

## **Experiment 6 - To implement Word Count problem using Pig**

### **What is Apache Pig**

Apache Pig is a high-level data flow platform for executing MapReduce programs of Hadoop. The language used for Pig is Pig Latin.

The Pig scripts get internally converted to Map Reduce jobs and get executed on data stored in HDFS. Apart from that, Pig can also execute its job in Apache Tez or Apache Spark.

Pig can handle any type of data, i.e., structured, semi-structured or unstructured and stores the corresponding results into Hadoop Data File System. Every task which can be achieved using PIG can also be achieved using java used in MapReduce.

### **Features of Apache Pig**

The various uses of Pig technology.

#### **1) Ease of programming**

Writing complex java programs for map reduce is quite tough for non-programmers. Pig makes this process easy. In the Pig, the queries are converted to MapReduce internally.

#### **2) Optimization opportunities**

It is how tasks are encoded permits the system to optimize their execution automatically, allowing the user to focus on semantics rather than efficiency.

#### **3) Extensibility**

A user-defined function is written in which the user can write their logic to execute over the data set.

#### **4) Flexible**

It can easily handle structured as well as unstructured data.

#### **5) In-built operators**

It contains various type of operators such as sort, filter and joins.

## Advantages of Apache Pig

- Less code - The Pig consumes less line of code to perform any operation.
- Reusability - The Pig code is flexible enough to reuse again.
- Nested data types - The Pig provides a useful concept of nested data types like tuple, bag, and map.

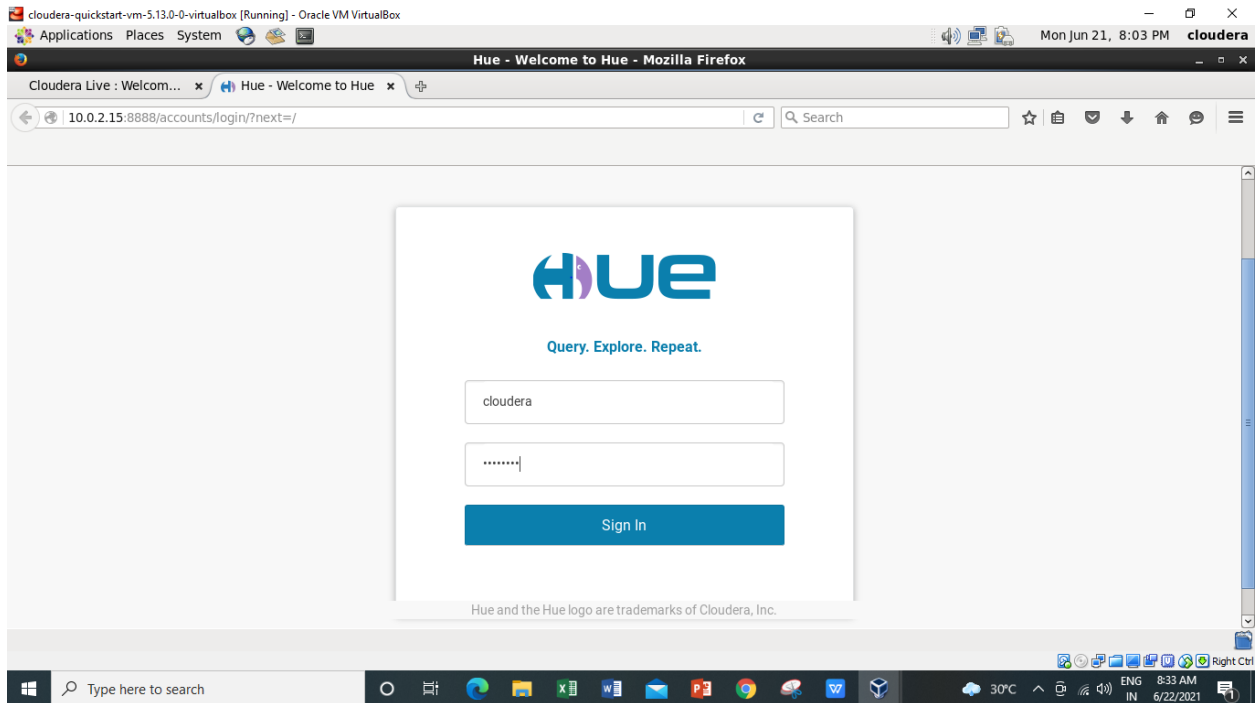
## To implement Word Count problem using Pig

### Steps:

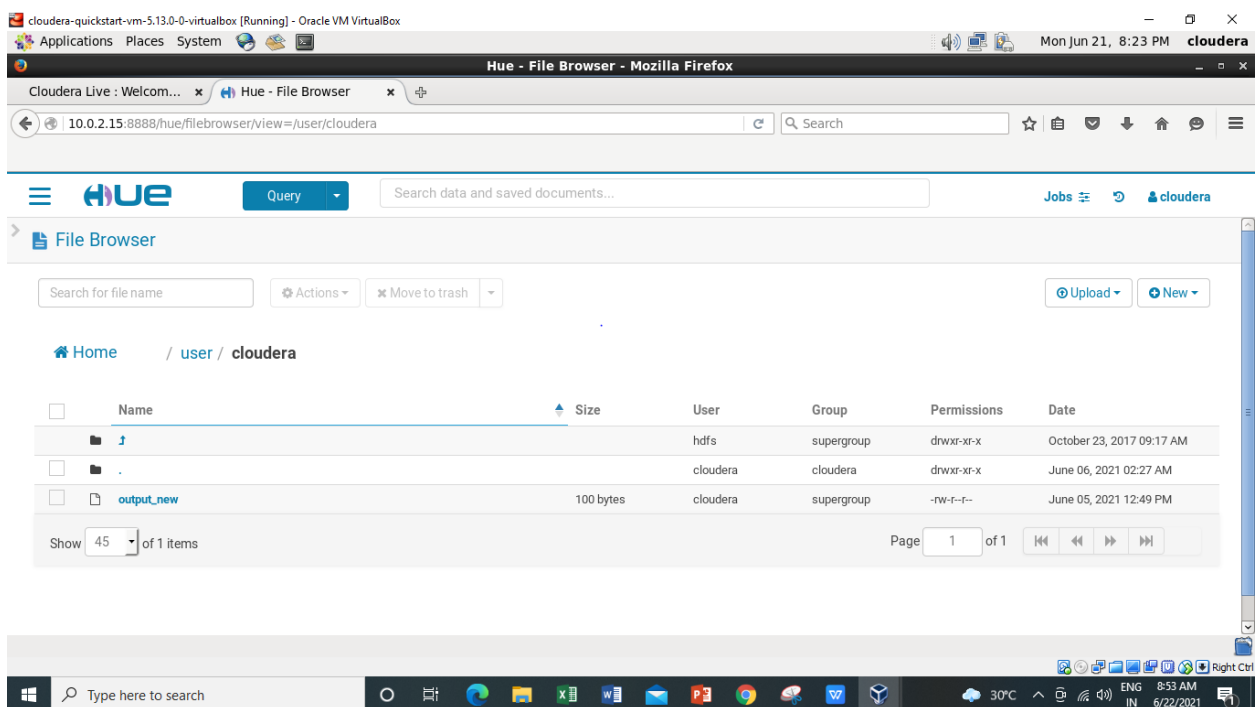
- 1) Start the cloudera.



- 2) Open the browser. And then open Hue and login.



### 3) Now open the directory /user/cloudera



### 4) Now we are creating the directory as **Training** inside /user/cloudera

cloudera-quickstart-vm-5.13.0-0-virtualbox [Running] - Oracle VM VirtualBox

Applications Places System

Hue - File Browser - Mozilla Firefox

Hue - File Browser x

Search or enter address Search

Hue Query Search data and saved documents...

Jobs cloudera

File Browser

Search for file name Actions Move to trash Upload New

File Directory

Home / user / cloudera

<input type="checkbox"/>	Name	Size	User	Group	Permissions	Date
<input type="checkbox"/>	↑		hdfs	supergroup	drwxr-xr-x	October 23, 2017 09:17 AM
<input type="checkbox"/>	.		cloudera	cloudera	drwxr-xr-x	June 06, 2021 02:27 AM
<input type="checkbox"/>	output_new	100 bytes	cloudera	supergroup	-rw-r--r--	June 05, 2021 12:49 PM

Show 45 of 1 items Page 1 of 1

Hue - File Browser - M...

Type here to search

29°C 10:51 PM 6/28/2021

cloudera-quickstart-vm-5.13.0-0-virtualbox [Running] - Oracle VM VirtualBox

Applications Places System

Hue - File Browser - Mozilla Firefox

Hue - File Browser x

Search or enter address Search

Hue Query Search data and saved documents...

Jobs cloudera

File Browser

Search for file name Actions Move to trash Upload New

File Directory

Home / user / cloudera

<input type="checkbox"/>	Name	Size	User	Group	Permissions	Date
<input type="checkbox"/>	↑		hdfs	supergroup	drwxr-xr-x	October 23, 2017 09:17 AM
<input type="checkbox"/>	.		cloudera	cloudera	drwxr-xr-x	June 06, 2021 02:27 AM
<input type="checkbox"/>	output_new	100 bytes	cloudera	supergroup	-rw-r--r--	June 05, 2021 12:49 PM

Show 45 of 1 items Page 1 of 1

Hue - File Browser - M...

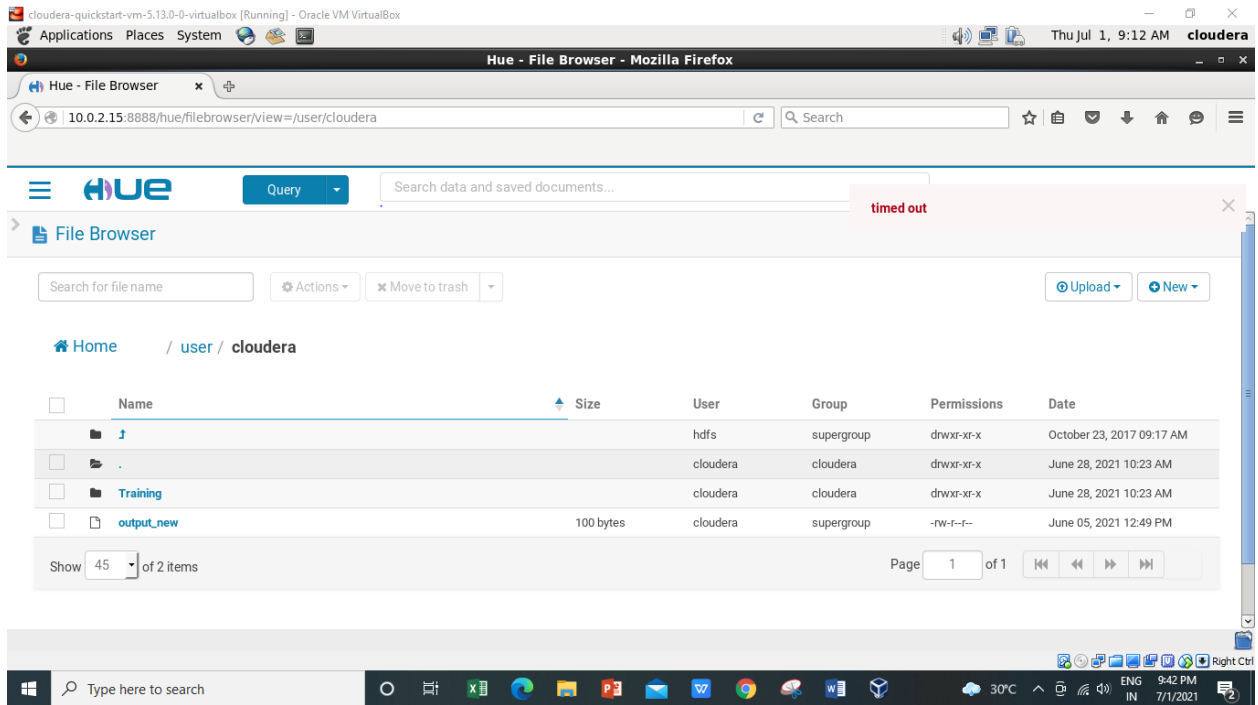
Type here to search

29°C 10:52 PM 6/28/2021

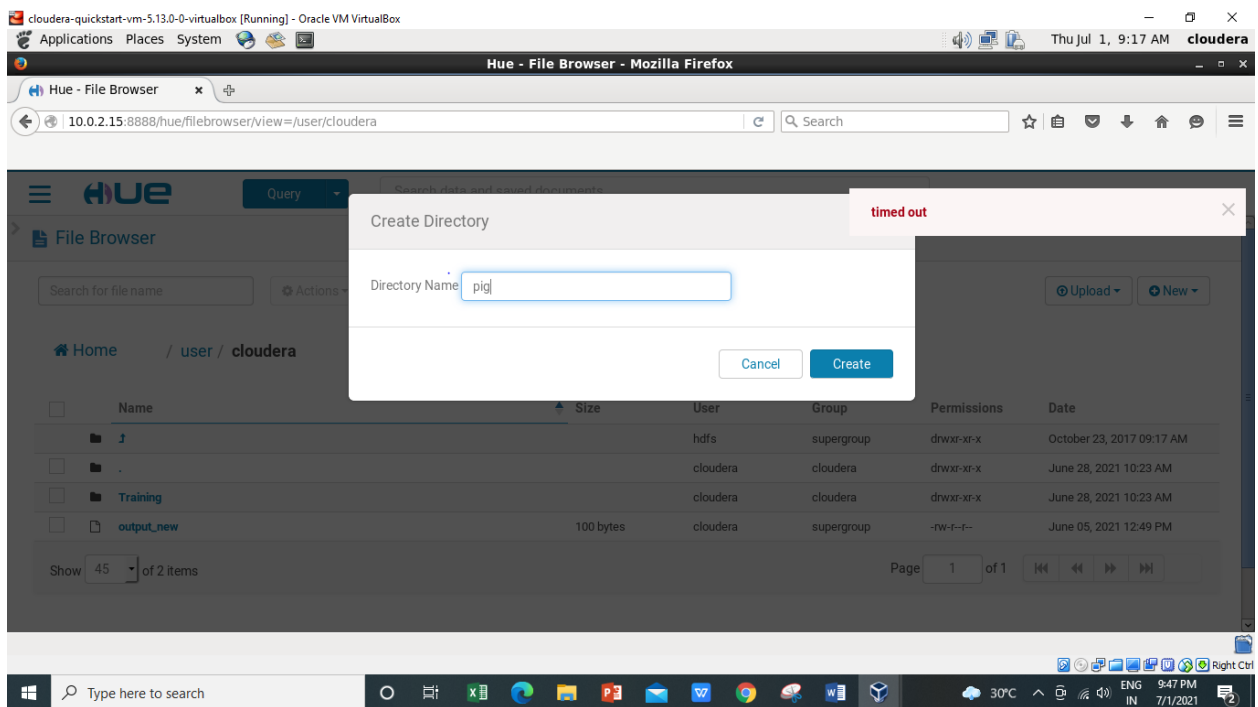
Create Directory

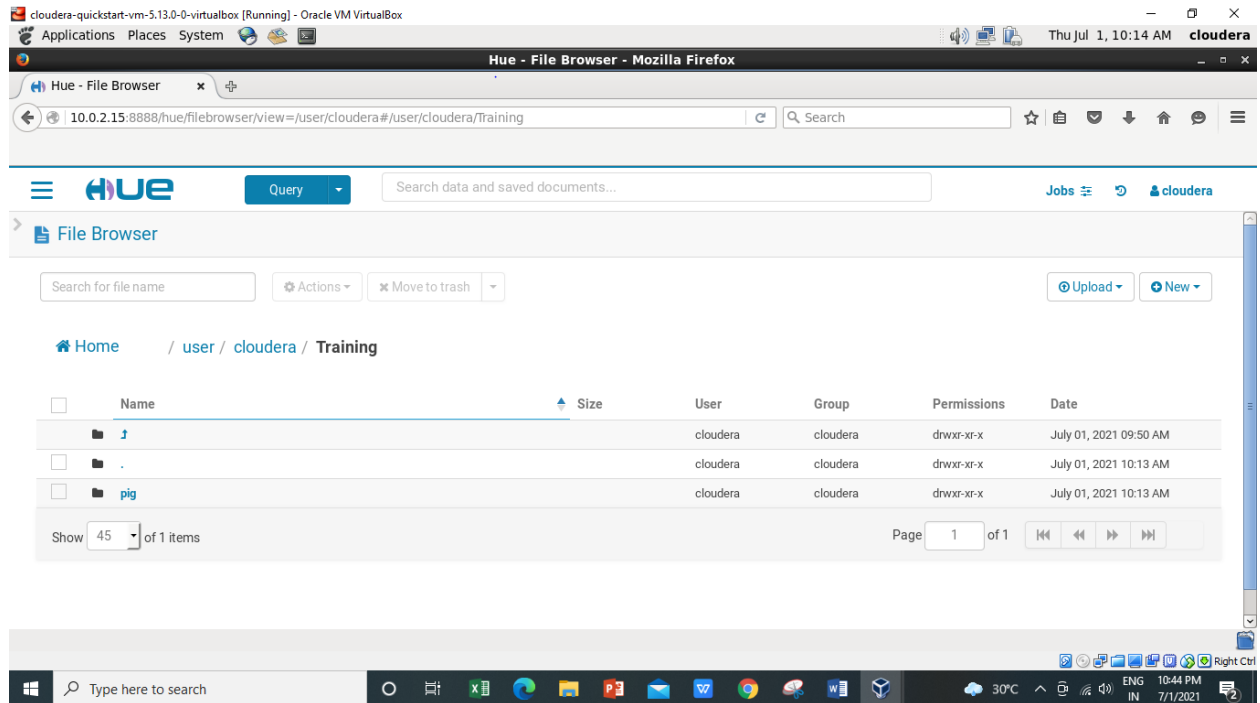
Directory Name Training

Cancel Create

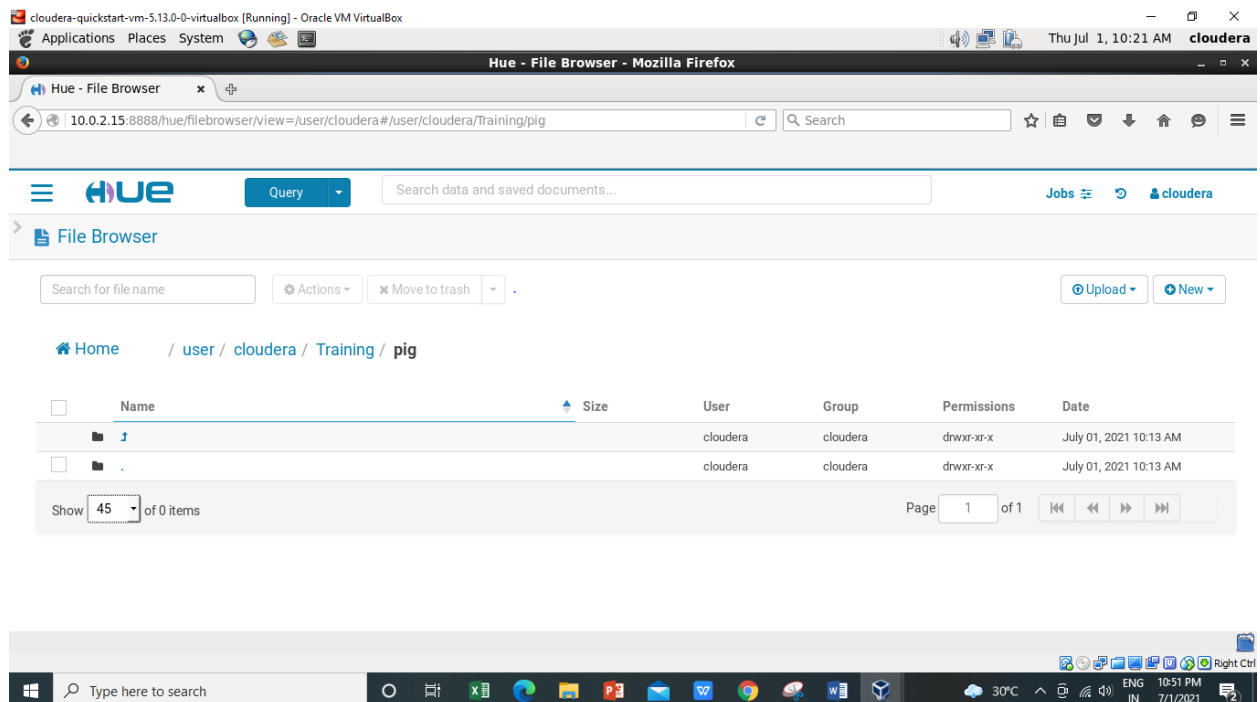


5) After creating Training directory now creating the **pig** directory inside Training.

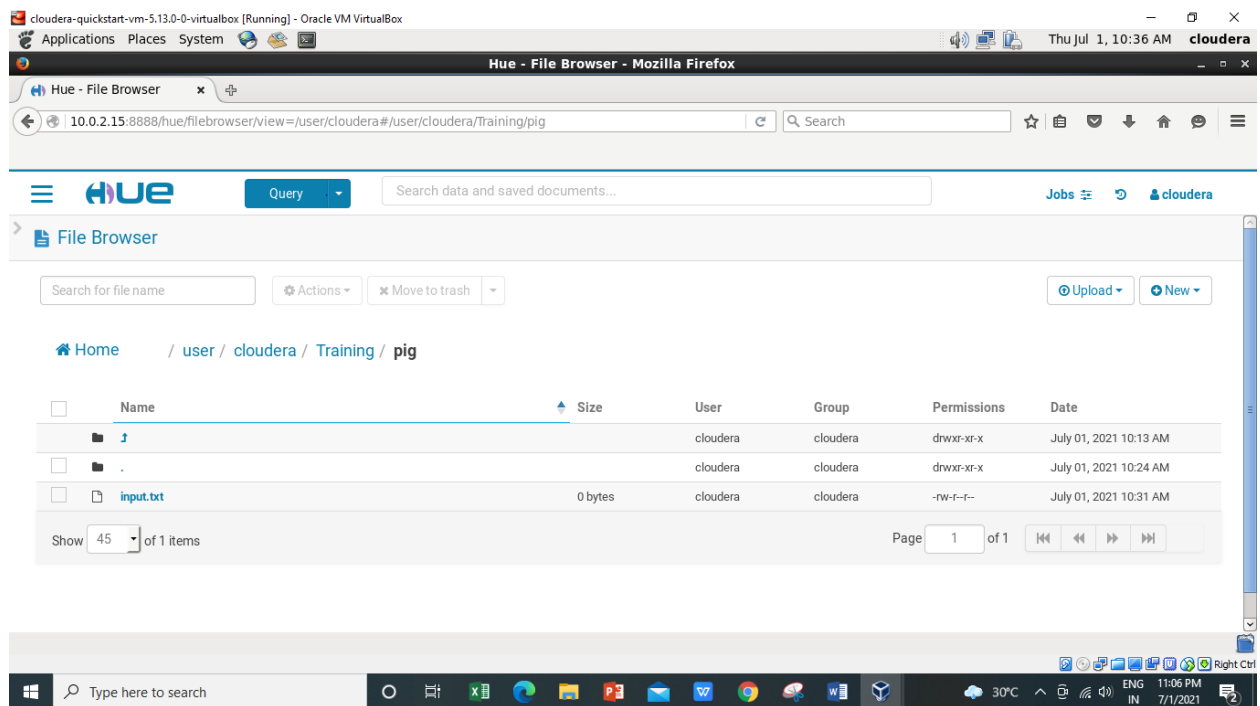
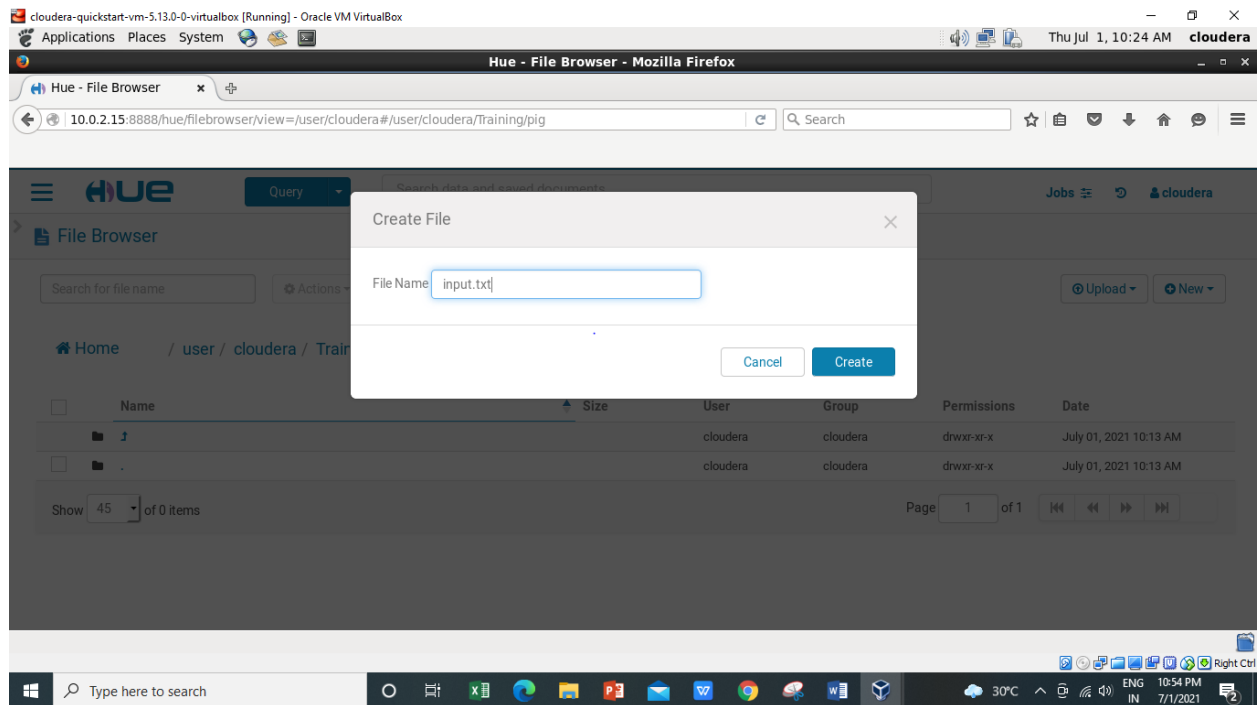




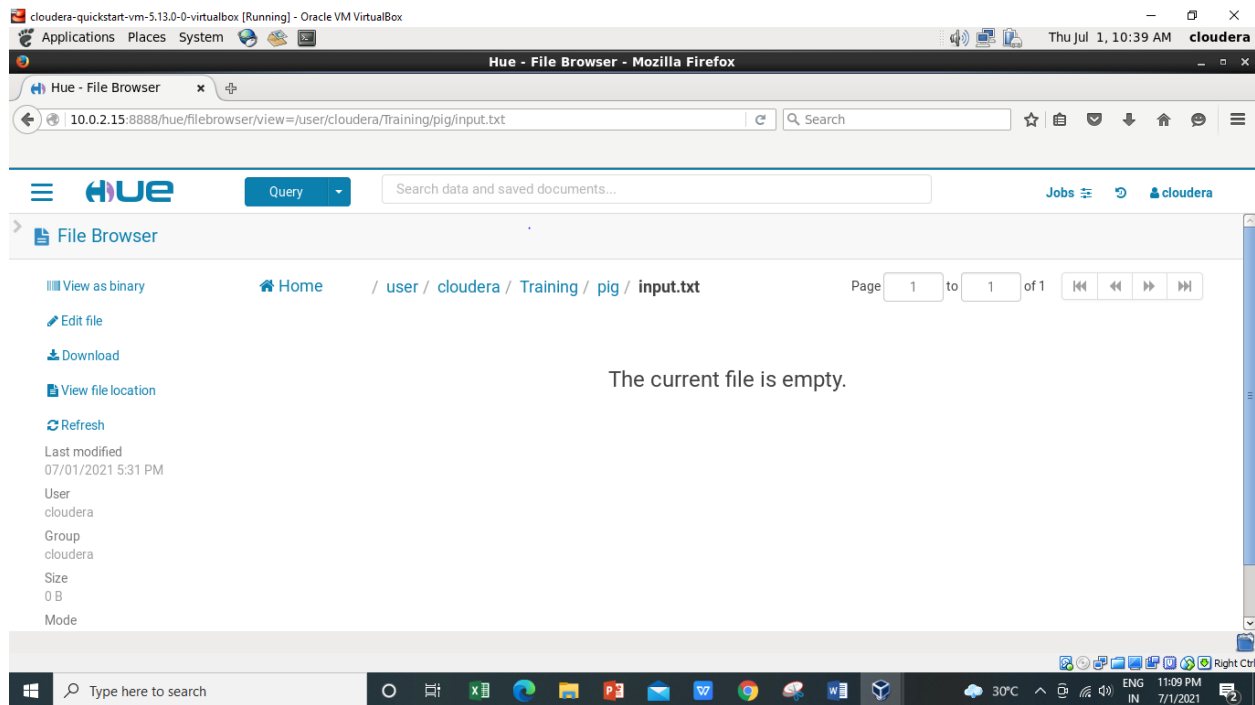
## 6) pig directory has been created inside /user/cloudera/Training



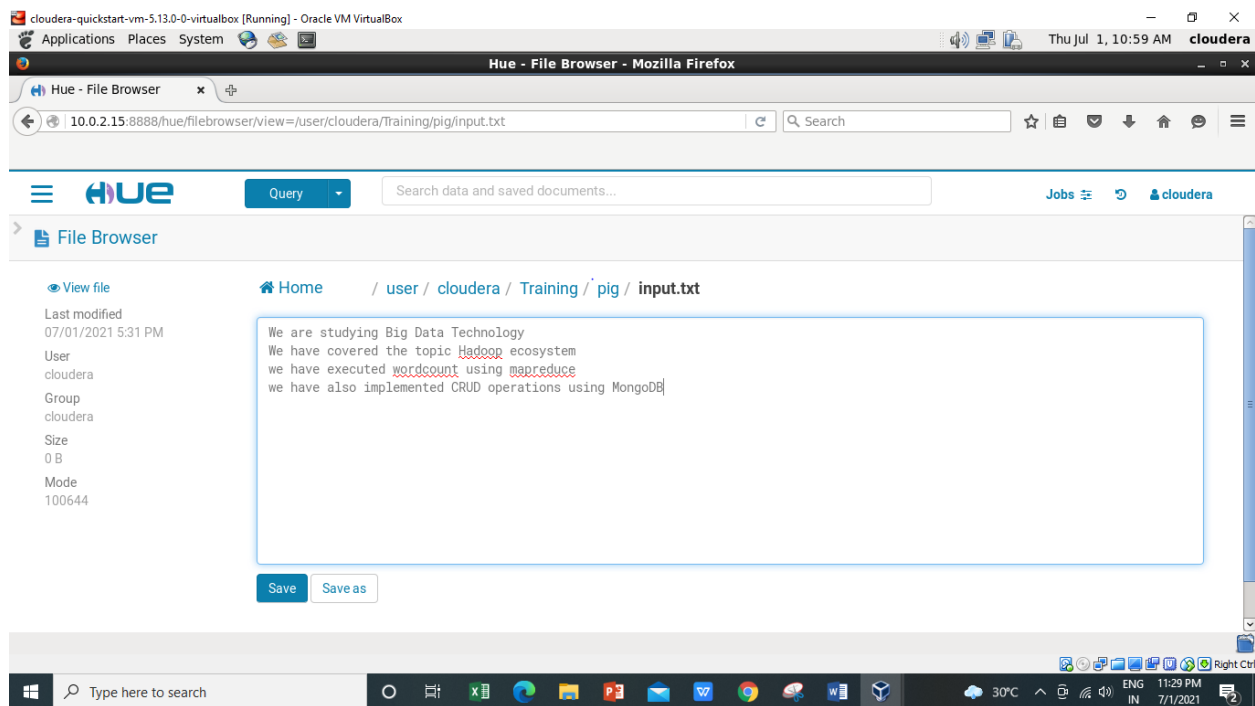
## 7) Creating input.txt file inside /usr/cloudera/training/pig directory



8) Adding some contents to this **input.txt** file.

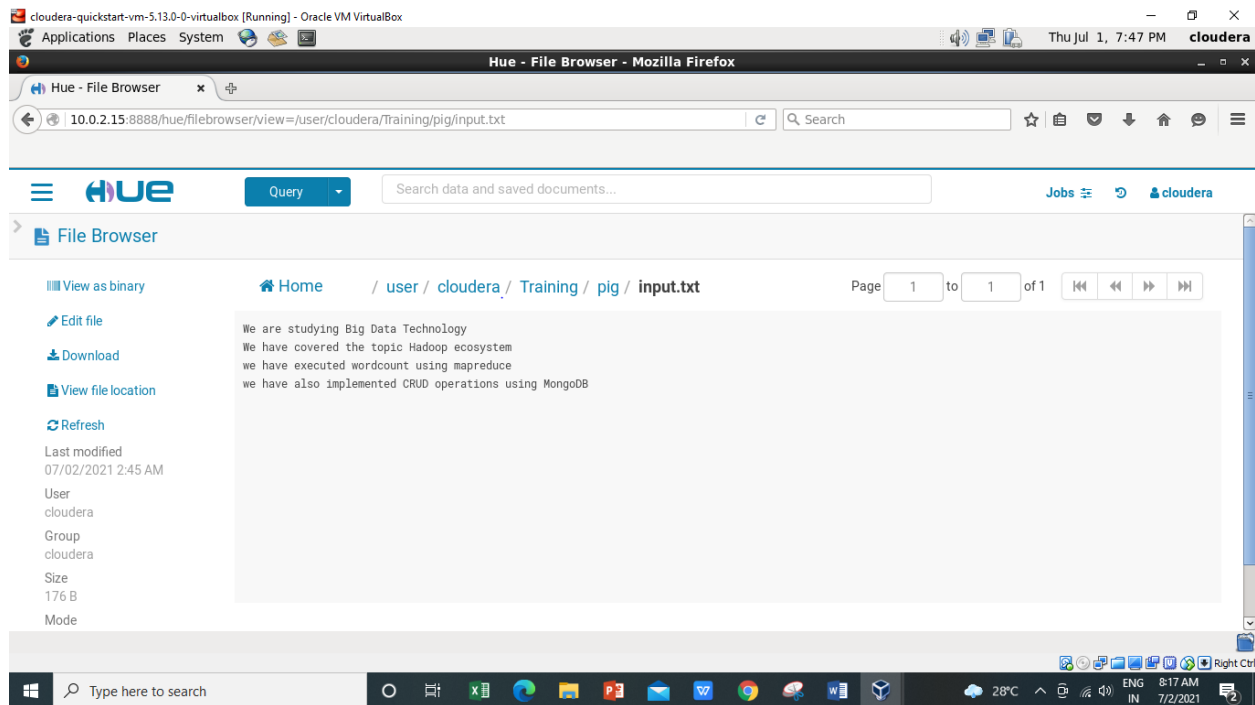


Click on **Edit file** option

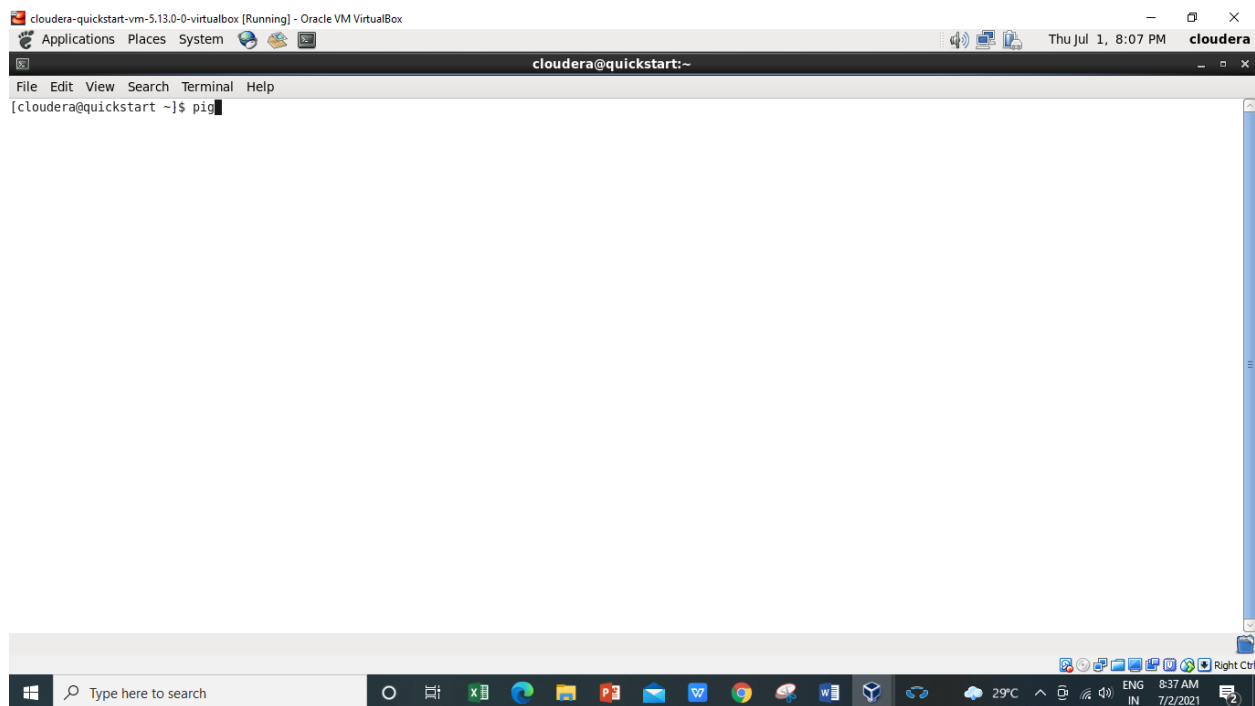


Save the **input.txt** file





9) Now Open the terminal. And start **Pig** by typing **pig** on terminal.



Now the pig started

10) Now we have to load that input file where ever it is stored. By typing the command

**Input1 = LOAD '/usr/cloudera/Training/pig/input.txt' AS (f1:chararray);**

```
grunt> input1 = LOAD '/user/cloudera/Training/pig/input.txt' AS (f1:chararray);
grunt> █
```

11) Now we are dumping the data. It will done the mapreduce task.

**DUMP input1;**

```
cloudera-quickstart-vm-5.13.0-0-virtualbox [Running] - Oracle VM VirtualBox
Applications Places System Thu Jul 1, 9:10 PM cloudera
cloudera@quickstart:~
File Edit View Search Terminal Help
grunt> input1 = LOAD '/user/cloudera/Training/pig/input.txt' AS (f1:chararray);
grunt> DUMP input1;
2021-07-01 20:48:47,378 [main] INFO org.apache.pig.tools.pigstats.ScriptState - Pig features used in the script: UNKNOWN
2021-07-01 20:48:47,382 [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]}
2021-07-01 20:48:47,397 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MRCompiler - file concatenation threshold: 100 optimistic? false
2021-07-01 20:48:47,400 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MultiQueryOptimizer - MR plan size before optimization: 1
2021-07-01 20:48:47,400 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MultiQueryOptimizer - MR plan size after optimization: 1
2021-07-01 20:48:48,081 [main] INFO org.apache.hadoop.yarn.client.RMProxy - Connecting to ResourceManager at /0.0.0.0:8032
2021-07-01 20:48:48,119 [main] INFO org.apache.pig.tools.pigstats.ScriptState - Pig script settings are added to the job
2021-07-01 20:48:48,133 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.JobControlCompiler - mapred.job.reduce.markreset.buffer.percent is not set, set to default 0.3
2021-07-01 20:48:55,017 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.JobControlCompiler - creating jar file Job7243832655036156589.jar
2021-07-01 20:48:55,027 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.JobControlCompiler - jar file Job7243832655036156589.jar created
2021-07-01 20:49:09,752 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.JobControlCompiler - Setting up single store job
2021-07-01 20:49:09,859 [main] INFO org.apache.pig.data.SchemaTupleFrontend - Key [pig.schematuple] is false, will not generate code.
2021-07-01 20:49:09,860 [main] INFO org.apache.pig.data.SchemaTupleFrontend - Starting process to move generated code to distributed cache
2021-07-01 20:49:10,003 [main] INFO org.apache.pig.data.SchemaTupleFrontend - Setting key [pig.schematuple.classes] with classes to deserialize []
2021-07-01 20:49:10,384 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - 1 map-reduce job(s) waiting for submission.
2021-07-01 20:49:10,385 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - mapred.job.tracker is deprecated. Instead, use mapreduce.jobtracker.address
2021-07-01 20:49:10,618 [JobControl] INFO org.apache.hadoop.yarn.client.RMProxy - Connecting to ResourceManager at /0.0.0.0:8032
2021-07-01 20:49:10,725 [JobControl] INFO org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Instead, use fs.defaultFS
2021-07-01 20:49:12,231 [JobControl] INFO org.apache.hadoop.mapreduce.lib.input.FileInputFormat - Total input paths to process : 1
2021-07-01 20:49:12,231 [JobControl] INFO org.apache.pig.backend.hadoop.executionengine.util.MapRedUtil - Total input paths to process : 1
2021-07-01 20:49:12,770 [JobControl] INFO org.apache.pig.backend.hadoop.executionengine.util.MapRedUtil - Total input paths (combined) to process : 1
2021-07-01 20:49:12,988 [JobControl] INFO org.apache.hadoop.mapreduce.JobSubmitter - number of splits:1
2021-07-01 20:49:15,827 [JobControl] INFO org.apache.hadoop.mapreduce.JobSubmitter - Submitting tokens for job: job_1625151715083_0002
2021-07-01 20:49:55,797 [JobControl] INFO org.apache.hadoop.yarn.client.impl.YarnClientImpl - Submitted application application_1625151715083_0002
2021-07-01 20:49:56,640 [JobControl] INFO org.apache.hadoop.mapreduce.Job - The url to track the job: http://quickstart.cloudera:8088/proxy/application_1625151715083_0002/
2021-07-01 20:49:56,641 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - HadoopJobId: job_1625151715083_0002
2021-07-01 20:49:56,641 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - Processing aliases input1
2021-07-01 20:49:56,641 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - detailed locations: M: input1[2,9],input1[-1,-1] C
```

```
cloudera-quickstart-vm-5.13.0-0-virtualbox [Running] - Oracle VM VirtualBox
Applications Places System Thu Jul 1, 9:16 PM cloudera
cloudera@quickstart:~
File Edit View Search Terminal Help
-1410184422/tmp-1396222661,

Input(s):
Successfully read 4 records (563 bytes) from: "/user/cloudera/Training/pig/input.txt"

Output(s):
Successfully stored 4 records (199 bytes) in: "hdfs://quickstart.cloudera:8020/tmp/temp-1410184422/tmp-1396222661"

Counters:
Total records written : 4
Total bytes written : 199
Spillable Memory Manager spill count : 0
Total bags proactively spilled: 0
Total records proactively spilled: 0

Job DAG:
job_1625151715083_0002

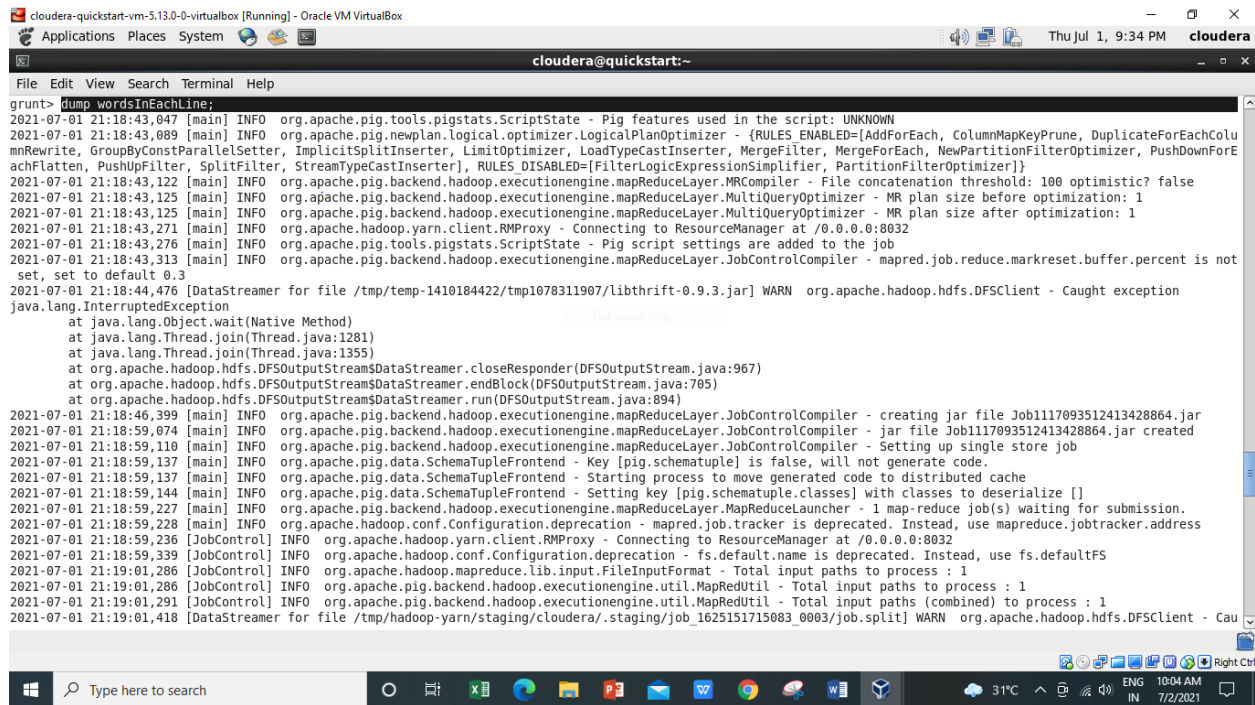
2021-07-01 21:05:21,987 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - Success!
2021-07-01 21:05:22,011 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Instead, use fs.defaultFS
2021-07-01 21:05:22,011 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - mapred.job.tracker is deprecated. Instead, use mapreduce.jobtracker.address
2021-07-01 21:05:22,021 [main] INFO org.apache.pig.data.SchemaTupleBackend - Key [pig.schematuple] was not set... will not generate code.
2021-07-01 21:05:22,284 [main] INFO org.apache.hadoop.mapreduce.lib.input.FileInputFormat - Total input paths to process : 1
2021-07-01 21:05:22,293 [main] INFO org.apache.pig.backend.hadoop.executionengine.util.MapRedUtil - Total input paths to process : 1

(We are studying Big Data Technology)
(We have covered the topic Hadoop ecosystem)
(we have executed wordcount using mapreduce)
(we have also implemented CRUD operations using MongoDB)
grunt>
```

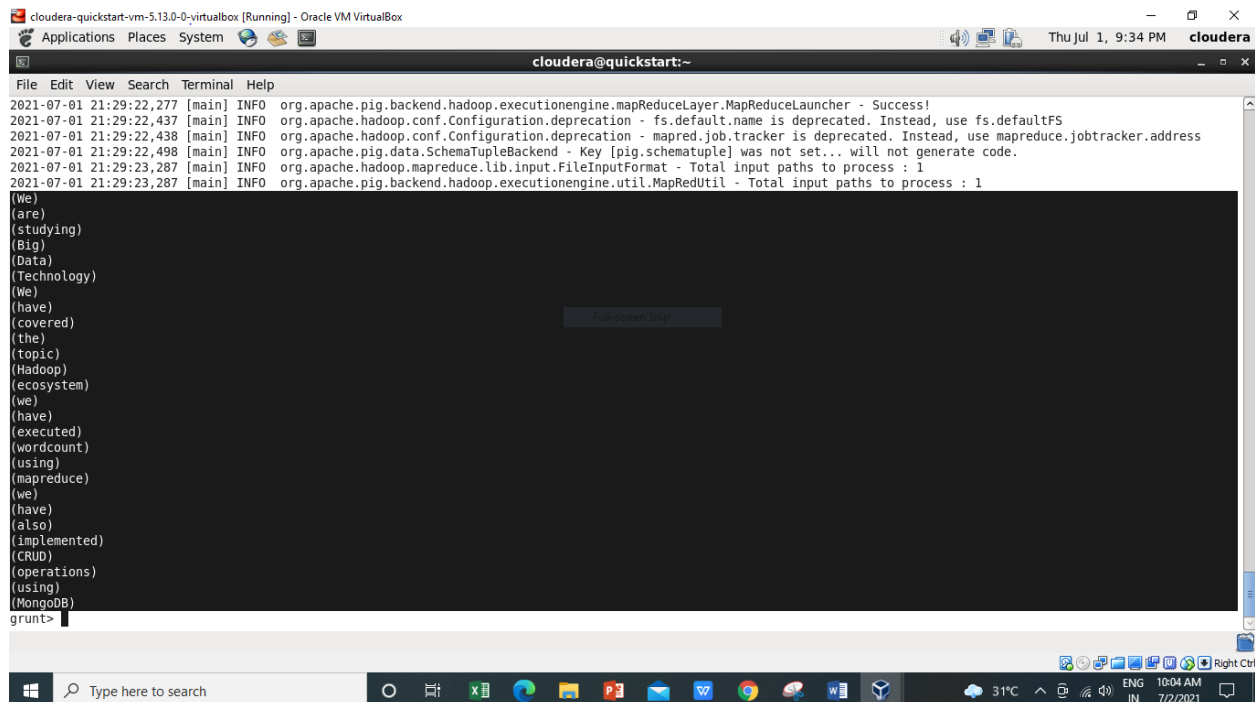
## 12) wordsInEachLine = FOREACH input1 GENERATE flatten(TOKENIZE(f1)) as word;

```
grunt> wordsInEachLine = FOREACH input1 GENERATE flatten(TOKENIZE(f1)) as word;
2021-07-01 21:14:45,902 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Instead, use fs.defaultFS
2021-07-01 21:14:45,902 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - mapred.job.tracker is deprecated. Instead, use mapreduce.jobtracker.address
grunt>
```

### 13) dump wordsInEachLine;



```
cloudera-quickstart-vm-5.13.0-0-virtualbox [Running] - Oracle VM VirtualBox
Applications Places System
cloudera@quickstart:~
File Edit View Search Terminal Help
grunt> dump wordsInEachLine;
2021-07-01 21:18:43,047 [main] INFO org.apache.pig.tools.pigstats.ScriptState - Pig features used in the script: UNKNOWN
2021-07-01 21:18:43,089 [main] INFO org.apache.pig.newplan.logical.optimizer.LogicalPlanOptimizer - {RULES_ENABLED=[AddForEach, ColumnMapKeyPrune, DuplicateForEachColu
mnRewrite, GroupByConstParallelSetter, ImplicitSplitInserter, LimitOptimizer, LoadTypeCastInserter, MergeFilter, NewPartitionFilterOptimizer, PushDownForE
achFlatten, PushUpFilter, SplitFilter, StreamTypeCastInserter], RULES_DISABLED=[FilterLogicExpressionSimplifier, PartitionFilterOptimizer]}
2021-07-01 21:18:43,122 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MRCCompiler - file concatenation threshold: 100 optimistic? false
2021-07-01 21:18:43,125 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MultiQueryOptimizer - MR plan size before optimization: 1
2021-07-01 21:18:43,271 [main] INFO org.apache.hadoop.yarn.client.RMProxy - Connecting to ResourceManager at /0.0.0.0:8032
2021-07-01 21:18:43,276 [main] INFO org.apache.pig.tools.pigstats.ScriptState - Pig script settings are added to the job
2021-07-01 21:18:43,313 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.JobControlCompiler - mapred.job.reduce.markreset.buffer.percent is not
set, set to default 0.3
2021-07-01 21:18:44,476 [DataStreamer for file /tmp/temp-1410184422/tmp1078311907/libthrift-0.9.3.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)
    at org.apache.hadoop.hdfs.DFSOutputStream$DataStreamer.run(DFSOutputStream.java:894)
2021-07-01 21:18:46,399 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.JobControlCompiler - creating jar file Job1117093512413428864.jar
2021-07-01 21:18:59,074 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.JobControlCompiler - jar file Job1117093512413428864.jar created
2021-07-01 21:18:59,110 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.JobControlCompiler - Setting up single store job
2021-07-01 21:18:59,137 [main] INFO org.apache.pig.data.SchemaTupleFrontend - Key [pig.schematuple] is false, will not generate code.
2021-07-01 21:18:59,137 [main] INFO org.apache.pig.data.SchemaTupleFrontend - Starting process to move generated code to distributed cache
2021-07-01 21:18:59,144 [main] INFO org.apache.pig.data.SchemaTupleFrontend - Setting key [pig.schematuple.classes] with classes to deserialize []
2021-07-01 21:18:59,227 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - 1 map-reduce job(s) waiting for submission.
2021-07-01 21:18:59,228 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - mapred.job.tracker is deprecated. Instead, use mapreduce.jobtracker.address
2021-07-01 21:18:59,236 [JobControl] INFO org.apache.hadoop.yarn.client.RMProxy - Connecting to ResourceManager at /0.0.0.0:8032
2021-07-01 21:18:59,339 [JobControl] INFO org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Instead, use fs.defaultFS
2021-07-01 21:19:01,286 [JobControl] INFO org.apache.hadoop.mapreduce.lib.input.FileInputFormat - Total input paths to process : 1
2021-07-01 21:19:01,286 [JobControl] INFO org.apache.pig.backend.hadoop.executionengine.util.MapRedUtil - Total input paths to process : 1
2021-07-01 21:19:01,291 [JobControl] INFO org.apache.pig.backend.hadoop.executionengine.util.MapRedUtil - Total input paths (combined) to process : 1
2021-07-01 21:19:01,418 [DataStreamer for file /tmp/hadoop-yarn/staging/cloudera/staging/job_1625151715083_0003/job.split] WARN org.apache.hadoop.hdfs.DFSClient - Cau
```

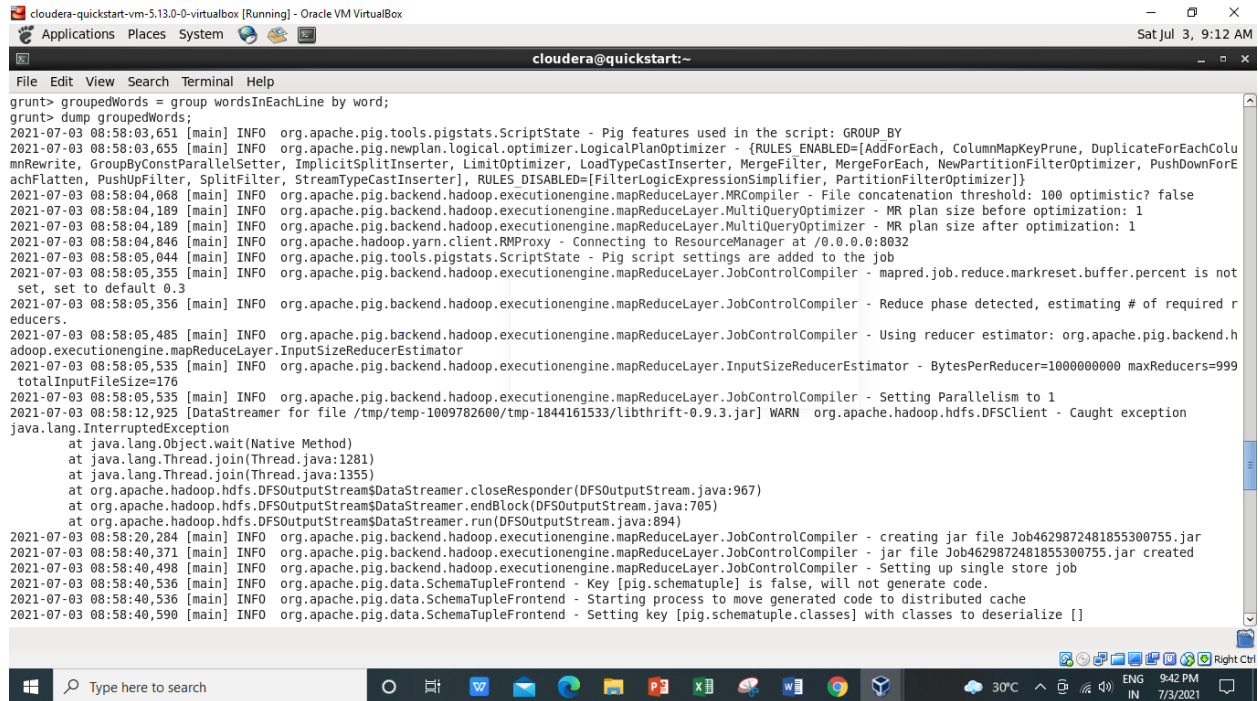


```
cloudera-quickstart-vm-5.13.0-0-virtualbox [Running] - Oracle VM VirtualBox
Applications Places System
cloudera@quickstart:~
File Edit View Search Terminal Help
2021-07-01 21:29:22,277 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - Success!
2021-07-01 21:29:22,437 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Instead, use fs.defaultFS
2021-07-01 21:29:22,438 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - mapred.job.tracker is deprecated. Instead, use mapreduce.jobtracker.address
2021-07-01 21:29:22,498 [main] INFO org.apache.pig.data.SchemaTupleBackend - Key [pig.schematuple] was not set... will not generate code.
2021-07-01 21:29:23,287 [main] INFO org.apache.hadoop.mapreduce.lib.input.FileInputFormat - Total input paths to process : 1
2021-07-01 21:29:23,287 [main] INFO org.apache.pig.backend.hadoop.executionengine.util.MapRedUtil - Total input paths to process : 1
(we)
(are)
(studying)
(Big)
(Data)
(Technology)
(we)
(have)
(covered)
(the)
(topic)
(Hadoop)
(ecosystem)
(we)
(have)
(executed)
(wordcount)
(using)
(mapreduce)
(we)
(have)
(also)
(implemented)
(CRUD)
(operations)
(using)
(MongoDB)
grunt>
```

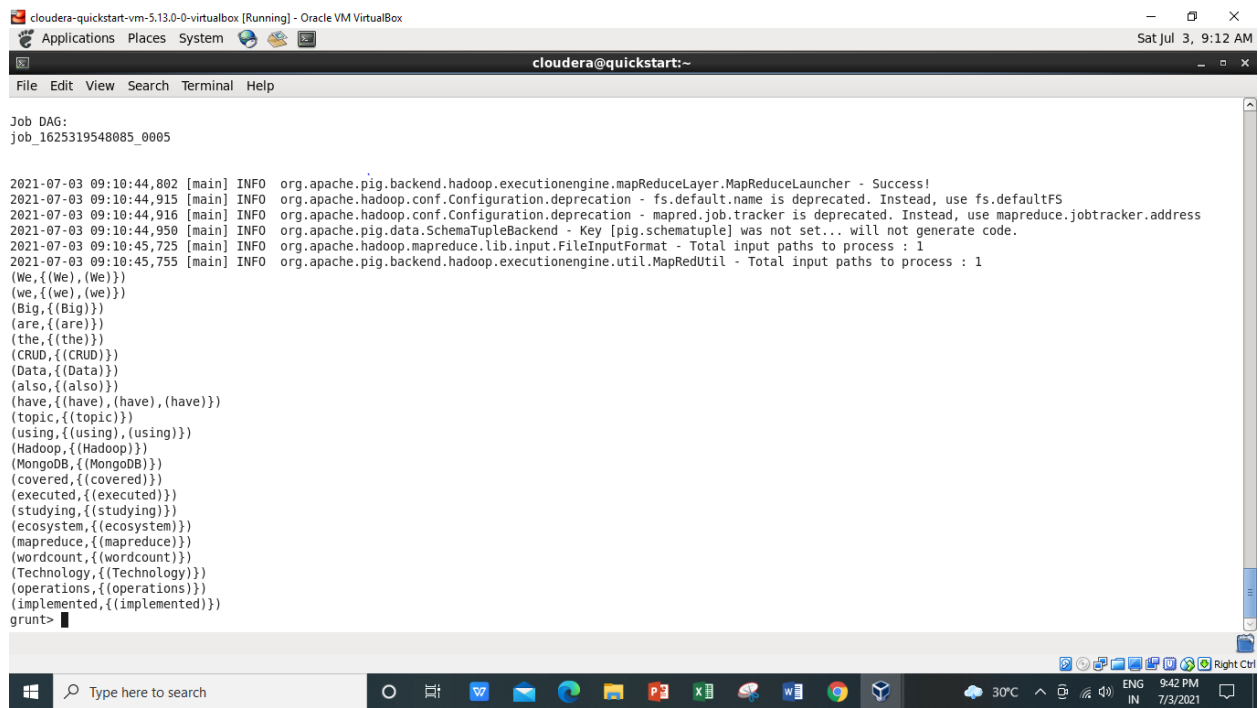
14) Now grouping the words present in each line.  
groupedWords = group wordsInEachLine by word;

```
grunt> groupedWords = group wordsInEachLine by word;
grunt> █
```

## dump groupedWords;



```
cloudera-quickstart-vm-5.13.0-0-virtualbox [Running] - Oracle VM VirtualBox
Applications Places System
cloudera@quickstart:~
File Edit View Search Terminal Help
grunt> groupedWords = group wordsInEachLine by word;
grunt> dump groupedWords;
2021-07-03 08:58:03,651 [main] INFO org.apache.pig.tools.pigstats.ScriptState - Pig features used in the script: GROUP_BY
2021-07-03 08:58:03,655 [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]}
2021-07-03 08:58:04,068 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MRCCompiler - File concatenation threshold: 100 optimistic? false
2021-07-03 08:58:04,189 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MultiQueryOptimizer - MR plan size before optimization: 1
2021-07-03 08:58:04,189 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MultiQueryOptimizer - MR plan size after optimization: 1
2021-07-03 08:58:04,846 [main] INFO org.apache.hadoop.yarn.client.RMProxy - Connecting to ResourceManager at /0.0.0.0:8032
2021-07-03 08:58:05,044 [main] INFO org.apache.pig.tools.pigstats.ScriptState - Pig script settings are added to the job
2021-07-03 08:58:05,355 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.JobControlCompiler - mapred.job.reduce.markreset.buffer.percent is not set, set to default 0.3
2021-07-03 08:58:05,356 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.JobControlCompiler - Reduce phase detected, estimating # of required reducers.
2021-07-03 08:58:05,485 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.JobControlCompiler - Using reducer estimator: org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.InputSizeReducerEstimator
2021-07-03 08:58:05,535 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.InputSizeReducerEstimator - BytesPerReducer=1000000000 maxReducers=999 totalInputFileSize=176
2021-07-03 08:58:05,535 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.JobControlCompiler - Setting Parallelism to 1
2021-07-03 08:58:12,925 [DataStreamer for file /tmp/temp-1009782600/tmp-1844161533/libthrift-0.9.3.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)
    at org.apache.hadoop.hdfs.DFSOutputStream$DataStreamer.run(DFSOutputStream.java:894)
2021-07-03 08:58:20,284 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.JobControlCompiler - creating jar file Job4629872481855300755.jar
2021-07-03 08:58:40,371 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.JobControlCompiler - jar file Job4629872481855300755.jar created
2021-07-03 08:58:40,498 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.JobControlCompiler - Setting up single store job
2021-07-03 08:58:40,536 [main] INFO org.apache.pig.data.SchemaTupleFrontend - Key [pig.schematuple] is false, will not generate code.
2021-07-03 08:58:40,536 [main] INFO org.apache.pig.data.SchemaTupleFrontend - Starting process to move generated code to distributed cache
2021-07-03 08:58:40,590 [main] INFO org.apache.pig.data.SchemaTupleFrontend - Setting key [pig.schematuple.classes] with classes to deserialize []
```



```
cloudera-quickstart-vm-5.13.0-0-virtualbox [Running] - Oracle VM VirtualBox
Applications Places System
cloudera@quickstart:~
File Edit View Search Terminal Help
Job DAG:
job_1625319548085_0005
2021-07-03 09:10:44,802 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - Success!
2021-07-03 09:10:44,915 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Instead, use fs.defaultFS
2021-07-03 09:10:44,916 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - mapred.job.tracker is deprecated. Instead, use mapreduce.jobtracker.address
2021-07-03 09:10:44,950 [main] INFO org.apache.pig.data.SchemaTupleBackend - Key [pig.schematuple] was not set... will not generate code.
2021-07-03 09:10:45,725 [main] INFO org.apache.hadoop.mapreduce.lib.input.FileInputFormat - Total input paths to process : 1
2021-07-03 09:10:45,755 [main] INFO org.apache.pig.backend.hadoop.executionengine.util.MapRedUtil - Total input paths to process : 1
(we, (we), (we))
(we, (we), (we))
(Big, {(Big)})
(are, {(are)})
(the, {(the)})
(CRUD, {(CRUD)})
(Data, {(Data)})
(also, {(also)})
(have, {(have), (have)})
(topic, {(topic)})
(using, {(using), (using)})
(Hadoop, {(Hadoop)})
(MongoDB, {(MongoDB)})
(covered, {(covered)})
(executed, {(executed)})
(studying, {(studying)})
(ecosystem, {(ecosystem)})
(mapreduce, {(mapreduce)})
(wordcount, {(wordcount)})
(Technology, {(Technology)})
(operations, {(operations)})
(implemented, {(implemented)})
grunt> █
```



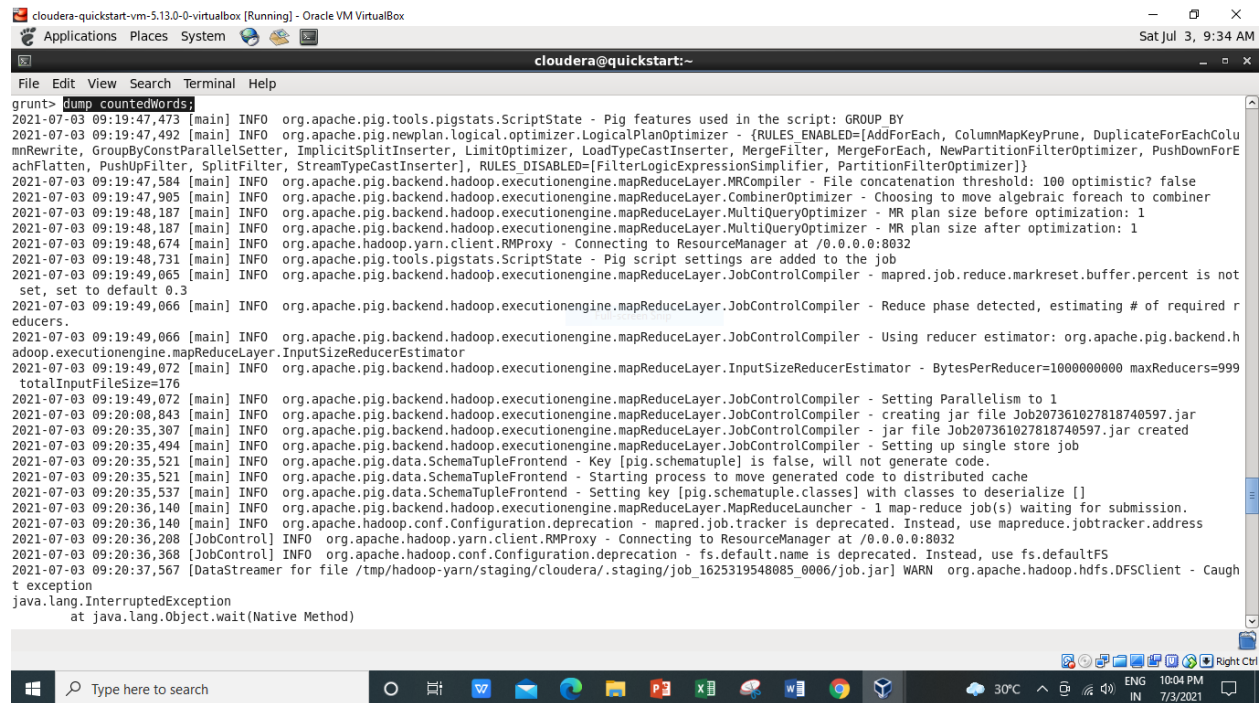
15) Now we count those words. For each group we count words in each line.

**countedWords = foreach groupedWords generate group, COUNT(wordsInEachLine);**

```
grunt> countedWords = foreach groupedWords generate group, COUNT(wordsInEachLine);
grunt> █
```

16) dump countedWords;

Now the Final Output we are getting as word count for every word.



```
cloudera-quickstart-vm-5.13.0-0-virtualbox [Running] - Oracle VM VirtualBox
Applications Places System
cloudera@quickstart:~
File Edit View Search Terminal Help
grunt> dump countedWords;
2021-07-03 09:19:47,473 [main] INFO org.apache.pig.tools.pigstats.ScriptState - Pig features used in the script: GROUP BY
2021-07-03 09:19:47,492 [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]}
2021-07-03 09:19:47,584 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MRCompiler - File concatenation threshold: 100 optimistic? false
2021-07-03 09:19:47,985 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.CombinerOptimizer - Choosing to move algebraic foreach to combiner
2021-07-03 09:19:48,187 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MultiQueryOptimizer - MR plan size before optimization: 1
2021-07-03 09:19:48,187 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MultiQueryOptimizer - MR plan size after optimization: 1
2021-07-03 09:19:48,674 [main] INFO org.apache.hadoop.yarn.client.RMPProxy - Connecting to ResourceManager at /0.0.0.0:8032
2021-07-03 09:19:48,731 [main] INFO org.apache.pig.tools.pigstats.ScriptState - Pig script settings are added to the job
2021-07-03 09:19:49,065 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.JobControlCompiler - mapred.job.reduce.markreset.buffer.percent is not set, set to default 0.3
2021-07-03 09:19:49,066 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.JobControlCompiler - Reduce phase detected, estimating # of required reducers.
2021-07-03 09:19:49,066 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.JobControlCompiler - Using reducer estimator: org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.InputSizeReducerEstimator
2021-07-03 09:19:49,072 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.InputSizeReducerEstimator - BytesPerReducer=1000000000 maxReducers=999 totalInputFileSize=176
2021-07-03 09:19:49,072 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.JobControlCompiler - Setting Parallelism to 1
2021-07-03 09:20:08,843 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.JobControlCompiler - creating jar file Job207361027818740597.jar
2021-07-03 09:20:35,307 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.JobControlCompiler - jar file Job207361027818740597.jar created
2021-07-03 09:20:35,494 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.JobControlCompiler - Setting up single store job
2021-07-03 09:20:35,521 [main] INFO org.apache.pig.data.SchemaTupleFrontend - Key [pig.schematuple] is false, will not generate code.
2021-07-03 09:20:35,521 [main] INFO org.apache.pig.data.SchemaTupleFrontend - Starting process to move generated code to distributed cache
2021-07-03 09:20:35,537 [main] INFO org.apache.pig.data.SchemaTupleFrontend - Setting key [pig.schematuple.classes] with classes to deserialize []
2021-07-03 09:20:36,140 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - 1 map-reduce job(s) waiting for submission.
2021-07-03 09:20:36,140 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - mapred.job.tracker is deprecated. Instead, use mapreduce.jobtracker.address
2021-07-03 09:20:36,208 [JobControl] INFO org.apache.hadoop.yarn.client.RMPProxy - Connecting to ResourceManager at /0.0.0.0:8032
2021-07-03 09:20:36,368 [JobControl] INFO org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Instead, use fs.defaultFS
2021-07-03 09:20:37,567 [DataStreamer for file /tmp/hadoop-yarn/staging/cloudera/.staging/job_1625319548085_0006/job.jar] WARN org.apache.hadoop.hdfs.DFSClient - Caught exception
java.lang.InterruptedException
    at java.lang.Object.wait(Native Method)
```

The screenshot shows a terminal window titled "cloudera-quickstart-vm-5.13.0-0-virtualbox [Running] - Oracle VM VirtualBox". The terminal displays the output of a Hadoop job. The logs show various INFO messages from the Hadoop ecosystem, including MapReduceLauncher, Configuration.deprecation, SchemaTupleBackend, and MapReduceUtil. The job DAG is identified as "job\_1625319548085\_0006". The word count output is displayed in a dark box, showing words and their frequencies: (We,2), (we,2), (Big,1), (are,1), (the,1), (CRUD,1), (Data,1), (also,1), (have,3), (topic,1), (using,2), (Hadoop,1), (MongoDB,1), (covered,1), (executed,1), (studying,1), (ecosystem,1), (mapreduce,1), (wordcount,1), (Technology,1), (operations,1), and (implemented,1). The prompt "grunt>" is visible at the bottom of the terminal window.

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

Job DAG:
job_1625319548085_0006

2021-07-03 09:33:44,640 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - Success!
2021-07-03 09:33:44,886 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Instead, use fs.defaultFS
2021-07-03 09:33:44,887 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - mapred.job.tracker is deprecated. Instead, use mapreduce.jobtracker.address
2021-07-03 09:33:44,928 [main] INFO org.apache.pig.data.SchemaTupleBackend - Key [pig.schematuple] was not set... will not generate code.
2021-07-03 09:33:45,722 [main] INFO org.apache.hadoop.mapreduce.lib.input.FileInputFormat - Total input paths to process : 1
2021-07-03 09:33:45,730 [main] INFO org.apache.pig.backend.hadoop.executionengine.util.MapRedUtil - Total input paths to process : 1

(We,2)
(we,2)
(Big,1)
(are,1)
(the,1)
(CRUD,1)
(Data,1)
(also,1)
(have,3)
(topic,1)
(using,2)
(Hadoop,1)
(MongoDB,1)
(covered,1)
(executed,1)
(studying,1)
(ecosystem,1)
(mapreduce,1)
(wordcount,1)
(Technology,1)
(operations,1)
(implemented,1)
grunt>
```

As we can see from above image the Word “We” start with capital W occurred twice, word “we” start with small w occurred twice, word “Big” occurred once, and so on.

17) Now Exit from the grunt shell using quit command.

```
grunt> quit
[cloudera@quickstart ~]$
```