**Assignment Submission Document**

**1. Reviewing the Existing Data and Designing a New Relational Model**

Goal:  
The objective was to review the unstructured JSON data provided and organize it into a structured, relational database model. This transformation ensures that the data is easier to query, manage, and scale in a data warehouse.

Steps Taken:

1. Data Files Reviewed:
   * Receipts: This file contains transaction-level data, such as purchase dates, total amounts spent and lists of purchased items.
   * Brands: This file provides information about product brands, including their names, categories, and barcodes.
   * Users: This file includes details about the users, such as their states, account creation dates, and activity status.
2. Challenges with the Data:
   * The JSON files were unstructured, making them difficult to work with efficiently.
   * Some fields were missing or inconsistent, such as receipts without purchase dates or brand categories being blank.
   * There was significant redundancy and no clear relationships between the data points.

Designing the Relational Model:  
To address these challenges, I designed a structured relational model with the following key tables:

* Users Table:
  + Purpose: Stores information about the users who scanned the receipts.
  + Fields: Includes user\_id (Primary Key), active status, state, created\_date, and last\_login.
* Receipts Table:
  + Purpose: Records transactions submitted by users.
  + Fields: Includes receipt\_id (Primary Key), user\_id (Foreign Key from Users), purchase\_date, total\_spent, bonus\_points\_awarded, and receipt\_status.
* Brands Table:
  + Purpose: Provides details about product brands.
  + Fields: Includes brand\_id (Primary Key), barcode, category, name, and a flag for whether the brand is a "top brand."
* Receipt Items Table:
  + Purpose: Stores detailed information about each item purchased in a receipt.
  + Fields: Includes item\_id (Primary Key), receipt\_id (Foreign Key from Receipts), barcode (Foreign Key from Brands), description, quantity, final\_price, and original\_price.

Relationships:  
The diagram below illustrates how these tables relate to each other:

* Users are linked to Receipts through the user\_id.
* Receipts are linked to individual items through the receipt\_id.
* Items are linked to Brands through the barcode.

This relational model ensures that the data is well-structured, relationships are clear, and queries can be run efficiently.

A diagram of a user flow

Description automatically generated

**2:** SQL queries

1. **What are the top 5 brands by receipts scanned for most recent month?**
   * Identified the top 5 brands based on the number of receipts scanned for the most recent month. However, the results were dominated by "Unknown Brand," which accounted for most of the receipts. This highlights a significant data quality issue where brand names are either missing or not mapped correctly.

**A screenshot of a computer

Description automatically generated**

1. **How does the ranking of the top 5 brands by receipts scanned for the recent month compare to the ranking for the previous month?**
   * Conducted a comparative analysis of the top 5 brands across the most recent and previous months. Similar to the first question, "Unknown Brand" consistently dominated both months, indicating incomplete or inconsistent data mapping for brand names.

**A screenshot of a computer program

Description automatically generated**

**A screenshot of a computer

Description automatically generated**

1. **When considering *average spend* from receipts with 'rewardsReceiptStatus’ of ‘Accepted’ or ‘Rejected’, which is greater?**
   * Calculated and compared the average spend for receipts with statuses of "Accepted" and "Rejected." It was observed that receipts with a status of "Accepted" tend to have a higher average spend.

**A screenshot of a computer

Description automatically generated**

1. **When considering *total number of items purchased* from receipts with 'rewardsReceiptStatus’ of ‘Accepted’ or ‘Rejected’, which is greater?**
   * Analyzed the total number of items purchased across "Accepted" and "Rejected" receipts. "Accepted" receipts consistently showed a higher number of purchased items, reflecting better data completeness and quality.

**A screenshot of a computer

Description automatically generated**

1. **Which brand has the most *spend* among users who were created within the past 6 months?**
   * Identified the brand with the highest total spending among users who created accounts within the past six months. This metric highlight emerging brand preferences among newer users

**A screenshot of a computer code

Description automatically generated**

**A screen shot of a computer

Description automatically generated**

1. **Which brand has the most *transactions* among users who were created within the past 6 months?**
   * Analyzed the total transactions by brand for users created in the past six months. The results highlighted the most frequently transacted brand for newer users, offering insights into their purchasing habits.

**A screenshot of a computer

Description automatically generated**

**Step 3: Data Quality Issues Identified**

**Note:**All example SQL queries used for analysis, data quality checks, and insights generation are included in the attached SQL file: consumer\_behavior.sql.  
This file contains detailed, commented queries for every step,

1. **Missing or Unmatched Barcodes**
   * **Description:** There are barcodes in transaction\_items that do not have a corresponding record in the product\_brands table.
   * **Significance:** Missing barcodes can lead to incomplete brand-level analytics and hinder proper product mapping.

**Example Queries:**

* + Identifying items with unmatched barcodes using LEFT JOIN.
  + Listing distinct barcodes from transaction\_items that are not present in product\_brands.

1. **Orphaned Records**
   * **Description:** Some receipts in transaction\_receipts have no associated items in transaction\_items.
   * **Significance:** Orphaned records indicate incomplete data relationships, affecting transaction-level analysis.

**Example Query:**

* + Identifying receipts with no items using a LEFT JOIN.

1. **Unknown Brand Dominance**
   * **Description:** A significant portion of receipts (98.8%) are attributed to "Unknown Brand," indicating missing or unlinked brand names.
   * **Significance:** This impacts brand-level insights and skews analytics. Temporal analysis revealed that "Unknown Brand" receipts spiked in January 2021.

**Example Queries:**

* + Analyzing receipt distribution by brand and percentage of total receipts.
  + Tracking monthly trends for "Unknown Brand."

1. **Missing Critical Fields**
   * **Description:** Critical fields such as categories, purchase dates, and rewards receipt items are missing in significant proportions.
   * **Significance:** Missing fields lead to incomplete data analysis and reduced business insights.
     + **Missing Categories:** 13% in product\_brands.
     + **Missing Purchase Dates:** 40% in transaction\_receipts.
     + **Missing Rewards Receipt Items:** 39% in transaction\_receipts.

**Example Queries:**

* + Summarizing missing fields using SUM and CASE statements.

1. **Invalid Data Entries**
   * **Description:** Some records contain invalid values such as:
     + Negative values for bonus\_points\_awarded or total\_spent.
     + Future purchase dates in transaction\_receipts.
   * **Significance:** Invalid data entries reduce trust in analytics and decision-making processes.

**Example Queries:**

* + Identifying records with invalid numeric fields or timestamps.

1. **Duplicate Records**
   * **Description:** Duplicate records were found in product\_brands, transaction\_receipts, and transaction\_items.
   * **Significance:** Duplicate records inflate counts and skew metrics, requiring deduplication.

**Example Query:**

* + Detecting duplicates using GROUP BY and HAVING COUNT(\*) > 1.

1. **Category Distribution Issues**
   * **Description:** A significant portion of product categories are empty or incomplete, with 13.28% missing in product\_brands.
   * **Significance:** Missing categories hinder product classification and analysis.

**Example Query:**

* + Analyzing category proportions using COUNT and PERCENTAGE.

.

**Step 4: Communication to Stakeholders**

**Email Draft:**

**Subject:** Data Quality Assessment and Recommendations

Hi Team,

I wanted to share the findings and proposed next steps from our data quality assessment of the Fetch Rewards dataset. Below are the key insights, questions, and a resolution plan to ensure the data aligns with our business objectives.

Key Findings

1. Data Quality Issues:
   * Missing Data: We found significant gaps in key fields. For instance, 13% of product categories and 40% of rewards receipt items are missing. Missing purchase dates were also a concern, affecting 448 records.
   * Invalid Data: Some records contain values that don’t make sense, like negative totals or future purchase dates. These likely result from data entry or processing errors.
   * "Unknown Brands": A large portion of receipts is tied to “Unknown Brand,” meaning the brand information wasn’t captured or matched properly. This affects nearly 99% of the dataset.
2. How We Identified the Issues:
   * We ran SQL queries to find missing, invalid, and inconsistent data.
   * We checked relationships between tables (e.g., receipts with no items, barcodes not matching products) to identify mismatches and anomalies.

Questions for Clarity

To resolve these issues effectively, we need your input:

1. Completeness: How critical are fields like category, purchase\_date, and rewards\_receipt\_items for analysis?
2. Validation Rules: Are there rules we should enforce, like preventing negative values or requiring non-null fields for key metrics?
3. Source System Details: Can you share how the data flows into our systems? For example, are there known issues in the APIs or processes that generate this data?

What We Need to Move Forward

1. Business Context: Understanding which fields are most important to the business will help us prioritize fixes. For example, are all product categories equally important, or are some more critical for reporting?
2. Technical Details: We need information about the source systems or ETL processes responsible for ingesting this data. This will help us identify root causes of missing or incorrect data.
3. Benchmarks: If we have examples of typical purchase amounts or common product categories, it will help us identify outliers or anomalies.

Challenges and Our Plan to Address Them

1. Performance and Scaling:
   * High data volumes, such as the 6,941 records in the transaction\_items table, may slow down queries.
   * Plan: Use indexing and partitioning to improve query speed.
2. Data Accuracy and Consistency:
   * Fixing historical data is critical, but we also need to prevent future issues.
   * Plan: Add validation checks to our data pipeline and set up automated alerts to catch problems early.
3. Storage and Cost Management:
   * Nested fields like rewards\_receipt\_items can increase storage and processing costs.
   * Plan: Flatten the data structure where possible and archive less-used data to reduce costs.

Next Steps

1. I’ll follow up with a proposal for cleaning the data based on the priorities you share.
2. If needed, we can set up a session to discuss the issues and the role these fields play in your reports or decisions.
3. Once we align on validation rules and requirements, we can implement automated checks to ensure ongoing data quality.

Let me know if there’s anything specific, you’d like me to focus on or if you have additional feedback. I look forward to hearing your thoughts!

Thanks,  
Vamsi Sai Tumu