

---

# Big Data Hadoop and Spark Developer

Lab Guide



Get Certified. Get Ahead.

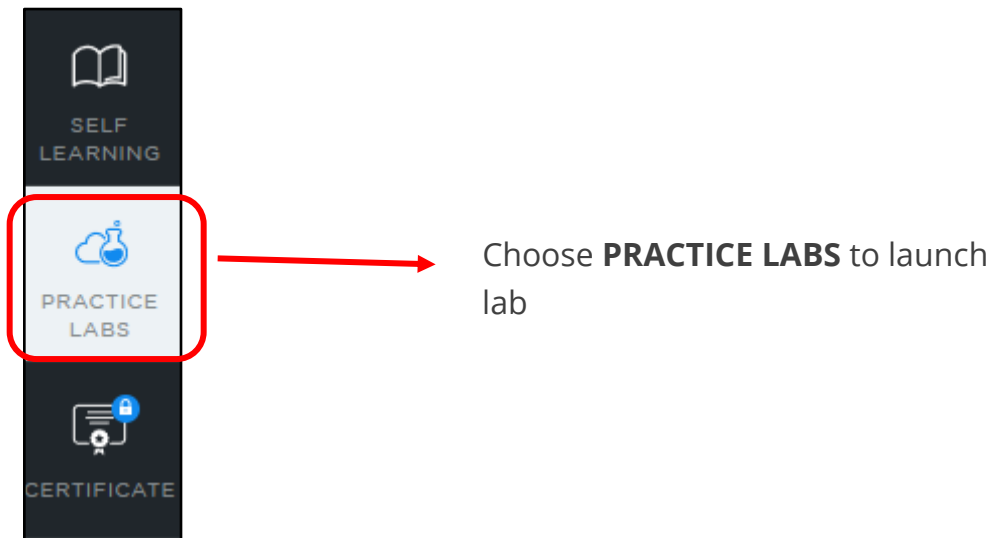
**Note: The screenshots are only for your reference. Your LMS may look different depending on your course content.**

This section will guide you to:

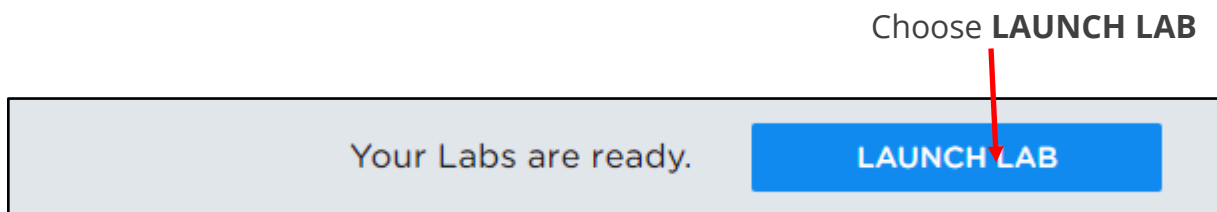
- Use labs for executing all the demos included in this course

**Step 1: Log in to Simplilearn LMS**

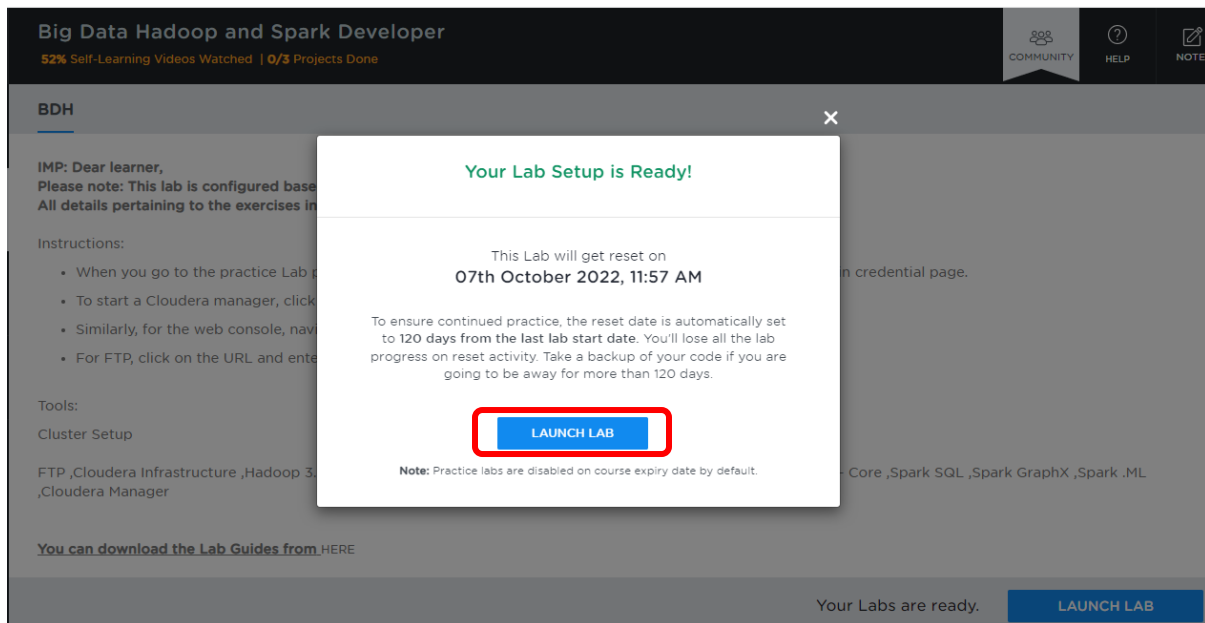
- Go to the respective course
- Starting **PRACTICE LABS** on LMS



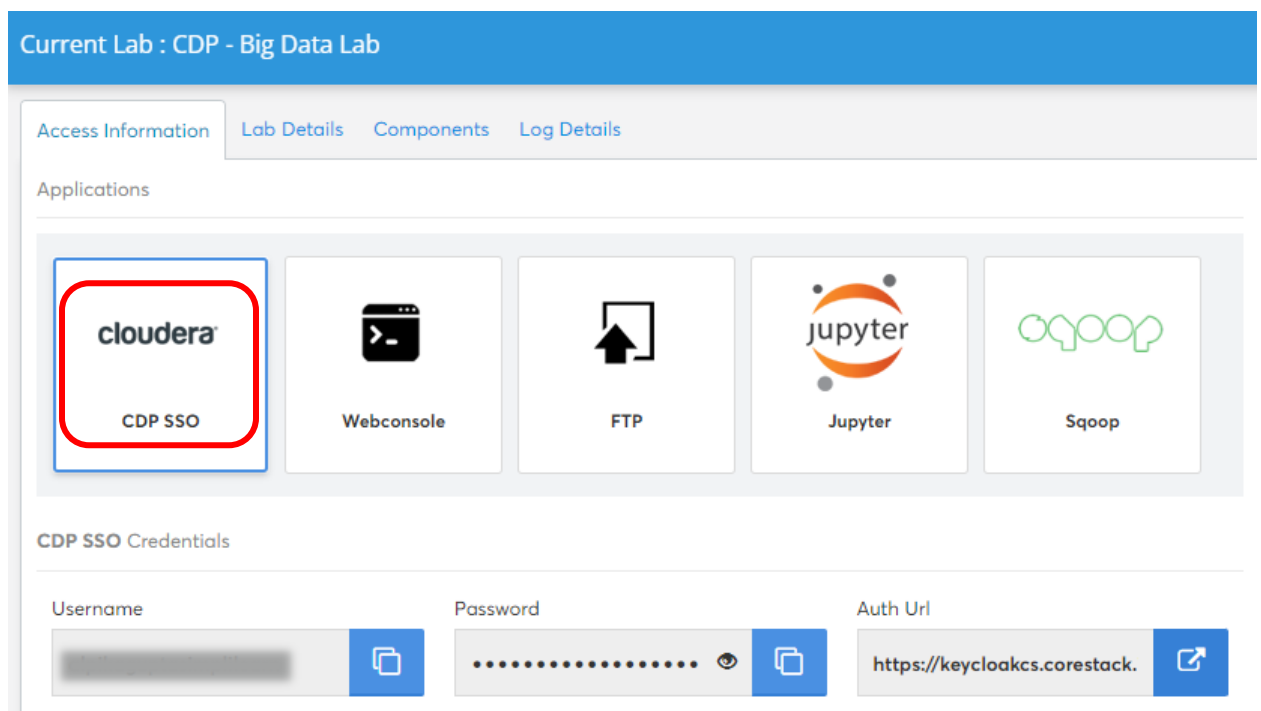
**Step 2: Click on the **LAUNCH LAB** button**



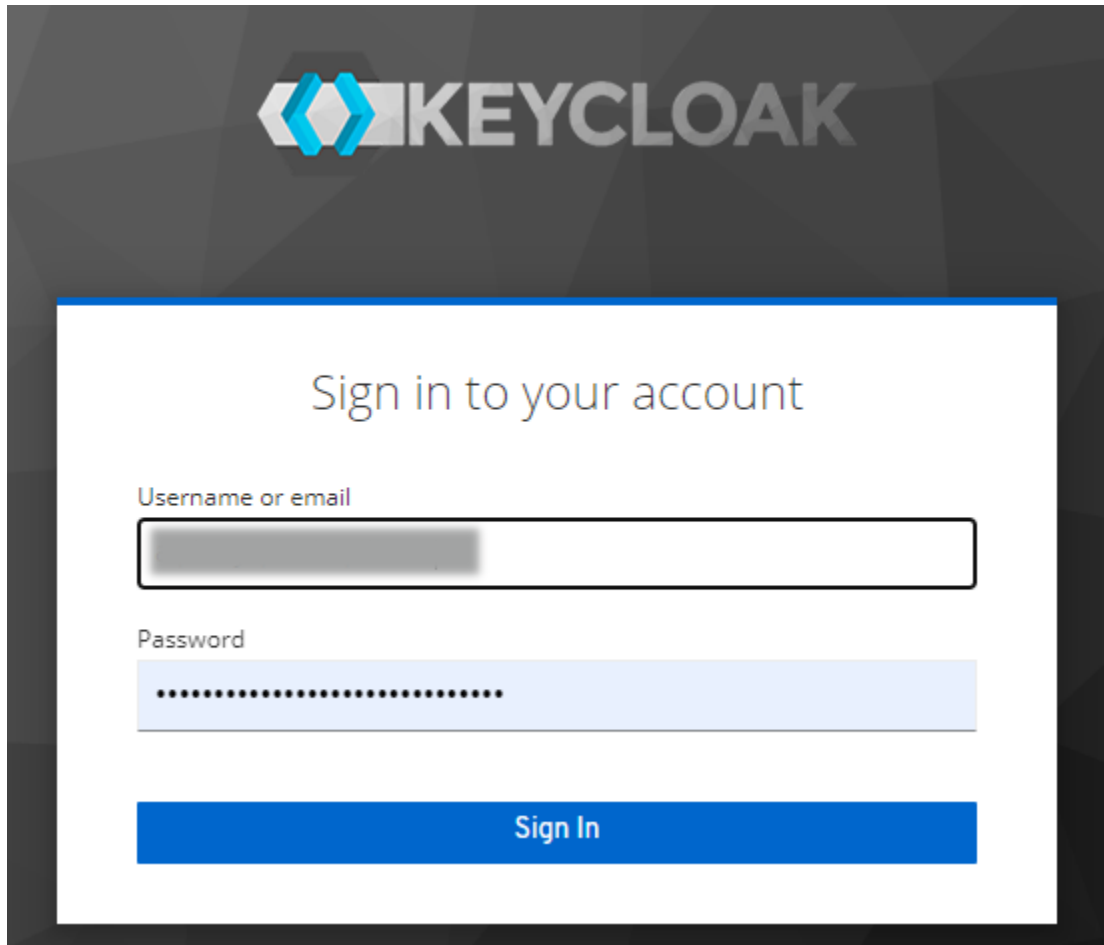
**Step 3:** A new window will open. Read the instructions and click on the **LAUNCH LAB** button. This will launch the Practice Labs for this course



**Step 4:** To log in to Cloudera, Hue, or any service provided by Cloudera. Select the Cloudera **CDP SSO** and click on the Auth URL as shown below. Copy the **Username** and the **Password** provided to log in



**Step 5:** Next, you will redirect to the login screen to enter your Username and Password

The image shows the Keycloak login interface. At the top, the Keycloak logo is displayed against a dark background with a geometric pattern. Below the logo, the text "Sign in to your account" is centered. Underneath, there are two input fields: "Username or email" and "Password". The "Username or email" field is a simple white box with a black border. The "Password" field is a light blue box with a black border and contains a series of dots representing masked characters. At the bottom of the form, there is a blue button labeled "Sign In".

KEYCLOAK

Sign in to your account

Username or email

Password

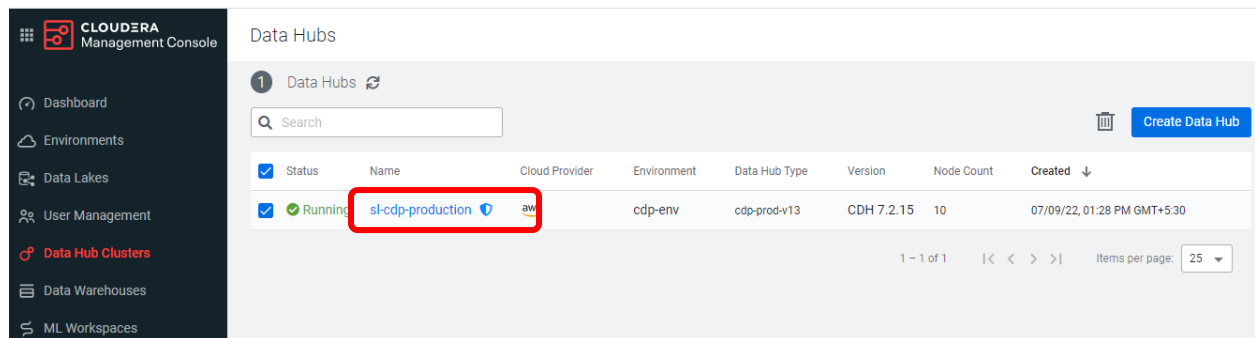
Sign In

Once you are successfully logged in the CDP page will be redirected to the below page as shown in Step 6:

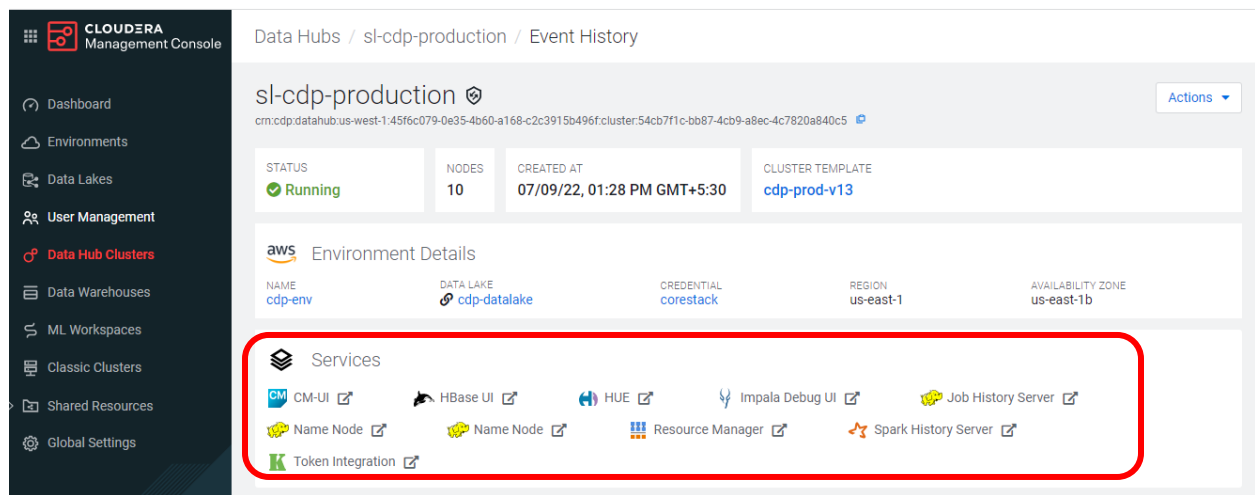
**Step 6:** Click on Data Hub Clusters



## Step 7: Click on the cluster box **sl-cdp-production**

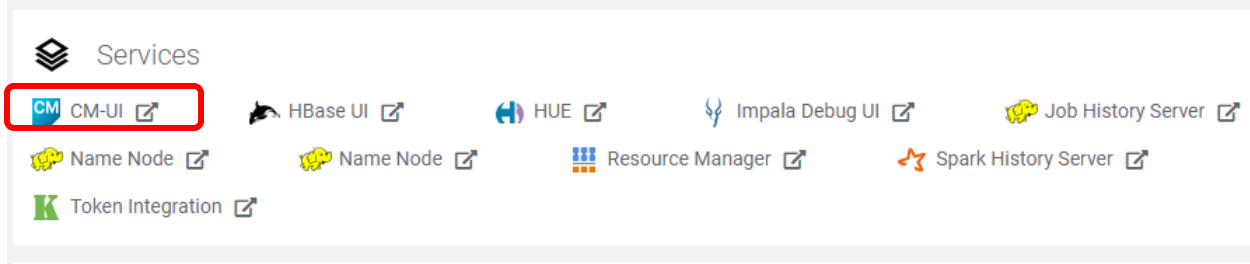


## Step 8: You will see the services provided by Cloudera



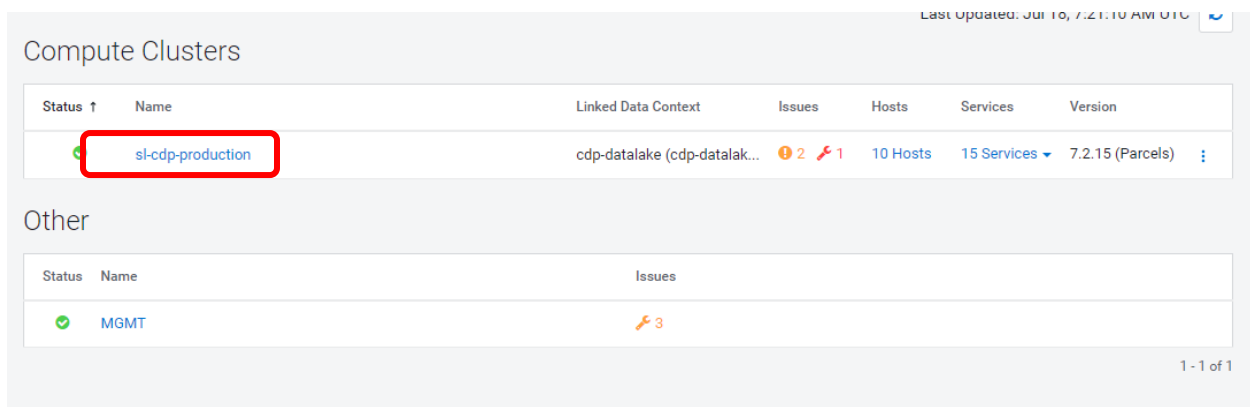
The box will direct you to Hue, the Cloudera Manager (CM-UI), and the services CDP provides.

**Step 9:** To find the hostname of Bigdata services, click on the link named **CM-UI**



**Step 10:** Once clicked the **CM-UI**, you will see the below image as shown in Step 11:

**Step 11:** Then, click on the **sl-cdp-production** under Compute Clusters



**Step 12:** Services provided will be listed under **Compute Clusters**. Click on the desired service. Suppose you need a ZooKeeper hostname. So, click on ZooKeeper

CLUSTERA  
Manager

Search

- Clusters
- Hosts
- Diagnostics
- Charts

Running Commands

Support

sl-cdp-production Actions

Status Health Issues Configuration 1

### Status

Data Context: cdp-datalake

- ranger
- atlas
- knox
- hive

Compute Cluster, Cloudera Runtime 7.2.15 (Parcels)

- 10 Hosts 7
- Hive 1
- Hive Metastore 1
- Kafka

### Charts

Cluster CPU

Cluster Disk IO

Cluster Network IO

HDFS IO

Completed Impala Queries

CLUSTERA  
Manager

Search

- Clusters
- Hosts
- Diagnostics
- Charts

Running Commands

Support

testdemomay1301mailinator

7.5.2

- Knox
- YARN Queue Manager
- ZooKeeper**
- bdh-datalake
- hbase 1
- hdfs
- hue
- impala
- oozie
- spark\_on\_yarn
- sqoop
- tez
- yarn

**Step 13:** You will see the ZooKeeper page as shown below:

**CloudERA Manager**

Search

Clusters  
Hosts  
Diagnostics  
Charts

Running Commands  
Support  
testdemomay1301mallinator

## ZooKeeper

30 minutes preceding

Status Instances Configuration Commands Charts Library Quick Links

### ZooKeeper Summary

ZooKeeper Current Zxid: Epoch: 0x00000003, xid: 0x000326f6

ZooKeeper Server Status: **Server, sl-cdp-production-master10 (Leader)**  
 Server, sl-cdp-production-master20 (Follower)  
 Server, sl-cdp-production-master30 (Follower)

### Health Tests

Show 2 Good

### Status Summary

Server: 3 Good Health  
 Hosts: 3 Good Health

### Charts

**CPU Cores Used**

0.004  
0.002  
cores

07 AM 07:15

Server (sl-cdp-... 0.0014 Server (sl-cdp-... 0.0015  
 Server (sl-cdp-pr... 0.001

**Canary Duration**

400ms  
200ms  
ms

07 AM 07:15

ZooKeeper, Canary Duration 474ms

**Step 14:** Click instances, this will list the hostnames available under ZooKeeper

sl-cdp-production

## ZooKeeper

Status **Instances** Configuration Commands Charts Library Quick Links

Search Filters Last Updated: Jul 18, 7:27:17 AM UTC

Filters

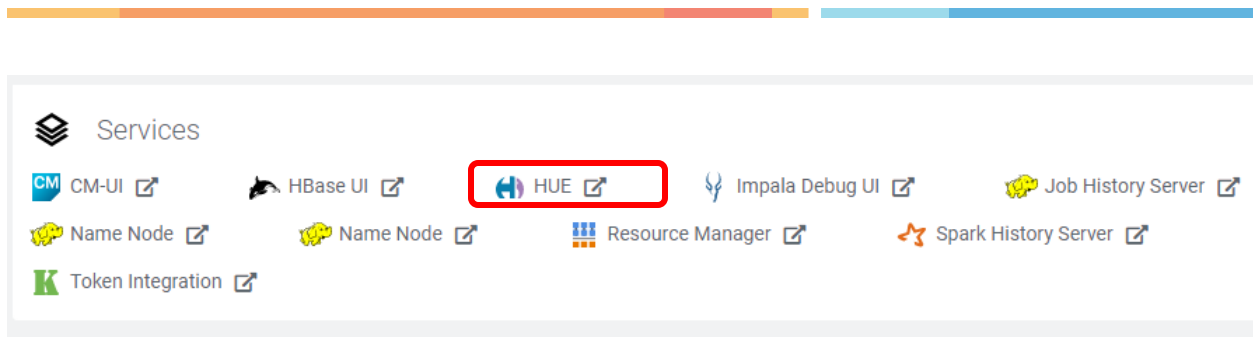
> HEALTH TEST

Status	Role Type	State	Hostname	Commission State	Role Group
✓	Server	Started	sl-cdp-production-master10.cdp-env.gne4-rutx.cloudera.site	Commissioned	Server Default Group
✓	Server	Started	sl-cdp-production-master20.cdp-env.gne4-rutx.cloudera.site	Commissioned	Server Default Group
✓	Server	Started	sl-cdp-production-master30.cdp-env.gne4-rutx.cloudera.site	Commissioned	Server Default Group

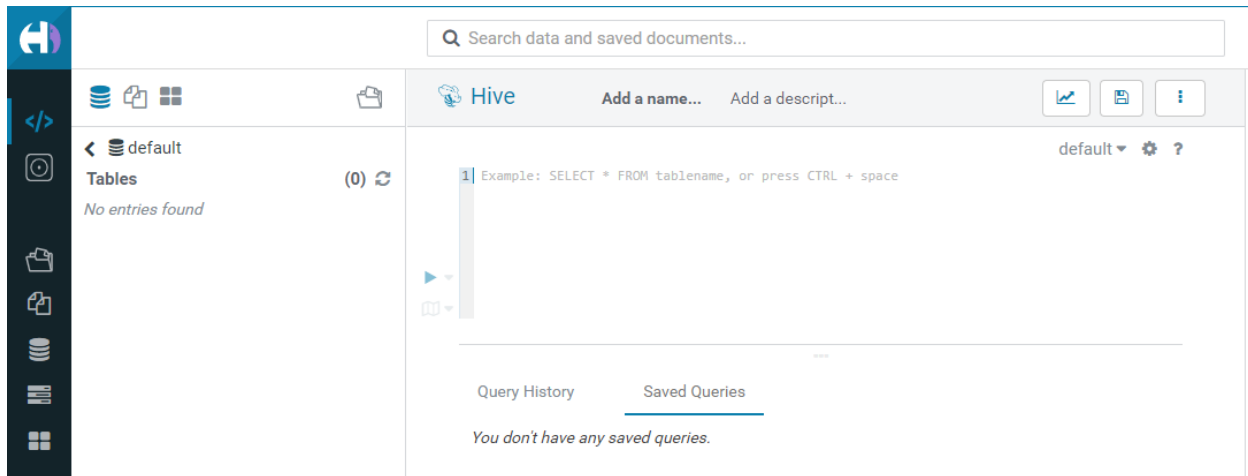
1 - 3 of 3

**Step 15:** Click on the **HUE** icon to log in

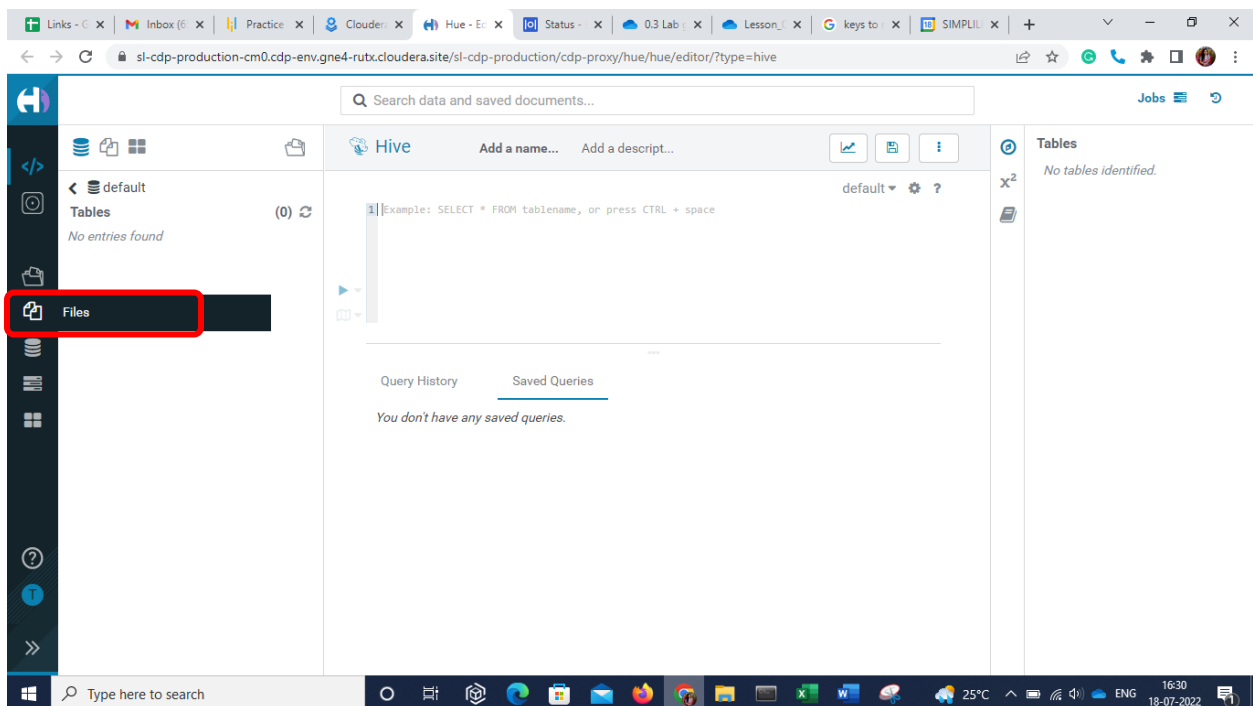




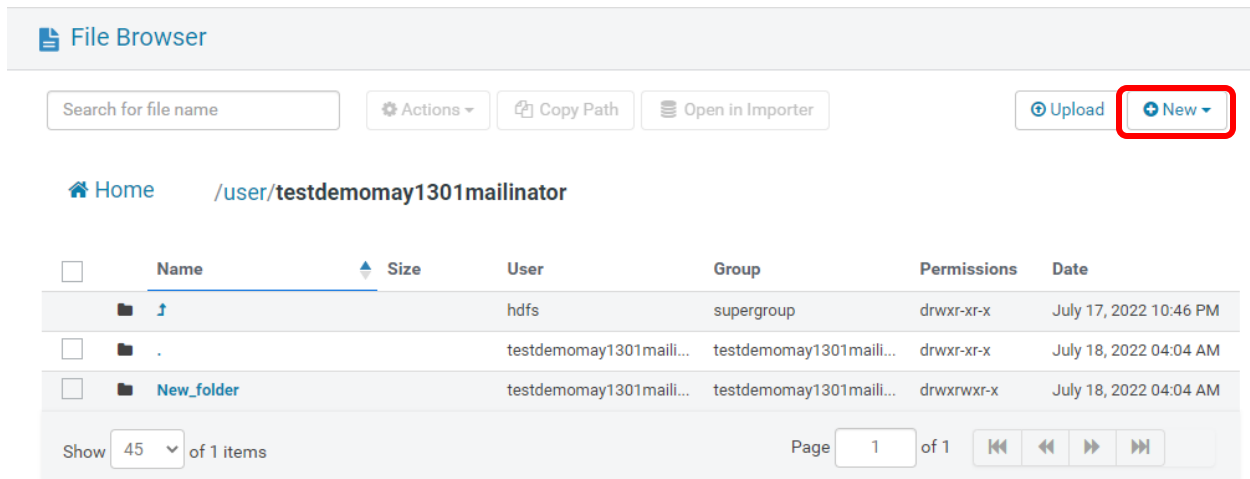
**Step 16:** You will be navigated to the dashboard as shown below.



**Step 17:** To upload the dataset in **HUE** click on the icon as shown below:



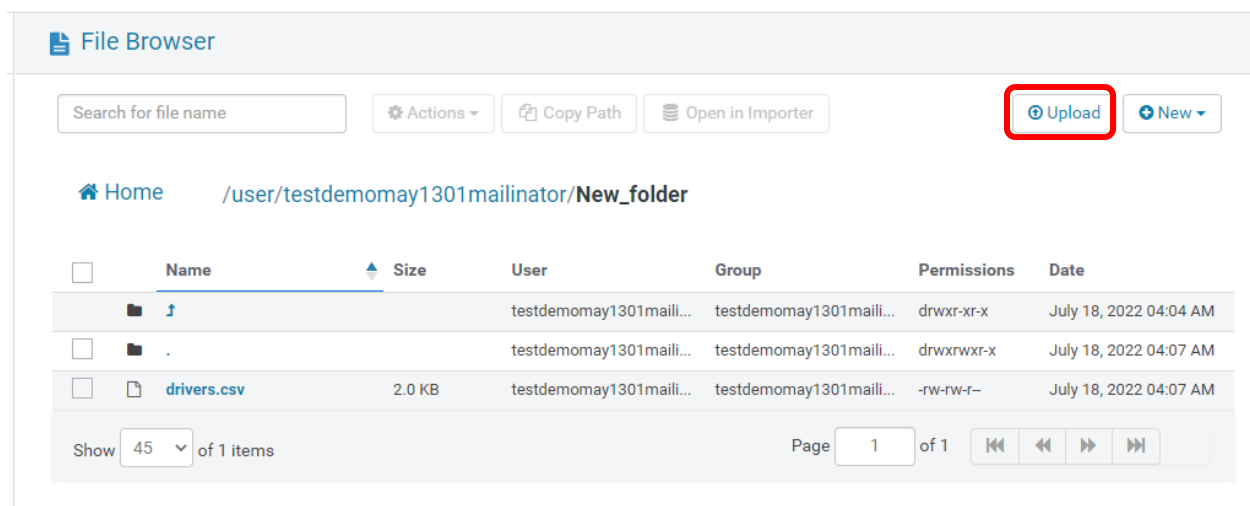
**Step 18:** Create a new directory by clicking on **New** and upload the dataset using the upload button as shown below:



The screenshot shows the 'File Browser' interface. At the top, there is a search bar and several action buttons: 'Actions', 'Copy Path', 'Open in Importer', 'Upload', and 'New'. The 'New' button is highlighted with a red box. Below the buttons, the breadcrumb path is '/user/testdemomay1301mailinator'. The main area displays a table of files and folders:

	Name	Size	User	Group	Permissions	Date
<input type="checkbox"/>	<b>↑</b>		hdfs	supergroup	drwxr-xr-x	July 17, 2022 10:46 PM
<input type="checkbox"/>	<b>.</b>		testdemomay1301maili...	testdemomay1301maili...	drwxr-xr-x	July 18, 2022 04:04 AM
<input type="checkbox"/>	<b>New_folder</b>		testdemomay1301maili...	testdemomay1301maili...	drwxrwxr-x	July 18, 2022 04:04 AM

At the bottom, there is a pagination bar showing 'Show 45 of 1 items' and 'Page 1 of 1' with navigation arrows.



The screenshot shows the 'File Browser' interface with the breadcrumb path '/user/testdemomay1301mailinator/New\_folder'. The 'Upload' button is highlighted with a red box. The main area displays a table of files and folders:

	Name	Size	User	Group	Permissions	Date
<input type="checkbox"/>	<b>↑</b>		testdemomay1301maili...	testdemomay1301maili...	drwxr-xr-x	July 18, 2022 04:04 AM
<input type="checkbox"/>	<b>.</b>		testdemomay1301maili...	testdemomay1301maili...	drwxrwxr-x	July 18, 2022 04:07 AM
<input type="checkbox"/>	<b>drivers.csv</b>	2.0 KB	testdemomay1301maili...	testdemomay1301maili...	-rw-rw-r--	July 18, 2022 04:07 AM

At the bottom, there is a pagination bar showing 'Show 45 of 1 items' and 'Page 1 of 1' with navigation arrows.


**Step 19:** Similarly, click on the Webconsole and navigate using **Auth Url**


**Step 20:** Copy the **Username** and the **Password** provided to log in to the **Webconsole**


Current Lab : CDP - Big Data Lab


Access Information Lab Details Components Log Details


Applications

  
CDP SSO

  
Webconsole


  
FTP

  
Jupyter


  
Sqoop

CDP SSO Credentials


Username



Password



Auth Url



**Step 21:** 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

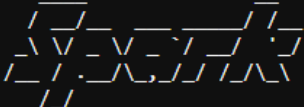
```
sl-cdp-production-en0 login: simplilearn
Password:
=====
*                               :
Password for simplilearn@CDP-ENV.GNE4-RUTX.CLOUDERA.SITE:
[simplilearn@sl-cdp-production-en0 ~]$
```

**Step 22:** Enter the **PySpark** console by running the below command:

**Command:**

pyspark3

```
Password for      simplilearn@CDP-ENV.GNE4-RUTX.CLOUDERA.SITE:  
[ simplilearn@sl-cdp-production-en0 ~]$ pyspark3  
Python 3.6.8 (default, Nov 16 2020, 16:55:22)  
[GCC 4.8.5 20150623 (Red Hat 4.8.5-44)] 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).  
/opt/cloudera/parcels/CDH-7.2.15-1.cdh7.2.15.p1.26792553/lib/spark3/python/pyspark/context.py:238: FutureWarning  
FutureWarning  
Welcome to
```



```
version 3.2.1.7.2.15.1-1
```

```
Using Python version 3.6.8 (default, Nov 16 2020 16:55:22)  
Spark context Web UI available at http://sl-cdp-production-en0.cdp-env.gne4-rutx.cloudera.site:4040  
Spark context available as 'sc' (master = local[*], app id = local-1658130010344).  
SparkSession available as 'spark'.  
>>>
```

**Step 23:** To enter the Python shell use the below command:

**Command:**

python3

```
sl-cdp-production-en1 login: simplilearn
Password:
=====
*                               :
Password for simplilearn@CDP-ENV.GNE4-RUTX.CLOUDERA.SITE:
[simplilearn@sl-cdp-production-en1 ~]$ python3
Python 3.6.8 (default, Nov 16 2020, 16:55:22)
[GCC 4.8.5 20150623 (Red Hat 4.8.5-44)] on linux
Type "help", "copyright", "credits" or "license" for more information.
>>>
```

**Step 24:** To come out from the Python shell press Ctrl+d. It would not be visible

```
sl-cdp-production-en1 login: simplilearn
Password:
=====
*                               :
Password for simplilearn@CDP-ENV.GNE4-RUTX.CLOUDERA.SITE:
[simplilearn@sl-cdp-production-en1 ~]$ python3
Python 3.6.8 (default, Nov 16 2020, 16:55:22)
[GCC 4.8.5 20150623 (Red Hat 4.8.5-44)] on linux
Type "help", "copyright", "credits" or "license" for more information.
>>>
[simplilearn@sl-cdp-production-en1 ~]$
```

**Step 25:** To enter the vi editor and to write any Python file or txt file use the below command:

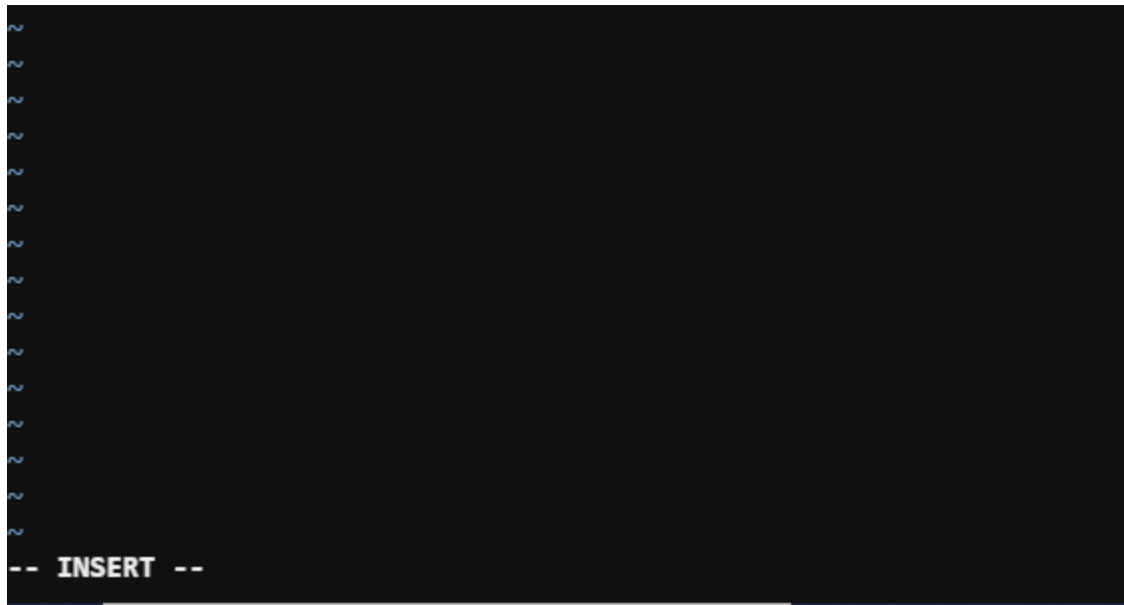
vi sample.py

Or

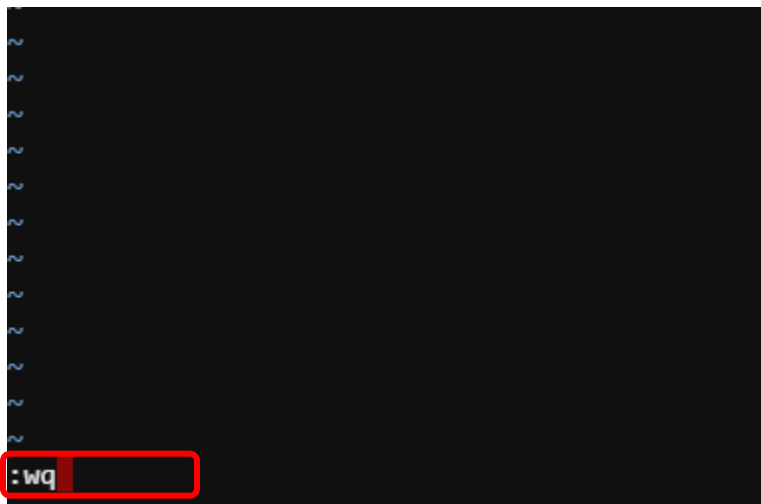
vi sample.txt

```
[alpikaguptasimplilearn@sl-cdp-production-en1 ~]$ python3
Python 3.6.8 (default, Nov 16 2020, 16:55:22)
[GCC 4.8.5 20150623 (Red Hat 4.8.5-44)] on linux
Type "help", "copyright", "credits" or "license" for more information.
>>>
[simplilearn@sl-cdp-production-en1 ~]$ vi sample.py
```

**Step 26:** Click on **i** on your keyboard to enter the insert mode



**Step 27:** To save and exit, click on the **ESC** key and type: wq



**Step 28:** To execute the Python script run the below command:

**Command:**

python3 sample.py

```
[ simplilearn@sl-cdp-production-en1 ~]$ python3
Python 3.6.8 (default, Nov 16 2020, 16:55:22)
[GCC 4.8.5 20150623 (Red Hat 4.8.5-44)] on linux
Type "help", "copyright", "credits" or "license" for more information.
>>>
[simplilearn@sl-cdp-production-en1 ~]$ vi sample.py
[simplilearn@sl-cdp-production-en1 ~]$ python3 sample.py
```

**Step 29:** To login into the Scala environment use the below command:

**Command:**

spark3-shell

```
[simlilearn@sl-cdp-production-en1 ~]$ spark3-shell
Setting default log level to "WARN".
To adjust logging level use sc.setLogLevel(newLevel). For SparkR, use setLogLevel(newLevel).
22/07/18 12:10:00 WARN util.Utils: Service 'SparkUI' could not bind on port 4040. Attempting port 4041.
22/07/18 12:10:00 WARN util.Utils: Service 'SparkUI' could not bind on port 4041. Attempting port 4042.
Spark context Web UI available at http://sl-cdp-production-en1.cdp-env.gne4-rutx.cloudera.site:4042
Spark context available as 'sc' (master = local[*], app id = local-1658146200453).
Spark session available as 'spark'.
Welcome to

  ____      _
 / ___|    / \
| |  | |  / _ \
| |  | | / ___ \
| |  | |/_/   \_\
| |  | |
| |  | |
|_|  |_|      version 3.2.1.7.2.15.1-1

Using Scala version 2.12.10 (OpenJDK 64-Bit Server VM, Java 1.8.0_332)
Type in expressions to have them evaluated.
Type :help for more information.


scala>
```


**Step 30:** To login into the FTP, click on **FTP** and click on the **Auth Url** to upload the dataset and copy the **Username** and the **Password** provided to log in to the **FTP**


Current Lab : CDP - Big Data Lab


Access Information Lab Details Components Log Details


Applications

  
CDP SSO

  
Webconsole


  
FTP

  
Jupyter


  
Sqoop

CDP SSO Credentials


Username



Password



Auth Url



### Cloud Lab FTP Server

Username:

Password:

Login

---

☐ Save login details

---

Language:  ▼



**Step 31:** You will be navigated to the screen as shown below:

■	Name	Size	Date	Time
□ □	<a href="#">sample.py</a>	0	18/07/22	07:43

New Folder

New File

Fetch File

Upload Files

Upload Folder

Host: localhost User: alpikaguptasimplilearn Upload Limit: 1GB

**Step 32:** Click on **Jupyter** and click on the **Auth Url** to log in and copy the **Username** and the **Password** provided to log in to the **Jupyter**

Current Lab : CDP - Big Data Lab

Access Information

Lab Details

Components

Log Details

Applications

cloudera

CDP SSO

>\_

Webconsole

↑

FTP

jupyter

Jupyter

oooop

Sqoop

CDP SSO Credentials

Username

Ⓞ

Password

.....

👁

Ⓞ

Auth Url

https://keycloakcs.corestack.

🔗

Sign in

Warning: JupyterHub seems to be served over an unsecured HTTP connection. We strongly recommend enabling HTTPS for JupyterHub.

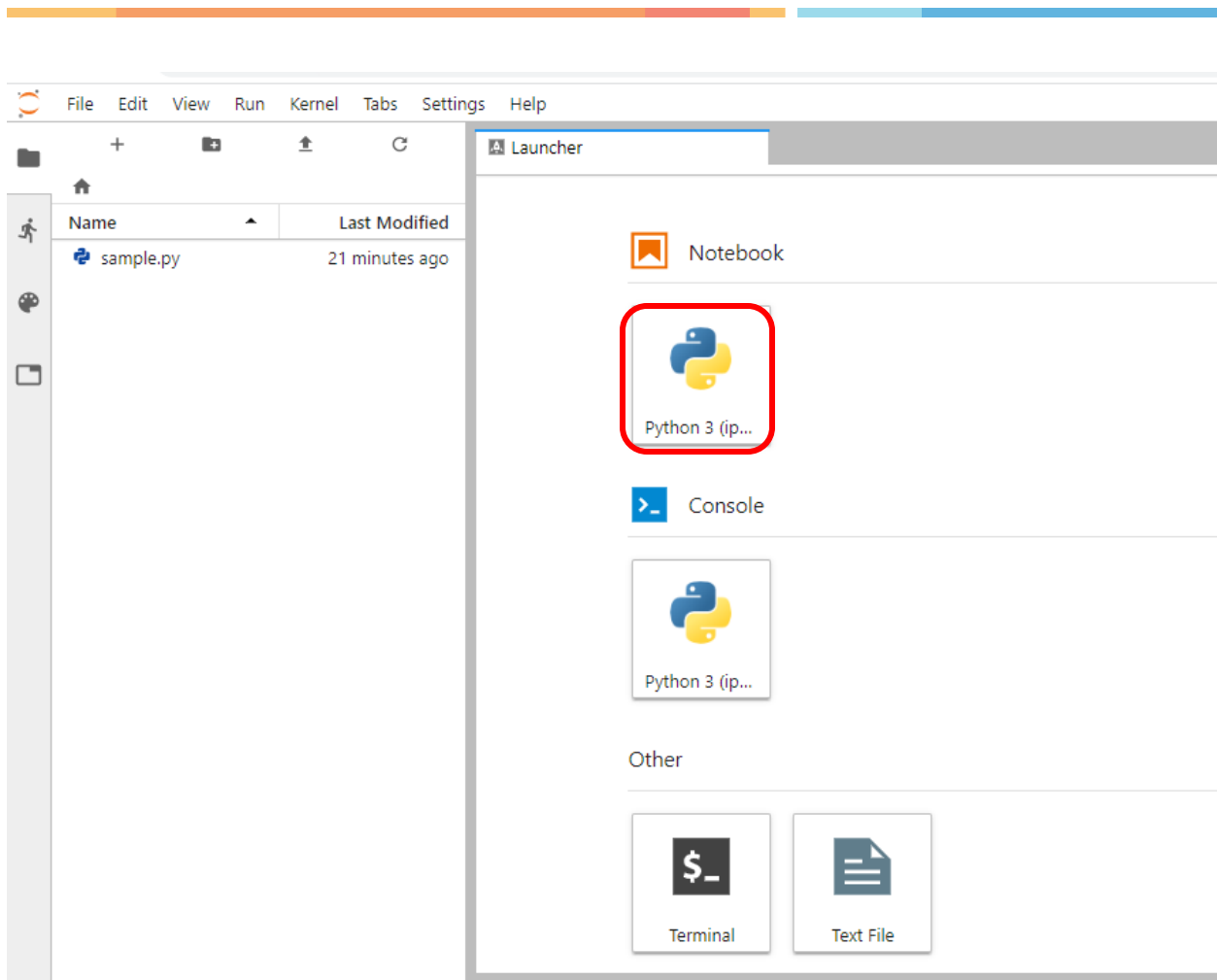
Username:

Password:

.....|

Sign in

**Step 33:** You will be able to see the Python interface as shown below and click on **Notebook Python 3**



**Step 34:** You will see the Jupyter notebook to write the Python code

