

fetch rewards coDING eXERCISE – dATA aNALYST



Bhavani Lagishetty

bhavanilagishetty6@gmail.com

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

# 

# Receipts Table: id\_oid is the primary key and userID is the foreign key

# Brands: id\_oid is the primary key

# Users: id is the primary key

# Rewardslist : is another table created from the nested values of the column rewards receipt item list from the receipts table.

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

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

**Approach**:

Firstly, to get the most recent month from the receipts table, I have used to concatenate function and combined both month and year from the datescanned column and filtered it which a where clause and max.

Next, with the month date found joined rewardslist table and brand table and found the other required fields.

**Query**:

SELECT STR\_TO\_DATE(r.dateScanned1, '%d-%m-%Y') as Most\_recent\_month , rl.barcode, count(rl.barcode) as Quantity, b.name as Brand\_name, count(b.name) as Highest\_scanned\_brand, b.brandCode

FROM fetch\_data.receipts r

left join fetch\_data.rewardlist rl on r.id\_oid = rl.\_id

left join fetch\_data.brands b on rl.barcode = b.brandCode

WHERE concat(month(STR\_TO\_DATE(dateScanned1, '%d-%m-%Y')), year(STR\_TO\_DATE(dateScanned1, '%d-%m-%Y'))) = (

SELECT distinct concat(month(STR\_TO\_DATE(dateScanned1, '%d-%m-%Y')), year(STR\_TO\_DATE(dateScanned1, '%d-%m-%Y')))

FROM fetch\_data.receipts ORDER BY concat(month(STR\_TO\_DATE(dateScanned1, '%d-%m-%Y')), year(STR\_TO\_DATE(dateScanned1, '%d-%m-%Y'))) DESC

LIMIT 1 )

and rl.\_id is not null

group by b.name, STR\_TO\_DATE(r.dateScanned1, '%d-%m-%Y'), rl.barcode, b.brandcode

order by count(b.name) desc

LIMIT 5

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

**Approach:**

I first found out the last 6 months as per the data available in users table, created date field and then took the userID from the table. Then matched those userid with the receipts table and found the brandcode which is available in brand table to find the brands. Limited the function to 1 as we need the most spend brand.

**Query**:

Select bd.name as Brand\_name, bd.brandCode from fetch\_data.brands bd

where bd.brandCode =

( select rwl.brandCode from fetch\_data.rewardlist rwl

where rwl.\_id in

( Select distinct re.id\_oid from fetch\_data.receipts re

where re.userId in

( SELECT distinct id\_oid as userID

FROM fetch\_data.users us

WHERE STR\_TO\_DATE(us.createdDate1, '%m-%d-%Y') >=

DATE\_ADD((SELECT MAX(STR\_TO\_DATE(createdDate1, '%m-%d-%Y')) FROM fetch\_data.users),INTERVAL -6 MONTH)

)

)

and rwl.brandCode is not null

group by rwl.brandCode

order by count(rwl.brandCode) desc

limit 1

)

# Third: Evaluate Data Quality Issues in the Data Provided

* Missing values: Absence of data in most of the fields are missing/blanks which is affecting the data analysis. We have about 4160 null values only in receipts table.

For example:

* + A column named top brand has many null values whereas that column is expected to provide a Boolean value such as true and false or 0 and
  + Brand code and Bar code columns in brands table and rewards table (part of receipts table) has many missing values because of which finding a brand was difficult.
* Duplicate data: Users table has around 282 duplicate records.
* Data consistency across tables: each table has different column names for the same kind of data, maintaining the same field name will make the data consistent.

# Fourth: Communicate with Stakeholders

**Subject**: Summary of Data Analysis and Recommendations

Dear Stakeholders;

I hope this message finds you well. I wanted to provide you with an overview of the recent analysis conducted on our data assets, as well as outline some key findings and next steps to optimize data quality and usability.

**Key Findings:**

While reviewing the data, we’ve identifies several areas of concern: like missing values, duplicate records, inconsistent data formatting through a combination of manual inspection and analysis techniques.

**Next Steps:**

According to the analysis I could do on the data, the best way to approach the data would be

* Data cleansing and validation which will involve collaborating closely with the data team to develop more validation rules to get a accurate data.
* Standardizing the column names across all datasets, this will streamline the data.
* Maintenance of the data over time to time will be a crucial step as that will make us understand what needs to be touched and what is going well.

**Performance and Scaling:**

The main challenge is the data volumes continue to grow day by data and addressing the issues will be difficult, hence we need to optimize the data storage and work towards the future infrastructure such as exploring the cloud-based tools and technologies, and the next would be working on data indexing will help in enhancing the query performance and reduce latency.

Thank you for providing this opportunity to go through the data and would appreciate if you can provide any feedback or insights you may have.

Regards,

Bhavani Lagishetty