

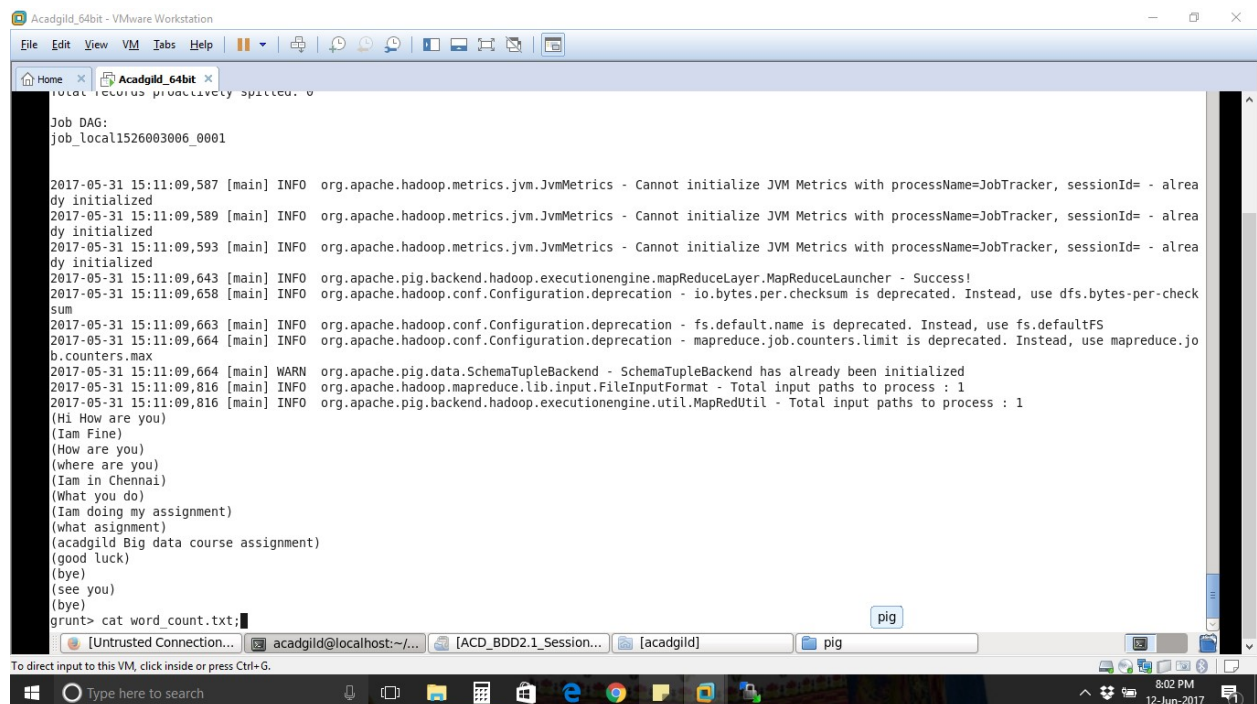
Assignment 8.1

Write a wordcount program to implement wordcount using Pig.

This is the Pig Latin Script for word count .

```
A = load 'word_count.txt';
B = foreach A generate flatten(TOKENIZE((chararray)$0)) as word;
C = group B by word;
D = foreach C generate COUNT(B), group;
store D into './wordcount';
```

1. Screen Shot of Loading the contents of word_count.txt file.

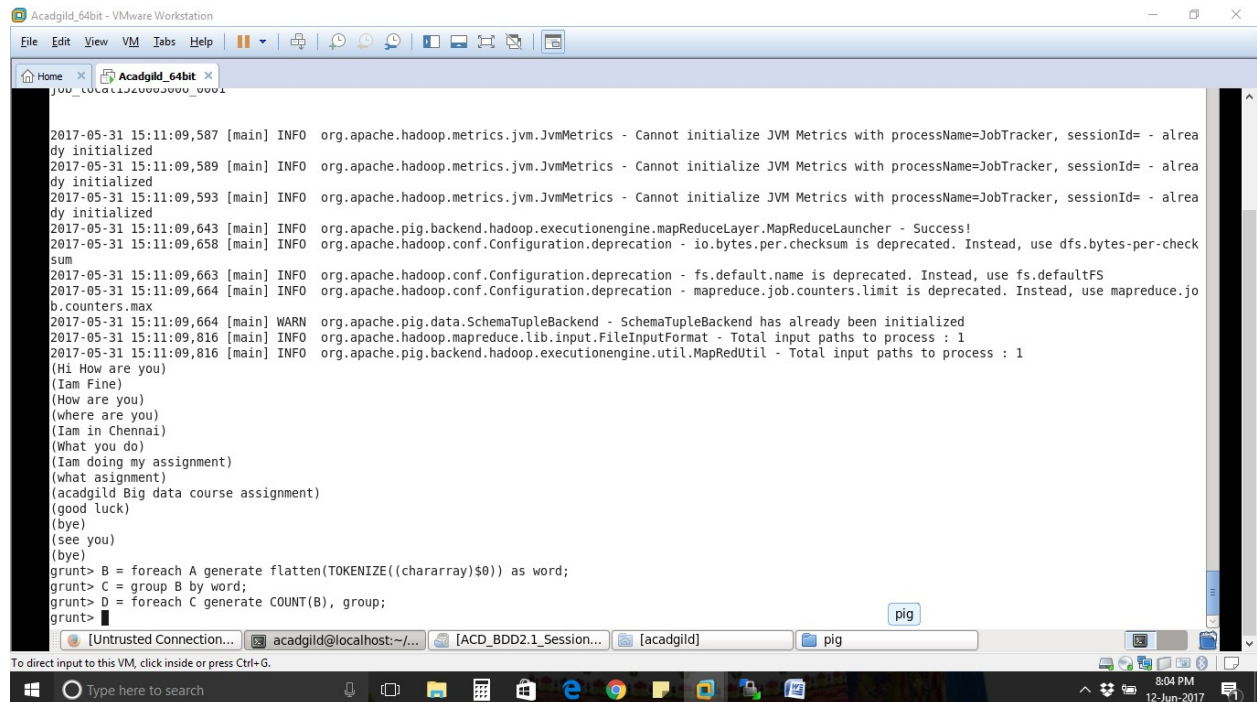


The screenshot shows a VMware Workstation window titled 'Acadgild_64bit - VMware Workstation'. Inside, a terminal window displays the output of a Pig script execution. The output includes a Job DAG, job ID 'job_local1526003006_0001', and various log messages from the Hadoop and Pig frameworks. The final output of the script is the contents of the 'word_count.txt' file, which lists various phrases and their counts. The terminal window also shows the command 'grunt> cat word_count.txt;' and the resulting output.

```
Job DAG:
job_local1526003006_0001

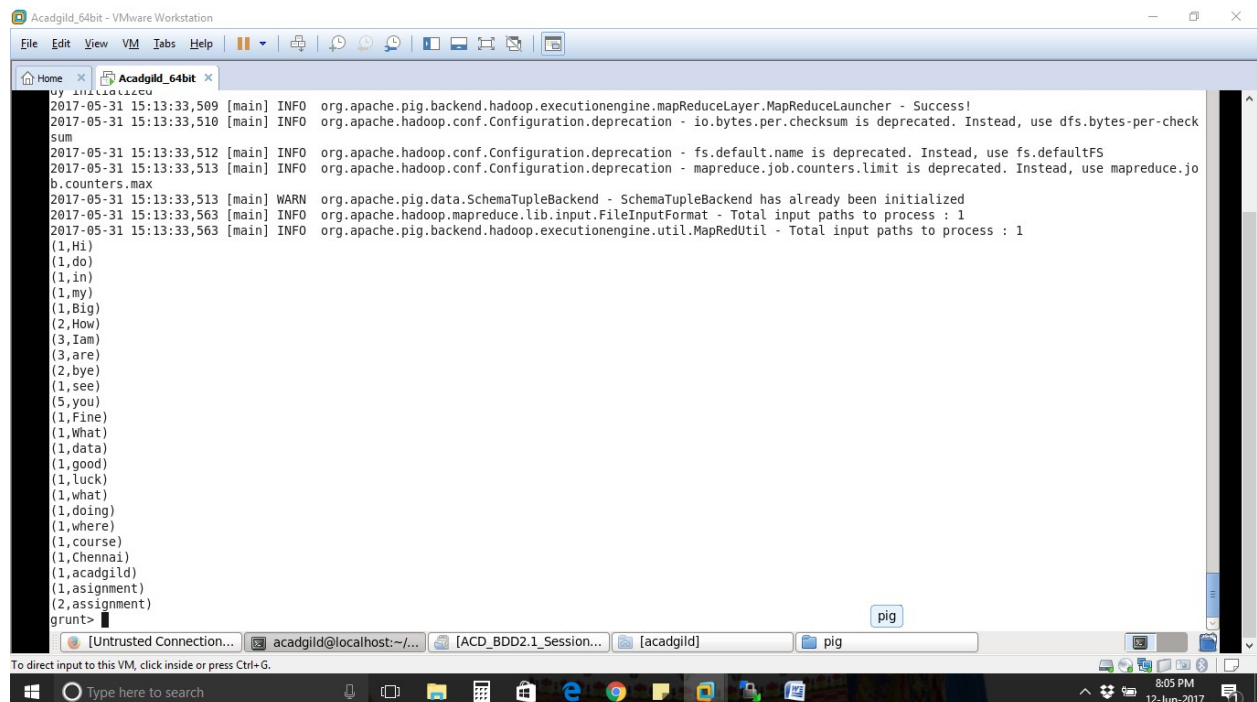
2017-05-31 15:11:09,587 [main] INFO org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM Metrics with processName=JobTracker, sessionId= - alrea
dy initialized
2017-05-31 15:11:09,589 [main] INFO org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM Metrics with processName=JobTracker, sessionId= - alrea
dy initialized
2017-05-31 15:11:09,593 [main] INFO org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM Metrics with processName=JobTracker, sessionId= - alrea
dy initialized
2017-05-31 15:11:09,643 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - Success!
2017-05-31 15:11:09,658 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - io.bytes.per.checksum is deprecated. Instead, use dfs.bytes-per-check
sum
2017-05-31 15:11:09,663 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Instead, use fs.defaultFS
2017-05-31 15:11:09,664 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - mapreduce.job.counters.limit is deprecated. Instead, use mapreduce.jo
b.counters.max
2017-05-31 15:11:09,664 [main] WARN org.apache.pig.data.SchemaTupleBackend - SchemaTupleBackend has already been initialized
2017-05-31 15:11:09,816 [main] INFO org.apache.hadoop.mapreduce.lib.input.FileInputFormat - Total input paths to process : 1
2017-05-31 15:11:09,816 [main] INFO org.apache.pig.backend.hadoop.executionengine.util.MapRedUtil - Total input paths to process : 1
(Hi How are you)
(Iam Fine)
(How are you)
(Where are you)
(Iam in Chennai)
(What you do)
(Iam doing my assignment)
(what assignment)
(acadgild Big data course assignment)
(good luck)
(bye)
(see you)
(bye)
grunt> cat word_count.txt;
```

Screenshot 2: Executing the Pig Latin script



```
2017-05-31 15:11:09,587 [main] INFO org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM Metrics with processName=JobTracker, sessionId= - alrea
dy initialized
2017-05-31 15:11:09,589 [main] INFO org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM Metrics with processName=JobTracker, sessionId= - alrea
dy initialized
2017-05-31 15:11:09,593 [main] INFO org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM Metrics with processName=JobTracker, sessionId= - alrea
dy initialized
2017-05-31 15:11:09,643 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - Success!
2017-05-31 15:11:09,658 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - io.bytes.per.checksum is deprecated. Instead, use dfs.bytes-per-check
sum
2017-05-31 15:11:09,663 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Instead, use fs.defaultFS
2017-05-31 15:11:09,664 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - mapreduce.job.counters.limit is deprecated. Instead, use mapreduce.jo
b.counters.max
2017-05-31 15:11:09,664 [main] WARN org.apache.pig.data.SchemaTupleBackend - SchemaTupleBackend has already been initialized
2017-05-31 15:11:09,816 [main] INFO org.apache.hadoop.mapreduce.lib.input.FileInputFormat - Total input paths to process : 1
2017-05-31 15:11:09,816 [main] INFO org.apache.pig.backend.hadoop.executionengine.util.MapRedUtil - Total input paths to process : 1
(Hi How are you)
(Iam Fine)
(How are you)
(where are you)
(Iam in Chennai)
(What you do)
(Iam doing my assignment)
(what assignment)
(acadgild Big data course assignment)
(good luck)
(bye)
(see you)
(bye)
grunt> B = foreach A generate flatten(TOKENIZE((chararray)$0)) as word;
grunt> C = group B by word;
grunt> D = foreach C generate COUNT(B), group;
grunt>
```

Screenshot 3 : Output of the word count program



```
2017-05-31 15:13:33,509 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - Success!
2017-05-31 15:13:33,510 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - io.bytes.per.checksum is deprecated. Instead, use dfs.bytes-per-check
sum
2017-05-31 15:13:33,512 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Instead, use fs.defaultFS
2017-05-31 15:13:33,513 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - mapreduce.job.counters.limit is deprecated. Instead, use mapreduce.jo
b.counters.max
2017-05-31 15:13:33,513 [main] WARN org.apache.pig.data.SchemaTupleBackend - SchemaTupleBackend has already been initialized
2017-05-31 15:13:33,563 [main] INFO org.apache.hadoop.mapreduce.lib.input.FileInputFormat - Total input paths to process : 1
2017-05-31 15:13:33,563 [main] INFO org.apache.pig.backend.hadoop.executionengine.util.MapRedUtil - Total input paths to process : 1
(1,Hi)
(1,do)
(1,in)
(1,my)
(1,Big)
(2,How)
(3,Iam)
(3,are)
(2,bye)
(1,see)
(5,you)
(1,Fine)
(1,What)
(1,data)
(1,good)
(1,luck)
(1,what)
(1,doing)
(1,where)
(1,course)
(1,Chennai)
(1,acadgild)
(1,assignment)
(2,assignment)
grunt>
```