databricks
Academy

DAWD 03-6 - Lab - Final Lab Assignment

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
- Query the ingested data
- Produce visualizations from the queries
- Create an Alert
- Produce a dashboard from the visualizations

Scenario

Here is the scenario for the final lab assignment:

- You work for a large retail outlet, called Specialty Electronics Outlet

- The outlet sells to customers in the United States
- Data streams from various retail locations on a continuous basis
- Data Engineers are responsible for making this data ready for you to analyze
- Data Scientists use this same data for running machine learning algorithms to make predictions

---

Your Task

Here is what you should accomplish:

- The lab is configured as a set of step-by-step instructions that will guide you through all the requirements
- The outcome of the lab will be a dashboard that is ready to be shared with stakeholders
- You will also configure Alerts to keep stakeholders informed of significant events
- Lastly, you will make Refresh Schedules to ensure all data remains up-to-date

---

Ingest Data

You have been asked to ingest sales data from a new data source. The source is a .json file that contains eight columns. The file contains header information and is located at `wasb://courseware@dbacademy.blob.core.windows.net/data-analysis-with-databricks/v02/retail-org/sales_orders/sales_orders_json`. The final table should be in Delta format (the default) and should be called `sales_orders`. Any intermediate tables should be dropped.

Complete the following:

1. Make the required changes to the query below
2. Run the query in Databricks SQL
3. Double-check that the `sales_orders` table is in Delta format using the Data Explorer
4. Check your work by entering your answer to the question
5. 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 sales_orders_json;
CREATE TABLE sales_orders_json
    USING FILL_IN
    OPTIONS (
        path 'wasb://courseware@dbacademy.blob.core.windows.net/data-analysis-with-
databricks/v02/retail-org/sales_orders/sales_orders_json',
        inferSchema "true"
);
DROP TABLE IF EXISTS sales_orders;
CREATE TABLE sales_orders AS
    SELECT * FROM FILL_IN;
DROP TABLE IF EXISTS sales_orders_json;
SELECT * FROM FILL_IN;
```

Show Answer Copy

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

Query the Ingested Data

Now that we have sales data ingested, we have been asked to find those customers who have spent more than $175,000. Notice in the dataset that column `ordered_products` is an array that contains json data. We need to explode the array in order to access the data inside. We need to access "price" and "qtY" and calculate the total amount spent on these items. We will then sum these amounts and group by the customers who purchased those products. We can do this all in one command by using CTEs.

Complete the following:

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

```
USE hive_metastore.class_013_odg7_da_dawd;
FILL_IN order_info AS
    (SELECT customer_id, explode(ordered_products) AS product_info FROM sales_orders),
total_sales AS
    (SELECT customer_id, sum(product_info["price"] * product_info["qty"]) AS total_spent FROM
order_info GROUP BY customer_id)
SELECT customer_name, total_spent FROM total_sales
    FILL_IN customers
    ON customers.customer_id = total_sales.customer_id
    FILL_IN total_spent > 175000
    FILL_IN BY total_spent DESC;
```

Show Answer Copy

How much has Bradsworth Digital Solutions, Inc spent?

Query Two

What are the sales in each of the loyalty segments? We can answer this question by joining the `sales_orders` table, the `customers` table, and the `loyalty_segments` table and using a CTE, the aggregate function `sum()`, and a column expression that multiplies two columns together to calculate sales.

Complete the following:

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

```
USE hive_metastore.class_013_odg7_da_dawd;
FILL_IN sold_products AS
(SELECT *, explode(ordered_products) as products FROM sales_orders
    FILL_IN customers
        ON customers.customer_id = sales_orders.customer_id
    JOIN loyalty_segments
        ON customers.loyalty_segment = loyalty_segments.loyalty_segment_id)
SELECT loyalty_segment_description, FILL_IN(products['price'] * products['qty']) AS sales FROM
sold_products
    GROUP BY loyalty_segment_description;
```

Show Answer Copy

What is the sales amount for customers in loyalty segment 3?

Produce Visualizations from the Queries

Now, let's produce visualizations for our queries. Let's first tackle the "High Dollar Customers" query. The company has put together a promotion. They want to reward their high dollar customers. The first to reach $250,000 in sales will get 20% off for a year. The second closest will get 15% off, and the third will get 10% off. We want to create two visualizations from this query that will help us see the progress of this marketing strategy. We will first make changes to the table visualization (the default).

Complete the following:

1. Click "Edit Visualization"
2. Click "customer_name" to open the settings for this column
3. Update the column name to "Customer Name" to make it look better
4. For "Description" type "These are our top customers."
5. Check your work by entering your answer to the question
6. Under "Font Conditions" click "Add Condition"
7. Open the first dropdown and select "total_spent"
8. Change "=" to ">", and input the value "200000" in the "Value" field
9. Click the checkerboard color tile, and select red
10. Click "total_spent" to open the settings for this column
11. Update the name from "total_spent" to "Total Spent"
12. Add the same font condition
13. Click "Save" in the lower-right column

The font conditions we set will change all rows that have a "Total Spent" greater than 200000 to red text. This will make them stand out.

14. Hover over the "+" symbol and click "Visualization"
15. Select "Counter" as the visualization type
16. For "Counter Label" type "Highest Sales Total"
17. For "Counter Value Column" make sure the column "total_spent" is selected
18. Click the "Format" tab
19. Optional: Change the decimal character and thousands separator
20. "total_spent" is a dollar figure, so add "$" to "Formatting String Prefix"
21. Click "Save" in the lower-right corner
22. Change the name of the visualization to "Highest Sales Total"

Let's move to the "Sales in Loyalty Segments" query.

23. Hover over the "+" symbol and click "Visualization"
24. Select "Bar" as the visualization type
25. For "X Column" select "loyalty_segment_description"
26. For "Y Column" add "sales"
27. Change the names to something that looks nicer on the "X Axis" and "Y Axis" tabs
28. Click "Save" in the lower-right corner
29. Change the name of the visualization to "Sales in Loyalty Segments"
30. Check your work by entering your answer to the question
31. After pressing `ENTER/RETURN`, green indicates a correct answer, and red indicates incorrect

What is the name of the right-most tab in the Bar visualization editor?

Create an Alert

Let's create an Alert with our "High Dollar Customers" query that will let us know when the threshold of $250,000 has been met.

Complete the following:

1. Click "Alerts" in the sidebar menu
2. Click "Create Alert"
3. From the Query dropdown, select the query: "High Dollar Customers"
4. Use the dropdown to change the "Value" column to `total_spent` and change "Threshold" to 250000

5. Change "Refresh" to "Every 1 day"
6. Check your work by entering your answer to the question
7. After pressing `ENTER/RETURN`, green indicates a correct answer, and red indicates incorrect

Will this Alert trigger even though there is no refresh schedule configured for the query itself?

Produce a Dashboard from the Visualizations

Now that we have configured a refresh schedule, let's set up an Alert that will notify us when the income generated from sales increases beyond a threshold.

Complete the following:

1. Click "Dashboards" in the sidebar menu
2. Click "Create Dashboard"
3. Name the dashboard "Loyalty Tracking"
4. Click "Add Visualization"
5. Click "High Dollar Customers"
6. Ensure "Table" is selected in "Choose Visualization"
7. Change "Title" to "High Dollar Customers"
8. Optional: write a description
9. Repeat steps 4-8 with the counter visualization we also added to "High Dollar Customers" and the bar chart in "Sales in Loyalty Segments"
10. Click "Add Textbox" and type "# Loyalty Segment Tracking"
11. Optional: Move the visualizations around by clicking and dragging each one
12. Optional: Resize each visualization by dragging the lower-right corner of the visualization
13. Optional: Click "Colors" to change the color palette used by visualizations in the dashboard
14. Click "Done Editing" in the upper-right corner
15. Run every query and refresh all visualizations all at once by clicking "Refresh"
16. Check your work by entering your answer to the question
17. After pressing `ENTER/RETURN`, green indicates a correct answer, and red indicates incorrect

If you wanted to configure a set interval to refresh this dashboard automatically, what button would you click?

1.2.13