

Big Data Hadoop and Spark Developer

Lesson-End Project Solution



Get Certified. Get Ahead.

Telecom Log Parsing

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: Click on the “**PRACTICE LABS**” tab on the left side and select “**LAUNCH LAB**”

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:

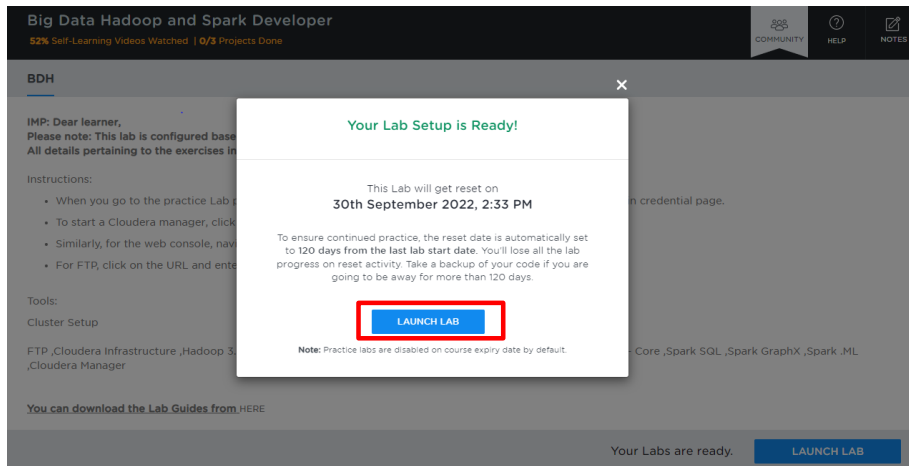
Cluster Setup

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

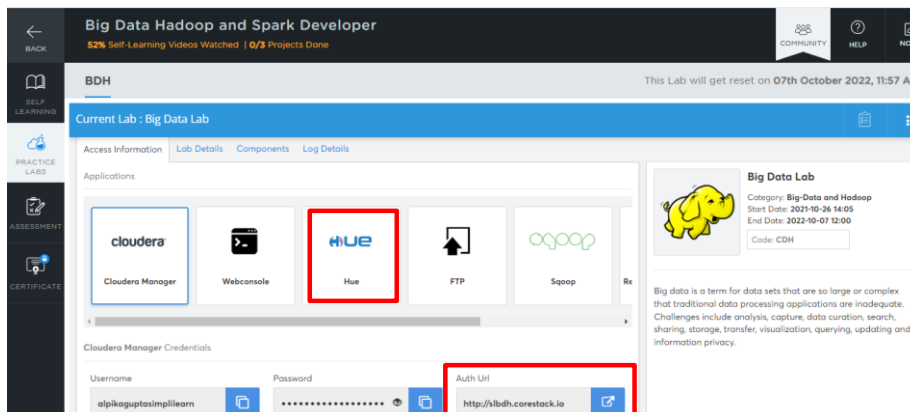
You can download the Lab Guides from [HERE](#)

Your Labs are ready. **LAUNCH LAB**

Step 5: Click on the “**LAUNCH LAB**” button



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



Step 7: Log in to HUE and create a directory named “data-files” and upload the CSV file into it

Home /user/testdemomay1301mailinator/data-files

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

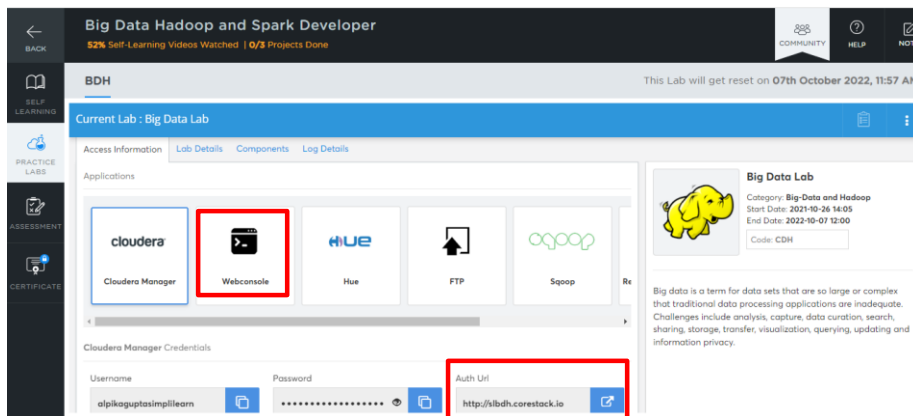
Step 8: Create a new directory in data-files directory named “**apache**” and upload the “**access.log**” file into the directory

Home /user/testdemomay1301mailinator/data-files/apache

<input type="checkbox"/>	Name	Size	User	Group	Permissions	Date
<input type="checkbox"/>	↑		testdemomay1301maili...	hadoop	drwxr-xr-x	June 03, 2022 04:2
<input type="checkbox"/>	.		testdemomay1301maili...	hadoop	drwxr-xr-x	May 26, 2022 05:3
<input type="checkbox"/>	access.log	471.1 MB	testdemomay1301maili...	hadoop	-rw-r--	May 26, 2022 05:3

Show 45 of 1 items Page 1 of 1

Step 9: Click on “**Webconsole**” and then on “**Auth Url**”



Step 10: Copy the "Username" and "Password" provided to log in to the Webconsole

Step 11: Paste the "Username" and "Password" on the console and click on Enter

Note: The password will not be visible when pasted on the console.

```
bdh-cluster2-edgenode10 login: testdemomay1301mailinator
Password:
Last login: Tue May 31 07:33:55 on pts/28

=====
*                               :

Password for testdemomay1301mailinator@BDH-ENV.GNE4-RUTX.CLOUDERA.SITE:
[testdemomay1301mailinator@bdh-cluster2-edgenode10 ~]$
```

Step 12: Log in to the PySpark shell

Command:

pyspark3


```
>>> server_log_rdd = spark.sparkContext.textFile("/user/testdemomay1301mailinator/data-files/apache/access.log")
>>> server_log_rdd.take(2)
['109.169.248.24/ - - [12/Dec/2015:18:25:11 +0100] GET /administrator/ HTTP/1.1 200 4263 - Mozilla/5.0 (Windows NT 6
]
>>> data = spark.sparkContext.textFile("/user/testdemomay1301mailinator/data-files/mobile-input-data.csv")
>>> data.take(5)
[('10', 'track_name', 'size_bytes', 'currency', 'price', 'rating_count_tot', 'rating_count_ver', 'user_rating', 'user_rating_
devices.num', 'ipad5c_urls.num', 'lang.num', 'vpp_license', '1', '281656475', 'PAC-MAN Premium', '100788224', 'USD', '3.99', '21292', '26',
, '281796108', 'Evernote - stay organized', '158578688', 'USD', '0.161065', '26', '4', '3.5', '8.2.2', '4+', 'Productivity', '37', '5', '23', '1', '3',
dar, Maps, Alerts', '100524032', 'USD', '0.188583', '2822', '3', '5', '4', '5', '5.0.0', '4+', 'Weather', '37', '5', '3', '1', '4', '282614216', 'eBay: Best
8512000', 'USD', '0.262241', '649', '4', '4', '5', '5.10.0', '12+', 'Shopping', '37', '5', '9', '1']
>>> server_log_rdd = server_log_rdd \
...     .filter(lambda line: line != '') \
...     .map(lambda line: line.split(" ")[8]) \
...     .filter(lambda ele: ele == "404") \
...     .count()
>>> server_log_rdd
227089
```