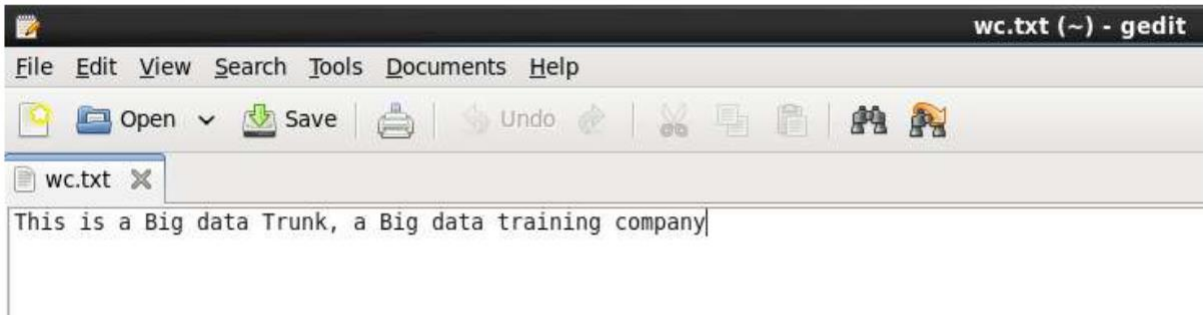


Aim: To execute word count program in pig. Using batch mode locally.

Create file with name wc.txt at '/home/cloudera/wc.txt'.

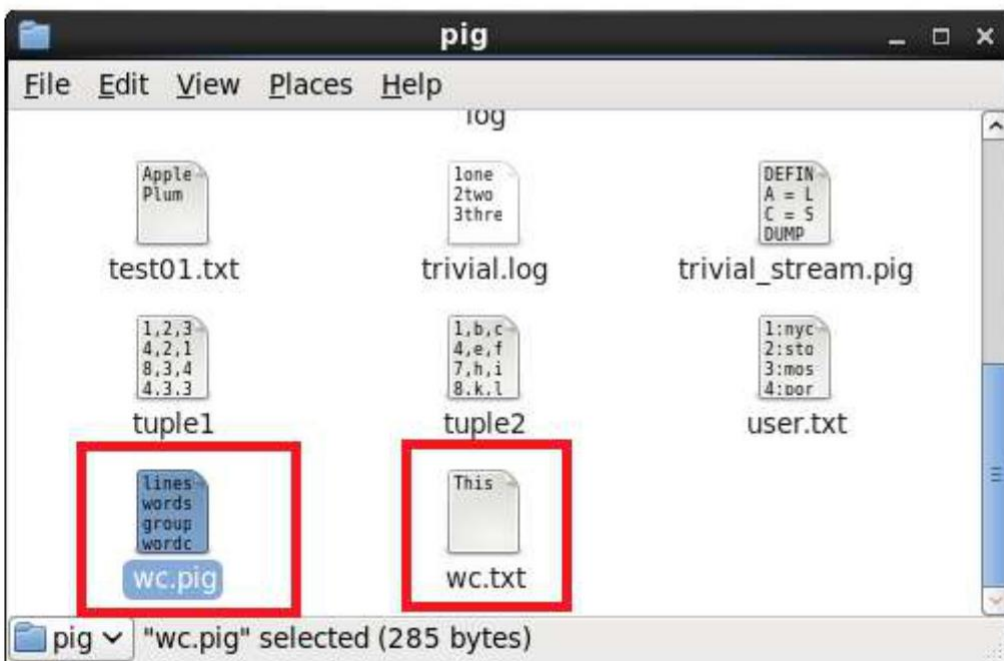
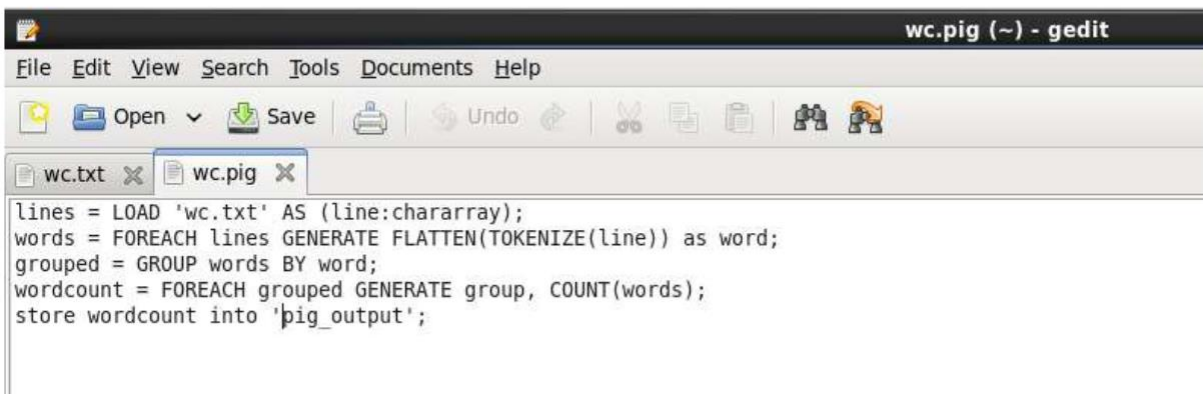
gedit wc.txt

```
cloudera@quickstart:~$ gedit wc.pig
cloudera@quickstart:~$ gedit wc.txt
cloudera@quickstart:~$
```



Create a file wc.pig at '/home/cloudera/wc.pig'.

gedit wc.pig



And put script

```
lines = LOAD 'wc.txt' AS (line:chararray);
```

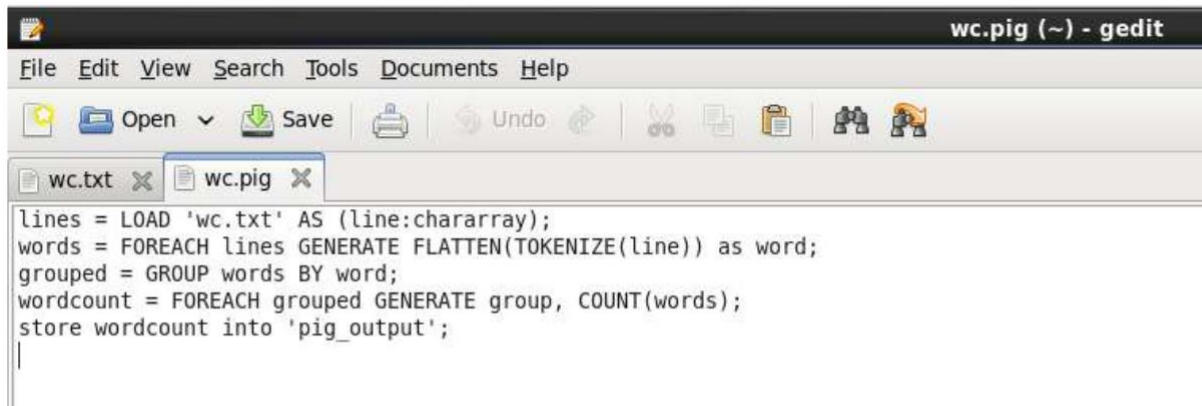
```
words = FOREACH lines GENERATE FLATTEN(TOKENIZE(line)) as word;
```

```
grouped = GROUP words BY word;
```

```
wordcount = FOREACH grouped GENERATE group, COUNT(words);
```

```
store wordcount into 'pig_output';
```

And Save it as wc.pig.

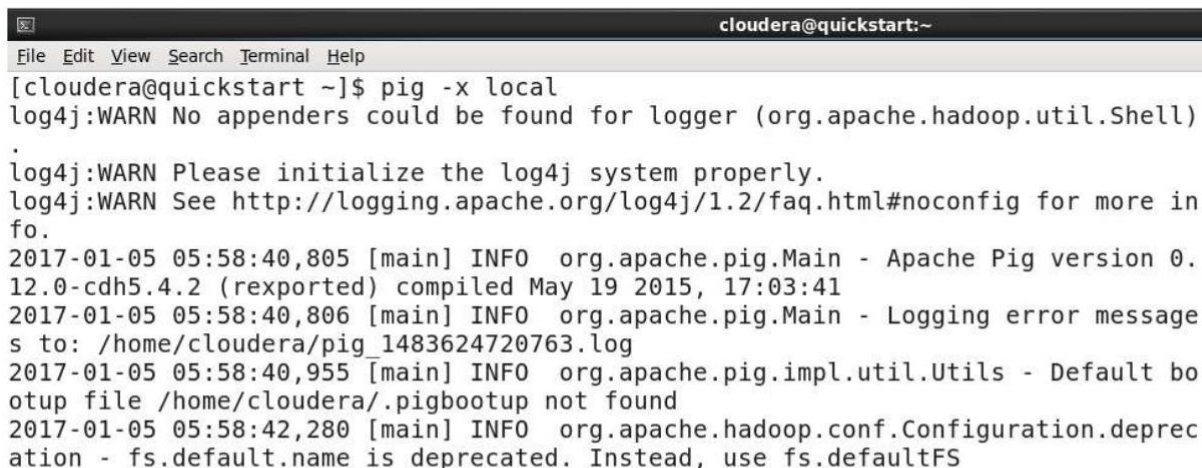


```
lines = LOAD 'wc.txt' AS (line:chararray);
words = FOREACH lines GENERATE FLATTEN(TOKENIZE(line)) as word;
grouped = GROUP words BY word;
wordcount = FOREACH grouped GENERATE group, COUNT(words);
store wordcount into 'pig_output';
```

Note: To store output in a file, the `pig_output` folder will be created by system as we know that mapreduce program will run in backend. So delete if you have already folder name `pig_output`.

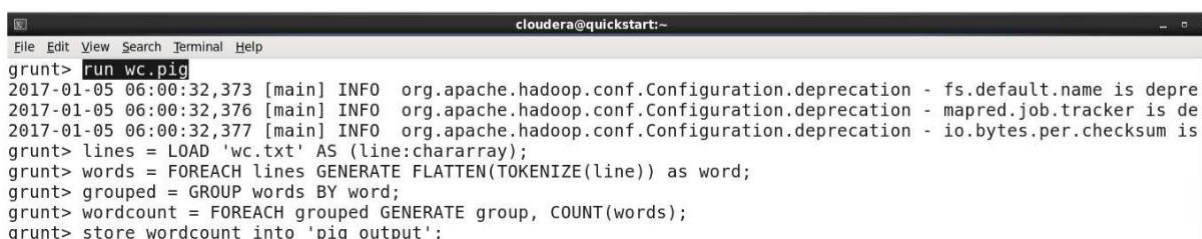
To run this script: go to the folder where you have your script and use `pig -x local` at terminal.

Pig -x local



```
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.
2017-01-05 05:58:40,805 [main] INFO org.apache.pig.Main - Apache Pig version 0.12.0-cdh5.4.2 (rexported) compiled May 19 2015, 17:03:41
2017-01-05 05:58:40,806 [main] INFO org.apache.pig.Main - Logging error messages to: /home/cloudera/pig_1483624720763.log
2017-01-05 05:58:40,955 [main] INFO org.apache.pig.impl.util.Utils - Default bootstrap file /home/cloudera/.pigbootstrap not found
2017-01-05 05:58:42,280 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Instead, use fs.defaultFS
```

run `wc.pig` file in pig terminal



```
cloudera@quickstart:~$ pig
grunt> run wc.pig
2017-01-05 06:00:32,373 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated
2017-01-05 06:00:32,376 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - mapred.job.tracker is deprecated
2017-01-05 06:00:32,377 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - io.bytes.per.checksum is deprecated
grunt> lines = LOAD 'wc.txt' AS (line:chararray);
grunt> words = FOREACH lines GENERATE FLATTEN(TOKENIZE(line)) as word;
grunt> grouped = GROUP words BY word;
grunt> wordcount = FOREACH grouped GENERATE group, COUNT(words);
grunt> store wordcount into 'pig_output';
```

021 Pig Wordcount lab

HadoopVersion	PigVersion	UserId	StartedAt	FinishedAt	Features	
2.6.0-cdh5.4.2	0.12.0-cdh5.4.2	cloudera	2017-01-05 06:00:37	2017-01-05 06:00:46	GROUP_BY	

Success!

Job Stats (time in seconds):

JobId	Alias	Feature	Outputs	
job_local2090714790_0001		grouped,lines,wordcount,words	GROUP_BY,COMBINER	file:///home/cloudera/

Input(s):

Successfully read records from: "file:///home/cloudera/wc.txt"

Output(s):

Successfully stored records in: "file:///home/cloudera/pig_output"

Job DAG:

job_local2090714790_0001

Now we can verify output at pig_output.

fs -ls pig_output

fs -cat pig_output/part*

```
grunt> fs -ls pig_output
Found 2 items
-rw-r--r-- 1 cloudera cloudera 0 2017-01-05 06:00 pig_output/_SUCCESS
-rw-r--r-- 1 cloudera cloudera 58 2017-01-05 06:00 pig_output/part-r-000000
grunt> fs -cat pig_output/part*
a 2
is 1
Big 2
This 1
data 2
Trunk 1
company 1
training 1
grunt>
```

Output here

Cloudera Live: Welco... cloudera@quickstart:~ cloudera@quickstart:~ Software Update wc.pig (~) - gedit