**Big Data: LabWork1**

**Author’s Name:** Shivani Sharma

**Course Number:** ALY6110

**Course Title:** Data Management and Big Data

**Academic Term:** Fall 2019 CPS Analytics

**Instructor’s Name:** Mr. Daya Rudhramoorthi

**Assignment Completion Date:** 04-26-2020

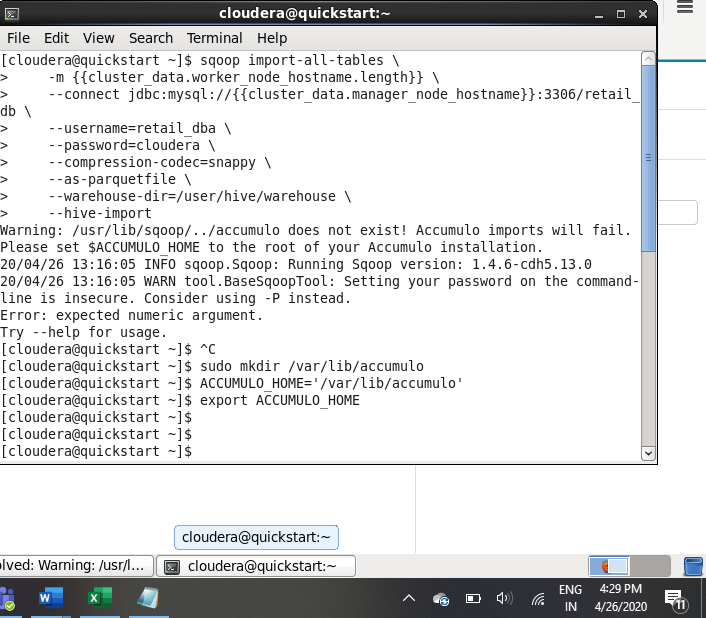


**Summary**

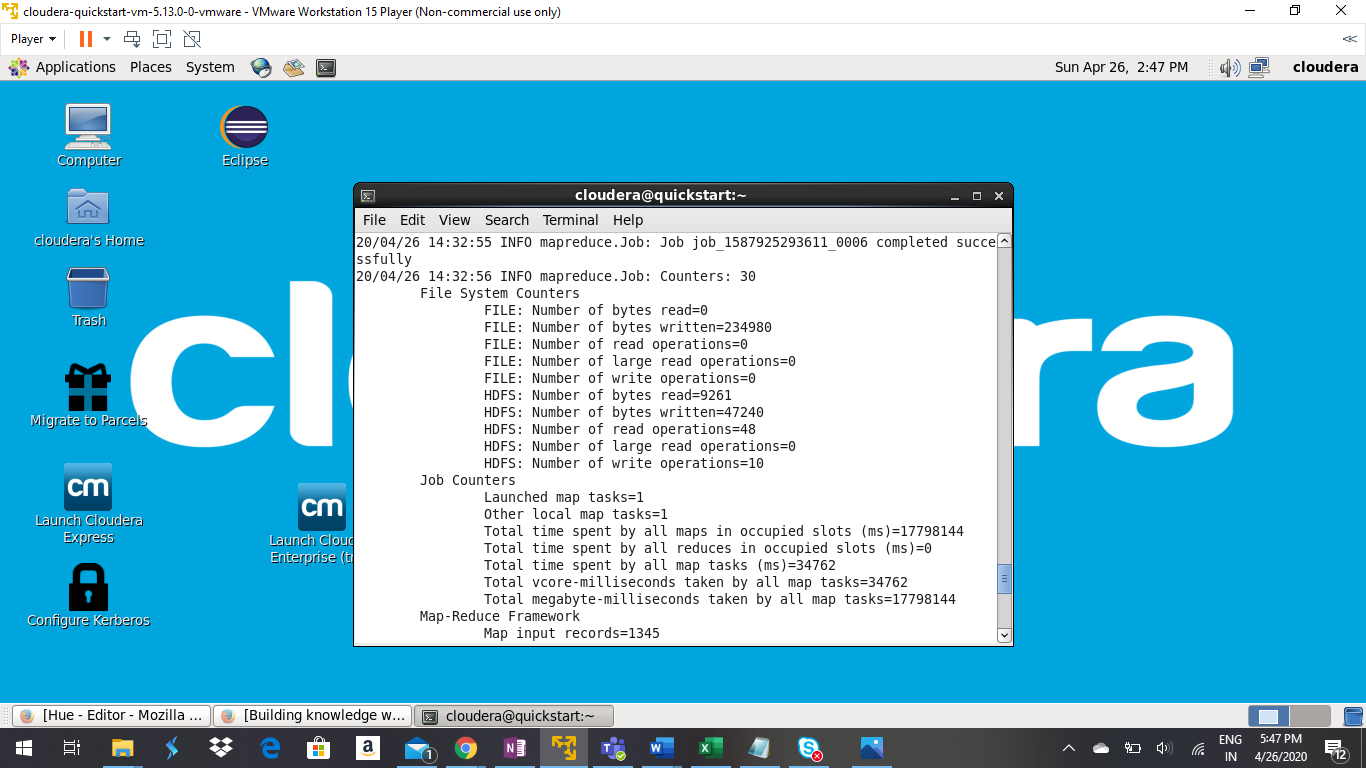
This assignment is based on the problem statement where we need to import a database via Sqoop into HDFS and then from HDFS we need to use the table into IMPALA . From the table categories we need to answer the problem statement. I would like to explain the problem statements in the analysis section.

**Analysis**

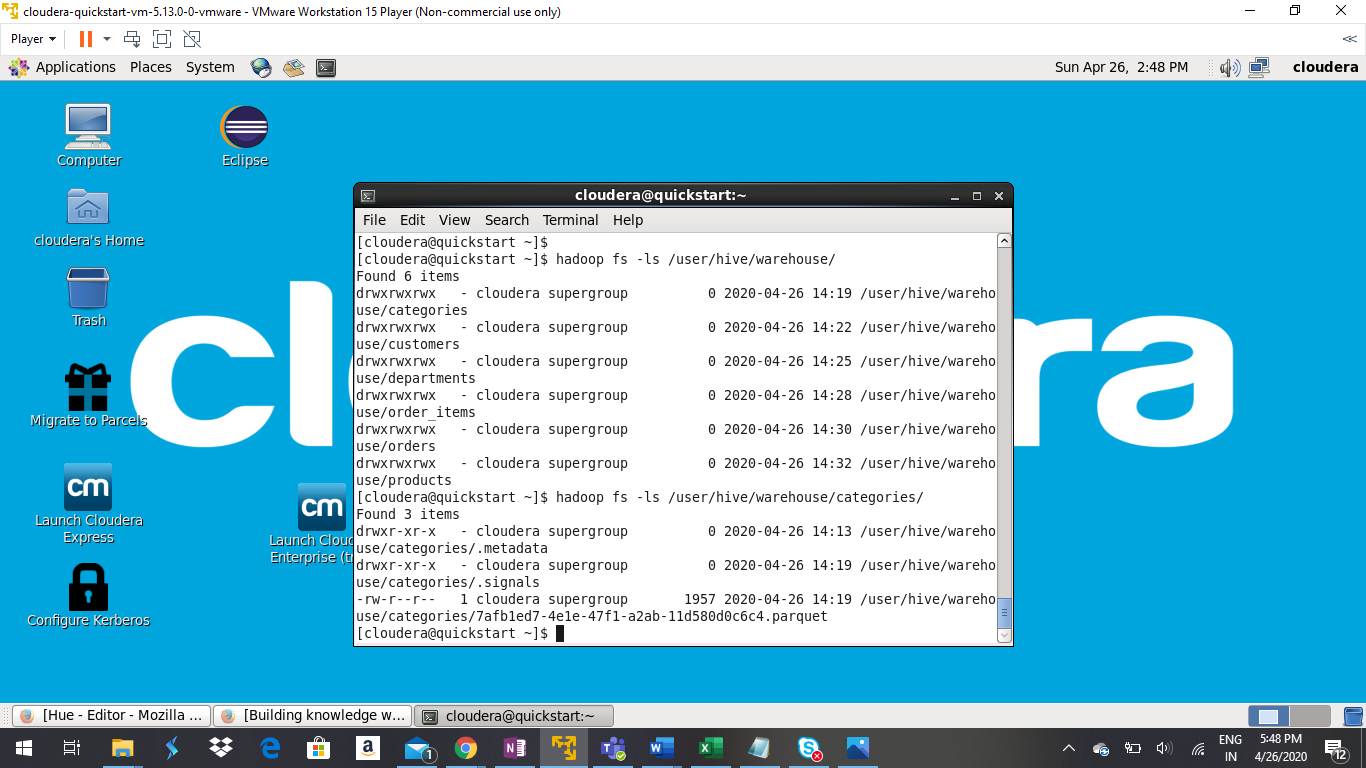
So, the first step for starting with the problem statement is the sqoop import. For this we need to open the terminal in our virtual machine and here is the logic for the same:



Now since I was getting this Accumulo error so I created the directory and set the home path and then opened another terminal and executed the sqoop command again and then the Map reduce job got started which imported the data file into HDFS. Following is the logic for the same.

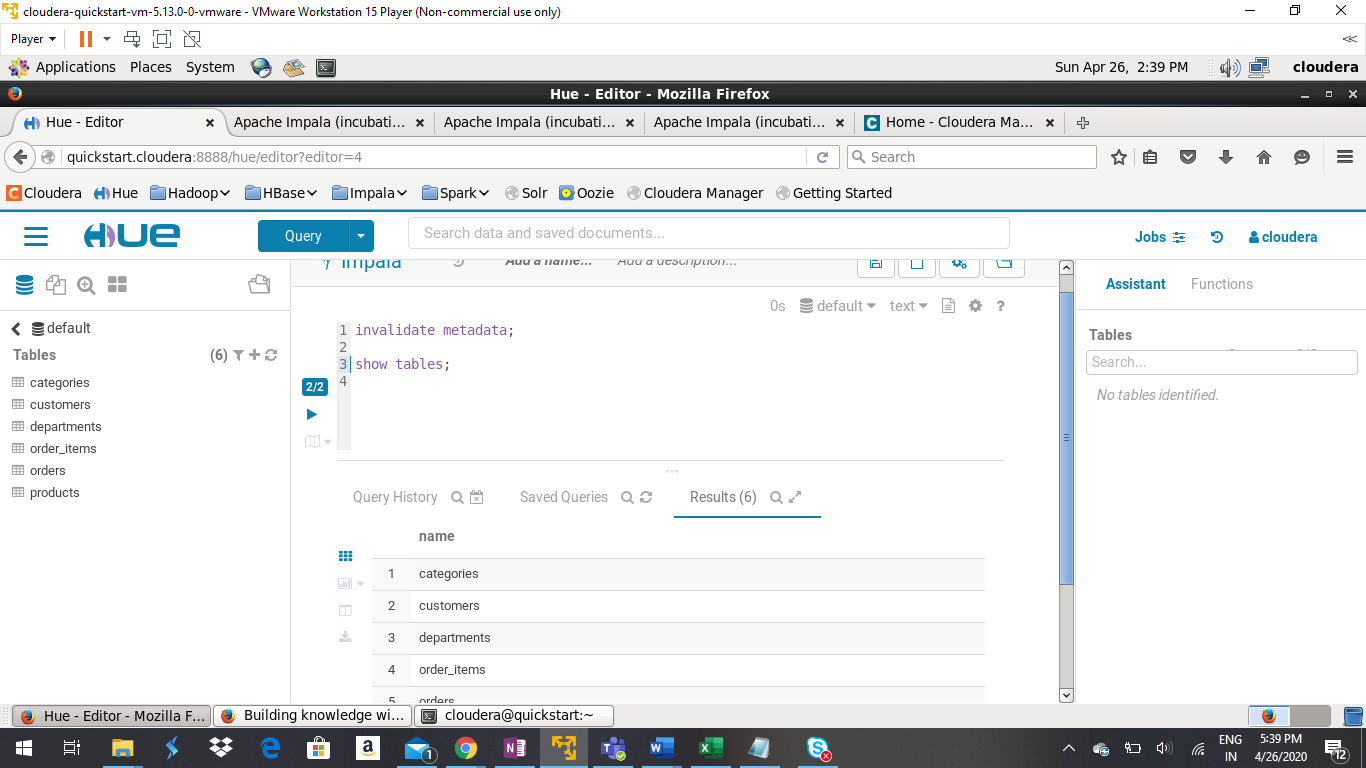


After this we need to validate that the data has arrived correctly in the HDFS and following is the logic for the same.

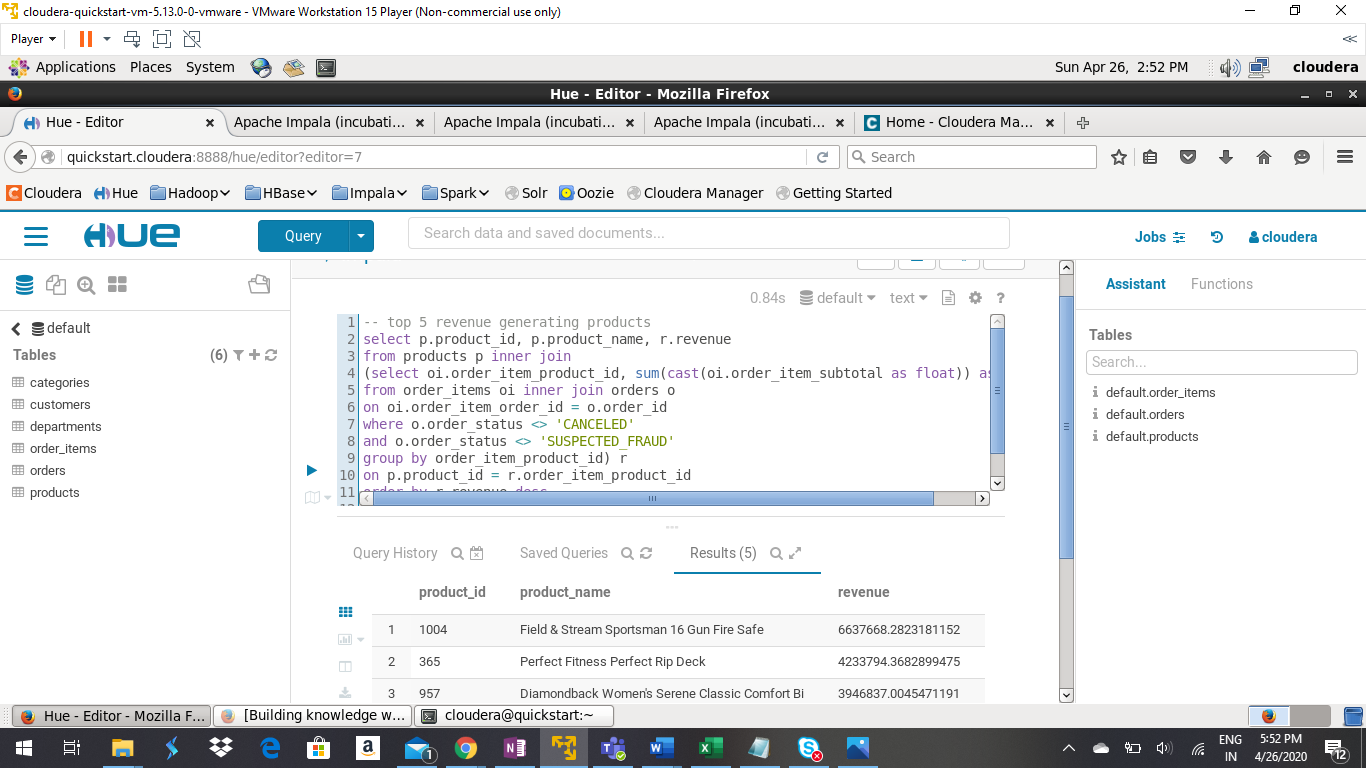


So, the file is present in HDFS now.

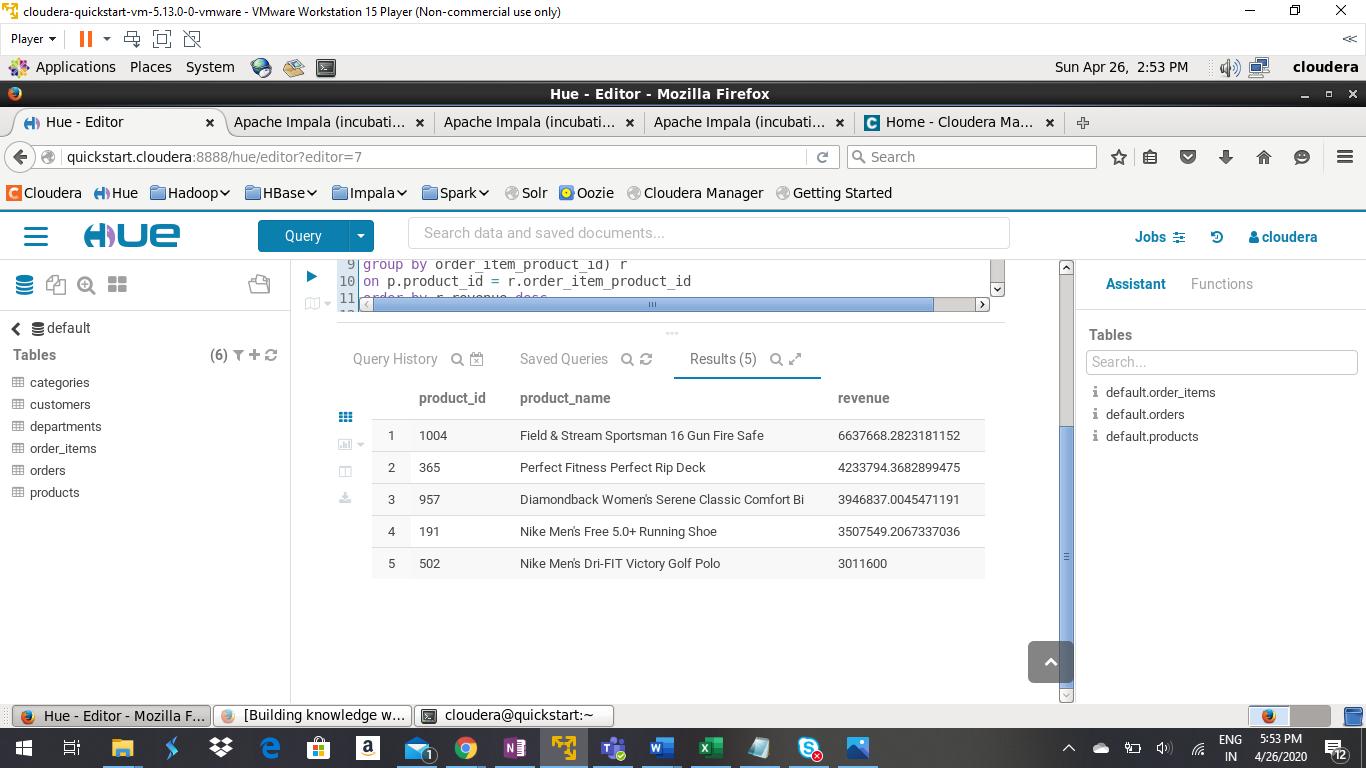
Now we have to use this data into impala. And for that we have to open the Hue dashboard and there we need to open Impala and refresh the database to show the tables. Following is the logic for the same:

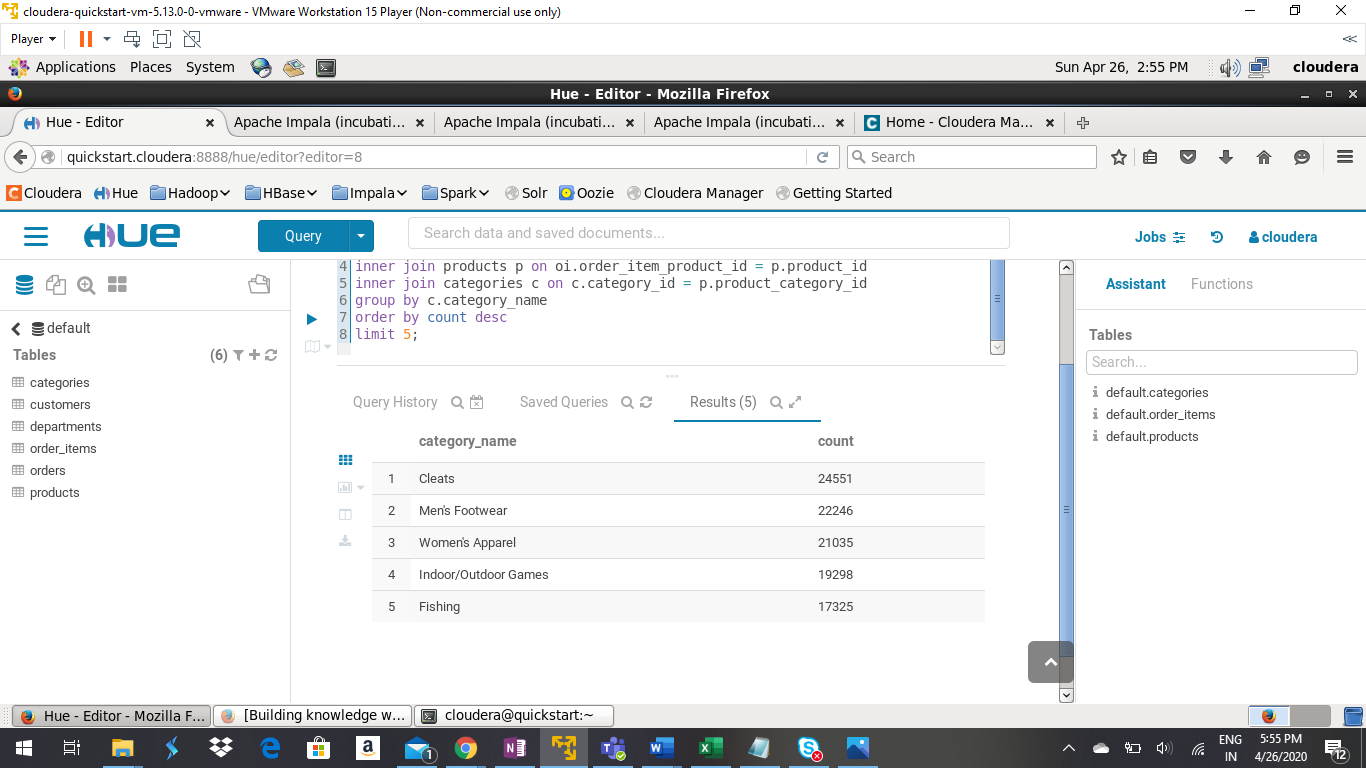


Now coming to the first problem we have to figure out the fifthth maximum revenue generating item from the given data table and here is the query that we have to execute in Impala for the same.

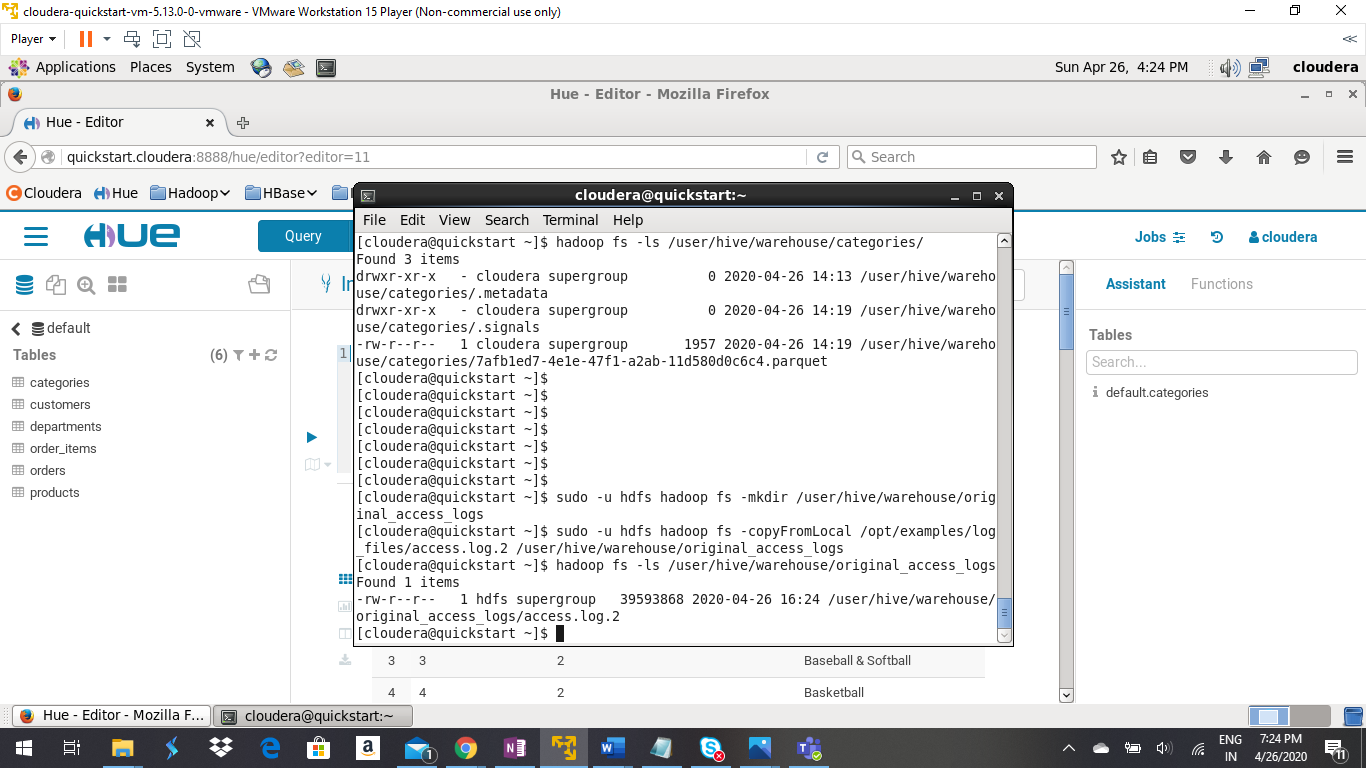


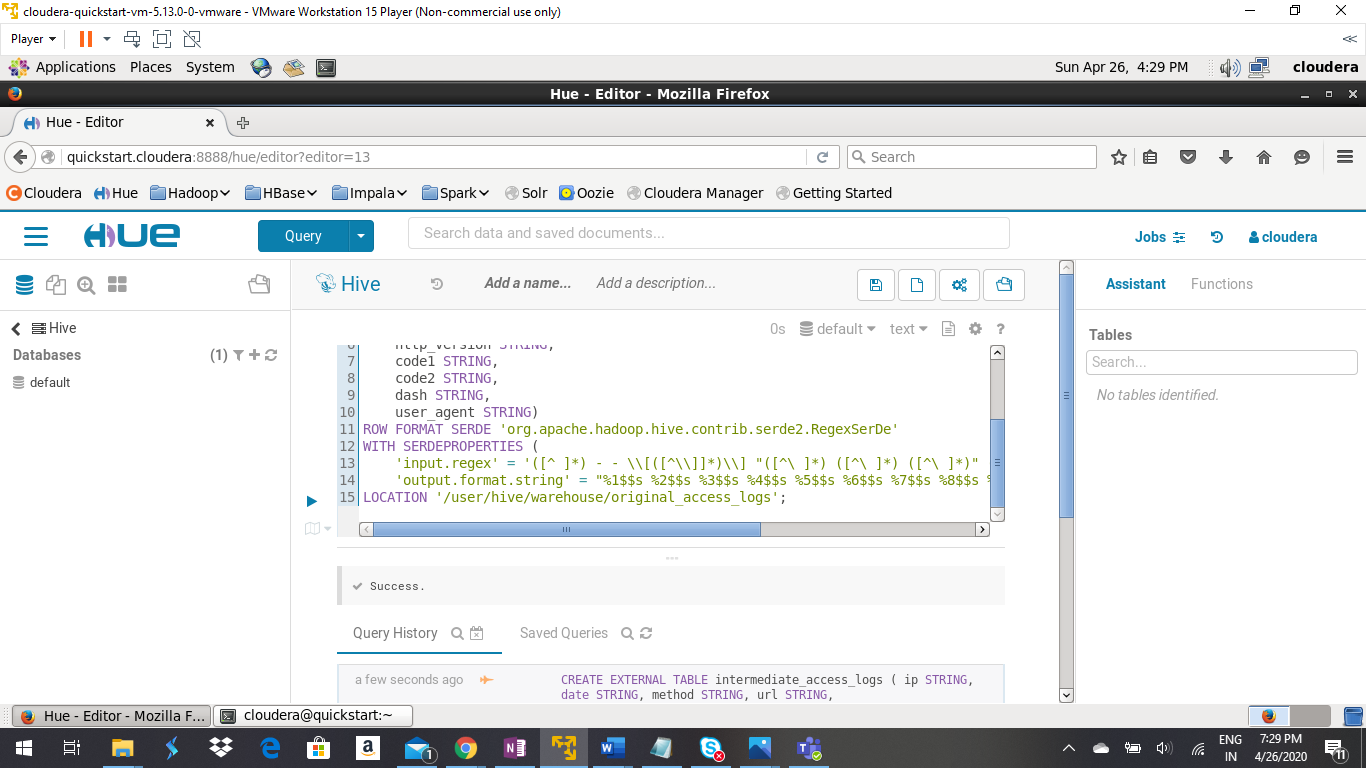
And the fifthth maximum revenue generating item is “Nike Men's Dri-FIT Victory Golf Polo”. And the revenue generated by this item is “3011600”.

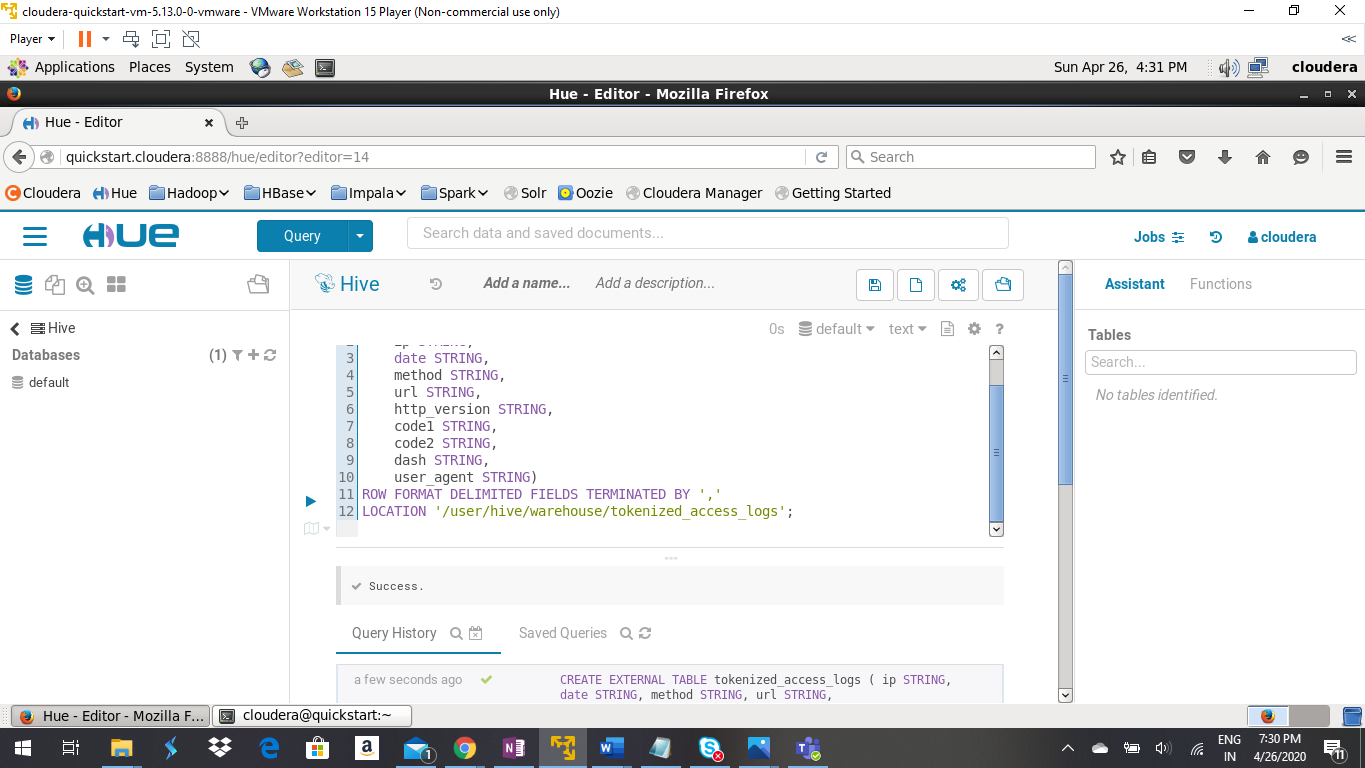


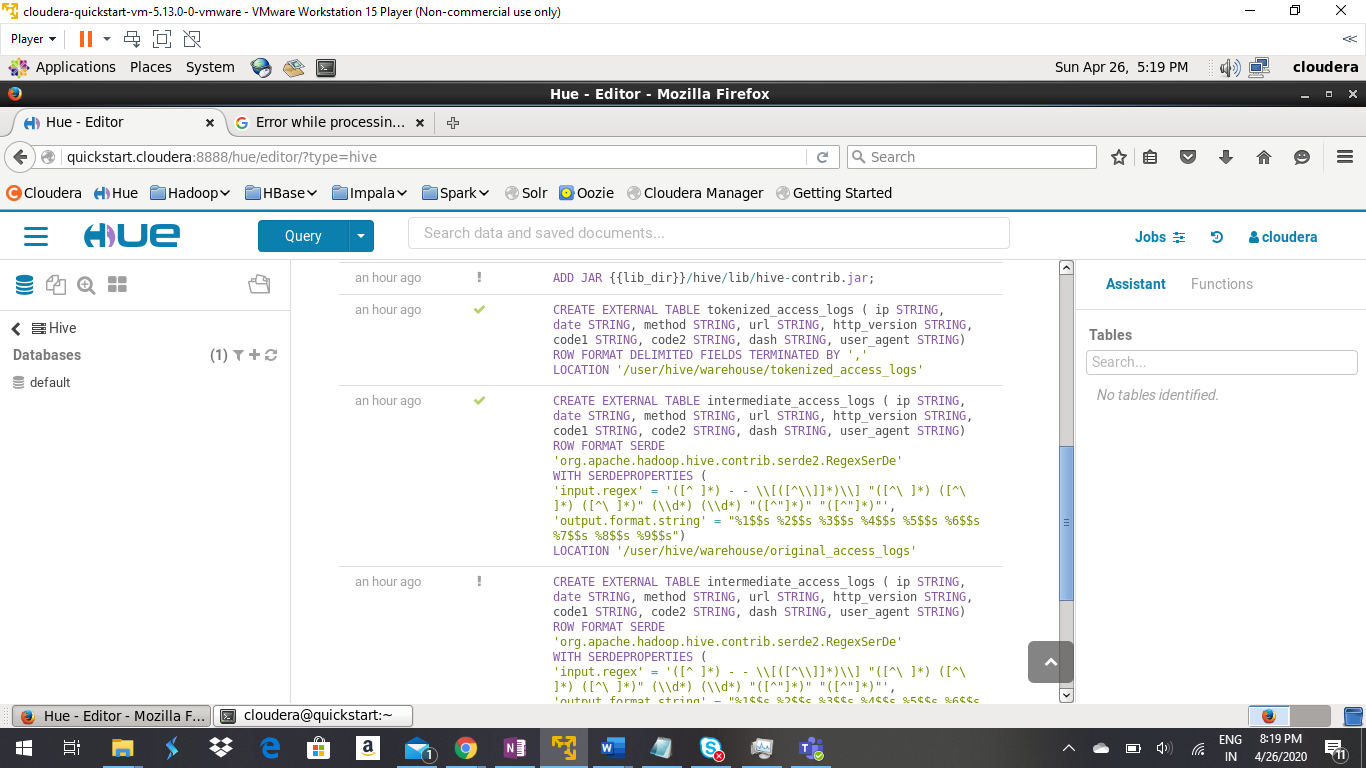


Now we have to find out the product which has been visited many times but its not purchased. So, we will find out the same by comparing the tables and from the results we can say that product 1 is the answer. There was a typo error in the name when we executed the query in impala. However, when we updated the same in hive and then again refreshed the metadata of the impala it got corrected and the price value was updated correctly and due to this reason, it was appearing in the viewed list but not in the purchased list. As the price was high earlier by mistake due to type error and then after updating it the sales for that product got increased. Following is the logic for the same.

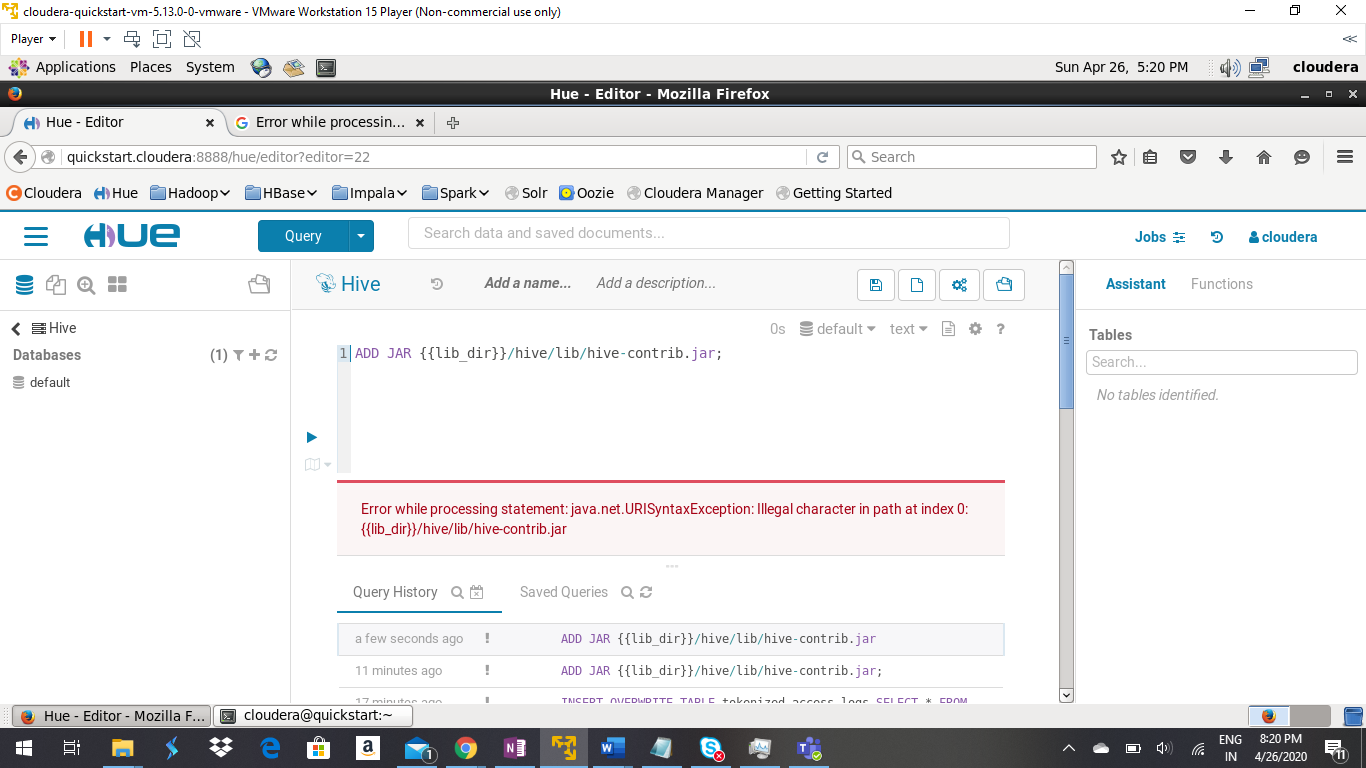


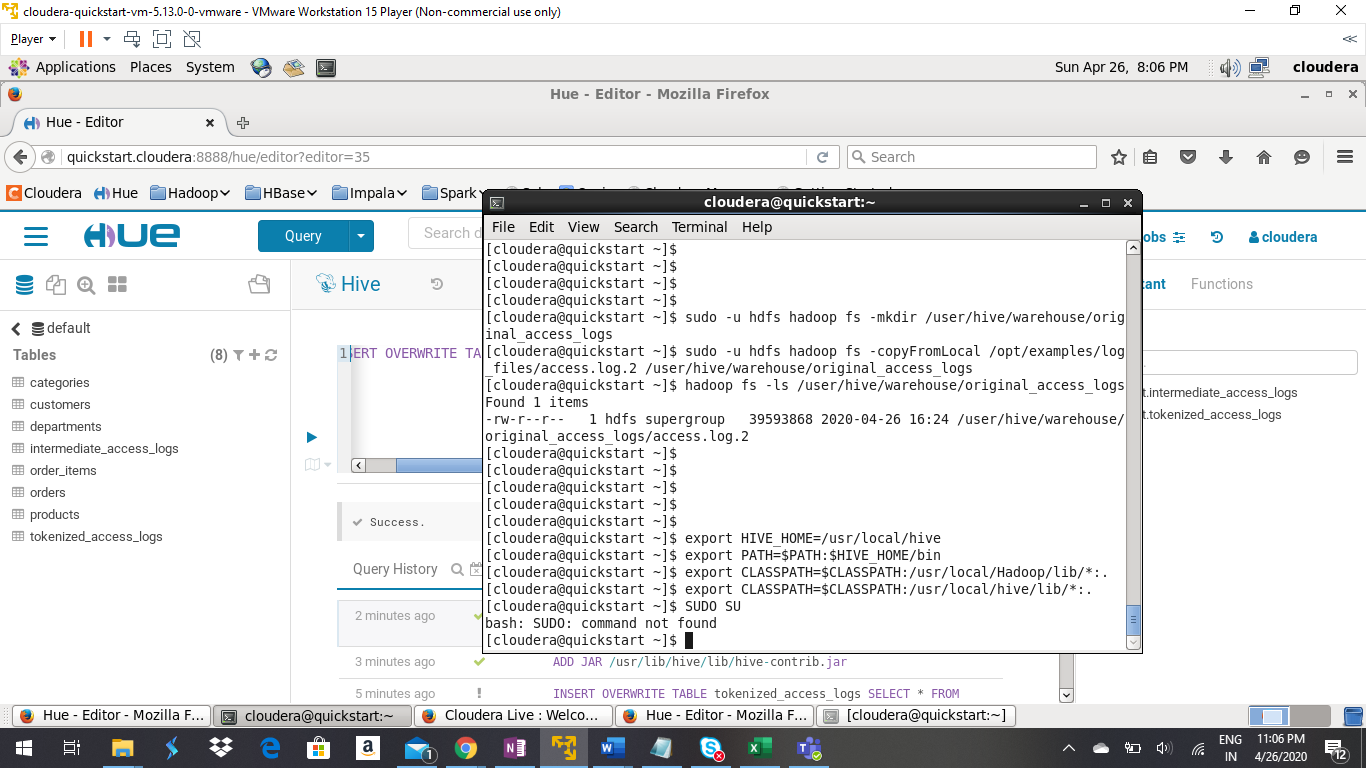


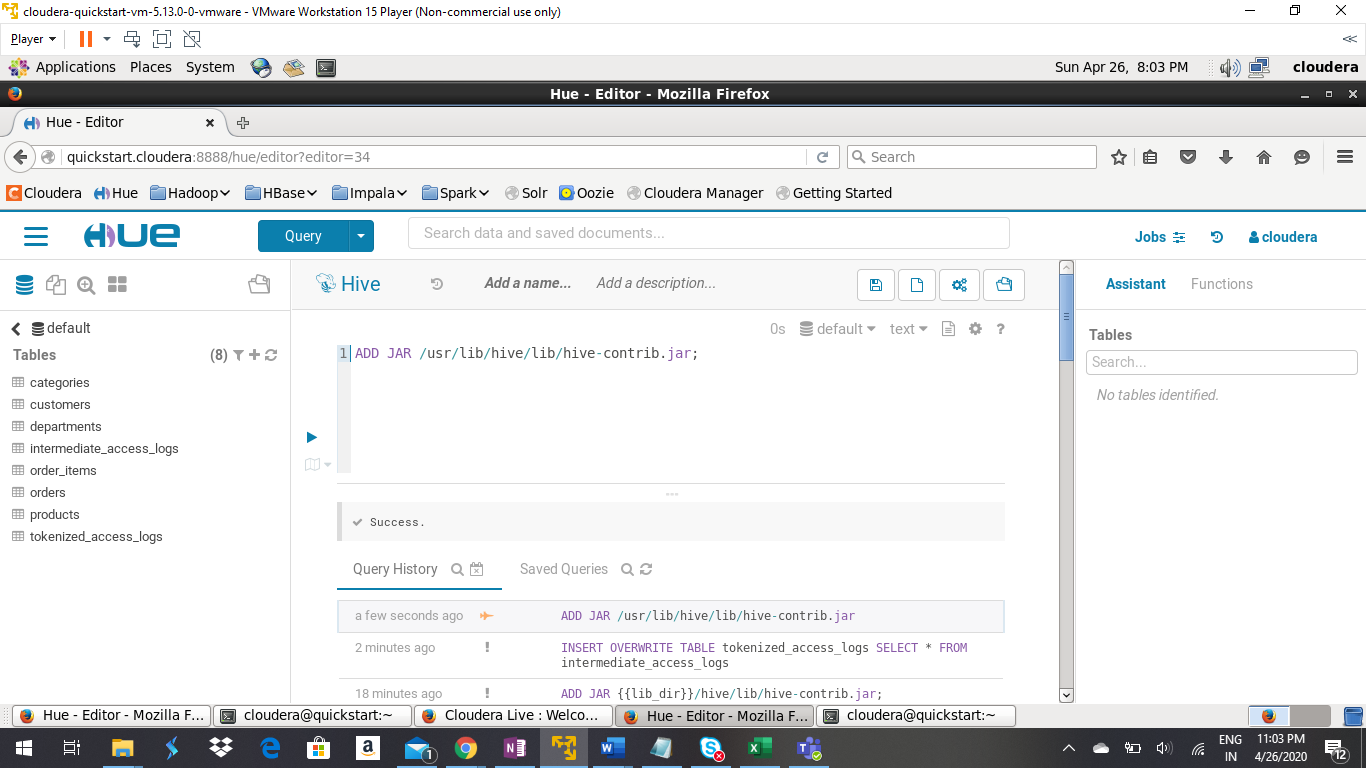




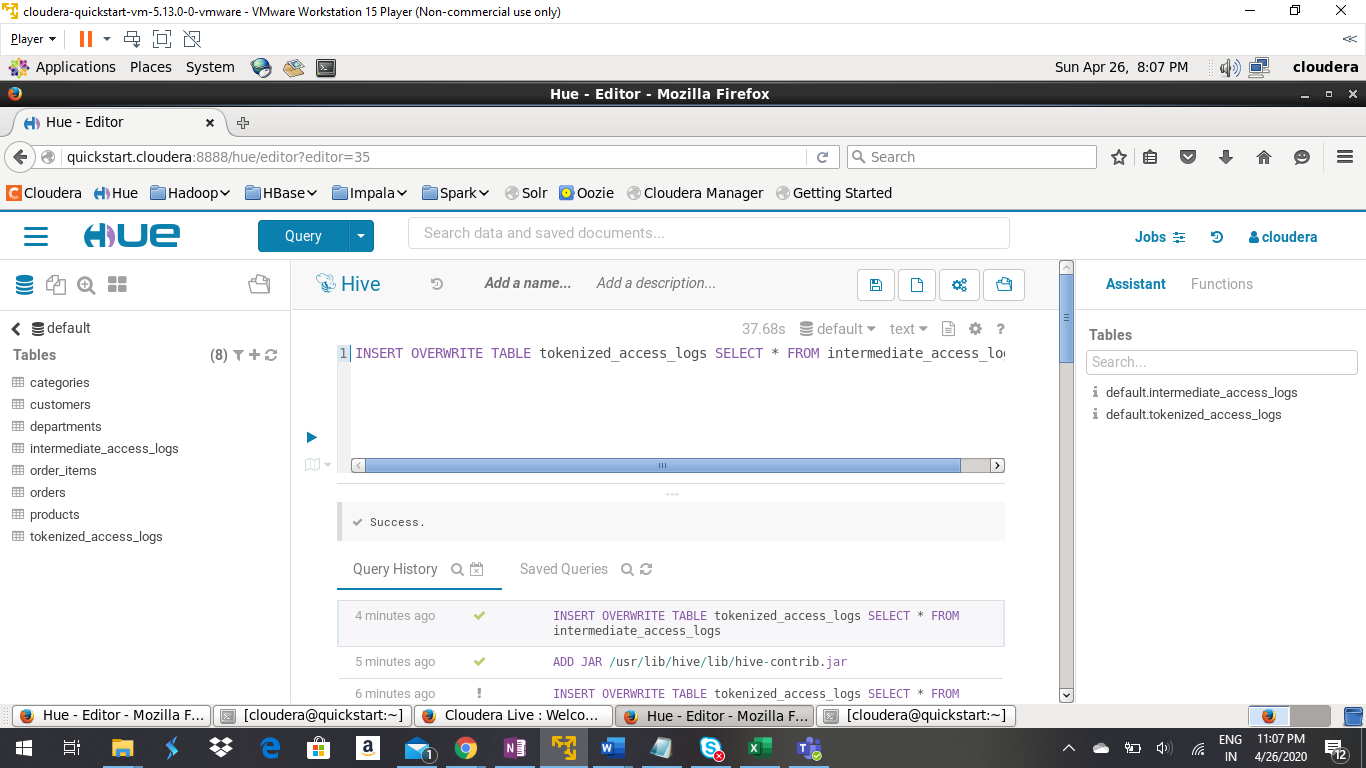
I got following error while trying to add the jar file. So I fixed it by adding the jar file as follows via terminal and then restarted the service and it got resolved.



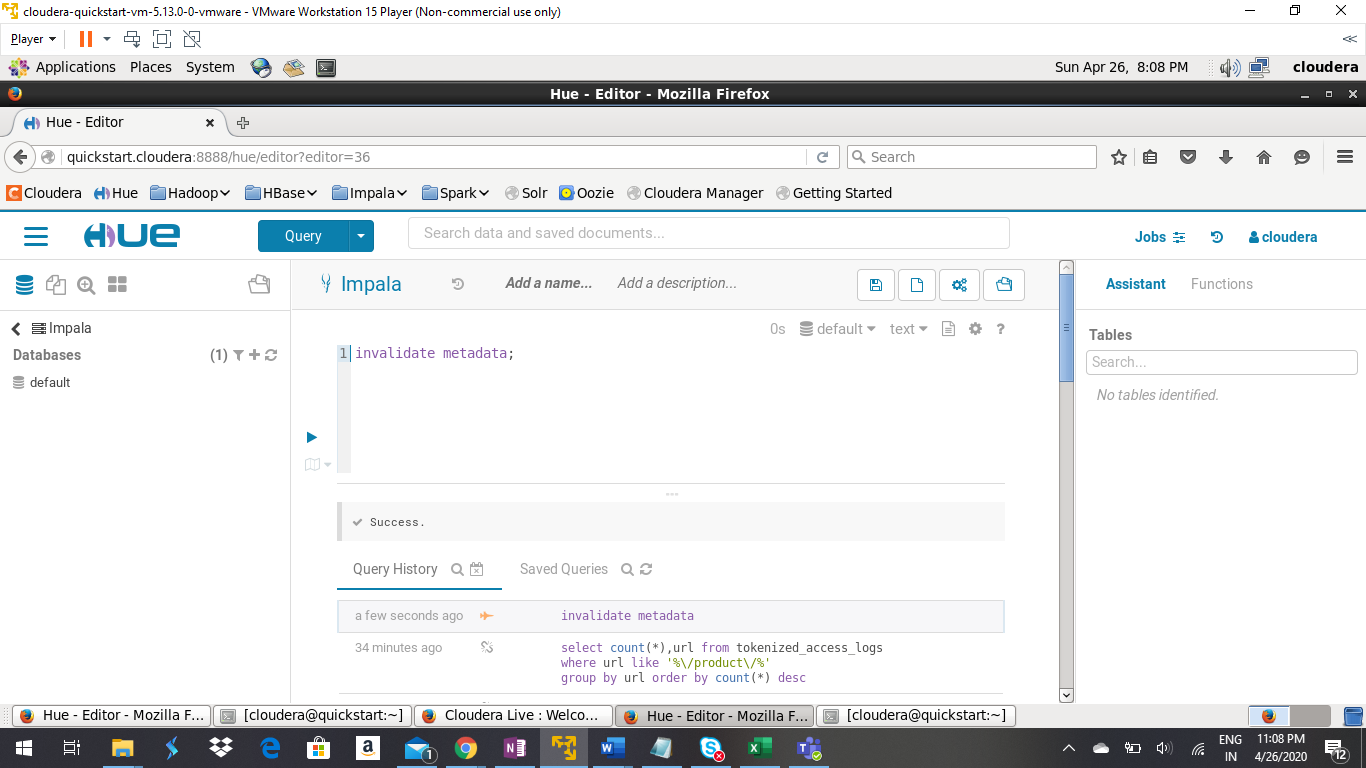


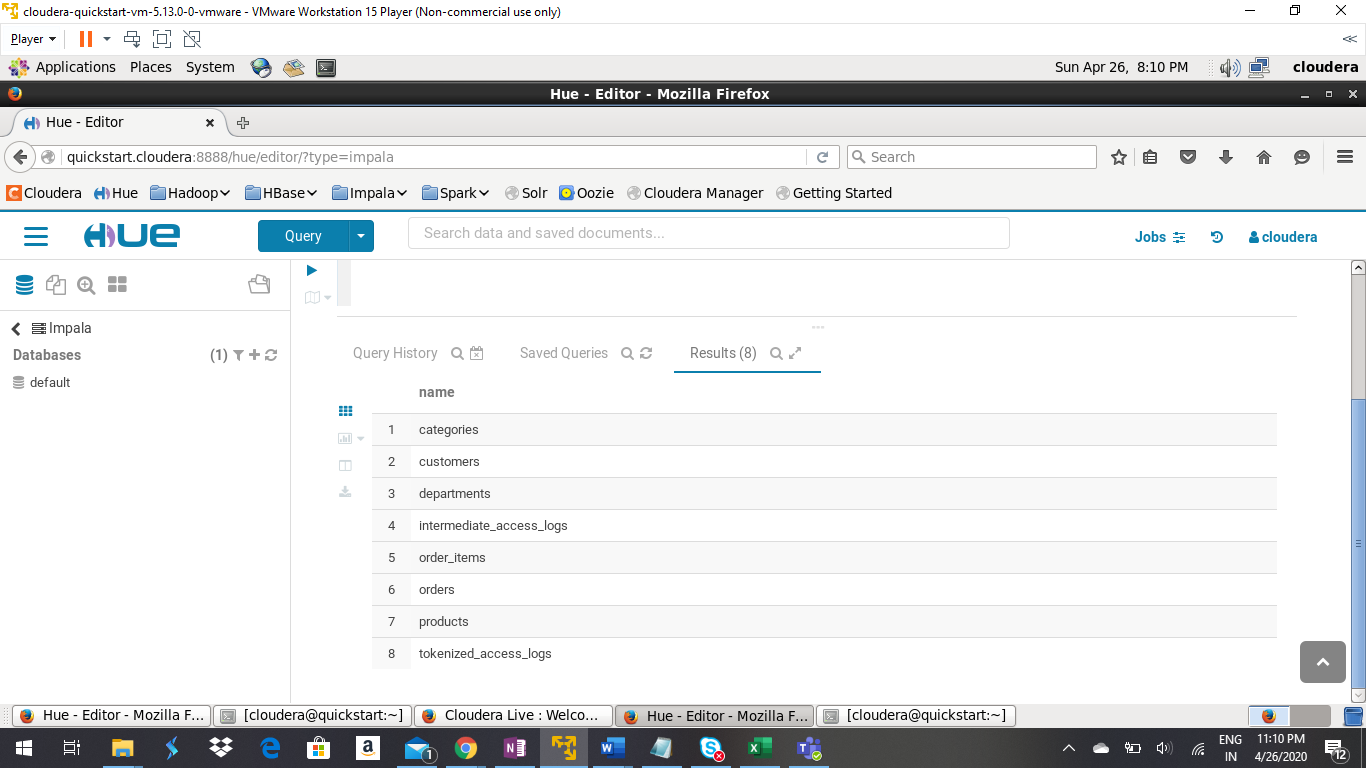


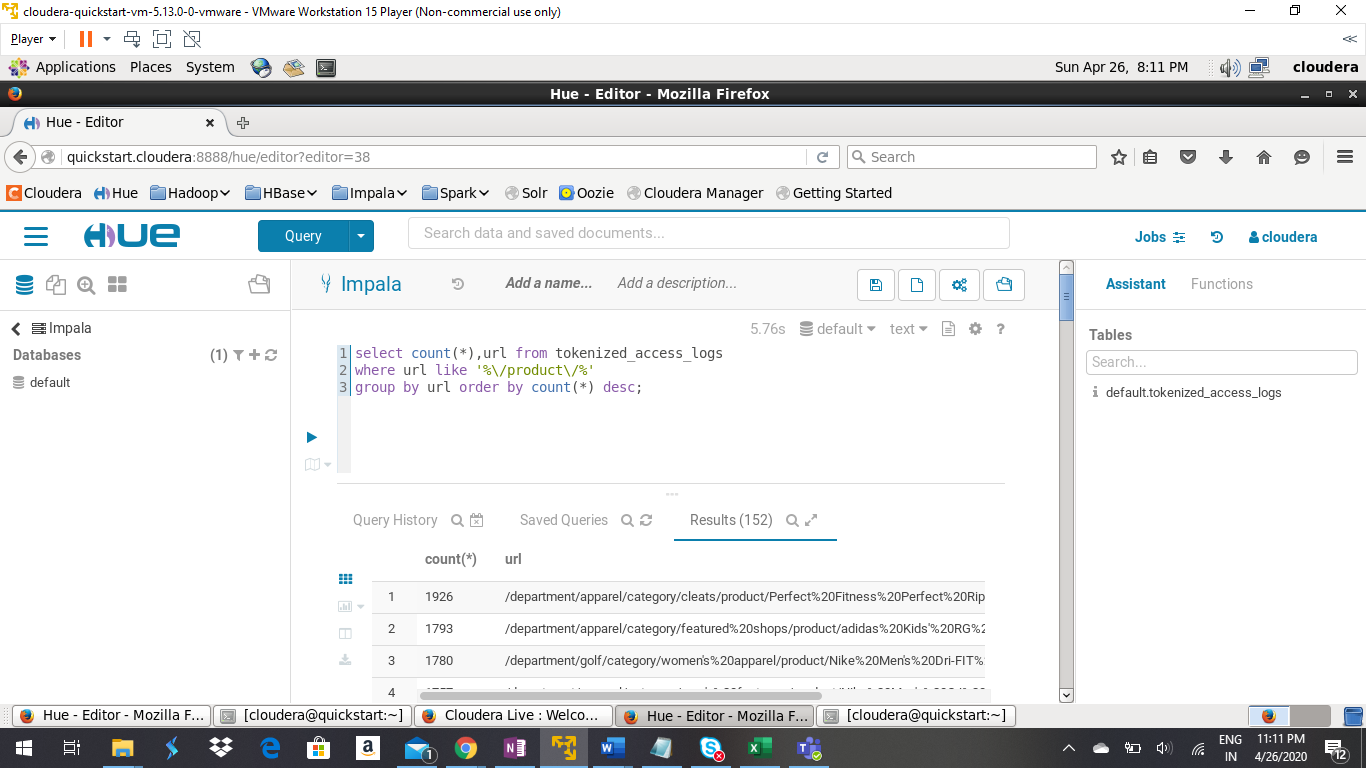
Then I executed the final query.

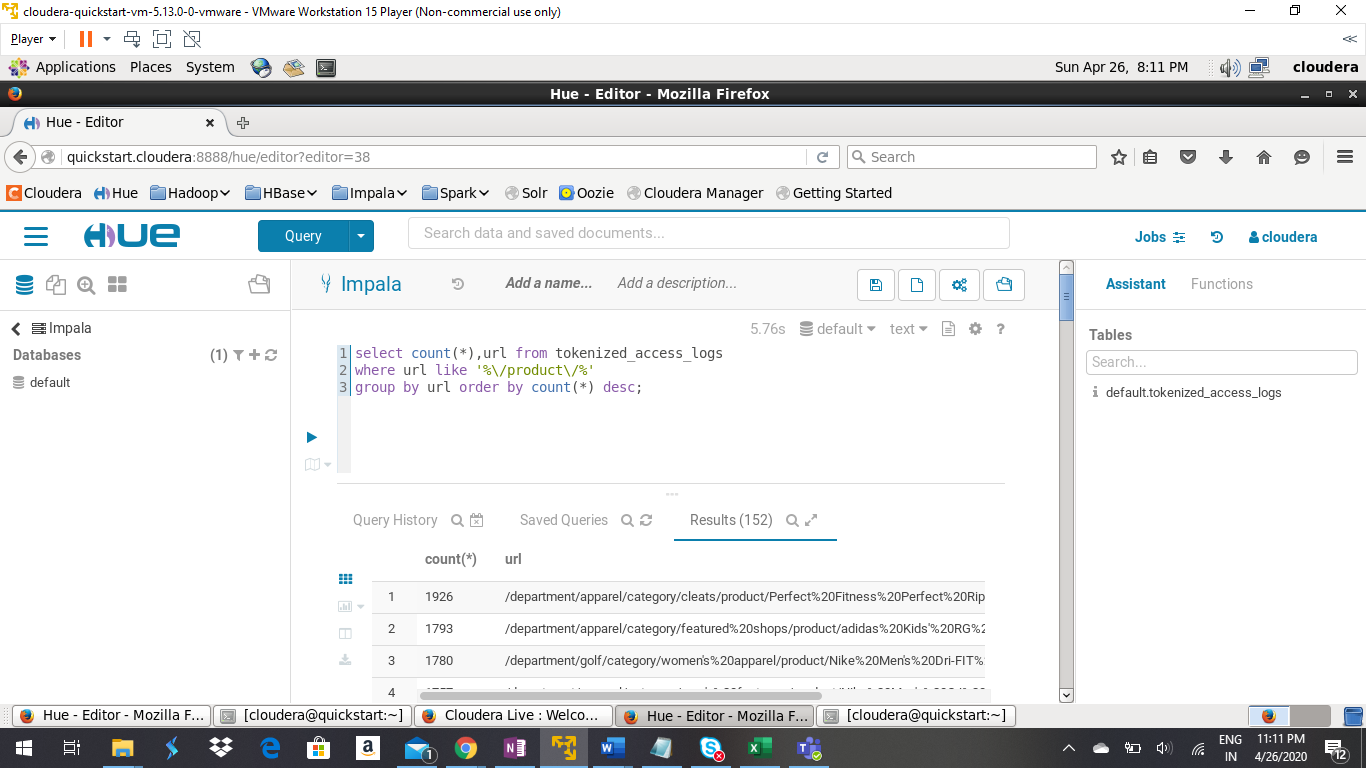


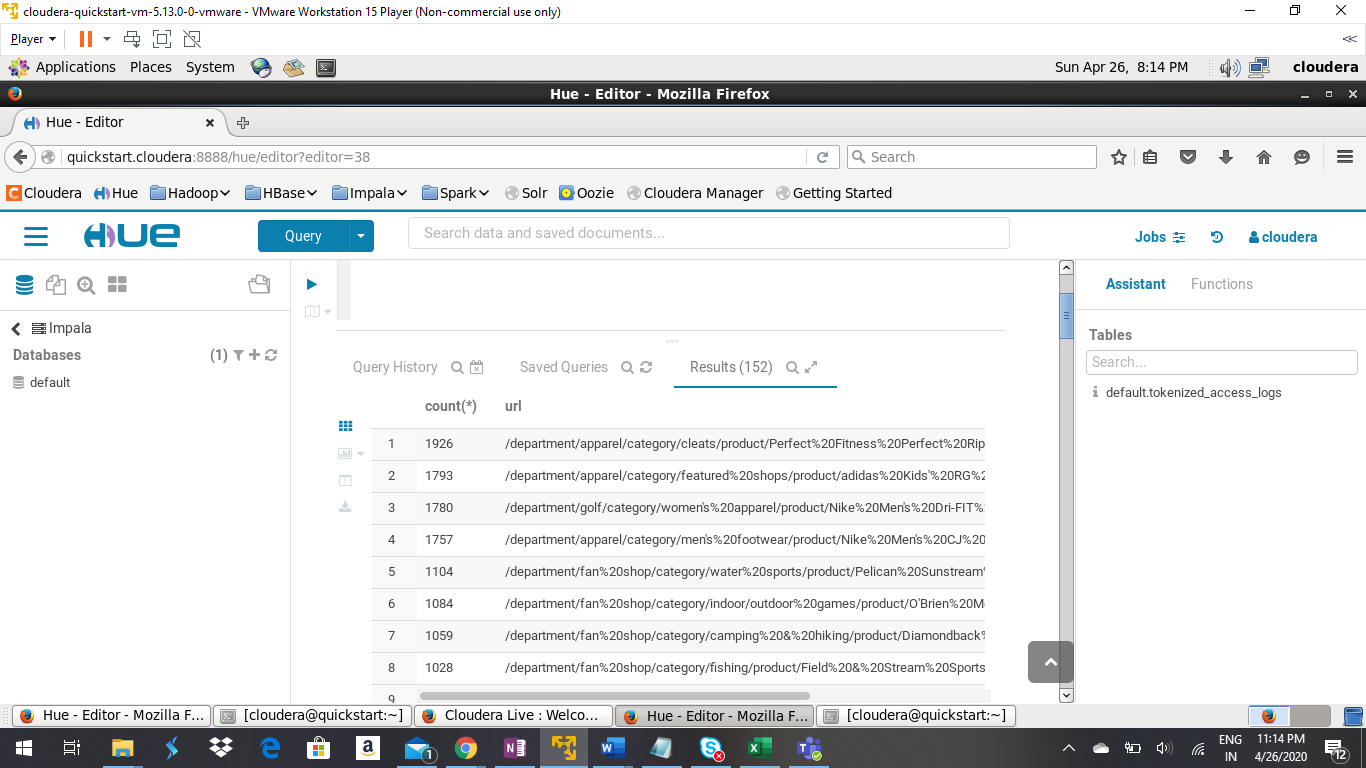
Then I refreshed the metadata of Impala in order to get the tables updated here too.











# References

1. <https://stackoverflow.com/questions/31902917/hive-add-file-jar-says-does-not-exist>
2. <https://stackoverflow.com/questions/37342061/error-on-my-first-sqoop-job>