Team 8: Michelle Xie, Elbek Gaybullaev​, Theophilus Ghartey​, Jingsi Hu​

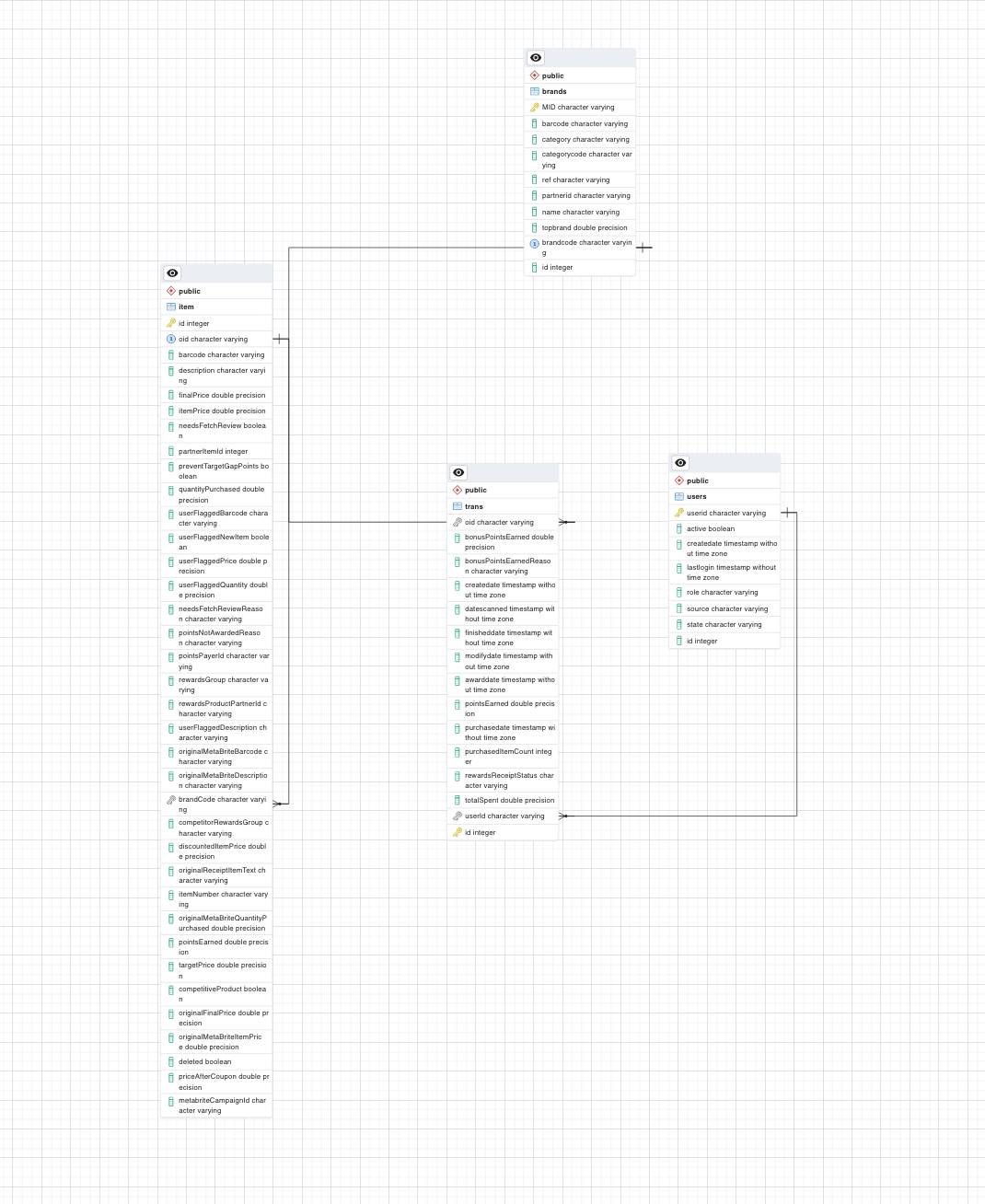
Dr. Fogel

BAN5573

30 April 2023

**Interview Case**

**Entity Relationship Diagram (ERD)**

****

**\*Note:** An “id” column was added in each of the tables a temporary primary key as we worked through the data to identify rows, possible duplications, and actual primary keys in the data.

**Business Questions**

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

SELECT b."NAME", COUNT(\*) AS num\_scans

FROM Trans t

JOIN Item i ON t."OID" = i."OID"

LEFT JOIN brands b ON i.barcode = b.barcode

WHERE t."DATESCANNED"::timestamp with time zone >= DATE\_TRUNC('month', CURRENT\_DATE)::timestamp with time zone

GROUP BY b."NAME"

ORDER BY num\_scans DESC

LIMIT 5;

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

SELECT b."NAME", COUNT(DISTINCT t."userId") AS num\_users

FROM Trans t

JOIN Item i ON t."OID" = i."OID"

JOIN brands b ON i.barcode = b.barcode

JOIN user u ON t."userId" = u."USERID"

WHERE u."CREATEDATE" >= DATE\_TRUNC('month', CURRENT\_DATE - INTERVAL '6 months')

GROUP BY b."NAME"

ORDER BY num\_users DESC

LIMIT 1;

1. When considering average spend from receipts with rewardsReceiptStatus of Accepted or Rejected, which is greater?

SELECT AVG("totalSpent") AS avg\_total\_spent\_finished, 'FINISHED' AS rewards\_receipt\_status

FROM trans

WHERE "rewardsReceiptStatus" = 'FINISHED'

UNION

SELECT AVG("totalSpent") AS avg\_total\_spent\_rejected, 'REJECTED' AS rewards\_receipt\_status

FROM trans

WHERE "rewardsReceiptStatus" = 'REJECTED';

1. When considering total number of items purchased from receipts with rewardReceiptsStatus of Accepted or Rejected, which is greater ?

SELECT COUNT("purchasedItemCount") AS total\_spent\_finished, 'FINISHED' AS rewards\_receipt\_status

FROM trans

WHERE "rewardsReceiptStatus" = 'FINISHED'

UNION

SELECT COUNT("purchasedItemCount") AS total\_spent\_rejected, 'REJECTED' AS rewards\_receipt\_status

FROM trans

WHERE "rewardsReceiptStatus" = 'REJECTED';

**Data Quality Issue (at least 1)**

Assuming we test the data quality issues of Users Data Schema were included with columns: "\_id," "state," "created date," "lastLogin", "role," and "active," one possible data quality issue could be missing or null values in the "lastLogin" column.

To identify this issue, we can use the following SQL query:

SELECT COUNT(\*) FROM users WHERE lastLogin IS NULL;

This query will count the rows in the "users" table where the "last login" column is null. If the result is more significant than zero, then we have missing values in the "lastLogin" column, which could indicate a data quality issue. Another possible issue could be inconsistent or invalid values in the "state" column. To identify this issue, we can use the following SQL query:

SELECT DISTINCT state FROM users;

This query will return a list of all unique values in the "state" column. If there are any unexpected values or inconsistencies in the list, we may have a data quality issue with the "state" column. For example, if we see deals like "NY", "New York", and "New York" in the list, then we may have inconsistent formatting or spelling of state names.

select \* from public.users u order by u.userid;

delete from public.users a using public.users b where a.userid = b.userid and a.lastlogin < b.lastlogin;

🡪 keeps the row with the most recent lastlogin value

delete from public.users a using public.users b where a.userid = b.userid and a.id > b.id;

🡪 keeps the row with the older ID

🡪 then you can make userid a primary key after cleanup

The two queries above were used. From the first query, we noticed that the user id was not unique. There were number rows with the same user id and information. To resolve this situation, the second query was used to delete the duplicates. Doing so, allows the user id to be unique and keeps the one with the most recent last login (assuming that it has the most up to date data).

**Communication with Stakeholder Questions**

1. What questions do you have about the data?

After working with the data, we have the following questions about the data:

* How was the data collected? What is the source of the data?
* What is the intended purpose of the data?
* Are their certain limitations that should be taken into consideration?
* Is there a reason that certain information is missing?

1. How did you discover the data quality issues?

**Data profiling**: involves analyzing the data to identify its structure, patterns, and quality characteristics. It could include reviewing data statistics such as min, max, and average values, checking data distributions, and analyzing missing or outliers.

**Data visualization:** Data visualization tools to create charts, graphs, and dashboards that make it easier to spot data quality issues. For example, to identify any correlation or trends between data points to identify data outliers.

**Data sampling:** Analyzing a subset of the data can be a valuable way to identify data quality issues quickly. By selecting a random sample of the data and analyzing it, data analysts can identify any inconsistencies or problems that may be present in the larger dataset.

**Data profiling tools:** There are various data profiling tools to automate identifying data quality issues. These tools can analyze data across multiple dimensions, such as completeness, consistency, accuracy, and highlight areas.

1. What do you need you need to know to resolve the data quality issues? **--**

Understanding the context and characteristics of the analyzed data is essential. This may require additional information about the data sources, collection processes, transformations, and usage.

To identify the root causes of data quality issues, some areas of inquiry that may be helpful include:

1. Data collection: How was the data collected? Were the data sources reliable and consistent? Were there any gaps or errors in the data collection process?
2. Data transformation: Were any manipulations applied to the data before it was stored or analyzed? Were these transformations performed correctly and consistently? Were there any data losses or data type conversions that could have affected data quality?
3. Data validation and verification: Were quality control measures in place to ensure data accuracy and consistency? Were any data validation or verification checks performed on the data? Were there any data outliers or anomalies that could have affected data quality?
4. Data governance: Was a data governance framework in place to manage and monitor data quality? Were there any data governance policies or procedures that could have affected data quality?

By understanding these factors, we are taking steps to address the root causes of data quality issues and implement solutions to improve the accuracy and reliability of the data. Possible solutions include:

* Modifying the data collection process.
* Improving data validation and verification procedures.
* Implementing data quality checks and alerts.
* Improving data governance policies and procedures.

1. What other information would you need to help you optimize the data assets you’re trying to create?

Context to what the columns labels represent. This includes knowing what OID represents, whether each transaction can consist of multiple items or if it is one to one in the data. This type of information would help better connect the 4 tables with one another and work on answering the business questions.

1. What performance and scaling concerns do you anticipate in production and how do you plan to address them?

There are no indicators of what the ids or columns represent, and which are unique. Therefore, making it hard to create an index. To address this type of issue we would need to better understand how the data is collect and meaning of what is being connect. From there the data can be cleaned up and processed to provide a more accurate insight.