
Big Data Hadoop and Spark Developer

Lesson-End Project Solution



Get Certified. Get Ahead.

Retail Business Analysis Using Spark Streaming

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 “**module.py**” script from the “**Course Resources**” section

Step 4: On the left side, click on the “**PRACTICE LABS**” tab and click on the “**LAUNCH LAB**” button

The screenshot displays the interface of a Learning Management System (LMS) for the course "Big Data Hadoop and Spark Developer". The top header shows the course title and progress: "52% Self-Learning Videos Watched | 0/3 Projects Done". The left sidebar contains navigation options: "SELF LEARNING", "PRACTICE LABS" (highlighted with a red box), "ASSESSMENT", and "CERTIFICATE". The main content area is titled "BDH" and contains the following text:

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:

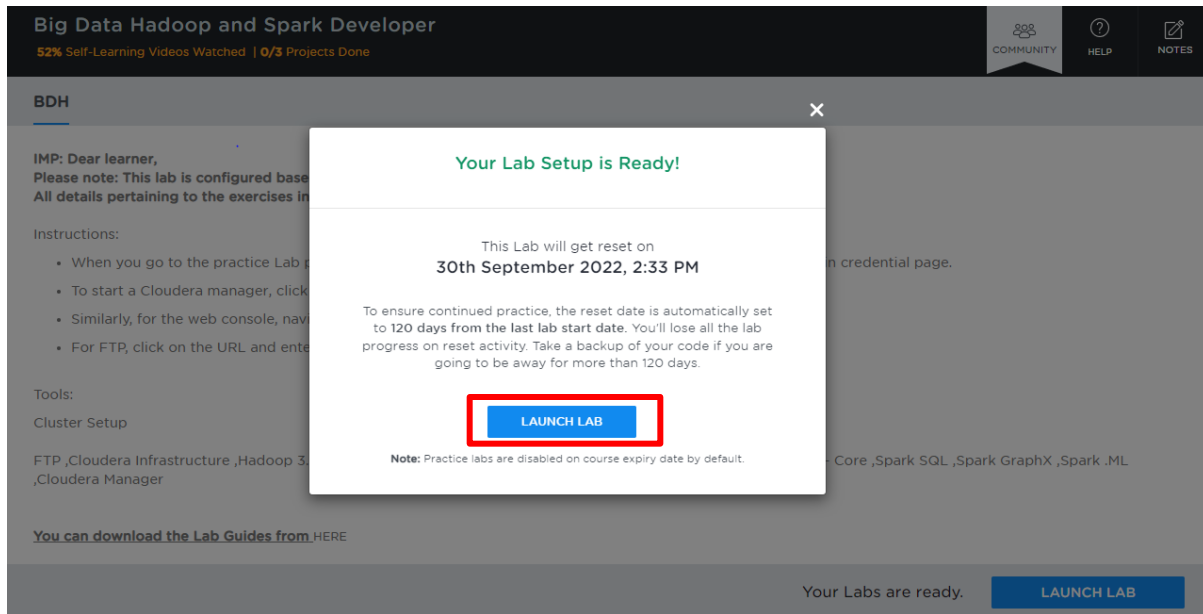
Cluster Setup

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

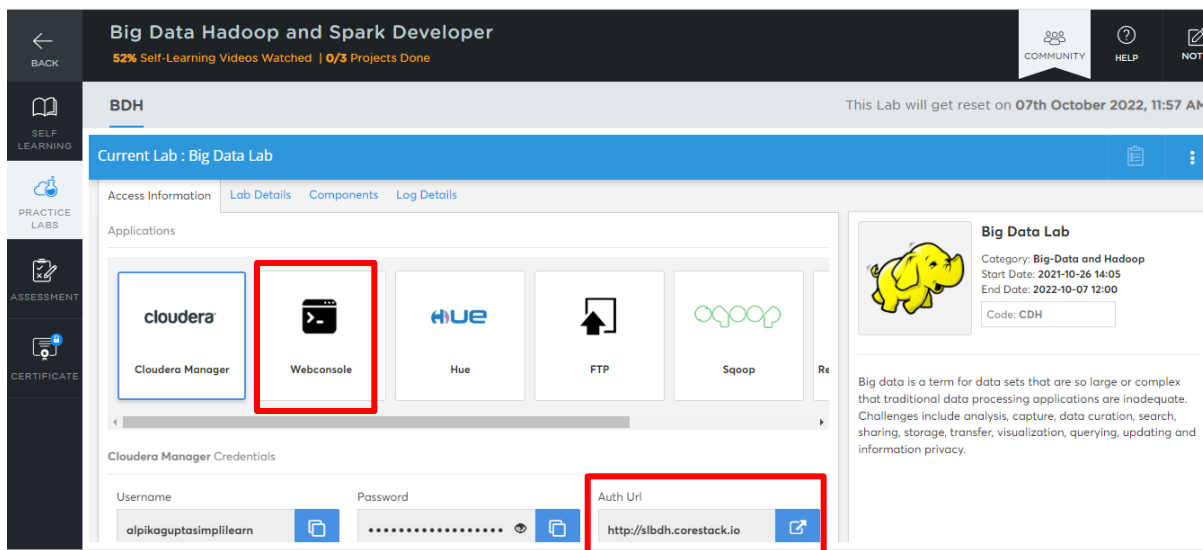
[You can download the Lab Guides from HERE](#)

At the bottom right, the text "Your Labs are ready." is followed by a blue "LAUNCH LAB" button, which is also highlighted with a red box.

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



Step 6: Click on the Webconsole and click on the “Auth Url”



Step 7: Copy the “Username” and the “Password” provided to log in to the Web console

Step 8: 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 10: Write the script into the “**module.py**” file which is available in the Course Resources section

```
import random
import time
from socket import *
from threading import Thread

import pyspark
from pyspark.streaming import StreamingContext

# Create a SparkContext with 2 threads in local mode
sc = pyspark.SparkContext("local[2]")

# Create a thread that reads streaming data from the socket and
# performs a wordcount on the data that arrived in the last 1 second
class Streamer(Thread):
    def __init__(self, sc):
        Thread.__init__(self)
        self.sc = sc

    def run(self):
        print("starting Streaming thread...")
        batchInterval = 1

        # Using the spark context, create a streaming context with a batch interval of 1 second
        ssc = StreamingContext(self.sc, batchInterval)
        # Create a socket DStream reading from localhost at port 4444
        socketDstream = ssc.socketTextStream("localhost", 9999)

        # WordCount
        wordcounts = socketDstream.flatMap(lambda line: line.split(" ")) \
            .map(lambda word: (word, 1)) \
            .reduceByKey(lambda a, b: a + b)

        # Print first 50 words counted in the last one second
        wordcounts.pprint(50)
"module.py" 79L, 3454C
```

Step 11: To come out from the script type `esc:wq`

Step 12: Run the below command to execute the “**module.py**” script

```
>>>
[testdemomay1301mailinator@bdh-cluster2-edgenode10 ~]$ vi module.py
[testdemomay1301mailinator@bdh-cluster2-edgenode10 ~]$ spark-submit module.py
```

Step 13: Next, you will see the streaming output

Time: 2022-05-26 12:49:02

(u'', 2)
(u'http://almhuetten-raith.at/administrator/', 1)
(u'- ', 5)
(u'rv:34.0)', 2)
(u'/administrator/', 1)
(u'POST', 1)
(u'NT', 2)
(u'200', 2)
(u'6.0;', 2)
(u'GET', 1)
(u'Gecko/20100101', 2)
(u'- . ', 2)
(u'+0100]', 2)
(u'83.167.113.100', 2)
(u'[12/Dec/2015:18:31:25', 2)
(u'/administrator/index.php', 1)