**A close up of a logo

Description automatically generated**

Week 3 Lab Assignment # 1

ALY 6110 Big Data and Data Management

CRN: 22973

Instructor: Daya Rudhramoorthi

Student: Yi Yang

02/02/2021

**Introduction**

This lab assignment is going through 2 tutorial exercise of Apache Hadoop and includes screenshots of running SQL codes in order to answer 3 questions.

**Analysis**

First of all, we need to set up the environment for Apache Hadoop. Since my laptop has met the hardware requirements to run Cloudera and my system is Windows, I downloaded and installed VMWare Workstation and Cloudera QuickStart in my laptop. After that, I change the setting of Virtual Machine into Memory is 12 GB, and the Processors is 2 GB in order to make the later analysis faster.

1. Once we started running the Virtual Machine, we now have access to Cloudera.

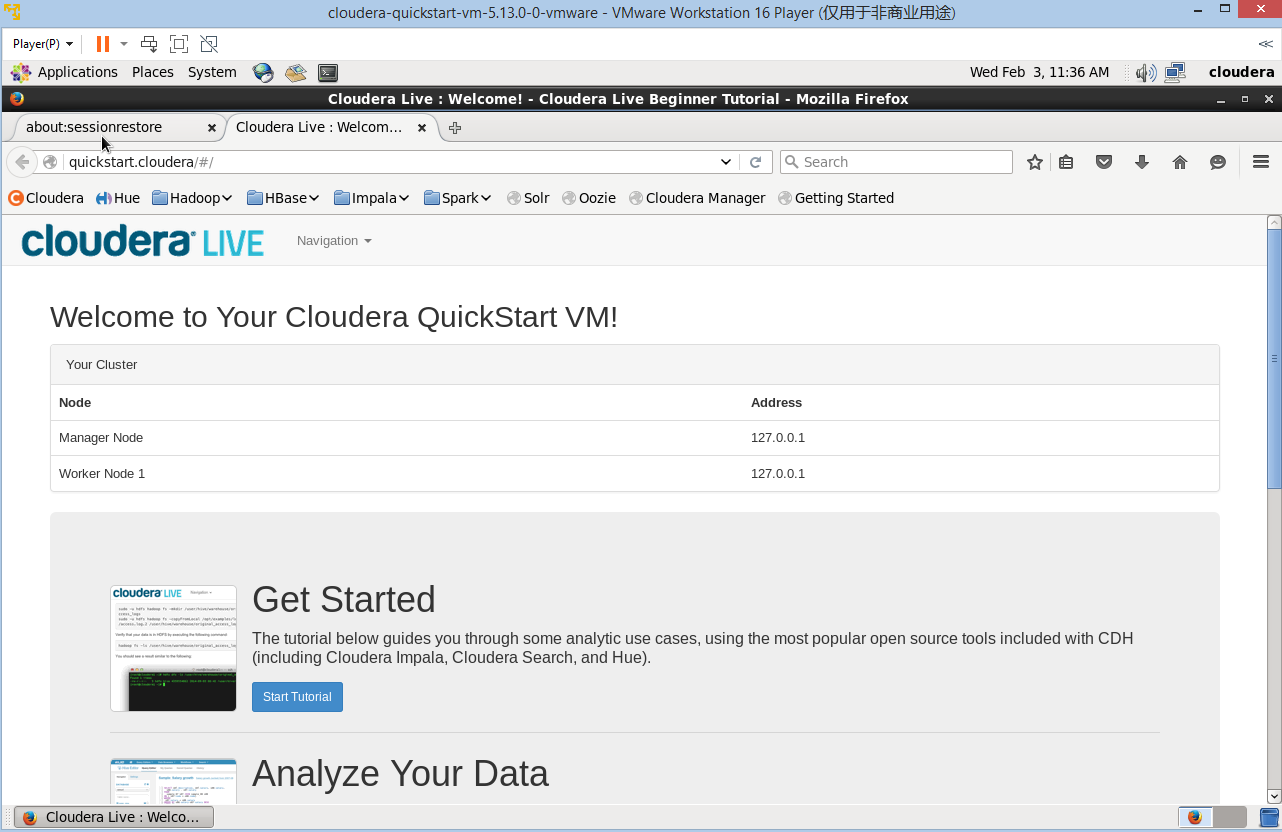


Figure 1

3. Then, we clicked Cloudera Manager and entered the name and password.

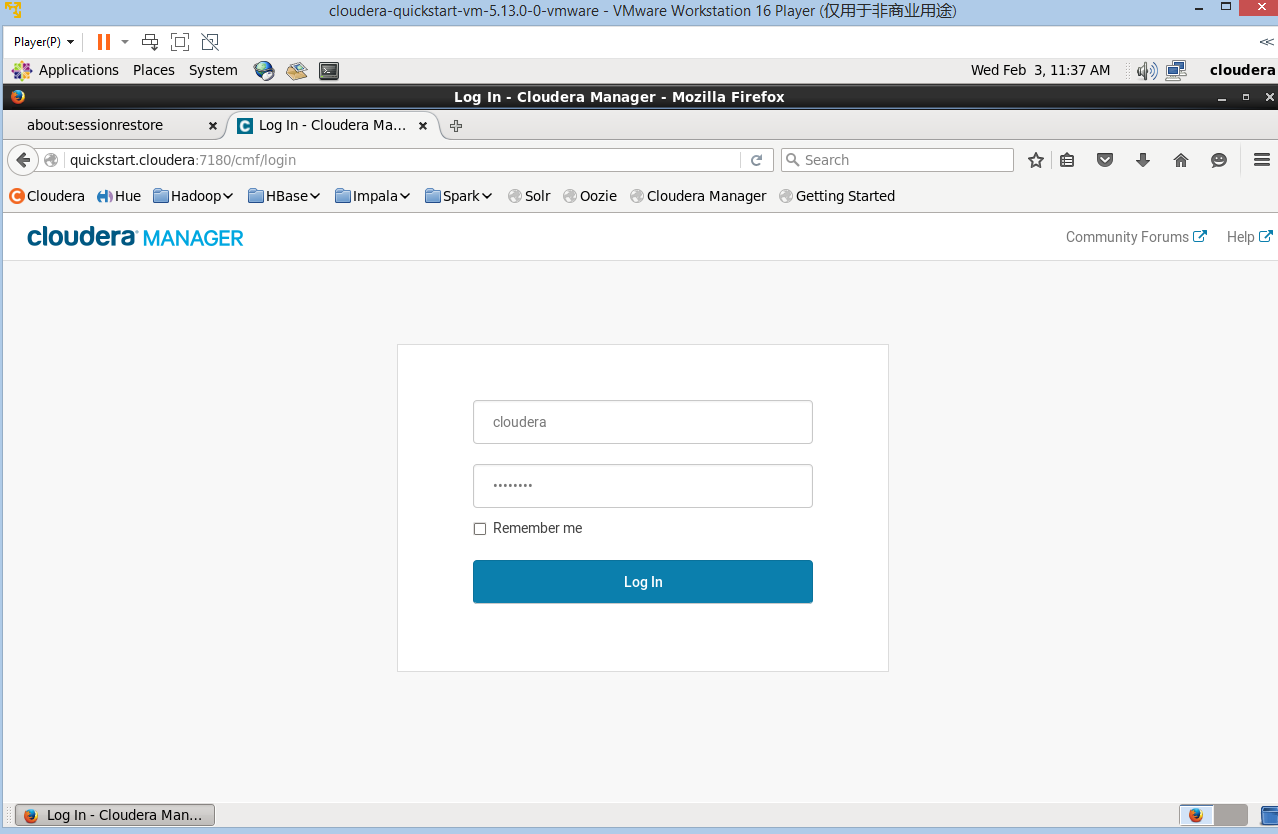


Figure 2

4. After logging into the Cloudera, we first restarted all the services.

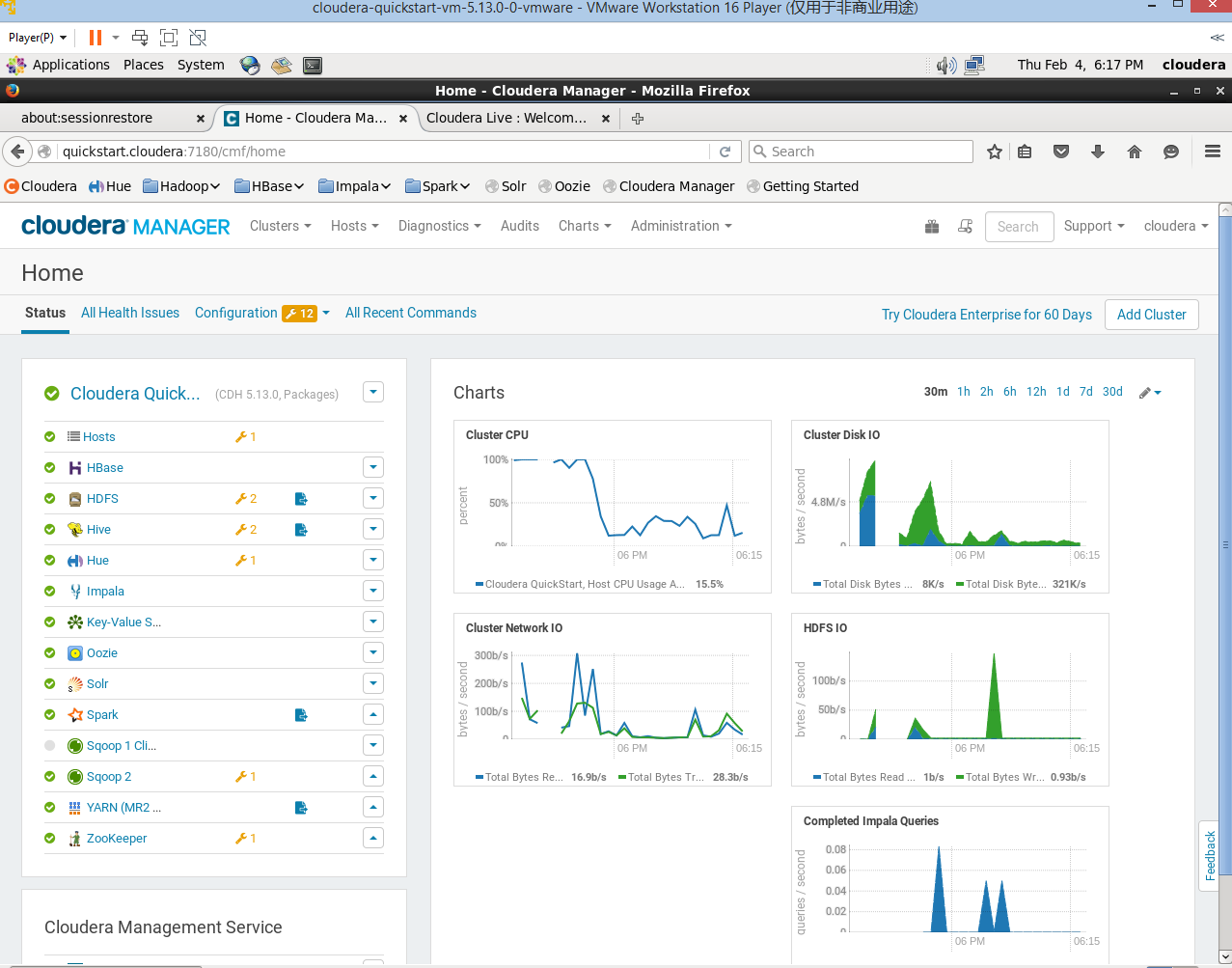


Figure 3

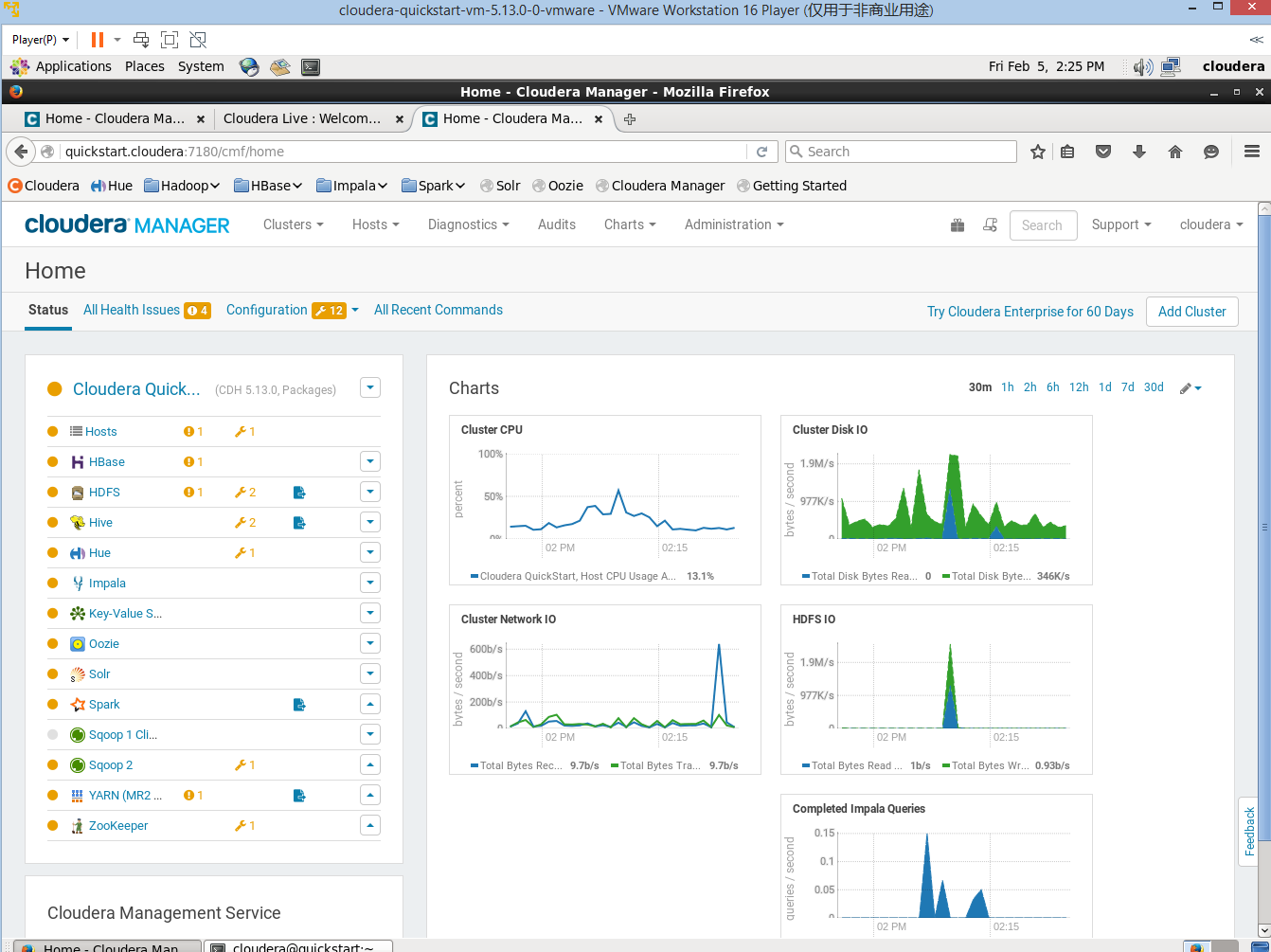


Figure 4

**Reflection:** I pasted both services are turning to all Green or all Orange during different time above. Since I have been starting and closing the Cloudera back and forth for several times, I notice that there are very few cases that all services in Cloudera are all turning into Green (Running/Healthy), most of the time, they are in Orange/Yellow (Some Warnings). In very few times, all services are turning into Green and stay a bit longer even though I refresh the cluster or restart all services once again.

I was worried about it first wondering if all or some services (Hue/Hive/Impala) turning into Orange may affect the results, but our TA confirmed with me that it is ok to have them into Orange and would not affect the results.

After making sure all or some services in Cloudera are starting, we started running some codes in the Cloudera Terminal.

*Tutorial Exercise 1*

For this exercise, we will answer business question like what products our customers like to buy/have the highest revenue.

1. We opened the Terminal, and put the Sqoop code to install MapReduce so that we have retrieved from datasets from MySQL database.

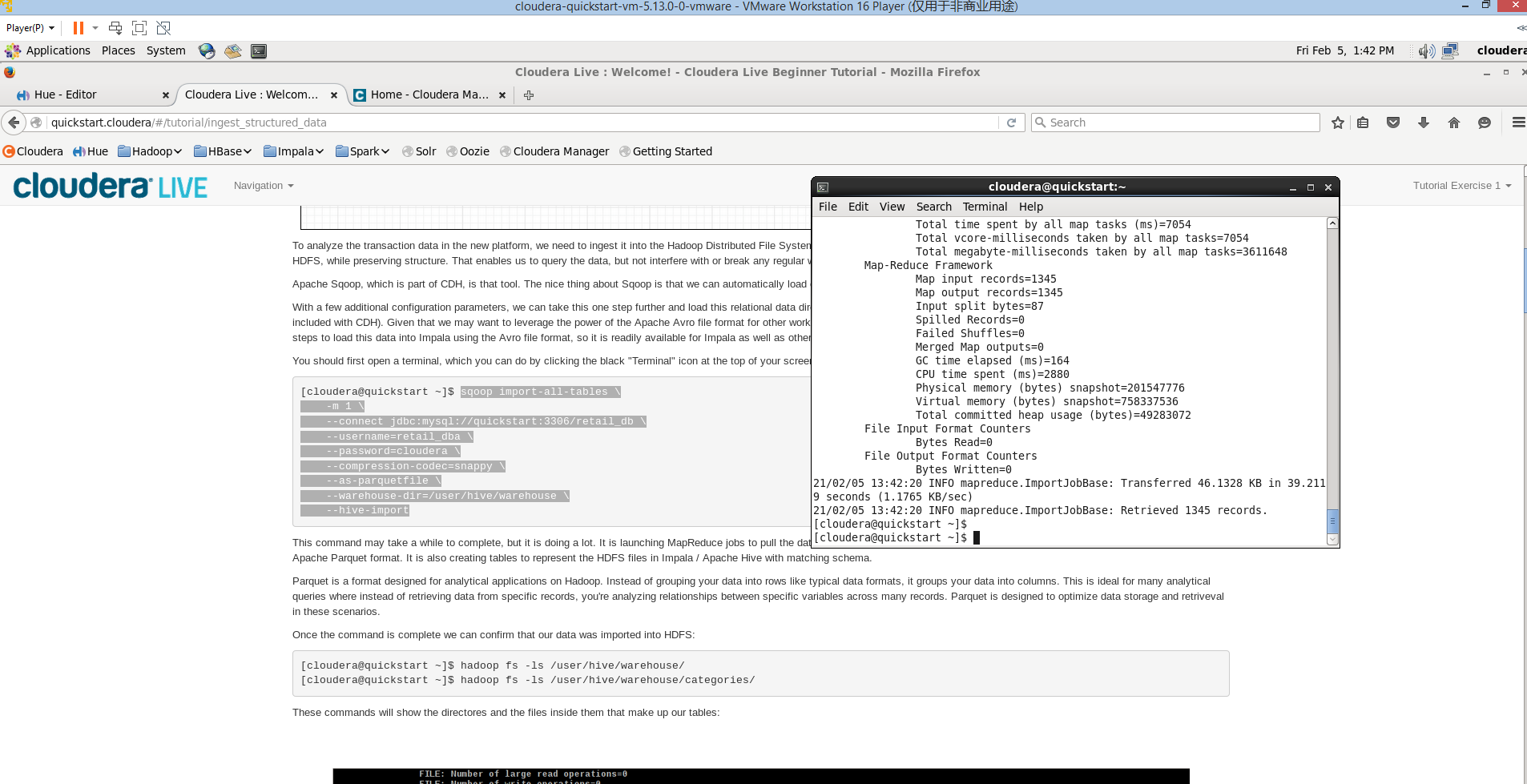


Figure 5

2. We now confirmed our data files exist in HDFS.

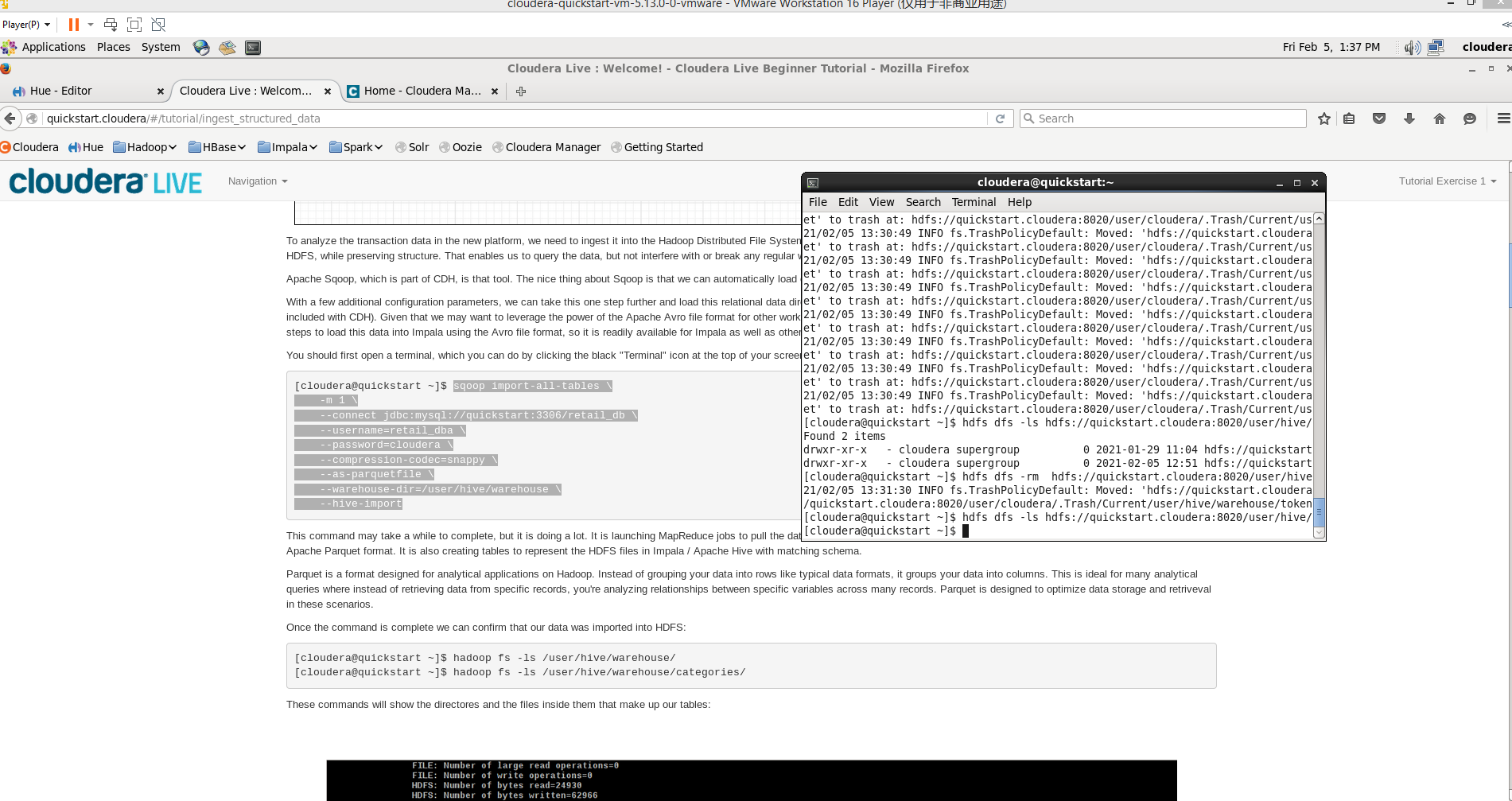


Figure 6

3. After loading, we went to Hue’s Impala Query Editor to show all tables.

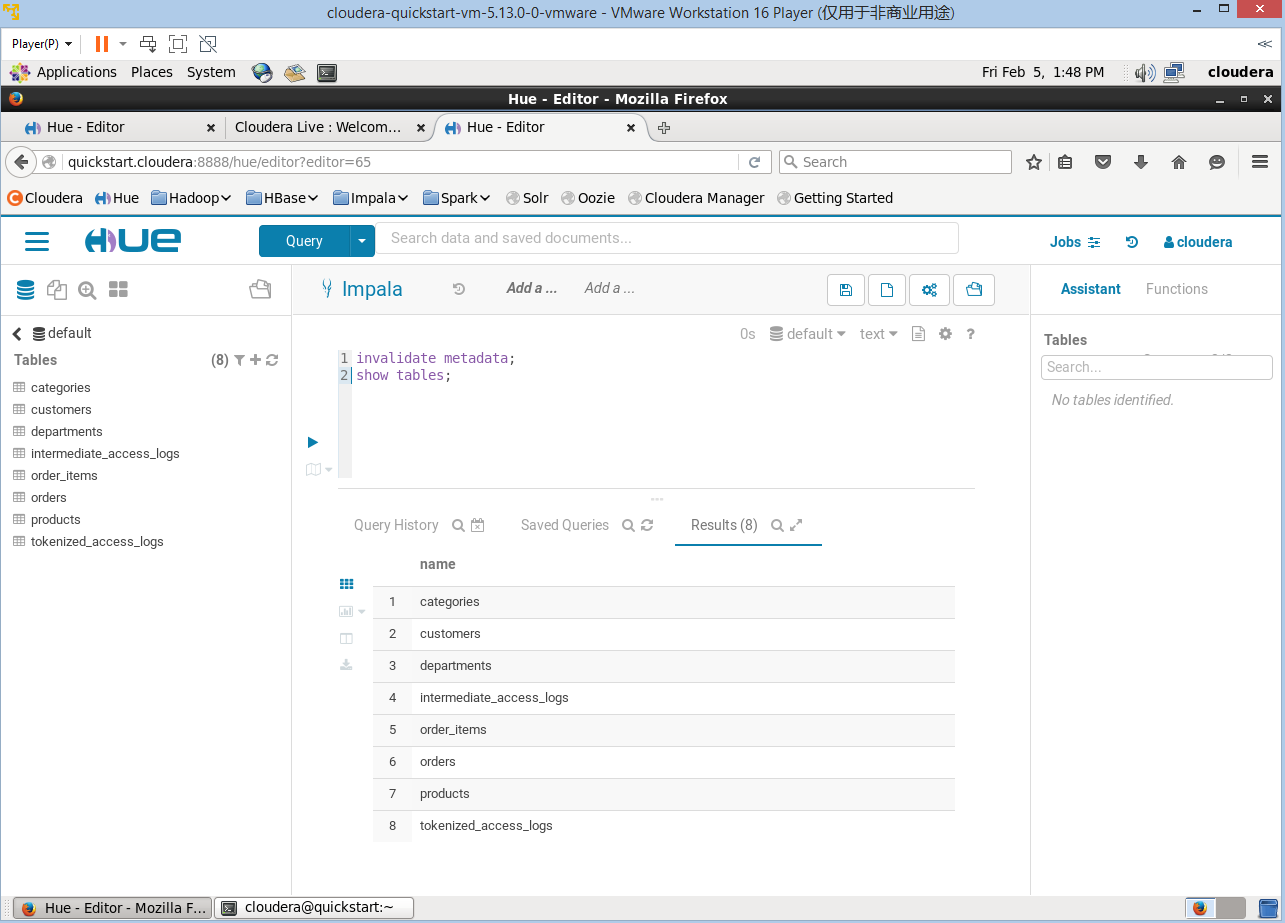


Figure 7

4. Now we used the following SQL queries to calculate total revenue per product.

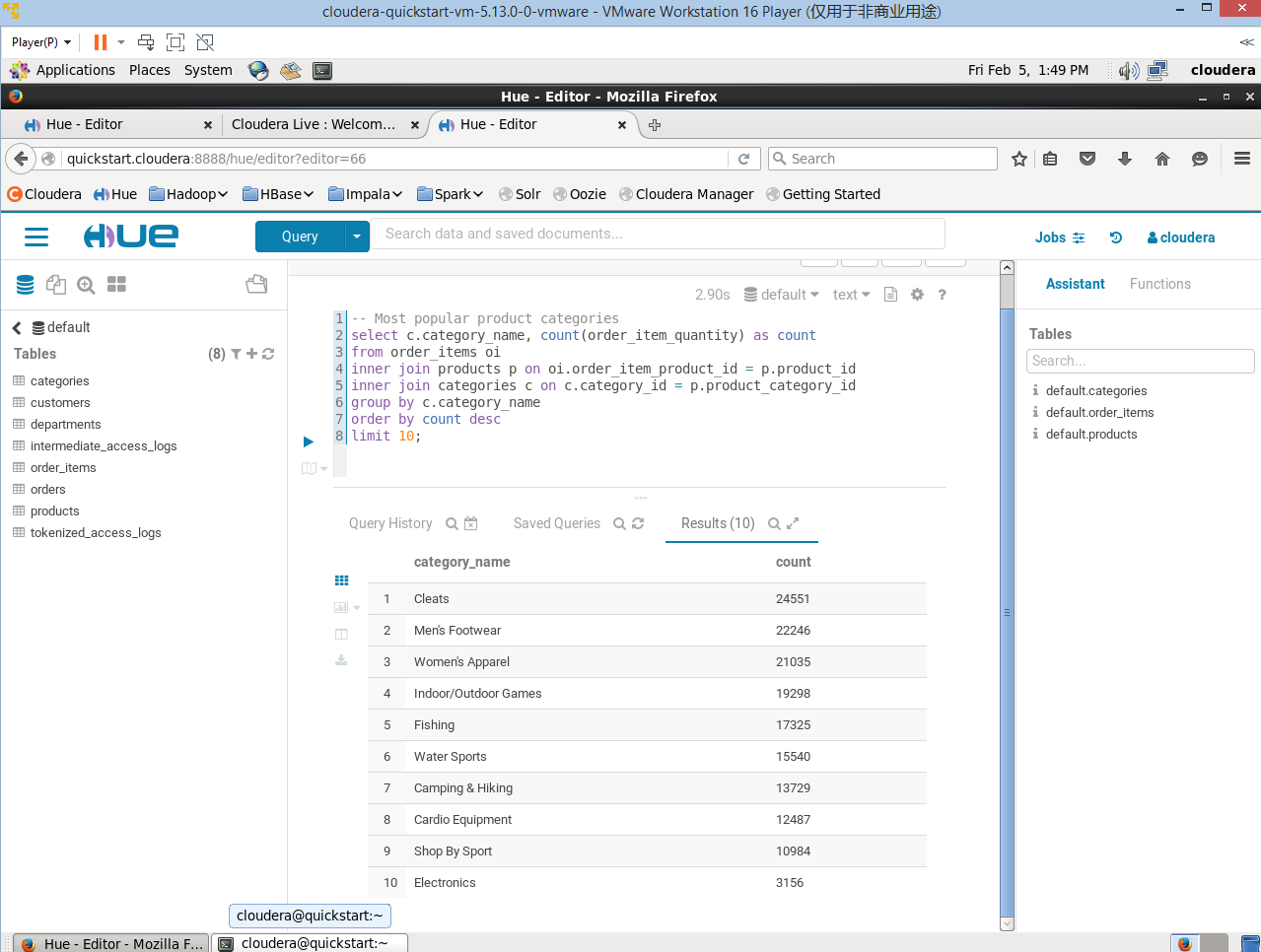


Figure 8

**Conclusion:** we can learn from the screenshot that the top 3 products that have the most counts are Cleats, Men’s Footwear, and Women’s Apparel.

5. Now we used the following SQL example queries to show the top 10 revenue generating products. These two versions indicated the process of getting the correct results.

Version 1 below (redundent rows and no desired results):

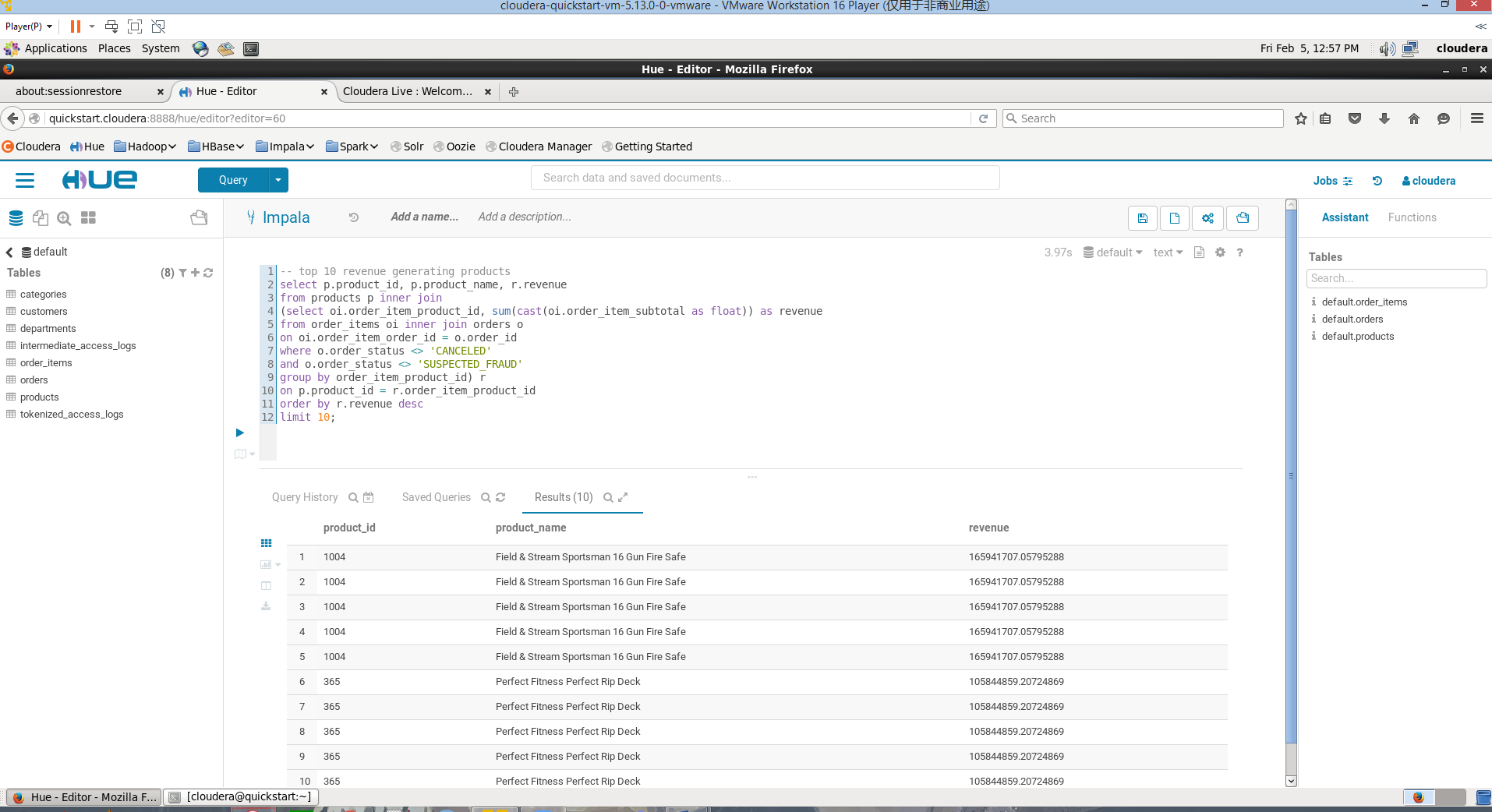


Figure 9

Version 2 below (all distinct rows and can achieve desired results):

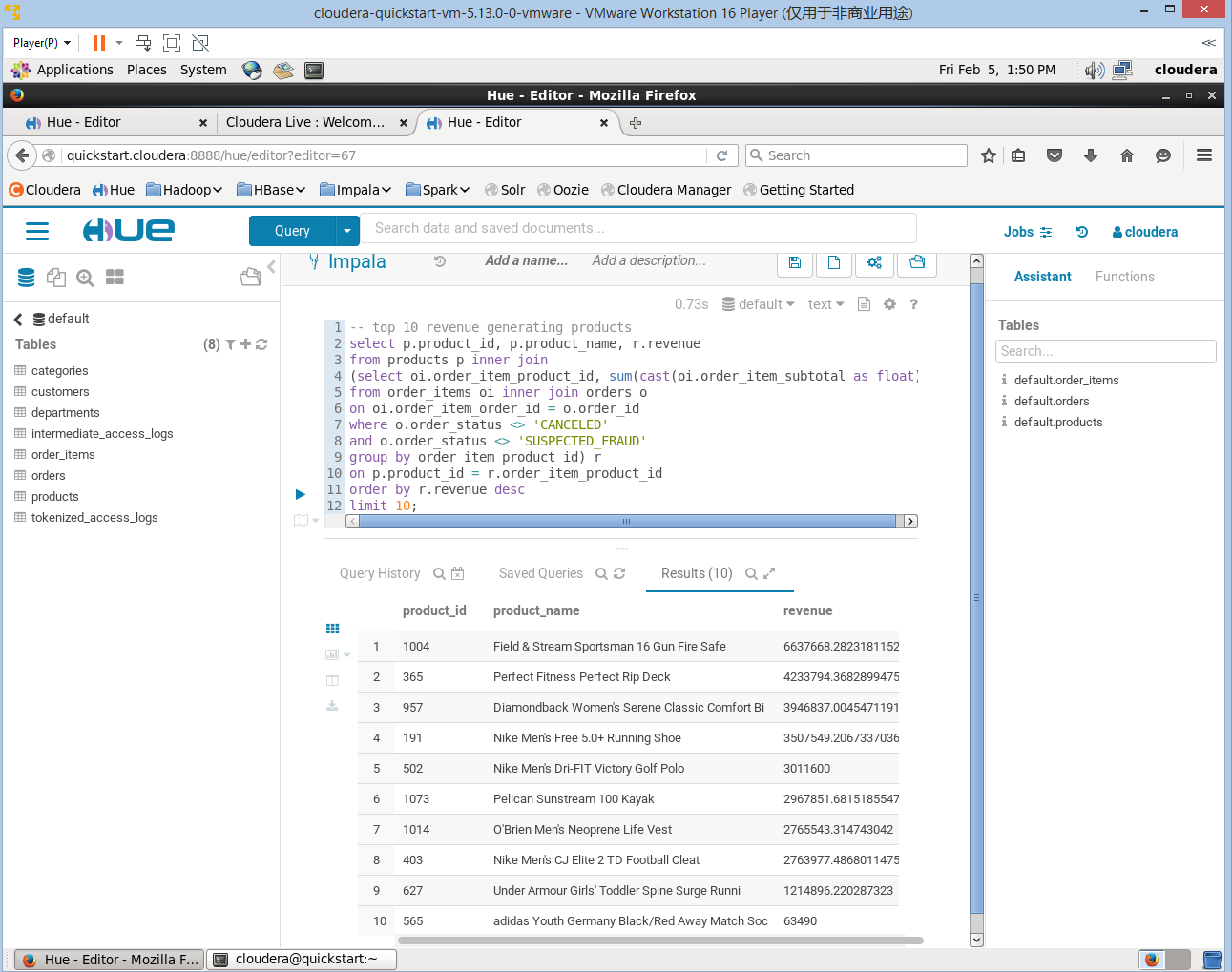


Figure 10

**Conclusion**: From Figure 10, Field & Stream Sportsman 16 Gun Fire Safe has the most revenue, followed by Perfect Fitness Perfect Rip Deck and Diamondback Women’s Serene Classic Comfort Bi.

**Reflection**: Between Feb 3rd and Feb 5th, I was running all the code on the Cloudera for 4 or 5 times, but I did not realize that every time I rerun the Sqoop code in the Terminal, I was appending all the tables together, so it was not surprising to see there are so many redundant rows in the Version 1 results.

I am glad that the TA pointed out this problem for me and proposed that I can remove all the tables and rerun the Sqoop code and other queries by using –rm command line in the Terminal for each table. After doing that, I successfully generated the desired table with 10 distinct products.

*Tutorial Exercise 2*

For this exercise, we can explore the business question that whether the most viewed products are most sold. We will use the web clickstream data to generate some findings.

1. We first used the pre-loaded sample access log data from the local filesystem into HDFS, and then verified that our data is in HDFS.

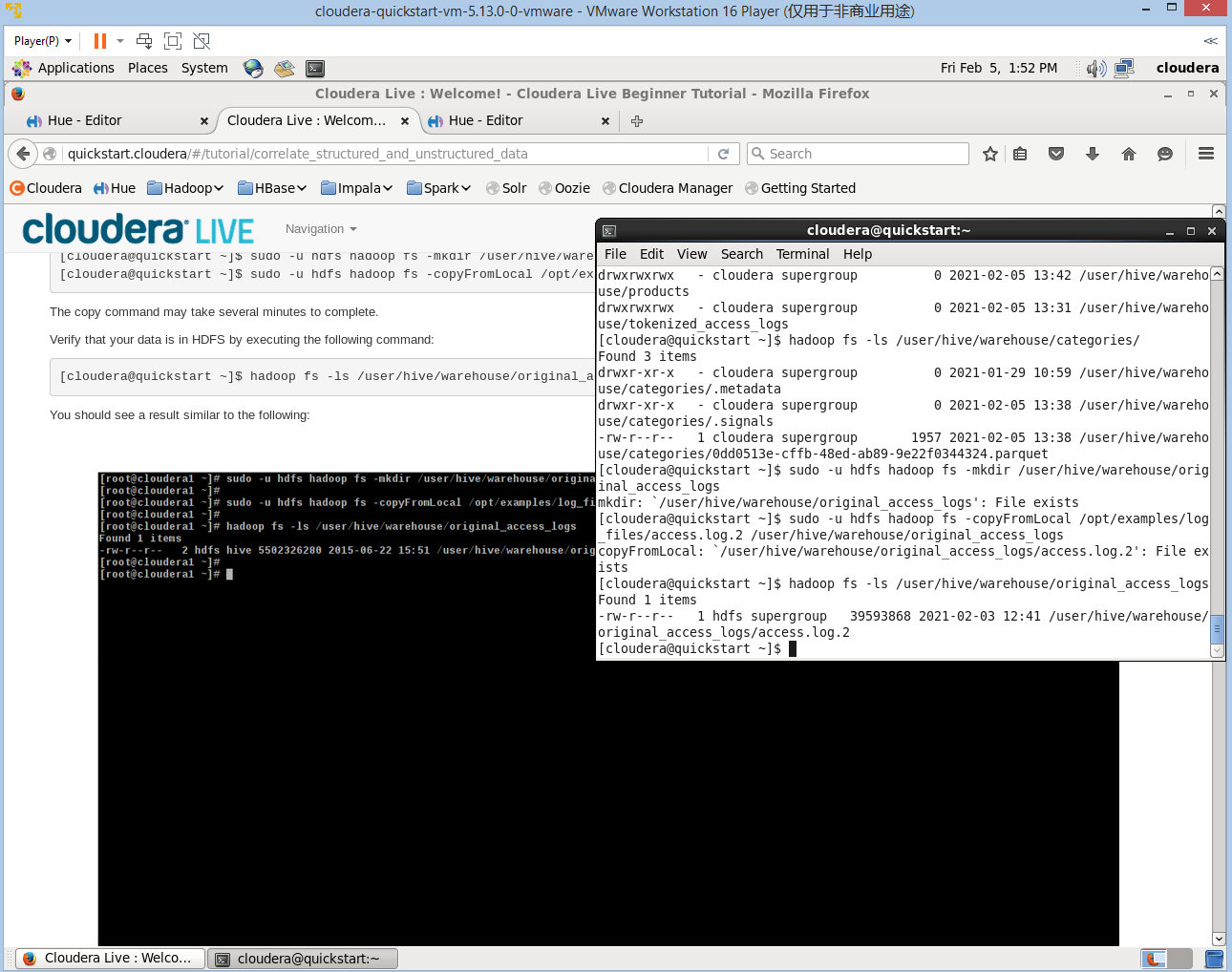


Figure 11

2. Then we built the intermediate table in Hive and transferred this to one that does not require any special serializers/deserializers.

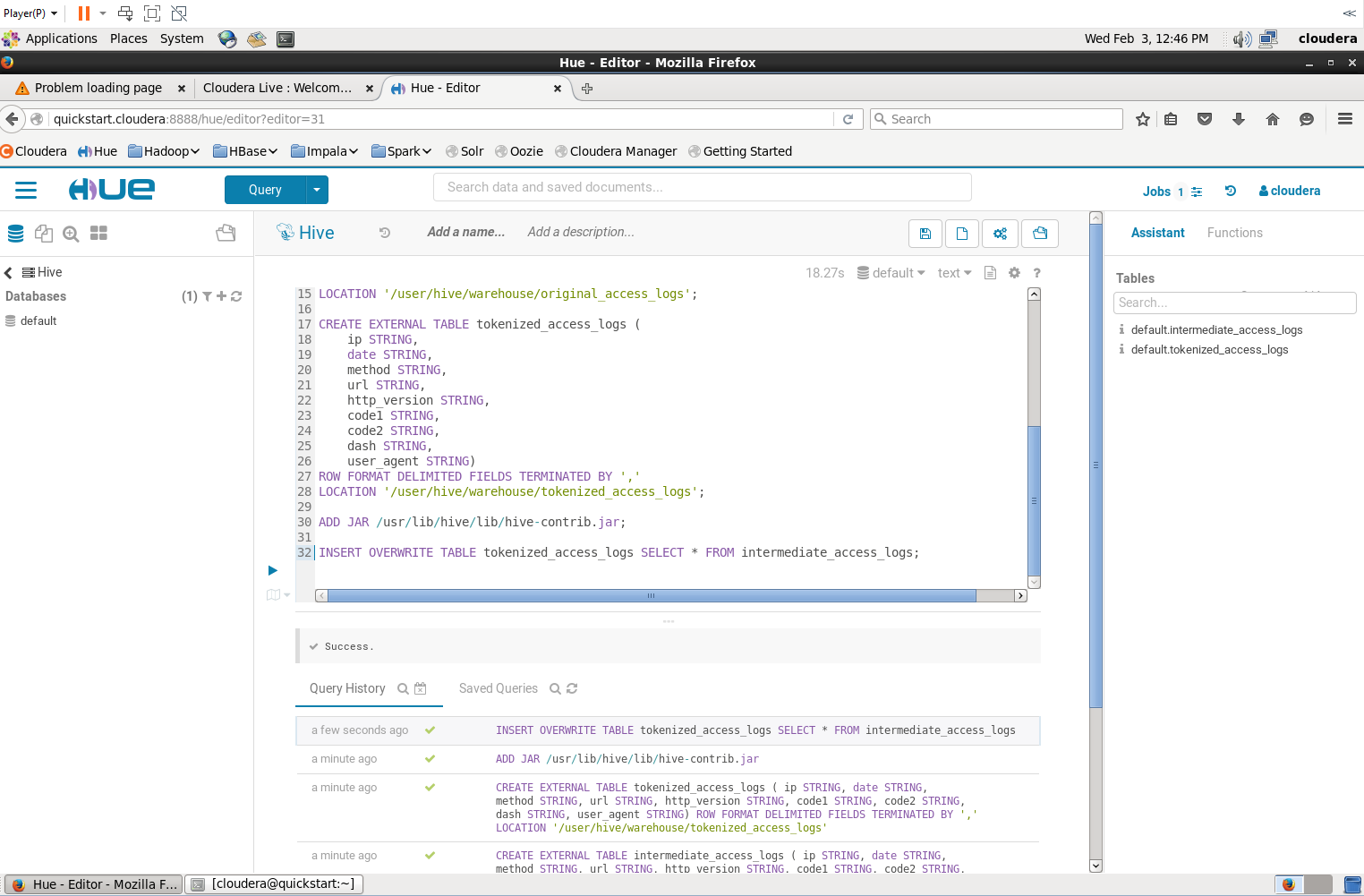


Figure 12

3. Then we told Impala that some tables have been created and run the queries in it to see the table that contains most sold products’ searched urls.

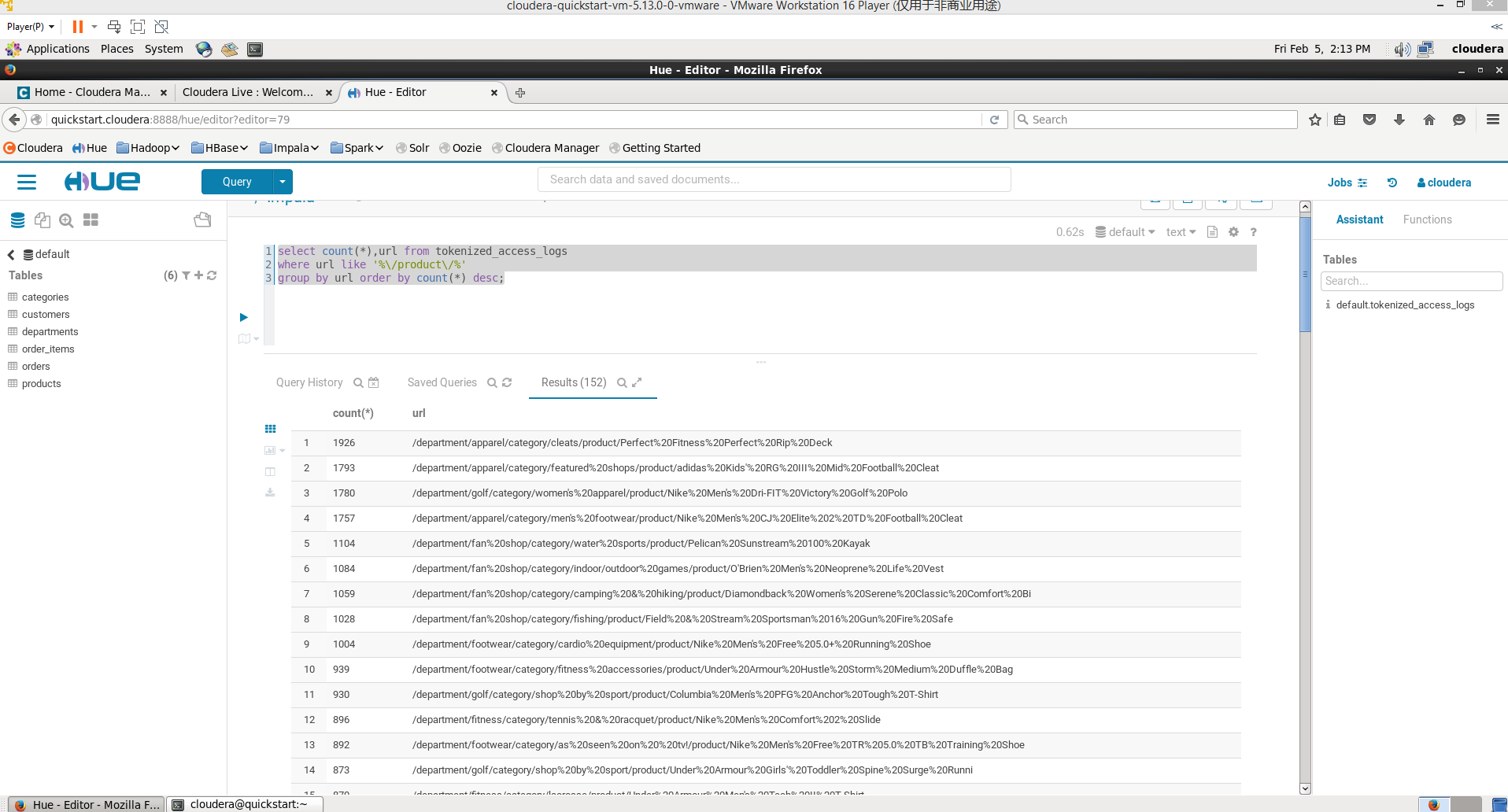


Figure 13

**Conclusion**: We can tell from Figure 13 that the most sold and searched product is Perfect Fitness Perfect Rip Deck, followed by adidas Kids RG ||| Mid Football Cleat and Nike Men’s 20Dri-Fit Victory Golf Polo.

*Answers*

* What is the 5th most revenue generating product?

**From Figure 10, the 5th most revenue generating product is Nike men’s Dry-Fit Victory Golf Polo.**

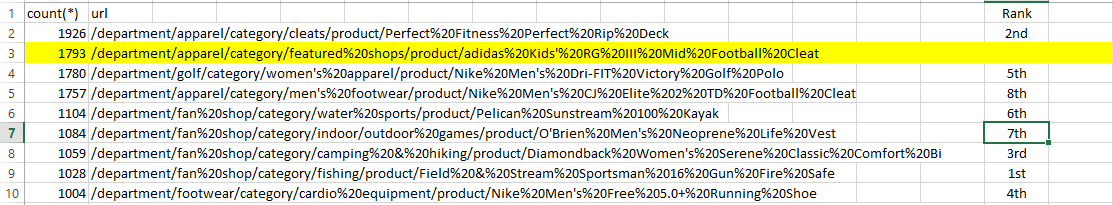
* How much revenue does the Nike men's dry fit polo earn?

**Also from Figure 10, Nike men’s dry fit polo earn the revenue of 3011600.**

* There is one product that did not show up in the previous result. It seems to be viewed a lot, but never purchased. Why?

**Comparing Figure 10 that contains top 10 products with most revenue to Figure 13 that contains at least 10 products with most views on the web, after crosschecking, we can tell that only Adidas Kids RG||| Mid Football Cleat that has 1793 count of views on the Internet did not show up in the Figure 10 which is the list of top 10 products most revenues. So lots of people viewed it, but never purchased.**

**The reason could be Kids’ football shoes are kind of hard or took time to finish purchasing than adults because of sizing problem, or people could borrow similar ones from family or friends or buy used ones somewhere else to save money since kids grow and shoe sizes change.**



**Summary**

Upon finishing this lab assignment, I have learned a lot along the way including setting up the environment, outsourcing like watching videos about Apache Hadoop to understand better, and asking the professor or the TA for help to resolve some issues. Fox example, the TA reminded me of a mistake that I made when I was running the queries in the Impala/Hive editor that is not selecting all the codes, meaning I just clicked the arrow running button which is just running the last line. Instead, I should select all the codes and then click the arrow button to get all the lines of codes running. This could make a difference to the results generated. And it is a nice experience to learn.

This learning experience of Apache Hadoop was very exciting and inspired me to learn deeper as it is a very powerful tool for me to continue learning Big Data Analytics.