databricks
Academy

DAWD 02-3 - Lab - Ingesting Data

The two fields below are used to customize queries used in this course. Enter your schema (database) name and username, and press "Enter" to populate necessary information in the queries on this page.

Schema Name:	hive_metastore.class_013_odg7_da_dawd
Username:	class+013@databricks.com

Lesson Objective

At the end of this lesson, you will be able to:

- Use Databricks SQL to ingest data

Ingest Data in Databricks SQL

In the previous lesson, we ingested data from a .csv file in two steps to create a Delta table. In this lab, we are going to run a single query, with a handful of statements, to complete the same process.

Databricks SQL will count the number of rows returned from a table automatically. However by default, only 1000 rows are returned. To have all rows returned from a table, deselect the checkbox `LIMIT 1000` just below the query editor.

Complete the following:

1. Copy the query below into the query editor
2. Run the query in Databricks SQL
3. Check your work by entering your answer to the question
4. After pressing `ENTER/RETURN`, green indicates a correct answer, and red indicates incorrect

```
USE hive_metastore.class_013_odg7_da_dawd;
DROP TABLE IF EXISTS customers_csv;
CREATE TABLE customers_csv
    USING csv
    OPTIONS (
        path='wasb://courseware@dbacademy.blob.core.windows.net/data-analysis-with-
databricks/v02/retail-org/customers/customers.csv',
        header="true",
        inferSchema="true"
);
CREATE OR REPLACE TABLE customers AS
    SELECT * FROM customers_csv;
DROP TABLE customers_csv;
SELECT * FROM customers;
```

Show Answer Copy

How many rows are in the `customers` table? (type only numbers)

Ingest Data Using COPY INTO

COPY INTO is often used to ingest streaming data. Files that have previously been ingested are ignored, and new files are ingested every time COPY INTO is run. We can set this up by configuring COPY INTO to point to a directory and not specifying and patterns or files.

In the query below, we are going to a create an empty Delta table, so we will specify the schema. We will then use COPY INTO to fill the table with data. After making changes to the query, run it more than once, and note that the number of rows does not change. As new .json files are added to the directory, they will be ingested into the table. Note: for this example, there are no new data files being placed in the directory, so the number of rows will not change.

Make changes to the query below so that COPY INTO pulls data from the directory 'wasb://courseware@dbacademy.blob.core.windows.net/data-analysis-with-databricks/v02/retail-org/sales_stream/sales_stream_json' and configure the file format as 'JSON'.

Complete the following:

1. Make the required changes to the query below
2. Run the query in Databricks SQL
3. Check your work by entering your answer to the question
4. After pressing `ENTER/RETURN`, green indicates a correct answer, and red indicates incorrect

```
USE hive_metastore.class_013_odg7_da_dawd;
CREATE OR REPLACE TABLE sales_stream (5minutes STRING,
                                      clicked_items ARRAY<ARRAY<STRING>>,
                                      customer_id STRING,
                                      customer_name STRING,
                                      datetime STRING,
                                      hour BIGINT,
                                      minute BIGINT,
                                      number_of_line_items STRING,
                                      order_datetime STRING,
                                      order_number BIGINT,
                                      ordered_products ARRAY<ARRAY<STRING>>,
                                      sales_person DOUBLE,
                                      ship_to_address STRING
);
COPY INTO sales_stream
    FROM 'FILL_IN'
    FILEFORMAT = FILL_IN;
SELECT * FROM sales_stream ORDER BY customer_id;
```

Show Answer Copy

What is the value of `customer_id` in the first row?

Grant/Revoke Privileges Using SQL

In this portion of the lab, we are going to grant all privileges on our brand new table, `sales_stream`, to all users in the workspace. We are then going to immediately revoke MODIFY from all users. By default, Databricks SQL has two groups: "admins" and "users", but Databricks admins can add more groups, as needed.

Complete the following:

1. Run the query below in Databricks SQL
2. Enter your answer to the question
3. After pressing `ENTER/RETURN`, green indicates a correct answer, and red indicates incorrect

```
USE hive_metastore.class_013_odg7_da_dawd;
GRANT SELECT ON TABLE `sales_stream` TO `users`;
SHOW GRANT ON `sales_stream`;
```

Show Answer Copy

How many privileges does the group "users" have on `sales_stream`?

Grant/Revoke Privileges Using the Data Explorer

We are now going to use the Data Explorer to perform the same tasks as above. Follow the instructions to find the answer to the question below.

Complete the following:

1. Click "Data" in the sidebar menu to go to the Data Explorer
2. Click "Default" to open the drop down, and select your schema
3. Select the table, "`sales_stream`" from the list
4. Select the "permissions" tab
5. Click "Grant"
6. Click inside the input box and select "All Users"
7. Click inside the input box again to dismiss the list
8. Select the checkbox next to "`MODIFY`" in the list of permissions
9. Click "OK"

Note which privileges have now been granted on the table `sales_stream`. To revoke these privileges, complete the following:

10. Select the checkboxes next to all privileges granted to "users"
11. Click "Revoke"
12. Read the warning, and click "Revoke"
13. When you are ready, enter your answer to the question below
14. After pressing `ENTER/RETURN`, green indicates a correct answer, and red indicates incorrect

How many privileges have you been granted to you on your schema?