Fetch Rewards Coding Exercise

1. Data Cleaning

The "Data Cleaning.ipynb" file is available in the GitHub repository. Here's a snapshot of the steps undertaken for cleaning the data:

* Transforming Data Structure: Converted stacked JSON data into a JSON array for streamlined processing.
* Data Extraction: Leveraged Pandas to extract data from the JSON array and transform it into a structured data frame.
* Data Type Correction: Resolved data type errors, whether the data appeared as lists or dictionaries, ensuring consistency.
* Timestamp Conversion: Translated Unix timestamps into human-readable date-time formats for better readability.
* Data Export: Successfully exported the cleaned data frame to a CSV file for easy readabilty.

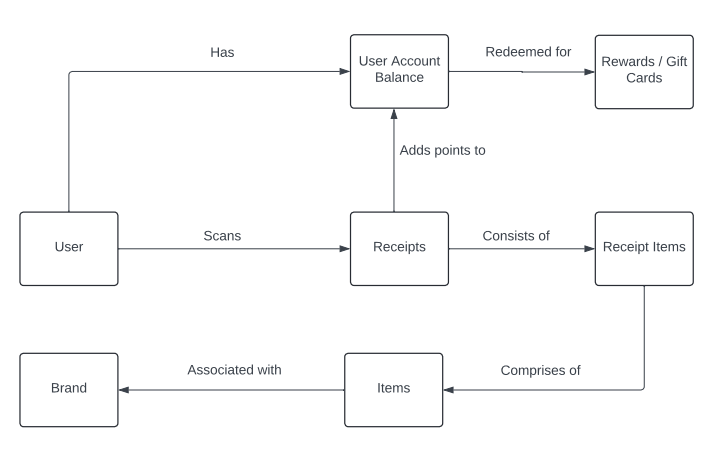
### 2. Flowchart for Receipt Scanning Process

After examining the CSV files, I developed a flowchart to illustrate the journey of users scanning receipts and the subsequent conversion to points in their accounts.

Here’s a concise overview of the process:

* A user scans their receipts.
* Each receipt consists of many receipt items.
* A receipt item comprises of an item
* A Brand is linked to specific items.
* Scanning a receipt adds points to the user's account balance.
* Users can redeem their account balance for gift cards or rewards.

This flowchart encapsulates the seamless process from scanning to reward redemption, highlighting each critical step in the user experience. The file can be found in the Github repository named as ‘Fetch Rewards Flowchart.pdf’.



**3. ER Diagram**

**ERD Assumptions:**  
  
1. CPG is being used for establishing brand hierarchy and is 1:1 with Brand and does not need to be a separate dimension table.  
2. Primary keys for the fact and dimension tables do not have null values.  
3. Both Dim\_Receipt and Dim\_Receipt\_Items should have the ability to track ‘PointsEarned’ and ‘ModifyDate’ to keep track of any changes happening on both Receipt and Receipt Item level.  
4. While we do not need to have Brand Id as foreign key in the the Fct\_Receipt\_Items table, it has been kept for convenience in querying to get answers for the sql questions.   
5. In the essence of time, while there are many more fields in the data set provided, the ERD has been designed with just the key fields required to answer the sql questions.   
  
**Pros and Cons of Current ER Diagram:**   
  
The current ER diagram is designed as a Snowflake schema which has the following pros and cons:  
**Pros:**  
Fast data retrieval  
High data quality  
High granularity  
Reduced data redundancy  
Optimized storage  
  
**Cons:**  
High overhead for initial setup  
Rigid data model  
Too many tables can require lots of joins

**Future recommendations:**  
  
1. To overcome rigidity in data modeling and added number of rows over time due to slowly changing dimensions, we can create a separate table to track change history of a given dimension. For eg., we can Item\_History and Brand\_History tables to keep track of changes in item prices and brand names over time.  
  
2. Create a Date dimension so we can perform analytics on different time periods like weeks, months, etc.  
  
3. Create Reporting tables on top of the dimensions and fact tables so that analysts don’t need to perform many table joins each time for querying the data.  
  
4. If we have further data on brand collections, we can create a separate dimension which would enable further brand analysis. For now, CPG is included as a part of Dim\_Brand.

A diagram of a computer

Description automatically generated with medium confidence

The ER diagram illustrates the relationships and structure of the data involved in the process of scanning receipts and converting them into user account balances within the Fetch Rewards system.

* A **User** can have multiple **Receipts**.
* A **Receipt** is associated with one **User**.
* A **Receipt** can have multiple **Receipt Items**.
* A **Receipt Item** is linked to one **Receipt**.
* A Receipt **Item** is associated with one **Item**.
* A Receipt Item is associated with one Brand.
* An **Item** is associated with one **Brand**.
* An Item can be associated with many Receipt Items.
* A Brand can be associated with many Items.
* A Brand can be associated with many Receipt Items.

**4. Write SQL queries to answer the business questions with the new structured data models**

**GENERAL DATA ASSUMPTIONS:**   
  
1. The test brands and items have been excluded from the data models.  
2. Null values have been handled and backfilled.  
3. Primary keys do not have null values.  
4. Data has been cleaned and does not include duplicate values.  
  
What are the top 5 brands by receipts scanned for most recent month?  
  
Assumptions:   
- Top brands are identified based on Receipt Items scanned in the given period and not the ‘topBrand’ flag in the Brands dimension.  
- Most recent month is not being calculated based on current date but based on the most recent month in data.



How does the ranking of the top 5 brands by receipts scanned for the recent month compare to previous month?   
  
Assumptions: - Top brands are identified based on Receipt Items scanned in the given period and not the ‘topBrand’ flag in the Brands dimension.  
- Most recent month is not being calculated based on current date but based on the most recent month in data.





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?

Assumption: We are considering past 6 months from current date.  
  


Which brand has the most *transactions* among users who were created within the past 6 months?  
  
Assumption: We are considering past 6 months from current date.  
We are considering transactions as the number of receipt item ids associated with a brand.



5. Evaluating data quality issues using SQL  
  
- **Identify missing data**

Here we are checking if there are null values in any of the fields



**- Identify duplicate records**  
Here we are checking if there are duplicate user ids.  
  


**- Identify Data Inconsistencies**  
  
Here we are checking if there are different versions of a given state. Eg: CA and California  
  


**- Referential Integrity Violations**  
  
Here we are checking if there is a receipt without a user id.



**- Identify Anomalies**   
Here we are checking if the user created date is greater than today  
  


**Example Data Issues Identified:**  
**Brands Table:**   
- Barcode values populated in BrandCode field  
- Test Brands included  
- Missing values in Brand Code and Category Code fields  
- Brand Name needs cleaning. Eg. Dr Pepper vs Dr. Pepper  
**Users table:**   
- Missing values in lastLogin, State and SignUp Source fields  
**Receipts table:**   
- Missing values in finishedDate, pointsAwardedDate, PurchasedDate, PurchasedItemCount and total Spent fields

**6. Email to the Stakeholder**

Hi Alex,

I hope you’re doing well! I wanted to share some findings from our recent look into the data for Fetch Rewards — covering receipts, users, brands and receipt items.   
  
**Here’s what we have found so far:**

1. **Missing Data**:
   * Some user records are missing values like Last Login Date and Signup Source fields
2. **Data Inconsistencies**:
   * Fields like Brand Name and Brand Code have inconsistencies that could mess with our reports.
3. **Duplicates**:
   * We found some duplicate entries in our brands data.
4. **Referential Issues**:
   * Some receipt items reference items that are not in our database.
5. **Date Anomalies**:
   * There are future dates in some fields that need correction.

**Questions for you:**

1. Which use cases are most affected by these issues?
2. Do we have external data to fill gaps or correct errors?
3. How often should we refresh our data?

**Next Steps:**

* We need to set default values for missing fields and confirm business rules.
* Access to external data and domain experts will help with verification and corrections.

**Performance and Scaling:**

* Cleaning the data will increase its volume, so we’ll be optimizing our database and considering scalable storage solutions.

Let me know if you have any questions or need more details. Looking forward to your guidance on the path forward!

Regards,

Vedika Shenoy  
Sr. Data Analyst | Fetch Rewards