### First: Review Existing Unstructured Data and Diagram a New Structured Relational Data Model

Reviewed the 3 sample json data files provided by Fetch and converted them into Structural format. Developed a simplified, structured, relational diagram to represent how i model the data in a data warehouse.

**Source Data**

Reviewed the following JSON datasets:

users.json.gz

receipts.json.gz

brands.json.gz

**Objective**

Transformed unstructured JSON into a structured relational model suitable for analysis and warehousing.

**Approach**

* Parsed JSON using Python.Normalized nested structures like rewardsReceiptItemList from receipts.
* Developed a **relational schema** that enforces referential integrity across receipts, users, brands, and items.

ER Diagram attached in the diagrams/ folder of the GitHub repo.

### Second: Write queries that directly answer predetermined questions from a business stakeholder

Code is stored in the Queries/ folder with explanatory comments.

### Third: Evaluate Data Quality Issues in the Data Provided

**Missing or Incomplete Data**  
Some receipts are missing important fields like purchaseDate or userId, which are essential for linking and time-based analysis.  
A number of items doesn’t have barcodes or prices, which makes it difficult to attribute them to a brand or calculate spend accurately.  
  
**Missing ACCEPTED Status**  
While analyzing the rewardsReceiptStatus field, I noticed the ACCEPTED status was completely missing from the sample. We only see statuses like FINISHED or REJECTED, which raises a question:  
Does FINISHED imply a successfully processed receipt? Or are we missing the most important "approved" data segment? This needs confirmation since many business questions rely on accepted receipts specifically.  
  
**Duplicate User IDs**  
Multiple identical \_id values found in users.json, indicating ingestion or duplication issues.  
  
**"Placeholder" Items**  
Many item descriptions are "ITEM NOT FOUND" or similar placeholders. It’s unclear whether these are just temporary system flags or real entries. Should they be excluded from analytics like top products or spend?  
  
**Inconsistent Data Types**  
Financial fields like totalSpent, pointsEarned, itemPrice are stored as strings instead of numbers. This requires casting before analysis  
  
**Barcodes Without Brands**  
Many receipt\_items reference barcodes not present in the brands dataset, breaking referential integrity.These may represent unbranded items or require updates in the master data.

### Fourth: Communicate with Stakeholders

**Subject:** Preliminary Data Assessment – Observations, Questions, and Next Steps

Hello Team,

I hope this message finds you well.

As part of the Fetch Analytics Engineer assessment, I have completed an initial technical review and modeling of the sample data provided. Below is a summary of key observations, open questions, and considerations for maintaining data quality and supporting scalable analytics.

### ****Summary of Findings****

**1.Missing or Ambiguous Receipt Statuses**  
The field rewardsReceiptStatus is critical for many business metrics. However, the sample data does not contain any receipts marked as 'ACCEPTED', only values such as 'FINISHED' and 'REJECTED'.

* Clarification is needed on whether 'FINISHED' is functionally equivalent to 'ACCEPTED', or if a segment of key receipts is missing.

**2.Placeholder or Incomplete Line Items**  
Numerous records in rewardsReceiptItemList contain generic or incomplete descriptions like "ITEM NOT FOUND", and many are missing barcodes and pricing details.

* Please advise whether these should be excluded, flagged, or included in performance and spend reporting.

**3.Duplicate User Records**  
A number of user entries share identical \_id and metadata.

* Is this an intentional behavior (For Ex: device syncing or user account merges), or should these be treated as duplicates during analysis?

**4.Inconsistent Data Types**  
Fields such as totalSpent, pointsEarned, and item-level pricing are stored as strings rather than numerical types. This will require standardization for analytical operations and joins.

**5.Referential Gaps**  
Several receipt items reference barcodes that do not have a corresponding record in the brands dataset. This affects the ability to attribute spend and analyze brand performance comprehensively.

### ****Questions for Clarification****

* Can you provide a data dictionary or documentation that outlines valid values for key fields such as rewardsReceiptStatus, role, topBrand, etc.?
* Are there any known ingestion anomalies or expected duplication logic that would explain repeated user or item records?
* How should "system-generated" or placeholder items be handled in reporting pipelines?

### ****Recommended Next Steps****

To support long-term scalability and analytics accuracy, I propose:

* Implementing referential integrity checks and data validation logic during ingestion (e.g., for missing purchaseDate, orphan barcodes).
* Creating database indexes on key fields (receipt\_id, userId, purchaseDate) to optimize query performance.
* Partitioning large fact tables (such as receipts and items) by purchaseDate as the dataset grows.
* Establishing a rulebook for handling placeholder entries and incomplete records.

Please let me know if a short sync would be helpful to review these observations in detail or if further clarification is needed. I would be happy to adjust the model or ingestion logic based on your team’s requirements.

Thank you for your time and collaboration.

Warm regards,  
Sreeja N