Big Data Hadoop and Spark Developer

Lesson-End Project Solution

Telecom Log Parsing

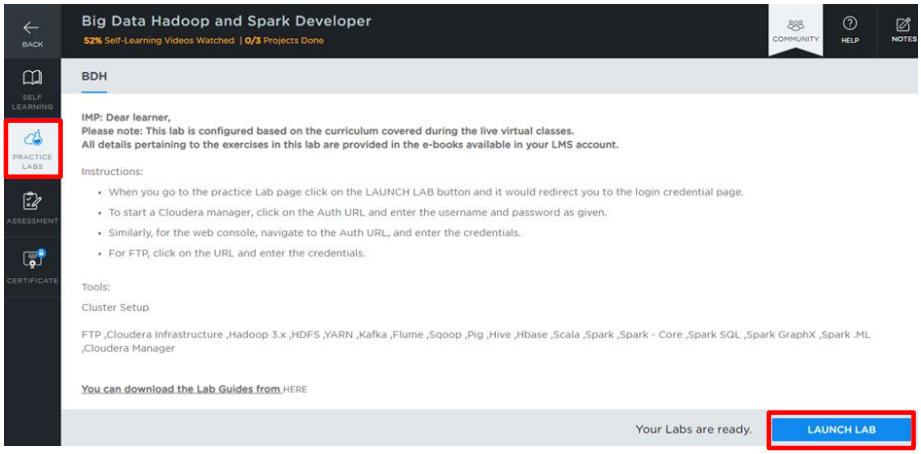
Steps to Perform:

Step 1: Log in to your LMS account

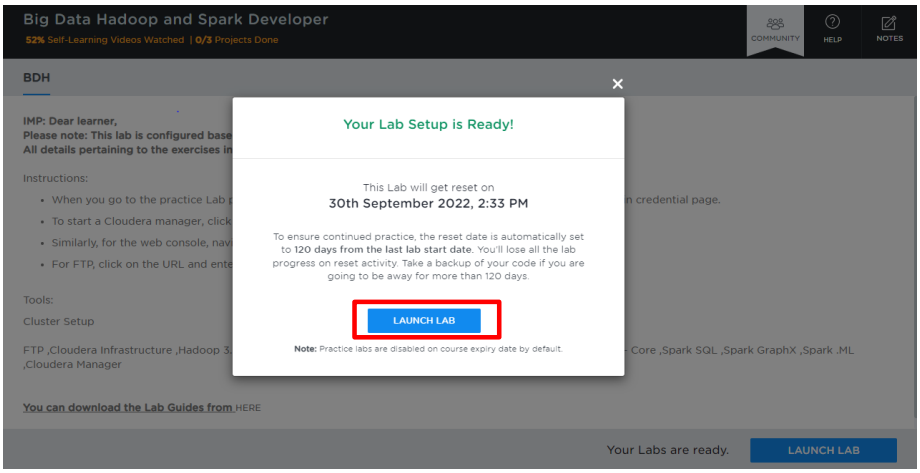
Step 2: Open the course "**Big Data Hadoop and Spark Developer**"

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

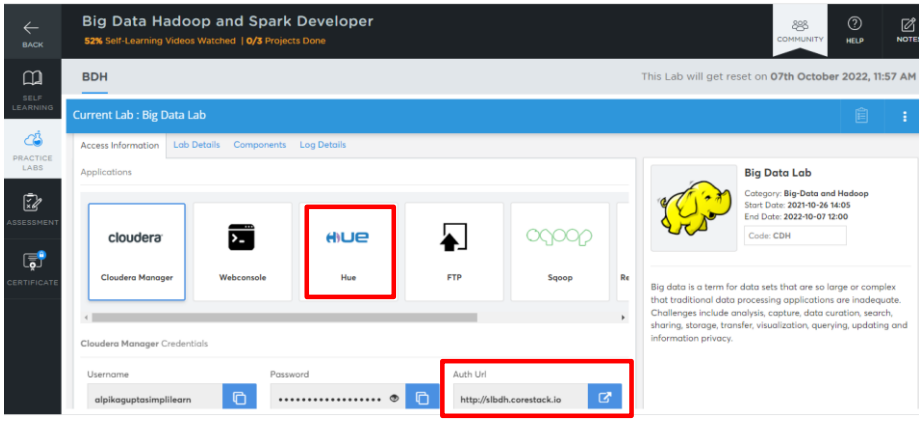
Step 4: Click on the "**PRACTICE LABS**" tab on the left side and select "**LAUNCH LAB**"



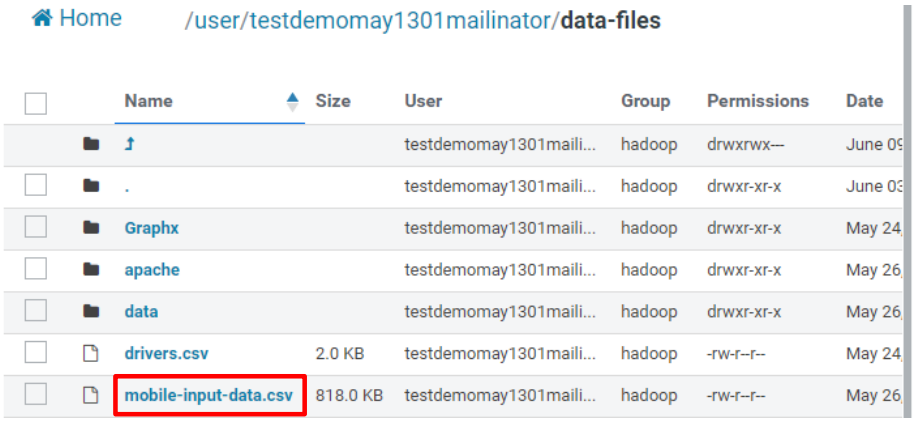
Step 5: Click on the "**LAUNCH LAB**" button



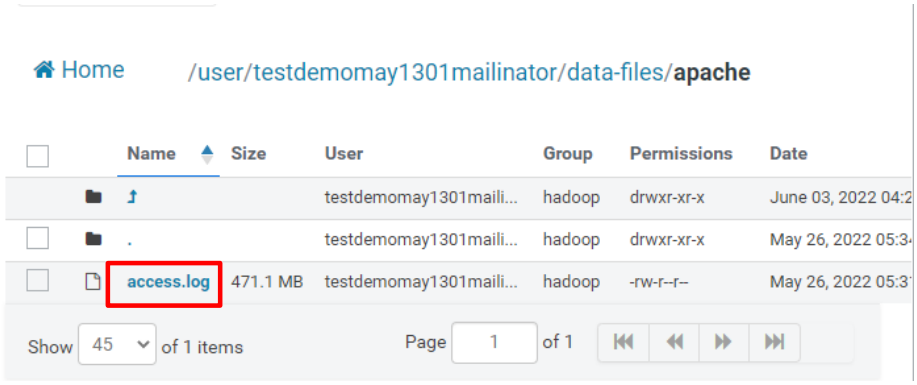
Step 6: Click on "**HUE**" to upload the datasets



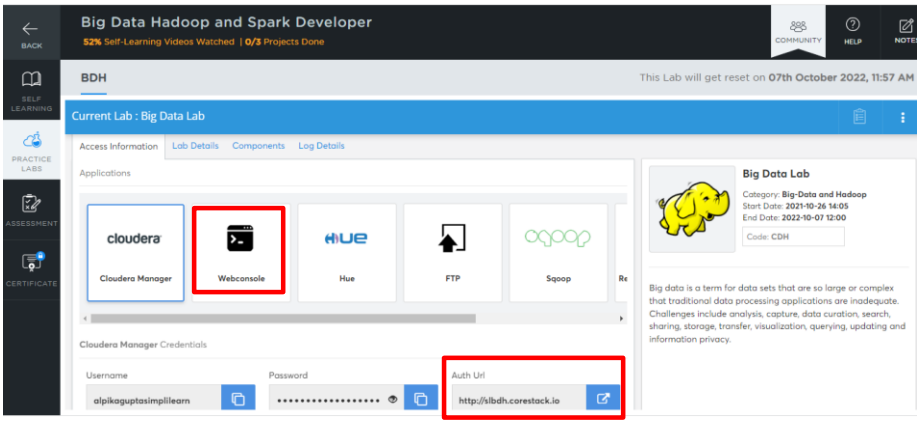
Step 7: Log in to **HUE** and create a directory named **"data-files"** and upload the CSV file into it



Step 8: Create a new directory in data-files directory named "**apache**" and upload the **"access.log"** file into the directory



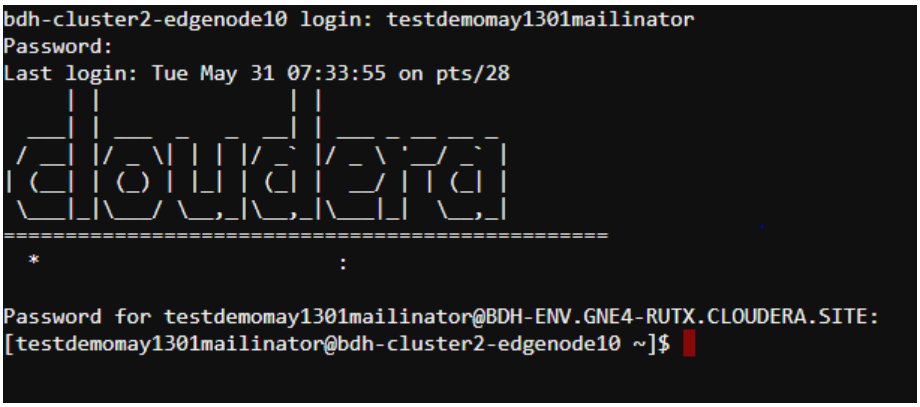
Step 9: Click on "**Webconsole**" and then on "**Auth Url**"



Step 10: Copy the "**Username**" and "**Password**" provided to log in to the **Webconsole**

Step 11: Paste the "**Username**" and "**Password**" on the console and click on Enter

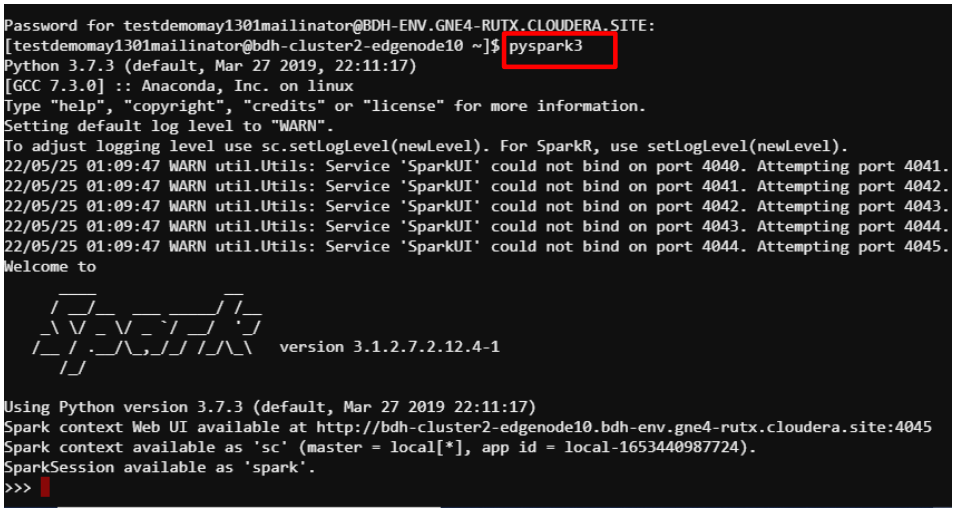
Note: The password will not be visible when pasted on the console.



Step 12: Log in to the PySpark shell

Command:

pyspark3



Step 13: Create an RDD to read the **"access.log"** data in PySpark Shell

Command:

Server_log_rdd = spark.sparkContext.textFile("/user/testdemomay1301mailinator/data-files/apache/access.log")

Step 14: Read the CSV file from the HDFS using the command below:

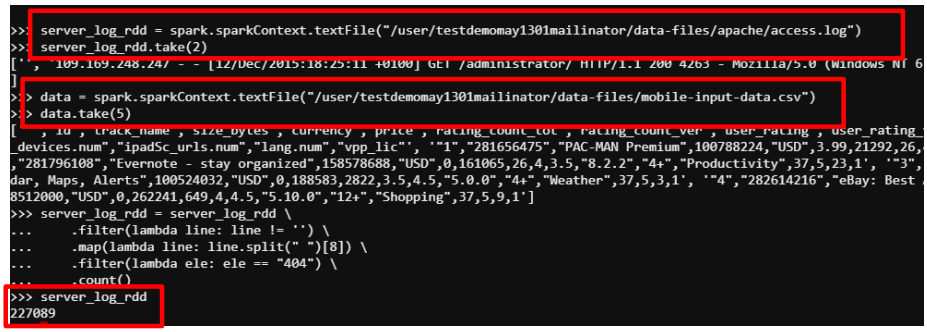
Command:

data = spark.sparkContext.textFile**(**("/user/testdemomay1301mailinator/data-files/mobile-input-data.csv")

Step 15: Create a lambda function to find the number of 404 HTTP codes present in **"access.log"**

Command:

```
server_log_rdd = server_log_rdd \
   .filter(lambda line: line != '') \
   .map(lambda line: line.split(" ")[8]) \
   .filter(lambda ele: ele == "404") \
   .count()
```


>>> server_log_rdd = spark.sparkContext.textFile("/user/testdemomay1301mailinator/data-files/apache/access.log")
>>> server_log_rdd.take(2)
>>> data = spark.sparkContext.textFile("/user/testdemomay1301mailinator/data-files/mobile-input-data.csv")
>>> data.take(5)
>>> server_log_rdd = server_log_rdd \
...   .filter(lambda line: line != '') \
...   .map(lambda line: line.split(" ")[8]) \
...   .filter(lambda ele: ele == "404") \
...   .count()
>>> server_log_rdd
227089