Big Data Hadoop and Spark Developer

Lesson-End Project Solution

Get Certified. Get Ahead.

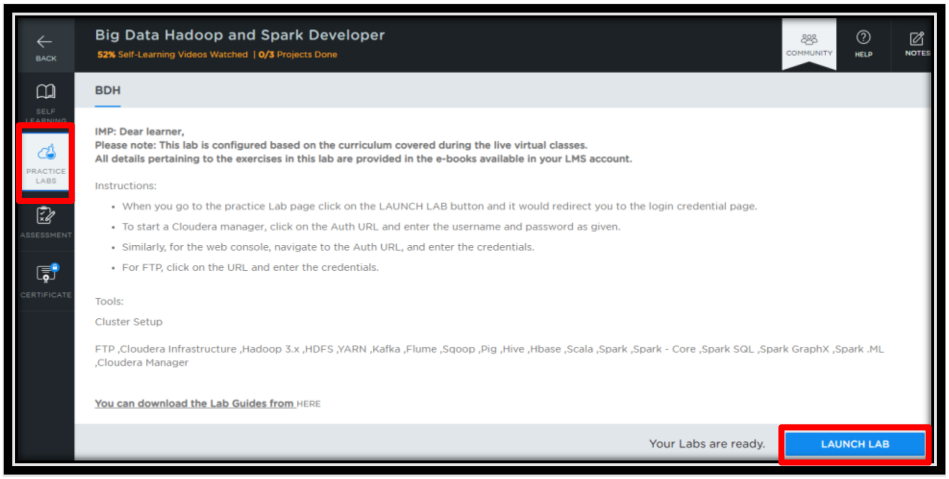
Post Office Data Analysis Using Hive

Steps to Perform:

Step 1: Download the lesson 6 dataset from the course resources and upload it to **“HDFS”**

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

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



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:

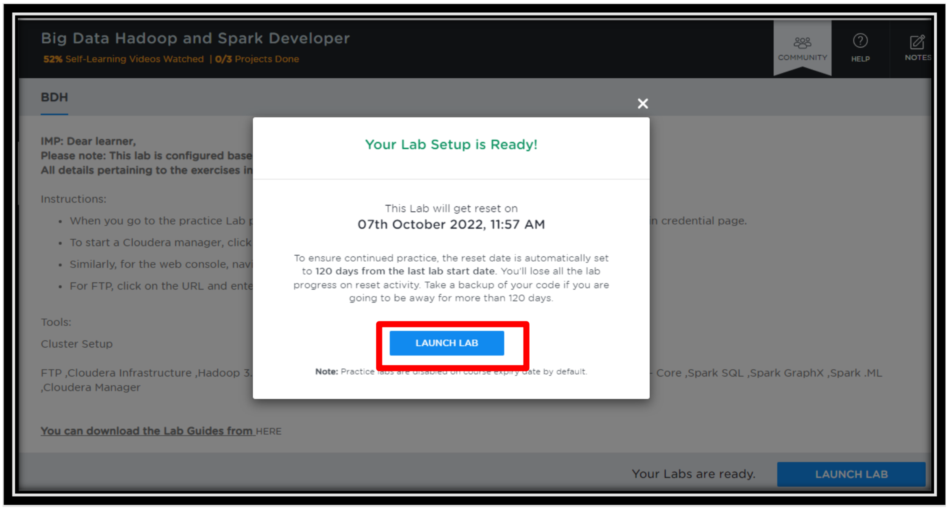
Cluster Setup

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.

1.3 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
×
LAUNCH LAB
LAUNCH LAB

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

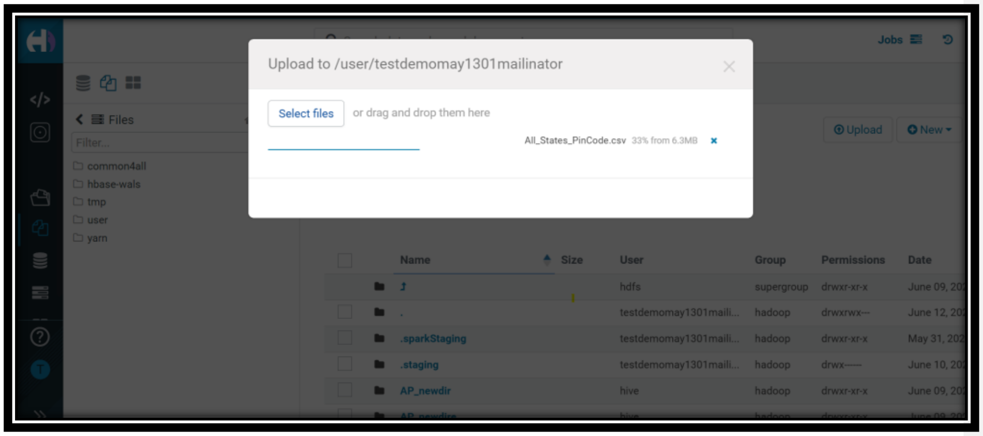
To ensure continued practice, the reset date is automatically set to 120 days from the last lab start date. You'll lose all the lab progress on reset activity. Take a backup of your code if you are going to be away for more than 120 days.

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

Your Labs are ready.

1.4 Log in to the **“HUE”** lab

1.5 Click on **“HDFS”** and upload the downloaded dataset

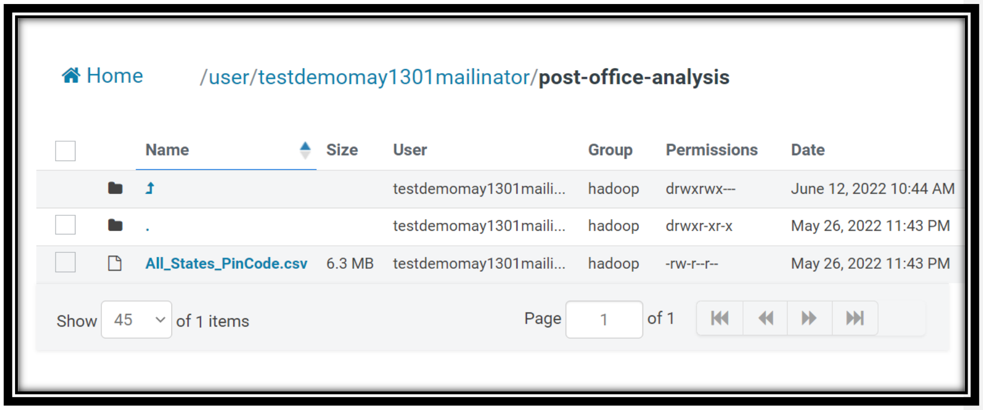


Upload to /user/testdemomay1301mailinator

or drag and drop them here

All_States_PinCode.csv 33% from 6.3MB

Name	Size	User	Group	Permissions	Date
⤴		hdfs	supergroup	drwxr-xr-x	June 09, 20
.		testdemomay1301maili...	hadoop	drwxrwx---	June 12, 20
.sparkStaging		testdemomay1301maili...	hadoop	drwxr-xr-x	May 31, 20
.staging		testdemomay1301maili...	hadoop	drwx------	June 10, 20
AP_newdir		hive	hadoop	drwxr-xr-x	June 09, 20



/user/testdemomay1301mailinator/**post-office-analysis**

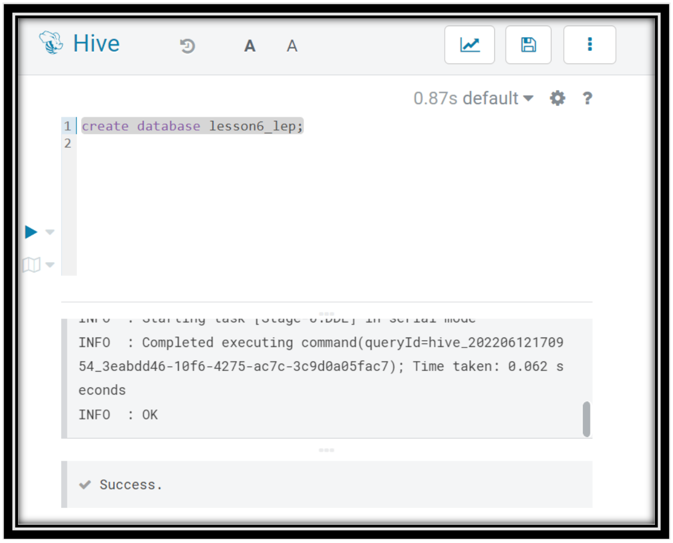
□		Name	Size	User	Group	Permissions	Date
		↥		testdemomay1301maili...	hadoop	drwxrwx---	June 12, 2022 10:44 AM
□		.		testdemomay1301maili...	hadoop	drwxr-xr-x	May 26, 2022 11:43 PM
□		All_States_PinCode.csv	6.3 MB	testdemomay1301maili...	hadoop	-rw-r--r--	May 26, 2022 11:43 PM

Show 45 of 1 items   Page 1 of 1

Step 2: Create a table on the Hive editor

2.1 Create a database

Command:
create database lesson6_lep;



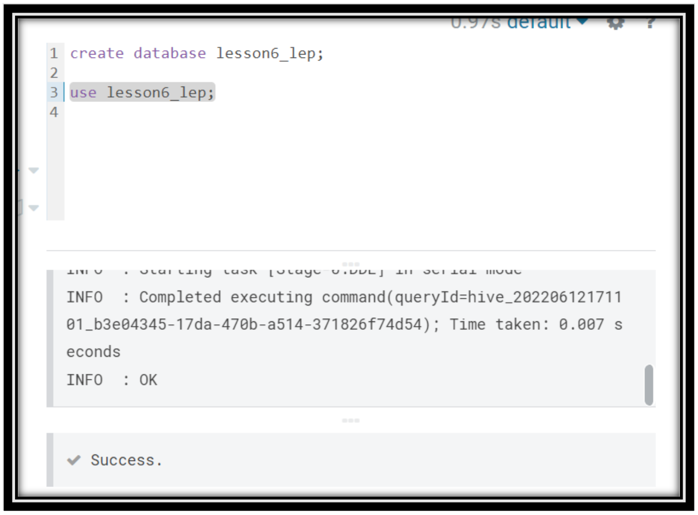
```
create database lesson6_lep;
```

```
INFO  : Starting task [Stage-0:DDL] in serial mode
INFO  : Completed executing command(queryId=hive_202206121709
54_3eabdd46-10f6-4275-ac7c-3c9d0a05fac7); Time taken: 0.062 s
econds
INFO  : OK
```

```
✓ Success.
```

2.2 Use the created database

Command:
use lesson6_lep;



```
create database lesson6_lep;

use lesson6_lep;
```

```
INFO  : Starting task [Stage-0:DDL] in serial mode
INFO  : Completed executing command(queryId=hive_202206121711
01_b3e04345-17da-470b-a514-371826f74d54); Time taken: 0.007 s
econds
INFO  : OK
```

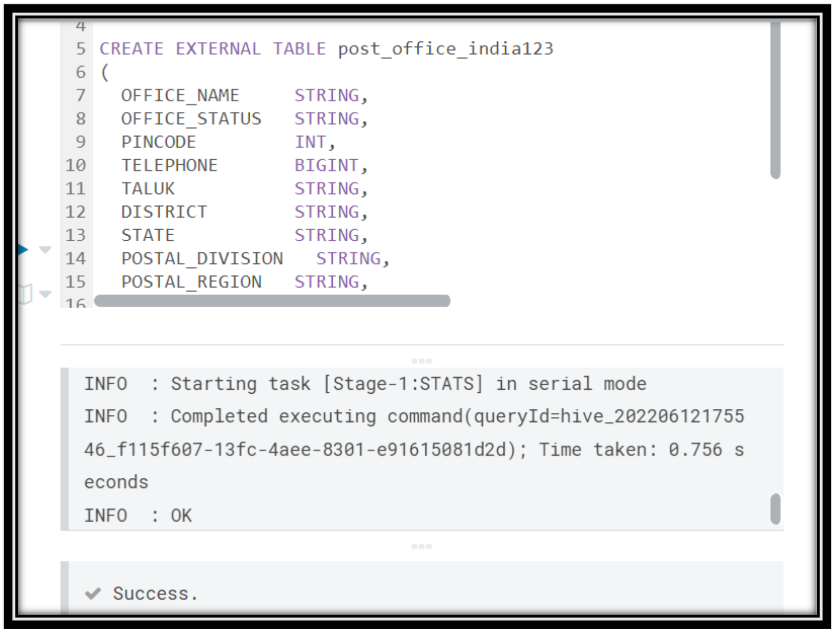
```
✔ Success.
```

2.3 Create the table

Command:

```
CREATE EXTERNAL TABLE post_office_india123
(
        OFFICE_NAME       STRING,
        OFFICE_STATUS     STRING,
        PINCODE           INT,
        TELEPHONE         BIGINT,
        TALUK        STRING,
        DISTRICT     STRING,
        STATE        STRING,
        POSTAL_DIVISION   STRING,
        POSTAL_REGION     STRING,
        POSTAL_CIRCLE     STRING
)
ROW FORMAT DELIMITED
FIELDS TERMINATED BY ','
```

STORED AS TEXTFILE;



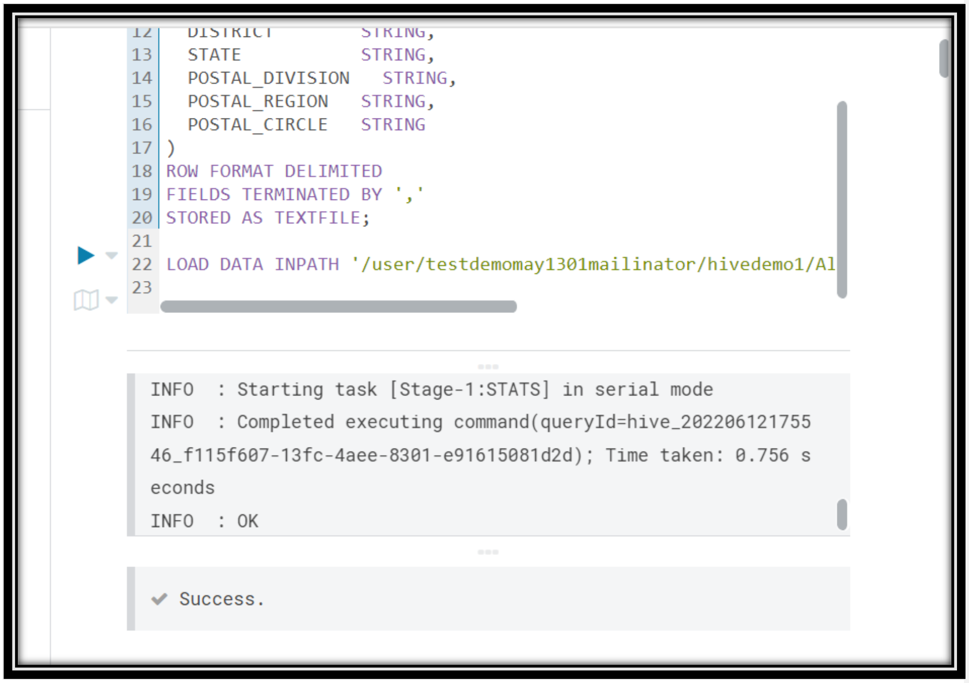
```
CREATE EXTERNAL TABLE post_office_india123
(
  OFFICE_NAME     STRING,
  OFFICE_STATUS   STRING,
  PINCODE         INT,
  TELEPHONE       BIGINT,
  TALUK           STRING,
  DISTRICT        STRING,
  STATE           STRING,
  POSTAL_DIVISION   STRING,
  POSTAL_REGION   STRING,
```

```
INFO  : Starting task [Stage-1:STATS] in serial mode
INFO  : Completed executing command(queryId=hive_202206121755
46_f115f607-13fc-4aee-8301-e91615081d2d); Time taken: 0.756 s
econds
INFO  : OK
```

✓ Success.

2.4 Load the CSV file into the table

Command:
LOAD DATA INPATH '/user/testdemomay1301mailinator/All_States_PinCode.csv' INTO TABLE post_office_india123;



```
  DISTRICT         STRING,
  STATE            STRING,
  POSTAL_DIVISION    STRING,
  POSTAL_REGION    STRING,
  POSTAL_CIRCLE    STRING
)
ROW FORMAT DELIMITED
FIELDS TERMINATED BY ','
STORED AS TEXTFILE;

LOAD DATA INPATH '/user/testdemomay1301mailinator/hivedemo1/Al
```

```
INFO  : Starting task [Stage-1:STATS] in serial mode
INFO  : Completed executing command(queryId=hive_202206121755
46_f115f607-13fc-4aee-8301-e91615081d2d); Time taken: 0.756 s
econds
INFO  : OK
```

✓ Success.

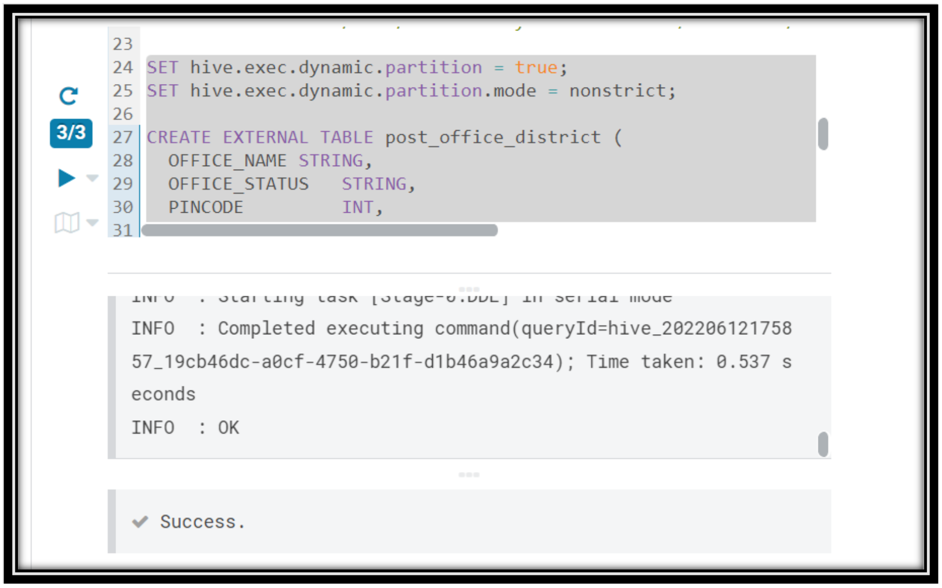
2.5 Create a partitioned table to fetch data easily

Configure the Hive to support dynamic partition creation and enter the following set commands.

```
SET hive.exec.dynamic.partition = true;
SET hive.exec.dynamic.partition.mode = nonstrict;
```

```
CREATE EXTERNAL TABLE post_office_district (
        OFFICE_NAME STRING,
        OFFICE_STATUS       STRING,
        PINCODE             INT,
        TELEPHONE   BIGINT,
        TALUK        STRING,
        DISTRICT STRING,
        POSTAL_DIVISION   STRING,
        POSTAL_REGION     STRING,
        POSTAL_CIRCLE       STRING
)
```

PARTITIONED BY (STATE STRING)
ROW FORMAT DELIMITED
FIELDS TERMINATED BY ','
STORED AS TEXTFILE;

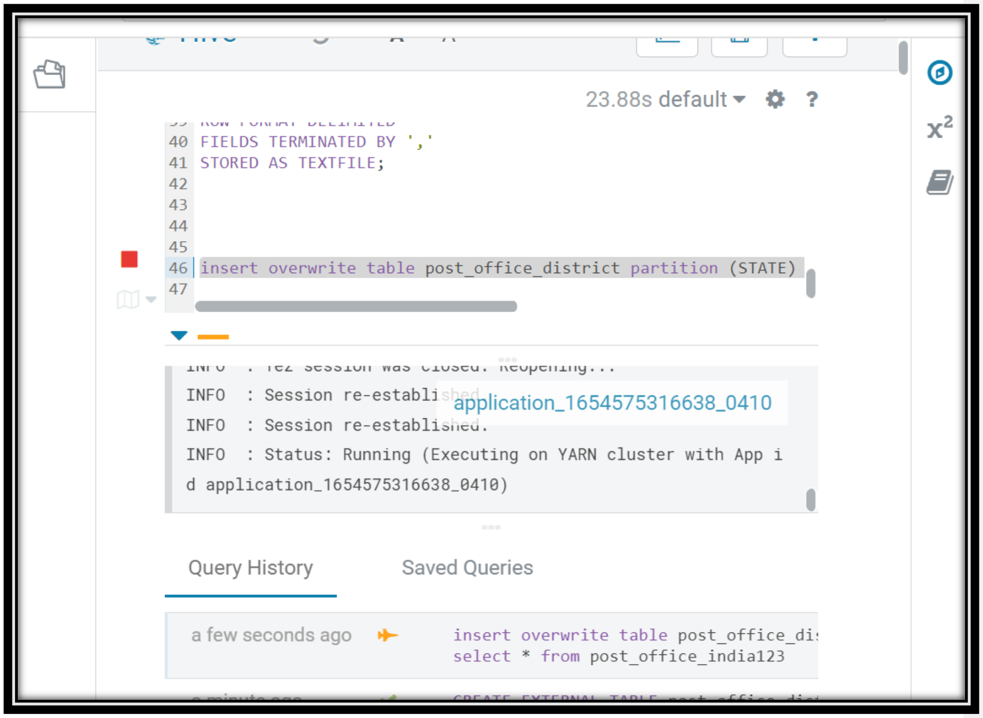


```
SET hive.exec.dynamic.partition = true;
SET hive.exec.dynamic.partition.mode = nonstrict;

CREATE EXTERNAL TABLE post_office_district (
  OFFICE_NAME STRING,
  OFFICE_STATUS   STRING,
  PINCODE         INT,
```

```
INFO  : Starting task [Stage-0:DDL] in serial mode
INFO  : Completed executing command(queryId=hive_202206121758
57_19cb46dc-a0cf-4750-b21f-d1b46a9a2c34); Time taken: 0.537 s
econds
INFO  : OK
```

✓ Success.

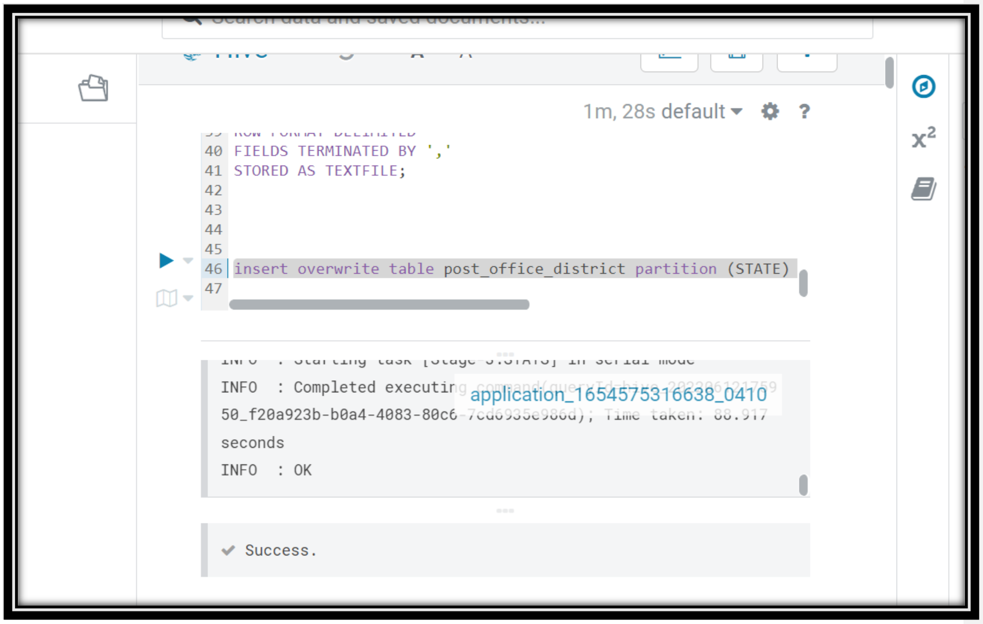
insert overwrite table post_office_district partition (STATE) select * from post_office_india123;



```
ROW FORMAT DELIMITED
FIELDS TERMINATED BY ','
STORED AS TEXTFILE;




insert overwrite table post_office_district partition (STATE)

```

```
INFO  : Tez session was closed. Reopening...
INFO  : Session re-establi
INFO  : Session re-established.
INFO  : Status: Running (Executing on YARN cluster with App i
d application_1654575316638_0410)
```

```
a few seconds ago      insert overwrite table post_office_di
                       select * from post_office_india123
a minute ago           CREATE EXTERNAL TABLE post_office_dis
```



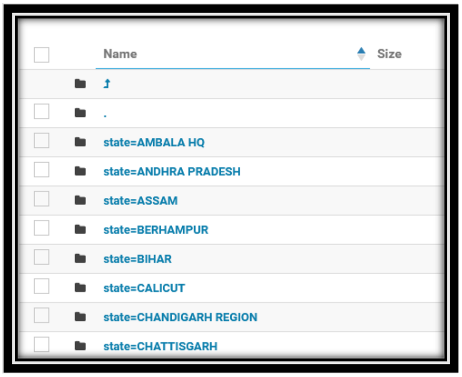
Search data and saved documents...
1m, 28s default
?
application_1654575316638_0410

```
ROW FORMAT DELIMITED
FIELDS TERMINATED BY ','
STORED AS TEXTFILE;




insert overwrite table post_office_district partition (STATE)

```

```
INFO  : Starting task [Stage-3:STATS] in serial mode
INFO  : Completed executing command(queryId=hive_20220612175950_f20a923b-b0a4-4083-80c6-7cd6935e986d); Time taken: 88.917
seconds
INFO  : OK
```

✓ Success.

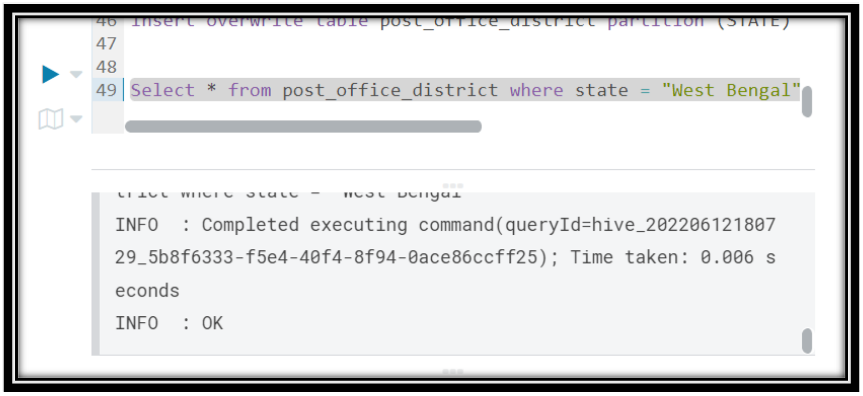
Note: The above command changes the setting only for a single session. This will create multiple partitions in the HDFS directory with each state as a subfolder.



		Name	Size
	🖿	↑	
□	🖿	.	
□	🖿	state=AMBALA HQ	
□	🖿	state=ANDHRA PRADESH	
□	🖿	state=ASSAM	
□	🖿	state=BERHAMPUR	
□	🖿	state=BIHAR	
□	🖿	state=CALICUT	
□	🖿	state=CHANDIGARH REGION	
□	🖿	state=CHATTISGARH	

2.7 Run the query where the state is West Bengal

Select * from post_office_district where state = 'West Bengal';



```
insert overwrite table post_office_district partition (STATE)


Select * from post_office_district where state = "West Bengal"
```

```
trict where state = 'West Bengal'
INFO  : Completed executing command(queryId=hive_202206121807
29_5b8f6333-f5e4-40f4-8f94-0ace86ccff25); Time taken: 0.006 s
econds
INFO  : OK
```

In comparison to the previous runs, which took 30 seconds to complete, this will return the results in seconds.