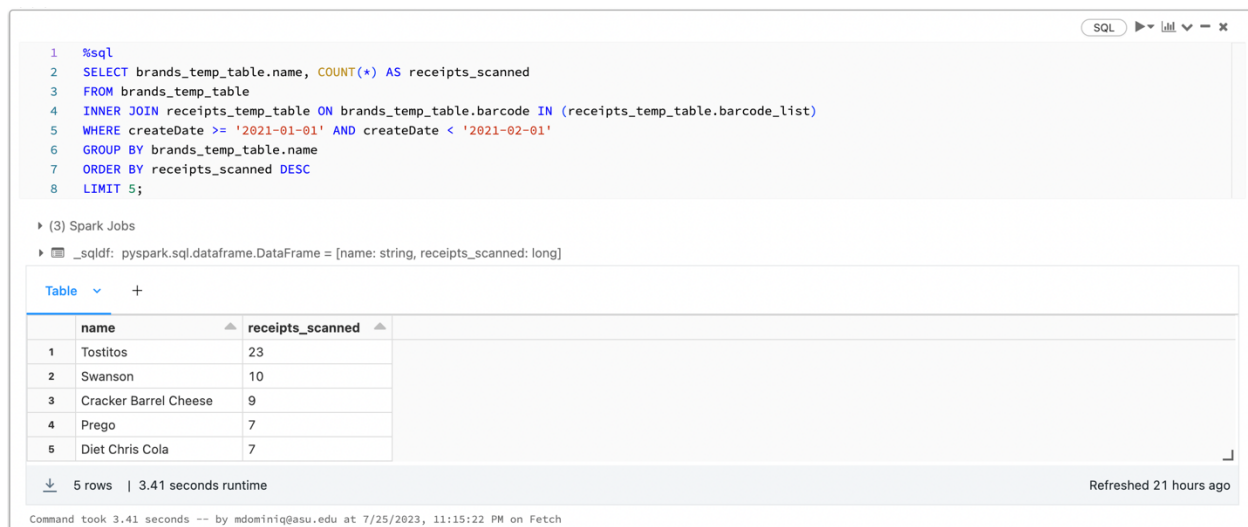


SQL Queries against the new structured Relational Data Model

- What are the top 5 brands by receipts scanned for most recent month?

```
SELECT brands_temp_table.name, COUNT(*) AS receipts_scanned
FROM brands_temp_table
INNER JOIN receipts_temp_table ON brands_temp_table.barcode IN
(receipts_temp_table.barcode_list)
WHERE createDate >= '2021-01-01' AND createDate < '2021-02-01'
GROUP BY brands_temp_table.name
ORDER BY receipts_scanned DESC
LIMIT 5;
```



The screenshot shows a Databricks SQL interface. At the top, a SQL query is entered in a text area. Below the query, the execution status is shown as "(3) Spark Jobs". A table icon indicates the query result is a table. Below this, a table view is displayed with 5 rows and 2 columns: "name" and "receipts_scanned". The table data is as follows:

	name	receipts_scanned
1	Tostitos	23
2	Swanson	10
3	Cracker Barrel Cheese	9
4	Prego	7
5	Diet Chris Cola	7

At the bottom of the table view, it says "5 rows | 3.41 seconds runtime". To the right, it says "Refreshed 21 hours ago". At the very bottom, a command line shows the command took 3.41 seconds and was executed by mdominiq@asu.edu at 7/25/2023, 11:15:22 PM on Fetch.

- When considering *average spend* from receipts with 'rewardsReceiptStatus' of 'Accepted' or 'Rejected', which is greater?

```
SELECT rewardsReceiptStatus, AVG(totalSpent) AS average_spend
FROM receipts_temp_table
WHERE rewardsReceiptStatus IN ('ACCEPTED', 'REJECTED')
GROUP BY rewardsReceiptStatus;
```

```

1 %sql
2 SELECT rewardsReceiptStatus, AVG(totalSpent) AS average_spend
3 FROM receipts_temp_table
4 WHERE rewardsReceiptStatus IN ('ACCEPTED', 'REJECTED')
5 GROUP BY rewardsReceiptStatus;

```

▶ (2) Spark Jobs

▶ _sqldf: pyspark.sql.dataframe.DataFrame = [rewardsReceiptStatus: string, average_spend: double]

	rewardsReceiptStatus	average_spend
1	REJECTED	30.043814432989713

1 row | 1.64 seconds runtime

Command took 1.64 seconds -- by mdominiq@asu.edu at 7/25/2023, 11:15:22 PM on Fetch

- When considering total number of items purchased from receipts with 'rewardsReceiptStatus' of 'Accepted' or 'Rejected', which is greater?

```

SELECT rewardsReceiptStatus, SUM(purchasedItemCount) AS
total_items_purchased
FROM receipts_temp_table
WHERE rewardsReceiptStatus IN ('ACCEPTED', 'REJECTED')
GROUP BY rewardsReceiptStatus;

```

```

1 %sql
2 SELECT rewardsReceiptStatus, SUM(purchasedItemCount) AS total_items_purchased
3 FROM receipts_temp_table
4 WHERE rewardsReceiptStatus IN ('ACCEPTED', 'REJECTED')
5 GROUP BY rewardsReceiptStatus;

```

▶ (2) Spark Jobs

▶ _sqldf: pyspark.sql.dataframe.DataFrame = [rewardsReceiptStatus: string, total_items_purchased: double]

	rewardsReceiptStatus	total_items_purchased
1	REJECTED	223

1 row | 1.28 seconds runtime

Command took 1.28 seconds -- by mdominiq@asu.edu at 7/25/2023, 11:15:22 PM on Fetch

- Which brand has the most *spend* among users who were created within the past 6 months?

```

SELECT b.name AS brand_name, SUM(r.totalSpent) AS total_spend
FROM brands_temp_table b
JOIN receipts_temp_table r ON b.barcode = r.barcode_list
JOIN users_temp_table u ON r.userId = u._id

```

```

WHERE u.createdDate >= DATEADD(month, -6, '2021-03-
01T21:32:28.000+0000')
GROUP BY b.name
ORDER BY total_spend DESC
LIMIT 1;

```

SQL

```

1 %sql
2 SELECT b.name AS brand_name, SUM(r.totalSpent) AS total_spend
3 FROM brands_temp_table b
4 JOIN receipts_temp_table r ON b.barcode = r.barcode_list
5 JOIN users_temp_table u ON r.userId = u._id
6 WHERE u.createdDate >= DATEADD(month, -6, '2021-03-01T21:32:28.000+0000')
7 GROUP BY b.name
8 ORDER BY total_spend DESC
9 LIMIT 1;

```

(4) Spark Jobs

_sqlidf: pyspark.sql.dataframe.DataFrame

brand_name: string

total_spend: double

Table

	brand_name	total_spend
1	Tostitos	23812.969999999998

1 row

3.08 seconds runtime

Refreshed 21 hours ago

Command took 3.08 seconds -- by mdominiq@asu.edu at 7/25/2023, 11:19:47 PM on Fetch

- Which brand has the most *transactions* among users who were created within the past 6 months?

```

SELECT b.name AS brand_name, COUNT(r.userId) AS transaction_count
FROM brands_temp_table b
LEFT JOIN receipts_temp_table r ON b.barcode = r.barcode_list
LEFT JOIN users_temp_table u ON r.userId = u._id
WHERE u.createdDate >= DATEADD(month, -6, '2021-03-
01T21:32:28.000+0000') OR u.createdDate IS NULL
GROUP BY b.name
ORDER BY transaction_count DESC
LIMIT 1;

```

SQL ▶ ⌵ ⌵ ⌵ ⌵ ⌵

```
1 %sql
2 SELECT b.name AS brand_name, COUNT(r.userId) AS transaction_count
3 FROM brands_temp_table b
4 LEFT JOIN receipts_temp_table r ON b.barcode = r.barcode_list
5 LEFT JOIN users_temp_table u ON r.userId = u._id
6 WHERE u.createdDate >= DATEADD(month, -6, '2021-03-01T21:32:28.000+0000') OR u.createdDate IS NULL
7 GROUP BY b.name
8 ORDER BY transaction_count DESC
9 LIMIT 1;
```

▶ (4) Spark Jobs

▶  _sqlidf: pyspark.sql.dataframe.DataFrame = [brand_name: string, transaction_count: long]

Table ⌵ +		
	brand_name ⬆	transaction_count ⬆
1	Tostitos	43

⬇ 1 row | 2.33 seconds runtime Refreshed 21 hours ago

Command took 2.33 seconds -- by mdominiq@asu.edu at 7/25/2023, 11:21:05 PM on Fetch