

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*

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

Cat > pig.txt

Pig -x mapreduce

I am learning Pig Using cloudera

I am learning Spark Using

cloudera I am learning Java

Using clouder

The screenshot shows a terminal window titled 'cloudera@quickstart:~' with the following commands and output:

```
[cloudera@quickstart ~]$ cat > pig.txt
I am Learning pig using Cloudera
I am Learning spark using Cloudera
I am Learning Java using Cloudera
^C
[cloudera@quickstart ~]$ hdfs dfs -mkdir in001
[cloudera@quickstart ~]$ hdfs dfs -put pig.txt /in001
[cloudera@quickstart ~]$ hdfs dfs -ls
Found 1 items
drwxr-xr-x - cloudera cloudera 0 2022-08-29 23:33 in001
[cloudera@quickstart ~]$ hdfs dfs -ls /in001
-rw-r--r-- 1 cloudera supergroup 102 2022-08-29 23:34 /in001/
[cloudera@quickstart ~]$ hdfs dfs -ls /in001/
-rw-r--r-- 1 cloudera supergroup 102 2022-08-29 23:34 /in001/
[cloudera@quickstart ~]$ ls
cm_api.py Downloads kerberos Pictures Videos
Desktop eclipse lib pig.txt workspace
Documents express-deployment.json Music Public
[cloudera@quickstart ~]$ hdfs dfs -ls /
-ls/: Unknown command
[cloudera@quickstart ~]$ hdfs dfs -ls /
Found 7 items
drwxr-xr-x - hdfs supergroup 0 2017-10-23 09:15 /benchmarks
drwxr-xr-x - hbase supergroup 0 2022-08-29 23:16 /hbase
-rw-r--r-- 1 cloudera supergroup 102 2022-08-29 23:34 /in001
drwxr-xr-x - solr solr 0 2017-10-23 09:18 /solr
drwxr-xr-x - hdfs supergroup 0 2022-08-29 23:16 /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 ~]$ hdfs dfs -mkdir /in002
[cloudera@quickstart ~]$ hdfs dfs -put pig.txt /in002
[cloudera@quickstart ~]$ hdfs dfs -ls /in002/
Found 1 items
-rw-r--r-- 1 cloudera supergroup 102 2022-08-29 23:48 /in002/pig.txt
[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 info.
2022-08-29 23:49:44,965 [main] INFO org.apache.pig.Main - Apache Pig version 0.12.0-cdh5.13.0 (reexported) compiled Oct 04 2017, 11:09:03
```

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

Inputs = LOAD ‘/in002/pig.txt’ AS (f1:chararray);

(I am learning Pig Using cloudera)

(I am learning Spark Using cloudera)

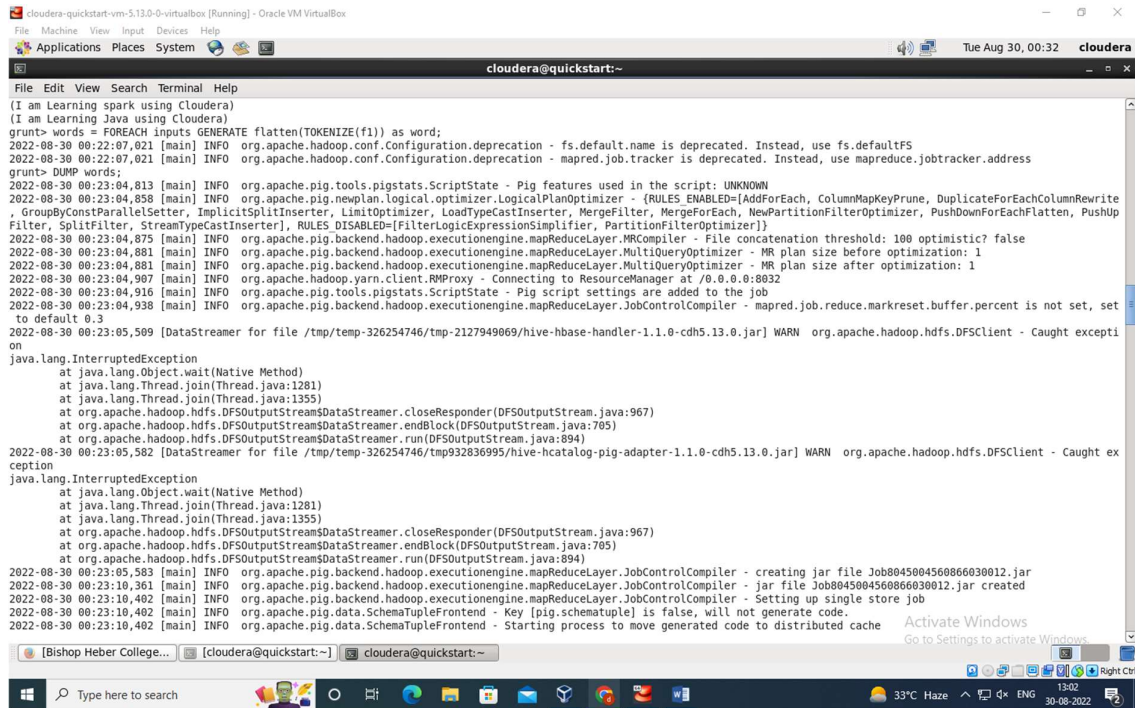
(I am learning Java Using cloudera)

```
cloudera-quickstart-vm-5.13.0-0-virtualbox [Running] - Oracle VM VirtualBox
File Machine View Input Devices Help
Applications Places System
cloudera@quickstart:~
File Edit View Search Terminal Help
Details at logfile: /home/cloudera/pig 1661843222660.log
grunt> [cloudera@quickstart ~]$ pig -x mapreduce
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 info.
2022-08-30 00:11:29,705 [main] INFO org.apache.pig.Main - Apache Pig version 0.12.0-cdh5.13.0 (reexported) compiled Oct 04 2017, 11:09:03
2022-08-30 00:11:29,706 [main] INFO org.apache.pig.Main - Logging error messages to: /home/cloudera/pig 1661843489673.log
2022-08-30 00:11:29,732 [main] INFO org.apache.pig.impl.util.Utils - Default bootstrap file /home/cloudera/.pigbootstrap not found
2022-08-30 00:11:30,988 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - mapred.job.tracker is deprecated. Instead, use mapreduce.jobtracker.address
2022-08-30 00:11:30,988 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Instead, use fs.defaultFS
2022-08-30 00:11:30,988 [main] INFO org.apache.pig.backend.hadoop.executionengine.HExecutionEngine - Connecting to hadoop file system at: hdfs://quickstart.cloudera:8020
2022-08-30 00:11:33,459 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - mapred.job.tracker is deprecated. Instead, use mapreduce.jobtracker.address
2022-08-30 00:11:33,459 [main] INFO org.apache.pig.backend.hadoop.executionengine.HExecutionEngine - Connecting to map-reduce job tracker at: localhost:8021
2022-08-30 00:11:33,460 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Instead, use fs.defaultFS
2022-08-30 00:11:33,600 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Instead, use fs.defaultFS
2022-08-30 00:11:33,603 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - mapred.job.tracker is deprecated. Instead, use mapreduce.jobtracker.address
2022-08-30 00:11:33,788 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Instead, use fs.defaultFS
2022-08-30 00:11:33,790 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - mapred.job.tracker is deprecated. Instead, use mapreduce.jobtracker.address
2022-08-30 00:11:33,970 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Instead, use fs.defaultFS
2022-08-30 00:11:33,974 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - mapred.job.tracker is deprecated. Instead, use mapreduce.jobtracker.address
2022-08-30 00:11:34,144 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Instead, use fs.defaultFS
2022-08-30 00:11:34,148 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - mapred.job.tracker is deprecated. Instead, use mapreduce.jobtracker.address
2022-08-30 00:11:34,325 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Instead, use fs.defaultFS
2022-08-30 00:11:34,328 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - mapred.job.tracker is deprecated. Instead, use mapreduce.jobtracker.address
2022-08-30 00:11:34,494 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Instead, use fs.defaultFS
2022-08-30 00:11:34,497 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - mapred.job.tracker is deprecated. Instead, use mapreduce.jobtracker.address
2022-08-30 00:11:34,659 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Instead, use fs.defaultFS
2022-08-30 00:11:34,666 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - mapred.job.tracker is deprecated. Instead, use mapreduce.jobtracker.address
2022-08-30 00:11:34,825 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Instead, use fs.defaultFS
2022-08-30 00:11:34,825 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - mapred.job.tracker is deprecated. Instead, use mapreduce.jobtracker.address
grunt> input = LOAD '/in002/pig.txt' AS (f1:chararray);
2022-08-30 00:12:20,317 [main] ERROR org.apache.pig.tools.grunt.Grunt - ERROR 1200: <Line 1, column 0> mismatched input 'input' expecting EOF
Details at logfile: /home/cloudera/pig 1661843489673.log
grunt> input = LOAD08020 'in002/pig.txt' AS (f1:chararray);
2022-08-30 00:13:24,245 [main] ERROR org.apache.pig.tools.grunt.Grunt - ERROR 1200: <Line 1, column 0> mismatched input 'input' expecting EOF
Details at logfile: /home/cloudera/pig 1661843489673.log
grunt> input = LOAD '/home/cloudera/pig.txt' AS (f1:chararray);
2022-08-30 00:13:37,907 [main] ERROR org.apache.pig.tools.grunt.Grunt - ERROR 1200: <Line 1, column 0> mismatched input 'input' expecting EOF
Details at logfile: /home/cloudera/pig 1661843489673.log
grunt> inputs = LOAD '/in002/pig.txt' AS (f1:chararray);
Activate Windows
Go to Settings to activate Windows.
```

```
cloudera-quickstart-vm-5.13.0-0-virtualbox [Running] - Oracle VM VirtualBox
File Machine View Input Devices Help
Applications Places System
cloudera@quickstart:~
File Edit View Search Terminal Help
HadoopVersion PigVersion UserId StartedAt FinishedAt Features
2.6.0-cdh5.13.0 0.12.0-cdh5.13.0 cloudera 2022-08-30 00:15:03 2022-08-30 00:15:54 UNKNOWN
Success!
Job Stats (time in seconds):
JobId Maps Reduces MaxMapTime MinMapTime AvgMapTime MedianMapTime MaxReduceTime MinReduceTime AvgReduceTime MedianReduceTime Alias Feature 0
utputs
job_166184066129_0001 1 0 11 11 11 11 n/a n/a n/a n/a inputs MAP_ONLY hdfs://quickstart.cloudera:8020/tmp/temp-32625474
6/tmp1402843552,
Input(s):
Successfully read 3 records (466 bytes) from: "/in002/pig.txt"
Output(s):
Successfully stored 3 records (121 bytes) in: "hdfs://quickstart.cloudera:8020/tmp/temp-326254746/tmp1402843552"
Counters:
Total records written : 3
Total bytes written : 121
Spillable Memory Manager spill count : 0
Total bags proactively spilled: 0
Total records proactively spilled: 0
Job DAG:
job_166184066129_0001
2022-08-30 00:15:54,291 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - Success!
2022-08-30 00:15:54,311 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Instead, use fs.defaultFS
2022-08-30 00:15:54,311 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - mapred.job.tracker is deprecated. Instead, use mapreduce.jobtracker.address
2022-08-30 00:15:54,313 [main] INFO org.apache.pig.data.SchemaTupleBackend - Key [pig.schemaTuple] was not set... will not generate code.
2022-08-30 00:15:54,343 [main] INFO org.apache.hadoop.mapreduce.lib.input.FileInputFormat - Total input paths to process : 1
2022-08-30 00:15:54,343 [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> words = FOREACH inputs GENERATE flatten(TOKENIZE(f1)) as word;
2022-08-30 00:22:07,021 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Instead, use fs.defaultFS
Activate Windows
Go to Settings to activate Windows.
```

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

```
grunt> words= FOREACH inputs GENERATE  
flatten(TOKENIZE(f1)) as word;
```



```
cloudera-quickstart-vm-5.13.0-0-virtualbox [Running] - Oracle VM VirtualBox  
File Machine View Input Devices Help  
Applications Places System  
cloudera@quickstart:~  
File Edit View Search Terminal Help  
(I am Learning Spark using Cloudera)  
(I am Learning Java using Cloudera)  
grunt> words= FOREACH inputs GENERATE flatten(TOKENIZE(f1)) as word;  
2022-08-30 00:22:07,021 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Instead, use fs.defaultFS  
2022-08-30 00:22:07,021 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - mapred.job.tracker is deprecated. Instead, use mapreduce.jobtracker.address  
grunt> DUMP words;  
2022-08-30 00:23:04,813 [main] INFO org.apache.pig.tools.pigstats.ScriptState - Pig features used in the script: UNKNOWN  
2022-08-30 00:23:04,858 [main] INFO org.apache.pig.newplan.logical.optimizer.LogicalPlanOptimizer - {RULES_ENABLED=[AddForEach, ColumnMapKeyPrune, DuplicateForEachColumnRewrite, GroupByConstParallelSetter, ImplicitSplitInserter, LimitOptimizer, LoadTypeCastInserter, MergeFilter, MergeForEach, NewPartitionFilterOptimizer, PushDownForEachFlatten, PushUpFilter, SplitFilter, StreamTypeCastInserter], RULES_DISABLED=[FilterLogicExpressionSimplifier, PartitionFilterOptimizer]}  
2022-08-30 00:23:04,875 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MRCompiler - File concatenation threshold: 100 optimistic? false  
2022-08-30 00:23:04,881 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MultiQueryOptimizer - MR plan size before optimization: 1  
2022-08-30 00:23:04,881 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MultiQueryOptimizer - MR plan size after optimization: 1  
2022-08-30 00:23:04,907 [main] INFO org.apache.hadoop.yarn.client.RMPProxy - Connecting to ResourceManager at /0.0.0.0:8032  
2022-08-30 00:23:04,916 [main] INFO org.apache.pig.tools.pigstats.ScriptState - Pig script settings are added to the job  
2022-08-30 00:23:04,938 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.JobControlCompiler - mapred.job.reduce.markreset.buffer.percent is not set, set to default 0.3  
2022-08-30 00:23:05,509 [DataStreamer for file /tmp/temp-326254746/tmp-2127949869/hive-hbase-handler-1.1.0-cdh5.13.0.jar] WARN org.apache.hadoop.hdfs.DFSClient - Caught exception  
java.lang.InterruptedRuntimeException  
    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$DataStreamer.closeResponder(DFSOutputStream.java:967)  
    at org.apache.hadoop.hdfs.DFSOutputStream$DataStreamer.endBlock(DFSOutputStream.java:705)  
    at org.apache.hadoop.hdfs.DFSOutputStream$DataStreamer.run(DFSOutputStream.java:894)  
2022-08-30 00:23:05,582 [DataStreamer for file /tmp/temp-326254746/tmp932836995/hive-hcatalog-pig-adapter-1.1.0-cdh5.13.0.jar] WARN org.apache.hadoop.hdfs.DFSClient - Caught exception  
java.lang.InterruptedRuntimeException  
    at java.lang.Object.wait(Native Method)  
    at java.lang.Thread.join(Thread.java:1281)  
    at java.lang.Thread.join(Thread.java:1355)  
    at org.apache.hadoop.hdfs.DFSOutputStream$DataStreamer.closeResponder(DFSOutputStream.java:967)  
    at org.apache.hadoop.hdfs.DFSOutputStream$DataStreamer.endBlock(DFSOutputStream.java:705)  
    at org.apache.hadoop.hdfs.DFSOutputStream$DataStreamer.run(DFSOutputStream.java:894)  
2022-08-30 00:23:05,583 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.JobControlCompiler - creating jar file Job8045004560866030012.jar  
2022-08-30 00:23:10,361 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.JobControlCompiler - jar file Job8045004560866030012.jar created  
2022-08-30 00:23:10,402 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.JobControlCompiler - Setting up single store job  
2022-08-30 00:23:10,402 [main] INFO org.apache.pig.data.SchemaTupleFrontend - Key [pig.schematuple] is false, will not generate code.  
2022-08-30 00:23:10,402 [main] INFO org.apache.pig.data.SchemaTupleFrontend - Starting process to move generated code to distributed cache  
Activate Windows  
Go to Settings to activate Windows  
[Bishop Heber College...] [cloudera@quickstart:~] [cloudera@quickstart:~]  
Type here to search  
33°C Haze 13:02 30-08-2022
```

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

```
grunt> groupwords = words by word;  
grunt> dump groupwords;  
grunt> describe groupwords;
```

```
cloudera-quickstart-vm-5.13.0-0-virtualbox [Running] - Oracle VM VirtualBox
File Machine View Input Devices Help
Applications Places System
cloudera@quickstart:~
File Edit View Search Terminal Help

Output(s):
Successfully stored 18 records (166 bytes) in: "hdfs://quickstart.cloudera:8020/tmp/326254746/tmp1992686162"

Counters:
Total records written : 18
Total bytes written : 166
Spillable Memory Manager spill count : 0
Total bags proactively spilled: 0
Total records proactively spilled: 0

Job DAG:
job_1661840066129_0002

2022-08-30 00:23:41,380 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - Success!
2022-08-30 00:23:41,380 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Instead, use fs.defaultFS
2022-08-30 00:23:41,380 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - mapred.job.tracker is deprecated. Instead, use mapreduce.jobtracker.address
2022-08-30 00:23:41,381 [main] INFO org.apache.pig.data.SchemaTupleBackend - Key [pig.schematuple] was not set... will not generate code.
2022-08-30 00:23:41,389 [main] INFO org.apache.hadoop.mapreduce.lib.input.FileInputFormat - Total input paths to process : 1
2022-08-30 00:23:41,389 [main] INFO org.apache.pig.backend.hadoop.executionengine.util.MapRedUtil - Total input paths to process : 1

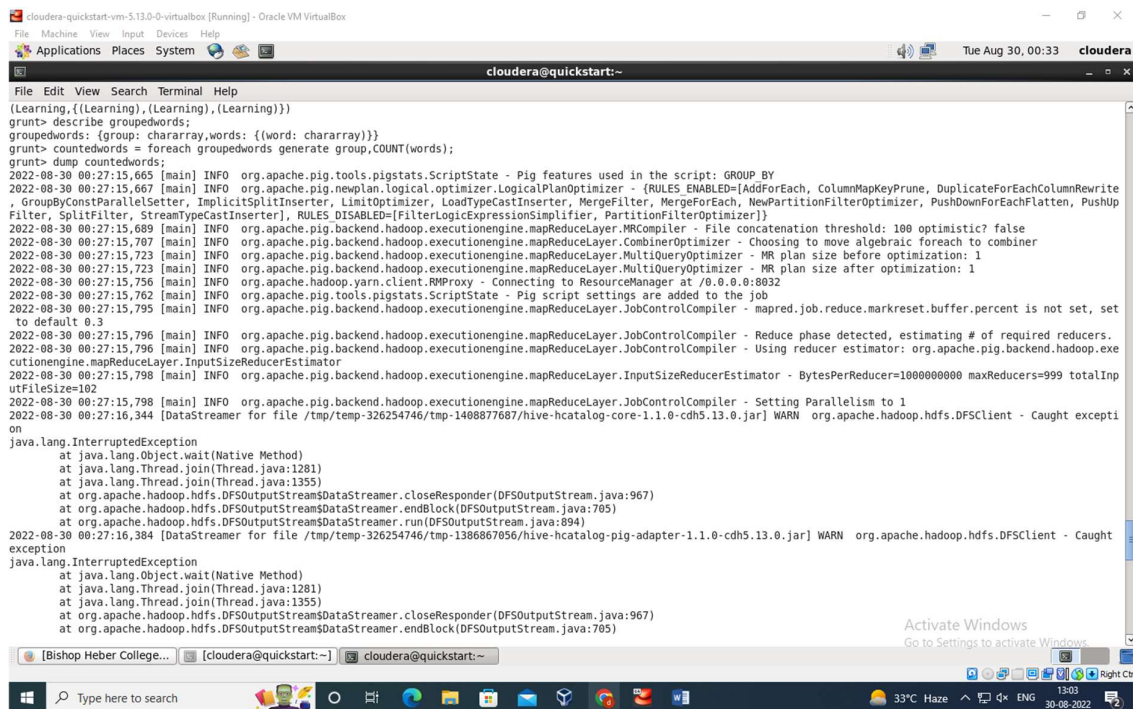
(1)
(am)
(Learning)
(pig)
(using)
(Cloudera)
(1)
(am)
(Learning)
(spark)
(using)
(Cloudera)
(1)
(am)
(Learning)
(Java)
(using)
(Cloudera)
grunt> groupedwords = group words by word;

Activate Windows
Go to Settings to activate Windows.
```

```
cloudera-quickstart-vm-5.13.0-0-virtualbox [Running] - Oracle VM VirtualBox
File Machine View Input Devices Help
Applications Places System
cloudera@quickstart:~
File Edit View Search Terminal Help

(Java)
(using)
(Cloudera)
grunt> groupedwords = group words by word;
grunt> dump groupedwords;
2022-08-30 00:24:57,129 [main] INFO org.apache.pig.tools.pigstats.ScriptState - Pig features used in the script: GROUP BY
2022-08-30 00:24:57,130 [main] INFO org.apache.pig.newplan.logical.optimizer.LogicalPlanOptimizer - {RULES_ENABLED=[AddForEach, ColumnMapKeyPrune, DuplicateForEachColumnRewrite, GroupByConstParallelSetter, ImplicitSplitInserter, LimitOptimizer, LoadTypeCastInserter, MergeFilter, MergeForEach, NewPartitionFilterOptimizer, PushDownForEachFlatten, PushUpFilter, SplitFilter, StreamTypeCastInserter], RULES_DISABLED=[FilterLogicExpressionSimplifier, PartitionFilterOptimizer]}
2022-08-30 00:24:57,140 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MRCompiler - File concatenation threshold: 100 optimistic? false
2022-08-30 00:24:57,148 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MultiQueryOptimizer - MR plan size before optimization: 1
2022-08-30 00:24:57,148 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MultiQueryOptimizer - MR plan size after optimization: 1
2022-08-30 00:24:57,190 [main] INFO org.apache.hadoop.yarn.client.RMProxy - Connecting to ResourceManager at /0.0.0.0:8032
2022-08-30 00:24:57,193 [main] INFO org.apache.pig.tools.pigstats.ScriptState - Pig script settings are added to the job
2022-08-30 00:24:57,209 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.JobControlCompiler - mapred.job.reduce.markreset.buffer.percent is not set, set to default 0.3
2022-08-30 00:24:57,210 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.JobControlCompiler - Reduce phase detected, estimating # of required reducers.
2022-08-30 00:24:57,212 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.JobControlCompiler - Using reducer estimator: org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.InputSizeReducerEstimator
2022-08-30 00:24:57,215 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.InputSizeReducerEstimator - BytesPerReducer=1000000000 maxReducers=999 totalInputFileSize=102
2022-08-30 00:24:57,215 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.JobControlCompiler - Setting Parallelism to 1
2022-08-30 00:24:57,716 [DataStreamer for file /tmp/temp-326254746/tmp-438167692/jdo-api-3.0.1.jar] WARN org.apache.hadoop.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.DFSOutputStreamDataStreamer.closeResponder(DFSOutputStream.java:967)
    at org.apache.hadoop.hdfs.DFSOutputStreamDataStreamer.endBlock(DFSOutputStream.java:705)
    at org.apache.hadoop.hdfs.DFSOutputStreamDataStreamer.run(DFSOutputStream.java:894)
2022-08-30 00:24:57,814 [DataStreamer for file /tmp/temp-326254746/tmp-832828479/hive-hcatalog-pig-adapter-1.1.0-cdh5.13.0.jar] WARN org.apache.hadoop.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.DFSOutputStreamDataStreamer.closeResponder(DFSOutputStream.java:967)
    at org.apache.hadoop.hdfs.DFSOutputStreamDataStreamer.endBlock(DFSOutputStream.java:705)
    at org.apache.hadoop.hdfs.DFSOutputStreamDataStreamer.run(DFSOutputStream.java:894)
2022-08-30 00:24:57,814 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.JobControlCompiler - creating jar file Job955791384431525036.jar

Go to Settings to activate Windows.
```

The screenshot shows a terminal window titled "cloudera@quickstart:~" with a file explorer in the background. The terminal displays the following Pig script and its execution logs:

```
(Learning, {(Learning), (Learning)}, (Learning)})
grunt> describe groupedwords;
groupedwords: (group: chararray, words: {(word: chararray)})
grunt> countwords = foreach groupedwords generate group,COUNT(words);
grunt> dump countwords;
```

The logs show various Pig features being used, including `GROUP BY`, `LogicalPlanOptimizer`, `File concatenation threshold`, `MR plan size before optimization`, `MR plan size after optimization`, `Reduce phase detected`, `Using reducer estimator`, `BytesPerReducer=1000000000`, `maxReducers=999`, `totalInputFileSize=102`, `Setting Parallelism to 1`, and `WARN org.apache.hadoop.hdfs.DFSClient - Caught exception`.

The error messages are:

```
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$DataStreamer.closeResponder(DFSOutputStream.java:967)
    at org.apache.hadoop.hdfs.DFSOutputStream$DataStreamer.endBlock(DFSOutputStream.java:705)
    at org.apache.hadoop.hdfs.DFSOutputStream$DataStreamer.run(DFSOutputStream.java:894)
2022-08-30 00:27:16,384 [DataStreamer for file /tmp/temp-326254746/tmp-1386867856/hive-hcatalog-core-1.1.0-cdh5.13.0.jar] WARN org.apache.hadoop.hdfs.DFSClient - Caught exception
java.lang.InterruptedException
    at java.lang.Object.wait(Native Method)
    at java.lang.Thread.join(Thread.java:1281)
    at java.lang.Thread.join(Thread.java:1355)
    at org.apache.hadoop.hdfs.DFSOutputStream$DataStreamer.closeResponder(DFSOutputStream.java:967)
    at org.apache.hadoop.hdfs.DFSOutputStream$DataStreamer.endBlock(DFSOutputStream.java:705)
```

Step 5: Count Words in the group.

```
grunt>countwords = foreach groupwords generate
group,COUNT(words);
grunt>DUMP countwords;
```

```
cloudera-quickstart-vm-5.13.0-0-virtualbox [Running] - Oracle VM VirtualBox
File Machine View Input Devices Help
Applications Places System
cloudera@quickstart:~
File Edit View Search Terminal Help

Output(s):
Successfully stored 8 records (250 bytes) in: "hdfs://quickstart.cloudera:8020/tmp/temp-326254746/tmp-2070150437"

Counters:
Total records written : 8
Total bytes written : 250
Spillable Memory Manager spill count : 0
Total bags proactively spilled: 0
Total records proactively spilled: 0

Job DAG:
job_1661840066129_0003

2022-08-30 00:25:43,457 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - Success!
2022-08-30 00:25:43,461 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Instead, use fs.defaultFS
2022-08-30 00:25:43,461 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - mapred.job.tracker is deprecated. Instead, use mapreduce.jobtracker.address
2022-08-30 00:25:43,462 [main] INFO org.apache.pig.data.SchemaTupleBackend - Key [pig.schematuple] was not set... will not generate code.
2022-08-30 00:25:43,476 [main] INFO org.apache.hadoop.mapreduce.lib.input.FileInputFormat - Total input paths to process : 1
2022-08-30 00:25:43,476 [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> describe groupedwords;
groupedwords: (group: chararray, words: {(word: chararray)})
grunt> countedwords = foreach groupedwords generate group,COUNT(words);
grunt> dump countedwords;
2022-08-30 00:27:15,665 [main] INFO org.apache.pig.tools.pigstats.ScriptState - Pig features used in the script: GROUP_BY
2022-08-30 00:27:15,667 [main] INFO org.apache.pig.newplan.logical.optimizer.LogicalPlanOptimizer - {RULES_ENABLED=[AddForEach, ColumnMapKeyPrune, DuplicateForEachColumnRewrite, GroupByConstParallelSetter, ImplicitSplitInserter, LimitOptimizer, LoadTypeCastInserter, MergeFilter, MergeForEach, NewPartitionFilterOptimizer, PushDownForEachFlatten, PushUpFilter, SplitFilter, StreamTypeCastInserter], RULES_DISABLED=[FilterLogicExpressionSimplifier, PartitionFilterOptimizer]}
2022-08-30 00:27:15,689 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MRCCompiler - File concatenation threshold: 100 optimistic? false
2022-08-30 00:27:15,707 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.CombinerOptimizer - Choosing to move algebraic foreach to combiner
2022-08-30 00:27:15,723 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MultiQueryOptimizer - MR plan size before optimization: 11 windows
Go to Settings to activate Windows
```

```
cloudera-quickstart-vm-5.13.0-0-virtualbox [Running] - Oracle VM VirtualBox
File Machine View Input Devices Help
Applications Places System
cloudera@quickstart:~
File Edit View Search Terminal Help

Success!

Job Stats (time in seconds):
JobId Maps Reduces MaxMapTime MinMapTime AvgMapTime MedianMapTime MaxReduceTime MinReduceTime AvgReduceTime MedianReduceTime Alias Feature 0
outputs
job_1661840066129_0004 1 1 10 10 10 9 9 9 9 countedwords,groupedwords,inputs,words GROUP_BY,COMBINER hdfs://qu
ickstart.cloudera:8020/tmp/temp-326254746/tmp-808209979,

Input(s):
Successfully read 3 records (466 bytes) from: "/in002/pig.txt"

Output(s):
Successfully stored 8 records (91 bytes) in: "hdfs://quickstart.cloudera:8020/tmp/temp-326254746/tmp-808209979"

Counters:
Total records written : 8
Total bytes written : 91
Spillable Memory Manager spill count : 0
Total bags proactively spilled: 0
Total records proactively spilled: 0

Job DAG:
job_1661840066129_0004

2022-08-30 00:28:02,117 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - Success!
2022-08-30 00:28:02,118 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Instead, use fs.defaultFS
2022-08-30 00:28:02,118 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - mapred.job.tracker is deprecated. Instead, use mapreduce.jobtracker.address
2022-08-30 00:28:02,119 [main] INFO org.apache.pig.data.SchemaTupleBackend - Key [pig.schematuple] was not set... will not generate code.
2022-08-30 00:28:02,127 [main] INFO org.apache.hadoop.mapreduce.lib.input.FileInputFormat - Total input paths to process : 1
2022-08-30 00:28:02,127 [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> █

Activate Windows
Go to Settings to activate Windows
```