Big Data Hadoop and Spark Developer

Lab Guide

simplilearn

Get Certified. Get Ahead.

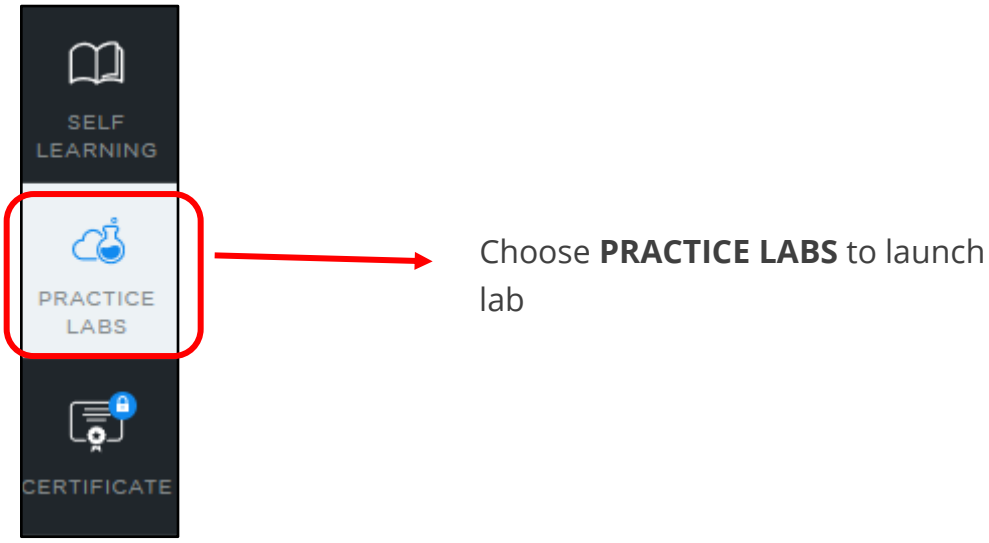
Note: The screenshots are only for your reference. Your LMS may look different depending on your course content.

This section will guide you to:

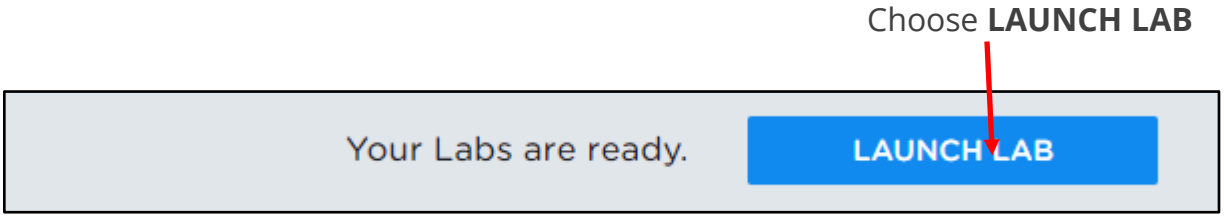
- Use labs for executing all the demos included in this course

Step 1: Log in to Simplilearn LMS

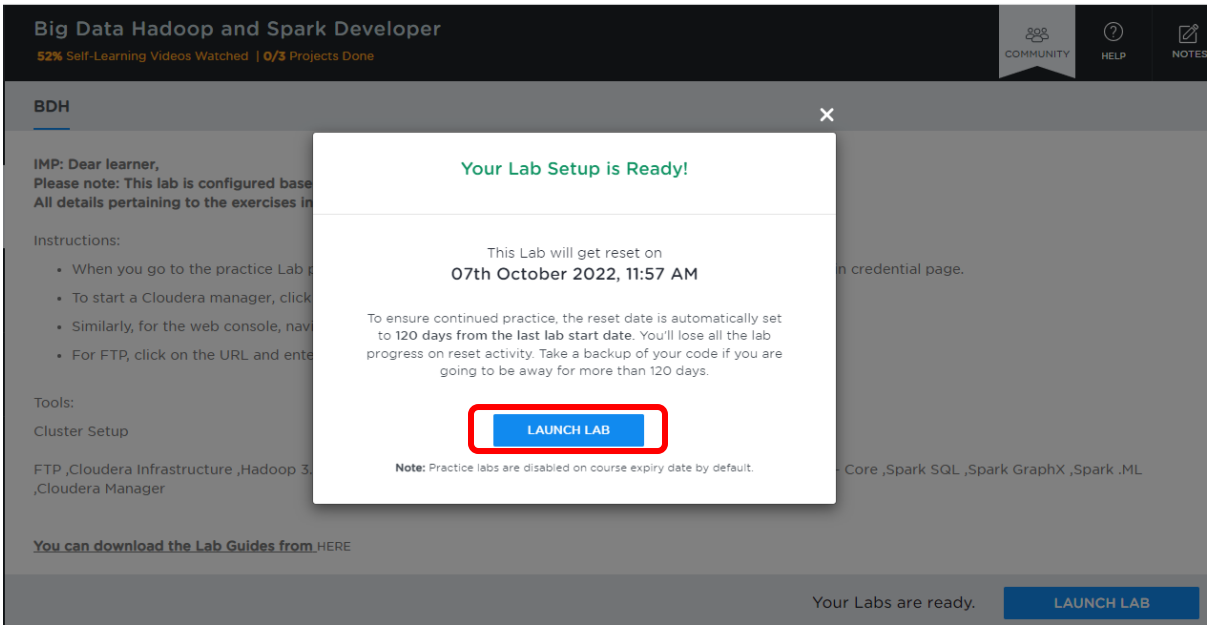
- Go to the respective course
- Starting **PRACTICE LABS** on LMS



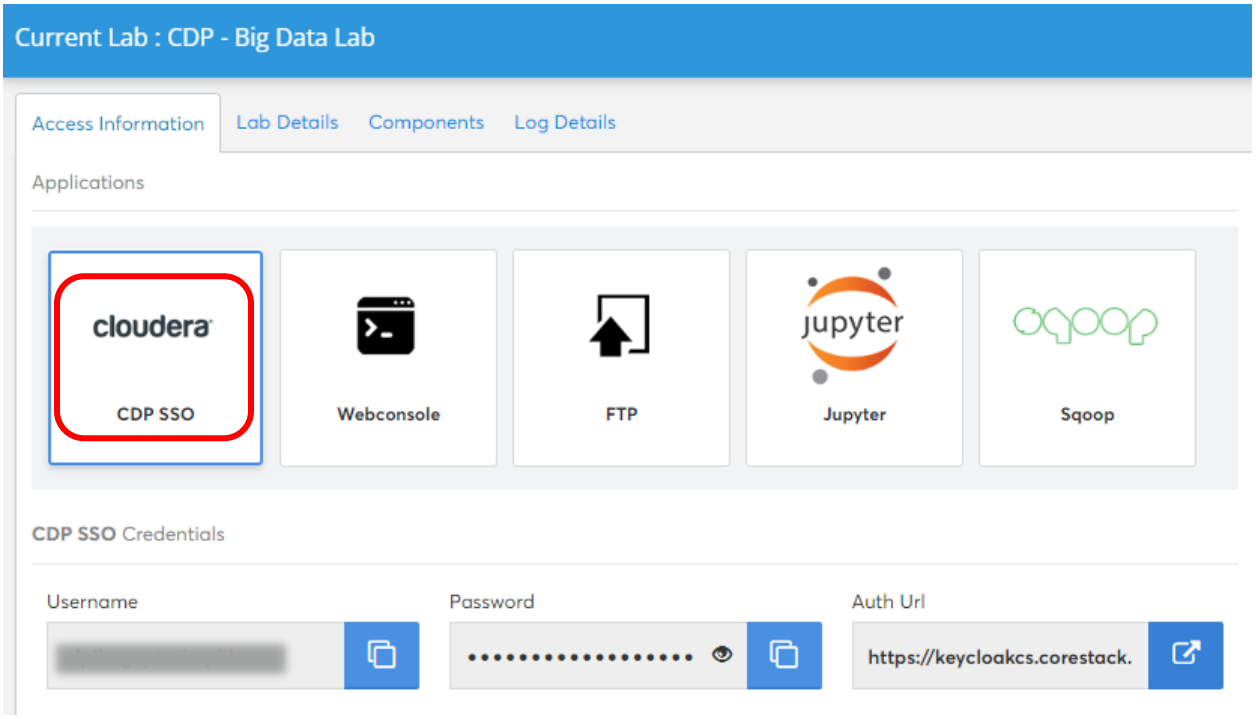
Step 2: Click on the **LAUNCH LAB** button



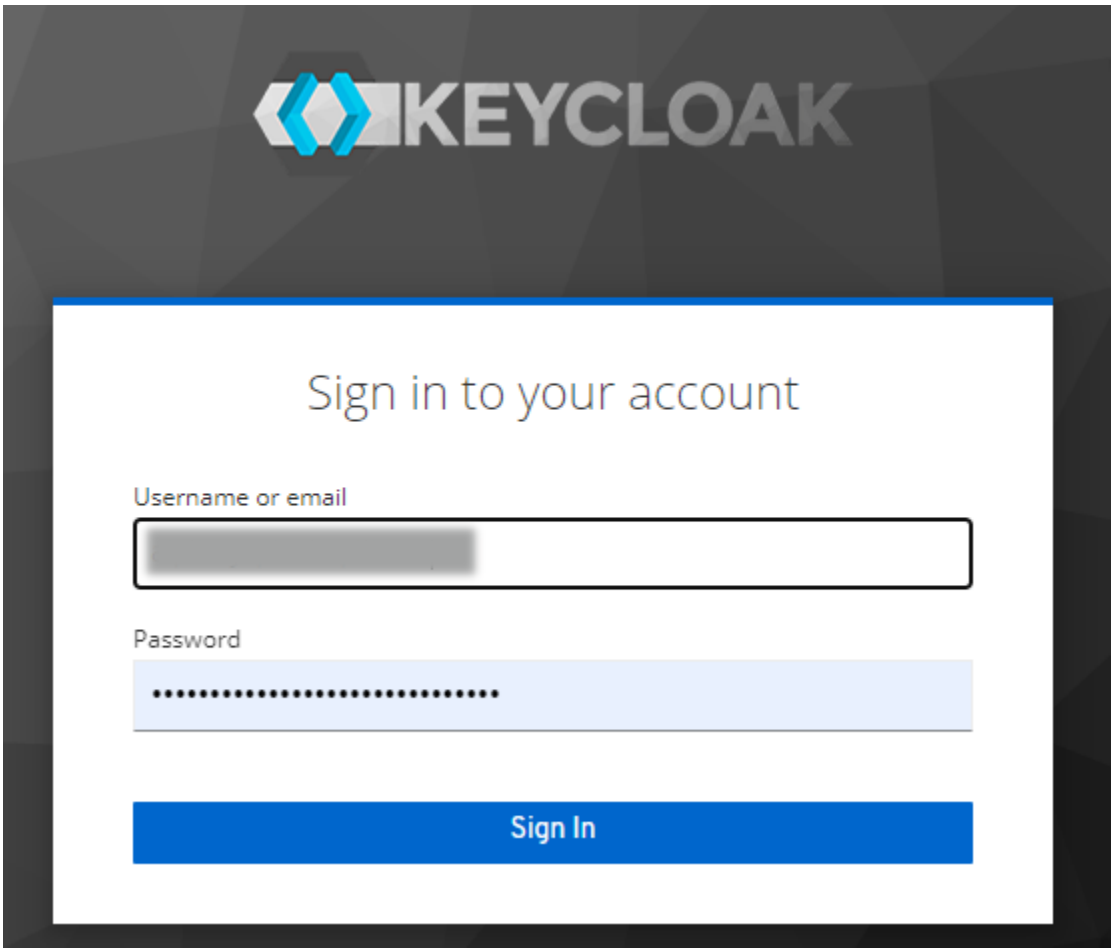
Step 3: A new window will open. Read the instructions and click on the **LAUNCH LAB** button**.** This will launch the Practice Labs for this course



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

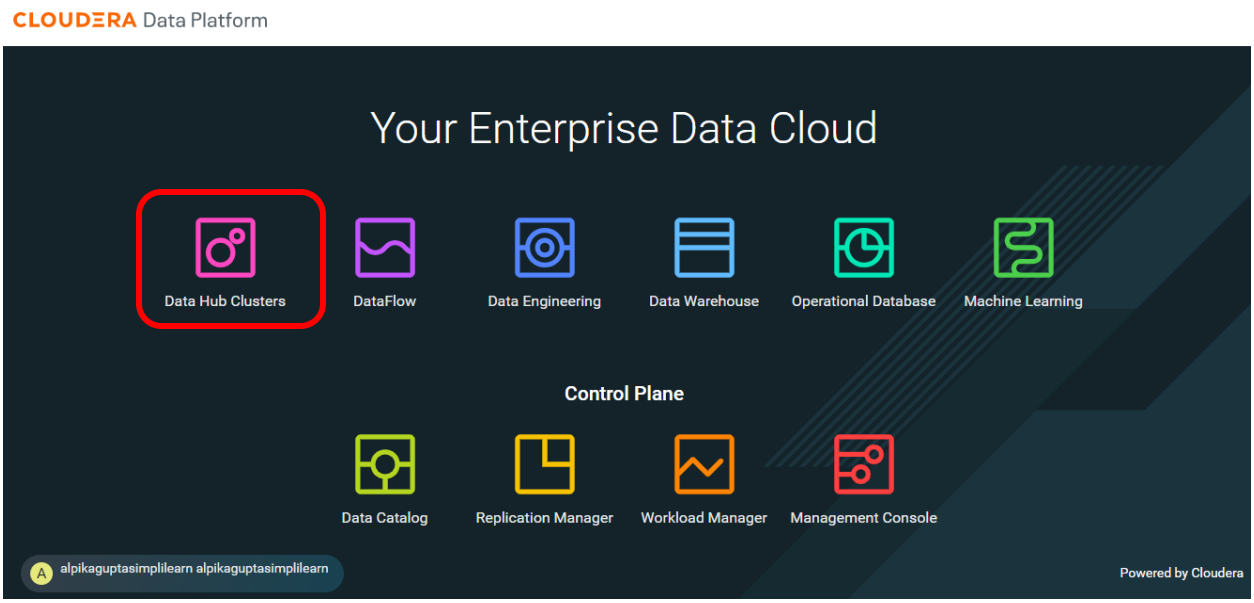


Step 5: Next, you will redirect to the login screen to enter your Username and Password

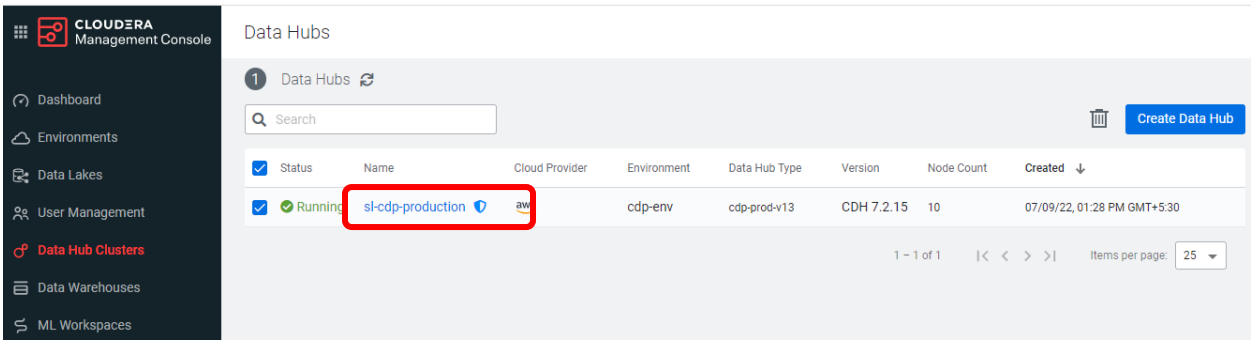


Once you are successfully logged in the CDP page will be redirected to the below page as shown in Step 6:

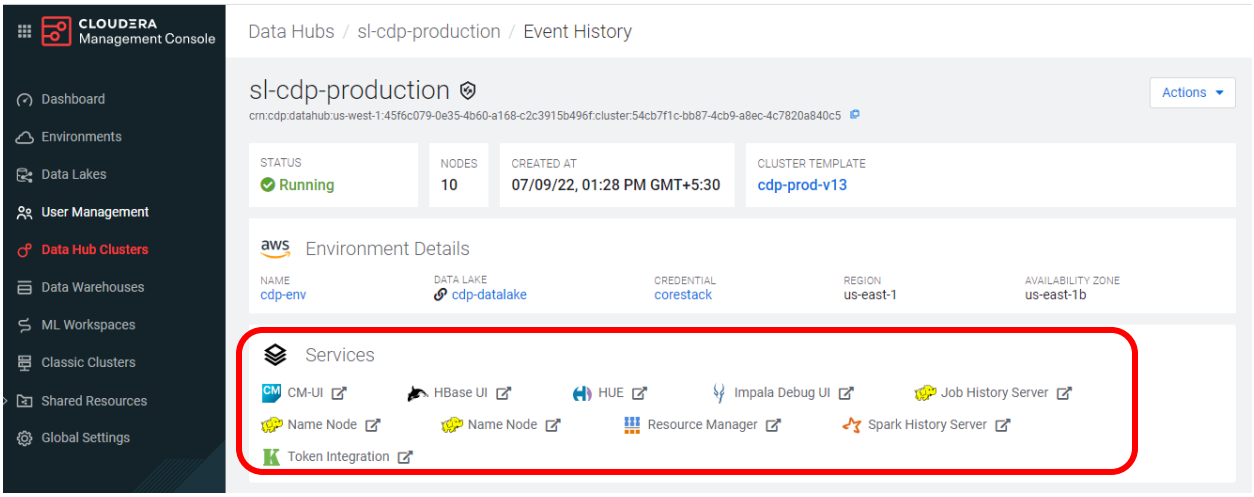
Step 6: Click on Data Hub Clusters



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

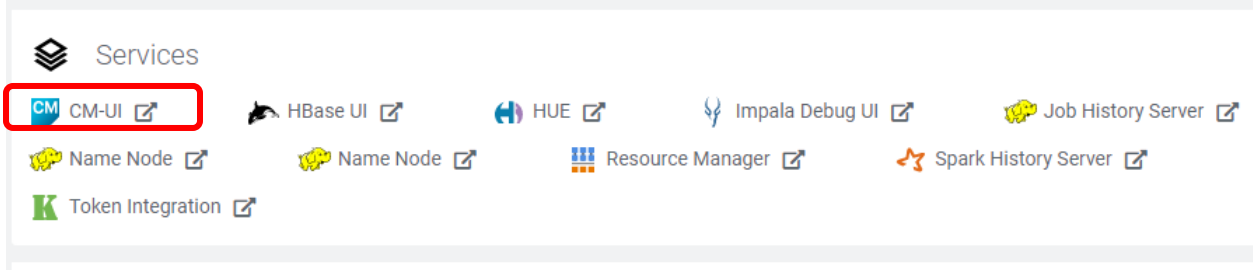


Step 8: You will see the services provided by Cloudera



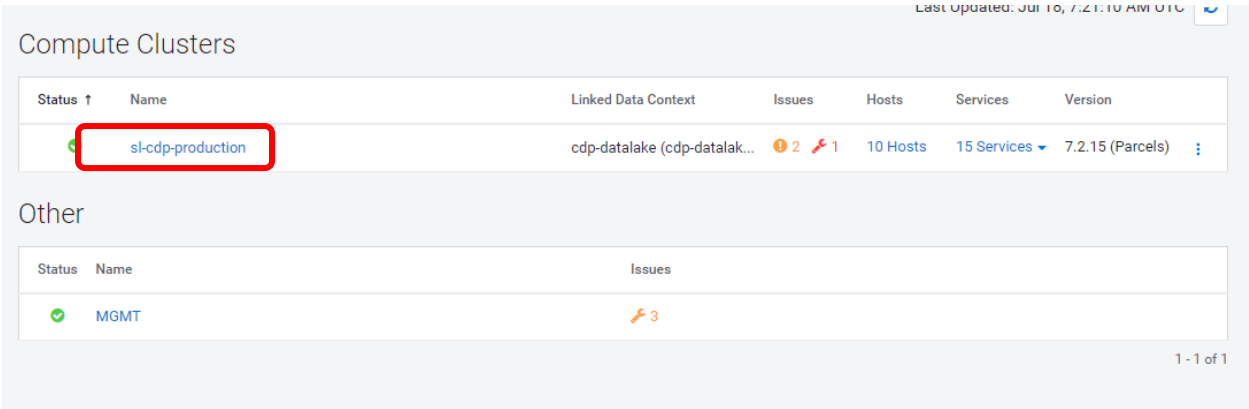
The box will direct you to Hue, the Cloudera Manager (CM-UI), and the services CDP provides.

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

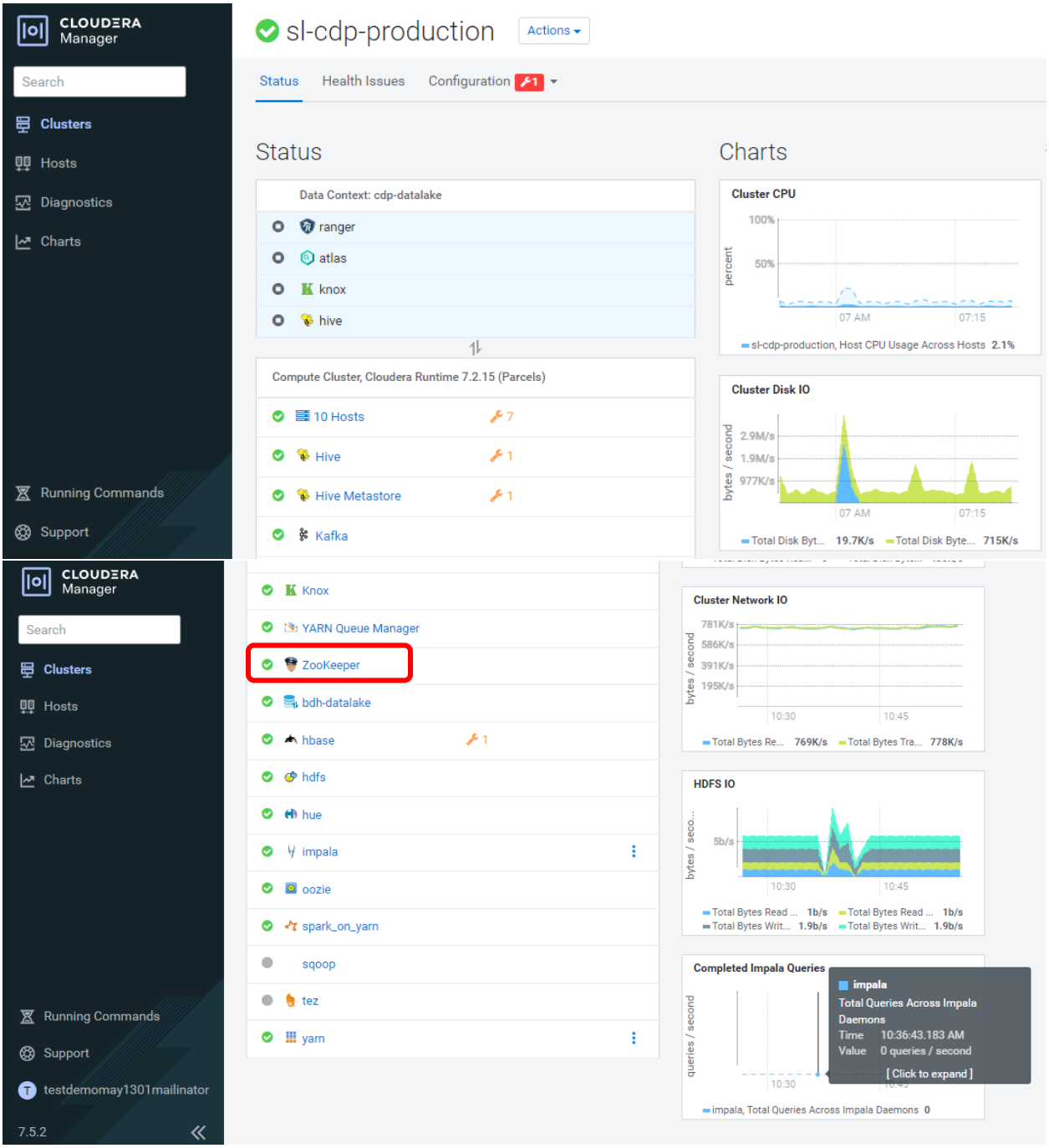


Step 10: Once clicked the **CM-UI**, you will see the below image as shown in Step 11:

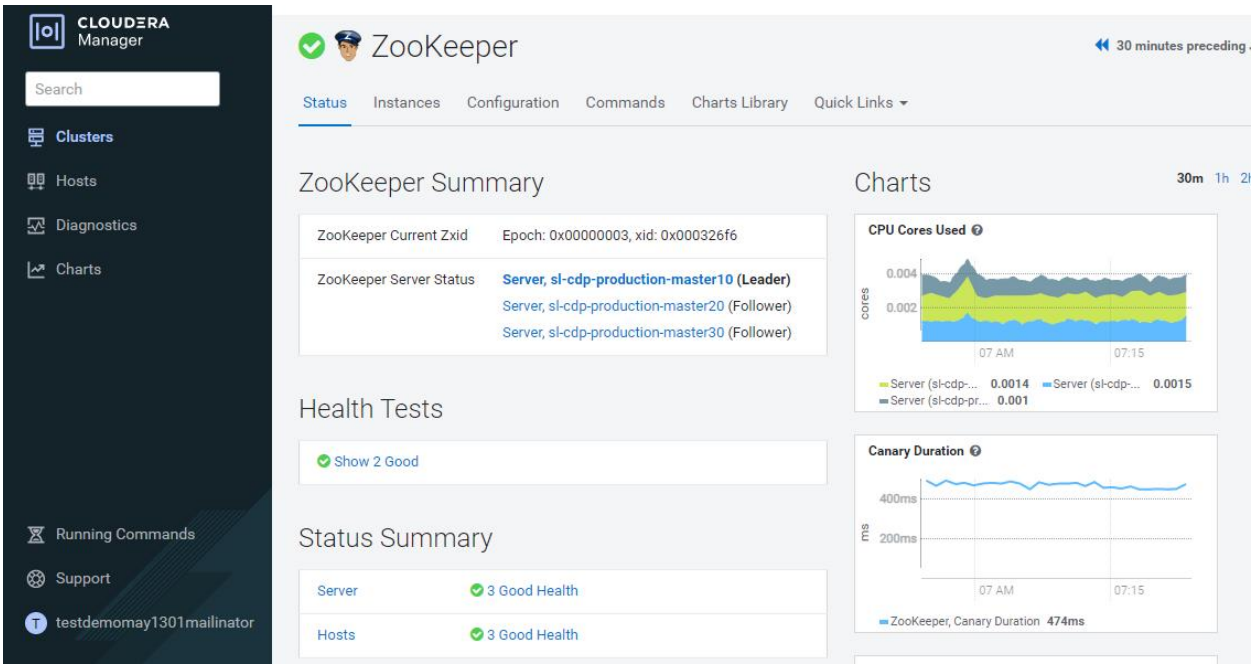
Step 11: Then, click on the **sl-cdp-production** under Compute Clusters



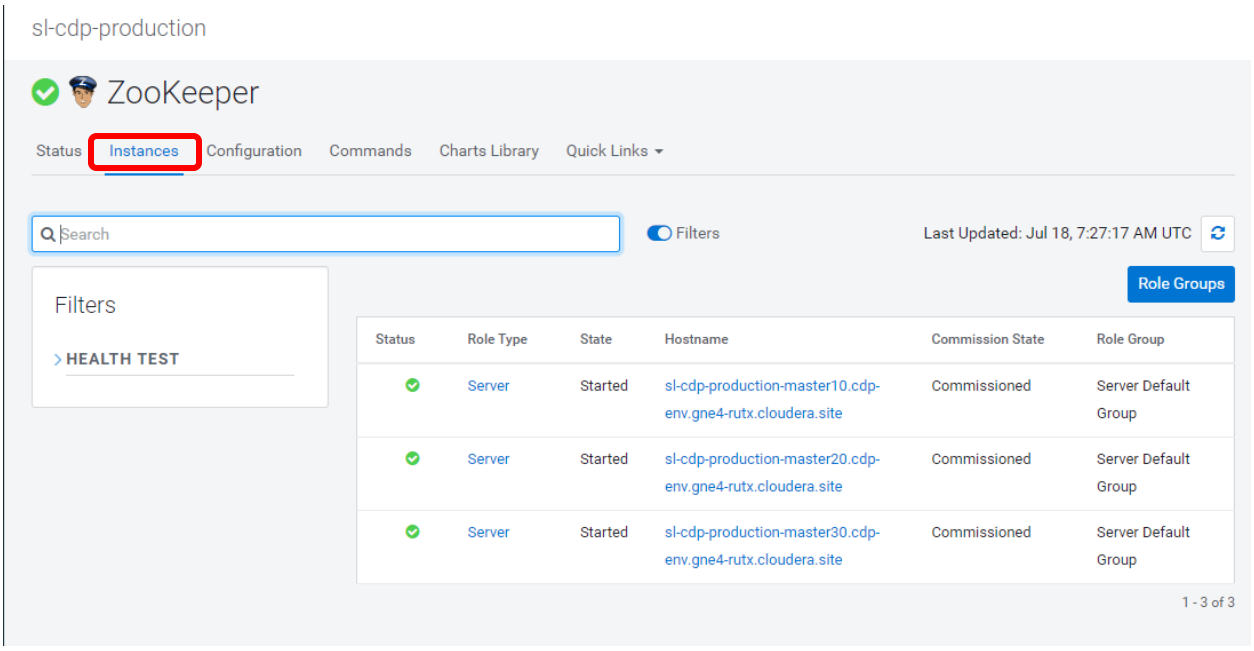
Step 12: Services provided will be listed under **Compute Clusters**. Click on the desired service. Suppose you need a ZooKeeper hostname. So, click on ZooKeeper



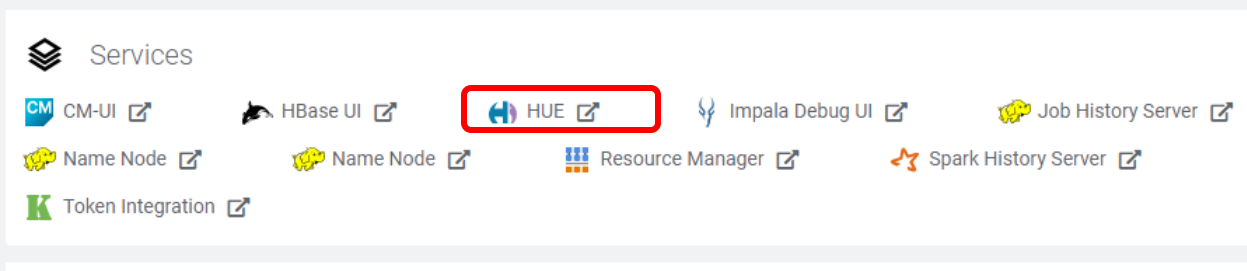
Step 13: You will see the ZooKeeper page as shown below:



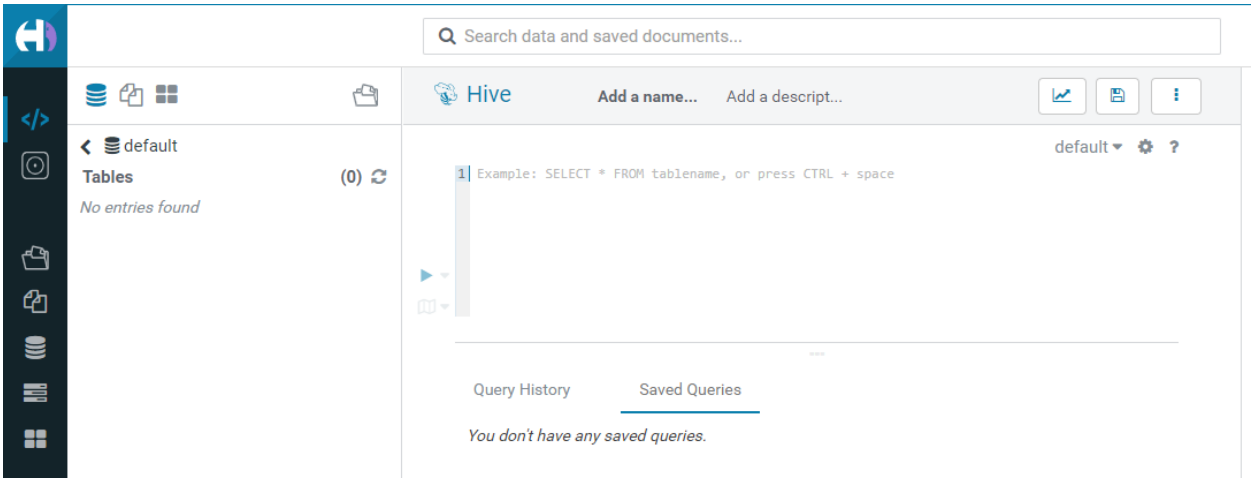
Step 14: Click instances, this will list the hostnames available under ZooKeeper



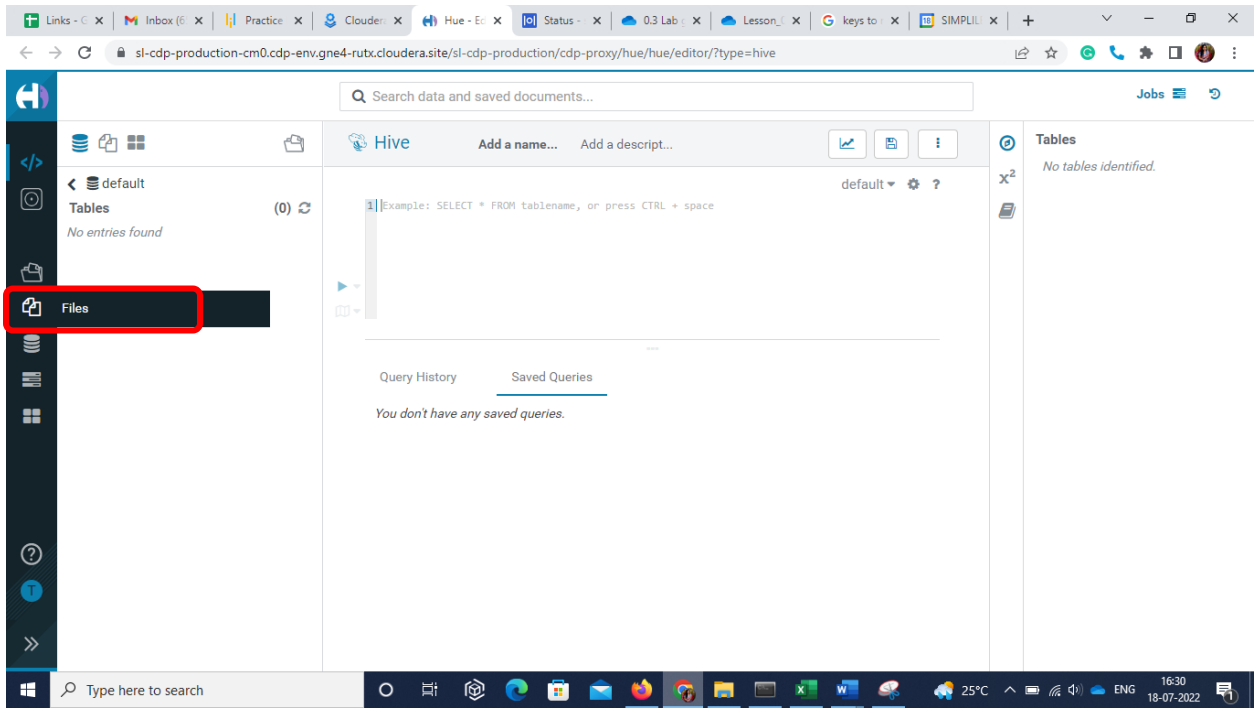
Step 15: Click on the **HUE** icon to log in



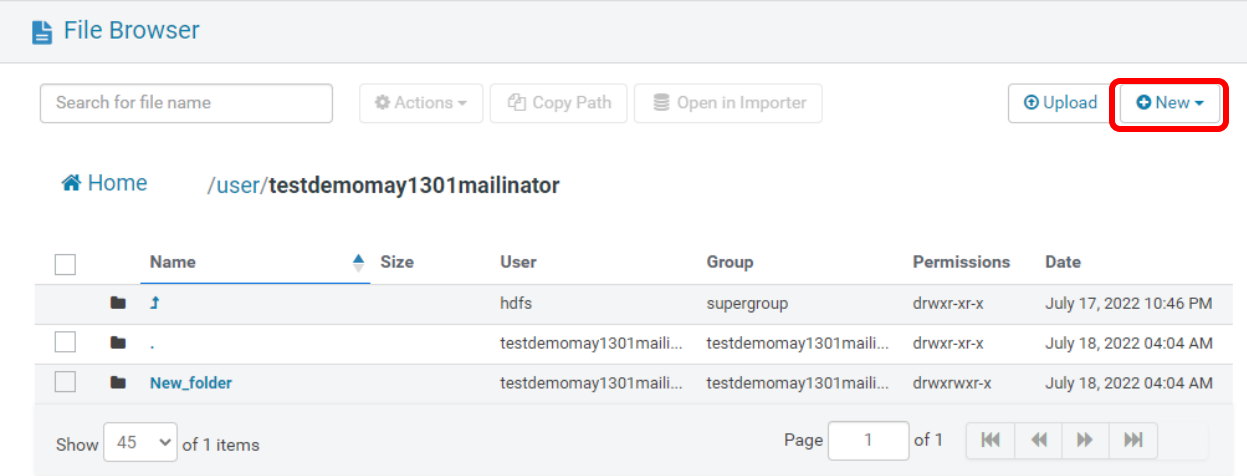
Step 16: You will be navigated to the dashboard as shown below.

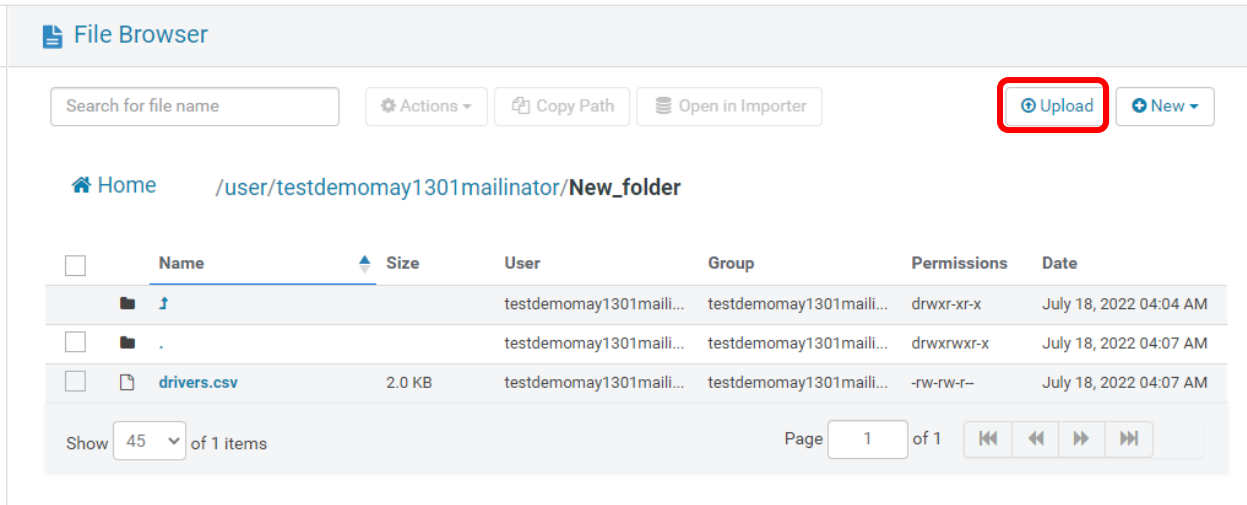


Step 17: To upload the dataset in **HUE** click on the icon as shown below:



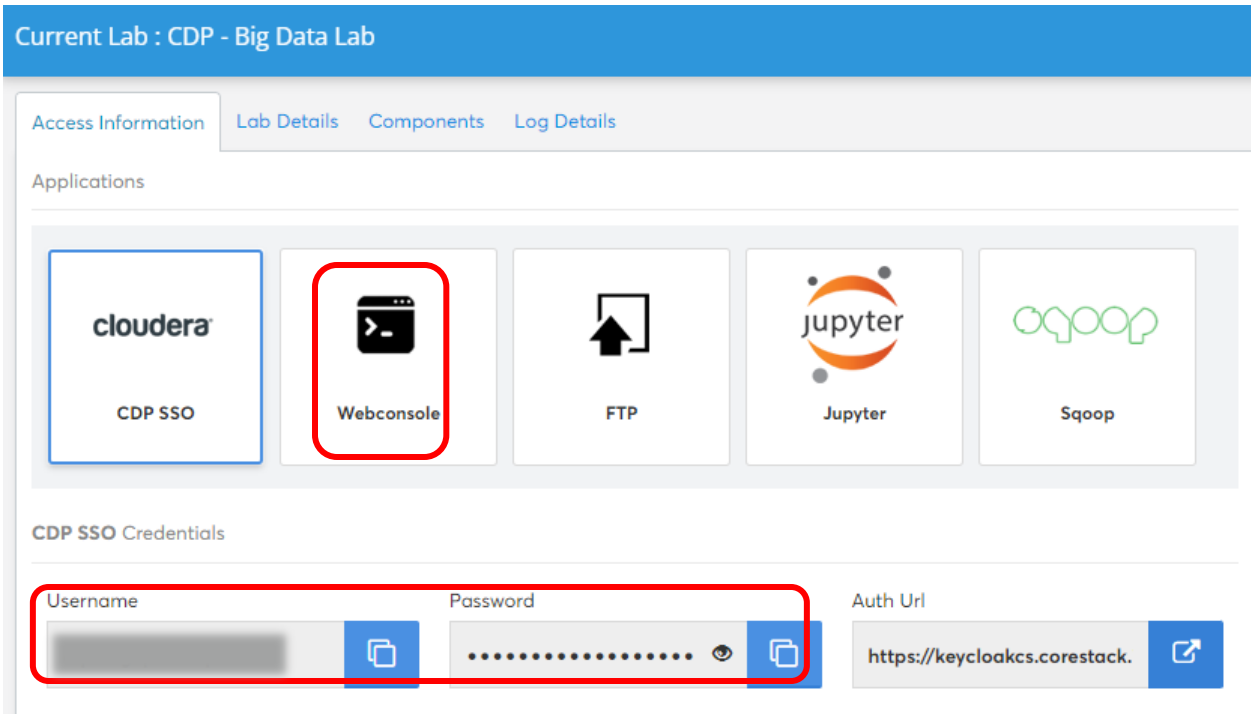
Step 18: Create a new directory by clicking on **New** and upload the dataset using the upload button as shown below:





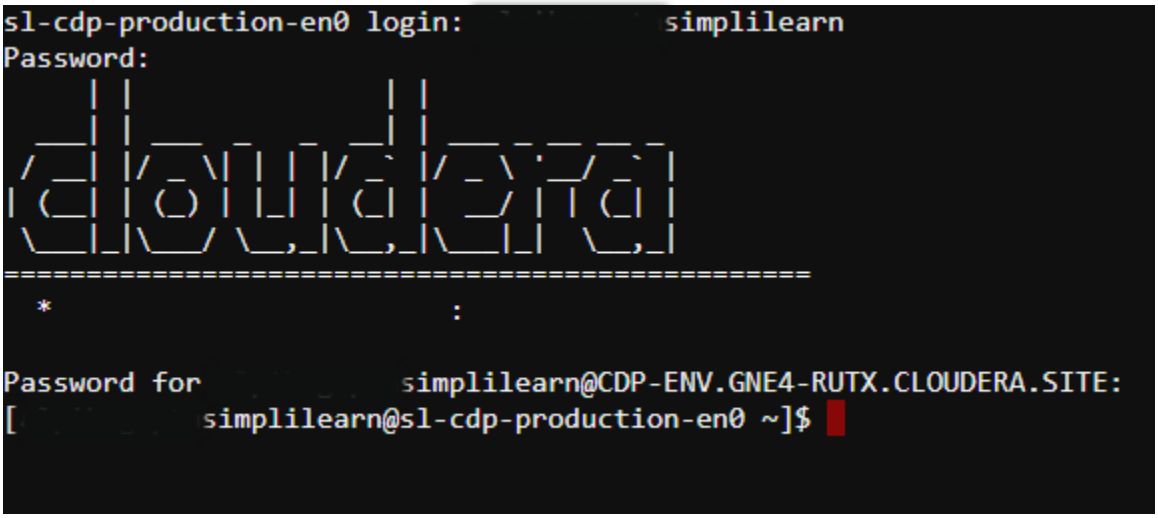
Step 19: Similarly, click on the Webconsole and navigate using **Auth Url**

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



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



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: Futur
  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

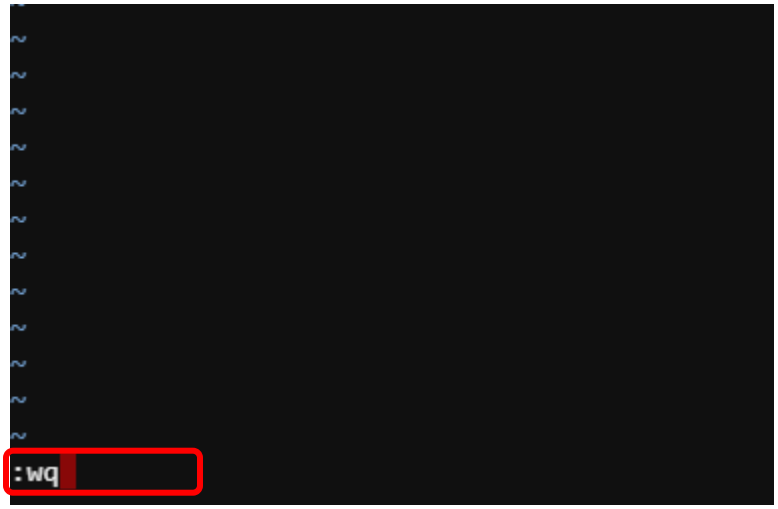
Or

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
```

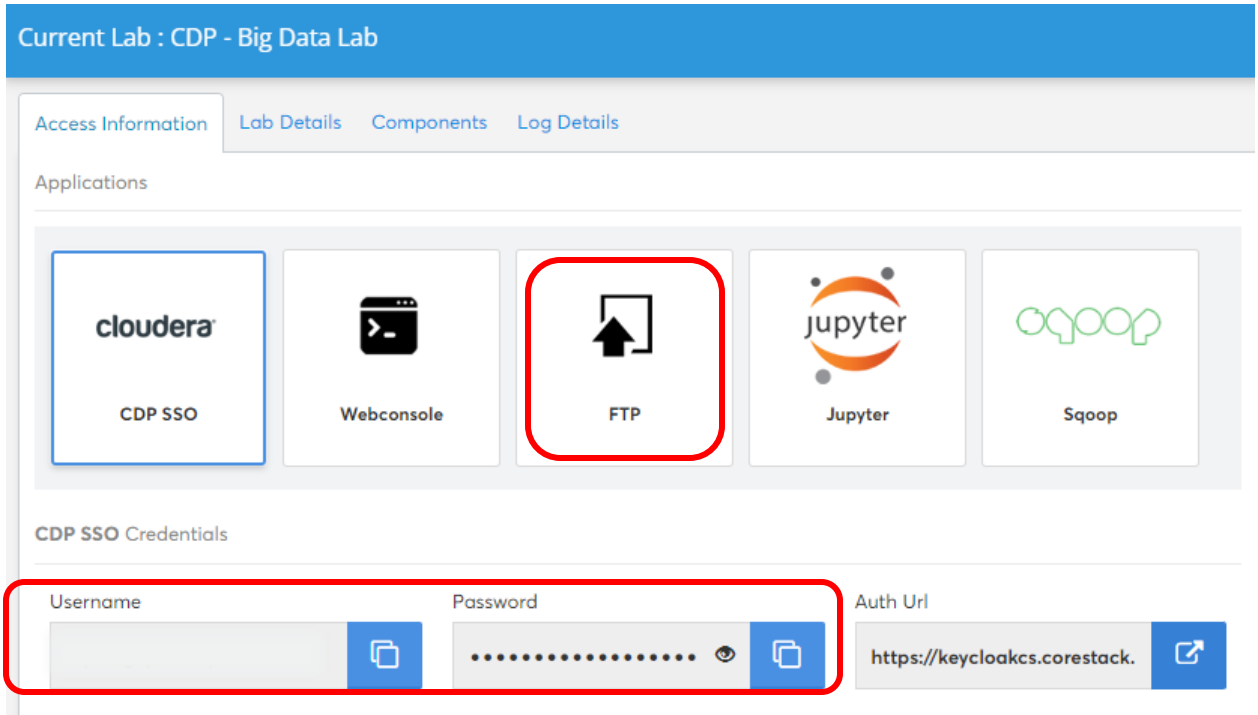
Step 29: To login into the Scala environment use the below command:

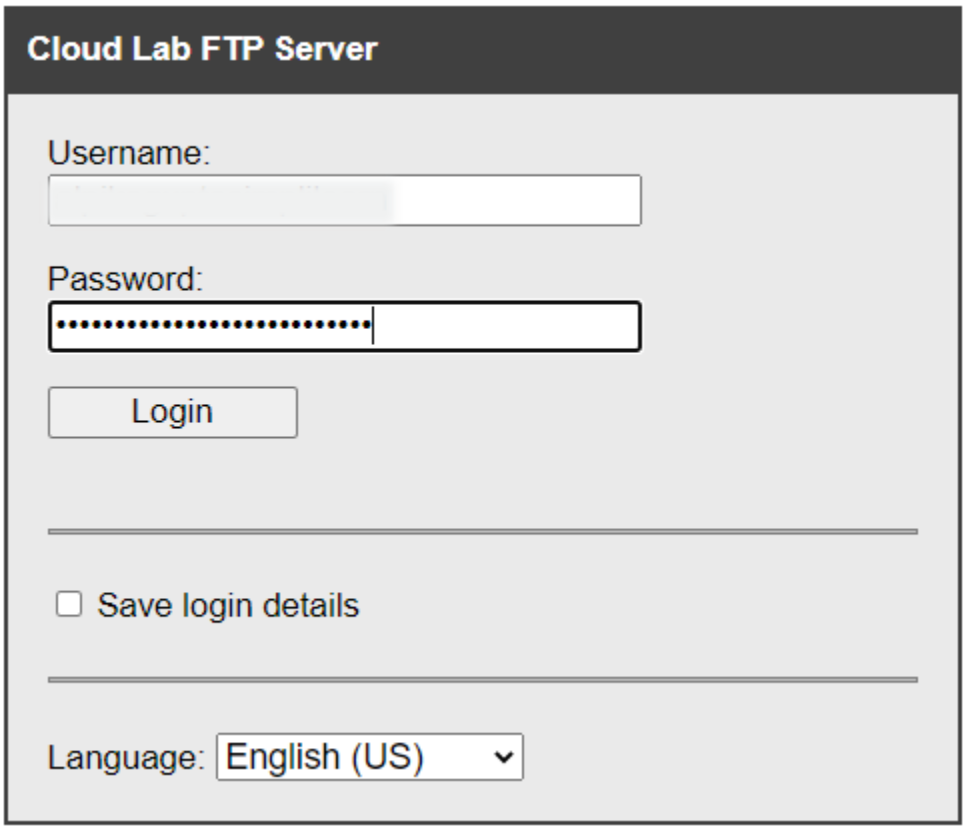
Command:

spark3-shell

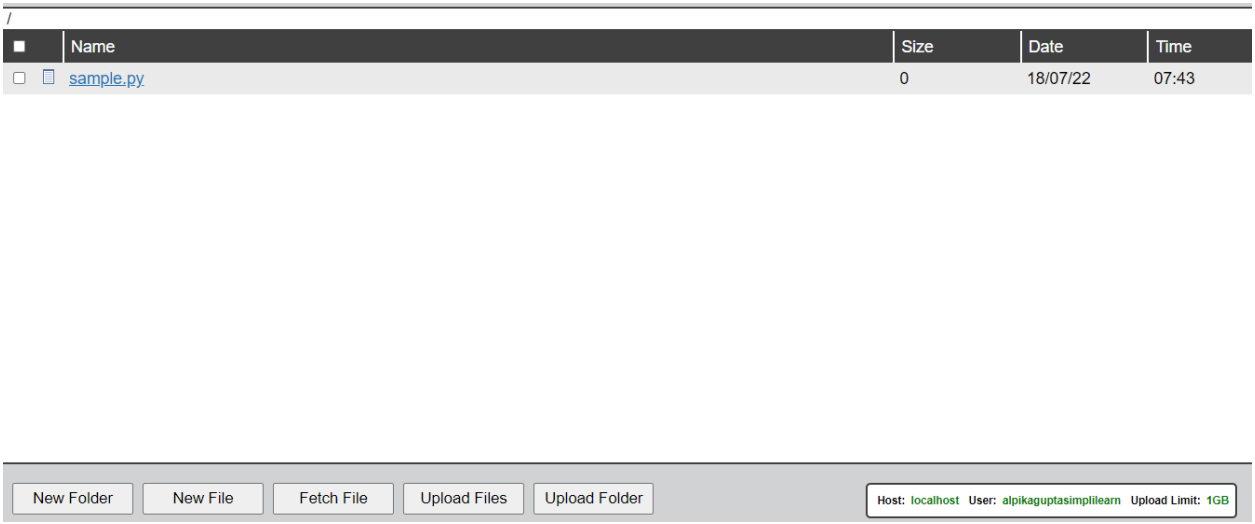
```
[            simplilearn@sl-cdp-production-en1 ~]$ spark3-shell
Setting default log level to "WARN".
To adjust logging level use sc.setLogLevel(newLevel). For SparkR, use setLogLevel(newLevel).
22/07/18 12:10:00 WARN util.Utils: Service 'SparkUI' could not bind on port 4040. Attempting port 4041.
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**

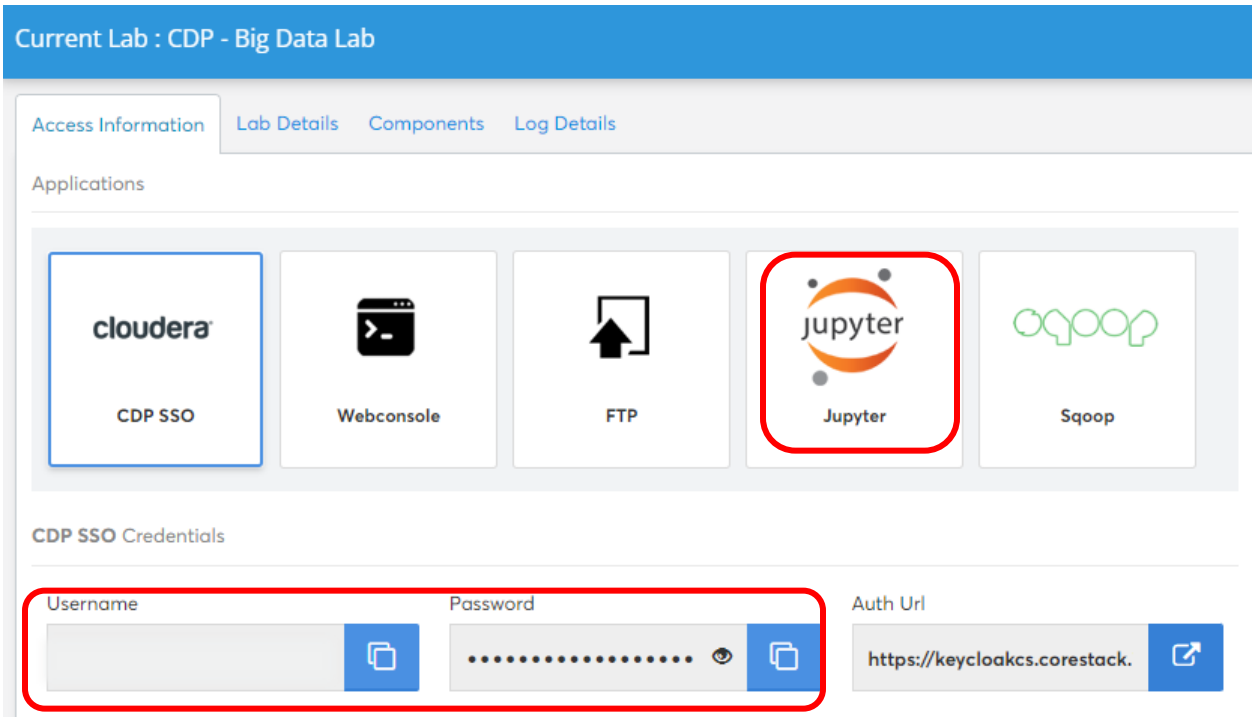


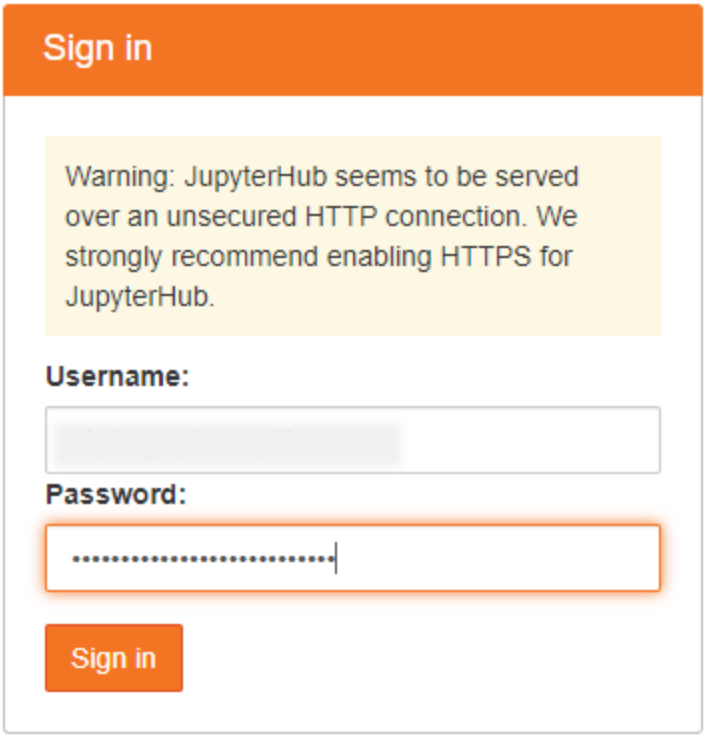


Step 31: You will be navigated to the screen as shown below:

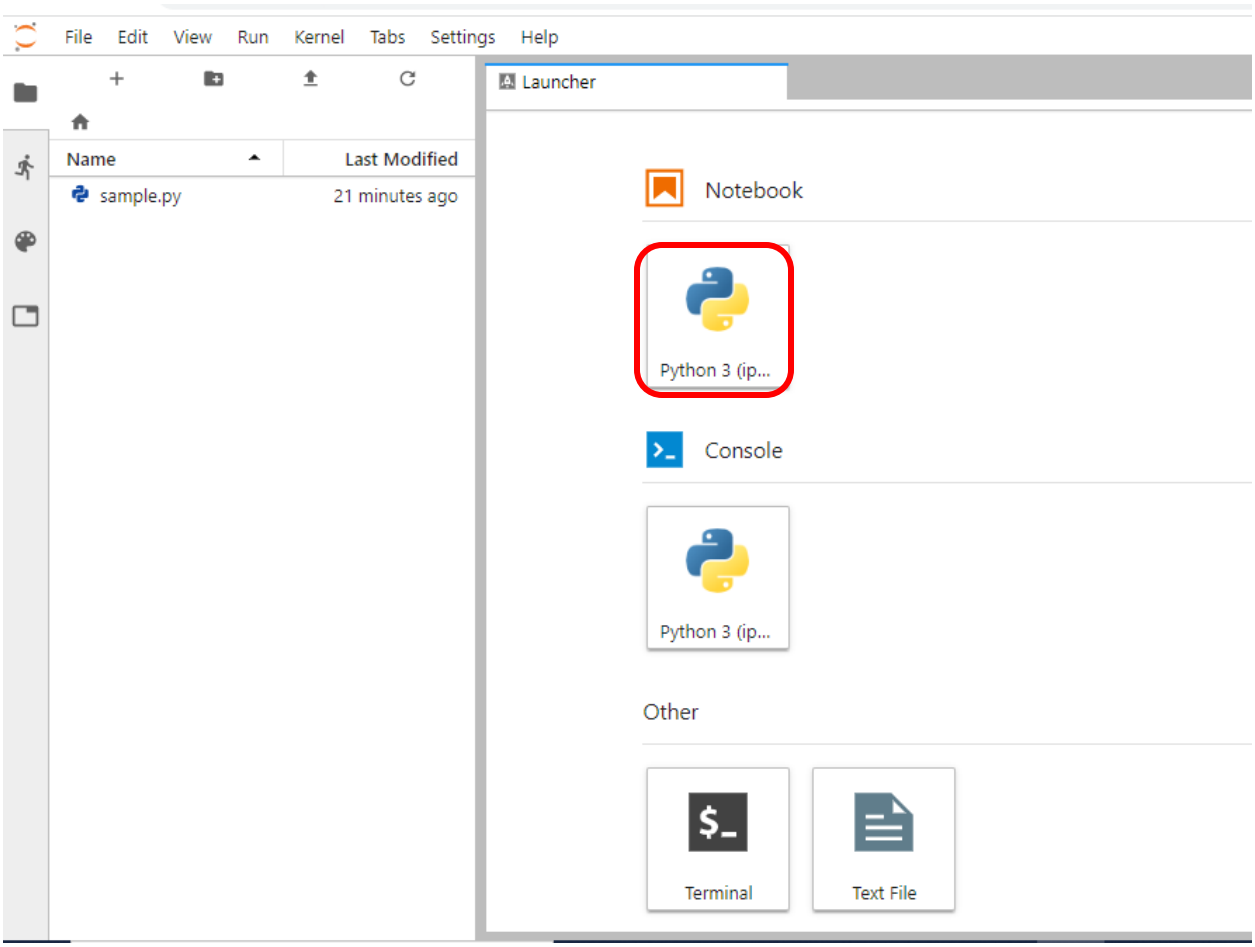


Step 32: Click on **Jupyter** and click on the **Auth Url** to log in and copy the **Username** and the **Password** provided to log in to the **Jupyter**





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