Big Data Hadoop and Spark Developer

Lab Guide

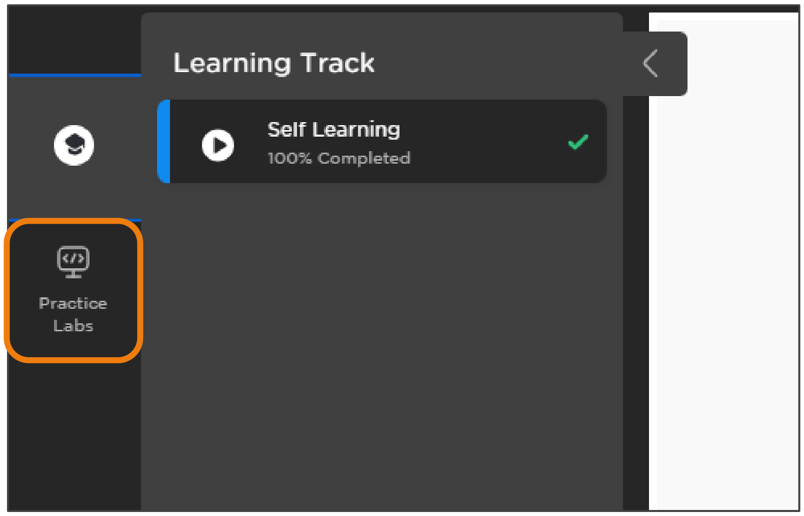
Get Certified. Get Ahead.

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

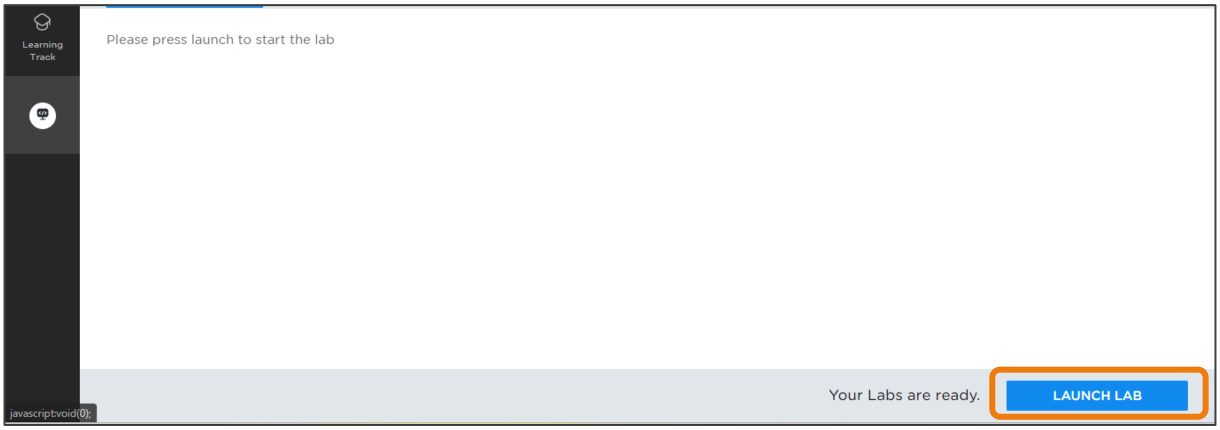
To execute the demos included in this course, follow the below steps:

Step 1: Log in to the Simplilearn LMS

- Go to the course
- Click on **Practice Labs** on the left



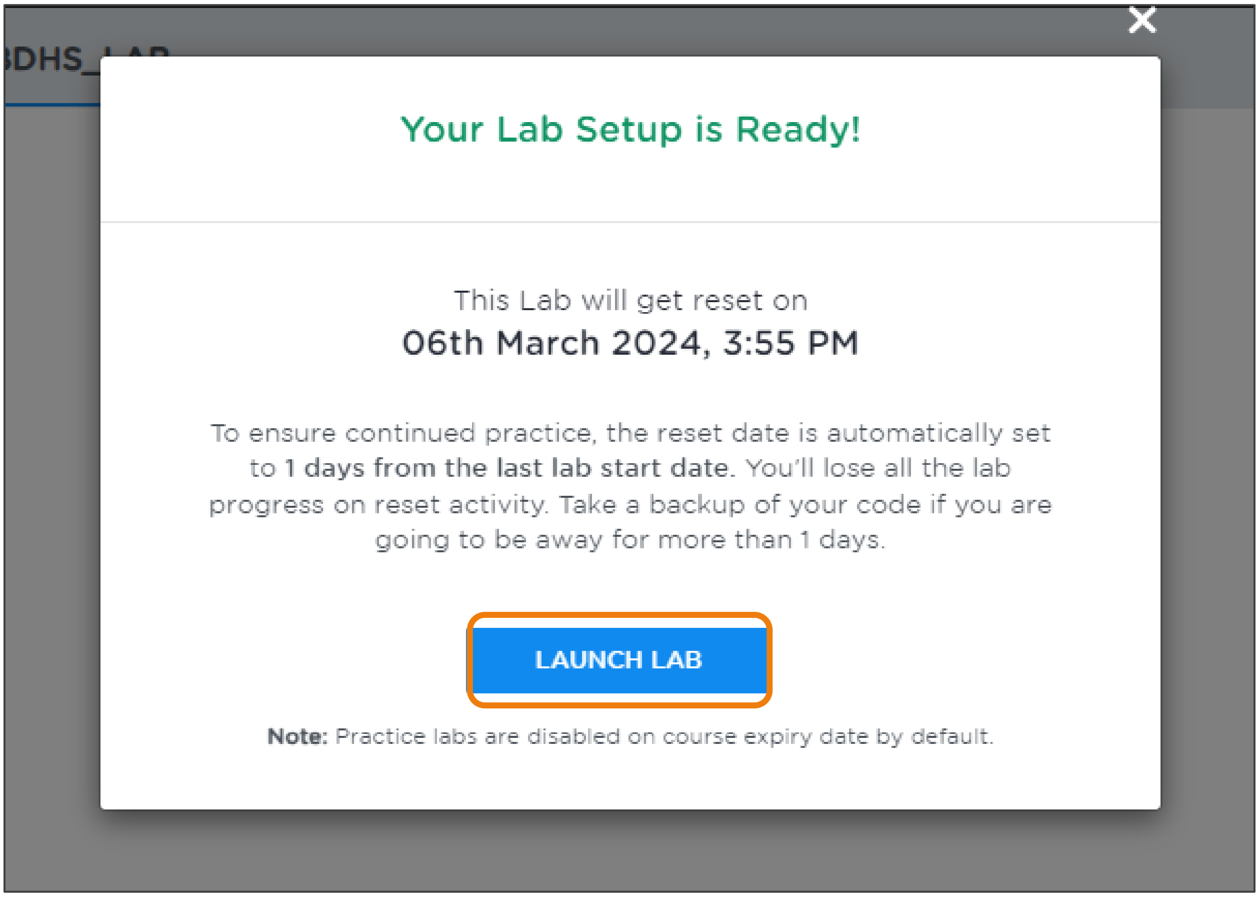
Step 2: Click on **LAUNCH LAB**



Please press launch to start the lab

Your Labs are ready.

Step 3: A small screen will pop up in the middle of your screen with important information about the lab. Again, click on the **LAUNCH LAB** button.



Your Lab Setup is Ready!

This Lab will get reset on

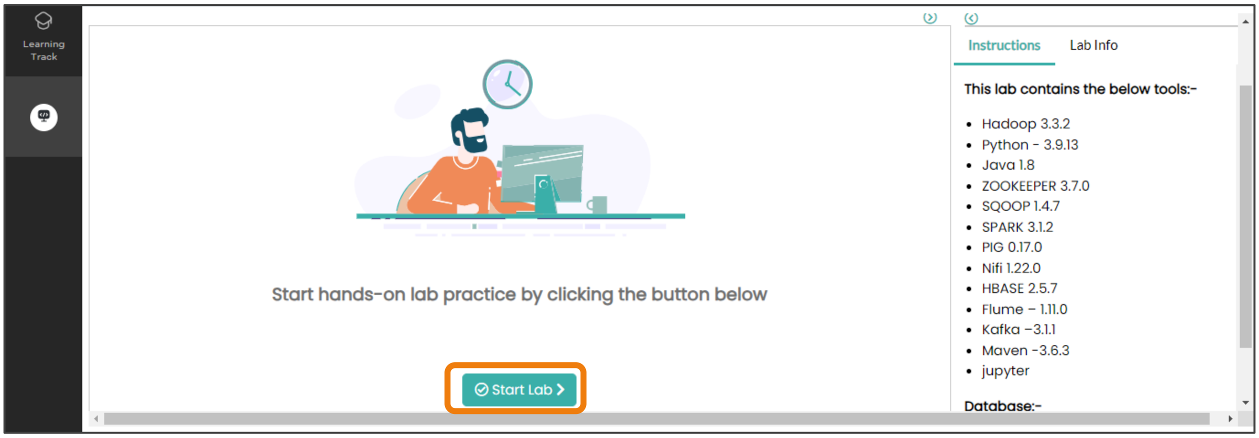
06th March 2024, 3:55 PM

To ensure continued practice, the reset date is automatically set to **1 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 1 days.

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

Note: It will take about three to five minutes for the lab environment to load.

Step 4: Click on the **Start Lab** button



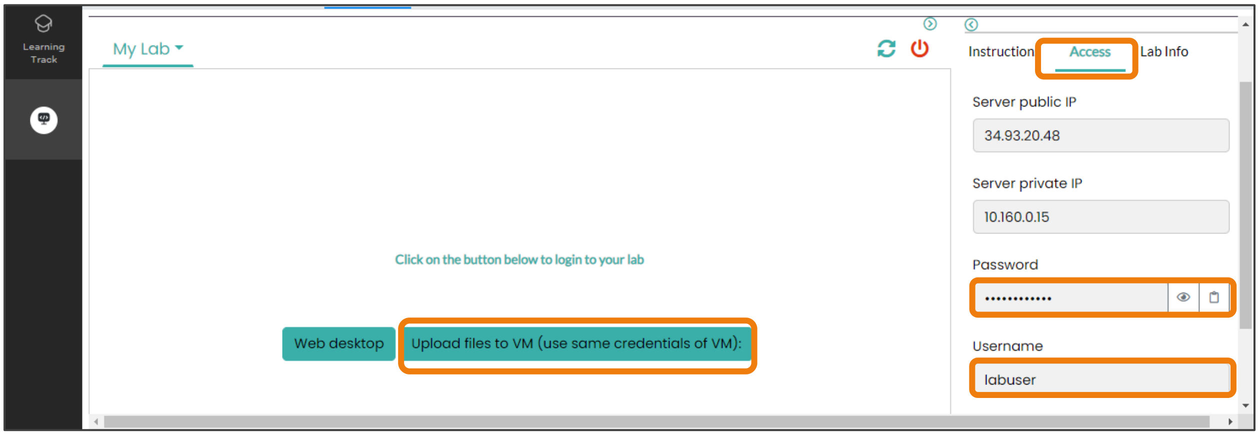
Start hands-on lab practice by clicking the button below

This lab contains the below tools:-

- Hadoop 3.3.2
- Python - 3.9.13
- Java 1.8
- ZOOKEEPER 3.7.0
- SQOOP 1.4.7
- SPARK 3.1.2
- PIG 0.17.0
- Nifi 1.22.0
- HBASE 2.5.7
- Flume – 1.11.0
- Kafka –3.1.1
- Maven -3.6.3
- jupyter

Database:-

Step 5: Once the environment has loaded, log in to the **FTP (Upload files to VM)** using the specified **Username** and **Password** from the **Access** tab



Click on the button below to login to your lab

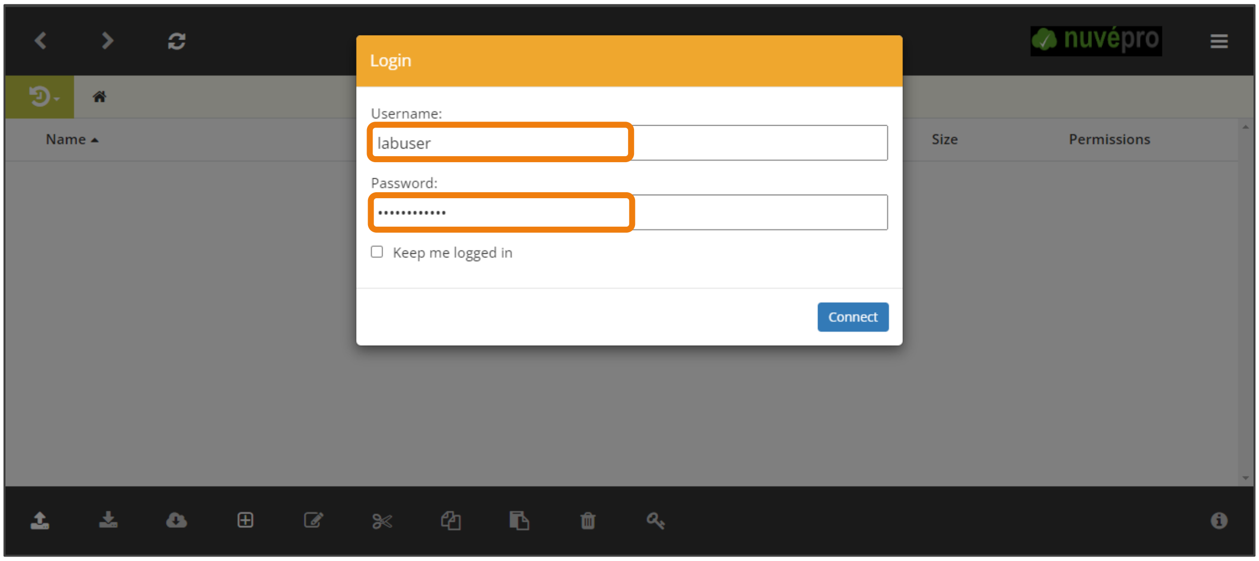
Server public IP: 34.93.20.48

Server private IP: 10.160.0.15

Password: ••••••••••••

Username: labuser

Step 6: You will now be directed to the login screen where you can enter your **Username** and **Password**

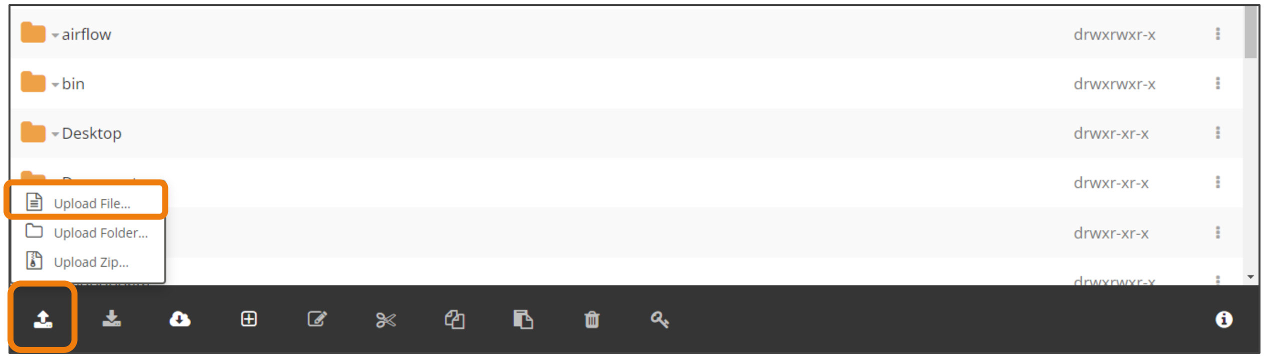


Username: labuser

Password: ••••••••••••

Note: Once you are successfully logged in, you will be redirected to the below page as shown in Step 7.

Step 7: Click on the **Upload File** button to upload the dataset



airflow drwxrwxr-x

bin drwxrwxr-x

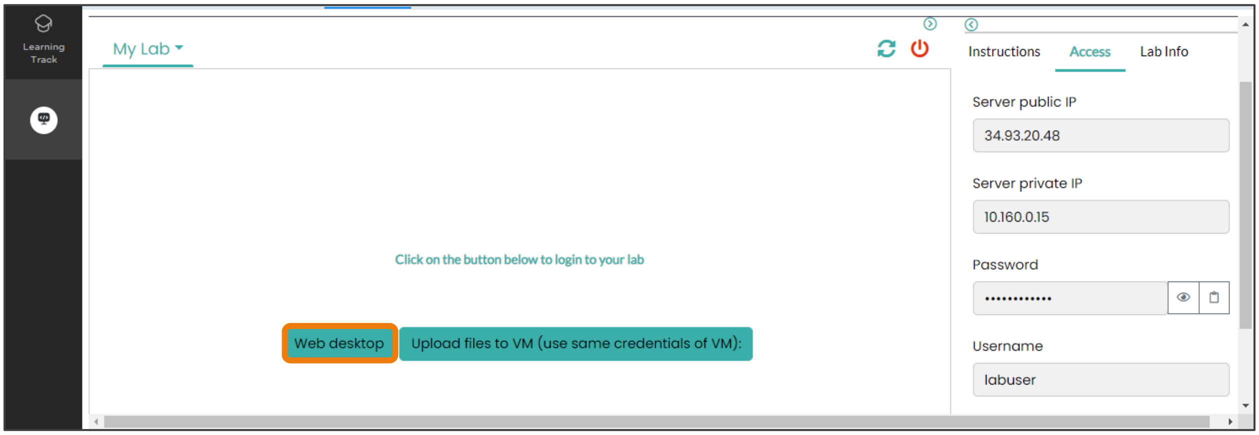
Desktop drwxr-xr-x

drwxr-xr-x

drwxr-xr-x

Note: Datasets are accessible in the lab for online use. Alternatively, for offline work, you can download and extract them from the **Reference Materials** in your LMS.

Step 8: To access the terminal, HUE, or any other provided service, click on **Web desktop**



Click on the button below to login to your lab

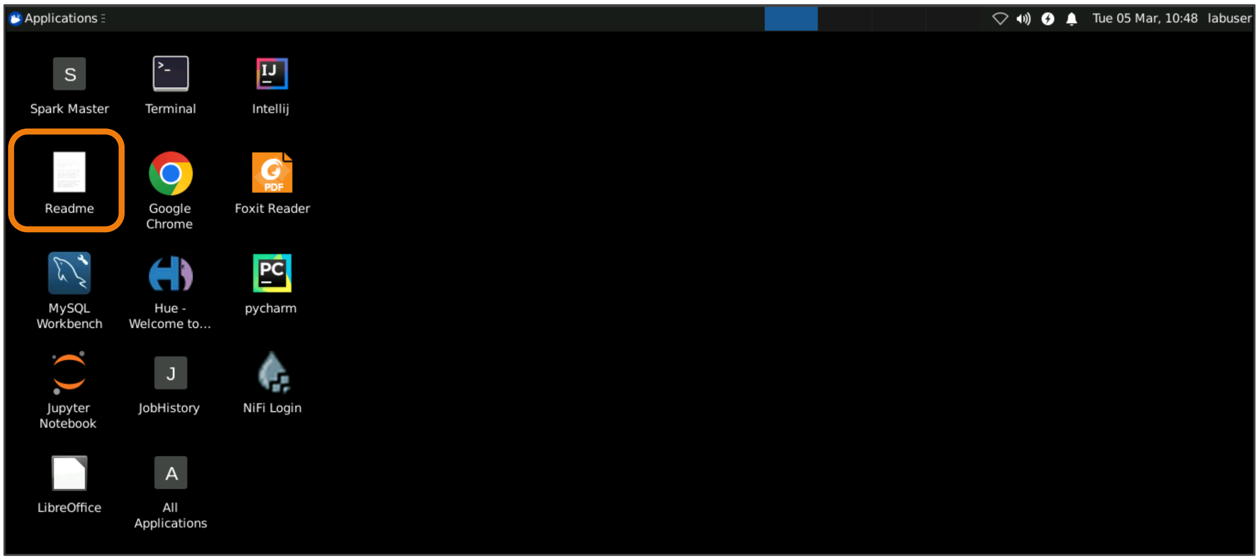
Server public IP: 34.93.20.48

Server private IP: 10.160.0.15

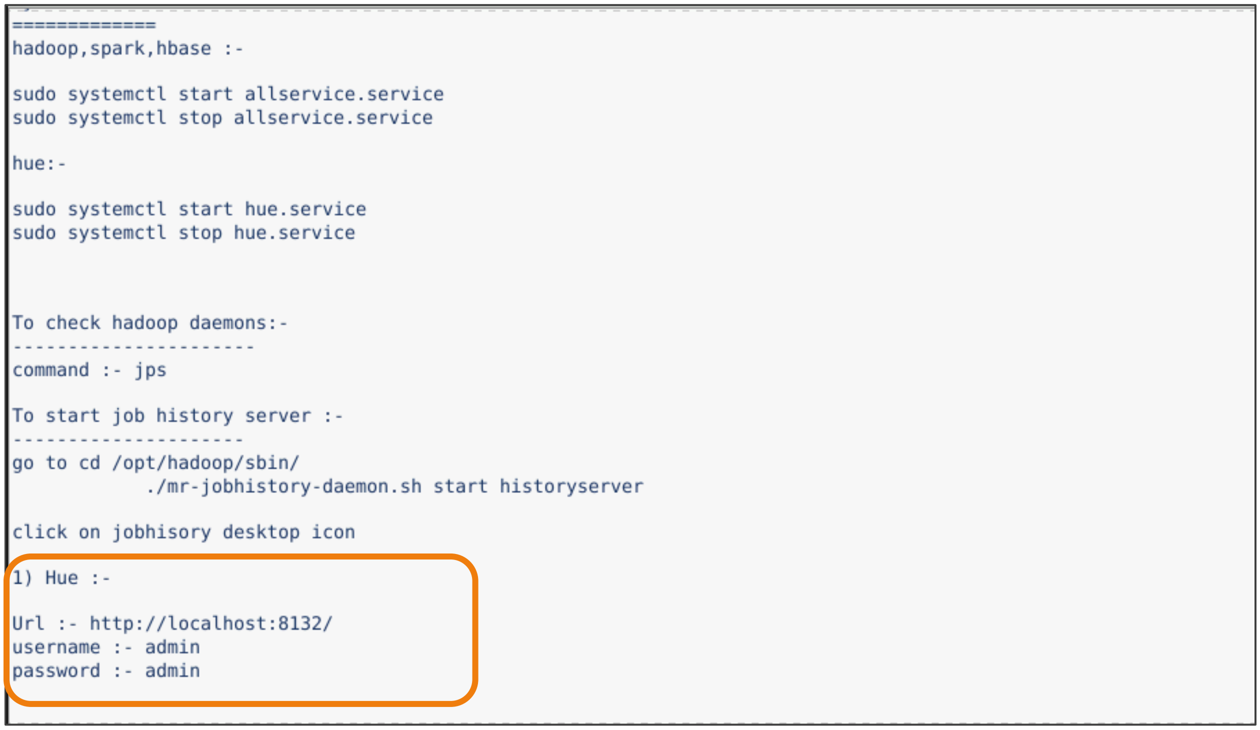
Password: ••••••••••••

Username: labuser

Step 9: Click on the **Readme** file to discover the installed tools, their URL/path, and the Username and Password for different services

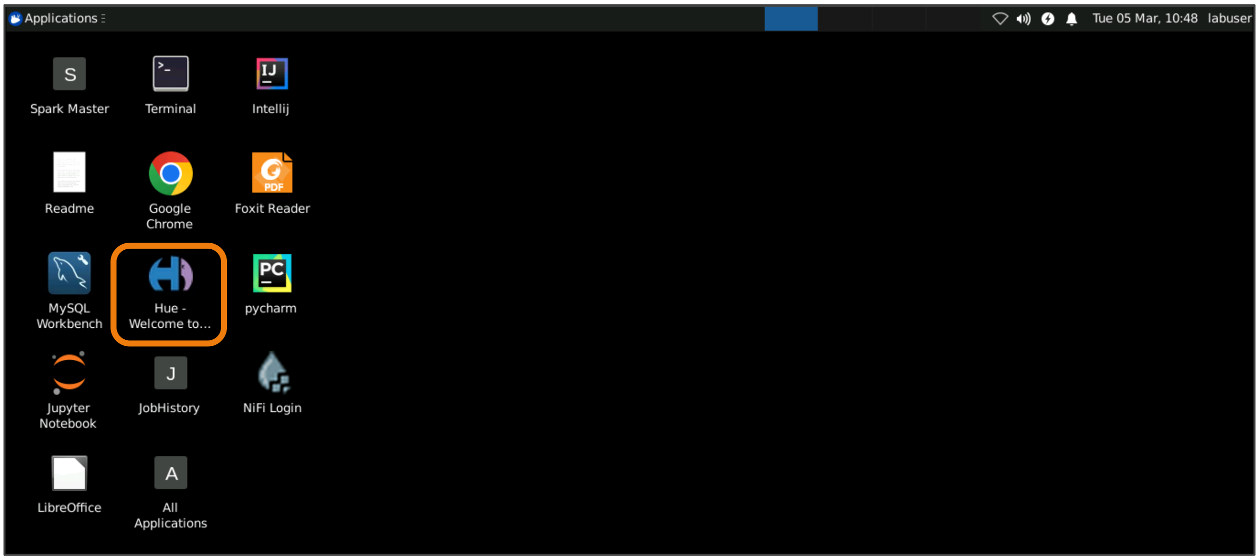


Step 10: Copy the **Username** and **Password** provided in Readme file to log in to **HUE**

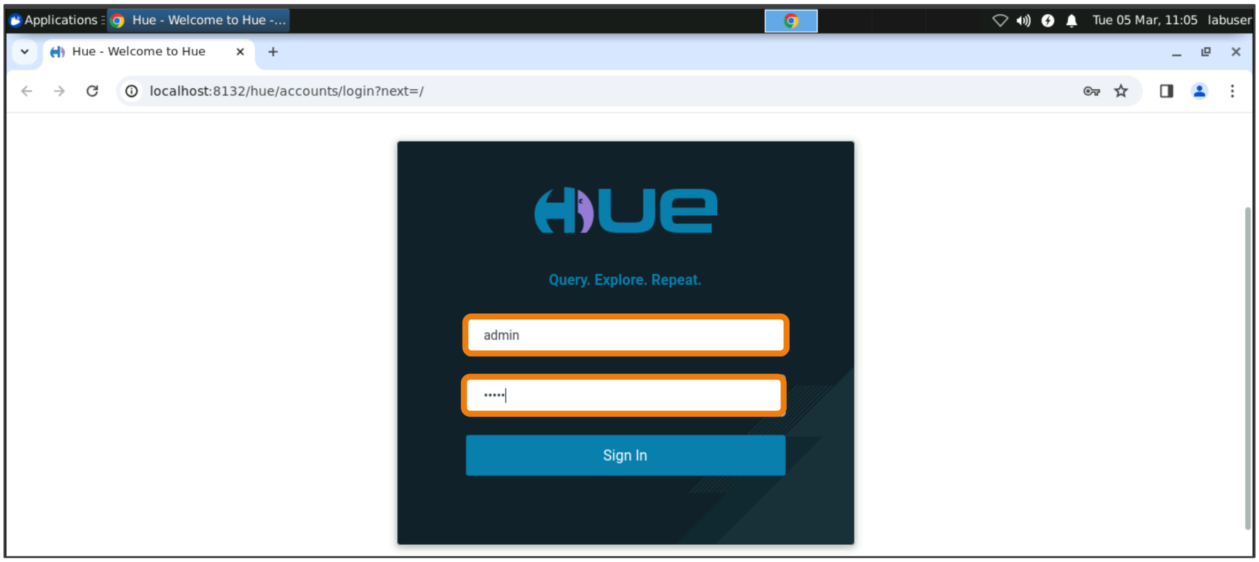


```
=============
hadoop,spark,hbase :-

sudo systemctl start allservice.service
sudo systemctl stop allservice.service

hue:-

sudo systemctl start hue.service
sudo systemctl stop hue.service


To check hadoop daemons:-
----------------------
command :- jps

To start job history server :-
---------------------
go to cd /opt/hadoop/sbin/
            ./mr-jobhistory-daemon.sh start historyserver

click on jobhisory desktop icon

1) Hue :-

Url :- http://localhost:8132/
username :- admin
password :- admin
```

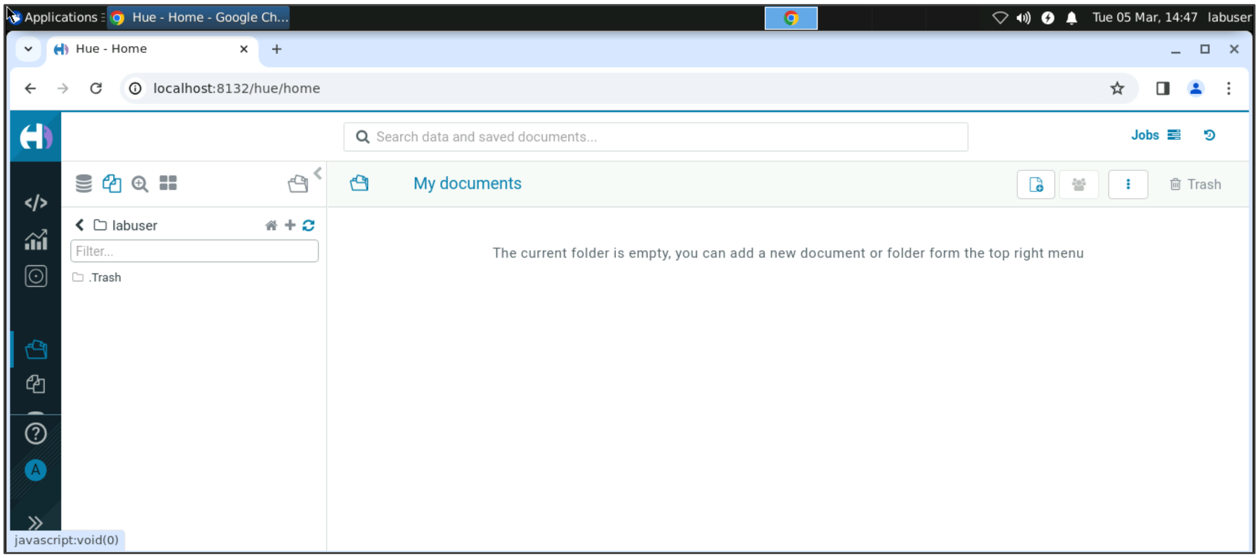
Step 11: Click on **HUE**



Step 12: Enter the **Username** and **Password** you copied to log in to **HUE**

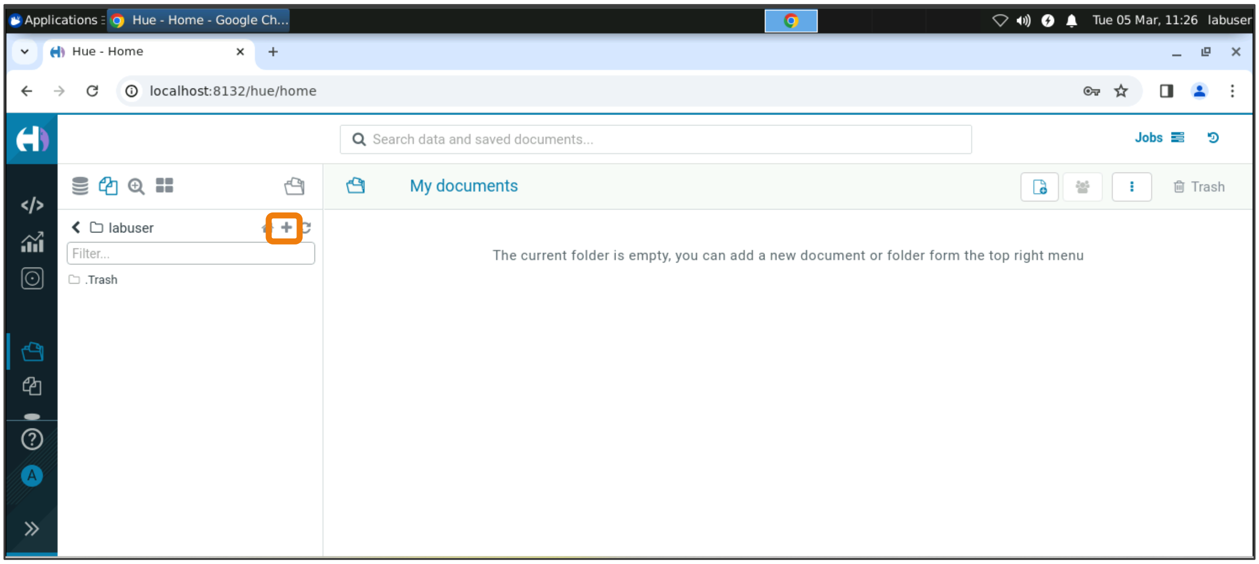

Applications
Hue - Welcome to Hue -...
Tue 05 Mar, 11:05 labuser
Hue - Welcome to Hue
localhost:8132/hue/accounts/login?next=/
HUE
Query. Explore. Repeat.
admin
•••••
Sign In

Note: You will be navigated to the dashboard as shown below:



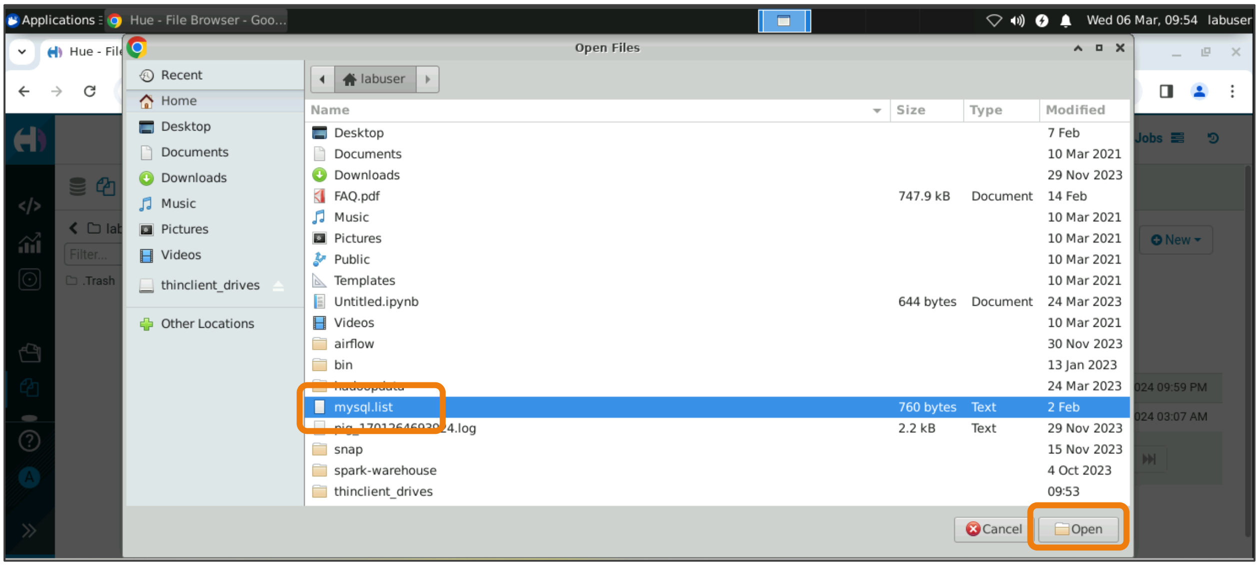
The current folder is empty, you can add a new document or folder form the top right menu

Step 13: Click on the shown **+** icon to upload the dataset in HUE



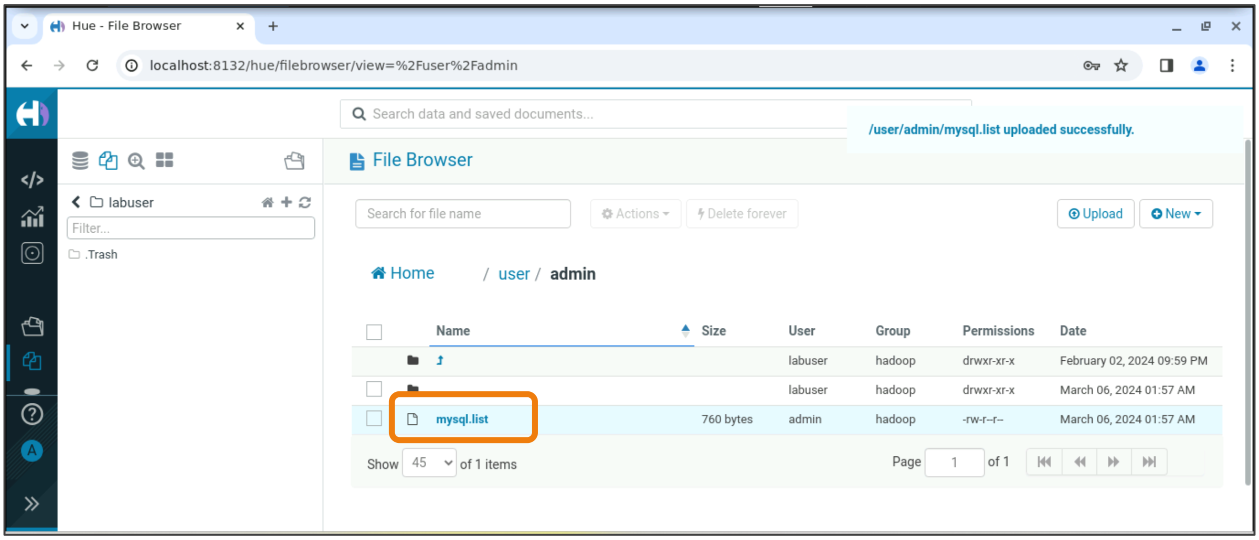
The current folder is empty, you can add a new document or folder form the top right menu

Step 14: Choose the dataset you want to upload, and then click on the **Open** button



Name	Size	Type	Modified
Desktop			7 Feb
Documents			10 Mar 2021
Downloads			29 Nov 2023
FAQ.pdf	747.9 kB	Document	14 Feb
Music			10 Mar 2021
Pictures			10 Mar 2021
Public			10 Mar 2021
Templates			10 Mar 2021
Untitled.ipynb	644 bytes	Document	24 Mar 2023
Videos			10 Mar 2021
airflow			30 Nov 2023
bin			13 Jan 2023
hadoopdata			24 Mar 2023
mysql.list	760 bytes	Text	2 Feb
pig_1701264693924.log	2.2 kB	Text	29 Nov 2023
snap			15 Nov 2023
spark-warehouse			4 Oct 2023
thinclient_drives			09:53

Note: After clicking the **Open** button, you will be able to see the uploaded dataset as shown below:



Hue - File Browser

localhost:8132/hue/filebrowser/view=%2Fuser%2Fadmin

/user/admin/mysql.list uploaded successfully.

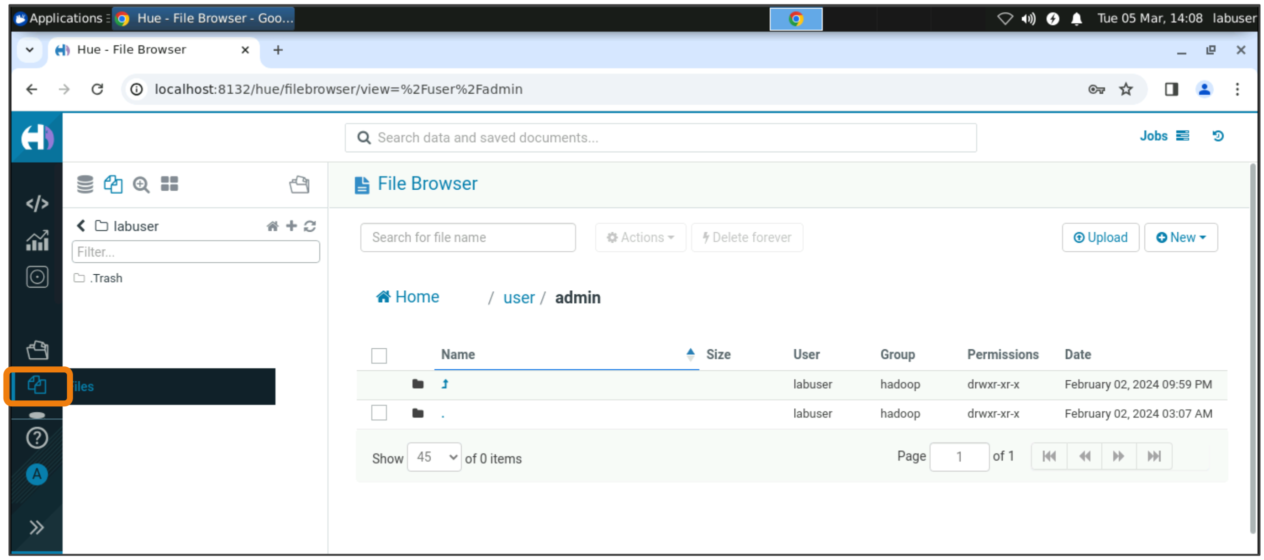
File Browser

Home / user / admin

	Name	Size	User	Group	Permissions	Date
	↥		labuser	hadoop	drwxr-xr-x	February 02, 2024 09:59 PM
□			labuser	hadoop	drwxr-xr-x	March 06, 2024 01:57 AM
□	mysql.list	760 bytes	admin	hadoop	-rw-r--r--	March 06, 2024 01:57 AM

Show 45 of 1 items   Page 1 of 1

Step 15: Click on the **Files** button to create a new file or directory within HUE



Applications   Hue - File Browser - Goo...   Tue 05 Mar, 14:08   labuser

Hue - File Browser

localhost:8132/hue/filebrowser/view=%2Fuser%2Fadmin

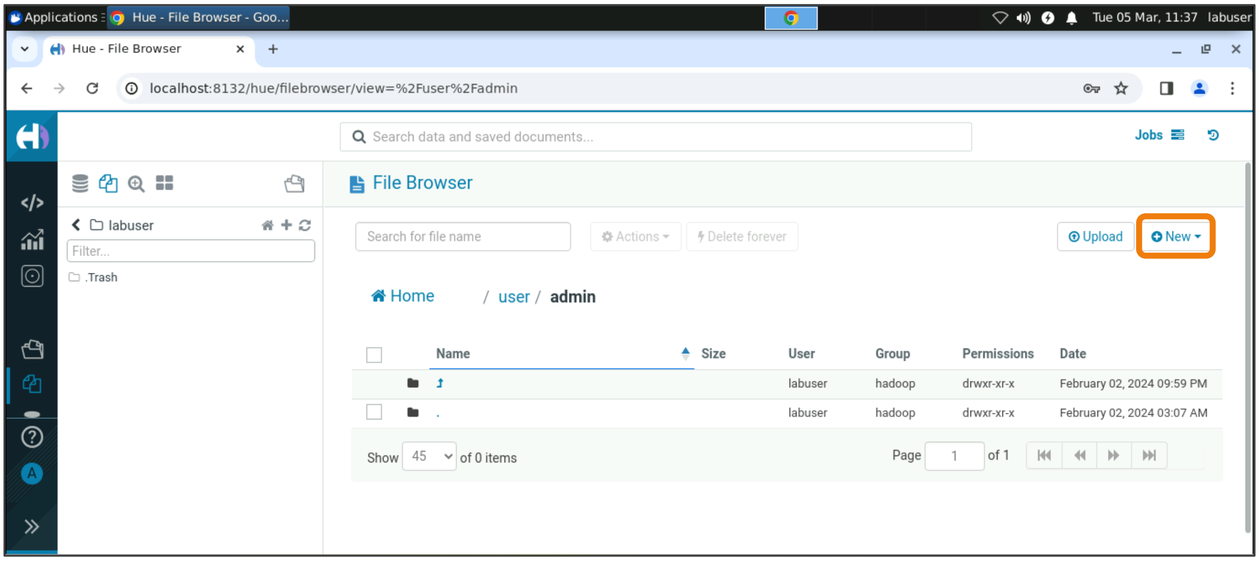
File Browser

Home / user / admin

	Name	Size	User	Group	Permissions	Date
	↥		labuser	hadoop	drwxr-xr-x	February 02, 2024 09:59 PM
□	.		labuser	hadoop	drwxr-xr-x	February 02, 2024 03:07 AM

Show 45 of 0 items   Page 1 of 1

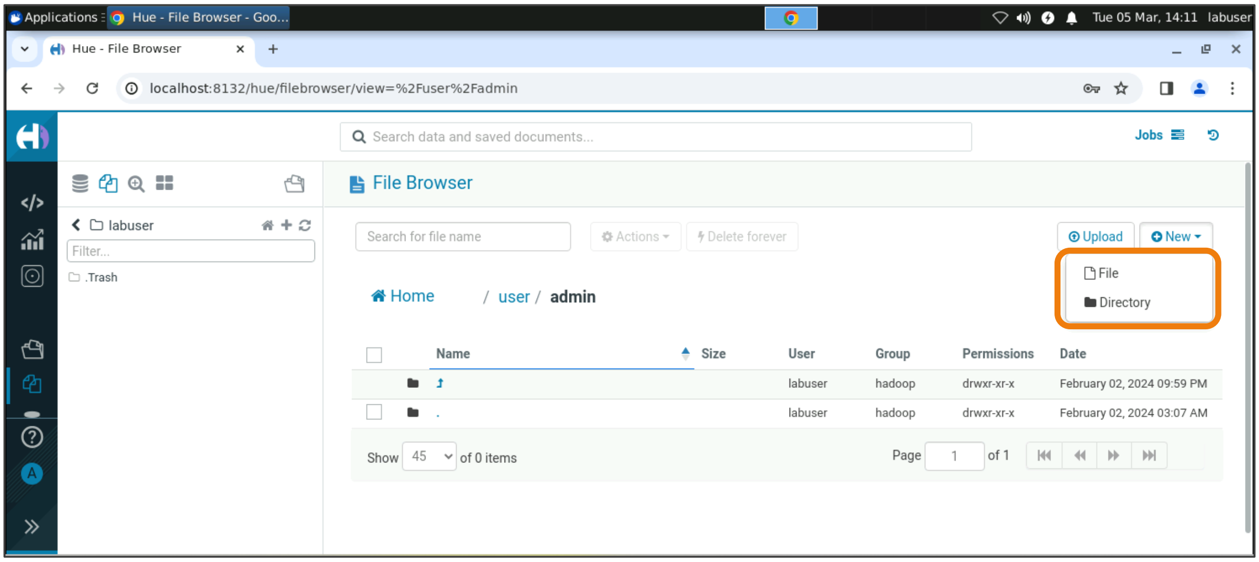
Step 16: Click on the **New** button


Applications
Hue - File Browser - Goo...
Tue 05 Mar, 11:37 labuser
Hue - File Browser
localhost:8132/hue/filebrowser/view=%2Fuser%2Fadmin
Search data and saved documents...
Jobs
File Browser
labuser
Filter...
.Trash
Search for file name
Actions
Delete forever
Upload
New
Home / user / admin

	Name	Size	User	Group	Permissions	Date
	↥		labuser	hadoop	drwxr-xr-x	February 02, 2024 09:59 PM
□	.		labuser	hadoop	drwxr-xr-x	February 02, 2024 03:07 AM

Show 45 of 0 items Page 1 of 1

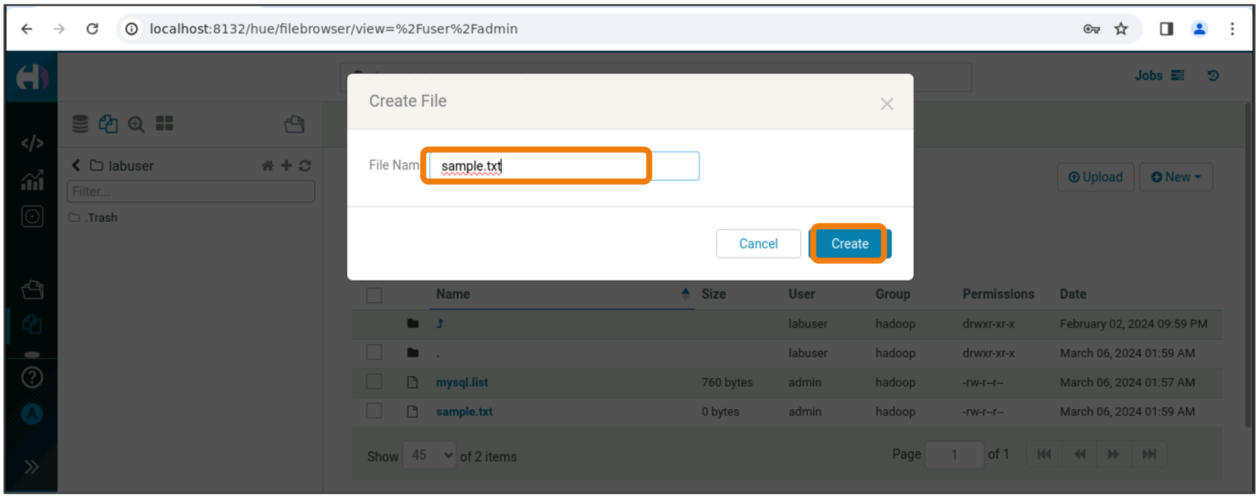
Step 17: Click on either **File** or **Directory** based on your needs



	Name	Size	User	Group	Permissions	Date
	↥		labuser	hadoop	drwxr-xr-x	February 02, 2024 09:59 PM
□	.		labuser	hadoop	drwxr-xr-x	February 02, 2024 03:07 AM

Show 45 of 0 items Page 1 of 1

Step 18: Enter the file name and click on the **Create** button



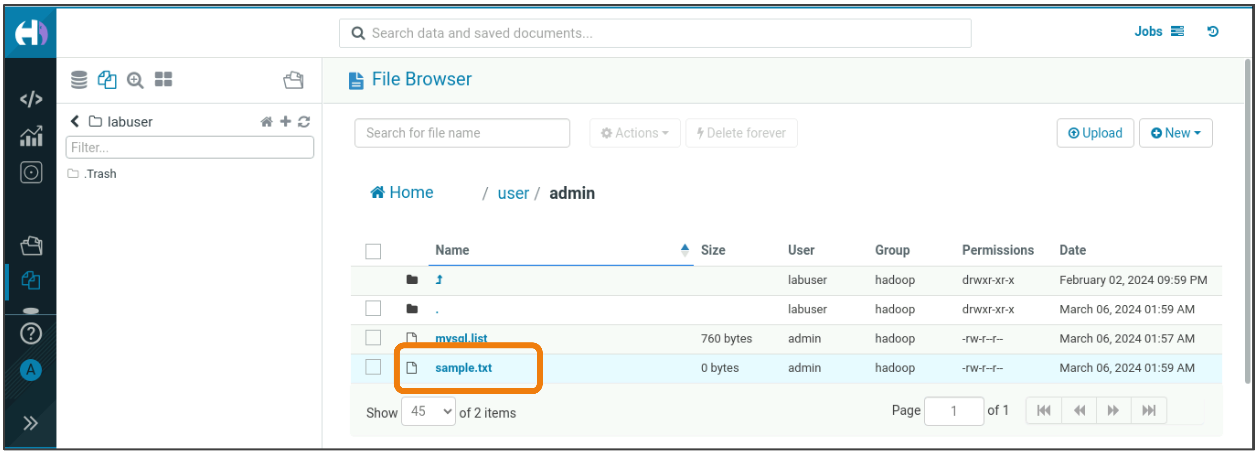
Create File

File Name: sample.txt

	Name	Size	User	Group	Permissions	Date
	↥		labuser	hadoop	drwxr-xr-x	February 02, 2024 09:59 PM
□	.		labuser	hadoop	drwxr-xr-x	March 06, 2024 01:59 AM
□	mysql.list	760 bytes	admin	hadoop	-rw-r--r--	March 06, 2024 01:57 AM
□	sample.txt	0 bytes	admin	hadoop	-rw-r--r--	March 06, 2024 01:59 AM

Show 45 of 2 items Page 1 of 1

Note: After clicking the **Create** button, you will be able to see the new file as shown below:

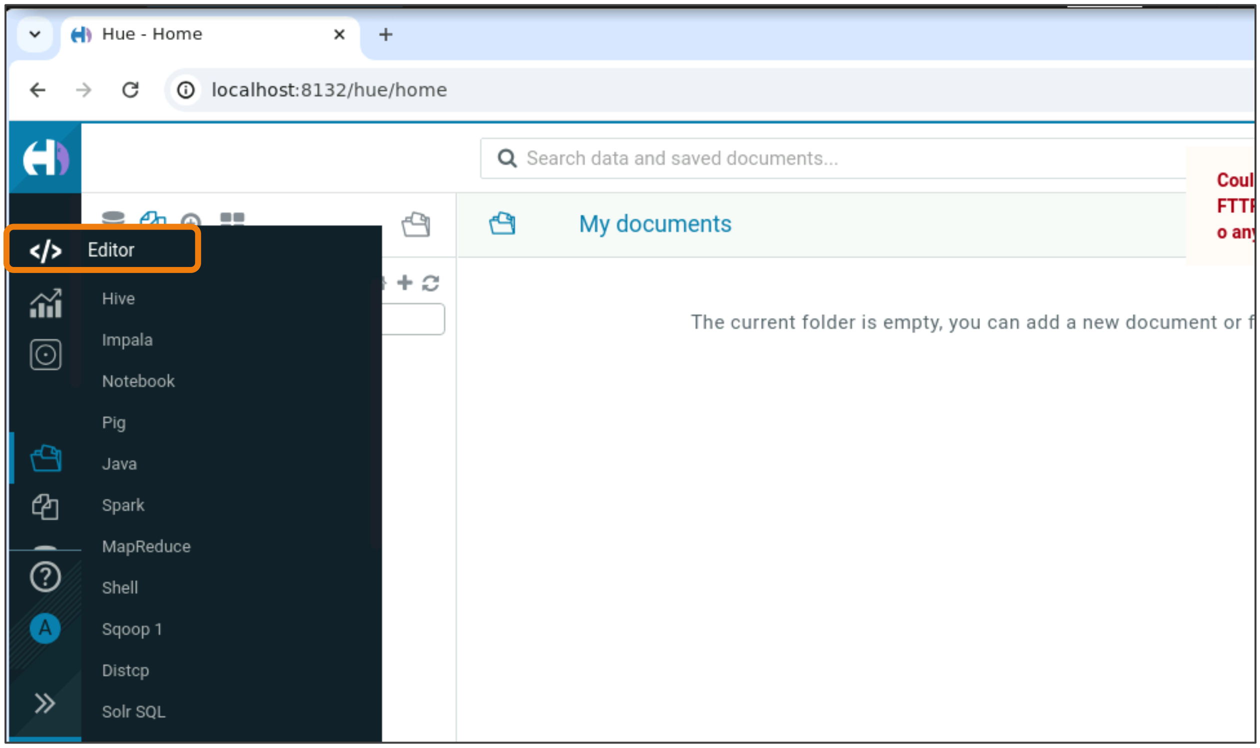


Home / user / admin

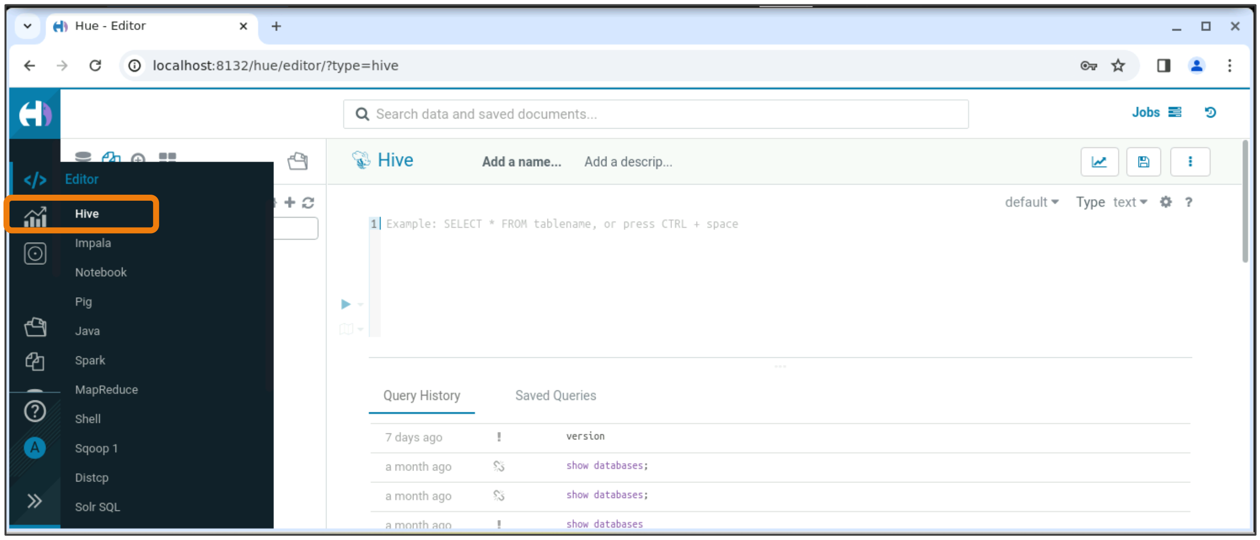
□	Name	Size	User	Group	Permissions	Date
	↥		labuser	hadoop	drwxr-xr-x	February 02, 2024 09:59 PM
□	.		labuser	hadoop	drwxr-xr-x	March 06, 2024 01:59 AM
□	mysql.list	760 bytes	admin	hadoop	-rw-r--r--	March 06, 2024 01:57 AM
□	sample.txt	0 bytes	admin	hadoop	-rw-r--r--	March 06, 2024 01:59 AM

Show 45 of 2 items Page 1 of 1

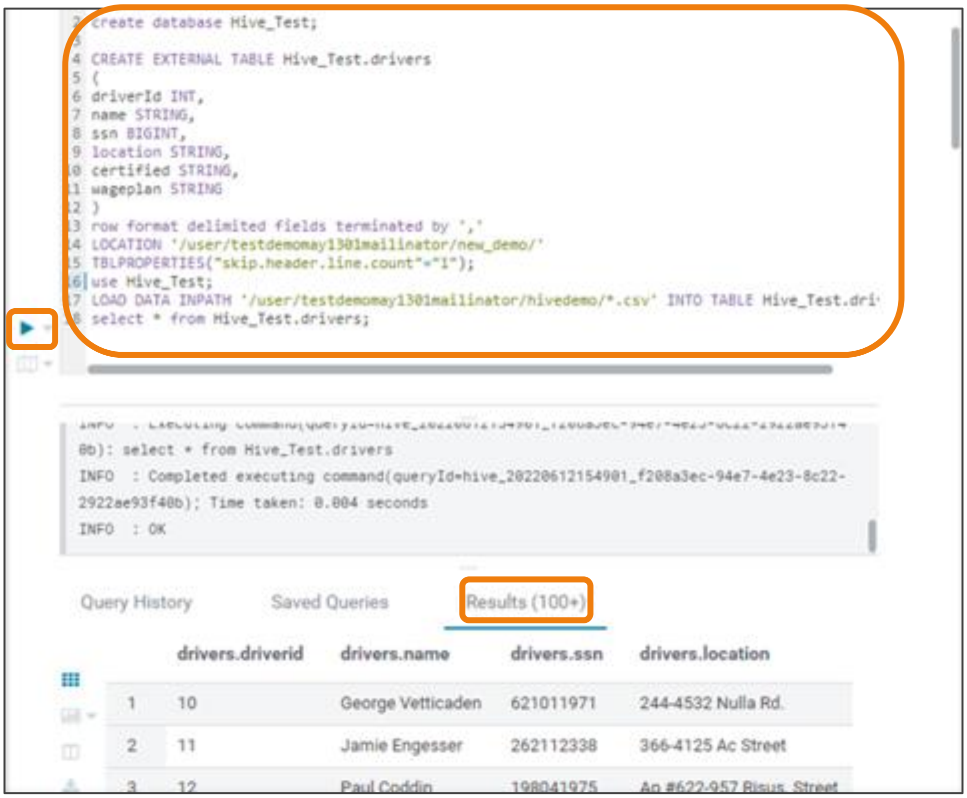
Step 19: Select the **Editor** button situated on the left side of the HUE dashboard to compose the **Hive query**



Step 20: Click on **Hive**



Step 21: Type the command into the editor, and then click on the **triangle** button located on the left side of the editor to execute the command. The output will be displayed in the **Results** column.

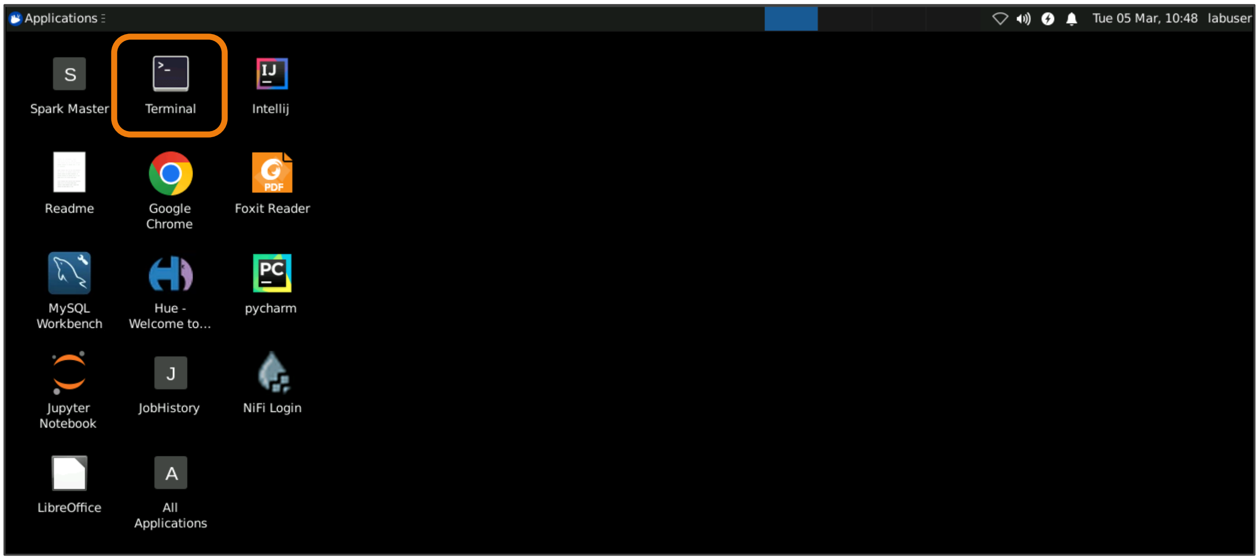


```
create database Hive_Test;

CREATE EXTERNAL TABLE Hive_Test.drivers
(
driverId INT,
name STRING,
ssn BIGINT,
location STRING,
certified STRING,
wageplan STRING
)
row format delimited fields terminated by ','
LOCATION '/user/testdemomay1301mailinator/new_demo/'
TBLPROPERTIES("skip.header.line.count"="1");
use Hive_Test;
LOAD DATA INPATH '/user/testdemomay1301mailinator/hivedemo/*.csv' INTO TABLE Hive_Test.dri
select * from Hive_Test.drivers;
```

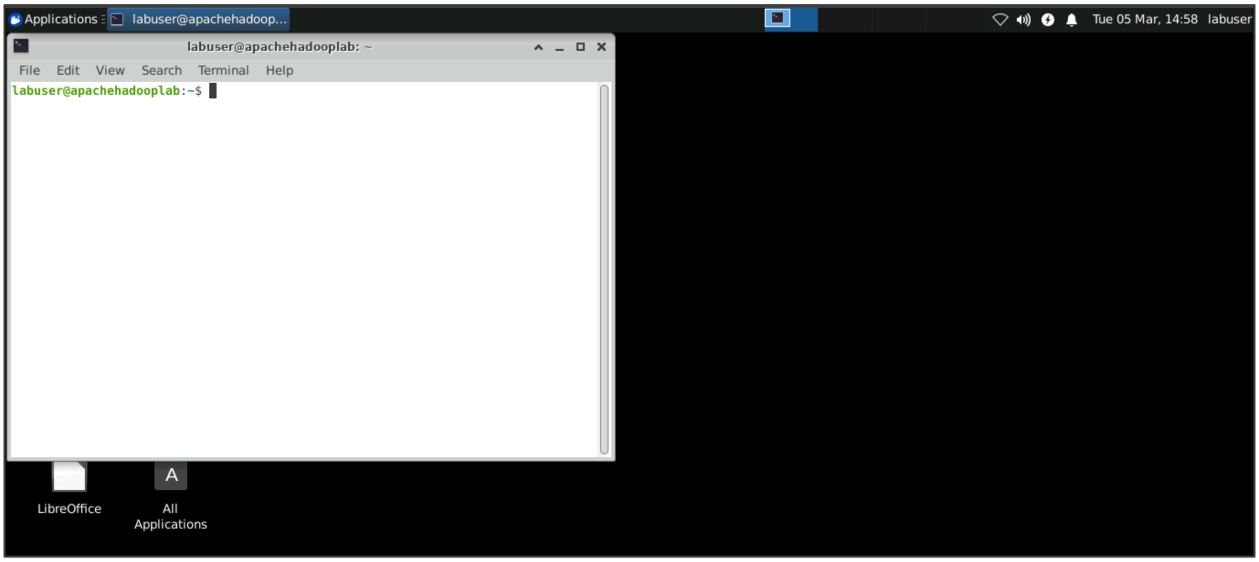
```
8b): select * from Hive_Test.drivers
INFO  : Completed executing command(queryId=hive_20220612154901_f208a3ec-94e7-4e23-8c22-
2922ae93f40b); Time taken: 0.004 seconds
INFO  : OK
```

	drivers.driverid	drivers.name	drivers.ssn	drivers.location
1	10	George Vetticaden	621011971	244-4532 Nulla Rd.
2	11	Jamie Engesser	262112338	366-4125 Ac Street
3	12	Paul Coddin	198041975	Ap #622-957 Risus. Street

Step 22: Click on **Terminal** to open the terminal window

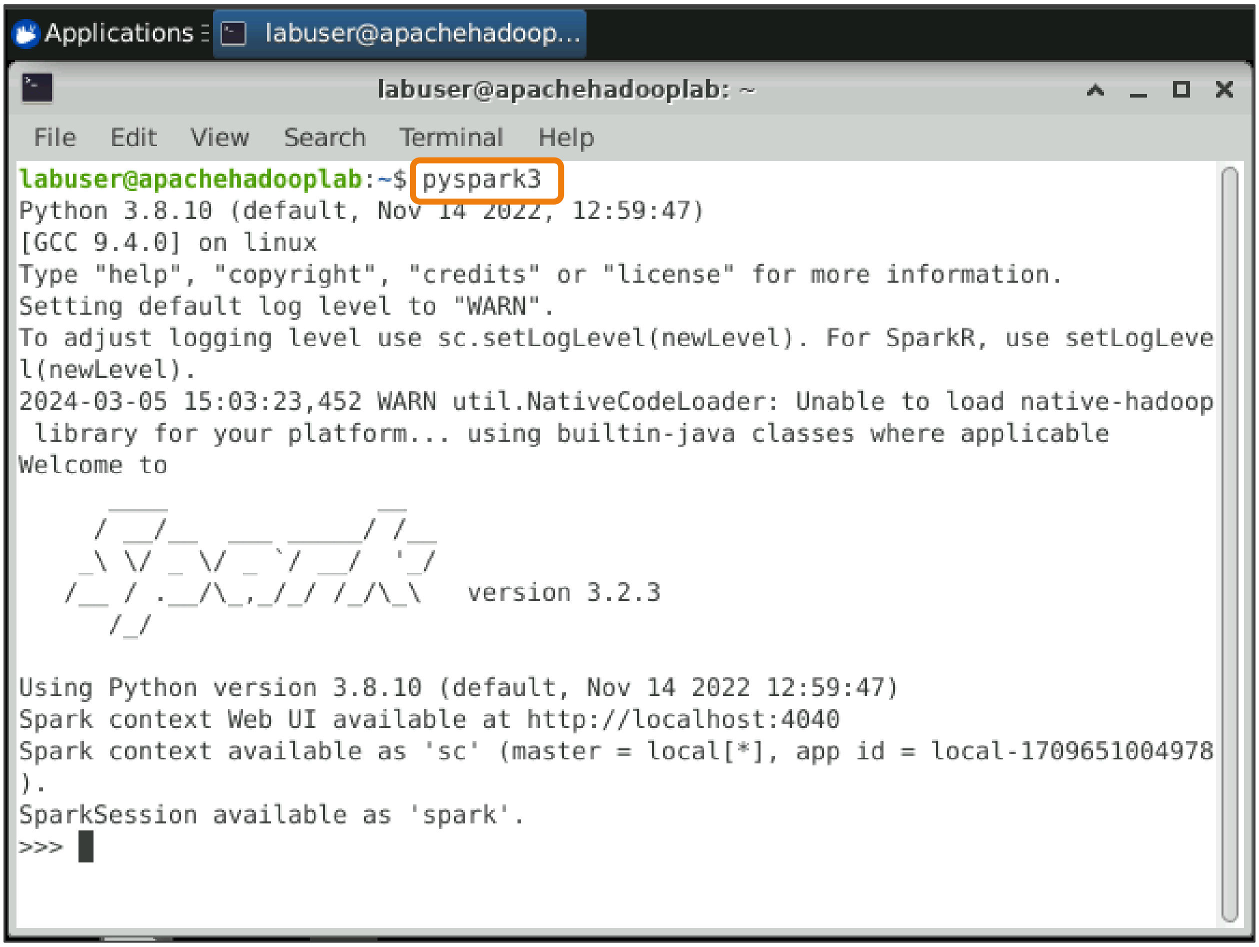


Note: You will be able to see the new terminal file as shown below:


Applications
labuser@apachehadoop...
Tue 05 Mar, 14:58 labuser
labuser@apachehadooplab: ~
File Edit View Search Terminal Help
LibreOffice
All Applications

```
labuser@apachehadooplab:~$
```

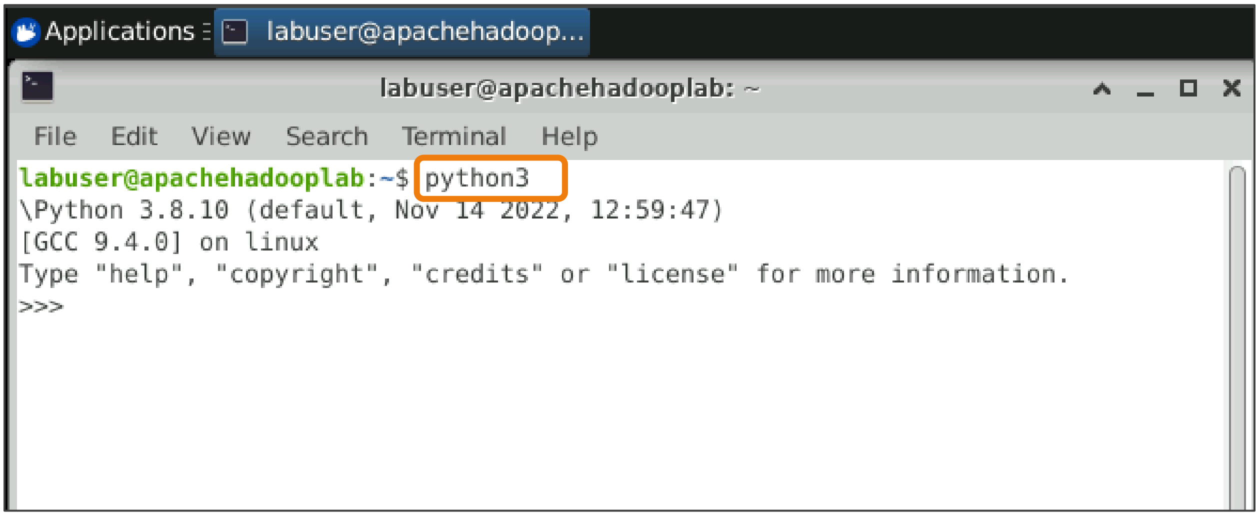
Step 23: To access the **PySpark console**, execute the following command: **pyspark3**



Applications
labuser@apachehadoop...
labuser@apachehadooplab: ~
File Edit View Search Terminal Help

```
labuser@apachehadooplab:~$ pyspark3
Python 3.8.10 (default, Nov 14 2022, 12:59:47)
[GCC 9.4.0] 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 setLogLeve
l(newLevel).
2024-03-05 15:03:23,452 WARN util.NativeCodeLoader: Unable to load native-hadoop
 library for your platform... using builtin-java classes where applicable
Welcome to
      ____              __
     / __/__  ___ _____/ /__
    _\ \/ _ \/ _ `/ __/  '_/
   /__ / .__/\_,_/_/ /_/\_\   version 3.2.3
      /_/

Using Python version 3.8.10 (default, Nov 14 2022 12:59:47)
Spark context Web UI available at http://localhost:4040
Spark context available as 'sc' (master = local[*], app id = local-1709651004978
).
SparkSession available as 'spark'.
>>> 
```

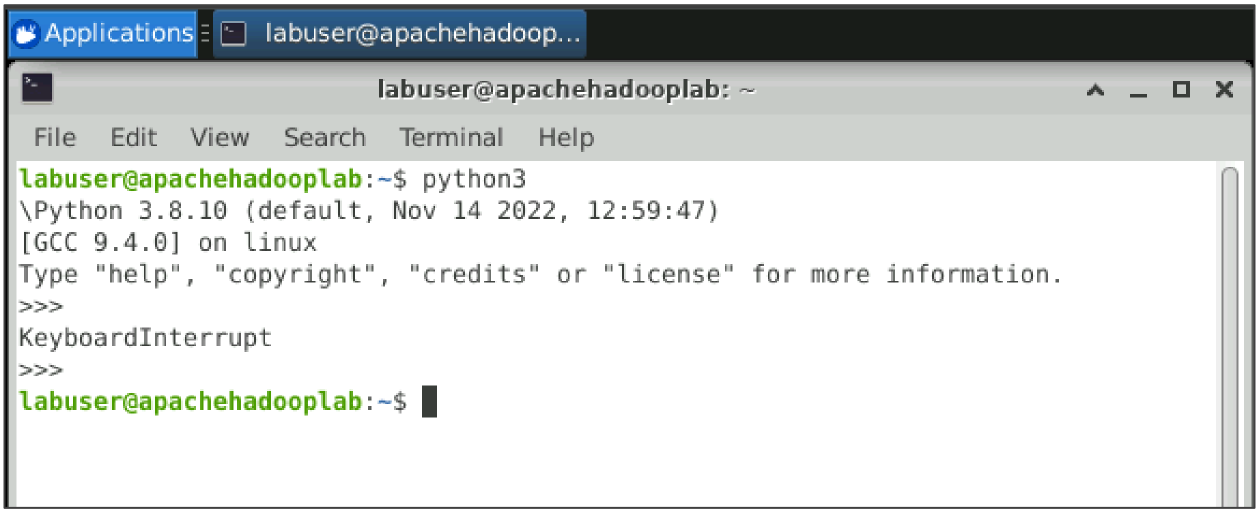
Step 24: To access the **Python shell**, execute the following command:
python3



Applications
labuser@apachehadoop...
labuser@apachehadooplab: ~
File Edit View Search Terminal Help

```
labuser@apachehadooplab:~$ python3
\Python 3.8.10 (default, Nov 14 2022, 12:59:47)
[GCC 9.4.0] on linux
Type "help", "copyright", "credits" or "license" for more information.
>>>
```

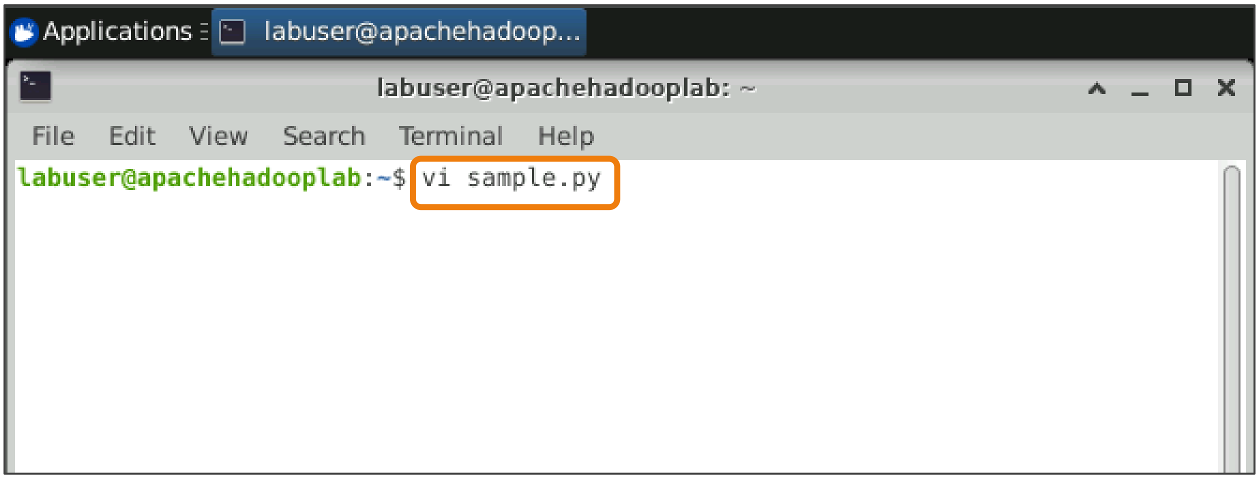
Step 25: Press **Ctrl+d** to exit the Python shell. This action will not be visibly displayed.



Applications
labuser@apachehadoop...
labuser@apachehadooplab: ~
File Edit View Search Terminal Help

```
labuser@apachehadooplab:~$ python3
\Python 3.8.10 (default, Nov 14 2022, 12:59:47)
[GCC 9.4.0] on linux
Type "help", "copyright", "credits" or "license" for more information.
>>>
KeyboardInterrupt
>>>
labuser@apachehadooplab:~$ 
```

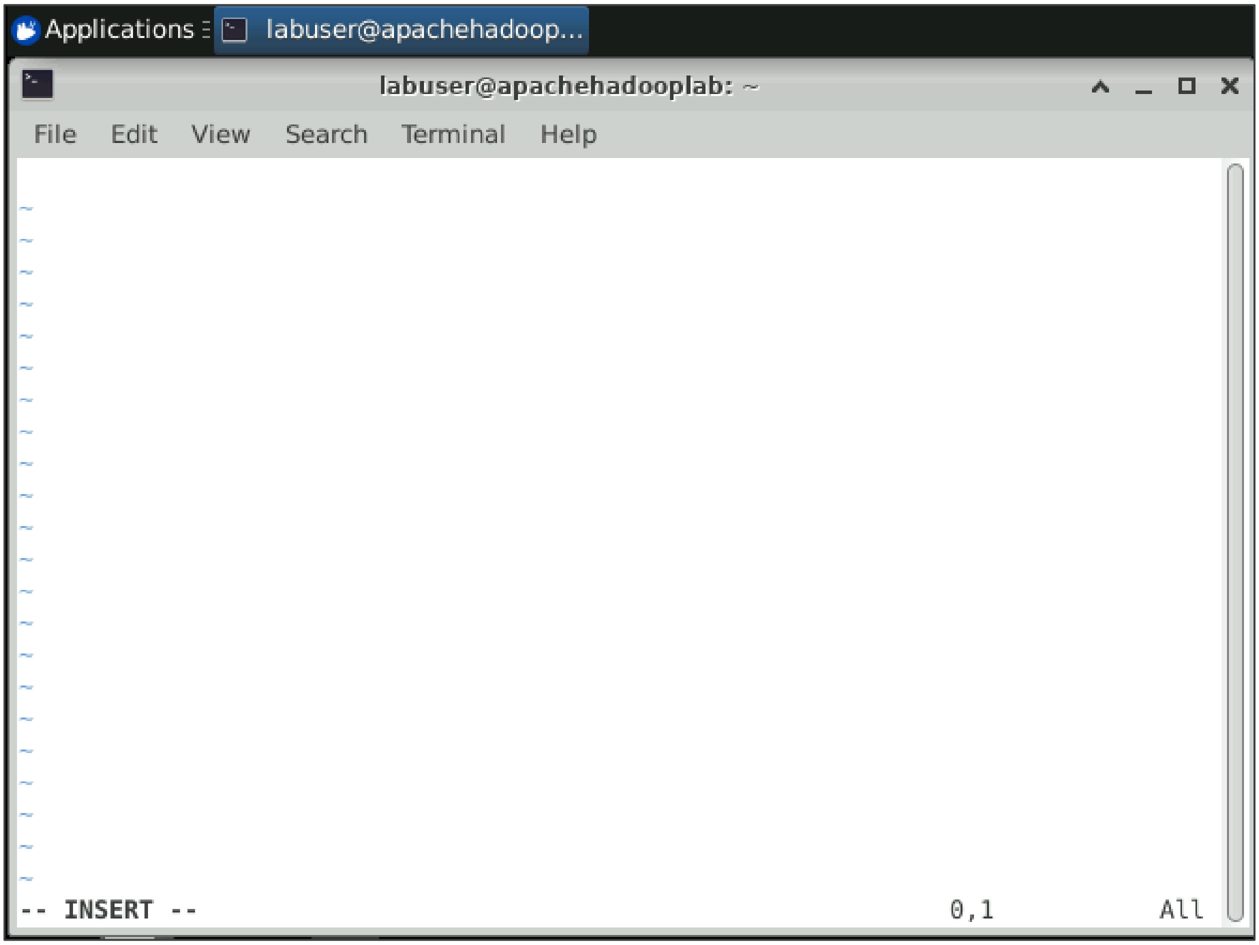
Step 26: To enter the **vi editor** and to write any Python file or txt file, use the command below:
vi sample.py
or
vi sample.txt



Applications
labuser@apachehadoop...
labuser@apachehadooplab: ~
File Edit View Search Terminal Help

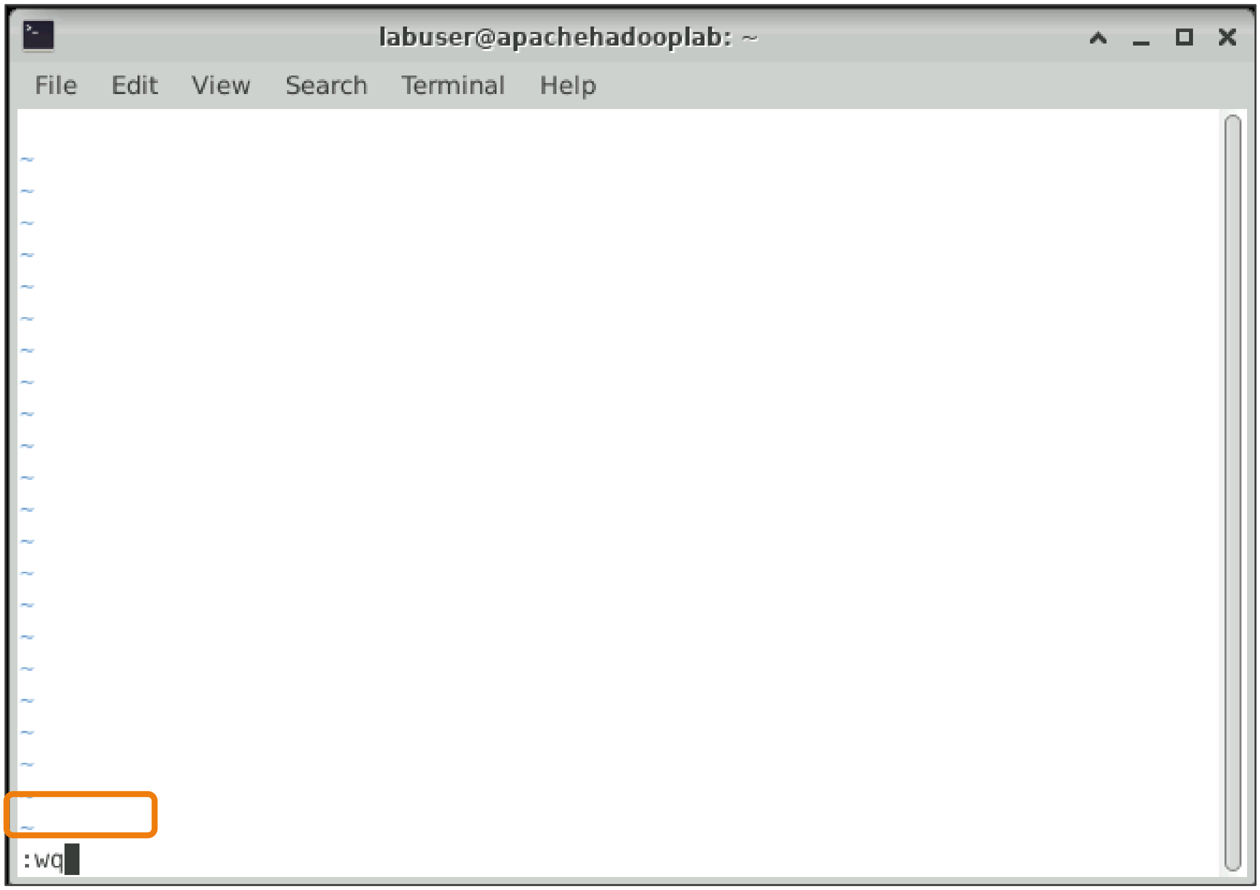
```
labuser@apachehadooplab:~$ vi sample.py
```

Step 27: Click on **i** on your keyboard to enter the **insert mode**



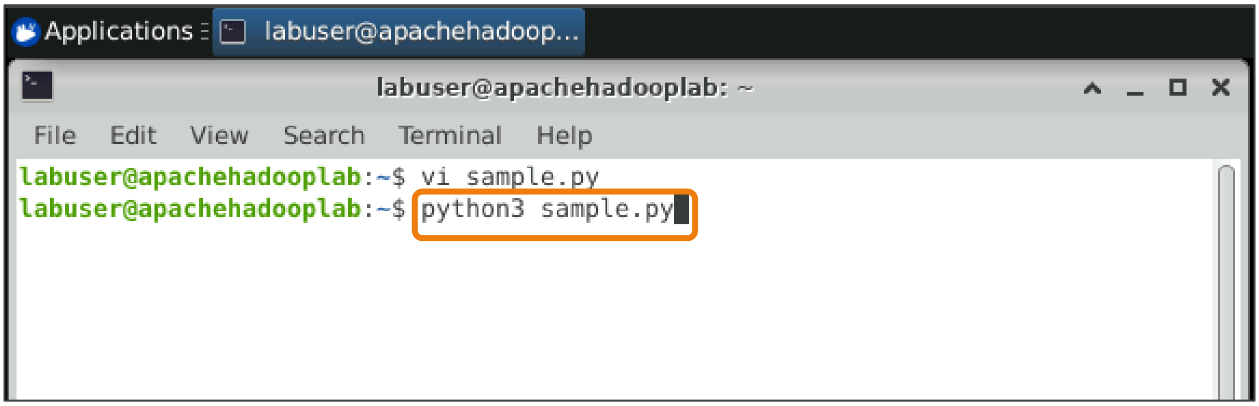
```
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
-- INSERT --                                                    0,1           All
```

Step 28: Click on the **ESC** key and then type **:wq** to save and exit


labuser@apachehadooplab: ~
File Edit View Search Terminal Help

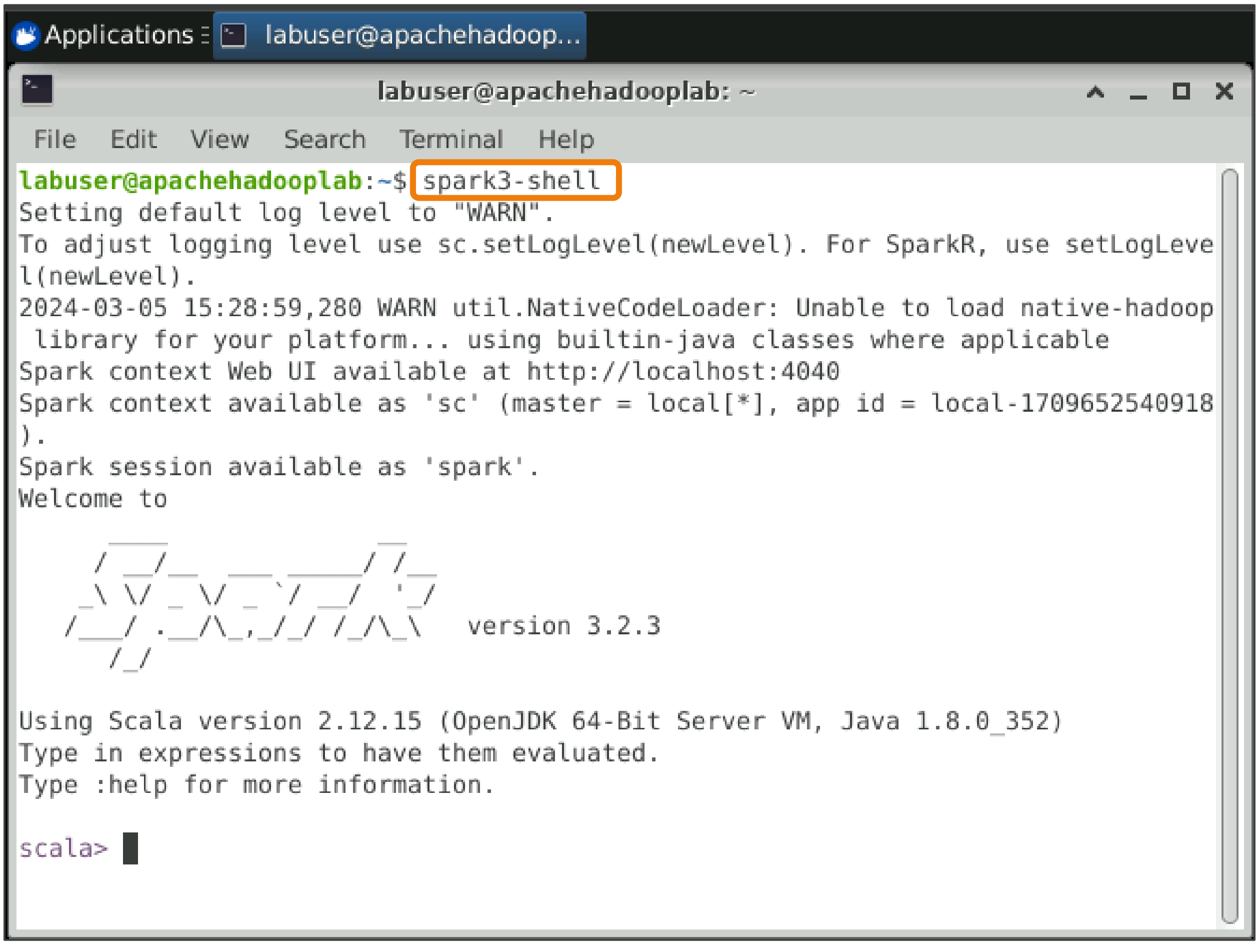
```
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
:wq
```

Step 29: To execute the **Python script**, run the command below:
python3 sample.py


Applications
labuser@apachehadoop...
labuser@apachehadooplab: ~
File Edit View Search Terminal Help

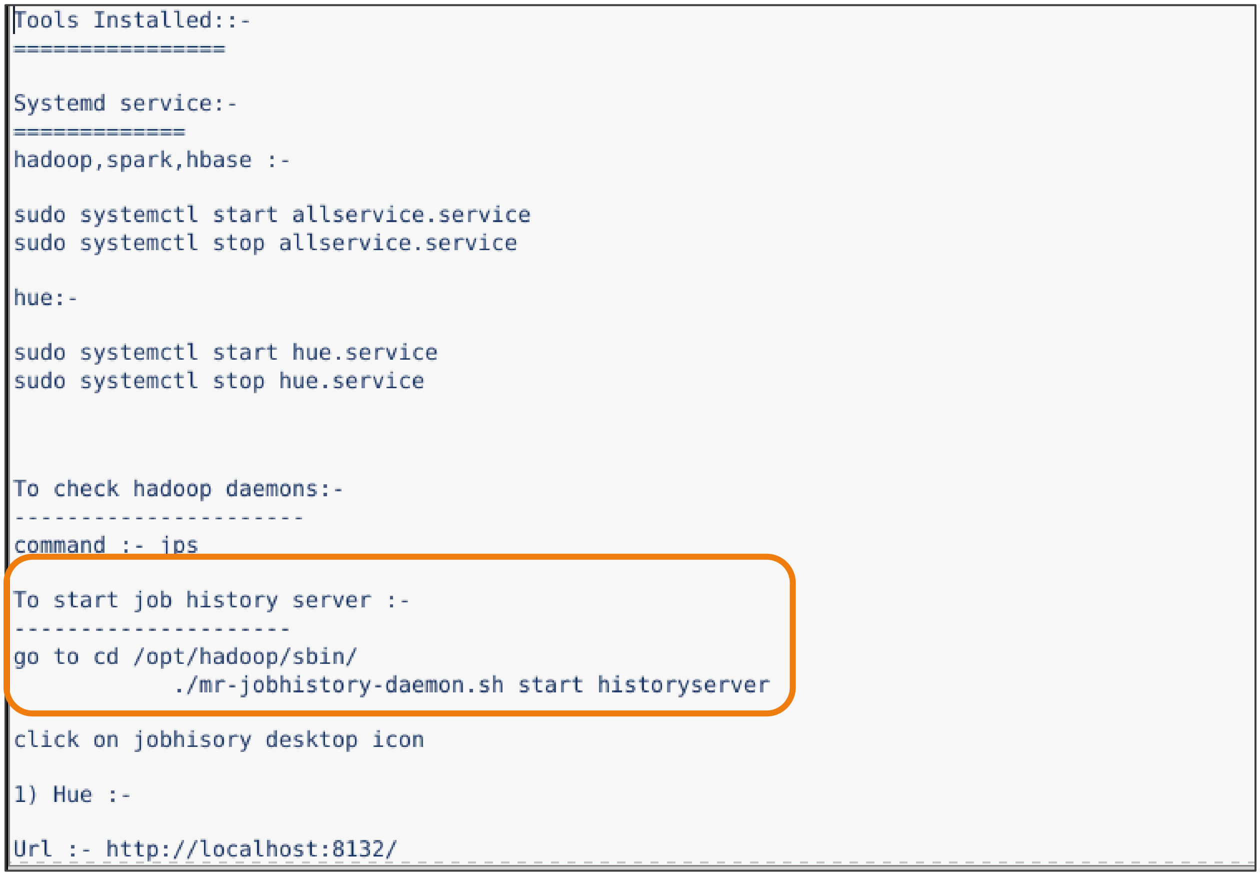
```
labuser@apachehadooplab:~$ vi sample.py
labuser@apachehadooplab:~$ python3 sample.py
```

Step 30: To enter the **Scala environment,** execute the following command: **spark3-shell**

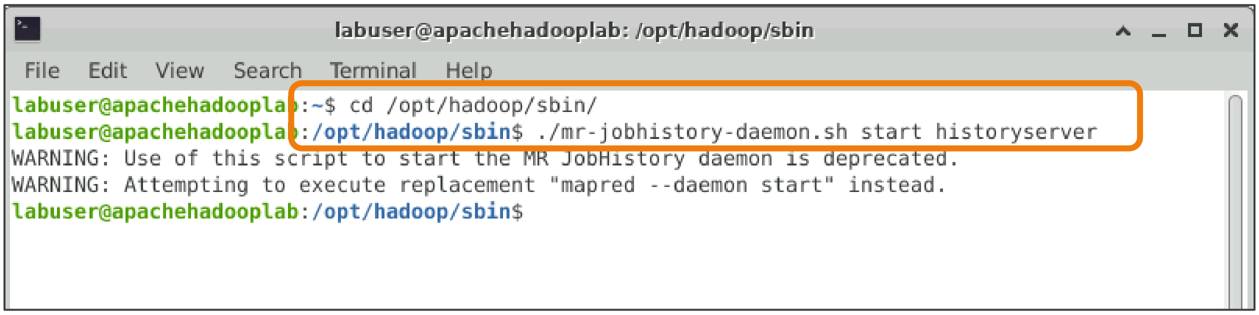


```
labuser@apachehadooplab:~$ spark3-shell
Setting default log level to "WARN".
To adjust logging level use sc.setLogLevel(newLevel). For SparkR, use setLogLeve
l(newLevel).
2024-03-05 15:28:59,280 WARN util.NativeCodeLoader: Unable to load native-hadoop
 library for your platform... using builtin-java classes where applicable
Spark context Web UI available at http://localhost:4040
Spark context available as 'sc' (master = local[*], app id = local-1709652540918
).
Spark session available as 'spark'.
Welcome to
      ____              __
     / __/__  ___ _____/ /__
    _\ \/ _ \/ _ `/ __/  '_/
   /___/ .__/\_,_/_/ /_/\_\   version 3.2.3
      /_/

Using Scala version 2.12.15 (OpenJDK 64-Bit Server VM, Java 1.8.0_352)
Type in expressions to have them evaluated.
Type :help for more information.

scala>
```

Step 31: To activate the **JobHistory server**, the initial step involves launching the **jobhistory daemon**. To achieve this, simply copy the command specified in the Readme file and execute it within your terminal.



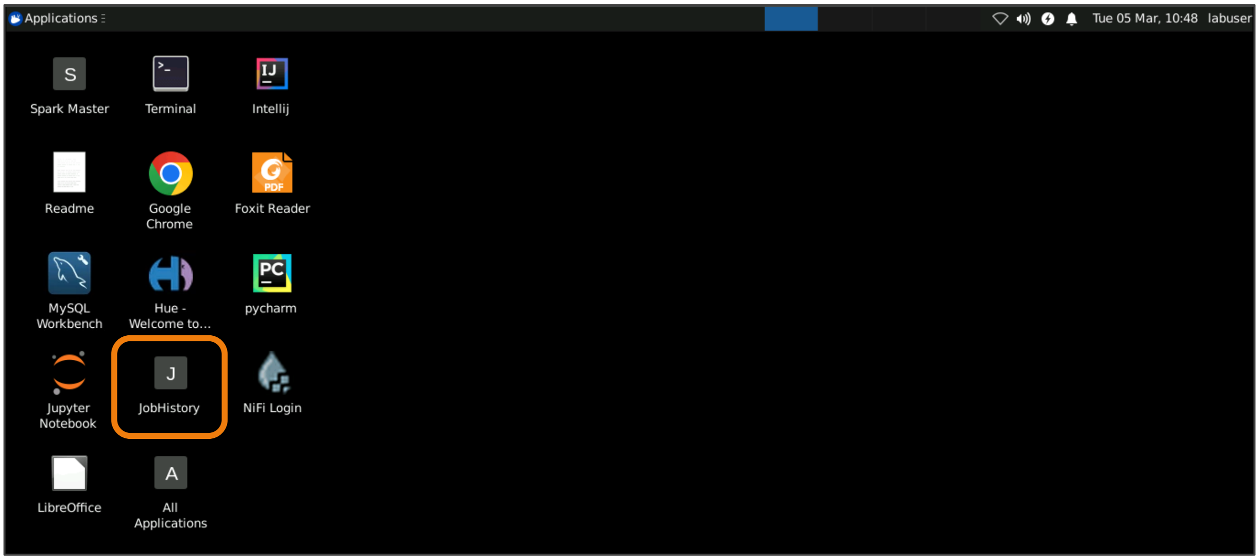
```
Tools Installed::-
================

Systemd service:-
=============
hadoop,spark,hbase :-

sudo systemctl start allservice.service
sudo systemctl stop allservice.service

hue:-

sudo systemctl start hue.service
sudo systemctl stop hue.service



To check hadoop daemons:-
----------------------
command :- jps

To start job history server :-
---------------------
go to cd /opt/hadoop/sbin/
            ./mr-jobhistory-daemon.sh start historyserver

click on jobhisory desktop icon

1) Hue :-

Url :- http://localhost:8132/
```



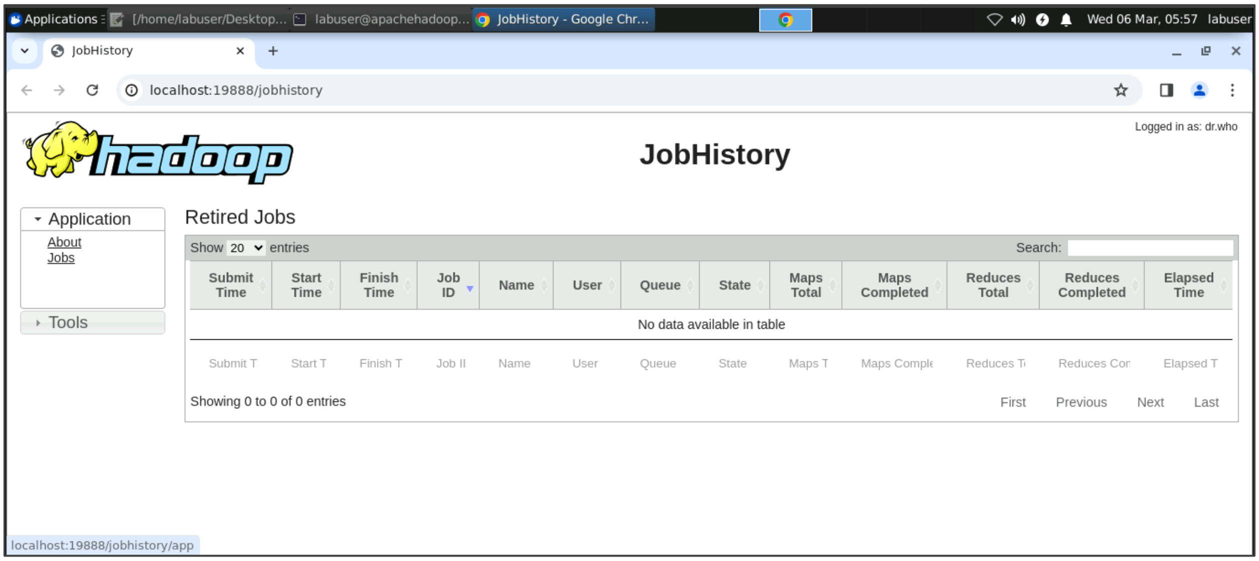
labuser@apachehadooplab: /opt/hadoop/sbin
File Edit View Search Terminal Help

```
labuser@apachehadooplab:~$ cd /opt/hadoop/sbin/
labuser@apachehadooplab:/opt/hadoop/sbin$ ./mr-jobhistory-daemon.sh start historyserver
WARNING: Use of this script to start the MR JobHistory daemon is deprecated.
WARNING: Attempting to execute replacement "mapred --daemon start" instead.
labuser@apachehadooplab:/opt/hadoop/sbin$
```

Step 32: Now, click on **JobHistory**



Note: After completing the activation of the jobhistory daemon, you can access the JobHistory dashboard to review your job history.


Applications
[/home/labuser/Desktop...
labuser@apachehadoop...
JobHistory - Google Chr...
Wed 06 Mar, 05:57
labuser
JobHistory
localhost:19888/jobhistory
hadoop
Application
About
Jobs
Tools
localhost:19888/jobhistory/app

Logged in as: dr.who

JobHistory

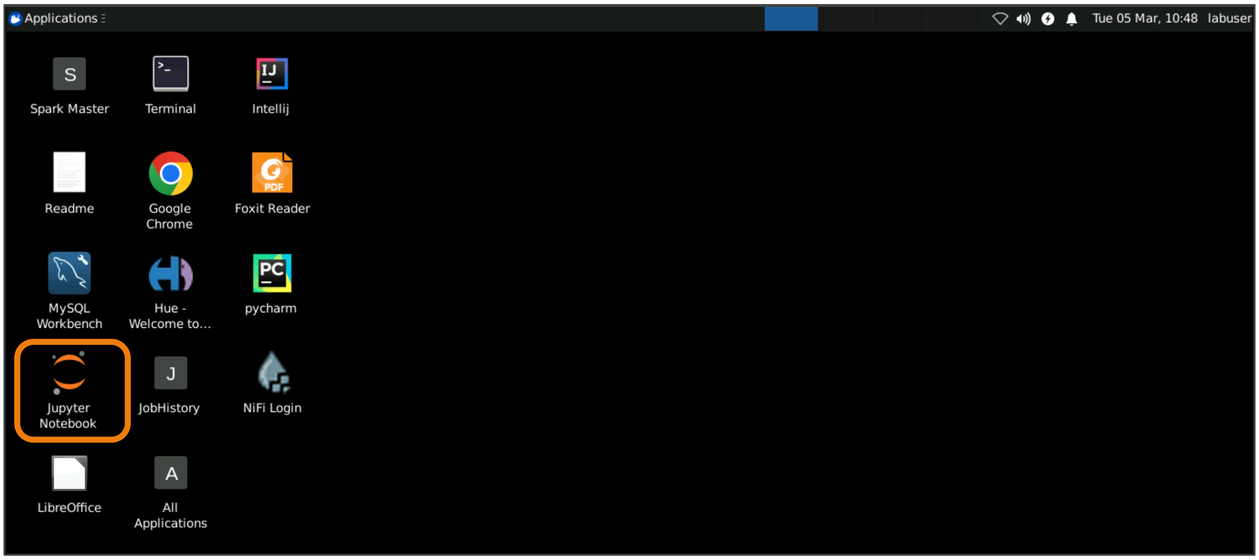
Retired Jobs

Show 20 entries   Search:

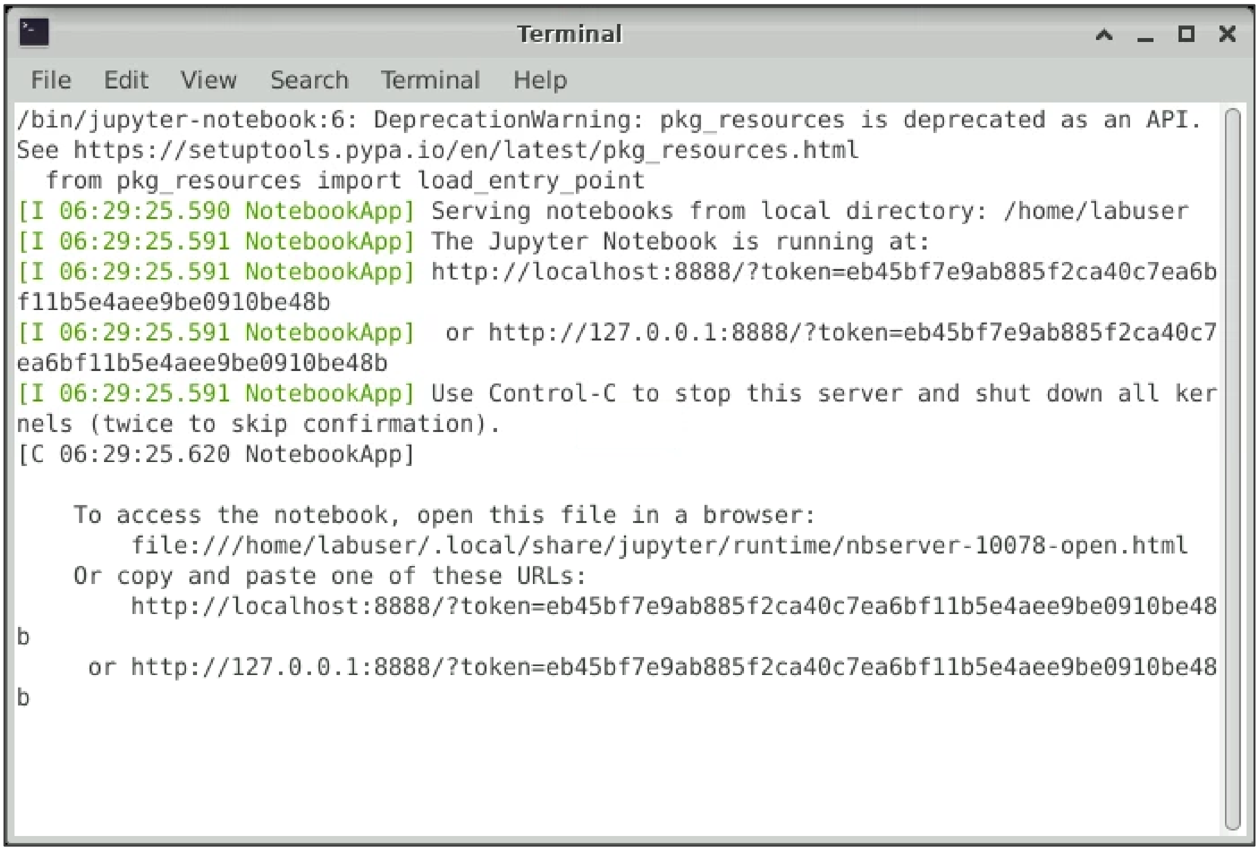
Submit Time	Start Time	Finish Time	Job ID	Name	User	Queue	State	Maps Total	Maps Completed	Reduces Total	Reduces Completed	Elapsed Time
No data available in table												

Showing 0 to 0 of 0 entries   First Previous Next Last

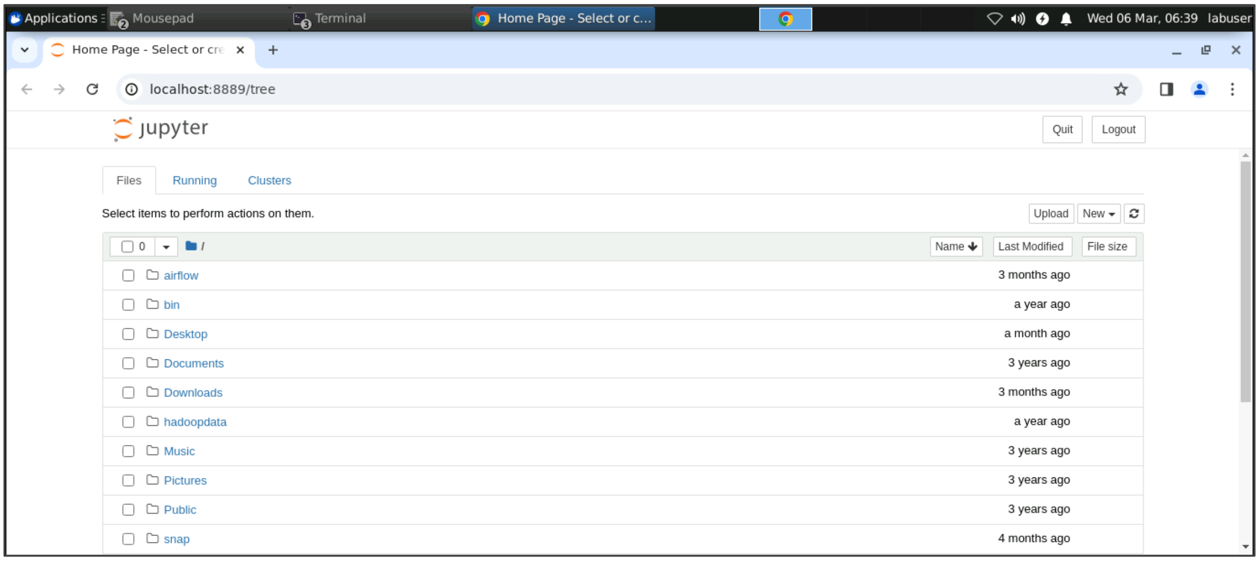
Step 33: Click on **Jupyter Notebook** to access the **Jupyter** environment



Note: You will be prompted with a terminal window, followed by the display of your Jupyter dashboard, where you can access your notebooks.



```
/bin/jupyter-notebook:6: DeprecationWarning: pkg_resources is deprecated as an API.
See https://setuptools.pypa.io/en/latest/pkg_resources.html
  from pkg_resources import load_entry_point
[I 06:29:25.590 NotebookApp] Serving notebooks from local directory: /home/labuser
[I 06:29:25.591 NotebookApp] The Jupyter Notebook is running at:
[I 06:29:25.591 NotebookApp] http://localhost:8888/?token=eb45bf7e9ab885f2ca40c7ea6b
f11b5e4aee9be0910be48b
[I 06:29:25.591 NotebookApp]  or http://127.0.0.1:8888/?token=eb45bf7e9ab885f2ca40c7
ea6bf11b5e4aee9be0910be48b
[I 06:29:25.591 NotebookApp] Use Control-C to stop this server and shut down all ker
nels (twice to skip confirmation).
[C 06:29:25.620 NotebookApp]

    To access the notebook, open this file in a browser:
        file:///home/labuser/.local/share/jupyter/runtime/nbserver-10078-open.html
    Or copy and paste one of these URLs:
        http://localhost:8888/?token=eb45bf7e9ab885f2ca40c7ea6bf11b5e4aee9be0910be48
b
     or http://127.0.0.1:8888/?token=eb45bf7e9ab885f2ca40c7ea6bf11b5e4aee9be0910be48
b
```



Select items to perform actions on them.

Name	Last Modified	File size
airflow	3 months ago	
bin	a year ago	
Desktop	a month ago	
Documents	3 years ago	
Downloads	3 months ago	
hadoopdata	a year ago	
Music	3 years ago	
Pictures	3 years ago	
Public	3 years ago	
snap	4 months ago	