College of Professional Studies

Northeastern University San Jose

MPS Analytics

Course: ALY6110 – Data Management and Big Data

Assignment:

Module 3 Lab 1

Submitted on:

June 17th, 2023

Submitted to:

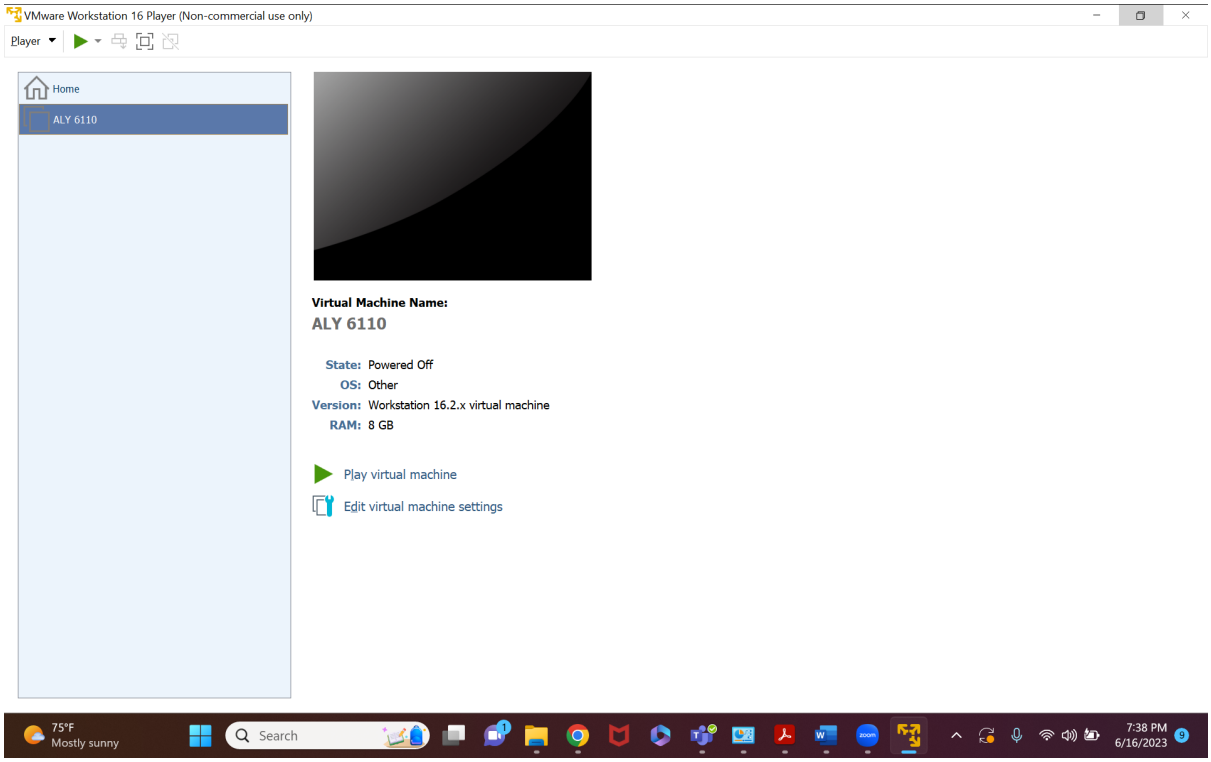
Professor: BEHZAD AHMADI

Submitted by:

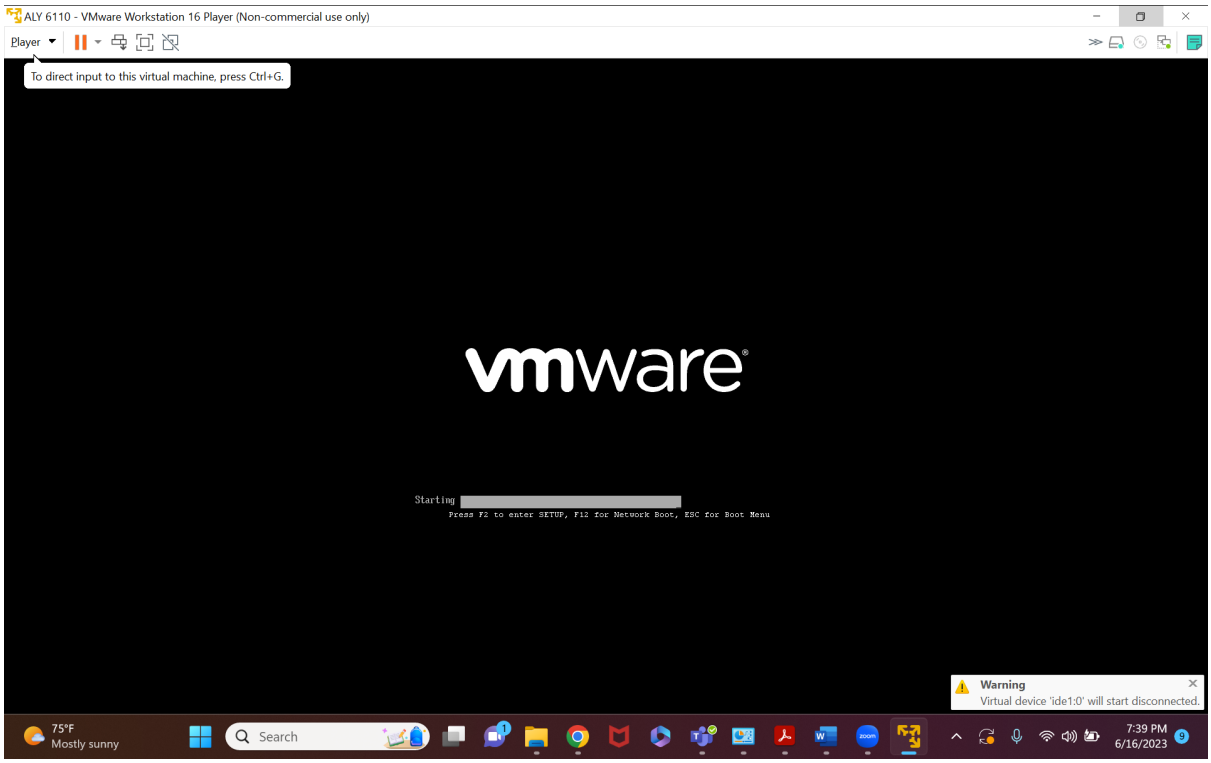
NIKSHITA RANGANATHAN

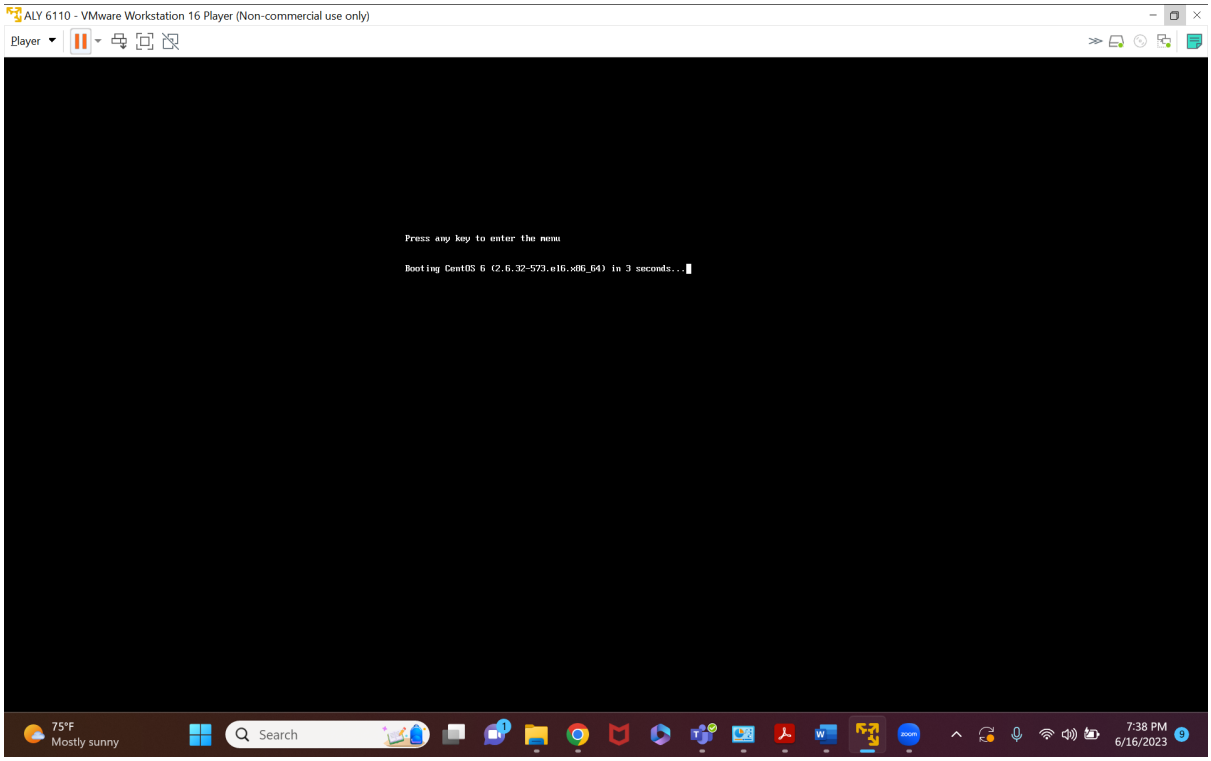
Exercise-1 Ingest and query relational data

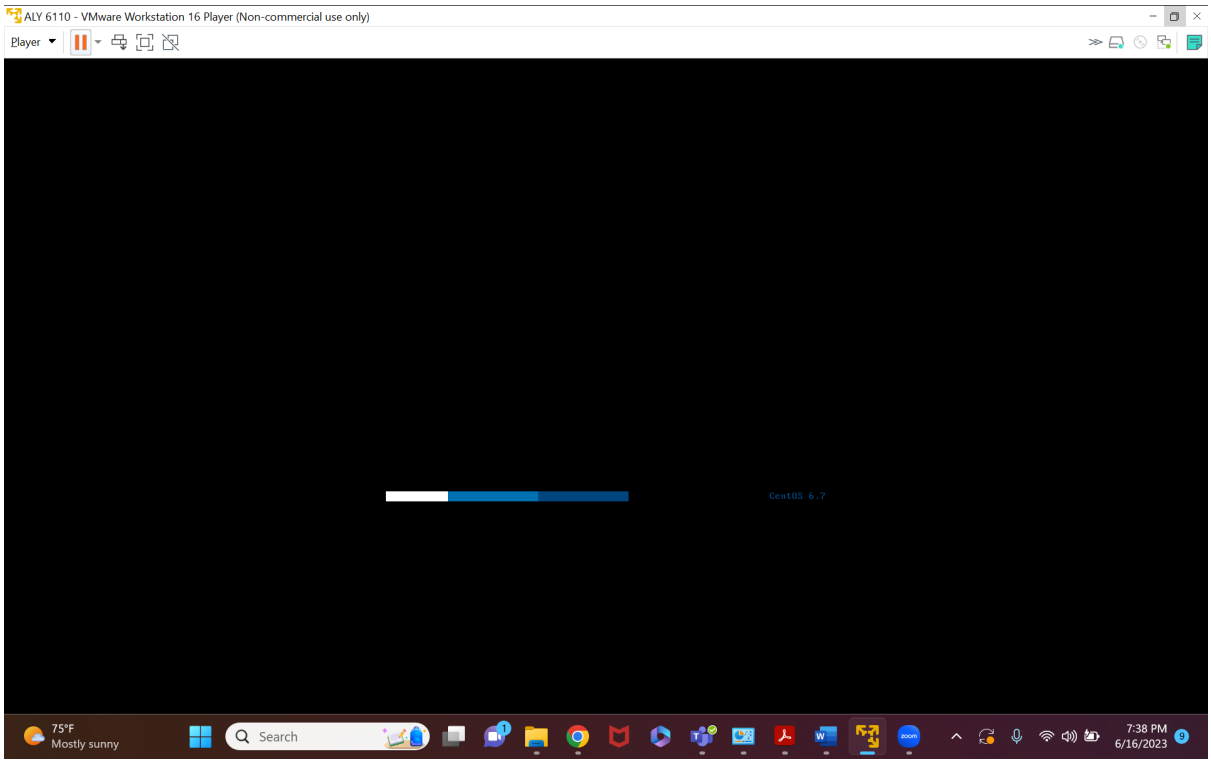
1. ALY 6110 Virtual Machine

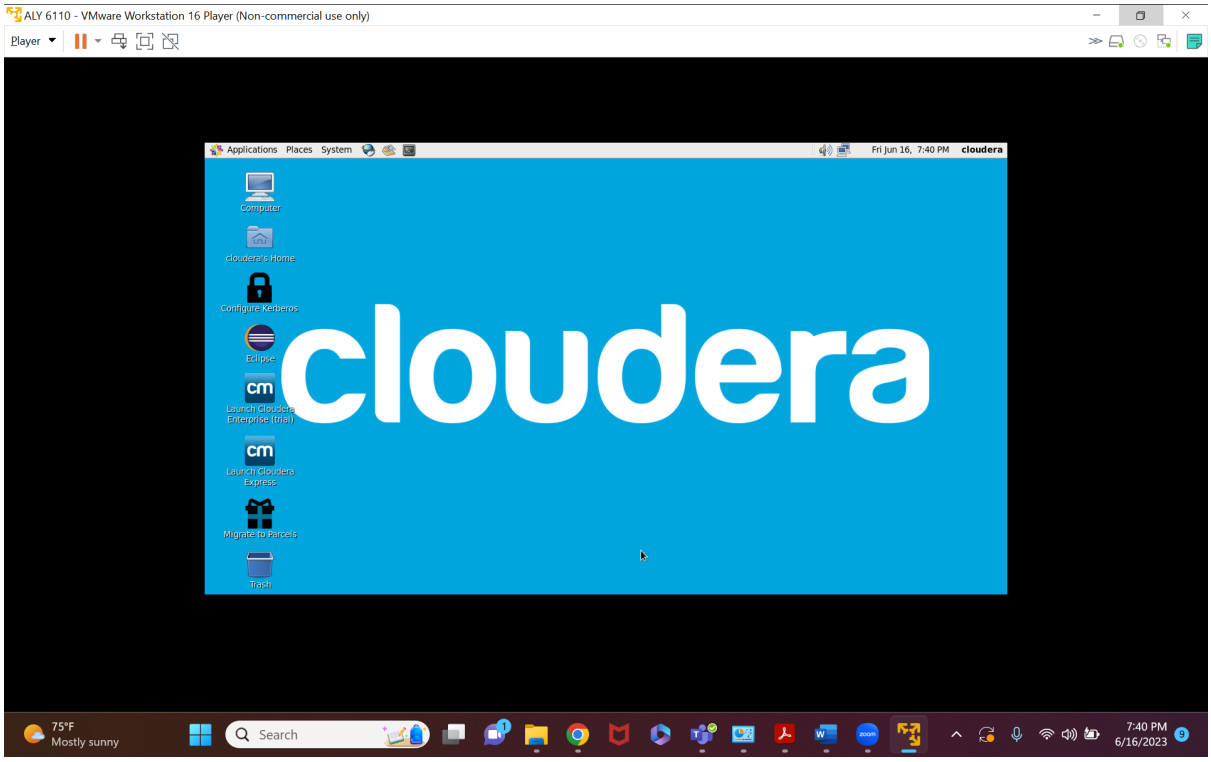


2. Starting the virtual machine

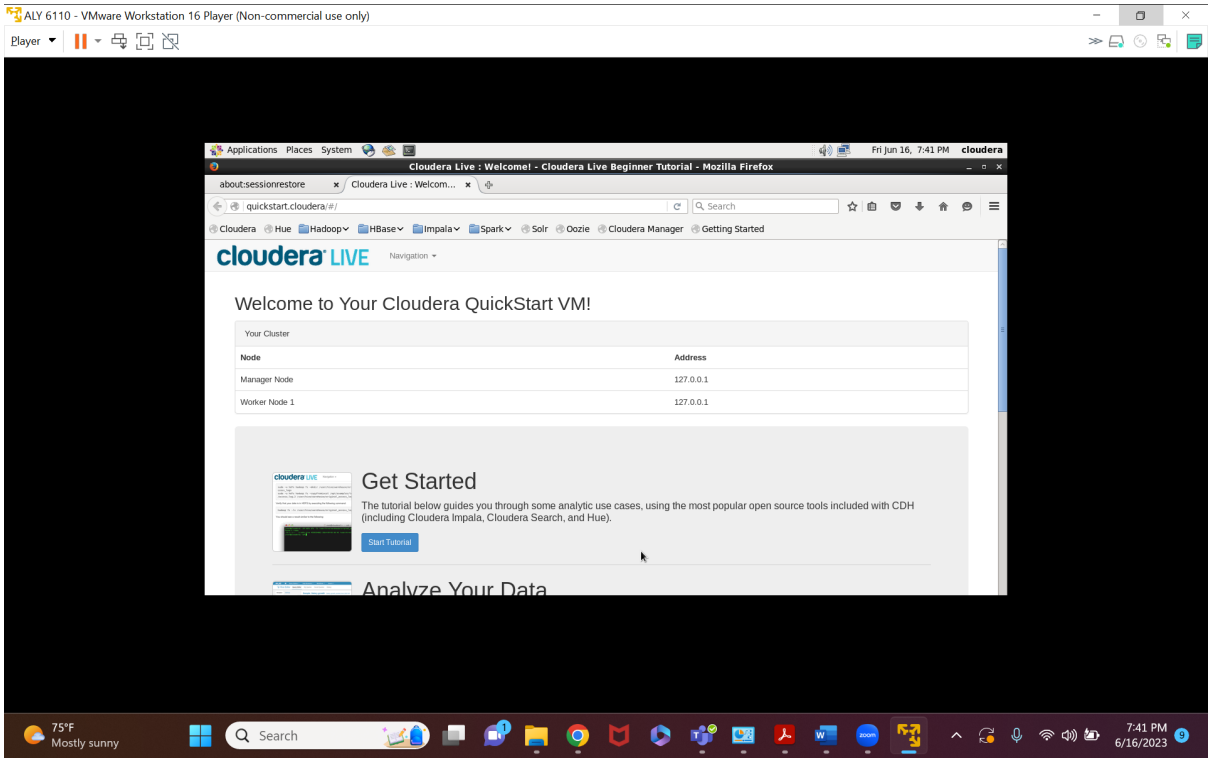



ALY 6110 - VMware Workstation 16 Player (Non-commercial use only)
Player
Press any key to enter the menu
Booting CentOS 6 (2.6.32-573.el6.x86_64) in 3 seconds...
75°F
Mostly sunny
Search
7:38 PM
6/16/2023

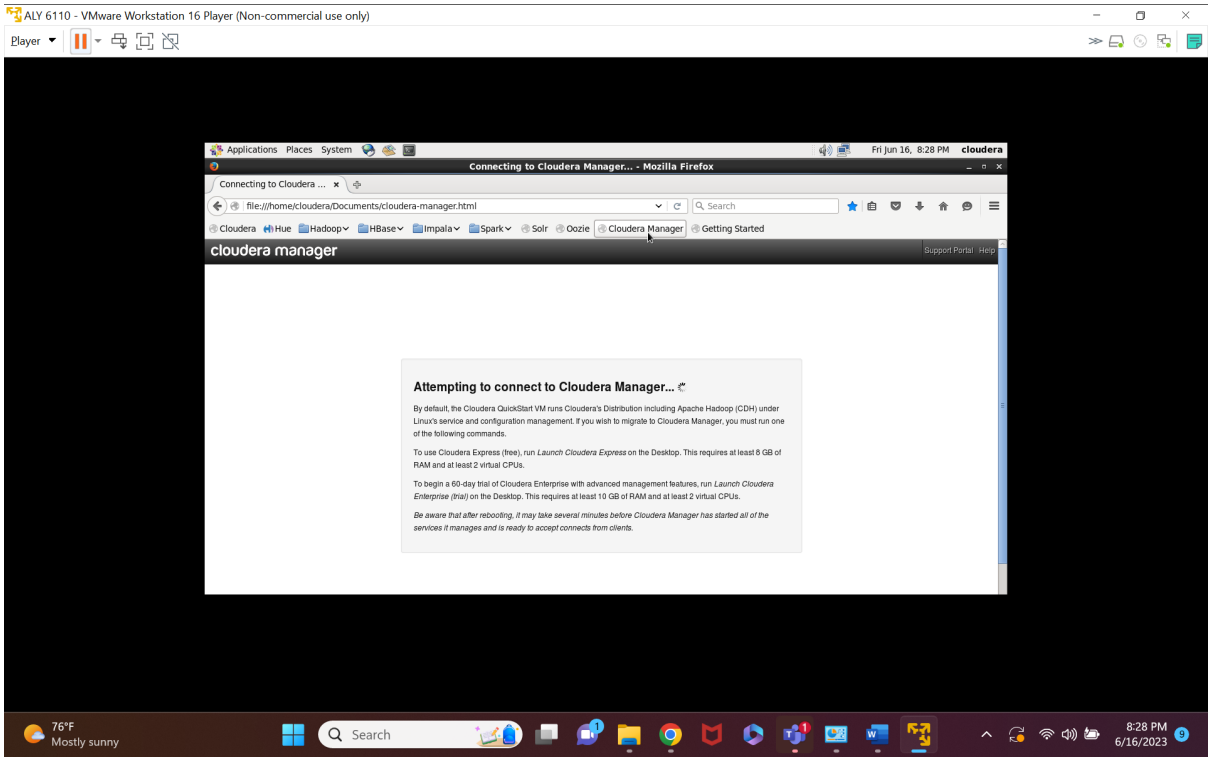

ALY 6110 - VMware Workstation 16 Player (Non-commercial use only)
Player
CentOS 6.7
75°F
Mostly sunny
Search
7:38 PM
6/16/2023

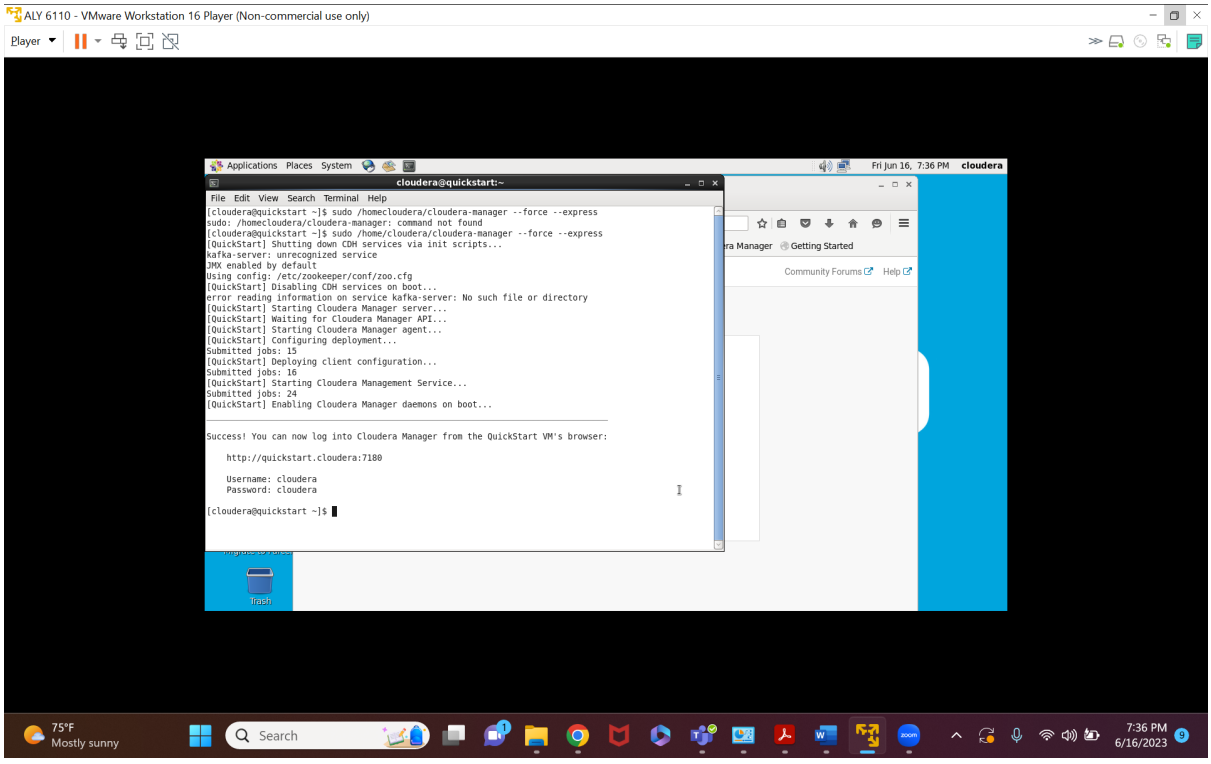


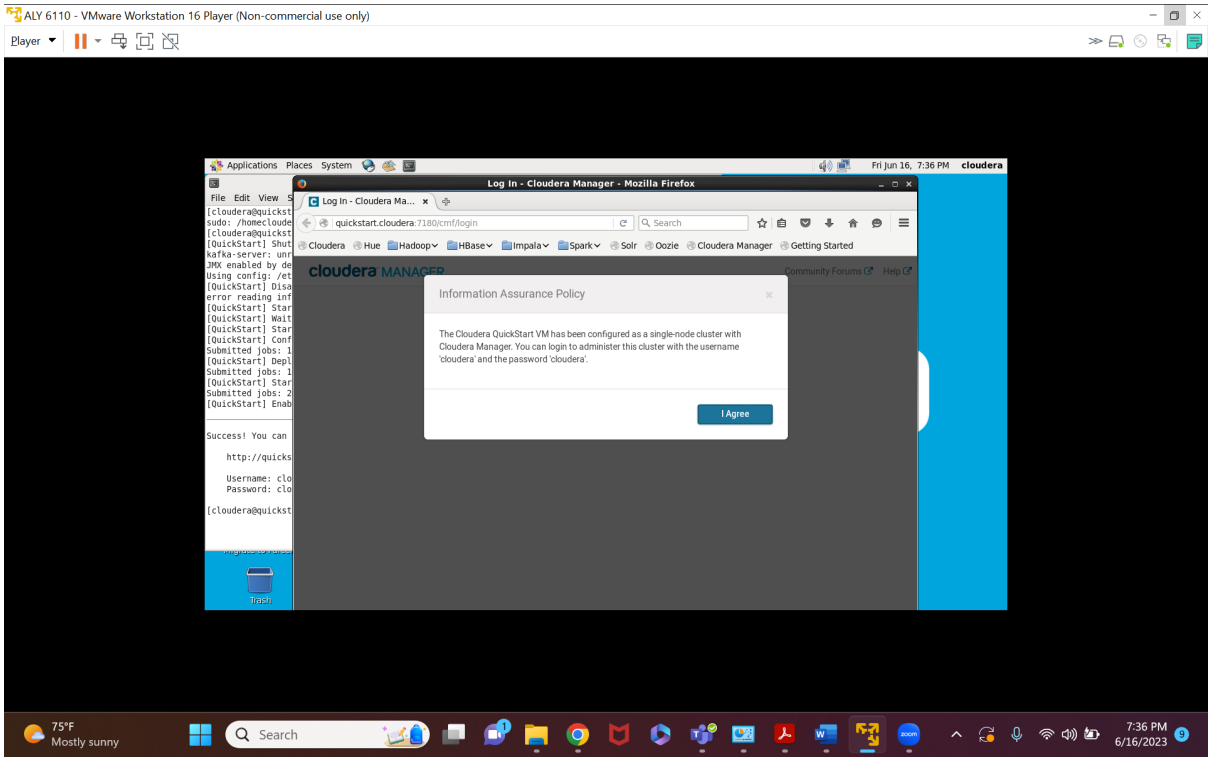
3. **Viewing the 'Cloudera' Homepage in Mozilla Firefox Browser. This interface serves as a gateway to explore and interact with the tools and features available in the 'Cloudera'.**



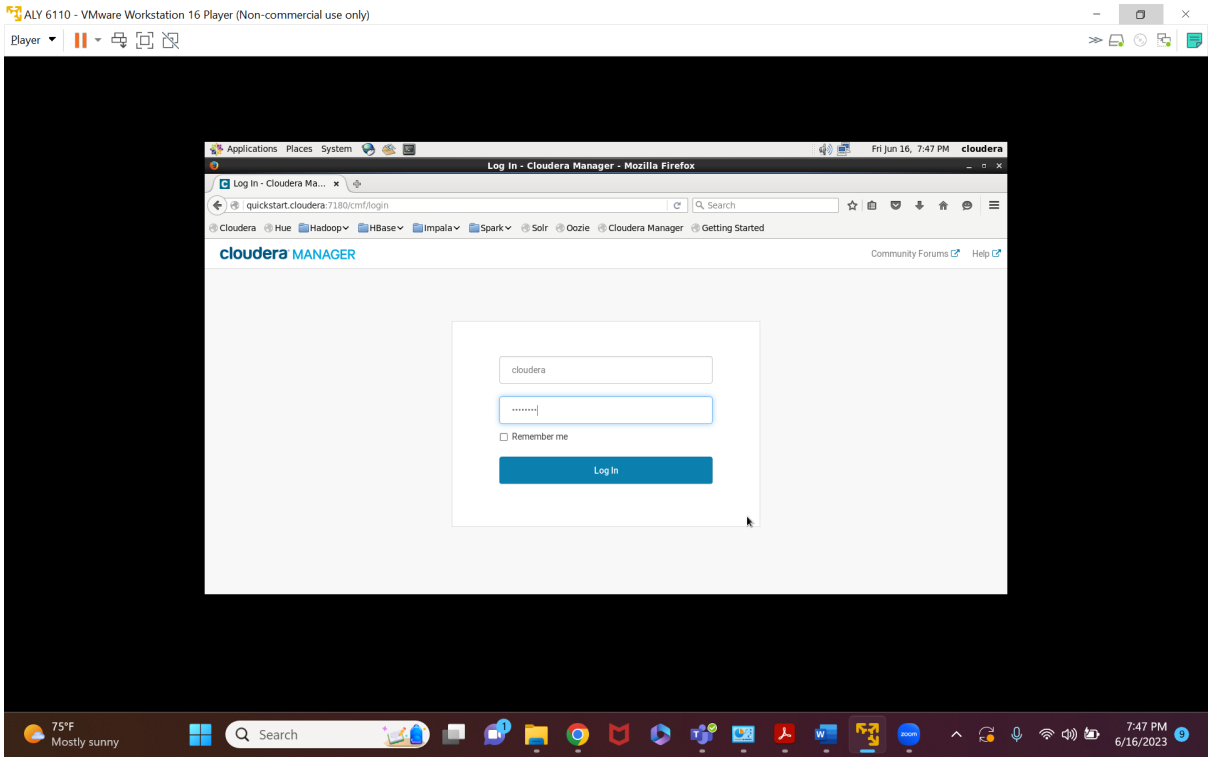
4. Opening Cloudera Manager



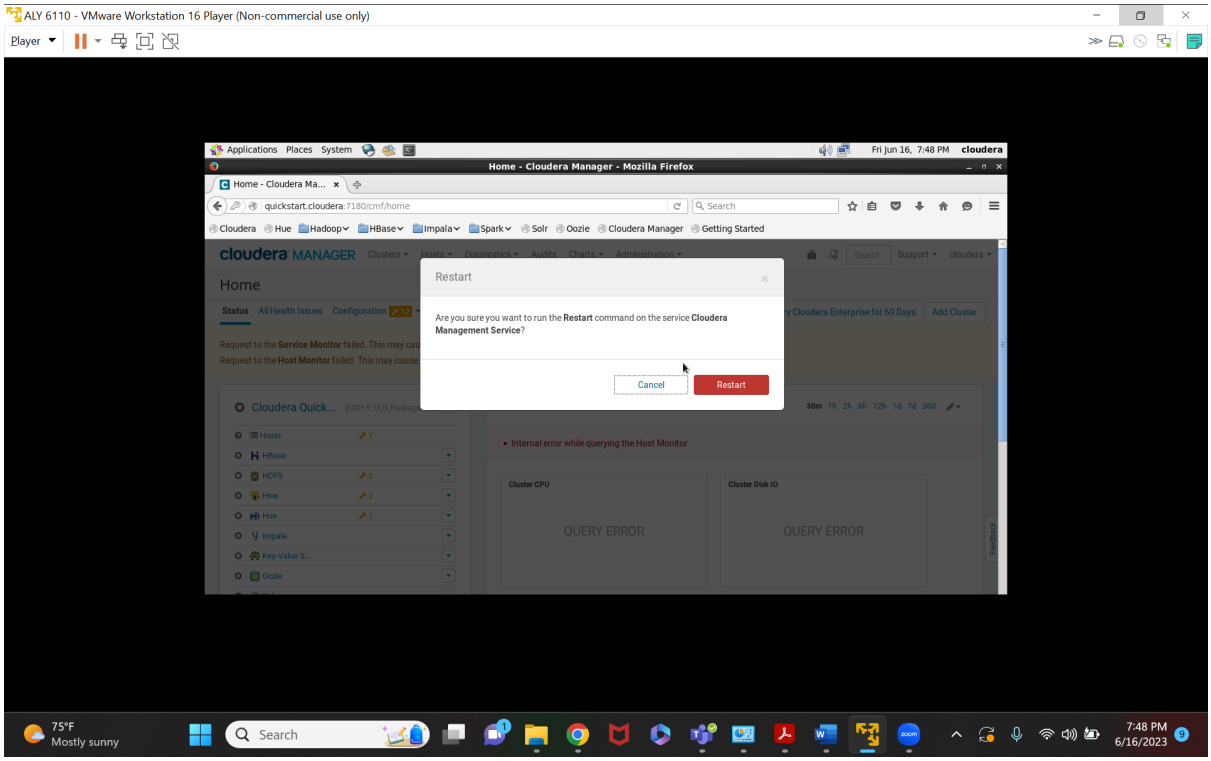


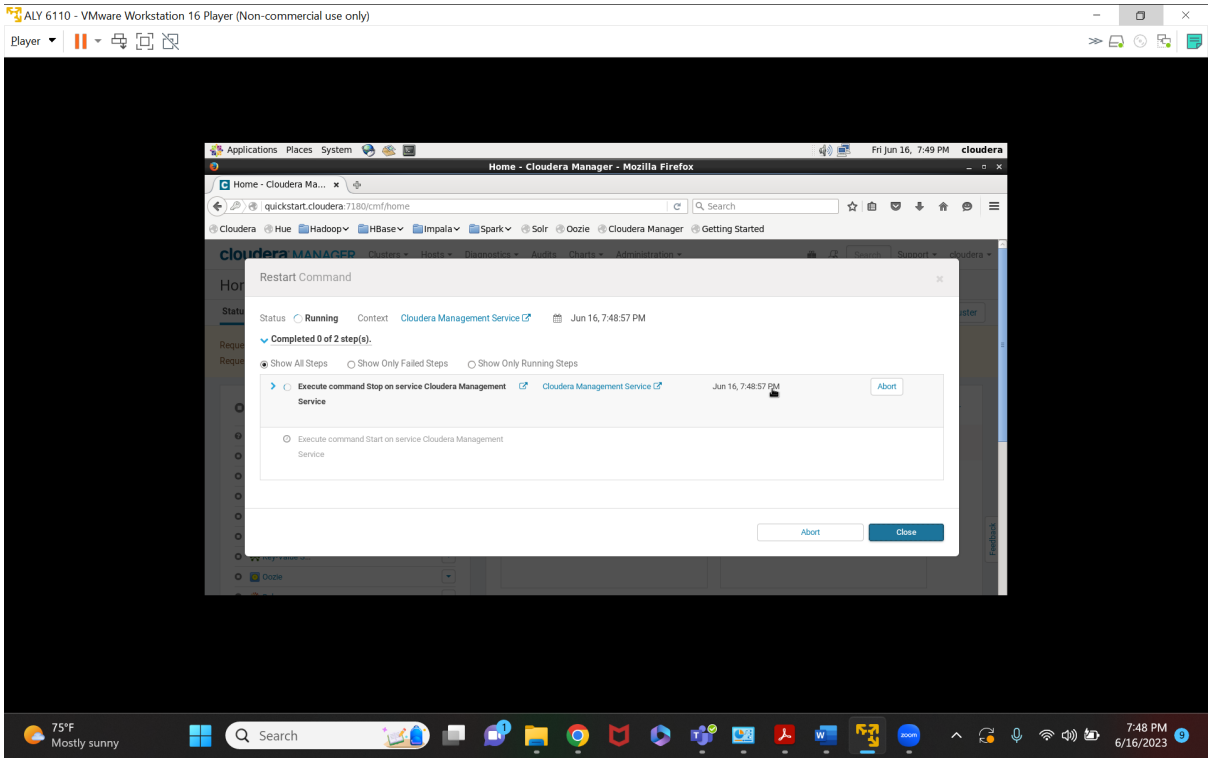


5. Logging in to Cloudera manager using username and password “cloudera”

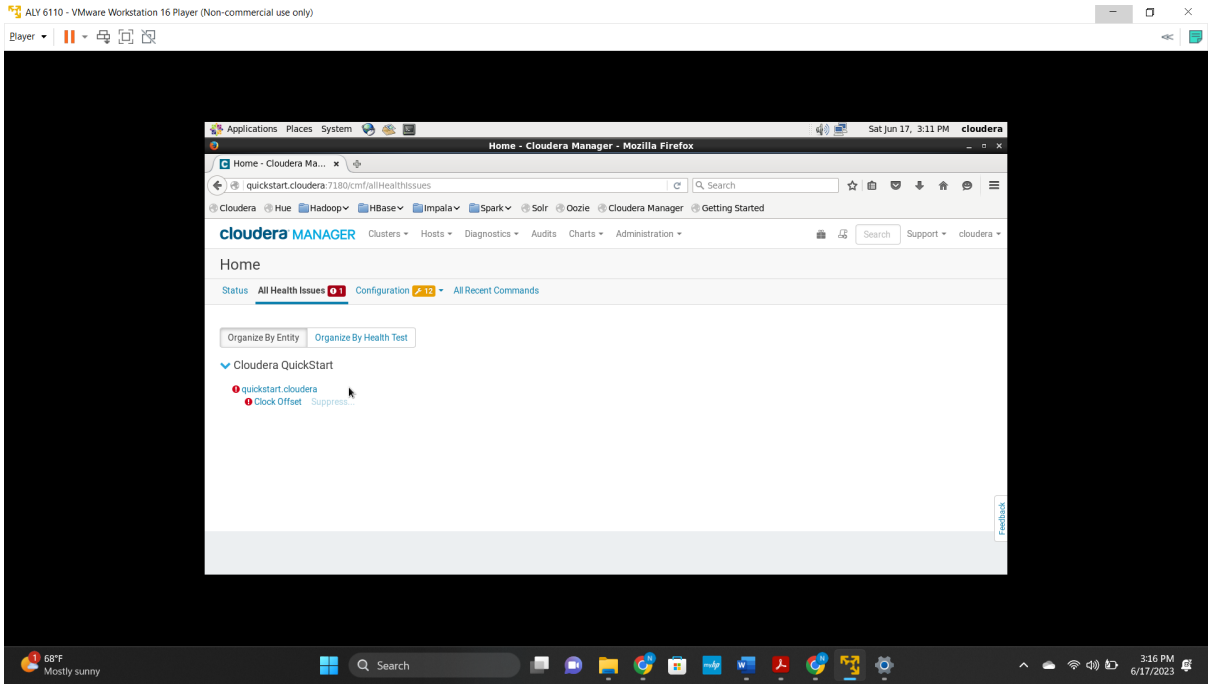


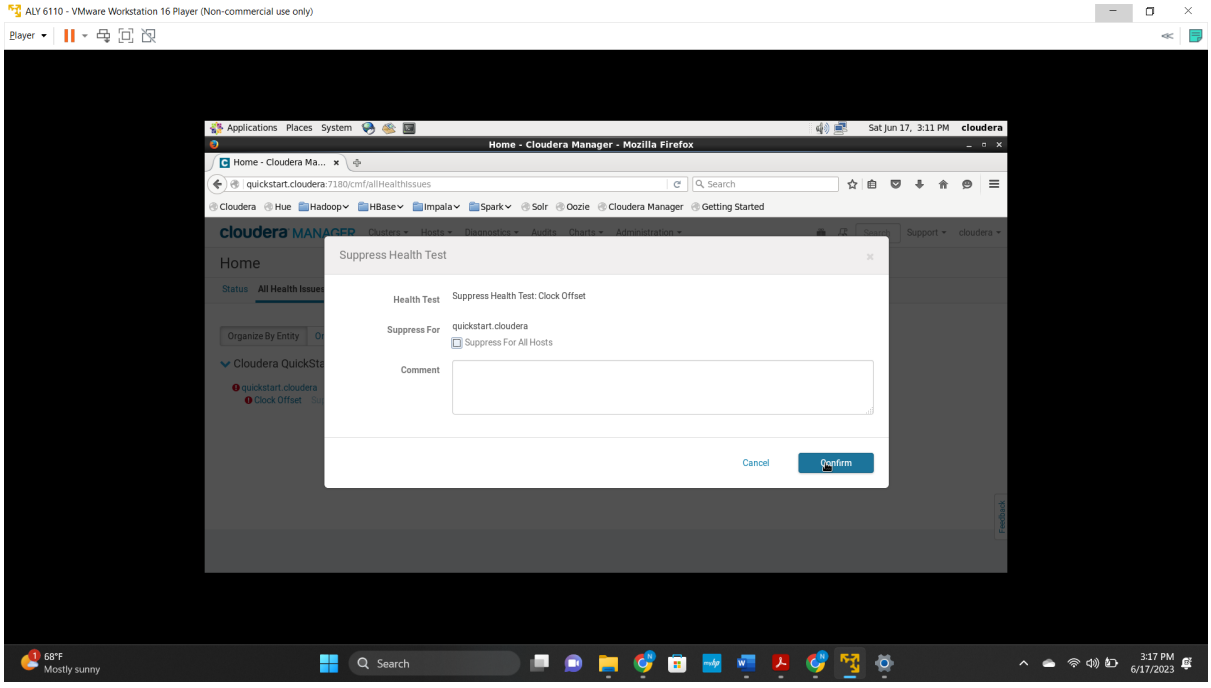
6. **Host monitor was not running so restarted the Cloudera management service. This allowed the host monitor to resume its monitoring tasks and ensure the proper functioning of the Cloudera cluster.**



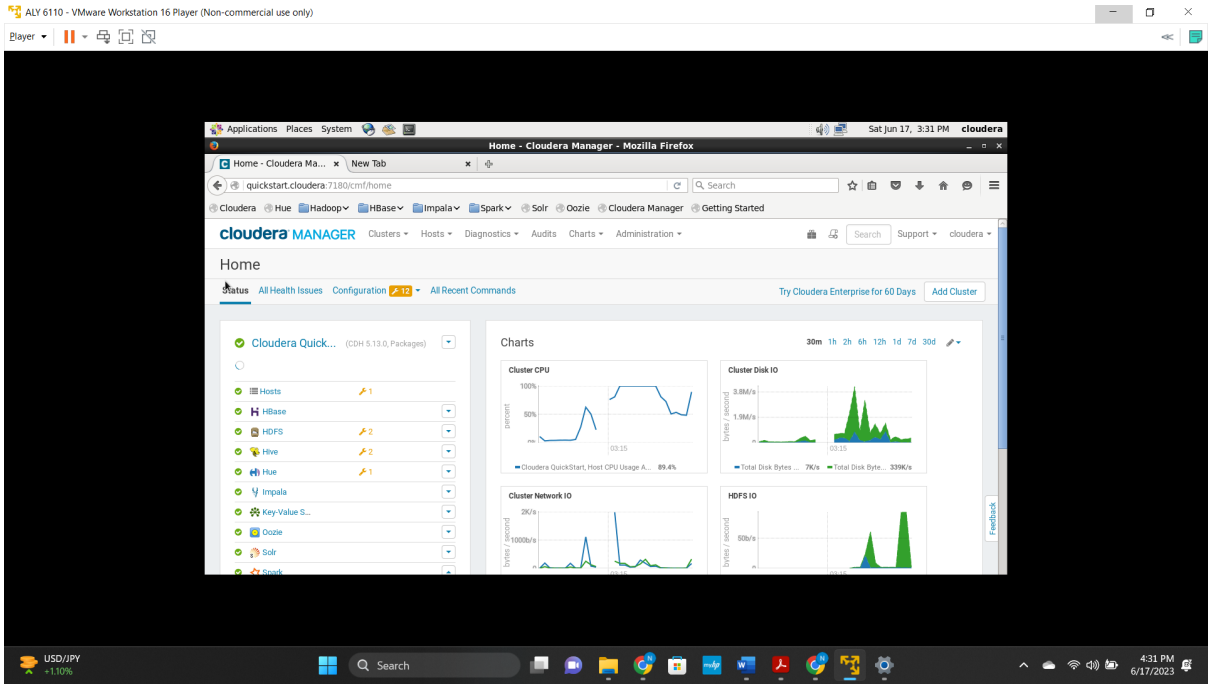


7. Suppressing the health tests

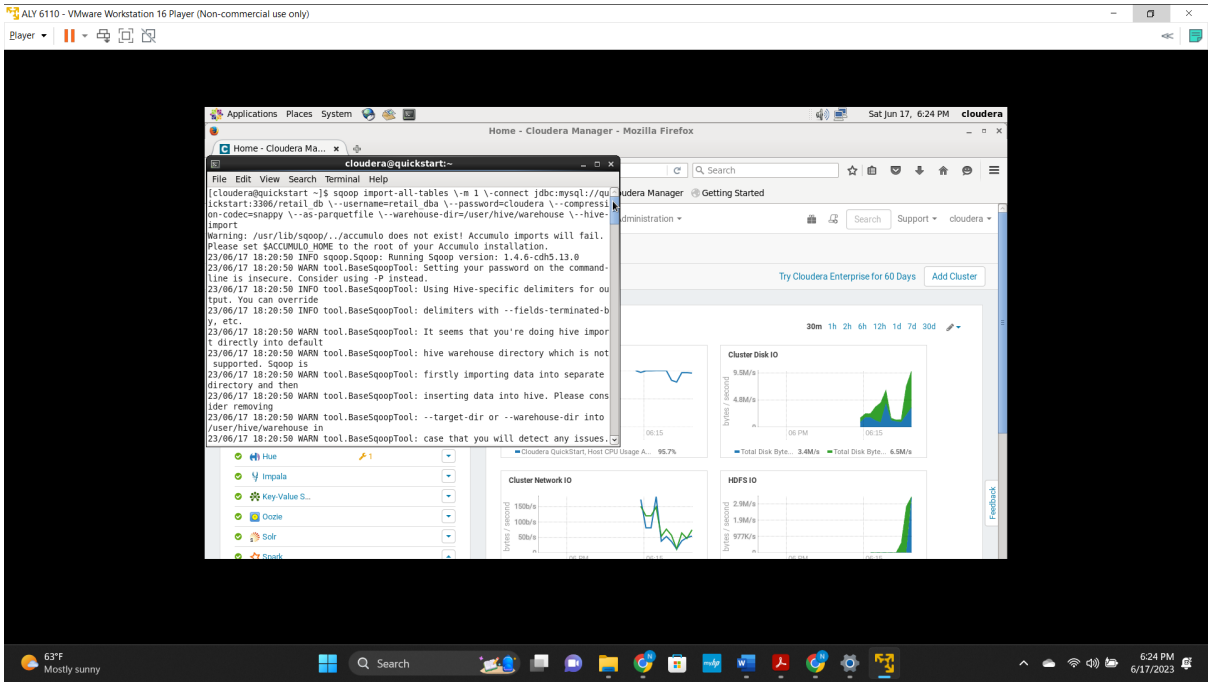


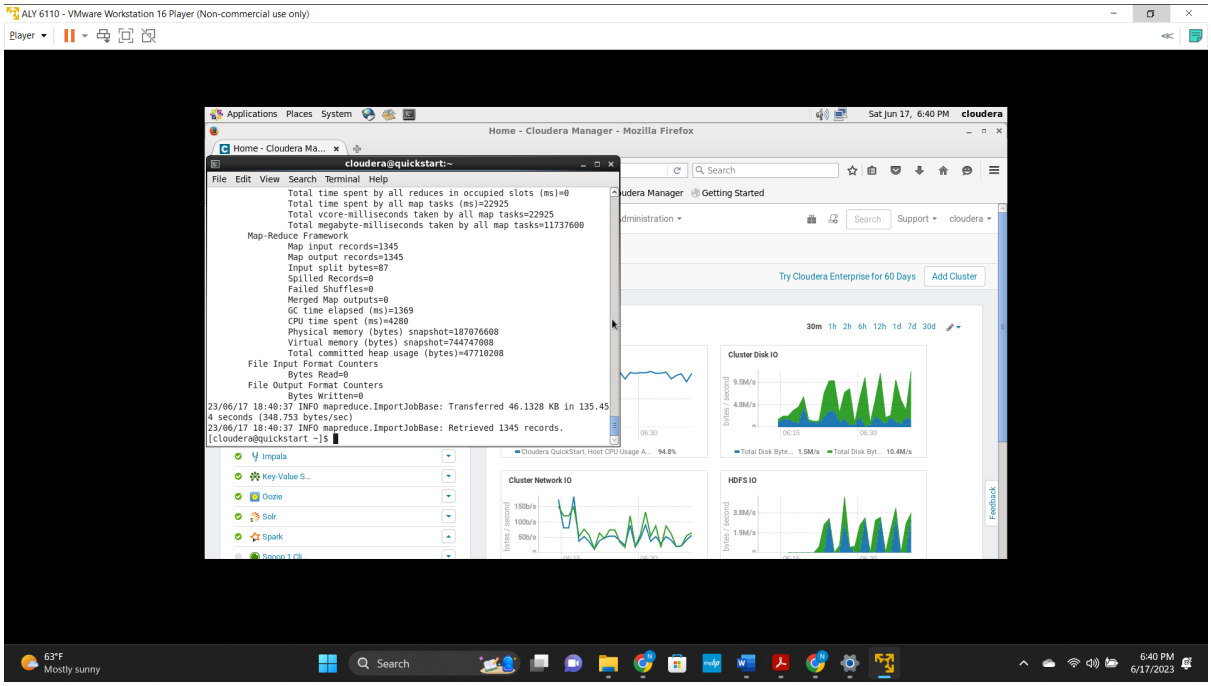


8. Cloudera manager with all services and charts running

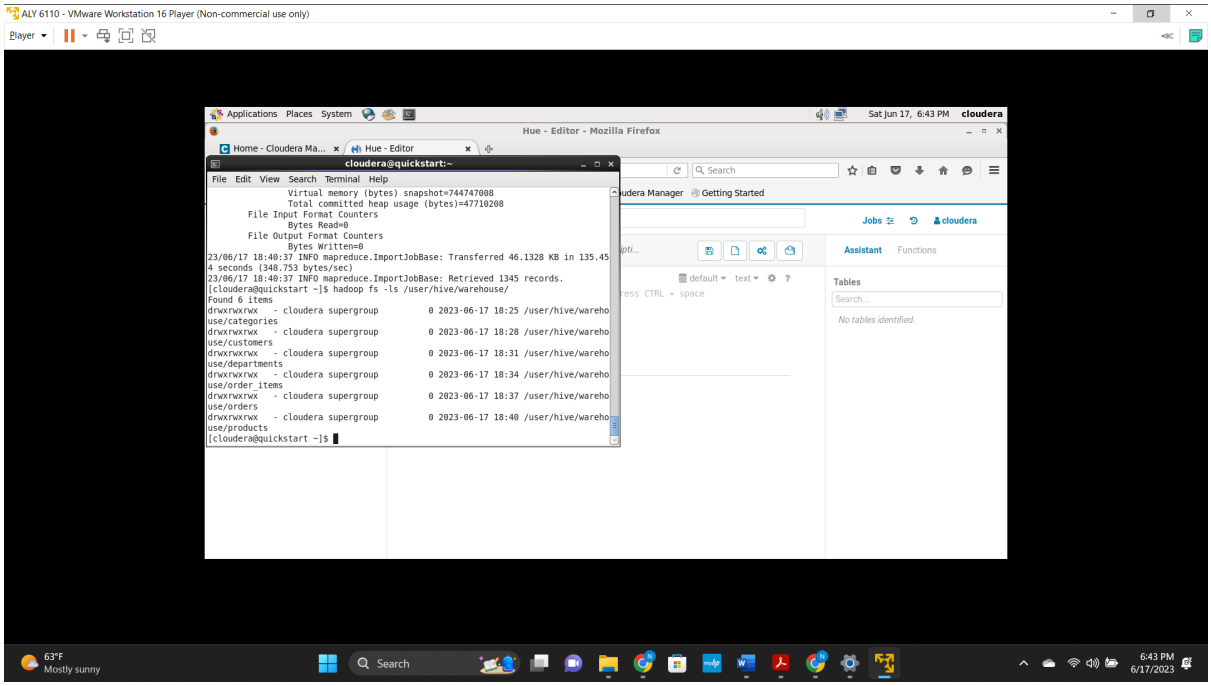


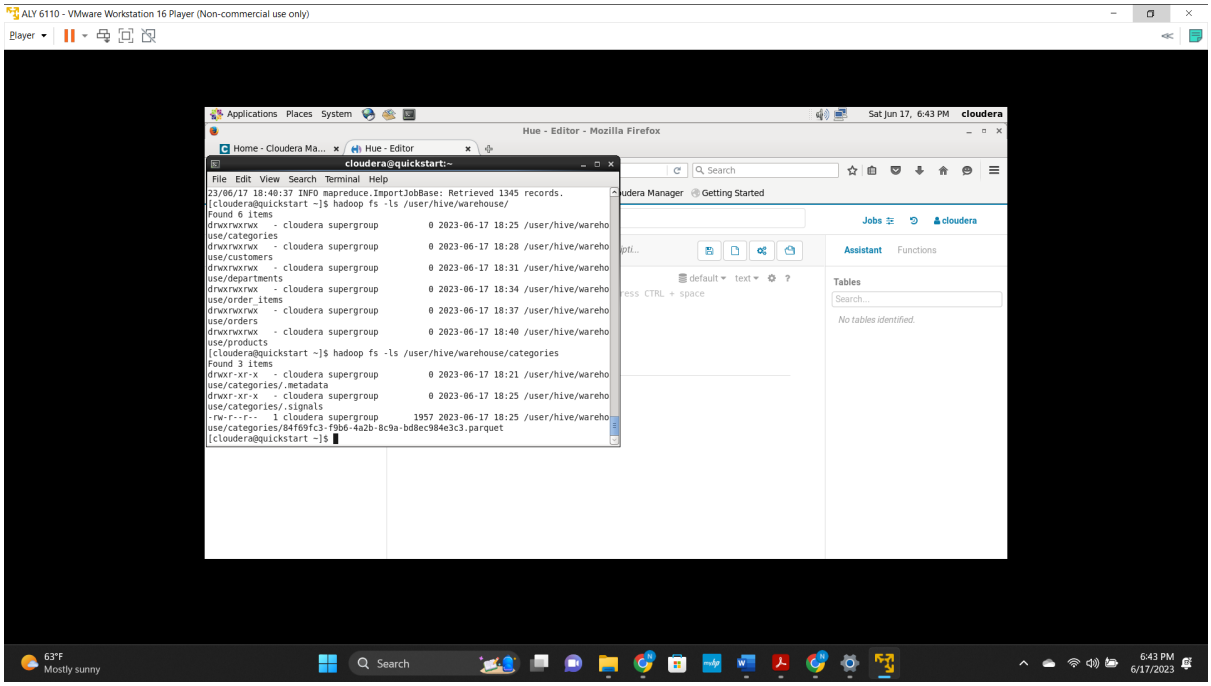
9. In the terminal, we executed the below sqoop command and ingested data from MySQL (RDBMS) to HDFS in Parquet format.



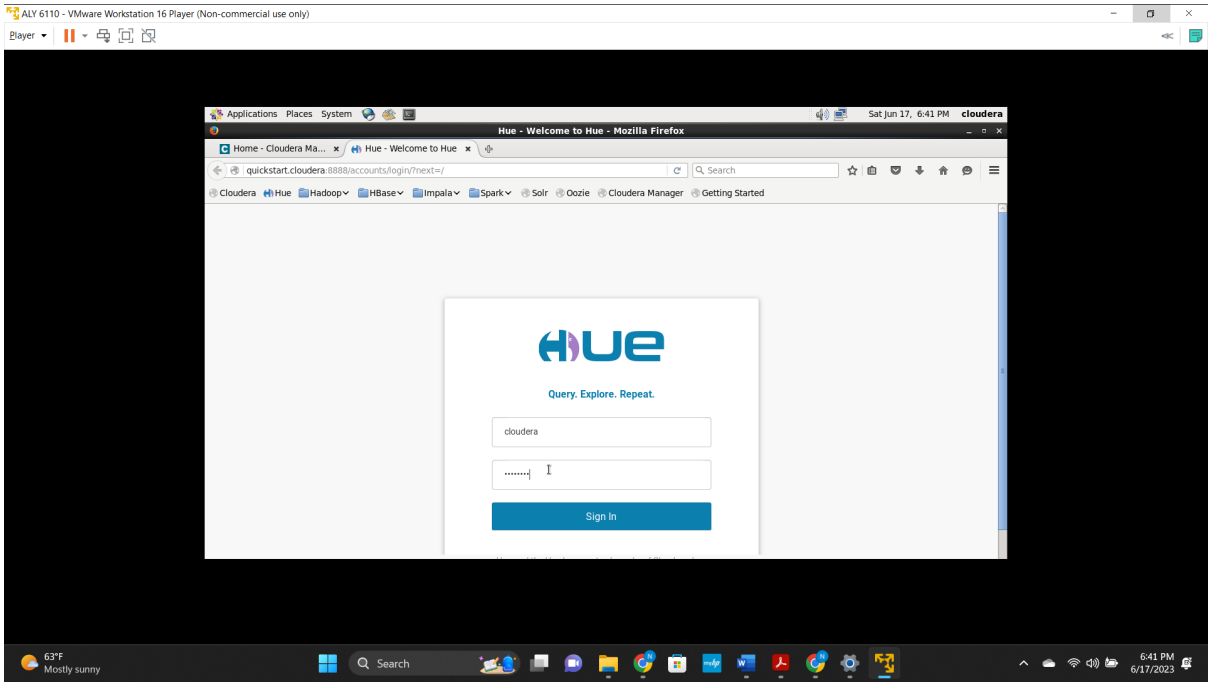


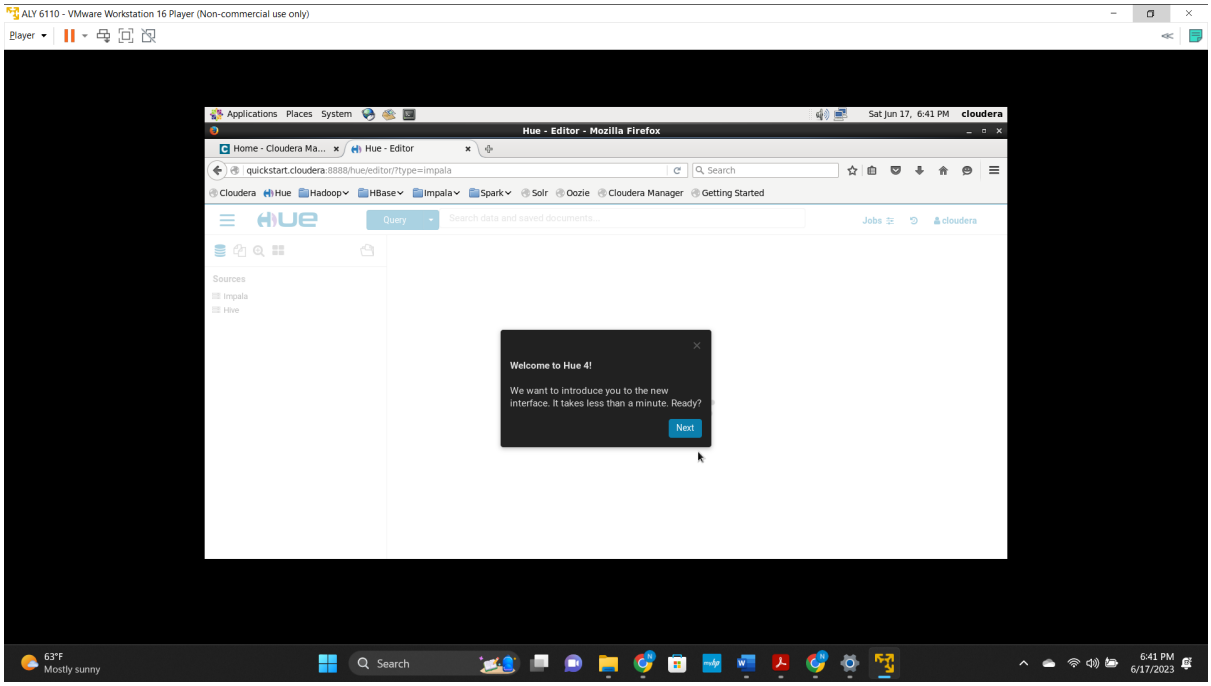
10. Once the sqoop command is executed, in the next step we check that the data files are present in HDFS.



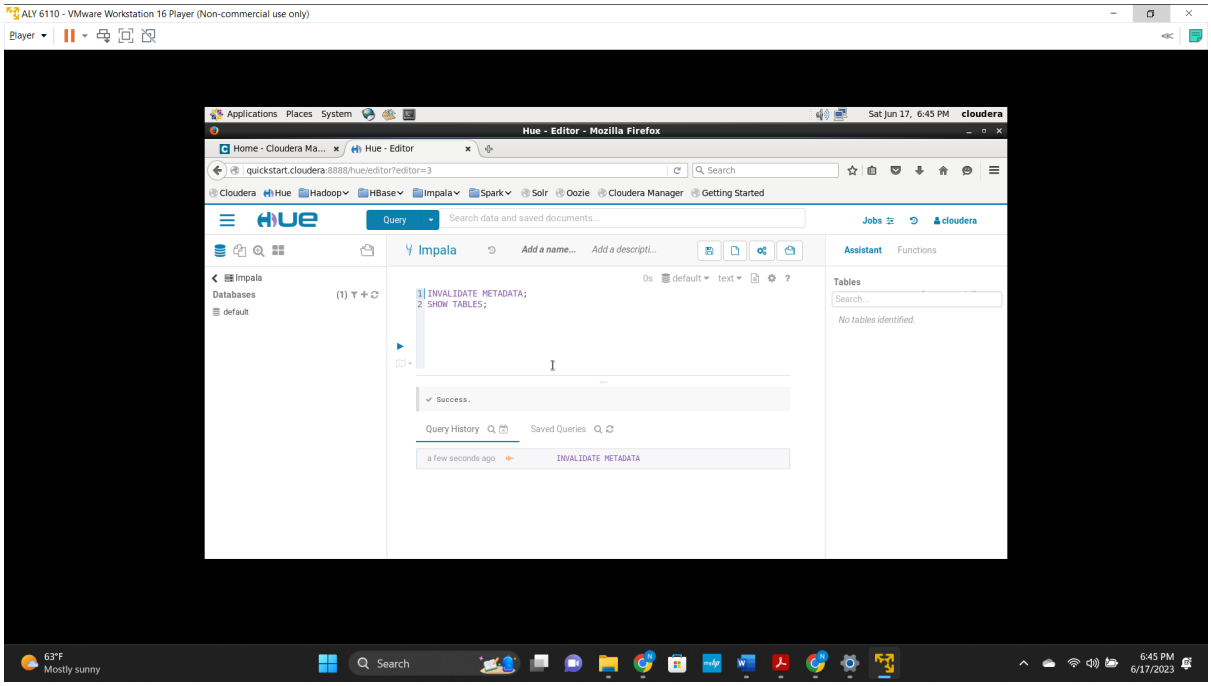


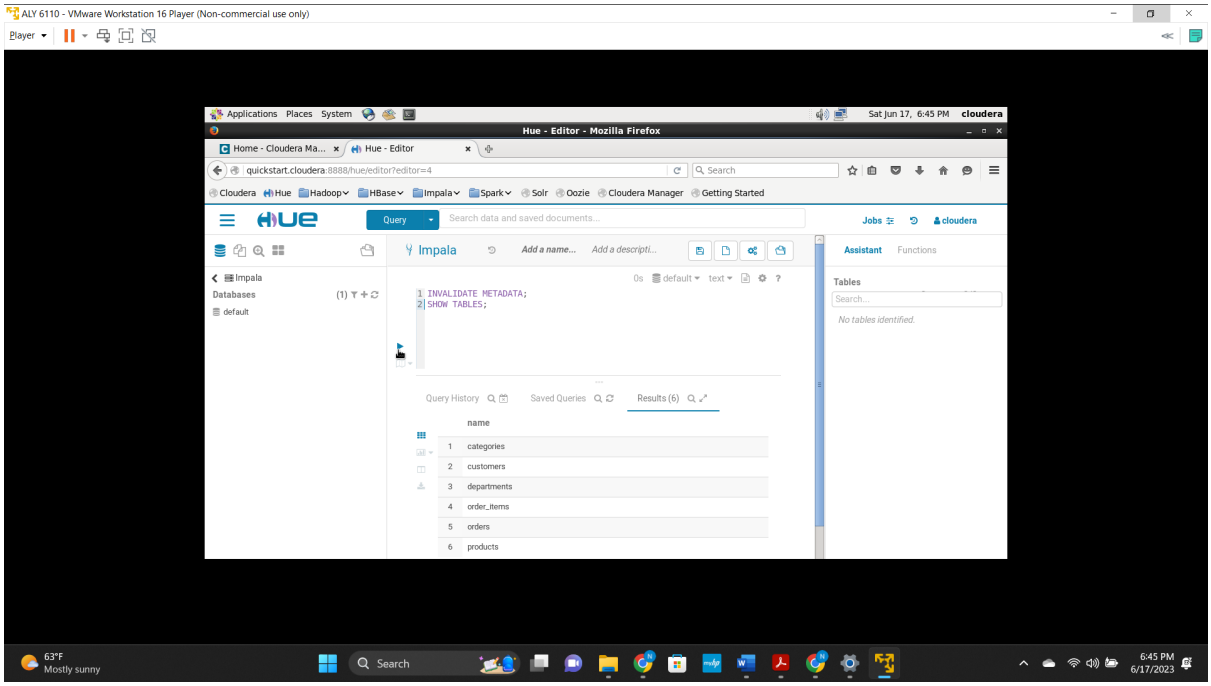
11. Then, we access Hue available on port 8888 of the Manager Node. User name and password are 'cloudera'





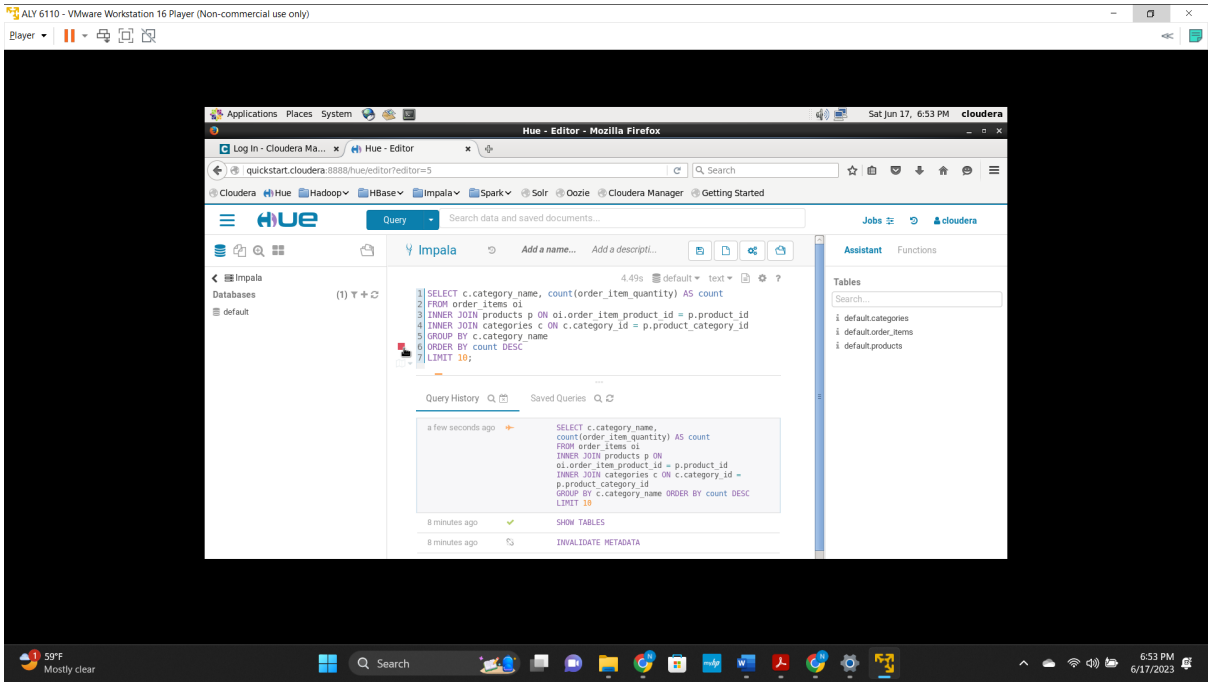
12. Invalidating the metadata using Impala queries and displaying all existing tables.

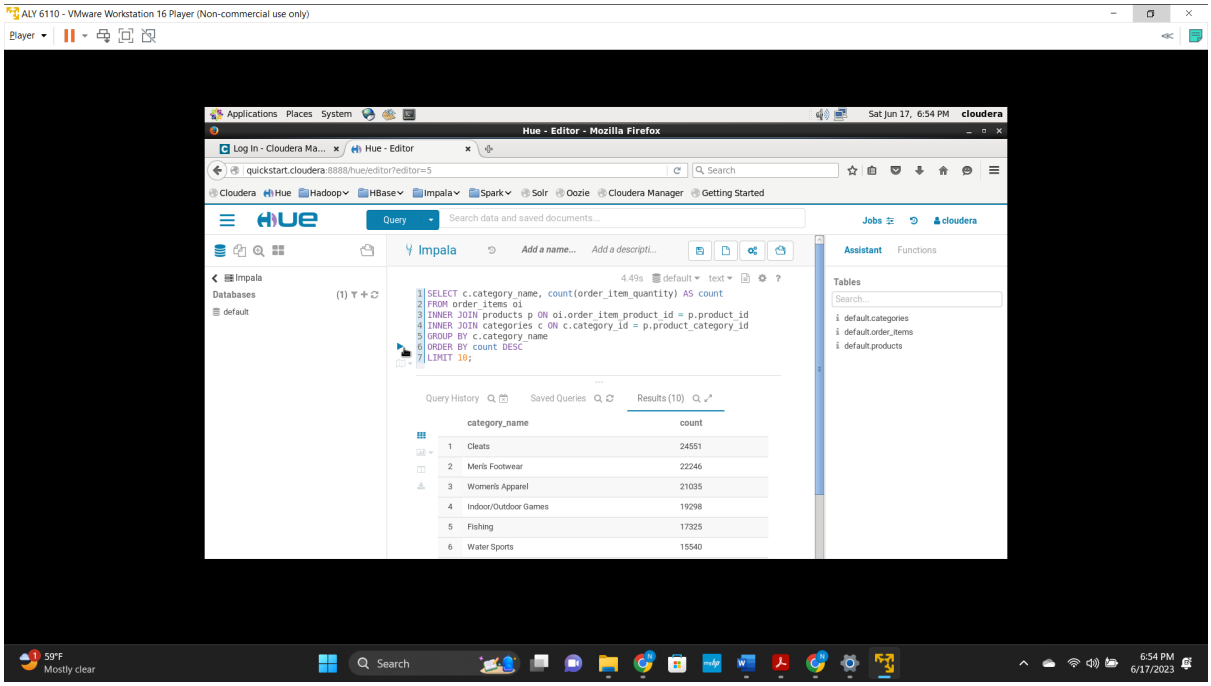




13. Executing the provided SQL example query to calculate total revenue per product and display the top 10 products generating the highest revenue.

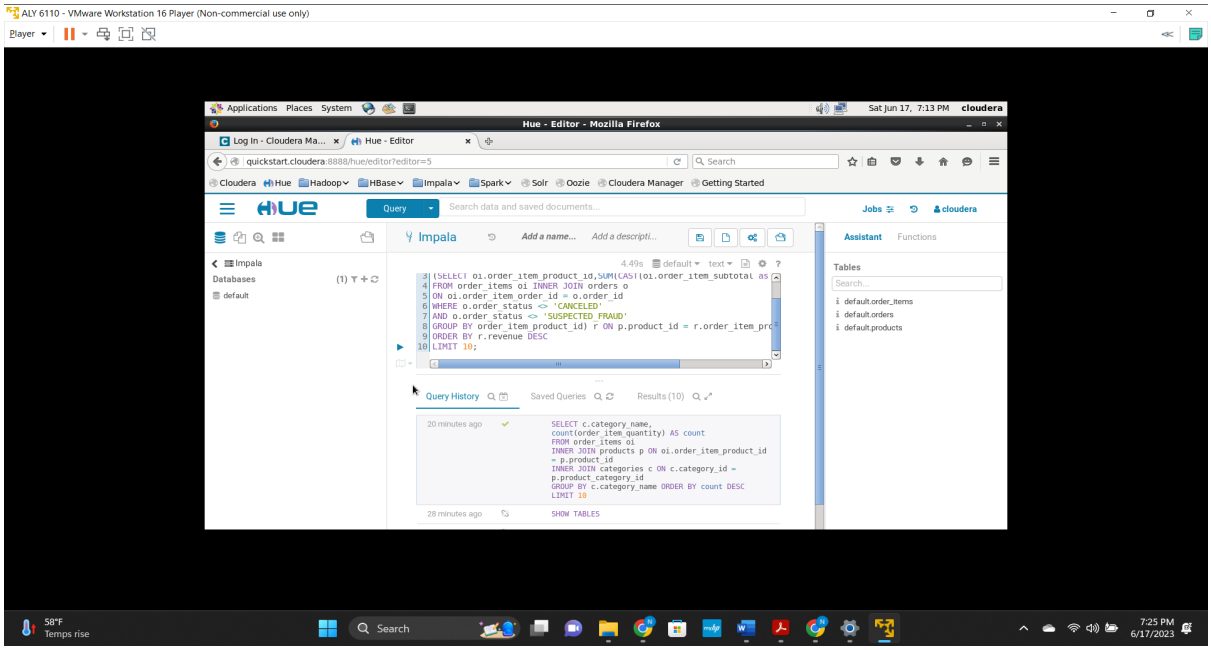
```
SELECT c.category_name, COUNT(order_item_quantity) AS count
FROM order_items oi
INNER JOIN products p ON oi.order_item_product_id = p.product_id
INNER JOIN  categories c ON c.category_id = p.product_ category_id
GROUP BY c.category_name
ORDER BY count DESC LIMIT 10;
```

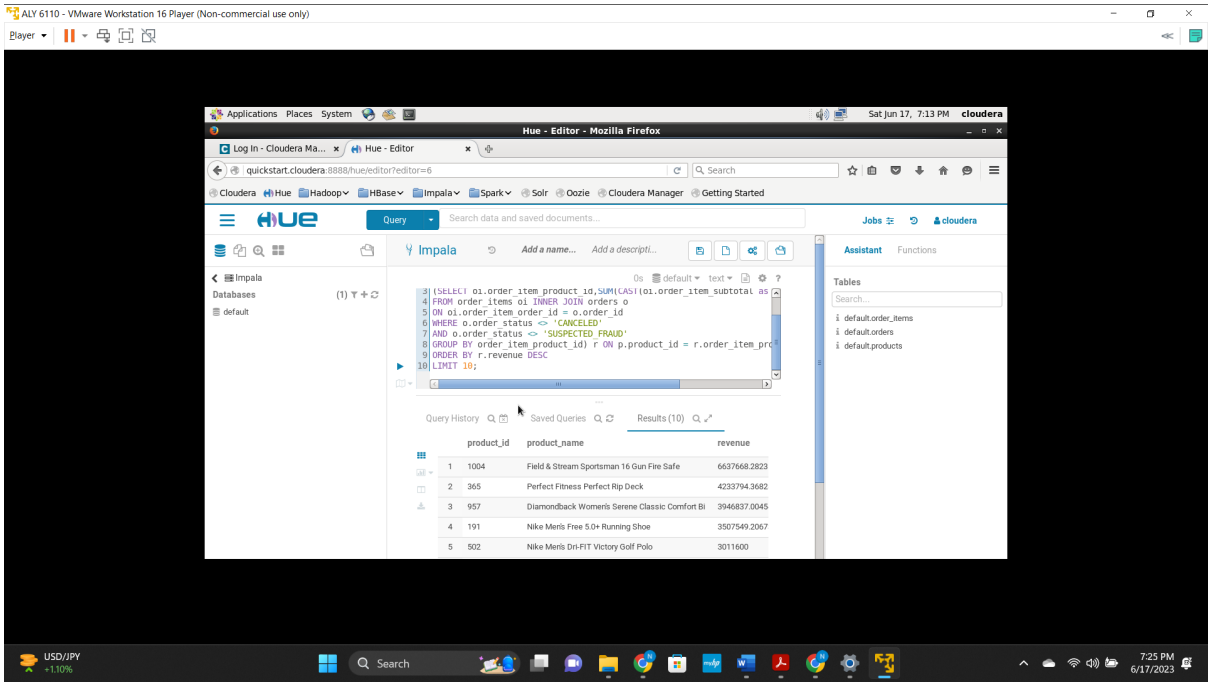




14. Query to calculate the aggregate revenue for products and displaying the top 10 products with the highest revenue.

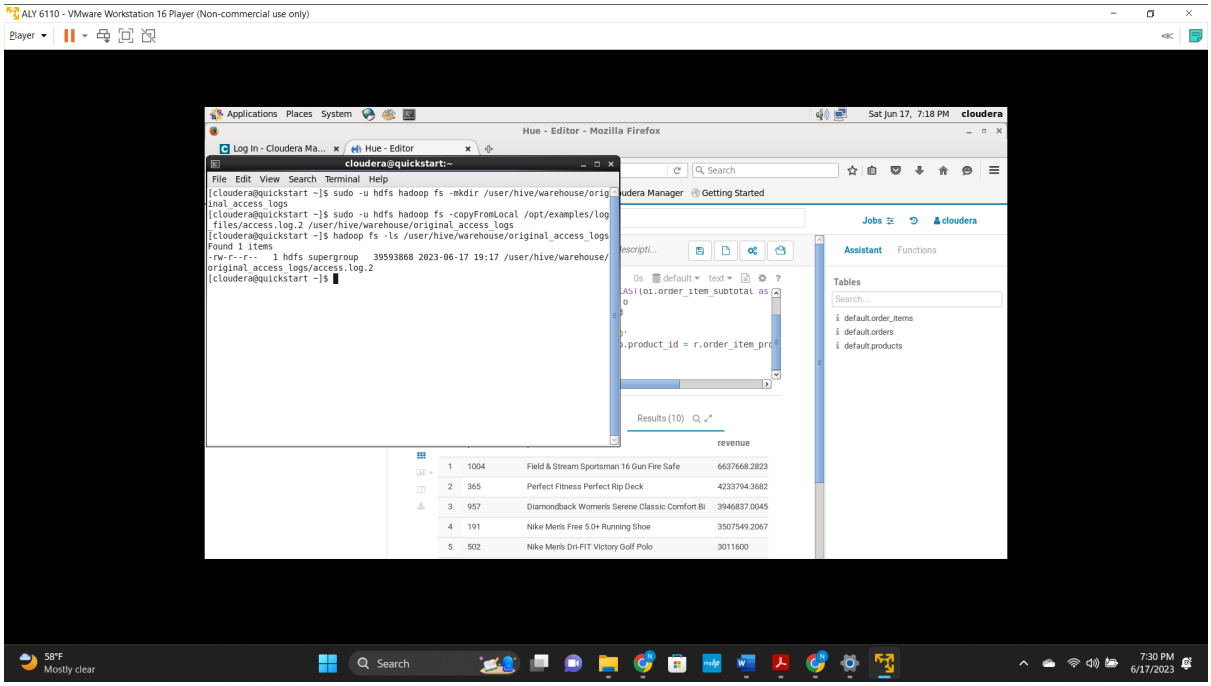
```
SELECT p.product_id, p.product_name, r.revenue
FROM products p
INNER JOIN (SELECT oi.order_item_product_id,
SUM(CAST(oi.order_item_subtotal AS float)) AS revenue FROM order_items oi
INNER JOIN orders o ON oi.order_item_order_id = o.order_id WHERE
o.order_status <> 'CANCELED' AND o.order_status <> '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;
```


ALY 6110 - VMware Workstation 16 Player (Non-commercial use only)
Player
Applications Places System
Sat Jun 17, 7:13 PM cloudera
Hue - Editor - Mozilla Firefox
Log In - Cloudera Ma...
Hue - Editor
quickstart.cloudera:8888/hue/editor?editor=5
Cloudera Hue Hadoop HBase Impala Spark Solr Oozie Cloudera Manager Getting Started
Query
Impala
Assistant Functions
Tables
default.order_items
default.orders
default.products
Query History
Saved Queries
Results (10)
SELECT c.category_name,
count(order_item_quantity) AS count
FROM order_items oi
INNER JOIN products p ON oi.order_item_product_id
= p.product_id
INNER JOIN categories c ON c.category_id =
p.product_category_id
GROUP BY c.category_name ORDER BY count DESC
LIMIT 10
SHOW TABLES
58°F
Temps rise
Search
7:25 PM
6/17/2023

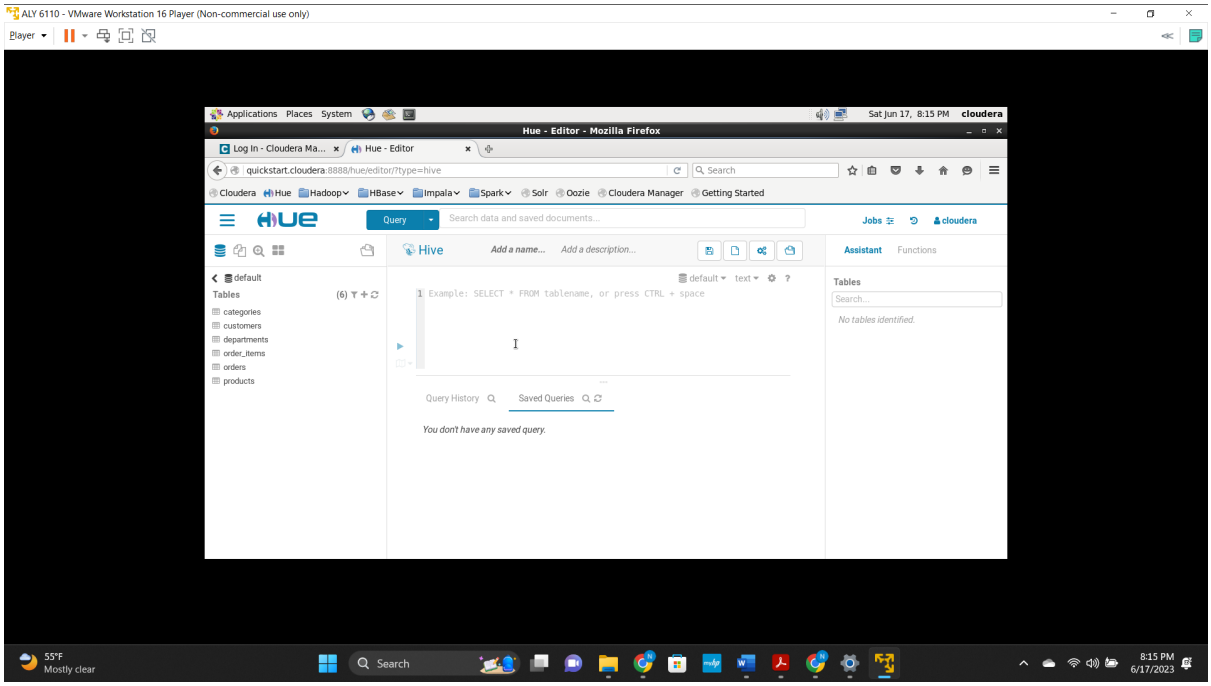

ALY 6110 - VMware Workstation 16 Player (Non-commercial use only)
Player
Applications Places System
Sat Jun 17, 7:13 PM cloudera
Hue - Editor - Mozilla Firefox
Log In - Cloudera Ma...
Hue - Editor
quickstart.cloudera:8888/hue/editor?editor=6
Cloudera Hue Hadoop HBase Impala Spark Solr Oozie Cloudera Manager Getting Started
Query
Impala
Assistant Functions
Tables
default.order_items
default.orders
default.products
Query History
Saved Queries
Results (10)
product_id product_name revenue
1 1004 Field & Stream Sportsman 16 Gun Fire Safe 6637668.2823
2 365 Perfect Fitness Perfect Rip Deck 4233794.3682
3 957 Diamondback Women's Serene Classic Comfort Bi 3946837.0045
4 191 Nike Men's Free 5.0+ Running Shoe 3507549.2067
5 502 Nike Men's Dri-FIT Victory Golf Polo 3011600
USD/JPY
+1.10%
Search
7:25 PM
6/17/2023

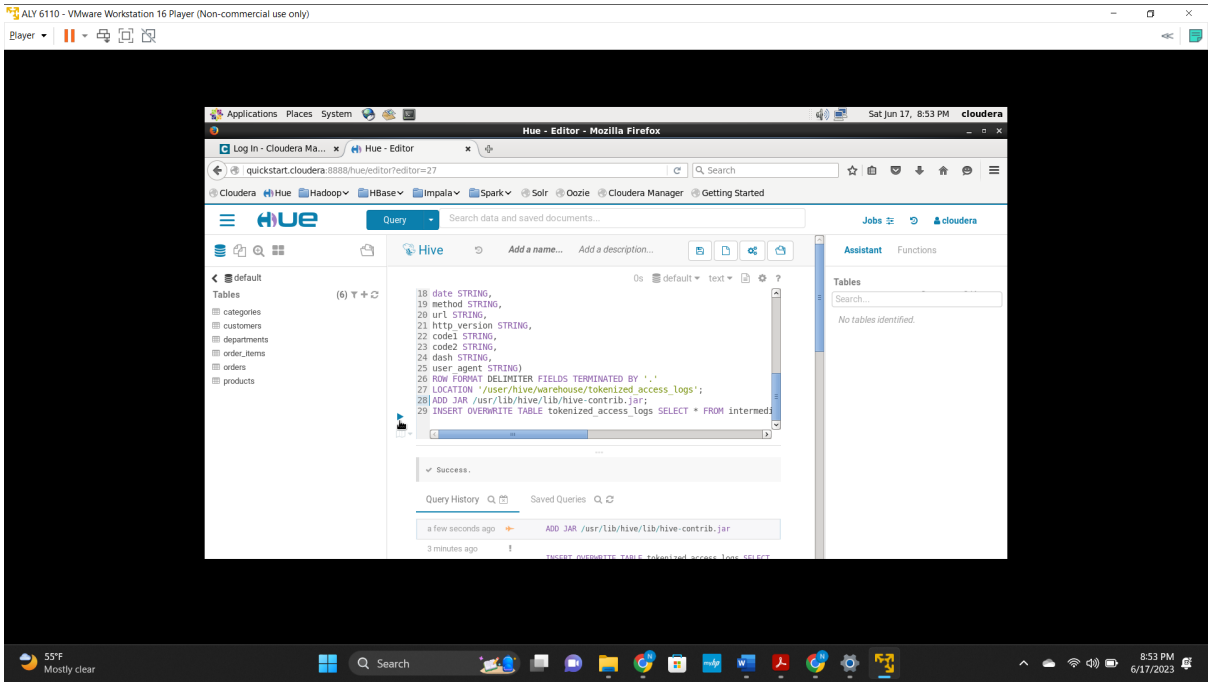
Exercise-2 Correlate structured data with unstructured data

15. Query to transfer the data from the local filesystem to HDFS and verifying its presence in HDFS

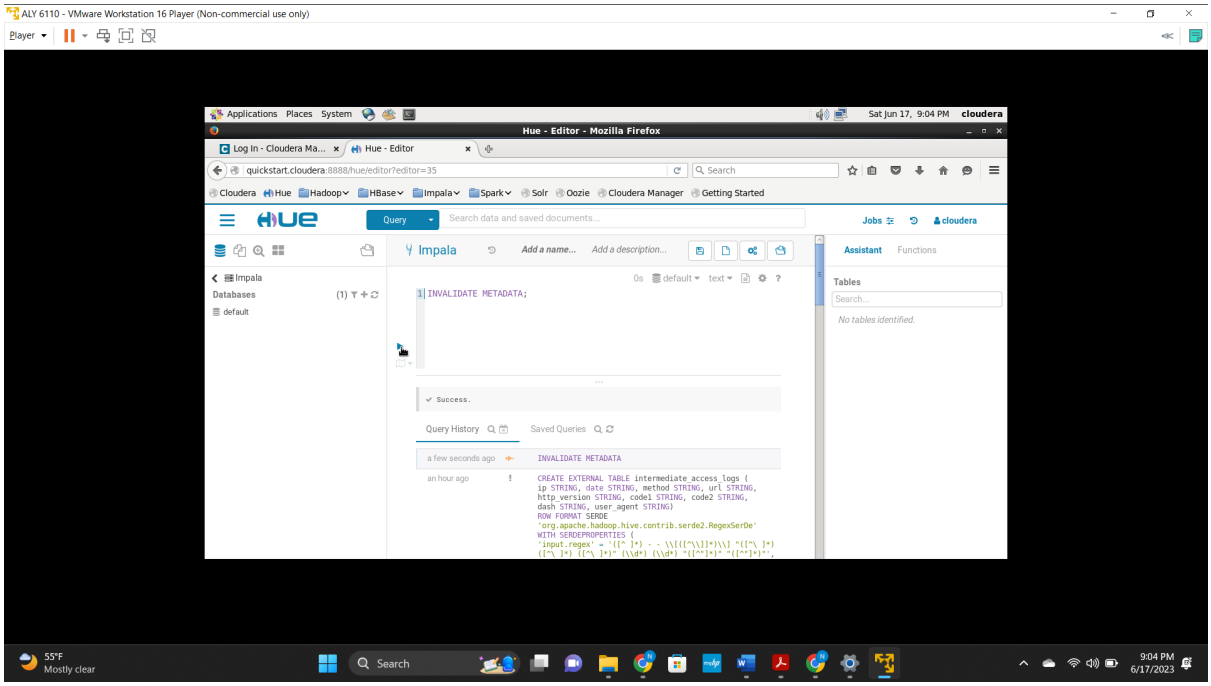


16. Query to create two tables in Hive using SerDes to parse the logs. For this, we change to Hive Query Editor first.

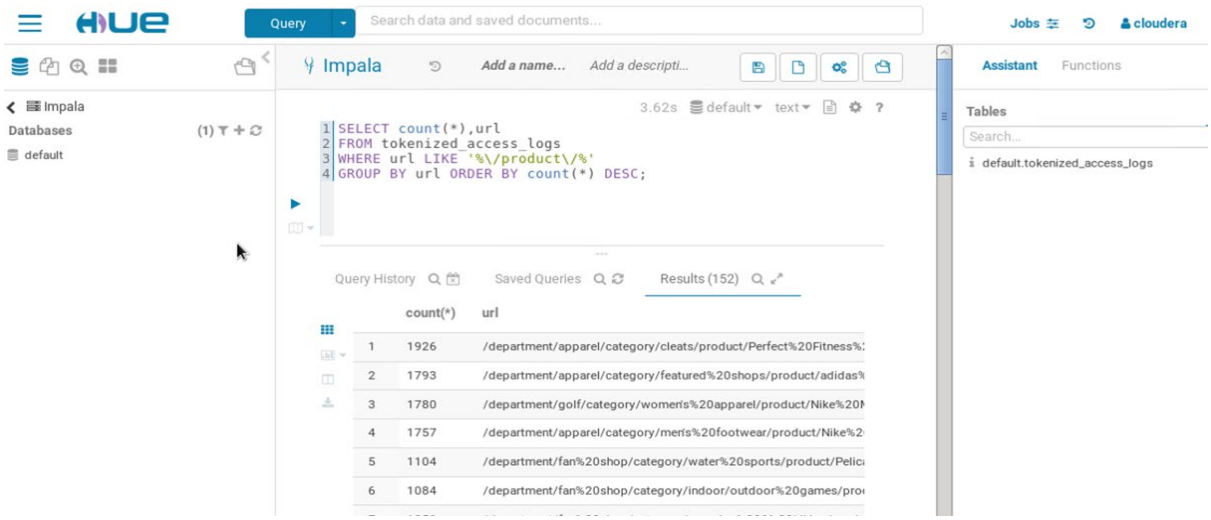




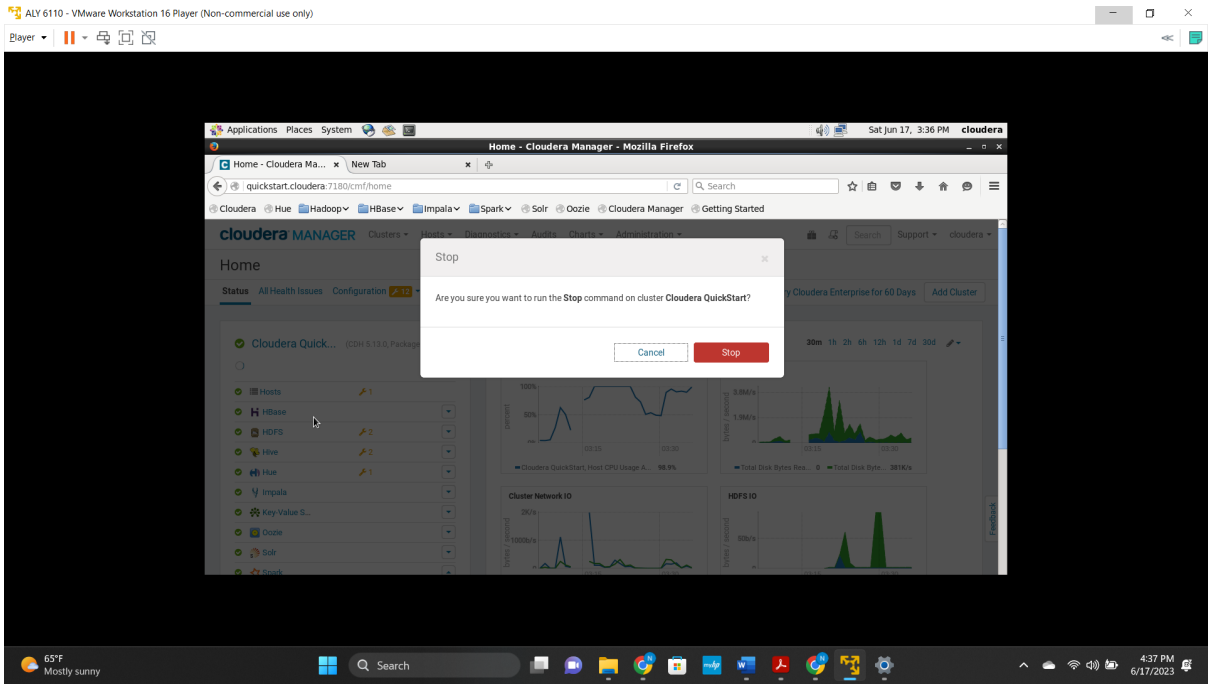
17. Invalidating the metadata after returning to the Impala editor.

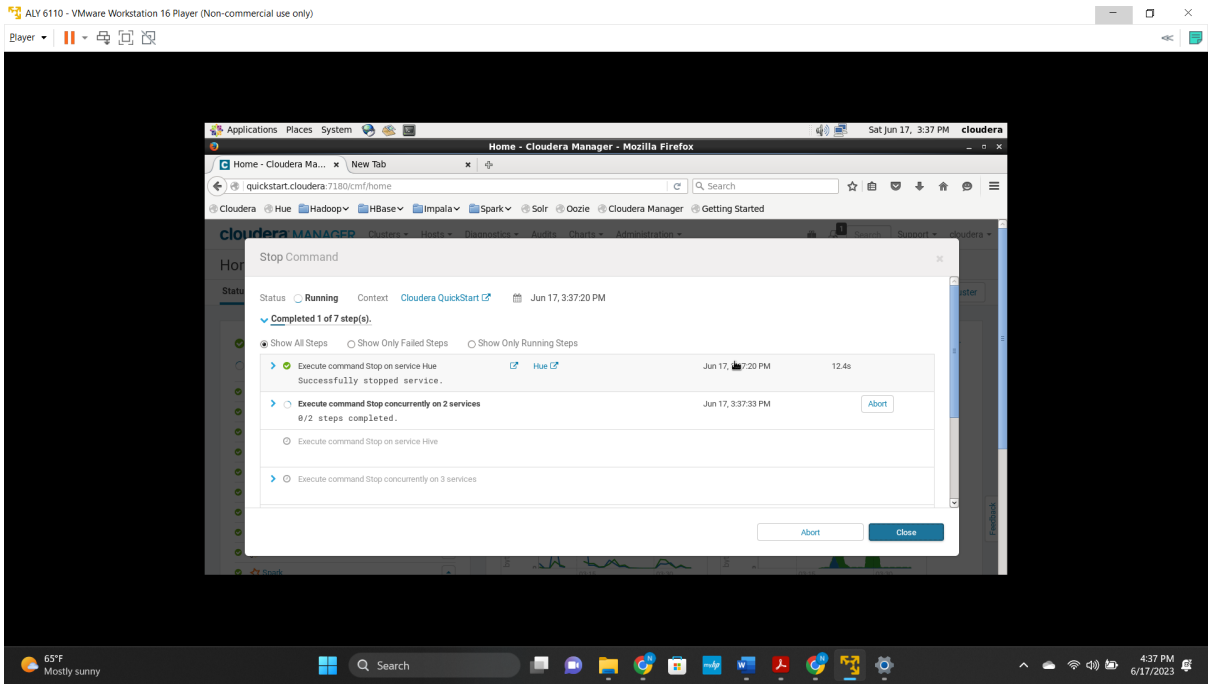


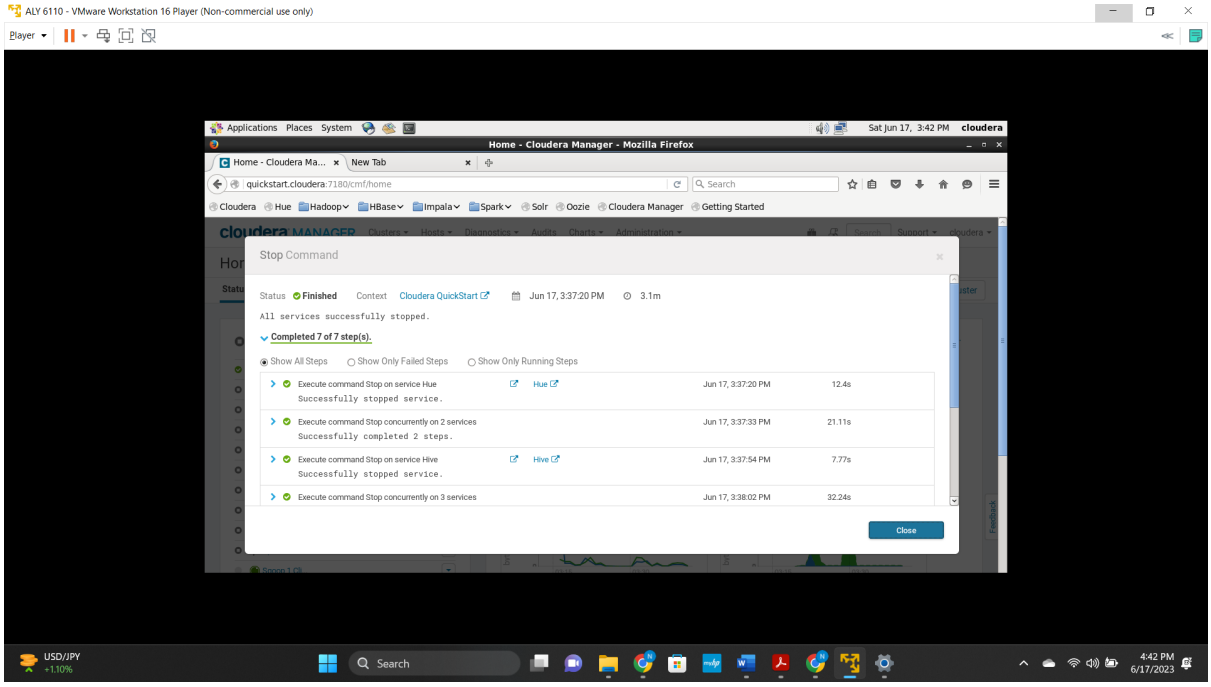
18. Query to determine the website that has been visited the most users



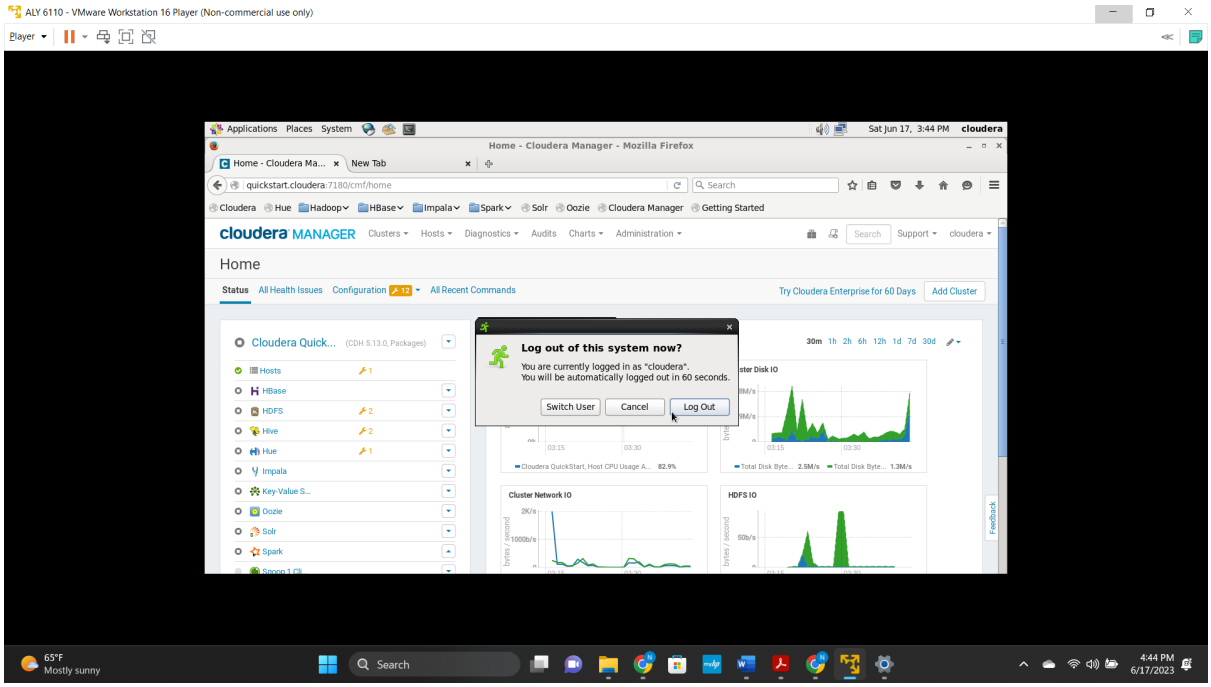
19. Stopping all services in Cloudera manager by clicking on stop button in the dropdown

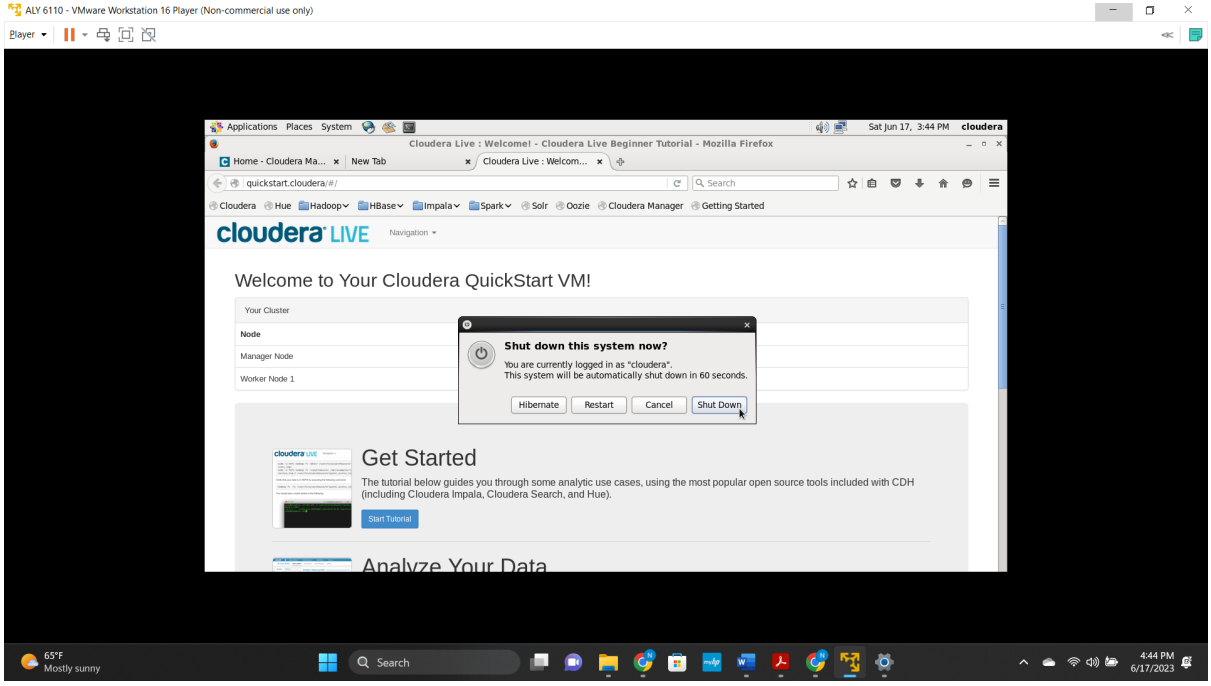



ALY 6110 - VMware Workstation 16 Player (Non-commercial use only)
Player
Applications Places System
Sat Jun 17, 3:37 PM cloudera
Home - Cloudera Manager - Mozilla Firefox
Home - Cloudera Ma...
New Tab
quickstart.cloudera:7180/cmf/home
Search
Cloudera Hue Hadoop HBase Impala Spark Solr Oozie Cloudera Manager Getting Started
Stop Command
Status Running Context Cloudera QuickStart Jun 17, 3:37:20 PM
Completed 1 of 7 step(s).
Show All Steps Show Only Failed Steps Show Only Running Steps
Execute command Stop on service Hue
Successfully stopped service.
Hue
Jun 17, 3:37:20 PM
12.4s
Execute command Stop concurrently on 2 services
0/2 steps completed.
Jun 17, 3:37:33 PM
Abort
Execute command Stop on service Hive
Execute command Stop concurrently on 3 services
Abort
Close
65°F
Mostly sunny
Search
4:37 PM
6/17/2023


ALY 6110 - VMware Workstation 16 Player (Non-commercial use only)
Player
Applications Places System
Sat Jun 17, 3:42 PM cloudera
Home - Cloudera Manager - Mozilla Firefox
Home - Cloudera Ma...
New Tab
quickstart.cloudera:7180/cmf/home
Search
Cloudera Hue Hadoop HBase Impala Spark Solr Oozie Cloudera Manager Getting Started
Stop Command
Status Finished Context Cloudera QuickStart Jun 17, 3:37:20 PM 3.1m
All services successfully stopped.
Completed 7 of 7 step(s).
Show All Steps Show Only Failed Steps Show Only Running Steps
Execute command Stop on service Hue
Successfully stopped service.
Hue
Jun 17, 3:37:20 PM
12.4s
Execute command Stop concurrently on 2 services
Successfully completed 2 steps.
Jun 17, 3:37:33 PM
21.11s
Execute command Stop on service Hive
Successfully stopped service.
Hive
Jun 17, 3:37:54 PM
7.77s
Execute command Stop concurrently on 3 services
Jun 17, 3:38:02 PM
32.24s
Close
USD/JPY
+1.10%
Search
4:42 PM
6/17/2023

20. Logging out from Cloudera and shutting down the VM

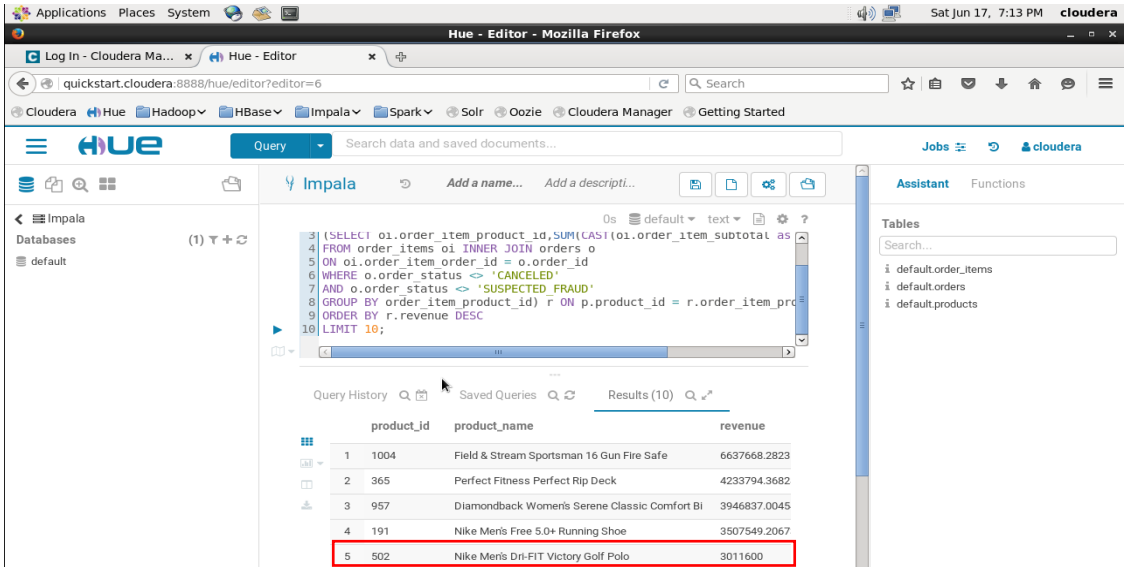




Answers to the questions

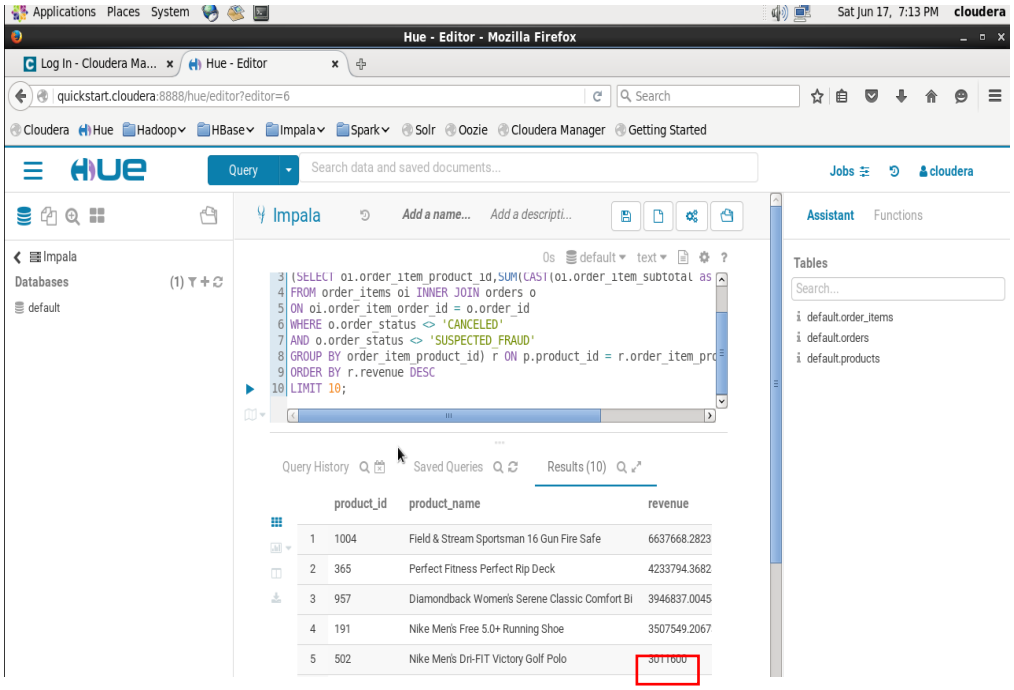
1. What is the 5th most revenue-generating product?

Nike's Men Dri-FIT Victory Golf Polo



2. How much revenue does the Nike men's dri-fit polo earn?

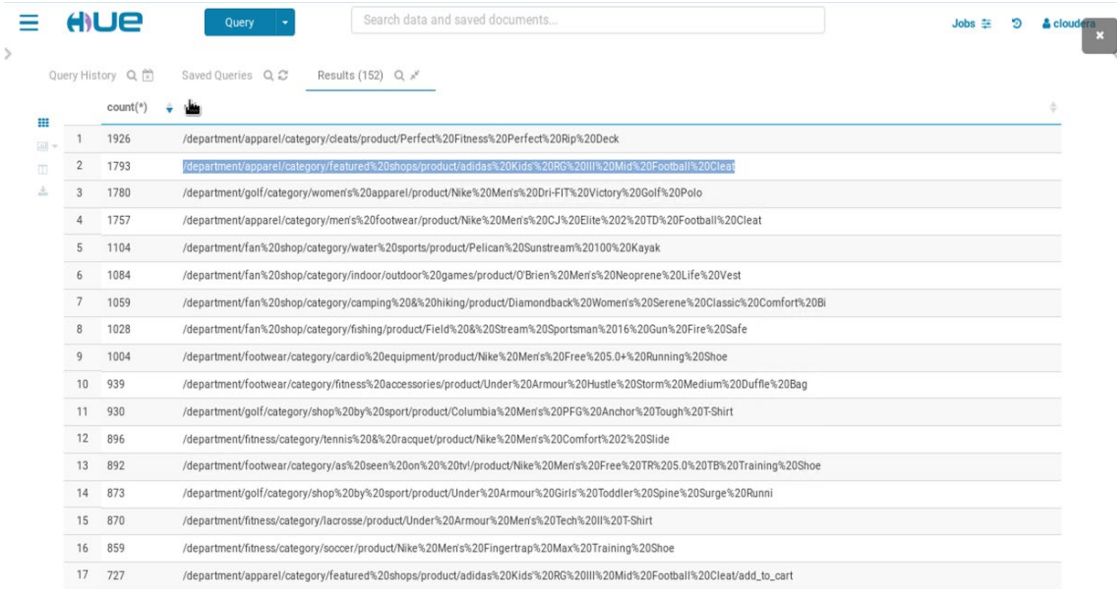
$ 3,011,600



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?**

   The Adidas Kids' RG III Mid Football Cleat appears to be a product that receives a significant amount of views on the website but is rarely purchased. This can be due to many reasons.

   - **Users could be simply browsing the website out of curiosity, exploring various products without immediate purchase intentions.**
   - **The price of the product might be relatively high, making it less affordable for potential buyers.**
   - **This product might experience periodic spikes in views due to seasonal interests or events.**



	count(*)	
1	1926	/department/apparel/category/cleats/product/Perfect%20Fitness%20Perfect%20Rip%20Deck
2	1793	/department/apparel/category/featured%20shops/product/adidas%20Kids'%20RG%20III%20Mid%20Football%20Cleat
3	1780	/department/golf/category/women's%20apparel/product/Nike%20Men's%20Dri-FIT%20Victory%20Golf%20Polo
4	1757	/department/apparel/category/men's%20footwear/product/Nike%20Men's%20CJ%20Elite%202%20TD%20Football%20Cleat
5	1104	/department/fan%20shop/category/water%20sports/product/Pelican%20Sunstream%20100%20Kayak
6	1084	/department/fan%20shop/category/indoor/outdoor%20games/product/O'Brien%20Men's%20Neoprene%20Life%20Vest
7	1059	/department/fan%20shop/category/camping%20&%20hiking/product/Diamondback%20Women's%20Serene%20Classic%20Comfort%20Bi
8	1028	/department/fan%20shop/category/fishing/product/Field%20&%20Stream%20Sportsman%2016%20Gun%20Fire%20Safe
9	1004	/department/footwear/category/cardio%20equipment/product/Nike%20Men's%20Free%205.0+%20Running%20Shoe
10	939	/department/footwear/category/fitness%20accessories/product/Under%20Armour%20Hustle%20Storm%20Medium%20Duffle%20Bag
11	930	/department/golf/category/shop%20by%20sport/product/Columbia%20Men's%20PFG%20Anchor%20Tough%20T-Shirt
12	896	/department/fitness/category/tennis%20&%20racquet/product/Nike%20Men's%20Comfort%202%20Slide
13	892	/department/footwear/category/as%20seen%20on%20%20tv!/product/Nike%20Men's%20Free%20TR%205.0%20TB%20Training%20Shoe
14	873	/department/golf/category/shop%20by%20sport/product/Under%20Armour%20Girls'%20Toddler%20Spine%20Surge%20Runni
15	870	/department/fitness/category/lacrosse/product/Under%20Armour%20Men's%20Tech%20III%20T-Shirt
16	859	/department/fitness/category/soccer/product/Nike%20Men's%20Fingertrap%20Max%20Training%20Shoe
17	727	/department/apparel/category/featured%20shops/product/adidas%20Kids'%20RG%20III%20Mid%20Football%20Cleat/add_to_cart