Program to implement wordcount using PIG

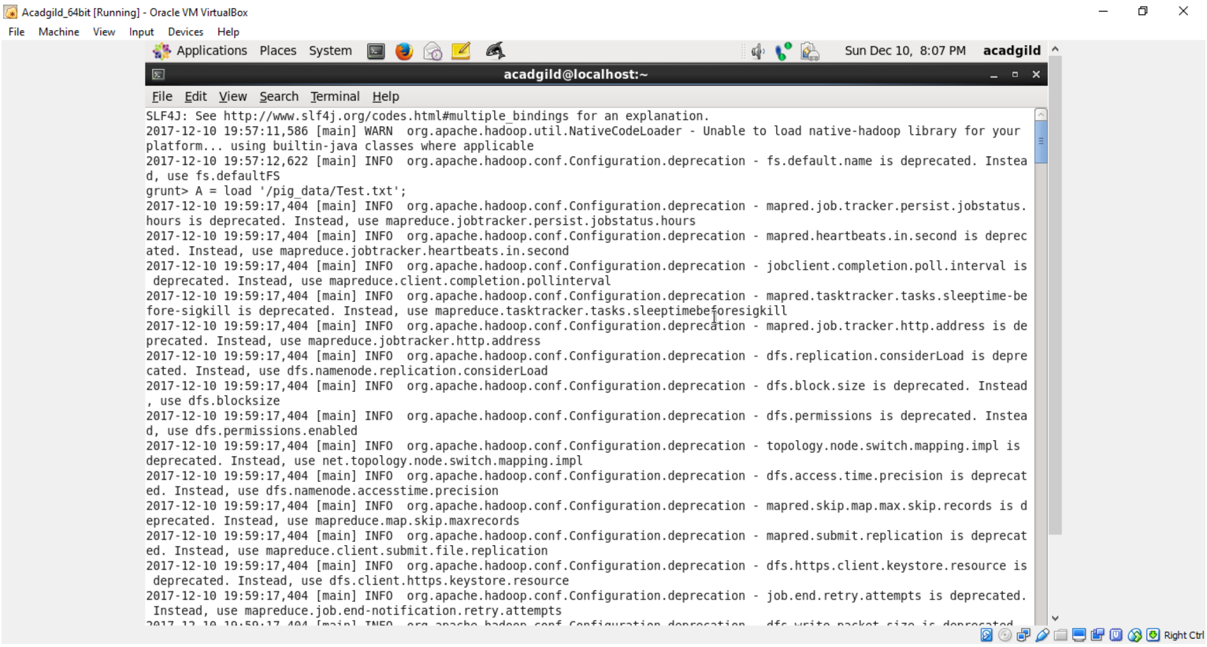
Commands :

```
A = load '/pig_data/test.txt';

B = foreach A generate flatten(TOKENIZE((chararray)$0)) as word;

C = group B by word;

D = foreach C generate group, COUNT(B);

dump D;
```

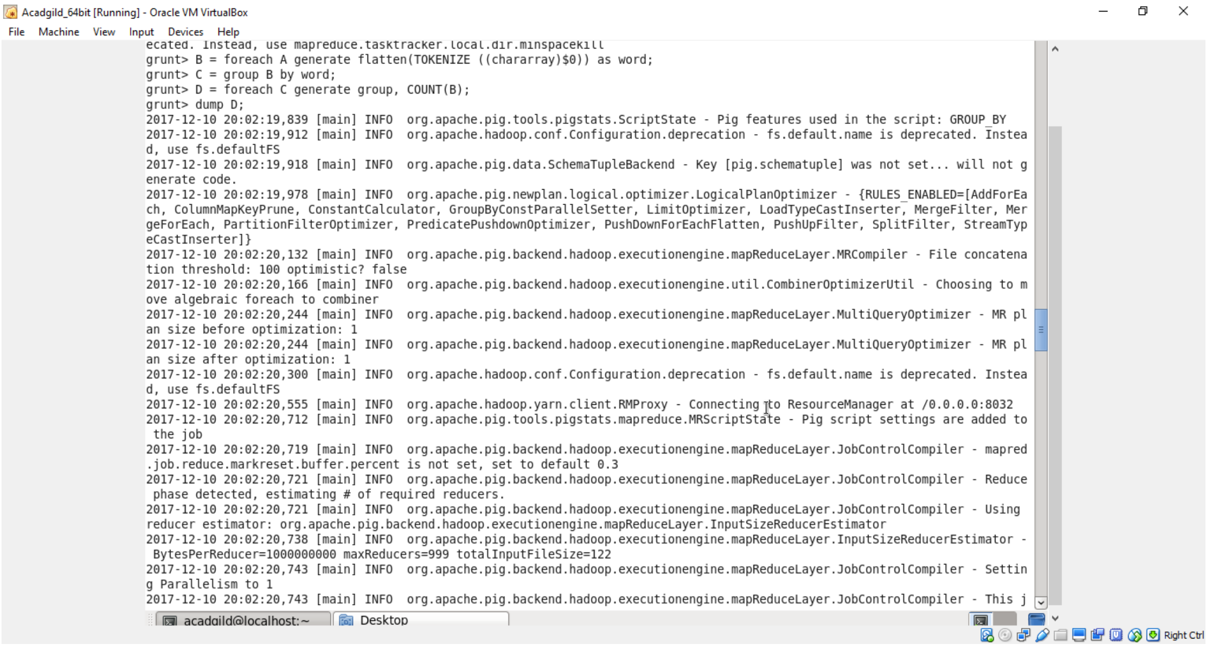
Content of input file :

```
Sample file to demonstrate word count program in pig
Sample file to demonstrate word count program in pig
Just a sample
```

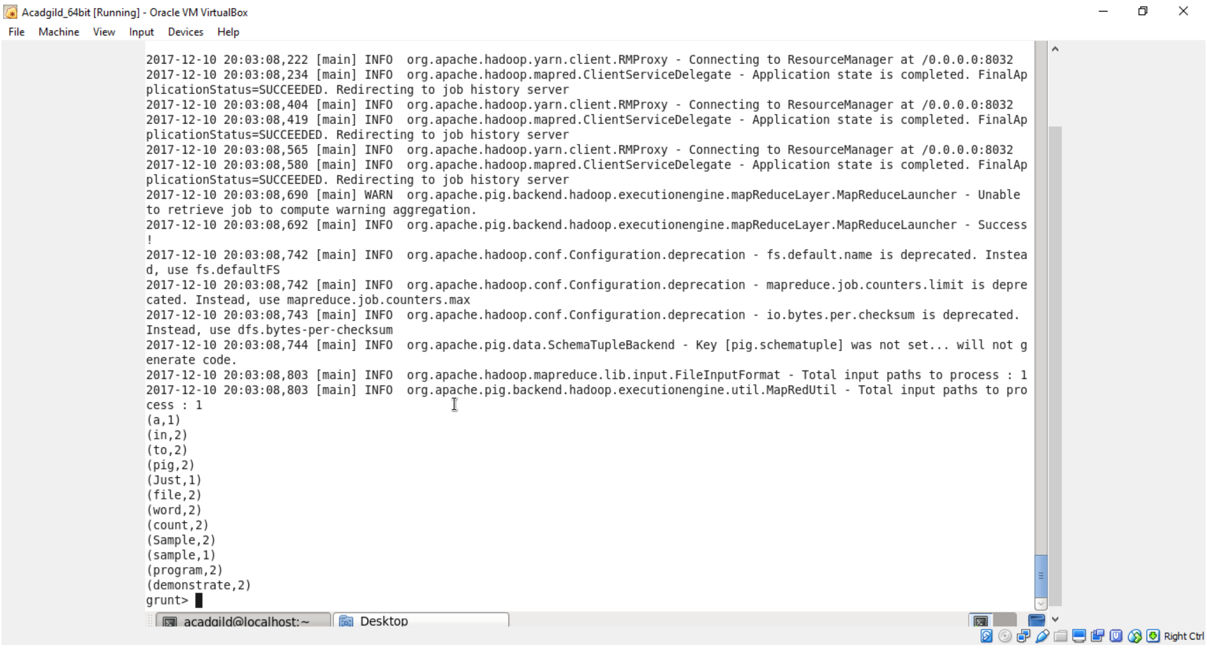
Command execution and output:


Acadgild_64bit [Running] - Oracle VM VirtualBox
File Machine View Input Devices Help
Applications Places System
Sun Dec 10, 8:07 PM acadgild
acadgild@localhost:~
File Edit View Search Terminal Help
Right Ctrl

```
SLF4J: See http://www.slf4j.org/codes.html#multiple_bindings for an explanation.
2017-12-10 19:57:11,586 [main] WARN  org.apache.hadoop.util.NativeCodeLoader - Unable to load native-hadoop library for your
platform... using builtin-java classes where applicable
2017-12-10 19:57:12,622 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Instea
d, use fs.defaultFS
grunt> A = load '/pig_data/Test.txt';
2017-12-10 19:59:17,404 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - mapred.job.tracker.persist.jobstatus.
hours is deprecated. Instead, use mapreduce.jobtracker.persist.jobstatus.hours
2017-12-10 19:59:17,404 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - mapred.heartbeats.in.second is deprec
ated. Instead, use mapreduce.jobtracker.heartbeats.in.second
2017-12-10 19:59:17,404 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - jobclient.completion.poll.interval is
 deprecated. Instead, use mapreduce.client.completion.pollinterval
2017-12-10 19:59:17,404 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - mapred.tasktracker.tasks.sleeptime-be
fore-sigkill is deprecated. Instead, use mapreduce.tasktracker.tasks.sleeptimebeforesigkill
2017-12-10 19:59:17,404 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - mapred.job.tracker.http.address is de
precated. Instead, use mapreduce.jobtracker.http.address
2017-12-10 19:59:17,404 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - dfs.replication.considerLoad is depre
cated. Instead, use dfs.namenode.replication.considerLoad
2017-12-10 19:59:17,404 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - dfs.block.size is deprecated. Instead
, use dfs.blocksize
2017-12-10 19:59:17,404 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - dfs.permissions is deprecated. Instea
d, use dfs.permissions.enabled
2017-12-10 19:59:17,404 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - topology.node.switch.mapping.impl is
deprecated. Instead, use net.topology.node.switch.mapping.impl
2017-12-10 19:59:17,404 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - dfs.access.time.precision is deprecat
ed. Instead, use dfs.namenode.accesstime.precision
2017-12-10 19:59:17,404 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - mapred.skip.map.max.skip.records is d
eprecated. Instead, use mapreduce.map.skip.maxrecords
2017-12-10 19:59:17,404 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - mapred.submit.replication is deprecat
ed. Instead, use mapreduce.client.submit.file.replication
2017-12-10 19:59:17,404 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - dfs.https.client.keystore.resource is
 deprecated. Instead, use dfs.client.https.keystore.resource
2017-12-10 19:59:17,404 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - job.end.retry.attempts is deprecated.
 Instead, use mapreduce.job.end-notification.retry.attempts
2017-12-10 19:59:17,404 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - dfs.write.packet.size is deprecated.
```


Acadgild_64bit [Running] - Oracle VM VirtualBox
File Machine View Input Devices Help
acadgild@localhost:~
Desktop
Right Ctrl

```
ecated. Instead, use mapreduce.tasktracker.local.dir.minspacekill
grunt> B = foreach A generate flatten(TOKENIZE ((chararray)$0)) as word;
grunt> C = group B by word;
grunt> D = foreach C generate group, COUNT(B);
grunt> dump D;
2017-12-10 20:02:19,839 [main] INFO  org.apache.pig.tools.pigstats.ScriptState - Pig features used in the script: GROUP_BY
2017-12-10 20:02:19,912 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Instea
d, use fs.defaultFS
2017-12-10 20:02:19,918 [main] INFO  org.apache.pig.data.SchemaTupleBackend - Key [pig.schematuple] was not set... will not g
enerate code.
2017-12-10 20:02:19,978 [main] INFO  org.apache.pig.newplan.logical.optimizer.LogicalPlanOptimizer - {RULES_ENABLED=[AddForEa
ch, ColumnMapKeyPrune, ConstantCalculator, GroupByConstParallelSetter, LimitOptimizer, LoadTypeCastInserter, MergeFilter, Mer
geForEach, PartitionFilterOptimizer, PredicatePushdownOptimizer, PushDownForEachFlatten, PushUpFilter, SplitFilter, StreamTyp
eCastInserter]}
2017-12-10 20:02:20,132 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MRCompiler - File concatena
tion threshold: 100 optimistic? false
2017-12-10 20:02:20,166 [main] INFO  org.apache.pig.backend.hadoop.executionengine.util.CombinerOptimizerUtil - Choosing to m
ove algebraic foreach to combiner
2017-12-10 20:02:20,244 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MultiQueryOptimizer - MR pl
an size before optimization: 1
2017-12-10 20:02:20,244 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MultiQueryOptimizer - MR pl
an size after optimization: 1
2017-12-10 20:02:20,300 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Instea
d, use fs.defaultFS
2017-12-10 20:02:20,555 [main] INFO  org.apache.hadoop.yarn.client.RMProxy - Connecting to ResourceManager at /0.0.0.0:8032
2017-12-10 20:02:20,712 [main] INFO  org.apache.pig.tools.pigstats.mapreduce.MRScriptState - Pig script settings are added to
 the job
2017-12-10 20:02:20,719 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.JobControlCompiler - mapred
.job.reduce.markreset.buffer.percent is not set, set to default 0.3
2017-12-10 20:02:20,721 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.JobControlCompiler - Reduce
 phase detected, estimating # of required reducers.
2017-12-10 20:02:20,721 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.JobControlCompiler - Using
reducer estimator: org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.InputSizeReducerEstimator
2017-12-10 20:02:20,738 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.InputSizeReducerEstimator -
 BytesPerReducer=1000000000 maxReducers=999 totalInputFileSize=122
2017-12-10 20:02:20,743 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.JobControlCompiler - Settin
g Parallelism to 1
2017-12-10 20:02:20,743 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.JobControlCompiler - This j
```


Acadgild_64bit [Running] - Oracle VM VirtualBox
File Machine View Input Devices Help
acadgild@localhost:~
Desktop
Right Ctrl

```
2017-12-10 20:03:08,222 [main] INFO  org.apache.hadoop.yarn.client.RMProxy - Connecting to ResourceManager at /0.0.0.0:8032
2017-12-10 20:03:08,234 [main] INFO  org.apache.hadoop.mapred.ClientServiceDelegate - Application state is completed. FinalAp
plicationStatus=SUCCEEDED. Redirecting to job history server
2017-12-10 20:03:08,404 [main] INFO  org.apache.hadoop.yarn.client.RMProxy - Connecting to ResourceManager at /0.0.0.0:8032
2017-12-10 20:03:08,419 [main] INFO  org.apache.hadoop.mapred.ClientServiceDelegate - Application state is completed. FinalAp
plicationStatus=SUCCEEDED. Redirecting to job history server
2017-12-10 20:03:08,565 [main] INFO  org.apache.hadoop.yarn.client.RMProxy - Connecting to ResourceManager at /0.0.0.0:8032
2017-12-10 20:03:08,580 [main] INFO  org.apache.hadoop.mapred.ClientServiceDelegate - Application state is completed. FinalAp
plicationStatus=SUCCEEDED. Redirecting to job history server
2017-12-10 20:03:08,690 [main] WARN  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - Unable
to retrieve job to compute warning aggregation.
2017-12-10 20:03:08,692 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - Success
!
2017-12-10 20:03:08,742 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Instea
d, use fs.defaultFS
2017-12-10 20:03:08,742 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - mapreduce.job.counters.limit is depre
cated. Instead, use mapreduce.job.counters.max
2017-12-10 20:03:08,743 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - io.bytes.per.checksum is deprecated.
Instead, use dfs.bytes-per-checksum
2017-12-10 20:03:08,744 [main] INFO  org.apache.pig.data.SchemaTupleBackend - Key [pig.schematuple] was not set... will not g
enerate code.
2017-12-10 20:03:08,803 [main] INFO  org.apache.hadoop.mapreduce.lib.input.FileInputFormat - Total input paths to process : 1
2017-12-10 20:03:08,803 [main] INFO  org.apache.pig.backend.hadoop.executionengine.util.MapRedUtil - Total input paths to pro
cess : 1
(a,1)
(in,2)
(to,2)
(pig,2)
(Just,1)
(file,2)
(word,2)
(count,2)
(Sample,2)
(sample,1)
(program,2)
(demonstrate,2)
grunt> 
```