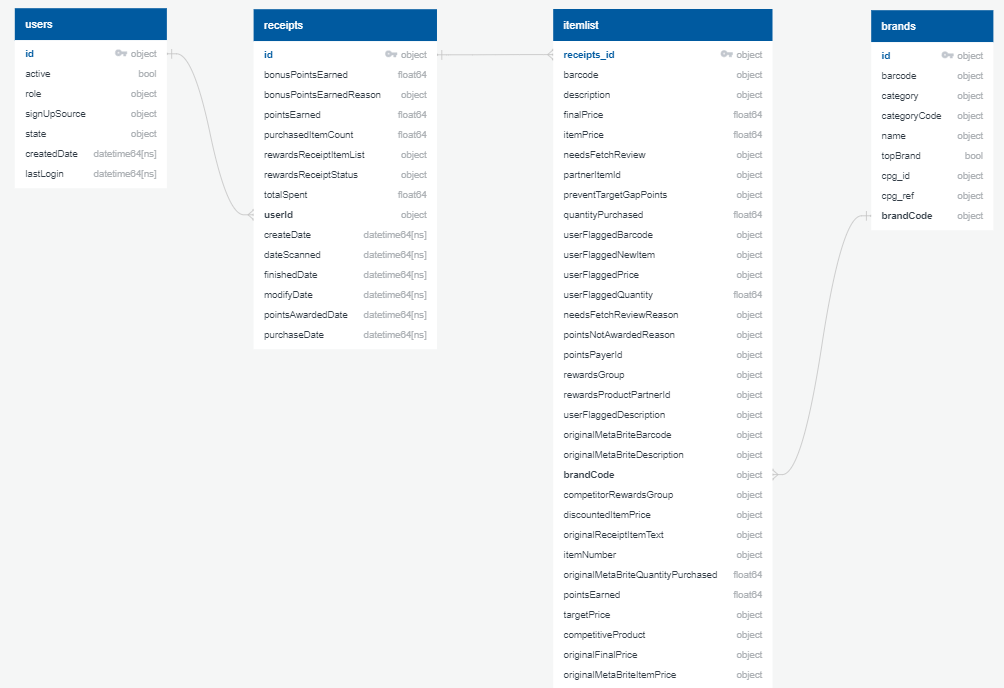
**1) Review Existing Unstructured Data and Diagram a New Structured Relational Data Model**

Online tool (<https://app.quickdatabasediagrams.com/>) is used to create the data model diagram.

Itemlist table is created after exploding the JSON data from ‘rewardsReceiptItemList’ column of ‘receipts’ dataset. Also, ‘ID’ column from ‘receipts’ table added to itemlist table as a ‘receipts\_ID’ for cross reference.



**2) Write a query that directly answers a predetermined question from a business stakeholder**

I used MS SQL Server dialect for queries. (refer to ‘SQLQuery.sql’ file). Cleaned datasets are exported to CSV files using Python then imported to MS SQL Server as tables.

**2.1) What are the top 5 brands by receipts scanned for most recent month?**

There is no brand associated with the receipts scanned in most recent month (March 2021). So, I ran query with previous month (February 2021).

Below are the top brands from February 2021

1. BRAND
2. MISSION
3. VIVA

SELECT TOP 5 IT.brandcode as 'Top Brands', COUNT(IT.brandCode) as 'Receipts scanned'

FROM itemlist as IT

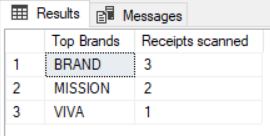
INNER JOIN receipts as RE ON IT.receipts\_id = RE.id

WHERE IT.brandCode IS NOT NULL AND

(MONTH(RE.dateScanned) = (SELECT MONTH(MAX(R1.datescanned))-1 FROM receipts as R1) AND YEAR(RE.dateScanned) = (SELECT YEAR(MAX(R2.datescanned)) FROM receipts as R2))

GROUP BY IT.brandcode

ORDER BY COUNT(IT.brandCode) DESC



**2.2) How does the ranking of the top 5 brands by receipts scanned for the recent month compare to the ranking for the previous month?**

To compare the ranking from current month with previous month I ran query with January 2021 data. No top 5 brand from February 2021 appeared in top 5 brand from January 2021.

Below are the top 5 brands from January 2021

1. HY-VEE
2. BEN AND JERRYS
3. PEPSI
4. KROGER
5. KLEENEX

SELECT TOP 5 IT.brandcode as 'Top Brands', COUNT(IT.brandCode) as 'Receipts scanned'

FROM itemlist as IT

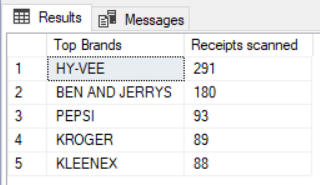
INNER JOIN receipts as RE ON IT.receipts\_id = RE.id

WHERE IT.brandCode IS NOT NULL AND

(MONTH(RE.dateScanned) = (SELECT MONTH(MAX(R1.datescanned))-2 FROM receipts as R1) AND YEAR(RE.dateScanned) = (SELECT YEAR(MAX(R2.datescanned)) FROM receipts as R2))

GROUP BY IT.brandcode

ORDER BY COUNT(IT.brandCode) DESC



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

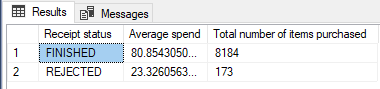
Average spend for ‘Finished’(Accepted) status is $80.85 which is higher than average spend for ‘Rejected’ $23.33

SELECT rewardsReceiptStatus as 'Receipt status', AVG(totalSpent) as 'Average spend', SUM(purchasedItemCount) as 'Total number of items purchased'

FROM receipts

WHERE rewardsReceiptStatus IN ('FINISHED', 'REJECTED')

GROUP BY rewardsReceiptStatus



**2.4) When considering total number of items purchased from receipts with 'rewardsReceiptStatus’ of ‘Accepted’ or ‘Rejected’, which is greater?**

Total number of items purchased for ‘Finished’(Accepted) status is 8184 which is higher than total number of items purchased for ‘Rejected’ 173

(Please refer to the above query)

**2.5) Which brand has the most spend among users who were created within the past 6 months?**

There are no new users created in last 6 months (or even 7 months). So I considered the users created in last 8 months.

‘HEMPLER’S” brand has the most spend among users who were created within the past 8 months.

SELECT TOP 1 IT.brandcode as 'Brand with most spend', SUM(IT.itemprice) as 'Total spend'

FROM Andon.pan.itemlist as IT

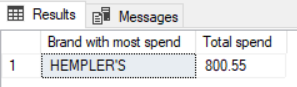
INNER JOIN Andon.pan.receipts as RE ON IT.receipts\_id = RE.id

INNER JOIN Andon.pan.users as US ON RE.userId = US.ID

WHERE IT.brandcode IS NOT NULL AND US.createdDate > dateadd(month,-8,CURRENT\_TIMESTAMP)

GROUP BY IT.brandcode

ORDER BY SUM(IT.itemprice) DESC



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

‘BEN AND JERRYS’ brand has the most transactions among users who were created within the past 8 months.

SELECT TOP 1 IT.brandcode as 'Brand with most transcations', COUNT(IT.brandcode) as 'Total transactions'

FROM Andon.pan.itemlist as IT

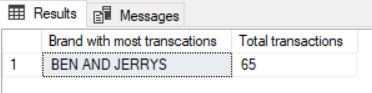
INNER JOIN Andon.pan.receipts as RE ON IT.receipts\_id = RE.id

INNER JOIN Andon.pan.users as US ON RE.userId = US.ID

WHERE IT.brandcode IS NOT NULL AND US.createdDate > dateadd(month,-8,CURRENT\_TIMESTAMP)

GROUP BY IT.brandcode

ORDER BY COUNT(IT.brandcode) DESC



**3) Evaluate Data Quality Issues in the Data Provided**

* There are duplicated rows is ‘users’ dataset. Duplicated rows need to be deleted to use user-ID as primary key to link with ‘receipts’ dataset.
* There are lots of missing value in each dataset. (eg. ‘brandcode’ column has lot of missing value in both ‘brands’ and ‘receipts/itemlist’ dataset). Which might yield wrong result while trying to filter top brands.
* There is no good link (PK/FK) between ‘brands’ and ‘receipts/itemlist’ dataset. (Brand-ID should be added to ‘receipts’ dataset instead of ‘brandcode’)
* Date columns in ‘users’ and ‘receipts’ dataset need to be converted to datetime format while importing to JSON files.

(Please refer to ‘FetchRewards.ipynb’ file for Python codes)

**4) Communicate with Stakeholders**

Hello,

After doing preliminary analysis, I see there are some issues with the json data exports. To build the production version and optimize the process, we first need to address these issues before moving forward.

Questions regarding the data.

* What are the KPI’s we want to monitor?
* What is the source of this dataset?
* How is the source data generated?

During preliminary analysis I found following quality issues with the dataset.

* Some datasets have duplicate records.
* All the datasets have missing information.
* There is no good key/link to connect brand and receipt/itemlist datasets.
* Dataset is not recent.

To resolve these quality issues, better link needs to be established between brand and receipt/itemlist datasets. For master datasets such as brands missing records should be possibly updated with correct information available. Duplicate records should be avoided while exporting from source data.

As number of users starts increasing number of receipts scanned will increase which results into larger size receipts dataset. This could affect the performance. If we have target KPI’s defined, then we can eliminate the redundant information from these datasets to keep dataset leaner.

Regards,

Pankaj Shinde

Data Analyst