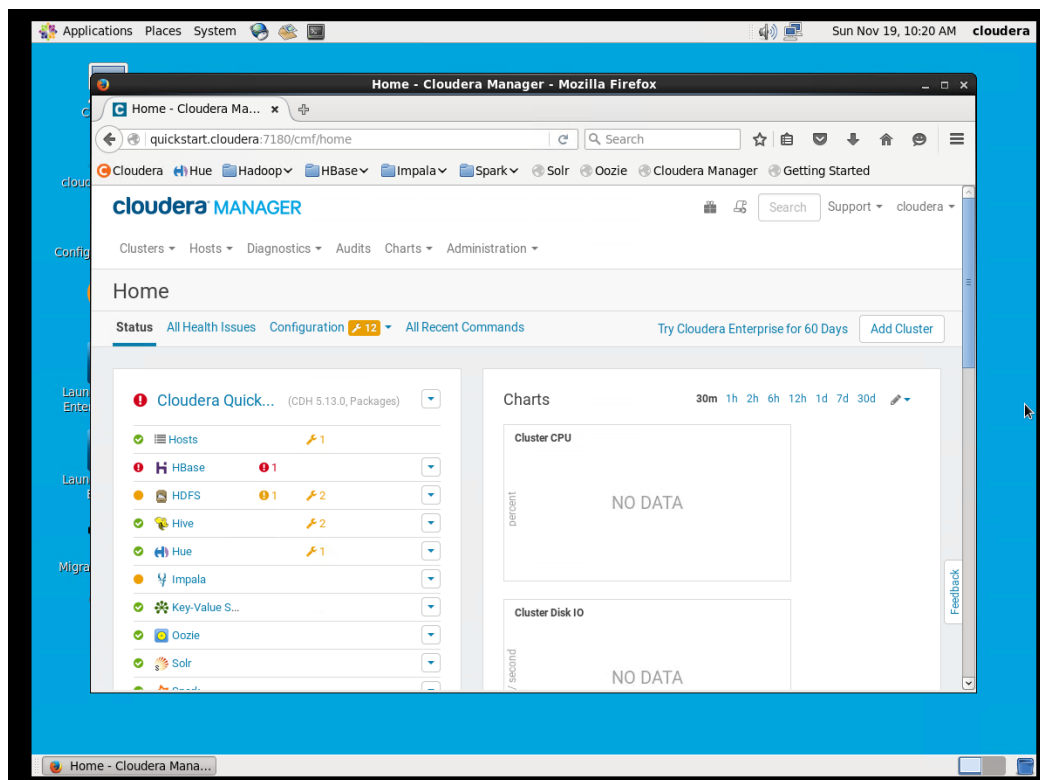


# Module 3 Lab — Individual Lab #1

By Mohammad Movahedi

## Exercise 1: Ingest and Query Relational Data

Step 0 : login and open Cloudera manager

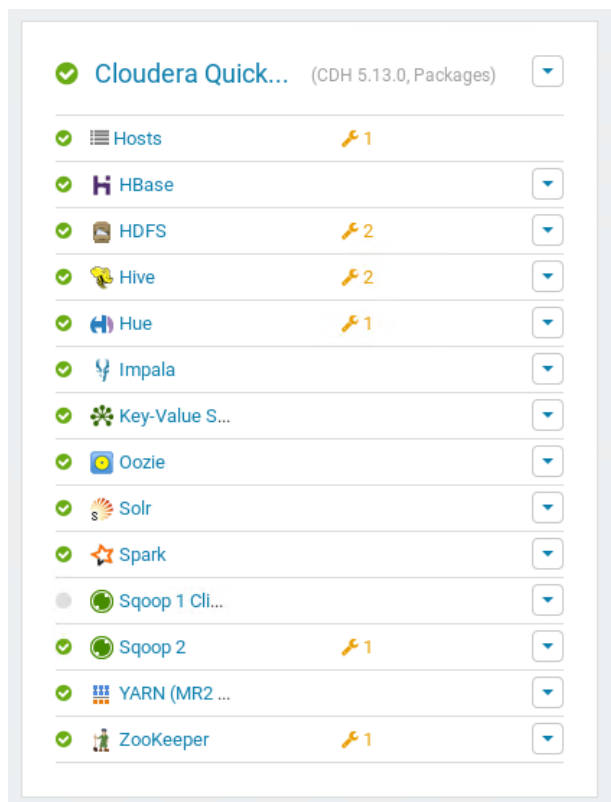


## Step 1: Verify Environment

Making Sure the services are running . as can be seen the following services in Cloudera Manager are up and running:

- Apache Impala
- Apache Hive
- HUE

- HDFS
- YARN



## Step 2: Ingest Data Using Apache Sqoop

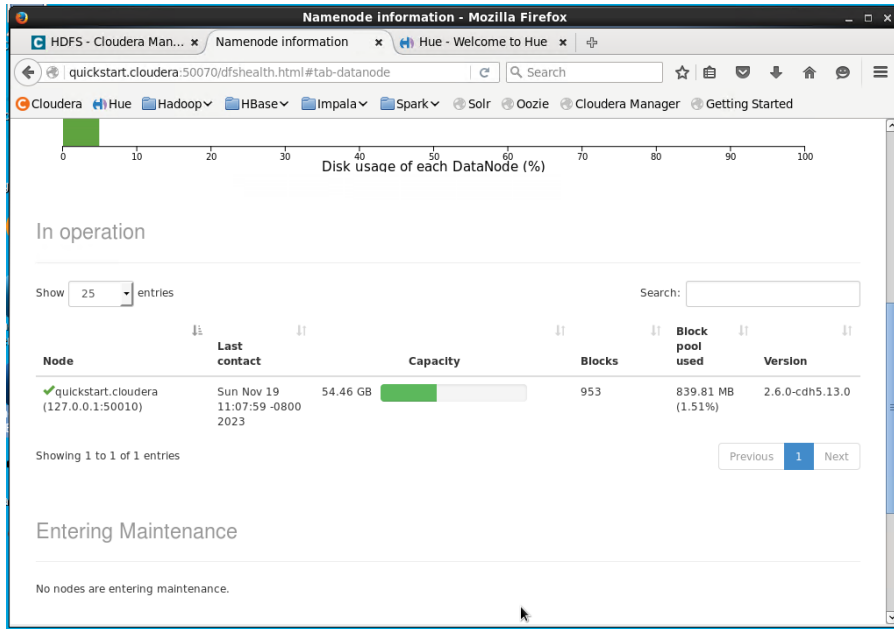
First I need to log in to the Master Node of your cluster. This is done via a terminal. Before I do that I look into directories

```
cloudera@quickstart:/  
File Edit View Search Terminal Help  
[cloudera@quickstart ~]$ cd /usr/lib  
[cloudera@quickstart lib]$ ls  
anaconda-runtime      hadoop-hdfs          llama                sentry  
avro                   hadoop-httpfs        locale              solr  
bigtop-tomcat          hadoop-kms           lsb                 spark  
bigtop-utils           hadoop-mapreduce     mahout              sqoop  
bonobo                 hadoop-yarn           mozilla              sqoop2  
ConsoleKit             hbase                oozie               vmware-tools  
crunch                 hbase-solr           parquet             vmware-vgauth  
cups                   hive                 pig                 whirr  
flume-ng               hive-hcatalog         python2.6            yum-plugins  
games                  hue                   rpm                 zookeeper  
gcc                    impala                search               zookeeper-native  
hadoop                 impala-shell          sendmail               
hadoop-0.20-mapreduce  kite                  sendmail.postfix  
[cloudera@quickstart lib]$ cd /opt/couldera  
bash: cd: /opt/couldera: No such file or directory  
[cloudera@quickstart lib]$ cd /opt/cloudera  
[cloudera@quickstart cloudera]$ ls  
csd parcel-cache parcel-repo parcels  
[cloudera@quickstart cloudera]$ cd ..  
[cloudera@quickstart opt]$ cd ..  
[cloudera@quickstart /]$ cd ..  
[cloudera@quickstart /]$
```

Then I launch cloudera express



Then I find the address to the live node



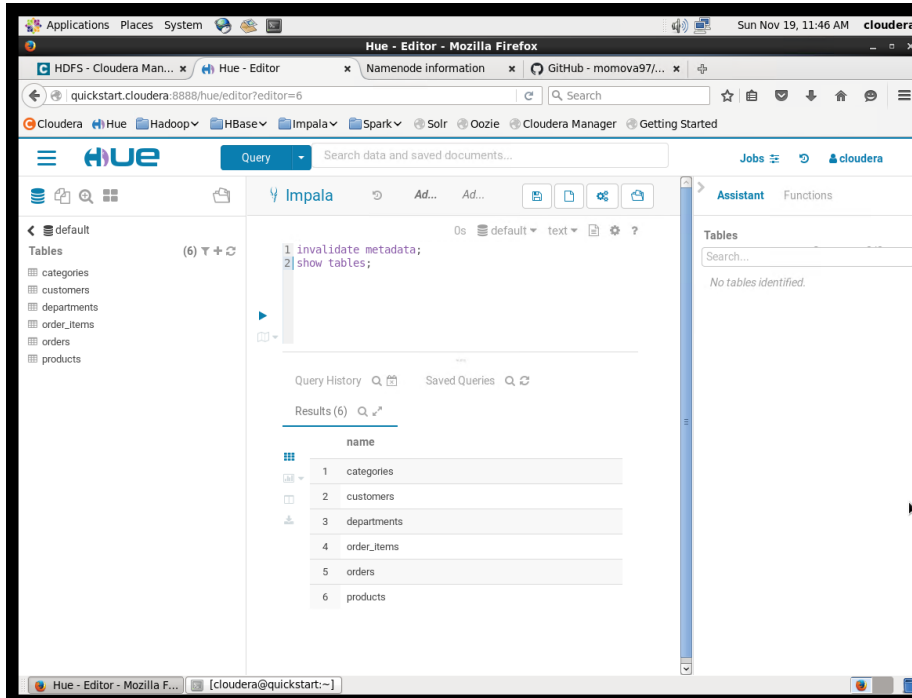
## Step 3: Verify Data in HDFS

Then I check to see if hadoop is running

```
cloudera@quickstart:~$ ls
Downloads  lib  Templates
codegen_categories.java  eclipse  Music  Videos
express-deployment.json  parcels  workspace
kerberos  Pictures  Public
Documents

cloudera@quickstart ~$ hadoop fs -ls /user/hive/warehouse/
Found 6 items
drwxr-xr-x - cloudera supergroup 0 2019-05-30 13:20 /user/hive/warehouse/use/categories
drwxr-xr-x - cloudera supergroup 0 2019-05-30 13:21 /user/hive/warehouse/use/customers
drwxr-xr-x - cloudera supergroup 0 2019-05-30 13:23 /user/hive/warehouse/use/departments
drwxr-xr-x - cloudera supergroup 0 2019-05-30 13:24 /user/hive/warehouse/use/order items
drwxr-xr-x - cloudera supergroup 0 2019-05-30 13:25 /user/hive/warehouse/use/orders
drwxr-xr-x - cloudera supergroup 0 2019-05-30 13:26 /user/hive/warehouse/use/products
[cloudera@quickstart ~]$ hadoop fs -ls /user/hive/warehouse/categories/
Found 5 items
drwxr-xr-x - cloudera supergroup 0 2019-05-07 06:40 /user/hive/warehouse/categories/.metadata
drwxr-xr-x - cloudera supergroup 0 2019-05-30 13:20 /user/hive/warehouse/use/categories/signals
-rw-r--r-- 1 cloudera supergroup 1957 2019-05-07 06:41 /user/hive/warehouse/categories/5a6f112-ea16-4844-86f3-1f271f217418.parquet
-rw-r--r-- 1 cloudera supergroup 1957 2019-05-30 13:20 /user/hive/warehouse/use/categories/63f3fcb-d4bc-4808-b7b0-2b73c19bed6f.parquet
-rw-r--r-- 1 cloudera supergroup 1957 2019-05-28 08:46 /user/hive/warehouse/use/categories/bb760f36-f2f2-42e7-8ea9-b19af1afdf85.parquet
[cloudera@quickstart ~]$
```

## Step 4: Query Data Using Impala

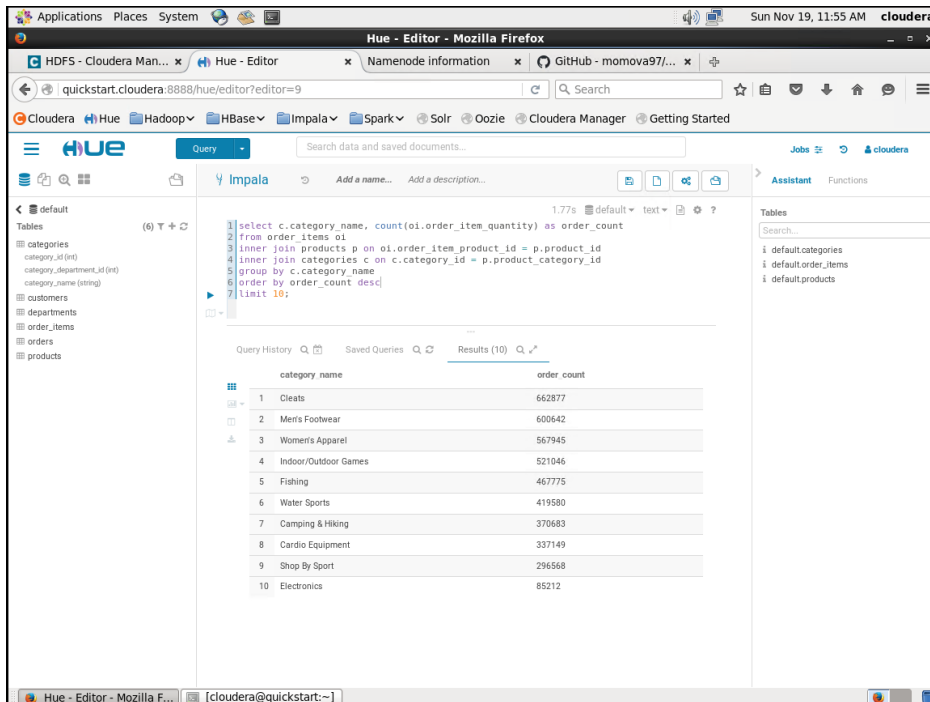


The screenshot shows the Hue Impala interface in a Mozilla Firefox browser. The query editor contains two lines of SQL: `1 invalidate metadata;` and `2 show tables;`. The results pane displays a table with one column, 'name', and six rows of data: categories, customers, departments, order\_items, orders, and products. The left sidebar shows a tree view of the database schema, and the right sidebar shows the 'Tables' section with a search bar.

name
1 categories
2 customers
3 departments
4 order_items
5 orders
6 products

As can be seen the data is loaded.

For the first example I got to ten most ordered categories



The screenshot shows the Hue Impala interface with a more complex SQL query. The query editor contains the following SQL: `1 select c.category_name, count(o.order_item_quantity) as order_count`, `2 from order_items o`, `3 inner join products p on o.order_item_product_id = p.product_id`, `4 inner join categories c on c.category_id = p.product_category_id`, `5 group by c.category_name`, `6 order by order_count desc`, and `7 limit 10;`. The results pane displays a table with two columns: 'category\_name' and 'order count'. The table lists the top 10 categories by order count, starting with 'Cleats' and ending with 'Electronics'.

category_name	order count
1 Cleats	662877
2 Men's Footwear	600642
3 Women's Apparel	567945
4 Indoor/Outdoor Games	521046
5 Fishing	467775
6 Water Sports	419580
7 Camping & Hiking	370683
8 Cardio Equipment	337149
9 Shop By Sport	296568
10 Electronics	85212

And in the next example I got top 10 most revenue generating products

The screenshot shows the Hue Editor interface in a Mozilla Firefox browser. The URL is `quickstart.cloudera:8888/hue/editor?editor=10`. The interface includes a sidebar with a file tree, a central query editor, and a results pane.

**Query:**

```
-- top 10 revenue generating products
1 SELECT p.product_id,
2 p.product_name,
3 r.revenue
4 FROM products p
5 INNER JOIN (
6 SELECT oi.order_item_product_id,
7 SUM(CAST(oi.order_item_subtotal AS FLOAT)) AS revenue
8 FROM order_items oi
```

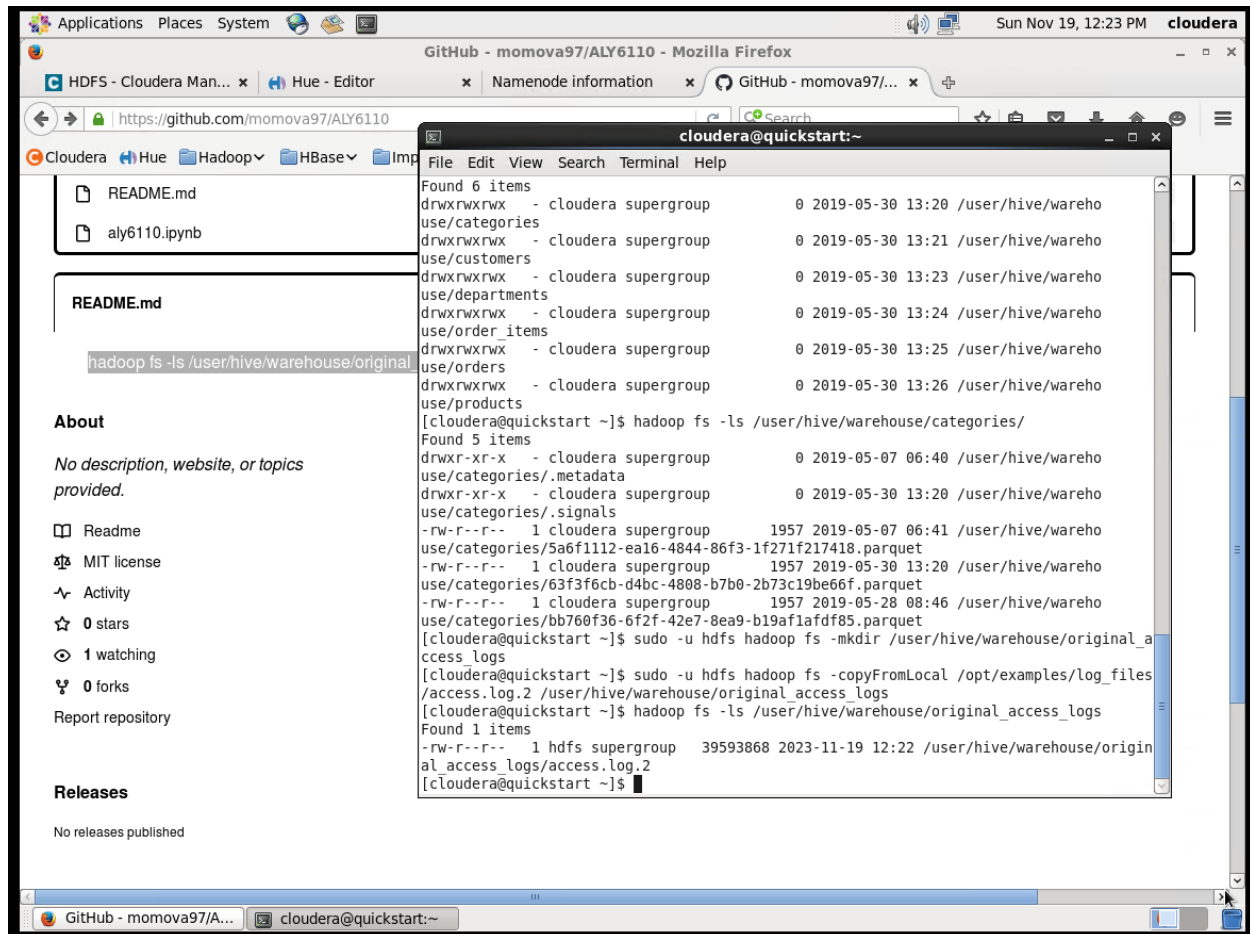
**Results (10):**

	product_id	product_name	revenue
1	1004	Field & Stream Sportsman 16 Gun Fire Safe	59739014.540863037
2	1004	Field & Stream Sportsman 16 Gun Fire Safe	59739014.540863037
3	1004	Field & Stream Sportsman 16 Gun Fire Safe	59739014.540863037
4	365	Perfect Fitness Perfect Rip Deck	38104149.314609528
5	365	Perfect Fitness Perfect Rip Deck	38104149.314609528
6	365	Perfect Fitness Perfect Rip Deck	38104149.314609528
7	957	Diamondback Women's Serene Classic Comfort Bi	35521533.040924072
8	957	Diamondback Women's Serene Classic Comfort Bi	35521533.040924072
9	957	Diamondback Women's Serene Classic Comfort Bi	35521533.040924072
10	191	Nike Meri's Free 5.0+ Running Shoe	31567942.860603333

## Exercise 2: Correlate Structured and Unstructured Data

### Step 1: Prepare Data

First I load the data into the hadoop



## Step 2: Create and Query Tables in Hive

I run the code given in the file to create two tables: `intermediate_access_logs` and `tokenized_access_logs`.

Applications Places System Sun Nov 19, 12:49 PM cloudera

Hue - Editor - Mozilla Firefox

HDFS - Cloudera Man... x Hue - Editor x Namenode information x GitHub - momova97/... x

quickstart.cloudera:8888/hue/editor?editor=30

Cloudera Hue Hadoop HBase Impala Spark Solr Oozie Cloudera Manager Getting Started

Query Search data and saved documents...

Jobs cloudera

Hive

Databases (1) +

default

31.46s default text

```
16
17 CREATE EXTERNAL TABLE tokenized_access_logs (
18   ip STRING,
19   date STRING,
20   method STRING,
21   url STRING,
22   http_version STRING,
23   code1 STRING,
24   code2 STRING,
25   dash STRING,
26   user_agent STRING)
27 ROW FORMAT DELIMITED
28 FIELDS TERMINATED BY ','
29 LOCATION '/user/hive/warehouse/tokenized_access_logs';
30
31 ADD JAR /usr/lib/hive/lib/hive-contrib.jar;
32
33 INSERT OVERWRITE TABLE tokenized_access_logs
34 SELECT * FROM intermediate_access_logs;
```

Success.

Query History Saved Queries

a minute ago	✓	INSERT OVERWRITE TABLE tokenized_access_logs SELECT * FROM intermediate_access_logs
a minute ago	✓	ADD JAR /usr/lib/hive/lib/hive-contrib.jar
2 minutes ago	!	ADD JAR usr/lib/hive/lib/hive-contrib.jar
2 minutes ago	!	ADD JAR usr/lib/hive/lib/hive-contrib.jar
4 minutes ago	!	ADD JAR usr/lib/hive/lib/hive-contrib.jar
5 minutes ago	!	ADD JAR usr/lib/hive/lib/hive-contrib.jar

Tables

Search...

- default.intermediate\_access\_logs
- default.tokenized\_access\_logs

Hue - Editor - Mozilla F... cloudera@quickstart:/...



## Step 3: Query Data Using Impala

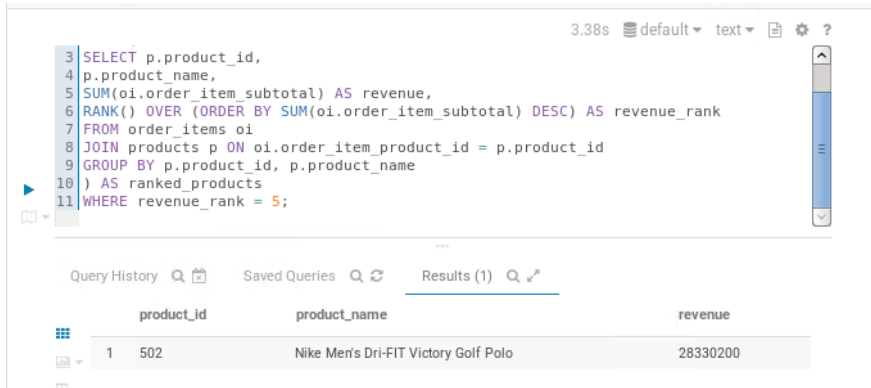
The screenshot shows the Hue web interface for Impala. The main editor area contains a query history table with the following data:

Time Ago	Status	Query
4 minutes ago	✓	SELECT COUNT(*) AS count, url FROM tokenized_access_logs WHERE url LIKE '%/product/%' GROUP BY url ORDER BY count DESC
6 minutes ago	!	SELECT count(*),url FROM tokenized_access_logs WHERE url LIKE '%/product/%' GROUP BY url ORDER BY count DESC;
8 minutes ago	!	SELECT COUNT(*), url FROM tokenized_access_logs WHERE url LIKE '%/product/%' GROUP BY url ORDER BY count DESC;
9 minutes ago	⚡	SELECT COUNT(*) AS count, url FROM tokenized_access_logs WHERE url LIKE '%/product/%' GROUP BY url ORDER BY count DESC
10 minutes ago	⚡	SELECT COUNT(*) AS count, url FROM tokenized_access_logs WHERE url LIKE '%/product/%' GROUP BY url ORDER BY count DESC
12 minutes ago	!	select count(*), url from tokenized_access_logs where url like '%/product/%' group by url order by count() desc;
13 minutes ago	!	select count(*) url from tokenized_access_logs where url like '%/product/%' group by url order by count() desc;
13 minutes ago	!	select count(url) from tokenized_access_logs where url like '%/product/%' group by url order by count() desc;
15 minutes ago	⚡	SHOW TABLES
16 minutes ago	⚡	INVALIDATE METADATA
an hour ago	✓	SHOW TABLES
an hour ago	✓	SHOW TABLES
an hour ago	⚡	INVALIDATE METADATA

The interface also includes a left sidebar with a table list (e.g., categories, customers, departments, tokenized\_access\_logs) and a right sidebar with a search bar and a 'Tables' section.

After this I answer the questions

**1,2- What is the 5th most revenue generating product? And how much revenue it generated**



```
3 SELECT p.product_id,
4 p.product_name,
5 SUM(oi.order_item_subtotal) AS revenue,
6 RANK() OVER (ORDER BY SUM(oi.order_item_subtotal) DESC) AS revenue_rank
7 FROM order_items oi
8 JOIN products p ON oi.order_item_product_id = p.product_id
9 GROUP BY p.product_id, p.product_name
10 ) AS ranked_products
11 WHERE revenue_rank = 5;
```

product_id	product_name	revenue
502	Nike Men's Dri-FIT Victory Golf Polo	28330200

The answer is product 502

And it generated 283302 \$

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

When a product gets a lot of views but isn't making sales, several factors might be at play:

- Pricing: It might be too expensive compared to other options.
- Availability Issues: Often out of stock, so people can't buy it.
- Product Page Problems: Unclear descriptions or poor images might be turning people off.
- Technical Issues: Problems with the website or app could be stopping sales.
- Bad Reviews: Negative feedback might be discouraging purchases.
- Market Trends/Seasonality: The product might not be in demand right now.
- Analytics Errors: The data you're seeing could be inaccurate.
- Comparison Shopping: People might be checking it out on your site but buying it somewhere else.

# Conclusion

In conclusion, this assignment involved creating and executing SQL queries to analyze product data, focusing on identifying revenue-generating products and investigating discrepancies between product views and purchases. We tailored queries to extract meaningful insights from complex datasets, revealing critical business dynamics like pricing strategies, customer engagement, and market trends. This exercise not only demonstrated the power of SQL in data analysis but also underscored the importance of understanding data context for effective decision-making.

# References

Azure.com. (2023). *Azure Lab Services*. [online] Available at:  
<https://labs.azure.com/virtualmachines> [Accessed 19 Nov. 2023].

Instructure.com. (2020). *Module 3 Lab — Individual Lab #1*. [online] Available at:  
<https://northeastern.instructure.com/courses/160780/assignments/2036226> [Accessed 19 Nov. 2023].

Getting Started with Hadoop Tutorial CLOUDER A DEPLOYMENT GUIDE. (n.d.).  
Available at:  
<https://www.cloudera.com/content/dam/www/marketing/documents/partners/ungated/cloudera-msazure-hadoop-deployment-guide.pdf>.