



**College of Professional Studies
Northeastern University San Jose**

MPS Analytics

Course: ALY6110 – Data Management and Big Data

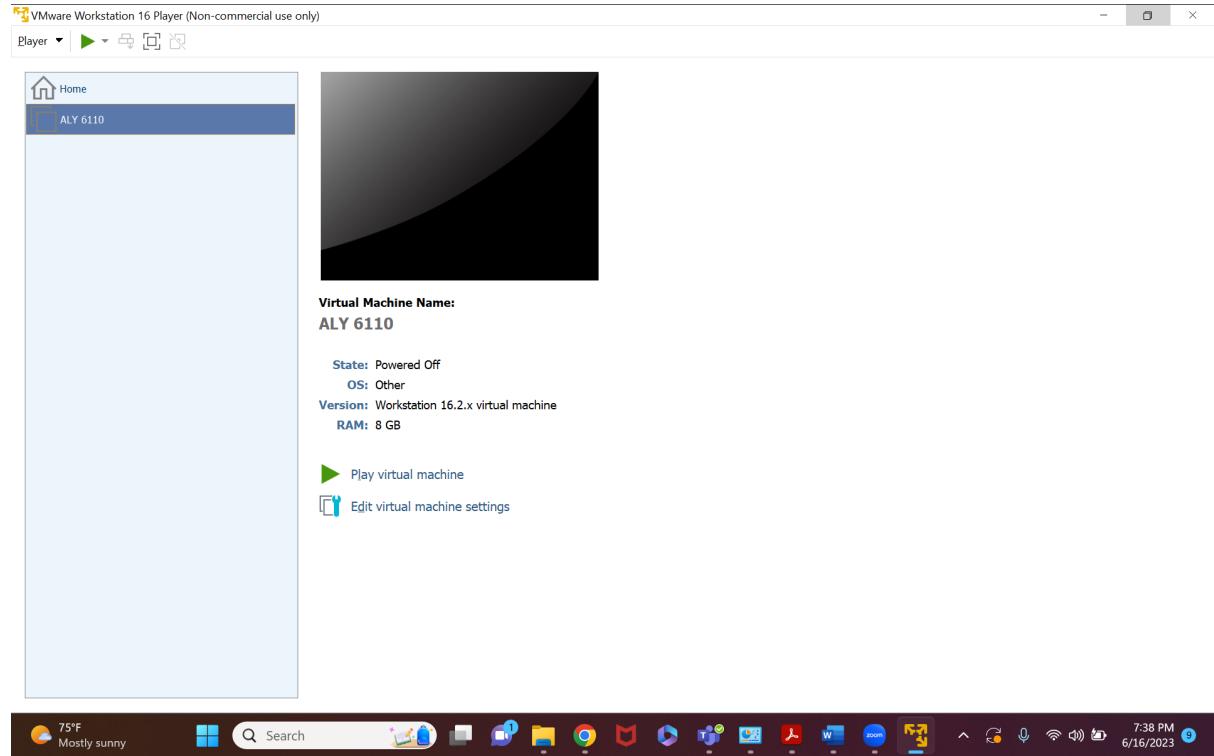
**Assignment:
Module 3 Lab 1**

**Submitted on:
June 17th, 2023**

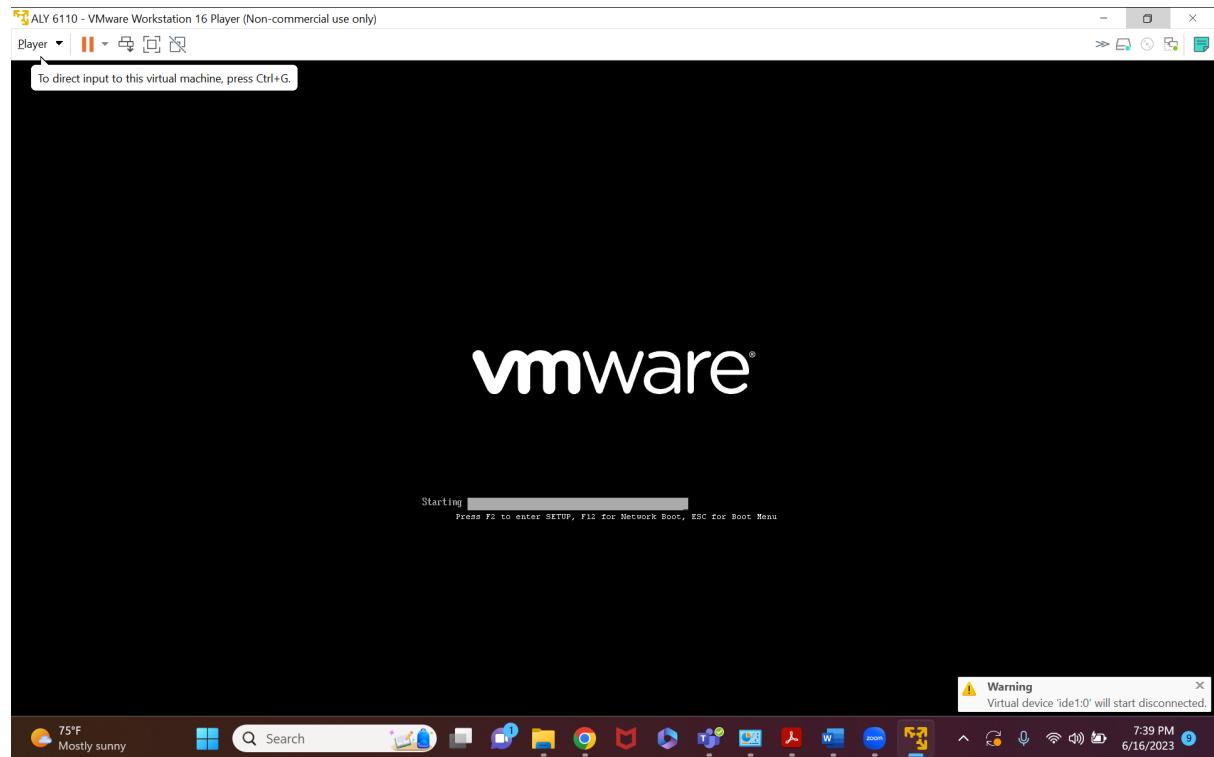
Submitted to: Professor: BEHZAD AHMADI **Submitted by:** NIKSHITA RANGANATHAN

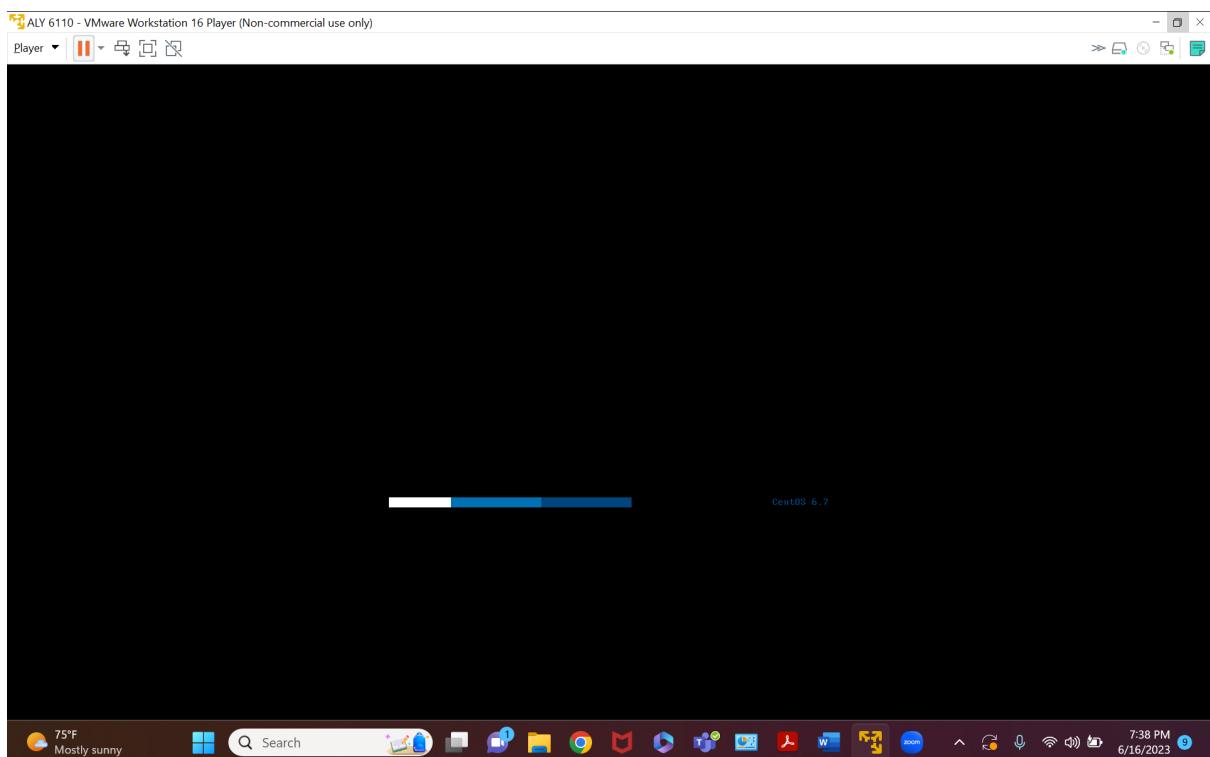
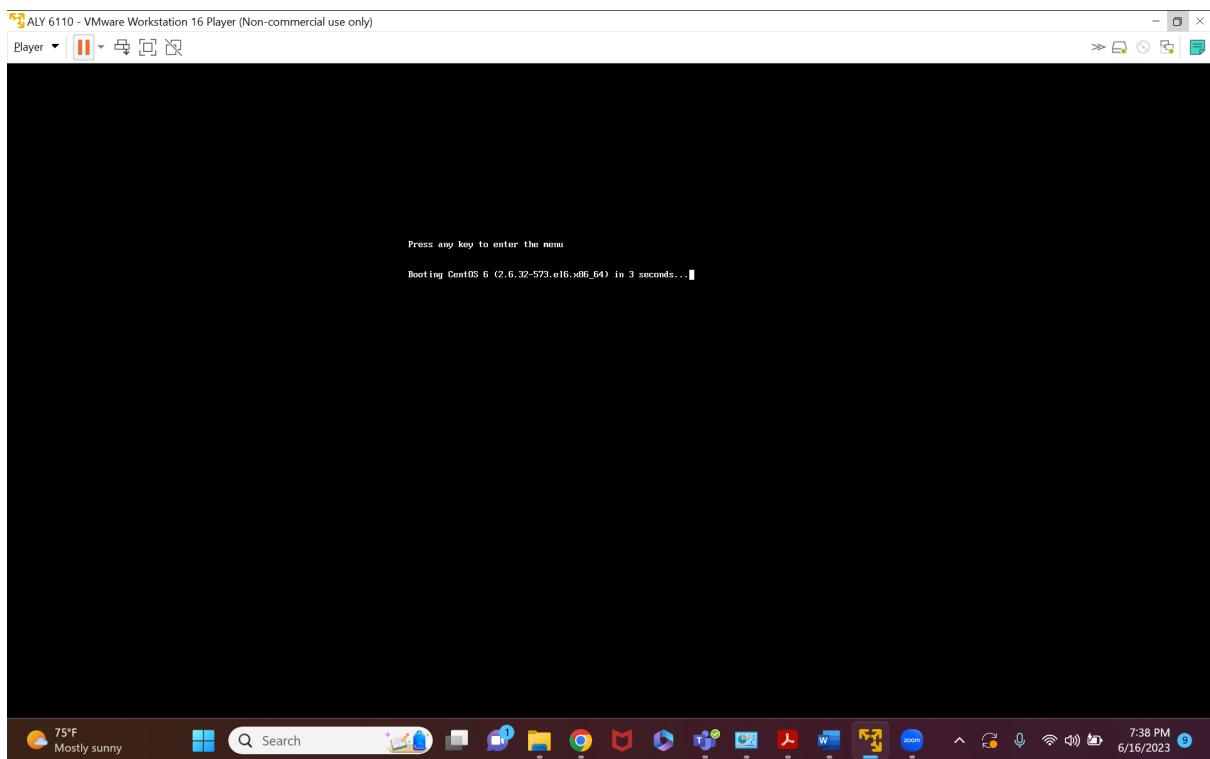
Exercise-1 Ingest and query relational data

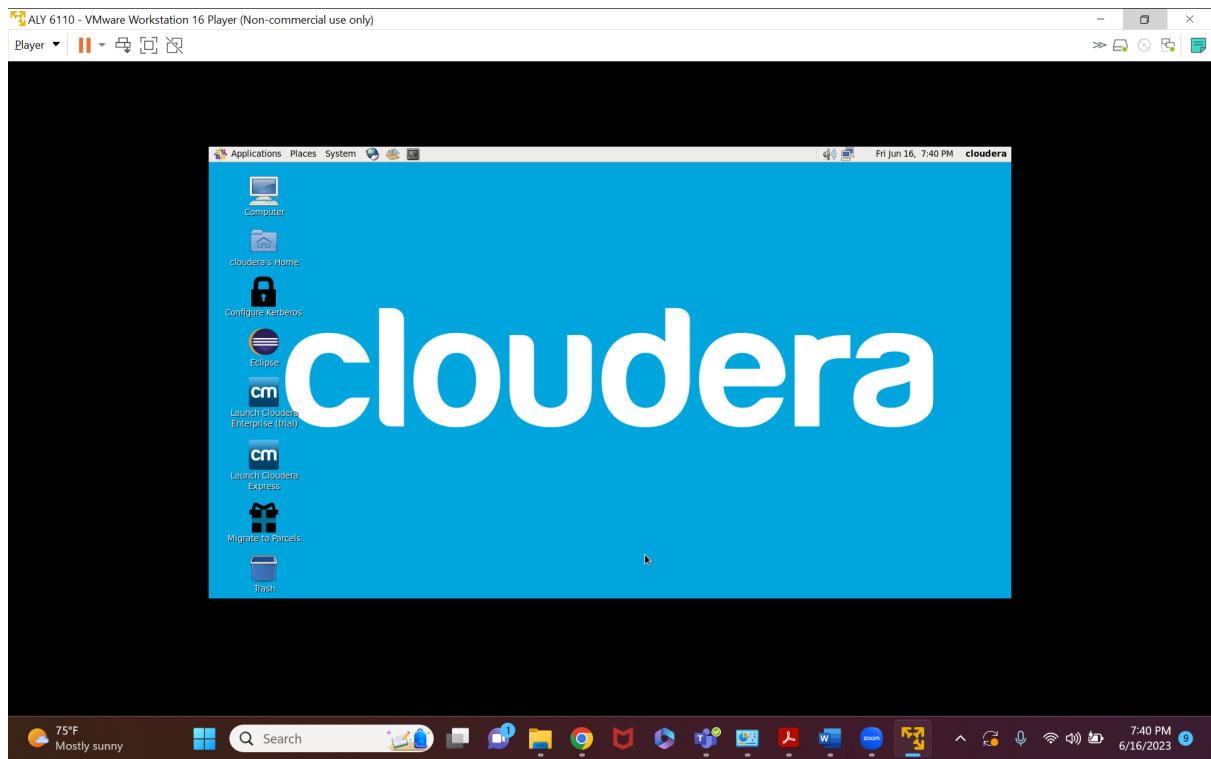
1. ALY 6110 Virtual Machine



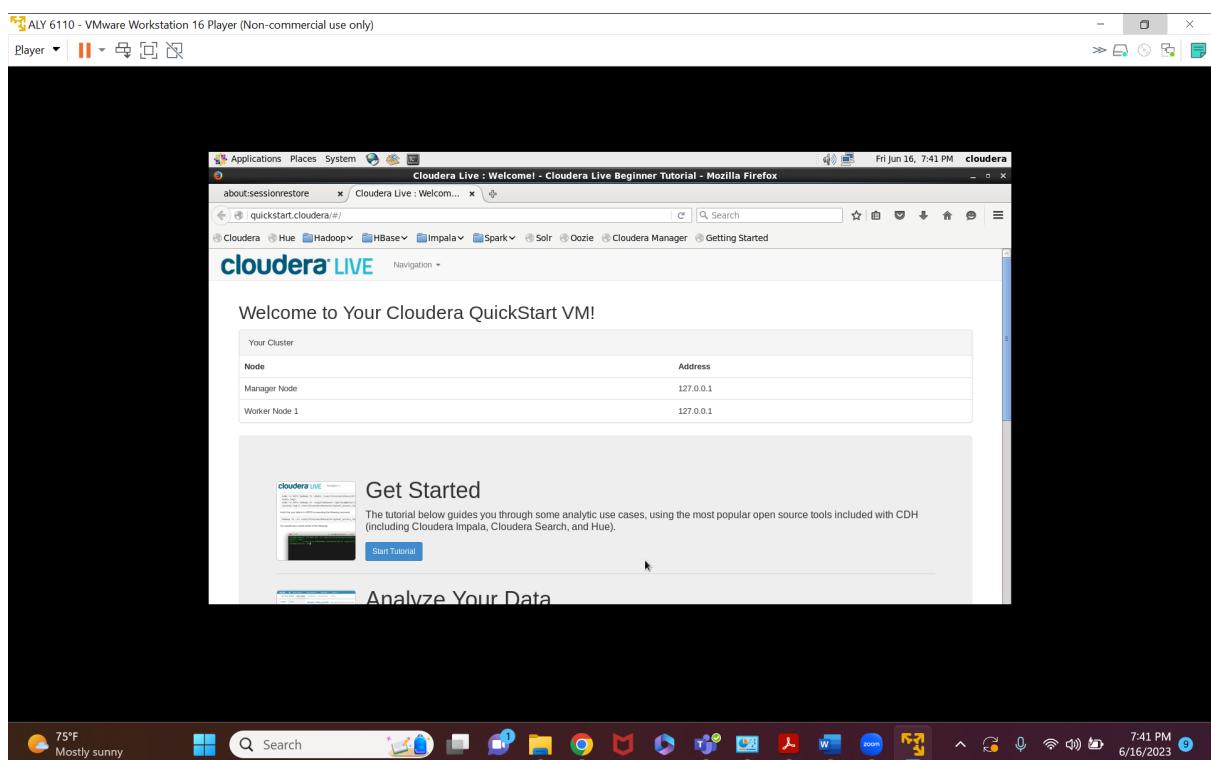
2. Starting the virtual machine



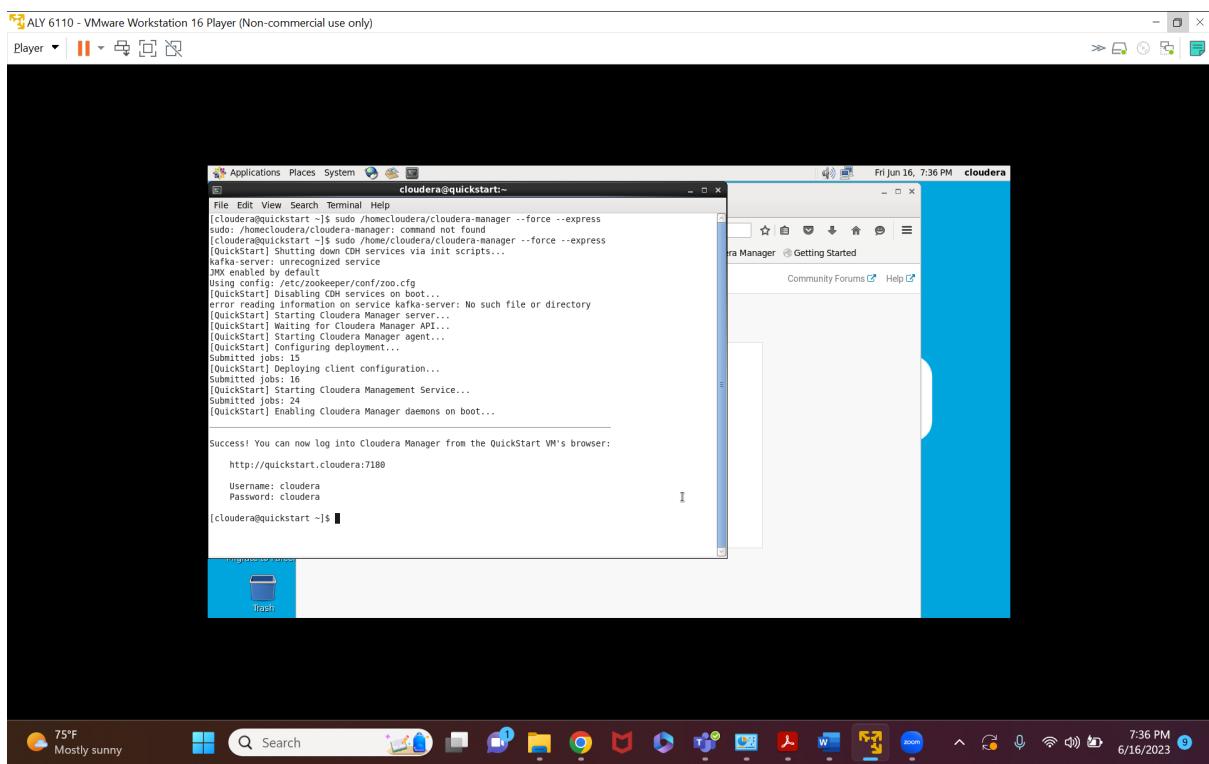
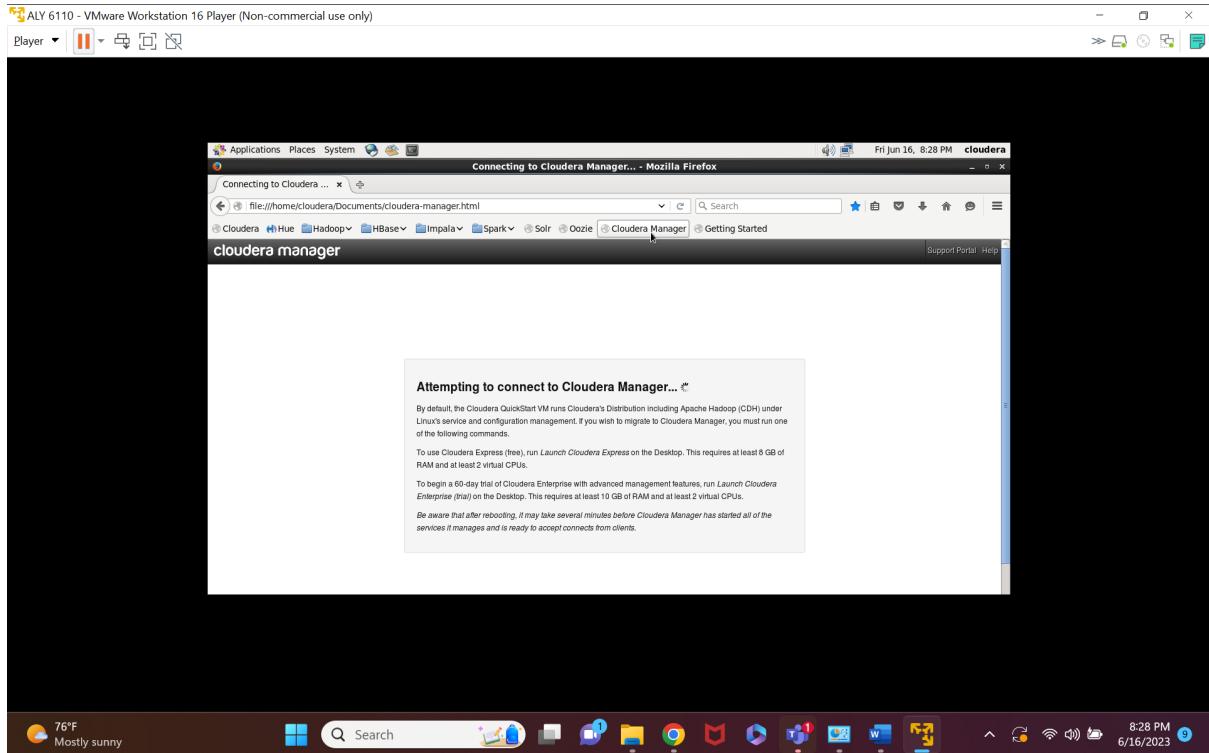


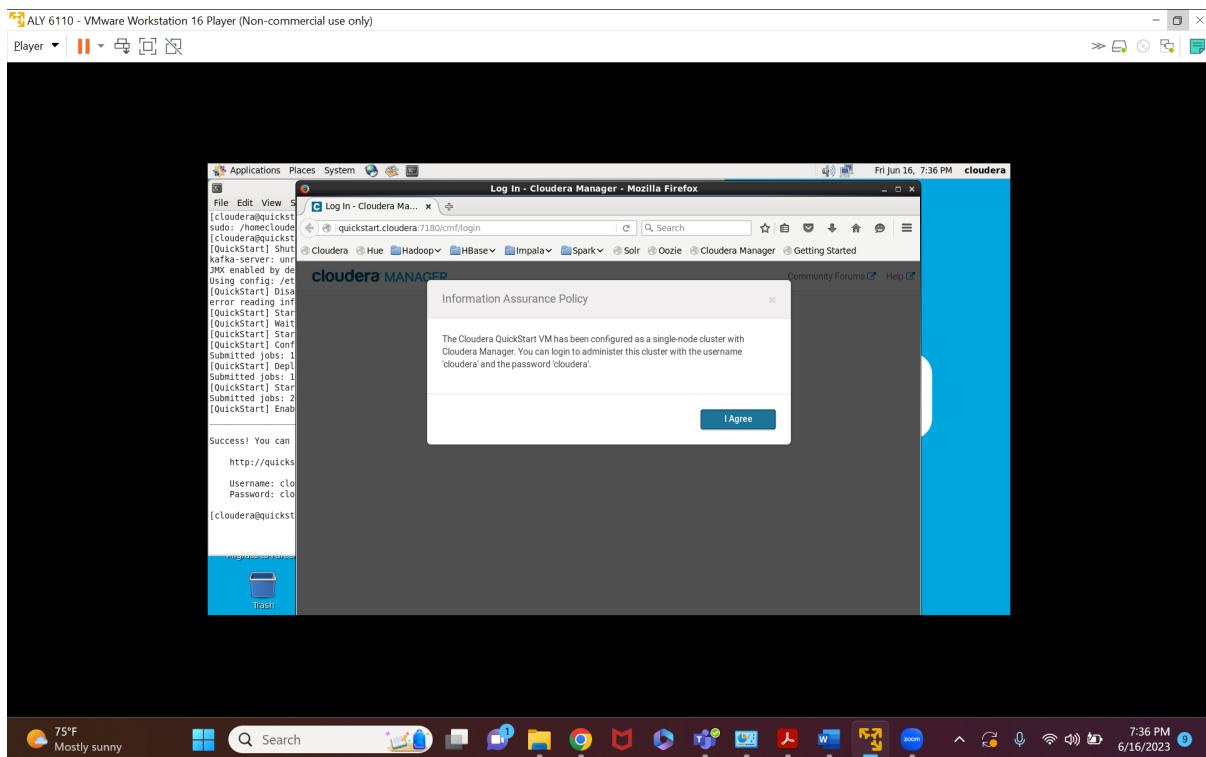


3. Viewing the 'Cloudera' Homepage in Mozilla Firefox Browser. This interface serves as a gateway to explore and interact with the tools and features available in the 'Cloudera'.

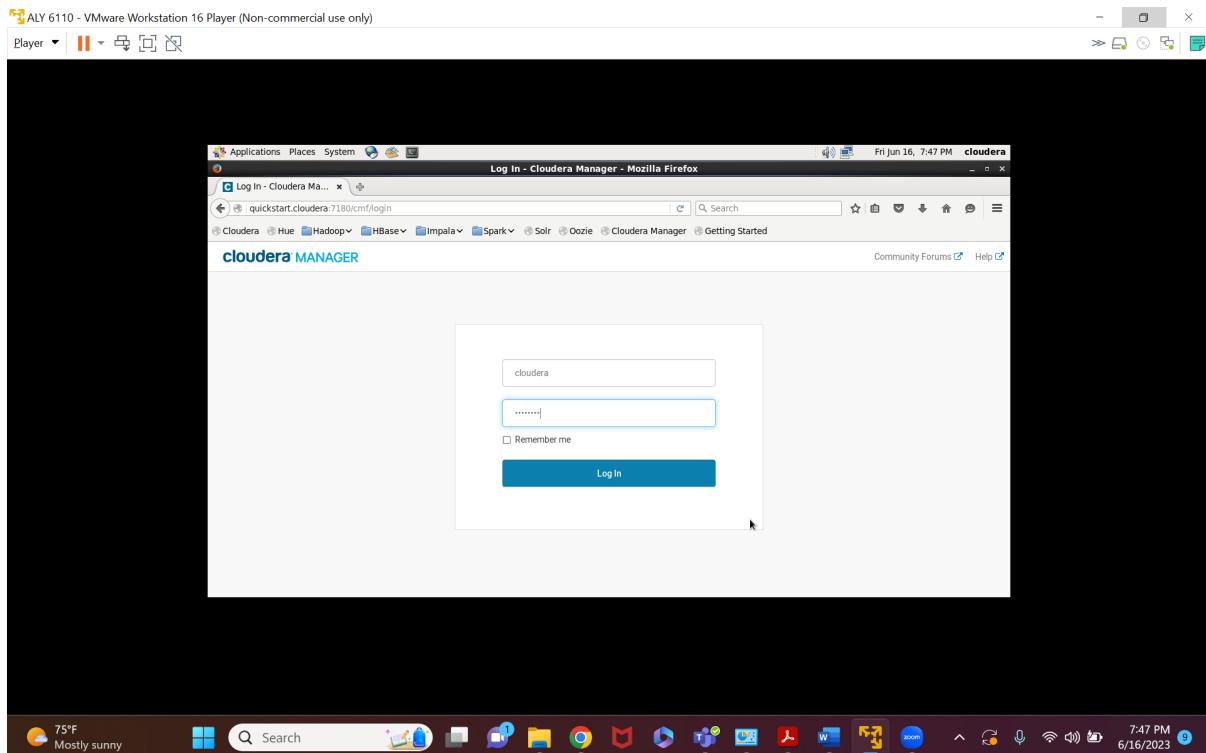


4. Opening Cloudera Manager

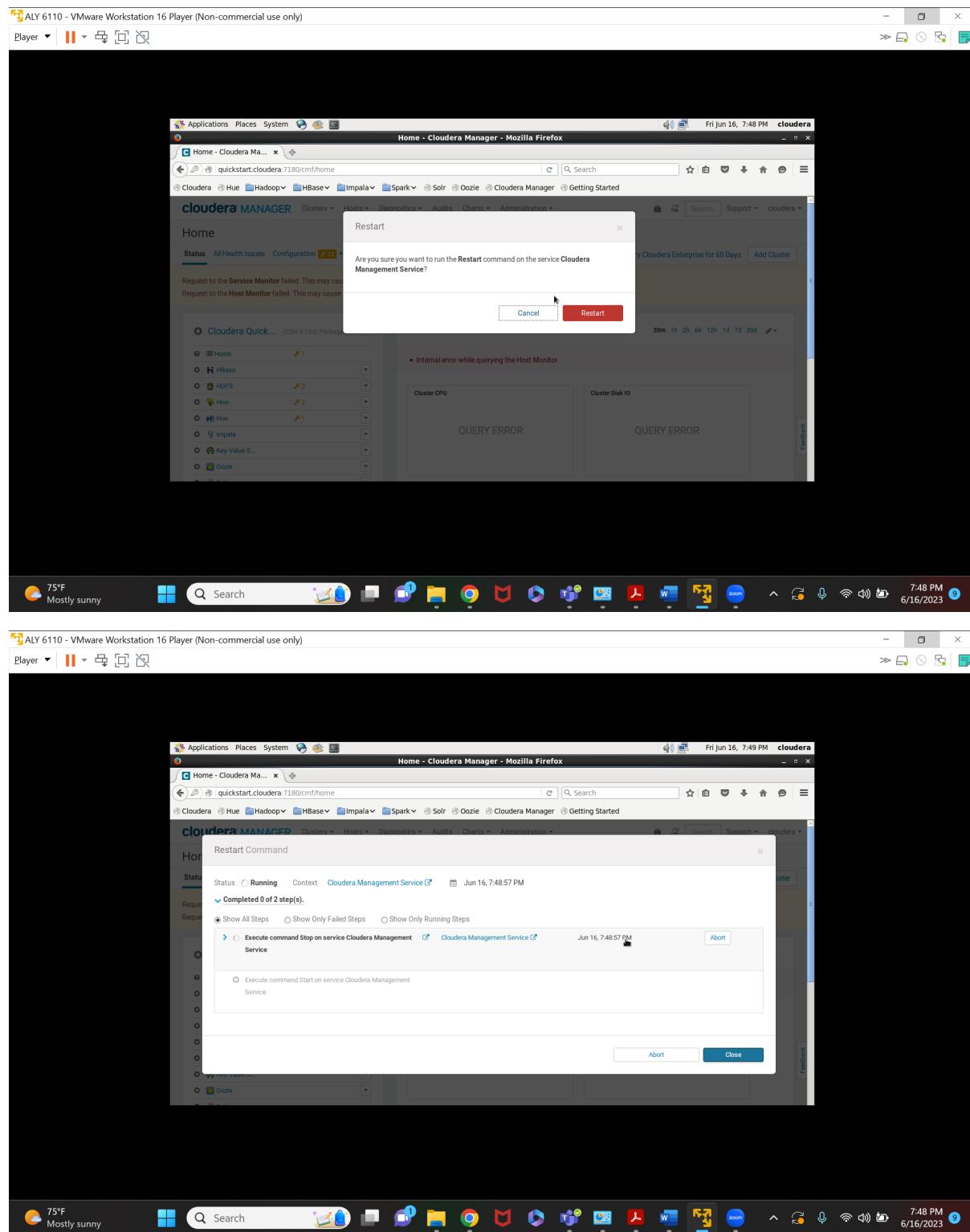




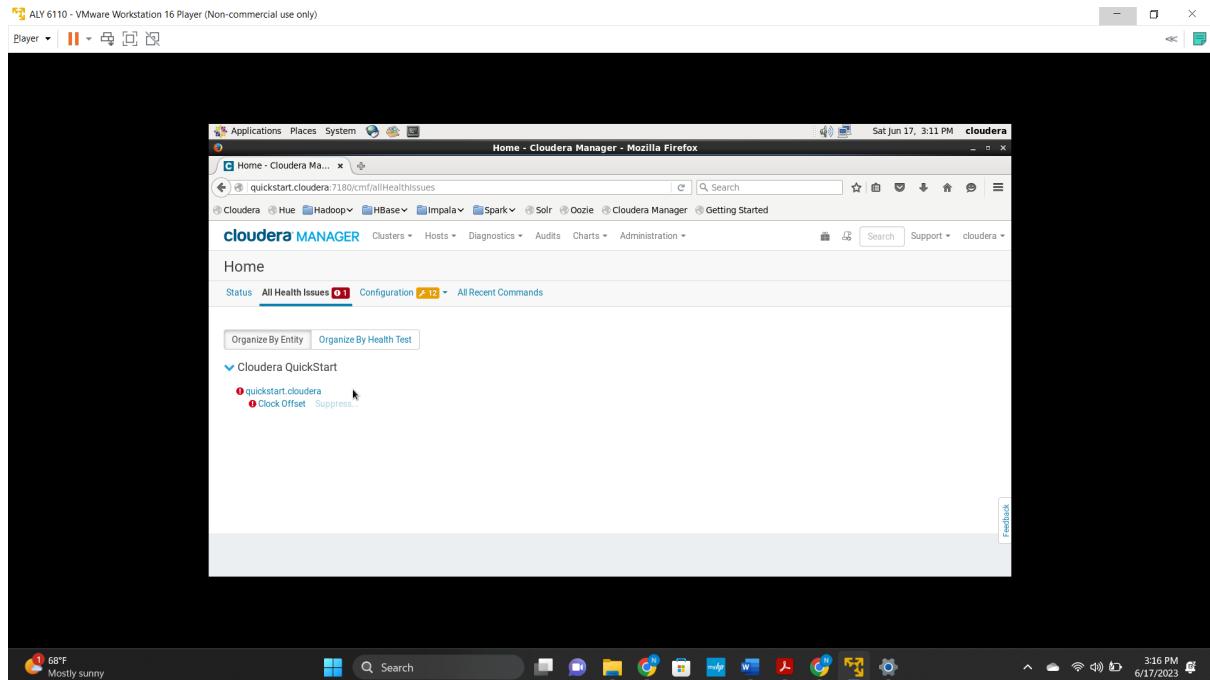
5. Logging in to Cloudera manager using username and password “cloudera”



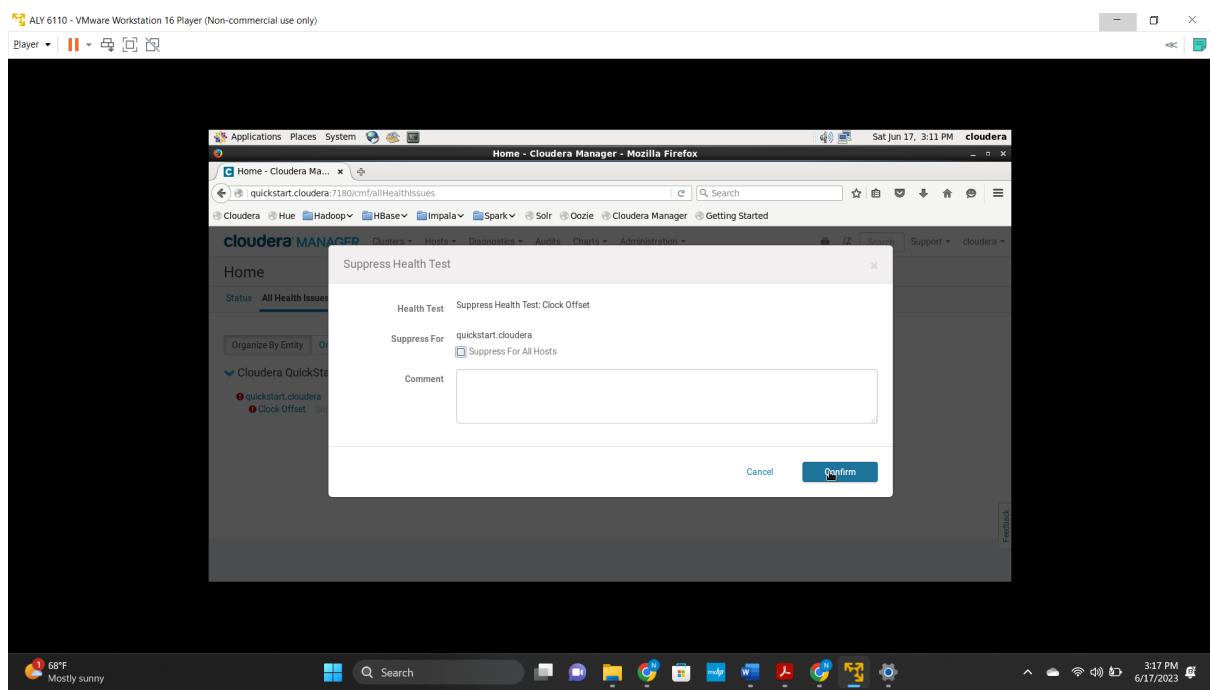
- 6. Host monitor was not running so restarted the Cloudera management service. This allowed the host monitor to resume its monitoring tasks and ensure the proper functioning of the Cloudera cluster.**



7. Suppressing the health tests

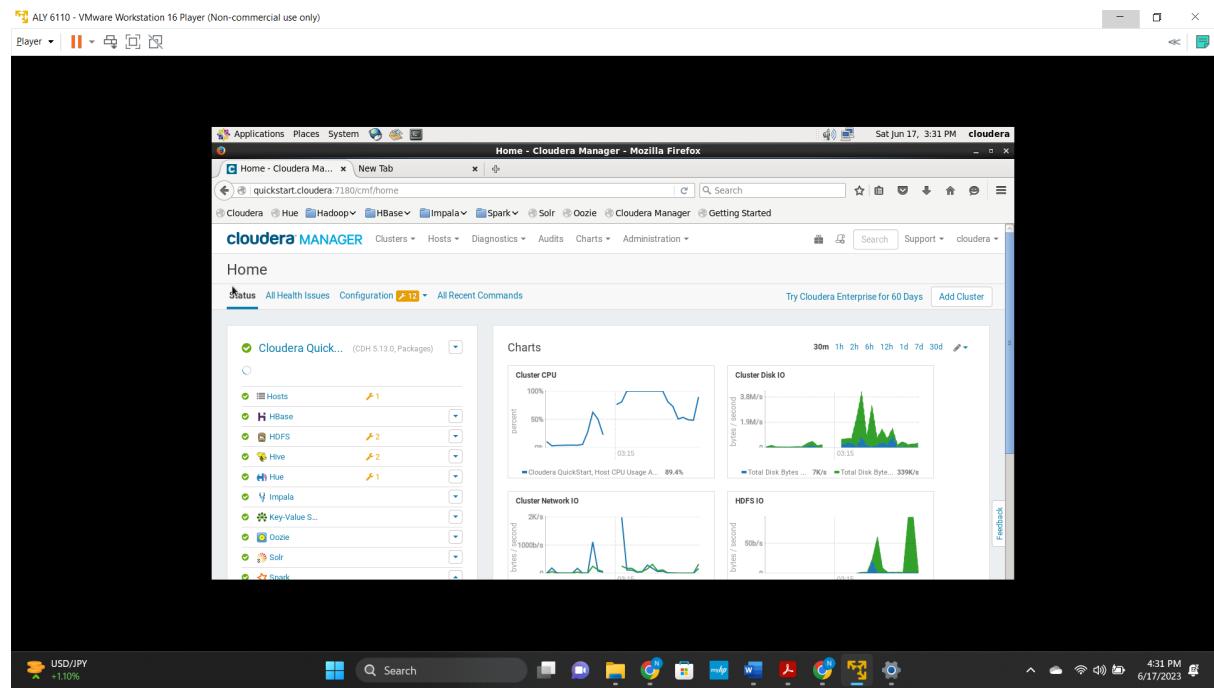


The screenshot shows the Cloudera Manager Home page in Mozilla Firefox. At the top, there's a navigation bar with links for Applications, Places, System, and a search bar. Below that is a header bar with the title "Home - Cloudera Manager - Mozilla Firefox", the date "Sat Jun 17, 3:11 PM", and the word "cloudera". The main content area has a heading "cloudera MANAGER" and a sub-heading "Home". It displays a status summary: "Status All Health Issues 0" (highlighted in red), "Configuration 12", and "All Recent Commands". Below this, there are two tabs: "Organize By Entity" and "Organize By Health Test". A section titled "Cloudera QuickStart" lists a single health issue: "quickstart.cloudera" with the error "Clock Offset". To the right of this issue is a link labeled "Suppress...". The bottom of the screen shows a Windows taskbar with various icons and the system tray.

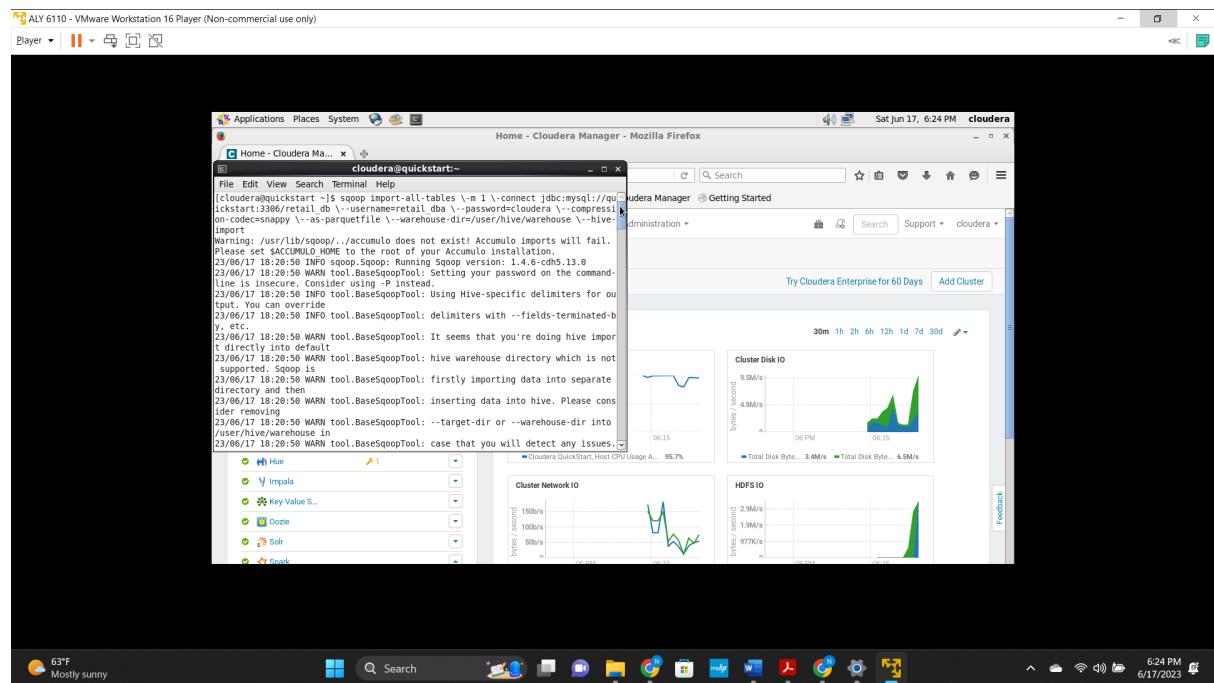


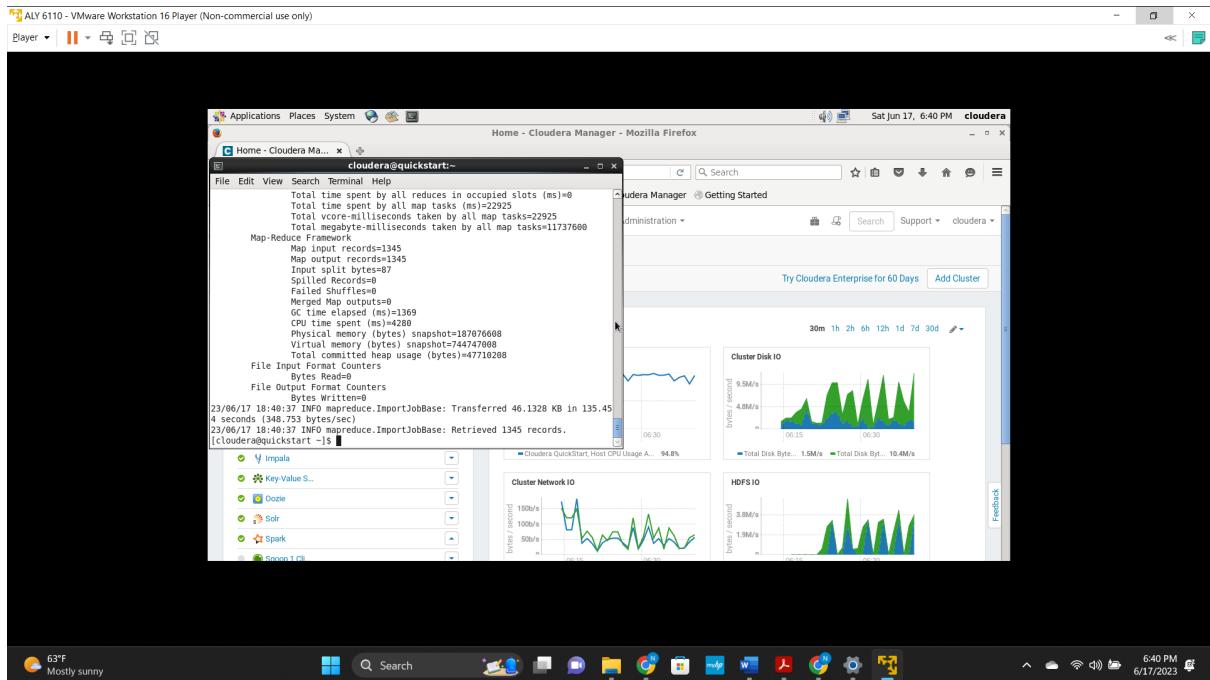
The screenshot shows a "Suppress Health Test" dialog box overlaid on the Cloudera Manager Home page. The dialog has a title "Suppress Health Test". Inside, it shows the "Health Test" set to "Suppress Health Test: Clock Offset", the "Suppress For" field containing "quickstart.cloudera" with an unchecked checkbox for "Suppress For All Hosts", and a "Comment" text area. At the bottom of the dialog are "Cancel" and "Confirm" buttons. The background of the dialog is semi-transparent, allowing the underlying Cloudera Manager interface to be seen.

8. Cloudera manager with all services and charts running

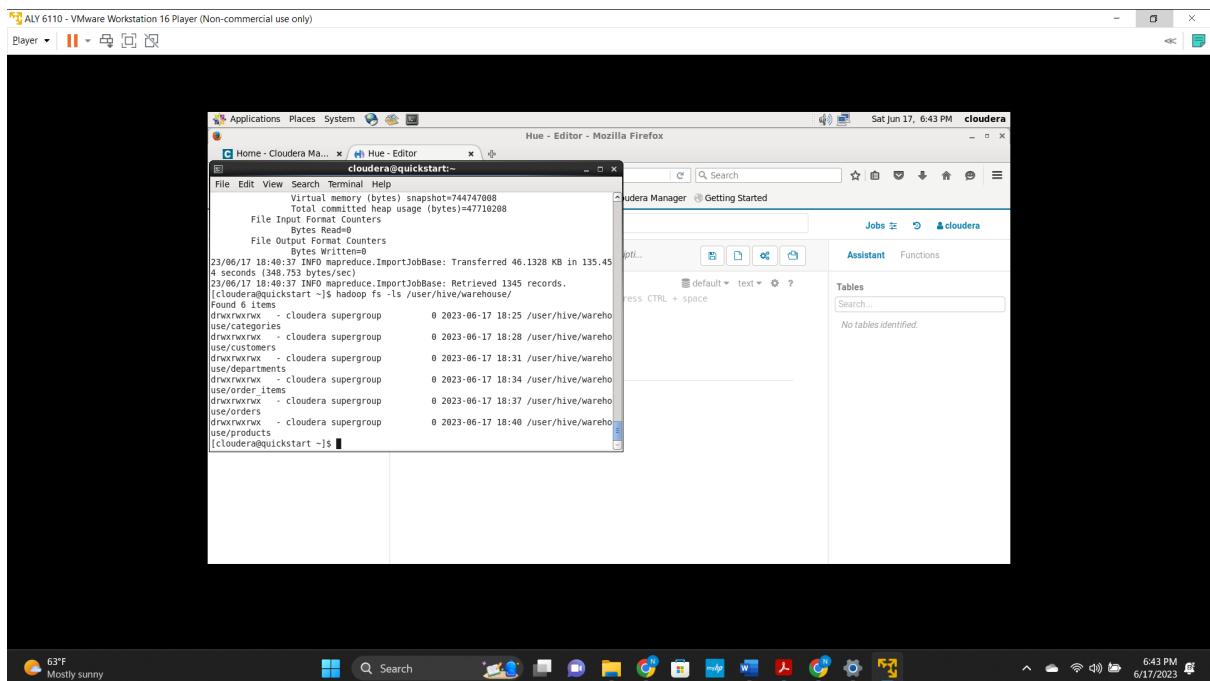


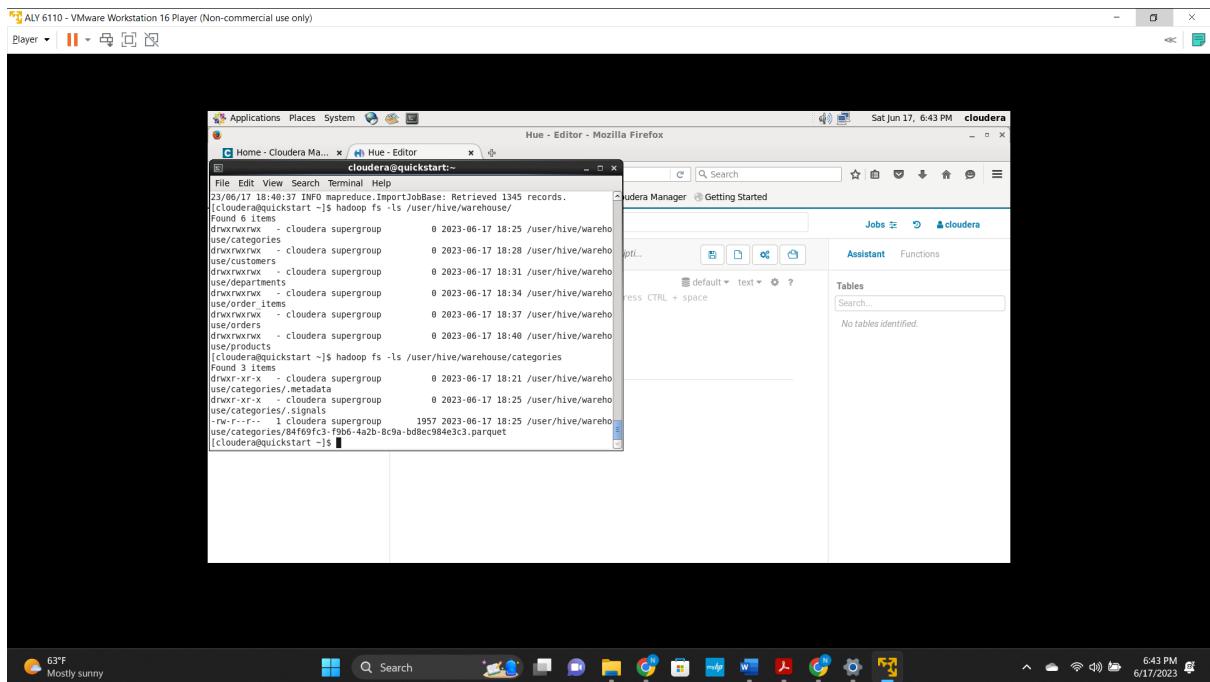
9. In the terminal, we executed the below sqoop command and ingested data from MySQL (RDBMS) to HDFS in Parquet format.



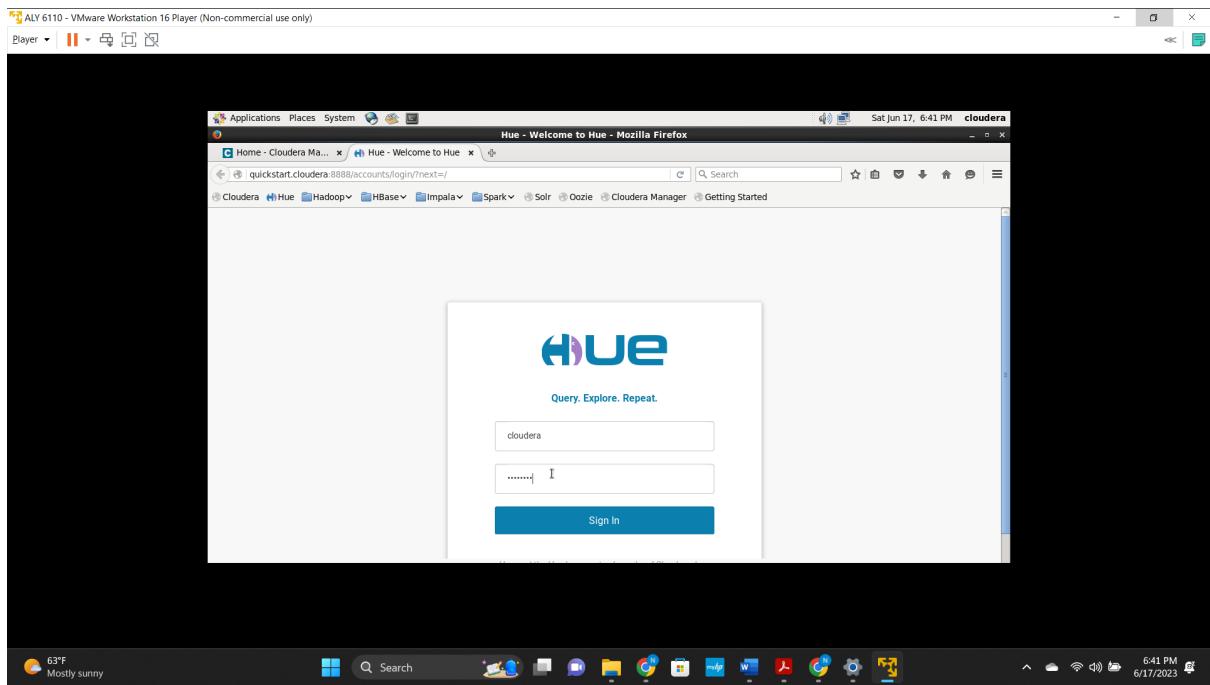


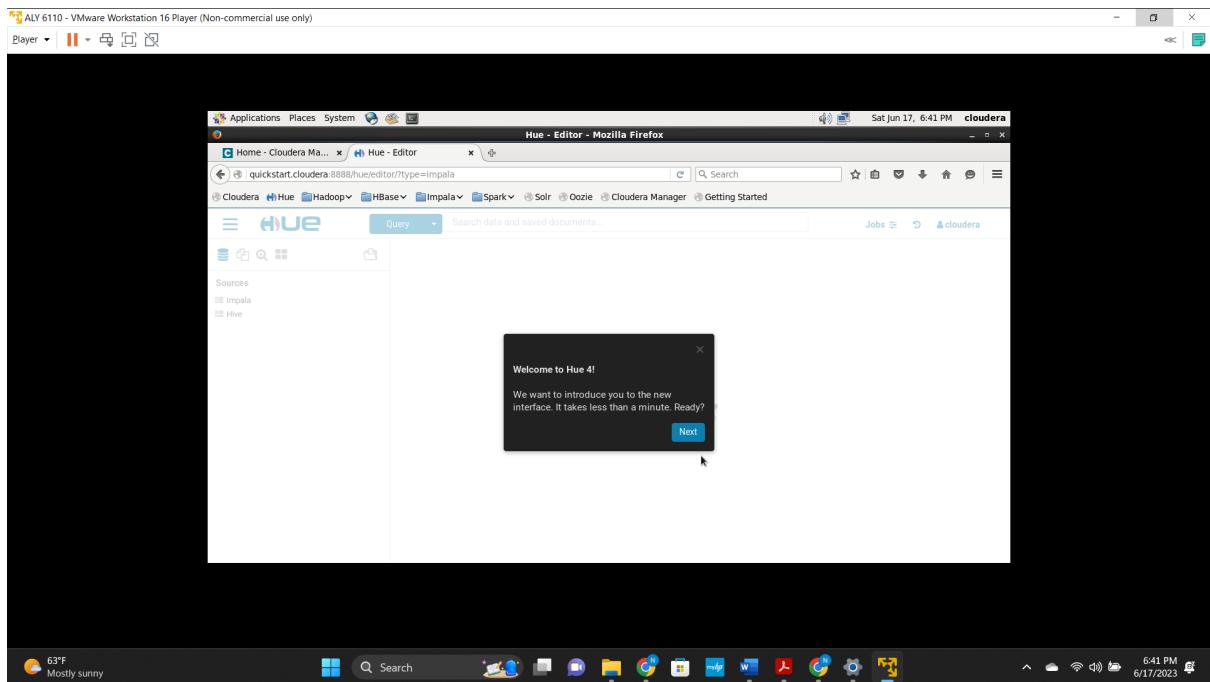
10. Once the sqoop command is executed, in the next step we check that the data files are present in HDFS.



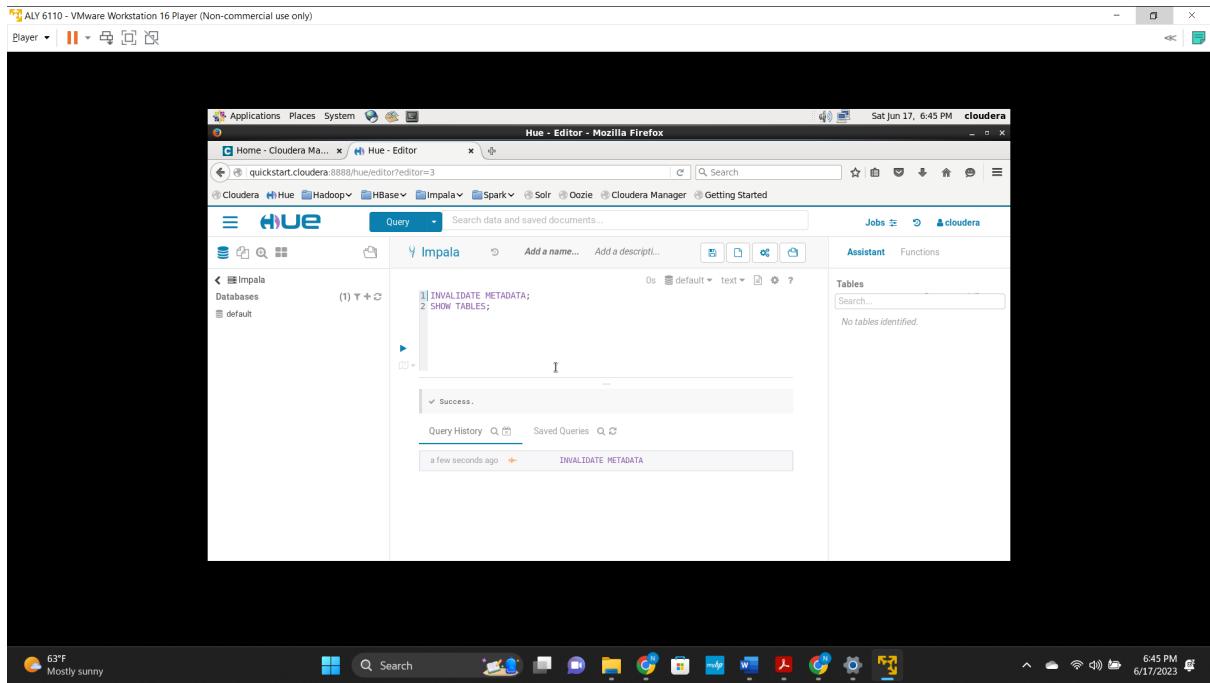


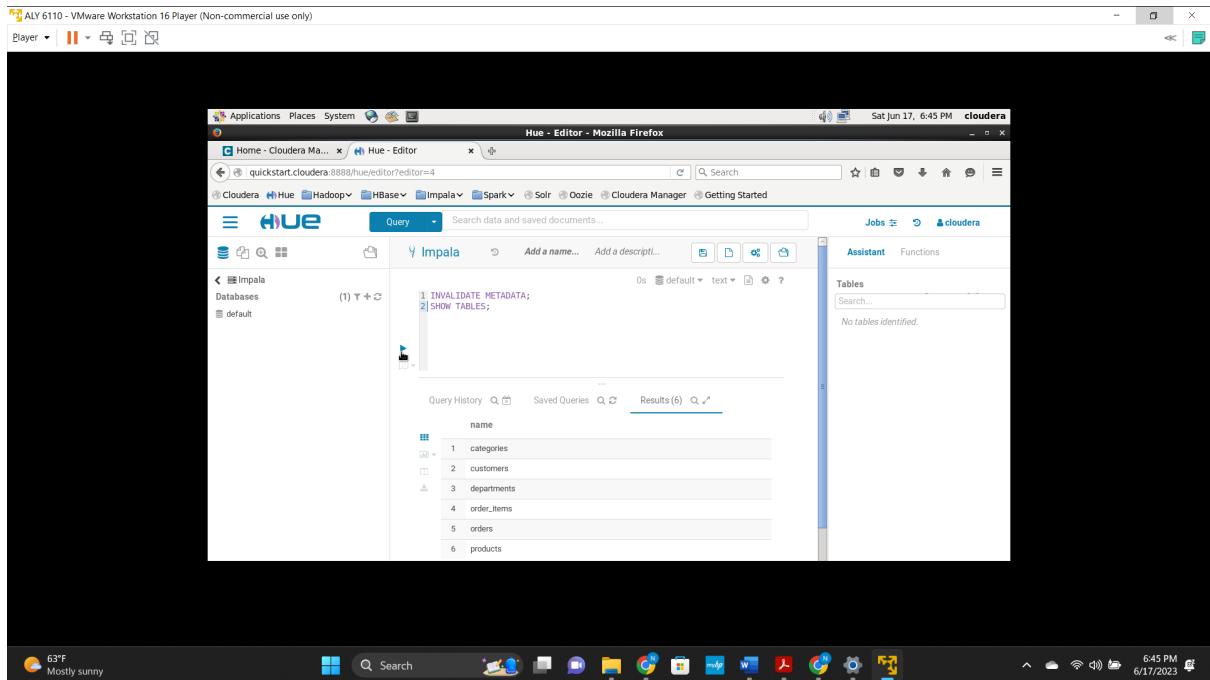
11. Then, we access Hue available on port 8888 of the Manager Node. User name and password are ‘cloudera’





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

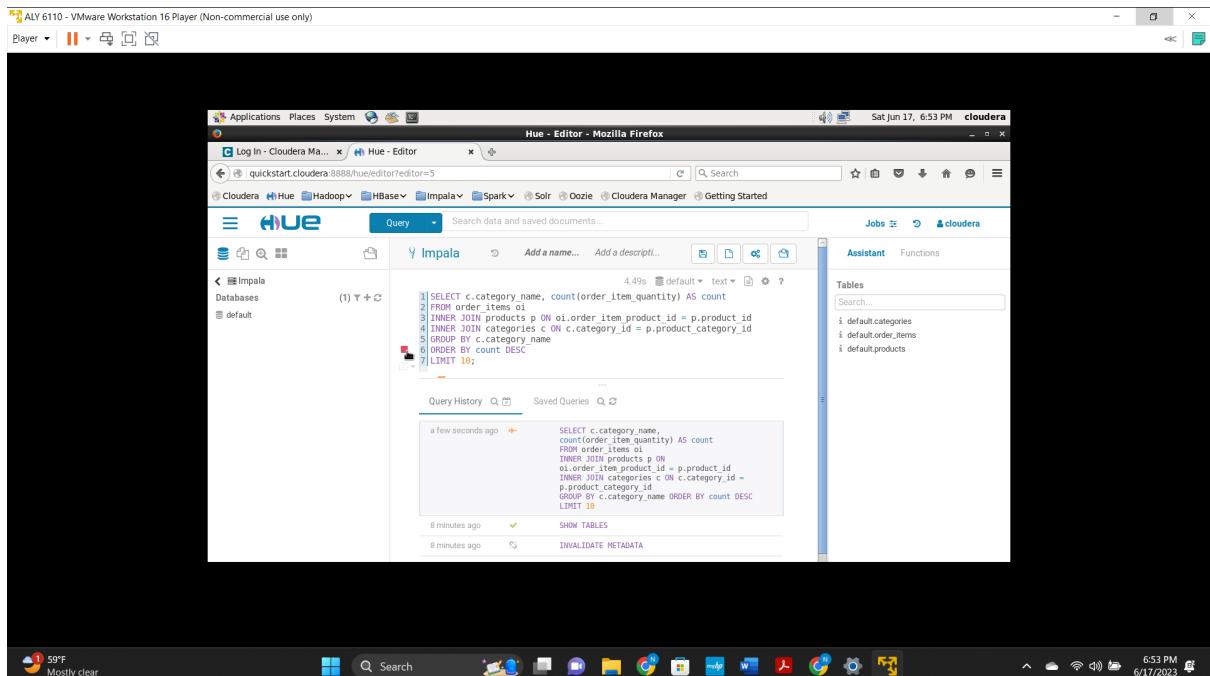


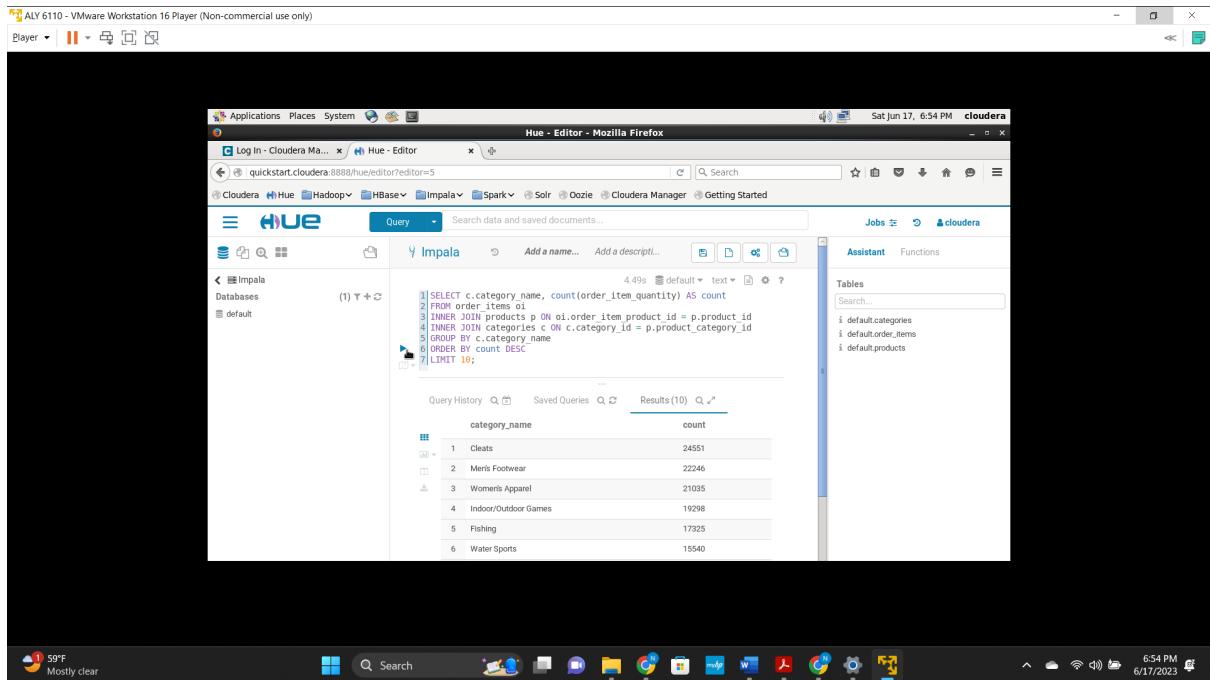


13. Executing the provided SQL example query to calculate total revenue per product and display the top 10 products generating the highest revenue.

```

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 | || |

The screenshot shows the Hue Editor interface in Mozilla Firefox. A query is being run against the Impala database. The query code is as follows:

```
3| SELECT ol.order_item_product_id, SUM(CAST(ol.order_item_subtotal AS
4| FROM order_items ol INNER JOIN orders o
5| ON ol.order_item_order_id = o.order_id
6| WHERE o.order_status = 'COMPLETED'
7| AND ol.order_status <> 'SUSPECTED_FRAUD'
8| GROUP BY ol.order_item_product_id r ON p.product_id = r.order_item_product_id
9| ORDER BY r.revenue DESC
10| LIMIT 10;
```

The results pane shows two queries:

- 20 minutes ago: SELECT c.category_name, count(order_item.quantity) AS count, FROM order_items ol INNER JOIN products p ON ol.order_item_product_id = p.product_id TIMES JOIN categories c ON c.category_id = p.product_category_id GROUP BY c.category_name ORDER BY count DESC LIMIT 10.
- 28 minutes ago: SHOW TABLES

The desktop taskbar at the bottom shows various application icons and the date/time: 7:25 PM 6/17/2023.

ALY 6110 - VMware Workstation 16 Player (Non-commercial use only)

Player | || |

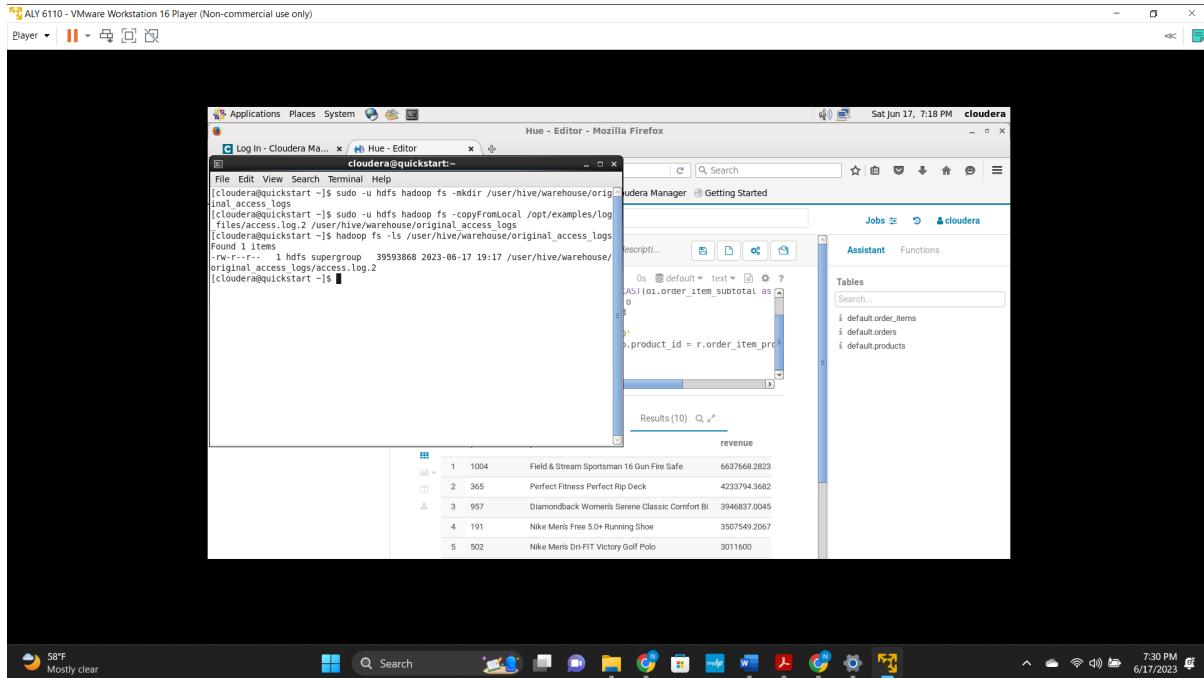
The screenshot shows the Hue Editor interface in Mozilla Firefox. A query has been run and the results are displayed in a table format. The query code is identical to the one in the previous screenshot.

product_id	product_name	revenue	
1	1004	Field & Stream Sportsman 16 Gun Safe	6637668.2823
2	365	Perfect Fitness Perfect Rip Deck	4233794.3682
3	957	Diamondback Womens Serene Classic Comfort BI	3946837.0045
4	191	Nike Men's Free 5.0+ Running Shoe	3507549.2067
5	502	Nike Merle Dri-FIT Victory Golf Polo	3011600

The desktop taskbar at the bottom shows various application icons and the date/time: 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



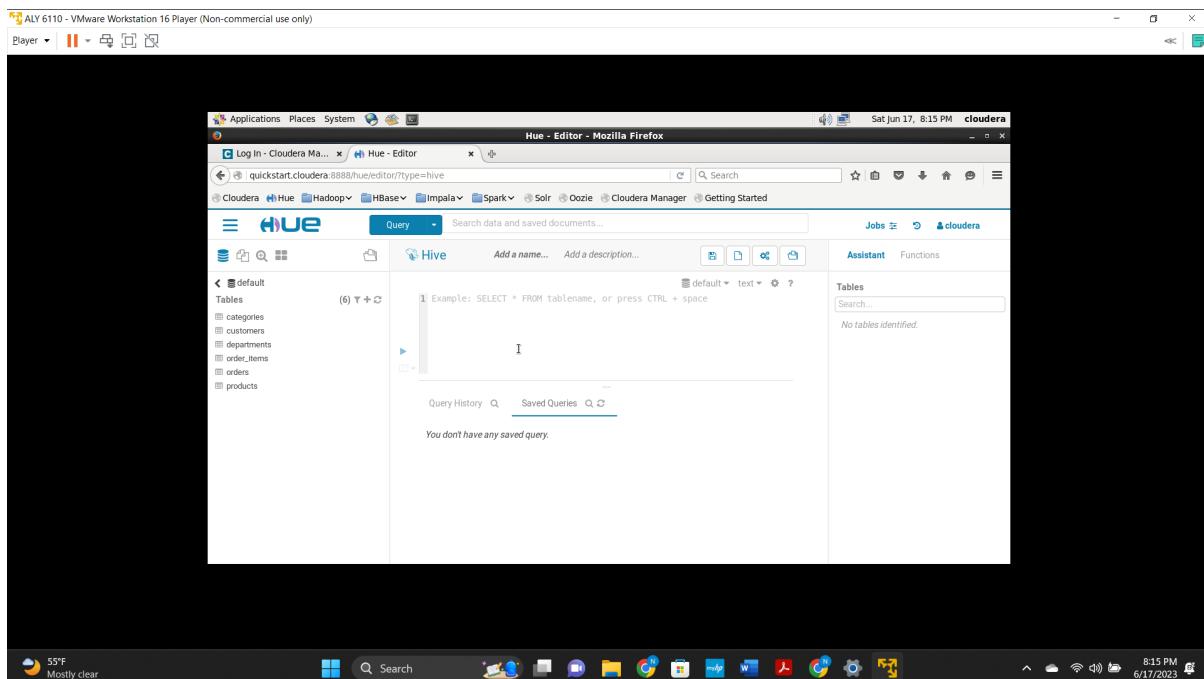
The screenshot shows a Linux desktop environment with several windows open. In the foreground, a terminal window titled 'Hue - Editor - Mozilla Firefox' displays the following command and its output:

```
[cloudera@quickstart ~]$ sudo -u hdfs hadoop fs -mkdir /user/hive/warehouse/original.access.logs
[cloudera@quickstart ~]$ sudo -u hdfs hadoop fs -copyFromLocal /opt/examples/logfiles/access.log.2 /user/hive/warehouse/original.access.logs
[cloudera@quickstart ~]$ hadoop fs -ls /user/hive/warehouse/original.access.logs
Found 1 items
-rw-r--r-- 1 hdfs supergroup 39593868 2023-06-17 19:17 /user/hive/warehouse/original.access.logs/access.log.2
[cloudera@quickstart ~]$
```

Below the terminal, the Hue Editor window shows a query results table:

	order_id	product_name	revenue
1	1004	Field & Stream Sportsman 16 Gun Fire Safe	6637668.2823
2	365	Perfect Fitness Perfect Rip Deck	4235794.3682
3	957	Diamondback Womens Serene Classic Comfort Bi	3946837.0045
4	191	Nike Merit Free 5.0+ Running Shoe	3507549.2067
5	502	Nike Merit Dri-FIT Victory Golf Polo	3011600

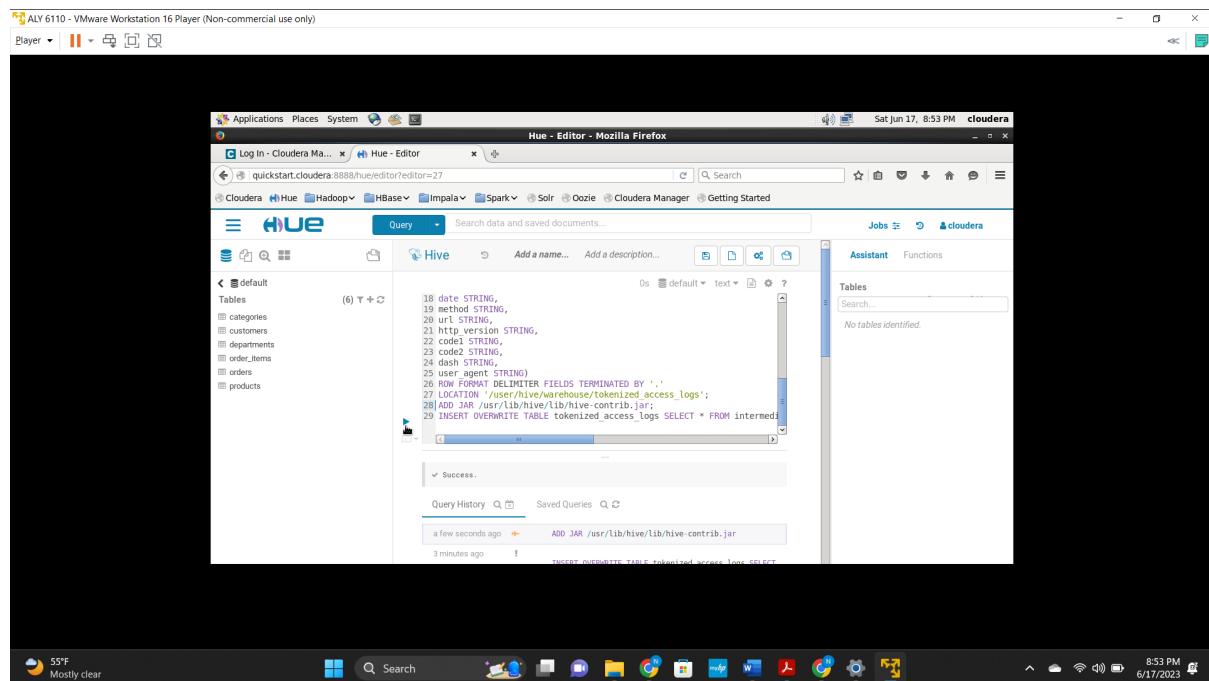
16. Query to create two tables in Hive using SerDes to parse the logs. For this, we change to Hive Query Editor first.



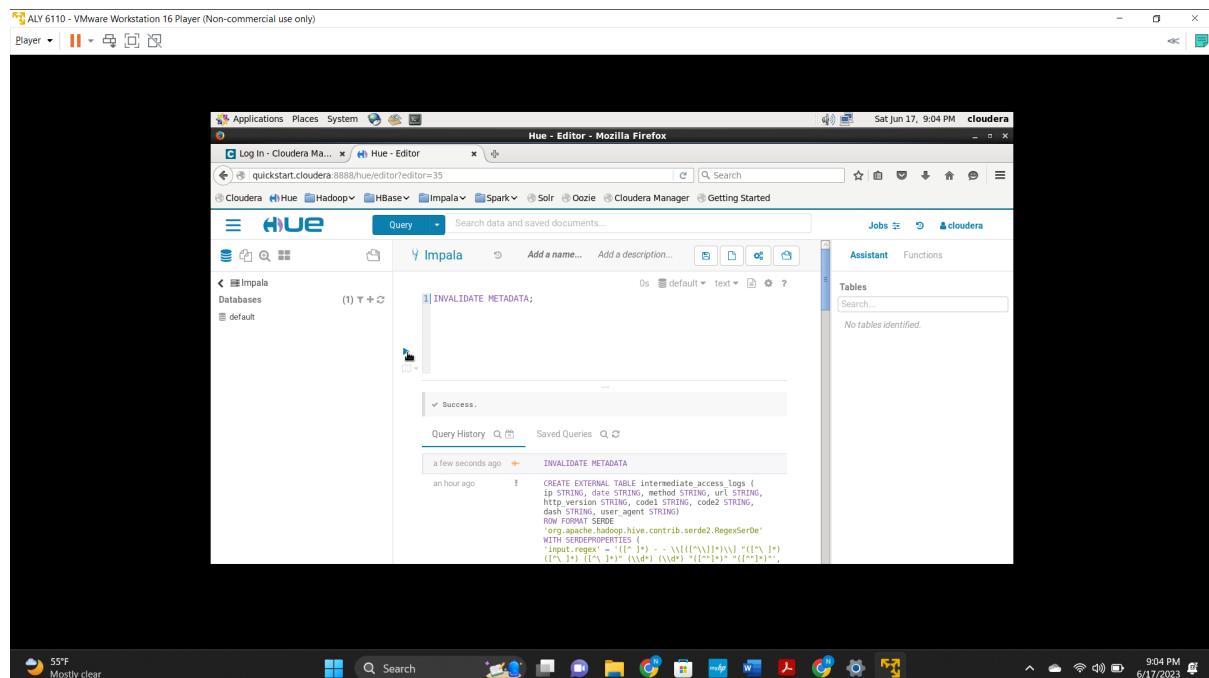
The screenshot shows a Linux desktop environment with the Hue Query Editor window open. The window title is 'Hue - Editor - Mozilla Firefox' and the URL is 'quickstart.cloudera:8888/hue/editor?type=hive'. The left sidebar shows a list of tables in the 'default' database: categories, customers, departments, order_items, orders, and products. The main pane contains a query editor with the following text:

```
1 Example: SELECT * FROM tablename, or press CTRL + space
```

The status bar at the bottom indicates the date and time as '6/17/2023 8:15 PM'.



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

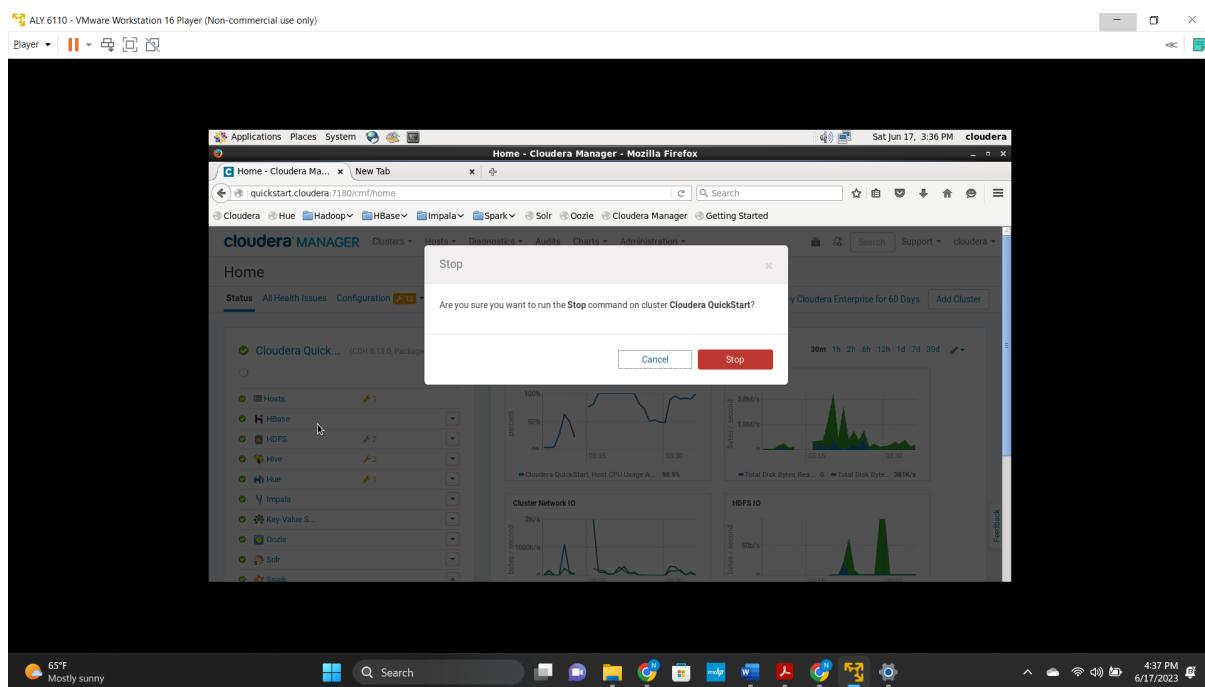


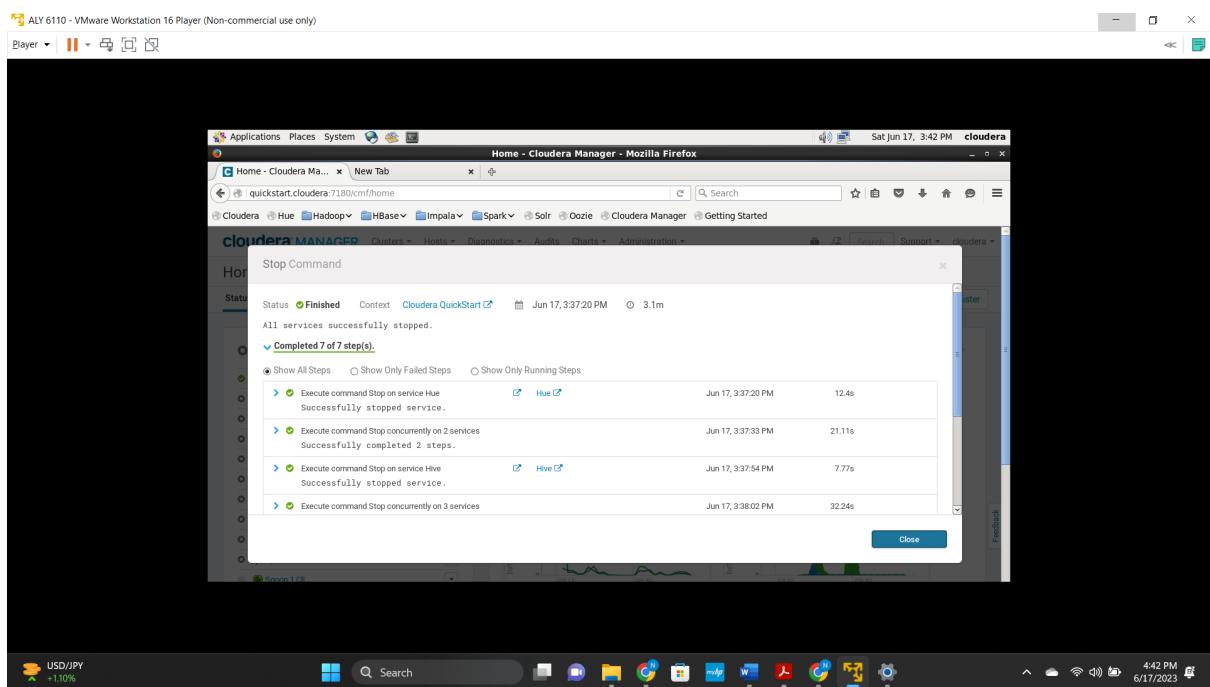
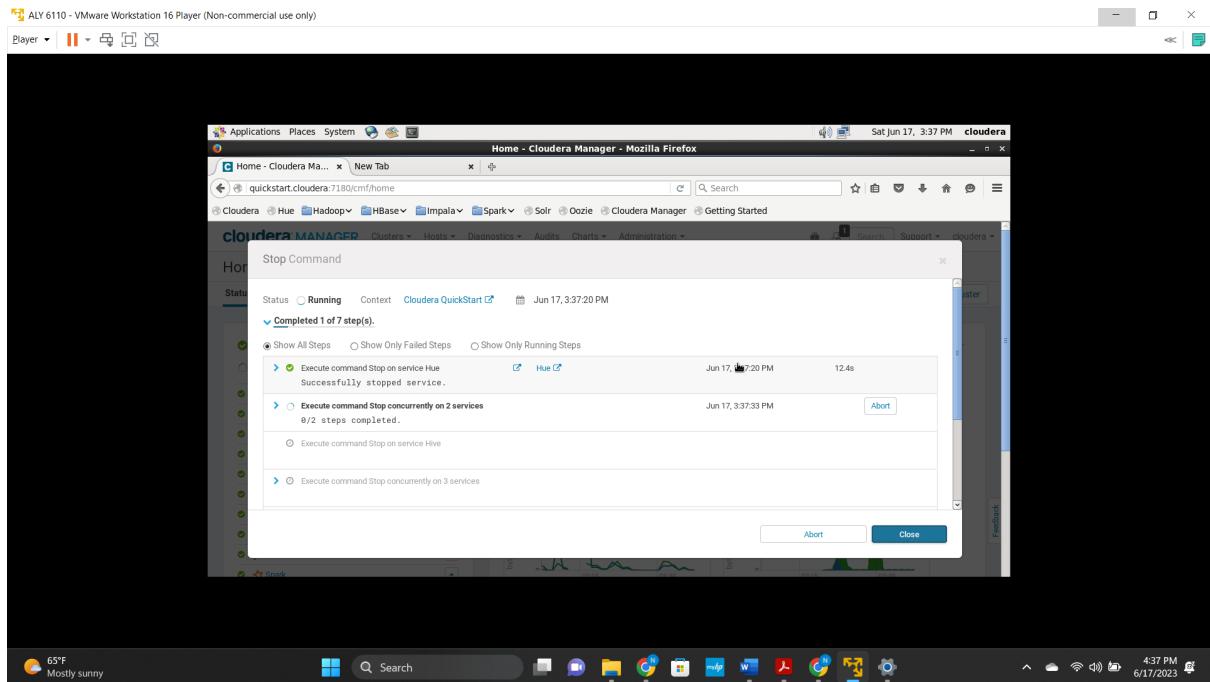
18. Query to determine the website that has been visited the most users

The screenshot shows the Hue web interface for Apache Impala. In the top navigation bar, 'Query' is selected. The main area displays a query result table titled 'Results (152)'. The table has two columns: 'count(*)' and 'url'. The data shows the following top URLs:

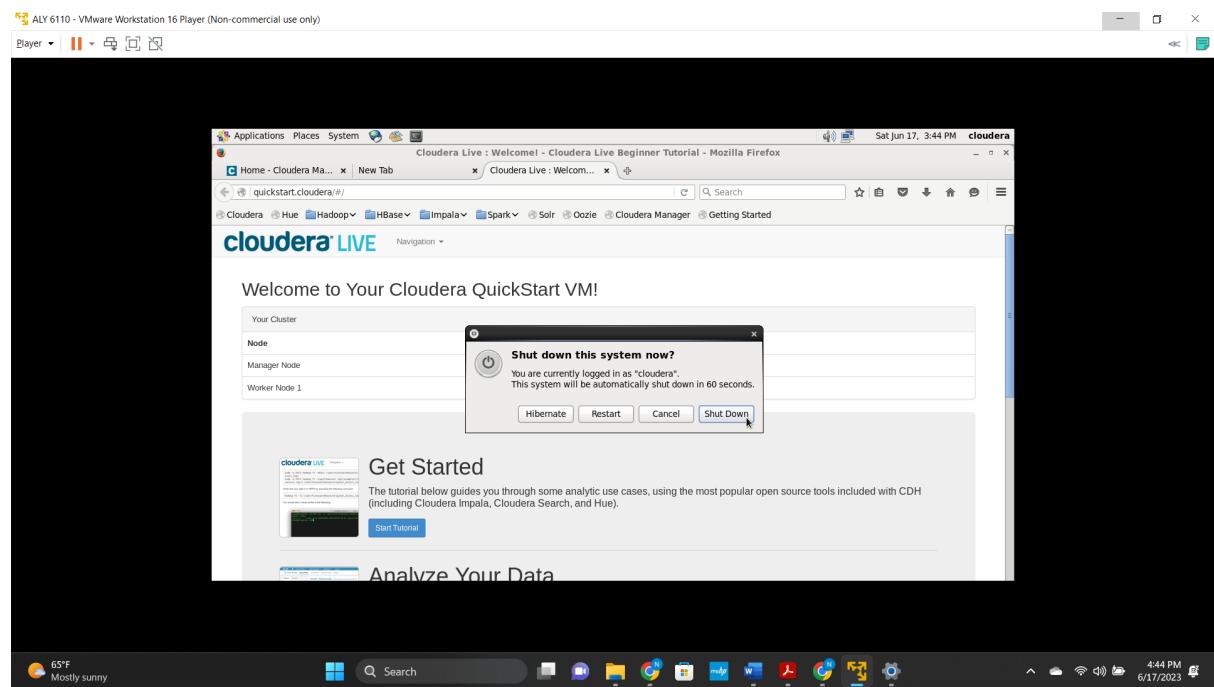
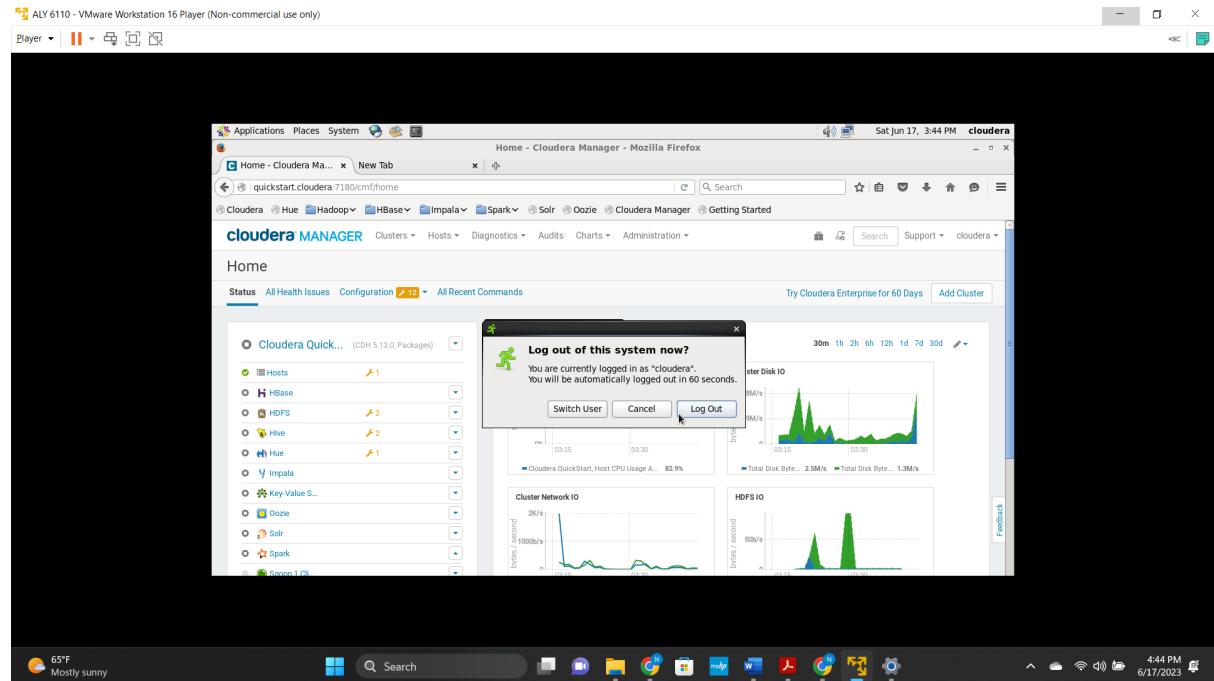
count(*)	url
1926	/department/apparel/category/cleats/product/Perfect%20Fitness%
1793	/department/apparel/category/featured%20shops/product/adidas%
1780	/department/golf/category/womens%20apparel/product/Nike%20Footwear
1757	/department/apparel/category/mens%20footwear/product/Nike%20Footwear
1104	/department/fan%20shop/category/water%20sports/product/Pelican%20Swimwear
1084	/department/fan%20shop/category/indoor%20outdoor%20games/products

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





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

The screenshot shows the Hue interface for running Impala queries. The query window contains the following SQL code:

```
3| SELECT oi.order_item_product_id, SUM(CAST(oi.order_item_subtotal AS
4| FROM order_items oi INNER JOIN orders o
5| ON oi.order_item_order_id = o.order_id
6| WHERE o.order_status <> 'CANCELED'
7| AND o.order_status <> 'SUSPECTED FRAUD'
8| GROUP BY order_item_product_id) r ON p.product_id = r.order_item_product_id
9| ORDER BY r.revenue DESC
10| LIMIT 10;
```

The results table shows the top 10 products by revenue, with the 5th row highlighted:

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 Meris Free 5.0+ Running Shoe	3507549.2067
5	502	Nike Men's Dri-FIT Victory Golf Polo	3011600

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

\$ 3,011,600

The screenshot shows the Hue interface for running Impala queries. The query window contains the same SQL code as the previous screenshot:

```
3| SELECT oi.order_item_product_id, SUM(CAST(oi.order_item_subtotal AS
4| FROM order_items oi INNER JOIN orders o
5| ON oi.order_item_order_id = o.order_id
6| WHERE o.order_status <> 'CANCELED'
7| AND o.order_status <> 'SUSPECTED FRAUD'
8| GROUP BY order_item_product_id) r ON p.product_id = r.order_item_product_id
9| ORDER BY r.revenue DESC
10| LIMIT 10;
```

The results table shows the top 10 products by revenue, with the 5th row highlighted:

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 Meris Free 5.0+ Running Shoe	3507549.2067
5	502	Nike Men's Dri-FIT Victory Golf Polo	3011600

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.

The screenshot shows the Hue search interface with the following details:

- Query History:** Shows a list of recent queries.
- Saved Queries:** Shows a list of saved queries.
- Results (152):** Shows the total number of results found.
- Search bar:** Contains the placeholder "Search data and saved documents...".
- Job status:** Shows "Jobs" with a count of 2, a progress bar, and a "cloudera" link.

Rank	Count	Path
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%20apparel/product/Nike%20Men%20Dri-FIT%20Victory%20Golf%20Polo
4	1757	/department/apparel/category/mens%20footwear/product/Nike%20Men%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%20outdoor%20games/product/O'Brien%20Men%20Neoprene%20Life%20Vest
7	1059	/department/fan%20shop/category/camping%20&%20hiking/product/Diamondback%20Women%20Serene%20Classic%20Comfort%20B
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%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%20PFG%20Anchor%20Tough%20T-Shirt
12	896	/department/fitness/category/tennis%20&%20racquet/product/Nike%20Men%20Comfort%202%20Slide
13	892	/department/footwear/category/as%20seen%20on%20tv/product/Nike%20Men%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%20Mens%20Tech%20II%20T-Shirt
16	859	/department/fitness/category/soccer/product/Nike%20Men%20Fingertrap%20Max%20Training%20Shoe
17	727	/department/apparel/category/featured%20shops/product/adidas%20Kids%20RG%20III%20Mid%20Football%20Cleat/add_to_cart