**PRACTICAL 9**

**To implement Word Count problem using Pig**

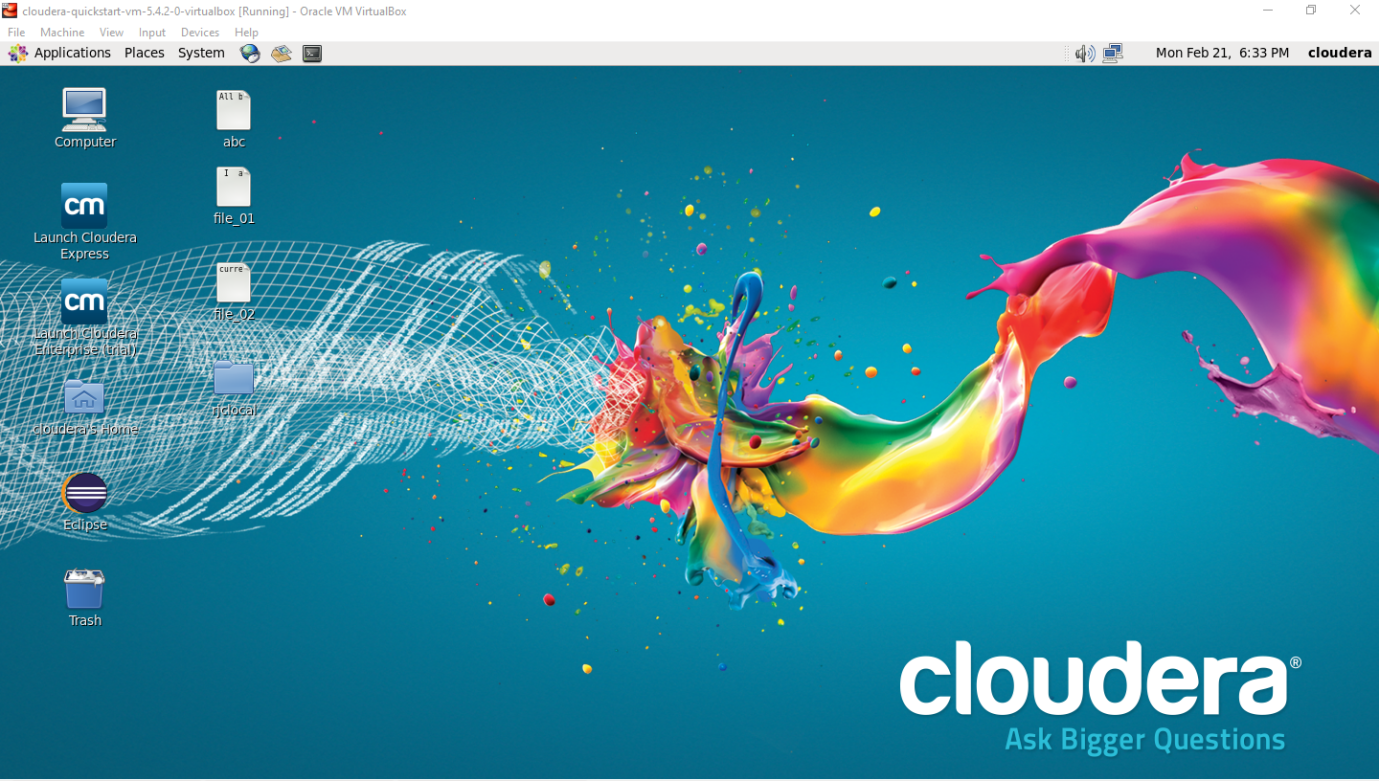
**Apache Pig**

* **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.

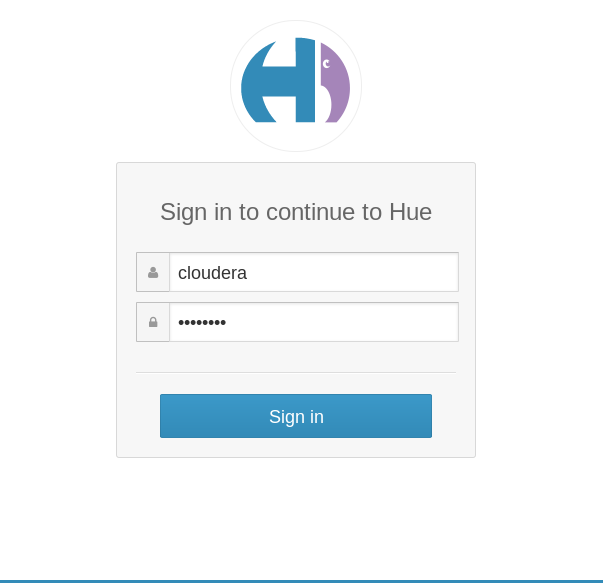
**To implement Word Count problem using Pig**

**Steps:**

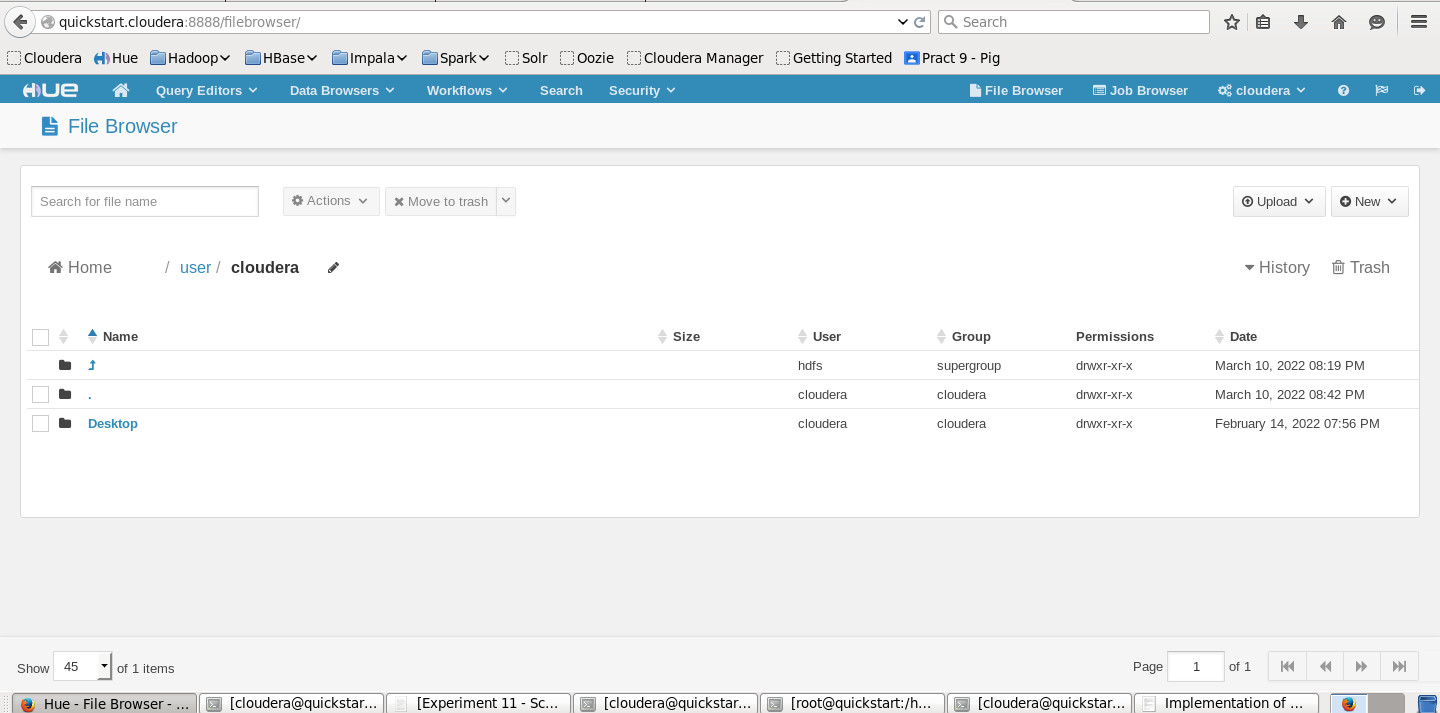
1. Start the cloudera.



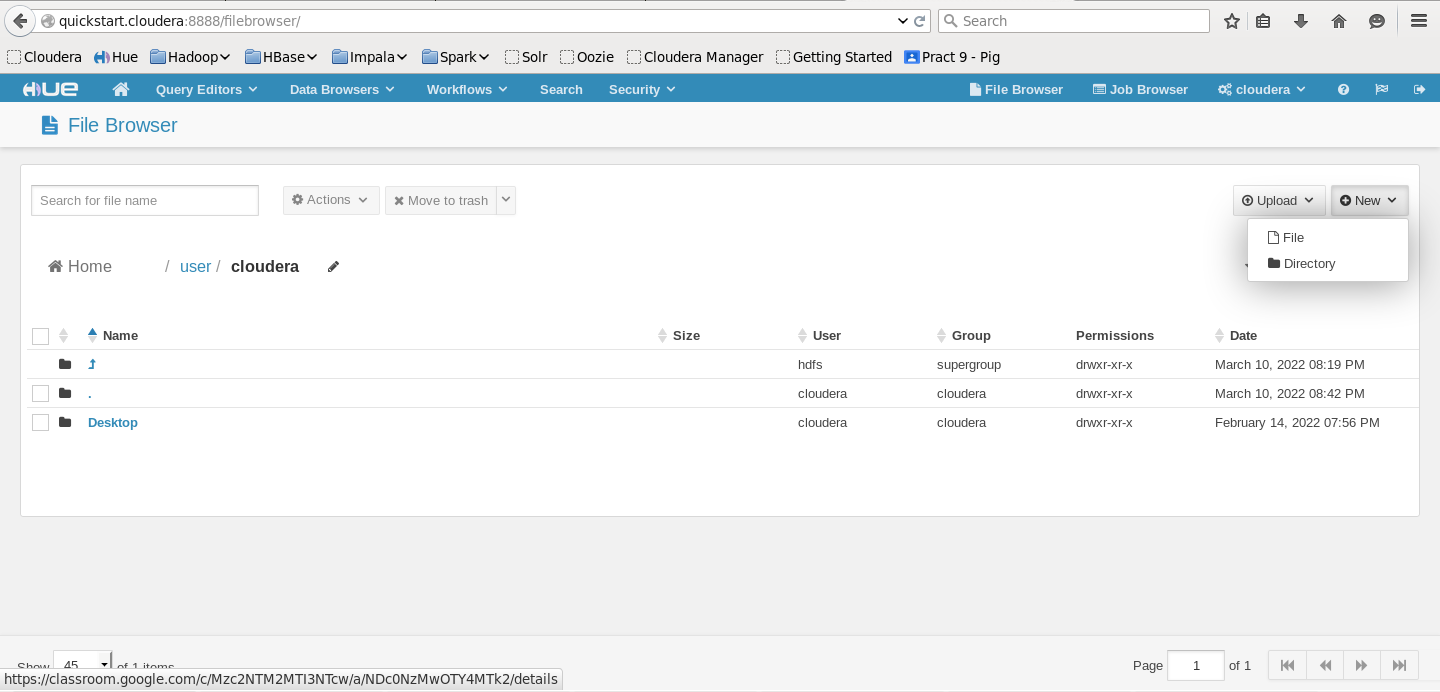
1. Open the browser. And then open Hue and login.



1. In Hue Go to file browser and Now open the directory /user/cloudera



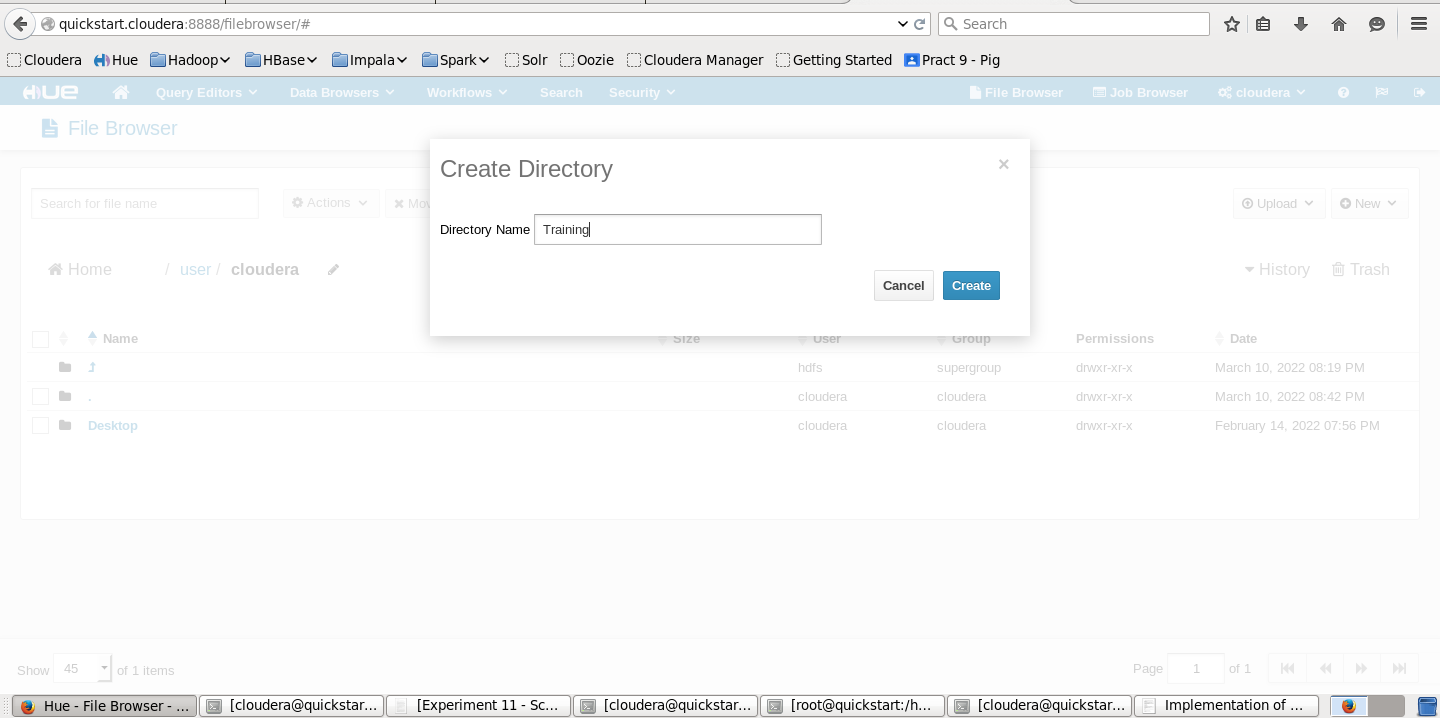
1. Now we are creating the directory as Training inside /user/cloudera

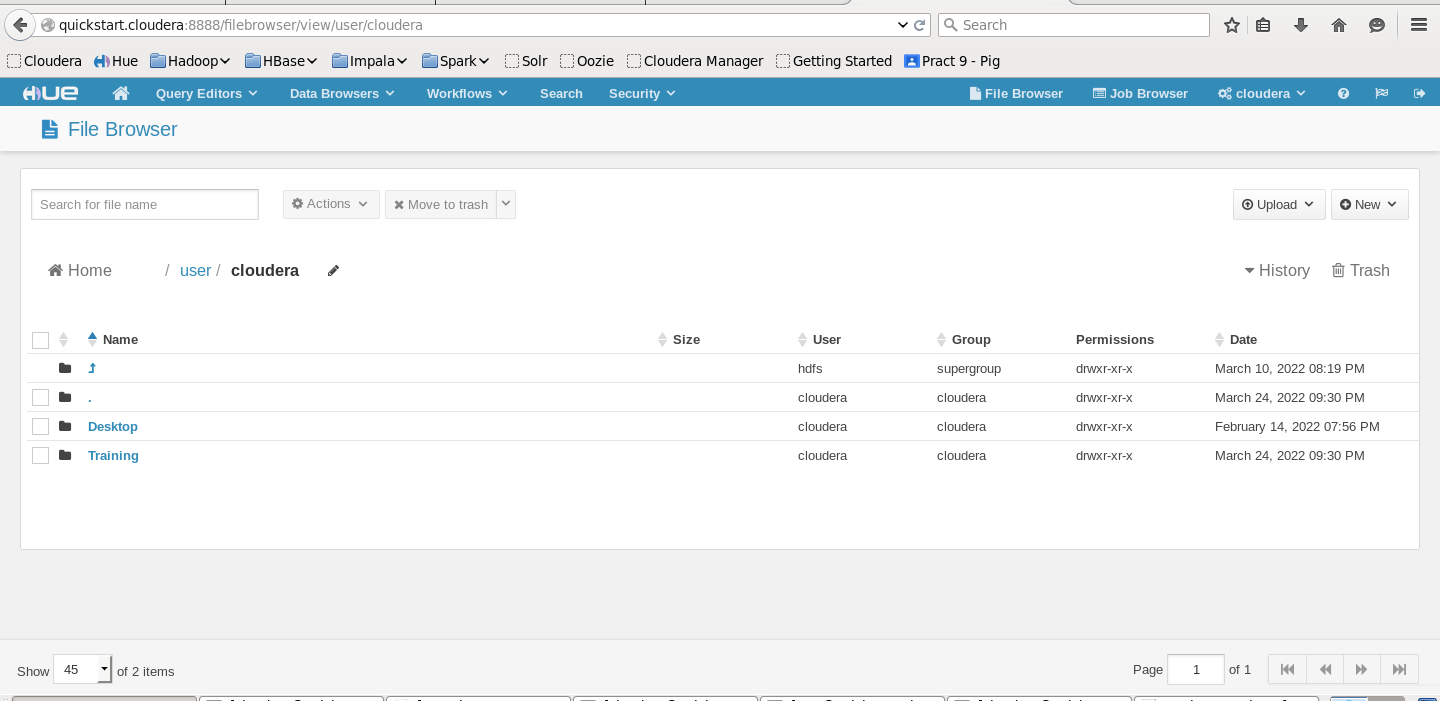


In File Browser we have New option in right corner

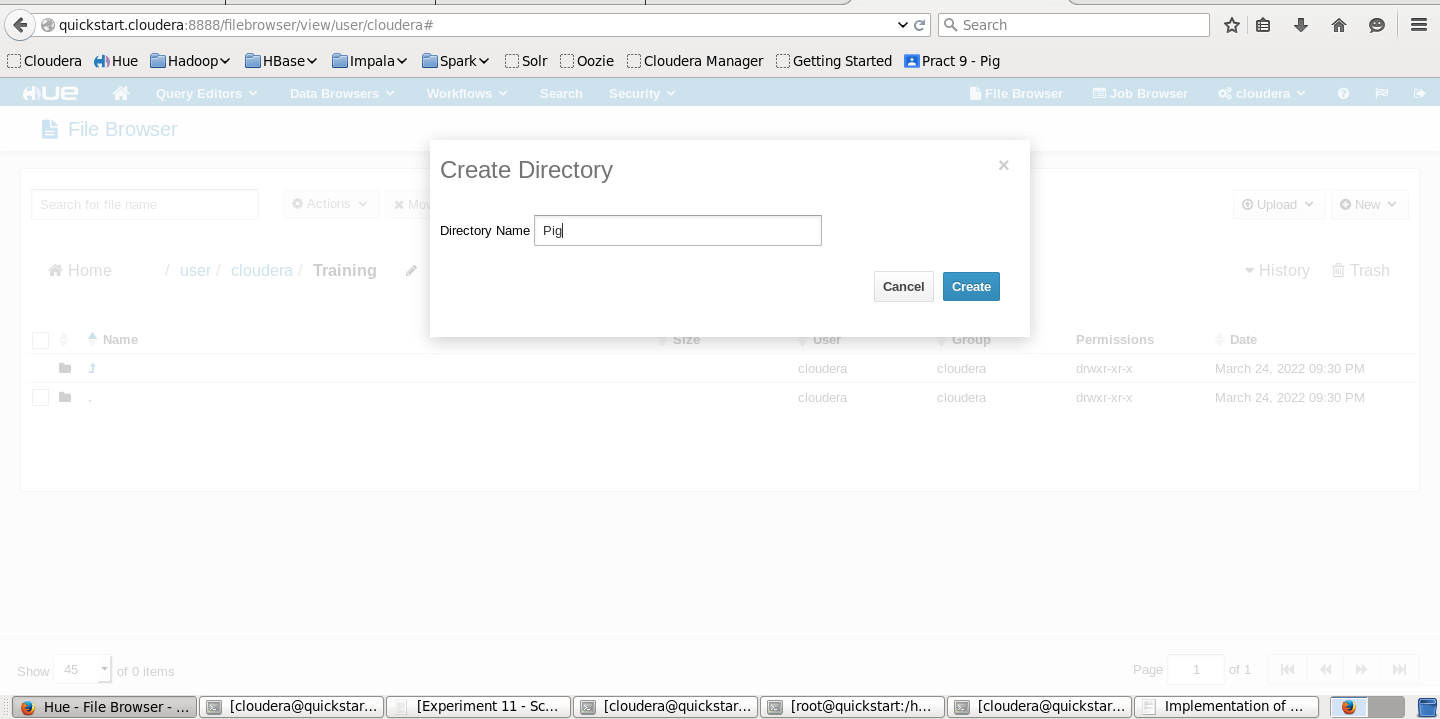
Click on New 🡪Directory

Give the directory name And click on Create

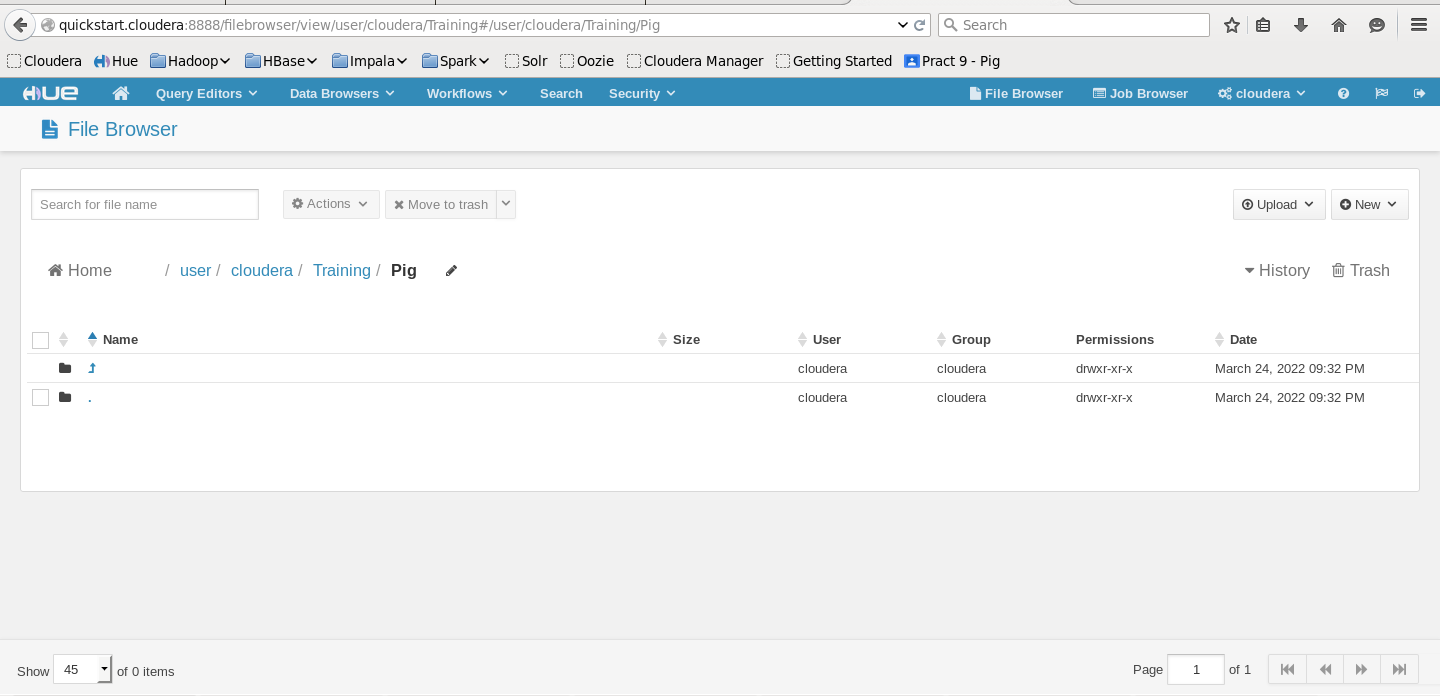




1. After creating Training directory now creating the Pig directory inside Training.

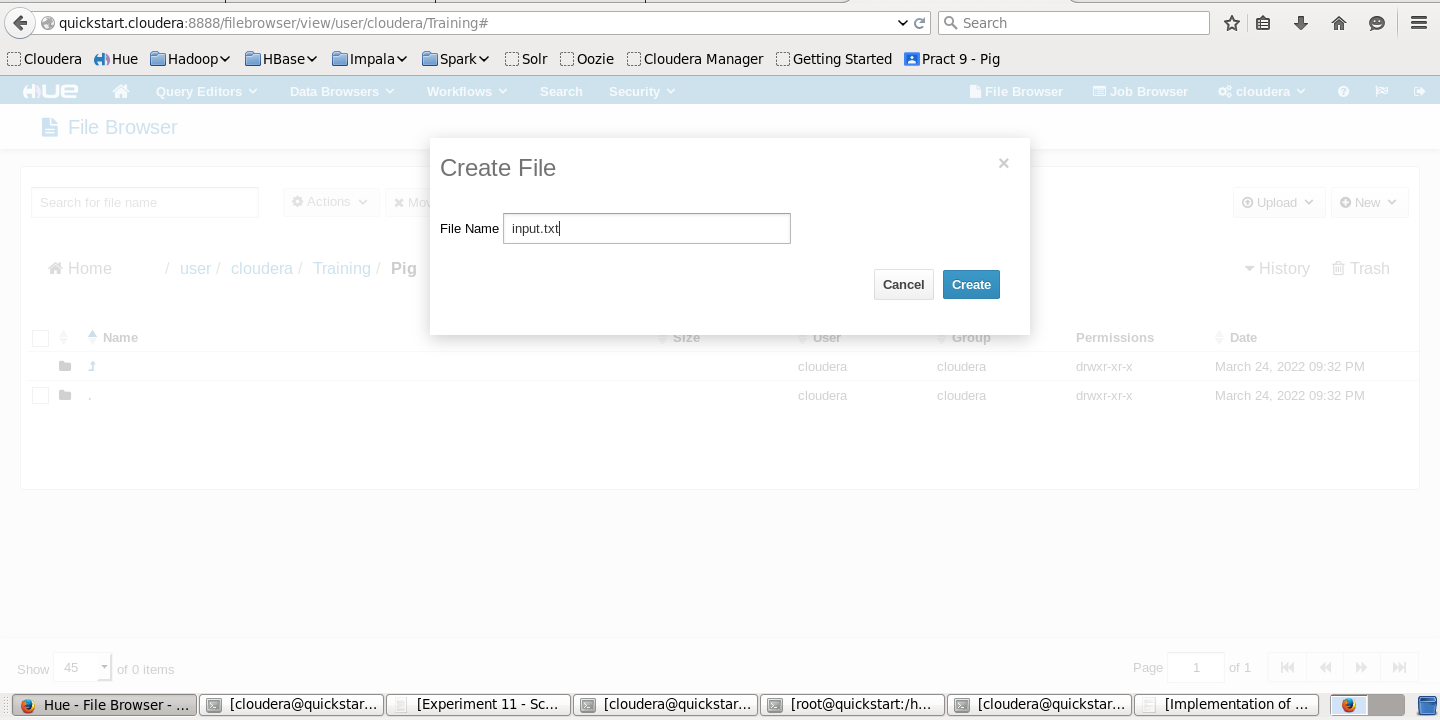


1. Pig directory has been created inside /user/cloudera/Training

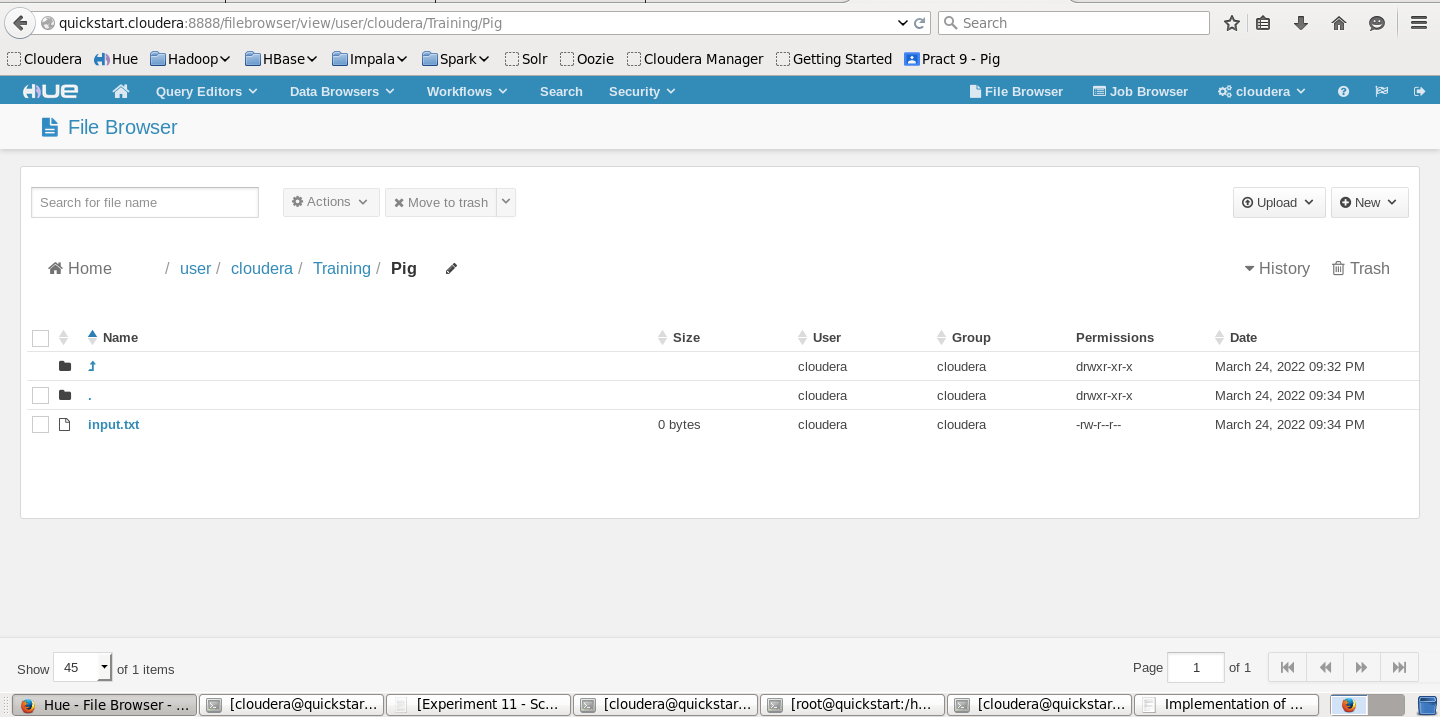


1. Creating input.txt file inside /usr/cloudera/Training/Pig directory

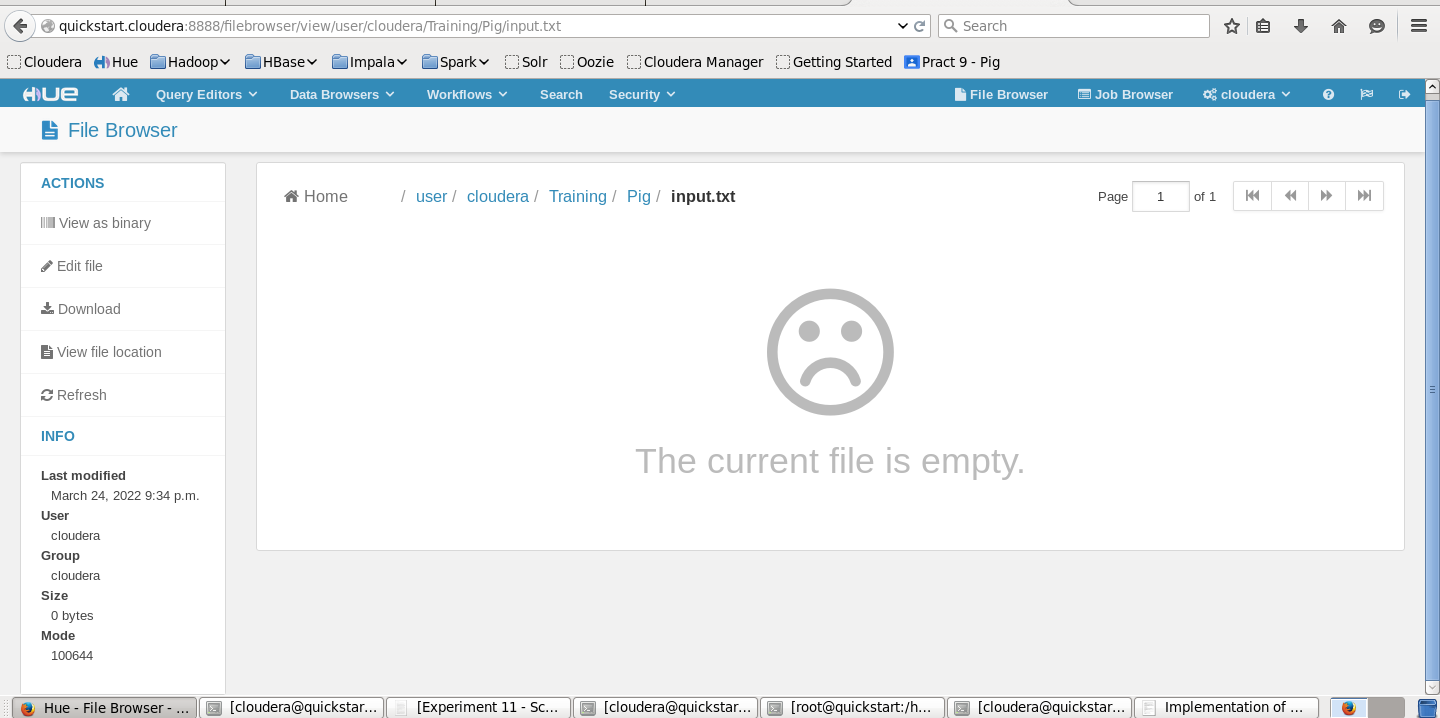
Again inside the Pig directory click on New and create file as’input.txt’



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

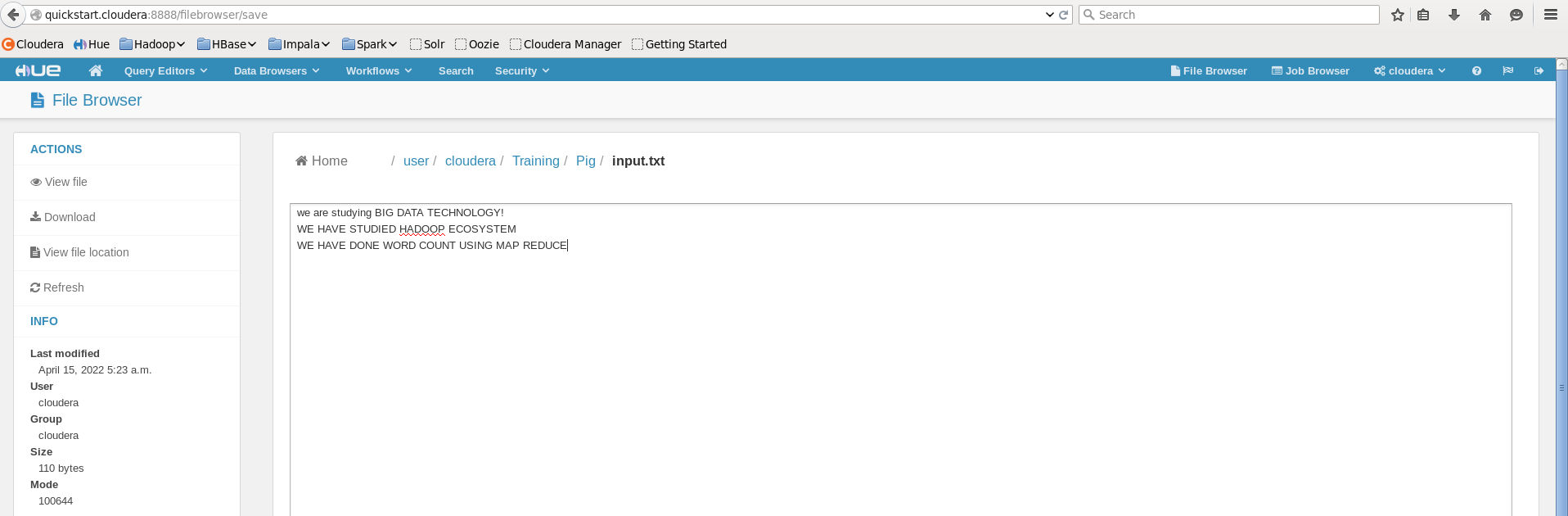


1. Adding some contents to this input.txt file.

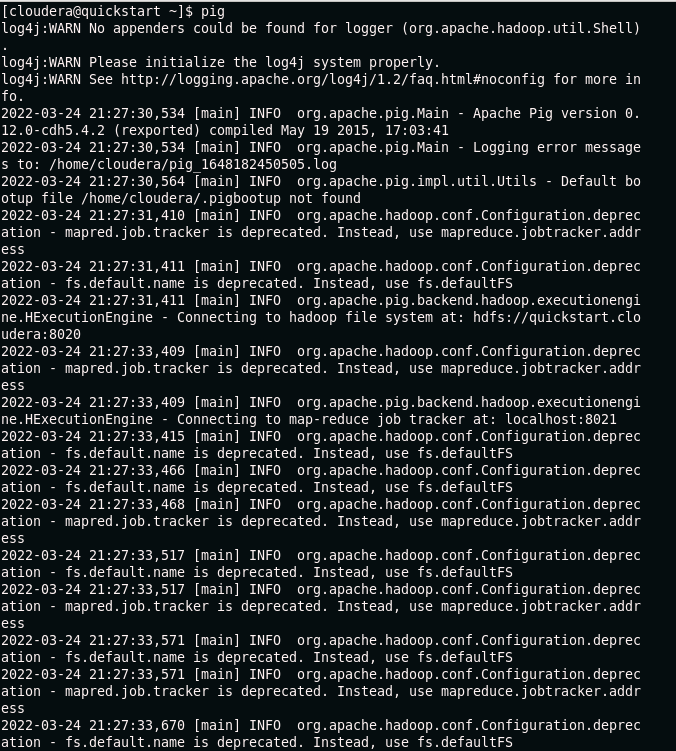


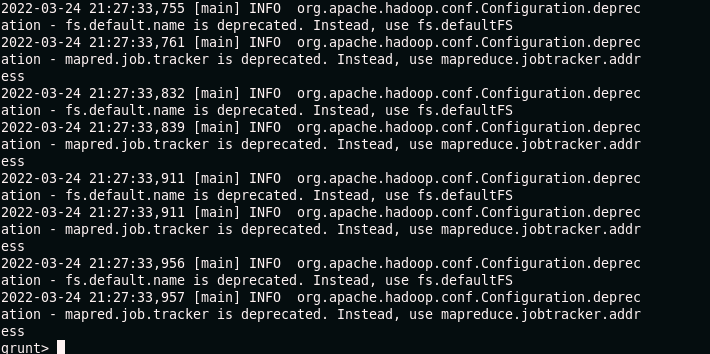
For adding content in the input file, Click on ‘Edit file’ option then add the content.

Save the input.txt file



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





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

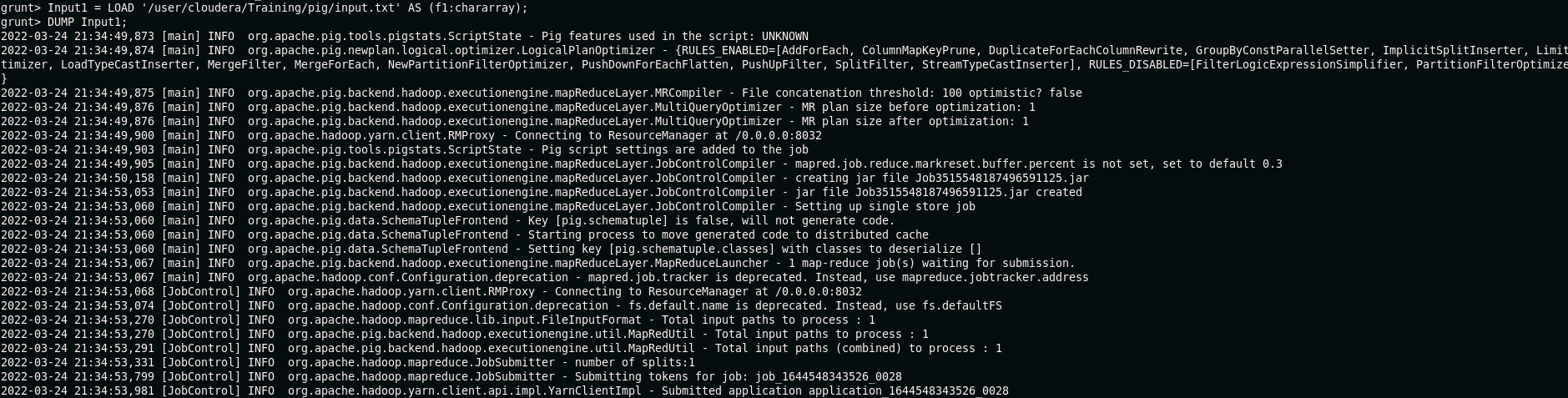
Input1 = LOAD ‘/usr/cloudera/Training/pig/input.txt’ AS (f1:chararray);

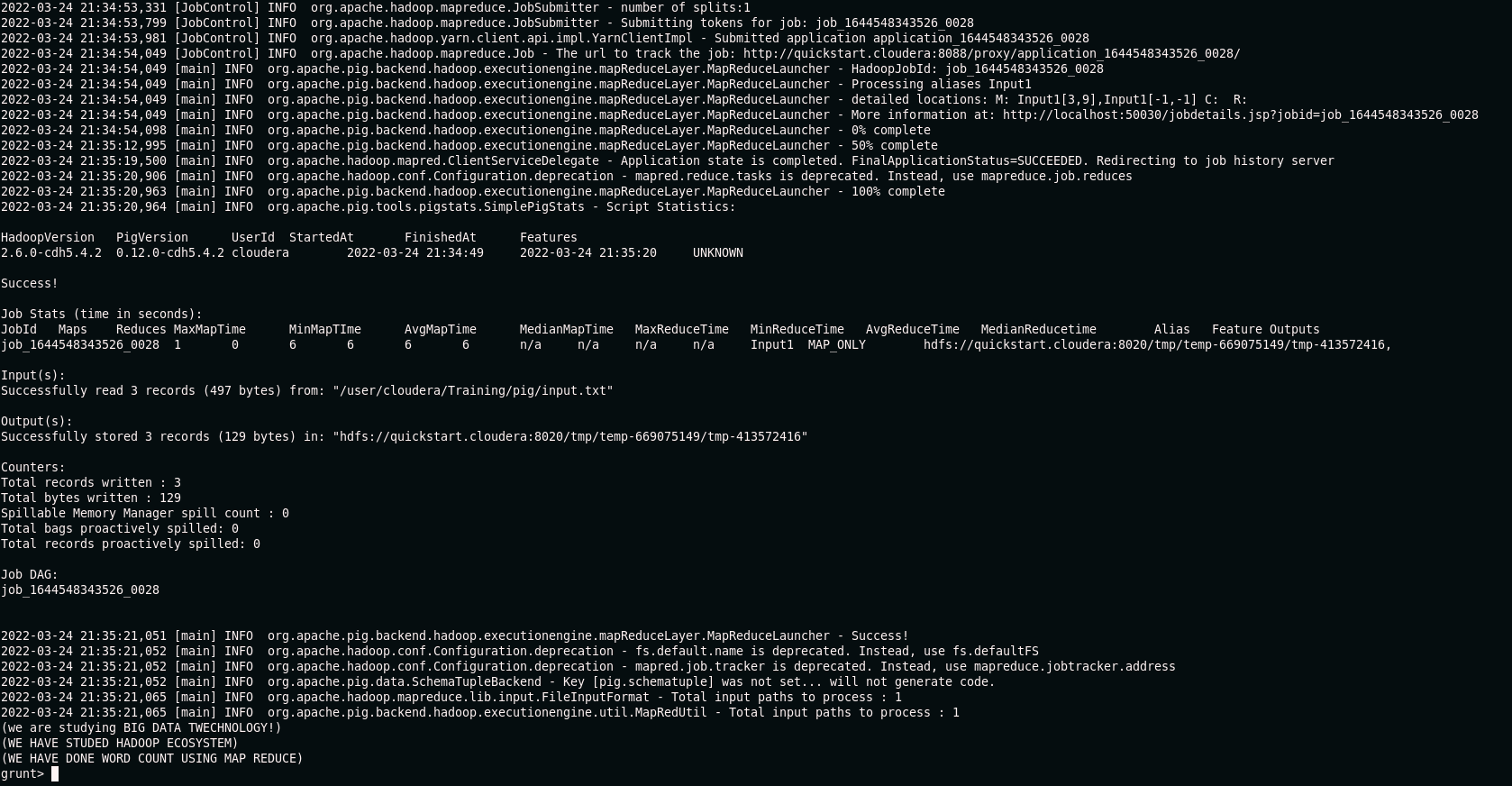


1. Now we are dumping the data. It will do the MapReduce task.

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;

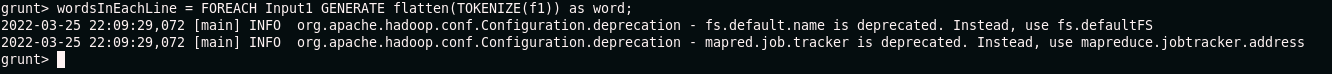




1. Here we are counting the words in each line for that we are using the following command

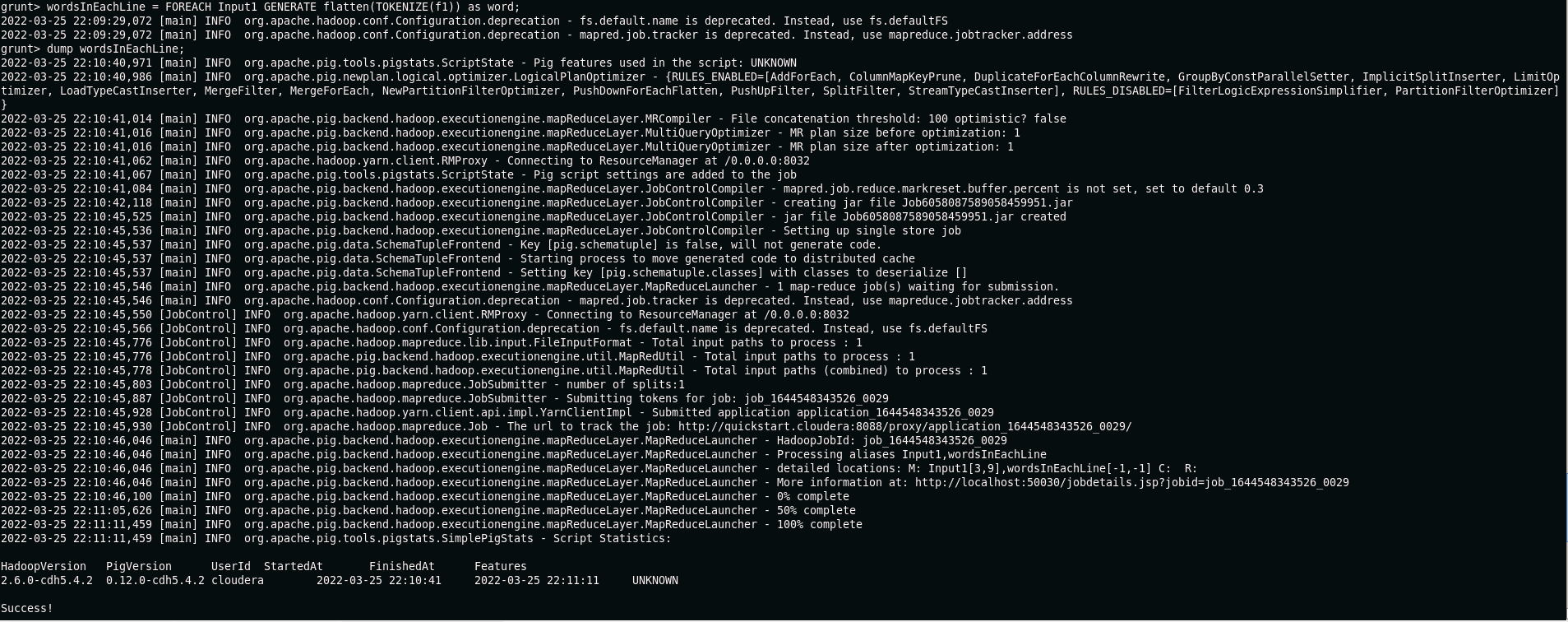
wordsInEachLine = FOREACH input1 GENERATE

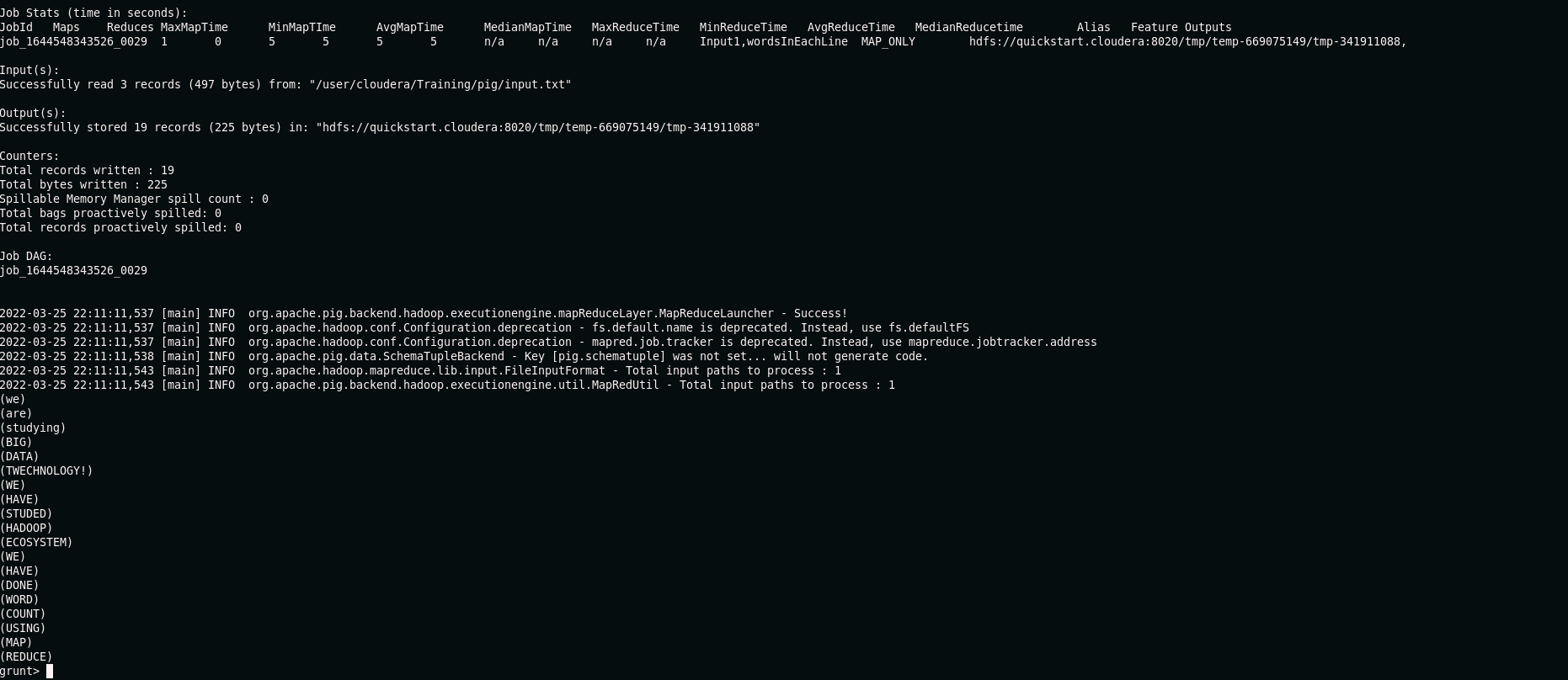
flatten(TOKENIZE(f1)) as word;



1. Again, we are dumping the data. It will do the MapReduce task.

dump wordsInEachLine;





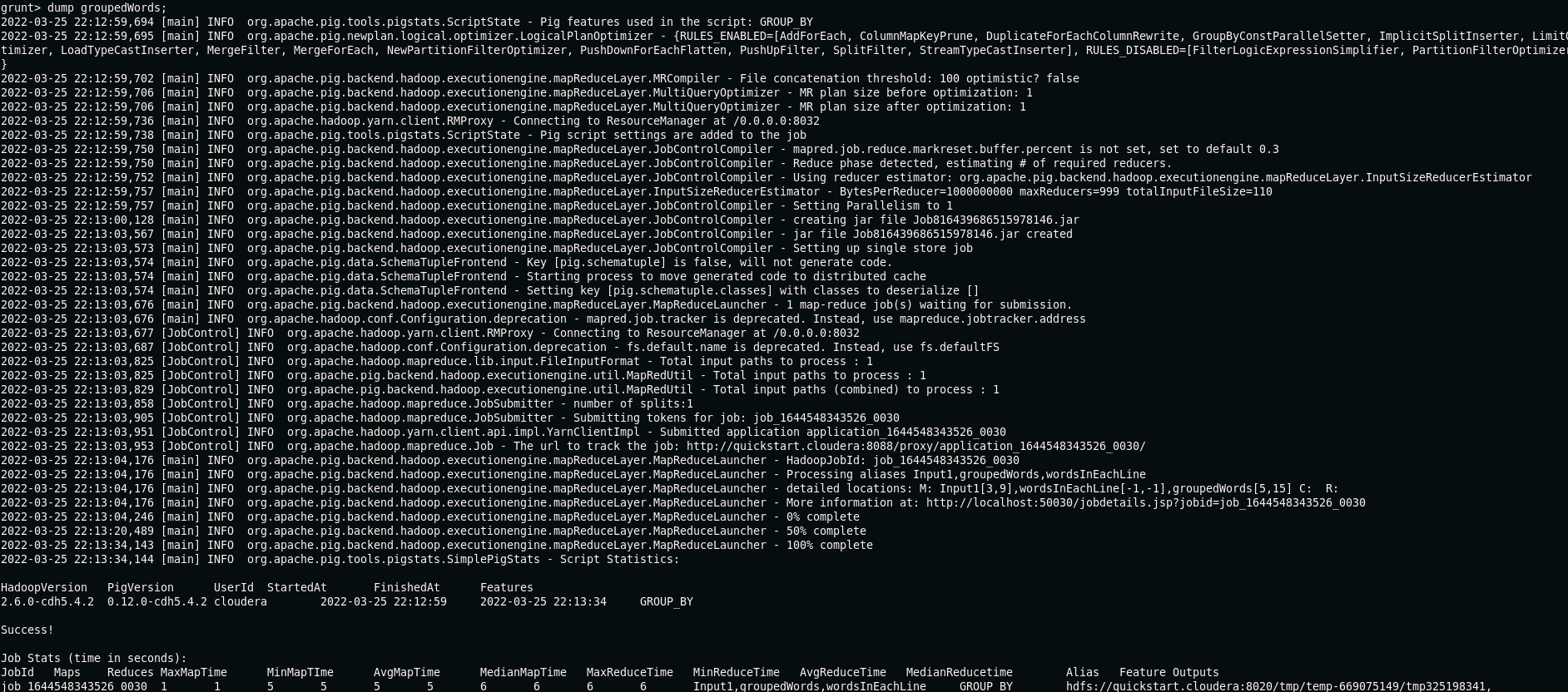
1. Now grouping the words present in each line.

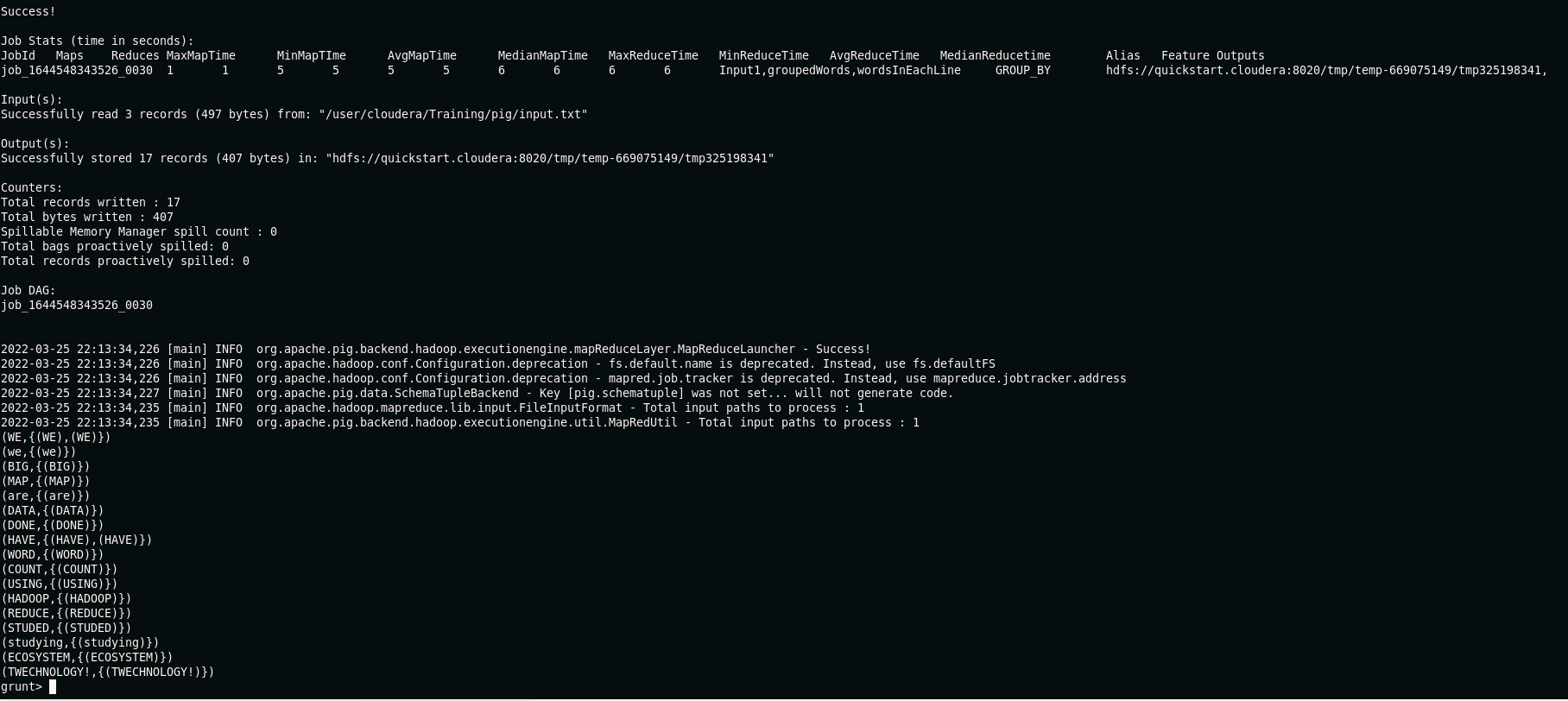
groupedWords = group wordsInEachLine by word;



And then dumping the data by the following command.

dump groupedWords;





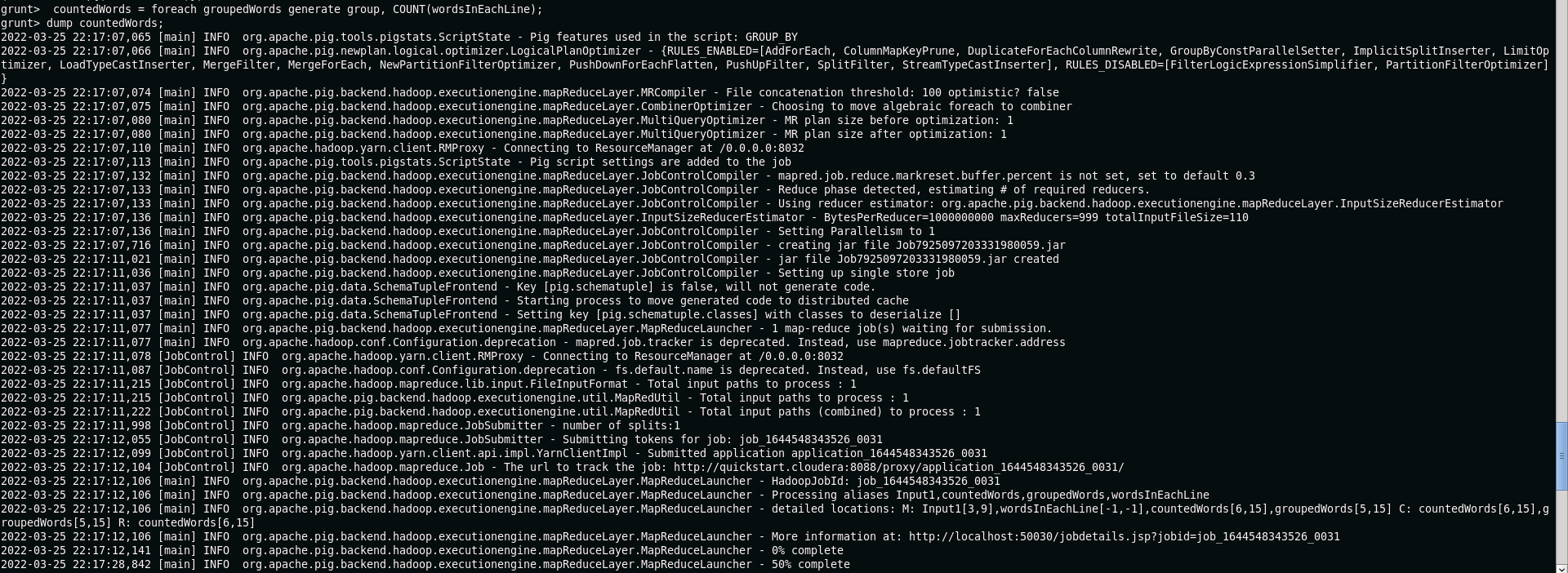
1. Now we count those words. For each group we count words in each line.

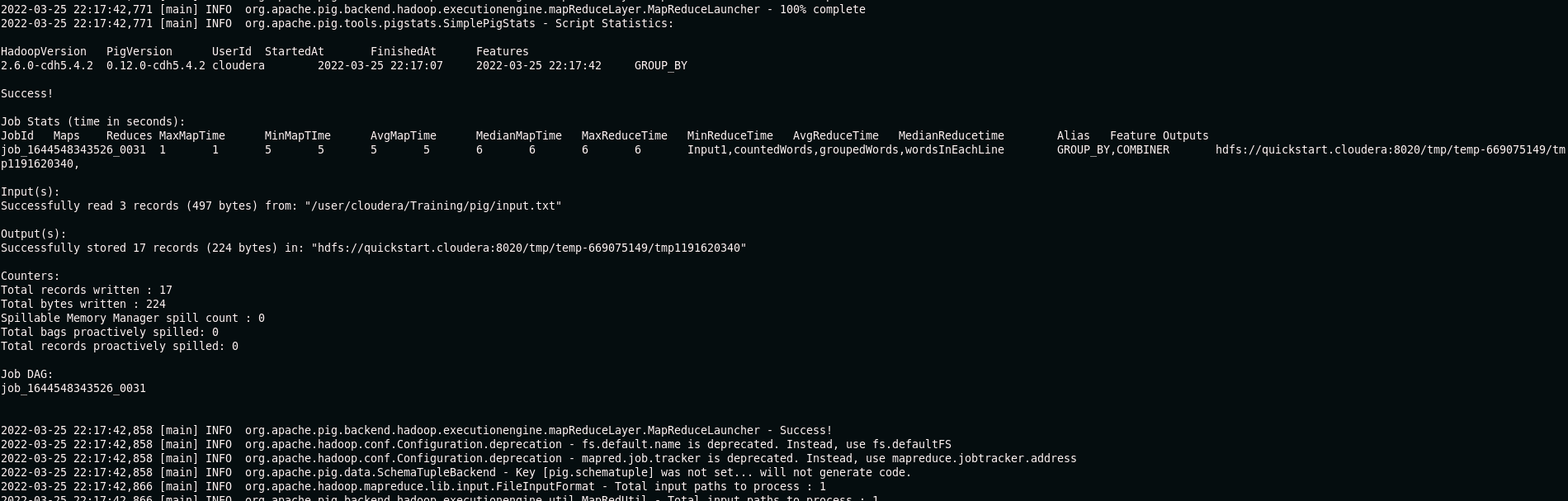
countedWords = foreach groupedWords generate group, COUNT(wordsInEachLine);

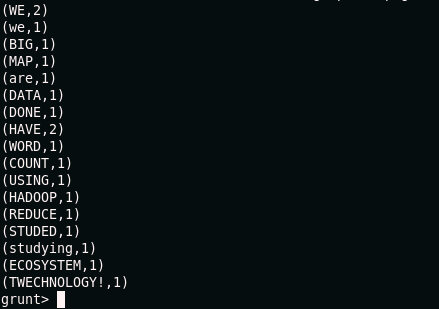


1. After every counting of words commands, we are dumping the data dump countedWords;

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







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.

1. Now Exit from the grunt shell using quit command.

