

---

# Big Data Hadoop and Spark Developer

Lesson-End Project Solution



Get Certified. Get Ahead.

# Retail Business Analytics

## Steps to Perform:

**Step 1:** Log in to your LMS account

**Step 2:** Open the course “**Big Data Hadoop and Spark Developer**”

**Step 3:** Download the datasets from the “**Course Resources**” section

**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

COMMUNITY HELP NOTES

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:

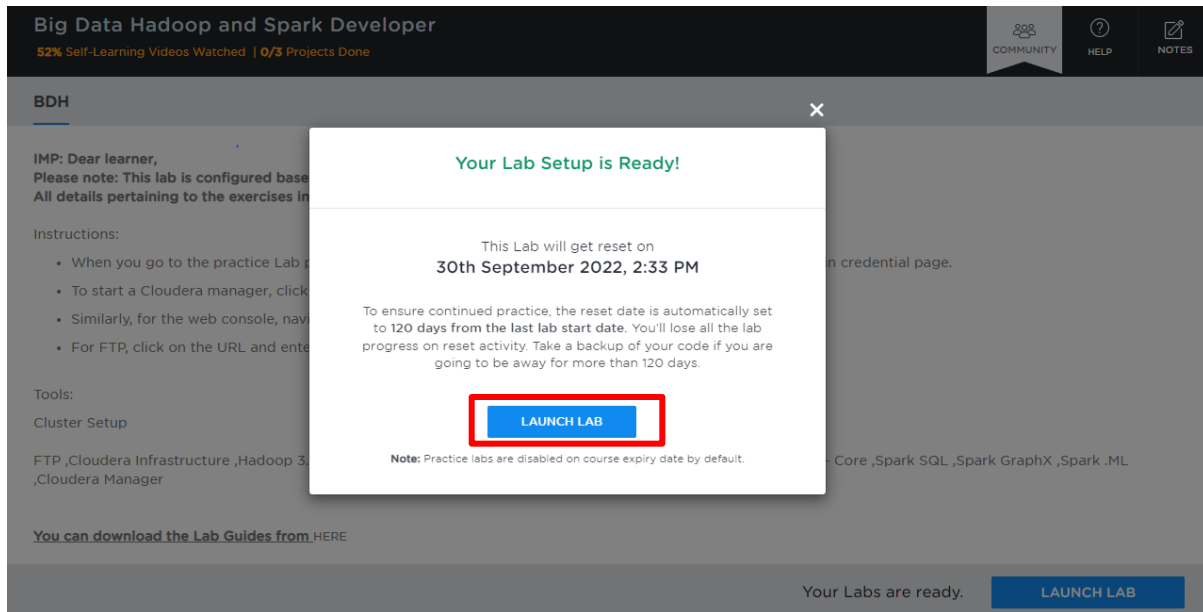
Cluster Setup

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

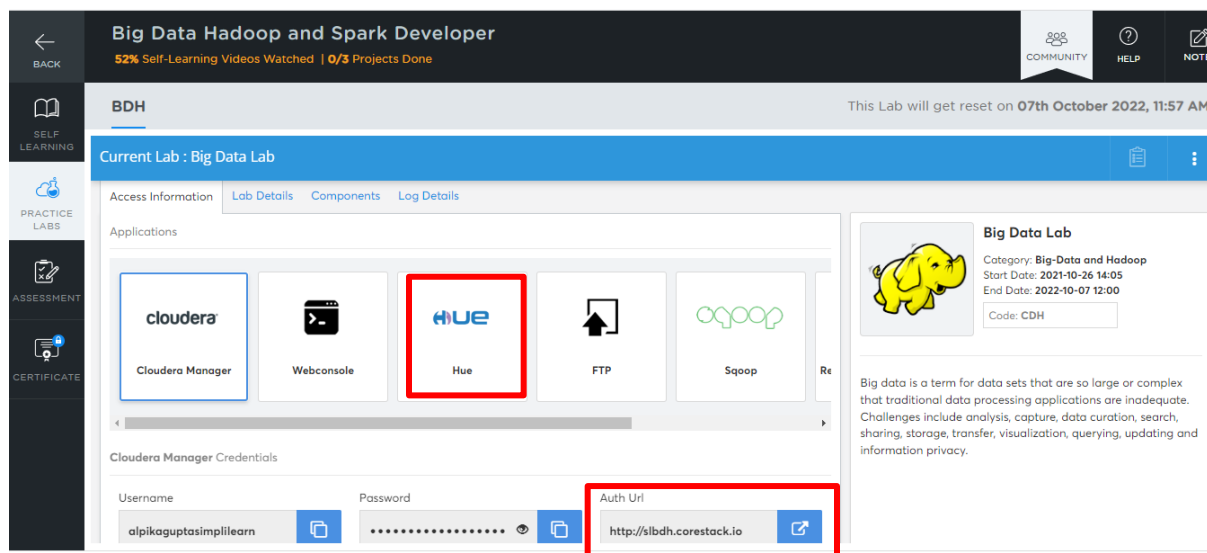
[You can download the Lab Guides from HERE](#)

Your Labs are ready. **LAUNCH LAB**

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



**Step 6:** Click on “HUE” to upload the datasets



**Step 7:** Log in to the “HUE” and click on create a directory named “data-files” and upload the “orders” dataset into it

Home

/user/testdemomay1301mailinator/data-files

<input type="checkbox"/>	Name	Size	User	Group	Permissions	Date
<input type="checkbox"/>	<a href="#">↑</a>		testdemomay1301maili...	hadoop	drwxrwx---	June 09
<input type="checkbox"/>	<a href="#">.</a>		testdemomay1301maili...	hadoop	drwxr-xr-x	June 03
<input type="checkbox"/>	<a href="#">Graphx</a>		testdemomay1301maili...	hadoop	drwxr-xr-x	May 24
<input type="checkbox"/>	<a href="#">apache</a>		testdemomay1301maili...	hadoop	drwxr-xr-x	May 26
<input type="checkbox"/>	<a href="#">data</a>		testdemomay1301maili...	hadoop	drwxr-xr-x	May 26
<input type="checkbox"/>	<a href="#">drivers.csv</a>	2.0 KB	testdemomay1301maili...	hadoop	-rw-r--r--	May 24
<input type="checkbox"/>	<a href="#">mobile-input-data.csv</a>	818.0 KB	testdemomay1301maili...	hadoop	-rw-r--r--	May 26
<input type="checkbox"/>	<a href="#">order_parquet.parquet</a>	476.8 KB	testdemomay1301maili...	hadoop	-rw-r--r--	May 23
<input type="checkbox"/>	<a href="#">orders</a>		testdemomay1301maili...	hadoop	drwxr-xr-x	May 26

**Step 8:** Click on the “Webconsole” and click on the “Auth Url”

←

BACK

Big Data Hadoop and Spark Developer

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

COMMUNITY

HELP

NOTES

BDH

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

Current Lab : Big Data Lab

Access Information Lab Details Components Log Details

Applications

cloudera

Cloudera Manager

Webconsole

hue

Hue

FTP

oozie

Sqoop

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 9:** Copy the “Username” and the “Password” provided to log in to the Web console

**Step 10:** 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.



```
.appName("Scenario1") \
.getOrCreate()
```

```
orders = spark.read.option("inferSchema", True) \
.csv("/user/testdemomay1301mailinator/data-files/orders") \
.toDF("order_id", "order_date", "order_customer_id", "order_status")
```

```
orders.createOrReplaceTempView("orders")
```

```
result = spark.sql("""
SELECT Substring(order_date, 1, 7) order_date,
order_status, Count(1) cnt
FROM orders
WHERE order_status = 'SUSPECTED_FRAUD'
GROUP BY Substring(order_date, 1, 7), order_status
ORDER BY order_date desc""").select("order_date", "cnt")
```

```
result.show(10)
```

**Step 13:** Now, you will be able to find the order\_date with a number of counts

```
>>> orders = spark.read.option("inferSchema", True) \
...     .csv("/user/testdemomay1301mailinator/data-files/orders") \
...     .toDF("order_id", "order_date", "order_customer_id", "order_status")
>>>
>>> orders.createOrReplaceTempView("orders")
>>>
>>> result = spark.sql("""
... SELECT Substring(order_date, 1, 7) order_date,
... order_status, Count(1) cnt
... FROM orders
... WHERE order_status = 'SUSPECTED_FRAUD'
... GROUP BY Substring(order_date, 1, 7), order_status
... ORDER BY order_date desc""").select("order_date", "cnt")
>>>
>>> result.show(10)
+-----+----+
|order_date|cnt|
+-----+----+
| 2014-07|101|
| 2014-06|131|
| 2014-05|130|
| 2014-04|112|
| 2014-03|138|
| 2014-02|119|
| 2014-01|131|
| 2013-12|126|
| 2013-11|150|
| 2013-10|108|
+-----+----+
only showing top 10 rows
```

