Assisted Practice 17: Data Exploration

Problem Scenario: Perform a data exploration and a descriptive analysis on the US companies' dataset.

Objective: In this demonstration, you will explore different commands to perform data exploration and descriptive analysis in PySpark.

Dataset Name: "Fortune 500 Companies US.csv"

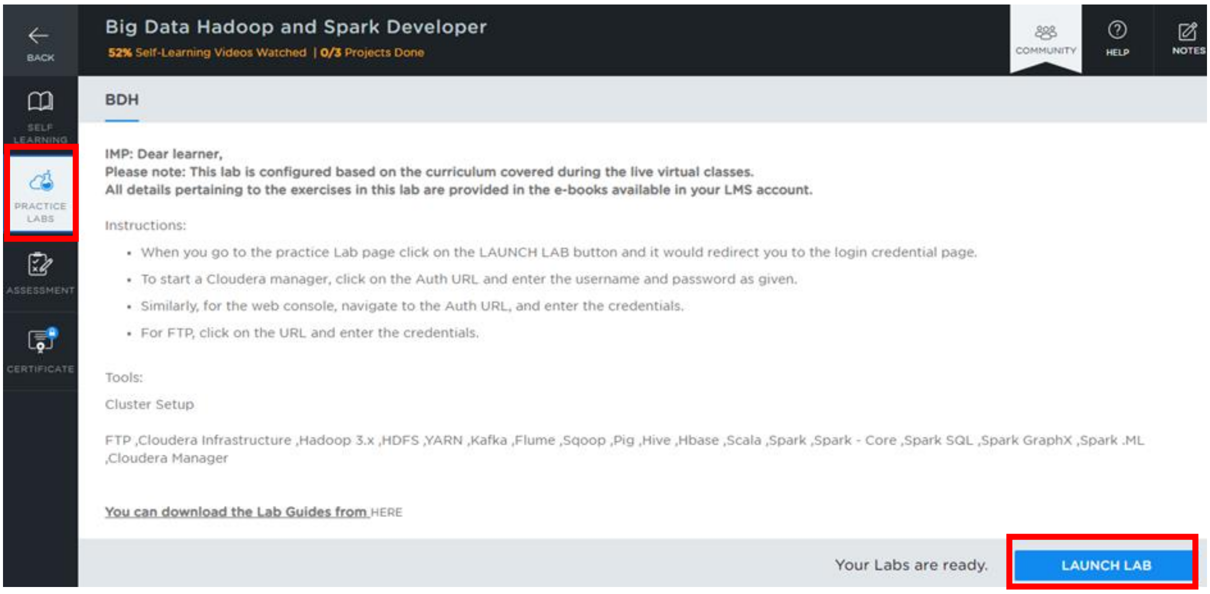
Steps to Perform:

Step 1: Download the dataset named **"Fortune 500 Companies US.csv"** 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 then 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:

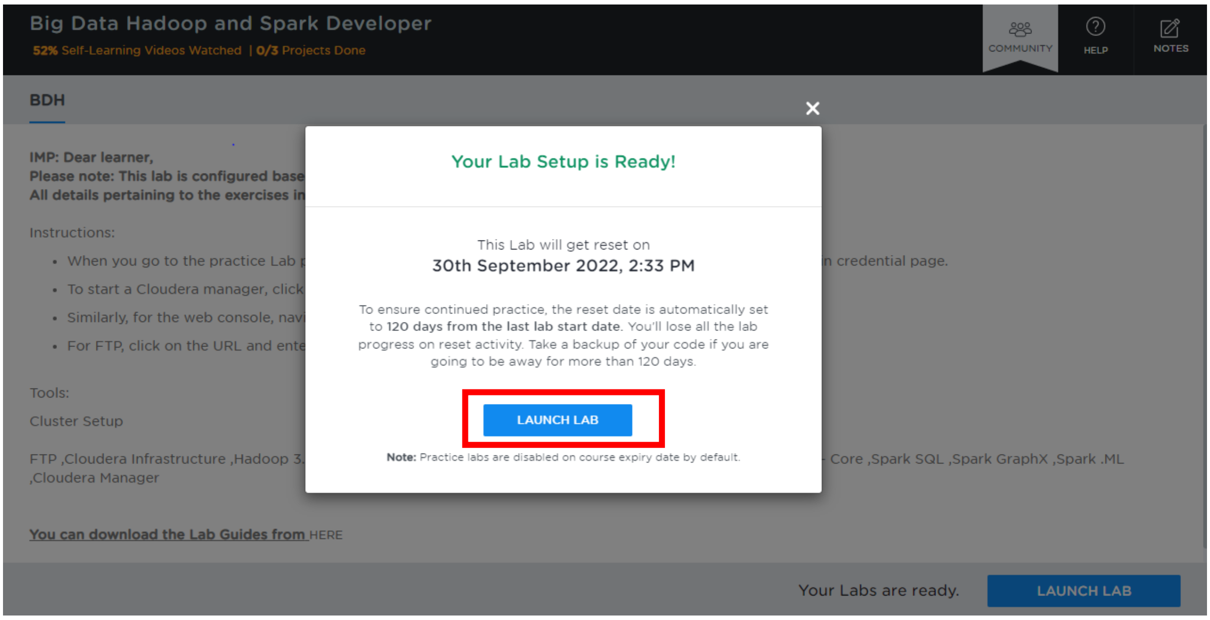
Cluster Setup

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.

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



Big Data Hadoop and Spark Developer

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

IMP: Dear learner,
Please note: This lab is configured base
All details pertaining to the exercises in

Instructions:

- When you go to the practice Lab p … n credential page.
- To start a Cloudera manager, click
- Similarly, for the web console, navi
- For FTP, click on the URL and ente

Tools:

Cluster Setup

FTP ,Cloudera Infrastructure ,Hadoop 3. … Core ,Spark SQL ,Spark GraphX ,Spark .ML ,Cloudera Manager

You can download the Lab Guides from HERE

Your Lab Setup is Ready!

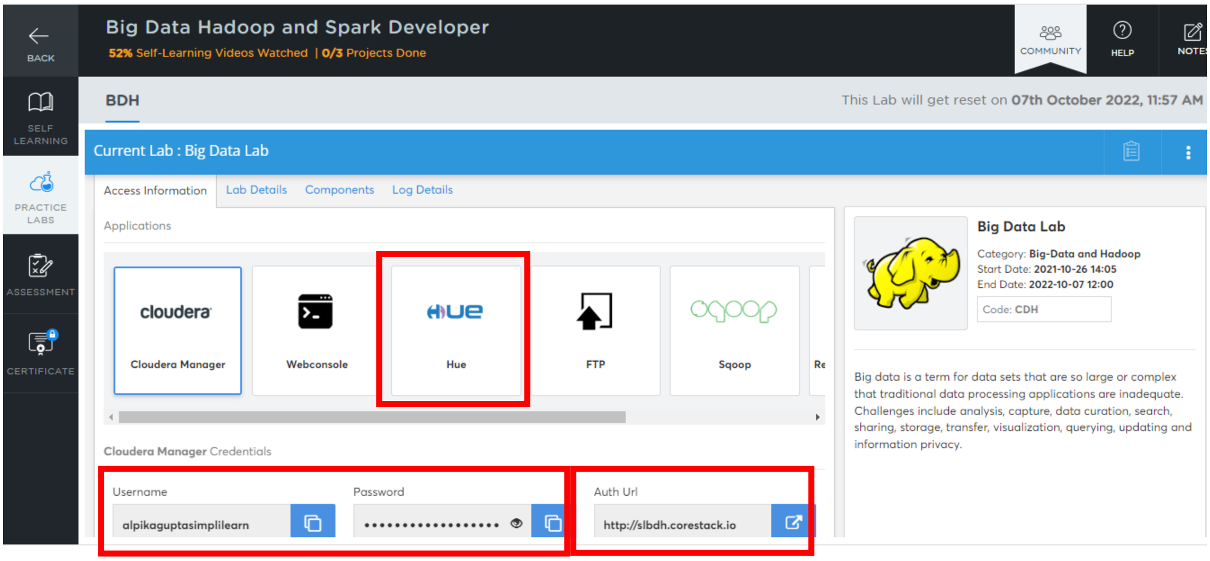
This Lab will get reset on
30th September 2022, 2:33 PM

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.

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

Your Labs are ready.

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

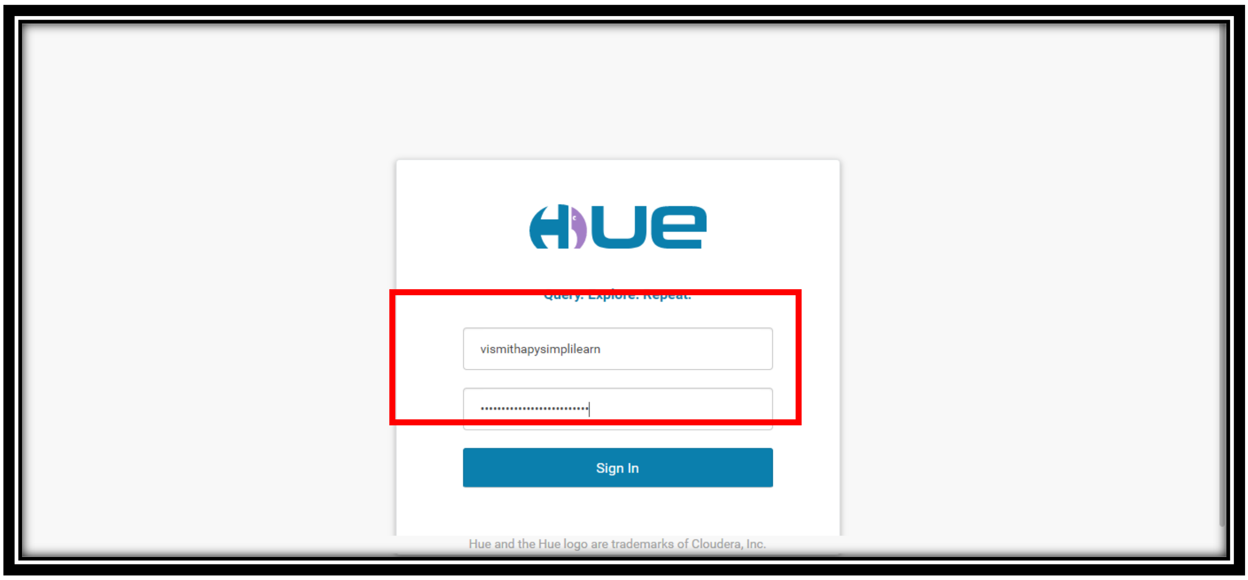
Big Data Lab

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.

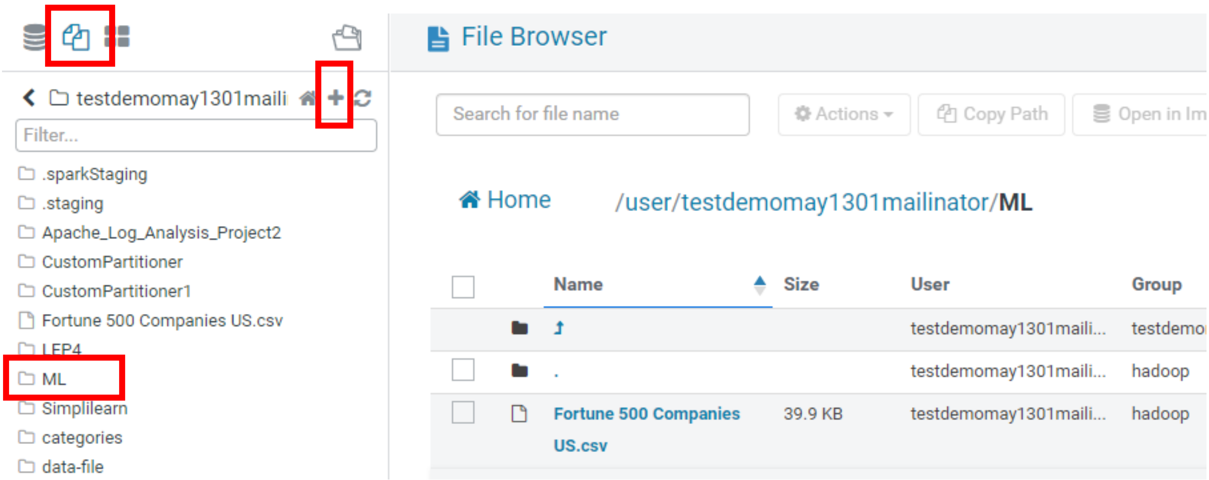
Step 7: Paste the **"Username"** and the **"Password"** on the log in window and click on **"Sign In"**



vismithapysimplilearn

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

Step 8: Create a directory named **“ML”** and click on the **“HDFS”** icon and then on the **“+”** symbol to upload the dataset



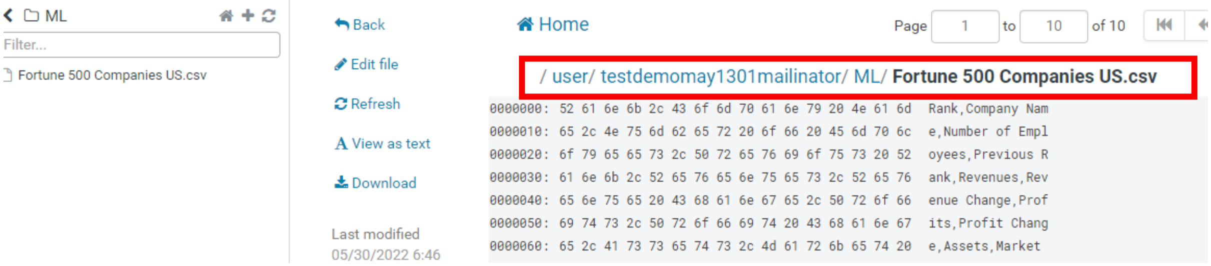
testdemomay1301maili

- .sparkStaging
- .staging
- Apache_Log_Analysis_Project2
- CustomPartitioner
- CustomPartitioner1
- Fortune 500 Companies US.csv
- LEP4
- ML
- Simplilearn
- categories
- data-file

Home /user/testdemomay1301mailinator/**ML**

		Name	Size	User	Group
		↥		testdemomay1301maili...	testdemo
□		.		testdemomay1301maili...	hadoop
□		Fortune 500 Companies US.csv	39.9 KB	testdemomay1301maili...	hadoop

Step 9: Copy the path of the dataset that has been uploaded

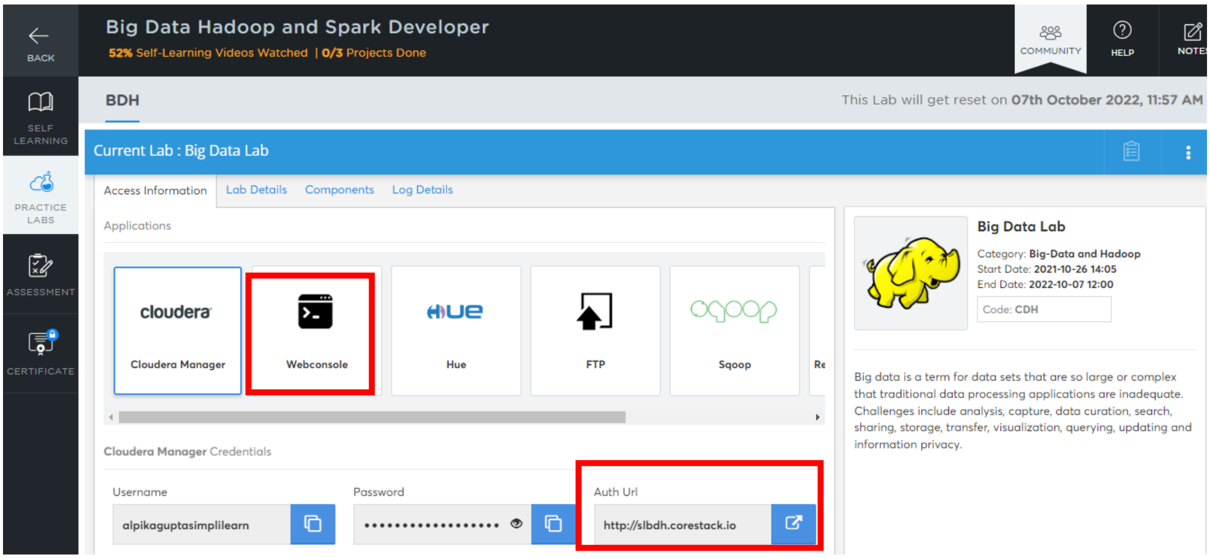


Last modified 05/30/2022 6:46

/ user/ testdemomay1301mailinator/ ML/ **Fortune 500 Companies US.csv**

```
0000000: 52 61 6e 6b 2c 43 6f 6d 70 61 6e 79 20 4e 61 6d  Rank,Company Nam
0000010: 65 2c 4e 75 6d 62 65 72 20 6f 66 20 45 6d 70 6c  e,Number of Empl
0000020: 6f 79 65 65 73 2c 50 72 65 76 69 6f 75 73 20 52  oyees,Previous R
0000030: 61 6e 6b 2c 52 65 76 65 6e 75 65 73 2c 52 65 76  ank,Revenues,Rev
0000040: 65 6e 75 65 20 43 68 61 6e 67 65 2c 50 72 6f 66  enue Change,Prof
0000050: 69 74 73 2c 50 72 6f 66 69 74 20 43 68 61 6e 67  its,Profit Chang
0000060: 65 2c 41 73 73 65 74 73 2c 4d 61 72 6b 65 74 20  e,Assets,Market
```

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



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: ••••••••••••••••••

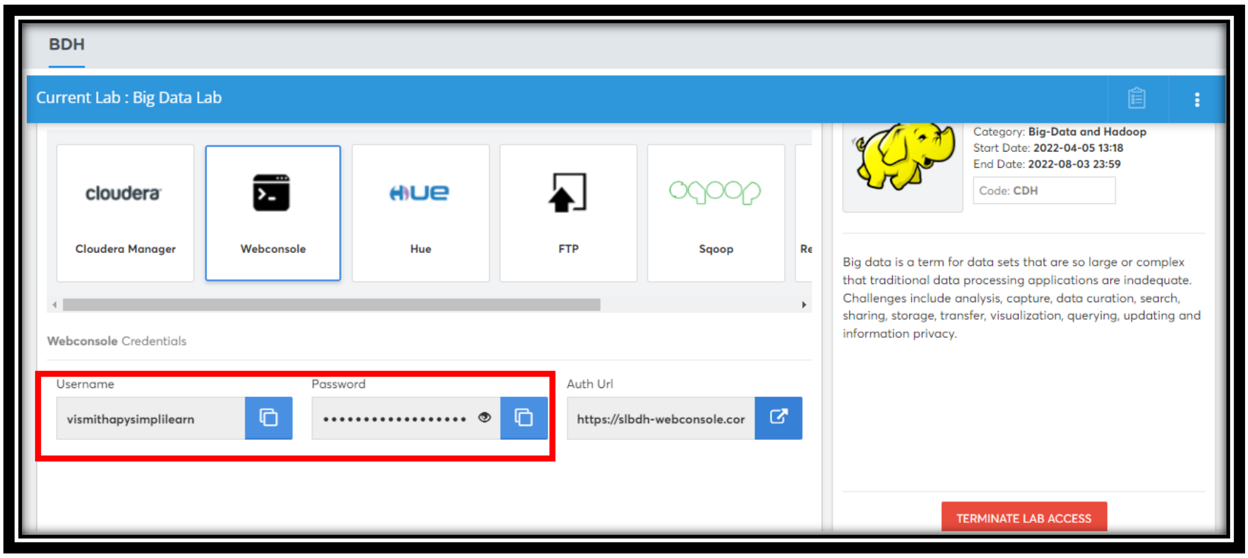
Auth Url: http://slbdh.corestack.io

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 11: Copy the “**Username**” and the “**Password**” provided to log in to the **“Webconsole”**



BDH

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

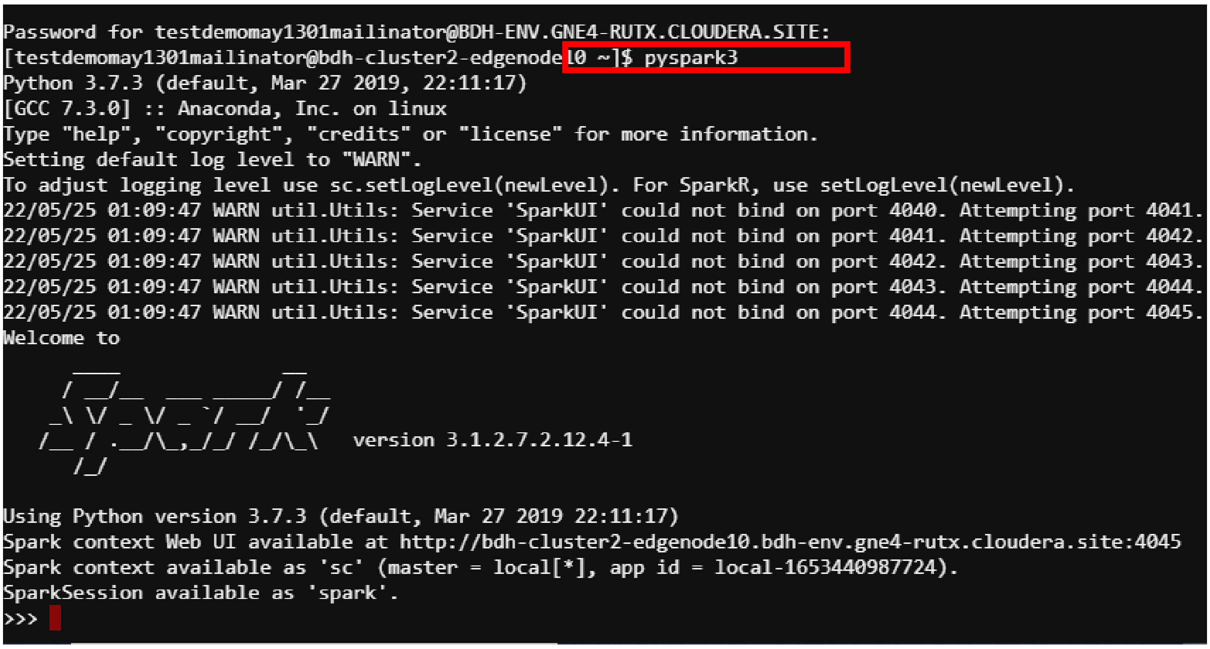
Step 12: Paste the “**Username**” and the “**Password**” on the console and click “Enter”

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: Import the necessary modules

Command:

from pyspark import SparkConf, SparkContext

from pyspark.sql import SQLContext

Step 15: Create a Spark Session, and then create a DataFrame from a CSV file to load data

Note: The path should be provided to the ML folder.

Command:

sc = SparkContext = SparkSession \

.builder \

.appName("Simlilearn Examples") \

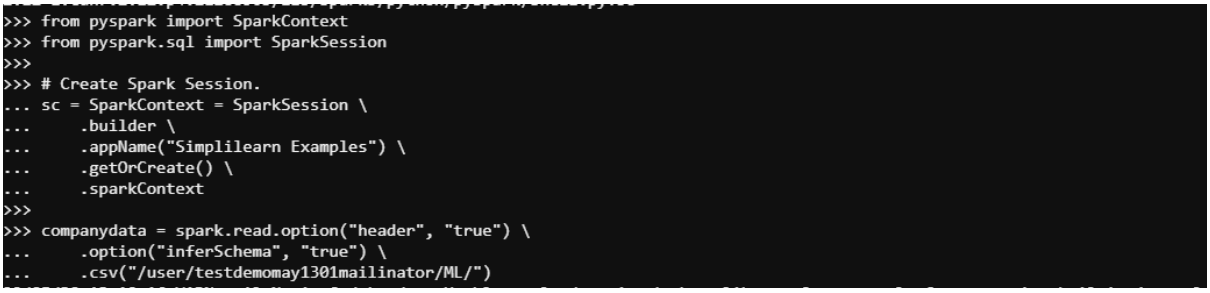
```
        .getOrCreate() \

        .sparkContext

companydata = spark.read.option(“header”, “true”) \

                         .option(“inferSchema”,”true”) \

                         .csv(“/user/testdemomay1301mailinator/ML”)
```

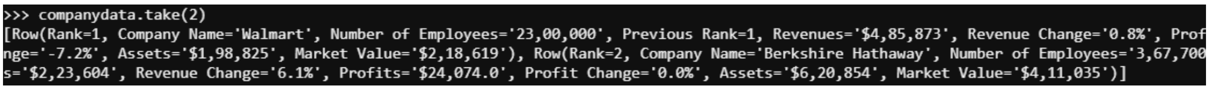


```
>>> from pyspark import SparkContext
>>> from pyspark.sql import SparkSession
>>>
>>> # Create Spark Session.
... sc = SparkContext = SparkSession \
...     .builder \
...     .appName("Simplilearn Examples") \
...     .getOrCreate() \
...     .sparkContext
>>>
>>> companydata = spark.read.option("header", "true") \
...     .option("inferSchema", "true") \
...     .csv("/user/testdemomay1301mailinator/ML/")
```

Step 16: View the loaded data using the below command:

Command:

```
companydata.take(2)
```



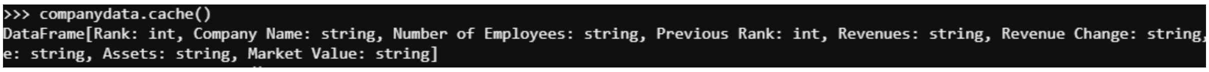
```
>>> companydata.take(2)
[Row(Rank=1, Company Name='Walmart', Number of Employees='23,00,000', Previous Rank=1, Revenues='$4,85,873', Revenue Change='0.8%', Prof
nge='-7.2%', Assets='$1,98,825', Market Value='$2,18,619'), Row(Rank=2, Company Name='Berkshire Hathaway', Number of Employees='3,67,700
s='$2,23,604', Revenue Change='6.1%', Profits='$24,074.0', Profit Change='0.0%', Assets='$6,20,854', Market Value='$4,11,035')]
```

Step 17: To check the data type of every column of a DataFrame and to print the schema of the DataFrame in a tree format, you can use the below commands:

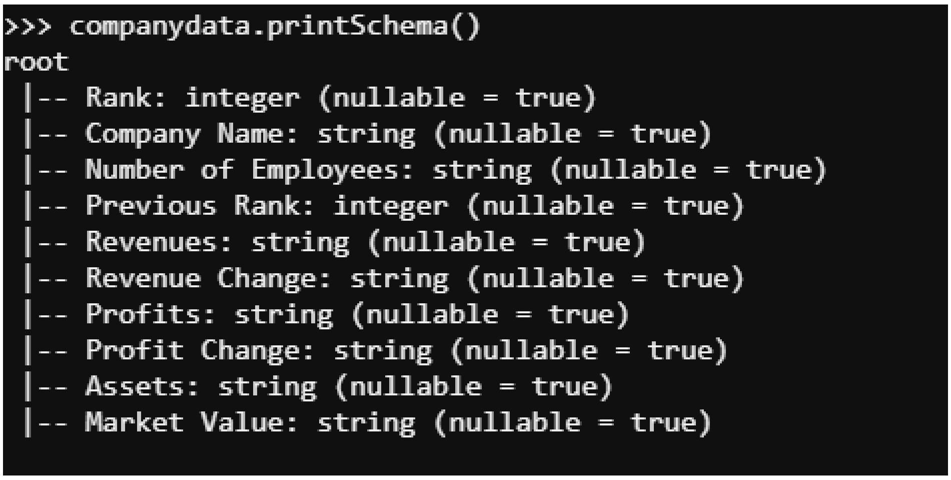
Command:

```
companydata.cache()

companydata.printSchema()
```



```
>>> companydata.cache()
DataFrame[Rank: int, Company Name: string, Number of Employees: string, Previous Rank: int, Revenues: string, Revenue Change: string,
e: string, Assets: string, Market Value: string]
```

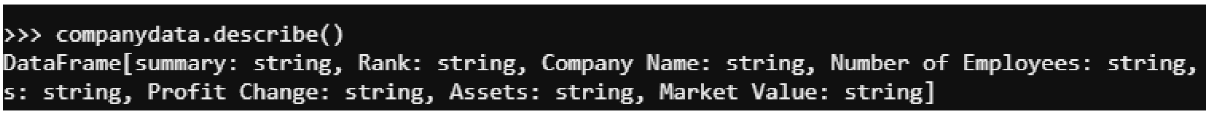


```
>>> companydata.printSchema()
root
 |-- Rank: integer (nullable = true)
 |-- Company Name: string (nullable = true)
 |-- Number of Employees: string (nullable = true)
 |-- Previous Rank: integer (nullable = true)
 |-- Revenues: string (nullable = true)
 |-- Revenue Change: string (nullable = true)
 |-- Profits: string (nullable = true)
 |-- Profit Change: string (nullable = true)
 |-- Assets: string (nullable = true)
 |-- Market Value: string (nullable = true)
```

Step 18: To perform a descriptive analysis of the company data you will use the below command:

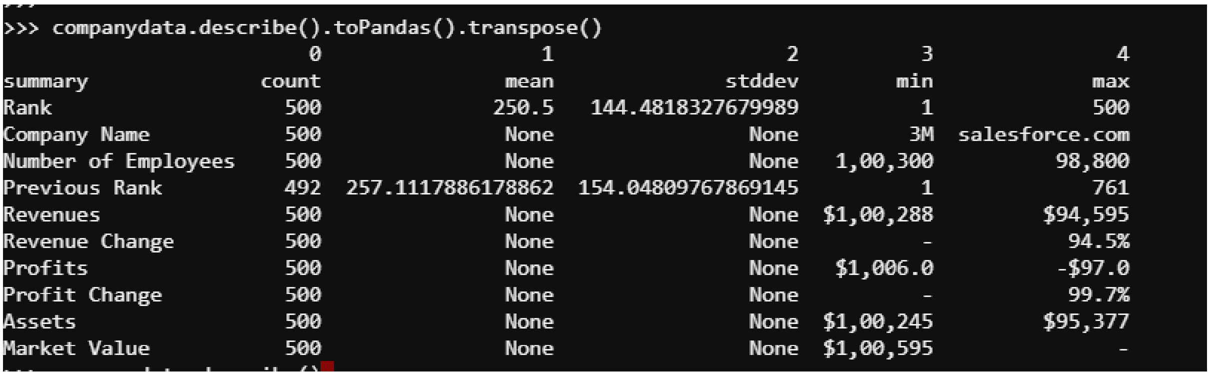
Command:

companydata.describe()



```
>>> companydata.describe()
DataFrame[summary: string, Rank: string, Company Name: string, Number of Employees: string,
s: string, Profit Change: string, Assets: string, Market Value: string]
```

companydata.describe().toPandas().transpose()



```
>>> companydata.describe().toPandas().transpose()
                         0                  1                   2          3               4
summary              count               mean              stddev        min             max
Rank                   500              250.5    144.4818327679989          1             500
Company Name           500               None                None         3M  salesforce.com
Number of Employees    500               None                None   1,00,300          98,800
Previous Rank          492  257.1117886178862  154.04809767869145          1             761
Revenues               500               None                None  $1,00,288         $94,595
Revenue Change         500               None                None          -           94.5%
Profits                500               None                None   $1,006.0          -$97.0
Profit Change          500               None                None          -           99.7%
Assets                 500               None                None  $1,00,245         $95,377
Market Value           500               None                None  $1,00,595               -
```