Big Data Hadoop and Spark Developer

Lesson-End Project Solution

Retail Business Analytics

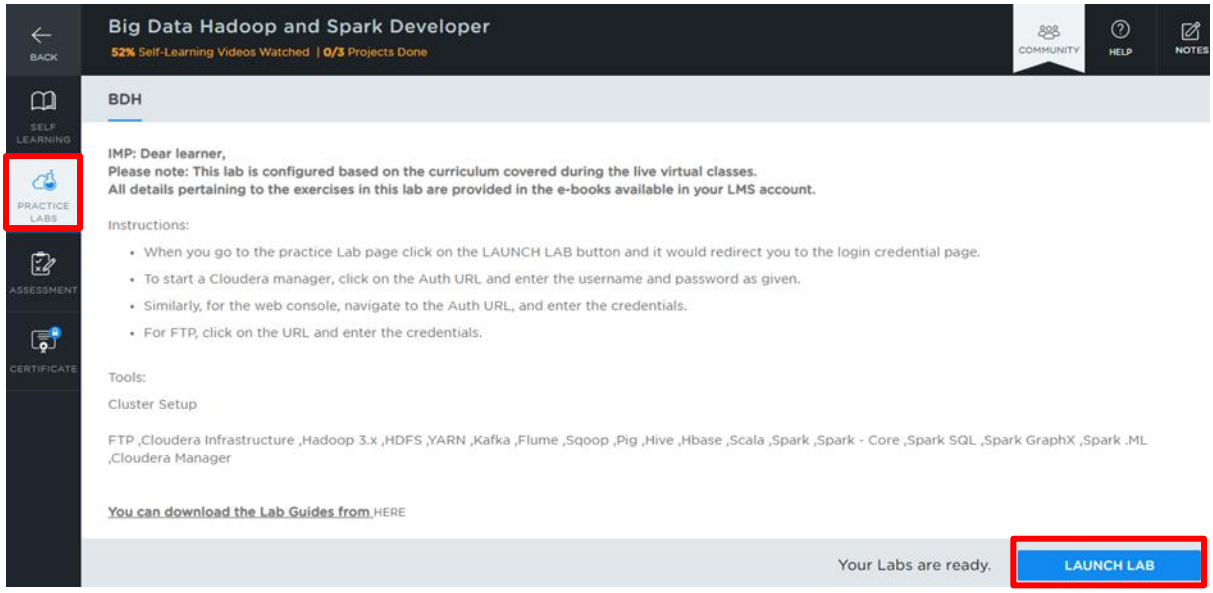
Steps to Perform:

Step 1: Log in to your LMS account

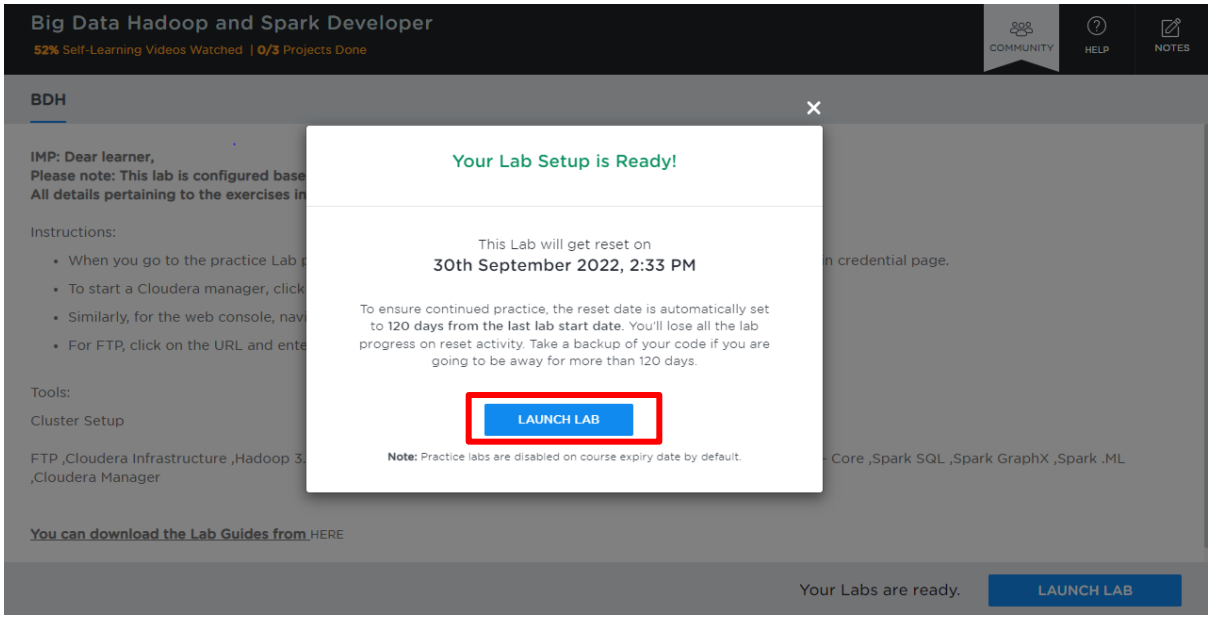
Step 2: Open the course "**Big Data Hadoop and Spark Developer**"

Step 3: Download the datasets from the "**Course Resources**" section

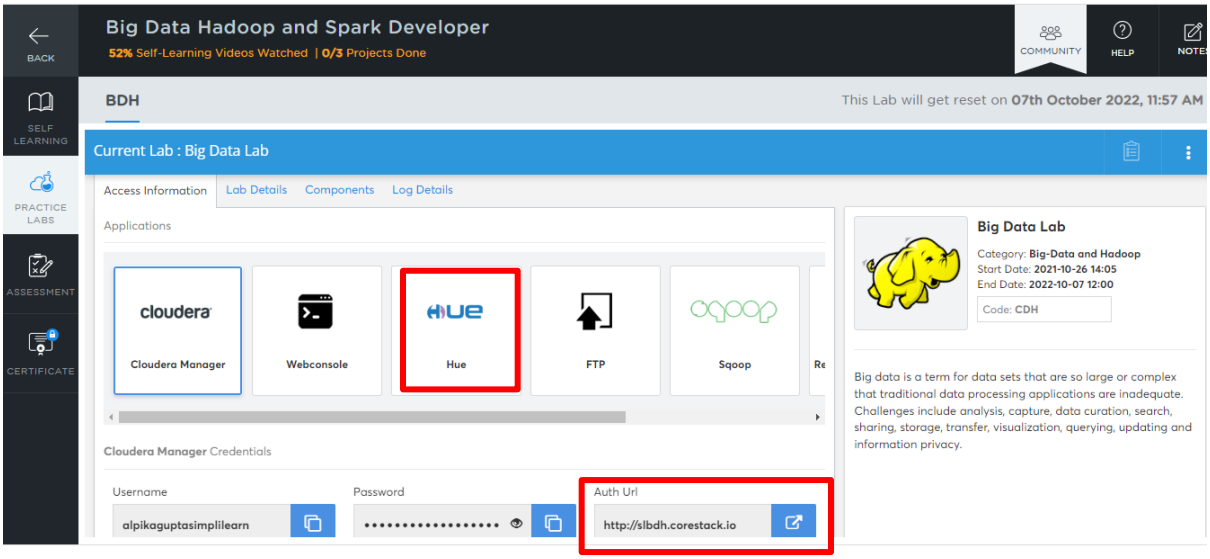
Step 4: On the left side, click on the "**PRACTICE LABS**" tab and click on the "**LAUNCH LAB**" button



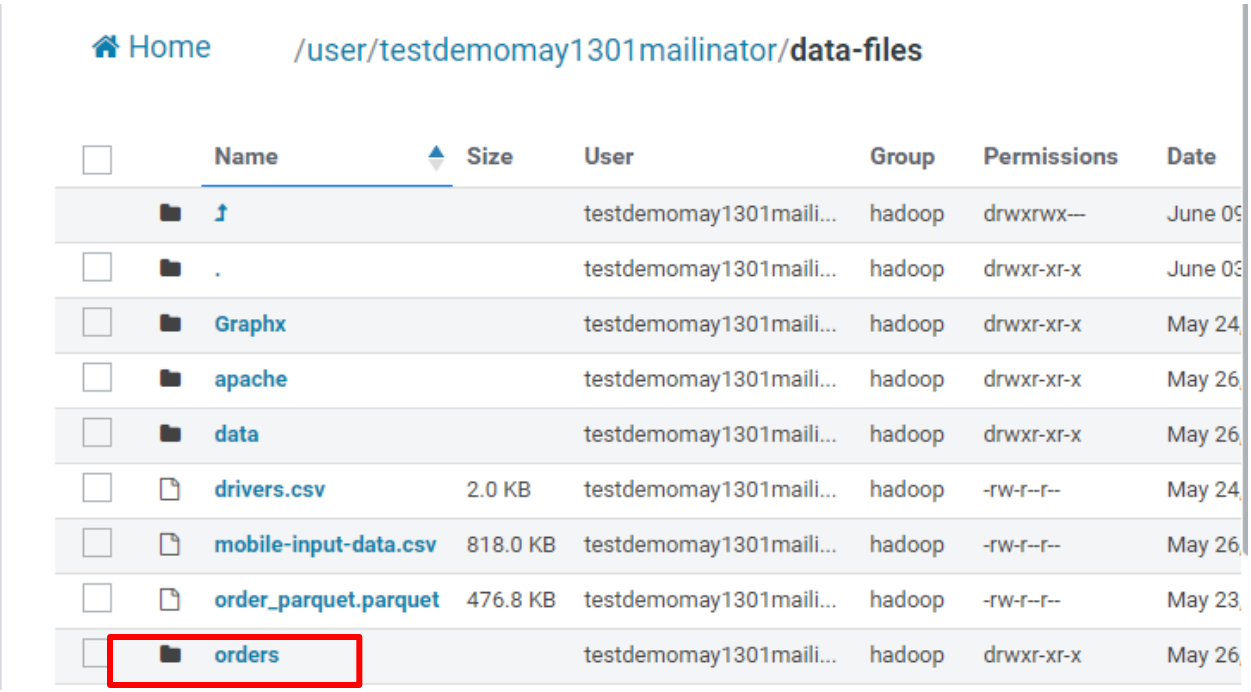
Step 5: Again, click on the "**LAUNCH LAB**" button



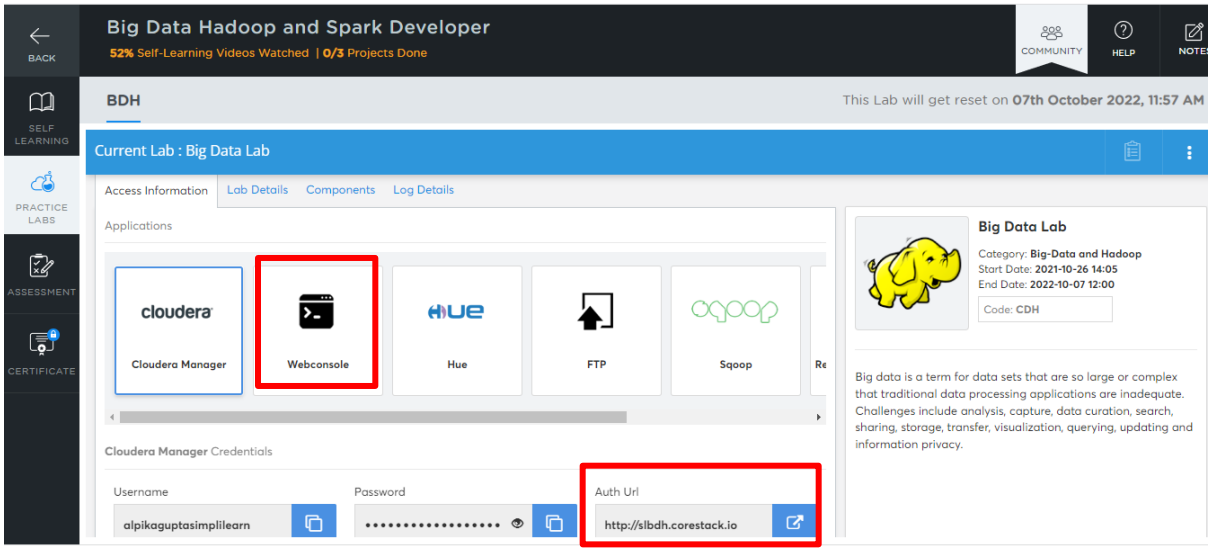
Step 6: Click on **"HUE"** to upload the datasets



Step 7: Log in to the "**HUE"** and click on create a directory named **"data-files"** and upload the **"orders"** dataset into it



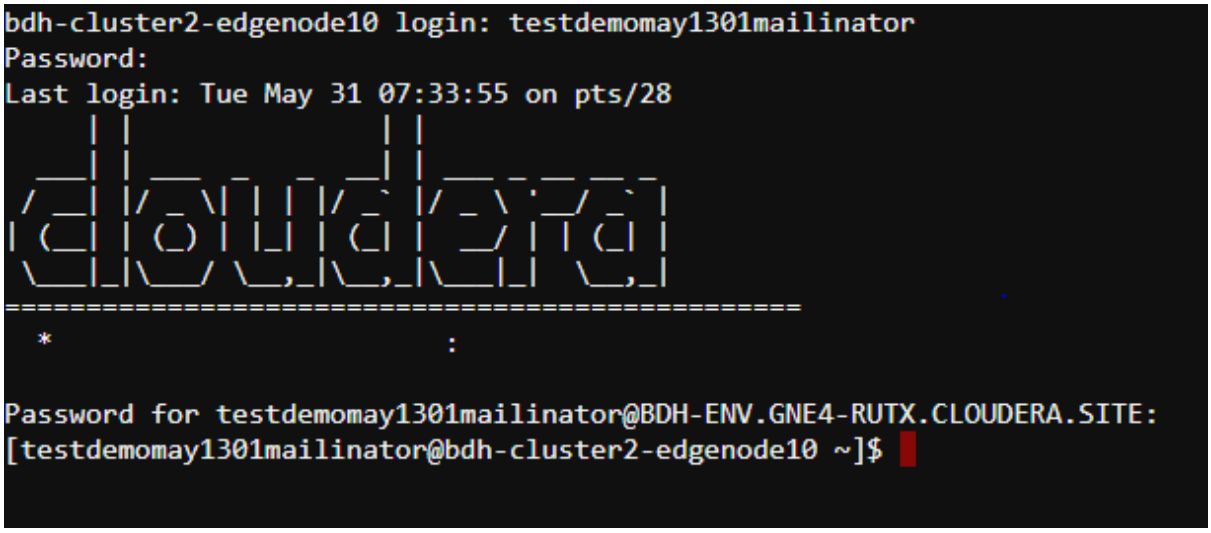
Step 8: Click on the "**Webconsole**" and click on the "**Auth Url**"



Step 9: Copy the "**Username**" and the "**Password**" provided to log in to the Web console

Step 10: Paste the "**Username**" and the "**Password**" on the console and click on Enter

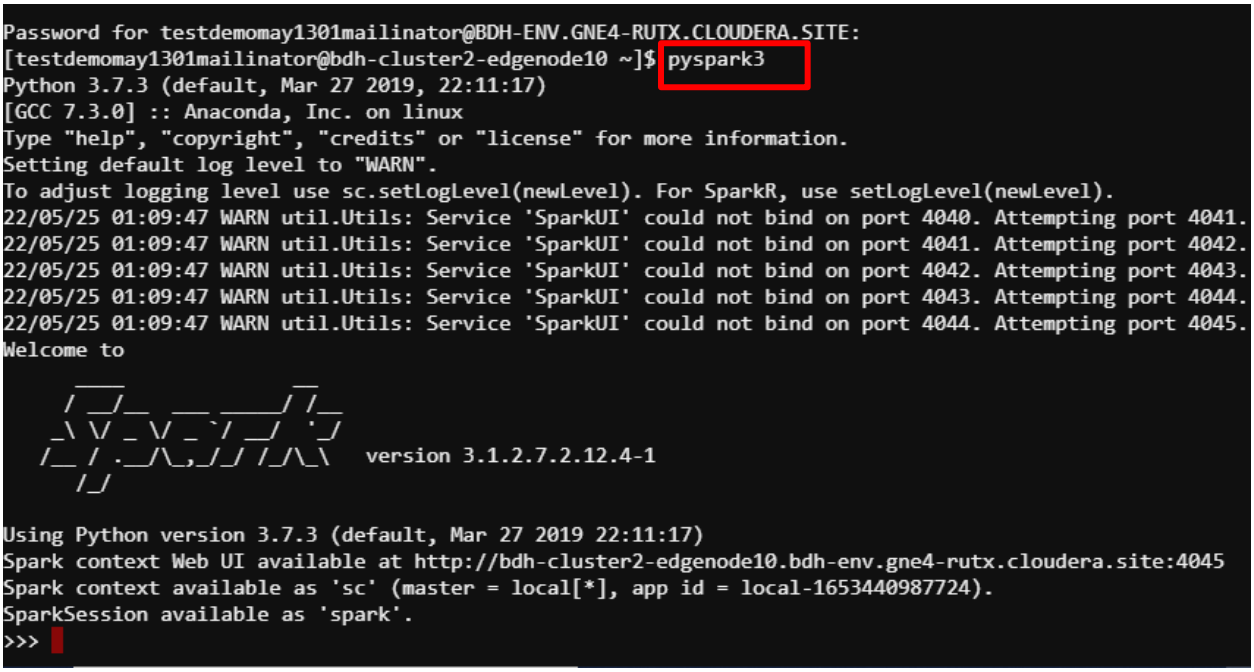
Note: The password will not be visible when pasted on the console.



Step 11: Log in to the PySpark shell

Command:

pyspark3



Step 12: Create a program in the PySpark shell

Program:

```
from pyspark.sql import SparkSession
spark: SparkSession = SparkSession \
    .builder \
```

```
    .appName("Scenario1") \
    .getOrCreate()

orders = spark.read.option("inferSchema", True) \
    .csv("/user/testdemomay1301mailinator/data-files/orders") \
    .toDF("order_id", "order_date", "order_customer_id", "order_status")

orders.createOrReplaceTempView("orders")

result = spark.sql("""
SELECT Substring(order_date, 1, 7) order_date,
order_status, Count(1) cnt
FROM   orders
WHERE  order_status = 'SUSPECTED_FRAUD'
GROUP  BY Substring(order_date, 1, 7), order_status
ORDER  BY order_date desc""").select("order_date", "cnt")

result.show(10)
```

Step 13: Now, you will be able to find the order_date with a number of counts

```
>>> orders = spark.read.option("inferSchema", True) \
...     .csv("/user/testdemomay1301mailinator/data-files/orders") \
...     .toDF("order_id", "order_date", "order_customer_id", "order_status")
>>>
>>> orders.createOrReplaceTempView("orders")
>>>
>>> result = spark.sql("""
... SELECT Substring(order_date, 1, 7) order_date,
... order_status, Count(1) cnt
... FROM   orders
... WHERE  order_status = 'SUSPECTED_FRAUD'
... GROUP  BY Substring(order_date, 1, 7), order_status
... ORDER  BY order_date desc""").select("order_date", "cnt")
>>>
>>> result.show(10)
+----------+---+
|order_date|cnt|
+----------+---+
|   2014-07|101|
|   2014-06|131|
|   2014-05|130|
|   2014-04|112|
|   2014-03|138|
|   2014-02|119|
|   2014-01|131|
|   2013-12|126|
|   2013-11|150|
|   2013-10|108|
+----------+---+
only showing top 10 rows
```