**Fetch Rewards Coding Exercise**

**Demonstrate how you reason about data and how you communicate your understanding of a specific data set to others.**

The Data:

3 unstructured JSON datasets are given.

1. Receipts - [Download receipts.json.gz](https://fetch-hiring.s3.amazonaws.com/data-analyst/ineeddata-data-modeling/receipts.json.gz)

The schema of the Receipts dataset is –

* **\_id**: uuid for this receipt
* **bonusPointsEarned**: Number of bonus points that were awarded upon receipt completion
* **bonusPointsEarnedReason**: event that triggered bonus points
* **createDate**: The date that the event was created
* **dateScanned**: Date that the user scanned their receipt
* **finishedDate**: Date that the receipt finished processing
* **modifyDate**: The date the event was modified
* **pointsAwardedDate**: The date we awarded points for the transaction
* **pointsEarned**: The number of points earned for the receipt
* **purchaseDate**: the date of the purchase
* **purchasedItemCount**: Count of number of items on the receipt
* **rewardsReceiptItemList**: The items that were purchased on the receipt
* **rewardsReceiptStatus**: status of the receipt through receipt validation and processing
* **totalSpent**: The total amount on the receipt
* **userId**: string id back to the User collection for the user who scanned the receipt

1. Users - [Download users.json.gz](https://fetch-hiring.s3.amazonaws.com/data-analyst/ineeddata-data-modeling/users.json.gz)

The schema of the Users dataset is -

* **\_id**: user Id
* **state**: state abbreviation
* **createdDate**: when the user created their account
* **lastLogin**: last time the user was recorded logging in to the app
* **role**: constant value set to 'CONSUMER'
* **active**: indicates if the user is active; only Fetch will de-activate an account with this flag

1. Brands - [Download brands.json.gz](https://fetch-hiring.s3.amazonaws.com/data-analyst/ineeddata-data-modeling/brands.json.gz)

The schema of the Brands dataset is –

* **\_id**: brand uuid
* **barcode**: the barcode on the item
* **brandCode**: String that corresponds with the brand column in a partner product file
* **category**: The category name for which the brand sells products in
* **categoryCode**: The category code that references a BrandCategory
* **cpg**: reference to CPG collection
* **topBrand**: Boolean indicator for whether the brand should be featured as a 'top brand'
* **name**: Brand name

**Requirements:**

1. **Review unstructured JSON data and diagram a new structured relational data model**

* Refer file ‘fetch.R’ for the R code used to load, explore and clean data.
* Receipts.csv, rewardsReceiptItemList.csv, brands.csv, users.csv are the cleaned, structured data files.
* Refer ‘er.pdf’/’er.png’/’er.mwb’ files to understand the entity relationship schema of the cleaned, structured tables (in mysql work bench)

1. **Generate a query that answers a predetermined business question**

* Refer to the ‘fetch.sql’ file to access the sql queries used to answer the predetermined questions.
* Refer to files ‘1.png’,’3.png’,’4.png’,’5.png’,’6.png’ for the sample outputs of the sql queries used.

1. **Generate a query to capture data quality issues against the new structured relational data model**

Refer to the R code in ‘fetch.R’ file to access the code used to clean and structure the data. Some of the data quality issues are:-

* In the receipts dataset all the ‘reward’ related columns are 50% null values.
* In the rewardsReceiptItemList dataset, most of the columns contain 90% null values.
* In the brands dataset columns category\_code,top\_brand,brand\_code have more than 50% of the null values
* Users dataset contains comparatively less null values
* Also, there are many special characters in some columns in rewardsReceiptItemList and brands datasets. These special characters are not compatible with mysql.
* rewardsReceiptItemList dataset contains lot of junk values (‘4011’) in the barcode column. So this column is not reliable.
* In the Brands dataset, the id column is supposed (logically) to be the primary key, but there are some duplicates in this column

1. **Write a short email or Slack message to the business stakeholder**

The given below are some of the questions that I ask regarding data quality:

* How is the data captured? The reason for the majority of the null values is because of no data or is it because of errors in data entry/storage process.
* Is the barcode column in rewardsreceiptitemlist dataset and brands dataset supposed to be same? If yes, then can each barcode can contain many items?
* Is the barcode for each brand unique? If yes, then why is the id column generated.
* Why are there duplicate observations in users dataset? Is it because each user registered twice? Is it a data entry mistake?
* Are these the only datasets available? Is there any separate entity for item level data capturing?