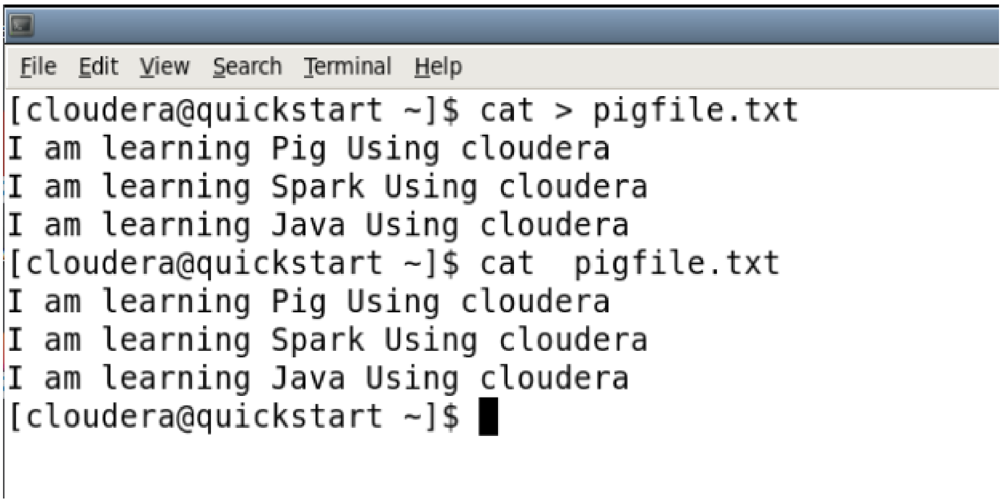
Exercise 07: Word Count using Pig Grouping

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 With following content.

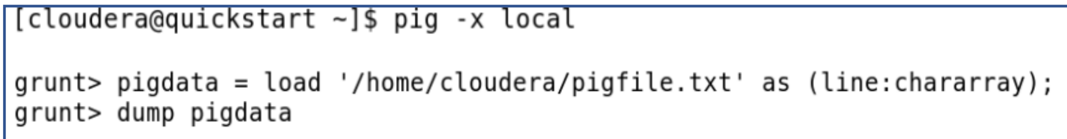


```
[cloudera@quickstart ~]$ cat > pigfile.txt
I am learning Pig Using cloudera
I am learning Spark Using cloudera
I am learning Java Using cloudera
[cloudera@quickstart ~]$ cat  pigfile.txt
I am learning Pig Using cloudera
I am learning Spark Using cloudera
I am learning Java Using cloudera
[cloudera@quickstart ~]$ 
```

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

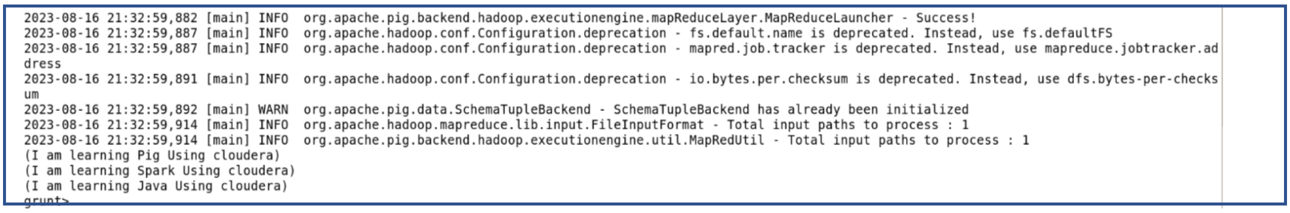
Step 2 : Load the file stored in hdfs with variable 'in1' and each line have to store in 'line' (Space separated file)

grunt>DUMP wordsinline;



```
[cloudera@quickstart ~]$ pig -x local

grunt> pigdata = load '/home/cloudera/pigfile.txt' as (line:chararray);
grunt> dump pigdata
```

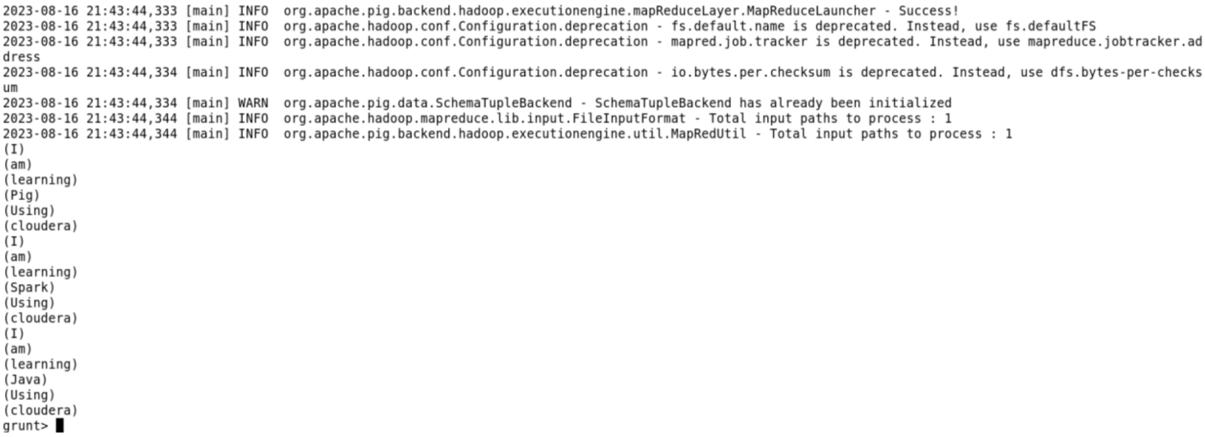


```
2023-08-16 21:32:59,882 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - Success!
2023-08-16 21:32:59,887 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Instead, use fs.defaultFS
2023-08-16 21:32:59,887 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - mapred.job.tracker is deprecated. Instead, use mapreduce.jobtracker.ad
dress
2023-08-16 21:32:59,891 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - io.bytes.per.checksum is deprecated. Instead, use dfs.bytes-per-checks
um
2023-08-16 21:32:59,892 [main] WARN  org.apache.pig.data.SchemaTupleBackend - SchemaTupleBackend has already been initialized
2023-08-16 21:32:59,914 [main] INFO  org.apache.hadoop.mapreduce.lib.input.FileInputFormat - Total input paths to process : 1
2023-08-16 21:32:59,914 [main] INFO  org.apache.pig.backend.hadoop.executionengine.util.MapRedUtil - Total input paths to process : 1
(I am learning Pig Using cloudera)
(I am learning Spark Using cloudera)
(I am learning Java Using cloudera)
grunt>
```

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

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

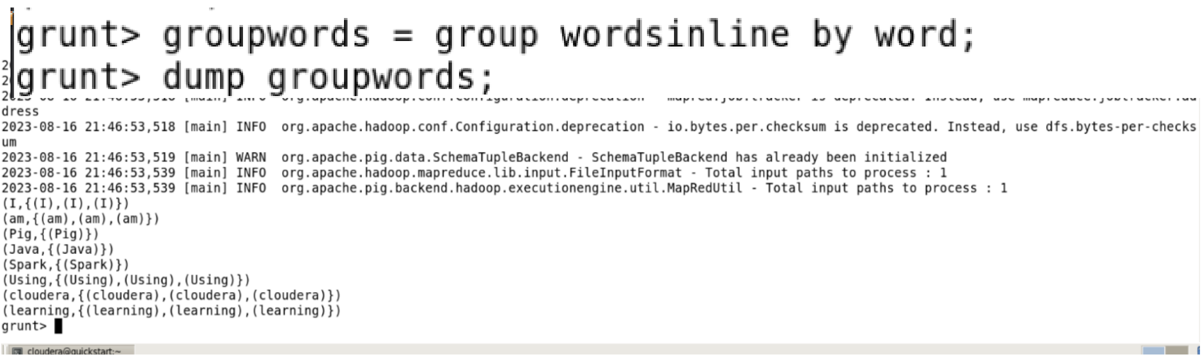
grunt>wordsinline = FOREACH input1 GENERATE flatten(TOKENIZE(line, ' ')) as word;



```
2023-08-16 21:43:44,333 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - Success!
2023-08-16 21:43:44,333 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Instead, use fs.defaultFS
2023-08-16 21:43:44,333 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - mapred.job.tracker is deprecated. Instead, use mapreduce.jobtracker.ad
dress
2023-08-16 21:43:44,334 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - io.bytes.per.checksum is deprecated. Instead, use dfs.bytes-per-checks
um
2023-08-16 21:43:44,334 [main] WARN  org.apache.pig.data.SchemaTupleBackend - SchemaTupleBackend has already been initialized
2023-08-16 21:43:44,344 [main] INFO  org.apache.hadoop.mapreduce.lib.input.FileInputFormat - Total input paths to process : 1
2023-08-16 21:43:44,344 [main] INFO  org.apache.pig.backend.hadoop.executionengine.util.MapRedUtil - Total input paths to process : 1
(I)
(am)
(learning)
(Pig)
(Using)
(cloudera)
(I)
(am)
(learning)
(Spark)
(Using)
(cloudera)
(I)
(am)
(learning)
(Java)
(Using)
(cloudera)
grunt> █
```

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

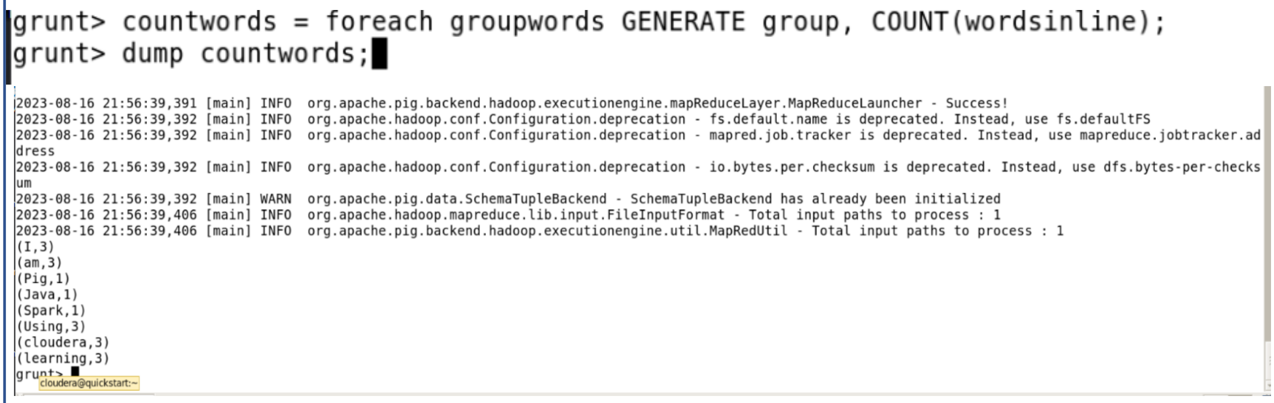
grunt>groupwords = _______ wordsinline by word;
grunt>dump groupwords; grunt>describe
groupwords;



```
grunt> groupwords = group wordsinline by word;
grunt> dump groupwords;
2023-08-16 21:46:53,518 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - mapred.job.tracker is deprecated. Instead, use mapreduce.jobtracker.ad
dress
2023-08-16 21:46:53,518 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - io.bytes.per.checksum is deprecated. Instead, use dfs.bytes-per-checks
um
2023-08-16 21:46:53,519 [main] WARN  org.apache.pig.data.SchemaTupleBackend - SchemaTupleBackend has already been initialized
2023-08-16 21:46:53,539 [main] INFO  org.apache.hadoop.mapreduce.lib.input.FileInputFormat - Total input paths to process : 1
2023-08-16 21:46:53,539 [main] INFO  org.apache.pig.backend.hadoop.executionengine.util.MapRedUtil - Total input paths to process : 1
(I,{(I),(I),(I)})
(am,{(am),(am),(am)})
(Pig,{(Pig)})
(Java,{(Java)})
(Spark,{(Spark)})
(Using,{(Using),(Using),(Using)})
(cloudera,{(cloudera),(cloudera),(cloudera)})
(learning,{(learning),(learning),(learning)})
grunt> 
```

Step 5: Count Words in the group.

grunt>countwords = foreach _______________________; grunt>DUMP
countwords;



```
grunt> countwords = foreach groupwords GENERATE group, COUNT(wordsinline);
grunt> dump countwords;
2023-08-16 21:56:39,391 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - Success!
2023-08-16 21:56:39,392 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Instead, use fs.defaultFS
2023-08-16 21:56:39,392 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - mapred.job.tracker is deprecated. Instead, use mapreduce.jobtracker.ad
dress
2023-08-16 21:56:39,392 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - io.bytes.per.checksum is deprecated. Instead, use dfs.bytes-per-checks
um
2023-08-16 21:56:39,392 [main] WARN  org.apache.pig.data.SchemaTupleBackend - SchemaTupleBackend has already been initialized
2023-08-16 21:56:39,406 [main] INFO  org.apache.hadoop.mapreduce.lib.input.FileInputFormat - Total input paths to process : 1
2023-08-16 21:56:39,406 [main] INFO  org.apache.pig.backend.hadoop.executionengine.util.MapRedUtil - Total input paths to process : 1
(I,3)
(am,3)
(Pig,1)
(Java,1)
(Spark,1)
(Using,3)
(cloudera,3)
(learning,3)
grunt> 
```