Big Data Hadoop and Spark Developer

Lesson-End Project Solution

simplilearn

Get Certified. Get Ahead.

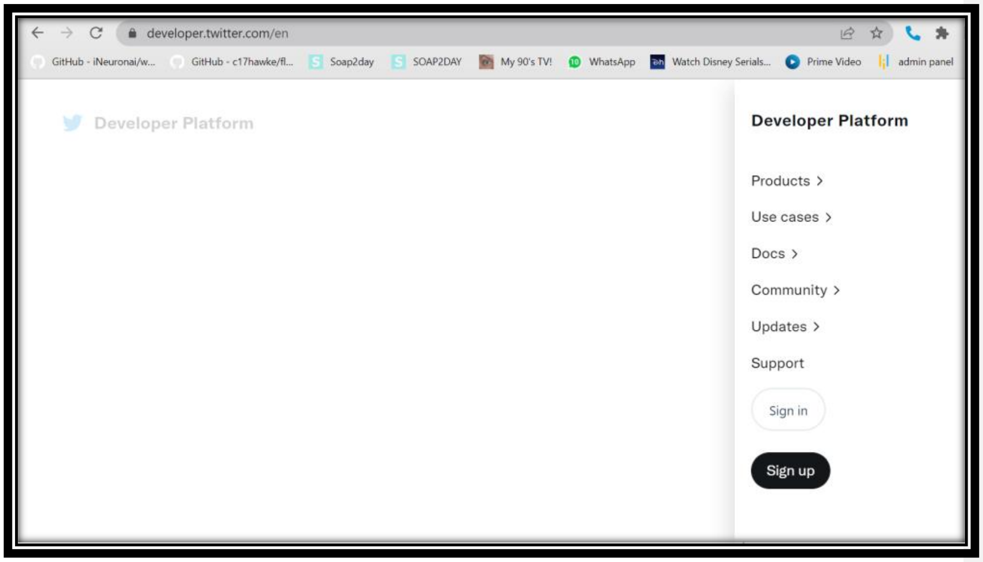
Twitter Data Ingestion with Flume

Steps to Perform:

Step 1: Set up the Twitter account

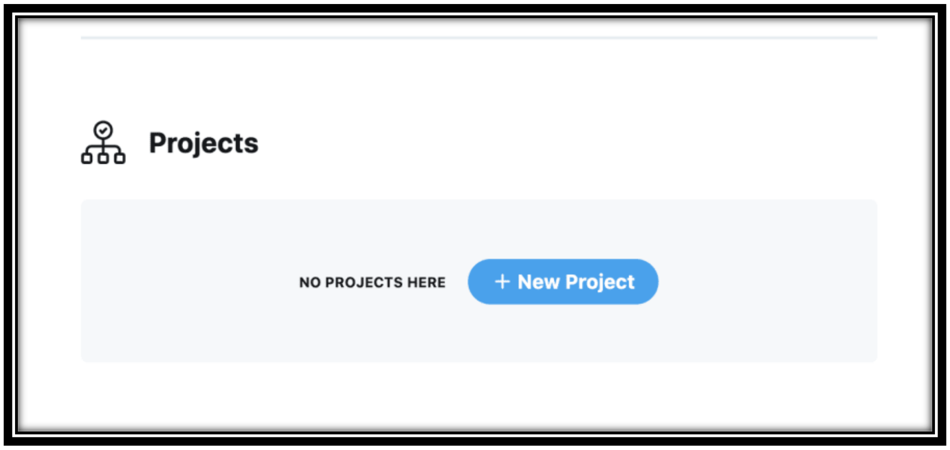
1.2.1 Sign up for a developer account from the link mentioned:

https://developer.twitter.com/en



1.2 Create a project and connect an App

1.2.1 In the developer portal, click on **Create a "New Project"**



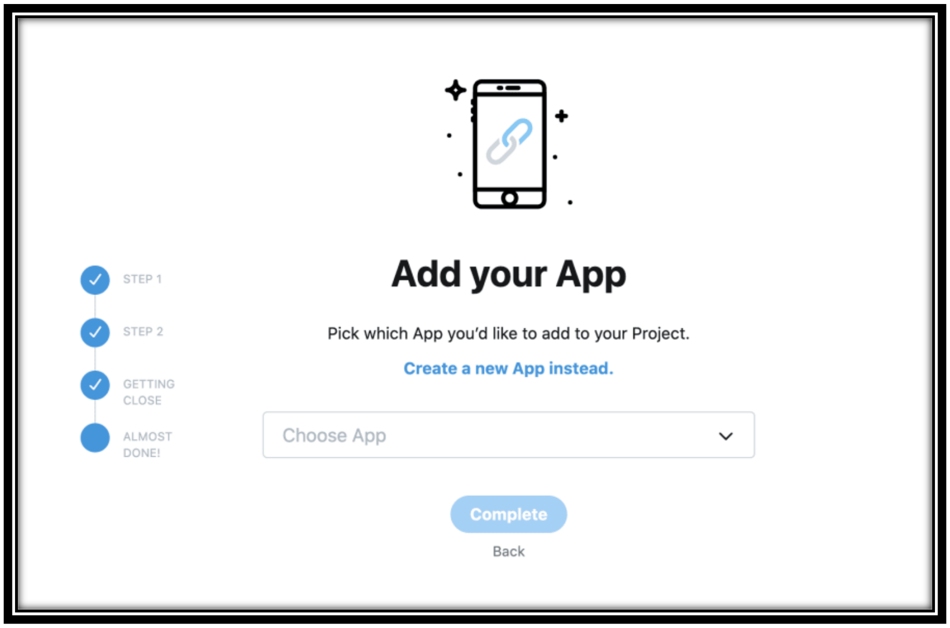
1.2.2 Provide a suitable name for the project

1.2.3 Select an appropriate use-case

1.2.4 Provide a project description

1.2.5 Create a new App or connect to an existing App

Note: An App is a container for API Keys that needs to make an HTTP request to the Twitter API.

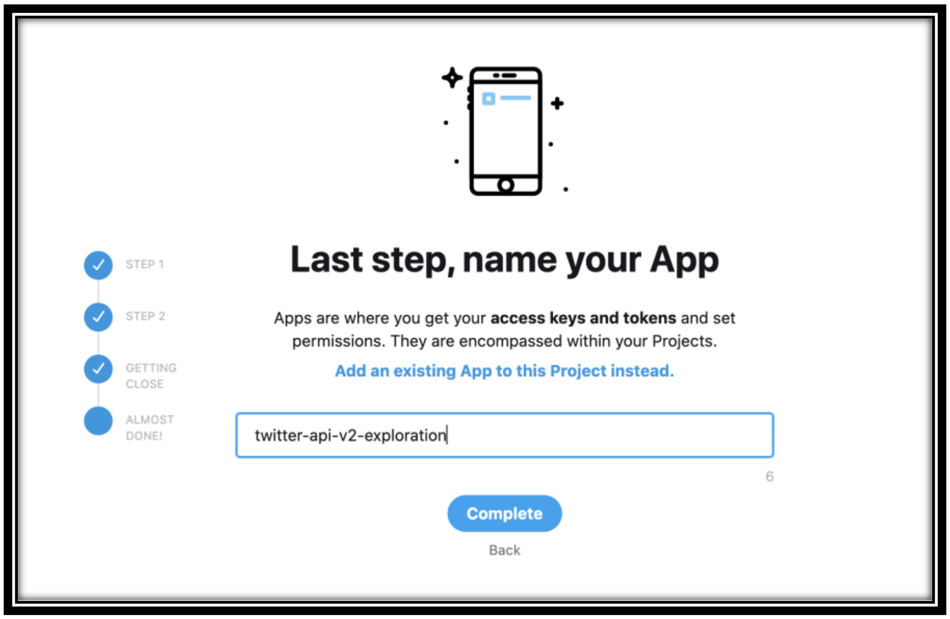


Add your App

Pick which App you'd like to add to your Project.

Create a new App instead.

1.2.6 Click on '**Create a new App instead'** and give your App a unique name



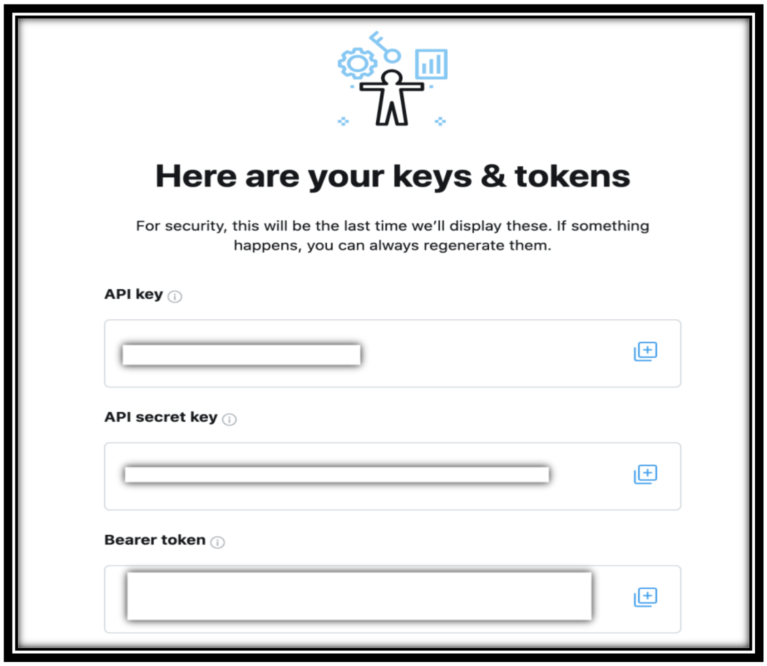
Last step, name your App

Apps are where you get your **access keys and tokens** and set permissions. They are encompassed within your Projects.

Add an existing App to this Project instead.

twitter-api-v2-exploration

Note: Once you click complete, you will get your API Keys and the Bearer Token that can be used to connect to the new endpoints in the Twitter API v2.



Here are your keys & tokens

For security, this will be the last time we'll display these. If something happens, you can always regenerate them.

API key

API secret key

Bearer token

1.2.7 Click the **(+)** next to API Key, API Secret Key, and Bearer Token and copy it in a safe place on your local machine

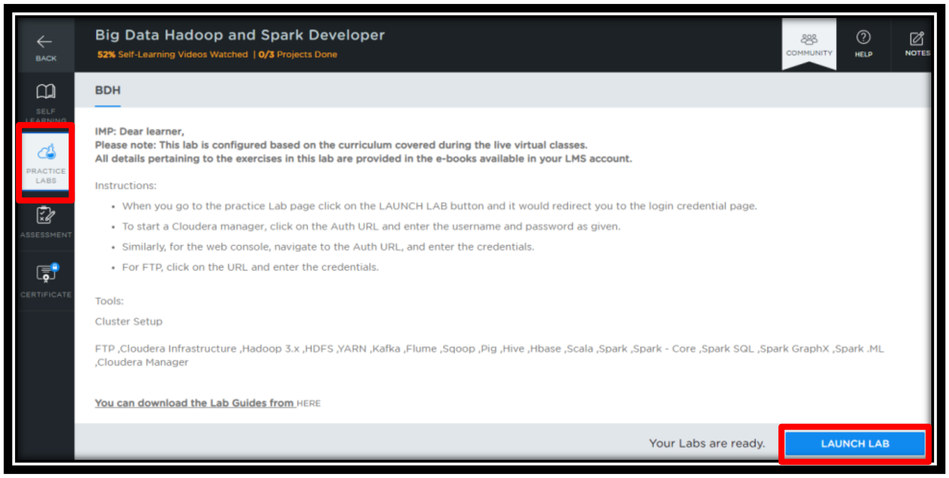
Note 1: You will need these to make the API calls in the next step.

Note 2: The keys in the screenshot above are hidden. However, in your own developer portal, you will be able to see the actual values for the API Key, API Secret Key, and Bearer Token.

Step 2: Set up the Flume configuration file on the “**Webconsole”**

2.1 Open the course “**Big Data Hadoop and Spark Developer”**

2.2 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:

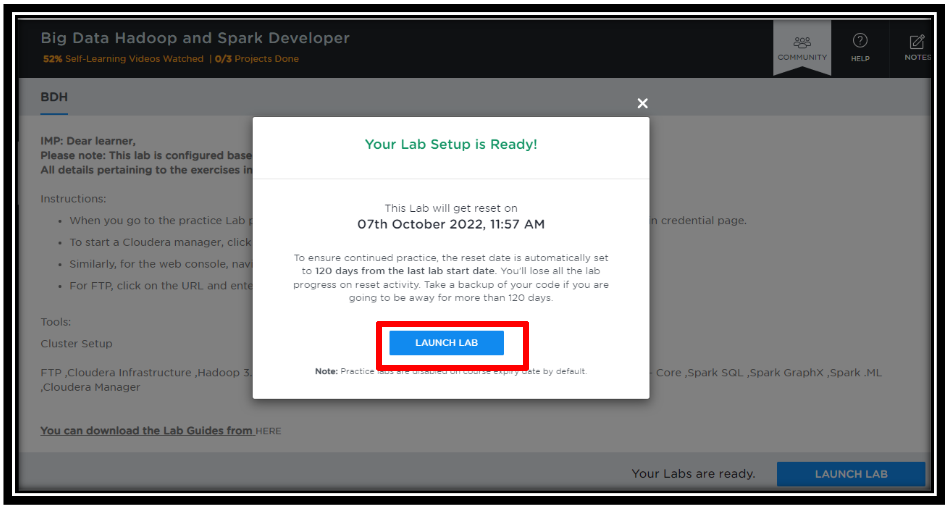
Cluster Setup

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.

2.3 Click on the **”LAUNCH LAB”** button



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 base
All details pertaining to the exercises in

Instructions:

- When you go to the practice Lab p … n credential page.
- To start a Cloudera manager, click
- Similarly, for the web console, navi
- For FTP, click on the URL and ente

Tools:

Cluster Setup

FTP ,Cloudera Infrastructure ,Hadoop 3. … - Core ,Spark SQL ,Spark GraphX ,Spark .ML ,Cloudera Manager

You can download the Lab Guides from HERE

Your Lab Setup is Ready!

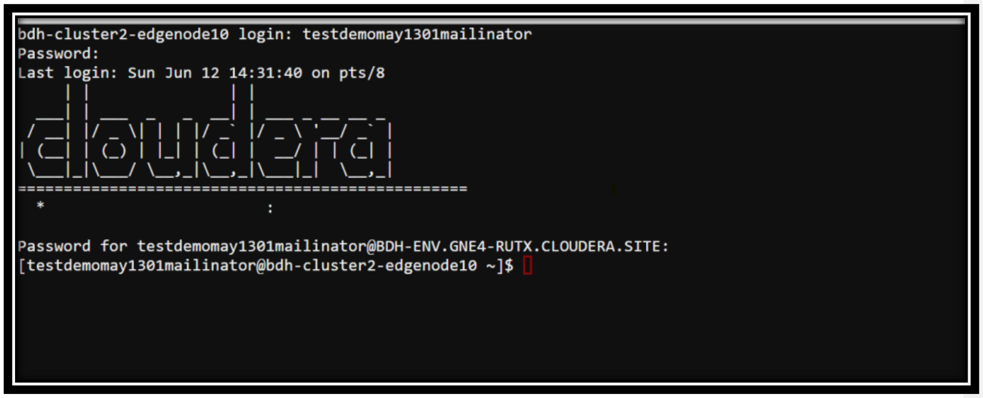
This Lab will get reset on
07th October 2022, 11:57 AM

To ensure continued practice, the reset date is automatically set to 120 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 120 days.

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

Your Labs are ready.

2.4 Log in to the **"Webconsole"** lab

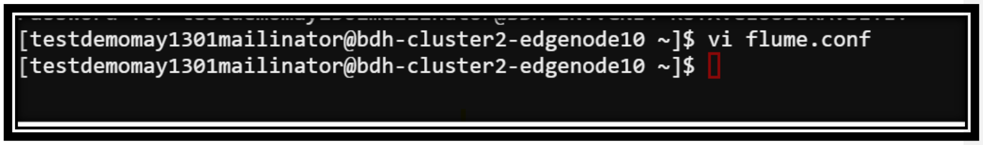


```
bdh-cluster2-edgenode10 login: testdemomay1301mailinator
Password:
Last login: Sun Jun 12 14:31:40 on pts/8
     | |               | |
  ___| |  ___  _   _  __| |  ___  _ __  __ _
 / __| | / _ \| | | |/ _` | / _ \| '__|/ _` |
| (__| || (_) | |_| | (_| ||  __/| |  | (_| |
 \___|_| \___/ \__,_|\__,_| \___||_|   \__,_|
=================================================
  *                         :

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

2.5 Run the command below to create a configuration file:

Command:

vi flume.conf



```
[testdemomay1301mailinator@bdh-cluster2-edgenode10 ~]$ vi flume.conf
[testdemomay1301mailinator@bdh-cluster2-edgenode10 ~]$ 
```

2.6 Add the below configurations with your secret access tokens received from the twitter API:

# Naming the components on the current agent.

TwitterAgent.sources = Twitter

TwitterAgent.channels = MemChannel

TwitterAgent.sinks = HDFS

# Describing/Configuring the source

TwitterAgent.sources.Twitter.type = org.apache.flume.source.twitter.TwitterSource

TwitterAgent.sources.Twitter.consumerKey = <your info>

TwitterAgent.sources.Twitter.consumerSecret = <your info>

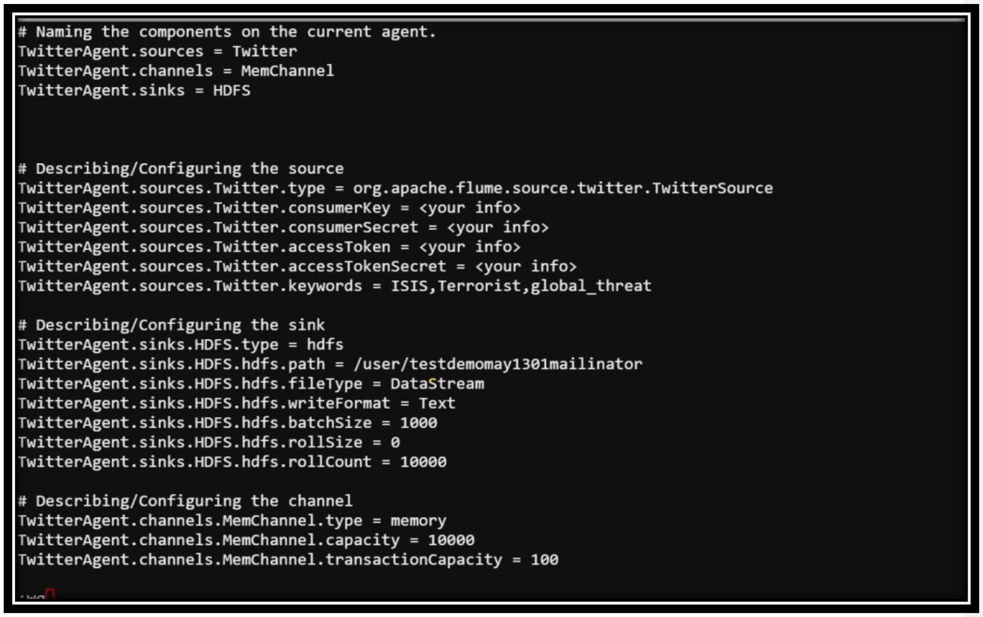
TwitterAgent.sources.Twitter.accessToken = <your info>

TwitterAgent.sources.Twitter.accessTokenSecret = <your info>

TwitterAgent.sources.Twitter.keywords = ISIS,Terrorist,global_threat

# Describing/Configuring the sink

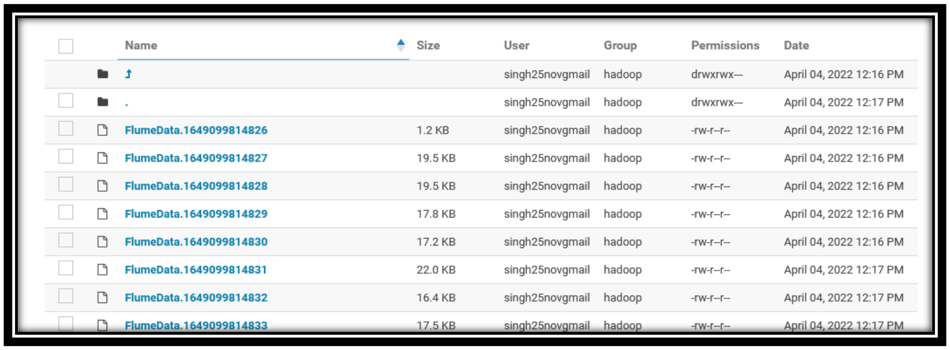
```
TwitterAgent.sinks.HDFS.type = hdfs
TwitterAgent.sinks.HDFS.hdfs.path = /user/testdemomay1301mailinator
TwitterAgent.sinks.HDFS.hdfs.fileType = DataStream
TwitterAgent.sinks.HDFS.hdfs.writeFormat = Text
TwitterAgent.sinks.HDFS.hdfs.batchSize = 1000
TwitterAgent.sinks.HDFS.hdfs.rollSize = 0
TwitterAgent.sinks.HDFS.hdfs.rollCount = 10000
# Describing/Configuring the channel
TwitterAgent.channels.MemChannel.type = memory
TwitterAgent.channels.MemChannel.capacity = 10000
TwitterAgent.channels.MemChannel.transactionCapacity = 100

# Binding the source and sink to the channel
TwitterAgent.sources.Twitter.channels = MemChannel
TwitterAgent.sinks.HDFS.channel = MemChannel
```



```
# Naming the components on the current agent.
TwitterAgent.sources = Twitter
TwitterAgent.channels = MemChannel
TwitterAgent.sinks = HDFS



# Describing/Configuring the source
TwitterAgent.sources.Twitter.type = org.apache.flume.source.twitter.TwitterSource
TwitterAgent.sources.Twitter.consumerKey = <your info>
TwitterAgent.sources.Twitter.consumerSecret = <your info>
TwitterAgent.sources.Twitter.accessToken = <your info>
TwitterAgent.sources.Twitter.accessTokenSecret = <your info>
TwitterAgent.sources.Twitter.keywords = ISIS,Terrorist,global_threat

# Describing/Configuring the sink
TwitterAgent.sinks.HDFS.type = hdfs
TwitterAgent.sinks.HDFS.hdfs.path = /user/testdemomay1301mailinator
TwitterAgent.sinks.HDFS.hdfs.fileType = DataStream
TwitterAgent.sinks.HDFS.hdfs.writeFormat = Text
TwitterAgent.sinks.HDFS.hdfs.batchSize = 1000
TwitterAgent.sinks.HDFS.hdfs.rollSize = 0
TwitterAgent.sinks.HDFS.hdfs.rollCount = 10000

# Describing/Configuring the channel
TwitterAgent.channels.MemChannel.type = memory
TwitterAgent.channels.MemChannel.capacity = 10000
TwitterAgent.channels.MemChannel.transactionCapacity = 100

:wq
```

2.7 Run the file on Console using the command below and verify the output on HDFS:

flume-ng agent -f /mnt/home/testdemomay1301mailinator/flume-demo/flume.conf -Dflume.root.logger=DEBUG,console -n TwitterAgent



□	Name	Size	User	Group	Permissions	Date
	↑		singh25novgmail	hadoop	drwxrwx---	April 04, 2022 12:16 PM
□	.		singh25novgmail	hadoop	drwxrwx---	April 04, 2022 12:17 PM
□	FlumeData.1649099814826	1.2 KB	singh25novgmail	hadoop	-rw-r--r--	April 04, 2022 12:16 PM
□	FlumeData.1649099814827	19.5 KB	singh25novgmail	hadoop	-rw-r--r--	April 04, 2022 12:16 PM
□	FlumeData.1649099814828	19.5 KB	singh25novgmail	hadoop	-rw-r--r--	April 04, 2022 12:16 PM
□	FlumeData.1649099814829	17.8 KB	singh25novgmail	hadoop	-rw-r--r--	April 04, 2022 12:16 PM
□	FlumeData.1649099814830	17.2 KB	singh25novgmail	hadoop	-rw-r--r--	April 04, 2022 12:16 PM
□	FlumeData.1649099814831	22.0 KB	singh25novgmail	hadoop	-rw-r--r--	April 04, 2022 12:17 PM
□	FlumeData.1649099814832	16.4 KB	singh25novgmail	hadoop	-rw-r--r--	April 04, 2022 12:17 PM
□	FlumeData.1649099814833	17.5 KB	singh25novgmail	hadoop	-rw-r--r--	April 04, 2022 12:17 PM