# CREATING AND EXECUTING PIG LATIN SCRIPT

## What is Pig in Hadoop?

Pig is a scripting platform that runs on Hadoop clusters designed to process and analyze large datasets. Pig is extensible, self-optimizing, and easily programmed.

Programmers can use Pig to write data transformations without knowing Java. Pig uses both structured and unstructured data as input to perform analytics and uses HDFS to store the results.

## Components of Pig

There are two major components of the Pig:

- Pig Latin script language
- A runtime engine

## Pig Latin script language:

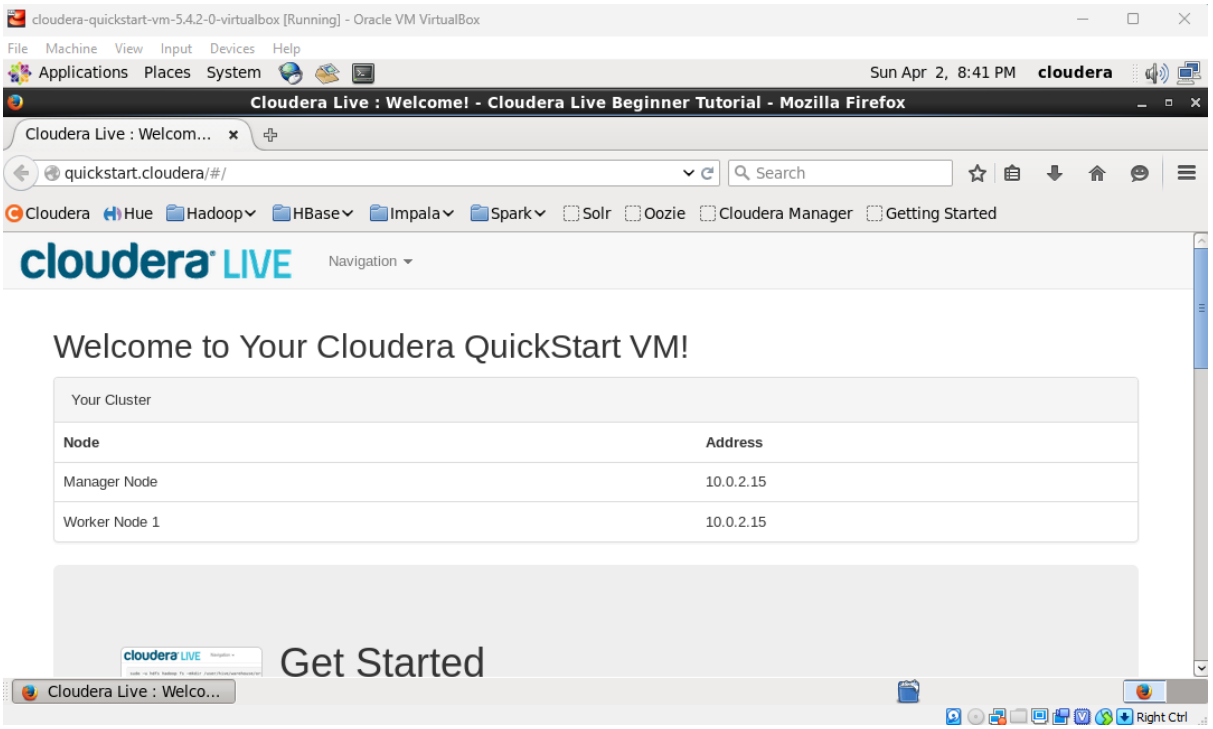
The Pig Latin script is a procedural data flow language. It contains syntax and commands that can be applied to implement business logic. Examples of Pig Latin are LOAD and STORE.

## A runtime engine:

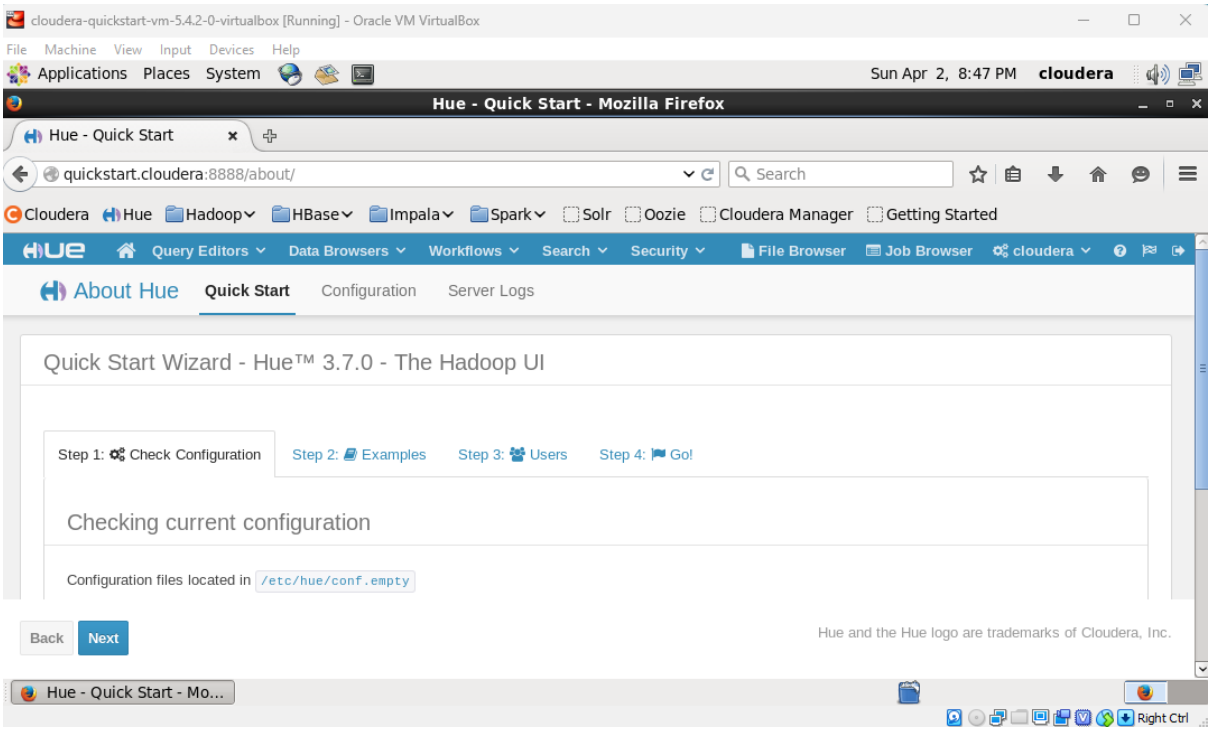
The runtime engine is a compiler that produces sequences of MapReduce programs. It uses HDFS to store and retrieve data. It is also used to interact with the Hadoop system (HDFS and MapReduce).

The runtime engine parses, validates, and compiles the script operations into a sequence of MapReduce jobs.

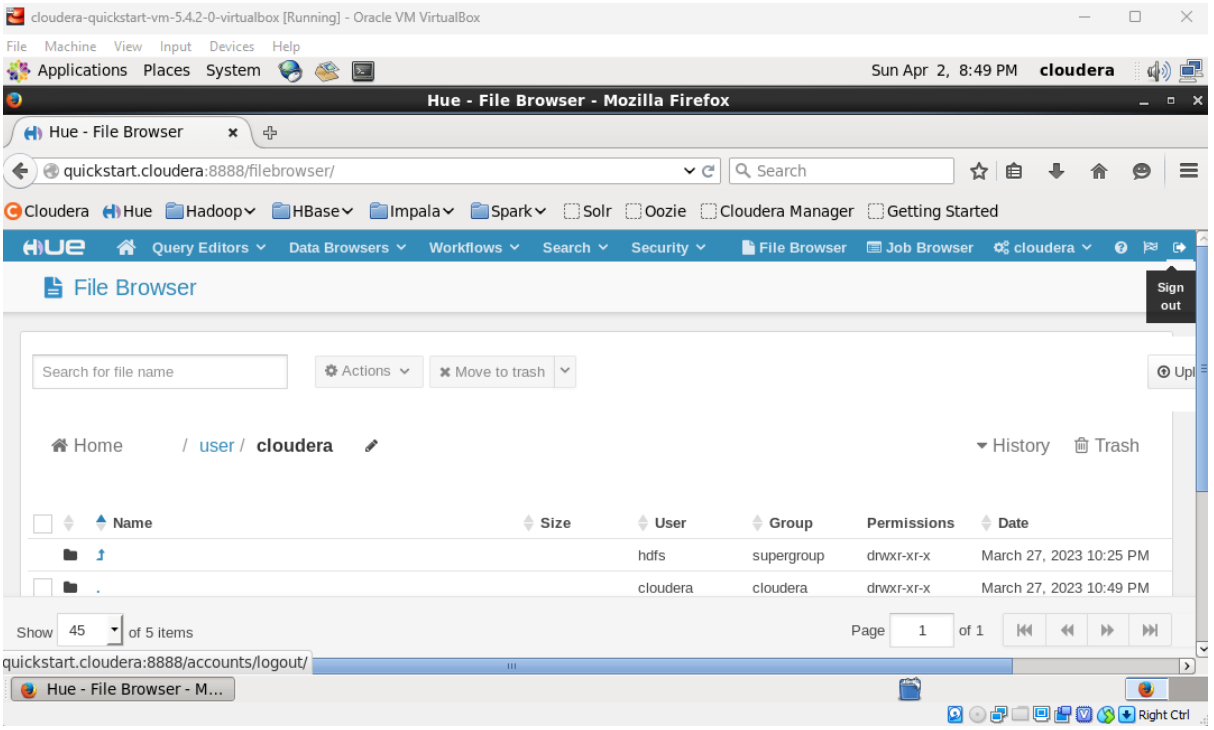
- **Open cloudera browser**



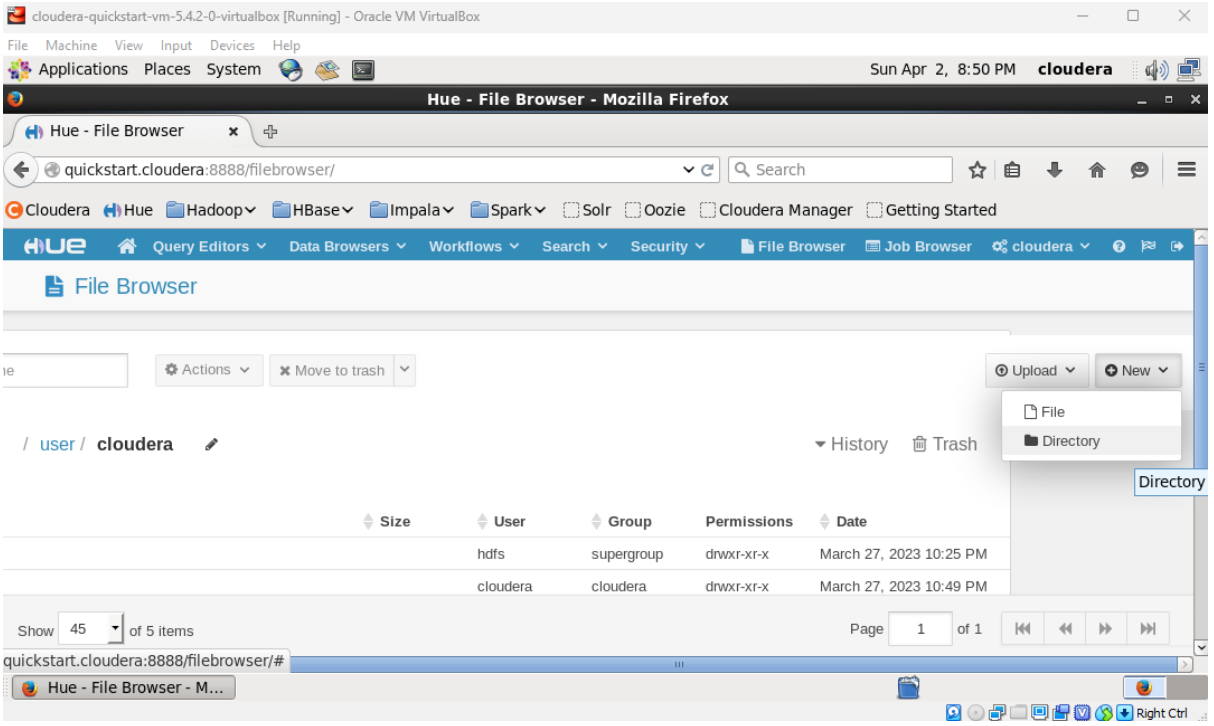

- **Locate Hue and get logged in with username cloudera and password cloudera.**

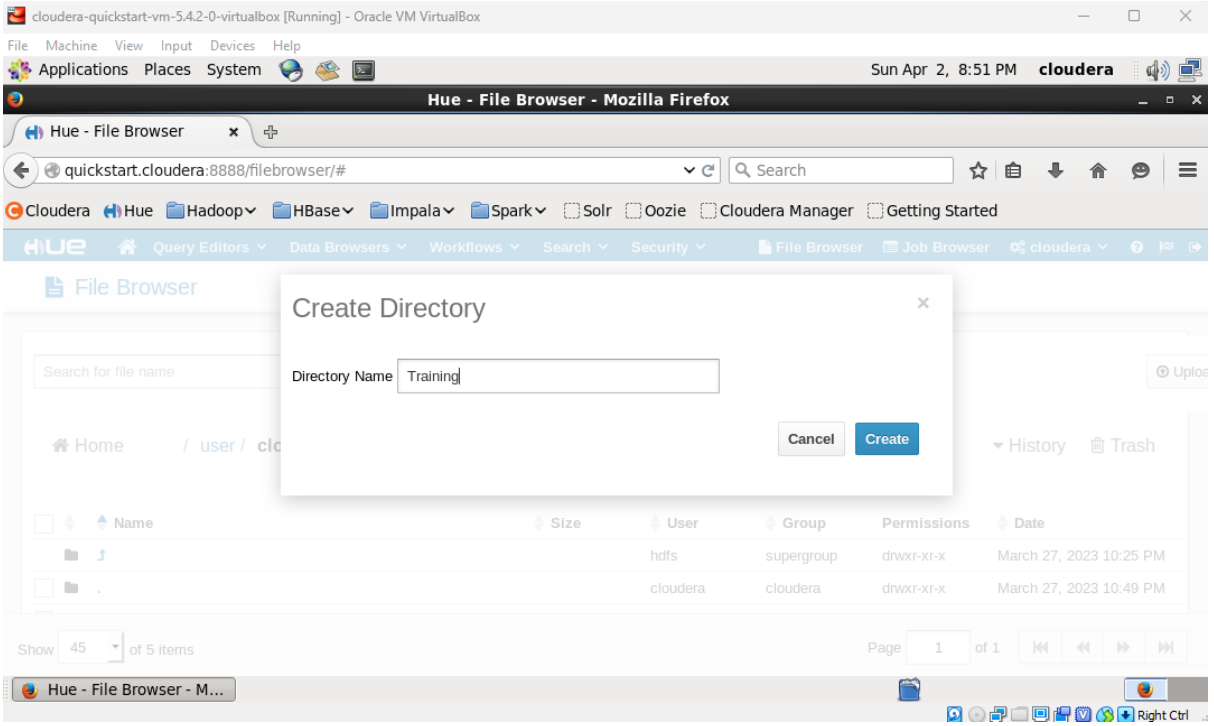


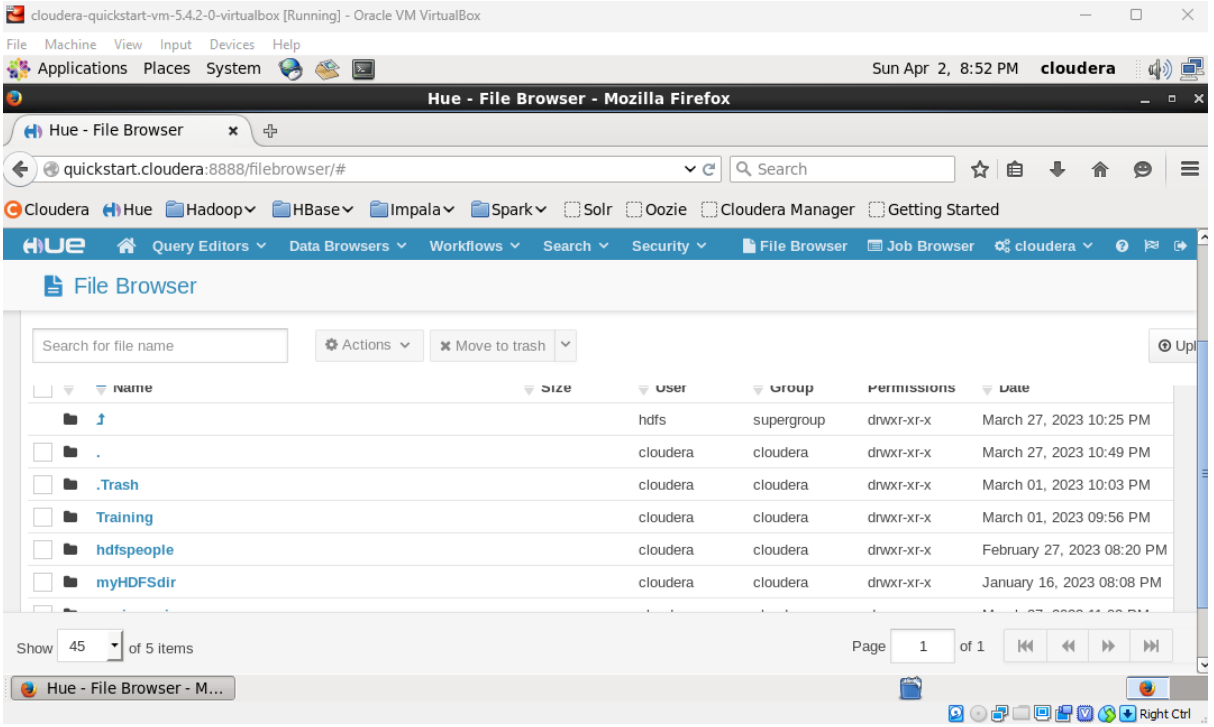
- **Locate file Browser and check for directory user/cloudera.**



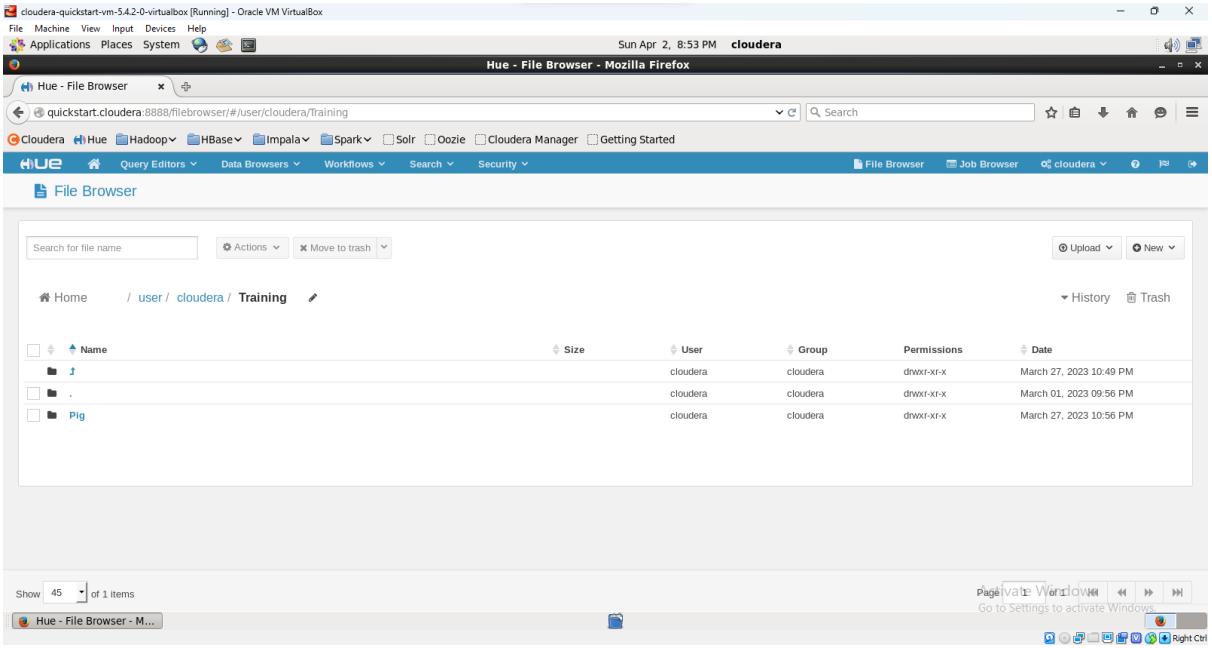

- **Create a new directory named 'training'.**



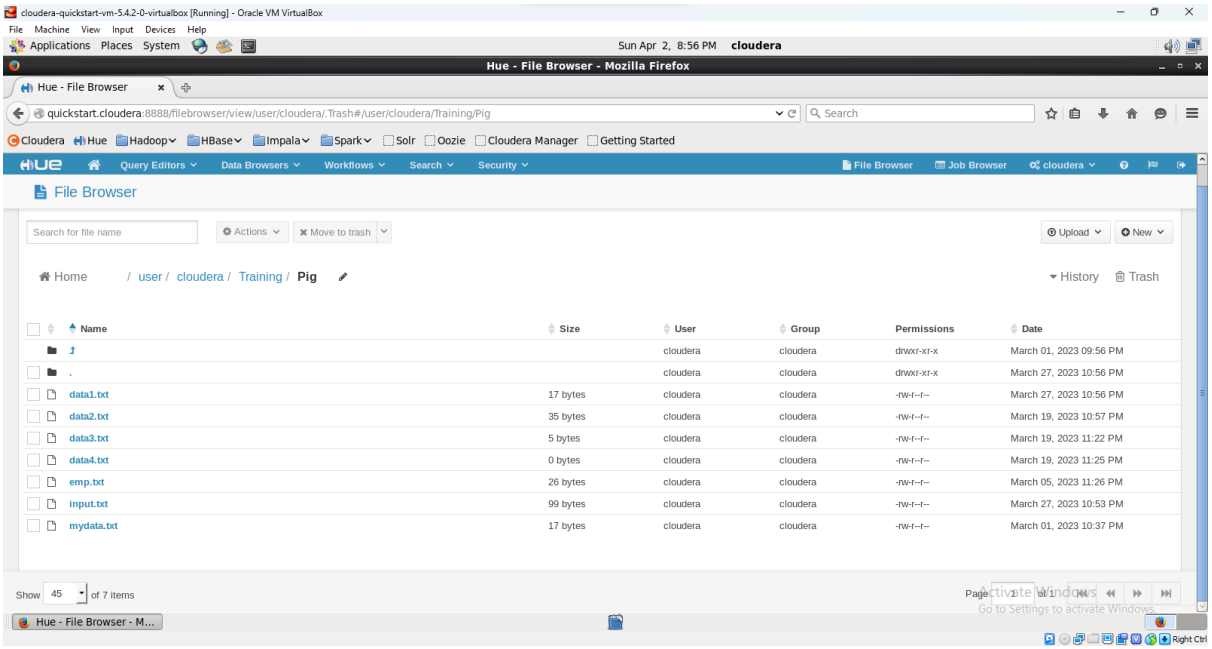

cloudera-quickstart-vm-5.4.2-0-virtualbox [Running] - Oracle VM VirtualBox
File Machine View Input Devices Help
Applications Places System
Sun Apr 2, 8:51 PM cloudera
Hue - File Browser - Mozilla Firefox
Hue - File Browser
quickstart.cloudera:8888/filebrowser/#
Search
Cloudera Hue Hadoop HBase Impala Spark Solr Oozie Cloudera Manager Getting Started
Query Editors Data Browsers Workflows Search Security File Browser Job Browser cloudera
File Browser
Create Directory
Directory Name Training
Cancel
Create
Search for file name
Home / user / clo
History Trash
Name Size User Group Permissions Date
hdfs supergroup drwxr-xr-x March 27, 2023 10:25 PM
cloudera cloudera drwxr-xr-x March 27, 2023 10:49 PM
Show 45 of 5 items
Page 1 of 1
Hue - File Browser - M...
Right Ctrl


cloudera-quickstart-vm-5.4.2-0-virtualbox [Running] - Oracle VM VirtualBox
File Machine View Input Devices Help
Applications Places System
Sun Apr 2, 8:52 PM cloudera
Hue - File Browser - Mozilla Firefox
Hue - File Browser
quickstart.cloudera:8888/filebrowser/#
Search
Cloudera Hue Hadoop HBase Impala Spark Solr Oozie Cloudera Manager Getting Started
Query Editors Data Browsers Workflows Search Security File Browser Job Browser cloudera
File Browser
Search for file name
Actions
Move to trash
Name Size User Group Permissions Date
hdfs supergroup drwxr-xr-x March 27, 2023 10:25 PM
. cloudera cloudera drwxr-xr-x March 27, 2023 10:49 PM
.Trash cloudera cloudera drwxr-xr-x March 01, 2023 10:03 PM
Training cloudera cloudera drwxr-xr-x March 01, 2023 09:56 PM
hdfspeople cloudera cloudera drwxr-xr-x February 27, 2023 08:20 PM
myHDFSdir cloudera cloudera drwxr-xr-x January 16, 2023 08:08 PM
Show 45 of 5 items
Page 1 of 1
Hue - File Browser - M...
Right Ctrl

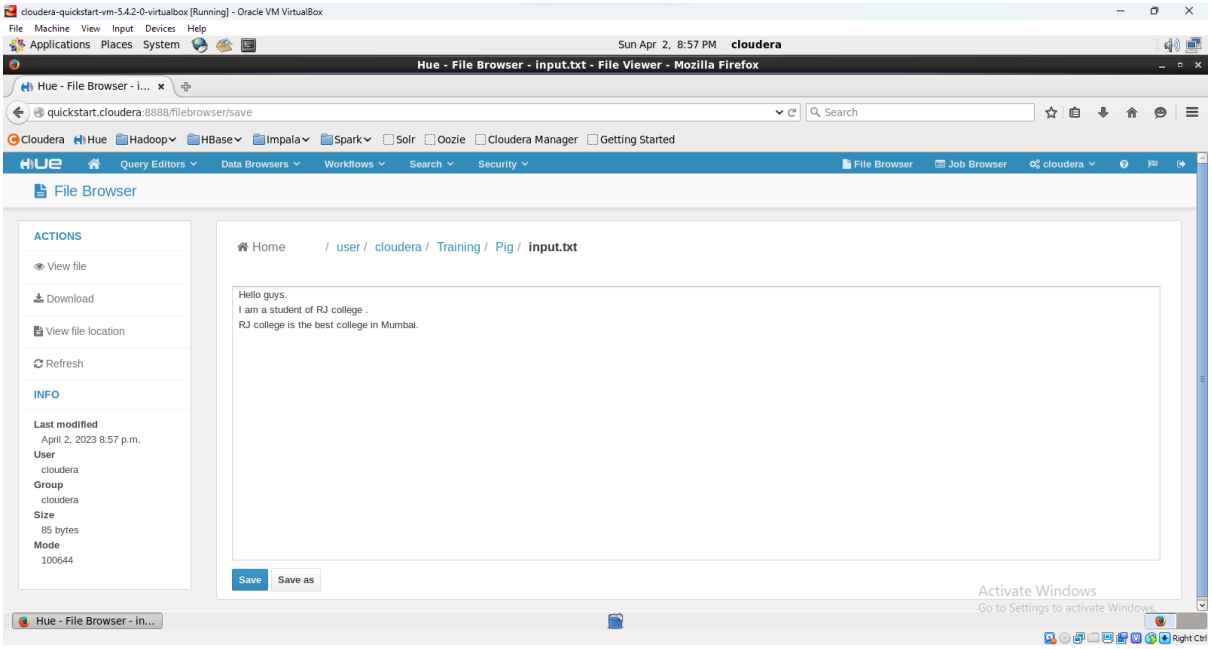
- **Create the pig directory inside the training directory.**



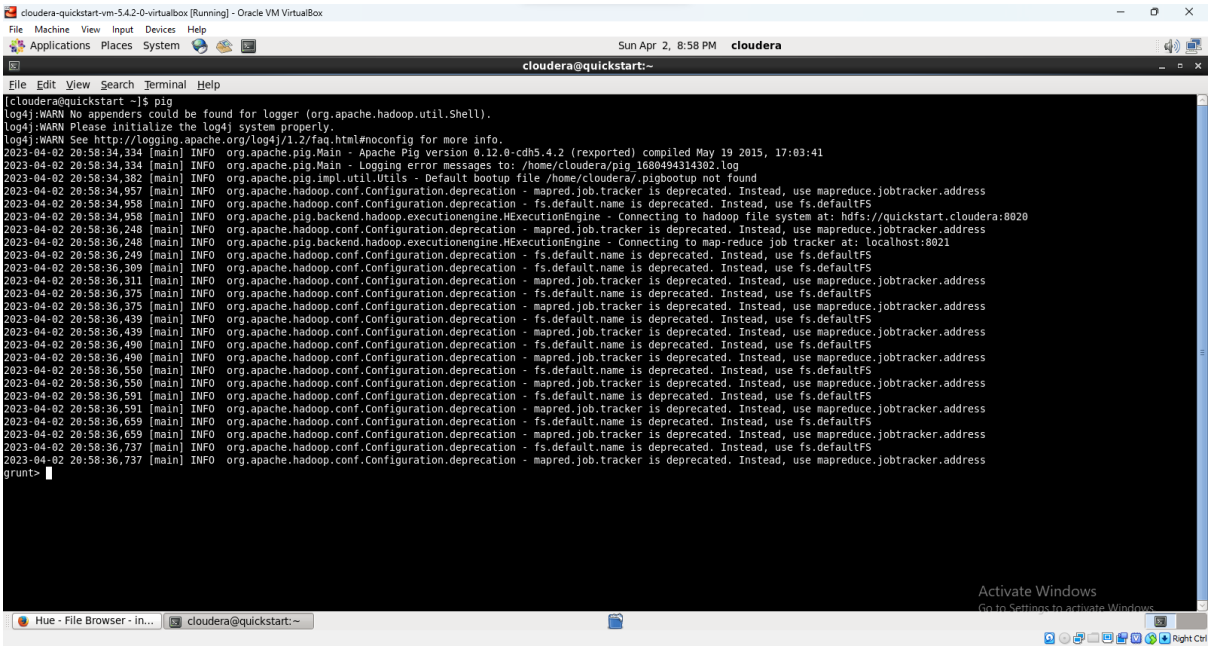

- **Create an input.txt file in the pig directory.**



- **Add some text into the input.txt and save the file.**




- **Open terminal and start pig tool by executing pig command.**

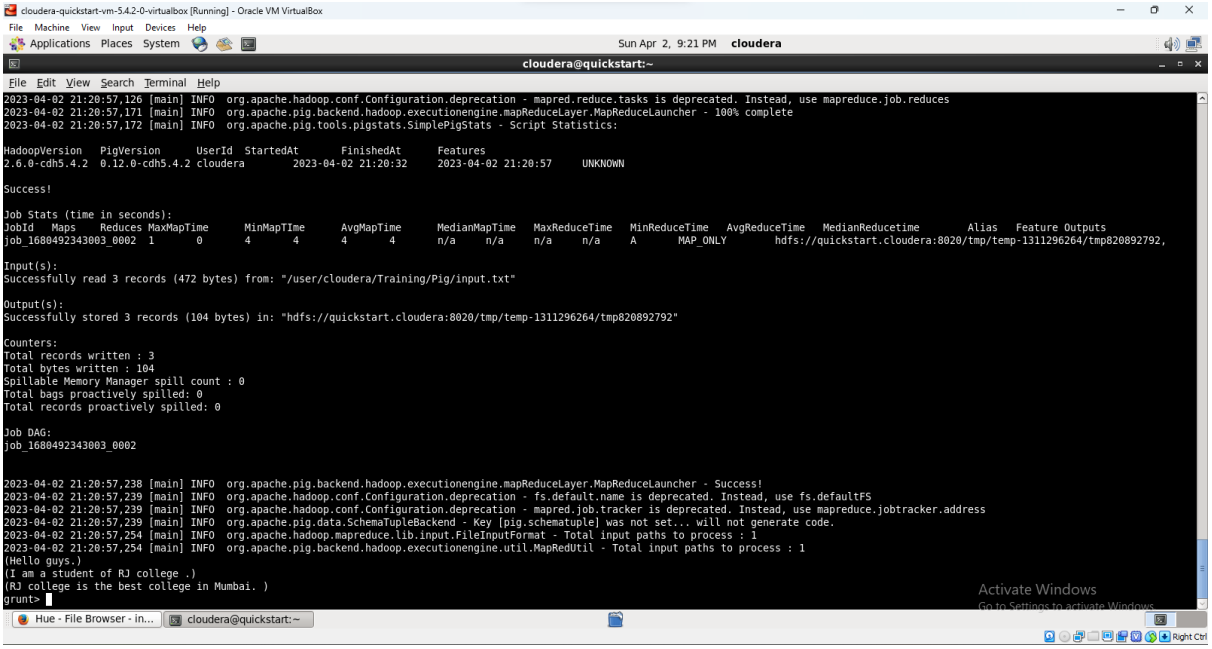



- **Load input.txt file in input variable.**

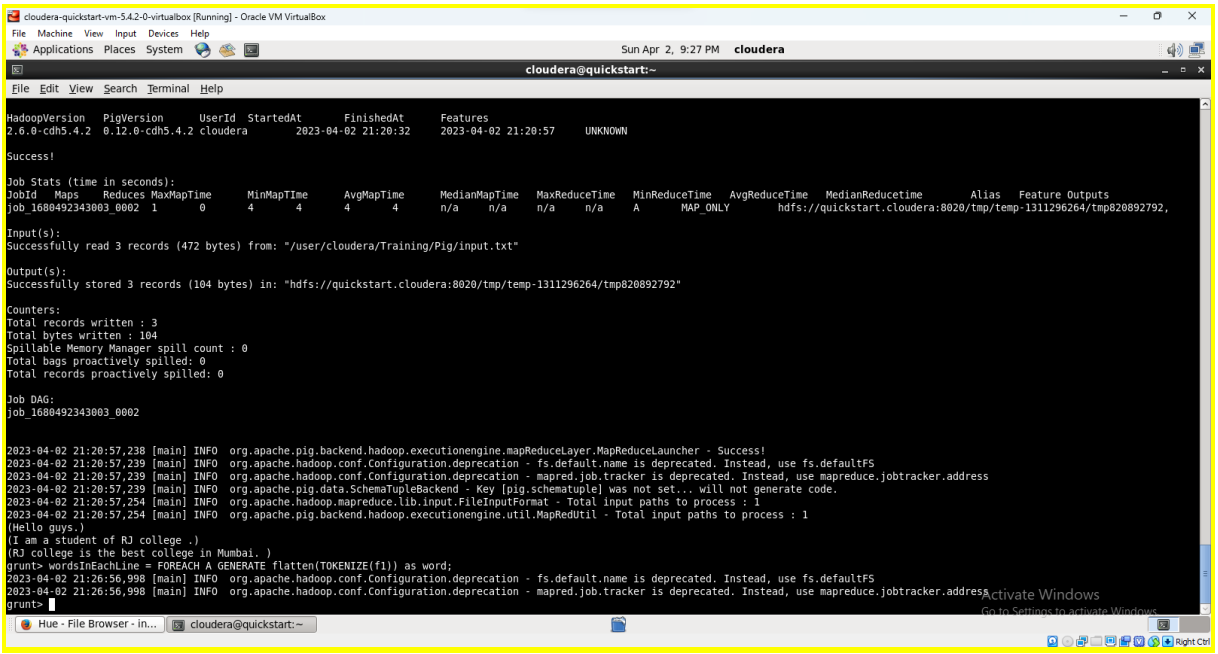
A = LOAD '/user/cloudera/Training/pig/input.txt' AS (f1:chararray);

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

- **Display the contents of the input variable.**
  **DUMP A**

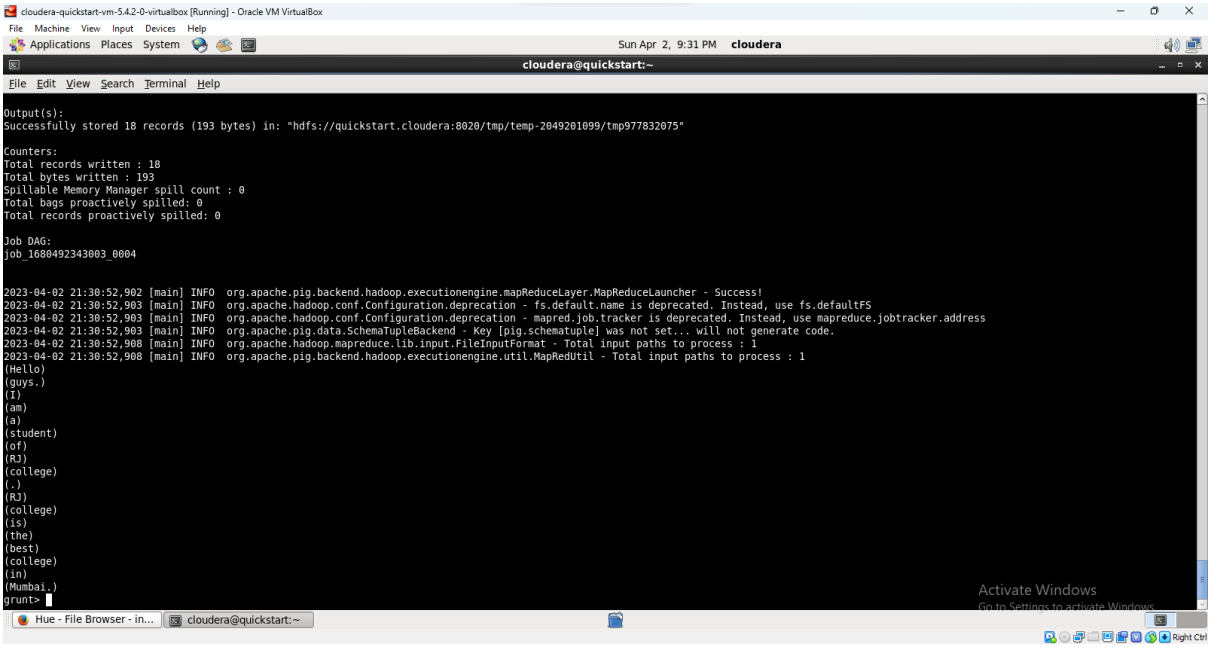



- Tokenize the text that is stored in variable input.
  wordsInEachLine = FOREACH A GENERATE flatten(TOKENIZE(f1)) as word;




dump wordsInEachLine;

```
grunt> DUMP wordsInEachLine;
```



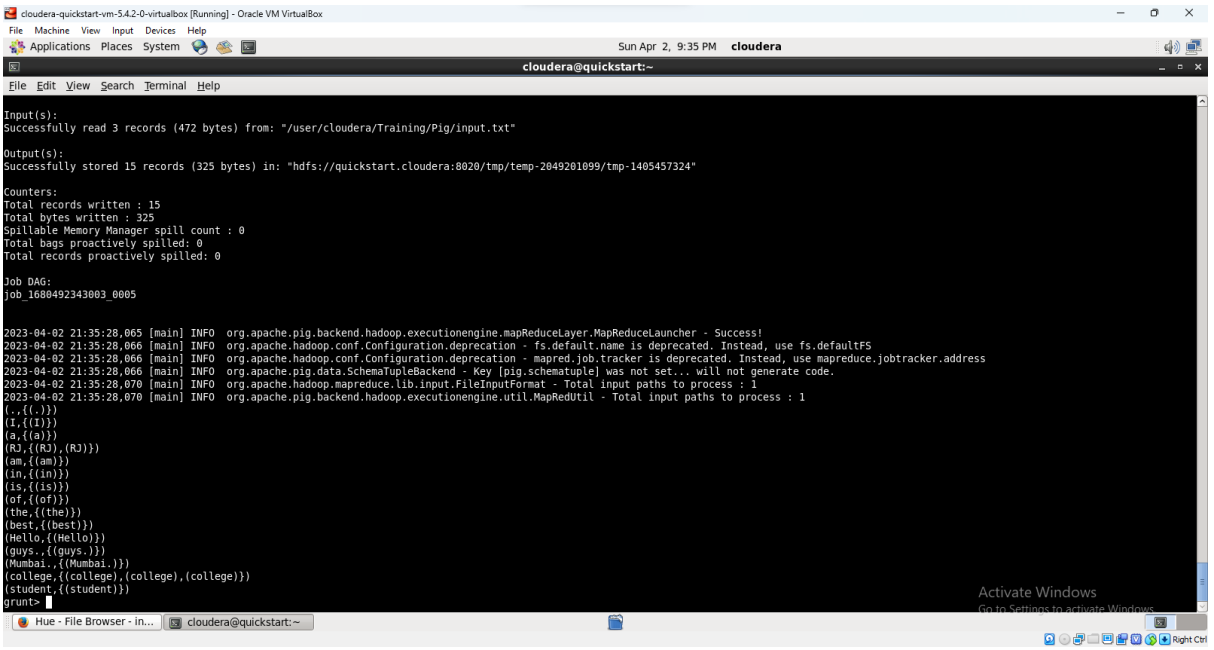

- Group all similar words.

groupedWords = group wordsInEachLine by word;

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

dump groupedWords;

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

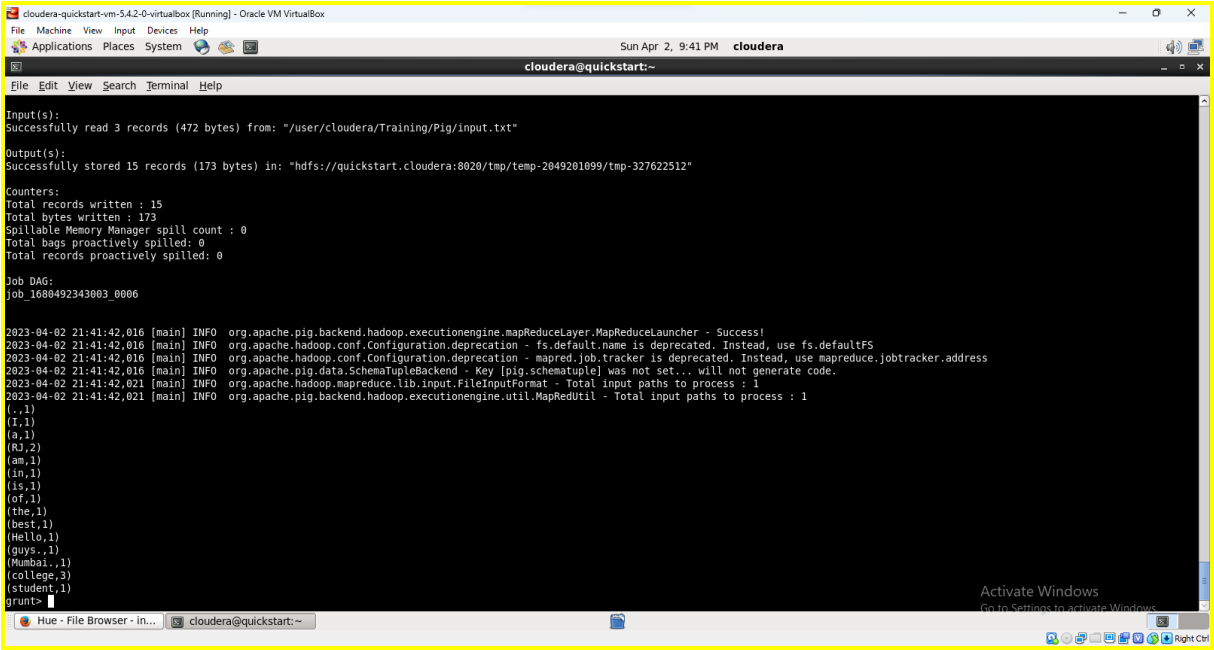


- Count the number of occurrences of each word.
  countedWords = FOREACH groupedWords GENERATE group, COUNT(wordsInEachLine);

```
grunt> countedWords = FOREACH groupedWords GENERATE group, COUNT(wordsInEachLine);
grunt>
```

dump countedWords;

```
grunt> countedWords = FOREACH groupedWords GENERATE group, COUNT(wordsInEachLine);
grunt> DUMP countedWords;
```




**Creating pig script**

```
/*
Wordcountex.pig
Counting the occurrences of word
*/
–Execute this script in mapreduce mode

words = LOAD '/user/cloudera/training/pig/input.txt' AS (line:chararray);
wordsInEachLine = FOREACH words GENERATE flatten(TOKENIZE(line)) as word;
groupedWords = group wordsInEachLine by word;
countedWords = foreach groupedWords generate group, COUNT(wordsInEachLine);
store countedWords into '/user/cloudera/training/pig/outputwordcount_1.txt' using PigStorage(',');
```

```
2023-04-14 00:08:06,545 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - 50% complete
2023-04-14 00:08:21,759 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - 100% complete
2023-04-14 00:08:21,766 [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        2023-04-14 00:07:37     2023-04-14 00:08:21     GROUP_BY

Success!

Job Stats (time in seconds):
JobId   Maps    Reduces MaxMapTime      MinMapTIme      AvgMapTime      MedianMapTime   MaxReduceTime   MinReduceTime   AvgReduceTime   MedianReducetime        Al
job_1681420822994_0002  1       1       6       6       6       6       7       7       7       7       countedWords,groupedWords,words,wordsInEachLine GROUP_BY,C

Input(s):
Successfully read 1 records (415 bytes) from: "/user/cloudera/training/pig/input.txt"

Output(s):
Successfully stored 6 records (40 bytes) in: "/user/cloudera/training/pig/outputwordcount_1.txt"

Counters:
Total records written : 6
Total bytes written : 40
Spillable Memory Manager spill count : 0
Total bags proactively spilled: 0
Total records proactively spilled: 0

Job DAG:
job_1681420822994_0002


2023-04-14 00:08:21,902 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - Success!
```

## Ex 2 : Loading matrix in a pig variable.

Mydata.txt

1 2 3
4 5 6
7 8 9

B=LOAD '/user/cloudera/Training/pig/mydata.txt' AS (c1:int,c2:int,c3:int);

```
grunt> B = LOAD '/user/cloudera/Training/Pig/data1.txt' AS (f1:chararray);
grunt>
```

DUMP B;

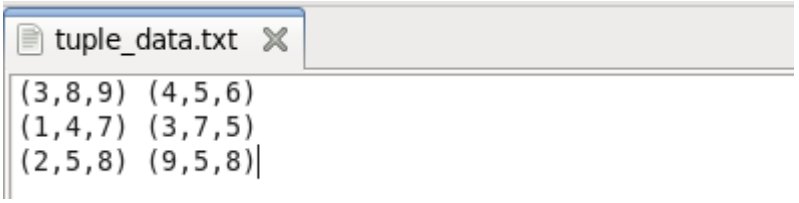
```
grunt> B = LOAD '/user/cloudera/Training/Pig/data1.txt' AS (f1:chararray);
grunt> DUMP B;
```

```
cloudera-quickstart-vm-5.4.2-0-virtualbox [Running] - Oracle VM VirtualBox
File  Machine  View  Input  Devices  Help
Applications  Places  System                    Sun Apr 2, 9:48 PM   cloudera
cloudera@quickstart:~
File  Edit  View  Search  Terminal  Help
2023-04-02 21:47:53,789 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - 50% complete
2023-04-02 21:47:58,012 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - 100% complete
2023-04-02 21:47:58,012 [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        2023-04-02 21:47:40     2023-04-02 21:47:58     UNKNOWN

Success!

Job Stats (time in seconds):
JobId   Maps    Reduces MaxMapTime      MinMapTIme      AvgMapTime      MedianMapTime   MaxReduceTime   MinReduceTime   AvgReduceTime   MedianReducetime        Alias   Feature Outputs
job_1680492343003_0007  1       0       3       3       3       3       n/a     n/a     n/a     n/a     B       MAP_ONLY        hdfs://quickstart.cloudera:8020/tmp/temp-2049201099/tmp443498406,

Input(s):
Successfully read 3 records (404 bytes) from: "/user/cloudera/Training/Pig/data1.txt"

Output(s):
Successfully stored 3 records (36 bytes) in: "hdfs://quickstart.cloudera:8020/tmp/temp-2049201099/tmp443498406"

Counters:
Total records written : 3
Total bytes written : 36
Spillable Memory Manager spill count : 0
Total bags proactively spilled: 0
Total records proactively spilled: 0

Job DAG:
job_1680492343003_0007


2023-04-02 21:47:58,074 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - Success!
2023-04-02 21:47:58,074 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Instead, use fs.defaultFS
2023-04-02 21:47:58,074 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - mapred.job.tracker is deprecated. Instead, use mapreduce.jobtracker.address
2023-04-02 21:47:58,074 [main] INFO  org.apache.pig.data.SchemaTupleBackend - Key [pig.schematuple] was not set... will not generate code.
2023-04-02 21:47:58,080 [main] INFO  org.apache.hadoop.mapreduce.lib.input.FileInputFormat - Total input paths to process : 1
2023-04-02 21:47:58,080 [main] INFO  org.apache.pig.backend.hadoop.executionengine.util.MapRedUtil - Total input paths to process : 1
(1 2 3)
(4 5 6)
(7 8 9)
grunt>
Hue - File Browser - da...   cloudera@quickstart:~
Activate Windows
Go to Settings to activate Windows.
Right Ctrl
```

Example 4: Creating pig script and loading tuple data.

1. Create a relation named 'tupledata' and enter the following data.

(3,8,9) (4,5,6)
(1,4,7) (3,7,5)
(2,5,8) (9,5,8)




a. Load tupledata into a pig variable named 'inputtuple'.

inputtuple = LOAD 'tupledata' AS (t1:tuple(t1a:int,
t1b:int,t1c:int),t2:tuple(t2a:int,t2b:int,t2c:int));
DUMP inputtuple;

```
2023-04-13 23:08:33,250 [main] INFO  org.apache.pig.backend.hadoop.executionengine.util.MapRedUtil - Total input paths to process : 1
((3,8,9),(4,5,6))
((1,4,7),(3,7,5))
((2,5,8),(9,5,8))
grunt>
```

2. Add the columns with a similar index of each tuple and store the result in addout.out file.

//try following cmd
addition = FOREACH A GENERATE t1a+t2a,t1b+t2b,t1c+t2c;
store addition into 'addout.out';

```
2023-04-13 23:10:26,858 [main] INFO  org.apache.pig.backend.hadoop.executionengine.util.MapRedUtil - Total input paths to process : 1
(7,13,15)
(4,11,12)
(11,10,16)
grunt>
```

A = LOAD 'data1' AS (a1:int,a2:int,a3:int);

DUMP A;
(1,2,3)
(4,2,1)

B = LOAD 'data2' AS (b1:int,b2:int);

DUMP B;
(2,4)
(8,9)
(1,3)

X = CROSS A, B;

DUMP X;

```
(4,2,1,1,3)
(4,2,1,8,9)
(4,2,1,2,4)
(1,2,3,1,3)
(1,2,3,8,9)
(1,2,3,2,4)
grunt>
```

d= DISTINCT A;

f1 = FILTER X BY (a1 == 1);

```
2023-04-13 23:31:30,136 [main] INFO  org.apache.pig.backend.hadoop.executionen
2023-04-13 23:31:30,137 [main] INFO  org.apache.hadoop.conf.Configuration.depr
2023-04-13 23:31:30,166 [main] INFO  org.apache.hadoop.conf.Configuration.depr
2023-04-13 23:31:30,166 [main] INFO  org.apache.hadoop.conf.Configuration.depr
2023-04-13 23:31:30,166 [main] WARN  org.apache.pig.data.SchemaTupleBackend -
2023-04-13 23:31:30,187 [main] INFO  org.apache.hadoop.mapreduce.lib.input.Fil
2023-04-13 23:31:30,187 [main] INFO  org.apache.pig.backend.hadoop.executionen
(1,2,3,1,3)
(1,2,3,8,9)
(1,2,3,2,4)
grunt>
```

Q. Filter all rows for 2nd column val+4th column val=5th column val
F2 = FILTER X BY (a2+b1 == b2);

```
2023-04-13 23:34:46,252 [main] INFO  org.apache.pig.backend.hadoop.executi
2023-04-13 23:34:46,253 [main] INFO  org.apache.hadoop.conf.Configuration.
2023-04-13 23:34:46,254 [main] INFO  org.apache.hadoop.conf.Configuration.
2023-04-13 23:34:46,254 [main] INFO  org.apache.hadoop.conf.Configuration.
2023-04-13 23:34:46,254 [main] WARN  org.apache.pig.data.SchemaTupleBacken
2023-04-13 23:34:46,278 [main] INFO  org.apache.hadoop.mapreduce.lib.input
2023-04-13 23:34:46,278 [main] INFO  org.apache.pig.backend.hadoop.executi
(4,2,1,1,3)
(4,2,1,2,4)
(1,2,3,1,3)
(1,2,3,2,4)
grunt>
```

Q. Filter all rows for 2nd column val+4th column val=5th column val and also check for 1st column val as 4
f1 = FILTER X BY (f2+f4 == f5) AND (f1==4);

```
2023-04-13 23:42:40,248 [main] INFO  org.apache.pig.backend.hadoop.executioneng
2023-04-13 23:42:40,249 [main] INFO  org.apache.hadoop.conf.Configuration.depre
2023-04-13 23:42:40,249 [main] INFO  org.apache.hadoop.conf.Configuration.depre
2023-04-13 23:42:40,249 [main] INFO  org.apache.hadoop.conf.Configuration.depre
2023-04-13 23:42:40,249 [main] WARN  org.apache.pig.data.SchemaTupleBackend - S
2023-04-13 23:42:40,303 [main] INFO  org.apache.hadoop.mapreduce.lib.input.File
2023-04-13 23:42:40,303 [main] INFO  org.apache.pig.backend.hadoop.executioneng
(4,2,1,1,3)
(4,2,1,2,4)
grunt>
```

XX = FOREACH A GENERATE f1;

```
2023-04-13 23:43:25,286 [main] I
2023-04-13 23:43:25,287 [main] I
2023-04-13 23:43:25,287 [main] I
2023-04-13 23:43:25,287 [main] I
2023-04-13 23:43:25,287 [main] W
2023-04-13 23:43:25,300 [main] I
2023-04-13 23:43:25,300 [main] I
(1)
(4)
grunt>
```

B = GROUP A BY f1

```
2023-04-13 23:44:13,988 [main] INFO  org.apache.pig
2023-04-13 23:44:13,988 [main] INFO  org.apache.had
2023-04-13 23:44:13,988 [main] INFO  org.apache.had
2023-04-13 23:44:13,988 [main] INFO  org.apache.had
2023-04-13 23:44:13,988 [main] WARN  org.apache.pig
2023-04-13 23:44:14,002 [main] INFO  org.apache.had
2023-04-13 23:44:14,002 [main] INFO  org.apache.pig
(1,{(1,2,3)})
(4,{(4,2,1)})
grunt>
```

DESCRIBE B

```
2023-04-13 23:44:13,988 [main] INFO  org.apache.hadoop.conf.Configuration
2023-04-13 23:44:13,988 [main] WARN  org.apache.pig.data.SchemaTupleBacke
2023-04-13 23:44:14,002 [main] INFO  org.apache.hadoop.mapreduce.lib.inpu
2023-04-13 23:44:14,002 [main] INFO  org.apache.pig.backend.hadoop.execut
(1,{(1,2,3)})
(4,{(4,2,1)})
grunt> DESCRIBE B;
B: {group: int,A: {(a1: int,a2: int,a3: int)}}
grunt>
```

ILLUSTRATE B

```
2023-04-13 23:45:27,348 [main] WARN  org.apache.pig.data.SchemaTupleBackend - SchemaTupleBackend has already been initialized
2023-04-13 23:45:27,350 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.PigMapReduce$Reduce - Aliases being
2023-04-13 23:45:27,351 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MRCompiler - File concatenation thre
2023-04-13 23:45:27,355 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MultiQueryOptimizer - MR plan size b
2023-04-13 23:45:27,355 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MultiQueryOptimizer - MR plan size a
2023-04-13 23:45:27,356 [main] INFO  org.apache.pig.tools.pigstats.ScriptState - Pig script settings are added to the job
2023-04-13 23:45:27,357 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.JobControlCompiler - mapred.job.redu
2023-04-13 23:45:27,357 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.JobControlCompiler - Reduce phase de
2023-04-13 23:45:27,357 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.JobControlCompiler - Using reducer e
2023-04-13 23:45:27,357 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.InputSizeReducerEstimator - BytesPer
2023-04-13 23:45:27,357 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.JobControlCompiler - Setting Paralle
2023-04-13 23:45:27,397 [main] WARN  org.apache.pig.data.SchemaTupleBackend - SchemaTupleBackend has already been initialized
2023-04-13 23:45:27,403 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.PigGenericMapReduce$Map - Aliases be
2023-04-13 23:45:27,410 [main] WARN  org.apache.pig.data.SchemaTupleBackend - SchemaTupleBackend has already been initialized
2023-04-13 23:45:27,414 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.PigMapReduce$Reduce - Aliases being
-----------------------------------------------
| A     | a1:int      | a2:int      | a3:int      |
-----------------------------------------------
|       | 1           | 2           | 3           |
|       | 1           | 2           | 3           |
-----------------------------------------------
-------------------------------------------------------------------------------------
| B     | group:int      | A:bag{:tuple(a1:int,a2:int,a3:int)}                        |
-------------------------------------------------------------------------------------
|       | 1              | {(1, 2, 3), (1, 2, 3)}                                     |
-------------------------------------------------------------------------------------

grunt>
```

1. Create the file named ‘data’ for the following data.

1,2,3
4,2,1
8,3,4
4,3,3
7,2,5
8,4,3

0. Load the contents of ‘data’ into relation A.

A = LOAD 'data' AS (a1:int,a2:int,a3:int);

```
(1,2,3)
(4,2,1)
(8,3,4)
(4,3,3)
(7,2,5)
(8,4,3)
grunt> A = LOAD '/home/cloudera/Desktop/data_mat.txt' using PigStorage(',') AS (a1:int,a2:int,a3:int);
```

0. Display the elements of the first column.

X = FOREACH A GENERATE f1;
DUMP X

```
2023-04-13 23:56:04,646 [main] INFO  org.apache.pig.
2023-04-13 23:56:04,647 [main] INFO  org.apache.hado
2023-04-13 23:56:04,647 [main] INFO  org.apache.hado
2023-04-13 23:56:04,647 [main] INFO  org.apache.hado
2023-04-13 23:56:04,647 [main] WARN  org.apache.pig.
2023-04-13 23:56:04,670 [main] INFO  org.apache.hado
2023-04-13 23:56:04,670 [main] INFO  org.apache.pig.
(1)
(4)
(8)
(4)
(7)
(8)
grunt>
```

0. Display all tuples of relation A.

```
X = FOREACH A GENERATE *;
DUMP X;
```

```
2023-04-13 23:57:34,849 [main] INFO  org.apache.pig.backend.hadoop.exec
2023-04-13 23:57:34,850 [main] INFO  org.apache.hadoop.conf.Configurati
2023-04-13 23:57:34,850 [main] INFO  org.apache.hadoop.conf.Configurati
2023-04-13 23:57:34,850 [main] INFO  org.apache.hadoop.conf.Configurati
2023-04-13 23:57:34,850 [main] WARN  org.apache.pig.data.SchemaTupleBac
2023-04-13 23:57:34,867 [main] INFO  org.apache.hadoop.mapreduce.lib.in
2023-04-13 23:57:34,867 [main] INFO  org.apache.pig.backend.hadoop.exec
(1,2,3)
(4,2,1)
(8,3,4)
(4,3,3)
(7,2,5)
(8,4,3)
grunt>
```

0. Display the first two columns of relation A.

```
X = FOREACH A GENERATE  a1,a2;
DUMP X;
```

```
2023-04-13 23:58:22,256 [main] INFO  org.apache.hadoop.con
2023-04-13 23:58:22,256 [main] INFO  org.apache.hadoop.con
2023-04-13 23:58:22,256 [main] INFO  org.apache.hadoop.con
2023-04-13 23:58:22,256 [main] WARN  org.apache.pig.data.S
2023-04-13 23:58:22,274 [main] INFO  org.apache.hadoop.map
2023-04-13 23:58:22,274 [main] INFO  org.apache.pig.backen
(1,2)
(4,2)
(8,3)
(4,3)
(7,2)
(8,4)
grunt>
```