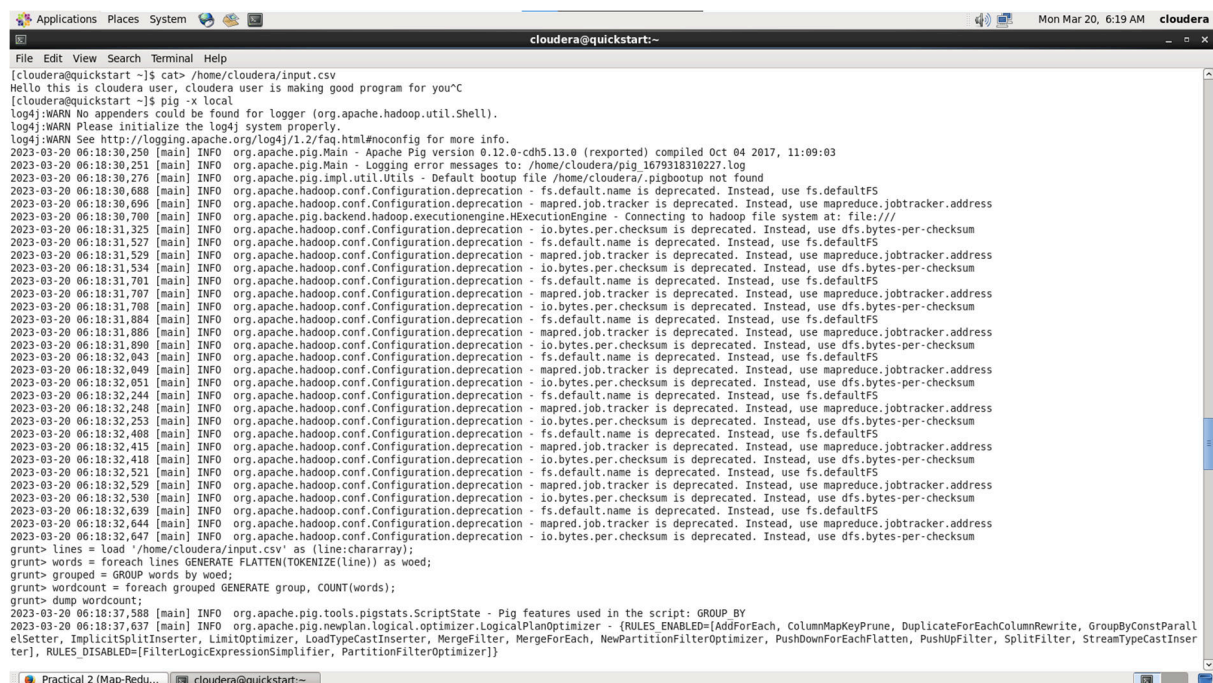


# Practical 3: PIG Script for Solving Counting Problems

Commands:

```
cat> /home/cloudera/input.csv
cat /home/cloudera/input.csv
pig -x local
lines = load '/home/cloudera/input.csv' as (line:chararray);
words = foreach lines GENERATE FLATTEN(TOKENIZE(line)) as woed;
grouped = GROUP words by woed;
wordcount = foreach grouped GENERATE group, COUNT(words);
dump wordcount;
```

## OUTPUT



```
Applications Places System cloudera@quickstart:~
File Edit View Search Terminal Help
[cloudera@quickstart ~]$ cat> /home/cloudera/input.csv
Hello this is cloudera user, cloudera user is making good program for you^C
[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.
2023-03-20 06:18:30,250 [main] INFO org.apache.pig.Main - Apache Pig version 0.12.0-cdh5.13.0 (rexpoted) compiled Oct 04 2017, 11:09:03
2023-03-20 06:18:30,251 [main] INFO org.apache.pig.Main - Logging error messages to: /home/cloudera/pig.1679318310227.log
2023-03-20 06:18:30,276 [main] INFO org.apache.pig.impl.util.Utils - Default bootstrap file /home/cloudera/pigbootstrap not found
2023-03-20 06:18:30,688 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Instead, use fs.defaultFS
2023-03-20 06:18:30,696 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - mapred.job.tracker is deprecated. Instead, use mapreduce.jobtracker.address
2023-03-20 06:18:30,700 [main] INFO org.apache.pig.backend.hadoop.executionengine.HExecutionEngine - Connecting to hadoop file system at: file:///
2023-03-20 06:18:31,325 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - io.bytes.per.checksum is deprecated. Instead, use dfs.bytes-per-checksum
2023-03-20 06:18:31,527 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Instead, use fs.defaultFS
2023-03-20 06:18:31,529 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - mapred.job.tracker is deprecated. Instead, use mapreduce.jobtracker.address
2023-03-20 06:18:31,534 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - io.bytes.per.checksum is deprecated. Instead, use dfs.bytes-per-checksum
2023-03-20 06:18:31,701 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Instead, use fs.defaultFS
2023-03-20 06:18:31,707 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - mapred.job.tracker is deprecated. Instead, use mapreduce.jobtracker.address
2023-03-20 06:18:31,708 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - io.bytes.per.checksum is deprecated. Instead, use dfs.bytes-per-checksum
2023-03-20 06:18:31,804 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Instead, use fs.defaultFS
2023-03-20 06:18:31,806 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - mapred.job.tracker is deprecated. Instead, use mapreduce.jobtracker.address
2023-03-20 06:18:31,890 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - io.bytes.per.checksum is deprecated. Instead, use dfs.bytes-per-checksum
2023-03-20 06:18:32,043 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Instead, use fs.defaultFS
2023-03-20 06:18:32,049 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - mapred.job.tracker is deprecated. Instead, use mapreduce.jobtracker.address
2023-03-20 06:18:32,051 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - io.bytes.per.checksum is deprecated. Instead, use dfs.bytes-per-checksum
2023-03-20 06:18:32,244 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Instead, use fs.defaultFS
2023-03-20 06:18:32,248 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - mapred.job.tracker is deprecated. Instead, use mapreduce.jobtracker.address
2023-03-20 06:18:32,253 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - io.bytes.per.checksum is deprecated. Instead, use dfs.bytes-per-checksum
2023-03-20 06:18:32,408 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Instead, use fs.defaultFS
2023-03-20 06:18:32,415 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - mapred.job.tracker is deprecated. Instead, use mapreduce.jobtracker.address
2023-03-20 06:18:32,521 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - io.bytes.per.checksum is deprecated. Instead, use dfs.bytes-per-checksum
2023-03-20 06:18:32,529 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Instead, use fs.defaultFS
2023-03-20 06:18:32,530 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - mapred.job.tracker is deprecated. Instead, use mapreduce.jobtracker.address
2023-03-20 06:18:32,639 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - io.bytes.per.checksum is deprecated. Instead, use dfs.bytes-per-checksum
2023-03-20 06:18:32,644 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Instead, use fs.defaultFS
2023-03-20 06:18:32,647 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - mapred.job.tracker is deprecated. Instead, use mapreduce.jobtracker.address
2023-03-20 06:18:32,647 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - io.bytes.per.checksum is deprecated. Instead, use dfs.bytes-per-checksum
grunt> lines = load '/home/cloudera/input.csv' as (line:chararray);
grunt> words = foreach lines GENERATE FLATTEN(TOKENIZE(line)) as woed;
grunt> grouped = GROUP words by woed;
grunt> wordcount = foreach grouped GENERATE group, COUNT(words);
grunt> dump wordcount;
2023-03-20 06:18:37,588 [main] INFO org.apache.pig.tools.pigstats.ScriptState - Pig features used in the script: GROUP_BY
2023-03-20 06:18:37,637 [main] INFO org.apache.pig.newplan.logical.optimizer.LogicalPlanOptimizer - {RULES_ENABLED=[AddForEach, ColumnMapKeyPrune, DuplicateForEachColumnRewrite, GroupByConstParall
elSetter, ImplicitSplitInserter, LimitOptimizer, LoadTypeCastInserter, MergeFilter, MergeForEach, NewPartitionFilterOptimizer, PushDownForEachFlatten, PushUpFilter, SplitFilter, StreamTypeCastInse
rter], RULES_DISABLED=[FilterLogicExpressionSimplifier, PartitionFilterOptimizer]}
```



```
Applications Places System cloudera@quickstart:~
File Edit View Search Terminal Help
cher for fetching Map Completion Events
2023-03-20 06:18:40,102 [localfetcher#1] INFO org.apache.hadoop.mapreduce.task.reduce.LocalFetcher - localfetcher#1 about to shuffle output of map attempt_local37669034_0001_m_000000_0 decomp: 2 l
en: 6 to MEMORY
2023-03-20 06:18:40,109 [localfetcher#1] INFO org.apache.hadoop.mapreduce.task.reduce.InMemoryMapOutput - Read 2 bytes from map-output for attempt_local37669034_0001_m_000000_0
2023-03-20 06:18:40,112 [localfetcher#1] INFO org.apache.hadoop.mapreduce.task.reduce.MergeManagerImpl - closeInMemoryFile -> map-output of size: 2, InMemoryMapOutputs.size()-> 1, commitMemory ->
0, useMemory -> 2
2023-03-20 06:18:40,113 [Readahead Thread #0] WARN org.apache.hadoop.io.ReadaheadPool - Failed readahead on ifile
EBADF: Bad file descriptor
at org.apache.hadoop.io.nativeio.NativeIO$POSIX.posixFadvise(Native Method)
at org.apache.hadoop.io.nativeio.NativeIO$POSIX.posixFadviseIfPossible(NativeIO.java:267)
at org.apache.hadoop.io.nativeio.NativeIO$POSIX$CacheManipulator.posixFadviseIfPossible(NativeIO.java:146)
at org.apache.hadoop.io.ReadaheadPool$ReadaheadRequestImpl.run(ReadaheadPool.java:206)
at java.util.concurrent.ThreadPoolExecutor.runWorker(ThreadPoolExecutor.java:1145)
at java.util.concurrent.ThreadPoolExecutor$Worker.run(ThreadPoolExecutor.java:615)
at java.lang.Thread.run(Thread.java:745)
2023-03-20 06:18:40,116 [EventFetcher for fetching Map Completion Events] INFO org.apache.hadoop.mapreduce.task.reduce.EventFetcher - EventFetcher is interrupted.. Returning
2023-03-20 06:18:40,117 [pool-3-thread-1] INFO org.apache.hadoop.mapred.LocalJobRunner - 1 / 1 copied.
2023-03-20 06:18:40,117 [pool-3-thread-1] INFO org.apache.hadoop.mapreduce.task.reduce.MergeManagerImpl - finalMerge called with 1 in-memory map-outputs and 0 on-disk map-outputs
2023-03-20 06:18:40,125 [pool-3-thread-1] INFO org.apache.hadoop.mapred.Merger - Merging 1 sorted segments
2023-03-20 06:18:40,125 [pool-3-thread-1] INFO org.apache.hadoop.mapred.Merger - Down to the last merge-pass, with 0 segments left of total size: 0 bytes
2023-03-20 06:18:40,126 [pool-3-thread-1] INFO org.apache.hadoop.mapreduce.task.reduce.MergeManagerImpl - Merged 1 segments, 2 bytes to disk to satisfy reduce memory limit
2023-03-20 06:18:40,126 [pool-3-thread-1] INFO org.apache.hadoop.mapreduce.task.reduce.MergeManagerImpl - Merging 1 files, 6 bytes from disk
2023-03-20 06:18:40,127 [pool-3-thread-1] INFO org.apache.hadoop.mapreduce.task.reduce.MergeManagerImpl - Merging 0 segments, 0 bytes from memory into reduce
2023-03-20 06:18:40,127 [pool-3-thread-1] INFO org.apache.hadoop.mapred.Merger - Merging 1 sorted segments
2023-03-20 06:18:40,127 [pool-3-thread-1] INFO org.apache.hadoop.mapred.Merger - Down to the last merge-pass, with 0 segments left of total size: 0 bytes
2023-03-20 06:18:40,128 [pool-3-thread-1] INFO org.apache.hadoop.mapred.LocalJobRunner - 1 / 1 copied.
2023-03-20 06:18:40,141 [pool-3-thread-1] INFO org.apache.hadoop.mapreduce.lib.output.FileOutputCommitter - File Output Committer Algorithm version is 1
2023-03-20 06:18:40,141 [pool-3-thread-1] INFO org.apache.hadoop.mapreduce.lib.output.FileOutputCommitter - FileOutputCommitter skip cleanup_temporary folders under output directory:false, ignore
cleanup failures: false
2023-03-20 06:18:40,144 [pool-3-thread-1] INFO org.apache.hadoop.conf.Configuration.deprecation - mapred.skip.on is deprecated. Instead, use mapreduce.job.skiprecords
2023-03-20 06:18:40,171 [pool-3-thread-1] WARN org.apache.pig.data.SchemaTupleBackend - SchemaTupleBackend has already been initialized
2023-03-20 06:18:40,181 [pool-3-thread-1] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.PigMapReduceReduce - Aliases being processed per job phase (AliasName[line,offset]): M:
lines[1,8], words[1,1], wordcount[4,12], grouped[3,10] R: wordcount[4,12]
2023-03-20 06:18:40,182 [pool-3-thread-1] INFO org.apache.hadoop.mapred.Task - Task:attempt_local37669034_0001_r_000000_0 is done. And is in the process of committing
2023-03-20 06:18:40,184 [pool-3-thread-1] INFO org.apache.hadoop.mapred.LocalJobRunner - 1 / 1 copied.
2023-03-20 06:18:40,184 [pool-3-thread-1] INFO org.apache.hadoop.mapred.Task - Task:attempt_local37669034_0001_r_000000_0 is allowed to commit now
2023-03-20 06:18:40,190 [pool-3-thread-1] INFO org.apache.hadoop.mapreduce.lib.output.FileOutputCommitter - Saved output of task 'attempt_local37669034_0001_r_000000_0' to file:/tmp/temp-893915745
/tmp-1577467803/ temporary/0/task_local37669034_0001_r_000000
2023-03-20 06:18:40,191 [pool-3-thread-1] INFO org.apache.hadoop.mapred.LocalJobRunner - reduce > reduce
2023-03-20 06:18:40,191 [pool-3-thread-1] INFO org.apache.hadoop.mapred.Task - Task:attempt_local37669034_0001_r_000000_0 done.
2023-03-20 06:18:40,191 [pool-3-thread-1] INFO org.apache.hadoop.mapred.LocalJobRunner - Finishing task: attempt_local37669034_0001_r_000000_0
2023-03-20 06:18:40,191 [Thread-8] INFO org.apache.hadoop.mapred.LocalJobRunner - reduce task executor complete.
2023-03-20 06:18:45,421 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - mapred.map.tasks is deprecated. Instead, use mapreduce.job.maps
2023-03-20 06:18:51,425 [main] WARN org.apache.pig.tools.pigstats.PigStatsUtil - Failed to get RunningJob for job job_local37669034_0001
2023-03-20 06:18:51,435 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - 100% complete
2023-03-20 06:18:51,436 [main] INFO org.apache.pig.tools.pigstats.SimplePigStats - Detected Local mode. Stats reported below may be incomplete

Practical 2 (Map-Redu... cloudera@quickstart:~
```

```
2023-03-20 06:18:51,446 [main] INFO org.apache.pig.tools.pigstats.SimplePigStats - Script Statistics:

HadoopVersion PigVersion UserId StartedAt FinishedAt Features
2.6.0-cdh5.13.0 0.12.0-cdh5.13.0 cloudera 2023-03-20 06:18:37 2023-03-20 06:18:51 GROUP_BY

Success!

Job Stats (time in seconds):
JobId Alias Feature Outputs
job_local37669034_0001 grouped,lines,wordcount,words GROUP_BY,COMBINER file:/tmp/temp-893915745/tmp-1577467803,

Input(s):
Successfully read records from: "/home/cloudera/input.csv"

Output(s):
Successfully stored records in: "file:/tmp/temp-893915745/tmp-1577467803"

Job DAG:
job_local37669034_0001

2023-03-20 06:18:57,463 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - Success!
2023-03-20 06:18:57,469 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Instead, use fs.defaultFS
2023-03-20 06:18:57,469 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - mapred.job.tracker is deprecated. Instead, use mapreduce.jobtracker.address
2023-03-20 06:18:57,478 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - io.bytes.per.checksum is deprecated. Instead, use dfs.bytes-per-checksum
2023-03-20 06:18:57,478 [main] WARN org.apache.pig.data.SchemaTupleBackend - SchemaTupleBackend has already been initialized
2023-03-20 06:18:57,494 [main] INFO org.apache.hadoop.mapreduce.lib.input.FileInputFormat - Total input paths to process : 1
grunt>
```

```
Practical 2 (Map-Redu... cloudera@quickstart:~

(.,1)
(is,2)
(for,1)
(you,1)
(good,1)
(this,1)
(cloudera,2)
(Hello,1)
(user,2)
(making,1)
(program,1)
grunt>
```