

Sqoop crash course Project

Developed by: Arturo Quintanilla

Description:

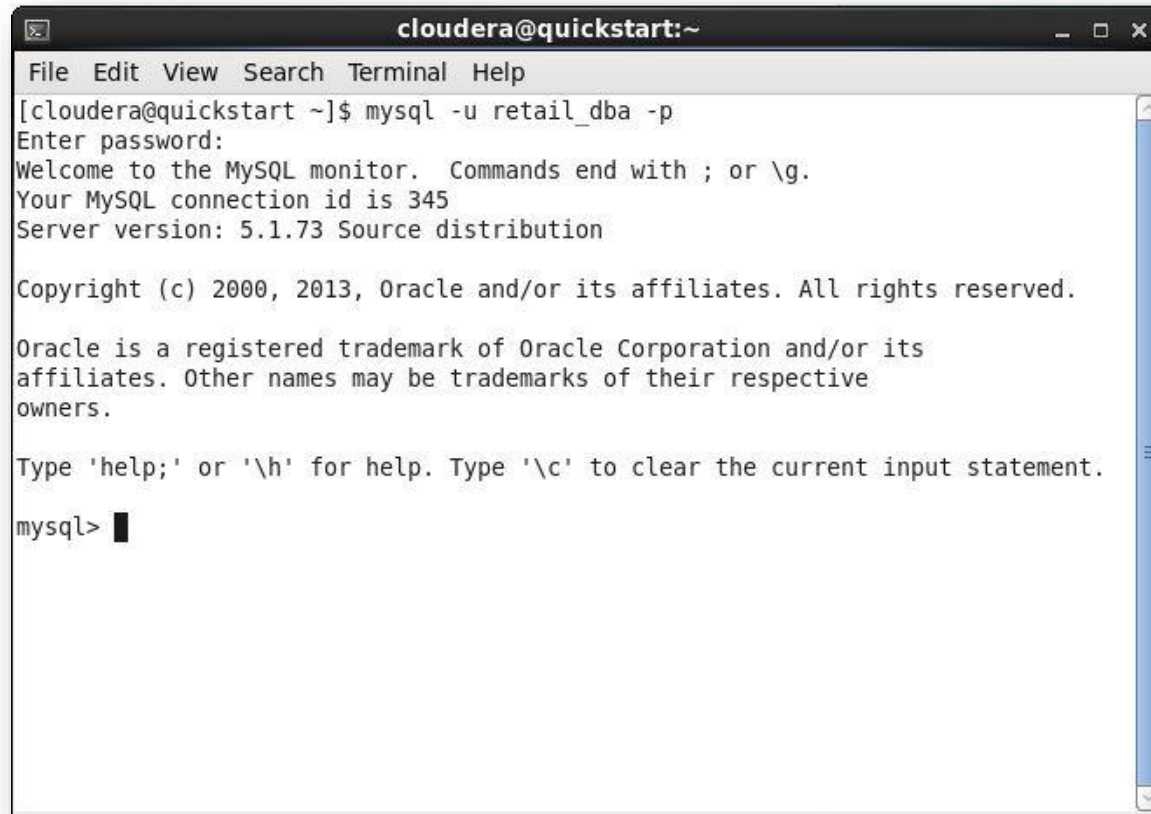
Problem Statement

- ▶ List databases: list tables and validate that you can connect to database using Sqoop with JDBC URL. As part of validation you should query from one of the tables in MySQL
- ▶ Create a HR_db Database in MySql and creat an Employee table (feel free to add fields, populate some sample data with a good amount of records)
- ▶ Eval connection with Sqoop command
- ▶ Importing HR_db tables to HDFS i- avro data file format to Hive
- ▶ Use 8 parallel threads (mappers)

Procedure: Part 1

List databases

First start mysql CLI from cloudera terminal.



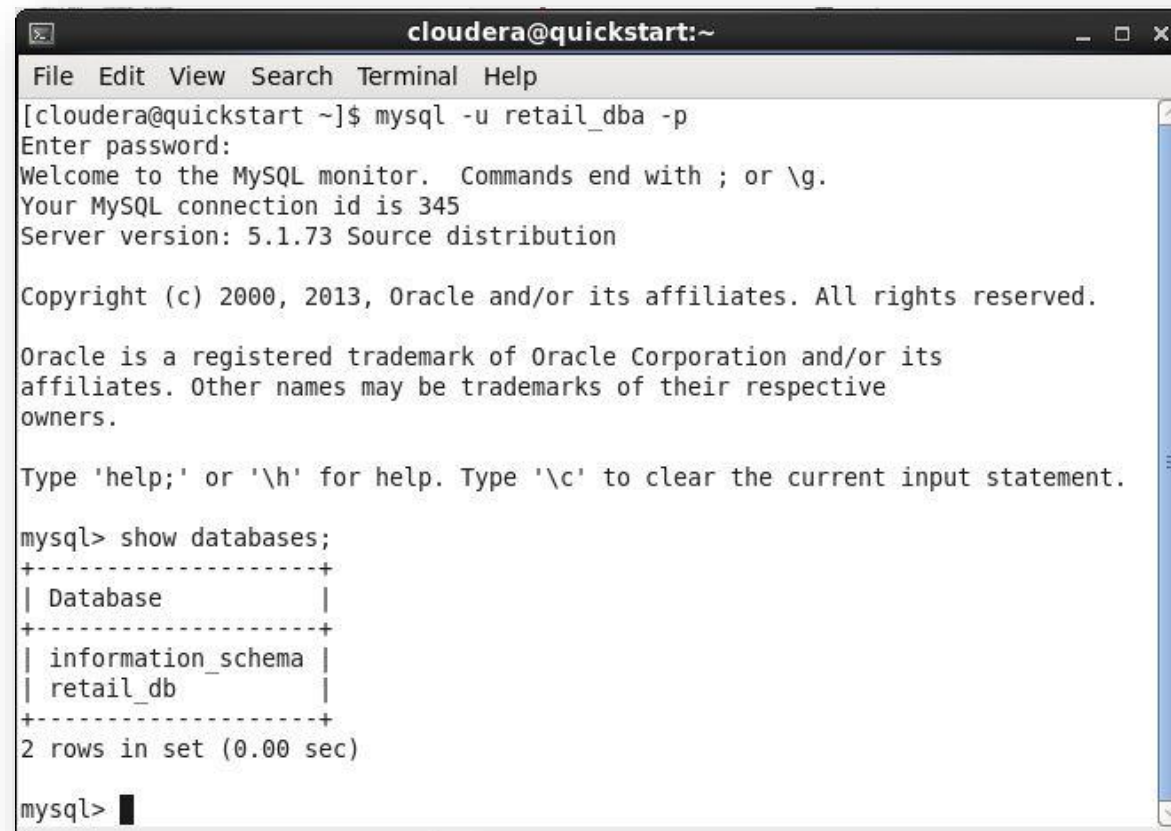
```
cloudera@quickstart:~
File Edit View Search Terminal Help
[cloudera@quickstart ~]$ mysql -u retail_dba -p
Enter password:
Welcome to the MySQL monitor.  Commands end with ; or \g.
Your MySQL connection id is 345
Server version: 5.1.73 Source distribution

Copyright (c) 2000, 2013, Oracle and/or its affiliates. All rights reserved.

Oracle is a registered trademark of Oracle Corporation and/or its
affiliates. Other names may be trademarks of their respective
owners.

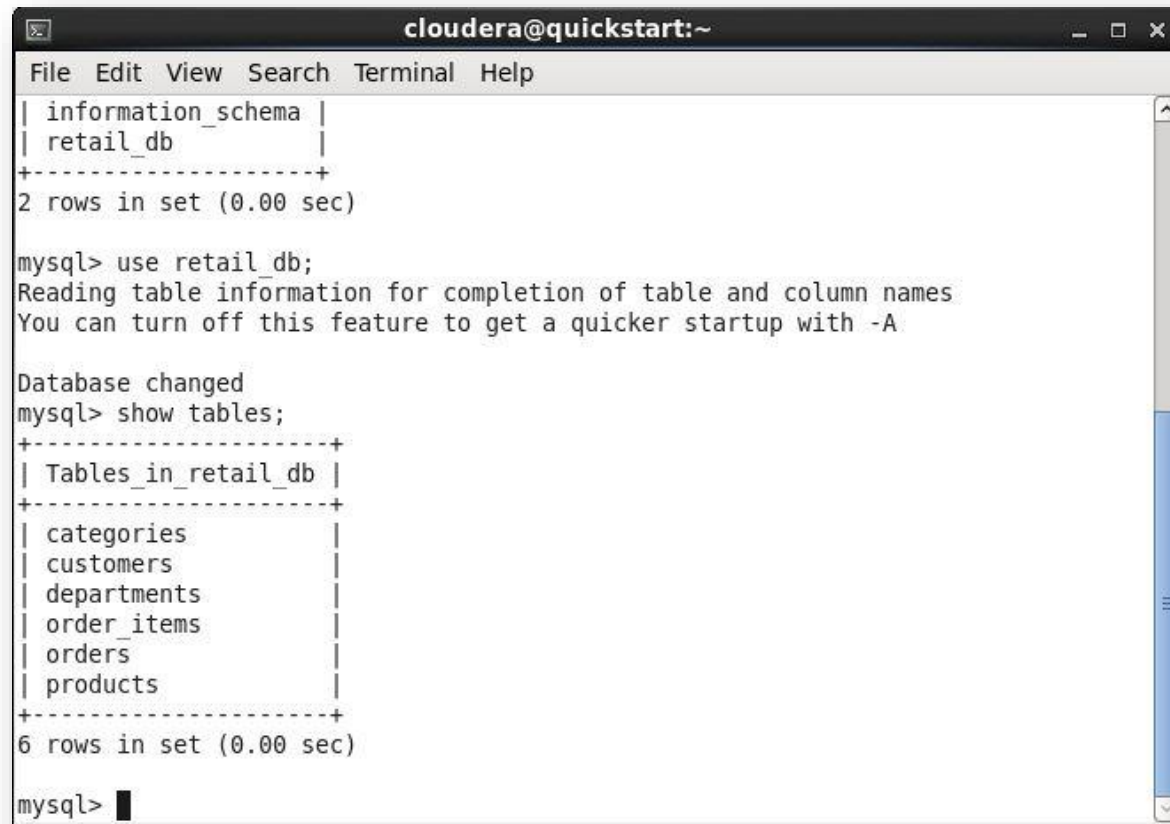
Type 'help;' or '\h' for help. Type '\c' to clear the current input statement.
mysql>
```

Use “show databases” command to view all available databases.

A terminal window titled 'cloudera@quickstart:~' with a menu bar (File, Edit, View, Search, Terminal, Help). The terminal shows the execution of 'mysql -u retail_dba -p', followed by a password prompt and MySQL welcome messages. The 'show databases;' command is executed, resulting in a table listing 'information_schema' and 'retail_db'.

```
cloudera@quickstart:~  
File Edit View Search Terminal Help  
[cloudera@quickstart ~]$ mysql -u retail_dba -p  
Enter password:  
Welcome to the MySQL monitor.  Commands end with ; or \g.  
Your MySQL connection id is 345  
Server version: 5.1.73 Source distribution  
  
Copyright (c) 2000, 2013, Oracle and/or its affiliates. All rights reserved.  
  
Oracle is a registered trademark of Oracle Corporation and/or its  
affiliates. Other names may be trademarks of their respective  
owners.  
  
Type 'help;' or '\h' for help. Type '\c' to clear the current input statement.  
  
mysql> show databases;  
+-----+  
| Database |  
+-----+  
| information_schema |  
| retail_db |  
+-----+  
2 rows in set (0.00 sec)  
  
mysql> █
```

Retail database will be used for this test. So we use the “use retail_db” command to set our working database and “show tables” command to view all available tables from our database.

A screenshot of a terminal window titled "cloudera@quickstart:~". The terminal shows a MySQL prompt where the user has entered "use retail_db;" and "show tables;". The output of "show tables;" lists six tables: categories, customers, departments, order_items, orders, and products. The terminal window has a menu bar with "File", "Edit", "View", "Search", "Terminal", and "Help".

```
cloudera@quickstart:~  
File Edit View Search Terminal Help  
+-----+  
| information_schema |  
| retail_db          |  
+-----+  
2 rows in set (0.00 sec)  
  
mysql> use retail_db;  
Reading table information for completion of table and column names  
You can turn off this feature to get a quicker startup with -A  
  
Database changed  
mysql> show tables;  
+-----+  
| Tables_in_retail_db |  
+-----+  
| categories          |  
| customers            |  
| departments          |  
| order_items          |  
| orders               |  
| products             |  
+-----+  
6 rows in set (0.00 sec)  
  
mysql> █
```

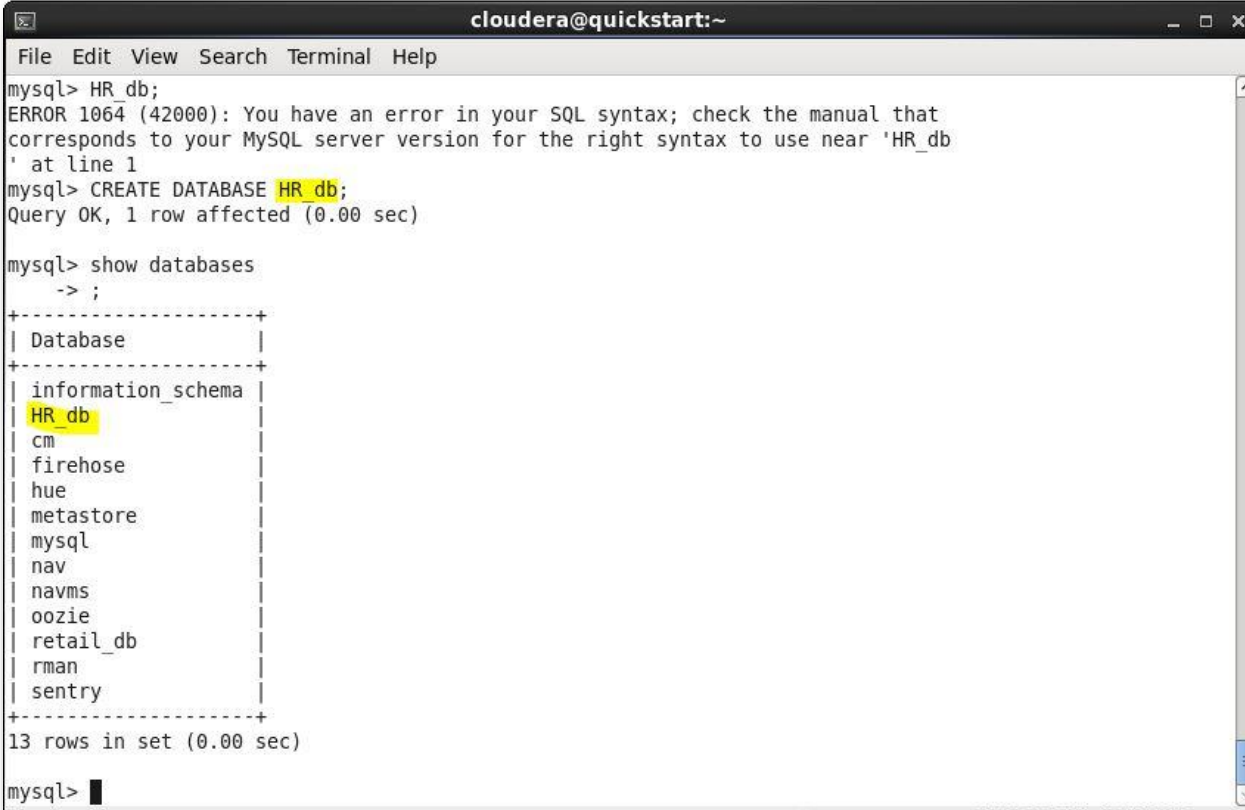
Finally for testing purposes a select query was execute to see all data available on the departments table.

```
cloudera@quickstart:~  
File Edit View Search Terminal Help  
+-----+-----+-----+-----+  
| 68878 | 2014-07-08 00:00:00 | 6753 | COMPLETE |  
| 68879 | 2014-07-09 00:00:00 | 778 | COMPLETE |  
| 68880 | 2014-07-13 00:00:00 | 1117 | COMPLETE |  
| 68881 | 2014-07-19 00:00:00 | 2518 | PENDING_PAYMENT |  
| 68882 | 2014-07-22 00:00:00 | 10000 | ON_HOLD |  
| 68883 | 2014-07-23 00:00:00 | 5533 | COMPLETE |  
+-----+-----+-----+-----+  
68883 rows in set (0.12 sec)  
  
mysql> select * from departments  
-> ;  
+-----+-----+  
| department_id | department_name |  
+-----+-----+  
| 2 | Fitness |  
| 3 | Footwear |  
| 4 | Apparel |  
| 5 | Golf |  
| 6 | Outdoors |  
| 7 | Fan Shop |  
+-----+-----+  
6 rows in set (0.00 sec)  
  
mysql> █
```

Procedure: Part 2

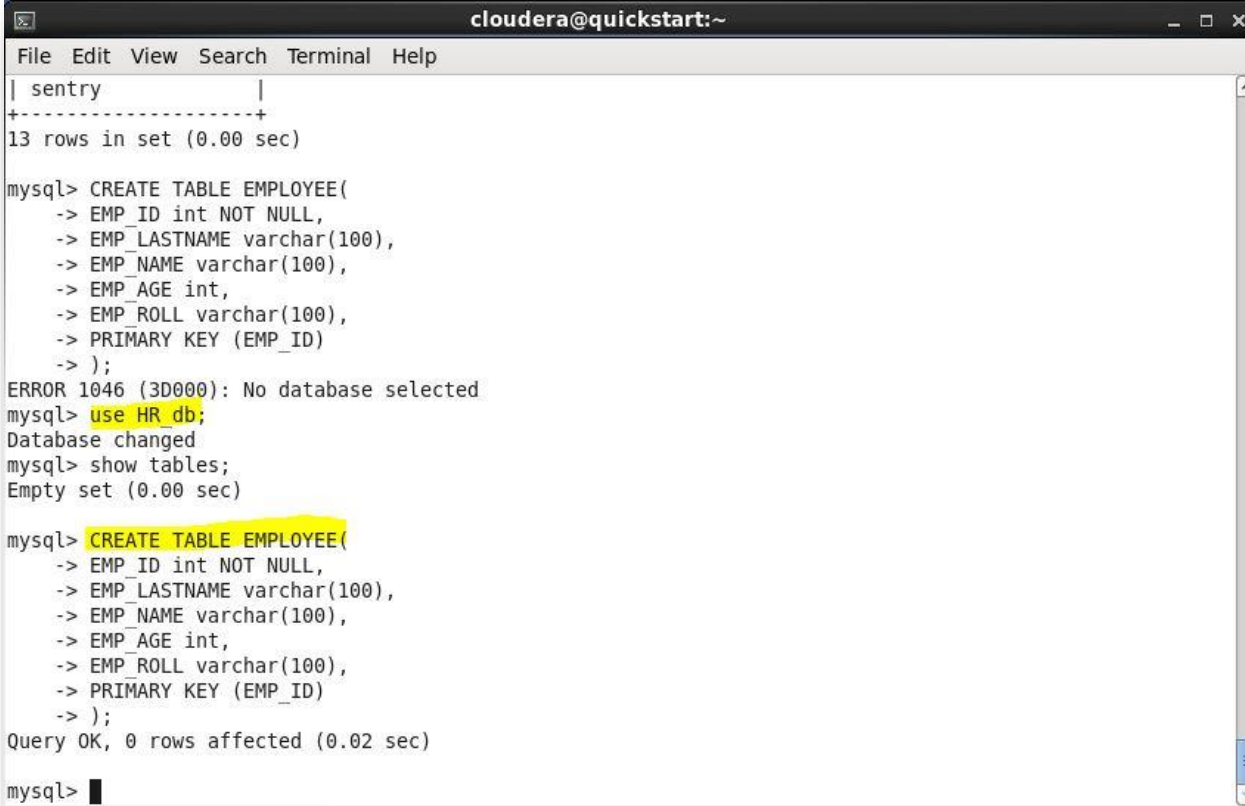
create databases

To create a new database in mysql just execute “CREATE DATABASE <name>;” In this case our database will be named “HR_db”.



```
cloudera@quickstart:~  
File Edit View Search Terminal Help  
mysql> HR_db;  
ERROR 1064 (42000): You have an error in your SQL syntax; check the manual that  
corresponds to your MySQL server version for the right syntax to use near 'HR_db  
' at line 1  
mysql> CREATE DATABASE HR_db;  
Query OK, 1 row affected (0.00 sec)  
  
mysql> show databases  
-> ;  
+-----+  
| Database |  
+-----+  
| information_schema |  
| HR_db |  
| cm |  
| firehose |  
| hue |  
| metastore |  
| mysql |  
| nav |  
| navms |  
| oozie |  
| retail_db |  
| rman |  
| sentry |  
+-----+  
13 rows in set (0.00 sec)  
  
mysql> █
```

Now that we have our database we'll create a table using the “CREATE TABLE <name>” command and also inserting data into it using “INSERT INTO <table> VALUES(var1,var2,va3);”

A terminal window titled 'cloudera@quickstart:~' with a menu bar (File, Edit, View, Search, Terminal, Help). The terminal shows a MySQL session. It starts with a query result for 'sentry' showing 13 rows. Then, the user enters 'mysql> CREATE TABLE EMPLOYEE(...);' which results in an error: 'ERROR 1046 (3D000): No database selected'. The user then enters 'mysql> use HR_db;' and receives 'Database changed'. Next, 'mysql> show tables;' returns 'Empty set (0.00 sec)'. Finally, the user enters 'mysql> CREATE TABLE EMPLOYEE(...);' again, and this time it succeeds with 'Query OK, 0 rows affected (0.02 sec)'. The prompt 'mysql>' is visible at the bottom.

```
cloudera@quickstart:~  
File Edit View Search Terminal Help  
| sentry |  
+-----+  
13 rows in set (0.00 sec)  
  
mysql> CREATE TABLE EMPLOYEE(  
-> EMP_ID int NOT NULL,  
-> EMP_LASTNAME varchar(100),  
-> EMP_NAME varchar(100),  
-> EMP_AGE int,  
-> EMP_ROLL varchar(100),  
-> PRIMARY KEY (EMP_ID)  
-> );  
ERROR 1046 (3D000): No database selected  
mysql> use HR_db;  
Database changed  
mysql> show tables;  
Empty set (0.00 sec)  
  
mysql> CREATE TABLE EMPLOYEE(  
-> EMP_ID int NOT NULL,  
-> EMP_LASTNAME varchar(100),  
-> EMP_NAME varchar(100),  
-> EMP_AGE int,  
-> EMP_ROLL varchar(100),  
-> PRIMARY KEY (EMP_ID)  
-> );  
Query OK, 0 rows affected (0.02 sec)  
  
mysql> 
```

Let's view our created table using the “show tables” command and also view it's respective records using the “SELECT * FROM <table>” command.

```
cloudera@quickstart:~  
File Edit View Search Terminal Help  
-> PRIMARY KEY (EMP_ID)  
-> );  
Query OK, 0 rows affected (0.02 sec)  
  
mysql> SHOW TABLES;  
+-----+  
| Tables_in_HR_db |  
+-----+  
| EMPLOYEE        |  
+-----+  
1 row in set (0.01 sec)  
  
mysql> INSERT INTO EMPLOYEE VALUES(1,"QUINTANILLA","ARTURO",22,"DATA SCIENTIST"),(2,"SANCHEZ","PEDRO",22,"DATA SCIENTIST"),(3,"WU","JACK",20,"FULLSTACK DEV"),(4,"PEKIN","NIKITA",22,"ETHICAL HACKER");  
Query OK, 4 rows affected (0.02 sec)  
Records: 4 Duplicates: 0 Warnings: 0  
  
mysql> SELECT * FROM EMPLOYEE;  
+-----+-----+-----+-----+-----+  
| EMP_ID | EMP_LASTNAME | EMP_NAME | EMP_AGE | EMP_ROLL |  
+-----+-----+-----+-----+-----+  
| 1 | QUINTANILLA | ARTURO | 22 | DATA SCIENTIST |  
| 2 | SANCHEZ | PEDRO | 22 | DATA SCIENTIST |  
| 3 | WU | JACK | 20 | FULLSTACK DEV |  
| 4 | PEKIN | NIKITA | 22 | ETHICAL HACKER |  
+-----+-----+-----+-----+-----+  
4 rows in set (0.00 sec)  
  
mysql> █
```

Now let's make a sqoop evaluation using the special key word "eval" with it's respective configuration and query which in this case it's going to be a "SELECT * FROM <table>".

```
cloudera@quickstart:~  
File Edit View Search Terminal Help  
[cloudera@quickstart ~]$ sqoop eval --connect "jdbc:mysql://quickstart.cloudera:3306/HR_db" --username=root --password=cloudera --query="SELECT * FROM EMPLOYEE"  
Warning: /usr/lib/sqoop/./accumulo does not exist! Accumulo imports will fail.  
Please set $ACCUMULO_HOME to the root of your Accumulo installation.  
18/05/06 14:55:36 INFO sqoop.Sqoop: Running Sqoop version: 1.4.6-cdh5.13.0  
18/05/06 14:55:36 WARN tool.BaseSqoopTool: Setting your password on the command-line is insecure. Consider using -P instead.  
18/05/06 14:55:36 INFO manager.MySQLManager: Preparing to use a MySQL streaming results et.  
-----  
| EMP_ID | EMP_LASTNAME | EMP_NAME | EMP_AGE | EMP_ROLL |  
-----  
| 1 | QUINTANILLA | ARTURO | 22 | DATA SCIENTIST |  
| 2 | SANCHEZ | PEDRO | 22 | DATA SCIENTIST |  
| 3 | WU | JACK | 20 | FULLSTACK DEVELOPER |  
| 4 | PEKIN | NIKITA | 22 | ETHICAL HACKER |  
-----  
[cloudera@quickstart ~]$
```

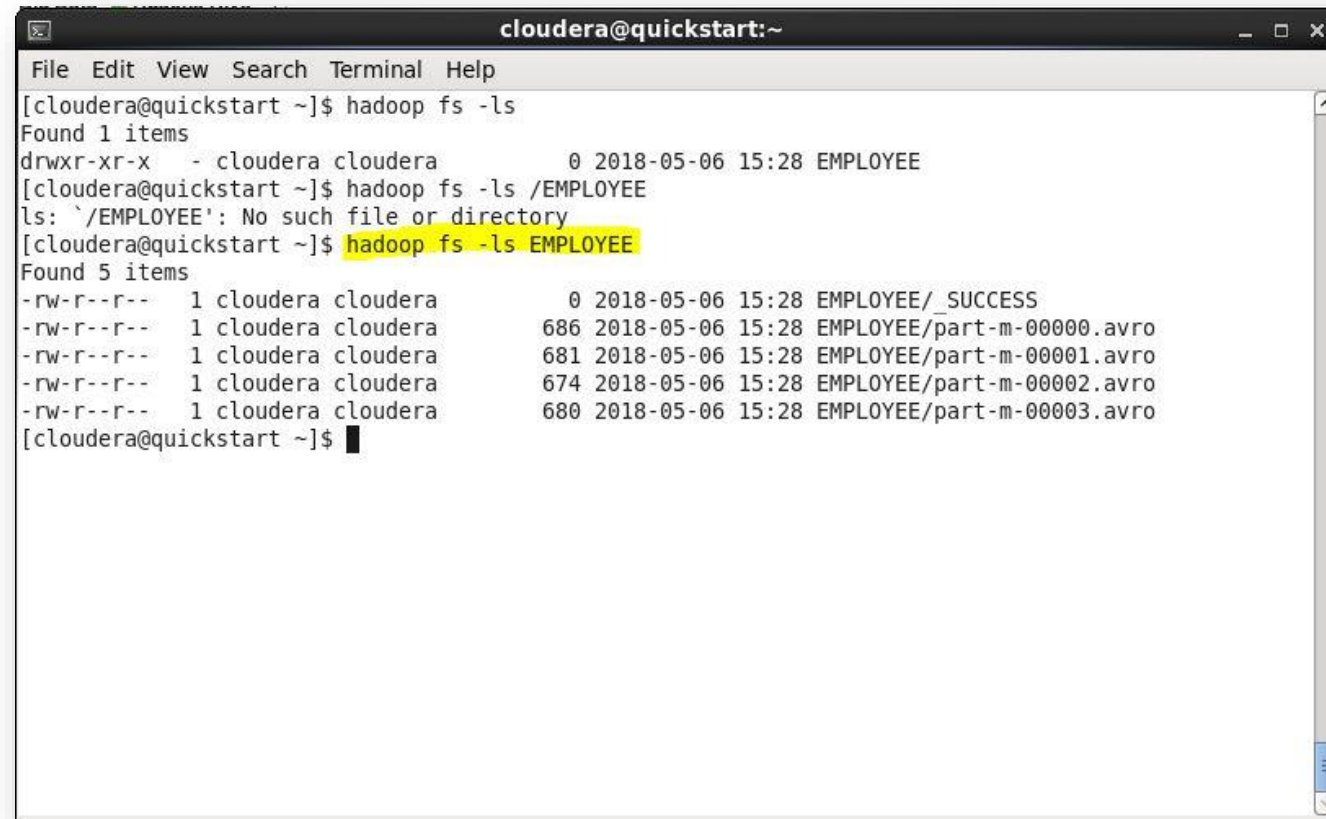
Now let's import our recently created table "EMPLOYEE" as an avro data file using 8 mappers.

```
cloudera@quickstart:~  
File Edit View Search Terminal Help  
[cloudera@quickstart ~]$ sqoop import --connect "jdbc:mysql://quickstart.cloudera:3306/HR_db" --use^  
rname=root --password=cloudera --table EMPLOYEE --m=8 --as-avrodatafile  
Warning: /usr/lib/sqoop/./accumulo does not exist! Accumulo imports will fail.  
Please set $ACCUMULO_HOME to the root of your Accumulo installation.  
18/05/06 15:26:51 INFO sqoop.Sqoop: Running Sqoop version: 1.4.6-cdh5.13.0  
18/05/06 15:26:51 WARN tool.BaseSqoopTool: Setting your password on the command-line is insecure. C  
onsider using -P instead.  
18/05/06 15:26:52 INFO manager.MySQLManager: Preparing to use a MySQL streaming resultset.  
18/05/06 15:26:52 INFO tool.CodeGenTool: Beginning code generation  
18/05/06 15:26:53 INFO manager.SqlManager: Executing SQL statement: SELECT t.* FROM `EMPLOYEE` AS t  
LIMIT 1  
18/05/06 15:26:53 INFO manager.SqlManager: Executing SQL statement: SELECT t.* FROM `EMPLOYEE` AS t  
LIMIT 1  
18/05/06 15:26:53 INFO orm.CompilationManager: HADOOP_MAPRED_HOME is /usr/lib/hadoop-mapreduce  
Note: /tmp/sqoop-cloudera/compile/8b7b1734ddd3900d24bdcd44201b72b2/EMPLOYEE.java uses or overrides  
a deprecated API.  
Note: Recompile with -Xlint:deprecation for details.  
18/05/06 15:27:00 INFO orm.CompilationManager: Writing jar file: /tmp/sqoop-cloudera/compile/8b7b17  
34ddd3900d24bdcd44201b72b2/EMPLOYEE.jar  
18/05/06 15:27:00 WARN manager.MySQLManager: It looks like you are importing from mysql.  
18/05/06 15:27:00 WARN manager.MySQLManager: This transfer can be faster! Use the --direct  
18/05/06 15:27:00 WARN manager.MySQLManager: option to exercise a MySQL-specific fast path.  
18/05/06 15:27:00 INFO manager.MySQLManager: Setting zero DATETIME behavior to convertToNull (mysql  
)  
18/05/06 15:27:00 INFO mapreduce.ImportJobBase: Beginning import of EMPLOYEE  
18/05/06 15:27:00 INFO Configuration.deprecation: mapred.job.tracker is deprecated. Instead, use ma
```


The importing was successful!

```
cloudera@quickstart:~  
File Edit View Search Terminal Help  
Other local map tasks=4  
Total time spent by all maps in occupied slots (ms)=143677  
Total time spent by all reduces in occupied slots (ms)=0  
Total time spent by all map tasks (ms)=143677  
Total vcore-milliseconds taken by all map tasks=143677  
Total megabyte-milliseconds taken by all map tasks=147125248  
Map-Reduce Framework  
Map input records=4  
Map output records=4  
Input split bytes=425  
Spilled Records=0  
Failed Shuffles=0  
Merged Map outputs=0  
GC time elapsed (ms)=2286  
CPU time spent (ms)=6220  
Physical memory (bytes) snapshot=838471680  
Virtual memory (bytes) snapshot=6277976064  
Total committed heap usage (bytes)=664797184  
File Input Format Counters  
Bytes Read=0  
File Output Format Counters  
Bytes Written=2721  
18/05/06 15:28:18 INFO mapreduce.ImportJobBase: Transferred 2.6572 KB in 72.5817 seconds (37.4888 b  
ytes/sec)  
18/05/06 15:28:18 INFO mapreduce.ImportJobBase: Retrieved 4 records.  
[cloudera@quickstart ~]$
```

Finally let's view our imported mysql table in HDFS.



A terminal window titled 'cloudera@quickstart:~' showing the following commands and output:

```
cloudera@quickstart:~$ hadoop fs -ls
Found 1 items
drwxr-xr-x  - cloudera cloudera          0 2018-05-06 15:28 EMPLOYEE
cloudera@quickstart ~]$ hadoop fs -ls /EMPLOYEE
ls: `/EMPLOYEE': No such file or directory
cloudera@quickstart ~]$ hadoop fs -ls EMPLOYEE
Found 5 items
-rw-r--r--  1 cloudera cloudera          0 2018-05-06 15:28 EMPLOYEE/_SUCCESS
-rw-r--r--  1 cloudera cloudera    686 2018-05-06 15:28 EMPLOYEE/part-m-00000.avro
-rw-r--r--  1 cloudera cloudera    681 2018-05-06 15:28 EMPLOYEE/part-m-00001.avro
-rw-r--r--  1 cloudera cloudera    674 2018-05-06 15:28 EMPLOYEE/part-m-00002.avro
-rw-r--r--  1 cloudera cloudera    680 2018-05-06 15:28 EMPLOYEE/part-m-00003.avro
cloudera@quickstart ~]$
```

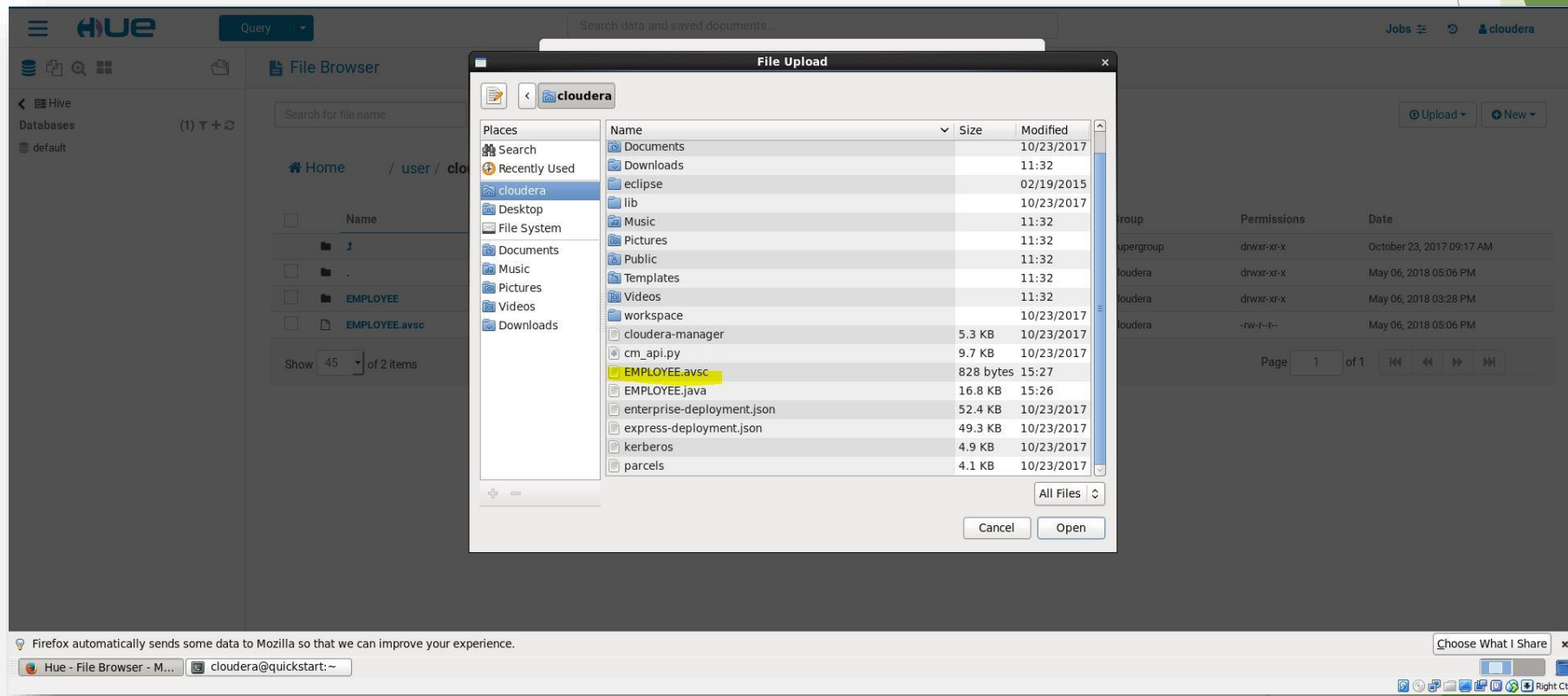
Procedure: Part 3

import avro data to hive

Once our table was imported from mysql to HDFS into avro format now we need to move it into our HDFS working directory.

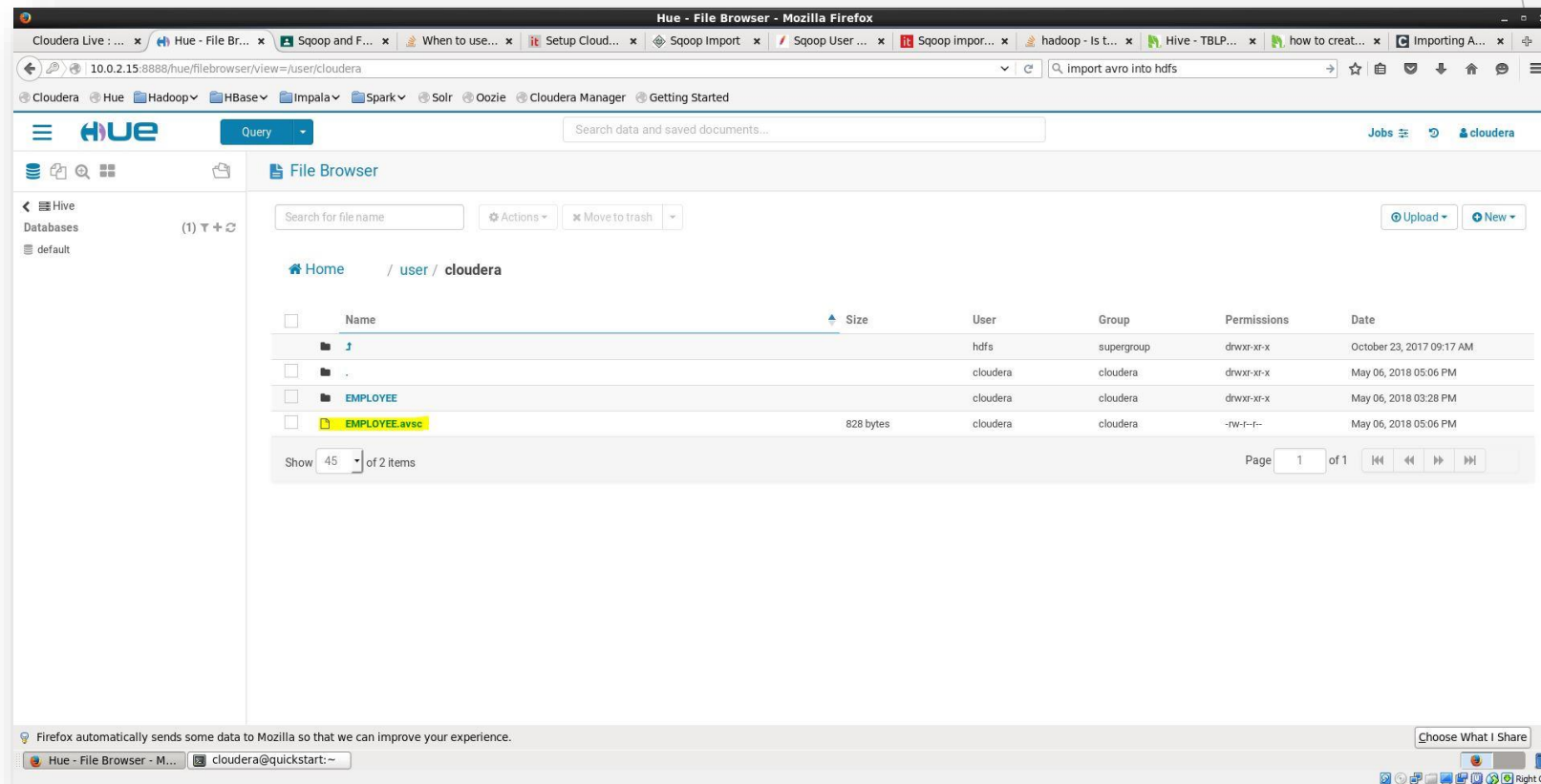
We can do the following by opening HUE in our cloudera browser and clicking the navbar and selecting files option. Finally click on the upload button to upload your file into HDFS.

This is done because sqoop imported our table data into HDFS but we need the avsc schema file which was saved on the home folder in cloudera file system.



Now we have access to our EMPLOYEE schema which will be used to import our avro data files to hive.

NOTE: This is done because the flag “--hive-import” is not compatible with avro or sequence data files.



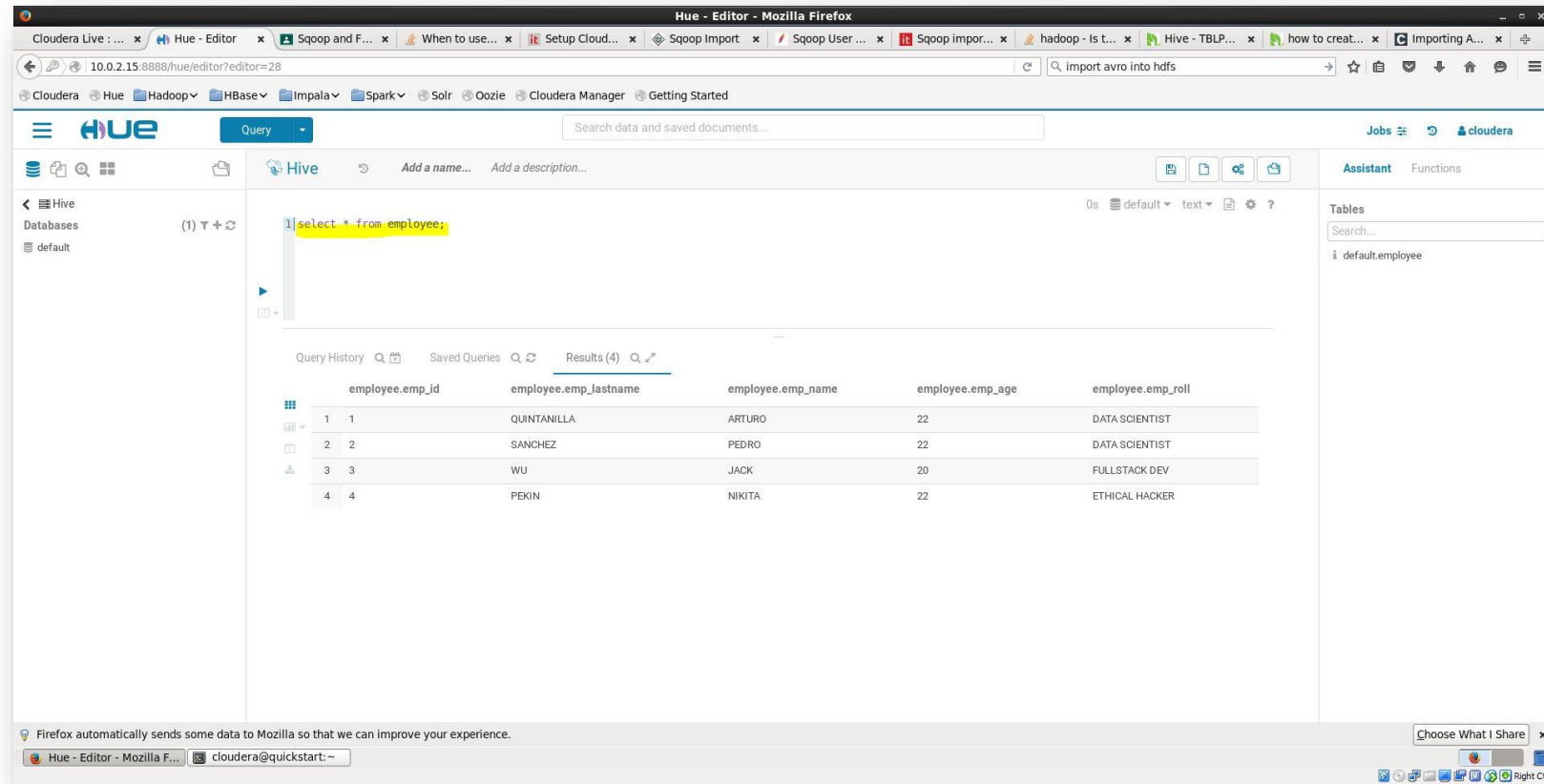
Now we execute the following query to create a new employee table where we are going to reference our data in hive.

The screenshot displays the Hue web interface for Cloudera Live. The browser tabs at the top include 'Cloudera Live', 'Hue - Editor', 'Sqoop and F...', 'When to use...', 'Setup Cloud...', 'Sqoop Import', 'Sqoop User', 'Sqoop impor...', 'hadoop - Is t...', 'Hive - TBLP...', 'how to creat...', and 'Importing A...'. The address bar shows '10.0.2.15:8888/hue/editor?editor=27'. The interface features a top navigation bar with 'Cloudera', 'Hue', 'Hadoop', 'HBase', 'Impala', 'Spark', 'Solr', 'Oozie', 'Cloudera Manager', and 'Getting Started'. A search bar is present with the text 'Search data and saved documents...'. The main workspace is divided into three sections: a left sidebar for 'Hive' databases, a central query editor, and a right sidebar for 'Assistant' and 'Functions'. The query editor contains the following SQL query:

```
1 CREATE TABLE EMPLOYEE
2 STORED AS AVRO
3 LOCATION "/user/cloudera/EMPLOYEE"
4 TBLPROPERTIES ('avro.schema.url'="/user/cloudera/EMPLOYEE.avsc");
```

A green 'Success' message is displayed below the query. The 'Query History' section shows a list of recent queries, with the most recent one being the successful execution of the above query. The bottom of the screen shows a Firefox notification bar and a taskbar with icons for 'Hue - Editor - Mozilla F...', 'cloudera@quickstart:~', and system icons.

Finally let's retrieve data from our recently created hive table.



The screenshot shows the Hue Editor interface in a Mozilla Firefox browser. The query editor at the top contains the SQL statement `select * from employee;`. Below the editor, the results are displayed in a table with 5 columns: `employee.emp_id`, `employee.lastname`, `employee.emp_name`, `employee.emp_age`, and `employee.emp_roll`. The table contains 4 rows of data. On the right side, the 'Tables' panel shows a search bar and a list containing `default.employee`.

	employee.emp_id	employee.lastname	employee.emp_name	employee.emp_age	employee.emp_roll
1	1	QUINTANILLA	ARTURO	22	DATA SCIENTIST
2	2	SANCHEZ	PEDRO	22	DATA SCIENTIST
3	3	WU	JACK	20	FULLSTACK DEV
4	4	PEKIN	NIKITA	22	ETHICAL HACKER

References

- ▶ https://www.cloudera.com/documentation/enterprise/5-4-x/topics/cdh_ig_sqoop.html
- ▶ <https://community.hortonworks.com/questions/32385/how-to-create-and-store-the-avro-files-in-hive-tab.html>
- ▶ <https://community.hortonworks.com/questions/15868/hive-tblproperties.html>
- ▶ https://stackoverflow.com/questions/21539694/is-there-an-equivalent-to-pwd-in-hdfs?utm_medium=organic&utm_source=google_rich_qa&utm_campaign=google_rich_qa
- ▶ <http://discuss.itversity.com/t/sqoop-import-into-hive-as-avro-datafile/7146/2>
- ▶ https://sqoop.apache.org/docs/1.4.2/SqoopUserGuide.html#_importing_data_into_hive
- ▶ https://www.tutorialspoint.com/sqoop/sqoop_import.htm
- ▶ <http://www.itversity.com/topic/setup-cloudera-quickstart-vm/>
- ▶ <https://stackoverflow.com/questions/31515498/when-to-use-sqoop-create-hive-table>