

PRACTICAL 9

To implement Word Count problem using Pig

Apache Pig

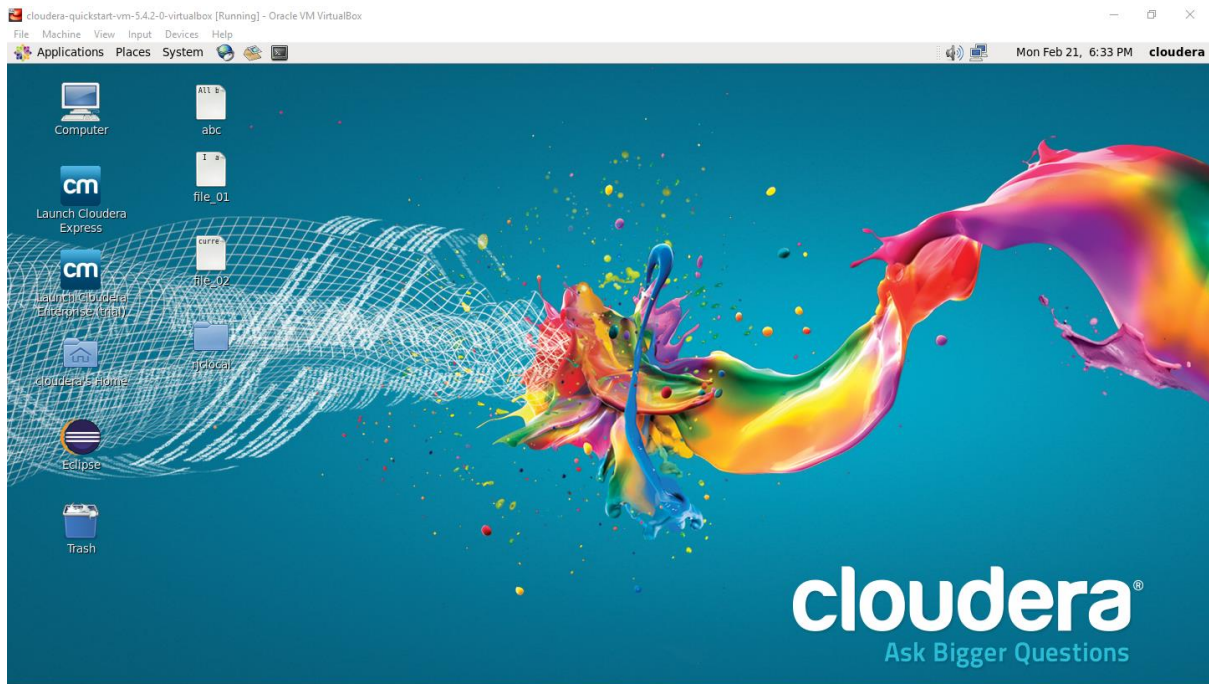
- **Apache Pig** is a platform for analyzing large data sets that consists of a high-level language for expressing data analysis programs, coupled with infrastructure for evaluating these programs. The salient property of Pig programs is that their structure is amenable to substantial parallelization, 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.

Name : - Zen Dsouza
Roll No: - 34
Subject - BDT

To implement Word Count problem using Pig

Steps:

1) Start the cloudera.



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

Name : - Zen Dsouza
Roll No: - 34
Subject - BDT



Sign in to continue to Hue

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

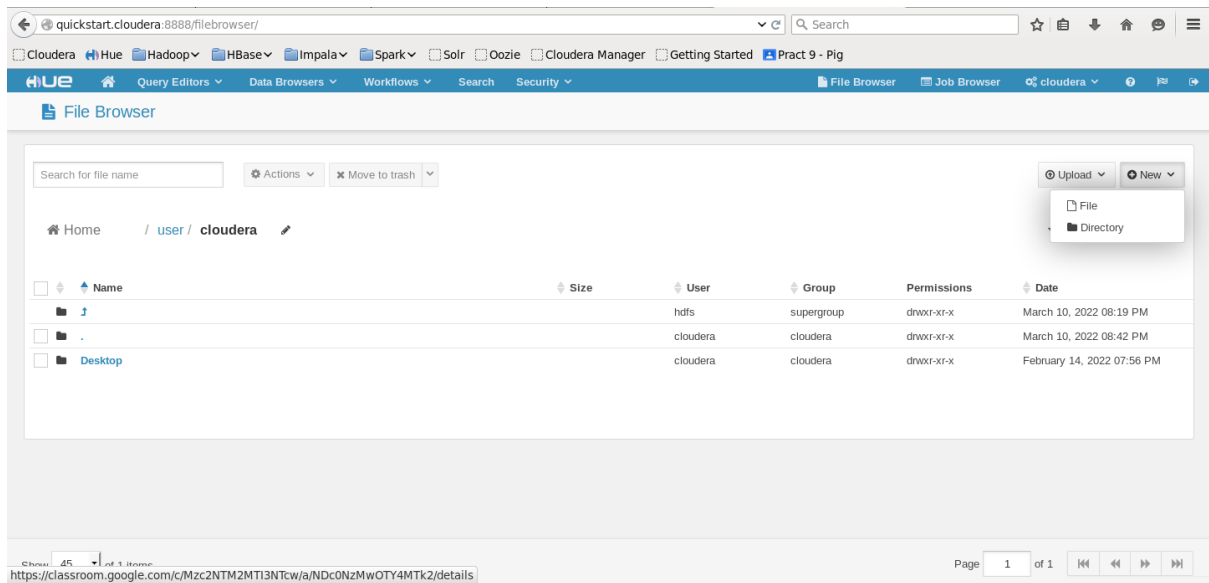
The screenshot shows the Hue File Browser interface. The breadcrumb path is Home / user / cloudera. Below the path, there is a table listing files and directories:

	Name	Size	User	Group	Permissions	Date
<input type="checkbox"/>	.		hdfs	supergroup	drwxr-xr-x	March 10, 2022 08:19 PM
<input type="checkbox"/>	..		cloudera	cloudera	drwxr-xr-x	March 10, 2022 08:42 PM
<input type="checkbox"/>	Desktop		cloudera	cloudera	drwxr-xr-x	February 14, 2022 07:56 PM

At the bottom, it shows 'Show 45 of 1 items' and 'Page 1 of 1'.

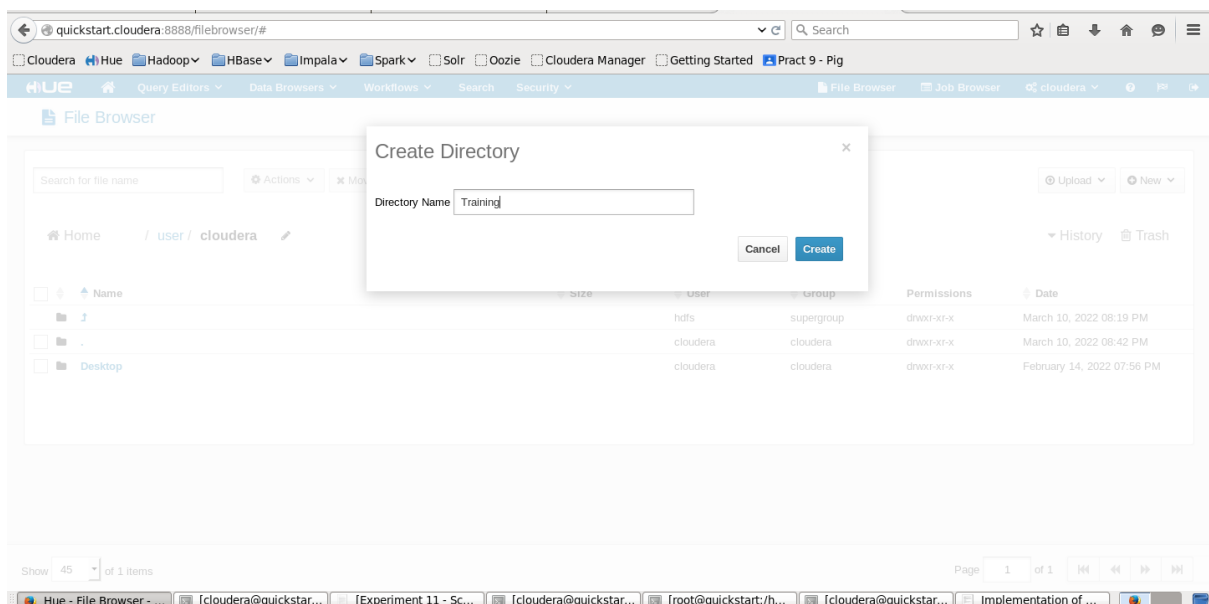
4) Now we are creating the directory as Training inside /user/cloudera

Name : - Zen Dsouza
Roll No: - 34
Subject - BDT

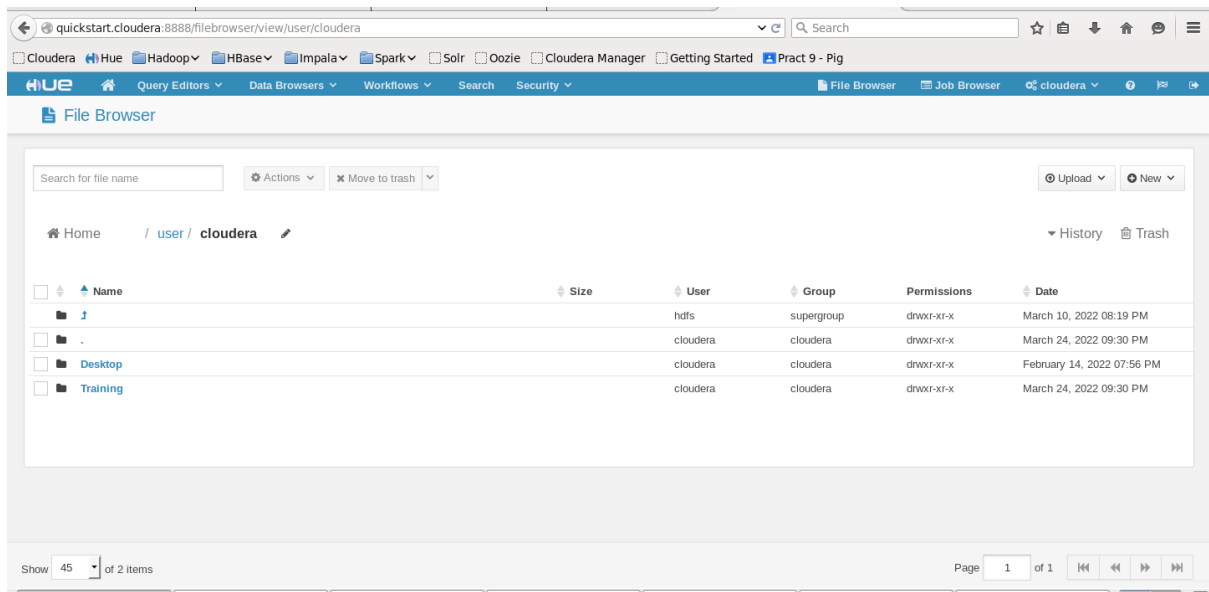


In File Browser we have New option in right corner
Click on New → Directory

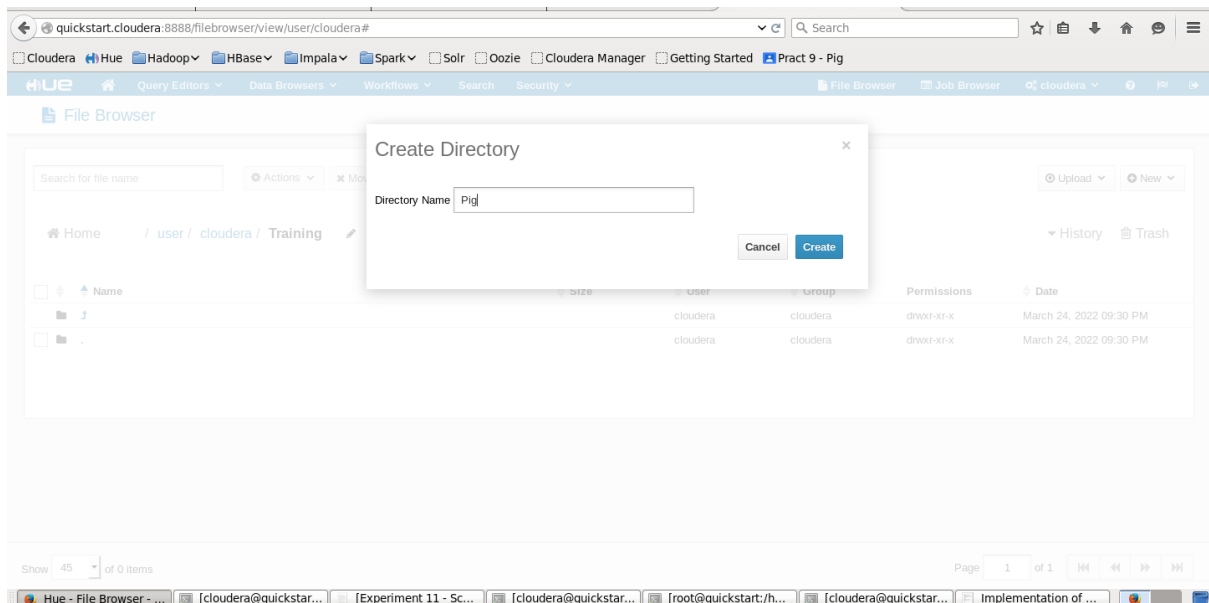
Give the directory name And click on Create



Name : - Zen Dsouza
Roll No: - 34
Subject - BDT

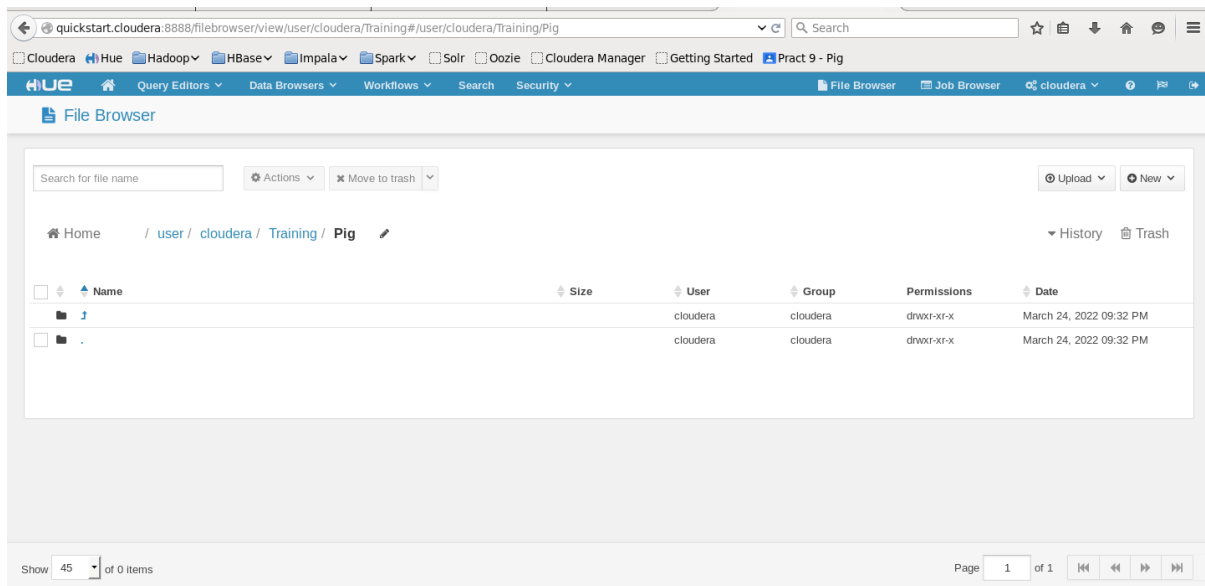


5) After creating Training directory now creating the Pig directory inside Training.



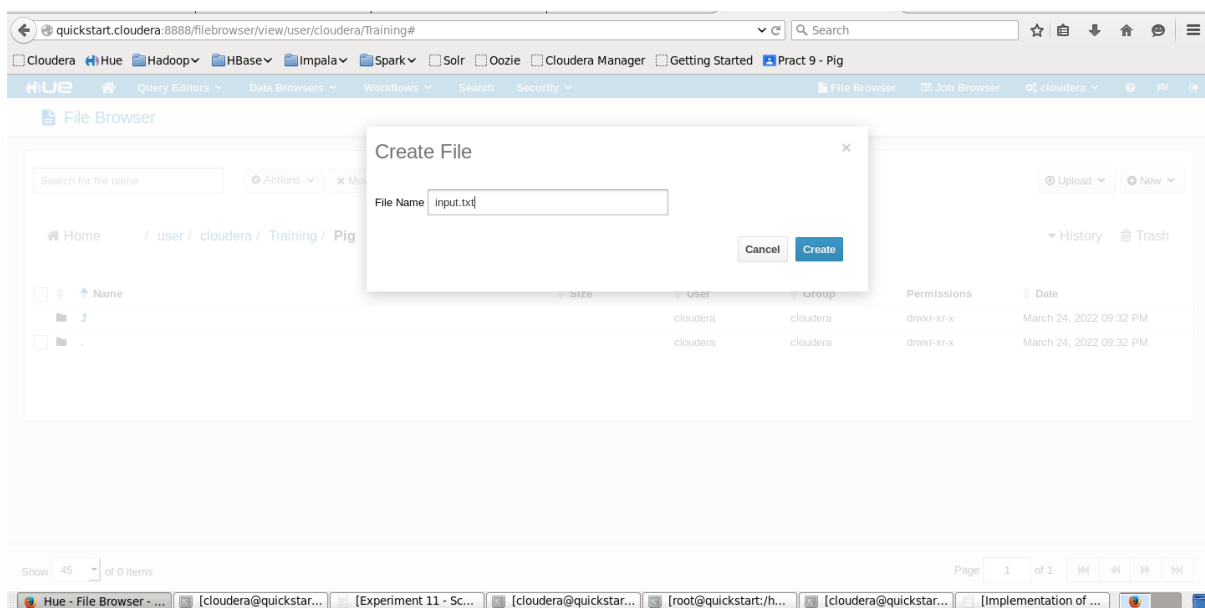
6) Pig directory has been created inside `/user/cloudera/Training`

Name : - Zen Dsouza
Roll No: - 34
Subject - BDT



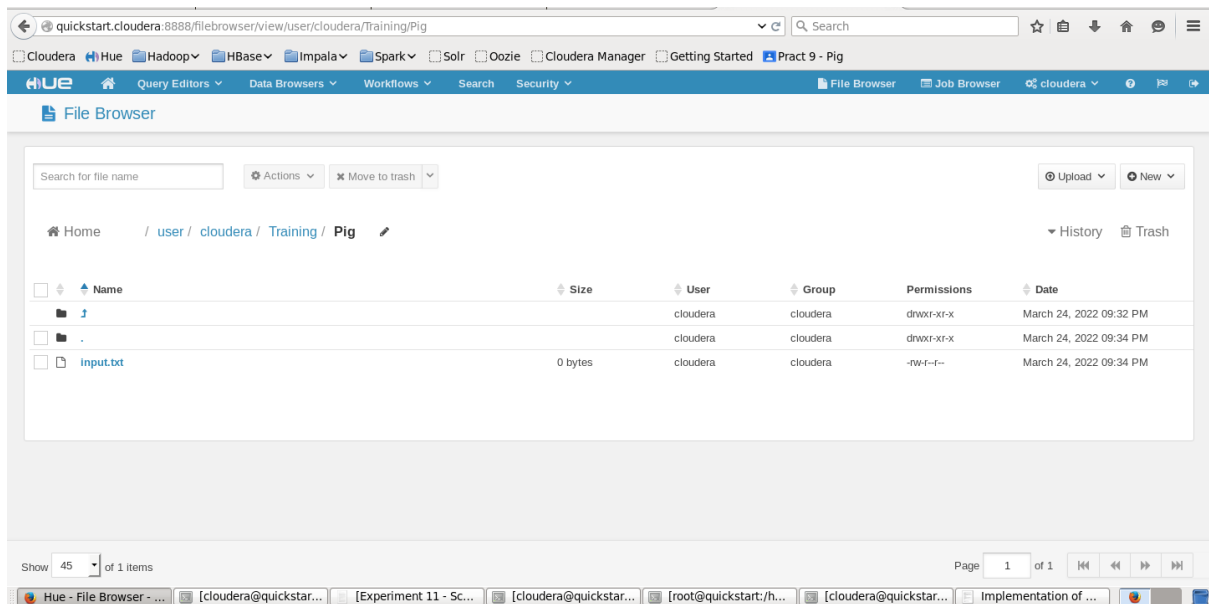
7) Creating input.txt file inside /usr/cloudera/Training/Pig directory

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

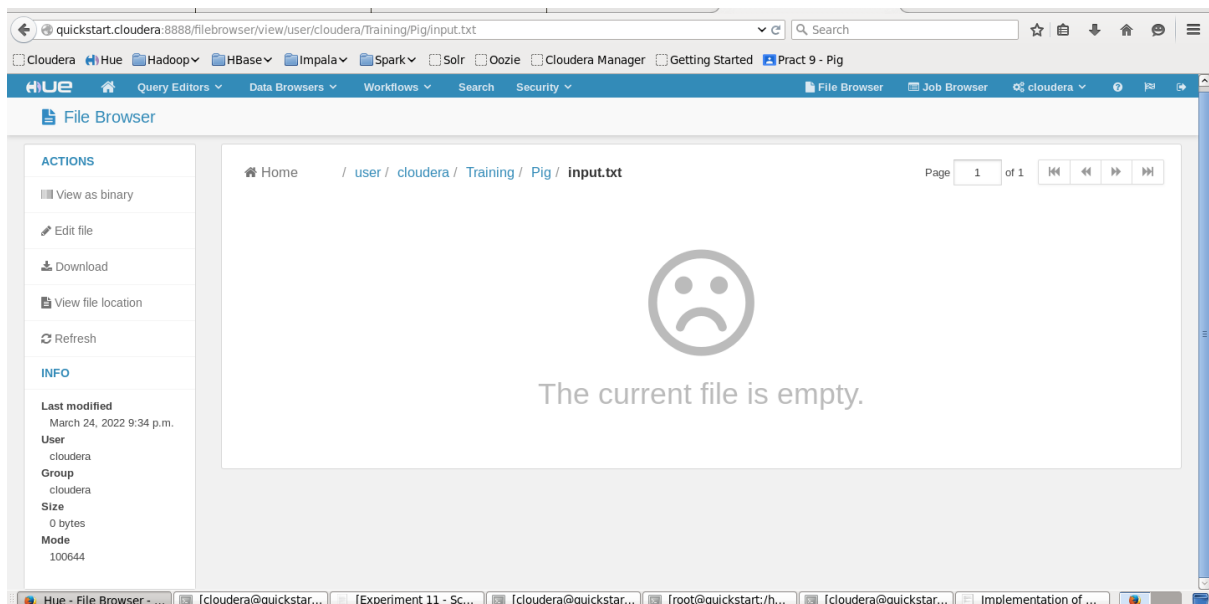


Once the file has been created click on 'input.txt' to add the content in it

Name : - Zen Dsouza
Roll No: - 34
Subject - BDT



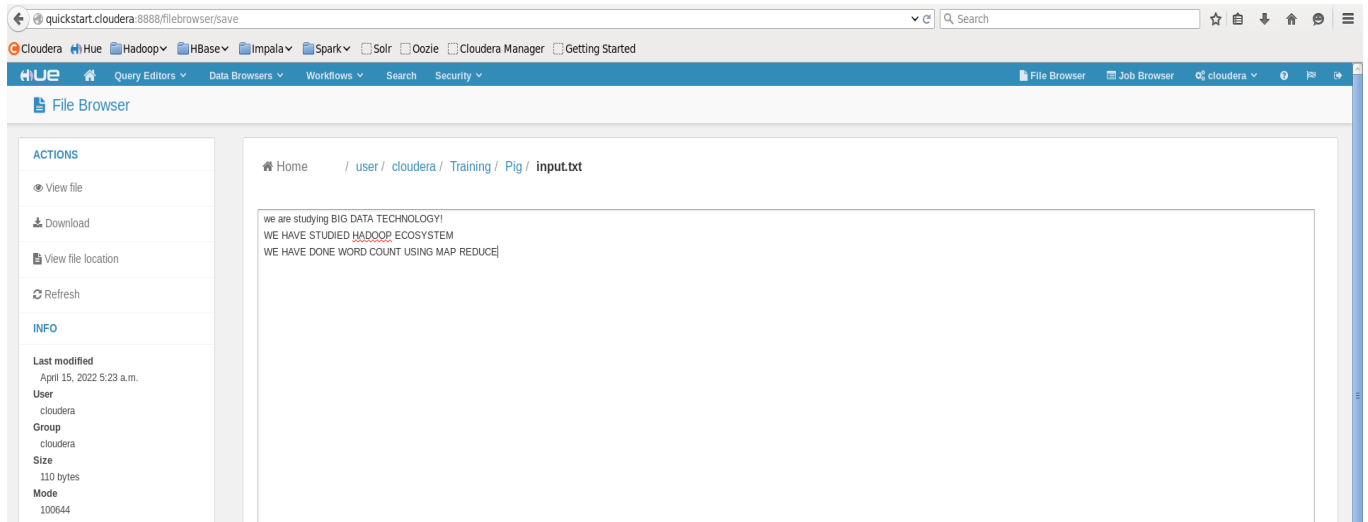
8) Adding some contents to this input.txt file.



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

Save the input.txt file

Name : - Zen Dsouza
Roll No: - 34
Subject - BDT



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

Name : - Zen Dsouza

Roll No: - 34

Subject - BDT

```
[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 messages 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
```

Name : - Zen Dsouza

Roll No: - 34

Subject - BDT

```
2022-03-24 21:27:33,755 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Instead, use fs.defaultFS
2022-03-24 21:27:33,761 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - mapred.job.tracker is deprecated. Instead, use mapreduce.jobtracker.address
2022-03-24 21:27:33,832 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Instead, use fs.defaultFS
2022-03-24 21:27:33,839 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - mapred.job.tracker is deprecated. Instead, use mapreduce.jobtracker.address
2022-03-24 21:27:33,911 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Instead, use fs.defaultFS
2022-03-24 21:27:33,911 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - mapred.job.tracker is deprecated. Instead, use mapreduce.jobtracker.address
2022-03-24 21:27:33,956 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Instead, use fs.defaultFS
2022-03-24 21:27:33,957 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - mapred.job.tracker is deprecated. Instead, use mapreduce.jobtracker.address
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, LimitInserter, LoadTypeCastInserter, MergeFilter, MergeForEach, NewPartitionFilterOptimizer, PushDownForEachFlatten, PushUpFilter, SplitFilter, StreamTypeCastInserter], RULES_DISABLED=[FilterLogicExpressionSimplifier, PartitionFilterOptimizer]]
2022-03-24 21:34:49,875 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MRCompiler - File concatenation threshold: 180 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.NMProxy - 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 21:34:50,150 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.JobControlCompiler - creating jar file Job3515548187496591125.jar
2022-03-24 21:34:53,053 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.JobControlCompiler - jar file Job3515548187496591125.jar created
2022-03-24 21:34:53,060 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.JobControlCompiler - Setting up single store job
2022-03-24 21:34:53,060 [main] INFO org.apache.pig.data.SchemaTupleFrontend - Key [pig.schematuple] is false, will not generate code
2022-03-24 21:34:53,060 [main] INFO org.apache.pig.data.SchemaTupleFrontend - Starting process to move generated code to distributed cache
2022-03-24 21:34:53,060 [main] INFO org.apache.pig.data.SchemaTupleFrontend - Setting key [pig.schematuple.classes] with classes to deserialize []
2022-03-24 21:34:53,067 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - 1 map-reduce job(s) waiting for submission.
2022-03-24 21:34:53,067 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - mapred.job.tracker is deprecated. Instead, use mapreduce.jobtracker.address
2022-03-24 21:34:53,068 [JobControl] INFO org.apache.hadoop.yarn.client.NMProxy - Connecting to ResourceManager at /0.0.0.0:8032
2022-03-24 21:34:53,074 [JobControl] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - 1 map-reduce job(s) waiting for submission.
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 21:34:53,281 [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 21:34:53,799 [JobControl] INFO org.apache.hadoop.mapreduce.JobSubmitter - Submitting tokens for job: job_1644548343526_0028
2022-03-24 21:34:53,981 [JobControl] INFO org.apache.hadoop.yarn.client.api.impl.YarnClientImpl - Submitted application application_1644548343526_0028
```

Subject - BDT

```

2022-03-24 21:34:53.331 [JobControl] INFO org.apache.hadoop.mapreduce.JobSubmitter - number of splits:1
2022-03-24 21:34:53.799 [JobControl] INFO org.apache.hadoop.mapreduce.JobSubmitter - Submitting tokens for job: job_1644548343526_0028
2022-03-24 21:34:53.581 [JobControl] INFO org.apache.hadoop.yarn.client.api.impl.YarnClientImpl - Submitted application application_1644548343526_0028
2022-03-24 21:34:54.049 [JobControl] INFO org.apache.hadoop.mapreduce.Job - The url to track the job: http://quickstart.cloudera:8080/proxy/application_1644548343526_0028/
2022-03-24 21:34:54.049 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - HadoopJobId: job_1644548343526_0028
2022-03-24 21:34:54.049 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - Processing aliases Input1
2022-03-24 21:34:54.049 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - detailed locations: W: Input1[3,9],Input1[-1,-1] C: R:
2022-03-24 21:34:54.049 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - More information at: http://localhost:50030/jobdetails.jsp?jobid=job_1644548343526_0028
2022-03-24 21:34:54.098 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - 0% complete
2022-03-24 21:35:12.295 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - 50% complete
2022-03-24 21:35:19.580 [main] INFO org.apache.hadoop.mapred.clientServiceDelegate - Application state is completed. FinalApplicationStatus=SUCCEEDED. Redirecting to job history server
2022-03-24 21:35:20.906 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - mapred.reduce.tasks is deprecated. Instead, use mapreduce.job.reduces
2022-03-24 21:35:20.963 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - 100% complete
2022-03-24 21:35:20.964 [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 2022-03-24 21:34:49 2022-03-24 21:35:28 UNKNOWN

Success!

Job Stats (time in seconds):
JobName Reduces MaxMapTime MinMapTime AvgMapTime MedianMapTime MaxReduceTime MinReduceTime AvgReduceTime MinReduceTime Alias Feature Outputs
job_1644548343526_0028 1 0 6 6 6 6 n/a n/a n/a n/a Input1 MAP_ONLY hdfs://quickstart.cloudera:8020/tmp/tmp-669075149/tmp-413572416,

Input(s):
Successfully read 3 records (497 bytes) from: "/user/cloudera/Training/pig/input.txt"

Output(s):
Successfully stored 3 records (129 bytes) in: "hdfs://quickstart.cloudera:8020/tmp/tmp-669075149/tmp-413572416"

Counters:
Total records written : 3
Total bytes written : 129
Spillable Memory Manager spill count : 0
Total bags proactively spilled: 0
Total records proactively spilled: 0

Job DAG:
job_1644548343526_0028

2022-03-24 21:35:21.051 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - Success!
2022-03-24 21:35:21.052 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Instead, use fs.defaultFS
2022-03-24 21:35:21.052 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - mapred.job.tracker is deprecated. Instead, use mapreduce.job.tracker.address
2022-03-24 21:35:21.052 [main] INFO org.apache.pig.data.SchemaTupleBackend - Key [pig.schematuple] was not set... will not generate code.
2022-03-24 21:35:21.065 [main] INFO org.apache.hadoop.mapreduce.lib.input.FileInputFormat - Total input paths to process : 1
2022-03-24 21:35:21.065 [main] INFO org.apache.pig.backend.hadoop.executionengine.util.MapRedUtil - Total input paths to process : 1
(we are studying BIG DATA TECHNOLOGY!)
(we HAVE STORED HADOOP ECOSYSTEM)
(HAVE DONE WORD COUNT USING MAP REDUCE)
grunt> █

```

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

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

```
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>
```

13) Again, we are dumping the data. It will do the MapReduce task.
dump wordsInEachLine;

```

2022-03-25 22:09:29,672 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Instead, use fs.defaults
2022-03-25 22:09:29,672 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - mapred.job.tracker is deprecated. Instead, use mapreduce.jobtracker.address
2022-03-25 22:09:29,672 [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.ScriptState - Pig features used in the script: UNKNOWN
2022-03-25 22:10:40,971 [main] INFO org.apache.hadoop.mapreduce.lib.input.FileInputFormat - Input paths are added to the job
2022-03-25 22:10:40,971 [main] INFO org.apache.hadoop.mapreduce.lib.output.FileOutputFormat - Output paths are added to the job
2022-03-25 22:10:40,971 [main] INFO org.apache.hadoop.mapreduce.lib.partition.PushdownOptimizers - PushdownOptimizers: FilterLogicExpressionSplitFilter, PartitionFilterOptimizer
2022-03-25 22:10:40,971 [main] INFO org.apache.hadoop.mapreduce.lib.partition.PushdownOptimizers - PushdownOptimizers: FilterLogicExpressionSplitFilter, PartitionFilterOptimizer
2022-03-25 22:10:41,814 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapreducecluster.MRCompiler - File concatenation threshold: 10 minutes? false
2022-03-25 22:10:41,816 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapreducecluster.MultiQueryOptimizer - MR plan size before optimization: 1
2022-03-25 22:10:41,816 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapreducecluster.MultiQueryOptimizer - MR plan size after optimization: 1
2022-03-25 22:10:41,862 [main] INFO org.apache.hadoop.yarn.client.NMProxy - Connecting to ResourceManager at /0.0.0.0:8032
2022-03-25 22:10:41,862 [main] INFO org.apache.pig.tools.pigstats.ScriptState - Pig script settings are added to the job
2022-03-25 22:10:41,864 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapreducecluster.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.mapreducecluster.JobControlCompiler - creating jar file Job04530875890545991.jar
2022-03-25 22:10:42,125 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapreducecluster.JobControlCompiler - Jar file Job04530875890545991.jar created
2022-03-25 22:10:42,326 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapreducecluster.JobControlCompiler - setting up single store job
2022-03-25 22:10:42,537 [main] INFO org.apache.pig.data.SchemaTupleFrontend - Key [pig.schematuple] is false, will not generate code
2022-03-25 22:10:42,537 [main] INFO org.apache.pig.data.SchemaTupleFrontend - Starting process to move generated code to distributed cache
2022-03-25 22:10:42,537 [main] INFO org.apache.pig.data.SchemaTupleFrontend - Setting key [pig.schematuple.classes] with classes to deserialize []
2022-03-25 22:10:42,546 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapreducecluster.Launcher - 1 map-reduce jobs(s) waiting for submission.
2022-03-25 22:10:42,546 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - mapred.job.tracker is deprecated. Instead, use mapreduce.jobtracker.address
2022-03-25 22:10:42,550 [JobControl] INFO org.apache.hadoop.yarn.client.NMProxy - Connecting to ResourceManager at /0.0.0.0:8032
2022-03-25 22:10:42,566 [JobControl] INFO org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Instead, use fs.defaults
2022-03-25 22:10:42,566 [JobControl] INFO org.apache.hadoop.conf.Configuration.deprecation - mapred.job.tracker is deprecated. Instead, use mapreduce.jobtracker.address
2022-03-25 22:10:42,576 [JobControl] INFO org.apache.pig.backend.hadoop.executionengine.util.MapRedUtil - Total input paths to process : 1
2022-03-25 22:10:42,578 [JobControl] INFO org.apache.pig.backend.hadoop.executionengine.util.MapRedUtil - Total input paths (combined) to process : 1
2022-03-25 22:10:42,583 [JobControl] INFO org.apache.pig.backend.hadoop.executionengine.util.MapRedUtil - Total number of splits: 1
2022-03-25 22:10:42,587 [JobControl] INFO org.apache.hadoop.mapreduce.JobSubmitter - Submitting tokens for job: job_1644548343526_0029
2022-03-25 22:10:42,598 [JobControl] INFO org.apache.hadoop.yarn.client.impl.YarnClientImpl - Submitted application application_1644548343526_0029
2022-03-25 22:10:42,598 [JobControl] INFO org.apache.hadoop.mapreduce.lib.input.FileInputFormat - The url to track job: http://cuihaijun@hahaiprivy/application_1644548343526_0029/
2022-03-25 22:10:42,606 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapreducecluster.Launcher - HadoopJobInfo: job_1644548343526_0029
2022-03-25 22:10:42,646 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapreducecluster.Launcher - Processing aliases Input_wordsInEachLine
2022-03-25 22:10:42,646 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapreducecluster.Launcher - Detailed locations: %t, Topic [%d], wordsInEachLine [%d]
2022-03-25 22:10:42,646 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapreducecluster.Launcher - More information at: http://localhost:38080/hbase[_details=-1,-1] c: b: r:
2022-03-25 22:10:42,646 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapreducecluster.Launcher - % complete
2022-03-25 22:11:10,190 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapreducecluster.Launcher - 50% complete
2022-03-25 22:11:10,190 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapreducecluster.Launcher - 100% complete
2022-03-25 22:11:11,459 [main] INFO org.apache.pig.tools.pigstats.SimplePigStats - Script Statistics:
2022-03-25 22:11:11,459 [main] INFO org.apache.pig.tools.pigstats.SimplePigStats - Script Statistics:

```


Name : - Zen Dsouza
Roll No: - 34
Subject - BDT

```
Job Stats (time in seconds):
JobId  Maps  Reduces  MaxMapTime  MinMapTime  AvgMapTime  MedianMapTime  MaxReduceTime  MinReduceTime  AvgReduceTime  MedianReduceTime  Alias  Feature Outputs
job_1644548343526_0029  1  0  5  5  5  5  n/a  n/a  n/a  n/a  hdfs://quickstart.cloudera:8020/tmp/temp-669075149/tmp-341911088,

Input(s):
Successfully read 3 records (497 bytes) from: "/user/cloudera/Training/pig/input.txt"

Output(s):
Successfully stored 19 records (225 bytes) in: "hdfs://quickstart.cloudera:8020/tmp/temp-669075149/tmp-341911088"

Counters:
Total records written : 19
Total bytes written : 225
Spillable Memory Manager spill count : 0
Total bags proactively spilled: 0
Total records proactively spilled: 0

Job DAG:
job_1644548343526_0029

2022-03-25 22:11:51.537 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - Success!
2022-03-25 22:11:51.537 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Instead, use fs.defaultFS
2022-03-25 22:11:51.537 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - mapred.job.tracker is deprecated. Instead, use mapreduce.jobtracker.address
2022-03-25 22:11:51.538 [main] INFO org.apache.pig.data.SchemaTupleBackend - Key [pig.schematuple] was not set., will not generate code.
2022-03-25 22:11:51.543 [main] INFO org.apache.hadoop.mapreduce.lib.input.FileInputFormat - Total input paths to process : 1
2022-03-25 22:11:51.543 [main] INFO org.apache.pig.backend.hadoop.executionengine.util.MapReduceUtil - Total input paths to process : 1
(we)
(are)
(studying)
(BIG)
(DATA)
(TECHNOLOGY)
(WE)
(HAVE)
(STUDIED)
(HADOOP)
(ECOSYSTEM)
(WE)
(HAVE)
(DONE)
(WORD)
(COUNT)
(USING)
(MAP)
(REDUCE)
grunt>
```

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

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

And then dumping the data by the following command.
dump groupedWords;

```
grunt> dump groupedWords;
2022-03-25 22:12:59.684 [main] INFO org.apache.pig.tools.pigstats.ScriptState - Pig features used in the script: GROUP BY
2022-03-25 22:12:59.685 [main] INFO org.apache.pig.newplan.logical.optimizer.LogicalPlanOptimizer - RULES_ENABLED=[AddForEach, ColumnMapKeyPrune, DuplicateForEachColumnWrite, GroupByConstParallelSetter, ImplicitSplitInserter, LimitInserter, LoadTypeCastInserter, MergeFilter, MergeJoin, MergeJoinFilterOptimizer, PushDownForEachFlatten, PushDownFilter, SplitFilter, StreamTypeCastInserter], RULES_DISABLED=[FilterLogicExpressionsSimplifier, PartitionFilterOptimizer]
2022-03-25 22:12:59.782 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MRCompiler - File concatenation threshold: 100 optimistic? false
2022-03-25 22:12:59.786 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MultiQueryOptimizer - MR plan size before optimization: 1
2022-03-25 22:12:59.786 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MultiQueryOptimizer - MR plan size after optimization: 1
2022-03-25 22:12:59.736 [main] INFO org.apache.hadoop.yarn.client.RMRProxy - Connecting to ResourceManager at /0.0.0.0:8032
2022-03-25 22:12:59.738 [main] INFO org.apache.pig.tools.pigstats.ScriptState - Pig script settings are added to the job
2022-03-25 22:12:59.750 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.JobControlCompiler - mapred.job.reduce.markreset.buffer.percent is not set, set to default 0.3
2022-03-25 22:12:59.750 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.JobControlCompiler - Reduce phase detected, estimating # of required reducers.
2022-03-25 22:12:59.752 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.JobControlCompiler - Using reducer estimator: org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.InputSizeReducerEstimator
2022-03-25 22:12:59.757 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.InputSizeReducerEstimator - BytesPerReducer=1000000000 maxReducers=999 totalInputFileSize=110
2022-03-25 22:12:59.757 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.JobControlCompiler - Setting Parallelism to 1
2022-03-25 22:13:00.128 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.JobControlCompiler - creating jar file Job816439686515978146.jar
2022-03-25 22:13:01.587 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.JobControlCompiler - Jar file Job816439686515978146.jar created
2022-03-25 22:13:03.573 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.JobControlCompiler - Setting up single store job
2022-03-25 22:13:03.574 [main] INFO org.apache.pig.data.SchemaTupleFrontend - Key [pig.schematuple] is false, will not generate code.
2022-03-25 22:13:03.574 [main] INFO org.apache.pig.data.SchemaTupleFrontend - Starting process to move generated code to distributed cache
2022-03-25 22:13:03.676 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - 1 map-reduce job(s) waiting for submission.
2022-03-25 22:13:03.676 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - mapred.job.tracker is deprecated. Instead, use mapreduce.jobtracker.address
2022-03-25 22:13:03.677 [JobControl] INFO org.apache.hadoop.yarn.client.RMRProxy - Connecting to ResourceManager at /0.0.0.0:8032
2022-03-25 22:13:03.687 [JobControl] INFO org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Instead, use fs.defaultFS
2022-03-25 22:13:03.825 [JobControl] INFO org.apache.pig.backend.hadoop.executionengine.util.MapReduceUtil - Total input paths to process : 1
2022-03-25 22:13:03.829 [JobControl] INFO org.apache.pig.backend.hadoop.executionengine.util.MapReduceUtil - Total input paths (combined) to process : 1
2022-03-25 22:13:03.858 [JobControl] INFO org.apache.hadoop.mapreduce.lib.input.FileInputFormat - Total input paths to process : 1
2022-03-25 22:13:03.985 [JobControl] INFO org.apache.hadoop.mapreduce.lib.output.JobSubmitter - Submitting tokens for job: job_1644548343526_0030
2022-03-25 22:13:03.951 [JobControl] INFO org.apache.hadoop.yarn.client.impl.YarnClientImpl - Submitted application application_1644548343526_0030
2022-03-25 22:13:03.953 [JobControl] INFO org.apache.hadoop.mapreduce.Job - The url to track the job: http://quickstart.cloudera:8030/proxy/application_1644548343526_0030/
2022-03-25 22:13:04.176 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - HadoopJobId: job_1644548343526_0030
2022-03-25 22:13:04.176 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - Processing aliases Input1,groupedWords,wordsInEachLine
2022-03-25 22:13:04.176 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - detailed locations: W Input[1:9],wordsInEachLine[1:-1],groupedWords[5,15] C: R:
2022-03-25 22:13:04.176 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - More information at: http://localhost:50030/jobdetails.jsp?jobId=job_1644548343526_0030
2022-03-25 22:13:04.246 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - 0% complete
2022-03-25 22:13:04.489 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - 50% complete
2022-03-25 22:13:14.143 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - 100% complete
2022-03-25 22:13:14.144 [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 22:12:59  2022-03-25 22:13:14  GROUP BY

Success!

Job Stats (time in seconds):
JobId  Maps  Reduces  MaxMapTime  MinMapTime  AvgMapTime  MedianMapTime  MaxReduceTime  MinReduceTime  AvgReduceTime  MedianReduceTime  Alias  Feature Outputs
job_1644548343526_0030  1  1  5  5  5  5  6  6  6  6  Input1,groupedWords,wordsInEachLine  GROUP BY  hdfs://quickstart.cloudera:8020/tmp/temp-669075149/tmp325198341,
```

Subject - BDT

BDT

Name : - Zen Dsouza

Roll No: - 34

Subject - BDT

```
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/tn
01191620340
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.hadoop.mapreduce.lib.input.FileInputFormat - 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 ~]$ █
```