Fetch Rewards Take Home Test

Ronald Daley

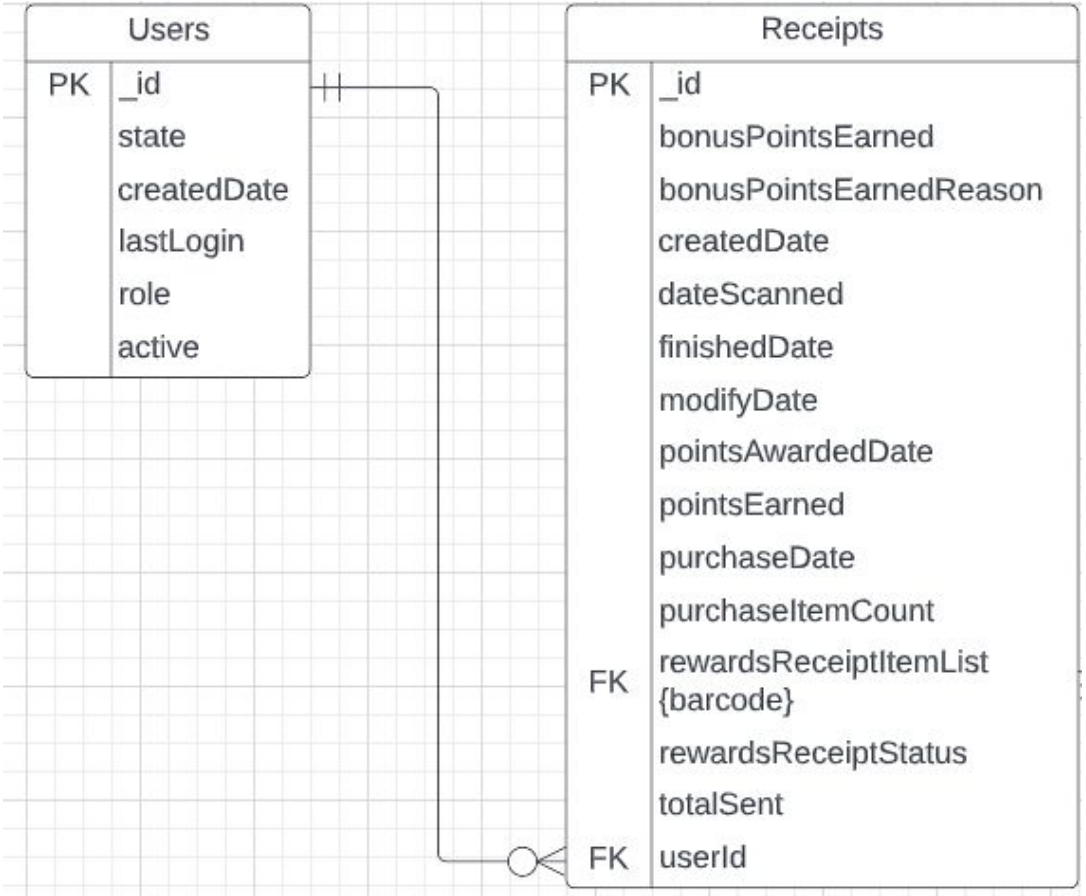
Requirement #1

Review Existing Unstructured Data and Diagram a New Structured Relational Data Model

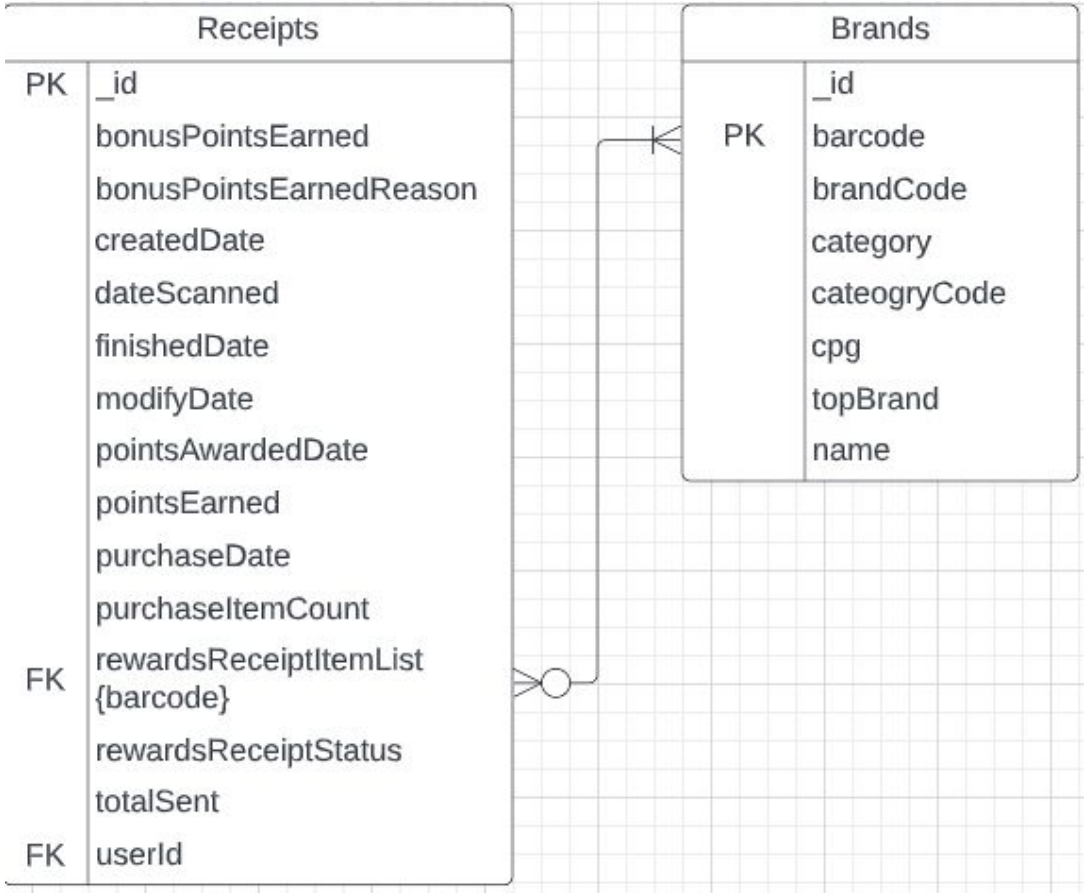
Review the 3 sample data files provided below. Develop a simplified, structured, relational diagram to represent how you would model the data in a data warehouse. The diagram should show each table's fields and the joinable keys. You can use pencil and paper, readme, or any digital drawing or diagramming tool with which you are familiar. If you can upload the text, image, or diagram into a git repository and we can read it, we will review it!

Final Take Schema

- **User - Receipt Relationship**
  - **Left to Right:**
    - What is the minimum number of receipts a user can have?
      - A user can exist but have 0 receipts
    - What is the maximum # of receipts a user can have?
      - A user can have infinite receipts.
    - **Conclusion: A user can have 0 to many receipts.**
  - **Right to Left:**
    - What is the minimum number of users a receipt can have?
      - 1
    - What is the maximum # of users a receipt can have?
      - 1
    - **Conclusion: Each receipt has one and only one customer that receives points.**



- **Receipt - Brand Relationship**
  - **Left to Right:**
    - What is the minimum number of brands a receipt can have?
      - 0
    - What is the maximum # of brands a receipt can have?
      - many
    - **Conclusion: A receipt can have none to many brands appear on it.**
  - **Right to Left:**
    - A Brand can be apart of how many receipts?
    - What is the minimum number of receipts a brand can appear on to receive point?
      - 1
    - What is the maximum # of receipts a brand can appear on?
      - many
    - **To receive points, A brand can appear on one to many receipts**



Assumptions

1. Each item in the brands data has a unique barcode number

Questions

2. Is there additional brands data?
    a. The provided only has 1167 brands in data and not all the brands appear in the receipts data.
3. What happens if the receipt does not have items (0 items) and thus no barcode?
    a. How are restaurants handled?
    b. Is there a special database?

Requirement #2

Write a query that directly answers a predetermined question from a business stakeholder

Write a SQL query against your new structured relational data model that answers one of the following bullet points below of your choosing. Commit it to the git repository along with the rest of the exercise.

Note: When creating your data model be mindful of the other requests being made by the business stakeholder. If you can capture more than one bullet point in your model while keeping it clean, efficient, and performant, that benefits you as well as your team.

- What are the top 5 brands by receipts scanned for most recent month?
- How does the ranking of the top 5 brands by receipts scanned for the recent month compare to the ranking for the previous month?
- **When considering *average spend* from receipts with 'rewardsReceiptStatus' of 'Accepted' or 'Rejected', which is greater?**
- **When considering *total number of items purchased* from receipts with 'rewardsReceiptStatus' of 'Accepted' or 'Rejected', which is greater?**
- Which brand has the most *spend* among users who were created within the past 6 months?
- Which brand has the most *transactions* among users who were created within the past 6 months?

SQL QUERY

```
SELECT receipts.rewardsReceiptStatus, sum (receipts.purchasedItemCount) as cnt_items,
        round(avg(receipts.totalSpent) as avg_spend,2)
FROM receipts
GROUP BY receipts.rewardsReceiptStatus
HAVING receipts.rewardsReceiptStatus = ‘FINISHED’
        OR receipts.rewardsReceiptStatus = ‘REJECTED’ ;
```

Requirement #3

Evaluate Data Quality Issues in the Data Provided

Using the programming language of your choice (SQL, Python, R, Bash, etc...) identify at least one data quality issue. We are not expecting a full blown review of all the data provided, but instead want to know how you explore and evaluate data of questionable provenance.

Commit your code and findings to the git repository along with the rest of the exercise.

Link to GitHub

Python Code

- Github Repo:
  - https://github.com/radaley1906/Fetch---THT

Requirement #4

Fourth: Communicate with Stakeholders

Construct an email or slack message that is understandable to a product or business leader who isn’t familiar with your day to day work. This part of the exercise should show off how you communicate and reason about data with others. Commit your answers to the git repository along with the rest of your exercise.

- What questions do you have about the data?
- How did you discover the data quality issues?
- What do you need to know to resolve the data quality issues?
- What other information would you need to help you optimize the data assets you're trying to create?
- What performance and scaling concerns do you anticipate in production and how do you plan to address them

Email to Data Quality Manager

Hi,

Hope all is well! My name is Ron and I am working with the Marketing team to assist them in building a dashboard that they can use to analyze customer brand activity, for their team to leverage when making data-driven business decisions. We are working together on this project to provide more insight to their team on their different customer behaviors when buying specific brands. I have reviewed the data provided by your team and would like to know if you are able to help clarify some information and provide some additional data. This will aid our team in providing the Marketing team a more complete dashboard.

As I was conducting my analysis, I noticed that there are a few data quality issues that I would like to bring to your attention. I imported and evaluated the user's data using Python. After viewing the data, I noticed a data quality issue that I would like to bring to your attention. The first thing I noticed is that there are duplicate user IDs in the _id column. My understanding is that the _id column is the primary key and each value should be unique. I believe that this is incorrect. As a potential solution for this issue, when your team is building or updating the table schema, **can they add a column or table constraint to the data entry that each new entry must be unique?** I believe this will ensure accuracy and reliability of the data, ensuring all values in the user id column are unique in the users data column.

Also, can you provide any additional brands' data? During my analysis, I also discovered that some of the brands uploaded with the receipts data were not found in the provided brands data and some had shorter barcodes that did not match. In order for our team to provide the Marketing Team with a complete analysis, we would like to have more information about the kind of brands their customers are using. Our team will need additional data to help with that. **Is there additional data available with more brands with additional barcodes?**

This additional clarification and information will help our team not only create a more complete story of the daily users, but also provide the Marketing team with a dashboard for their team to leverage when making data-driven business decisions. If you are not able to assist with this, would you be able to forward my email to someone who could help?

Thanks for all your help,

Ron