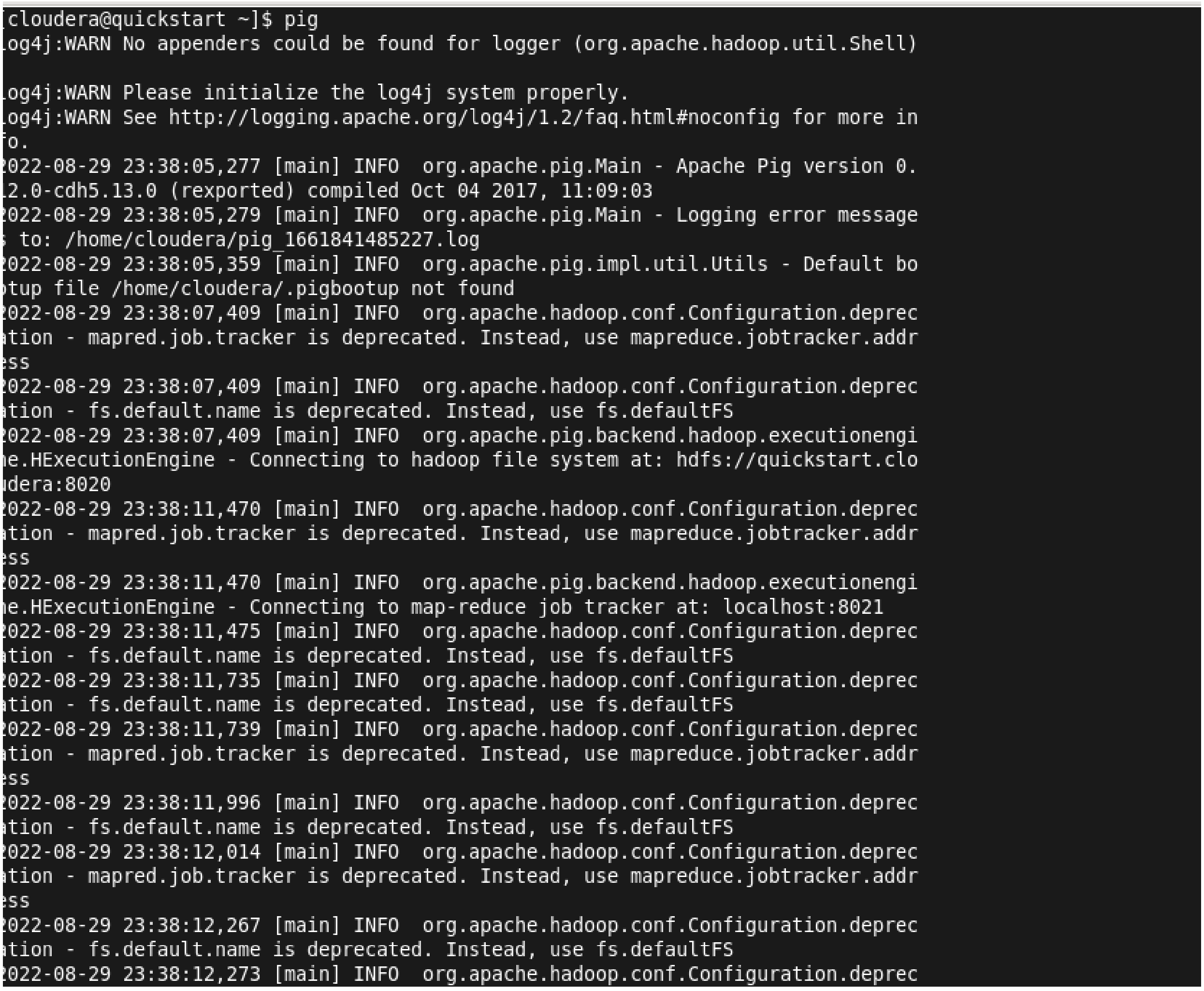
Exercise 04: Word Count using Pig Grouping

Here, we will be running Apache Pig Sample scripts using grunts. It is to just see the power of Apache Pig.

Step 1A: Start Grunt shell.

Open terminal and type *pig*



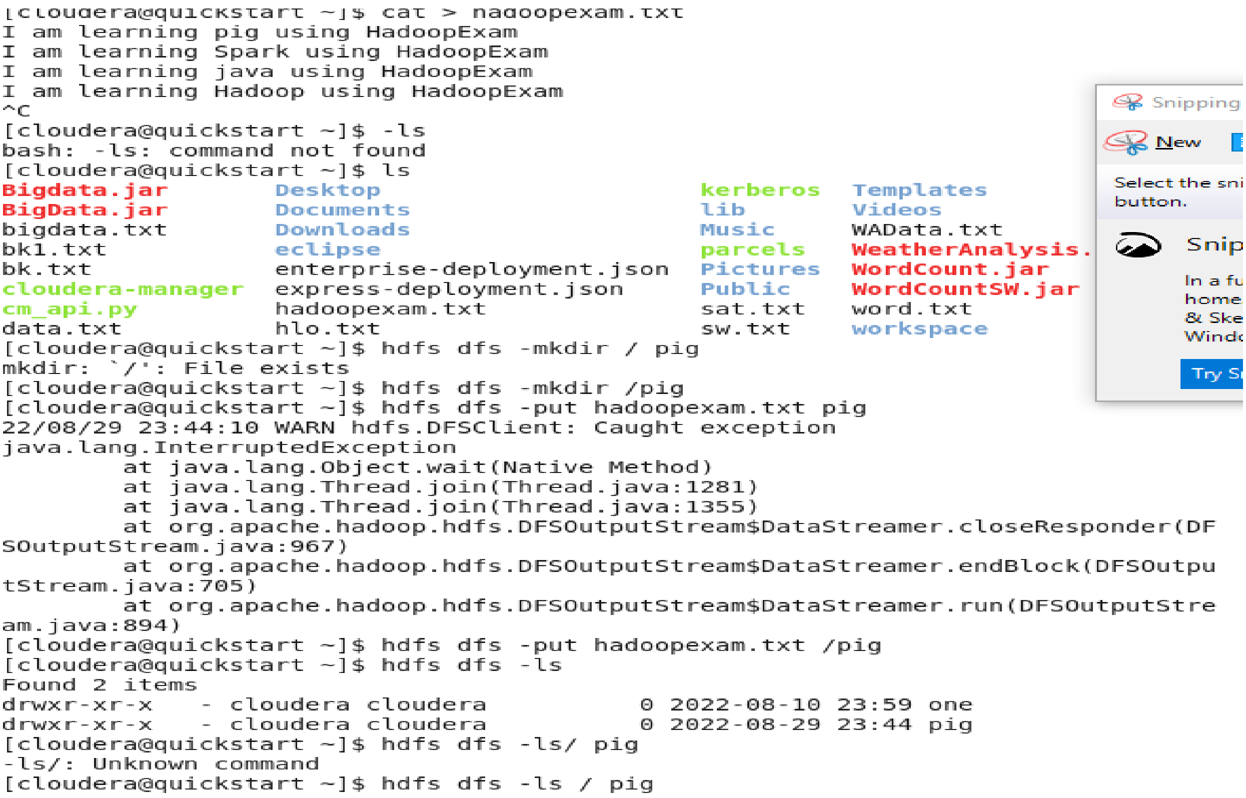
```
[cloudera@quickstart ~]$ pig
.og4j:WARN No appenders could be found for logger (org.apache.hadoop.util.Shell)

.og4j:WARN Please initialize the log4j system properly.
.og4j:WARN See http://logging.apache.org/log4j/1.2/faq.html#noconfig for more in
fo.
2022-08-29 23:38:05,277 [main] INFO  org.apache.pig.Main - Apache Pig version 0.
12.0-cdh5.13.0 (rexported) compiled Oct 04 2017, 11:09:03
2022-08-29 23:38:05,279 [main] INFO  org.apache.pig.Main - Logging error message
s to: /home/cloudera/pig_1661841485227.log
2022-08-29 23:38:05,359 [main] INFO  org.apache.pig.impl.util.Utils - Default bo
otup file /home/cloudera/.pigbootup not found
2022-08-29 23:38:07,409 [main] INFO  org.apache.hadoop.conf.Configuration.deprec
ation - mapred.job.tracker is deprecated. Instead, use mapreduce.jobtracker.addr
ess
2022-08-29 23:38:07,409 [main] INFO  org.apache.hadoop.conf.Configuration.deprec
ation - fs.default.name is deprecated. Instead, use fs.defaultFS
2022-08-29 23:38:07,409 [main] INFO  org.apache.pig.backend.hadoop.executionengi
ne.HExecutionEngine - Connecting to hadoop file system at: hdfs://quickstart.clo
udera:8020
2022-08-29 23:38:11,470 [main] INFO  org.apache.hadoop.conf.Configuration.deprec
ation - mapred.job.tracker is deprecated. Instead, use mapreduce.jobtracker.addr
ess
2022-08-29 23:38:11,470 [main] INFO  org.apache.pig.backend.hadoop.executionengi
ne.HExecutionEngine - Connecting to map-reduce job tracker at: localhost:8021
2022-08-29 23:38:11,475 [main] INFO  org.apache.hadoop.conf.Configuration.deprec
ation - fs.default.name is deprecated. Instead, use fs.defaultFS
2022-08-29 23:38:11,735 [main] INFO  org.apache.hadoop.conf.Configuration.deprec
ation - fs.default.name is deprecated. Instead, use fs.defaultFS
2022-08-29 23:38:11,739 [main] INFO  org.apache.hadoop.conf.Configuration.deprec
ation - mapred.job.tracker is deprecated. Instead, use mapreduce.jobtracker.addr
ess
2022-08-29 23:38:11,996 [main] INFO  org.apache.hadoop.conf.Configuration.deprec
ation - fs.default.name is deprecated. Instead, use fs.defaultFS
2022-08-29 23:38:12,014 [main] INFO  org.apache.hadoop.conf.Configuration.deprec
ation - mapred.job.tracker is deprecated. Instead, use mapreduce.jobtracker.addr
ess
2022-08-29 23:38:12,267 [main] INFO  org.apache.hadoop.conf.Configuration.deprec
ation - fs.default.name is deprecated. Instead, use fs.defaultFS
2022-08-29 23:38:12,273 [main] INFO  org.apache.hadoop.conf.Configuration.deprec
```

Step 1B: Create a file at /user/cloudera/pigfile.txt Withfollowing content.

I am learning Pig Using cloudera

I am learning Spark Using cloudera

I am learning Java Using cloudera

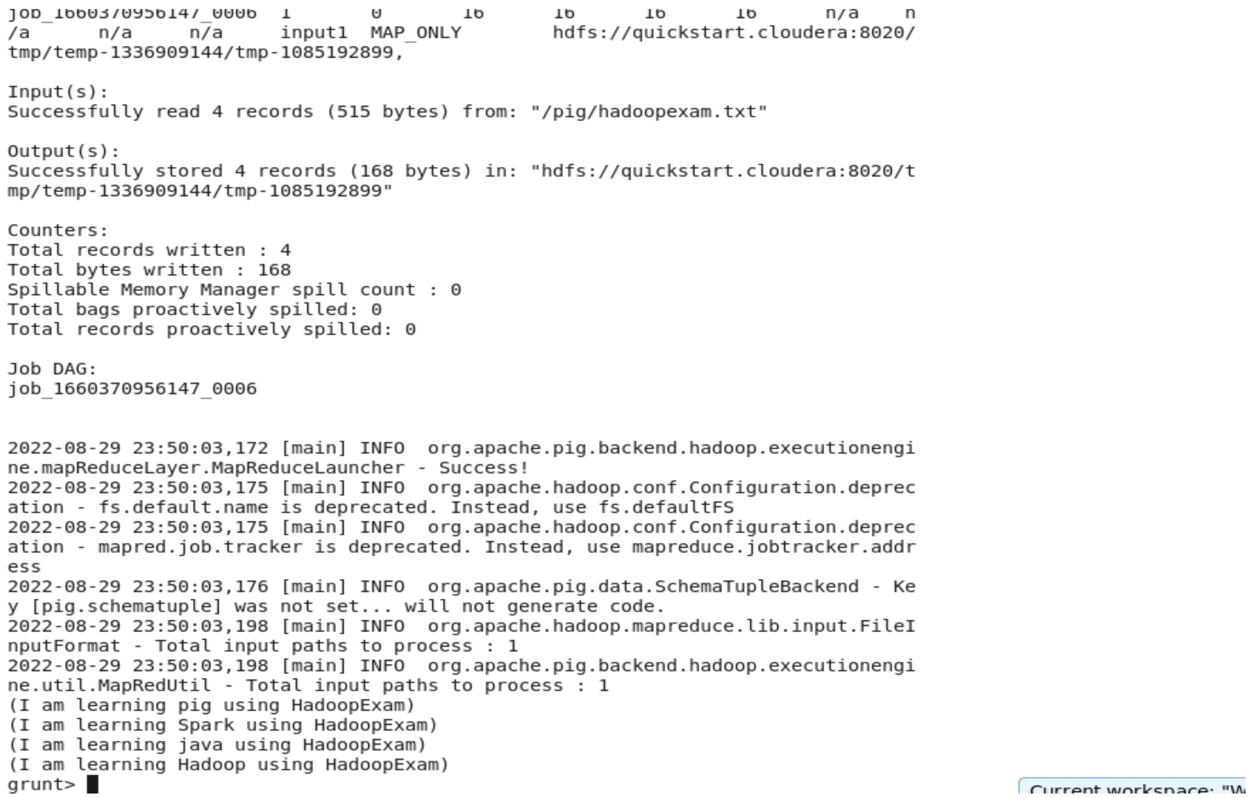


```
[cloudera@quickstart ~]$ cat > hadoopexam.txt
I am learning pig using HadoopExam
I am learning Spark using HadoopExam
I am learning java using HadoopExam
I am learning Hadoop using HadoopExam
^C
[cloudera@quickstart ~]$ -ls
bash: -ls: command not found
[cloudera@quickstart ~]$ ls
Bigdata.jar           Desktop                      kerberos   Templates
BigData.jar           Documents                    lib        Videos
bigdata.txt           Downloads                    Music      WAData.txt
bk1.txt               eclipse                      parcels    WeatherAnalysis.
bk.txt                enterprise-deployment.json   Pictures   WordCount.jar
cloudera-manager      express-deployment.json      Public     WordCountSW.jar
cm_api.py             hadoopexam.txt               sat.txt    word.txt
data.txt              hlo.txt                      sw.txt     workspace
[cloudera@quickstart ~]$ hdfs dfs -mkdir / pig
mkdir: `/': File exists
[cloudera@quickstart ~]$ hdfs dfs -mkdir /pig
[cloudera@quickstart ~]$ hdfs dfs -put hadoopexam.txt pig
22/08/29 23:44:10 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$DataStreamer.closeResponder(DF
SOutputStream.java:967)
        at org.apache.hadoop.hdfs.DFSOutputStream$DataStreamer.endBlock(DFSOutpu
tStream.java:705)
        at org.apache.hadoop.hdfs.DFSOutputStream$DataStreamer.run(DFSOutputStre
am.java:894)
[cloudera@quickstart ~]$ hdfs dfs -put hadoopexam.txt /pig
[cloudera@quickstart ~]$ hdfs dfs -ls
Found 2 items
drwxr-xr-x   - cloudera cloudera          0 2022-08-10 23:59 one
drwxr-xr-x   - cloudera cloudera          0 2022-08-29 23:44 pig
[cloudera@quickstart ~]$ hdfs dfs -ls/ pig
-ls/: Unknown command
[cloudera@quickstart ~]$ hdfs dfs -ls / pig
```

Step 2 : Load the file stored in hdfs with variable ‘in1’ and each line store

‘line’ (Space separated file)

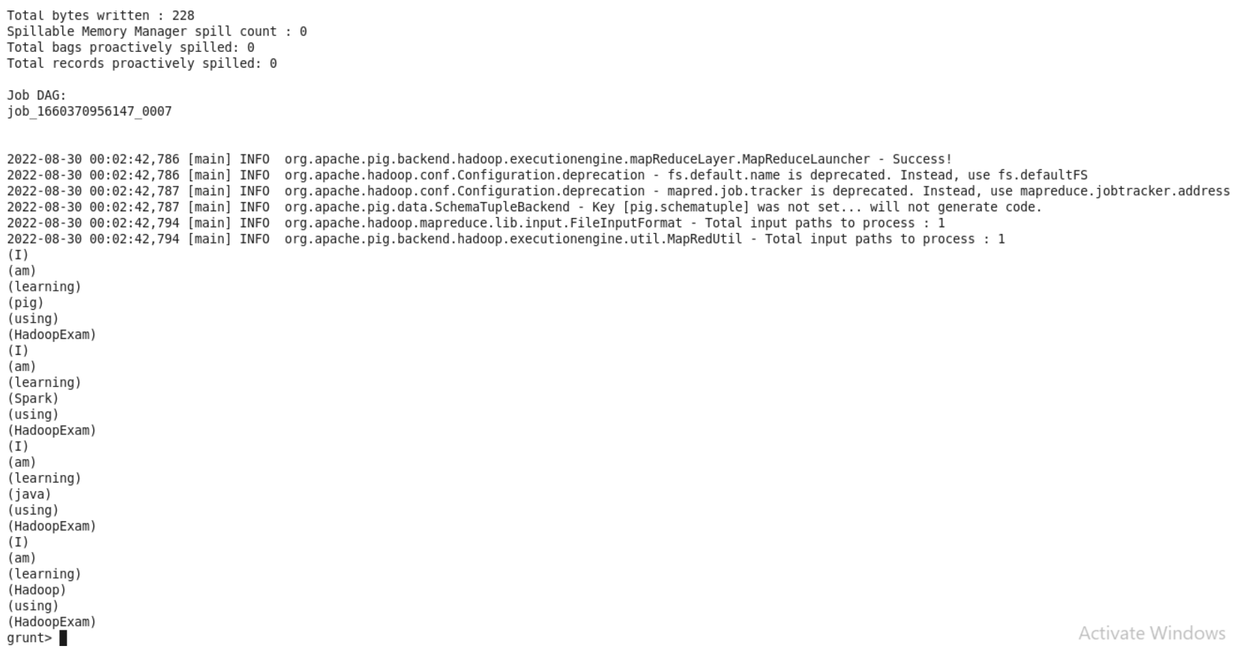
(I am learning Pig Using cloudera)
(I am learning Spark Using cloudera)
(I am learning Java Using cloudera)



```
job_1660370956147_0006  1       0       16      16      16      16      n/a     n
/a      n/a     n/a     input1  MAP_ONLY        hdfs://quickstart.cloudera:8020/
tmp/temp-1336909144/tmp-1085192899,

Input(s):
Successfully read 4 records (515 bytes) from: "/pig/hadoopexam.txt"

Output(s):
Successfully stored 4 records (168 bytes) in: "hdfs://quickstart.cloudera:8020/t
mp/temp-1336909144/tmp-1085192899"

Counters:
Total records written : 4
Total bytes written : 168
Spillable Memory Manager spill count : 0
Total bags proactively spilled: 0
Total records proactively spilled: 0

Job DAG:
job_1660370956147_0006


2022-08-29 23:50:03,172 [main] INFO  org.apache.pig.backend.hadoop.executionengi
ne.mapReduceLayer.MapReduceLauncher - Success!
2022-08-29 23:50:03,175 [main] INFO  org.apache.hadoop.conf.Configuration.deprec
ation - fs.default.name is deprecated. Instead, use fs.defaultFS
2022-08-29 23:50:03,175 [main] INFO  org.apache.hadoop.conf.Configuration.deprec
ation - mapred.job.tracker is deprecated. Instead, use mapreduce.jobtracker.addr
ess
2022-08-29 23:50:03,176 [main] INFO  org.apache.pig.data.SchemaTupleBackend - Ke
y [pig.schematuple] was not set... will not generate code.
2022-08-29 23:50:03,198 [main] INFO  org.apache.hadoop.mapreduce.lib.input.FileI
nputFormat - Total input paths to process : 1
2022-08-29 23:50:03,198 [main] INFO  org.apache.pig.backend.hadoop.executionengi
ne.util.MapRedUtil - Total input paths to process : 1
(I am learning pig using HadoopExam)
(I am learning Spark using HadoopExam)
(I am learning java using HadoopExam)
(I am learning Hadoop using HadoopExam)
grunt> █
```

Step 3: flatten the words in each line from variable ‘in1’ and save separated words into variable ‘wordsinline’

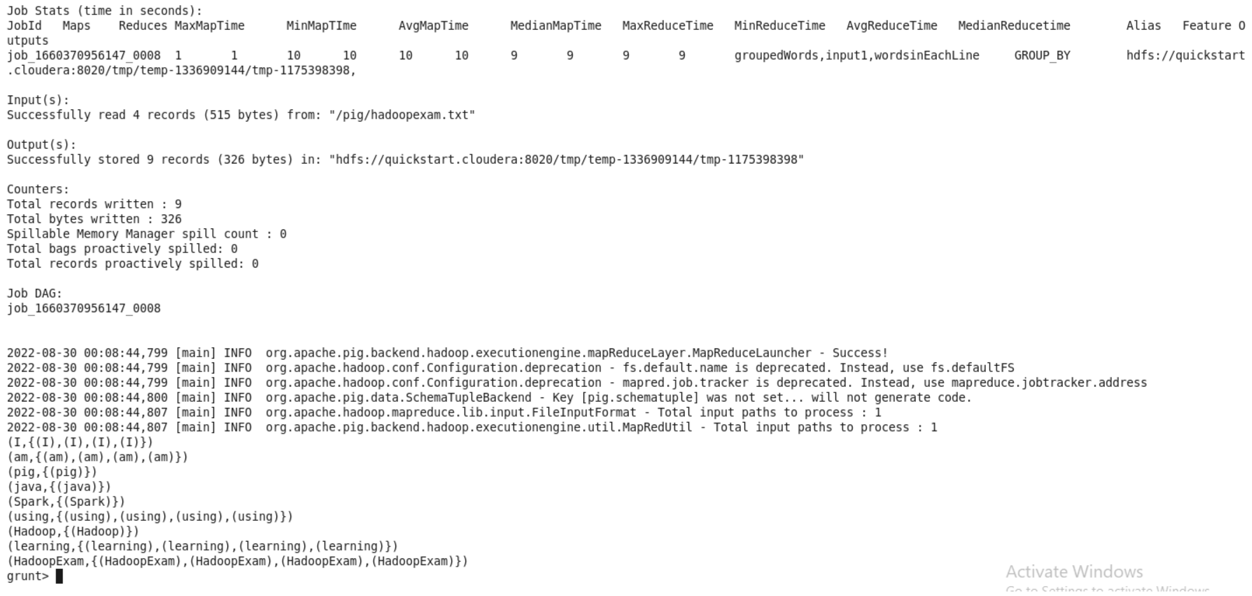
grunt>wordsinline = FOREACH input1 GENERATE flatten(TOKENIZE(line, ‘ ‘)) as word;
grunt>DUMPwordsinline;



```
Total bytes written : 228
Spillable Memory Manager spill count : 0
Total bags proactively spilled: 0
Total records proactively spilled: 0

Job DAG:
job_1660370956147_0007


2022-08-30 00:02:42,786 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - Success!
2022-08-30 00:02:42,786 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Instead, use fs.defaultFS
2022-08-30 00:02:42,787 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - mapred.job.tracker is deprecated. Instead, use mapreduce.jobtracker.address
2022-08-30 00:02:42,787 [main] INFO  org.apache.pig.data.SchemaTupleBackend - Key [pig.schematuple] was not set... will not generate code.
2022-08-30 00:02:42,794 [main] INFO  org.apache.hadoop.mapreduce.lib.input.FileInputFormat - Total input paths to process : 1
2022-08-30 00:02:42,794 [main] INFO  org.apache.pig.backend.hadoop.executionengine.util.MapRedUtil - Total input paths to process : 1
(I)
(am)
(learning)
(pig)
(using)
(HadoopExam)
(I)
(am)
(learning)
(Spark)
(using)
(HadoopExam)
(I)
(am)
(learning)
(java)
(using)
(HadoopExam)
(I)
(am)
(learning)
(Hadoop)
(using)
(HadoopExam)
grunt> 
```

Step 4: Group the similar words and save into variable ‘groupwords’

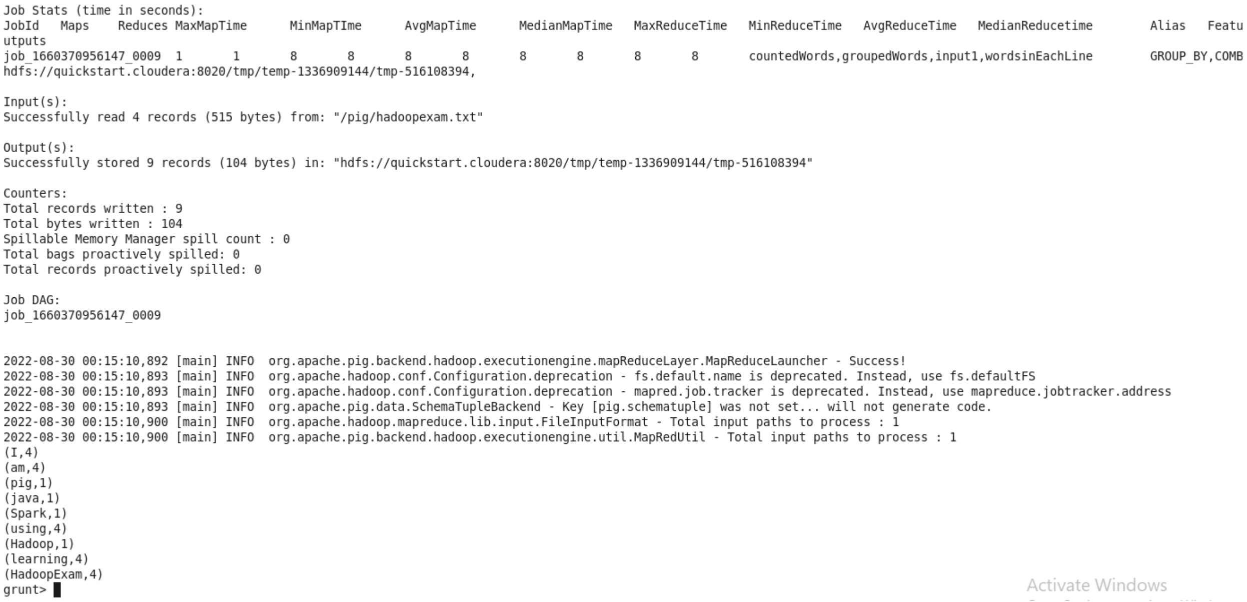
grunt>groupwords =________wordsinline by word;
grunt>dump groupwords;
grunt>describe groupwords;



```
Job Stats (time in seconds):
JobId   Maps    Reduces MaxMapTime      MinMapTIme      AvgMapTime      MedianMapTime   MaxReduceTime   MinReduceTime   AvgReduceTime   MedianReducetime        Alias   Feature O
utputs
job_1660370956147_0008  1       1       10      10      10      10      9       9       9       9       groupedWords,input1,wordsinEachLine     GROUP_BY        hdfs://quickstart
.cloudera:8020/tmp/temp-1336909144/tmp-1175398398,

Input(s):
Successfully read 4 records (515 bytes) from: "/pig/hadoopexam.txt"

Output(s):
Successfully stored 9 records (326 bytes) in: "hdfs://quickstart.cloudera:8020/tmp/temp-1336909144/tmp-1175398398"

Counters:
Total records written : 9
Total bytes written : 326
Spillable Memory Manager spill count : 0
Total bags proactively spilled: 0
Total records proactively spilled: 0

Job DAG:
job_1660370956147_0008


2022-08-30 00:08:44,799 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - Success!
2022-08-30 00:08:44,799 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Instead, use fs.defaultFS
2022-08-30 00:08:44,799 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - mapred.job.tracker is deprecated. Instead, use mapreduce.jobtracker.address
2022-08-30 00:08:44,800 [main] INFO  org.apache.pig.data.SchemaTupleBackend - Key [pig.schematuple] was not set... will not generate code.
2022-08-30 00:08:44,807 [main] INFO  org.apache.hadoop.mapreduce.lib.input.FileInputFormat - Total input paths to process : 1
2022-08-30 00:08:44,807 [main] INFO  org.apache.pig.backend.hadoop.executionengine.util.MapRedUtil - Total input paths to process : 1
(I,{(I),(I),(I),(I)})
(am,{(am),(am),(am),(am)})
(pig,{(pig)})
(java,{(java)})
(Spark,{(Spark)})
(using,{(using),(using),(using),(using)})
(Hadoop,{(Hadoop)})
(learning,{(learning),(learning),(learning),(learning)})
(HadoopExam,{(HadoopExam),(HadoopExam),(HadoopExam),(HadoopExam)})
grunt> 
```

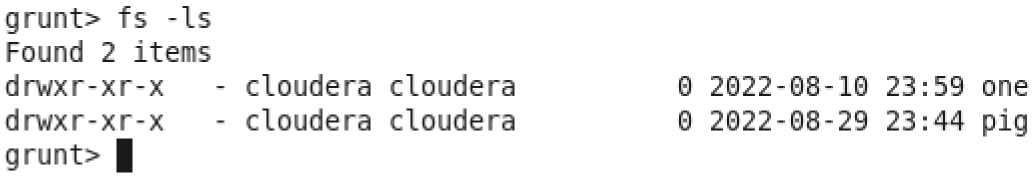
Step 5: Count Words in the group.

grunt>countwords= foreach________________________;
grunt>DUMPcountwords;



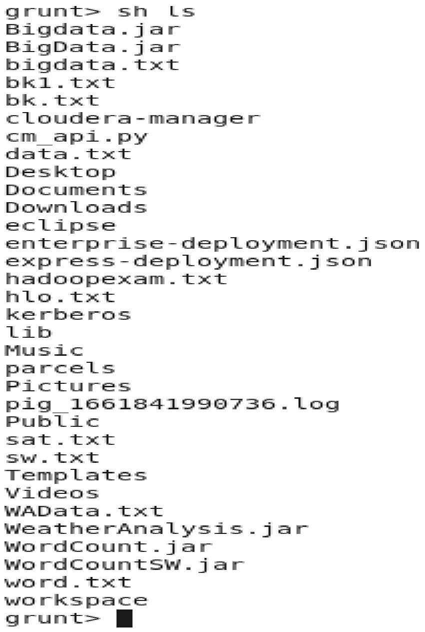
```
Job Stats (time in seconds):
JobId   Maps    Reduces MaxMapTime      MinMapTIme      AvgMapTime      MedianMapTime   MaxReduceTime   MinReduceTime   AvgReduceTime   MedianReducetime        Alias   Featu
utputs
job_1660370956147_0009  1       1       8       8       8       8       8       8       8       8       countedWords,groupedWords,input1,wordsinEachLine        GROUP_BY,COMB
hdfs://quickstart.cloudera:8020/tmp/temp-1336909144/tmp-516108394,

Input(s):
Successfully read 4 records (515 bytes) from: "/pig/hadoopexam.txt"

Output(s):
Successfully stored 9 records (104 bytes) in: "hdfs://quickstart.cloudera:8020/tmp/temp-1336909144/tmp-516108394"

Counters:
Total records written : 9
Total bytes written : 104
Spillable Memory Manager spill count : 0
Total bags proactively spilled: 0
Total records proactively spilled: 0

Job DAG:
job_1660370956147_0009


2022-08-30 00:15:10,892 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - Success!
2022-08-30 00:15:10,893 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Instead, use fs.defaultFS
2022-08-30 00:15:10,893 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - mapred.job.tracker is deprecated. Instead, use mapreduce.jobtracker.address
2022-08-30 00:15:10,893 [main] INFO  org.apache.pig.data.SchemaTupleBackend - Key [pig.schematuple] was not set... will not generate code.
2022-08-30 00:15:10,900 [main] INFO  org.apache.hadoop.mapreduce.lib.input.FileInputFormat - Total input paths to process : 1
2022-08-30 00:15:10,900 [main] INFO  org.apache.pig.backend.hadoop.executionengine.util.MapRedUtil - Total input paths to process : 1
(I,4)
(am,4)
(pig,1)
(java,1)
(Spark,1)
(using,4)
(Hadoop,1)
(learning,4)
(HadoopExam,4)
grunt>
```

STEP-6



```
grunt> fs -ls
Found 2 items
drwxr-xr-x   - cloudera cloudera          0 2022-08-10 23:59 one
drwxr-xr-x   - cloudera cloudera          0 2022-08-29 23:44 pig
grunt>
```

STEP-7



```
grunt> sh ls
Bigdata.jar
BigData.jar
bigdata.txt
bk1.txt
bk.txt
cloudera-manager
cm_api.py
data.txt
Desktop
Documents
Downloads
eclipse
enterprise-deployment.json
express-deployment.json
hadoopexam.txt
hlo.txt
kerberos
lib
Music
parcels
Pictures
pig_1661841990736.log
Public
sat.txt
sw.txt
Templates
Videos
WAData.txt
WeatherAnalysis.jar
WordCount.jar
WordCountSW.jar
word.txt
workspace
grunt>
```