Bharathwin MA
205229105
Big Data Management and Analytics

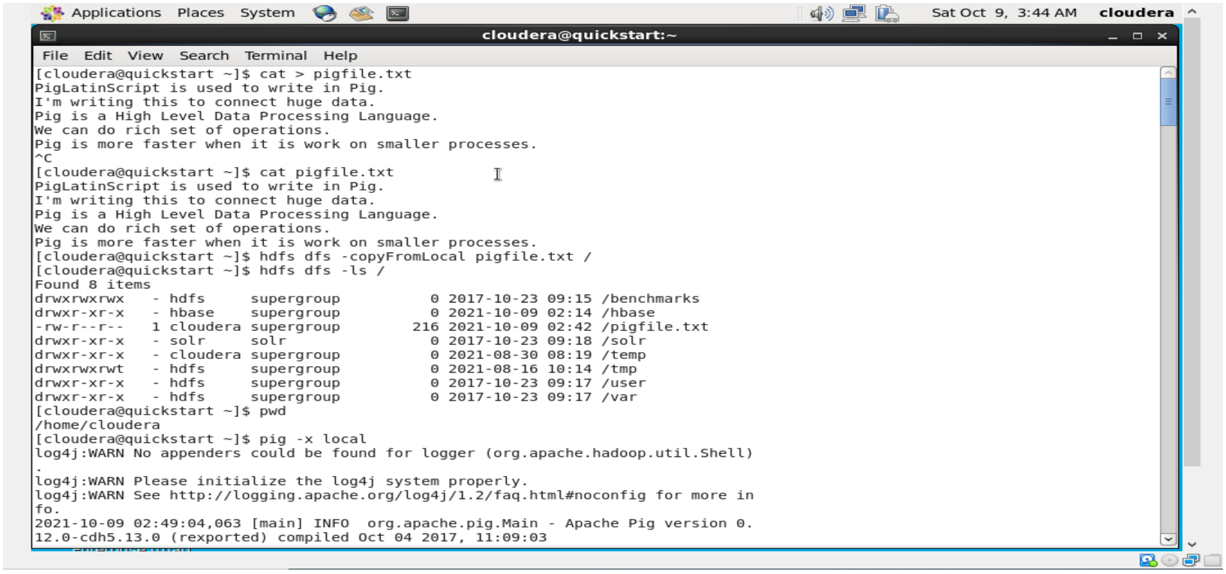
Exercise 07: Word Count using Pig Grouping

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.

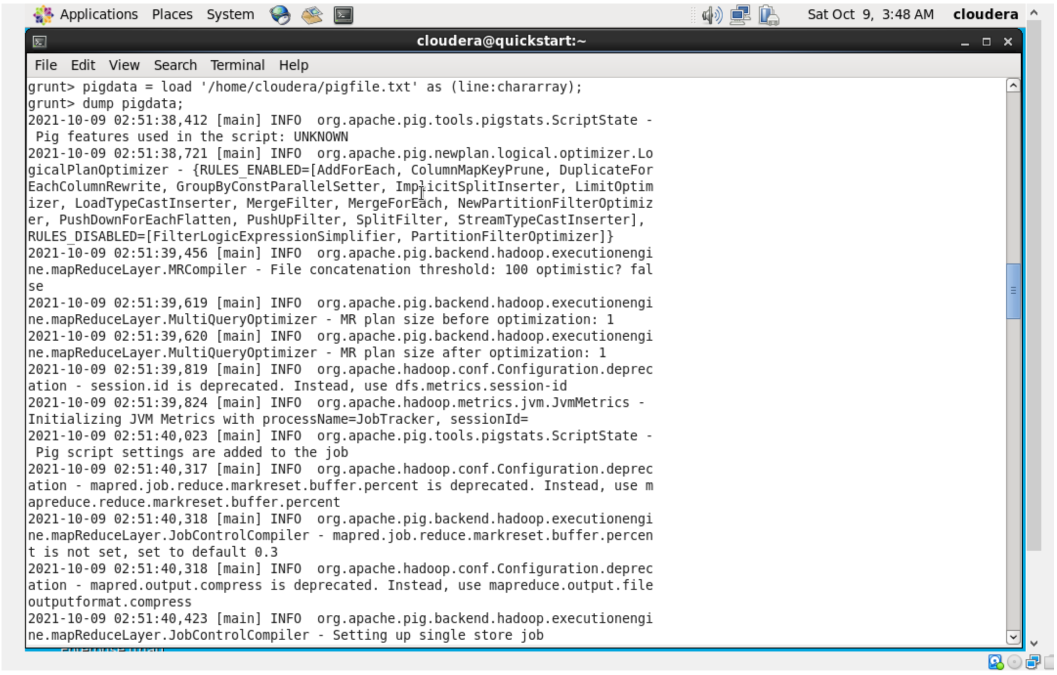
Open terminal and type *pig*

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

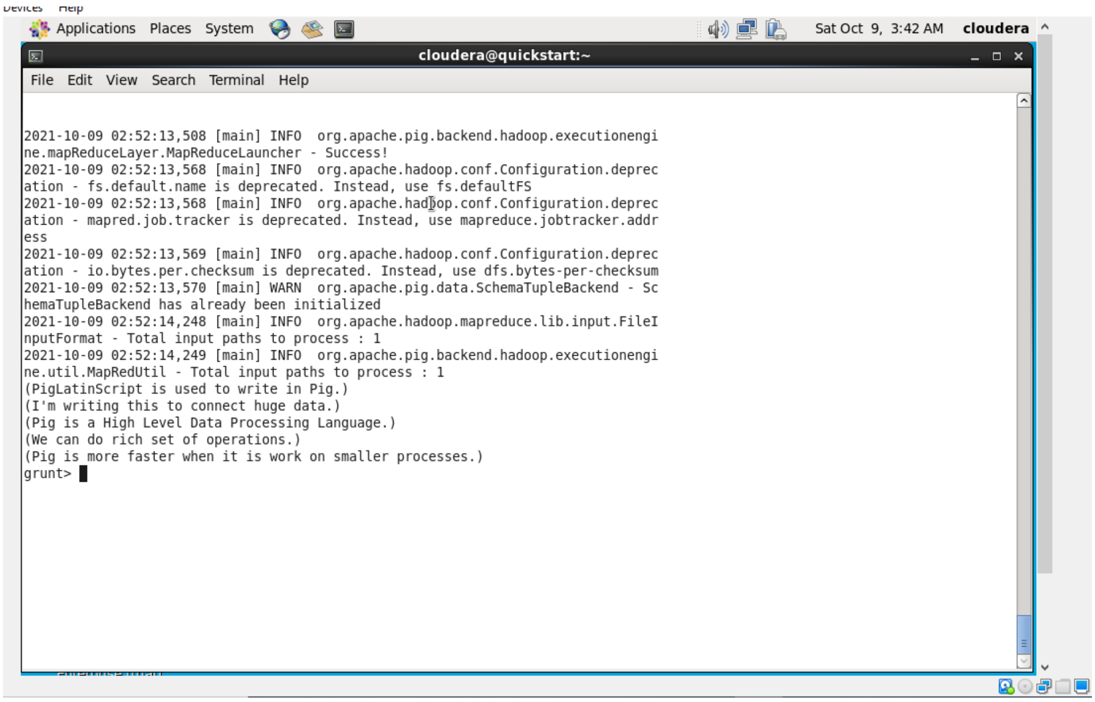


```
[cloudera@quickstart ~]$ cat > pigfile.txt
PigLatinScript is used to write in Pig.
I'm writing this to connect huge data.
Pig is a High Level Data Processing Language.
We can do rich set of operations.
Pig is more faster when it is work on smaller processes.
^C
[cloudera@quickstart ~]$ cat pigfile.txt
PigLatinScript is used to write in Pig.
I'm writing this to connect huge data.
Pig is a High Level Data Processing Language.
We can do rich set of operations.
Pig is more faster when it is work on smaller processes.
[cloudera@quickstart ~]$ hdfs dfs -copyFromLocal pigfile.txt /
[cloudera@quickstart ~]$ hdfs dfs -ls /
Found 8 items
drwxrwxrwx   - hdfs     supergroup          0 2017-10-23 09:15 /benchmarks
drwxr-xr-x   - hbase    supergroup          0 2021-10-09 02:14 /hbase
-rw-r--r--   1 cloudera supergroup        216 2021-10-09 02:42 /pigfile.txt
drwxr-xr-x   - solr     solr                0 2017-10-23 09:18 /solr
drwxr-xr-x   - cloudera supergroup          0 2021-08-30 08:19 /temp
drwxrwxrwt   - hdfs     supergroup          0 2021-08-16 10:14 /tmp
drwxr-xr-x   - hdfs     supergroup          0 2017-10-23 09:17 /user
drwxr-xr-x   - hdfs     supergroup          0 2017-10-23 09:17 /var
[cloudera@quickstart ~]$ pwd
/home/cloudera
[cloudera@quickstart ~]$ pig -x local
log4j:WARN No appenders could be found for logger (org.apache.hadoop.util.Shell)
.
log4j:WARN Please initialize the log4j system properly.
log4j:WARN See http://logging.apache.org/log4j/1.2/faq.html#noconfig for more in
fo.
2021-10-09 02:49:04,063 [main] INFO  org.apache.pig.Main - Apache Pig version 0.
12.0-cdh5.13.0 (rexported) compiled Oct 04 2017, 11:09:03
```

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


Applications Places System
Sat Oct 9, 3:48 AM
cloudera
cloudera@quickstart:~
File Edit View Search Terminal Help

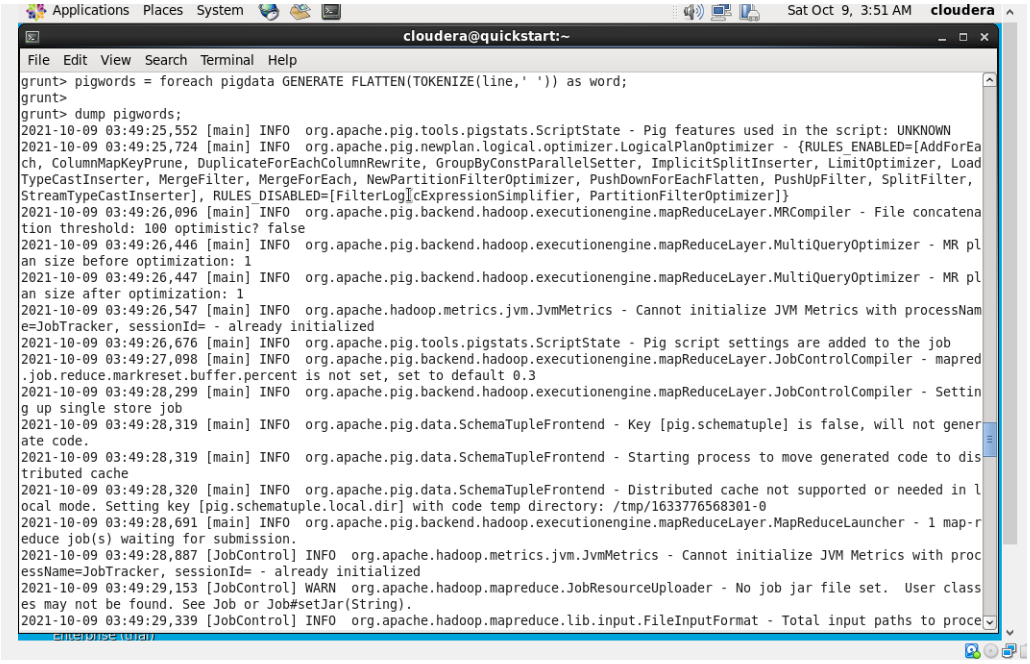
```
grunt> pigdata = load '/home/cloudera/pigfile.txt' as (line:chararray);
grunt> dump pigdata;
2021-10-09 02:51:38,412 [main] INFO  org.apache.pig.tools.pigstats.ScriptState -
 Pig features used in the script: UNKNOWN
2021-10-09 02:51:38,721 [main] INFO  org.apache.pig.newplan.logical.optimizer.Lo
gicalPlanOptimizer - {RULES_ENABLED=[AddForEach, ColumnMapKeyPrune, DuplicateFor
EachColumnRewrite, GroupByConstParallelSetter, ImplicitSplitInserter, LimitOptim
izer, LoadTypeCastInserter, MergeFilter, MergeForEach, NewPartitionFilterOptimiz
er, PushDownForEachFlatten, PushUpFilter, SplitFilter, StreamTypeCastInserter],
RULES_DISABLED=[FilterLogicExpressionSimplifier, PartitionFilterOptimizer]}
2021-10-09 02:51:39,456 [main] INFO  org.apache.pig.backend.hadoop.executionengi
ne.mapReduceLayer.MRCompiler - File concatenation threshold: 100 optimistic? fal
se
2021-10-09 02:51:39,619 [main] INFO  org.apache.pig.backend.hadoop.executionengi
ne.mapReduceLayer.MultiQueryOptimizer - MR plan size before optimization: 1
2021-10-09 02:51:39,620 [main] INFO  org.apache.pig.backend.hadoop.executionengi
ne.mapReduceLayer.MultiQueryOptimizer - MR plan size after optimization: 1
2021-10-09 02:51:39,819 [main] INFO  org.apache.hadoop.conf.Configuration.deprec
ation - session.id is deprecated. Instead, use dfs.metrics.session-id
2021-10-09 02:51:39,824 [main] INFO  org.apache.hadoop.metrics.jvm.JvmMetrics -
Initializing JVM Metrics with processName=JobTracker, sessionId=
2021-10-09 02:51:40,023 [main] INFO  org.apache.pig.tools.pigstats.ScriptState -
 Pig script settings are added to the job
2021-10-09 02:51:40,317 [main] INFO  org.apache.hadoop.conf.Configuration.deprec
ation - mapred.job.reduce.markreset.buffer.percent is deprecated. Instead, use m
apreduce.reduce.markreset.buffer.percent
2021-10-09 02:51:40,318 [main] INFO  org.apache.pig.backend.hadoop.executionengi
ne.mapReduceLayer.JobControlCompiler - mapred.job.reduce.markreset.buffer.percen
t is not set, set to default 0.3
2021-10-09 02:51:40,318 [main] INFO  org.apache.hadoop.conf.Configuration.deprec
ation - mapred.output.compress is deprecated. Instead, use mapreduce.output.file
outputformat.compress
2021-10-09 02:51:40,423 [main] INFO  org.apache.pig.backend.hadoop.executionengi
ne.mapReduceLayer.JobControlCompiler - Setting up single store job
```


Devices Help
Applications Places System
Sat Oct 9, 3:42 AM
cloudera
cloudera@quickstart:~
File Edit View Search Terminal Help

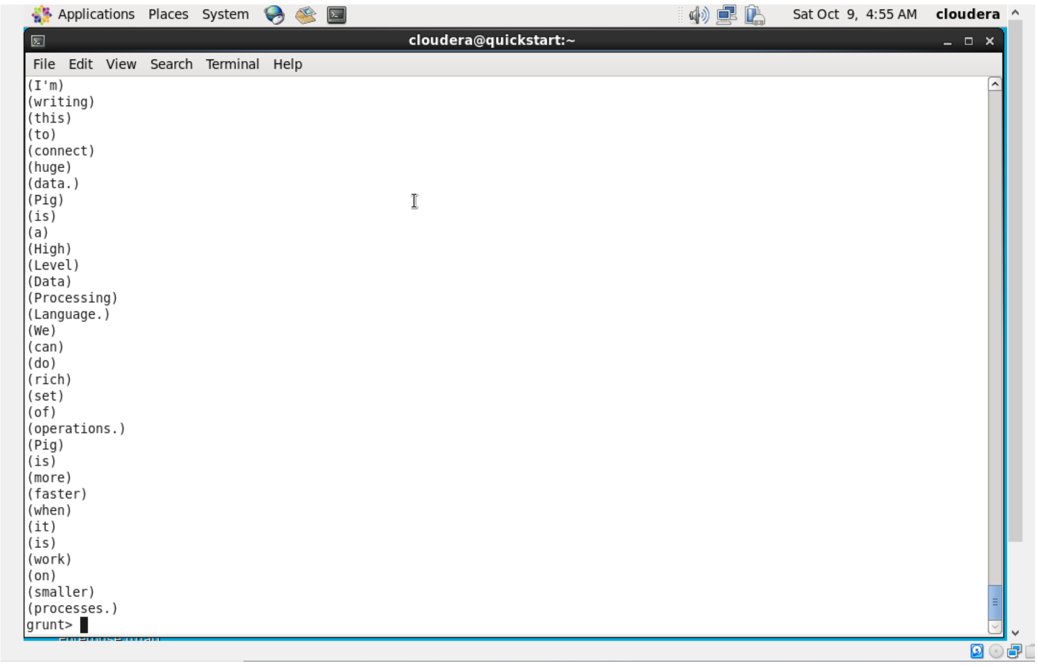
```
2021-10-09 02:52:13,508 [main] INFO  org.apache.pig.backend.hadoop.executionengi
ne.mapReduceLayer.MapReduceLauncher - Success!
2021-10-09 02:52:13,568 [main] INFO  org.apache.hadoop.conf.Configuration.deprec
ation - fs.default.name is deprecated. Instead, use fs.defaultFS
2021-10-09 02:52:13,568 [main] INFO  org.apache.hadoop.conf.Configuration.deprec
ation - mapred.job.tracker is deprecated. Instead, use mapreduce.jobtracker.addr
ess
2021-10-09 02:52:13,569 [main] INFO  org.apache.hadoop.conf.Configuration.deprec
ation - io.bytes.per.checksum is deprecated. Instead, use dfs.bytes-per-checksum
2021-10-09 02:52:13,570 [main] WARN  org.apache.pig.data.SchemaTupleBackend - Sc
hemaTupleBackend has already been initialized
2021-10-09 02:52:14,248 [main] INFO  org.apache.hadoop.mapreduce.lib.input.FileI
nputFormat - Total input paths to process : 1
2021-10-09 02:52:14,249 [main] INFO  org.apache.pig.backend.hadoop.executionengi
ne.util.MapRedUtil - Total input paths to process : 1
(PigLatinScript is used to write in Pig.)
(I'm writing this to connect huge data.)
(Pig is a High Level Data Processing Language.)
(We can do rich set of operations.)
(Pig is more faster when it is work on smaller processes.)
grunt> 
```

Step 3: flatten the words in each line and save separated words into a variable

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



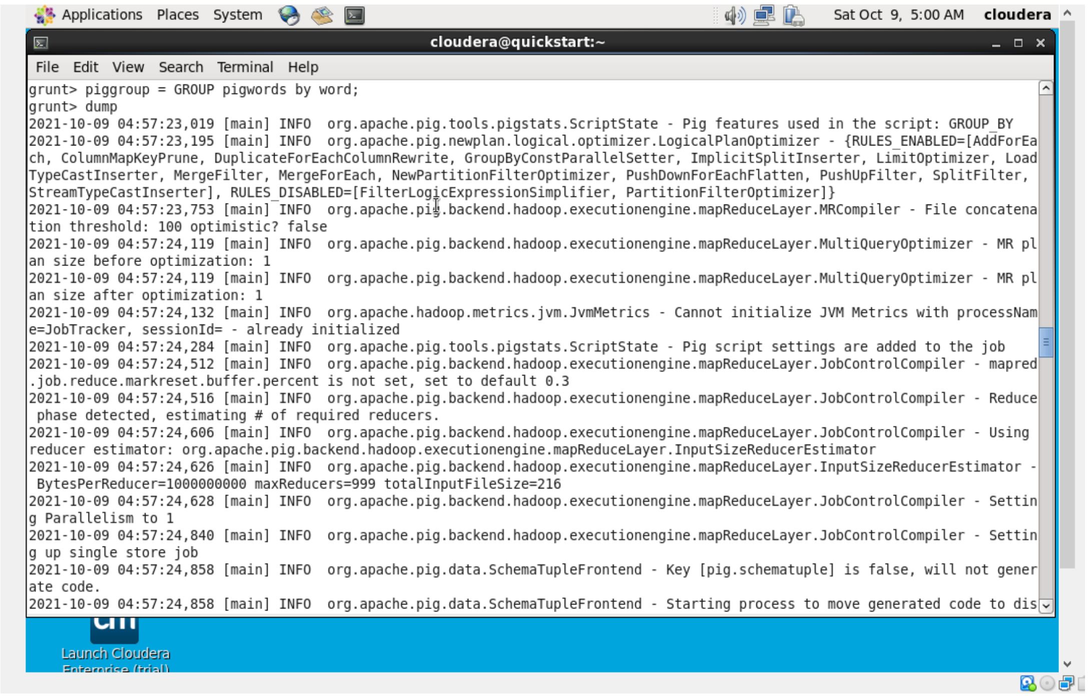
```
grunt> pigwords = foreach pigdata GENERATE FLATTEN(TOKENIZE(line,' ')) as word;
grunt>
grunt> dump pigwords;
2021-10-09 03:49:25,552 [main] INFO  org.apache.pig.tools.pigstats.ScriptState - Pig features used in the script: UNKNOWN
2021-10-09 03:49:25,724 [main] INFO  org.apache.pig.newplan.logical.optimizer.LogicalPlanOptimizer - {RULES_ENABLED=[AddForEa
ch, ColumnMapKeyPrune, DuplicateForEachColumnRewrite, GroupByConstParallelSetter, ImplicitSplitInserter, LimitOptimizer, Load
TypeCastInserter, MergeFilter, MergeForEach, NewPartitionFilterOptimizer, PushDownForEachFlatten, PushUpFilter, SplitFilter,
StreamTypeCastInserter], RULES_DISABLED=[FilterLogicExpressionSimplifier, PartitionFilterOptimizer]}
2021-10-09 03:49:26,096 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MRCompiler - File concatena
tion threshold: 100 optimistic? false
2021-10-09 03:49:26,446 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MultiQueryOptimizer - MR pl
an size before optimization: 1
2021-10-09 03:49:26,447 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MultiQueryOptimizer - MR pl
an size after optimization: 1
2021-10-09 03:49:26,547 [main] INFO  org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM Metrics with processNam
e=JobTracker, sessionId= - already initialized
2021-10-09 03:49:26,676 [main] INFO  org.apache.pig.tools.pigstats.ScriptState - Pig script settings are added to the job
2021-10-09 03:49:27,098 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.JobControlCompiler - mapred
.job.reduce.markreset.buffer.percent is not set, set to default 0.3
2021-10-09 03:49:28,299 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.JobControlCompiler - Settin
g up single store job
2021-10-09 03:49:28,319 [main] INFO  org.apache.pig.data.SchemaTupleFrontend - Key [pig.schematuple] is false, will not gener
ate code.
2021-10-09 03:49:28,319 [main] INFO  org.apache.pig.data.SchemaTupleFrontend - Starting process to move generated code to dis
tributed cache
2021-10-09 03:49:28,320 [main] INFO  org.apache.pig.data.SchemaTupleFrontend - Distributed cache not supported or needed in l
ocal mode. Setting key [pig.schematuple.local.dir] with code temp directory: /tmp/1633776568301-0
2021-10-09 03:49:28,691 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - 1 map-r
educe job(s) waiting for submission.
2021-10-09 03:49:28,887 [JobControl] INFO  org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM Metrics with proc
essName=JobTracker, sessionId= - already initialized
2021-10-09 03:49:29,153 [JobControl] WARN  org.apache.hadoop.mapreduce.JobResourceUploader - No job jar file set.  User class
es may not be found. See Job or Job#setJar(String).
2021-10-09 03:49:29,339 [JobControl] INFO  org.apache.hadoop.mapreduce.lib.input.FileInputFormat - Total input paths to proce
```



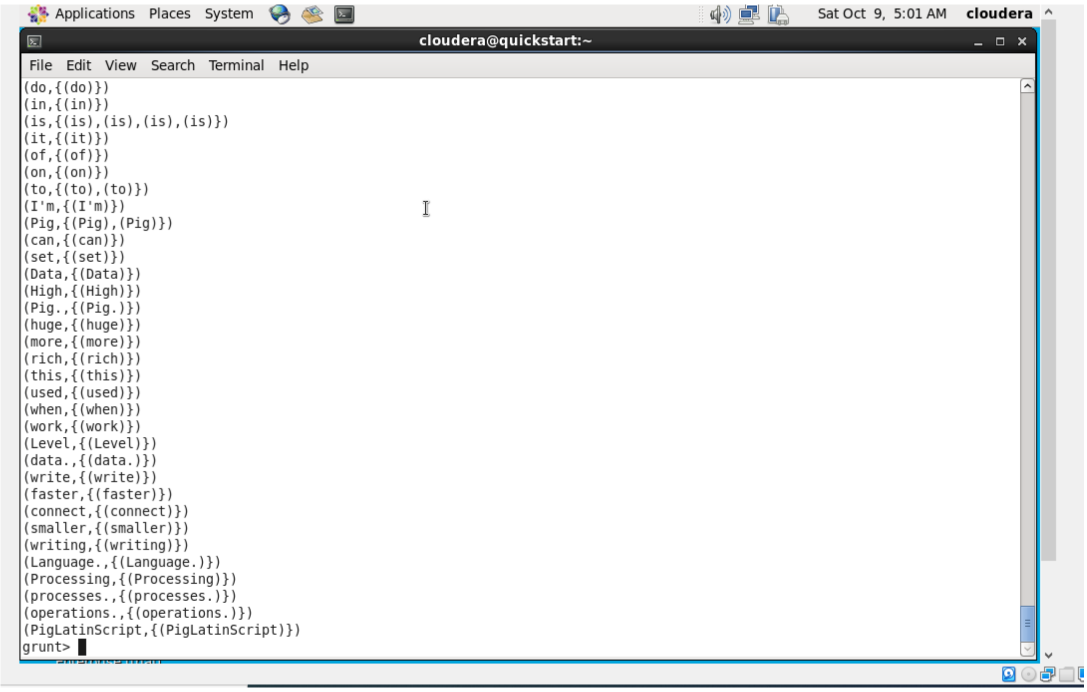
```
(I'm)
(writing)
(this)
(to)
(connect)
(huge)
(data.)
(Pig)
(is)
(a)
(High)
(Level)
(Data)
(Processing)
(Language.)
(We)
(can)
(do)
(rich)
(set)
(of)
(operations.)
(Pig)
(is)
(more)
(faster)
(when)
(it)
(is)
(work)
(on)
(smaller)
(processes.)
grunt> 
```

Step 4: Group the similar words and save into a variable

grunt>groupwords = ______ wordsinline by word;
grunt>dump groupwords;
grunt>describe groupwords;



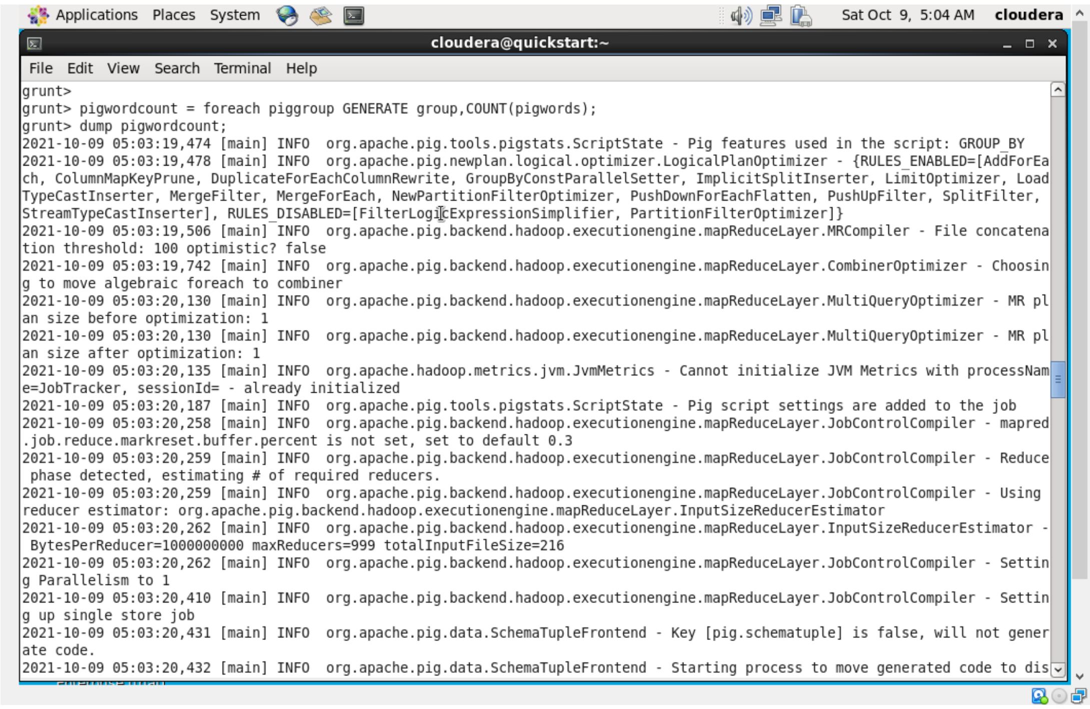
```
grunt> piggroup = GROUP pigwords by word;
grunt> dump
2021-10-09 04:57:23,019 [main] INFO  org.apache.pig.tools.pigstats.ScriptState - Pig features used in the script: GROUP_BY
2021-10-09 04:57:23,195 [main] INFO  org.apache.pig.newplan.logical.optimizer.LogicalPlanOptimizer - {RULES_ENABLED=[AddForEa
ch, ColumnMapKeyPrune, DuplicateForEachColumnRewrite, GroupByConstParallelSetter, ImplicitSplitInserter, LimitOptimizer, Load
TypeCastInserter, MergeFilter, MergeForEach, NewPartitionFilterOptimizer, PushDownForEachFlatten, PushUpFilter, SplitFilter,
StreamTypeCastInserter], RULES_DISABLED=[FilterLogicExpressionSimplifier, PartitionFilterOptimizer]}
2021-10-09 04:57:23,753 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MRCompiler - File concatena
tion threshold: 100 optimistic? false
2021-10-09 04:57:24,119 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MultiQueryOptimizer - MR pl
an size before optimization: 1
2021-10-09 04:57:24,119 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MultiQueryOptimizer - MR pl
an size after optimization: 1
2021-10-09 04:57:24,132 [main] INFO  org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM Metrics with processNam
e=JobTracker, sessionId= - already initialized
2021-10-09 04:57:24,284 [main] INFO  org.apache.pig.tools.pigstats.ScriptState - Pig script settings are added to the job
2021-10-09 04:57:24,512 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.JobControlCompiler - mapred
.job.reduce.markreset.buffer.percent is not set, set to default 0.3
2021-10-09 04:57:24,516 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.JobControlCompiler - Reduce
 phase detected, estimating # of required reducers.
2021-10-09 04:57:24,606 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.JobControlCompiler - Using
reducer estimator: org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.InputSizeReducerEstimator
2021-10-09 04:57:24,626 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.InputSizeReducerEstimator -
 BytesPerReducer=1000000000 maxReducers=999 totalInputFileSize=216
2021-10-09 04:57:24,628 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.JobControlCompiler - Settin
g Parallelism to 1
2021-10-09 04:57:24,840 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.JobControlCompiler - Settin
g up single store job
2021-10-09 04:57:24,858 [main] INFO  org.apache.pig.data.SchemaTupleFrontend - Key [pig.schematuple] is false, will not gener
ate code.
2021-10-09 04:57:24,858 [main] INFO  org.apache.pig.data.SchemaTupleFrontend - Starting process to move generated code to dis
```



```
(do,{(do)})
(in,{(in)})
(is,{(is),(is),(is),(is)})
(it,{(it)})
(of,{(of)})
(on,{(on)})
(to,{(to),(to)})
(I'm,{(I'm)})
(Pig,{(Pig),(Pig)})
(can,{(can)})
(set,{(set)})
(Data,{(Data)})
(High,{(High)})
(Pig.,{(Pig.)})
(huge,{(huge)})
(more,{(more)})
(rich,{(rich)})
(this,{(this)})
(used,{(used)})
(when,{(when)})
(work,{(work)})
(Level,{(Level)})
(data.,{(data.)})
(write,{(write)})
(faster,{(faster)})
(connect,{(connect)})
(smaller,{(smaller)})
(writing,{(writing)})
(Language.,{(Language.)})
(Processing,{(Processing)})
(processes.,{(processes.)})
(operations.,{(operations.)})
(PigLatinScript,{(PigLatinScript)})
grunt> 
```

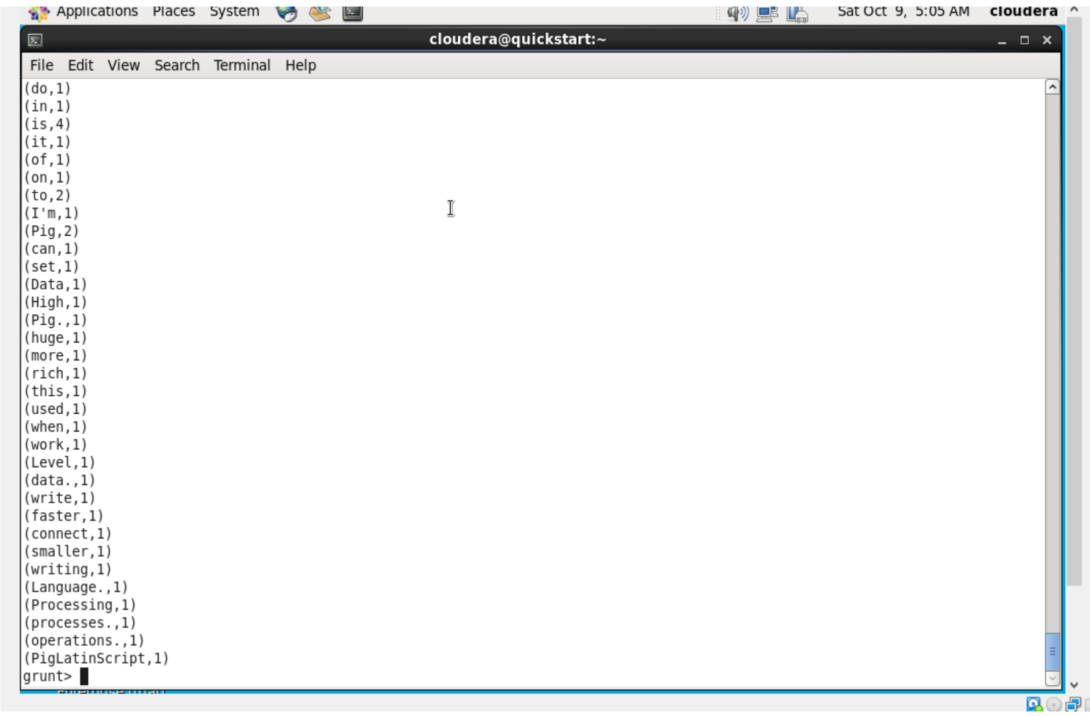
Step 5: Count Words in the group.

grunt>countwords = foreach ____________________;
grunt>DUMP countwords;



Applications Places System
Sat Oct 9, 5:04 AM cloudera
cloudera@quickstart:~
File Edit View Search Terminal Help

```
grunt>
grunt> pigwordcount = foreach piggroup GENERATE group,COUNT(pigwords);
grunt> dump pigwordcount;
2021-10-09 05:03:19,474 [main] INFO  org.apache.pig.tools.pigstats.ScriptState - Pig features used in the script: GROUP_BY
2021-10-09 05:03:19,478 [main] INFO  org.apache.pig.newplan.logical.optimizer.LogicalPlanOptimizer - {RULES_ENABLED=[AddForEa
ch, ColumnMapKeyPrune, DuplicateForEachColumnRewrite, GroupByConstParallelSetter, ImplicitSplitInserter, LimitOptimizer, Load
TypeCastInserter, MergeFilter, MergeForEach, NewPartitionFilterOptimizer, PushDownForEachFlatten, PushUpFilter, SplitFilter,
StreamTypeCastInserter], RULES_DISABLED=[FilterLogicExpressionSimplifier, PartitionFilterOptimizer]}
2021-10-09 05:03:19,506 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MRCompiler - File concatena
tion threshold: 100 optimistic? false
2021-10-09 05:03:19,742 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.CombinerOptimizer - Choosin
g to move algebraic foreach to combiner
2021-10-09 05:03:20,130 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MultiQueryOptimizer - MR pl
an size before optimization: 1
2021-10-09 05:03:20,130 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MultiQueryOptimizer - MR pl
an size after optimization: 1
2021-10-09 05:03:20,135 [main] INFO  org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM Metrics with processNam
e=JobTracker, sessionId= - already initialized
2021-10-09 05:03:20,187 [main] INFO  org.apache.pig.tools.pigstats.ScriptState - Pig script settings are added to the job
2021-10-09 05:03:20,258 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.JobControlCompiler - mapred
.job.reduce.markreset.buffer.percent is not set, set to default 0.3
2021-10-09 05:03:20,259 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.JobControlCompiler - Reduce
 phase detected, estimating # of required reducers.
2021-10-09 05:03:20,259 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.JobControlCompiler - Using
reducer estimator: org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.InputSizeReducerEstimator
2021-10-09 05:03:20,262 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.InputSizeReducerEstimator -
 BytesPerReducer=1000000000 maxReducers=999 totalInputFileSize=216
2021-10-09 05:03:20,262 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.JobControlCompiler - Settin
g Parallelism to 1
2021-10-09 05:03:20,410 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.JobControlCompiler - Settin
g up single store job
2021-10-09 05:03:20,431 [main] INFO  org.apache.pig.data.SchemaTupleFrontend - Key [pig.schematuple] is false, will not gener
ate code.
2021-10-09 05:03:20,432 [main] INFO  org.apache.pig.data.SchemaTupleFrontend - Starting process to move generated code to dis
```



Applications Places System
Sat Oct 9, 5:05 AM cloudera
cloudera@quickstart:~
File Edit View Search Terminal Help

```
(do,1)
(in,1)
(is,4)
(it,1)
(of,1)
(on,1)
(to,2)
(I'm,1)
(Pig,2)
(can,1)
(set,1)
(Data,1)
(High,1)
(Pig.,1)
(huge,1)
(more,1)
(rich,1)
(this,1)
(used,1)
(when,1)
(work,1)
(Level,1)
(data.,1)
(write,1)
(faster,1)
(connect,1)
(smaller,1)
(writing,1)
(Language.,1)
(Processing,1)
(processes.,1)
(operations.,1)
(PigLatinScript,1)
grunt>
```