# FETCH ASSESSMENT

Saanil Khanna

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

**A screenshot of a computer

Description automatically generated**

The ER diagram represents the relationships between three tables: receipts, users, and brands.

**receipts Table:**

The receipts table stores information about receipts.

Each receipt is identified by a unique \_id.

The table includes columns such as bonusPointsEarned, bonusPointsEarnedReason, createDate, dateScanned, finishedDate, modifyDate, pointsAwardedDate, pointsEarned, purchaseDate, purchasedItemCount, rewardsReceiptStatus, totalSpent, userId, and rewardsReceiptItemList.

The userId column establishes a relationship with the users table based on the users.\_id primary key.

**users Table:**

The users table contains information about users.

Each user is identified by a unique \_id.

The table includes columns such as state, createdDate, lastLogin, role, and active.

The users.\_id primary key is referenced by the receipts.userId foreign key, creating a one-to-many relationship between users and receipts.

**brands Table:**

The brands table stores information about brands.

Each brand is identified by a unique \_id.

The table includes columns such as barcode, brandCode, category, categoryCode, cpg, topBrand, and name.

The rewardsReceiptItemList column in the receipts table references the brands.cpg column, establishing a relationship based on the cpg value from the rewards receipt item list.

Overall, the ER diagram illustrates the connections between the receipts, users, and brands tables. It shows that receipts are associated with users through the userId relationship, and receipts are also linked to brands based on the cpg value from the rewards receipt item list.

**Structuring the data:**

A screen shot of a computer program

Description automatically generated

1. Data Reading:

* The code reads the JSON data files (brands.json.gz, receipts.json.gz, and users.json.gz) using spark.read.json.
* Temporary views are created for each dataset using createOrReplaceTempView.

1. DataFrame Creation:

* DataFrames are created for each dataset (brands, receipts, and users) using the loaded JSON data.

1. DataFrame Operations:

* The df DataFrame is created by selecting all columns from the users table and renaming the first column as user\_id\_new.
* Another DataFrame df is created by selecting all columns from the brands table and renaming the first column as cpg\_id.

1. Temporary View Creation:

* Temporary views are created for the updated DataFrames (users and brands) using createOrReplaceTempView.

A computer screen shot of text

Description automatically generated

1. SQL Query and DataFrame Transformations:

* The code executes an SQL query using Spark SQL to perform joins between the Receipts, Users, and Brands tables based on specific column conditions.
* The resulting DataFrame undergoes several transformations using withColumn and from\_unixtime functions.
* Columns are renamed, and UNIX timestamps are converted to the appropriate timestamp type.
* The \_id column is dropped using drop.

1. Deduplication:

* dropDuplicates() is applied to remove any duplicate rows from the DataFrame.

1. Column Selection and Rearrangement:

* The desired columns are selected in the desired order using select.

1. Temporary View Creation:

* A temporary view named "shopping" is created with the final DataFrame using createOrReplaceTempView.

1. Conversion to Pandas DataFrame and Saving as CSV:

* The toPandas method is called on the Spark DataFrame df to convert it into a Pandas DataFrame.
* The resulting Pandas DataFrame is assigned to the variable pandas\_df.
* The to\_csv method is used on the Pandas DataFrame pandas\_df to save it as a CSV file.

The CSV file is named "Final.csv", and the index=False parameter is passed to exclude the row index from the saved CSV file.

### **Second: Write a query that directly answers a predetermined question from a business stakeholder**

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

Code:

A computer screen shot of a code

Description automatically generated

Output: Top 5 Brands by receipts scanned for most recent month

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 compared to the ranking for the previous month?

Code:

A screenshot of a computer code

Description automatically generated

Output:

A screenshot of a computer

Description automatically generated

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

rewardsReceiptStatus column does not have any 'Accepted' value, and 'Finished' status is assumed to be equivalent to 'Accepted'

Code:

A screen shot of a computer

Description automatically generated

Output: Accepted is Greater

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?

Code:

A screenshot of a computer

Description automatically generated

Output: Accepted is Greater

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?

Code:

A screen shot of a computer code

Description automatically generated

Output: Brand Name “Chester’s” had the most spend among users within the past 6 months

A screenshot of a computer

Description automatically generated

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

Code:

A screen shot of a computer code

Description automatically generated

Output: Brand name “Chester’s” had the most transactions among the users who were created within the past 6 months

A screenshot of a computer

Description automatically generated

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

**Data Quality Issues:**

* **Data Format:** Data was stored in JSON format. This required transformation into a structured format (SQL or CSV) for easier analysis. However, due to the complex nature of the data and the need for proper formatting, this transformation process consumed a significant amount of time and resources.
* **Nested Array:** One of the columns had an array of elements. In order to perform analysis on individual elements of the array, it was necessary to use techniques like exploding the array in PySpark to convert it into separate rows.
* **Inadequate Data Explanation:** The provided data lacked clear and comprehensive explanations of the columns and their meanings. This made it difficult to understand the context and semantics of the data.
* **Undefined Relationships:** The relationships between tables were not clearly defined. It was necessary to understand the relationships between different tables in order to perform accurate analysis and joins.
* **Data Cleanliness:** The data had some quality issues, such as missing or inconsistent values, duplicate records, or improperly formatted data. These issues needed to be addressed and cleaned to ensure accurate analysis and results.
* **Date Formatting:** The date columns were in Unix timestamp format, which required conversion to a human-readable date format for better understanding and analysis. Multiple transformations were needed to convert Unix timestamps into the desired date format.

These data quality issues highlight the importance of ensuring data consistency, completeness, accuracy, and proper formatting. Addressing these issues is crucial to obtain reliable and meaningful insights from the data and to perform accurate analysis and decision-making.

### **Fourth: Communicate with Stakeholders**

**Subject:** Data Quality Assessment and Optimization for Business Insights

Dear [Business Leader's Name],

I hope this message finds you well. I wanted to discuss some insights and observations regarding the data we've been working with, as well as potential areas for optimization. As a data professional, my goal is to ensure the accuracy and reliability of our data assets to support informed decision-making. Below, I address some key points for consideration:

**Data Quality Assessment:**

While analyzing the data, I encountered a few data quality issues that are important to address. These include:

* Data was stored in JSON format. This required transformation into a structured format (SQL or CSV) for easier analysis. However, due to the complex nature of the data and the need for proper formatting, this transformation process consumed a significant amount of time and resources.
* Nested Array: One of the columns contained an array of elements, necessitating the use of techniques like exploding the array to analyze individual elements.
* Inadequate data explanations: The provided data lacked comprehensive explanations of columns and their relationships, making it challenging to interpret and analyze the data accurately.
* Date formatting: Date columns were stored as Unix timestamps, requiring conversion to a human-readable format for better understanding and analysis.

**Resolving Data Quality Issues:**

To resolve these data quality issues, I would appreciate your insights on the following:

* Clarification on the meaning and context of the data columns to ensure accurate interpretation and analysis.
* Assistance in identifying and addressing any missing or inconsistent data values.
* Collaborating with the data team to determine the correct format and transformation for the date columns.
* Defining the relationships between different data tables to ensure accurate joins and analysis.

**Optimizing Data Assets:**

To optimize the data assets and maximize their value, it would be helpful to have additional information, such as:

* Key performance indicators (KPIs) or business metrics that are crucial for decision-making.
* Any specific analysis or reporting requirements to tailor the data assets accordingly. Insights into potential data sources or external datasets that could enhance our analyses.

**Performance and Scaling Considerations:**

As we move towards production, it is important to consider performance and scaling concerns, such as:

* Increasing data volumes: Assessing the scalability of our data processing and storage systems to handle larger datasets efficiently.
* Query performance: Optimizing our data infrastructure and query execution to ensure timely retrieval of insights.
* Real-time analysis: Exploring technologies and approaches for real-time data analysis, if necessary.

I am committed to delivering reliable and actionable insights to support your decision-making process. Your guidance and input on the above points would greatly assist in resolving data quality issues and optimizing our data assets.

Please let me know your availability for a brief discussion to address any questions or concerns you may have. I am looking forward to our collaboration in leveraging data for better business outcomes.

Thank you for your attention, and I remain at your disposal.

Best regards,

Saanil Khanna