LAB 04
Apache Spark

Name: MUHAMAD RAFIQ IQBAL BIN SAMSUDIN
Matric :17206926

A . HDFS DATA UPLOAD

1. Make Input file name Sparkdata using command

   hdfs dfs -mkdir /Sparkdata

2. Check the list of files in hdfs

```
[cloudera@quickstart Desktop]$ hdfs dfs -ls /
Found 13 items
drwxr-xr-x   - cloudera supergroup          0 2024-11-10 20:00 /Sparkdata
drwxrwxrwx   - hdfs     supergroup          0 2017-04-05 04:27 /benchmarks
drwxr-xr-x   - hbase    supergroup          0 2024-10-27 23:24 /hbase
drwxr-xr-x   - cloudera supergroup          0 2024-10-18 07:27 /inputfolder1
drwxr-xr-x   - cloudera supergroup          0 2024-10-21 00:21 /inputfolder2
drwxr-xr-x   - cloudera supergroup          0 2024-11-10 20:09 /my_spark_output1
drwxr-xr-x   - cloudera supergroup          0 2024-10-18 07:30 /out1
drwxr-xr-x   - cloudera supergroup          0 2024-10-21 00:23 /out2
drwxr-xr-x   - solr     solr                0 2017-04-05 04:29 /solr
-rw-r--r--   1 cloudera supergroup    4418139 2024-11-10 19:10 /sparkdata
drwxrwxrwt   - hdfs     supergroup          0 2024-10-17 01:46 /tmp
drwxr-xr-x   - hdfs     supergroup          0 2017-04-05 04:29 /user
drwxr-xr-x   - hdfs     supergroup          0 2017-04-05 04:29 /var
[cloudera@quickstart Desktop]$ hdfs dfs-mkdir /Sparkdata
```

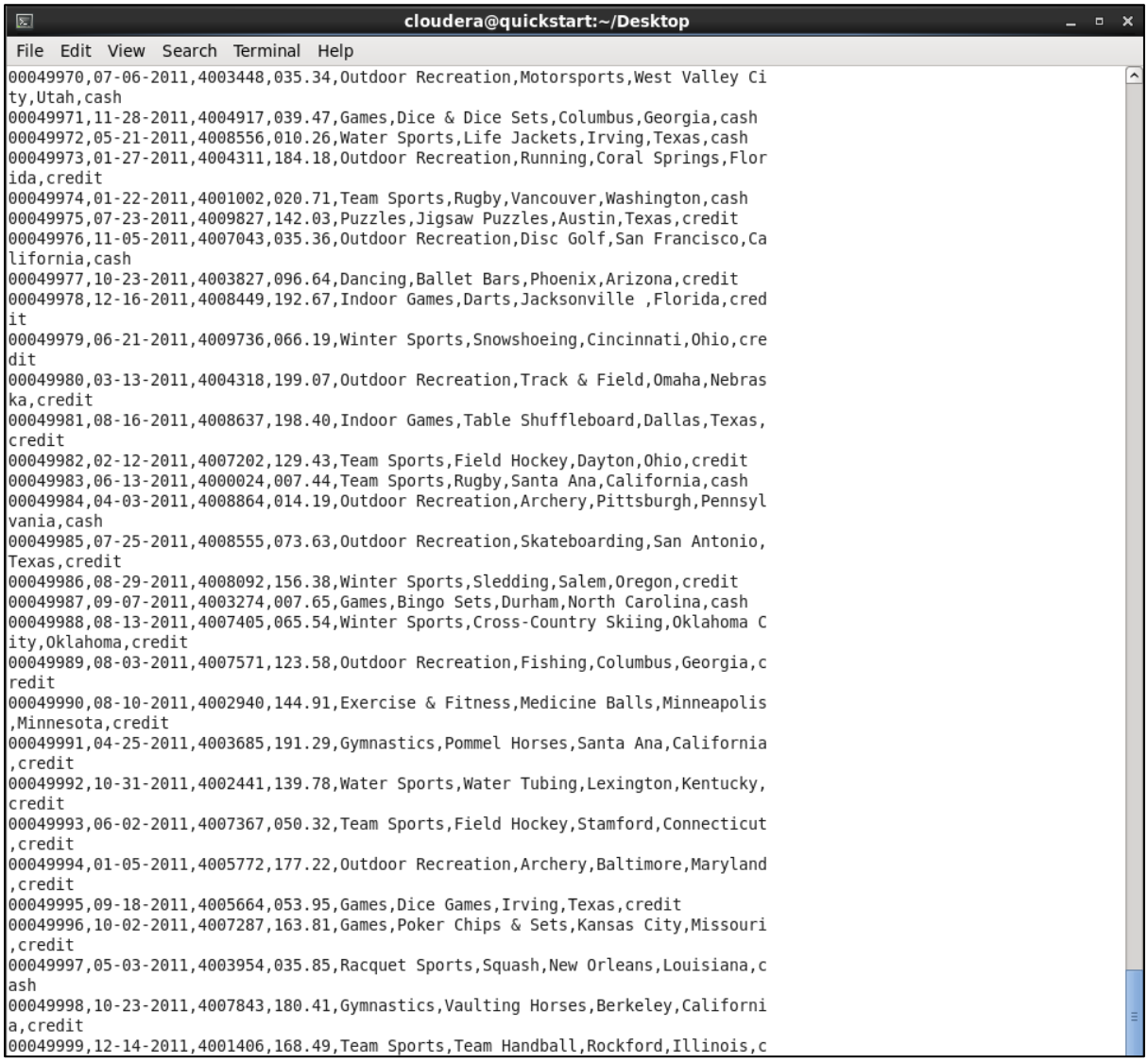
3. Upload the hivedata.txt into the input file (Sparkdata) by using command below.

```
[cloudera@quickstart Desktop]$ hdfs dfs -put /home/cloudera/Desktop/hivedata.txt /Sparkdata
```

4. Check the file content in Sparkdata using command below.

```
[cloudera@quickstart Desktop]$ hdfs dfs -ls /
Found 13 items
drwxr-xr-x   - cloudera supergroup          0 2024-11-10 20:00 /Sparkdata
drwxrwxrwx   - hdfs     supergroup          0 2017-04-05 04:27 /benchmarks
drwxr-xr-x   - hbase    supergroup          0 2024-10-27 23:24 /hbase
drwxr-xr-x   - cloudera supergroup          0 2024-10-18 07:27 /inputfolder1
drwxr-xr-x   - cloudera supergroup          0 2024-10-21 00:21 /inputfolder2
drwxr-xr-x   - cloudera supergroup          0 2024-11-10 20:09 /my_spark_output1
drwxr-xr-x   - cloudera supergroup          0 2024-10-18 07:30 /out1
drwxr-xr-x   - cloudera supergroup          0 2024-10-21 00:23 /out2
drwxr-xr-x   - solr     solr                0 2017-04-05 04:29 /solr
-rw-r--r--   1 cloudera supergroup    4418139 2024-11-10 19:10 /sparkdata
drwxrwxrwt   - hdfs     supergroup          0 2024-10-17 01:46 /tmp
drwxr-xr-x   - hdfs     supergroup          0 2017-04-05 04:29 /user
drwxr-xr-x   - hdfs     supergroup          0 2017-04-05 04:29 /var
[cloudera@quickstart Desktop]$ hdfs dfs -cat /Sparkdata
cat: `/Sparkdata': Is a directory
[cloudera@quickstart Desktop]$ hdfs dfs -cat /Sparkdata
cat: `/Sparkdata': Is a directory
[cloudera@quickstart Desktop]$ hdfs dfs -cat /Sparkdata/Hivedata.txt
cat: `/Sparkdata/Hivedata.txt': No such file or directory
[cloudera@quickstart Desktop]$ hdfs dfs -cat /Sparkdata/hivedata.txt
```

The output :


cloudera@quickstart:~/Desktop
File Edit View Search Terminal Help
00049970,07-06-2011,4003448,035.34,Outdoor Recreation,Motorsports,West Valley Ci
ty,Utah,cash
00049971,11-28-2011,4004917,039.47,Games,Dice & Dice Sets,Columbus,Georgia,cash
00049972,05-21-2011,4008556,010.26,Water Sports,Life Jackets,Irving,Texas,cash
00049973,01-27-2011,4004311,184.18,Outdoor Recreation,Running,Coral Springs,Flor
ida,credit
00049974,01-22-2011,4001002,020.71,Team Sports,Rugby,Vancouver,Washington,cash
00049975,07-23-2011,4009827,142.03,Puzzles,Jigsaw Puzzles,Austin,Texas,credit
00049976,11-05-2011,4007043,035.36,Outdoor Recreation,Disc Golf,San Francisco,Ca
lifornia,cash
00049977,10-23-2011,4003827,096.64,Dancing,Ballet Bars,Phoenix,Arizona,credit
00049978,12-16-2011,4008449,192.67,Indoor Games,Darts,Jacksonville ,Florida,cred
it
00049979,06-21-2011,4009736,066.19,Winter Sports,Snowshoeing,Cincinnati,Ohio,cre
dit
00049980,03-13-2011,4004318,199.07,Outdoor Recreation,Track & Field,Omaha,Nebras
ka,credit
00049981,08-16-2011,4008637,198.40,Indoor Games,Table Shuffleboard,Dallas,Texas,
credit
00049982,02-12-2011,4007202,129.43,Team Sports,Field Hockey,Dayton,Ohio,credit
00049983,06-13-2011,4000024,007.44,Team Sports,Rugby,Santa Ana,California,cash
00049984,04-03-2011,4008864,014.19,Outdoor Recreation,Archery,Pittsburgh,Pennsyl
vania,cash
00049985,07-25-2011,4008555,073.63,Outdoor Recreation,Skateboarding,San Antonio,
Texas,credit
00049986,08-29-2011,4008092,156.38,Winter Sports,Sledding,Salem,Oregon,credit
00049987,09-07-2011,4003274,007.65,Games,Bingo Sets,Durham,North Carolina,cash
00049988,08-13-2011,4007405,065.54,Winter Sports,Cross-Country Skiing,Oklahoma C
ity,Oklahoma,credit
00049989,08-03-2011,4007571,123.58,Outdoor Recreation,Fishing,Columbus,Georgia,c
redit
00049990,08-10-2011,4002940,144.91,Exercise & Fitness,Medicine Balls,Minneapolis
,Minnesota,credit
00049991,04-25-2011,4003685,191.29,Gymnastics,Pommel Horses,Santa Ana,California
,credit
00049992,10-31-2011,4002441,139.78,Water Sports,Water Tubing,Lexington,Kentucky,
credit
00049993,06-02-2011,4007367,050.32,Team Sports,Field Hockey,Stamford,Connecticut
,credit
00049994,01-05-2011,4005772,177.22,Outdoor Recreation,Archery,Baltimore,Maryland
,credit
00049995,09-18-2011,4005664,053.95,Games,Dice Games,Irving,Texas,credit
00049996,10-02-2011,4007287,163.81,Games,Poker Chips & Sets,Kansas City,Missouri
,credit
00049997,05-03-2011,4003954,035.85,Racquet Sports,Squash,New Orleans,Louisiana,c
ash
00049998,10-23-2011,4007843,180.41,Gymnastics,Vaulting Horses,Berkeley,Californi
a,credit
00049999,12-14-2011,4001406,168.49,Team Sports,Team Handball,Rockford,Illinois,c

B . Command Execution of MapReduce

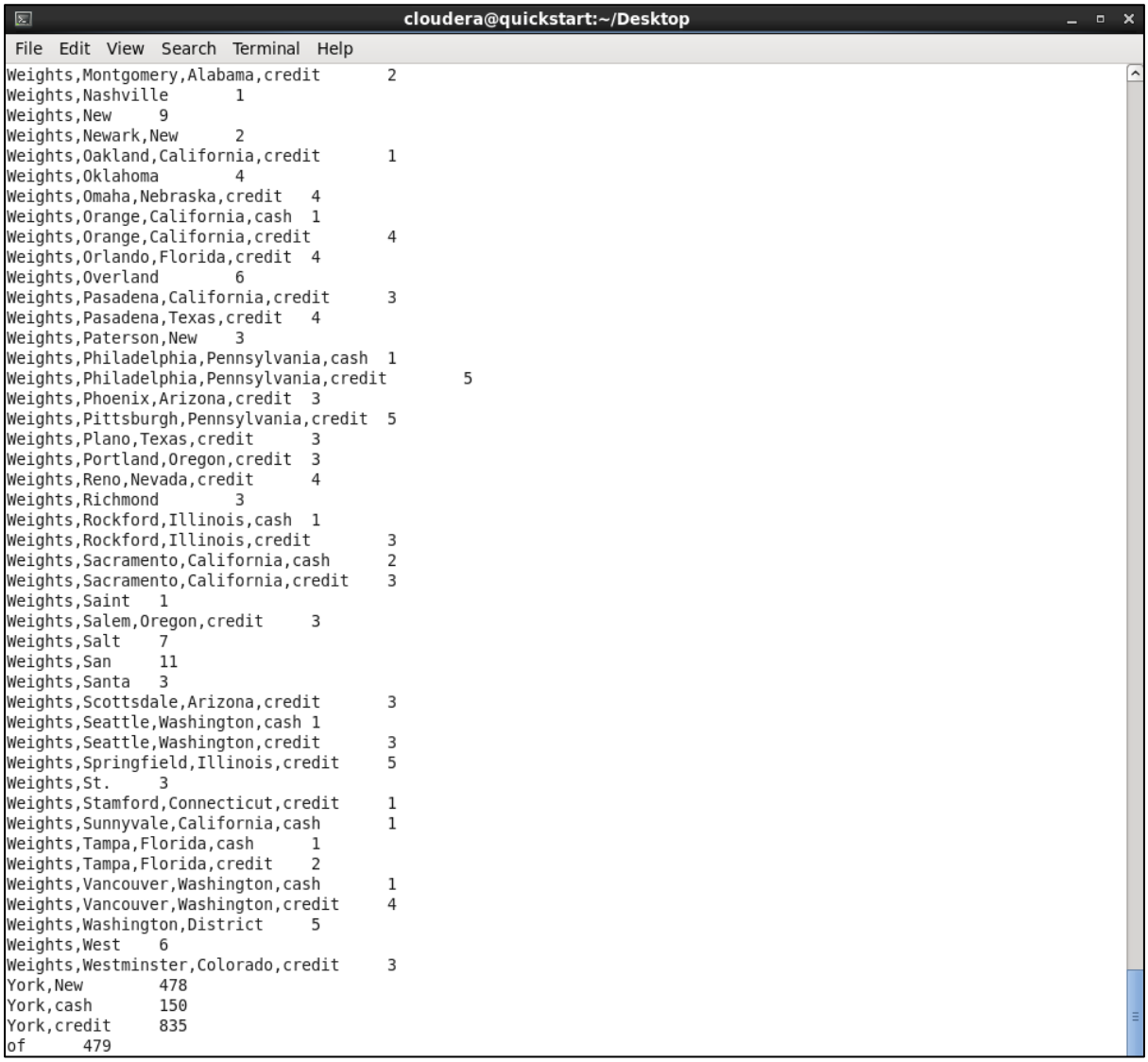
5. Run Hadoop MapReduce Command and output it in file name my_spark_output1

```
[cloudera@quickstart Desktop]$ hadoop jar /home/cloudera/WordCount.jar WordCount /Sparkdata/hivedata.txt /my_spark_ou
tput1
24/11/10 20:08:51 INFO client.RMProxy: Connecting to ResourceManager at /0.0.0.0:8032
24/11/10 20:08:52 WARN mapreduce.JobResourceUploader: Hadoop command-line option parsing not performed. Implement the
 Tool interface and execute your application with ToolRunner to remedy this.
24/11/10 20:08:52 INFO input.FileInputFormat: Total input paths to process : 1
24/11/10 20:08:52 WARN hdfs.DFSClient: Caught exception
java.lang.InterruptedException
        at java.lang.Object.wait(Native Method)
        at java.lang.Thread.join(Thread.java:1281)
        at java.lang.Thread.join(Thread.java:1355)
        at org.apache.hadoop.hdfs.DFSOutputStream$DataStreamer.closeResponder(DFSOutputStream.java:951)
        at org.apache.hadoop.hdfs.DFSOutputStream$DataStreamer.endBlock(DFSOutputStream.java:689)
        at org.apache.hadoop.hdfs.DFSOutputStream$DataStreamer.run(DFSOutputStream.java:878)
24/11/10 20:08:52 INFO mapreduce.JobSubmitter: number of splits:1
24/11/10 20:08:52 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_1729154725495_0003
24/11/10 20:08:52 INFO impl.YarnClientImpl: Submitted application application_1729154725495_0003
24/11/10 20:08:52 INFO mapreduce.Job: The url to track the job: http://quickstart.cloudera:8088/proxy/application_172
9154725495_0003/
24/11/10 20:08:52 INFO mapreduce.Job: Running job: job_1729154725495_0003
24/11/10 20:09:00 INFO mapreduce.Job: Job job_1729154725495_0003 running in uber mode : false
24/11/10 20:09:00 INFO mapreduce.Job:  map 0% reduce 0%
S24/11/10 20:09:07 INFO mapreduce.Job:  map 100% reduce 0%
```

6. After Done Check the output file (my_spark_output1)

```
[cloudera@quickstart Desktop]$ hdfs dfs -ls /
Found 13 items
drwxr-xr-x   - cloudera supergroup          0 2024-11-10 20:00 /Sparkdata
drwxrwxrwx   - hdfs     supergroup          0 2017-04-05 04:27 /benchmarks
drwxr-xr-x   - hbase    supergroup          0 2024-10-27 23:24 /hbase
drwxr-xr-x   - cloudera supergroup          0 2024-10-18 07:27 /inputfolder1
drwxr-xr-x   - cloudera supergroup          0 2024-10-21 00:21 /inputfolder2
drwxr-xr-x   - cloudera supergroup          0 2024-11-10 20:09 /my_spark_output1
drwxr-xr-x   - cloudera supergroup          0 2024-10-18 07:30 /out1
drwxr-xr-x   - cloudera supergroup          0 2024-10-21 00:23 /out2
drwxr-xr-x   - solr     solr                0 2017-04-05 04:29 /solr
-rw-r--r--   1 cloudera supergroup    4418139 2024-11-10 19:10 /sparkdata
drwxrwxrwt   - hdfs     supergroup          0 2024-10-17 01:46 /tmp
drwxr-xr-x   - hdfs     supergroup          0 2017-04-05 04:29 /user
drwxr-xr-x   - hdfs     supergroup          0 2017-04-05 04:29 /var
[cloudera@quickstart Desktop]$ hdfs dfs -cat /my_spark_output1
cat: `/my_spark_output1': Is a directory
[cloudera@quickstart Desktop]$ ^C
[cloudera@quickstart Desktop]$ hdfs dfs -ls /my_spark_output1
Found 2 items
-rw-r--r--   1 cloudera supergroup          0 2024-11-10 20:09 /my_spark_output1/_SUCCESS
-rw-r--r--   1 cloudera supergroup    2789387 2024-11-10 20:09 /my_spark_output1/part-r-00000
[cloudera@quickstart Desktop]$ hdfs dfs -cat /my_spark_output1/part-r-00000
```

Output:


cloudera@quickstart:~/Desktop
File Edit View Search Terminal Help
Weights,Montgomery,Alabama,credit 2
Weights,Nashville 1
Weights,New 9
Weights,Newark,New 2
Weights,Oakland,California,credit 1
Weights,Oklahoma 4
Weights,Omaha,Nebraska,credit 4
Weights,Orange,California,cash 1
Weights,Orange,California,credit 4
Weights,Orlando,Florida,credit 4
Weights,Overland 6
Weights,Pasadena,California,credit 3
Weights,Pasadena,Texas,credit 4
Weights,Paterson,New 3
Weights,Philadelphia,Pennsylvania,cash 1
Weights,Philadelphia,Pennsylvania,credit 5
Weights,Phoenix,Arizona,credit 3
Weights,Pittsburgh,Pennsylvania,credit 5
Weights,Plano,Texas,credit 3
Weights,Portland,Oregon,credit 3
Weights,Reno,Nevada,credit 4
Weights,Richmond 3
Weights,Rockford,Illinois,cash 1
Weights,Rockford,Illinois,credit 3
Weights,Sacramento,California,cash 2
Weights,Sacramento,California,credit 3
Weights,Saint 1
Weights,Salem,Oregon,credit 3
Weights,Salt 7
Weights,San 11
Weights,Santa 3
Weights,Scottsdale,Arizona,credit 3
Weights,Seattle,Washington,cash 1
Weights,Seattle,Washington,credit 3
Weights,Springfield,Illinois,credit 5
Weights,St. 3
Weights,Stamford,Connecticut,credit 1
Weights,Sunnyvale,California,cash 1
Weights,Tampa,Florida,cash 1
Weights,Tampa,Florida,credit 2
Weights,Vancouver,Washington,cash 1
Weights,Vancouver,Washington,credit 4
Weights,Washington,District 5
Weights,West 6
Weights,Westminster,Colorado,credit 3
York,New 478
York,cash 150
York,credit 835
of 479

C . Spark Output File Execution

7. Run spark Shell

```
[cloudera@quickstart Desktop]$ spark-shell
Setting default log level to "WARN".
To adjust logging level use sc.setLogLevel(newLevel).
SLF4J: Class path contains multiple SLF4J bindings.
SLF4J: Found binding in [jar:file:/usr/lib/zookeeper/lib/slf4j-log4j12-1.7.5.jar!/org/slf4j/impl/StaticLoggerBinder.c
lass]
SLF4J: Found binding in [jar:file:/usr/lib/flume-ng/lib/slf4j-log4j12-1.7.5.jar!/org/slf4j/impl/StaticLoggerBinder.cl
ass]
SLF4J: Found binding in [jar:file:/usr/lib/parquet/lib/slf4j-log4j12-1.7.5.jar!/org/slf4j/impl/StaticLoggerBinder.cla
ss]
SLF4J: See http://www.slf4j.org/codes.html#multiple_bindings for an explanation.
SLF4J: Actual binding is of type [org.slf4j.impl.Log4jLoggerFactory]
Welcome to
      ____              __
     / __/__  ___ _____/ /__
    _\ \/ _ \/ _ `/ __/  '_/
   /___/ .__/\_,_/_/ /_/\_\   version 1.6.0
      /_/

Using Scala version 2.10.5 (Java HotSpot(TM) 64-Bit Server VM, Java 1.7.0_67)
Type in expressions to have them evaluated.
Type :help for more information.
24/11/10 20:17:45 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin
-java classes where applicable
24/11/10 20:17:45 WARN util.Utils: Your hostname, quickstart.cloudera resolves to a loopback address: 127.0.0.1; usin
g 10.0.2.15 instead (on interface eth0)
24/11/10 20:17:45 WARN util.Utils: Set SPARK_LOCAL_IP if you need to bind to another address
Spark context available as sc (master = local[*], app id = local-1731298668024).
24/11/10 20:18:00 WARN metastore.ObjectStore: Version information not found in metastore. hive.metastore.schema.verif
ication is not enabled so recording the schema version 1.1.0
24/11/10 20:18:01 WARN metastore.ObjectStore: Failed to get database default, returning NoSuchObjectException
24/11/10 20:18:03 WARN shortcircuit.DomainSocketFactory: The short-circuit local reads feature cannot be used because
 libhadoop cannot be loaded.
SQL context available as sqlContext.

scala>
```

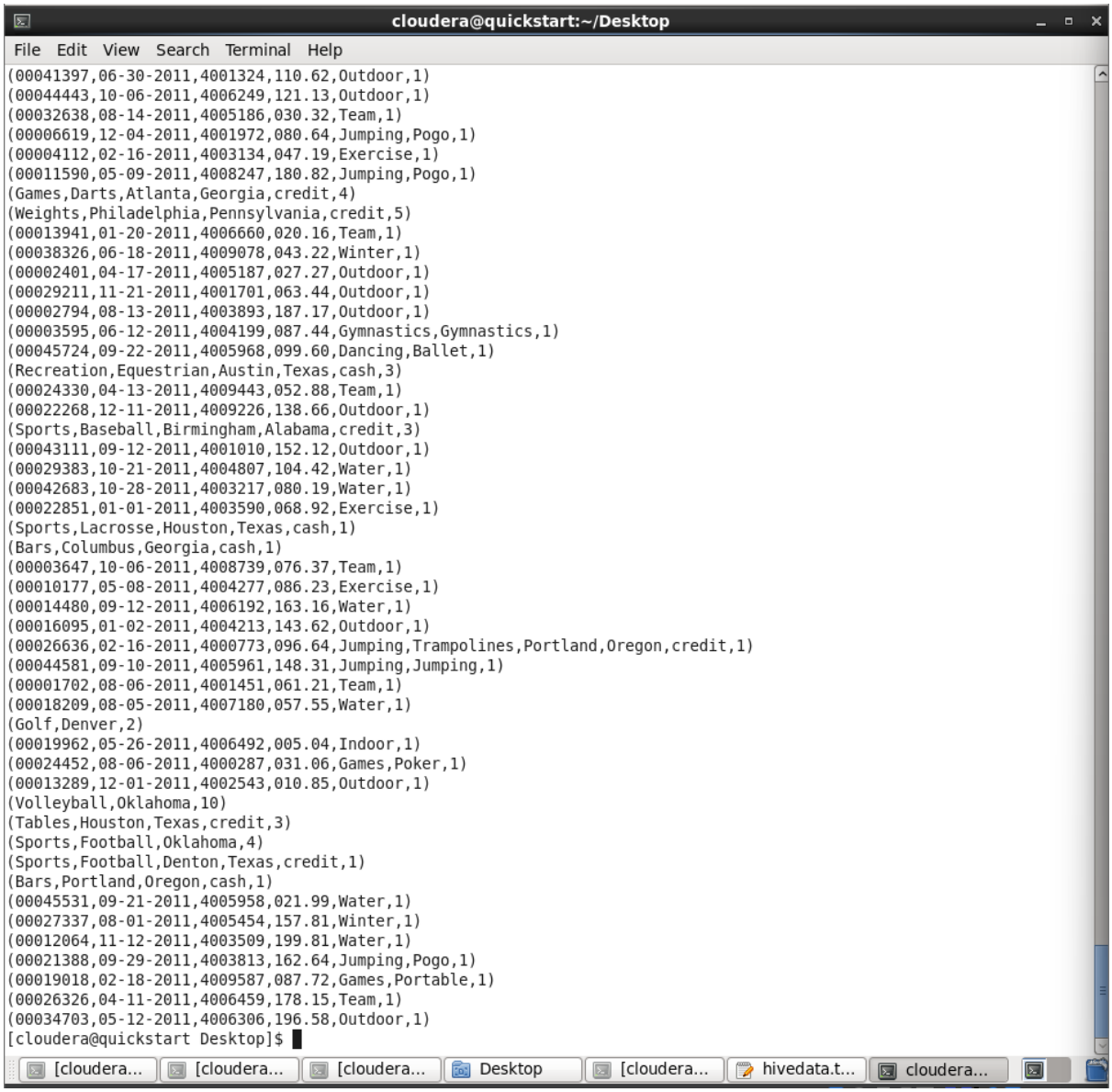
8. Run the Command and save it in my_spark_output2

```
scala> var linesRDD = sc.textFile("hdfs:///Sparkdata")
linesRDD: org.apache.spark.rdd.RDD[String] = hdfs:///Sparkdata MapPartitionsRDD[1] at textFile at <console>:27

scala> var wordsRDD = linesRDD.flatMap( .split(" "))
wordsRDD: org.apache.spark.rdd.RDD[String] = MapPartitionsRDD[2] at flatMap at <console>:29

scala> var wordsKvRdd = wordsRDD.map(( ,1))
wordsKvRdd: org.apache.spark.rdd.RDD[(String, Int)] = MapPartitionsRDD[3] at map at <console>:31

scala> var wordCounts = wordsKvRdd.reduceByKey( + )
wordCounts: org.apache.spark.rdd.RDD[(String, Int)] = ShuffledRDD[4] at reduceByKey at <console>:33

scala> wordCounts.saveAsTextFile("my_spark_output2")
```

9. Open the output file

```
cloudera@quickstart:~/Desktop
File  Edit  View  Search  Terminal  Help
[cloudera@quickstart Desktop]$ hdfs dfs -ls
Found 2 items
drwxr-xr-x   - cloudera cloudera          0 2024-11-10 20:21 my_spark_output1
drwxr-xr-x   - cloudera cloudera          0 2024-11-10 20:31 my_spark_output2
[cloudera@quickstart Desktop]$ hdfs dfs -ls my_spark_output2
Found 2 items
-rw-r--r--   1 cloudera cloudera          0 2024-11-10 20:31 my_spark_output2/_S
UCCESS
-rw-r--r--   1 cloudera cloudera    2915879 2024-11-10 20:31 my_spark_output2/pa
rt-00000
[cloudera@quickstart Desktop]$ hdfs dfs -cat my_spark_output2/part-00000
```

Output


cloudera@quickstart:~/Desktop
File Edit View Search Terminal Help
(00041397,06-30-2011,4001324,110.62,Outdoor,1)
(00044443,10-06-2011,4006249,121.13,Outdoor,1)
(00032638,08-14-2011,4005186,030.32,Team,1)
(00006619,12-04-2011,4001972,080.64,Jumping,Pogo,1)
(00004112,02-16-2011,4003134,047.19,Exercise,1)
(00011590,05-09-2011,4008247,180.82,Jumping,Pogo,1)
(Games,Darts,Atlanta,Georgia,credit,4)
(Weights,Philadelphia,Pennsylvania,credit,5)
(00013941,01-20-2011,4006660,020.16,Team,1)
(00038326,06-18-2011,4009078,043.22,Winter,1)
(00002401,04-17-2011,4005187,027.27,Outdoor,1)
(00029211,11-21-2011,4001701,063.44,Outdoor,1)
(00002794,08-13-2011,4003893,187.17,Outdoor,1)
(00003595,06-12-2011,4004199,087.44,Gymnastics,Gymnastics,1)
(00045724,09-22-2011,4005968,099.60,Dancing,Ballet,1)
(Recreation,Equestrian,Austin,Texas,cash,3)
(00024330,04-13-2011,4009443,052.88,Team,1)
(00022268,12-11-2011,4009226,138.66,Outdoor,1)
(Sports,Baseball,Birmingham,Alabama,credit,3)
(00043111,09-12-2011,4001010,152.12,Outdoor,1)
(00029383,10-21-2011,4004807,104.42,Water,1)
(00042683,10-28-2011,4003217,080.19,Water,1)
(00022851,01-01-2011,4003590,068.92,Exercise,1)
(Sports,Lacrosse,Houston,Texas,cash,1)
(Bars,Columbus,Georgia,cash,1)
(00003647,10-06-2011,4008739,076.37,Team,1)
(00010177,05-08-2011,4004277,086.23,Exercise,1)
(00014480,09-12-2011,4006192,163.16,Water,1)
(00016095,01-02-2011,4004213,143.62,Outdoor,1)
(00026636,02-16-2011,4000773,096.64,Jumping,Trampolines,Portland,Oregon,credit,1)
(00044581,09-10-2011,4005961,148.31,Jumping,Jumping,1)
(00001702,08-06-2011,4001451,061.21,Team,1)
(00018209,08-05-2011,4007180,057.55,Water,1)
(Golf,Denver,2)
(00019962,05-26-2011,4006492,005.04,Indoor,1)
(00024452,08-06-2011,4000287,031.06,Games,Poker,1)
(00013289,12-01-2011,4002543,010.85,Outdoor,1)
(Volleyball,Oklahoma,10)
(Tables,Houston,Texas,credit,3)
(Sports,Football,Oklahoma,4)
(Sports,Football,Denton,Texas,credit,1)
(Bars,Portland,Oregon,cash,1)
(00045531,09-21-2011,4005958,021.99,Water,1)
(00027337,08-01-2011,4005454,157.81,Winter,1)
(00012064,11-12-2011,4003509,199.81,Water,1)
(00021388,09-29-2011,4003813,162.64,Jumping,Pogo,1)
(00019018,02-18-2011,4009587,087.72,Games,Portable,1)
(00026326,04-11-2011,4006459,178.15,Team,1)
(00034703,05-12-2011,4006306,196.58,Outdoor,1)
[cloudera@quickstart Desktop]$
[cloudera...
[cloudera...
[cloudera...
Desktop
[cloudera...
hivedata.t...
cloudera...