# Exercise 07: Word Count using Pig Grouping

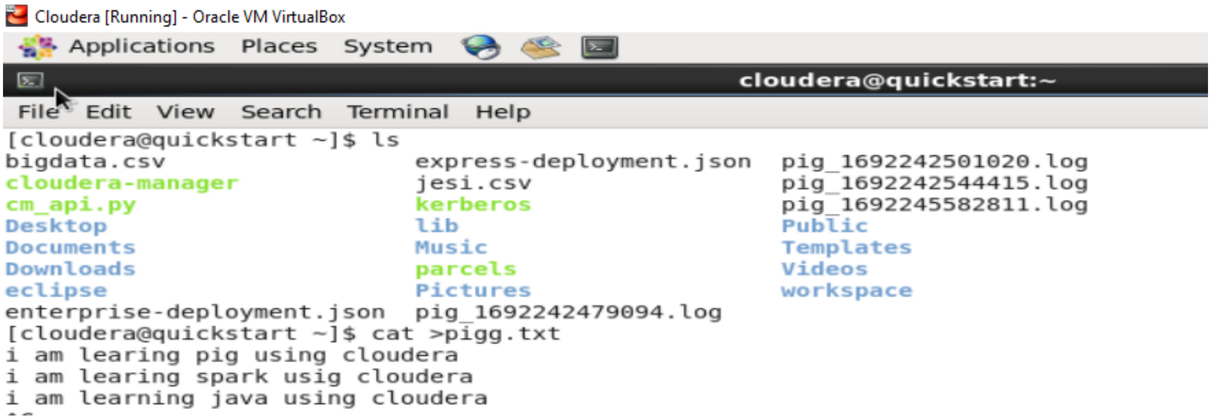
## NAME: ANNAPOORNIMA S (225229101)

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

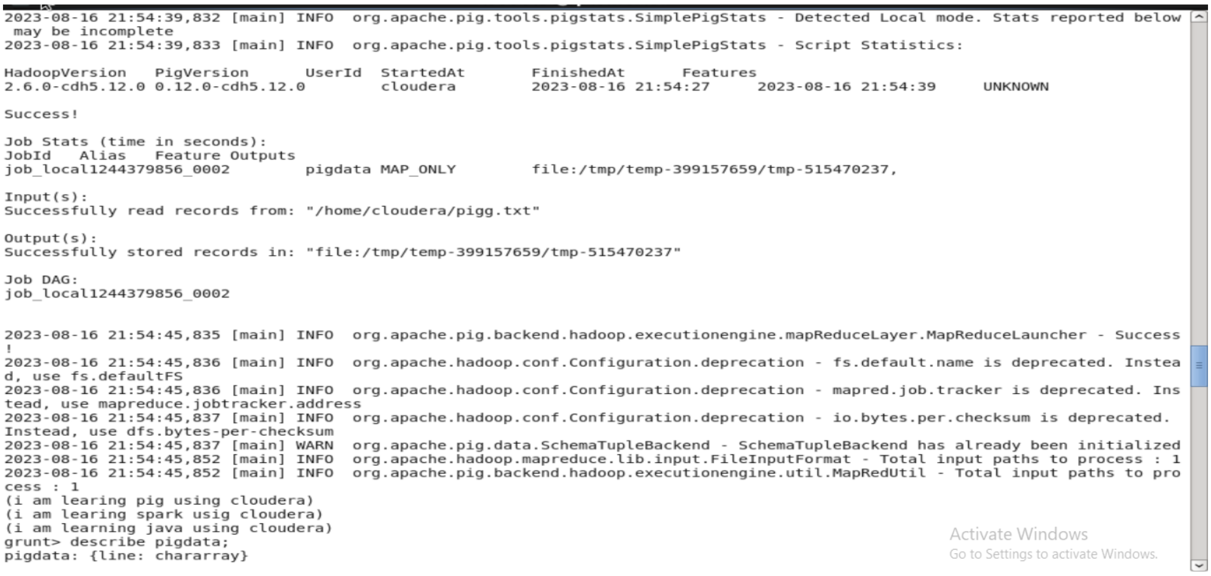
### Step 1B: Create a file at /user/cloudera/pigfile.txt With following content.



```
[cloudera@quickstart ~]$ ls
bigdata.csv                 express-deployment.json  pig_1692242501020.log
cloudera-manager            jesi.csv                 pig_1692242544415.log
cm_api.py                   kerberos                 pig_1692245582811.log
Desktop                     lib                      Public
Documents                   Music                    Templates
Downloads                   parcels                  Videos
eclipse                     Pictures                 workspace
enterprise-deployment.json  pig_1692242479094.log
[cloudera@quickstart ~]$ cat >pigg.txt
i am learing pig using cloudera
i am learing spark usig cloudera
i am learning java using cloudera
```

### Step 2 : Load the file stored in hdfs with variable ‘in1’ and each line have to store in ‘line’ (Space separated file)

**pigdata = load '/home/cloudera/pigg.txt' as (line:chararray);**
**dump pigdata;**

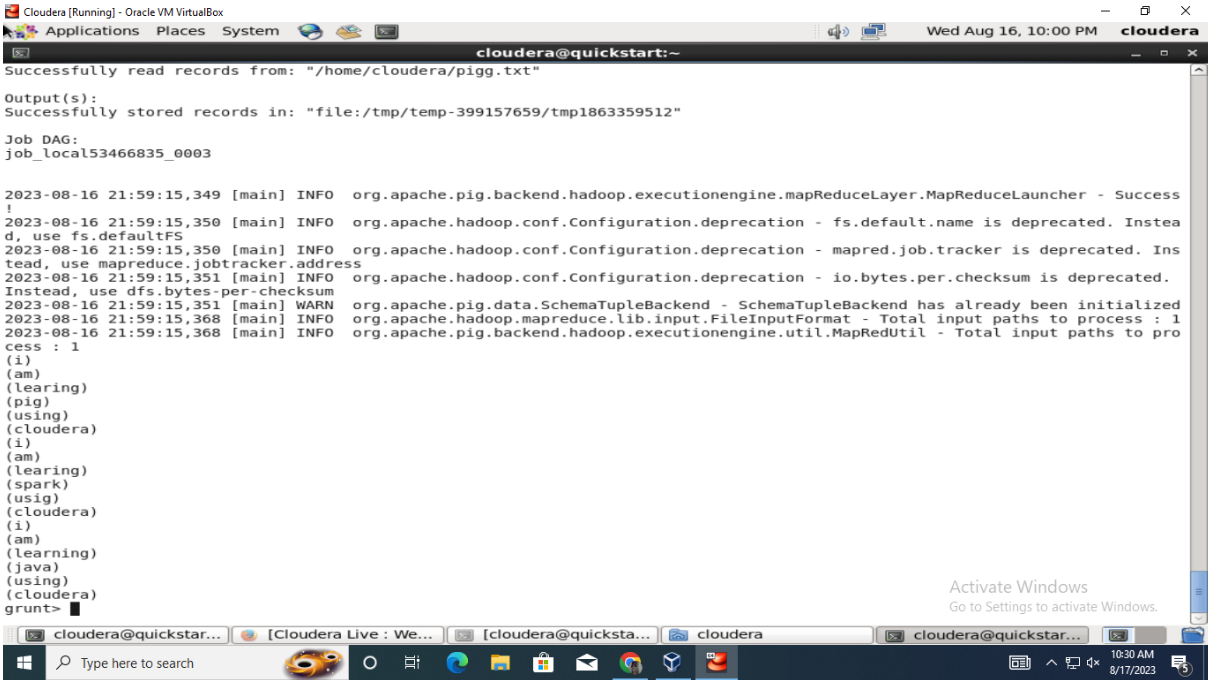


```
2023-08-16 21:54:39,832 [main] INFO  org.apache.pig.tools.pigstats.SimplePigStats - Detected Local mode. Stats reported below
 may be incomplete
2023-08-16 21:54:39,833 [main] INFO  org.apache.pig.tools.pigstats.SimplePigStats - Script Statistics:

HadoopVersion   PigVersion      UserId  StartedAt       FinishedAt      Features
2.6.0-cdh5.12.0 0.12.0-cdh5.12.0        cloudera        2023-08-16 21:54:27     2023-08-16 21:54:39     UNKNOWN

Success!

Job Stats (time in seconds):
JobId   Alias   Feature Outputs
job_local1244379856_0002        pigdata MAP_ONLY        file:/tmp/temp-399157659/tmp-515470237,

Input(s):
Successfully read records from: "/home/cloudera/pigg.txt"

Output(s):
Successfully stored records in: "file:/tmp/temp-399157659/tmp-515470237"

Job DAG:
job_local1244379856_0002


2023-08-16 21:54:45,835 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - Success
!
2023-08-16 21:54:45,836 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Instea
d, use fs.defaultFS
2023-08-16 21:54:45,836 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - mapred.job.tracker is deprecated. Ins
tead, use mapreduce.jobtracker.address
2023-08-16 21:54:45,837 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - io.bytes.per.checksum is deprecated.
Instead, use dfs.bytes-per-checksum
2023-08-16 21:54:45,837 [main] WARN  org.apache.pig.data.SchemaTupleBackend - SchemaTupleBackend has already been initialized
2023-08-16 21:54:45,852 [main] INFO  org.apache.hadoop.mapreduce.lib.input.FileInputFormat - Total input paths to process : 1
2023-08-16 21:54:45,852 [main] INFO  org.apache.pig.backend.hadoop.executionengine.util.MapRedUtil - Total input paths to pro
cess : 1
(i am learing pig using cloudera)
(i am learing spark usig cloudera)
(i am learning java using cloudera)
grunt> describe pigdata;
pigdata: {line: chararray}
```

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

**pigwords = foreach pigdata generate flatten (TOKENIZE(line,' ')) as word;**

**dump pigwords;**

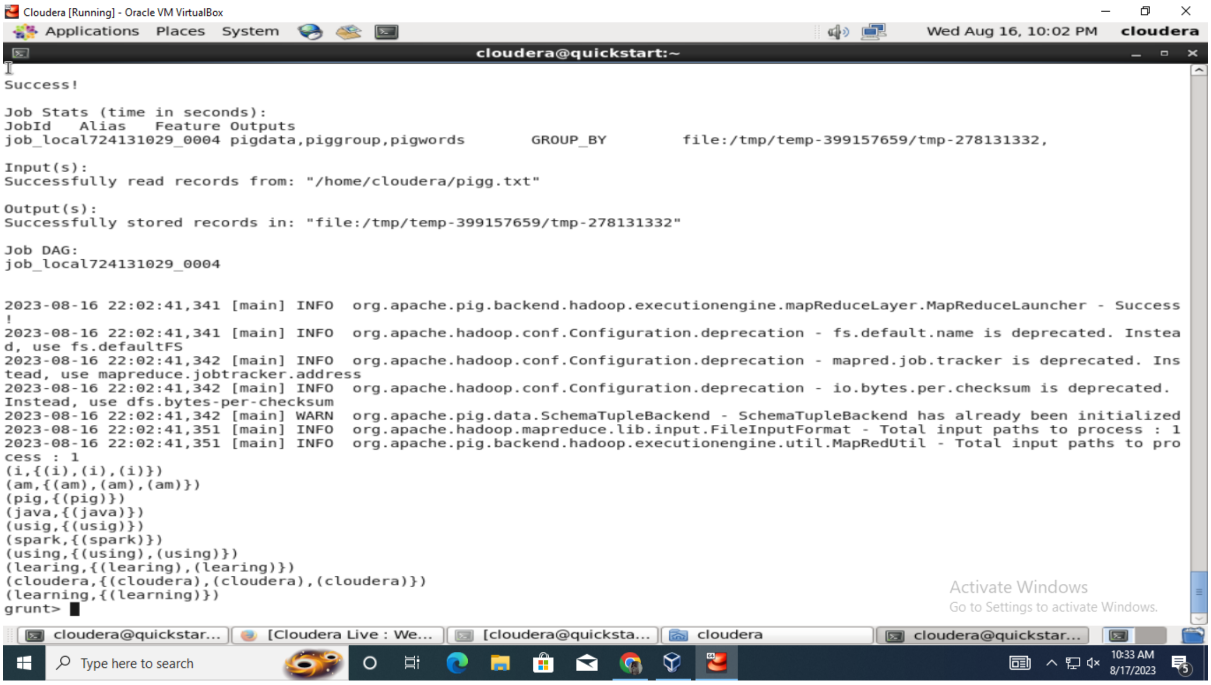


```
Successfully read records from: "/home/cloudera/pigg.txt"

Output(s):
Successfully stored records in: "file:/tmp/temp-399157659/tmp1863359512"

Job DAG:
job_local53466835_0003


2023-08-16 21:59:15,349 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - Success
!
2023-08-16 21:59:15,350 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Instea
d, use fs.defaultFS
2023-08-16 21:59:15,350 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - mapred.job.tracker is deprecated. Ins
tead, use mapreduce.jobtracker.address
2023-08-16 21:59:15,351 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - io.bytes.per.checksum is deprecated.
Instead, use dfs.bytes-per-checksum
2023-08-16 21:59:15,351 [main] WARN  org.apache.pig.data.SchemaTupleBackend - SchemaTupleBackend has already been initialized
2023-08-16 21:59:15,368 [main] INFO  org.apache.hadoop.mapreduce.lib.input.FileInputFormat - Total input paths to process : 1
2023-08-16 21:59:15,368 [main] INFO  org.apache.pig.backend.hadoop.executionengine.util.MapRedUtil - Total input paths to pro
cess : 1
(i)
(am)
(learing)
(pig)
(using)
(cloudera)
(i)
(am)
(learing)
(spark)
(usig)
(cloudera)
(i)
(am)
(learning)
(java)
(using)
(cloudera)
grunt> 
```

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

**piggroup = GROUP pigwords by word;**

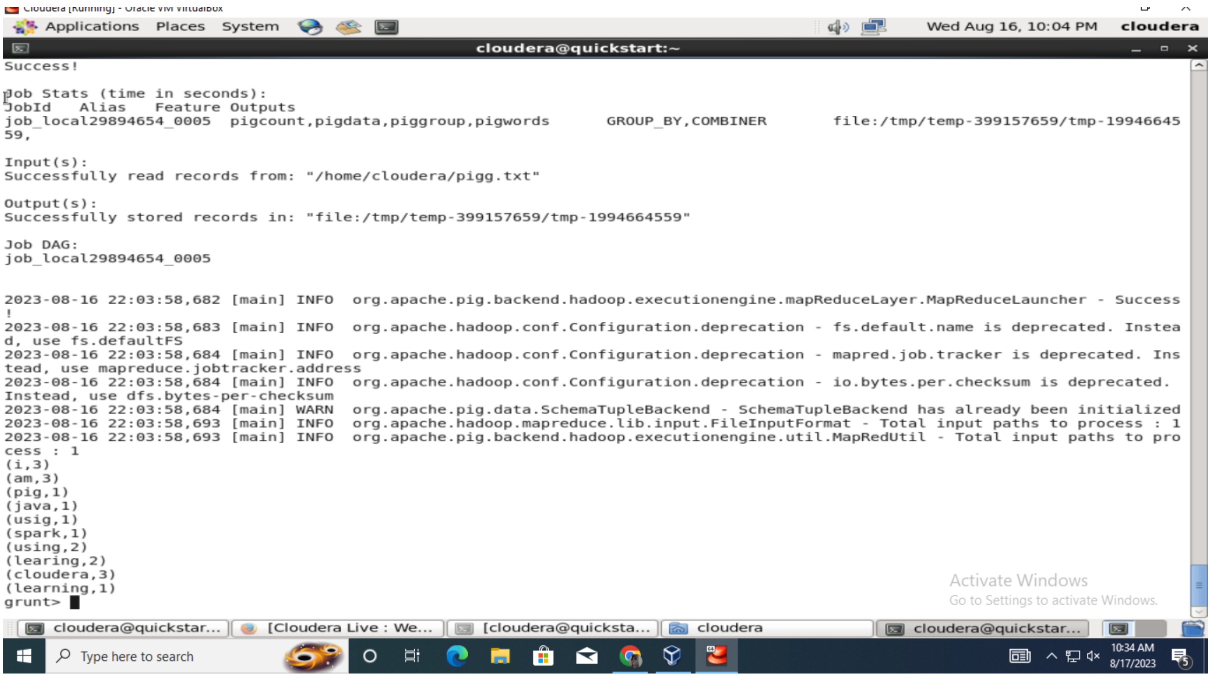
**dump piggroup;**



```
Success!

Job Stats (time in seconds):
JobId   Alias   Feature Outputs
job_local724131029_0004 pigdata,piggroup,pigwords       GROUP_BY        file:/tmp/temp-399157659/tmp-278131332,

Input(s):
Successfully read records from: "/home/cloudera/pigg.txt"

Output(s):
Successfully stored records in: "file:/tmp/temp-399157659/tmp-278131332"

Job DAG:
job_local724131029_0004


2023-08-16 22:02:41,341 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - Success
!
2023-08-16 22:02:41,341 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Instea
d, use fs.defaultFS
2023-08-16 22:02:41,342 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - mapred.job.tracker is deprecated. Ins
tead, use mapreduce.jobtracker.address
2023-08-16 22:02:41,342 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - io.bytes.per.checksum is deprecated.
Instead, use dfs.bytes-per-checksum
2023-08-16 22:02:41,342 [main] WARN  org.apache.pig.data.SchemaTupleBackend - SchemaTupleBackend has already been initialized
2023-08-16 22:02:41,351 [main] INFO  org.apache.hadoop.mapreduce.lib.input.FileInputFormat - Total input paths to process : 1
2023-08-16 22:02:41,351 [main] INFO  org.apache.pig.backend.hadoop.executionengine.util.MapRedUtil - Total input paths to pro
cess : 1
(i,{(i),(i),(i)})
(am,{(am),(am),(am)})
(pig,{(pig)})
(java,{(java)})
(usig,{(usig)})
(spark,{(spark)})
(using,{(using),(using)})
(learing,{(learing),(learing)})
(cloudera,{(cloudera),(cloudera),(cloudera)})
(learning,{(learning)})
grunt> █
```

## Step 5: Count Words in the group.

**pigcount = foreach piggroup GENERATE group, COUNT(pigwords);**
**dump pigcount;**



```
Success!

Job Stats (time in seconds):
JobId   Alias   Feature Outputs
job_local29894654_0005  pigcount,pigdata,piggroup,pigwords      GROUP_BY,COMBINER       file:/tmp/temp-399157659/tmp-19946645
59,

Input(s):
Successfully read records from: "/home/cloudera/pigg.txt"

Output(s):
Successfully stored records in: "file:/tmp/temp-399157659/tmp-1994664559"

Job DAG:
job_local29894654_0005


2023-08-16 22:03:58,682 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - Success
!
2023-08-16 22:03:58,683 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Instea
d, use fs.defaultFS
2023-08-16 22:03:58,684 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - mapred.job.tracker is deprecated. Ins
tead, use mapreduce.jobtracker.address
2023-08-16 22:03:58,684 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - io.bytes.per.checksum is deprecated.
Instead, use dfs.bytes-per-checksum
2023-08-16 22:03:58,684 [main] WARN  org.apache.pig.data.SchemaTupleBackend - SchemaTupleBackend has already been initialized
2023-08-16 22:03:58,693 [main] INFO  org.apache.hadoop.mapreduce.lib.input.FileInputFormat - Total input paths to process : 1
2023-08-16 22:03:58,693 [main] INFO  org.apache.pig.backend.hadoop.executionengine.util.MapRedUtil - Total input paths to pro
cess : 1
(i,3)
(am,3)
(pig,1)
(java,1)
(usig,1)
(spark,1)
(using,2)
(learing,2)
(cloudera,3)
(learning,1)
grunt> █
```