

**Assignment: Lab Assignment #1**

Quoc Tuong Dong

Analytics, College of Professional Studies, Northeastern University

ALY6110 23240 Data Management & Big Data

Instructor: Doctor Reza Sadoddin

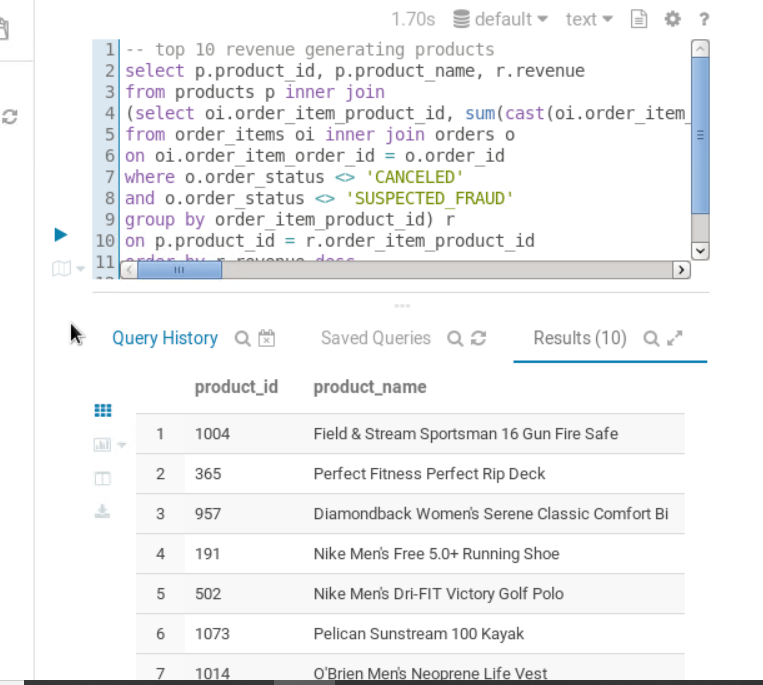
Due Date: March 29, 2020

In order to complete Lab 1 assignment, I followed the tutorial of Exercise 1 and 2.

**Exercise 1: “Ingest and query relational data”**

I first loaded the relational data from MySQL into HDFS, while preserving the structure and then launched the Sqoop job. After running the command, I verified that the data files existed in HDFS. Then, I used Hive and Impala to create tables by defining a schema over exiting files with “CREATE EXTERNAL TABLE” statements. I opened the Impala Query editor to try the first two codes. The last code is to answer the first two questions:

1. The 5th most revenue generating product is: **“Nike Men Dri-FIT Victory Golf Polo”**
2. Nike Men Dri-FIT Victory Golf Polo ‘s revenue is: **3011600**

-- top 10 revenue generating products

select p.product\_id, p.product\_name, r.revenue

from products p inner join

(select oi.order\_item\_product\_id, sum(cast(oi.order\_item\_subtotal as float)) as revenue

from order\_items oi inner join orders o

on oi.order\_item\_order\_id = o.order\_id

where o.order\_status <> ‘CANCELED’

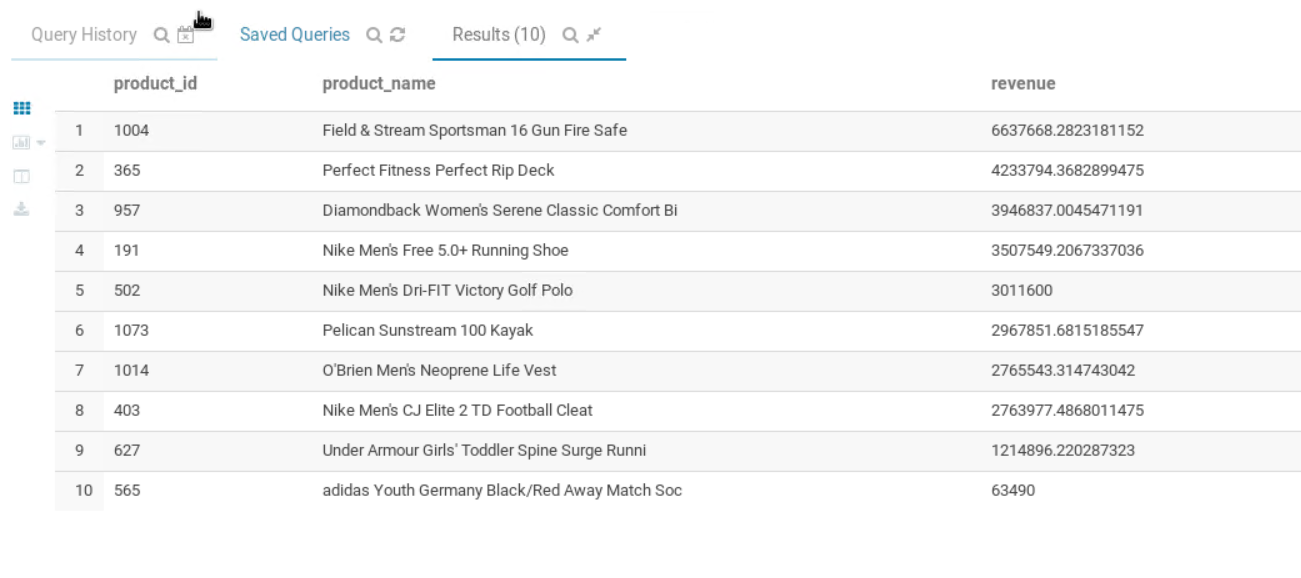
and o.order\_status <> ‘SUSPECTED\_FRAUD’

group by order\_item\_product\_id) r

on p.product\_id = r.order\_item\_product\_id

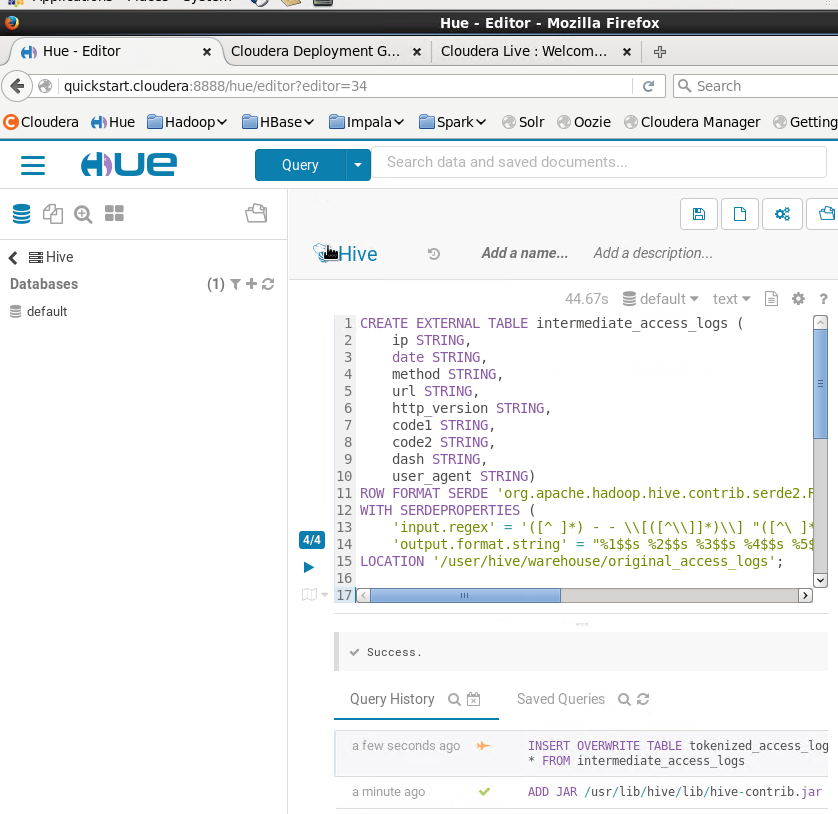
order by r.revenue desc

limit 10;

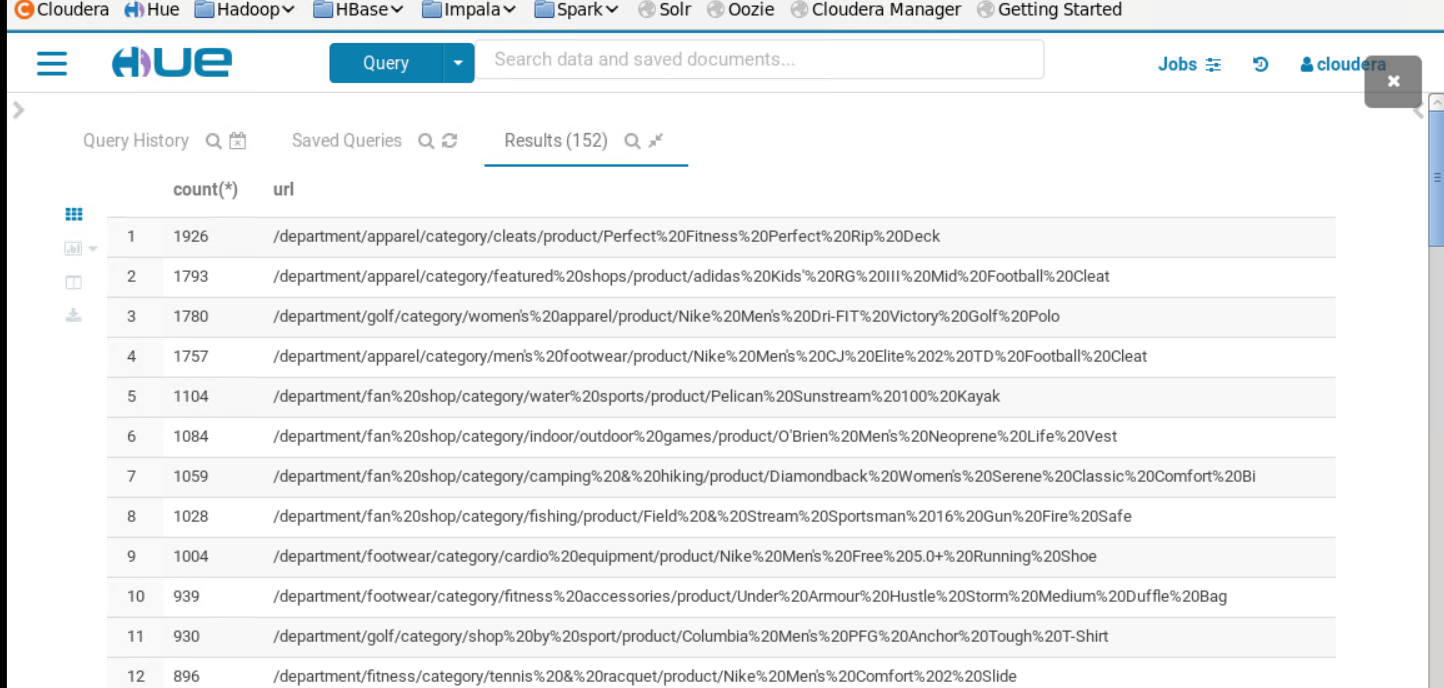


**Exercise 2: “Ingest and query relational data”**

For this exercise I move the data from the local file system, into HDFS using the sudo command from the tutorial. Then I verified the data is in HDFS or not with the hadoop fs command. Next, I followed the Hive Query Editor app in Hue to execute the code



Then after that I went back to Impala Query Editor to put the last two codes inside, which will give us this result of the most visited product’s url



1. From a respective point of view**, “Adidas Kids III Mild Football Cleat”** is missing in the Revenue table compared to here. So when I looked at the tutorial, it is said that the there is a typo in the price for the item. Such issue misled the users to think it is more expensive than expected and thus do not want to purchase the product at all. After correcting the price, the tutorial said that they could see the SKU skyrocketed rapidly.