

## Practical\_9

### To implement Word Count problem using Pig

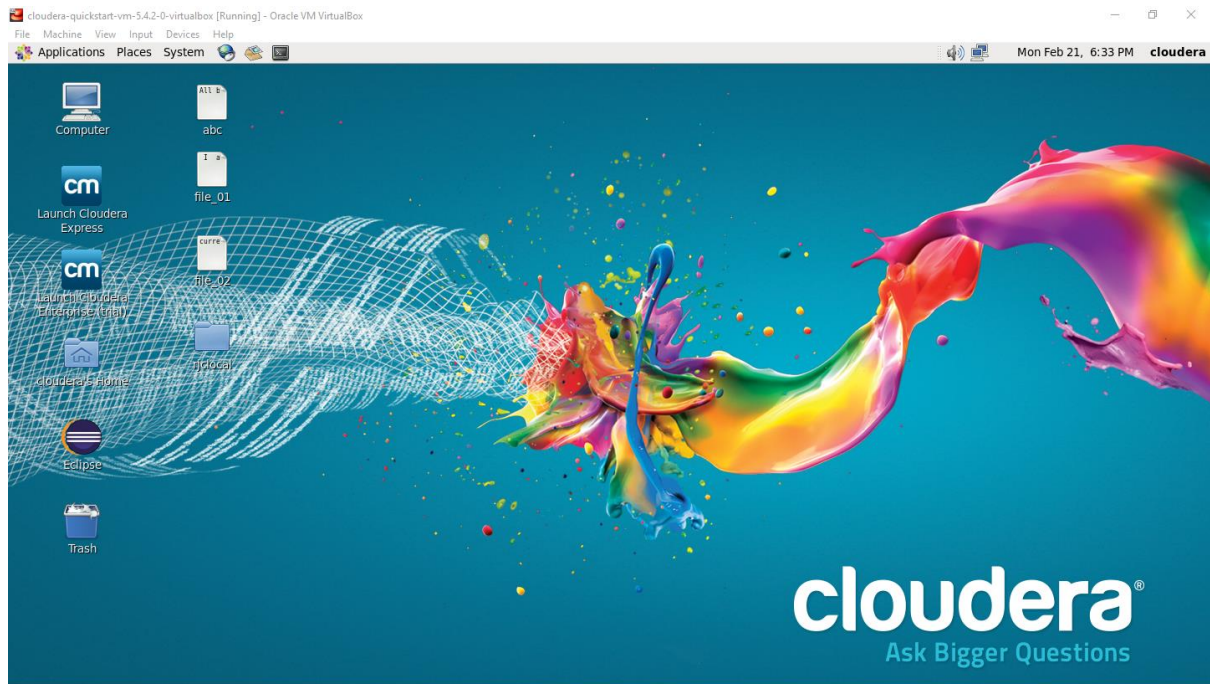
#### Apache Pig

- **Apache Pig** is a platform for analyzing large data sets that consists of a high-level language for expressing data analysis programs, coupled with infrastructure for evaluating these programs. The salient property of Pig programs is that their structure is amenable to substantial parallelization, which in turns enables them to handle very large data sets.
  - 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.
- Pig's language layer currently consists of a textual language called Pig Latin, which has the following key properties:
- **Ease of programming.** It is trivial to achieve parallel execution of simple, "embarrassingly parallel" data analysis tasks. Complex tasks comprised of multiple interrelated data transformations are explicitly encoded as data flow sequences, making them easy to write, understand, and maintain.
  - **Optimization opportunities.** The way in which tasks are encoded permits the system to optimize their execution automatically, allowing the user to focus on semantics rather than efficiency.
  - **Extensibility.** Users can create their own functions to do special-purpose processing.

## To implement Word Count problem using Pig

### Steps:

1) Start the cloudera.



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

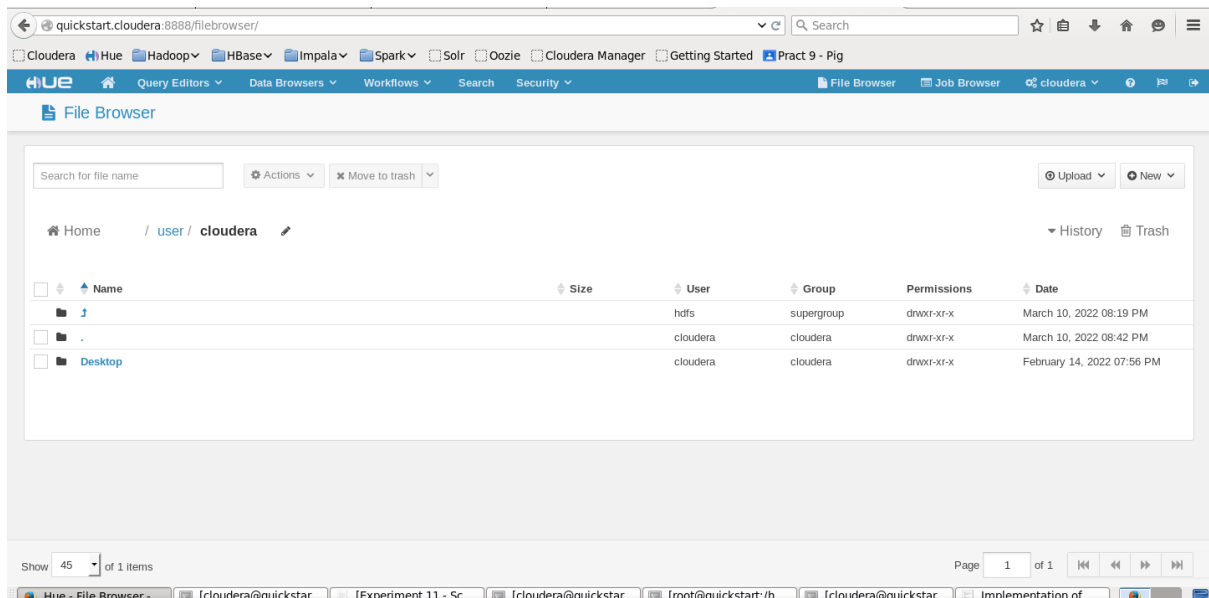
**Name: Pooja Pathak**

**Roll No: 14**



Sign in to continue to Hue

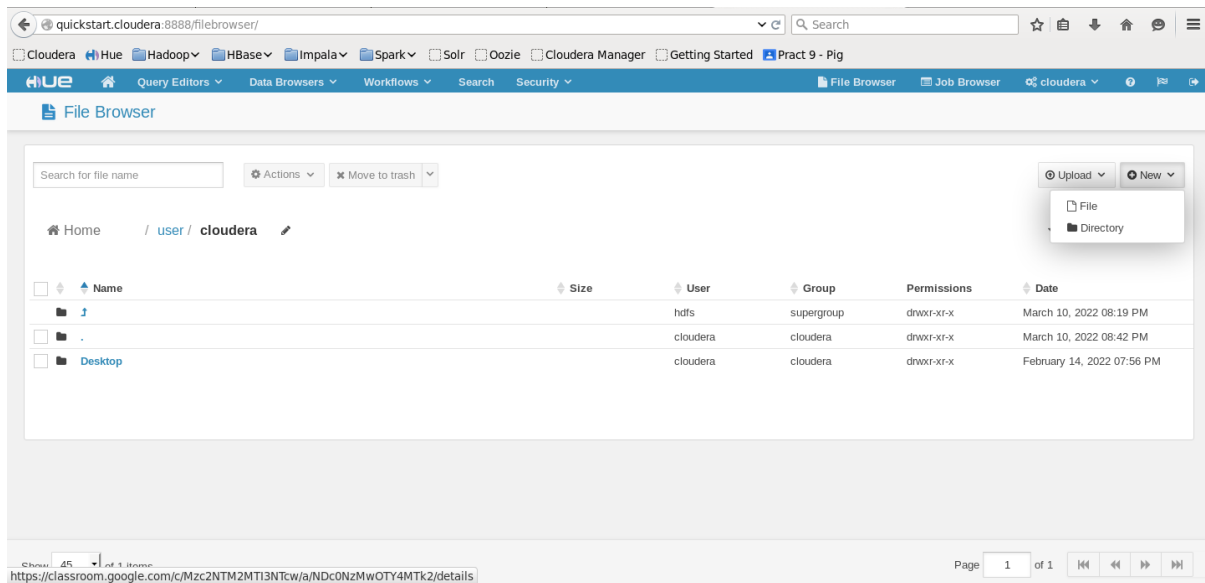
3) In Hue Go to file browser and Now open the directory /user/cloudera



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

Name: Pooja Pathak

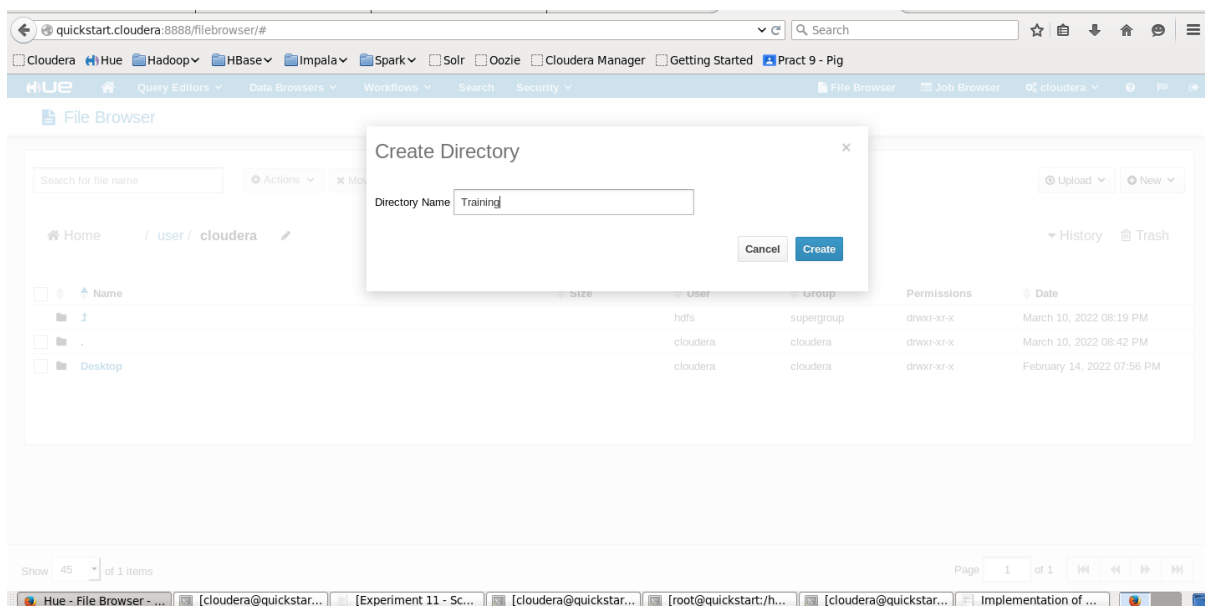
Roll No: 14



In File Browser we have New option in right corner

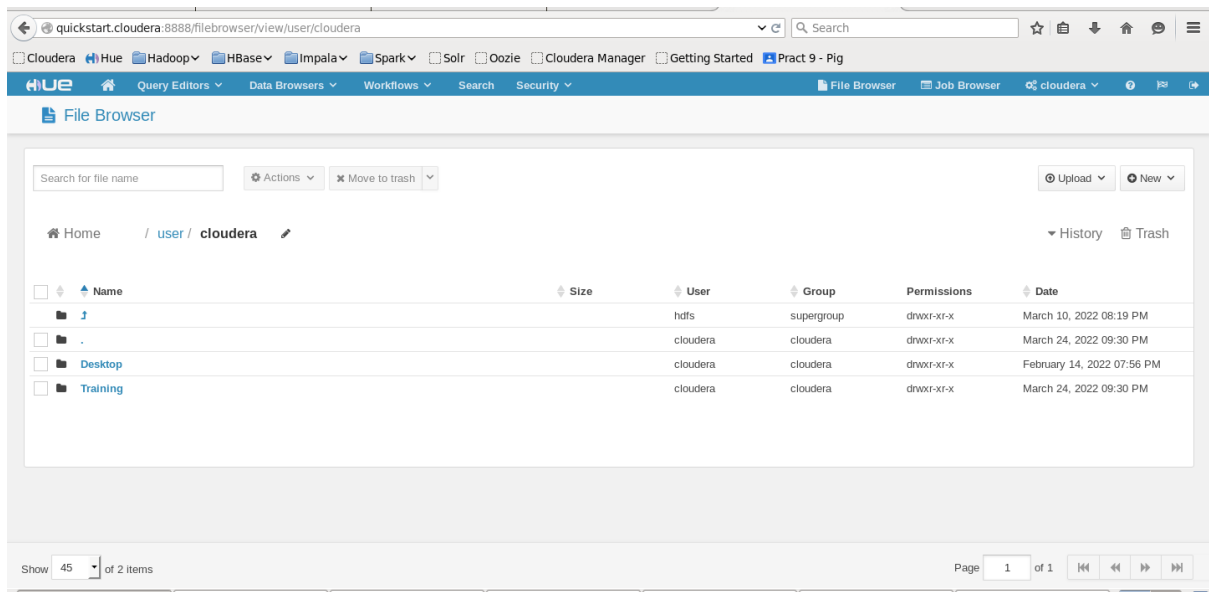
Click on New → Directory

Give the directory name And click on Create

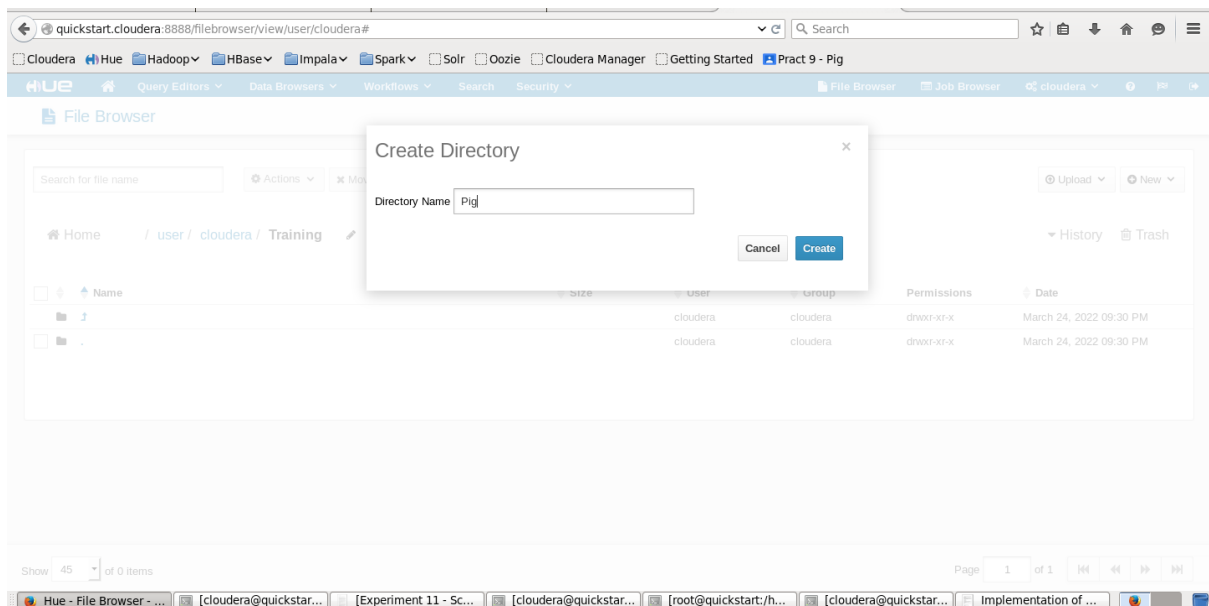


**Name: Pooja Pathak**

**Roll No: 14**



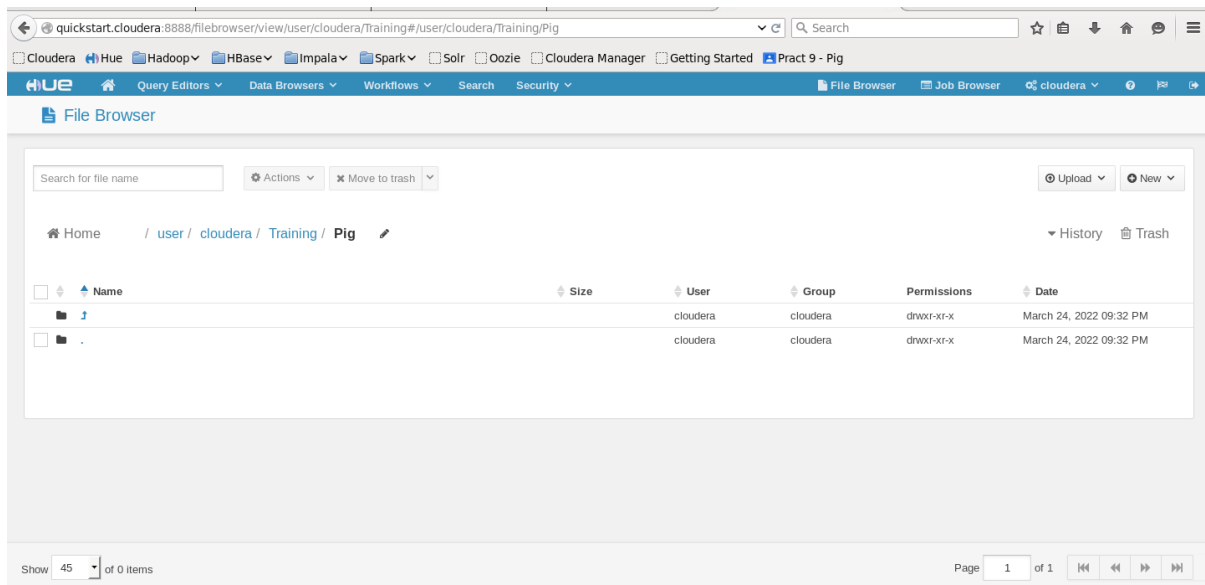
5) After creating Training directory now creating the Pig directory inside Training.



6) Pig directory has been created inside `/user/cloudera/Training`

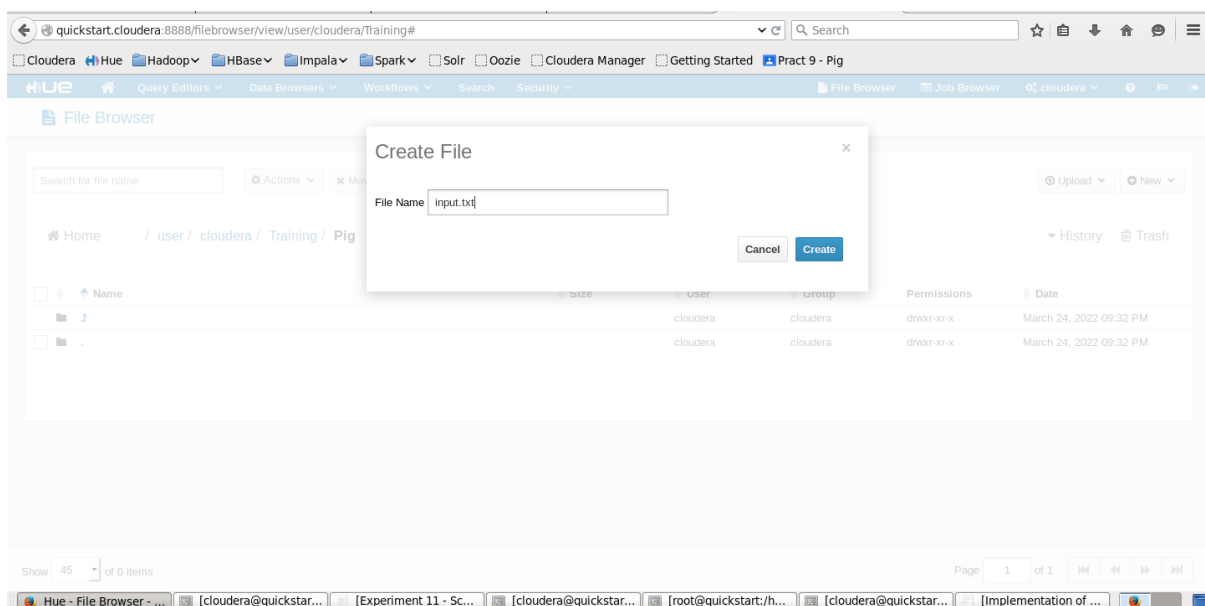
**Name: Pooja Pathak**

**Roll No: 14**



7) Creating input.txt file inside /usr/cloudera/Training/Pig directory

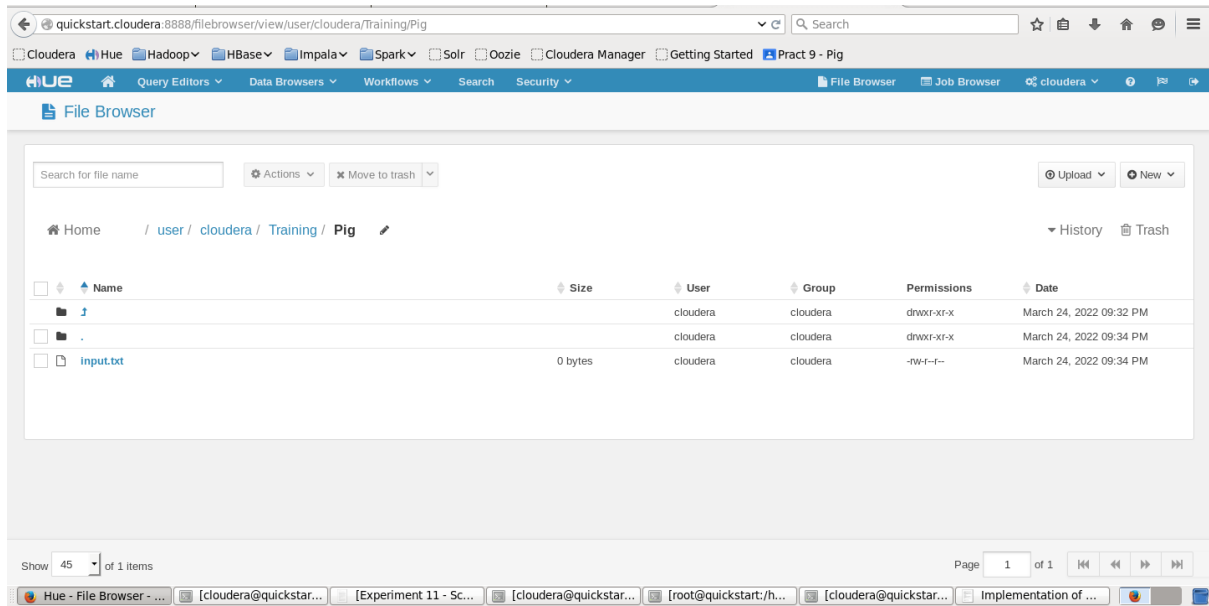
Again inside the Pig directory click on New and create file as 'input.txt'



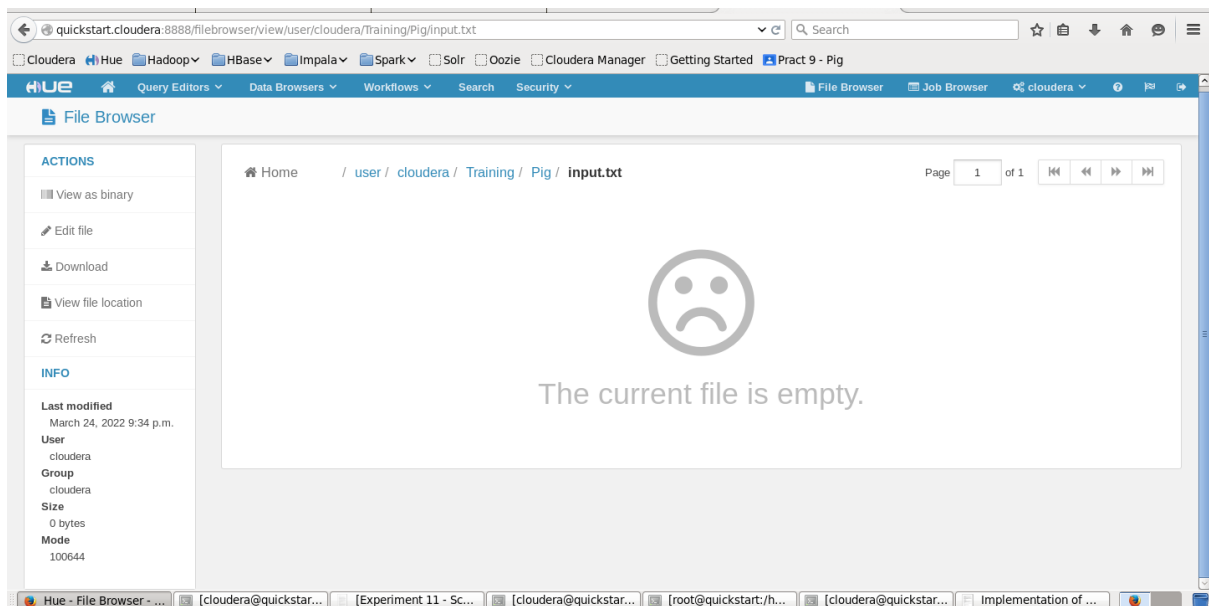
Once the file has been created click on 'input.txt' to add the content in it

Name: Pooja Pathak

Roll No: 14



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

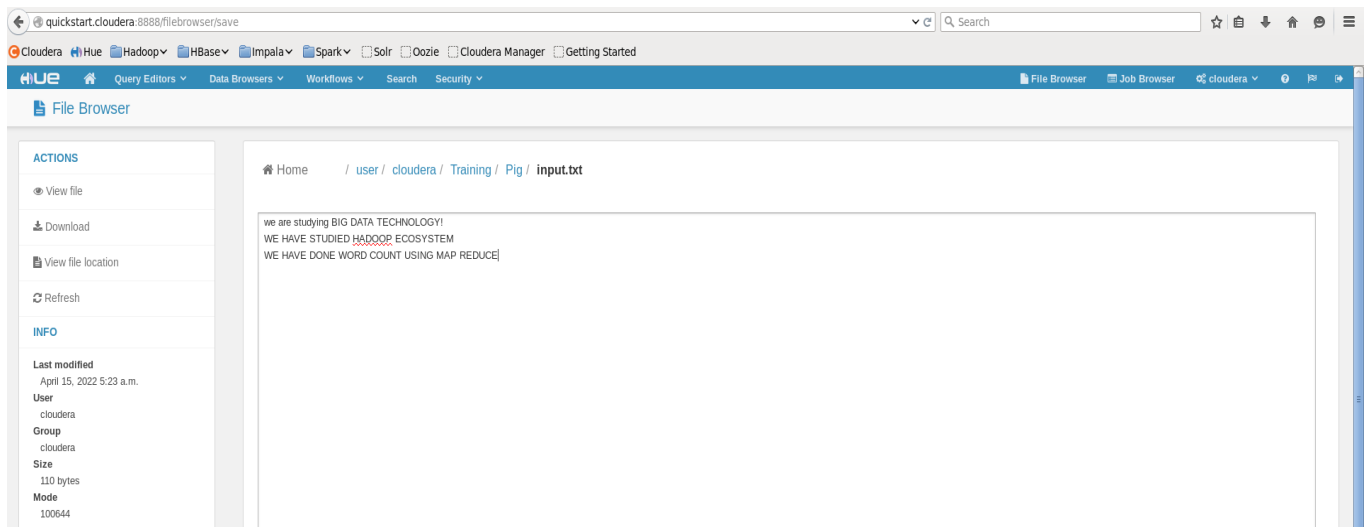


For adding content in the input file, Click on 'Edit file' option then add the content.

Save the input.txt file

**Name: Pooja Pathak**

**Roll No: 14**



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



```
[cloudera@quickstart ~]$ pig
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.
2022-03-24 21:27:30,534 [main] INFO org.apache.pig.Main - Apache Pig version 0.12.0-cdh5.4.2 (reexported) compiled May 19 2015, 17:03:41
2022-03-24 21:27:30,534 [main] INFO org.apache.pig.Main - Logging error message to: /home/cloudera/pig_1648182450505.log
2022-03-24 21:27:30,564 [main] INFO org.apache.pig.impl.util.Utils - Default bootstrap file /home/cloudera/.pigbootstrap not found
2022-03-24 21:27:31,410 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - mapred.job.tracker is deprecated. Instead, use mapreduce.jobtracker.address
2022-03-24 21:27:31,411 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Instead, use fs.defaultFS
2022-03-24 21:27:31,411 [main] INFO org.apache.pig.backend.hadoop.executionengine.HExecutionEngine - Connecting to hadoop file system at: hdfs://quickstart.cloudera:8020
2022-03-24 21:27:33,409 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - mapred.job.tracker is deprecated. Instead, use mapreduce.jobtracker.address
2022-03-24 21:27:33,409 [main] INFO org.apache.pig.backend.hadoop.executionengine.HExecutionEngine - Connecting to map-reduce job tracker at: localhost:8021
2022-03-24 21:27:33,415 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Instead, use fs.defaultFS
2022-03-24 21:27:33,466 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Instead, use fs.defaultFS
2022-03-24 21:27:33,468 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - mapred.job.tracker is deprecated. Instead, use mapreduce.jobtracker.address
2022-03-24 21:27:33,517 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Instead, use fs.defaultFS
2022-03-24 21:27:33,517 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - mapred.job.tracker is deprecated. Instead, use mapreduce.jobtracker.address
2022-03-24 21:27:33,571 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Instead, use fs.defaultFS
2022-03-24 21:27:33,571 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - mapred.job.tracker is deprecated. Instead, use mapreduce.jobtracker.address
2022-03-24 21:27:33,670 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Instead, use fs.defaultFS
```

```

2022-03-24 21:27:33,755 [main] INFO org.apache.hadoop.conf.Configuration.deprec
ation - fs.default.name is deprecated. Instead, use fs.defaultFS
2022-03-24 21:27:33,761 [main] INFO org.apache.hadoop.conf.Configuration.deprec
ation - mapred.job.tracker is deprecated. Instead, use mapreduce.jobtracker.addr
ess
2022-03-24 21:27:33,832 [main] INFO org.apache.hadoop.conf.Configuration.deprec
ation - fs.default.name is deprecated. Instead, use fs.defaultFS
2022-03-24 21:27:33,839 [main] INFO org.apache.hadoop.conf.Configuration.deprec
ation - mapred.job.tracker is deprecated. Instead, use mapreduce.jobtracker.addr
ess
2022-03-24 21:27:33,911 [main] INFO org.apache.hadoop.conf.Configuration.deprec
ation - fs.default.name is deprecated. Instead, use fs.defaultFS
2022-03-24 21:27:33,911 [main] INFO org.apache.hadoop.conf.Configuration.deprec
ation - mapred.job.tracker is deprecated. Instead, use mapreduce.jobtracker.addr
ess
2022-03-24 21:27:33,956 [main] INFO org.apache.hadoop.conf.Configuration.deprec
ation - fs.default.name is deprecated. Instead, use fs.defaultFS
2022-03-24 21:27:33,957 [main] INFO org.apache.hadoop.conf.Configuration.deprec
ation - mapred.job.tracker is deprecated. Instead, use mapreduce.jobtracker.addr
ess
grunt>

```

- 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 '/usr/cloudera/Training/pig/input.txt' AS (f1:chararray);
grunt>

```

- 11) Now we are dumping the data. It will do the MapReduce task. The Dump operator is used to run the Pig Latin statements and display the results on the screen. It is generally used for debugging Purpose.

DUMP input1;

```

grunt> Input1 = LOAD '/usr/cloudera/Training/pig/input.txt' AS (f1:chararray);
grunt> DUMP Input1;
2022-03-24 21:34:49,873 [main] INFO org.apache.pig.tools.pigstats.ScriptState - Pig features used in the script: UNKNOWN
2022-03-24 21:34:49,874 [main] INFO org.apache.pig.newplan.logical.optimizer.LogicalPlanOptimizer - (RULES_ENABLED=AddForEach, ColumnMapKeyPrune, DuplicateForEachColumnRewrite, GroupByConstParallelSetter, ImplicitSplitInserter, Limit
Optimizer, LoadTypeCastInserter, MergeFilter, MergeForEach, NewPartitionFilterOptimizer, PushDownForEachFlatten, PushUpFilter, SplitFilter, StreamTypeCastInserter), RULES_DISABLED=FilterLogicExpressionSimplifier, PartitionFilterOptimiz
er)
2022-03-24 21:34:49,875 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MRCompiler - File concatenation threshold: 100 optimistic? false
2022-03-24 21:34:49,876 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MultiQueryOptimizer - MR plan size before optimization: 1
2022-03-24 21:34:49,876 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MultiQueryOptimizer - MR plan size after optimization: 1
2022-03-24 21:34:49,900 [main] INFO org.apache.hadoop.yarn.client.RMProxy - Connecting to ResourceManager at /0.0.0.0:8032
2022-03-24 21:34:49,903 [main] INFO org.apache.pig.tools.pigstats.ScriptState - Pig script settings are added to the job
2022-03-24 21:34:49,905 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.JobControlCompiler - mapred.job.reduce.markreset.buffer.percent is not set, set to default 0.3
2022-03-24 21:34:50,156 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.JobControlCompiler - creating jar file Job3515548187496591125.jar
2022-03-24 21:34:53,653 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.JobControlCompiler - jar file Job3515548187496591125.jar created
2022-03-24 21:34:53,660 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.JobControlCompiler - Setting up single store job
2022-03-24 21:34:53,660 [main] INFO org.apache.pig.data.SchemaTupleFrontend - Key [pig.schematuple] is false, will not generate code
2022-03-24 21:34:53,667 [main] INFO org.apache.pig.data.SchemaTupleFrontend - Starting process to move generated code to distributed cache
2022-03-24 21:34:53,668 [main] INFO org.apache.pig.data.SchemaTupleFrontend - Setting key [pig.schematuple.classes] with classes to deserialize []
2022-03-24 21:34:53,667 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - 1 map-reduce job(s) waiting for submission.
2022-03-24 21:34:53,674 [JobControl] INFO org.apache.hadoop.conf.Configuration.deprecation - mapred.job.tracker is deprecated. Instead, use mapreduce.jobtracker.address
2022-03-24 21:34:53,674 [JobControl] INFO org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Instead, use fs.defaultFS
2022-03-24 21:34:53,270 [JobControl] INFO org.apache.hadoop.mapreduce.lib.input.FileInputFormat - Total input paths to process : 1
2022-03-24 21:34:53,270 [JobControl] INFO org.apache.pig.backend.hadoop.executionengine.util.MapReduceUtil - Total input paths to process : 1
2022-03-24 21:34:53,291 [JobControl] INFO org.apache.pig.backend.hadoop.executionengine.util.MapReduceUtil - Total input paths (combined) to process : 1
2022-03-24 21:34:53,331 [JobControl] INFO org.apache.hadoop.mapreduce.JobSubmitter - number of splits:1
2022-03-24 21:34:53,399 [JobControl] INFO org.apache.hadoop.mapreduce.JobSubmitter - Submitting tokens for job: job_1644540343526_0028
2022-03-24 21:34:53,501 [JobControl] INFO org.apache.hadoop.yarn.client.api.impl.YarnClientImpl - Submitted application application_1644540343526_0028

```

**Roll No: 14**

12) Here we are counting the words in each line for that we are using the following command

```

grunts wordsInEachLine = FOREACH Input1 GENERATE flatten(TOKENIZE(f1)) as word;
2022-03-25 22:09:29,072 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Instead, use fs.defaultFS
2022-03-25 22:09:29,072 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - mapred.job.tracker is deprecated. Instead, use mapreduce.jobtracker.address
grunts

```

```

2022-03-25 22:09:29,072 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Instead, use fs.defaults
2022-03-25 22:09:29,072 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - mapred.job.tracker is deprecated. Instead, use mapreduce.jobtracker.address
2022-03-25 22:09:29,072 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - mapred.job.tracker is deprecated. Instead, use mapreduce.jobtracker.address
2022-03-25 22:10:40,971 [main] INFO org.apache.pig.tools.pigstats.ScriptStats - Pig features used in the script: UNKNOWN
2022-03-25 22:10:40,971 [main] INFO org.apache.pig.mapred.logical.optimizer.LocalPigOptimizer - RULES ENABLED:AdaptFilter,ColumnMajorGroupPrune, DuplicateForEachColumnRewrite, GroupConcatParallelSetter, ImplicitSplitInserter, LimitBy
2022-03-25 22:10:40,971 [main] INFO org.apache.pig.mapred.logical.optimizer.LocalPigOptimizer - RULES DISABLED:FilterConcatExpressionsInserter, PartitionFilterOptimizer,
2022-03-25 22:10:40,971 [main] INFO org.apache.pig.mapred.logical.optimizer.LocalPigOptimizer - PushDownAggregatesInserter, PushDownFilter, SplitFilter, StreamConcatInserter,
2022-03-25 22:10:41,014 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapreduceLayer.MRCompiler - File concatenation threshold: 160 optimistic: false
2022-03-25 22:10:41,016 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapreduceLayer.MultiQueryOptimizer - MR plan size before optimization: 1
2022-03-25 22:10:41,016 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapreduceLayer.MultiQueryOptimizer - MR plan size after optimization: 1
2022-03-25 22:10:41,062 [main] INFO org.apache.hadoop.yarn.client.RMProxy - Connecting to ResourceManager at /0.0.0.0:8032
2022-03-25 22:10:41,062 [main] INFO org.apache.pig.tools.pigstats.ScriptStats - Pig script settings are added to the job
2022-03-25 22:10:41,084 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapreduceLayer.JobControlCompiler - mapred.job.reduce.markrest.buffer.percent is not set, set to default 0.3
2022-03-25 22:10:42,118 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapreduceLayer.JobControlCompiler - creating jar file job0654808738905845951.jar
2022-03-25 22:10:42,118 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapreduceLayer.JobControlCompiler - jar file job0654808738905845951.jar created
2022-03-25 22:10:43,725 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapreduceLayer.JobControlCompiler - Setting up single store job
2022-03-25 22:10:45,537 [main] INFO org.apache.pig.data.SchemaTupleFrontend - Key [pig.schematuple] is false, will not generate code.
2022-03-25 22:10:45,537 [main] INFO org.apache.pig.data.SchemaTupleFrontend - Starting process to move generated code to distributed cache
2022-03-25 22:10:45,537 [main] INFO org.apache.pig.data.SchemaTupleFrontend - Setting key [pig.schematuple.classes] with classes to deserialize []
2022-03-25 22:10:45,546 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapreduceLayer.MapReduceLauncher - 1 map-reduce jobs (waiting for submission.
2022-03-25 22:10:45,546 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Instead, use fs.defaults
2022-03-25 22:10:45,546 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - mapred.job.tracker is deprecated. Instead, use mapreduce.jobtracker.address
2022-03-25 22:10:45,559 [JobControl] INFO org.apache.hadoop.yarn.client.RMProxy - Connecting to ResourceManager at /0.0.0.0:8032
2022-03-25 22:10:45,566 [JobControl] INFO org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Instead, use fs.defaults
2022-03-25 22:10:45,566 [JobControl] INFO org.apache.hadoop.conf.Configuration.deprecation - mapred.job.tracker is deprecated. Instead, use mapreduce.jobtracker.address
2022-03-25 22:10:45,776 [JobControl] INFO org.apache.pig.backend.hadoop.executionengine.util.MapReduceUtil - Total input paths to process : 1
2022-03-25 22:10:45,778 [JobControl] INFO org.apache.pig.backend.hadoop.executionengine.util.MapReduceUtil - Total input paths (combined) to process : 1
2022-03-25 22:10:45,780 [JobControl] INFO org.apache.pig.backend.mapreduce.JobSubmitter - number of splits: 1
2022-03-25 22:10:45,787 [JobControl] INFO org.apache.hadoop.mapreduce.JobSubmitter - Submitting tokens for job: job_1644548343526_0029
2022-03-25 22:10:45,928 [JobControl] INFO org.apache.hadoop.yarn.client.api.impl.YarnClientImpl - Submitted application application_1644548343526_0029
2022-03-25 22:10:46,002 [JobControl] INFO org.apache.pig.backend.hadoop.executionengine.mapreduceLayer.JobSubmitter - http://quickstart:8030/proxy/application_1644548343526_0029/
2022-03-25 22:10:46,046 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapreduceLayer.MapReduceLauncher - mapreduce.jobid = job_1644548343526_0029
2022-03-25 22:10:46,046 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapreduceLayer.MapReduceLauncher - Processing aliases InputWordsInSequence
2022-03-25 22:10:46,046 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapreduceLayer.MapReduceLauncher - Detailed locations: M Input1[3], wordsInSequence[1]-1, c: M
2022-03-25 22:10:46,046 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapreduceLayer.MapReduceLauncher - More information at: http://localhost:8080/jobdetails_?jobid=job_1644548343526_0029
2022-03-25 22:10:46,100 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapreduceLayer.MapReduceLauncher - 0% complete
2022-03-25 22:10:46,100 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapreduceLayer.MapReduceLauncher - 50% complete
2022-03-25 22:10:46,100 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapreduceLayer.MapReduceLauncher - 100% complete
2022-03-25 22:11:11,459 [main] INFO org.apache.pig.tools.pigstats.SimplePigStats - Script Statistics:
HadoopVersion: PIGVersion: UserStd: StartedAt: FinishedAt: Features:
2.6.0-cdh.4.2 0.12.0-cdh.4.2 1.2 Cloudera 2022-03-25 22:10:41 2022-03-25 22:11:11 UNKNOWN

```



**Roll No: 14**

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

And then dumping the data by the following command.

```
dump groupedWords;
```

MSc Part-I Sem-II

BDT

**Roll No: 14**

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

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

16) After every counting of words commands, we are dumping the data  
dump countedWords;

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

[illegible]

Name: Pooja Pathak

Roll No: 14

```
2022-03-25 22:17:42,771 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - 100% complete
2022-03-25 22:17:42,771 [main] INFO org.apache.pig.tools.pigstats.SimplePigStats - Script Statistics:
HadoopVersion PigVersion UserId StartedAt FinishedAt Features
2.6.0-cdh5.4.2 0.12.0-cdh5.4.2 cloudera 2022-03-25 22:17:07 2022-03-25 22:17:42 GROUP_BY
Success!
Job Stats (time in seconds):
JobId Maps Reduces MaxMapTime MinMapTime AvgMapTime MedianMapTime MaxReduceTime MinReduceTime AvgReduceTime MedianReduceTime Alias Feature Outputs
job_1644548343526_0031 1 1 5 5 5 5 6 6 6 6 Input1,CountedWords,groupedWords,wordsInEachLine GROUP_BY,COMBINER hdfs://quickstart.cloudera:8020/tmp/temp-669075149/tmp1191620340
1191620340
Input(s):
Successfully read 3 records (497 bytes) from: "/user/cloudera/Training/pig/input.txt"
Output(s):
Successfully stored 17 records (224 bytes) in: "hdfs://quickstart.cloudera:8020/tmp/temp-669075149/tmp1191620340"
Counters:
Total records written : 17
Total bytes written : 224
Spillable Memory Manager spill count : 0
Total bags proactively spilled: 0
Total records proactively spilled: 0
Job DAG:
job_1644548343526_0031
2022-03-25 22:17:42,858 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - Success!
2022-03-25 22:17:42,858 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Instead, use fs.defaultFS
2022-03-25 22:17:42,858 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - mapred.job.tracker is deprecated. Instead, use mapreduce.jobtracker.address
2022-03-25 22:17:42,858 [main] INFO org.apache.pig.data.SchemaTupleBackend - Key [pig.schematuple] was not set... Will not generate code.
2022-03-25 22:17:42,866 [main] INFO org.apache.hadoop.mapreduce.lib.input.FileInputFormat - Total input paths to process : 1
2022-03-25 22:17:42,866 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer - Total input paths to process : 1
```

```
(WE,2)
(we,1)
(BIG,1)
(MAP,1)
(are,1)
(DATA,1)
(DONE,1)
(HAVE,2)
(WORD,1)
(COUNT,1)
(USING,1)
(HADOOP,1)
(REDUCE,1)
(STUDED,1)
(studying,1)
(ECOSYSTEM,1)
(TWECHNOLOGY!,1)
grunt>
```

As we can see from above image the Word “a” occurred twice, word “for, data” start with small w occurred twice, word “I” occurred once, and so on.

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

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