***Practical 8: To implement Word Count problem using Pig***

**Steps involved implement Word Count problem using Pig below in the following order-**

**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.

1. **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.

1. **Extensibility**

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

1. **Flexible**

It can easily handle structured as well as unstructured data.

1. **In-built operators**

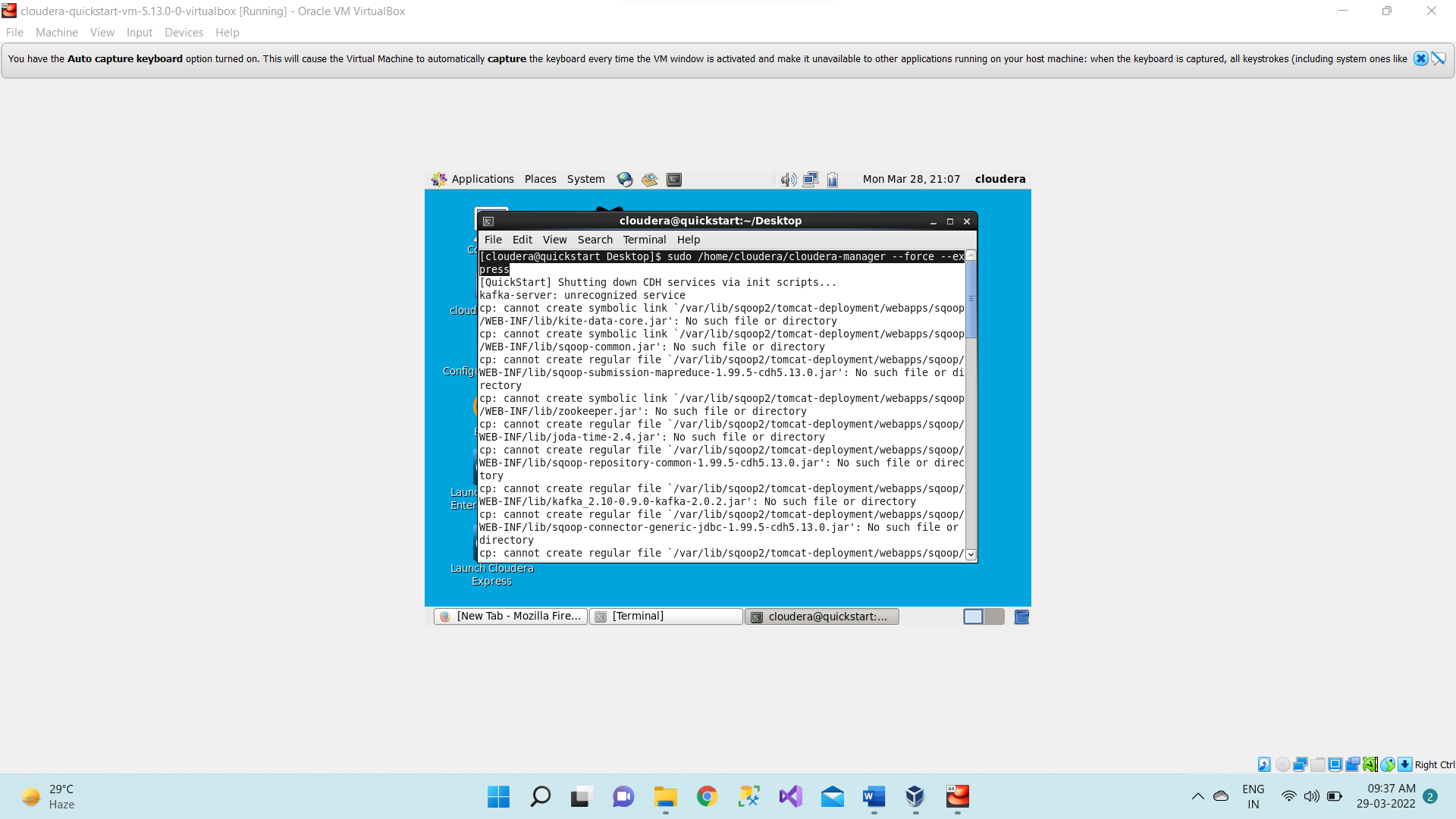
It contains various type of operators such as sort, filter and joins.

**Advantages of Apache Pig**

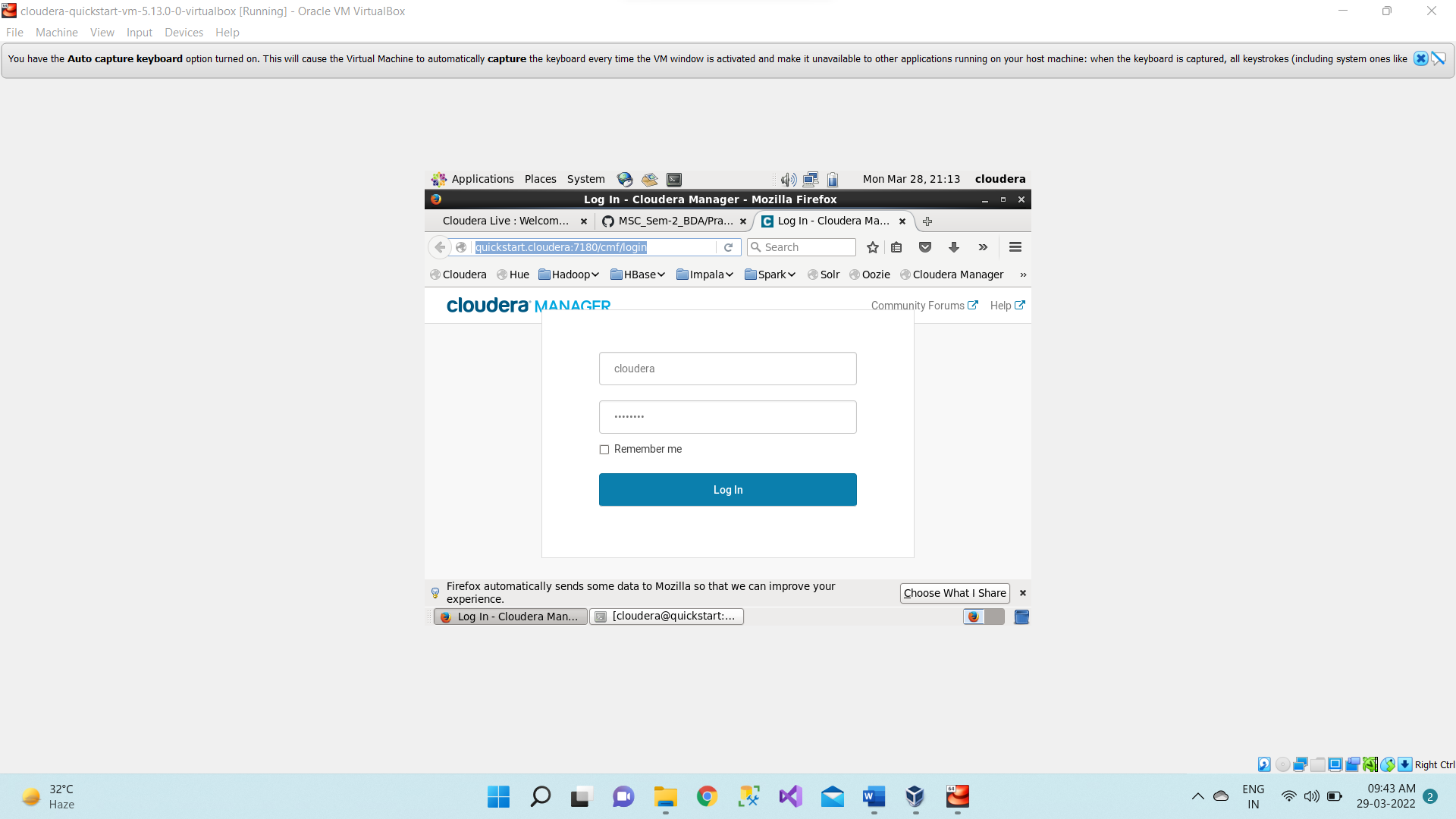
* **Less code -** The Pig consumes less line of code to perform any operation.
  + **Reusability -** The Pig code is flexible enough to reuse again.
  + **Nested data types -** The Pig provides a useful concept of nested data types like tuple, bag.

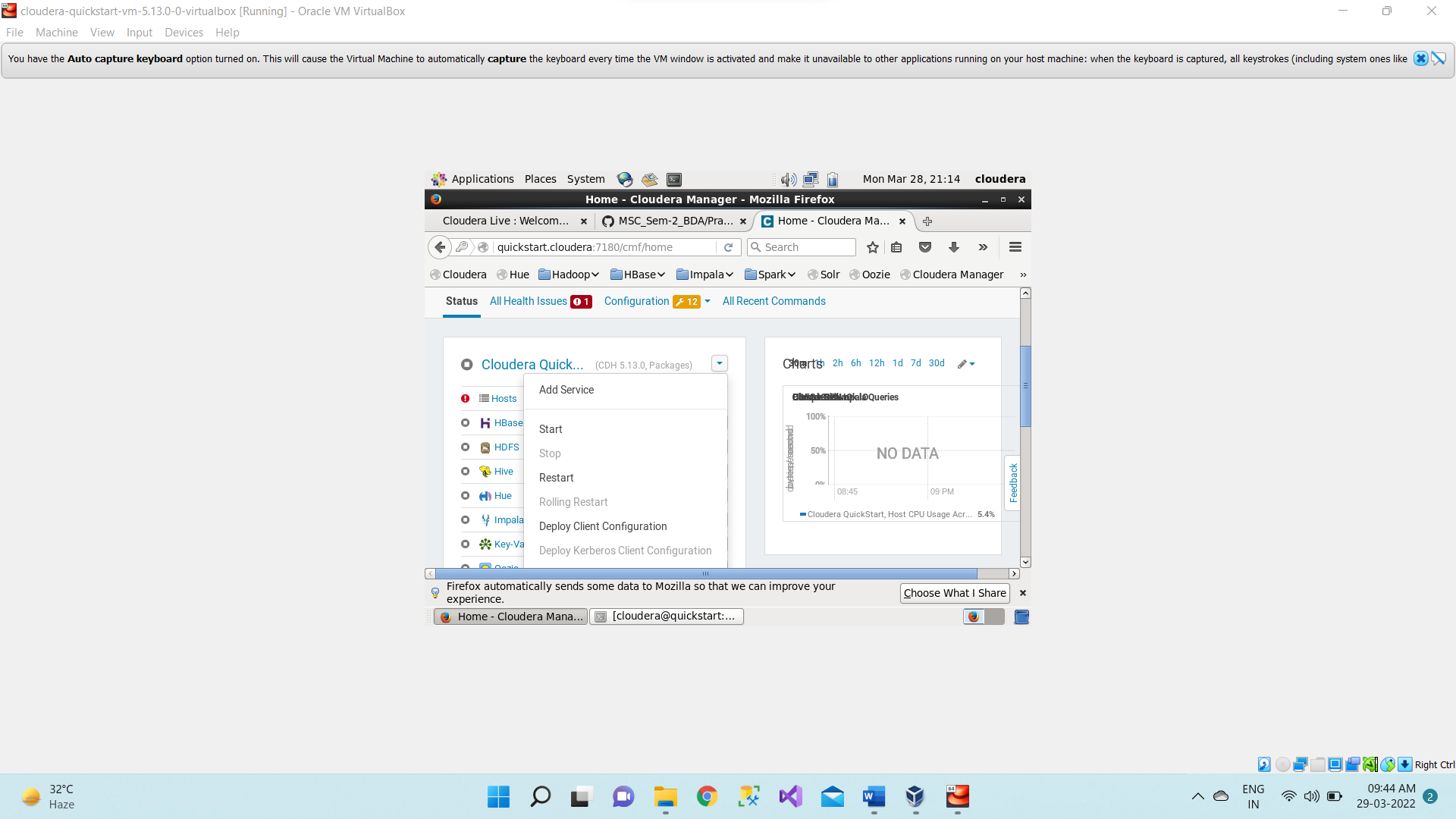
Before the perform the practical first perform **3 steps** of the given following-

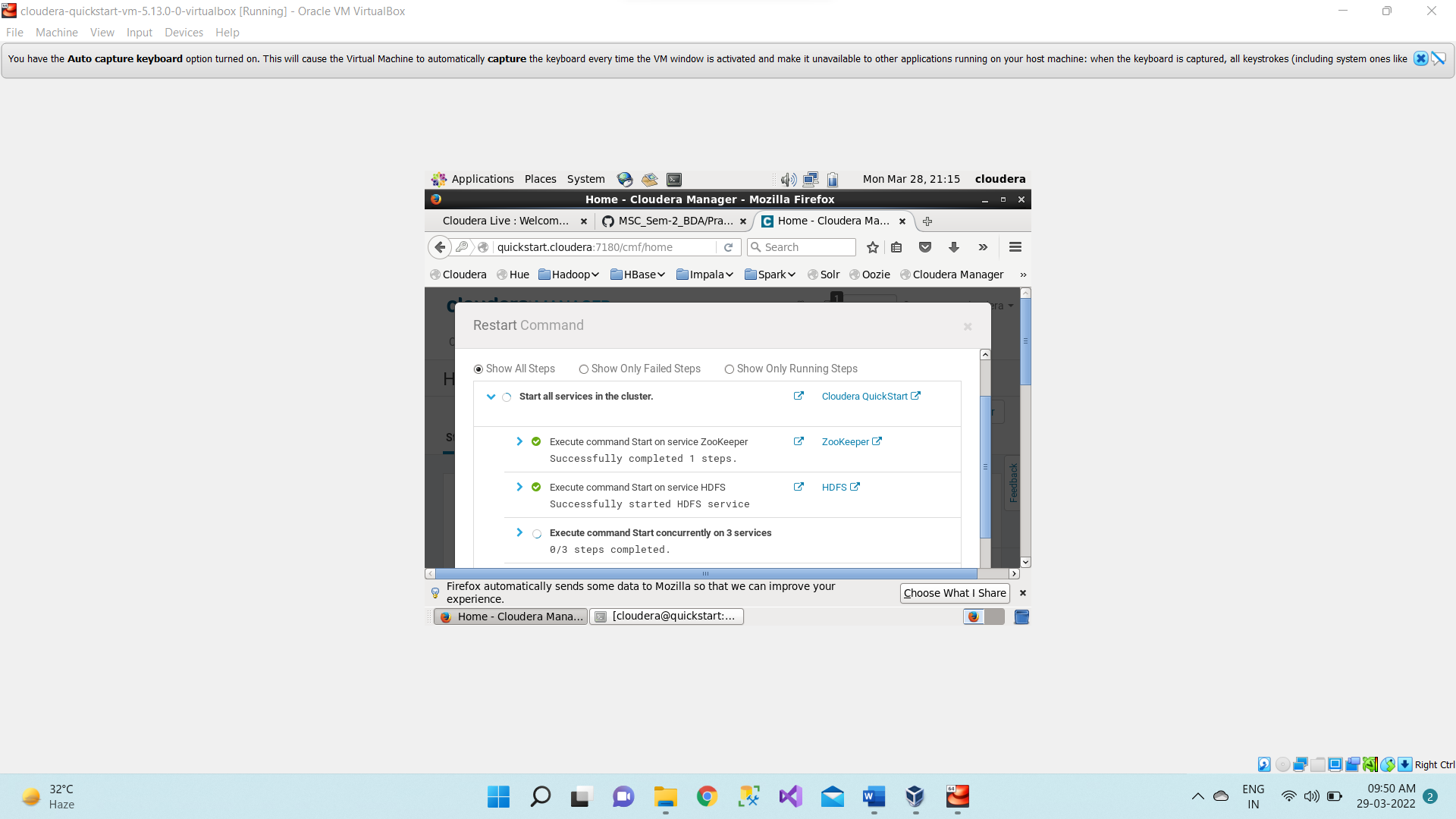
1. **sudo /home/cloudera/cloudera-manager --force –express**



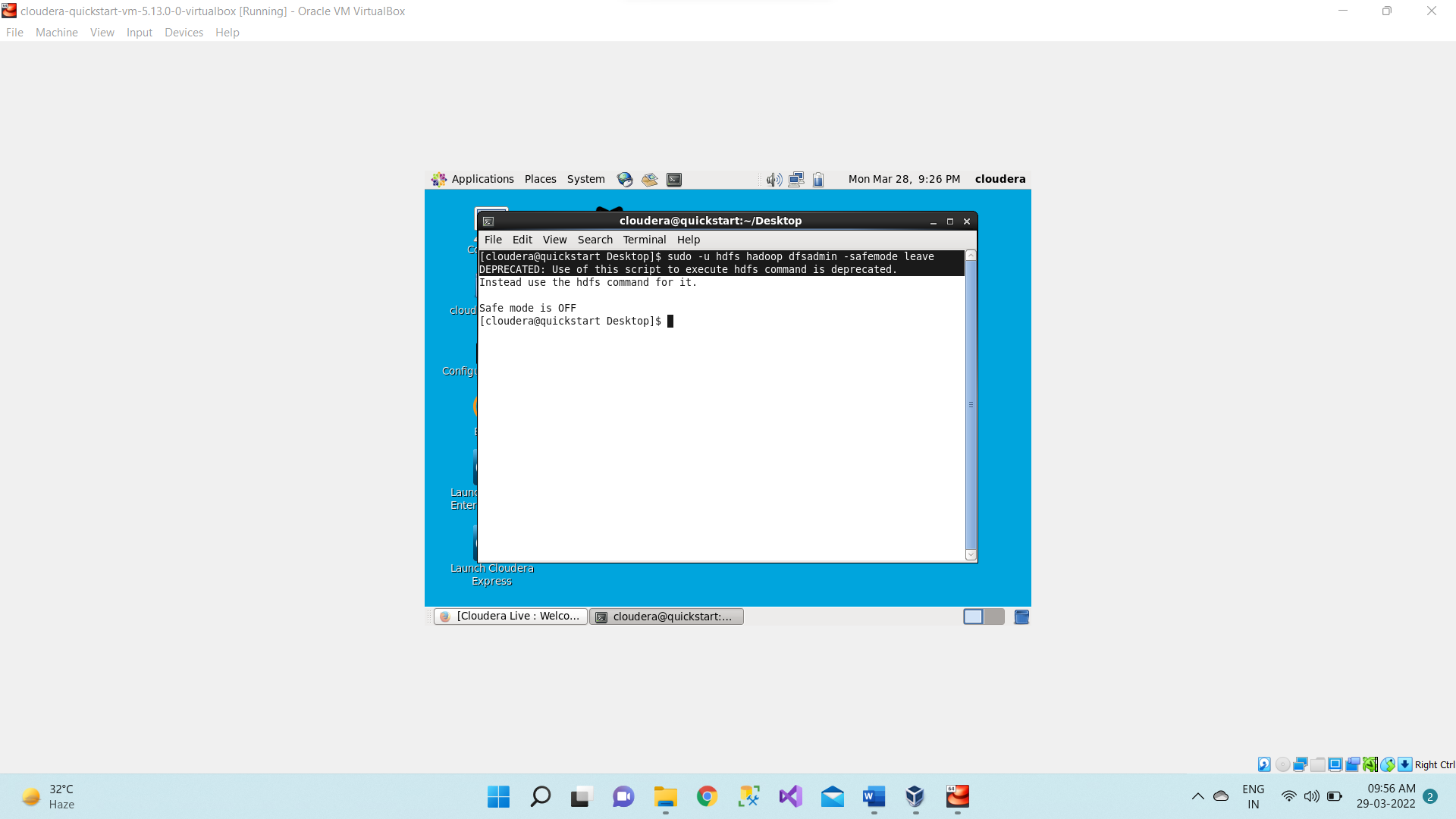
1. **Start all services**

****

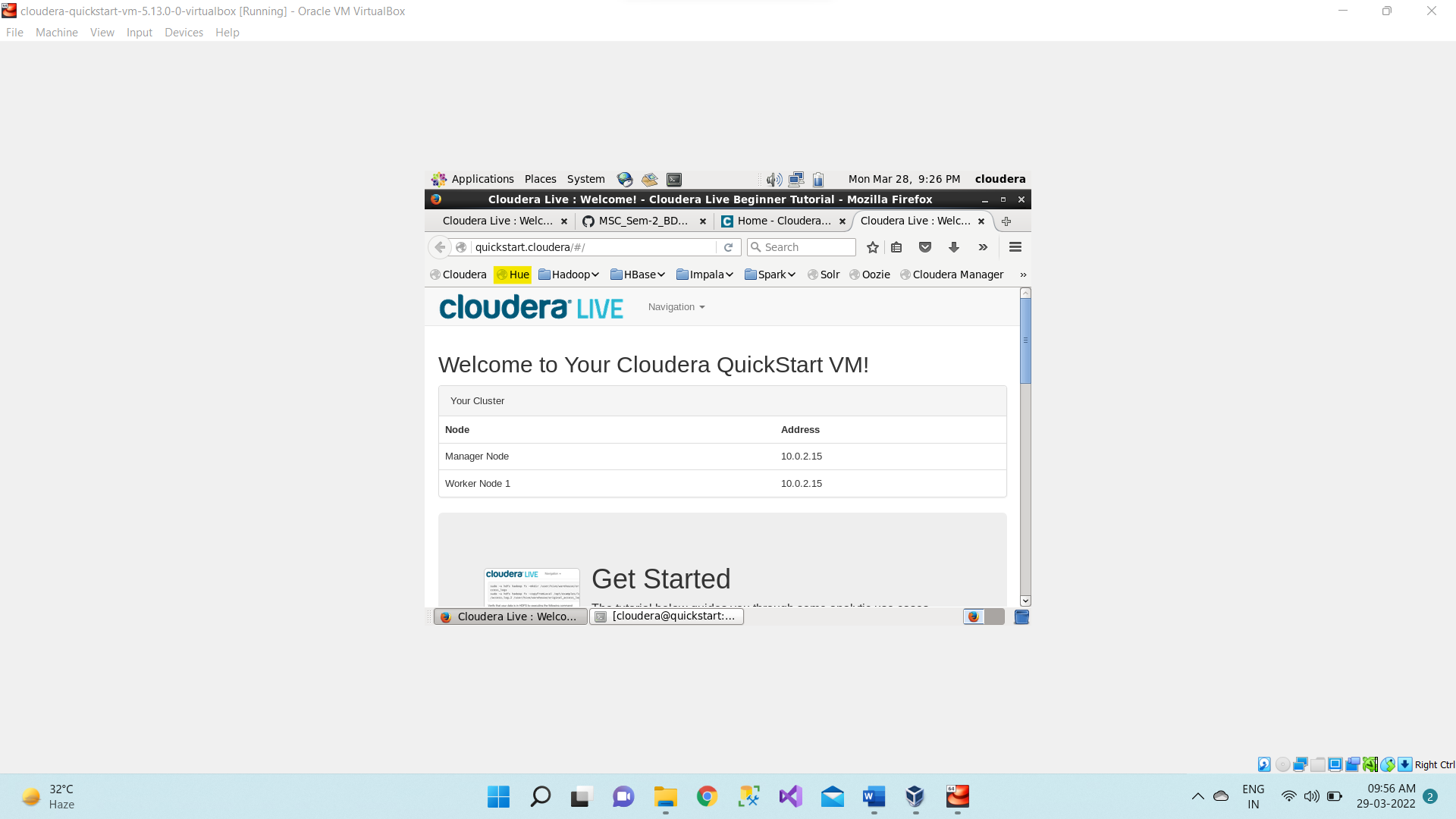
****

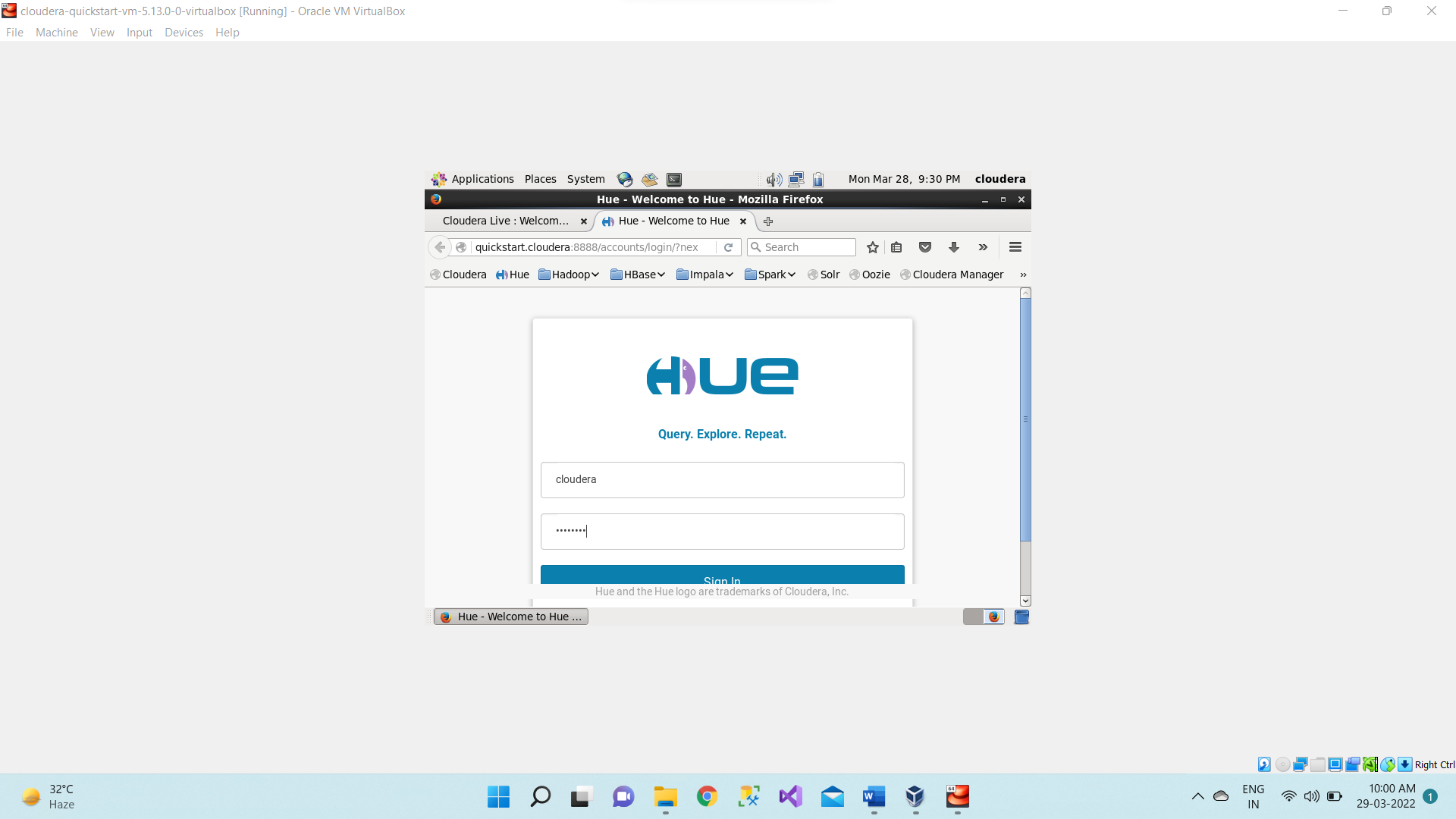
****

1. **sudo -u hdfs hadoop dfsadmin -safemode leave**

****

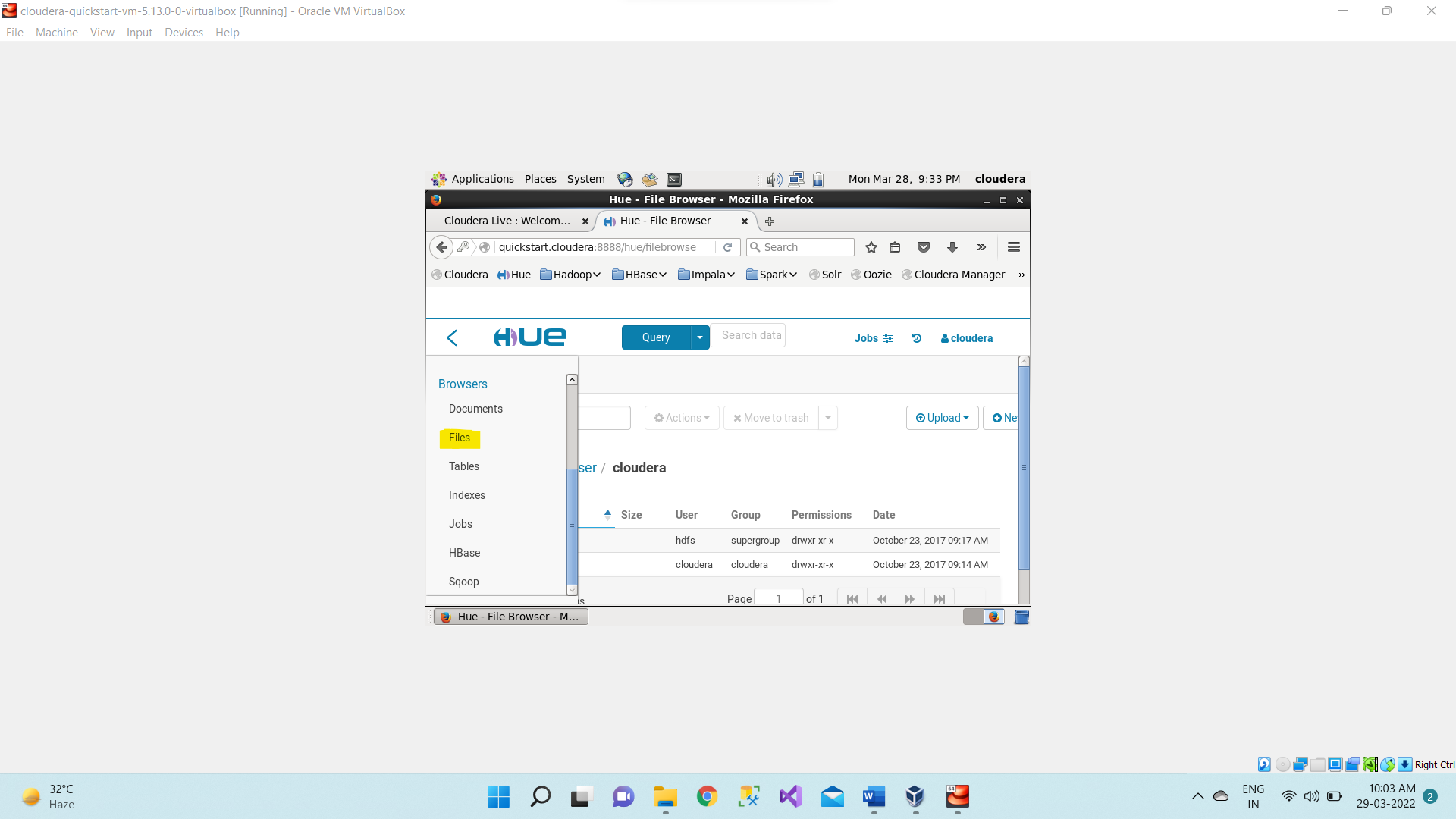
1. Open the browser. And then open Hue and login.

****

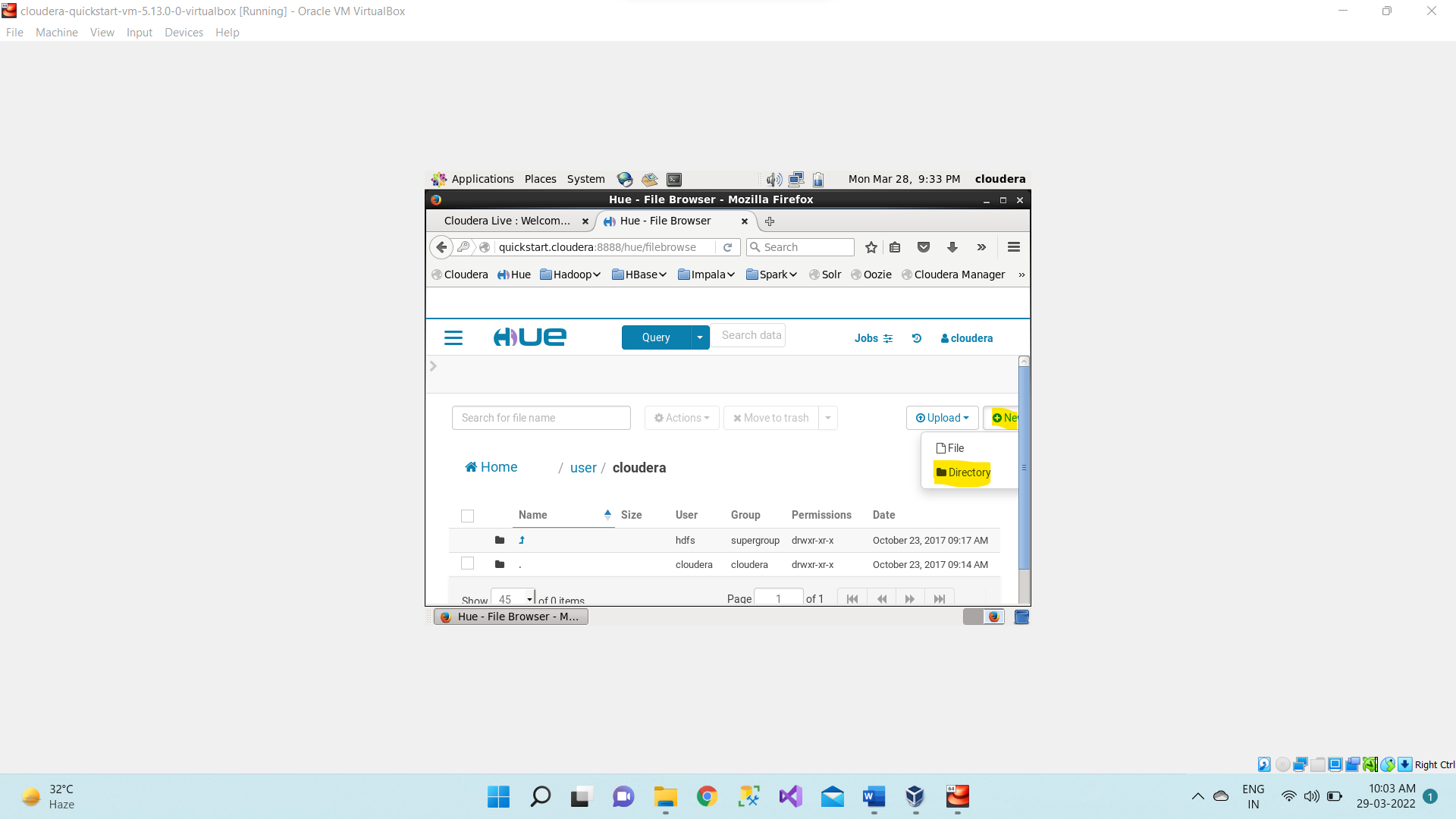
****

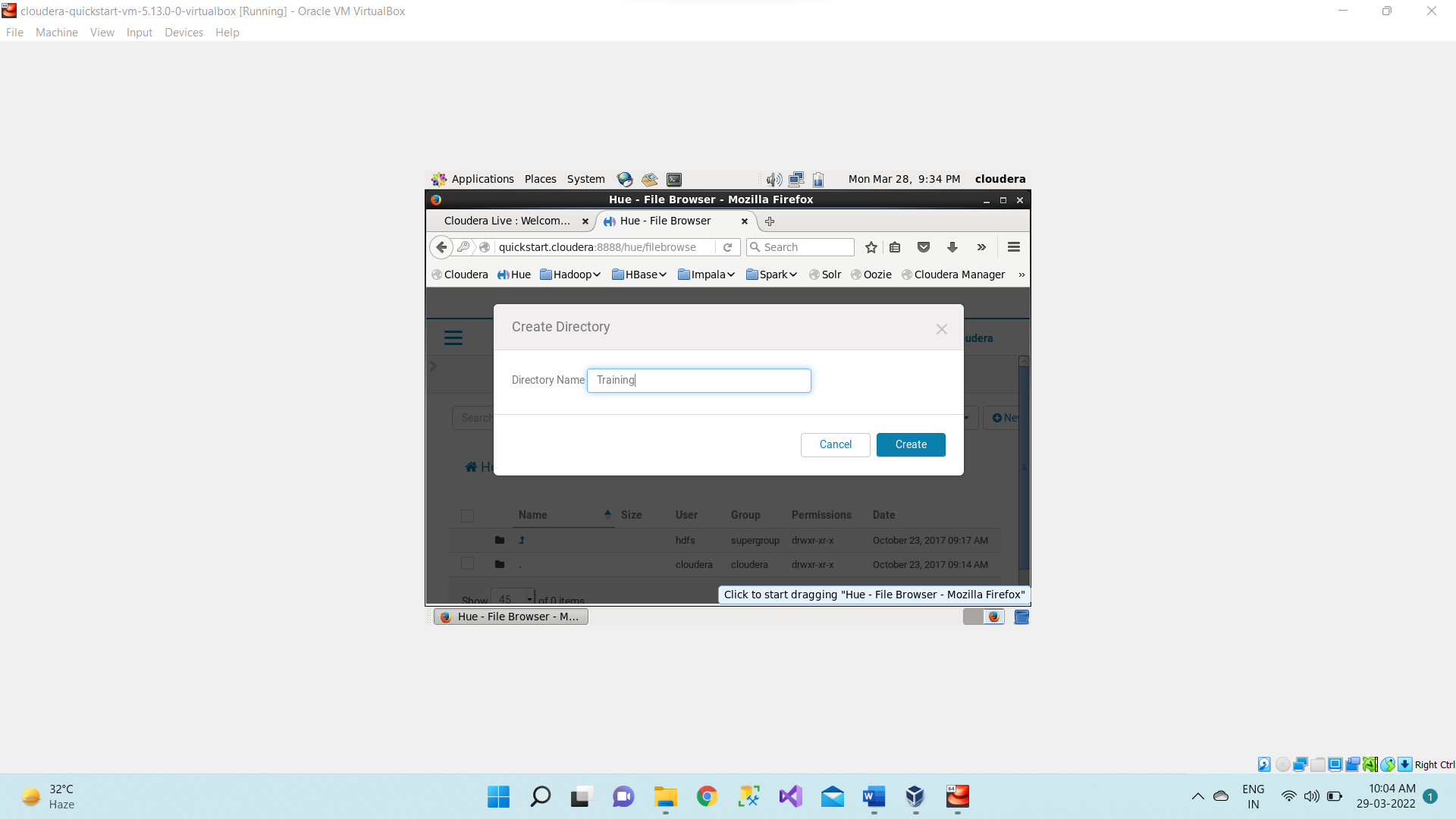
1. Now open the directory /user/cloudera

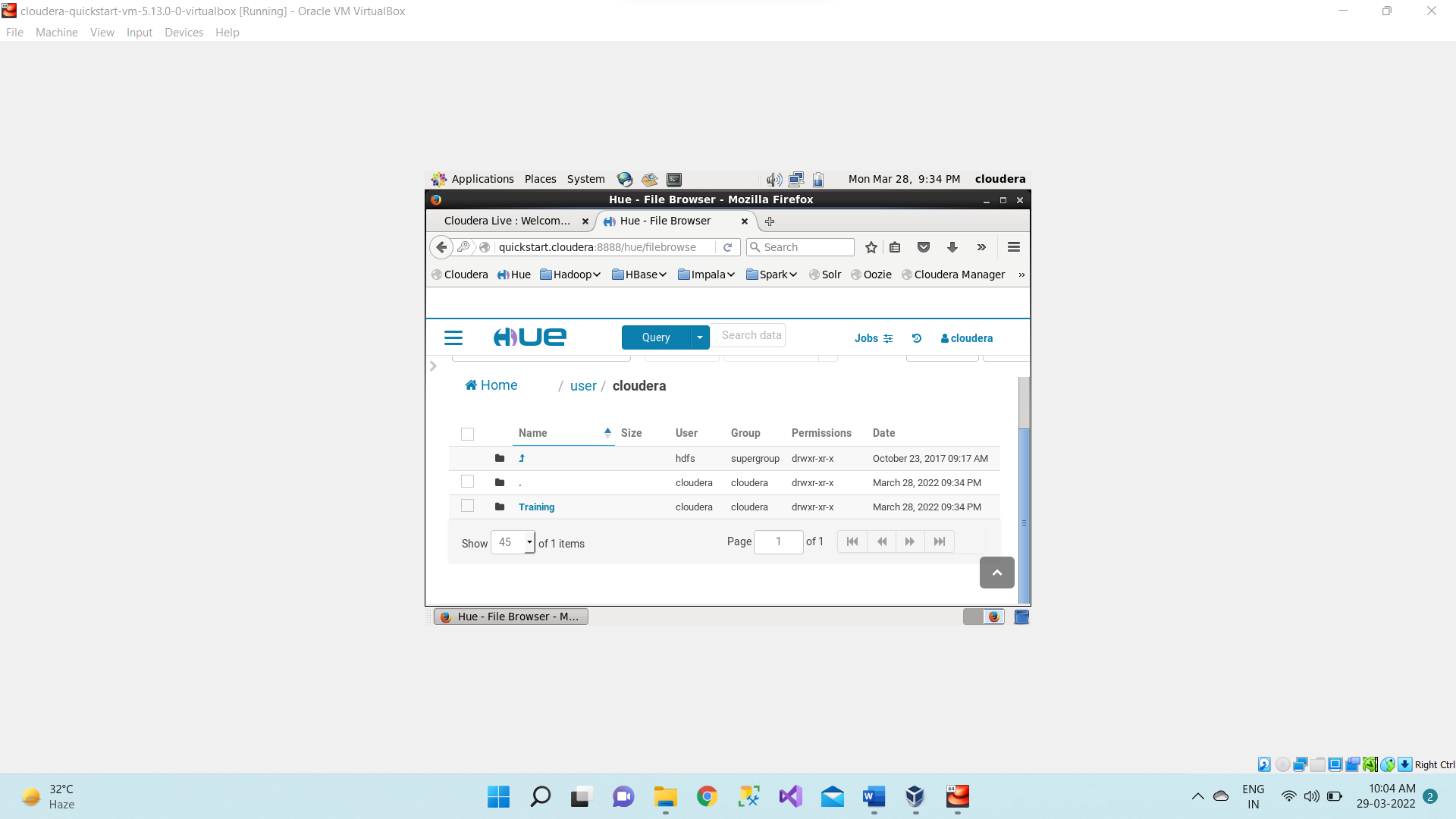
Type /hue/filebrowser

****

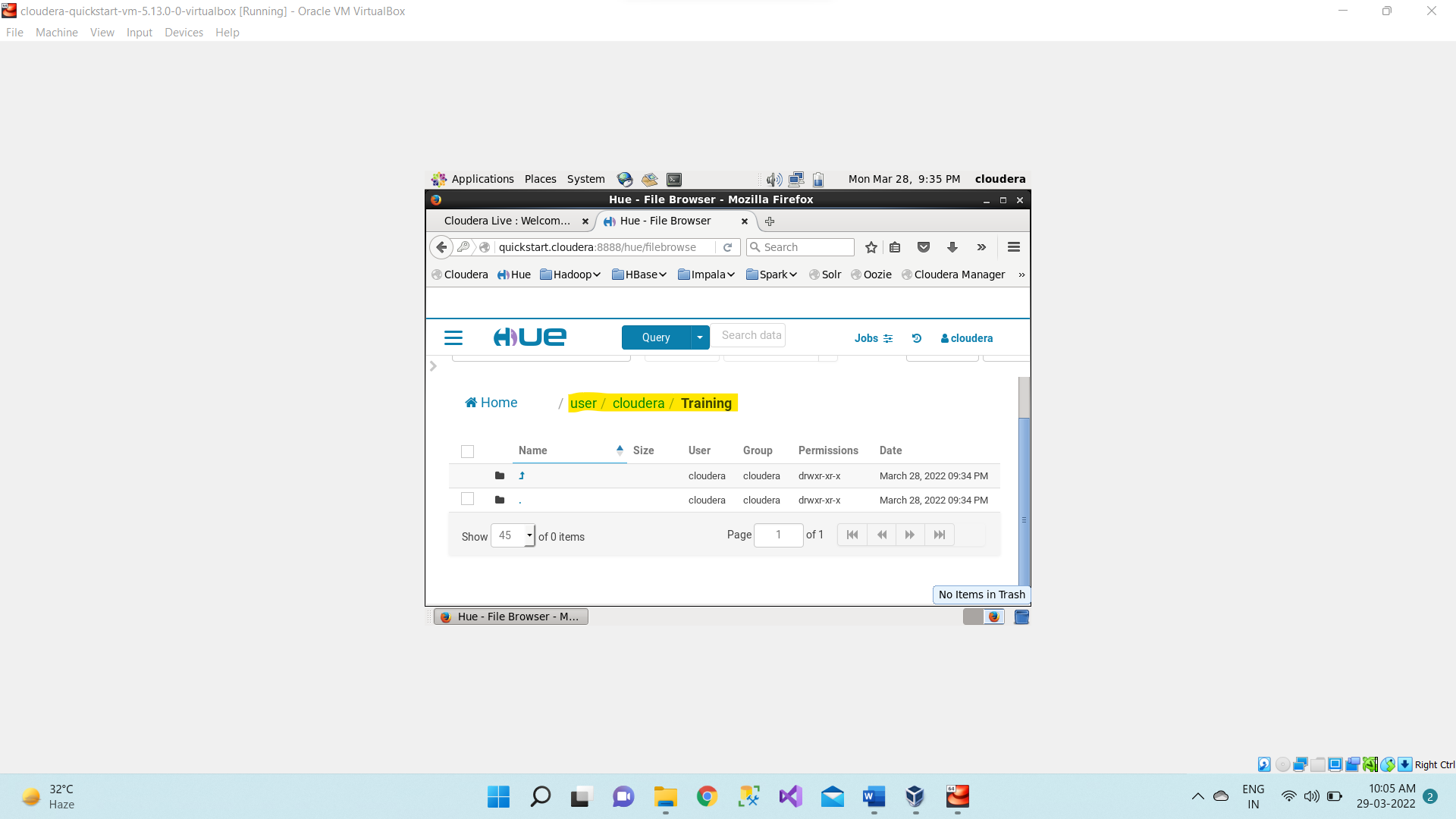
1. Now we are creating the directory as Training inside /user/cloudera

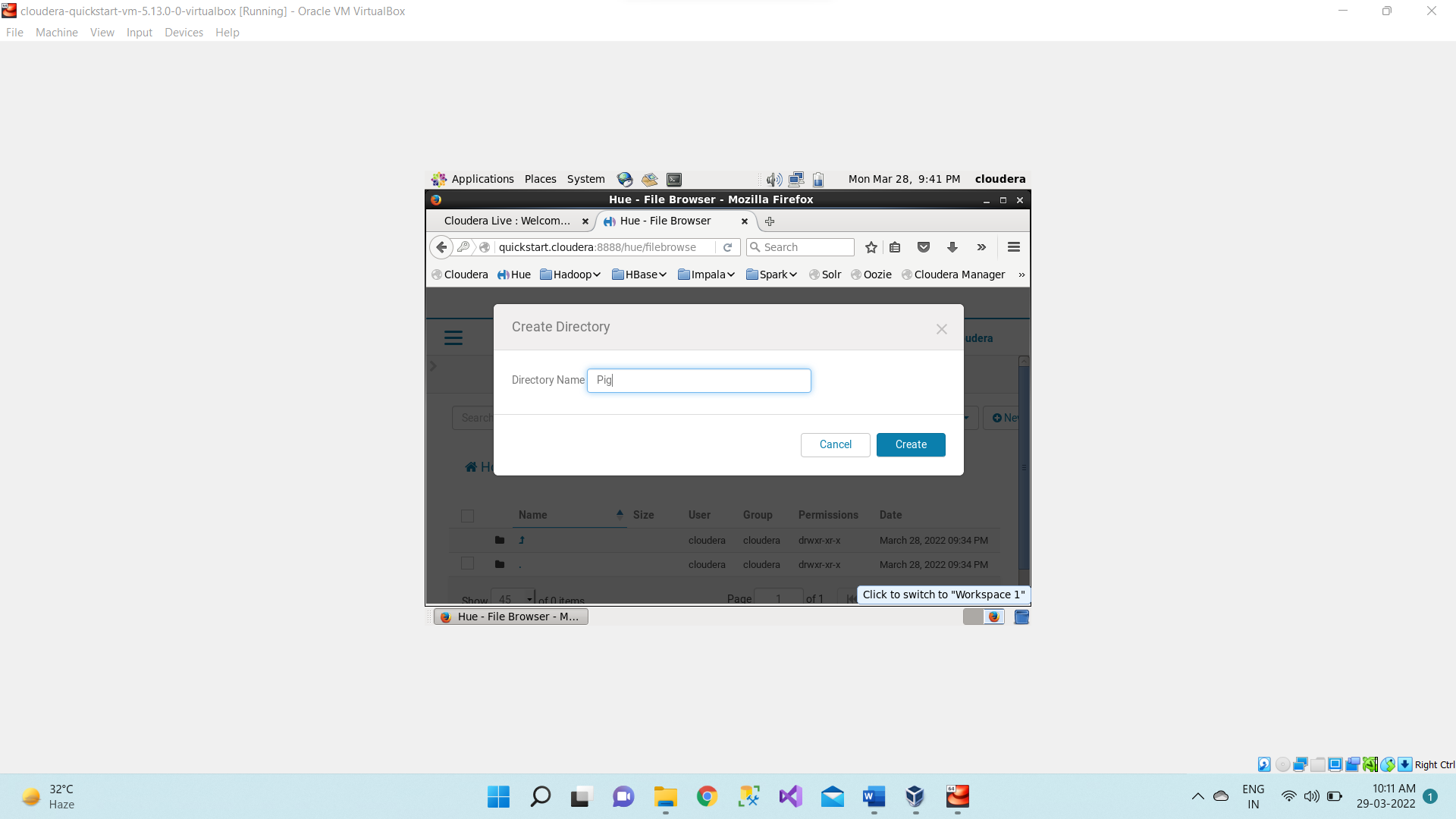
****

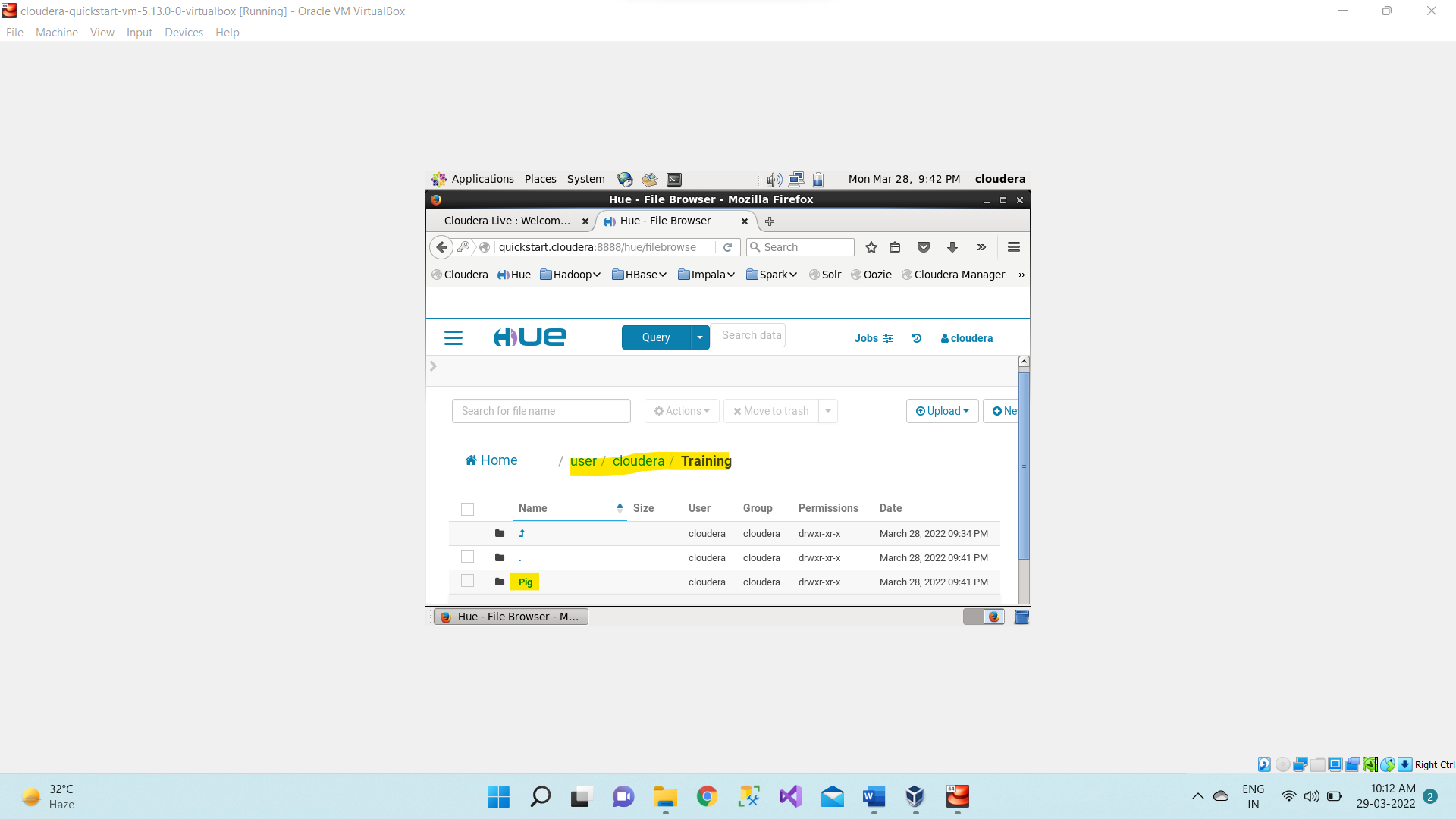
****

****

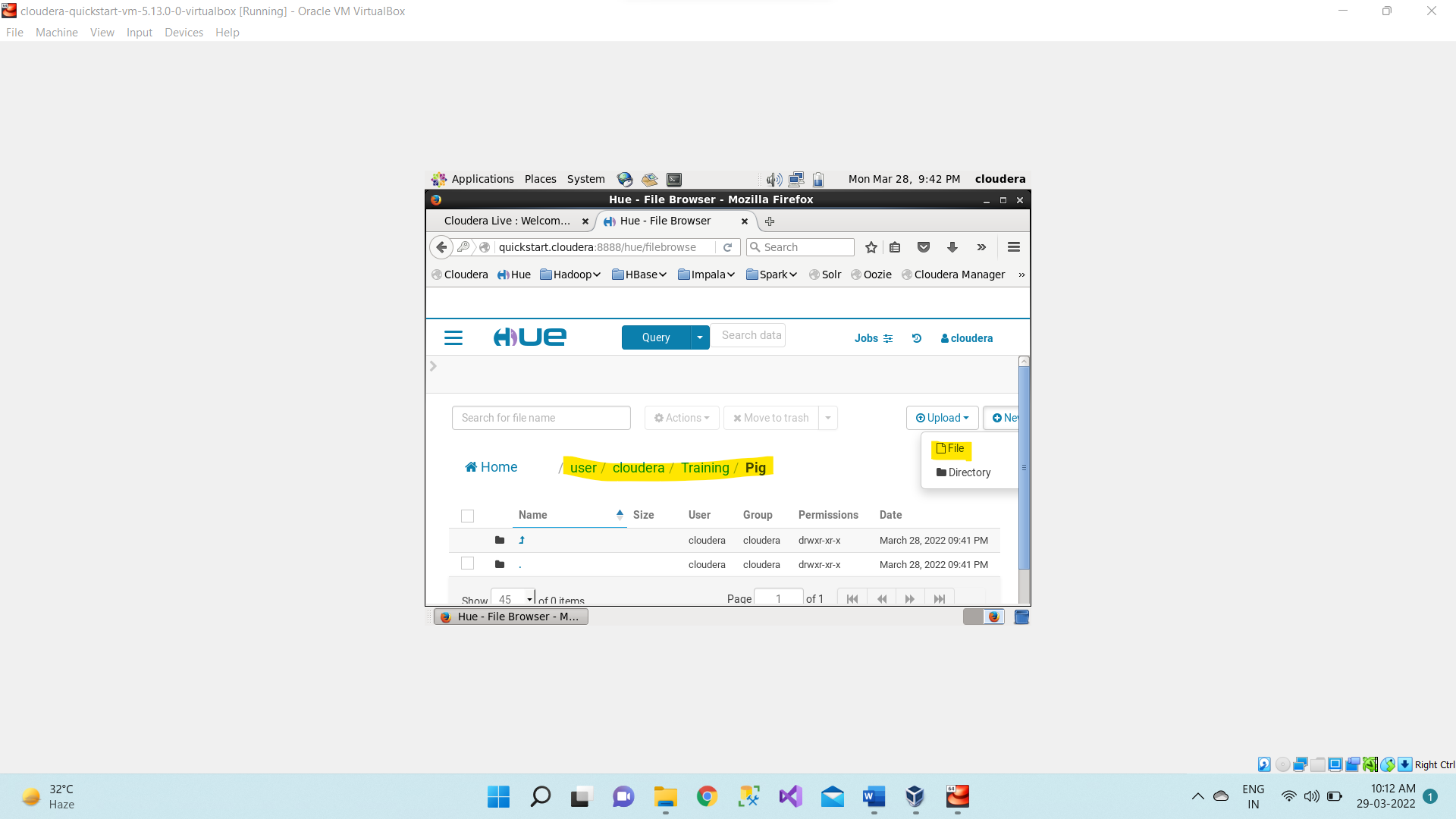
1. After creating Training directory now creating the pig directory inside /user/cloudera/Training

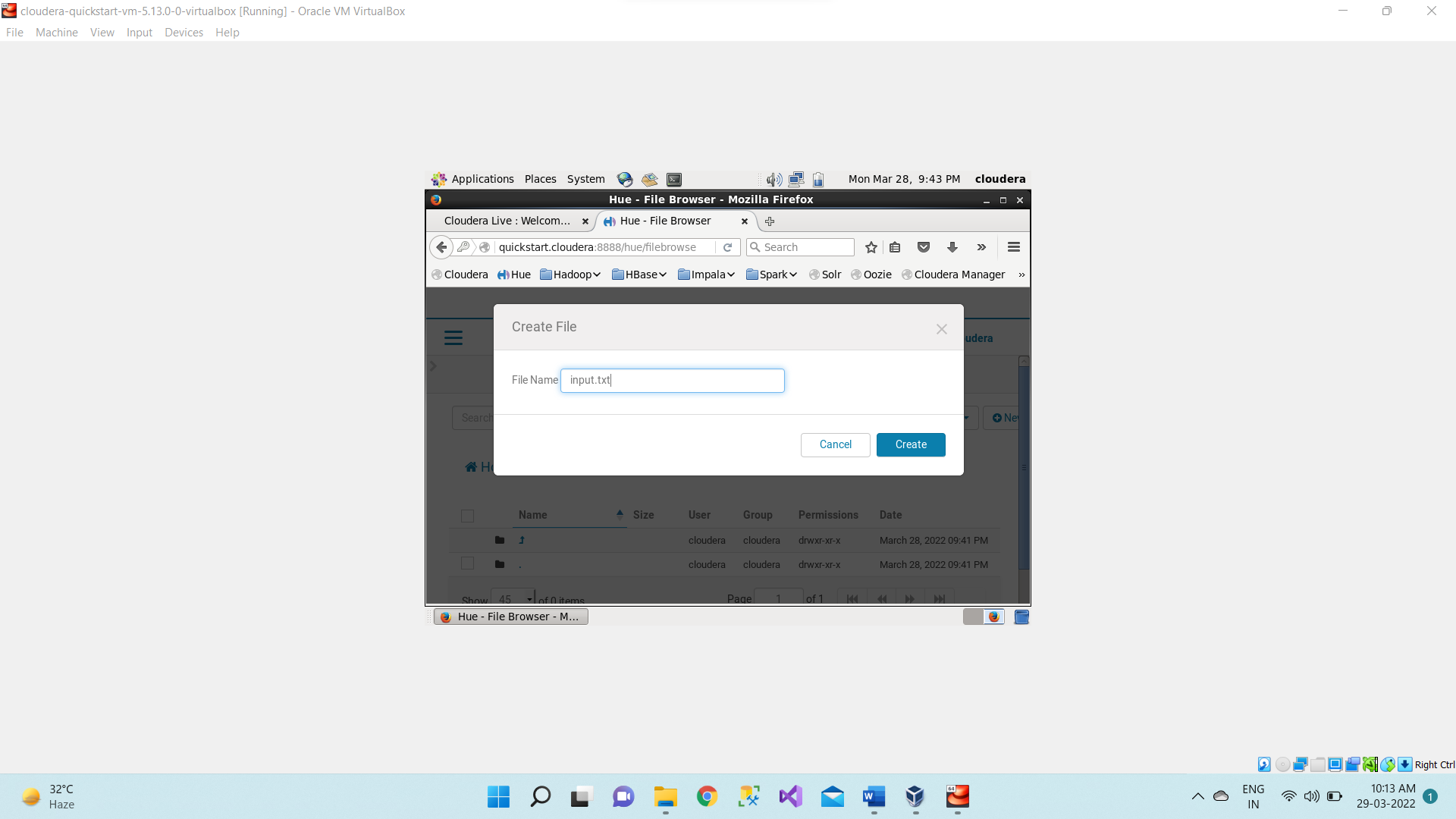
****

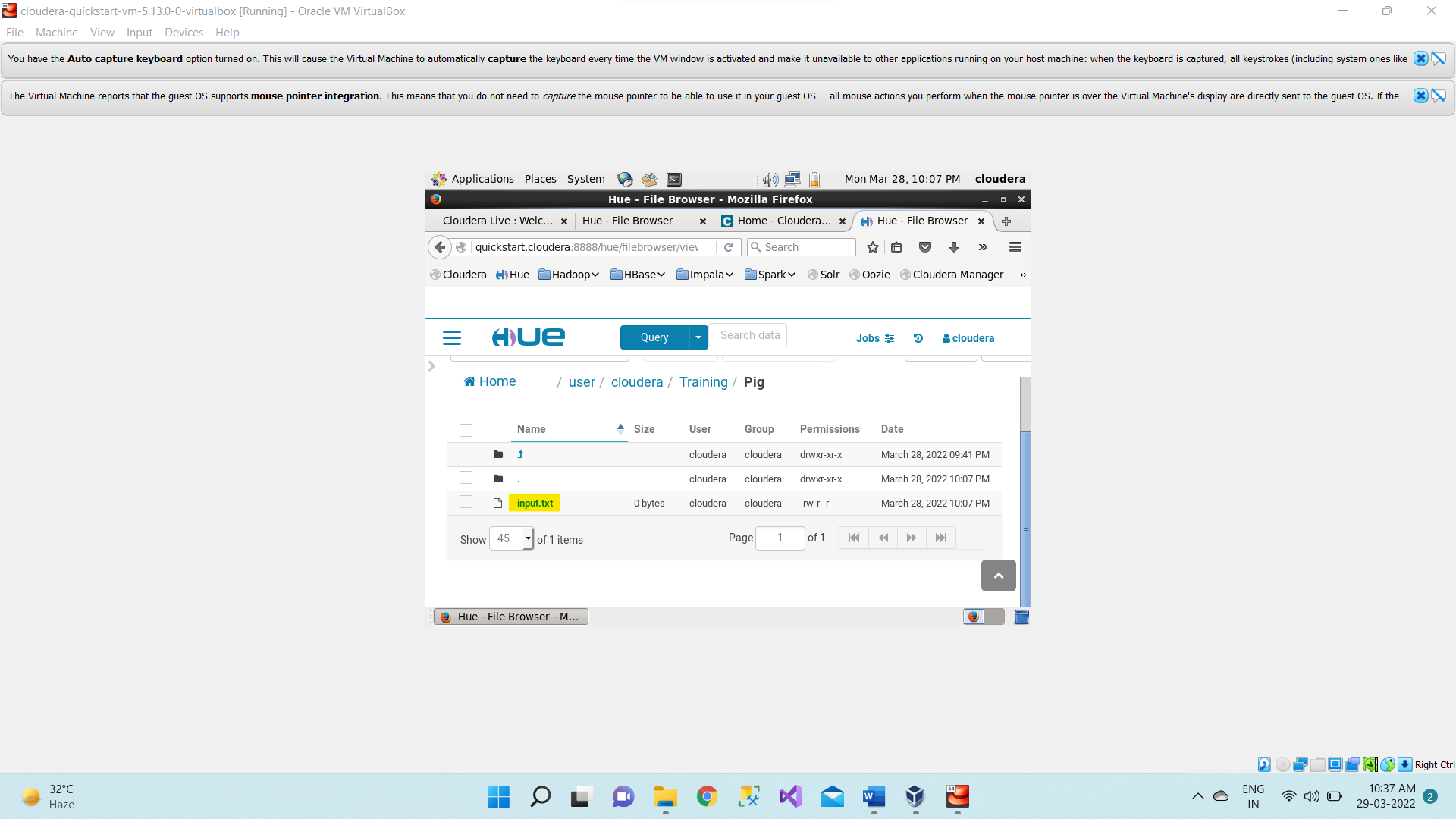
****

pig directory has been created inside /user/cloudera/Training ****

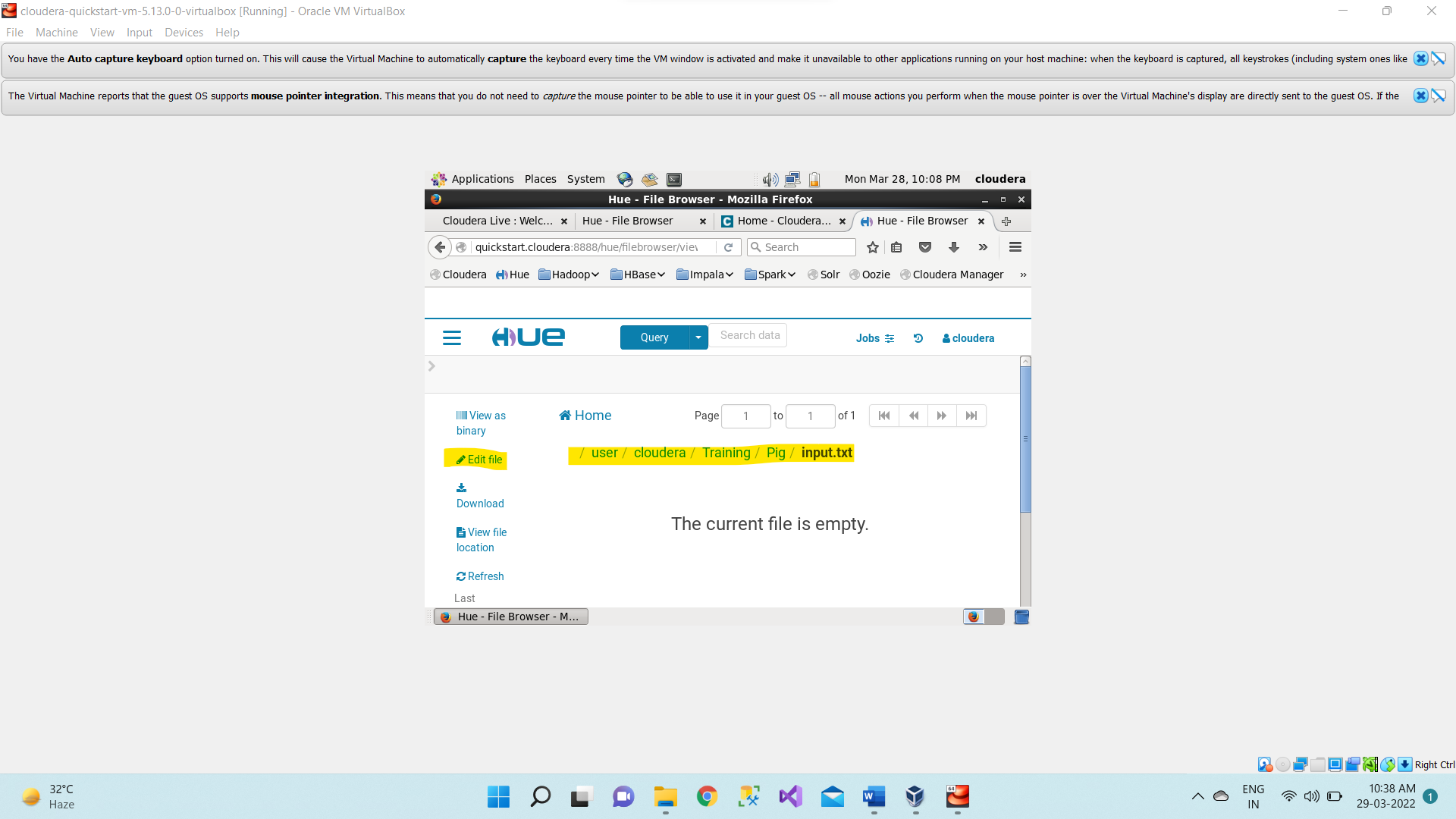
1. Creating **input.txt** file inside /usr/cloudera/training/**pig** directory

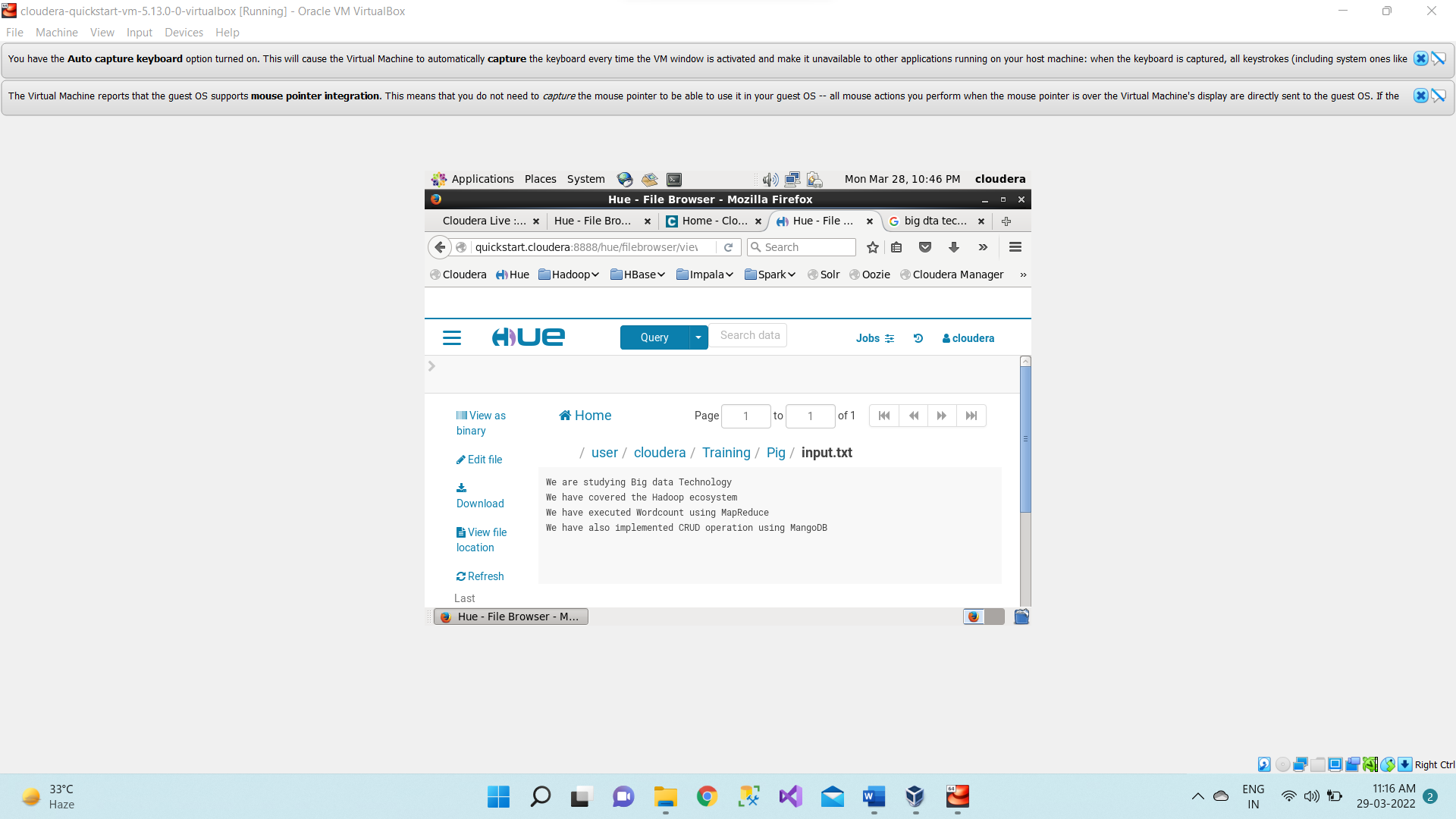
****

****

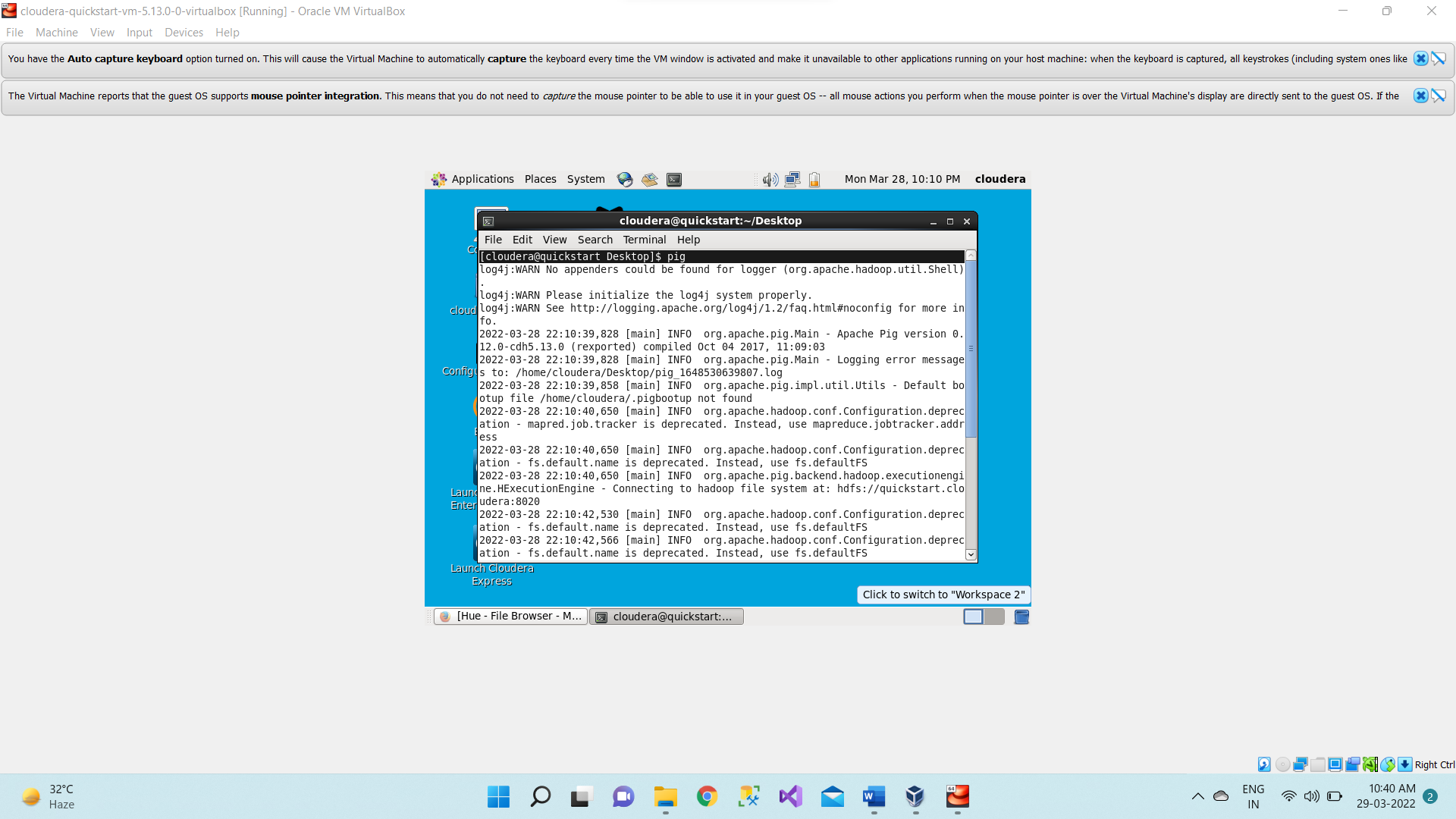
****

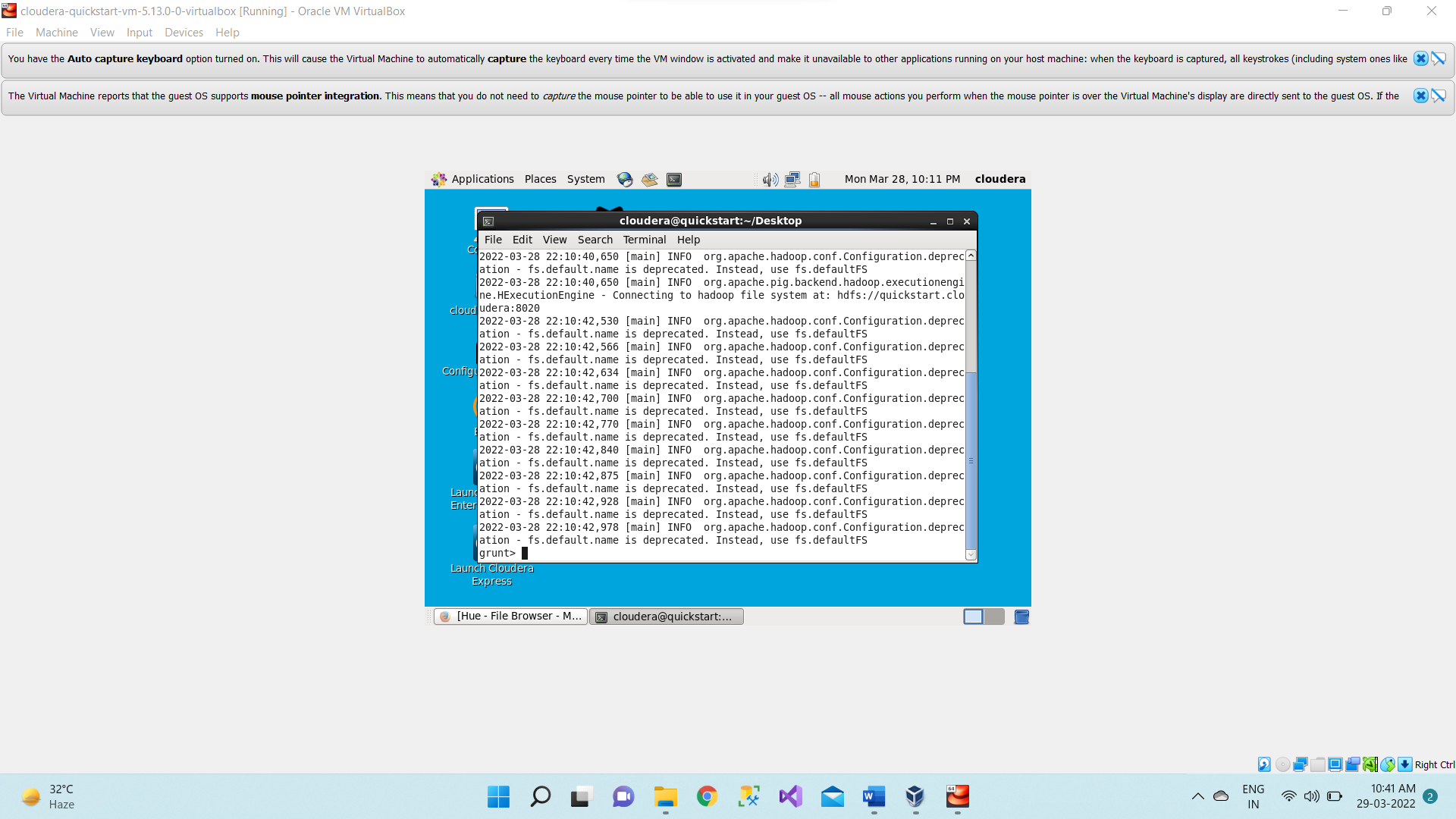
1. Adding some contents to this **input.txt** file.

****



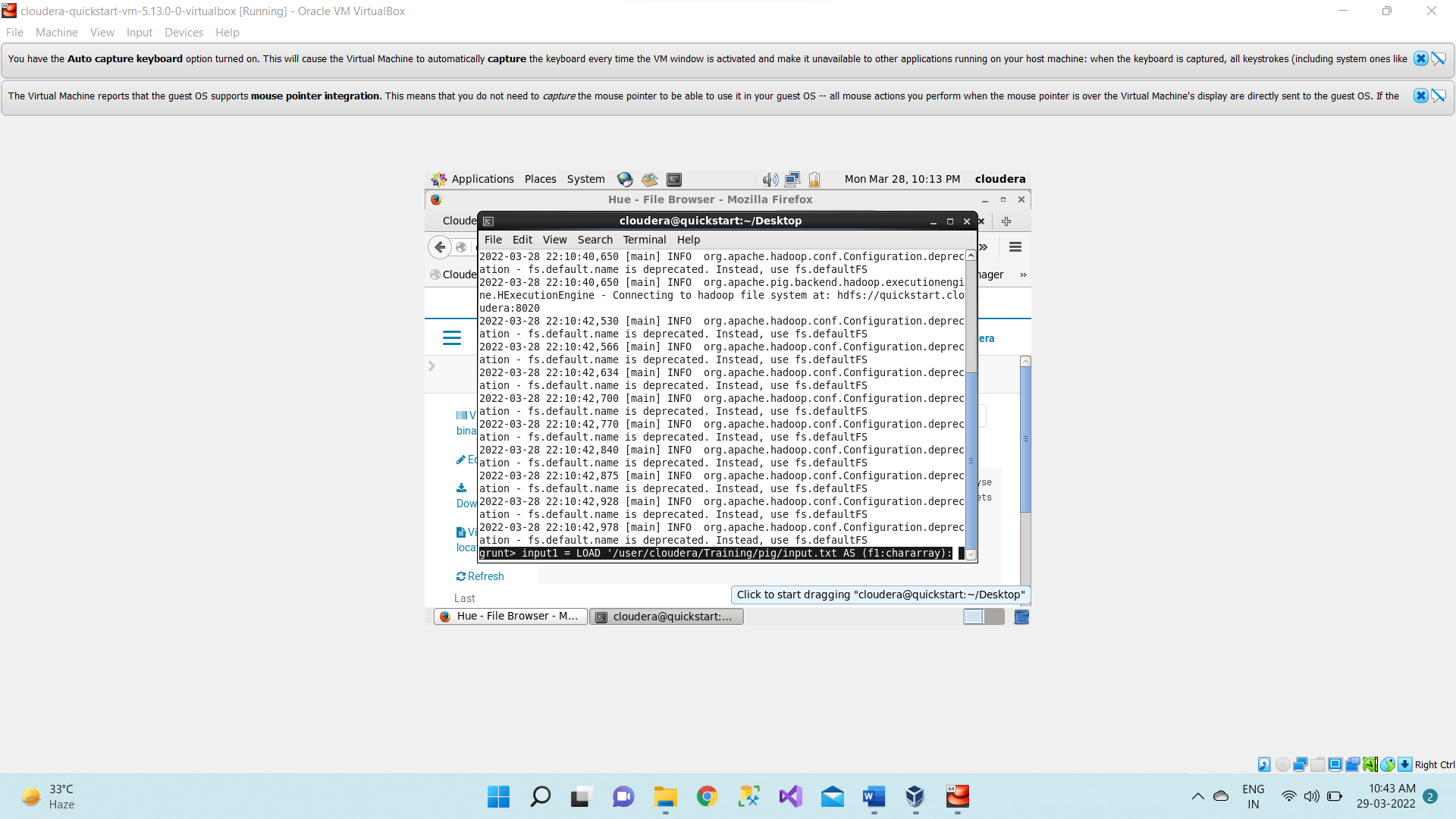
1. Now Open the **terminal**. And start **Pig** by typing pig on terminal.



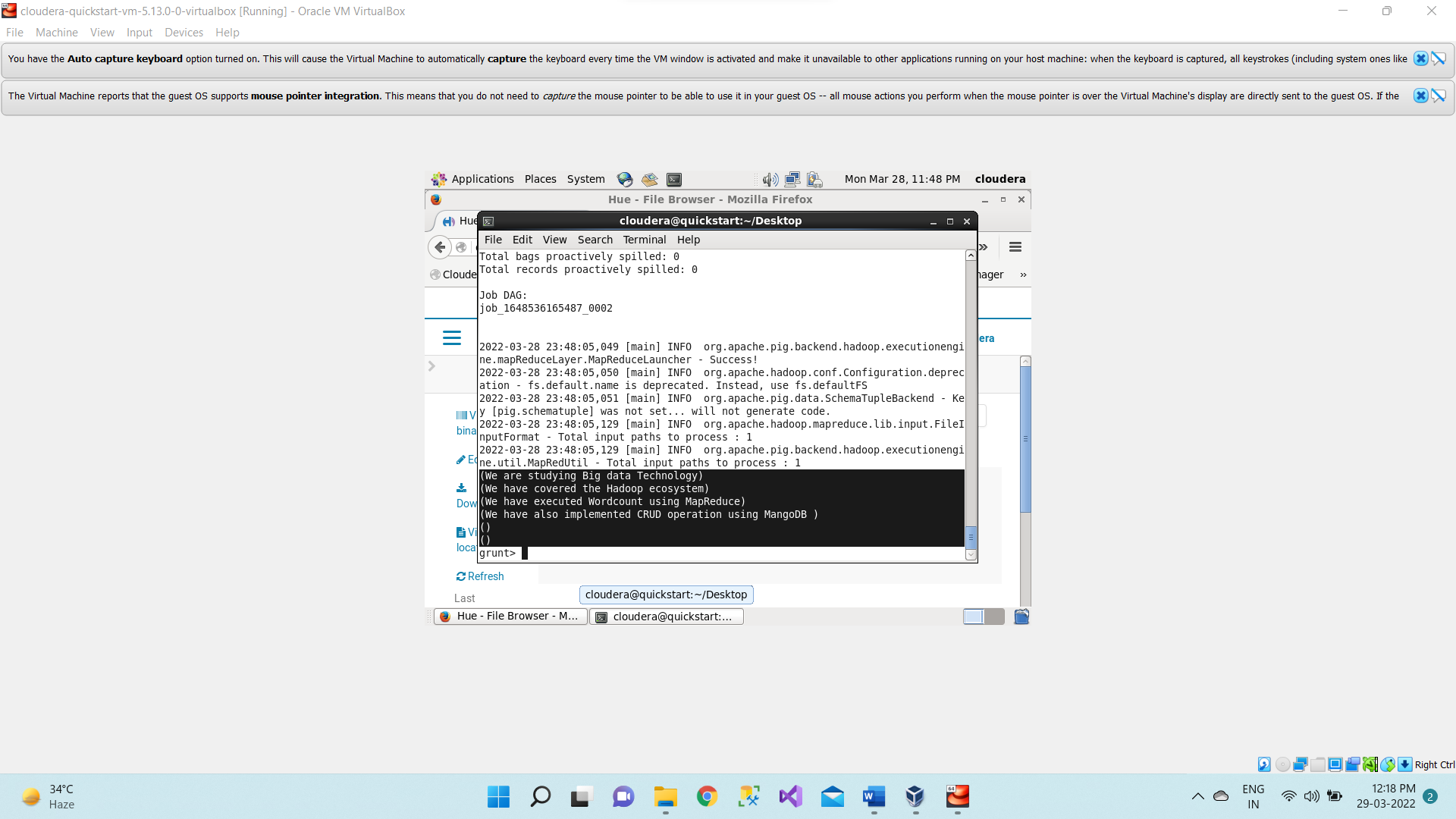
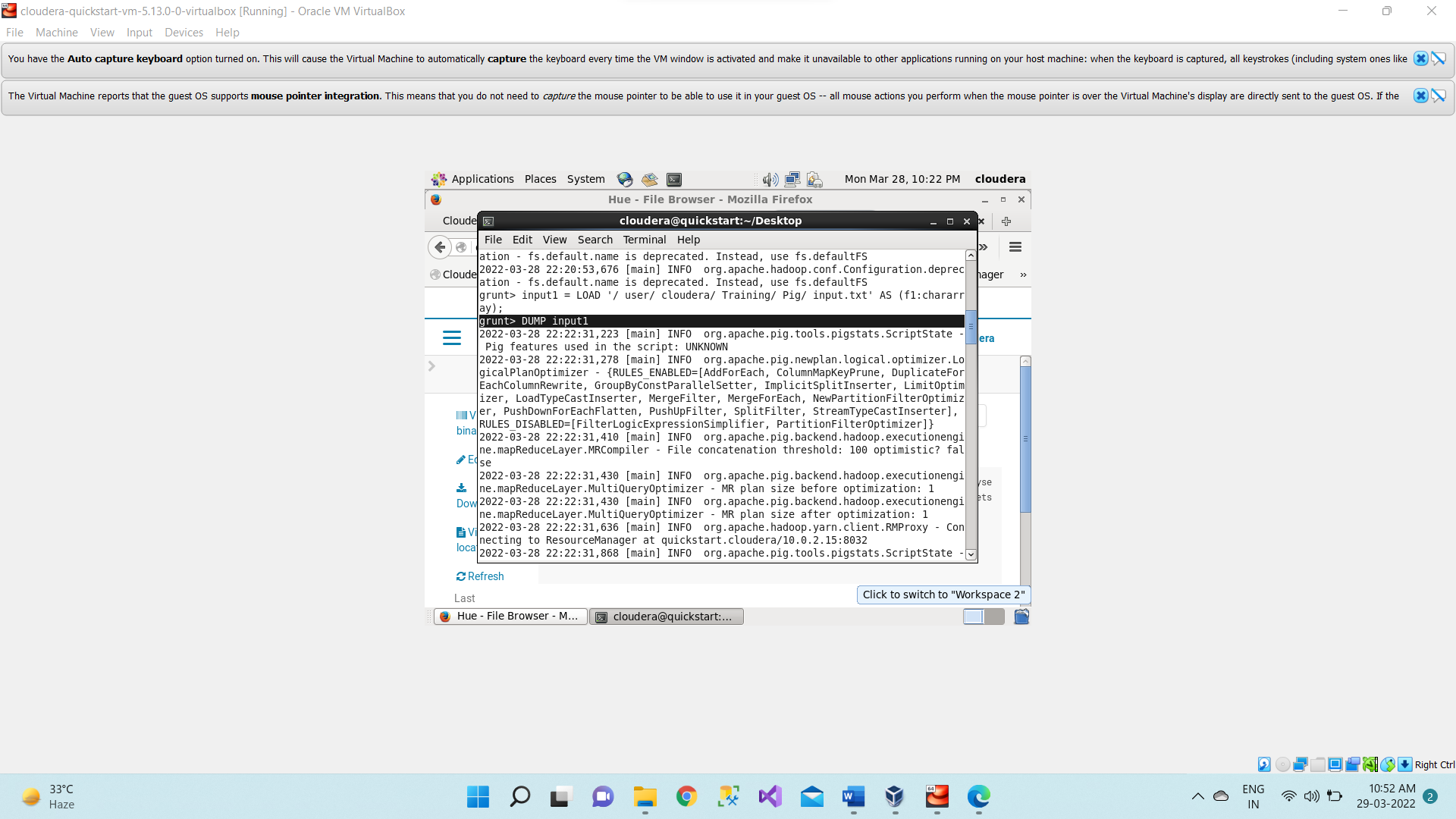
****

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

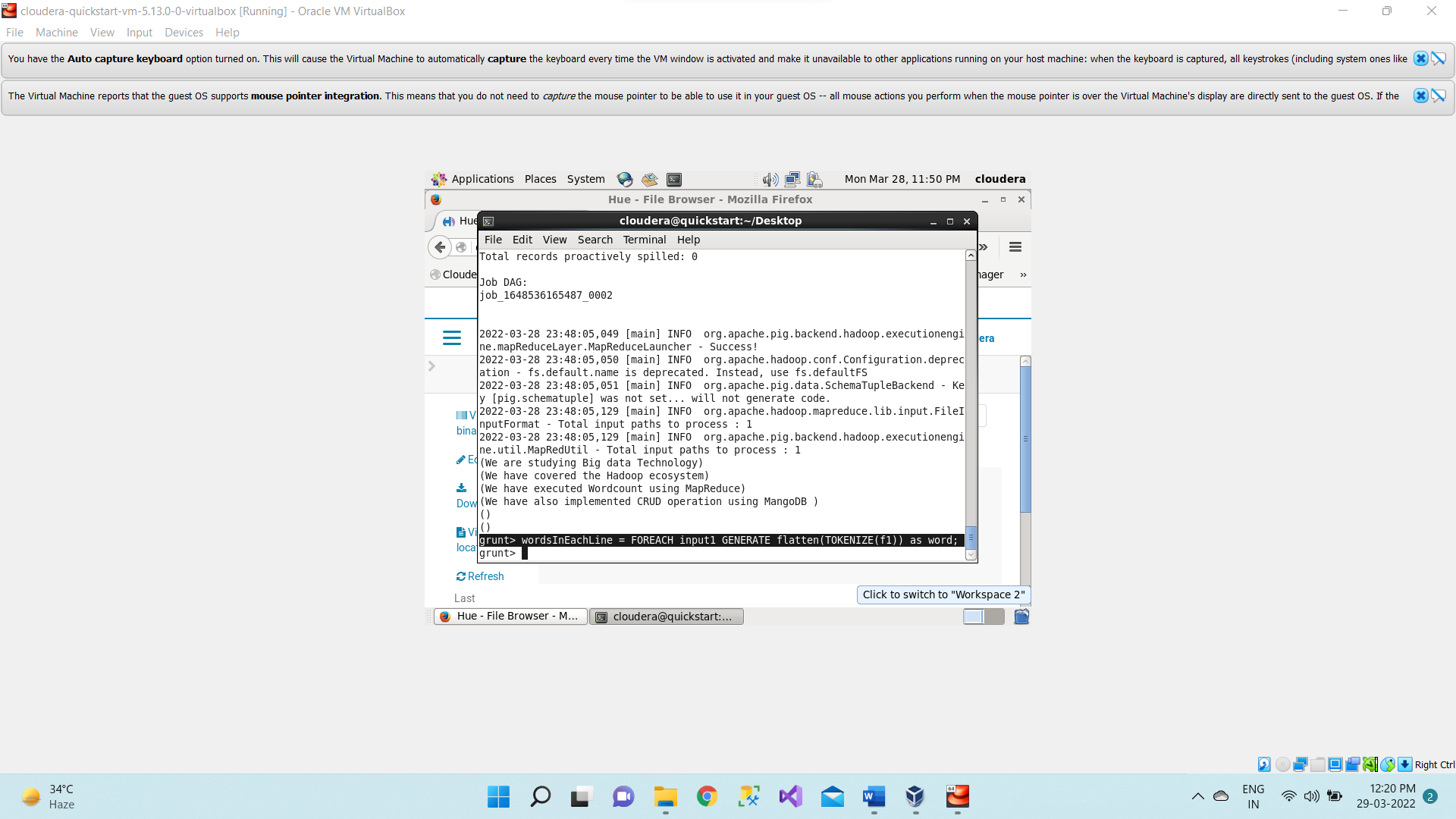
**input1 = LOAD '/user/cloudera/Training/pig/input.txt' AS (f1:chararray);**

****

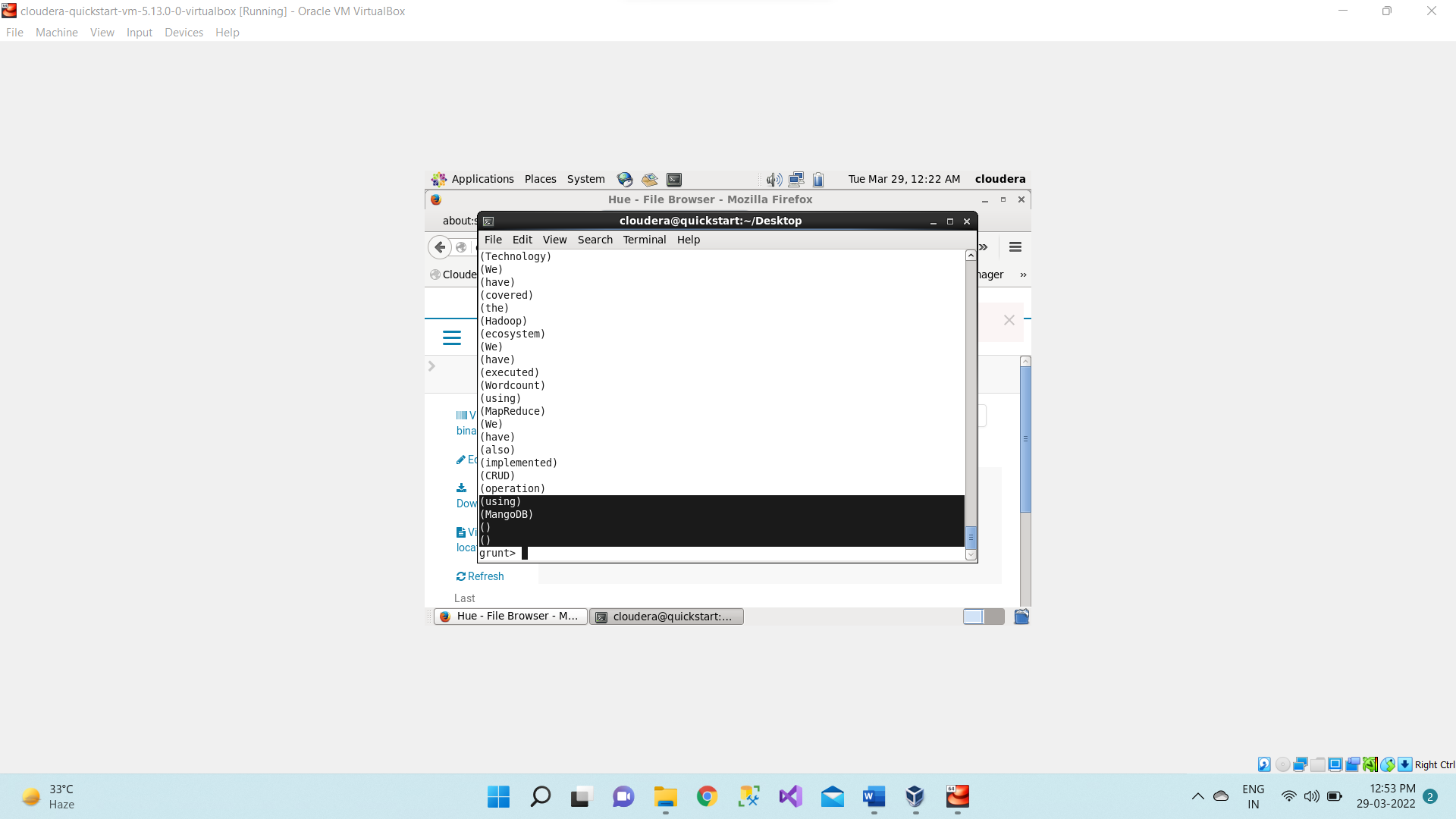
1. Now we are dumping the data. It will be done the MapReduce task.

**DUMP input1**

1. **wordsInEachLine = FOREACH input1 GENERATE flatten(TOKENIZE(f1)) as word;**

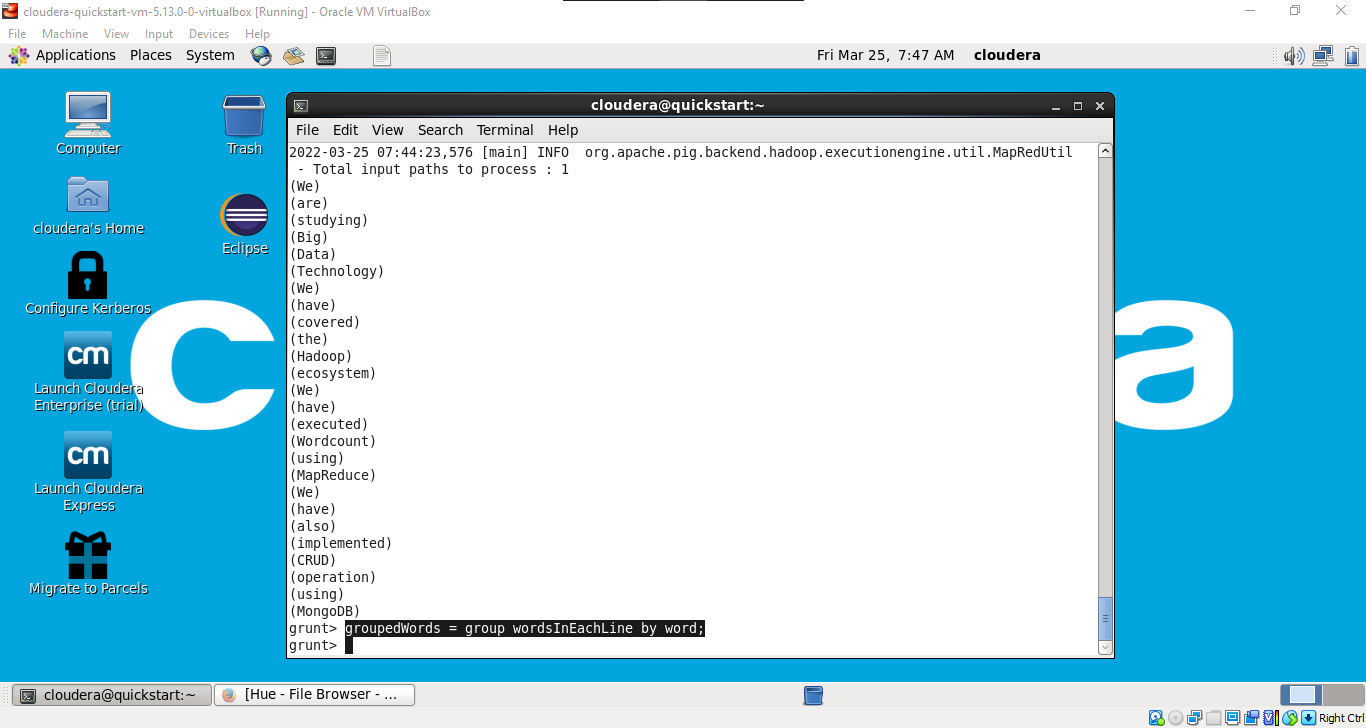
****

1. **DUMP wordsInEachLine;**

****

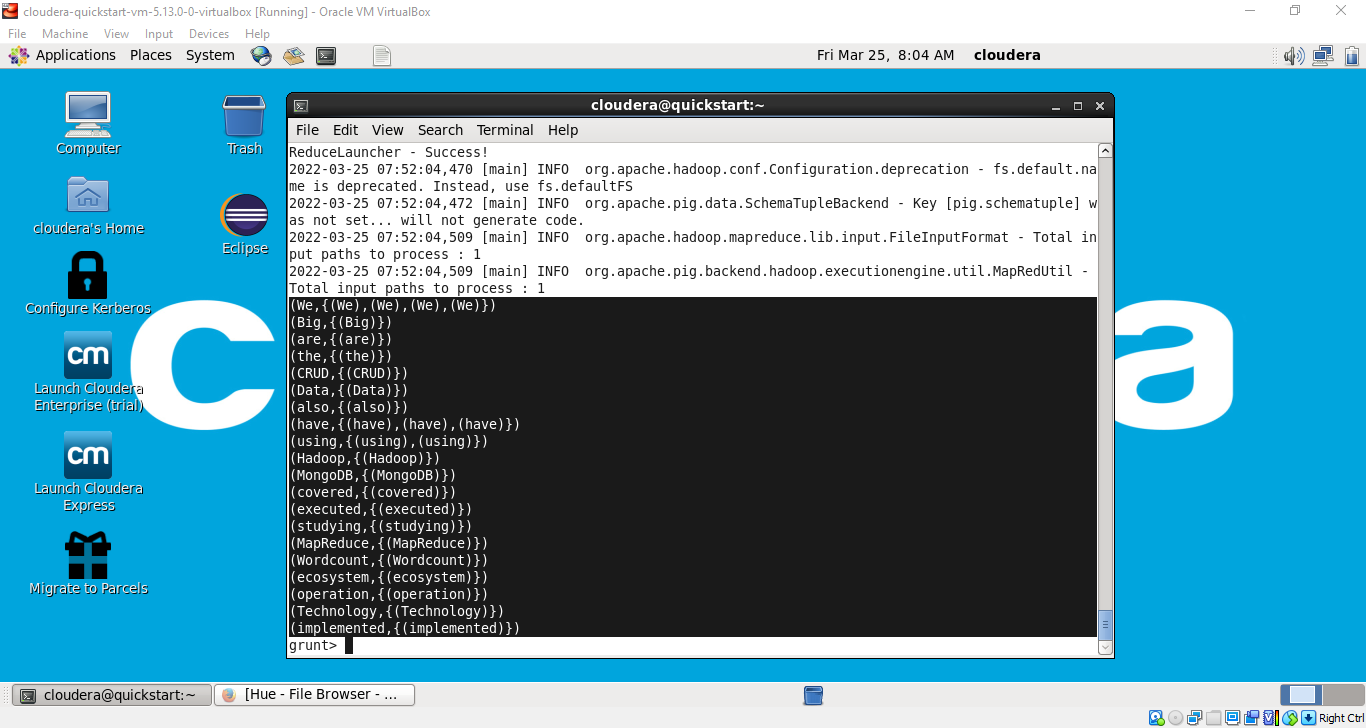
1. Now grouping the words present in each line.

**groupedWords = group wordsInEachLine by word;**

****

1. **DUMP groupedWords;**

****

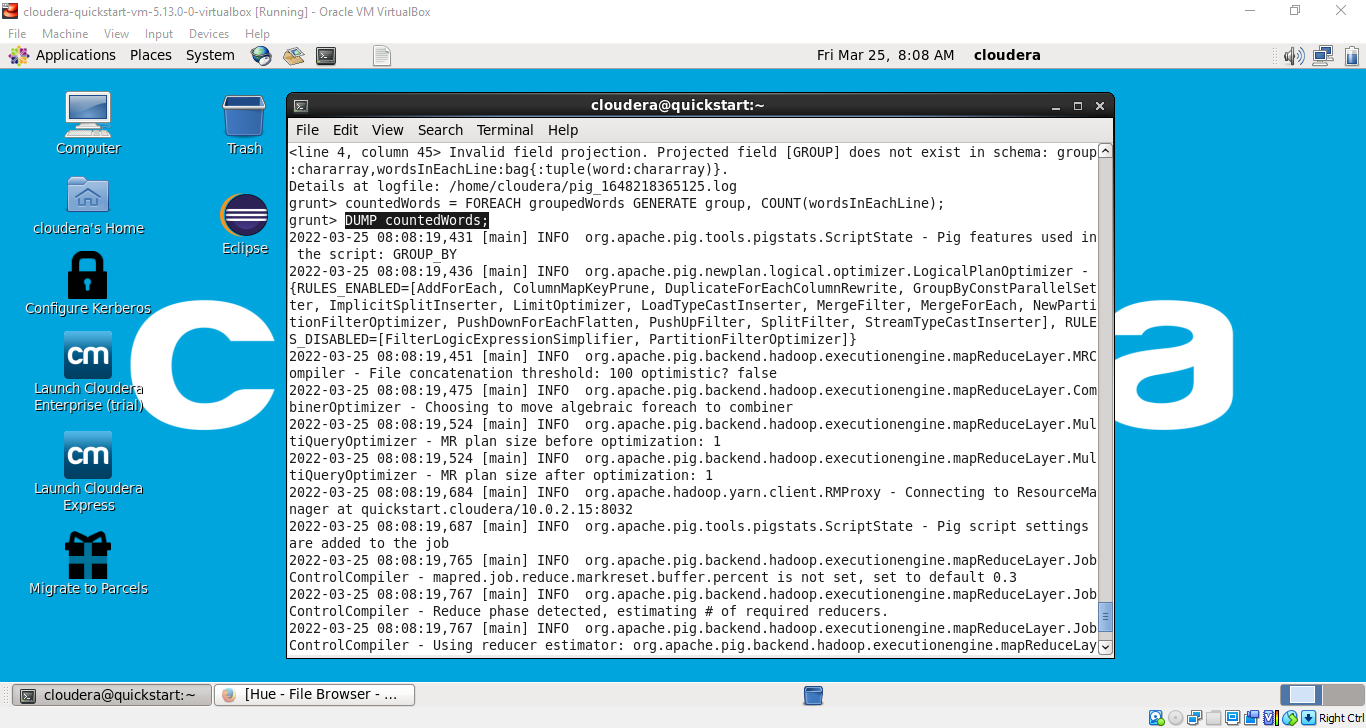
****

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

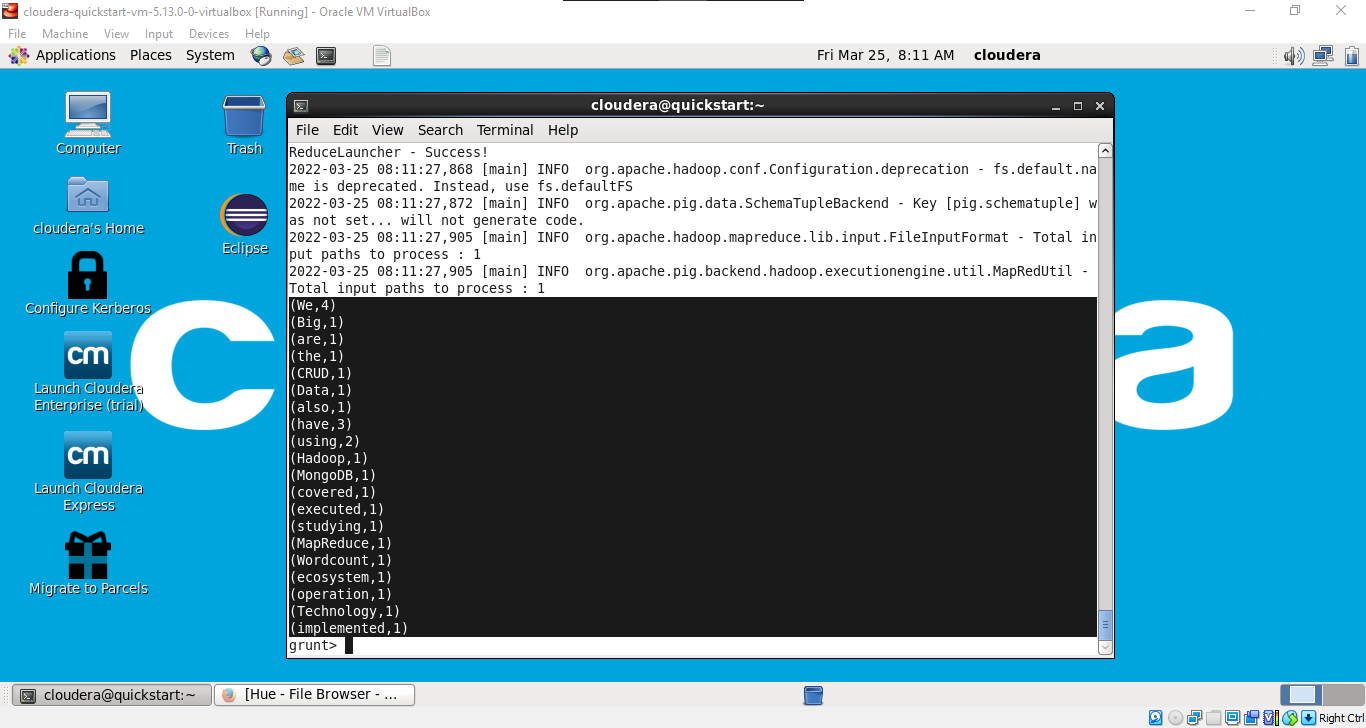
**countedWords = FOREACH groupedWords GENERATE group, COUNT(wordsInEachLine);**

1. **DUMP countedWords;**

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

****

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

****

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