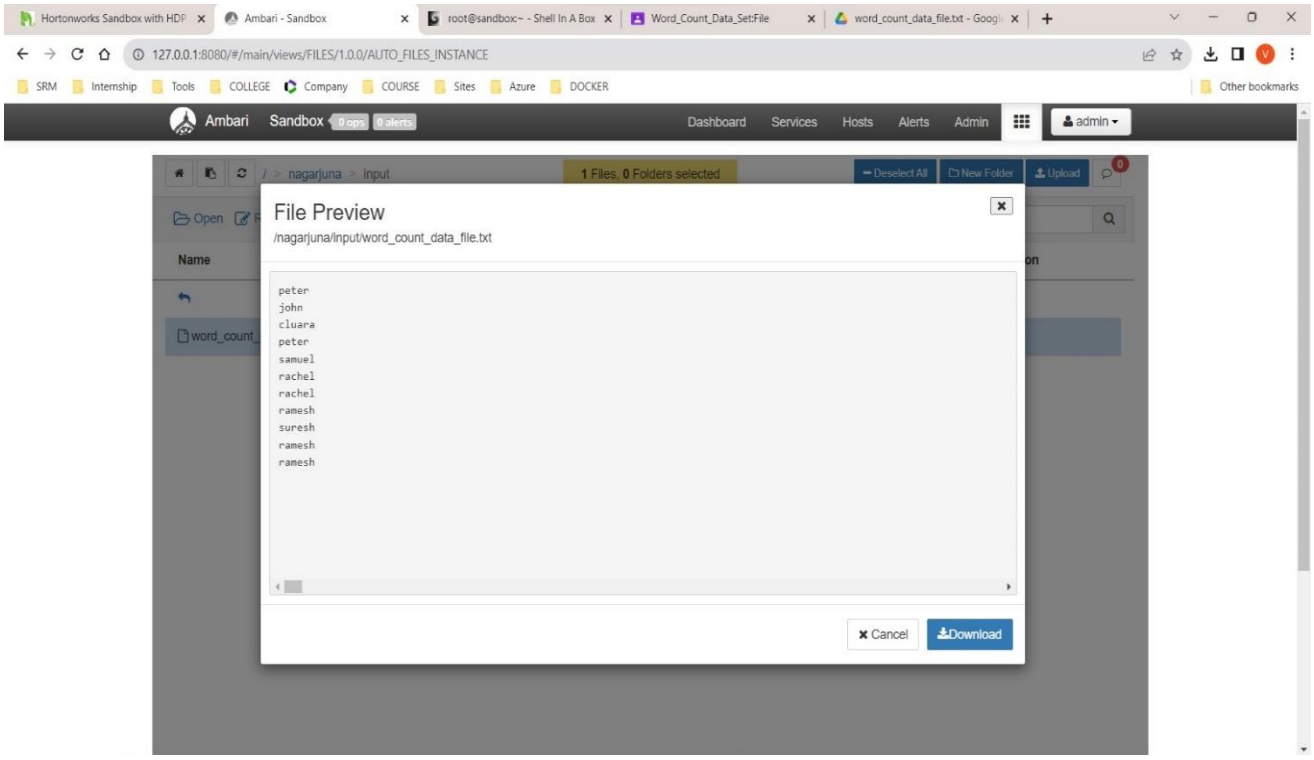
CODE:

```
import java.io.IOException;
import java.util.StringTokenizer;

import org.apache.hadoop.conf.Configuration;
import org.apache.hadoop.fs.Path;
import org.apache.hadoop.io.IntWritable;
import org.apache.hadoop.io.Text;
import org.apache.hadoop.mapreduce.Job;
import org.apache.hadoop.mapreduce.Mapper;
import org.apache.hadoop.mapreduce.Reducer;
import org.apache.hadoop.mapreduce.lib.input.FileInputFormat;
import org.apache.hadoop.mapreduce.lib.output.FileOutputFormat;

public class WordcountExample {

  public static class TokenizerMapper
       extends Mapper<Object, Text, Text, IntWritable>{

    private final static IntWritable one = new IntWritable(1);
    private Text word = new Text();

    public void map(Object key, Text value, Context context
                    ) throws IOException, InterruptedException {
      StringTokenizer itr = new StringTokenizer(value.toString());
```

```
    while (itr.hasMoreTokens()) {
      word.set(itr.nextToken());
      context.write(word, one);
    }
  }
}

public static class IntSumReducer
     extends Reducer<Text,IntWritable,Text,IntWritable> {
  private IntWritable result = new IntWritable();

  public void reduce(Text key, Iterable<IntWritable> values,
                     Context context
                     ) throws IOException, InterruptedException {
    int sum = 0;
    for (IntWritable val : values) {
      sum += val.get();
    }
    result.set(sum);
    context.write(key, result);
  }
}

public static void main(String[] args) throws Exception {
  Configuration conf = new Configuration();
  Job job = Job.getInstance(conf, "word count");
  job.setJarByClass(WordcountExample.class);
  job.setMapperClass(TokenizerMapper.class);
  job.setReducerClass(IntSumReducer.class);
  job.setOutputKeyClass(Text.class);
  job.setOutputValueClass(IntWritable.class);
```

```
    FileInputFormat.addInputPath(job, new Path(args[0]));
    FileOutputFormat.setOutputPath(job, new Path(args[1]));
    System.exit(job.waitForCompletion(true) ? 0 : 1);
  }
}
```

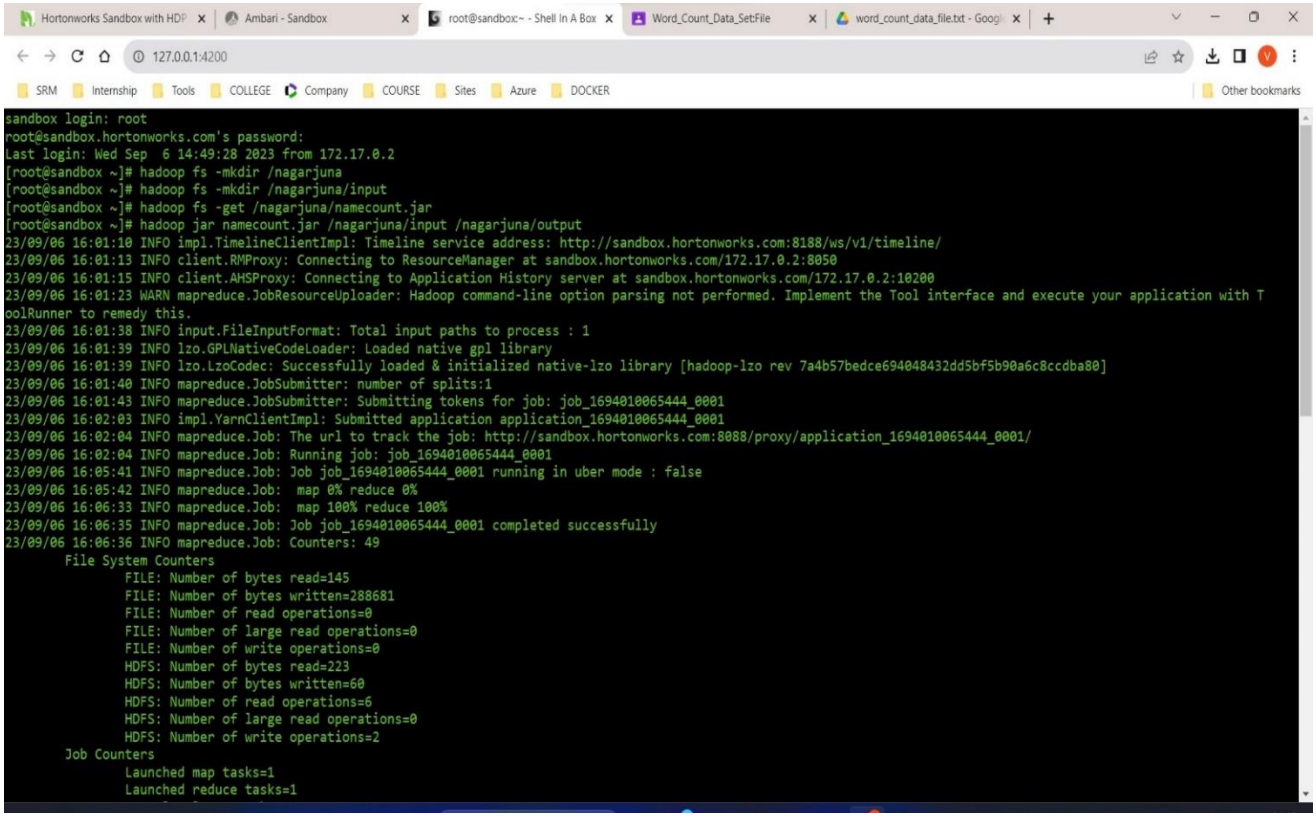
INPUT:

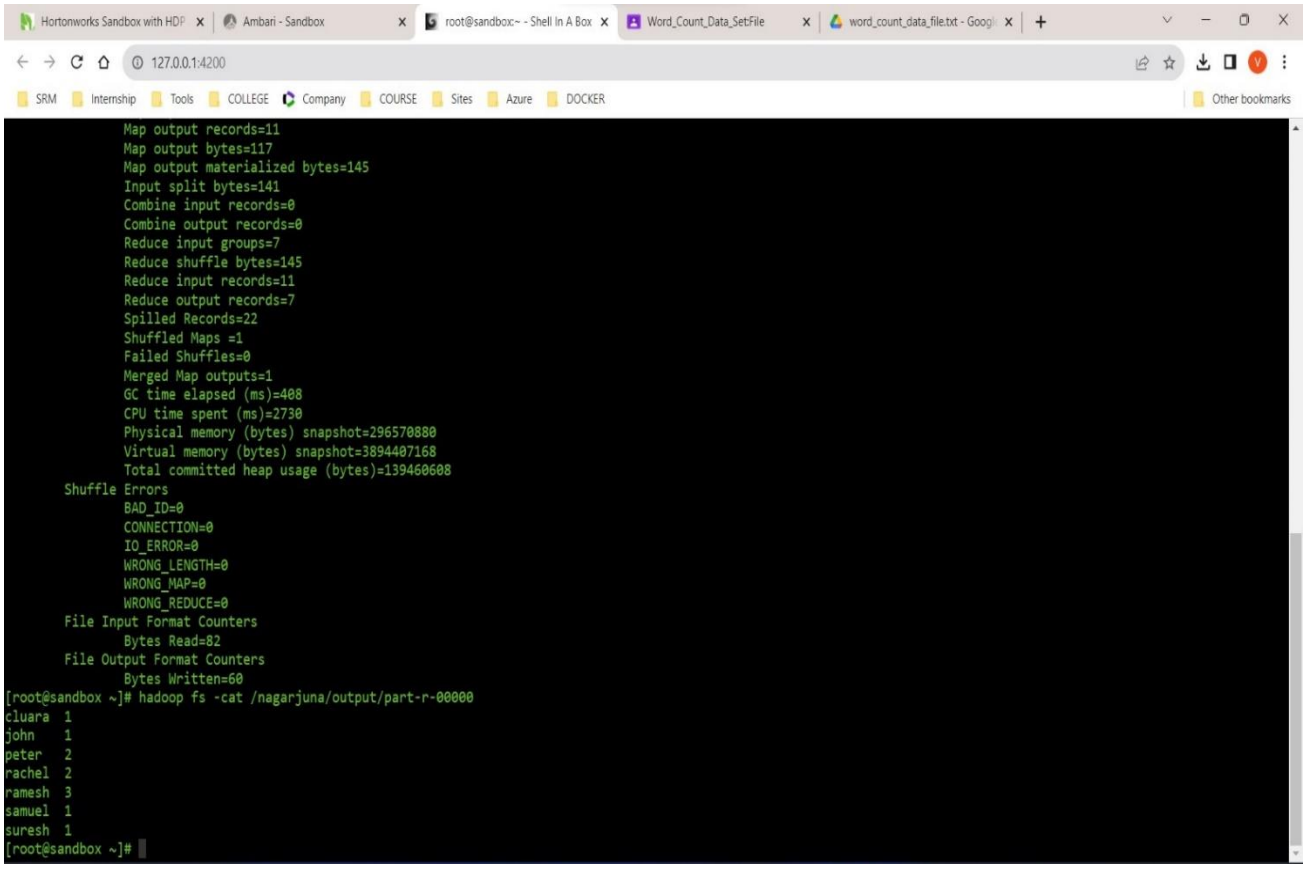


COMMANDS:

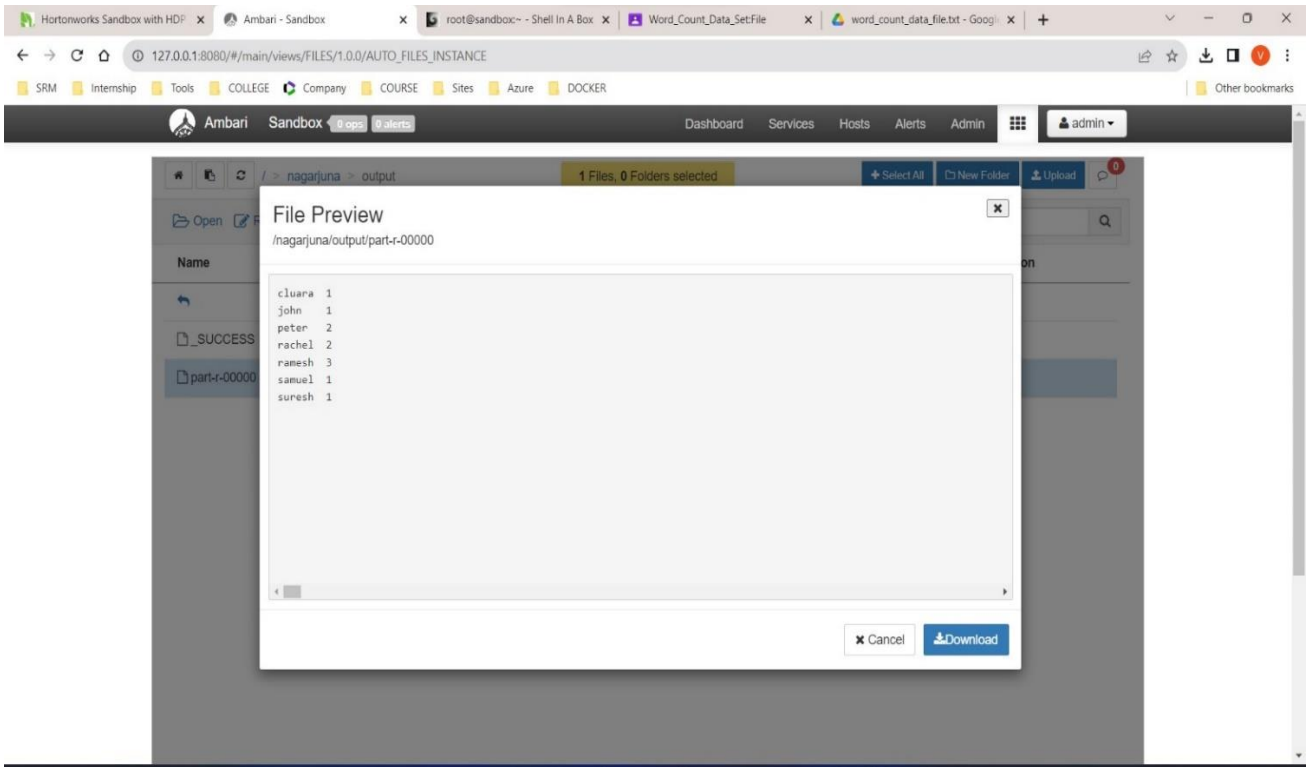
- hadoop fs -mkdir /nagarjuna
- hadoop fs -mkdir /nagarjuna/input
- hadoop fs -get /nagarjuna/namecount.jar
- hadoop jar namecount.jar /nagarjuna/input /nagarjuna/output
- hadoop fs -cat /nagarjuna/output/part-r-00000

EXECUTION:


Hortonworks Sandbox with HDP
Ambari - Sandbox
root@sandbox:~ - Shell In A Box
Word_Count_Data_Set:File
word_count_data_file.txt - Google
127.0.0.1:4200
SRM
Internship
Tools
COLLEGE
Company
COURSE
Sites
Azure
DOCKER
Other bookmarks
sandbox login: root
root@sandbox.hortonworks.com's password:
Last login: Wed Sep 6 14:49:28 2023 from 172.17.0.2
[root@sandbox ~]# hadoop fs -mkdir /nagarjuna
[root@sandbox ~]# hadoop fs -mkdir /nagarjuna/input
[root@sandbox ~]# hadoop fs -get /nagarjuna/namecount.jar
[root@sandbox ~]# hadoop jar namecount.jar /nagarjuna/input /nagarjuna/output
23/09/06 16:01:10 INFO impl.TimelineClientImpl: Timeline service address: http://sandbox.hortonworks.com:8188/ws/v1/timeline/
23/09/06 16:01:13 INFO client.RMProxy: Connecting to ResourceManager at sandbox.hortonworks.com/172.17.0.2:8050
23/09/06 16:01:15 INFO client.AHSProxy: Connecting to Application History server at sandbox.hortonworks.com/172.17.0.2:10200
23/09/06 16:01:23 WARN mapreduce.JobResourceUploader: Hadoop command-line option parsing not performed. Implement the Tool interface and execute your application with T
oolRunner to remedy this.
23/09/06 16:01:38 INFO input.FileInputFormat: Total input paths to process : 1
23/09/06 16:01:39 INFO lzo.GPLNativeCodeLoader: Loaded native gpl library
23/09/06 16:01:39 INFO lzo.LzoCodec: Successfully loaded & initialized native-lzo library [hadoop-lzo rev 7a4b57bedce694048432dd5bf5b90a6c8ccdba80]
23/09/06 16:01:40 INFO mapreduce.JobSubmitter: number of splits:1
23/09/06 16:01:43 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_1694010065444_0001
23/09/06 16:02:03 INFO impl.YarnClientImpl: Submitted application application_1694010065444_0001
23/09/06 16:02:04 INFO mapreduce.Job: The url to track the job: http://sandbox.hortonworks.com:8088/proxy/application_1694010065444_0001/
23/09/06 16:02:04 INFO mapreduce.Job: Running job: job_1694010065444_0001
23/09/06 16:05:41 INFO mapreduce.Job: Job job_1694010065444_0001 running in uber mode : false
23/09/06 16:05:42 INFO mapreduce.Job: map 0% reduce 0%
23/09/06 16:06:33 INFO mapreduce.Job: map 100% reduce 100%
23/09/06 16:06:35 INFO mapreduce.Job: Job job_1694010065444_0001 completed successfully
23/09/06 16:06:36 INFO mapreduce.Job: Counters: 49
File System Counters
FILE: Number of bytes read=145
FILE: Number of bytes written=288681
FILE: Number of read operations=0
FILE: Number of large read operations=0
FILE: Number of write operations=0
HDFS: Number of bytes read=223
HDFS: Number of bytes written=60
HDFS: Number of read operations=6
HDFS: Number of large read operations=0
HDFS: Number of write operations=2
Job Counters
Launched map tasks=1
Launched reduce tasks=1

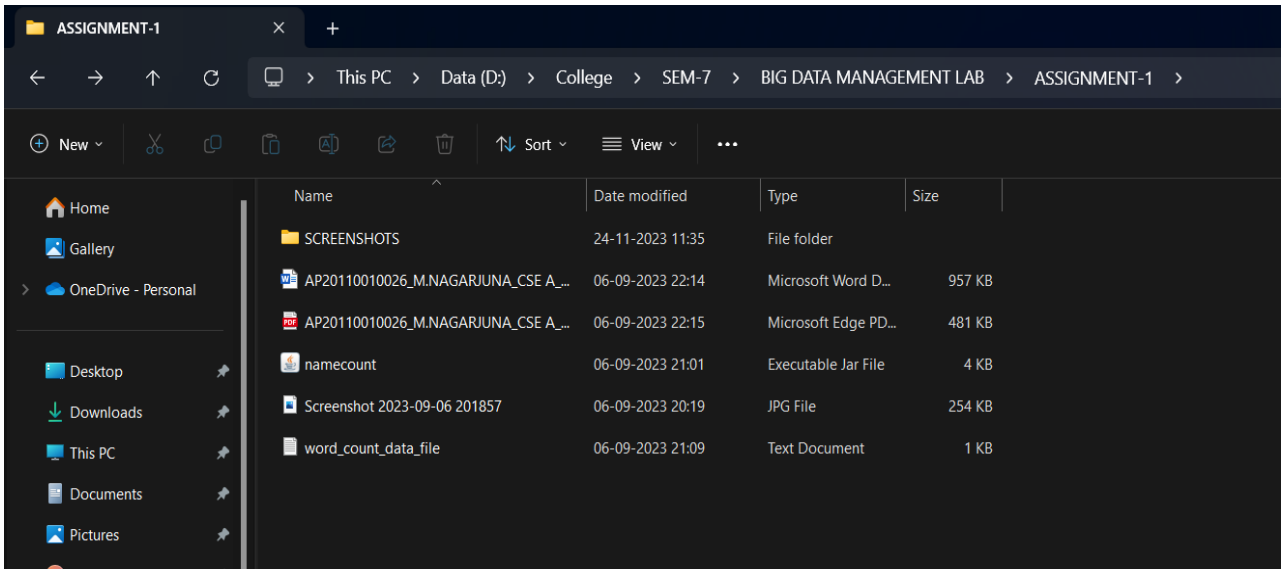

Hortonworks Sandbox with HDP
Ambari - Sandbox
root@sandbox:~ - Shell In A Box
Word_Count_Data_Set:File
word_count_data_file.txt - Google
127.0.0.1:4200
SRM
Internship
Tools
COLLEGE
Company
COURSE
Sites
Azure
DOCKER
Other bookmarks
Map output records=11
Map output bytes=117
Map output materialized bytes=145
Input split bytes=141
Combine input records=0
Combine output records=0
Reduce input groups=7
Reduce shuffle bytes=145
Reduce input records=11
Reduce output records=7
Spilled Records=22
Shuffled Maps =1
Failed Shuffles=0
Merged Map outputs=1
GC time elapsed (ms)=408
CPU time spent (ms)=2730
Physical memory (bytes) snapshot=296570880
Virtual memory (bytes) snapshot=3894407168
Total committed heap usage (bytes)=139460608
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=82
File Output Format Counters
Bytes Written=60
[root@sandbox ~]# hadoop fs -cat /nagarjuna/output/part-r-00000
cluara 1
john 1
peter 2
rachel 2
ramesh 3
samuel 1
suresh 1
[root@sandbox ~]#

OUTPUT:



PATH:

D:\College\SEM-7\BIG DATA MANAGEMENT LAB\ASSIGNMENT-1



ASSIGNMENT-2

PROGRAM:

Implementation of MapReduce algorithm for Matrix Multiplication

CODE:

MatrixMapper:

```
import java.io.IOException;
import org.apache.hadoop.conf.Configuration;
import org.apache.hadoop.io.LongWritable;
import org.apache.hadoop.io.Text;
import org.apache.hadoop.mapreduce.Mapper;
public class MatrixMapper extends Mapper<LongWritable, Text,
Text, Text> {
public void map(LongWritable key, Text value, Context
context) throws IOException, InterruptedException {
Configuration conf = context.getConfiguration();
int m = Integer.parseInt(conf.get("m"));
int p = Integer.parseInt(conf.get("p"));
String line = value.toString();
String[] indicesAndValue = line.split(",");
Text outputKey = new Text();
Text outputValue = new Text();
if (indicesAndValue[0].equals("A")) {
for (int k = 0; k < p; k++) {
outputKey.set(indicesAndValue[1] + "," + k);
outputValue.set("A," + indicesAndValue[2] + "," +
indicesAndValue[3]);
context.write(outputKey, outputValue);
}
} else {
for (int i = 0; i < m; i++) {
outputKey.set(i + "," + indicesAndValue[2]);
outputValue.set("B," + indicesAndValue[1] + "," +
indicesAndValue[3]);
context.write(outputKey, outputValue);
}
}
}
}
```

MatrixMultiplication:

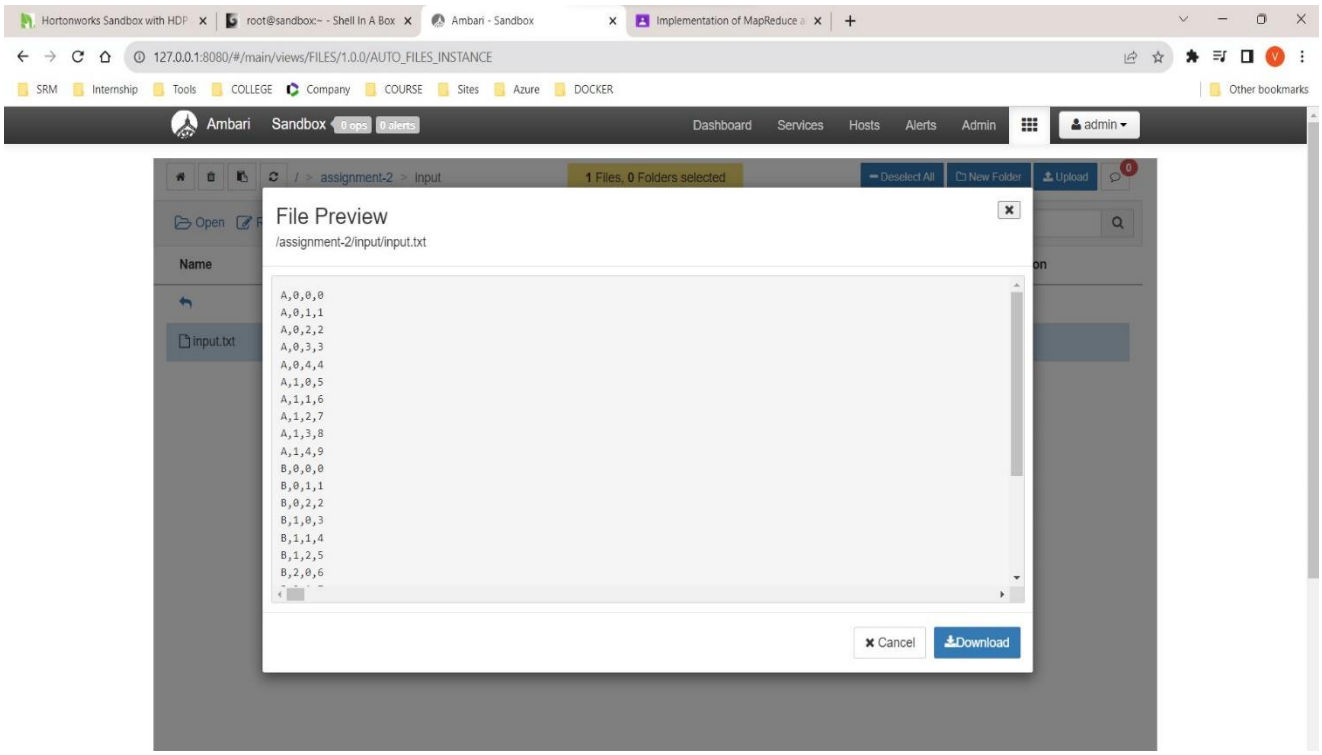
```
import org.apache.hadoop.fs.Path;
import org.apache.hadoop.conf.*;
import org.apache.hadoop.io.*;
import org.apache.hadoop.mapreduce.*;
import org.apache.hadoop.mapreduce.lib.input.FileInputFormat;
import org.apache.hadoop.mapreduce.lib.input.TextInputFormat;
import org.apache.hadoop.mapreduce.lib.output.FileOutputFormat;
import org.apache.hadoop.mapreduce.lib.output.TextOutputFormat;
public class MatrixMultiplication {
public static void main(String[] args) throws Exception {
Configuration conf = new Configuration();
// A is an m-by-n matrix; B is an n-by-p matrix.
conf.set("m", "2");
conf.set("n", "5");
conf.set("p", "3");
Job job = new Job(conf, "MatrixMultiplication");
job.setJarByClass(MatrixMultiplication.class);
job.setOutputKeyClass(Text.class);
job.setOutputValueClass(Text.class);
job.setMapperClass(MatrixMapper.class);
job.setReducerClass(MatrixReducer.class);
job.setInputFormatClass(TextInputFormat.class);
job.setOutputFormatClass(TextOutputFormat.class);
FileInputFormat.addInputPath(job, new Path(args[0]));
FileOutputFormat.setOutputPath(job, new Path(args[1]));
job.waitForCompletion(true);
}
}
```

MatrixReducer:

```
import java.io.IOException;
import java.util.HashMap;
import org.apache.hadoop.io.Text;
import org.apache.hadoop.mapreduce.Reducer;
public class MatrixReducer extends Reducer<Text, Text, Text, Text> {
public void reduce(Text key, Iterable<Text> values, Context
context) throws IOException, InterruptedException {
String[] value;
HashMap<Integer, Float> hashA = new HashMap<Integer,
Float>();
HashMap<Integer, Float> hashB = new HashMap<Integer,
Float>();
for (Text val : values) {
value = val.toString().split(",");
if (value[0].equals("A")) {
hashA.put(Integer.parseInt(value[1]),
Float.parseFloat(value[2]));
} else {
hashB.put(Integer.parseInt(value[1]),
```

```
Float.parseFloat(value[2]));
}
}
int n =
Integer.parseInt(context.getConfiguration().get("n"));
float result = 0.0f;
float a_ij;
float b_jk;
for (int j = 0; j < n; j++) {
a_ij = hashA.containsKey(j) ? hashA.get(j) : 0.0f;
b_jk = hashB.containsKey(j) ? hashB.get(j) : 0.0f;
result += a_ij * b_jk;
}
if (result != 0.0f) {
context.write(null, new Text(key.toString() + "," +
Float.toString(result)));
}
}
}
```

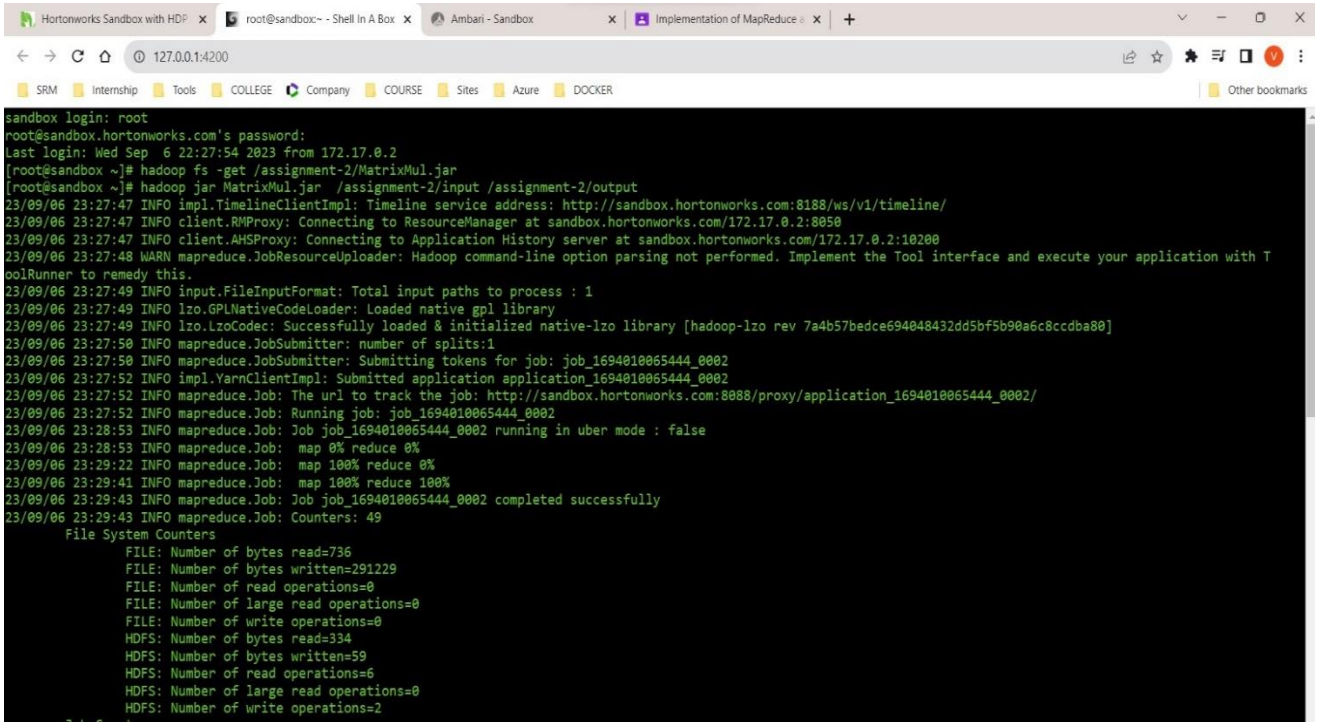
INPUT:

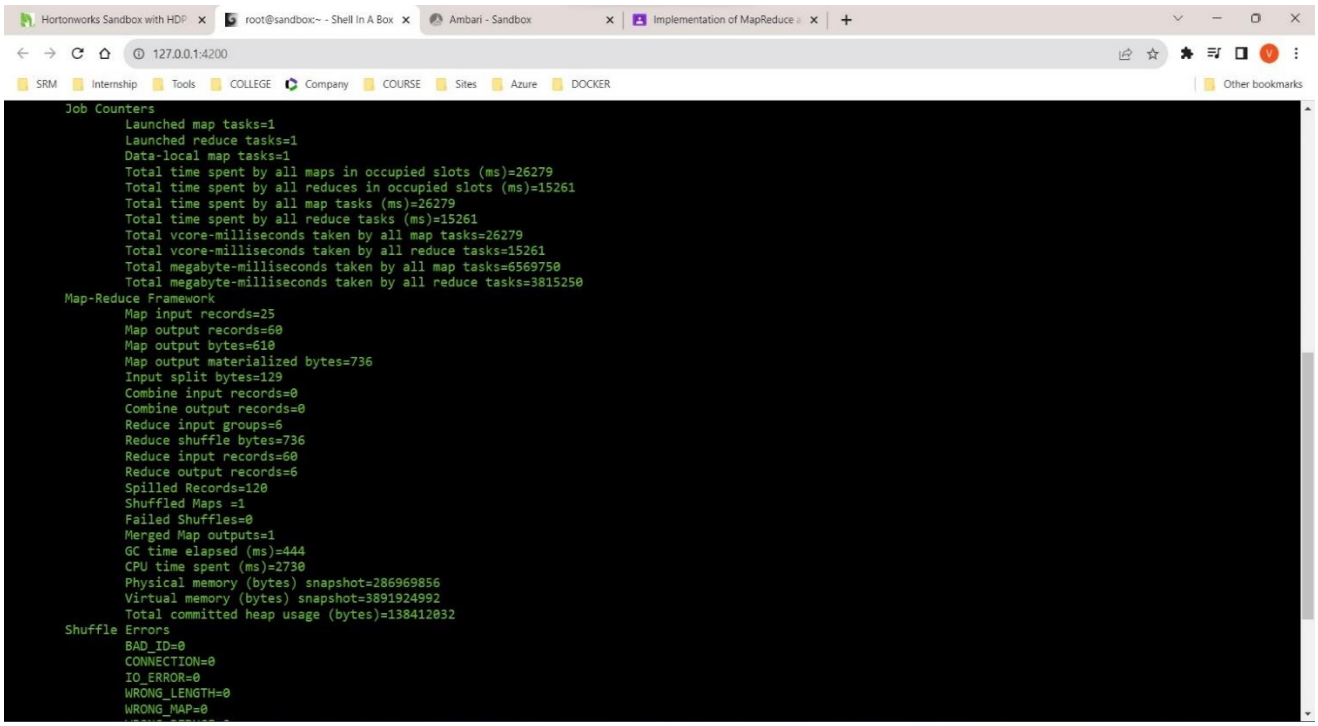


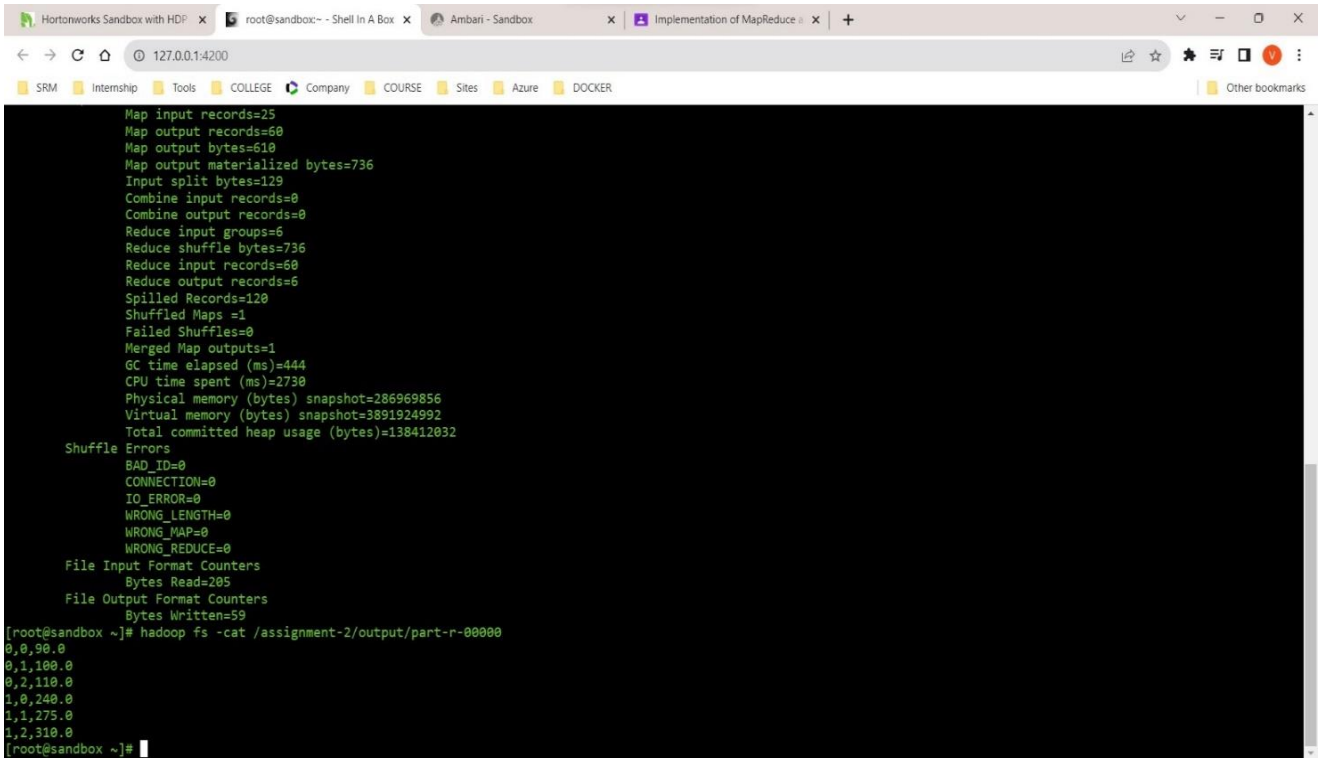
COMMAND:

- hadoop fs -get /assignment-2/MatrixMul.jar
- hadoop jar MatrixMul.jar /assignment-2/input /assignment-2/output
- hadoop fs -cat /assignment-2/output/part-r-00000

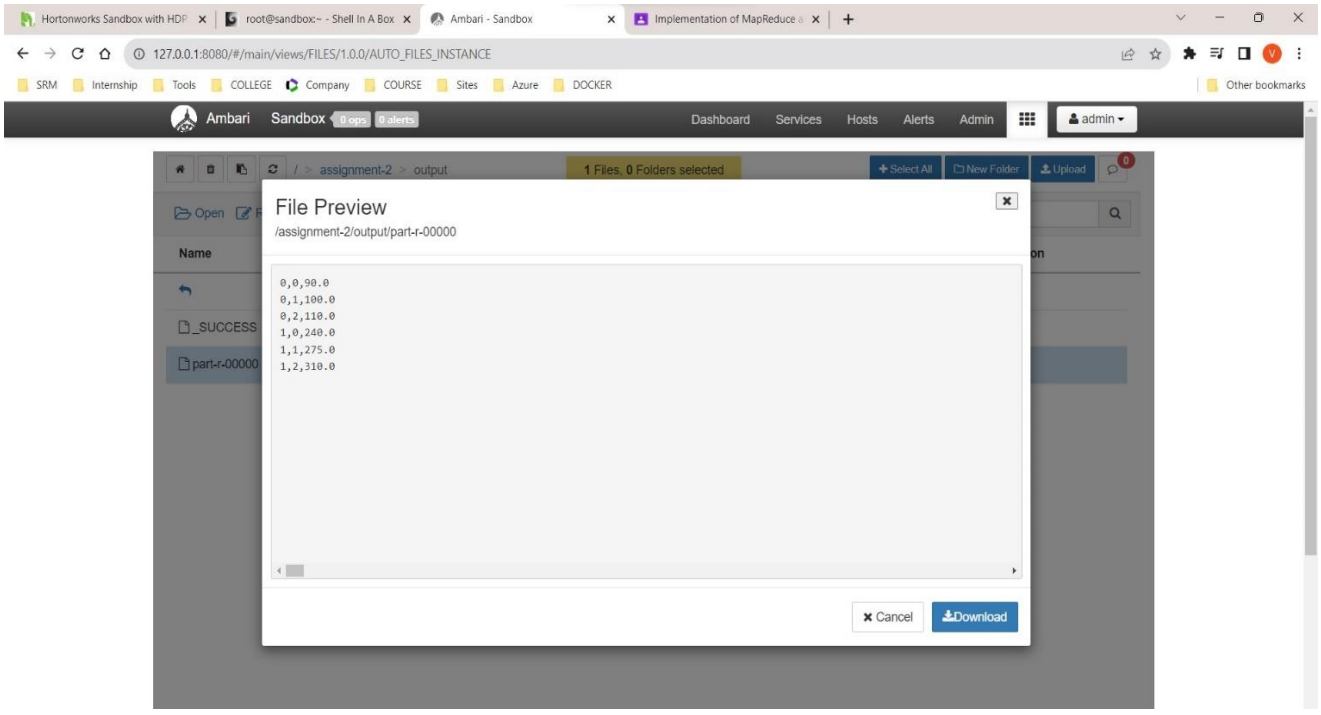
EXECUTION:





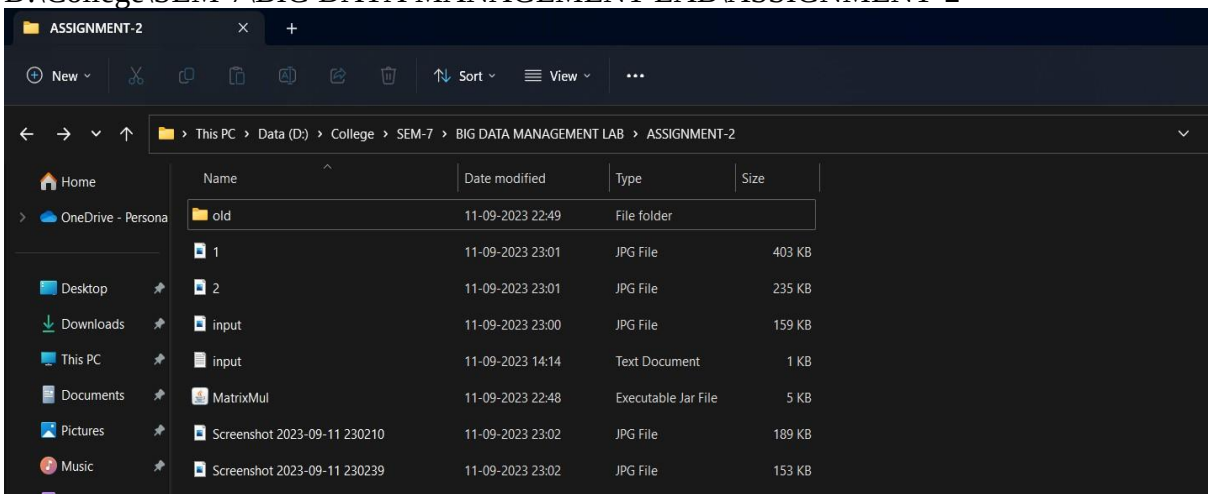

Hortonworks Sandbox with HDP
root@sandbox:~ - Shell In A Box
Ambari - Sandbox
Implementation of MapReduce
127.0.0.1:4200
SRM
Internship
Tools
COLLEGE
Company
COURSE
Sites
Azure
DOCKER
Other bookmarks
Map input records=25
Map output records=60
Map output bytes=610
Map output materialized bytes=736
Input split bytes=129
Combine input records=0
Combine output records=0
Reduce input groups=6
Reduce shuffle bytes=736
Reduce input records=60
Reduce output records=6
Spilled Records=120
Shuffled Maps =1
Failed Shuffles=0
Merged Map outputs=1
GC time elapsed (ms)=444
CPU time spent (ms)=2730
Physical memory (bytes) snapshot=286969856
Virtual memory (bytes) snapshot=3891924992
Total committed heap usage (bytes)=138412032
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=205
File Output Format Counters
Bytes Written=59
[root@sandbox ~]# hadoop fs -cat /assignment-2/output/part-r-00000
0,0,90.0
0,1,100.0
0,2,110.0
1,0,240.0
1,1,275.0
1,2,310.0
[root@sandbox ~]#

OUTPUT:


Hortonworks Sandbox with HDP
root@sandbox:~ - Shell In A Box
Ambari - Sandbox
Implementation of MapReduce
127.0.0.1:8080/#/main/views/FILES/1.0.0/AUTO_FILES_INSTANCE
SRM
Internship
Tools
COLLEGE
Company
COURSE
Sites
Azure
DOCKER
Other bookmarks
Ambari
Sandbox
0 ops
0 alerts
Dashboard
Services
Hosts
Alerts
Admin
admin
assignment-2
output
1 Files, 0 Folders selected
Select All
New Folder
Upload
Open
Name
_SUCCESS
part-r-00000
File Preview
/assignment-2/output/part-r-00000
0,0,90.0
0,1,100.0
0,2,110.0
1,0,240.0
1,1,275.0
1,2,310.0
Cancel
Download

PATH:

D:\College\SEM-7\BIG DATA MANAGEMENT LAB\ASSIGNMENT-2



ASSIGNMENT-3

PROGRAM:

Weather data analysis for analyzing hot and cold days using MapReduce.

CODES:

MaxTemperatureMapper:

```
import java.io.IOException;

import org.apache.hadoop.io.IntWritable;
import org.apache.hadoop.io.LongWritable;
import org.apache.hadoop.io.Text;
import org.apache.hadoop.mapreduce.Mapper;

public class MaxTemperatureMapper
    extends Mapper<LongWritable, Text, Text, IntWritable> {

  private static final int MISSING = 9999;

  @Override
  public void map(LongWritable key, Text value, Context context)
      throws IOException, InterruptedException {

    String line = value.toString();
    String year = line.substring(15, 19);
    int airTemperature;
    if (line.charAt(87) == '+') { // parseInt doesn't like leading plus signs
      airTemperature = Integer.parseInt(line.substring(88, 92));
    } else {
      airTemperature = Integer.parseInt(line.substring(87, 92));
    }
    String quality = line.substring(92, 93);
    if (airTemperature != MISSING && quality.matches("[01459]")) {
      context.write(new Text(year), new IntWritable(airTemperature));
    }
  }
}
```

MaxTemperatureReducer:

```
import java.io.IOException;

import org.apache.hadoop.io.IntWritable;
import org.apache.hadoop.io.Text;
import org.apache.hadoop.mapreduce.Reducer;

public class MaxTemperatureReducer
    extends Reducer<Text, IntWritable, Text, IntWritable> {
```

```
  @Override
  public void reduce(Text key, Iterable<IntWritable> values, Context context)
      throws IOException, InterruptedException {

    int maxValue = Integer.MIN_VALUE;
    for (IntWritable value : values) {
      maxValue = Math.max(maxValue, value.get());
    }
    context.write(key, new IntWritable(maxValue));
  }
}
```

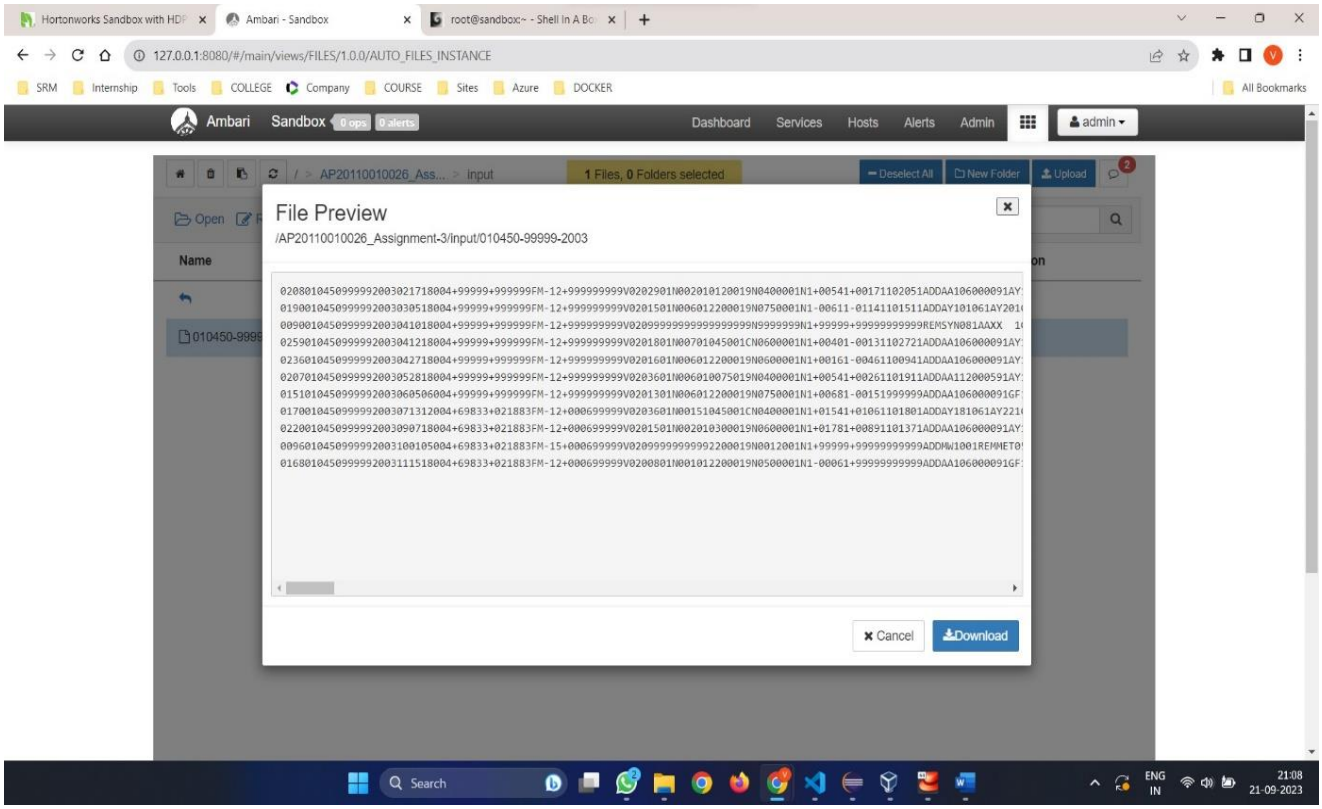
MaxTemperature:

```
import org.apache.hadoop.fs.Path;
import org.apache.hadoop.io.IntWritable;
import org.apache.hadoop.io.Text;
import org.apache.hadoop.mapreduce.Job;
import org.apache.hadoop.mapreduce.lib.input.FileInputFormat;
import org.apache.hadoop.mapreduce.lib.output.FileOutputFormat;

public class MaxTemperature {

  public static void main(String[] args) throws Exception {
    if (args.length != 2) {
      System.err.println("Usage: MaxTemperature <input path> <output path>");
      System.exit(-1);
    }

    Job job = new Job();
    job.setJarByClass(MaxTemperature.class);
    job.setJobName("Max temperature");
    FileInputFormat.addInputPath(job, new Path(args[0]));
    FileOutputFormat.setOutputPath(job, new Path(args[1]));

    job.setMapperClass(MaxTemperatureMapper.class);
    job.setReducerClass(MaxTemperatureReducer.class);

    job.setOutputKeyClass(Text.class);
    job.setOutputValueClass(IntWritable.class);

    System.exit(job.waitForCompletion(true) ? 0 : 1);
  }
}
```

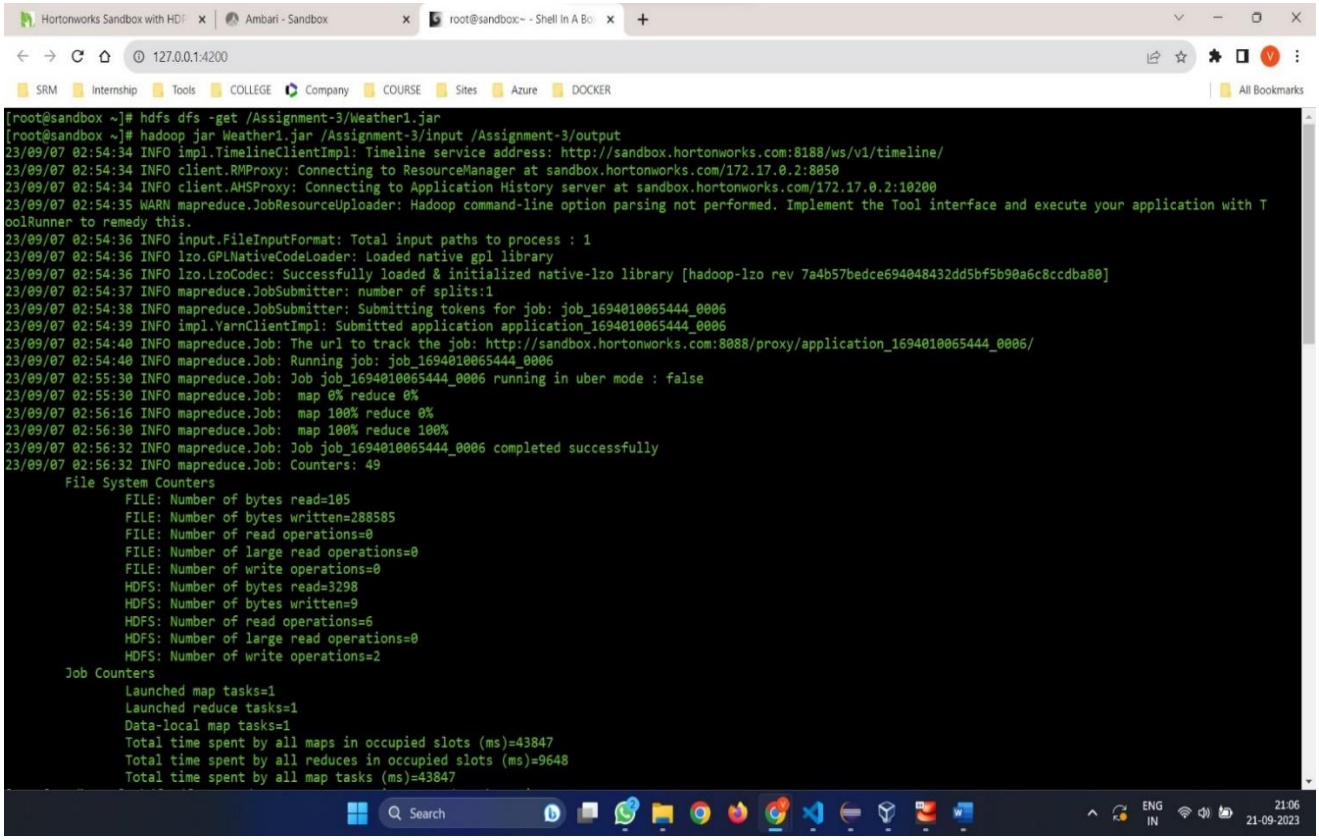
COMMANDS:

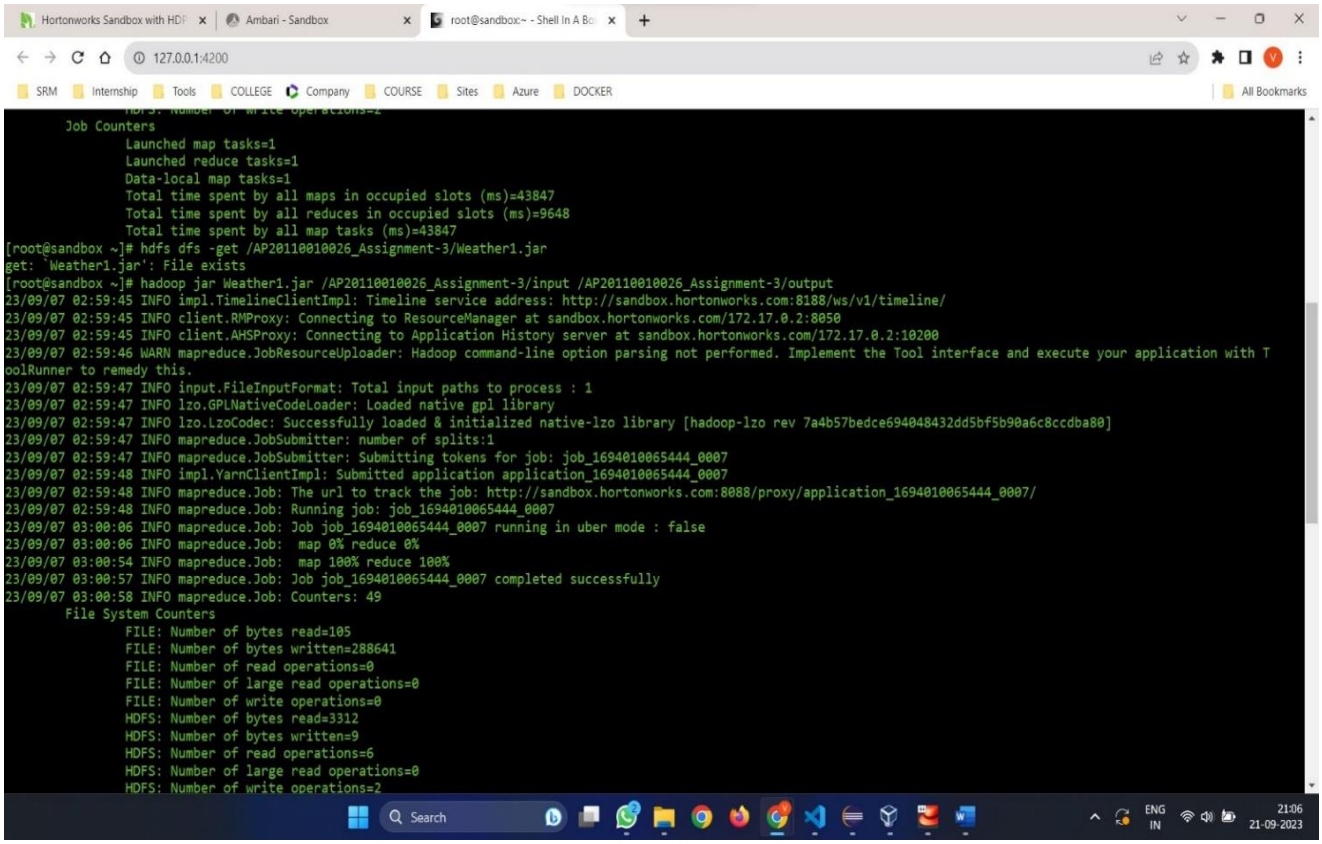
- hdfs dfs -get /Assignment-3/Weather1.jar
- hadoop jar Weather1.jar /Assignment-3/input /Assignment-3/output

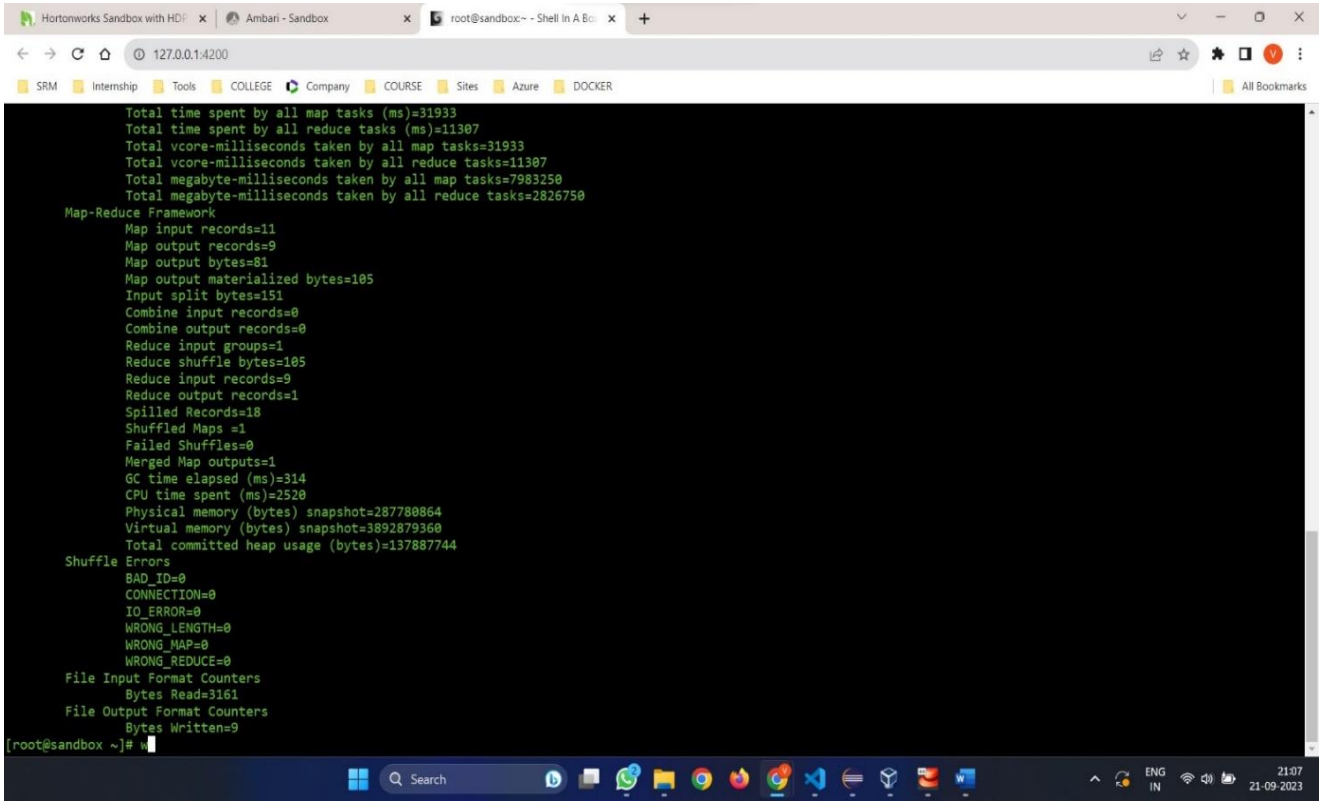
INPUT:



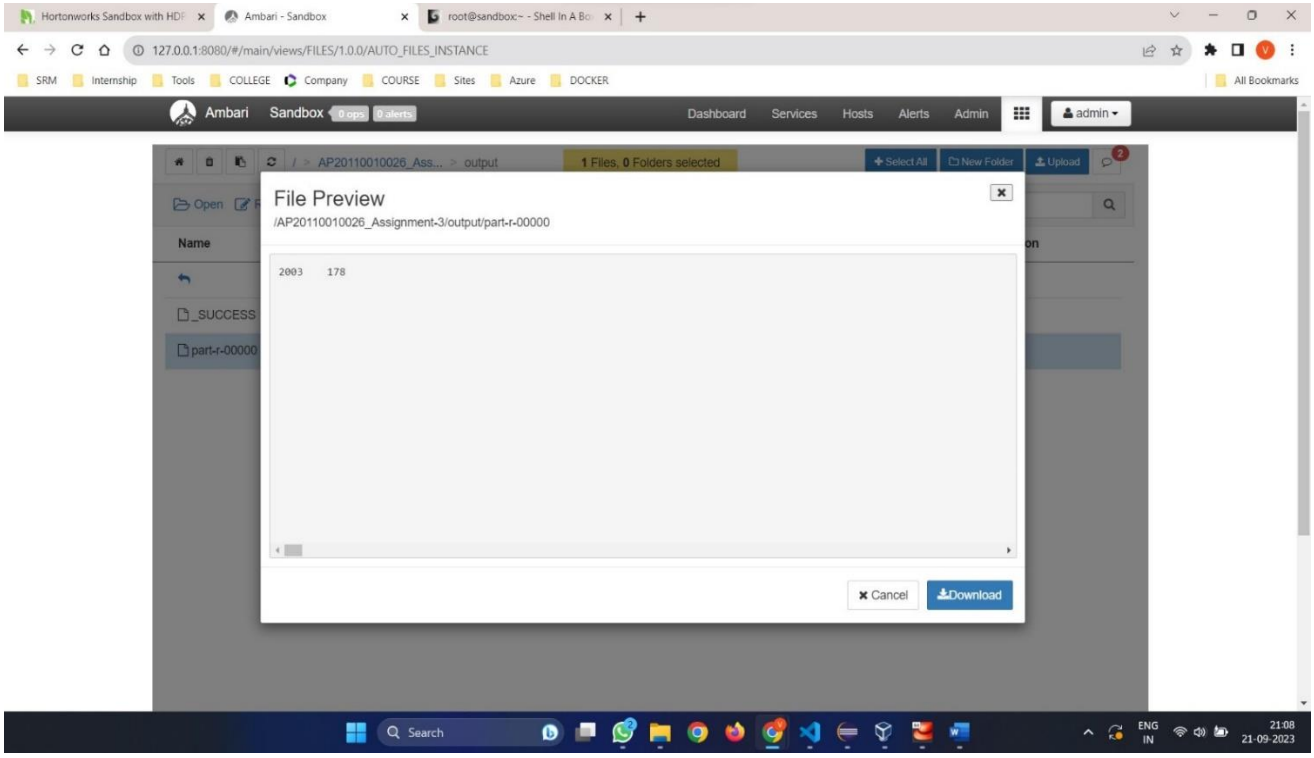
EXECUTION:




Hortonworks Sandbox with HDF
Ambari - Sandbox
root@sandbox:~ - Shell In A Box
127.0.0.1:4200
SRM
Internship
Tools
COLLEGE
Company
COURSE
Sites
Azure
DOCKER
All Bookmarks
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)=43847
Total time spent by all reduces in occupied slots (ms)=9648
Total time spent by all map tasks (ms)=43847
[root@sandbox ~]# hdfs dfs -get /AP20110010026_Assignment-3/Weather1.jar
get: `Weather1.jar': File exists
[root@sandbox ~]# hadoop jar Weather1.jar /AP20110010026_Assignment-3/input /AP20110010026_Assignment-3/output
23/09/07 02:59:45 INFO impl.TimelineClientImpl: Timeline service address: http://sandbox.hortonworks.com:8188/ws/v1/timeline/
23/09/07 02:59:45 INFO client.RMProxy: Connecting to ResourceManager at sandbox.hortonworks.com/172.17.0.2:8050
23/09/07 02:59:45 INFO client.AHSProxy: Connecting to Application History server at sandbox.hortonworks.com/172.17.0.2:10200
23/09/07 02:59:46 WARN mapreduce.JobResourceUploader: Hadoop command-line option parsing not performed. Implement the Tool interface and execute your application with T
oolRunner to remedy this.
23/09/07 02:59:47 INFO input.FileInputFormat: Total input paths to process : 1
23/09/07 02:59:47 INFO lzo.GPLNativeCodeLoader: Loaded native gpl library
23/09/07 02:59:47 INFO lzo.LzoCodec: Successfully loaded & initialized native-lzo library [hadoop-lzo rev 7a4b57bedce694048432dd5bf5b90a6c8ccdba80]
23/09/07 02:59:47 INFO mapreduce.JobSubmitter: number of splits:1
23/09/07 02:59:47 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_1694010065444_0007
23/09/07 02:59:48 INFO impl.YarnClientImpl: Submitted application application_1694010065444_0007
23/09/07 02:59:48 INFO mapreduce.Job: The url to track the job: http://sandbox.hortonworks.com:8088/proxy/application_1694010065444_0007/
23/09/07 02:59:48 INFO mapreduce.Job: Running job: job_1694010065444_0007
23/09/07 03:00:06 INFO mapreduce.Job: Job job_1694010065444_0007 running in uber mode : false
23/09/07 03:00:06 INFO mapreduce.Job: map 0% reduce 0%
23/09/07 03:00:54 INFO mapreduce.Job: map 100% reduce 100%
23/09/07 03:00:57 INFO mapreduce.Job: Job job_1694010065444_0007 completed successfully
23/09/07 03:00:58 INFO mapreduce.Job: Counters: 49
File System Counters
FILE: Number of bytes read=105
FILE: Number of bytes written=288641
FILE: Number of read operations=0
FILE: Number of large read operations=0
FILE: Number of write operations=0
HDFS: Number of bytes read=3312
HDFS: Number of bytes written=9
HDFS: Number of read operations=6
HDFS: Number of large read operations=0
HDFS: Number of write operations=2
Search
ENG
IN
21:06
21-09-2023

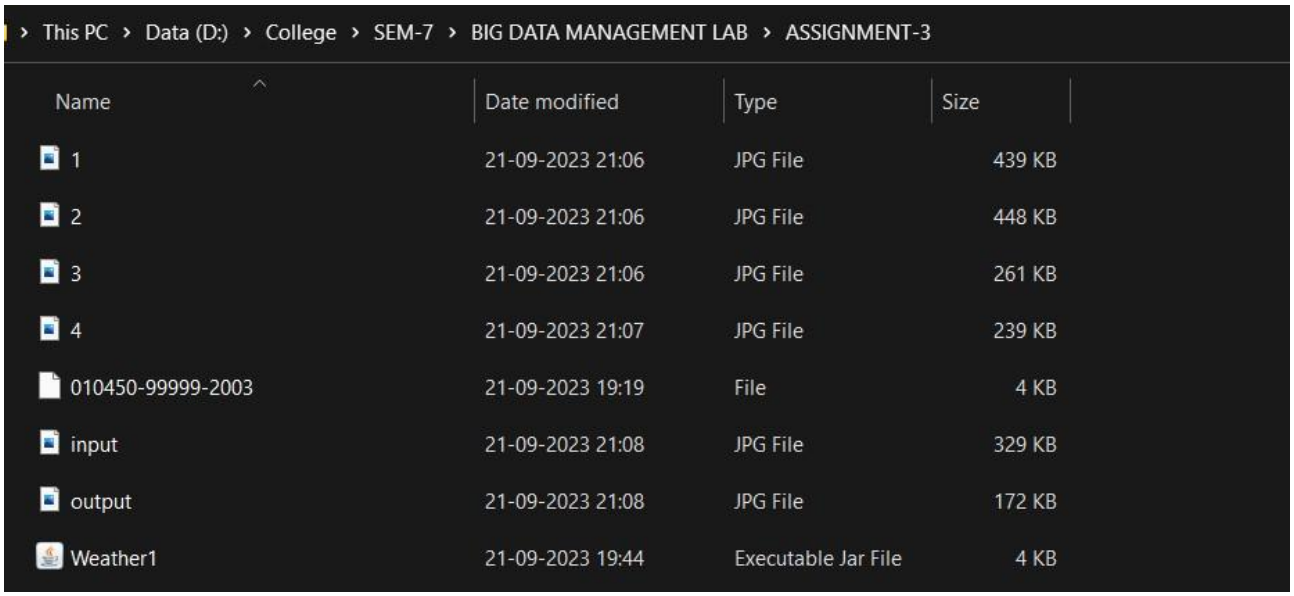

Hortonworks Sandbox with HDF
Ambari - Sandbox
root@sandbox:~ - Shell In A Box
127.0.0.1:4200
SRM
Internship
Tools
COLLEGE
Company
COURSE
Sites
Azure
DOCKER
All Bookmarks
Total time spent by all map tasks (ms)=31933
Total time spent by all reduce tasks (ms)=11307
Total vcore-milliseconds taken by all map tasks=31933
Total vcore-milliseconds taken by all reduce tasks=11307
Total megabyte-milliseconds taken by all map tasks=7983250
Total megabyte-milliseconds taken by all reduce tasks=2826750
Map-Reduce Framework
Map input records=11
Map output records=9
Map output bytes=81
Map output materialized bytes=105
Input split bytes=151
Combine input records=0
Combine output records=0
Reduce input groups=1
Reduce shuffle bytes=105
Reduce input records=9
Reduce output records=1
Spilled Records=18
Shuffled Maps =1
Failed Shuffles=0
Merged Map outputs=1
GC time elapsed (ms)=314
CPU time spent (ms)=2520
Physical memory (bytes) snapshot=287780864
Virtual memory (bytes) snapshot=3892879360
Total committed heap usage (bytes)=137887744
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=3161
File Output Format Counters
Bytes Written=9
[root@sandbox ~]# w
Search
ENG
IN
21:07
21-09-2023

OUTPUT:



PATH:

D:\College\SEM-7\BIG DATA MANAGEMENT LAB\ASSIGNMENT-3



Name	Date modified	Type	Size
1	21-09-2023 21:06	JPG File	439 KB
2	21-09-2023 21:06	JPG File	448 KB
3	21-09-2023 21:06	JPG File	261 KB
4	21-09-2023 21:07	JPG File	239 KB
010450-99999-2003	21-09-2023 19:19	File	4 KB
input	21-09-2023 21:08	JPG File	329 KB
output	21-09-2023 21:08	JPG File	172 KB
Weather1	21-09-2023 19:44	Executable Jar File	4 KB

ASSIGNMENT-4

PROGRAM:

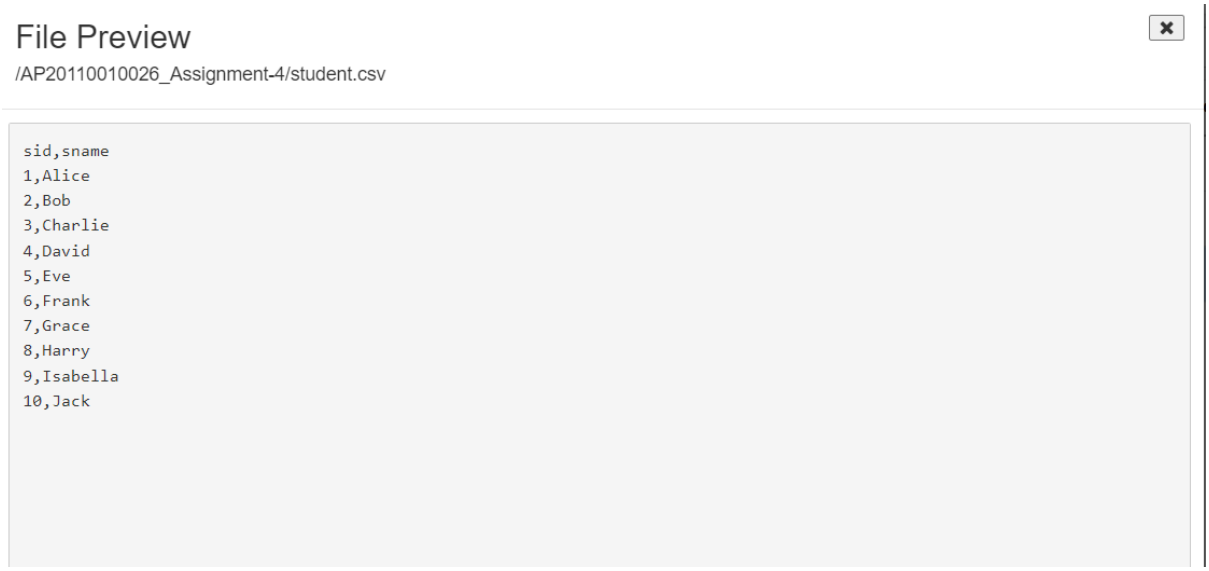
"A Student is Studying Many Subjects".

From the above sentences identify the entities and take any attributes;
Implement JOINS using HIVE programming.

TABLES:

student.csv

sid	sname
1	Alice
2	Bob
3	Charlie
4	David
5	Eve
6	Frank
7	Grace
8	Harry
9	Isabella
10	Jack



course.csv

cid	cname	sid
101	Math	1
102	English	2
103	Science	3
104	History	4
105	Computer	5
106	Art	6
107	Music	7
108	Physical Ed	8
109	Foreign La	9
110	Philosophy	10
102	English	1
104	History	5
104	History	2
107	Music	6

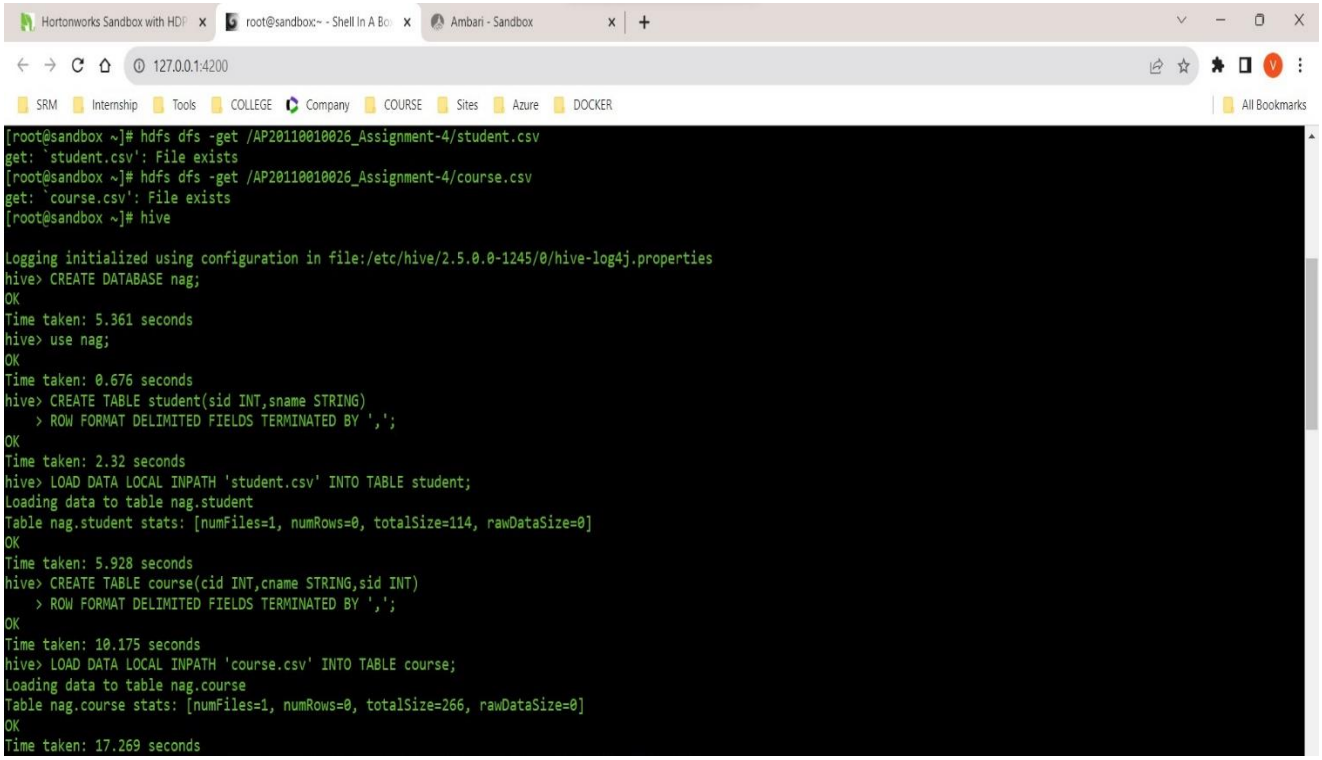
File Preview

/AP20110010026_Assignment-4/course.csv

```
cid,cname,sid
101,Math,1
102,English,2
103,Science,3
104,History,4
105,Computer Science,5
106,Art,6
107,Music,7
108,Physical Education,8
109,Foreign Language,9
110,Philosophy,10
102,English,1
104,History,5
104,History,2
107,Music,6
```

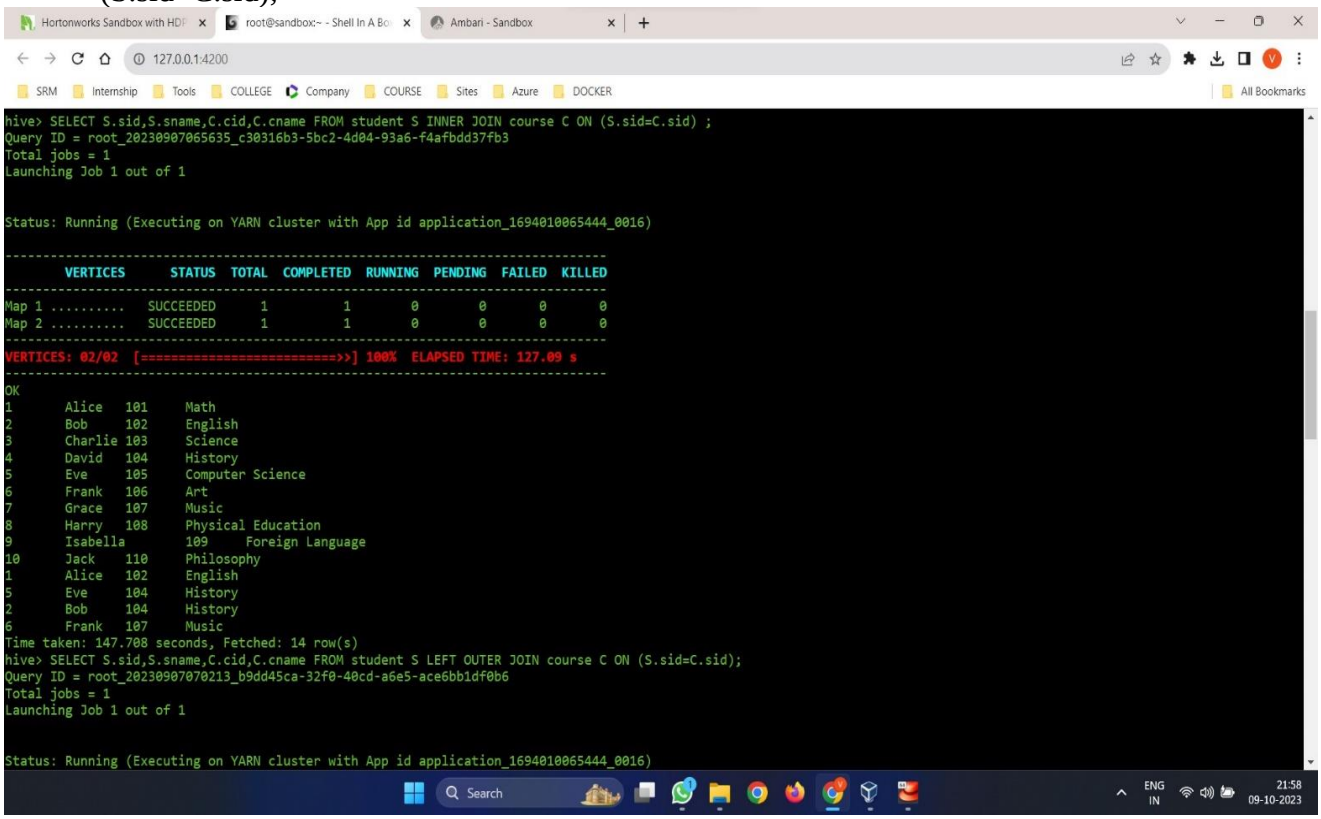
CREATION OF DATABASE,TABLES AND LOADING THE DATA INTO TABLES:

- CREATE DATABASE nag;
- use nag;
- CREATE TABLE student (sid INT, sname STRING)
- ROW FORMAT DELIMITED FIELDS TERMINATED BY ',' ;
- LOAD DATA LOCAL INPATH 'student.csv' INTO TABLE student;
- CREATE TABLE course (cid INT, cname STRING, sid INT)
- ROW FORMAT DELIMITED FIELDS TERMINATED BY '';
- LOAD DATA LOCAL INPATH 'course.csv' INTO TABLE course;



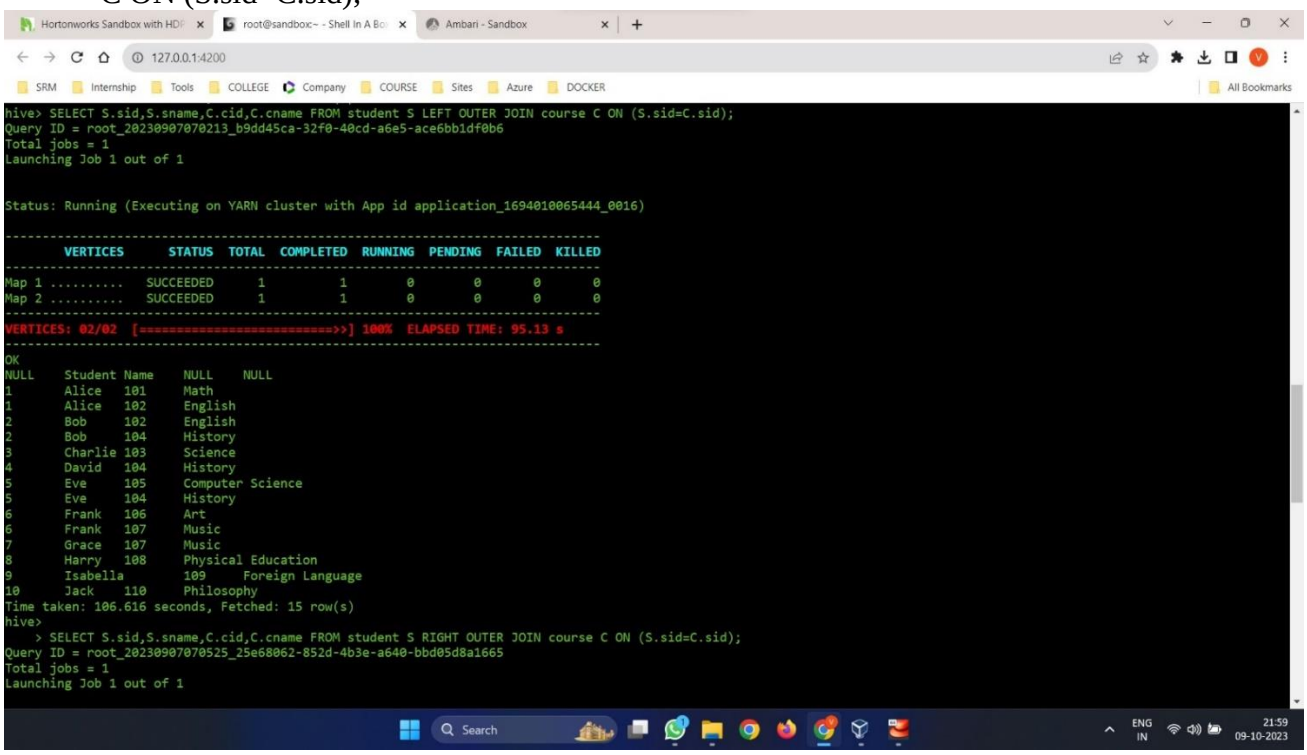
INNER JOIN:

- SELECT S.sid.S.sname,C.cid,C.cname FROM student S INNER JOIN course C ON (S.sid=C.sid);



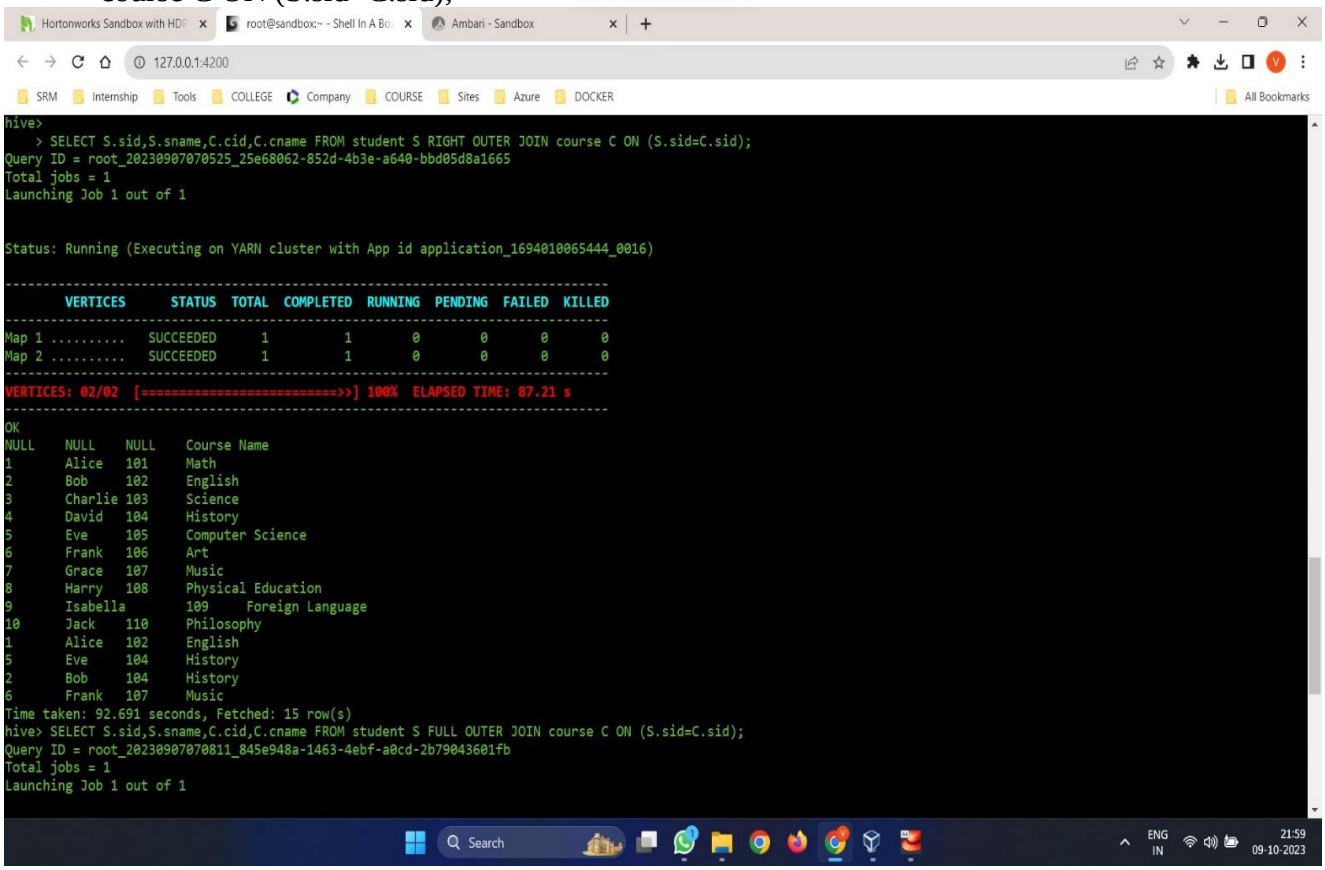
LEFT OUTER JOIN:

- SELECT S.sid, S.sname,C.cid,C.cname FROM student S LEFT OUTER JOIN course C ON (S.sid=C.sid);



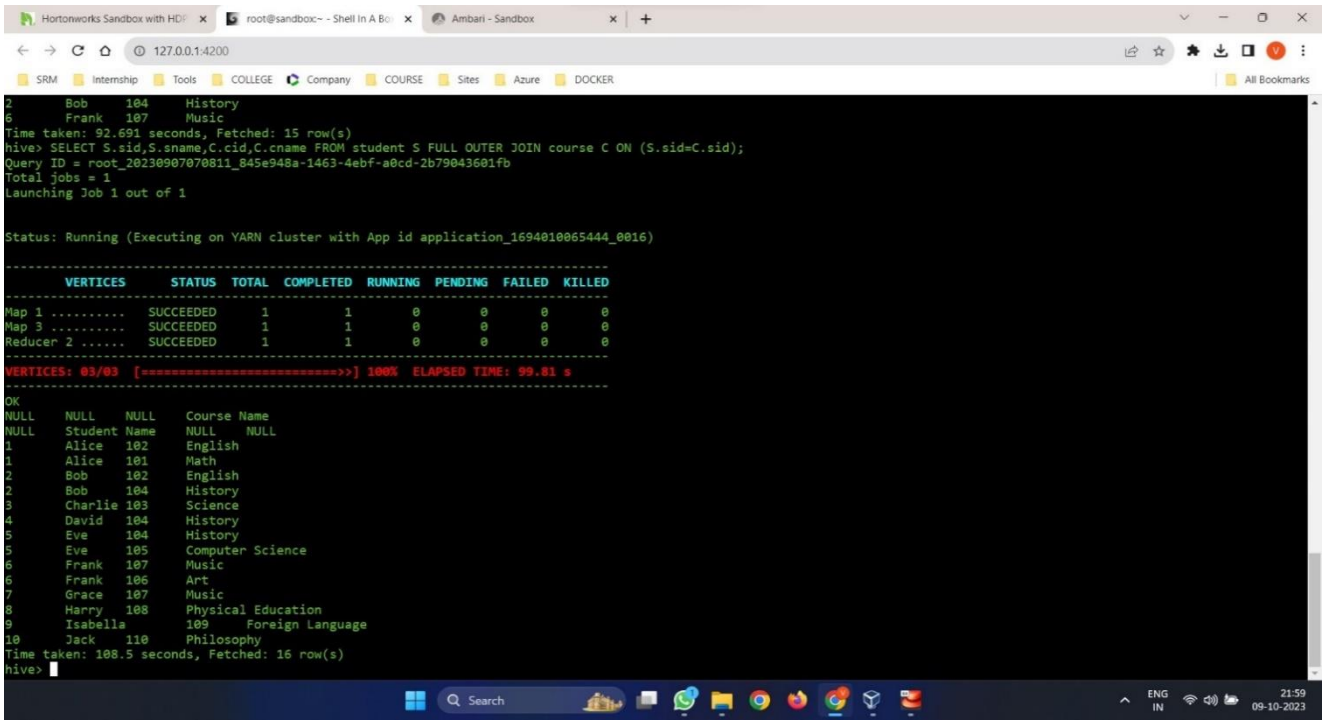
RIGHT OUTER JOIN:

- SELECT S.sid, S.sname, C.cid, C.cname FROM student S RIGHT OUTER JOIN course C ON (S.sid=C.sid);



FULL OUTER JOIN:

- SELECT S.sid,S.sname, C.cid, C.cname FROM student S FULL OUTER JOIN course C ON (S.sid=C.sid);



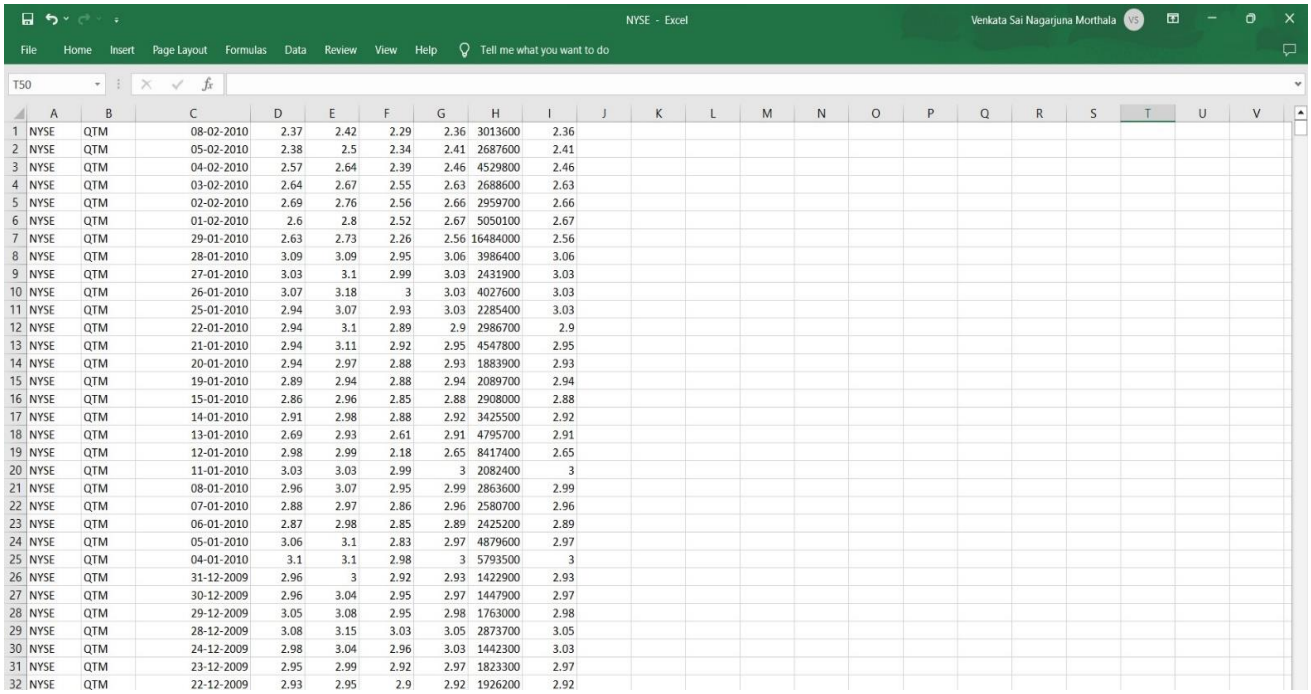
ASSIGNMENT-5

PROGRAM:

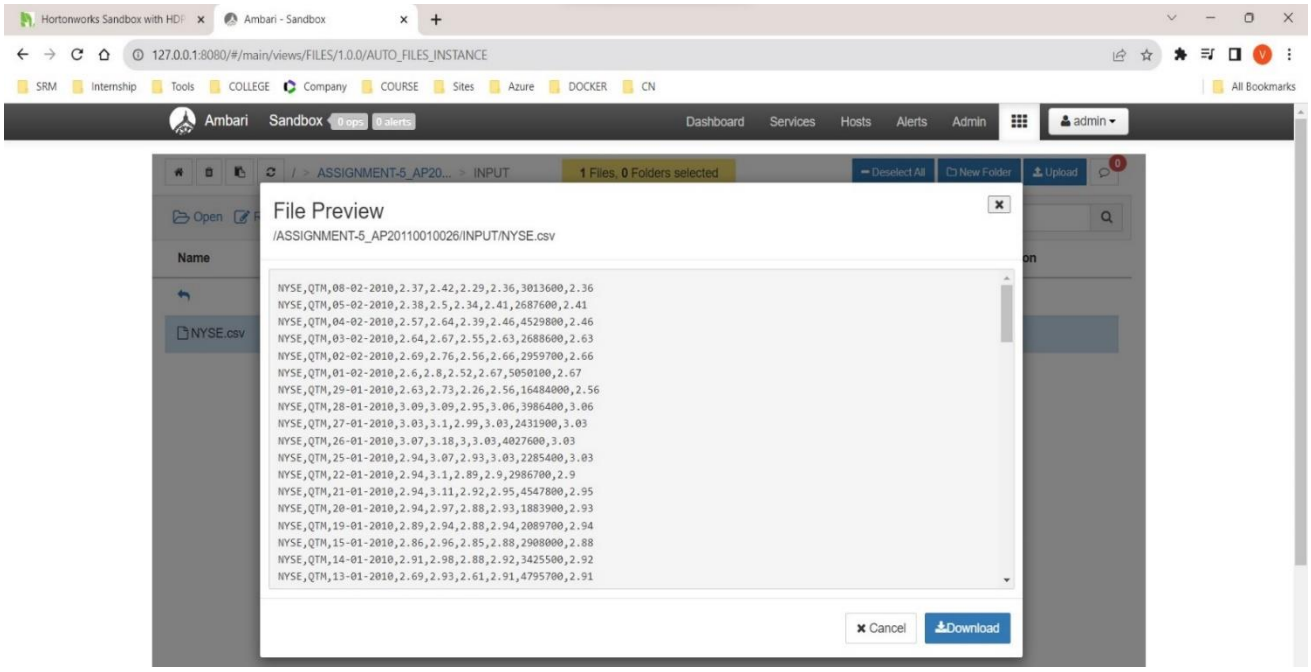
USE HIVEQL TO ANALYZE THE STOCK EXCHANGE DATASET AND CALCULATE THE COVARIANCE BETWEEN THE STOCKS FOR EACH MONTH

INPUT:

NYSE.csv



	A	B	C	D	E	F	G	H	I
1	NYSE	QTM	08-02-2010	2.37	2.42	2.29	2.36	3013600	2.36
2	NYSE	QTM	05-02-2010	2.38	2.5	2.34	2.41	2687600	2.41
3	NYSE	QTM	04-02-2010	2.57	2.64	2.39	2.46	4529800	2.46
4	NYSE	QTM	03-02-2010	2.64	2.67	2.55	2.63	2688600	2.63
5	NYSE	QTM	02-02-2010	2.69	2.76	2.56	2.66	2959700	2.66
6	NYSE	QTM	01-02-2010	2.6	2.8	2.52	2.67	5050100	2.67
7	NYSE	QTM	29-01-2010	2.63	2.73	2.26	2.56	16484000	2.56
8	NYSE	QTM	28-01-2010	3.09	3.09	2.95	3.06	3986400	3.06
9	NYSE	QTM	27-01-2010	3.03	3.1	2.99	3.03	2431900	3.03
10	NYSE	QTM	26-01-2010	3.07	3.18	3	3.03	4027600	3.03
11	NYSE	QTM	25-01-2010	2.94	3.07	2.93	3.03	2285400	3.03
12	NYSE	QTM	22-01-2010	2.94	3.1	2.89	2.9	2986700	2.9
13	NYSE	QTM	21-01-2010	2.94	3.11	2.92	2.95	4547800	2.95
14	NYSE	QTM	20-01-2010	2.94	2.97	2.88	2.93	1883900	2.93
15	NYSE	QTM	19-01-2010	2.89	2.94	2.88	2.94	2089700	2.94
16	NYSE	QTM	15-01-2010	2.86	2.96	2.85	2.88	2908000	2.88
17	NYSE	QTM	14-01-2010	2.91	2.98	2.88	2.92	3425500	2.92
18	NYSE	QTM	13-01-2010	2.69	2.93	2.61	2.91	4795700	2.91
19	NYSE	QTM	12-01-2010	2.98	2.99	2.18	2.65	8417400	2.65
20	NYSE	QTM	11-01-2010	3.03	3.03	2.99	3	2082400	3
21	NYSE	QTM	08-01-2010	2.96	3.07	2.95	2.99	2863600	2.99
22	NYSE	QTM	07-01-2010	2.88	2.97	2.86	2.96	2580700	2.96
23	NYSE	QTM	06-01-2010	2.87	2.98	2.85	2.89	2425200	2.89
24	NYSE	QTM	05-01-2010	3.06	3.1	2.83	2.97	4879600	2.97
25	NYSE	QTM	04-01-2010	3.1	3.1	2.98	3	5793500	3
26	NYSE	QTM	31-12-2009	2.96	3	2.92	2.93	1422900	2.93
27	NYSE	QTM	30-12-2009	2.96	3.04	2.95	2.97	1447900	2.97
28	NYSE	QTM	29-12-2009	3.05	3.08	2.95	2.98	1763000	2.98
29	NYSE	QTM	28-12-2009	3.08	3.15	3.03	3.05	2873700	3.05
30	NYSE	QTM	24-12-2009	2.98	3.04	2.96	3.03	1442300	3.03
31	NYSE	QTM	23-12-2009	2.95	2.99	2.92	2.97	1823300	2.97
32	NYSE	QTM	22-12-2009	2.93	2.95	2.9	2.92	1926200	2.92



COMMANDS:

*CREATING DIRECTORY

hdfs dfs -mkdir /ASSIGNMENT-5_AP20110010026

hdfs dfs -mkdir /ASSIGNMENT-5_AP20110010026/INPUT

*GETTING FILE TO THE LOCAL SYSTEM

hdfs dfs -get /ASSIGNMENT-5_AP20110010026/INPUT/NYSE.csv

*CREATING DATABASE:

hive> create database stocks;

hive> use stocks;

*CREATING TABLE:

hive> create table nyse(exchange_data STRING,stock_symbol STRING,stock_date STRING,stock_price_open DOUBLE,stock_price_high DOUBLE,stock_price_low DOUBLE,stock_price_close DOUBLE,stock_volume DOUBLE,stock_price_adj_close DOUBLE)

> row format delimited fields terminated by ',';

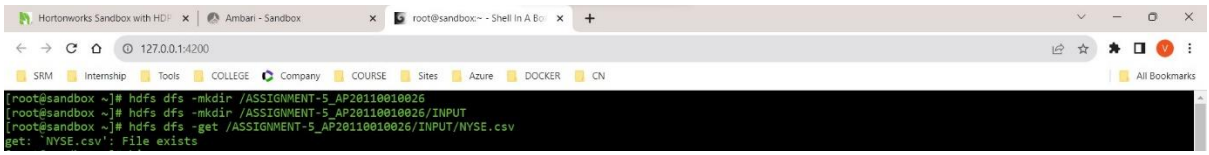
*LOADING DATA

hive> load data local inpath 'NYSE.csv' into table nyse;

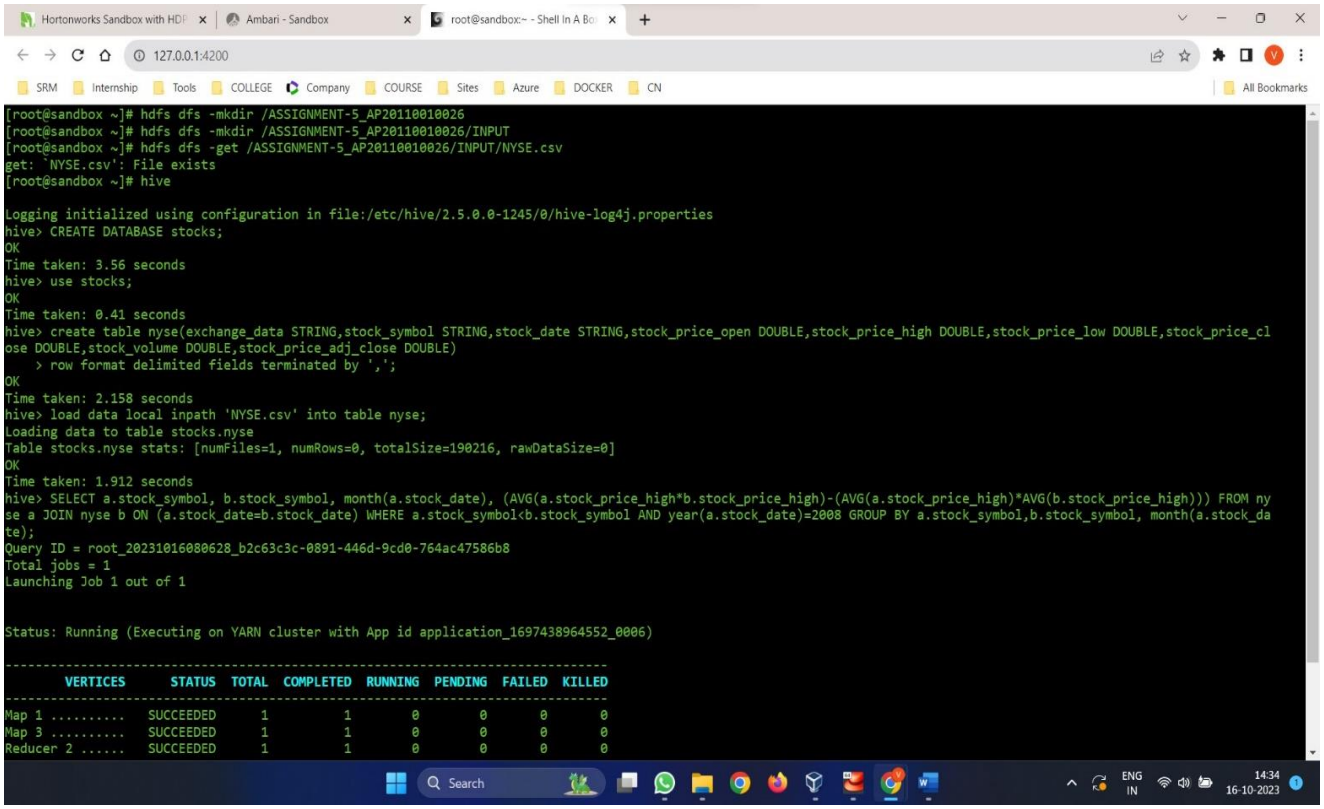
*USE THE FOLLOWING QUERY TO CALCULATE THE COVARIANCE BETWEEN STOCKS.

hive> SELECT a.stock_symbol, b.stock_symbol, month(a.stock_date), (AVG(a.stock_price_high*b.stock_price_high)- (AVG(a.stock_price_high)*AVG(b.stock_price_high))) FROM nyse a JOIN nyse b ON (a.stock_date=b.stock_date) WHERE a.stock_symbol<b.stock_symbol AND year(a.stock_date)=2008 GROUP BY a.stock_symbol,b.stock_symbol, month(a.stock_date);

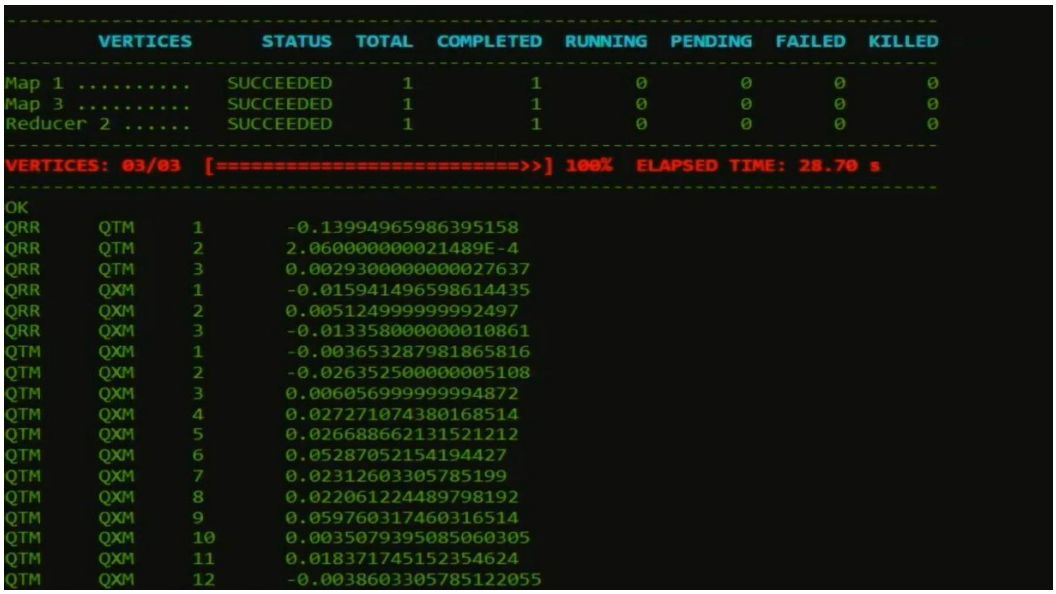
CREATING DIRECTORY AND GETTING THE FILE TO THE LOCAL SYSTEM:



HIVE CREATING DATABASE AND TABLE:



OUTPUT:



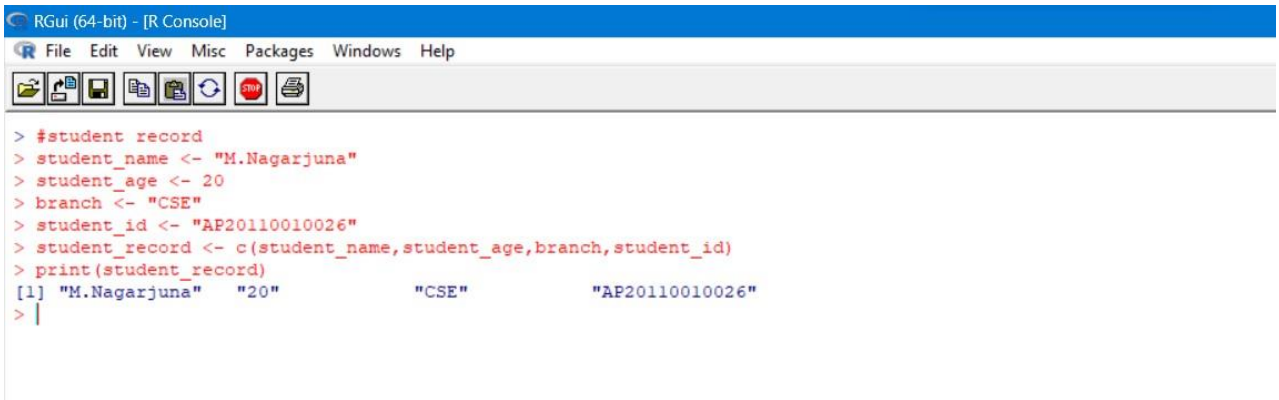
ASSIGNMENT-6

1. Write a R program to create a student record using the Vector concept.

CODE:

```
#student record
 student_name <- "M.Nagarjuna"
 student_age <- 20
 branch <- "CSE"
 student_id <- "AP20110010026"
 student_record <- c(student_name,student_age,branch,student_id)
 print(student_record)
```

OUTPUT:



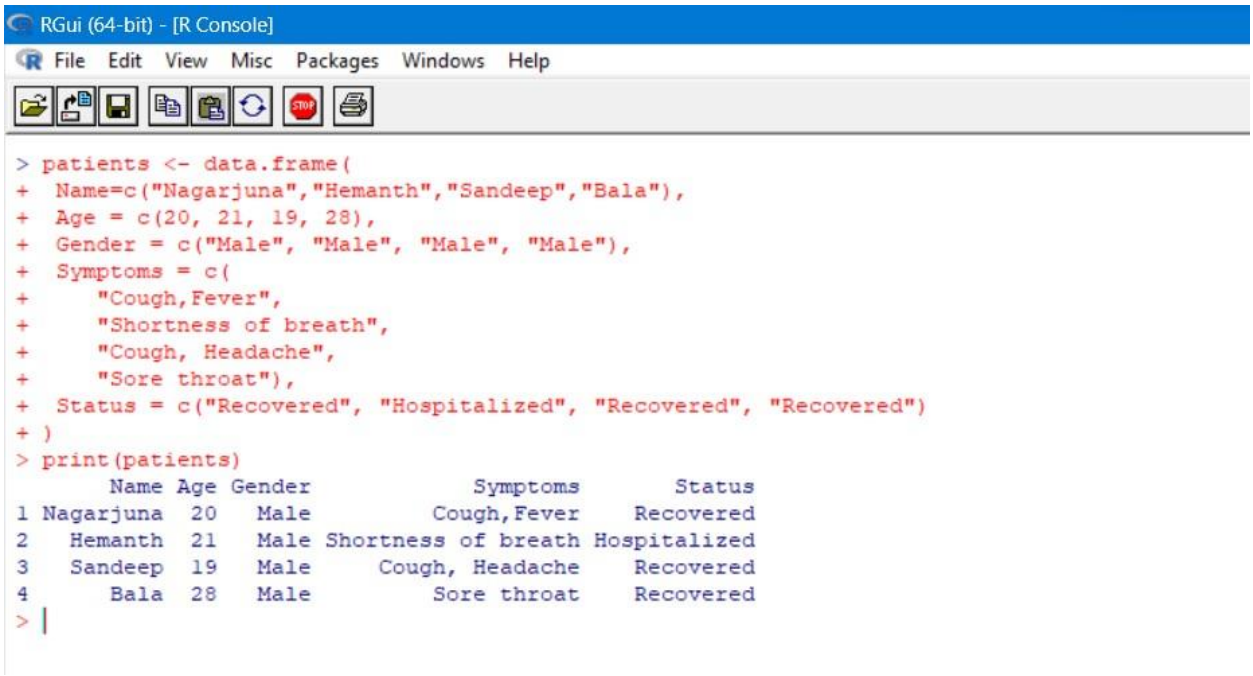
2. Write a R program to create medical patients status using data frame

i) Patient age ii) Gender iii) Symptoms iv) Patient Status

CODE:

```
patients <- data.frame(
 Name=c("Nagarjuna","Hemanth","Sandeep","Bala"),
 Age = c(20, 21, 19, 28),
 Gender = c("Male", "Male", "Male", "Male"),
 Symptoms = c(
    "Cough,Fever",
    "Shortness of breath",
    "Cough, Headache",
    "Sore throat"),
 Status = c("Recovered", "Hospitalized", "Recovered", "Recovered")
)
print(patients)
```

OUTPUT:



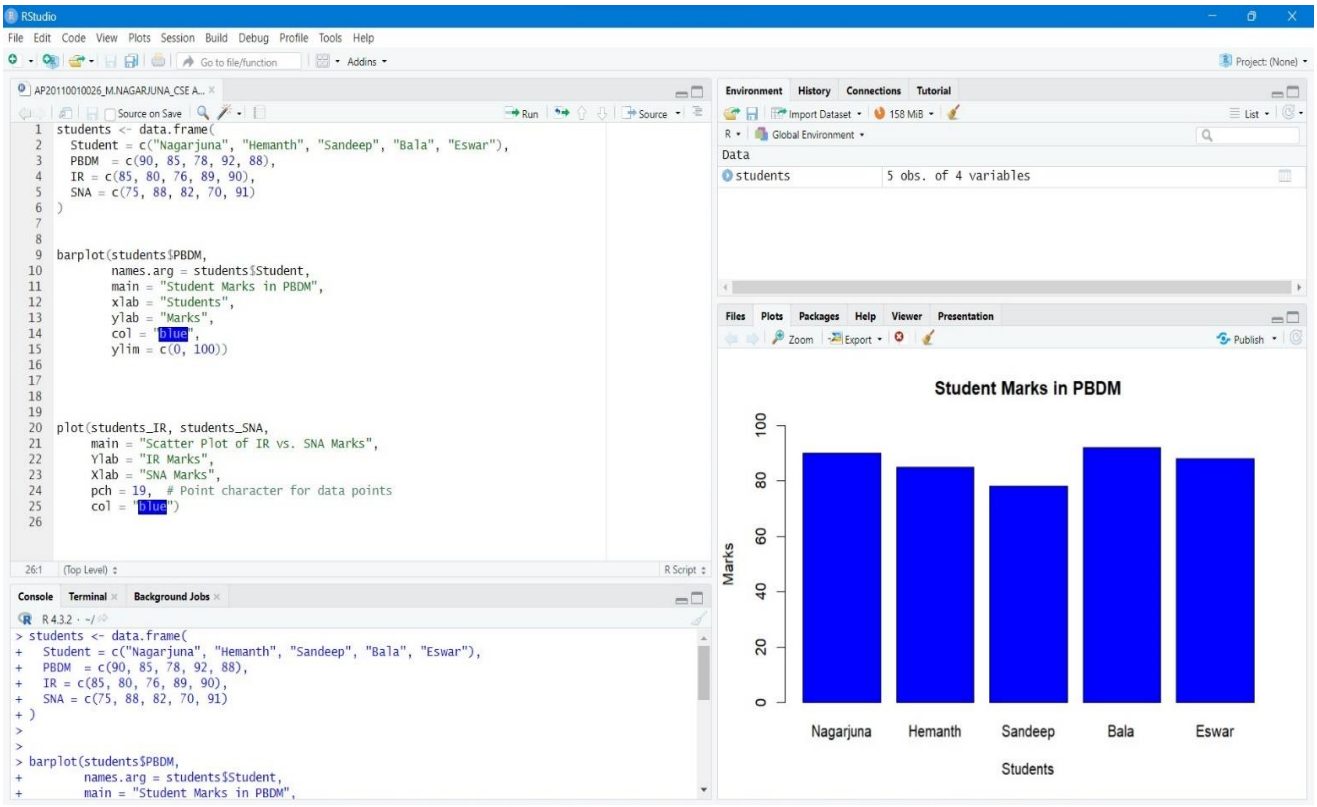
3. Write a R program to visualize student marks of various subjects using Bar-chart and Scatter plot.

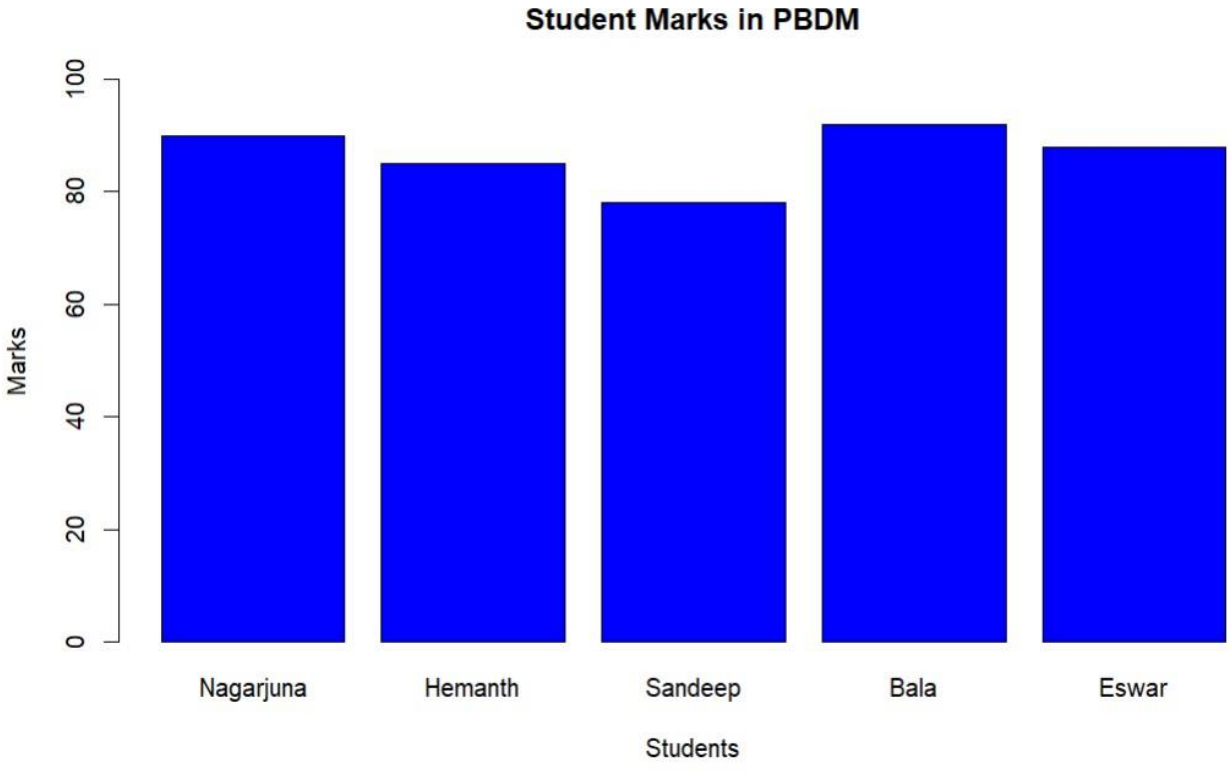
CODE:

```
students <- data.frame(
  Student = c("Nagarjuna", "Hemanth", "Sandeep", "Bala", "Eswar"),
  PBDM  = c(90, 85, 78, 92, 88),
  IR = c(85, 80, 76, 89, 90),
  SNA = c(75, 88, 82, 70, 91)
)
barplot(students$PBDM,
        names.arg = students$Student,
        main = "Student Marks in PBDM",
        xlab = "Students",
        ylab = "Marks",
        col = "blue",
        ylim = c(0, 100))
plot(students$IR, students$SNA,
     main = "Scatter Plot of IR vs. SNA Marks",
     Ylab = "IR Marks",
     Xlab = "SNA Marks",
     pch = 19,  # Point character for data points
     col = "blue")
```

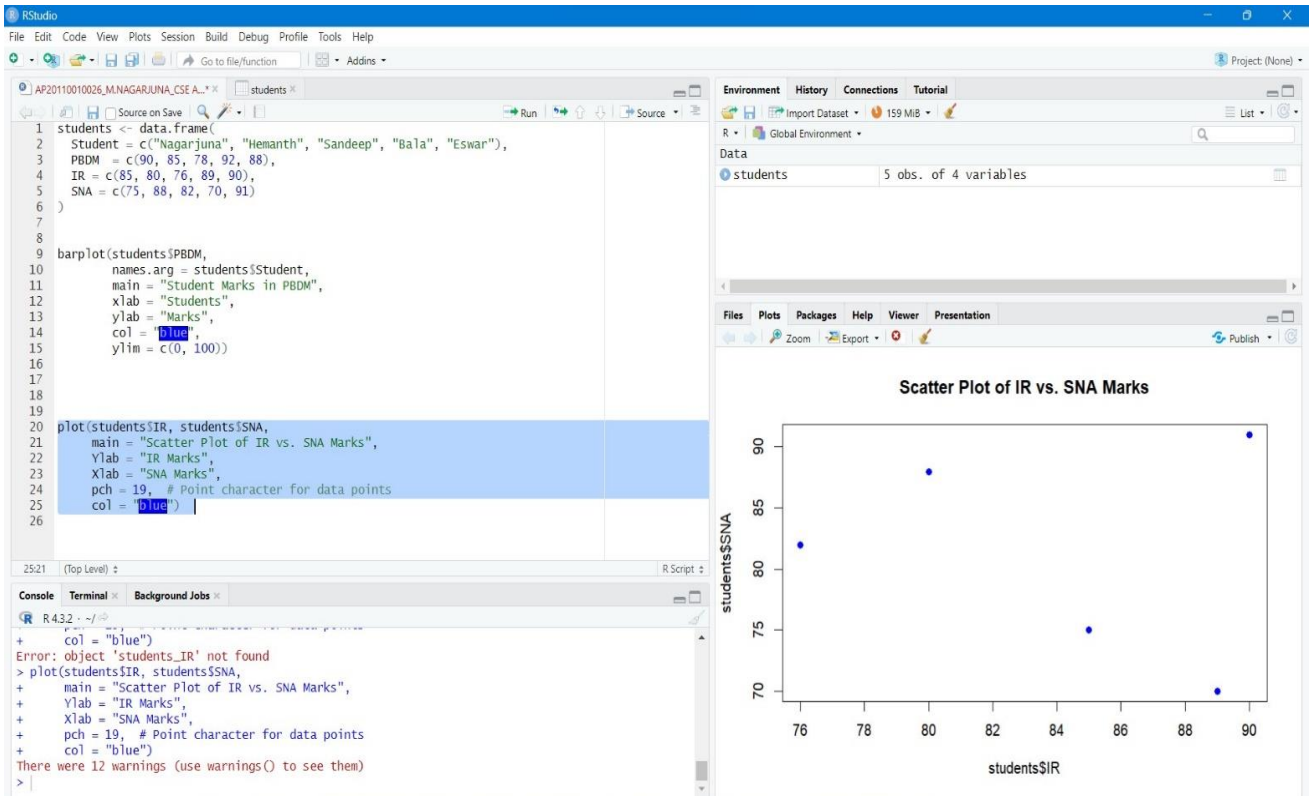
OUTPUT:

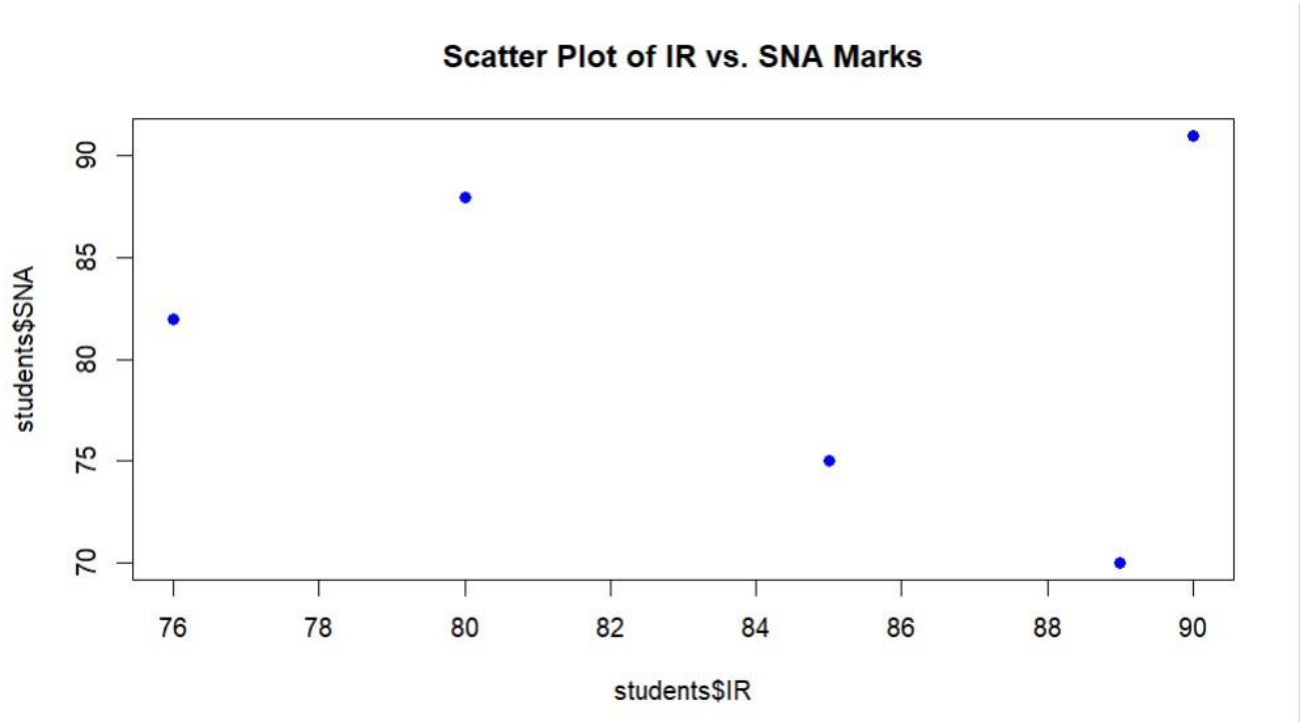
BARCHART:





SCATTER_PLOT:





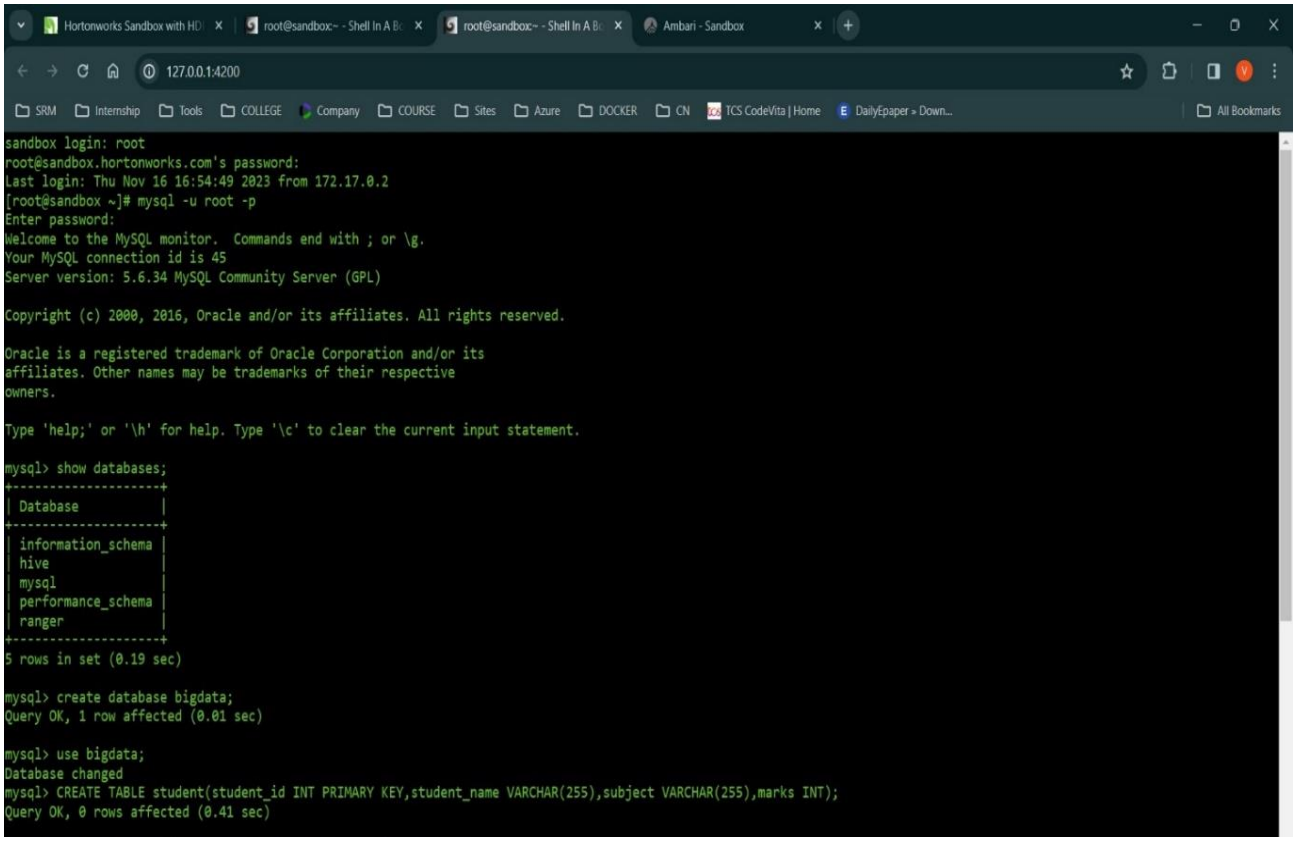
ASSIGNMENT-7

PROGRAM:

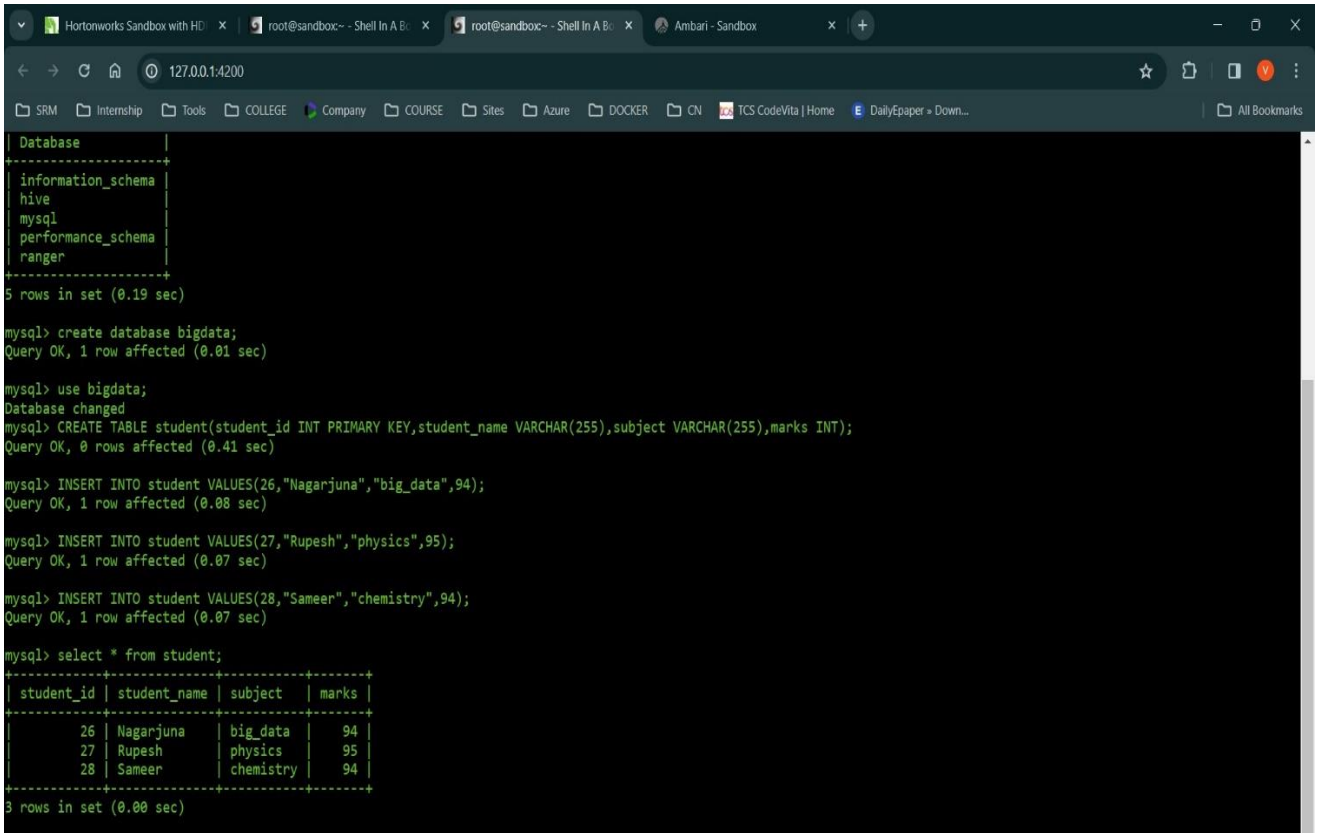
Transfer data between Hadoop and relational database servers using Sqoop.

EXECUTION:

- mysql -u root -p
- create database bigdata;
- use bigdata;

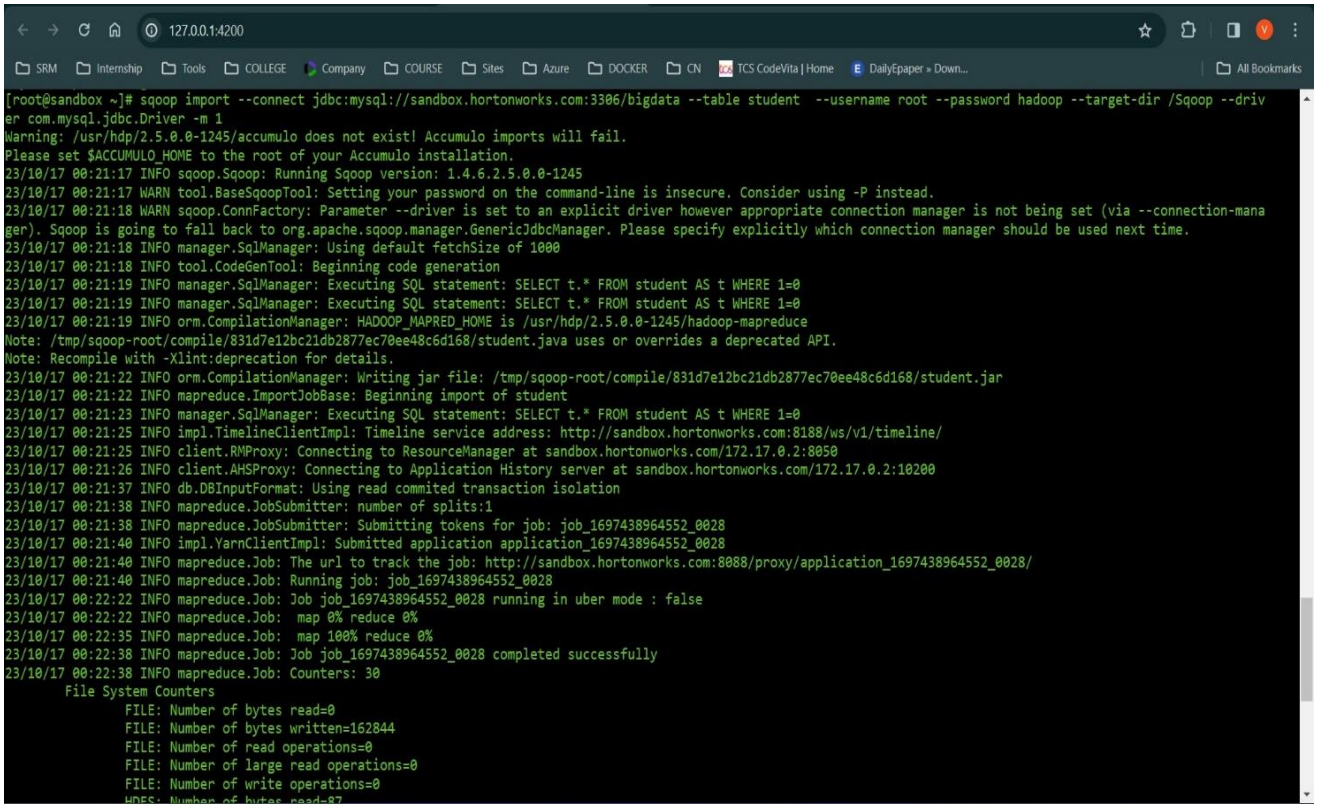


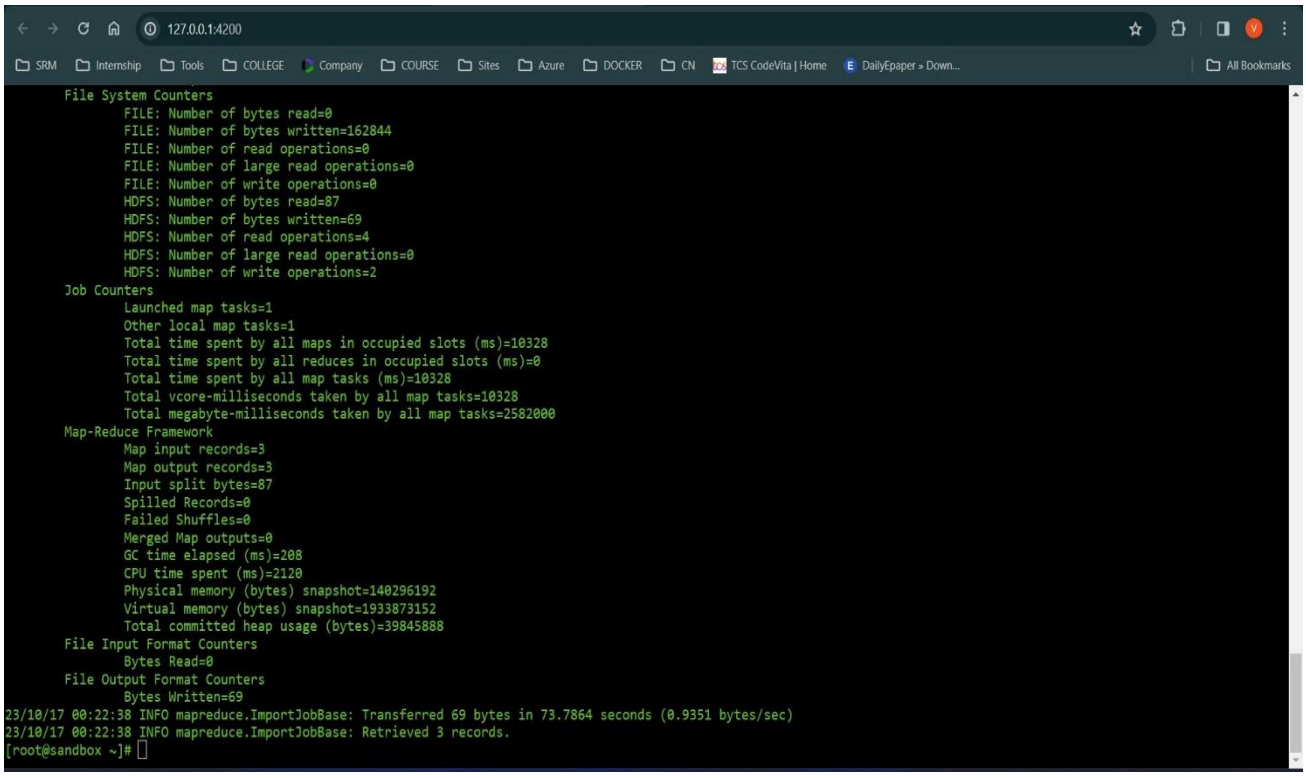
- INSERT INTO student VALUES(26, "Nagarjuna", "big_data",94);
- INSERT INTO student VALUES (27, "Rupesh", "physics", 95);
- INSERT INTO student VALUES(28, "Sameer","chemistry",94);



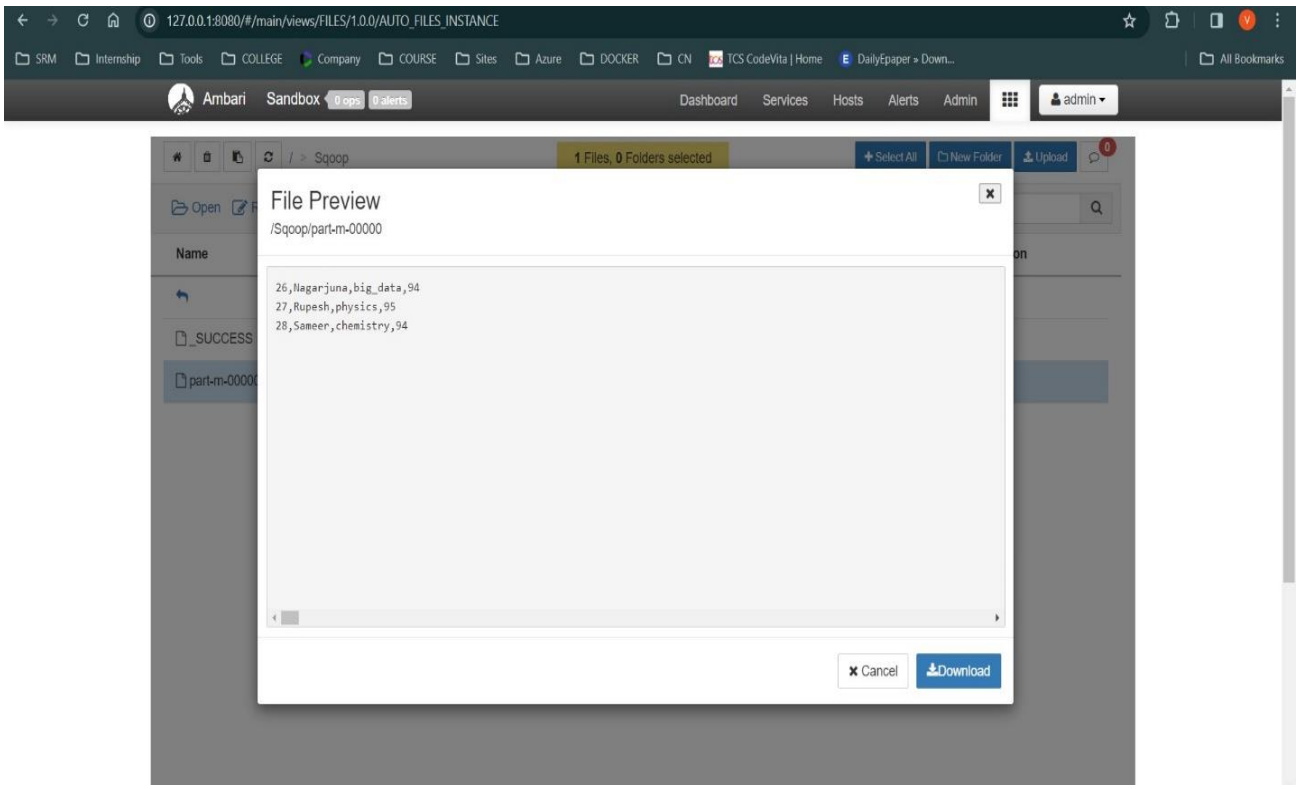
COMMAND:

sqoop import --connect jdbc:mysql://sandbox.hortonworks.com:3306/bigdata --table student --username root --password hadoop --target-dir /Sqoop --driver com.mysql.jdbc.Driver -m 1





OUTPUT:


127.0.0.1:8080/#/main/views/FILES/1.0.0/AUTO_FILES_INSTANCE
Ambari
Sandbox
Dashboard
Services
Hosts
Alerts
Admin
admin
Sqoop
1 Files, 0 Folders selected
Select All
New Folder
Upload
Open
Name
_SUCCESS
part-m-00000
File Preview
/Sqoop/part-m-00000
26,Nagarjuna,big_data,94
27,Rupesh,physics,95
28,Sameer,chemistry,94
Cancel
Download


File Preview
/Sqoop/part-m-00000
26,Nagarjuna,big_data,94
27,Rupesh,physics,95
28,Sameer,chemistry,94
Cancel
Download