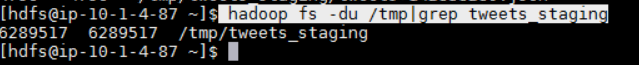
Session-2 - Lab 2 – Creating Hive Table

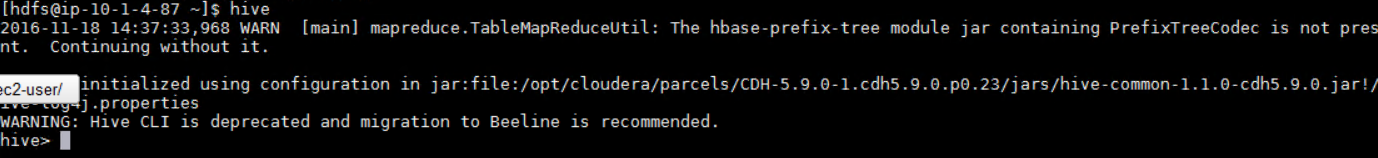
# Create Hive table on Tweets

* + 1. In this step we will create external hive table which we will leverage to extract the data and gather data quality and usefulness.
    2. First verify that we our data collection stream is capturing data from “twitter firehose”

|  |
| --- |
| hadoop fs -du /tmp|grep tweets\_staging |

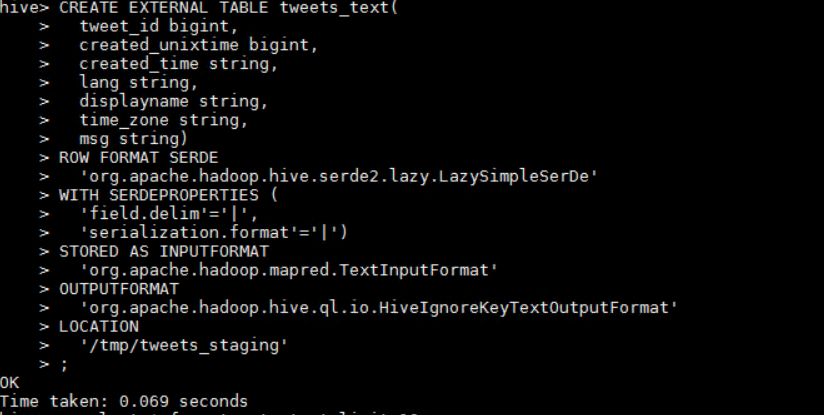


* + 1. If NiFi stream is operational, you should see change in size of dataset.
    2. To create hive table, first we need to login to hive or beeline(Hive cli replacement)
       1. Login to hive by typing “hive”



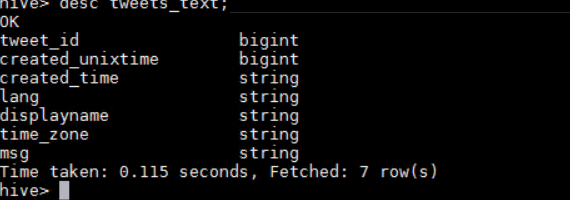
* + - 1. By default there is database named “default”, we will use default for this lab. Let’s create new table called “tweets\_text.”

|  |
| --- |
| CREATE EXTERNAL TABLE tweets\_text(  tweet\_id bigint,  created\_unixtime bigint,  created\_time string,  lang string,  displayname string,  time\_zone string,  msg string)  ROW FORMAT SERDE  'org.apache.hadoop.hive.serde2.lazy.LazySimpleSerDe'  WITH SERDEPROPERTIES (  'field.delim'='|',  'serialization.format'='|')  STORED AS INPUTFORMAT  'org.apache.hadoop.mapred.TextInputFormat'  OUTPUTFORMAT  'org.apache.hadoop.hive.ql.io.HiveIgnoreKeyTextOutputFormat'  LOCATION  '/tmp/tweets\_staging'  ; |



|  |
| --- |
| desc tweets\_text; |

* + - 1. Verify the table creation was successful:



* + - 1. Notice that data is already in text and readable format, NiFi processed and cleaned the data prior to writing to HDFS. Also, data can be written in its raw format, but tools like NiFi workflow makes it simple for users to scrub the data.

# Query data via CLI

* + 1. Now that we have data ingested into HDFS, let’s execute some basic queries to find out data structure.
       1. Query table to find total row counts

|  |
| --- |
| **Select count(\*) from tweets\_text;** |

* + - 1. Query table to find total tweets count by “displayname”

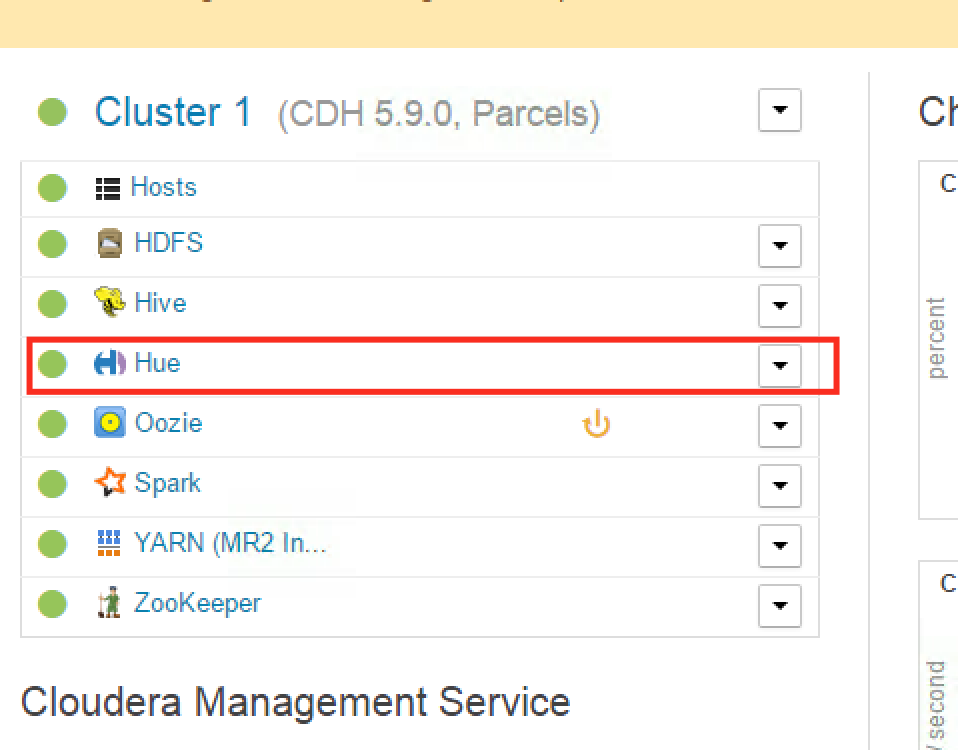
|  |
| --- |
| **select count(\*) , displayname from tweets\_text group by displayname;** |

* + - 1. Query table and find out users with more than 50 tweets

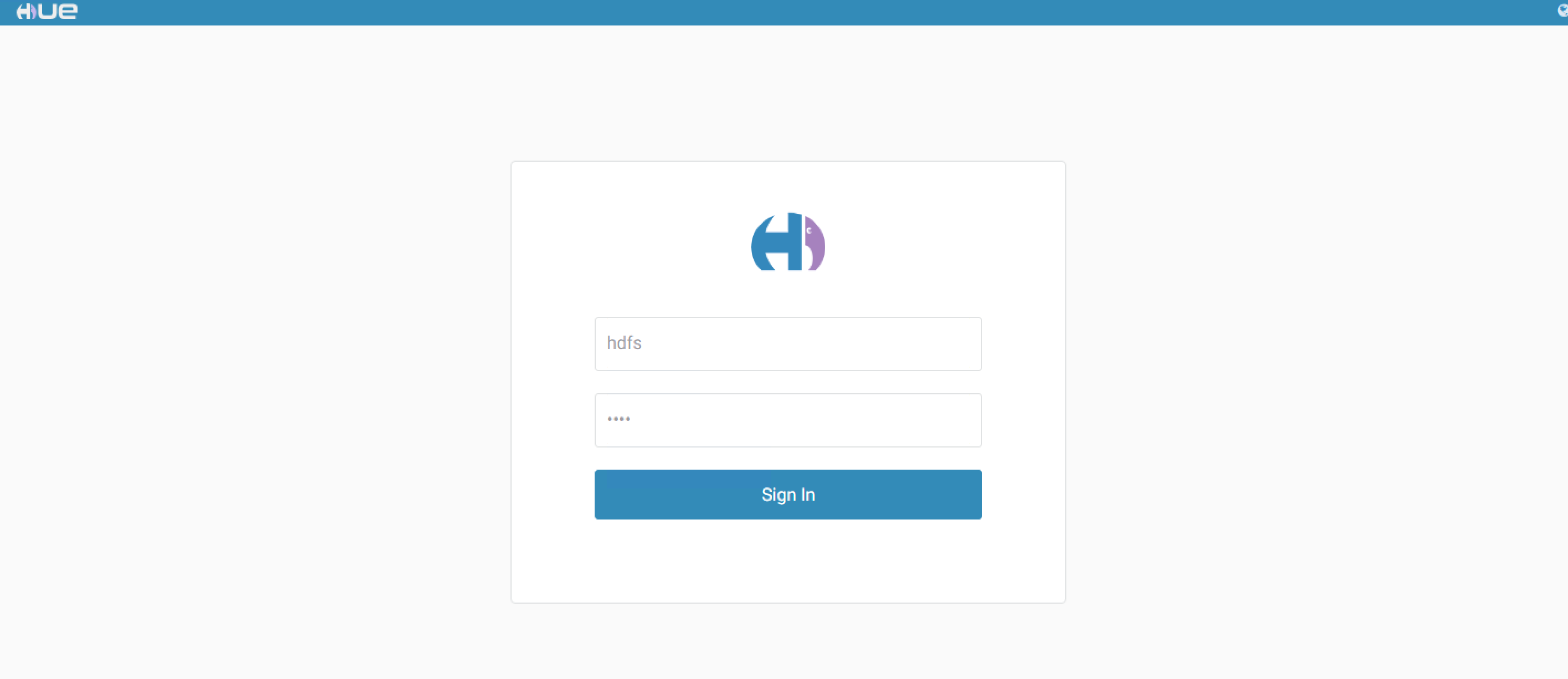
|  |
| --- |
| **select count(\*) , displayname from tweets\_text group by displayname having count(\*) > 50;** |

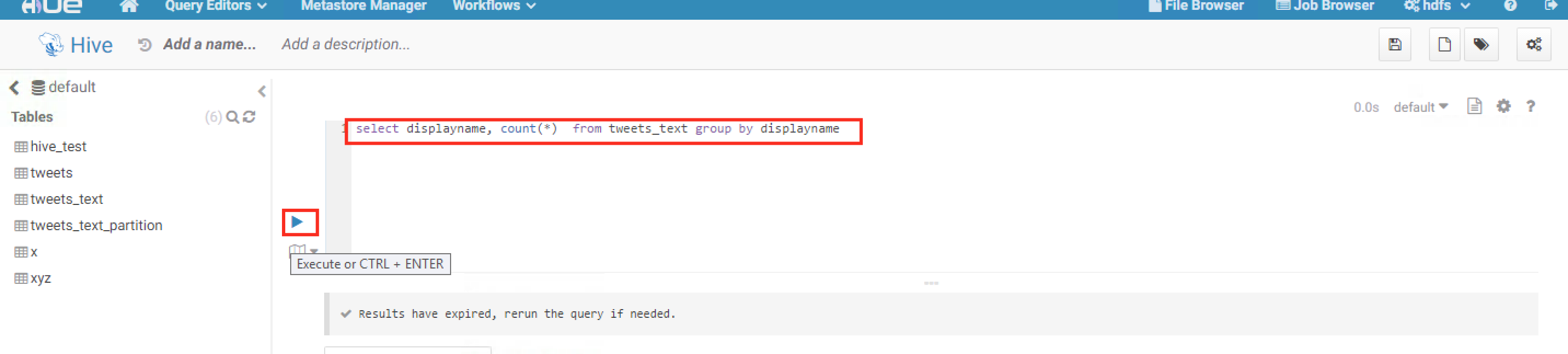
# Query data via CLI and Hue

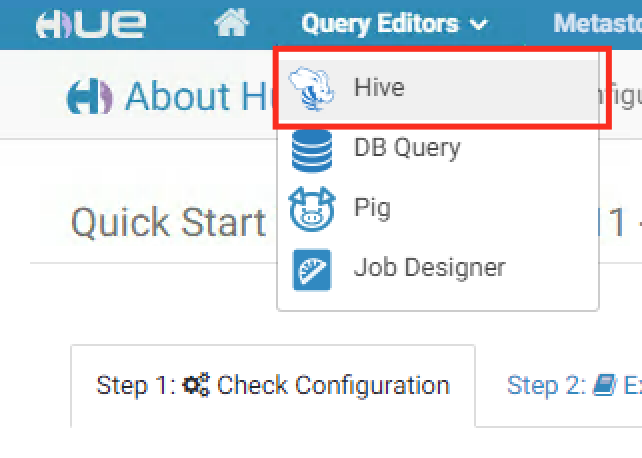
* + 1. Let’s query data via HUE(Hadoop User Experience), by using hue we can take data and convert it into pie charts, plot charts and so on.
    2. Hue is good starter tool for users that are not experienced or comfortable with command line.
    3. Login to Hue using hdfs user (First login becomes super user for HUE) via cloudera manager
       - Login to CM



* Login to HUE with hdfs user, password can be whatever you like



* + - * From drop down menu, select HIVE
      * 



Execute the queries from previous steps.