Assisted Practice 15.4: Create an RDD in Spark

Problem Scenario: Create an RDD with a real-world retail business dataset of different categories.

Objective: In this demonstration, you will read the data from HDFS and print the distinct categories.

Dataset Name: "part-m-00000"

Tasks to Perform:

1. Download the "part-m-00000" dataset from the categories folder from the course resource section and upload it into the HDFS using "Hue"
2. Login into the "webconsole" and open the PySpark shell
3. Create an RDD using textFile and update the path of the dataset
4. Create a lambda function that will split the line
5. Print each element using collect() method

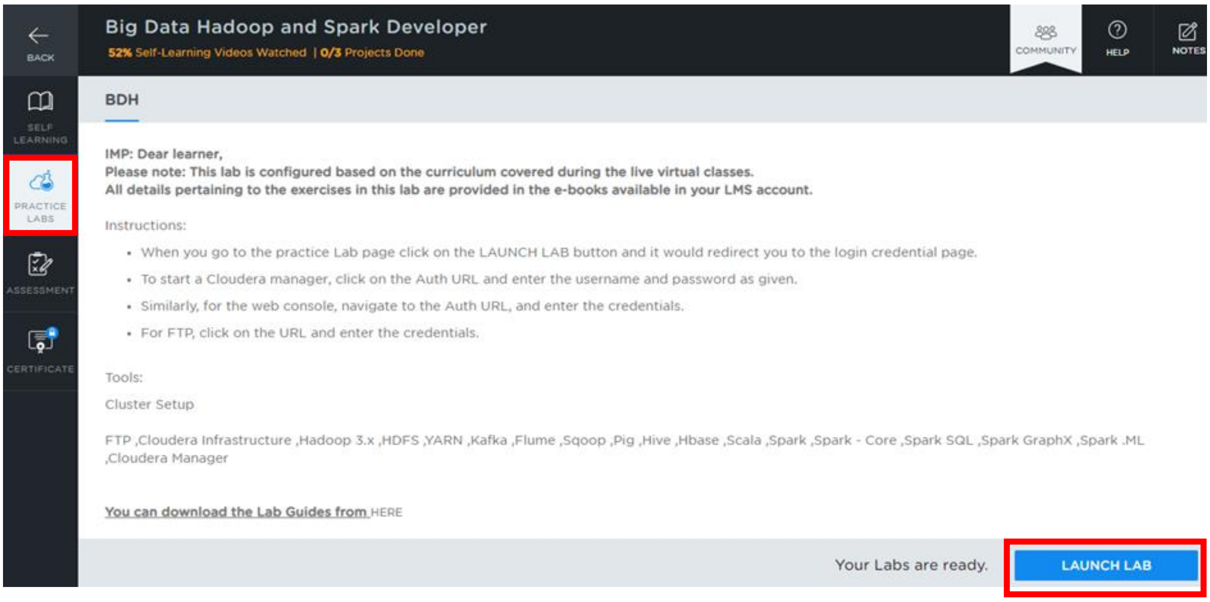
Steps to Perform:

Step 1: Download the dataset with the name **"part-m-00000"** from the course resources section.

Step 2: Log in to your LMS account

Step 3: Open the course **"Big Data Hadoop and Spark developer"**

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



Big Data Hadoop and Spark Developer

52% Self-Learning Videos Watched | 0/3 Projects Done

BDH

IMP: Dear learner,
Please note: This lab is configured based on the curriculum covered during the live virtual classes.
All details pertaining to the exercises in this lab are provided in the e-books available in your LMS account.

Instructions:

- When you go to the practice Lab page click on the LAUNCH LAB button and it would redirect you to the login credential page.
- To start a Cloudera manager, click on the Auth URL and enter the username and password as given.
- Similarly, for the web console, navigate to the Auth URL, and enter the credentials.
- For FTP, click on the URL and enter the credentials.

Tools:

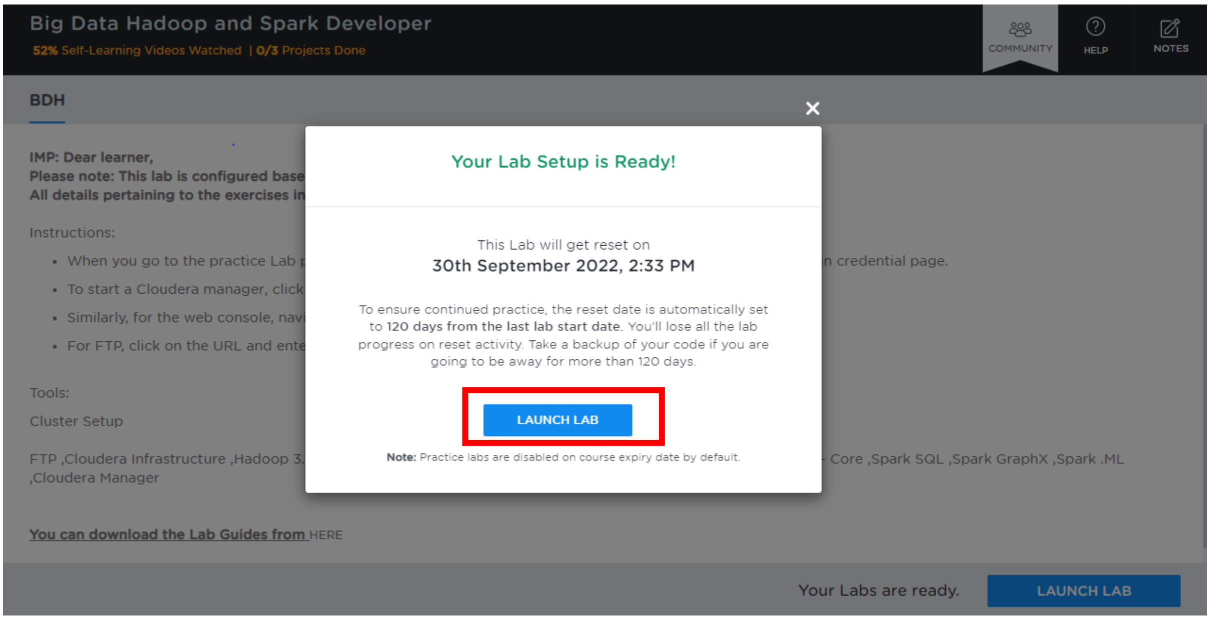
Cluster Setup

FTP ,Cloudera Infrastructure ,Hadoop 3.x ,HDFS ,YARN ,Kafka ,Flume ,Sqoop ,Pig ,Hive ,Hbase ,Scala ,Spark ,Spark - Core ,Spark SQL ,Spark GraphX ,Spark .ML ,Cloudera Manager

You can download the Lab Guides from HERE

Your Labs are ready. LAUNCH LAB

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



Your Lab Setup is Ready!

This Lab will get reset on
30th September 2022, 2:33 PM

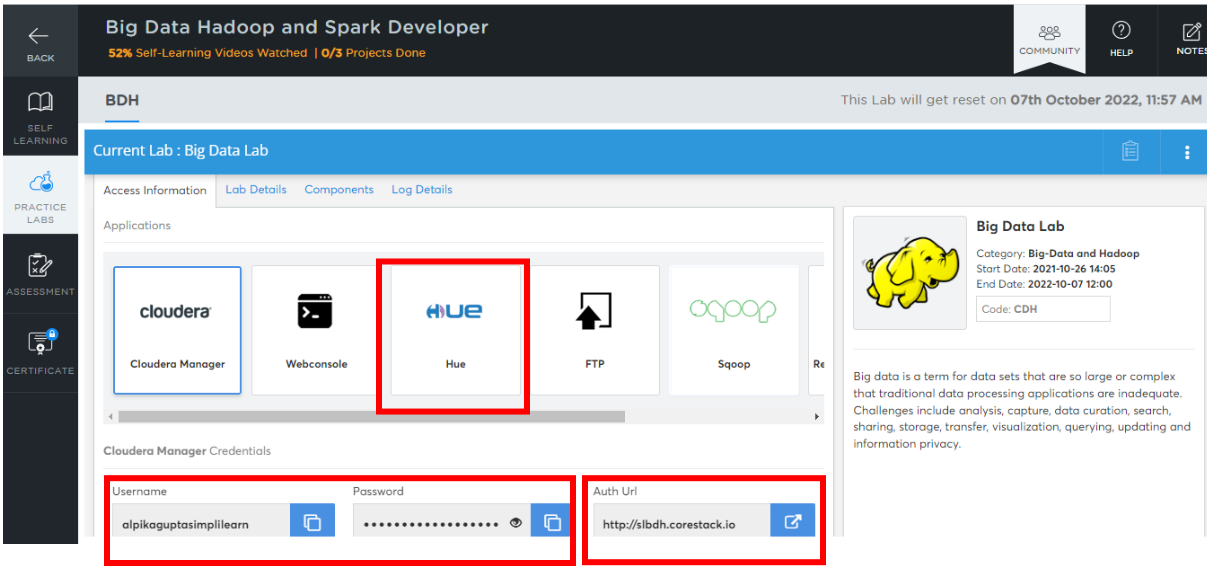
To ensure continued practice, the reset date is automatically set to **120 days from the last lab start date**. You'll lose all the lab progress on reset activity. Take a backup of your code if you are going to be away for more than 120 days.

LAUNCH LAB

Note: Practice labs are disabled on course expiry date by default.

Your Labs are ready. LAUNCH LAB

Step 6: Click on **"Hue"** and click on the **"Auth Url"** to upload the dataset and copy the "**Username**" and the "**Password**" provided to log in to the **"Hue"**



Big Data Hadoop and Spark Developer

52% Self-Learning Videos Watched | 0/3 Projects Done

BDH

This Lab will get reset on 07th October 2022, 11:57 AM

Current Lab : Big Data Lab

Cloudera Manager Credentials

Username: alpikaguptasimplilearn

Password: ••••••••••••••••••

Auth Url: http://slbdh.corestack.io

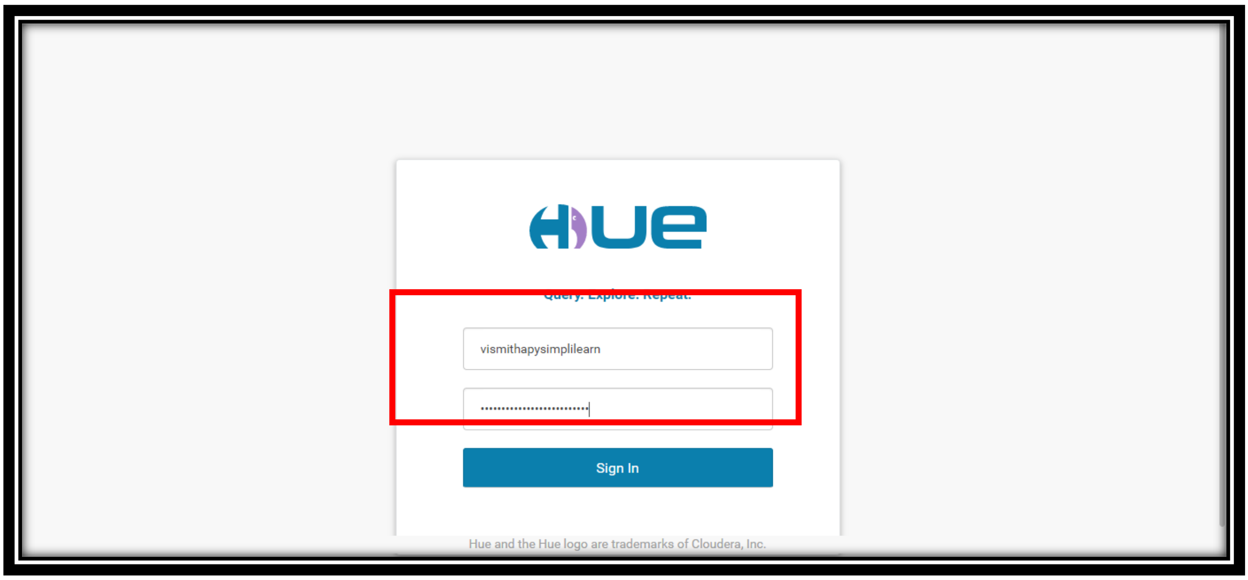
Big Data Lab

Category: Big-Data and Hadoop
Start Date: 2021-10-26 14:05
End Date: 2022-10-07 12:00

Code: CDH

Big data is a term for data sets that are so large or complex that traditional data processing applications are inadequate. Challenges include analysis, capture, data curation, search, sharing, storage, transfer, visualization, querying, updating and information privacy.

Step 7: Paste the **"Username"** and the **"Password"** on the login window and click on sign in

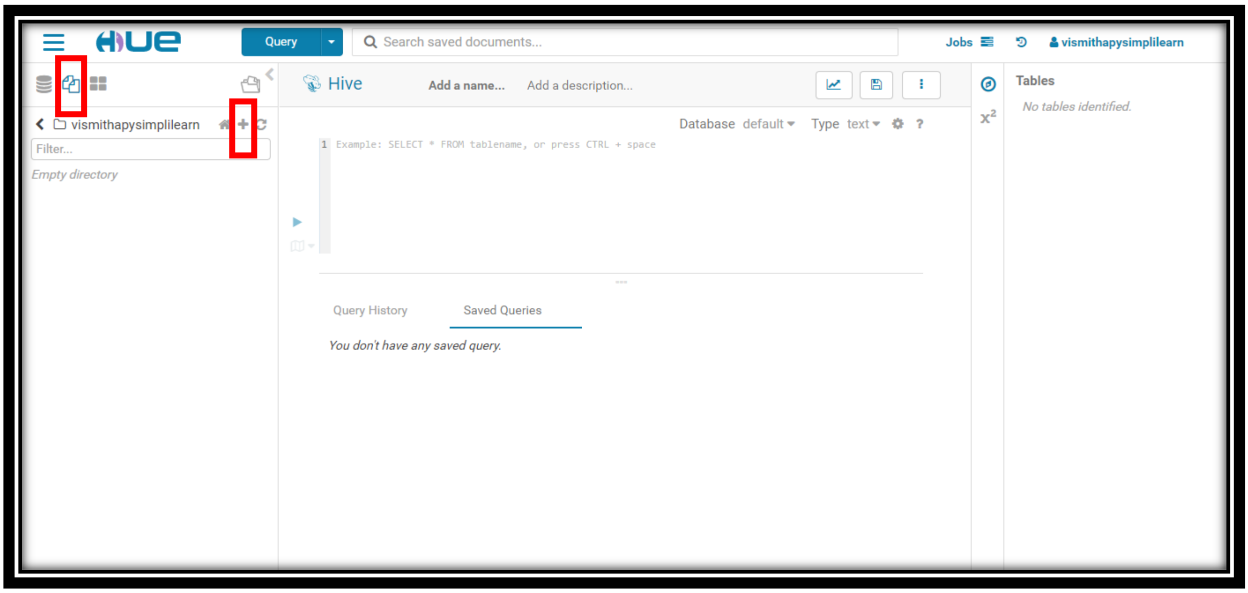


vismithapysimplilearn

••••••••••••••••••••••••••

Hue and the Hue logo are trademarks of Cloudera, Inc.

Step 8: Click on **"HDFS"** icon and click on the **"+"** symbol to upload the dataset



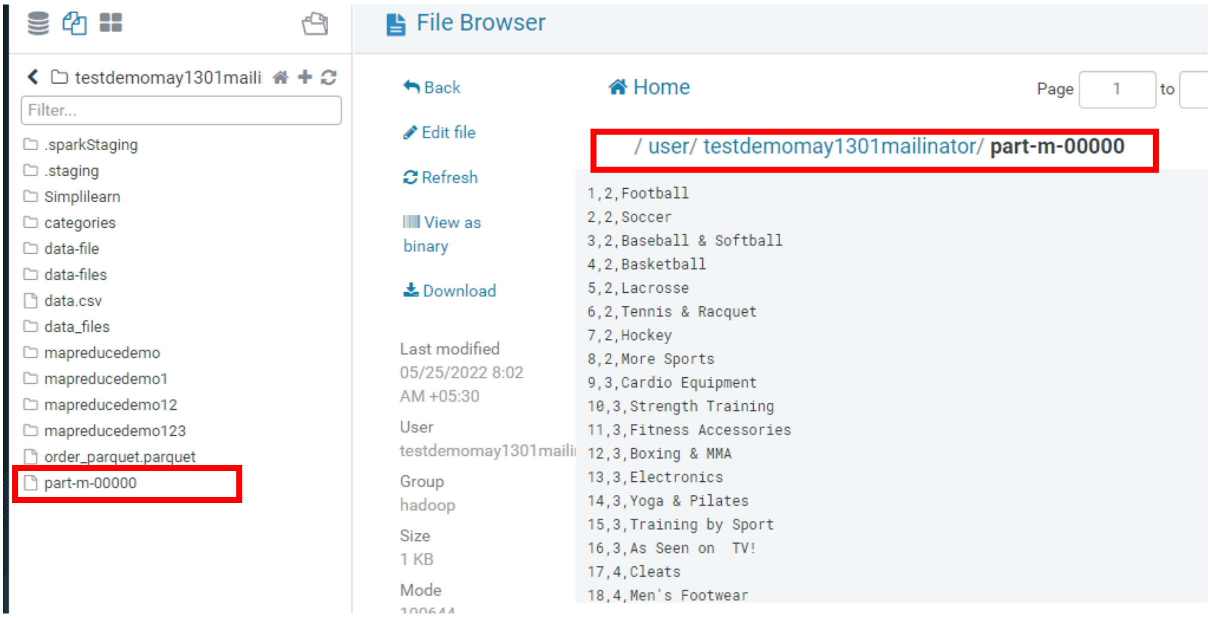
No tables identified.

Empty directory

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

You don't have any saved query.

Step 9: Select the downloaded dataset file and upload it to **"HDFS."** In addition, by right-clicking, copy the path from the dataset that has been uploaded.



/ user/ testdemomay1301mailinator/ **part-m-00000**

Last modified
05/25/2022 8:02 AM +05:30

User
testdemomay1301maili

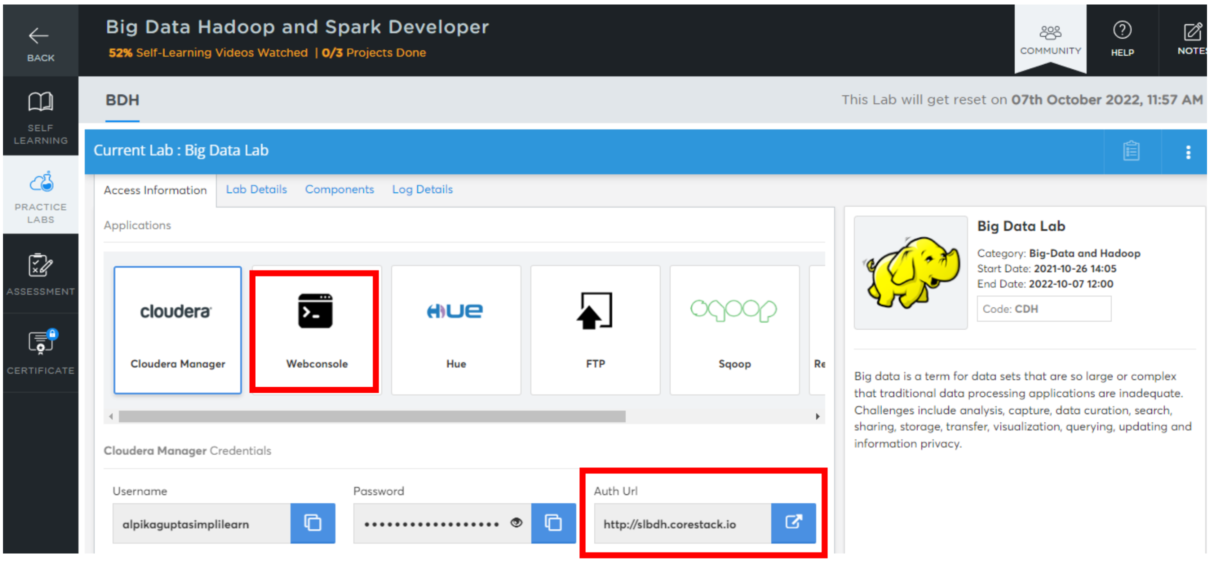
Group
hadoop

Size
1 KB

Mode
100644

```
1,2,Football
2,2,Soccer
3,2,Baseball & Softball
4,2,Basketball
5,2,Lacrosse
6,2,Tennis & Racquet
7,2,Hockey
8,2,More Sports
9,3,Cardio Equipment
10,3,Strength Training
11,3,Fitness Accessories
12,3,Boxing & MMA
13,3,Electronics
14,3,Yoga & Pilates
15,3,Training by Sport
16,3,As Seen on  TV!
17,4,Cleats
18,4,Men's Footwear
```

Step 10: Go back to the lab window and click on "**Webconsole**" and click on the **"Auth Url"**



Big Data Hadoop and Spark Developer

52% Self-Learning Videos Watched | 0/3 Projects Done

This Lab will get reset on 07th October 2022, 11:57 AM

Current Lab : Big Data Lab

Big Data Lab

Category: Big-Data and Hadoop
Start Date: 2021-10-26 14:05
End Date: 2022-10-07 12:00
Code: CDH

Big data is a term for data sets that are so large or complex that traditional data processing applications are inadequate. Challenges include analysis, capture, data curation, search, sharing, storage, transfer, visualization, querying, updating and information privacy.

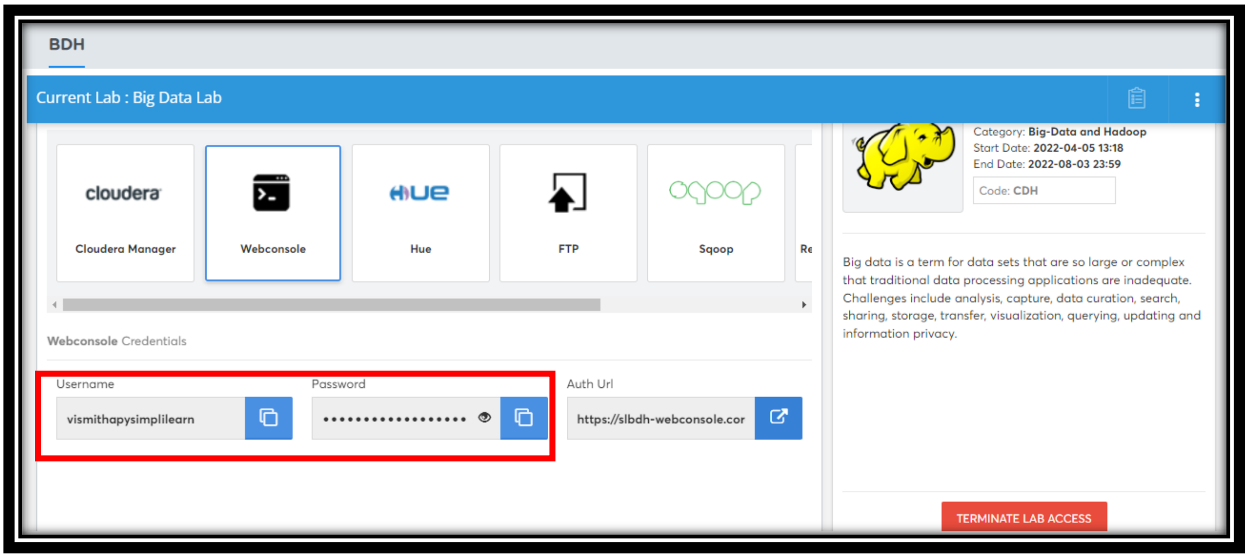
Cloudera Manager Credentials

Username: alpikaguptasimplilearn

Password: ••••••••••••••••••

Auth Url: http://slbdh.corestack.io

Step 11: Copy the "**Username**" and the "**Password**" provided to log in to the "**Webconsole**"



Current Lab : Big Data Lab

Category: Big-Data and Hadoop
Start Date: 2022-04-05 13:18
End Date: 2022-08-03 23:59
Code: CDH

Big data is a term for data sets that are so large or complex that traditional data processing applications are inadequate. Challenges include analysis, capture, data curation, search, sharing, storage, transfer, visualization, querying, updating and information privacy.

Webconsole Credentials

Username: vismithapysimplilearn

Password: ••••••••••••••••••

Auth Url: https://slbdh-webconsole.cor

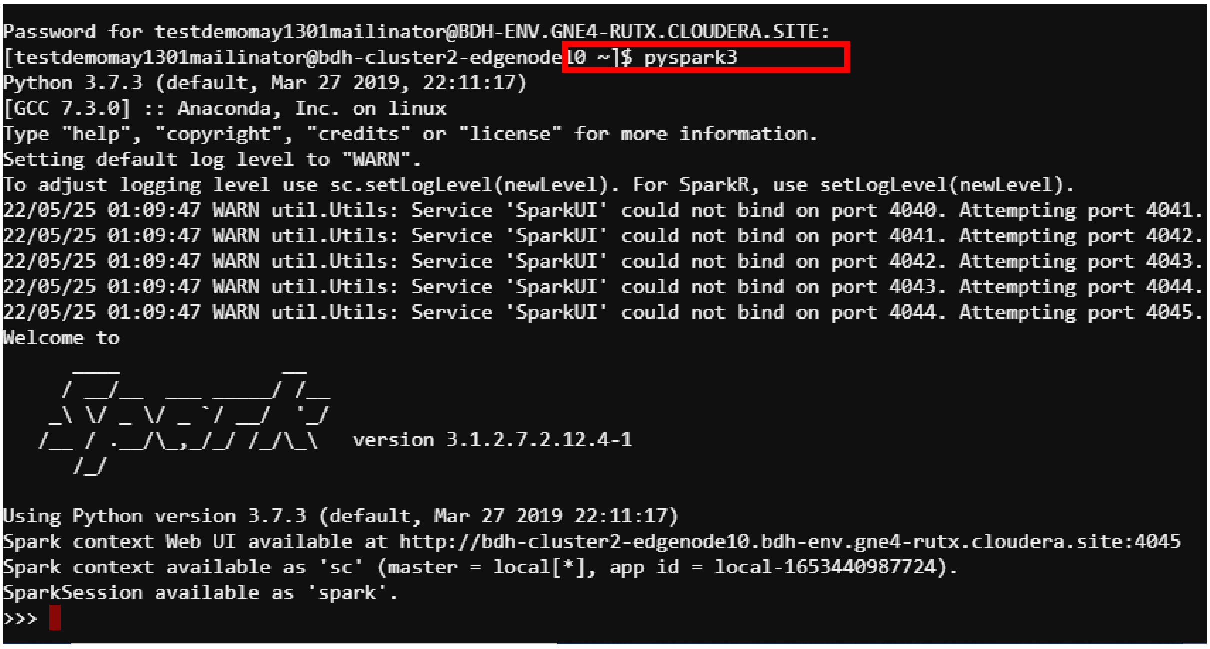
Step 12: Paste the "**Username**" and the "**Password**" on the console and click on enter.

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

Step 13: Enter the **"PySpark"** console by running the below command.

Command:

pyspark3



```
Password for testdemomay1301mailinator@BDH-ENV.GNE4-RUTX.CLOUDERA.SITE:
[testdemomay1301mailinator@bdh-cluster2-edgenode10 ~]$ pyspark3
Python 3.7.3 (default, Mar 27 2019, 22:11:17)
[GCC 7.3.0] :: Anaconda, Inc. on linux
Type "help", "copyright", "credits" or "license" for more information.
Setting default log level to "WARN".
To adjust logging level use sc.setLogLevel(newLevel). For SparkR, use setLogLevel(newLevel).
22/05/25 01:09:47 WARN util.Utils: Service 'SparkUI' could not bind on port 4040. Attempting port 4041.
22/05/25 01:09:47 WARN util.Utils: Service 'SparkUI' could not bind on port 4041. Attempting port 4042.
22/05/25 01:09:47 WARN util.Utils: Service 'SparkUI' could not bind on port 4042. Attempting port 4043.
22/05/25 01:09:47 WARN util.Utils: Service 'SparkUI' could not bind on port 4043. Attempting port 4044.
22/05/25 01:09:47 WARN util.Utils: Service 'SparkUI' could not bind on port 4044. Attempting port 4045.
Welcome to
      ____              __
     / __/__  ___ _____/ /__
    _\ \/ _ \/ _ `/ __/  '_/
   /__ / .__/\_,_/_/ /_/\_\   version 3.1.2.7.2.12.4-1
      /_/

Using Python version 3.7.3 (default, Mar 27 2019 22:11:17)
Spark context Web UI available at http://bdh-cluster2-edgenode10.bdh-env.gne4-rutx.cloudera.site:4045
Spark context available as 'sc' (master = local[*], app id = local-1653440987724).
SparkSession available as 'spark'.
>>>
```

Step 14: Create an RDD using textFile and update the path of the dataset.

Command:

catRDD = spark.sparkContext.textFile("/ user/testdemomay1301mailinator/ part-m-00000 ")

Step 15: Create a lambda function that will split the line.

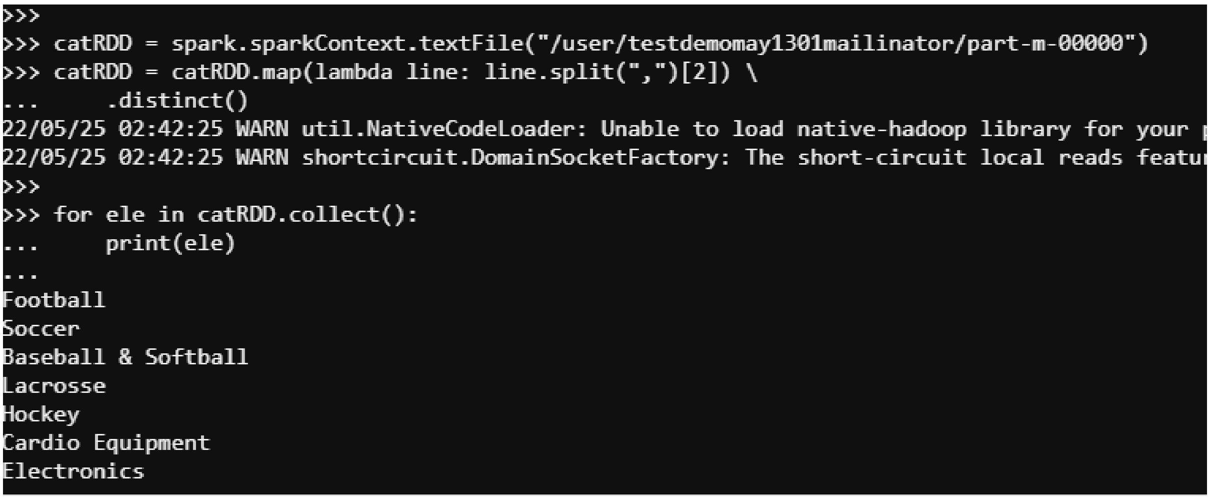
Command:

catRDD = catRDD.map(lambda line: line.split(",")[2]) \.distinct()

Step 16: Print each element using the collect() method. You will see the output as shown below:

Command:

```
for ele in catRDD.collect():
    print(ele)
```



```
>>>
>>> catRDD = spark.sparkContext.textFile("/user/testdemomay1301mailinator/part-m-00000")
>>> catRDD = catRDD.map(lambda line: line.split(",")[2]) \
...     .distinct()
22/05/25 02:42:25 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your p
22/05/25 02:42:25 WARN shortcircuit.DomainSocketFactory: The short-circuit local reads featu
>>>
>>> for ele in catRDD.collect():
...     print(ele)
...
Football
Soccer
Baseball & Softball
Lacrosse
Hockey
Cardio Equipment
Electronics
```