ASSIGNMENT BY SAHIL KHURANA

Issue #1

Data Hadoop & Spark Training - ACADGILD Assignment 3.3

18-Nov-17

BIG DATA

DATA SET DESCRIPTION

Name of the file - television.txt

```
Samsung|Optima|14|Madhya Pradesh|132401|14200
Onida|Lucid|18|Uttar Pradesh|232401|16200
Akai|Decent|16|Kerala|922401|12200
Lava|Attention|20|Assam|454601|24200
Zen|Super|14|Maharashtra|619082|9200
Samsung|Optima|14|Madhya Pradesh|132401|14200
Onida|Lucid|18|Uttar Pradesh|232401|16200
Onida|Decent|14|Uttar Pradesh|232401|16200
Onida|NA|16|Kerala|922401|12200
Lava|Attention|20|Assam|454601|24200
Zen|Super|14|Maharashtra|619082|9200
Samsung|Optima|14|Madhya Pradesh|132401|14200
NA|Lucid|18|Uttar Pradesh|232401|16200
Samsung|Decent|16|Kerala|922401|12200
Lava|Attention|20|Assam|454601|24200
Samsung|Super|14|Maharashtra|619082|9200
Samsung|Super|14|Maharashtra|619082|9200
Samsung|Super|14|Maharashtra|619082|9200
```

PROBLEM STATEMENT

We have a dataset of sales of different TV sets across different locations.

Records look like:

Samsung|Optima|14|Madhya Pradesh|132401|14200

The fields are arranged like:

Company Name|Product Name|Size in inches|State|Pin Code|Price

There are some invalid records which contain 'NA' in either Company Name or Product Name.

Task1. Write a Map Reduce program to filter out the invalid records. Map only job will fit for this context.

Note- Task 1 is completed in Assignment 3.2. The output of Assignment 3.2 will be taken as input of Assignment 3.3

Task2. Write a Map Reduce program to calculate the total units sold for each Company.

Task3. Write a Map Reduce program to calculate the total units sold in each state for Onida company.

Task2

Mapper Code: Task2Mapper.java

```
/*
 * All the comments in the program are highlighted in Green.
 * @author Sahil Khurana <sahilkhurana369@gmail.com
 */
```

```
package Assignment_3_Task2;                                  // Package Declared

import java.io.IOException;                                  // IOException class import to handle exceptions

import org.apache.hadoop.io.IntWritable;

import org.apache.hadoop.io.LongWritable;

import org.apache.hadoop.io.Text;

import org.apache.hadoop.mapreduce.*;

// class is extended to have the arguments keyIn as LongWritable and ValueIn as Text and KeyOut as Text and ValueOut as IntWritable.

public class Task2Mapper extends Mapper<LongWritable, Text, Text, IntWritable> {

// The variables declared are: a Text variable Company_Name_Key

Text Company_Name_Key = new Text();
// The variables declared are: a IntWritable variable Company_Name_Value which will store the value of the MapReduce deals with Key and Value pairs.

IntWritable Company_Name_Value = new IntWritable();

// overriding the map method which will run one time for every line.

@Override

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

// storing the line in a string variable and splitting the line by using comma "|" delimiter and storing the values in a String Array so that all the columns in a row are stored in the string array.

String[] myLineArray = value.toString().split("\\|");

Company_Name_Key.set(myLineArray[0]);

Company_Name_Value.set(1);

// obtaining key and value context.

context.write(Company_Name_Key, Company_Name_Value);

          }             // map class closed


}             // class Task2Mapper closed
```

Reducer Code: Task2Reducer.java

```
/*
 * All the comments in the program are highlighted in Green.
 * @author Sahil Khurana <sahilkhurana369@gmail.com
 */
package Assignment_3_Task2;                                                            // Package declared

import java.io.IOException;

import org.apache.hadoop.io.IntWritable;

import org.apache.hadoop.io.Text;

import org.apache.hadoop.mapreduce.Reducer;

// class is extended to have the arguments keyIn as LongWritable and ValueIn as Text and KeyOut as Text and ValueOut as IntWritable.

public class Task2Reducer    extends Reducer<Text, IntWritable, Text, IntWritable> {

// The variables declared are: a IntWritable variable Company_Name_Value which will store the value of the MapReduce deals with Key and Value pairs.

IntWritable Company_Name_Value = new IntWritable();

// The variables declared are: a IntWritable variable Company_Name_Value which will store the value of the MapReduce deals with Key and Value pairs.

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

// declaring an integer sum which will store the sum of all the Company_Name_Value

int sum = 0;

// for each loop is taken which will run each time for the values inside the "Iterable values" which are coming from the shuffle and sort phase after the mapper phase.
for (IntWritable value : values) {

sum += value.get();}                                                  // storing and calculating the sum of the values.

Company_Name_Value.set(sum);                                          // setting the sum

context.write(key, Company_Name_Value);                               // obtaining key and value context.

        }                                                             // map class closed

}                                                                     // class Task2Reducer closed
```

Driver Code: Task2.java

```
/*
 * All the comments in the program are highlighted in Green.
 * @author Sahil Khurana <sahilkhurana369@gmail.com
 */
```

```
package Assignment_3_Task2;                                      // Package declared

import org.apache.hadoop.conf.Configuration;                     // Import the Configuration of system parameters.

import org.apache.hadoop.fs.Path;                                // used to Names a file or directory in a AbstractFileSystem for hdfs.

import org.apache.hadoop.io.Text;                                // This class stores text using standard UTF8 encoding.

import org.apache.hadoop.mapreduce.Job;                          // The job submitter's view of the Job.

import org.apache.hadoop.mapreduce.lib.input.FileInputFormat;    // FilterInputFormat is a convenience class that wraps InputFormat.

import org.apache.hadoop.mapreduce.lib.output.FileOutputFormat;  // FilterOutputFormat is a convenience class that wraps   OutputFormat.

import org.apache.hadoop.mapreduce.lib.input.TextInputFormat;    // TextInputFormat is a convenience class that wraps InputFormat.

import org.apache.hadoop.mapreduce.lib.output.TextOutputFormat;  // TextOutputFormat is a convenience class that wraps OutputFormat.

import org.apache.hadoop.io.IntWritable;

public class Task2 {                                           // class declared

@SuppressWarnings("deprecation")                               // The @SuppressWarnings annotation disables certain compiler

                                                               // warnings. In this case, the warning about deprecated code ("deprecation")

public static void main(String[] args) throws Exception {      // main class started

Configuration conf = new Configuration();                      // Create a configuration object for the job

Job job = new Job(conf, "Assignment_3.3_Task2");               // created the new object name Job

job.setJarByClass(Task2.class);                                // Set a name of the Jar

job.setMapOutputKeyClass(Text.class);                          // Set the output Key type for the Mapper

job.setMapOutputValueClass(IntWritable.class);                 // Set the output Value type for the Mapper

job.setOutputKeyClass(Text.class);                             // Set the output Key type for the Reducer

job.setOutputValueClass(IntWritable.class);                    // Set the output Value type for the Reducer

job.setMapperClass(Task2Mapper.class);                         // Set the Mapper Class

job.setReducerClass(Task2Reducer.class);                       // Set the Reducer Class

job.setInputFormatClass(TextInputFormat.class);                // Set the format of the input that will be provided to the program
job.setOutputFormatClass(TextOutputFormat.class);          // Set the format of the output for the program
FileInputFormat.addInputPath(job, new Path(args[0]));      // Set input directories using command line arguments,arg[0] on HDFS
FileOutputFormat.setOutputPath(job, new Path(args[1]));    // Set output directories using command line arguments,arg[0]  on HDFS
job.waitForCompletion(true);                               // Submit the job, then poll for progress until the job is complete
  }                                                        // main method closed
}                                                          // class Task1 closed
```

Steps to Run Mapreduce program on command prompt for Task2

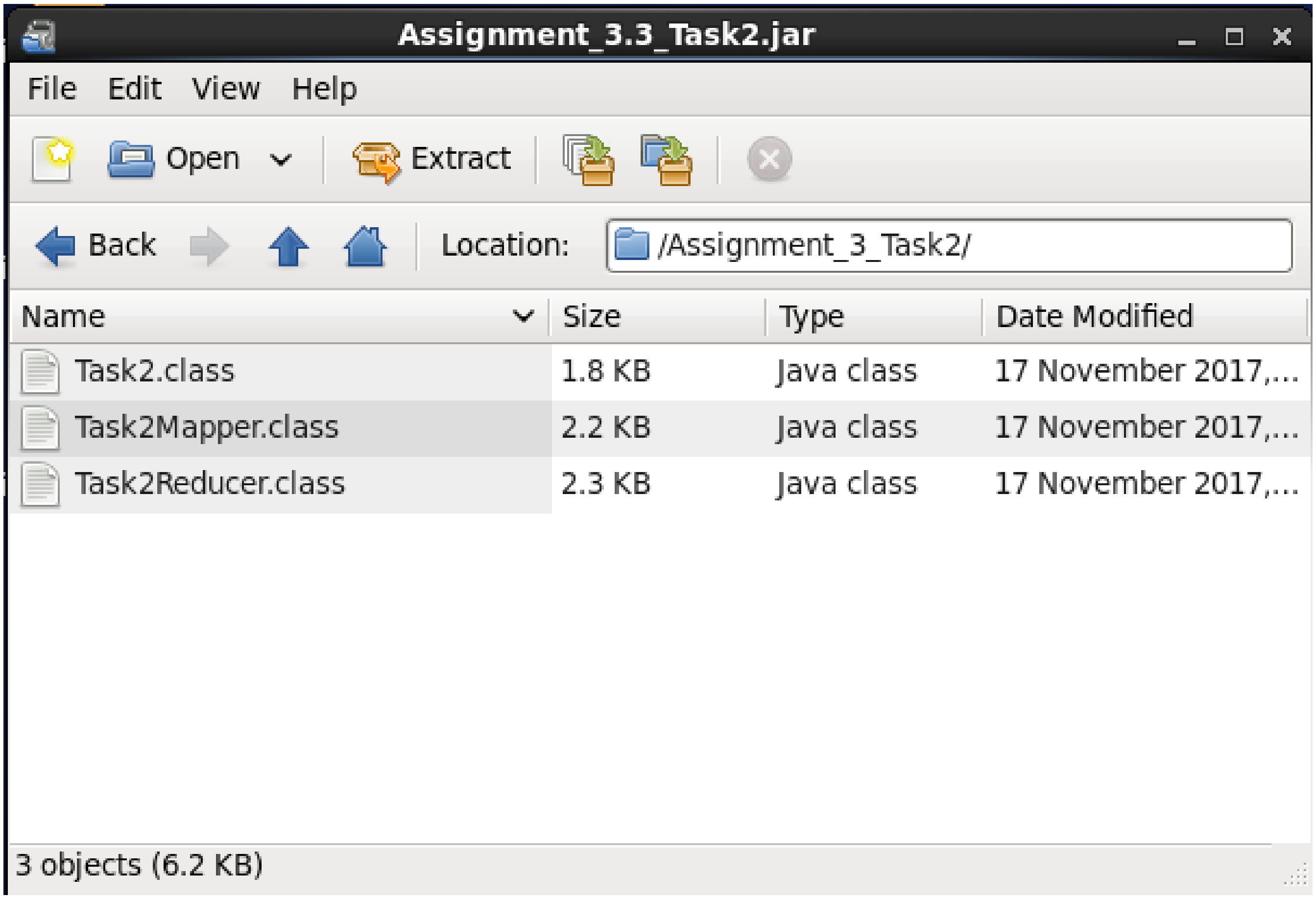
Step 1:

Create a Jar, which will contain

Driver Code: Task2.java

Mapper Code: Task2Mapper.java

Reducer Code: Task2Reducer.java

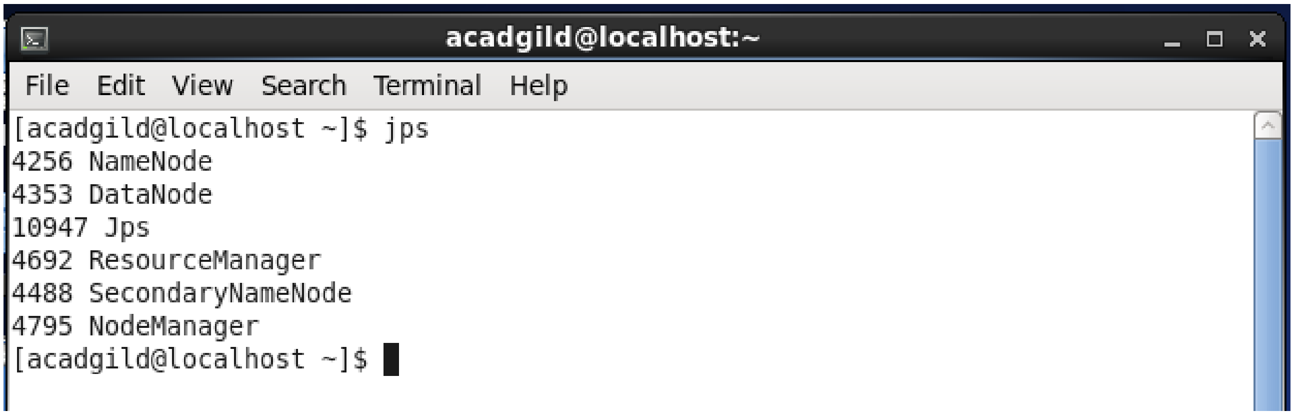


Name	Size	Type	Date Modified
Task2.class	1.8 KB	Java class	17 November 2017,...
Task2Mapper.class	2.2 KB	Java class	17 November 2017,...
Task2Reducer.class	2.3 KB	Java class	17 November 2017,...

3 objects (6.2 KB)

Step 2:

Check the all the Hadoop services are running or not by typing jps command

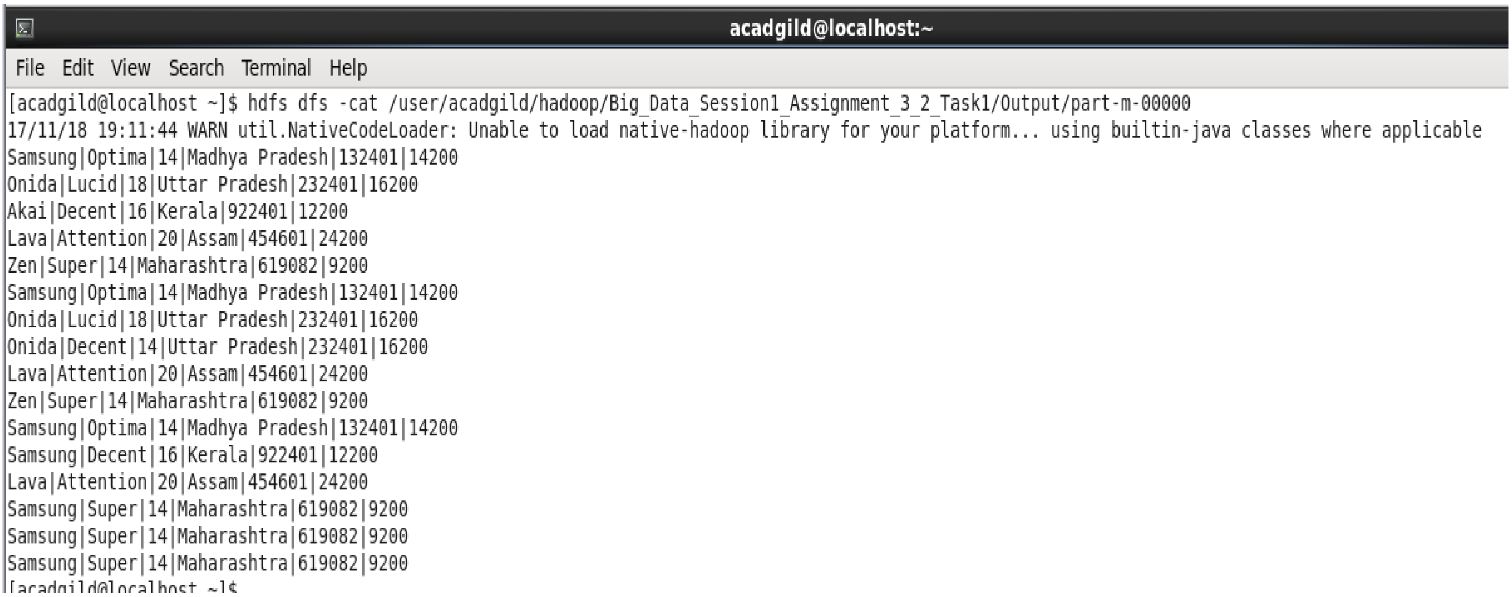


```
[acadgild@localhost ~]$ jps
4256 NameNode
4353 DataNode
10947 Jps
4692 ResourceManager
4488 SecondaryNameNode
4795 NodeManager
[acadgild@localhost ~]$
```

Step 3:

The output of Assignment 3.2 will be taken as input of Assignment 3.3

hdfs dfs -cat /user/acadgild/hadoop/Big_Data_Session1_Assignment_3_2_Task1/Output/part-m-00000



```
[acadgild@localhost ~]$ hdfs dfs -cat /user/acadgild/hadoop/Big_Data_Session1_Assignment_3_2_Task1/Output/part-m-00000
17/11/18 19:11:44 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applicable
Samsung|Optima|14|Madhya Pradesh|132401|14200
Onida|Lucid|18|Uttar Pradesh|232401|16200
Akai|Decent|16|Kerala|922401|12200
Lava|Attention|20|Assam|454601|24200
Zen|Super|14|Maharashtra|619082|9200
Samsung|Optima|14|Madhya Pradesh|132401|14200
Onida|Lucid|18|Uttar Pradesh|232401|16200
Onida|Decent|14|Uttar Pradesh|232401|16200
Lava|Attention|20|Assam|454601|24200
Zen|Super|14|Maharashtra|619082|9200
Samsung|Optima|14|Madhya Pradesh|132401|14200
Samsung|Decent|16|Kerala|922401|12200
Lava|Attention|20|Assam|454601|24200
Samsung|Super|14|Maharashtra|619082|9200
Samsung|Super|14|Maharashtra|619082|9200
Samsung|Super|14|Maharashtra|619082|9200
[acadgild@localhost ~]$
```

Step 4:

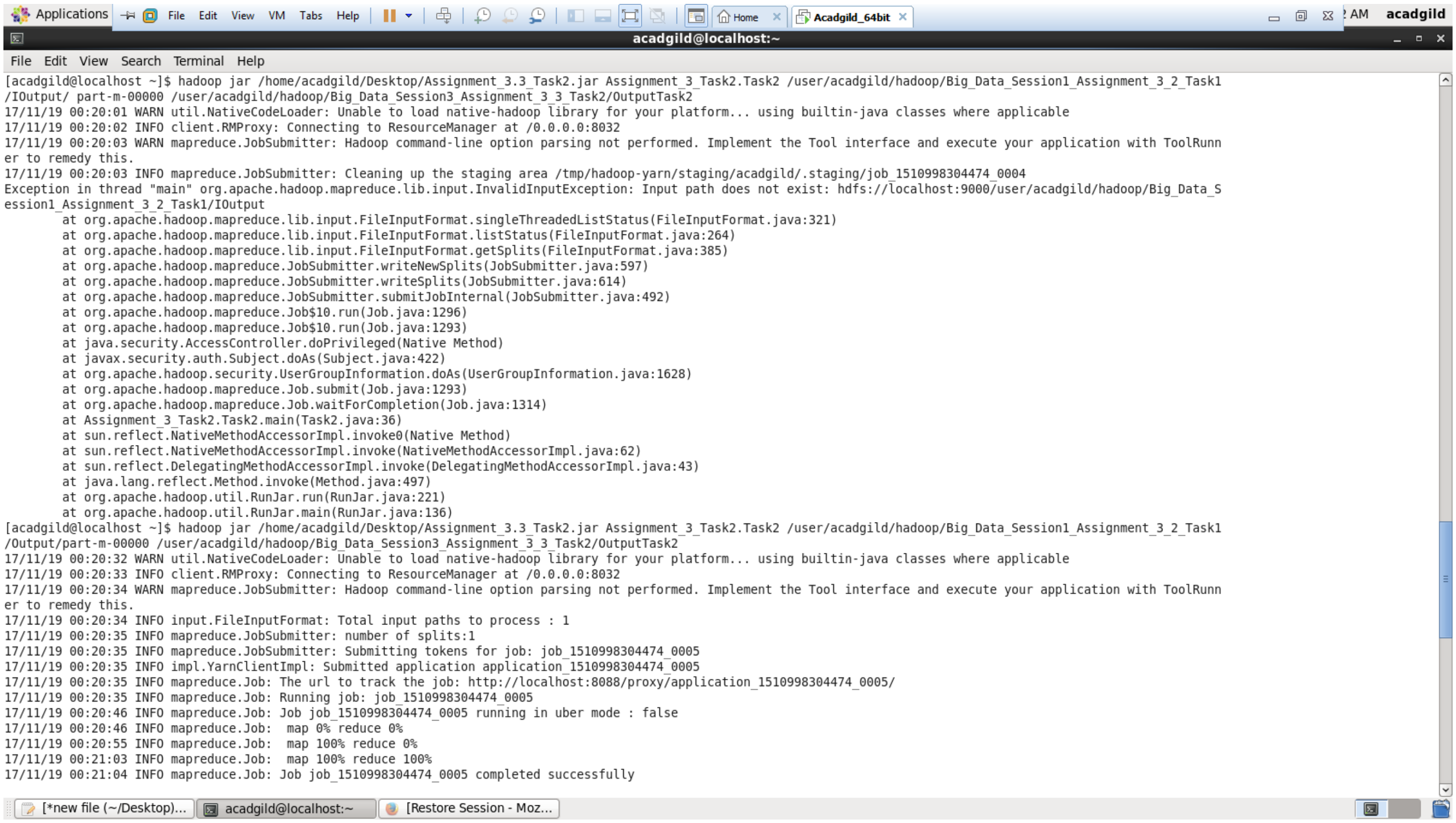
Run the MapReduce Job

Generic Command:-

hadoop jar <JAR file Path > <PackageName.MainClass> <Inputfile> <outputDir>

In case of Assignment 3.3 Task2:-

```
hadoop jar /home/acadgild/Desktop/Assignment_3.3_Task2.jar Assignment_3_Task2.Task2
/user/acadgild/hadoop/Big_Data_Session1_Assignment_3_2_Task1/Output/part-m-00000
/user/acadgild/hadoop/Big_Data_Session3_Assignment_3_3_Task2/OutputTask2
```



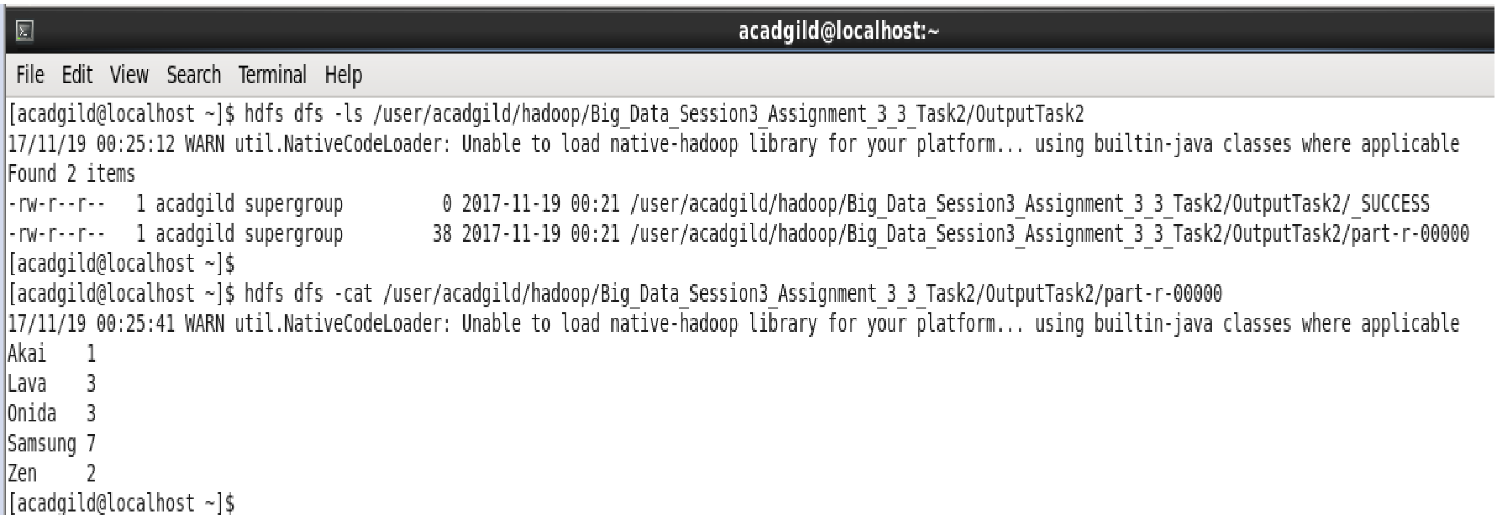
```
[acadgild@localhost ~]$ hadoop jar /home/acadgild/Desktop/Assignment_3.3_Task2.jar Assignment_3_Task2.Task2 /user/acadgild/hadoop/Big_Data_Session1_Assignment_3_2_Task1
/IOutput/ part-m-00000 /user/acadgild/hadoop/Big_Data_Session3_Assignment_3_3_Task2/OutputTask2
17/11/19 00:20:01 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applicable
17/11/19 00:20:02 INFO client.RMProxy: Connecting to ResourceManager at /0.0.0.0:8032
17/11/19 00:20:03 WARN mapreduce.JobSubmitter: Hadoop command-line option parsing not performed. Implement the Tool interface and execute your application with ToolRunn
er to remedy this.
17/11/19 00:20:03 INFO mapreduce.JobSubmitter: Cleaning up the staging area /tmp/hadoop-yarn/staging/acadgild/.staging/job_1510998304474_0004
Exception in thread "main" org.apache.hadoop.mapreduce.lib.input.InvalidInputException: Input path does not exist: hdfs://localhost:9000/user/acadgild/hadoop/Big_Data_S
ession1_Assignment_3_2_Task1/IOutput
        at org.apache.hadoop.mapreduce.lib.input.FileInputFormat.singleThreadedListStatus(FileInputFormat.java:321)
        at org.apache.hadoop.mapreduce.lib.input.FileInputFormat.listStatus(FileInputFormat.java:264)
        at org.apache.hadoop.mapreduce.lib.input.FileInputFormat.getSplits(FileInputFormat.java:385)
        at org.apache.hadoop.mapreduce.JobSubmitter.writeNewSplits(JobSubmitter.java:597)
        at org.apache.hadoop.mapreduce.JobSubmitter.writeSplits(JobSubmitter.java:614)
        at org.apache.hadoop.mapreduce.JobSubmitter.submitJobInternal(JobSubmitter.java:492)
        at org.apache.hadoop.mapreduce.Job$10.run(Job.java:1296)
        at org.apache.hadoop.mapreduce.Job$10.run(Job.java:1293)
        at java.security.AccessController.doPrivileged(Native Method)
        at javax.security.auth.Subject.doAs(Subject.java:422)
        at org.apache.hadoop.security.UserGroupInformation.doAs(UserGroupInformation.java:1628)
        at org.apache.hadoop.mapreduce.Job.submit(Job.java:1293)
        at org.apache.hadoop.mapreduce.Job.waitForCompletion(Job.java:1314)
        at Assignment_3_Task2.Task2.main(Task2.java:36)
        at sun.reflect.NativeMethodAccessorImpl.invoke0(Native Method)
        at sun.reflect.NativeMethodAccessorImpl.invoke(NativeMethodAccessorImpl.java:62)
        at sun.reflect.DelegatingMethodAccessorImpl.invoke(DelegatingMethodAccessorImpl.java:43)
        at java.lang.reflect.Method.invoke(Method.java:497)
        at org.apache.hadoop.util.RunJar.run(RunJar.java:221)
        at org.apache.hadoop.util.RunJar.main(RunJar.java:136)
[acadgild@localhost ~]$ hadoop jar /home/acadgild/Desktop/Assignment_3.3_Task2.jar Assignment_3_Task2.Task2 /user/acadgild/hadoop/Big_Data_Session1_Assignment_3_2_Task1
/Output/part-m-00000 /user/acadgild/hadoop/Big_Data_Session3_Assignment_3_3_Task2/OutputTask2
17/11/19 00:20:32 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applicable
17/11/19 00:20:33 INFO client.RMProxy: Connecting to ResourceManager at /0.0.0.0:8032
17/11/19 00:20:34 WARN mapreduce.JobSubmitter: Hadoop command-line option parsing not performed. Implement the Tool interface and execute your application with ToolRunn
er to remedy this.
17/11/19 00:20:34 INFO input.FileInputFormat: Total input paths to process : 1
17/11/19 00:20:35 INFO mapreduce.JobSubmitter: number of splits:1
17/11/19 00:20:35 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_1510998304474_0005
17/11/19 00:20:35 INFO impl.YarnClientImpl: Submitted application application_1510998304474_0005
17/11/19 00:20:35 INFO mapreduce.Job: The url to track the job: http://localhost:8088/proxy/application_1510998304474_0005/
17/11/19 00:20:35 INFO mapreduce.Job: Running job: job_1510998304474_0005
17/11/19 00:20:46 INFO mapreduce.Job: Job job_1510998304474_0005 running in uber mode : false
17/11/19 00:20:46 INFO mapreduce.Job:  map 0% reduce 0%
17/11/19 00:20:55 INFO mapreduce.Job:  map 100% reduce 0%
17/11/19 00:21:03 INFO mapreduce.Job:  map 100% reduce 100%
17/11/19 00:21:04 INFO mapreduce.Job: Job job_1510998304474_0005 completed successfully
```

```
[acadgild@localhost ~]$ hadoop jar /home/acadgild/Desktop/Assignment_3.3_Task2.jar Assignment_3_Task2.Task2
/user/acadgild/hadoop/Big_Data_Session1_Assignment_3_2_Task1/IOutput/ part-m-00000
/user/acadgild/hadoop/Big_Data_Session3_Assignment_3_3_Task2/OutputTask2

17/11/19 00:20:01 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where
applicable

17/11/19 00:20:02 INFO client.RMProxy: Connecting to ResourceManager at /0.0.0.0:8032

17/11/19 00:20:03 WARN mapreduce.JobSubmitter: Hadoop command-line option parsing not performed. Implement the Tool interface and
execute your application with ToolRunner to remedy this.

17/11/19 00:20:03 INFO mapreduce.JobSubmitter: Cleaning up the staging area /tmp/hadoop-
yarn/staging/acadgild/.staging/job_1510998304474_0004

Exception in thread "main" org.apache.hadoop.mapreduce.lib.input.InvalidInputException: Input path does not exist:
hdfs://localhost:9000/user/acadgild/hadoop/Big_Data_Session1_Assignment_3_2_Task1/IOutput

        at org.apache.hadoop.mapreduce.lib.input.FileInputFormat.singleThreadedListStatus(FileInputFormat.java:321)

        at org.apache.hadoop.mapreduce.lib.input.FileInputFormat.listStatus(FileInputFormat.java:264)

        at org.apache.hadoop.mapreduce.lib.input.FileInputFormat.getSplits(FileInputFormat.java:385)

        at org.apache.hadoop.mapreduce.JobSubmitter.writeNewSplits(JobSubmitter.java:597)
        at org.apache.hadoop.mapreduce.JobSubmitter.writeSplits(JobSubmitter.java:614)

        at org.apache.hadoop.mapreduce.JobSubmitter.submitJobInternal(JobSubmitter.java:492)

        at org.apache.hadoop.mapreduce.Job$10.run(Job.java:1296)

        at org.apache.hadoop.mapreduce.Job$10.run(Job.java:1293)

        at java.security.AccessController.doPrivileged(Native Method)

        at javax.security.auth.Subject.doAs(Subject.java:422)

        at org.apache.hadoop.security.UserGroupInformation.doAs(UserGroupInformation.java:1628)

        at org.apache.hadoop.mapreduce.Job.submit(Job.java:1293)

        at org.apache.hadoop.mapreduce.Job.waitForCompletion(Job.java:1314)

        at Assignment_3_Task2.Task2.main(Task2.java:36)

        at sun.reflect.NativeMethodAccessorImpl.invoke0(Native Method)

        at sun.reflect.NativeMethodAccessorImpl.invoke(NativeMethodAccessorImpl.java:62)

        at sun.reflect.DelegatingMethodAccessorImpl.invoke(DelegatingMethodAccessorImpl.java:43)

        at java.lang.reflect.Method.invoke(Method.java:497)

        at org.apache.hadoop.util.RunJar.run(RunJar.java:221)

        at org.apache.hadoop.util.RunJar.main(RunJar.java:136)

[acadgild@localhost ~]$ hadoop jar /home/acadgild/Desktop/Assignment_3.3_Task2.jar Assignment_3_Task2.Task2
/user/acadgild/hadoop/Big_Data_Session1_Assignment_3_2_Task1/Output/part-m-00000
/user/acadgild/hadoop/Big_Data_Session3_Assignment_3_3_Task2/OutputTask2

17/11/19 00:20:32 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where
applicable

17/11/19 00:20:33 INFO client.RMProxy: Connecting to ResourceManager at /0.0.0.0:8032

17/11/19 00:20:34 WARN mapreduce.JobSubmitter: Hadoop command-line option parsing not performed. Implement the Tool interface and
execute your application with ToolRunner to remedy this.

17/11/19 00:20:34 INFO input.FileInputFormat: Total input paths to process : 1

17/11/19 00:20:35 INFO mapreduce.JobSubmitter: number of splits:1

17/11/19 00:20:35 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_1510998304474_0005

17/11/19 00:20:35 INFO impl.YarnClientImpl: Submitted application application_1510998304474_0005

17/11/19 00:20:35 INFO mapreduce.Job: The url to track the job: http://localhost:8088/proxy/application_1510998304474_0005/

17/11/19 00:20:35 INFO mapreduce.Job: Running job: job_1510998304474_0005

17/11/19 00:20:46 INFO mapreduce.Job: Job job_1510998304474_0005 running in uber mode : false

17/11/19 00:20:46 INFO mapreduce.Job:  map 0% reduce 0%

17/11/19 00:20:55 INFO mapreduce.Job:  map 100% reduce 0%

17/11/19 00:21:03 INFO mapreduce.Job:  map 100% reduce 100%

17/11/19 00:21:04 INFO mapreduce.Job: Job job_1510998304474_0005 completed successfully

17/11/19 00:21:04 INFO mapreduce.Job: Counters: 49

        File System Counters

                FILE: Number of bytes read=204

                FILE: Number of bytes written=213453

                FILE: Number of read operations=0
                FILE: Number of large read operations=0
                FILE: Number of write operations=0
                HDFS: Number of bytes read=828
                HDFS: Number of bytes written=38
                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
                Data-local map tasks=1
                Total time spent by all maps in occupied slots (ms)=5715
                Total time spent by all reduces in occupied slots (ms)=5418
                Total time spent by all map tasks (ms)=5715
                Total time spent by all reduce tasks (ms)=5418
                Total vcore-seconds taken by all map tasks=5715
                Total vcore-seconds taken by all reduce tasks=5418
                Total megabyte-seconds taken by all map tasks=5852160
                Total megabyte-seconds taken by all reduce tasks=5548032
        Map-Reduce Framework
                Map input records=16
                Map output records=16
                Map output bytes=166
                Map output materialized bytes=204
                Input split bytes=166
                Combine input records=0
                Combine output records=0
                Reduce input groups=5
                Reduce shuffle bytes=204
                Reduce input records=16
                Reduce output records=5
                Spilled Records=32
                Shuffled Maps =1
                Failed Shuffles=0
                Merged Map outputs=1
                GC time elapsed (ms)=164
                CPU time spent (ms)=1800
                Physical memory (bytes) snapshot=297373696
                Virtual memory (bytes) snapshot=4115832832
                Total committed heap usage (bytes)=165810176
        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=662
        File Output Format Counters
                Bytes Written=38
```

Step 5:

After execution, the result will be stored on HDFS location:-

/user/acadgild/hadoop/Big_Data_Session3_Assignment_3_3_Task2/OutputTask2



```
[acadgild@localhost ~]$ hdfs dfs -ls /user/acadgild/hadoop/Big_Data_Session3_Assignment_3_3_Task2/OutputTask2
17/11/19 00:25:12 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applicable
Found 2 items
-rw-r--r--   1 acadgild supergroup          0 2017-11-19 00:21 /user/acadgild/hadoop/Big_Data_Session3_Assignment_3_3_Task2/OutputTask2/_SUCCESS
-rw-r--r--   1 acadgild supergroup         38 2017-11-19 00:21 /user/acadgild/hadoop/Big_Data_Session3_Assignment_3_3_Task2/OutputTask2/part-r-00000
[acadgild@localhost ~]$
[acadgild@localhost ~]$ hdfs dfs -cat /user/acadgild/hadoop/Big_Data_Session3_Assignment_3_3_Task2/OutputTask2/part-r-00000
17/11/19 00:25:41 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applicable
Akai    1
Lava    3
Onida   3
Samsung 7
Zen     2
[acadgild@localhost ~]$
```

part-r-00000

Akai 1

Lava 3

Onida 3

Samsung 7

Zen 2

Task3

Mapper Code: Task3Mapper.java

```
/*
 * All the comments in the program are highlighted in Green.
 * @author Sahil Khurana <sahilkhurana369@gmail.com
 */
```

```
package Assignment_3_Task3;                                                    // Package Declared

import java.io.IOException;                                                    // IOException class import to handle exceptions

import org.apache.hadoop.io.IntWritable;

import org.apache.hadoop.io.LongWritable;

import org.apache.hadoop.io.Text;

import org.apache.hadoop.mapreduce.*;

// class is extended to have the arguments keyIn as LongWritable and ValueIn as Text and KeyOut as Text and ValueOut as IntWritable.

public class Task3Mapper extends Mapper<LongWritable, Text, Text, IntWritable> {

// The variables declared are: a Text variable Company_Name_Key

Text Company_Name_Key = new Text();

// The variables declared are: a IntWritable variable Company_Name_Value which will store the value of the MapReduce deals with Key and
Value pairs.

IntWritable Company_Name_Value = new IntWritable();

// overriding the map method which will run one time for every line.

@Override

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

// storing the line in a string variable and splitting the line by using comma "|" delimiter and storing the values in a String Array so that all the
columns in a row are stored in the string array.

String[] lineArray = value.toString().split("\\|");

// to obtain key and value of ONIDA

if(lineArray[0].equalsIgnoreCase("ONIDA")) {

Onida_Key.set("ONIDA" + "\t" + lineArray[3]);

Onida_Value.set(1);

context.write(Onida_Key, Onida_Value);        }      // obtaining key and value context.

              }    // map class closed

}   // class Task2Mapper closed
```

Reducer Code: Task3Reducer.java

```
/*
 * All the comments in the program are highlighted in Green.
 * @author Sahil Khurana <sahilkhurana369@gmail.com
 */
```

```
package Assignment_3_Task3;                                   // Package declared

import java.io.IOException;

import org.apache.hadoop.io.IntWritable;

import org.apache.hadoop.io.Text;

import org.apache.hadoop.mapreduce.Reducer;

// class is extended to have the arguments keyIn as LongWritable and ValueIn as Text and KeyOut as Text and ValueOut as IntWritable.

public class Task3Reducer    extends Reducer<Text, IntWritable, Text, IntWritable> {

// The variables declared are: a IntWritable variable Company_Name_Value which will store the value of the MapReduce deals with Key and Value pairs.

IntWritable Company_Name_Value = new IntWritable();

// The variables declared are: a IntWritable variable Company_Name_Value which will store the value of the MapReduce deals with Key and Value pairs.

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

// declaring an integer sum which will store the sum of all the Company_Name_Value

int sum = 0;

// for each loop is taken which will run each time for the values inside the "Iterable values" which are coming from the shuffle and sort phase after the mapper phase.

for (IntWritable value : values) {

sum += value.get();}                                          // storing and calculating the sum of the values.

Company_Name_Value.set(sum);                                  // setting the sum

context.write(key, Company_Name_Value);                       // obtaining key and value context.

            }                                                 // map class closed

}                                                             // class Task2Reducer closed
```

Driver Code: Task3.java

```
/*
 * All the comments in the program are highlighted in Green.
 * @author Sahil Khurana <sahilkhurana369@gmail.com
 */
```

```
package Assignment_3_Task3;                          // Package declared

import org.apache.hadoop.conf.Configuration;         // Import the Configuration of system parameters.

import org.apache.hadoop.fs.Path;                    // used to Names a file or directory in a AbstractFileSystem for hdfs.

import org.apache.hadoop.io.Text;                    // This class stores text using standard UTF8 encoding.
import org.apache.hadoop.mapreduce.Job;                              // The job submitter's view of the Job.

import org.apache.hadoop.mapreduce.lib.input.FileInputFormat;        // FilterInputFormat is a convenience class that wraps InputFormat.

import org.apache.hadoop.mapreduce.lib.output.FileOutputFormat;      // FilterOutputFormat is a convenience class that wraps    OutputFormat.

import org.apache.hadoop.mapreduce.lib.input.TextInputFormat;        // TextInputFormat is a convenience class that wraps InputFormat.

import org.apache.hadoop.mapreduce.lib.output.TextOutputFormat;      // TextOutputFormat is a convenience class that wraps OutputFormat.

import org.apache.hadoop.io.IntWritable;

public class Task3 {                                                 // class declared

@SuppressWarnings("deprecation")                                     // The @SuppressWarnings annotation disables certain compiler

                                                                     // warnings. In this case, the warning about deprecated code ("deprecation")

public static void main(String[] args) throws Exception {            // main class started

Configuration conf = new Configuration();                            // Create a configuration object for the job

Job job = new Job(conf, "Assignment_3.3_Task3");                     // created the new object name Job

job.setJarByClass(Task3.class);                                      // Set a name of the Jar

job.setMapOutputKeyClass(Text.class);                                // Set the output Key type for the Mapper

job.setMapOutputValueClass(IntWritable.class);                       // Set the output Value type for the Mapper

job.setOutputKeyClass(Text.class);                                   // Set the output Key type for the Reducer

job.setOutputValueClass(IntWritable.class);                          // Set the output Value type for the Reducer

job.setMapperClass(Task2Mapper.class);                               // Set the Mapper Class

job.setReducerClass(Task2Reducer.class);                             // Set the Reducer Class

job.setInputFormatClass(TextInputFormat.class);                      // Set the format of the input that will be provided to the program

job.setOutputFormatClass(TextOutputFormat.class);                    // Set the format of the output for the program

FileInputFormat.addInputPath(job, new Path(args[0]));                // Set input directories using command line arguments,arg[0] on HDFS

FileOutputFormat.setOutputPath(job, new Path(args[1]));              // Set output directories using command line arguments,arg[0]  on HDFS

job.waitForCompletion(true);                                         // Submit the job, then poll for progress until the job is complete

  }                                                                  // main method closed

}                                                                    // class Task1 closed
```

Steps to Run Mapreduce program on command prompt for Task3

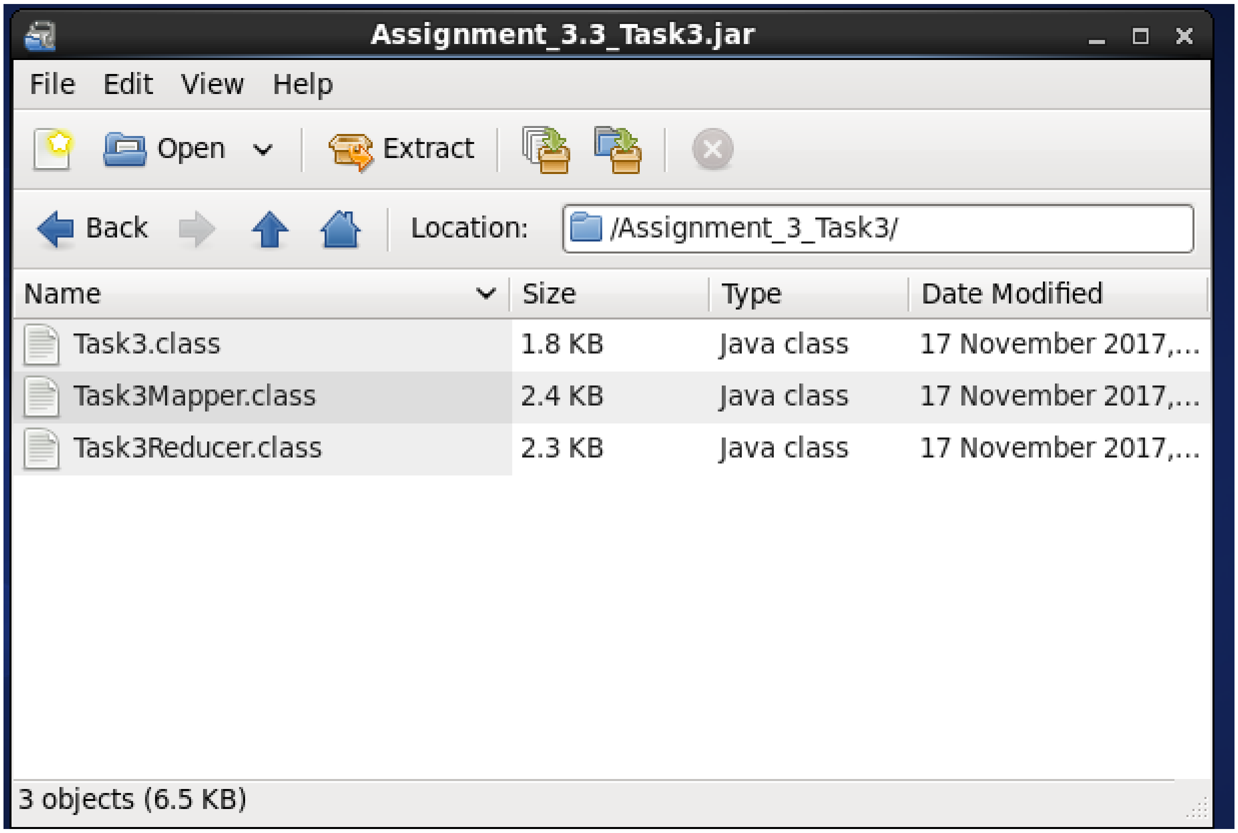
Step 1:

Create a Jar, which will contain

Driver Code: Task3.java

Mapper Code: Task3Mapper.java

Reducer Code: Task3Reducer.java



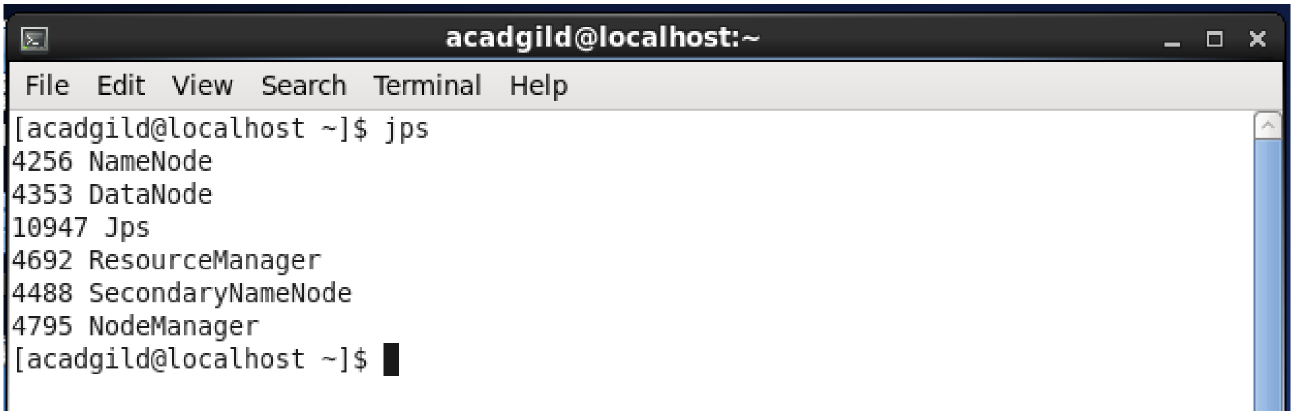
/Assignment_3_Task3/

Name	Size	Type	Date Modified
Task3.class	1.8 KB	Java class	17 November 2017,...
Task3Mapper.class	2.4 KB	Java class	17 November 2017,...
Task3Reducer.class	2.3 KB	Java class	17 November 2017,...

3 objects (6.5 KB)

Step 2:

Check the all the Hadoop services are running or not by typing jps command

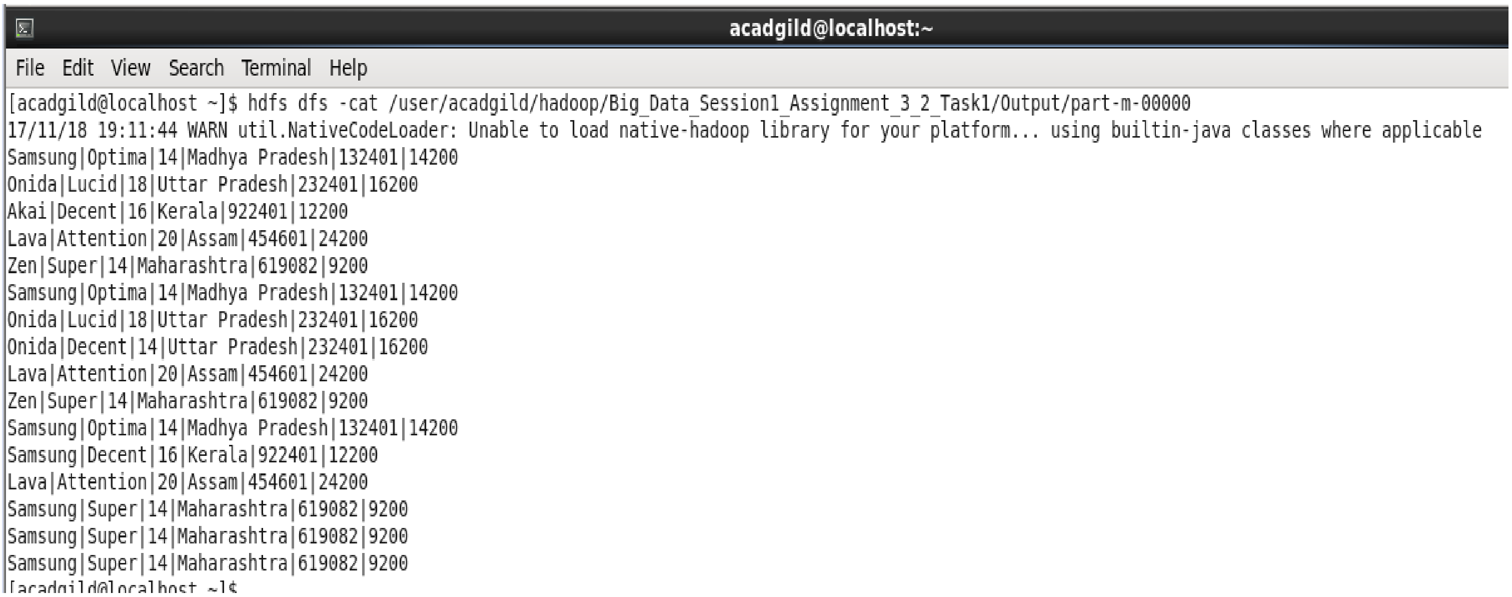


```
[acadgild@localhost ~]$ jps
4256 NameNode
4353 DataNode
10947 Jps
4692 ResourceManager
4488 SecondaryNameNode
4795 NodeManager
[acadgild@localhost ~]$
```

Step 3:

The output of Assignment 3.2 will be taken as input of Assignment 3.3

hdfs dfs -cat /user/acadgild/hadoop/Big_Data_Session1_Assignment_3_2_Task1/Output/part-m-00000



```
[acadgild@localhost ~]$ hdfs dfs -cat /user/acadgild/hadoop/Big_Data_Session1_Assignment_3_2_Task1/Output/part-m-00000
17/11/18 19:11:44 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applicable
Samsung|Optima|14|Madhya Pradesh|132401|14200
Onida|Lucid|18|Uttar Pradesh|232401|16200
Akai|Decent|16|Kerala|922401|12200
Lava|Attention|20|Assam|454601|24200
Zen|Super|14|Maharashtra|619082|9200
Samsung|Optima|14|Madhya Pradesh|132401|14200
Onida|Lucid|18|Uttar Pradesh|232401|16200
Onida|Decent|14|Uttar Pradesh|232401|16200
Lava|Attention|20|Assam|454601|24200
Zen|Super|14|Maharashtra|619082|9200
Samsung|Optima|14|Madhya Pradesh|132401|14200
Samsung|Decent|16|Kerala|922401|12200
Lava|Attention|20|Assam|454601|24200
Samsung|Super|14|Maharashtra|619082|9200
Samsung|Super|14|Maharashtra|619082|9200
Samsung|Super|14|Maharashtra|619082|9200
[acadgild@localhost ~]$
```

Step 4:

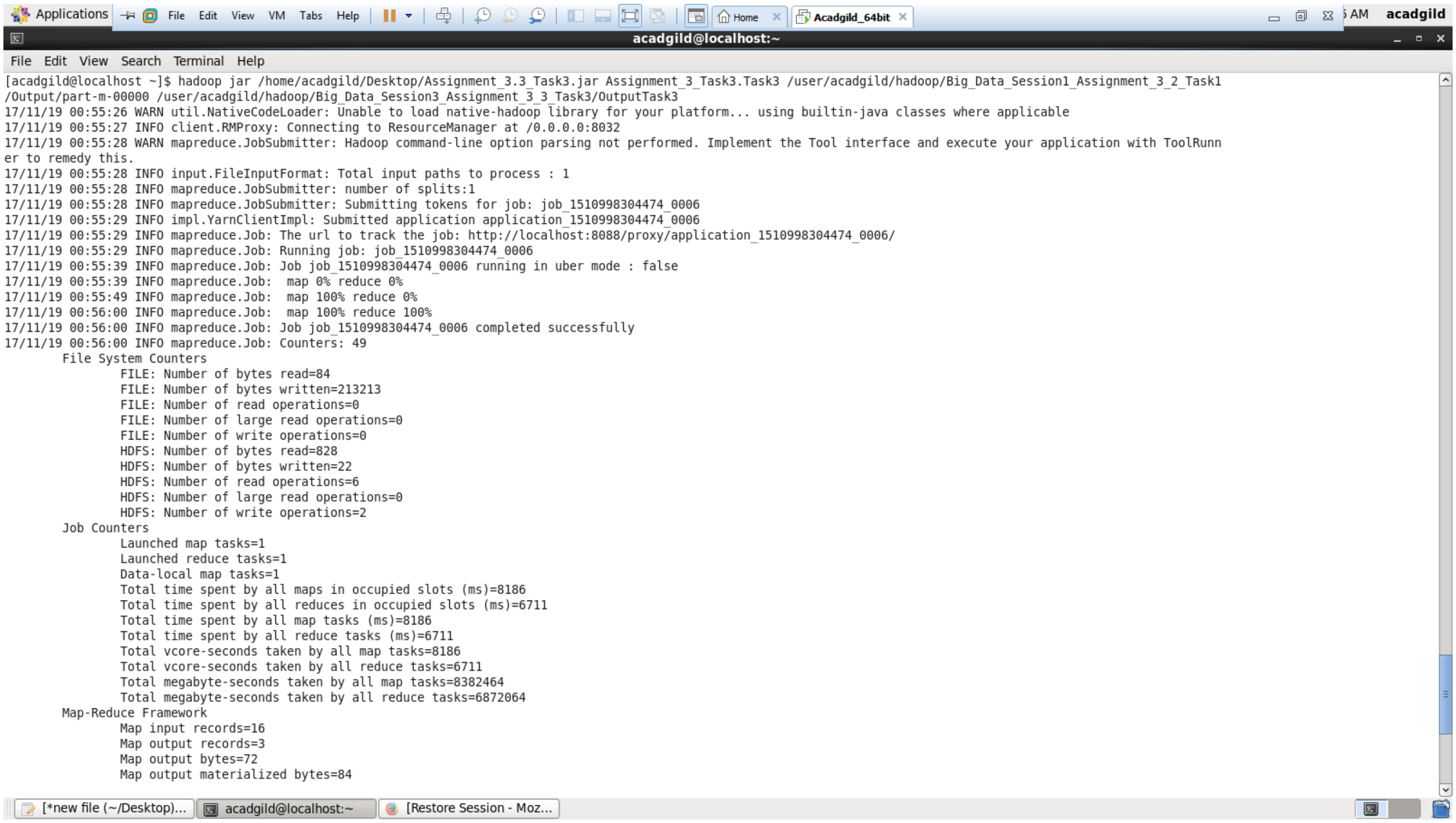
Run the MapReduce Job

Generic Command:-

hadoop jar <JAR file Path > <PackageName.MainClass> <Inputfile> <outputDir>

In case of Assignment 3.3 Task2:-

hadoop jar /home/acadgild/Desktop/Assignment_3.3_Task3.jar Assignment_3_Task3.Task3 /user/acadgild/hadoop/Big_Data_Session1_Assignment_3_2_Task1/Output/part-m-00000 /user/acadgild/hadoop/Big_Data_Session3_Assignment_3_3_Task3/OutputTask3


Applications
File Edit View VM Tabs Help
Home
Acadgild_64bit
AM
acadgild
acadgild@localhost:~
File Edit View Search Terminal Help
[*new file (~/Desktop)...
acadgild@localhost:~
[Restore Session - Moz...

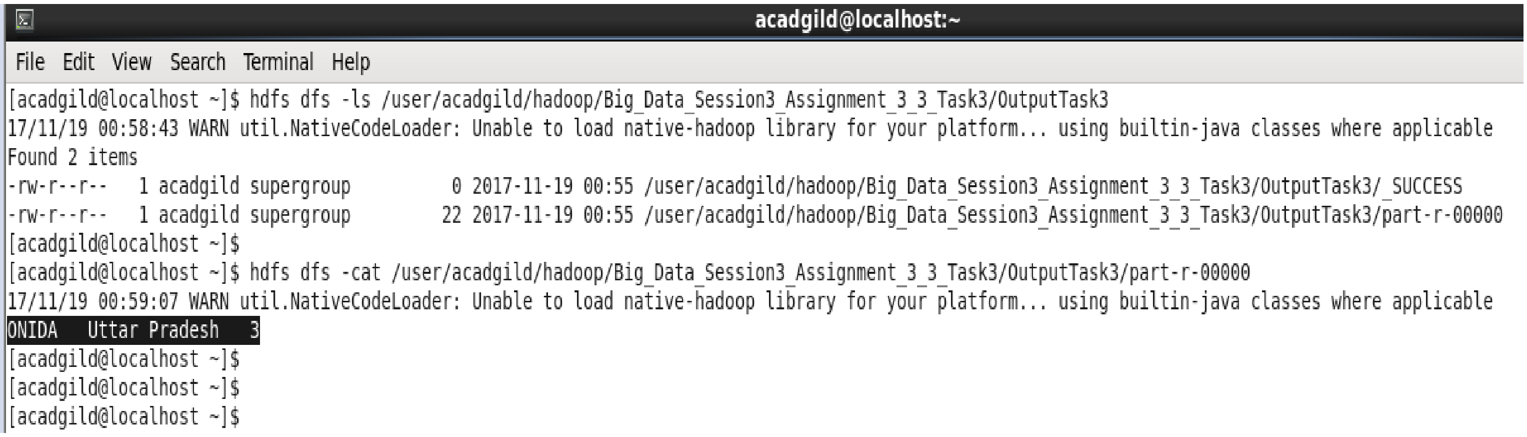
```
[acadgild@localhost ~]$ hadoop jar /home/acadgild/Desktop/Assignment_3.3_Task3.jar Assignment_3_Task3.Task3 /user/acadgild/hadoop/Big_Data_Session1_Assignment_3_2_Task1
/Output/part-m-00000 /user/acadgild/hadoop/Big_Data_Session3_Assignment_3_3_Task3/OutputTask3
17/11/19 00:55:26 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applicable
17/11/19 00:55:27 INFO client.RMProxy: Connecting to ResourceManager at /0.0.0.0:8032
17/11/19 00:55:28 WARN mapreduce.JobSubmitter: Hadoop command-line option parsing not performed. Implement the Tool interface and execute your application with ToolRunn
er to remedy this.
17/11/19 00:55:28 INFO input.FileInputFormat: Total input paths to process : 1
17/11/19 00:55:28 INFO mapreduce.JobSubmitter: number of splits:1
17/11/19 00:55:28 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_1510998304474_0006
17/11/19 00:55:29 INFO impl.YarnClientImpl: Submitted application application_1510998304474_0006
17/11/19 00:55:29 INFO mapreduce.Job: The url to track the job: http://localhost:8088/proxy/application_1510998304474_0006/
17/11/19 00:55:29 INFO mapreduce.Job: Running job: job_1510998304474_0006
17/11/19 00:55:39 INFO mapreduce.Job: Job job_1510998304474_0006 running in uber mode : false
17/11/19 00:55:39 INFO mapreduce.Job:  map 0% reduce 0%
17/11/19 00:55:49 INFO mapreduce.Job:  map 100% reduce 0%
17/11/19 00:56:00 INFO mapreduce.Job:  map 100% reduce 100%
17/11/19 00:56:00 INFO mapreduce.Job: Job job_1510998304474_0006 completed successfully
17/11/19 00:56:00 INFO mapreduce.Job: Counters: 49
        File System Counters
                FILE: Number of bytes read=84
                FILE: Number of bytes written=213213
                FILE: Number of read operations=0
                FILE: Number of large read operations=0
                FILE: Number of write operations=0
                HDFS: Number of bytes read=828
                HDFS: Number of bytes written=22
                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
                Data-local map tasks=1
                Total time spent by all maps in occupied slots (ms)=8186
                Total time spent by all reduces in occupied slots (ms)=6711
                Total time spent by all map tasks (ms)=8186
                Total time spent by all reduce tasks (ms)=6711
                Total vcore-seconds taken by all map tasks=8186
                Total vcore-seconds taken by all reduce tasks=6711
                Total megabyte-seconds taken by all map tasks=8382464
                Total megabyte-seconds taken by all reduce tasks=6872064
        Map-Reduce Framework
                Map input records=16
                Map output records=3
                Map output bytes=72
                Map output materialized bytes=84
```

```
[acadgild@localhost ~]$ hadoop jar /home/acadgild/Desktop/Assignment_3.3_Task3.jar Assignment_3_Task3.Task3
/user/acadgild/hadoop/Big_Data_Session1_Assignment_3_2_Task1/Output/part-m-00000
/user/acadgild/hadoop/Big_Data_Session3_Assignment_3_3_Task3/OutputTask3

17/11/19 00:55:26 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where
applicable

17/11/19 00:55:27 INFO client.RMProxy: Connecting to ResourceManager at /0.0.0.0:8032

17/11/19 00:55:28 WARN mapreduce.JobSubmitter: Hadoop command-line option parsing not performed. Implement the Tool interface and
execute your application with ToolRunner to remedy this.

17/11/19 00:55:28 INFO input.FileInputFormat: Total input paths to process : 1

17/11/19 00:55:28 INFO mapreduce.JobSubmitter: number of splits:1

17/11/19 00:55:28 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_1510998304474_0006

17/11/19 00:55:29 INFO impl.YarnClientImpl: Submitted application application_1510998304474_0006

17/11/19 00:55:29 INFO mapreduce.Job: The url to track the job: http://localhost:8088/proxy/application_1510998304474_0006/

17/11/19 00:55:29 INFO mapreduce.Job: Running job: job_1510998304474_0006

17/11/19 00:55:39 INFO mapreduce.Job: Job job_1510998304474_0006 running in uber mode : false

17/11/19 00:55:39 INFO mapreduce.Job:  map 0% reduce 0%

17/11/19 00:55:49 INFO mapreduce.Job:  map 100% reduce 0%

17/11/19 00:56:00 INFO mapreduce.Job:  map 100% reduce 100%

17/11/19 00:56:00 INFO mapreduce.Job: Job job_1510998304474_0006 completed successfully

17/11/19 00:56:00 INFO mapreduce.Job: Counters: 49

        File System Counters

                FILE: Number of bytes read=84

                FILE: Number of bytes written=213213

                FILE: Number of read operations=0

                FILE: Number of large read operations=0

                FILE: Number of write operations=0

                HDFS: Number of bytes read=828

                HDFS: Number of bytes written=22

                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

                Data-local map tasks=1

                Total time spent by all maps in occupied slots (ms)=8186

                Total time spent by all reduces in occupied slots (ms)=6711

                Total time spent by all map tasks (ms)=8186

                Total time spent by all reduce tasks (ms)=6711

                Total vcore-seconds taken by all map tasks=8186
                Total vcore-seconds taken by all reduce tasks=6711
                Total megabyte-seconds taken by all map tasks=8382464
                Total megabyte-seconds taken by all reduce tasks=6872064
        Map-Reduce Framework
                Map input records=16
                Map output records=3
                Map output bytes=72
                Map output materialized bytes=84
                Input split bytes=166
                Combine input records=0
                Combine output records=0
                Reduce input groups=1
                Reduce shuffle bytes=84
                Reduce input records=3
                Reduce output records=1
                Spilled Records=6
                Shuffled Maps =1
                Failed Shuffles=0
                Merged Map outputs=1
                GC time elapsed (ms)=458
                CPU time spent (ms)=2590
                Physical memory (bytes) snapshot=296165376
                Virtual memory (bytes) snapshot=4113870848
                Total committed heap usage (bytes)=165810176
        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=662
        File Output Format Counters
                Bytes Written=22
```

Step 5:

After execution, the result will be stored on HDFS location:-

/user/acadgild/hadoop/Big_Data_Session3_Assignment_3_3_Task3/OutputTask3



```
[acadgild@localhost ~]$ hdfs dfs -ls /user/acadgild/hadoop/Big_Data_Session3_Assignment_3_3_Task3/OutputTask3
17/11/19 00:58:43 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applicable
Found 2 items
-rw-r--r--   1 acadgild supergroup          0 2017-11-19 00:55 /user/acadgild/hadoop/Big_Data_Session3_Assignment_3_3_Task3/OutputTask3/_SUCCESS
-rw-r--r--   1 acadgild supergroup         22 2017-11-19 00:55 /user/acadgild/hadoop/Big_Data_Session3_Assignment_3_3_Task3/OutputTask3/part-r-00000
[acadgild@localhost ~]$
[acadgild@localhost ~]$ hdfs dfs -cat /user/acadgild/hadoop/Big_Data_Session3_Assignment_3_3_Task3/OutputTask3/part-r-00000
17/11/19 00:59:07 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applicable
ONIDA   Uttar Pradesh   3
[acadgild@localhost ~]$
[acadgild@localhost ~]$
[acadgild@localhost ~]$
```

part-r-00000

ONIDA   Uttar Pradesh 3