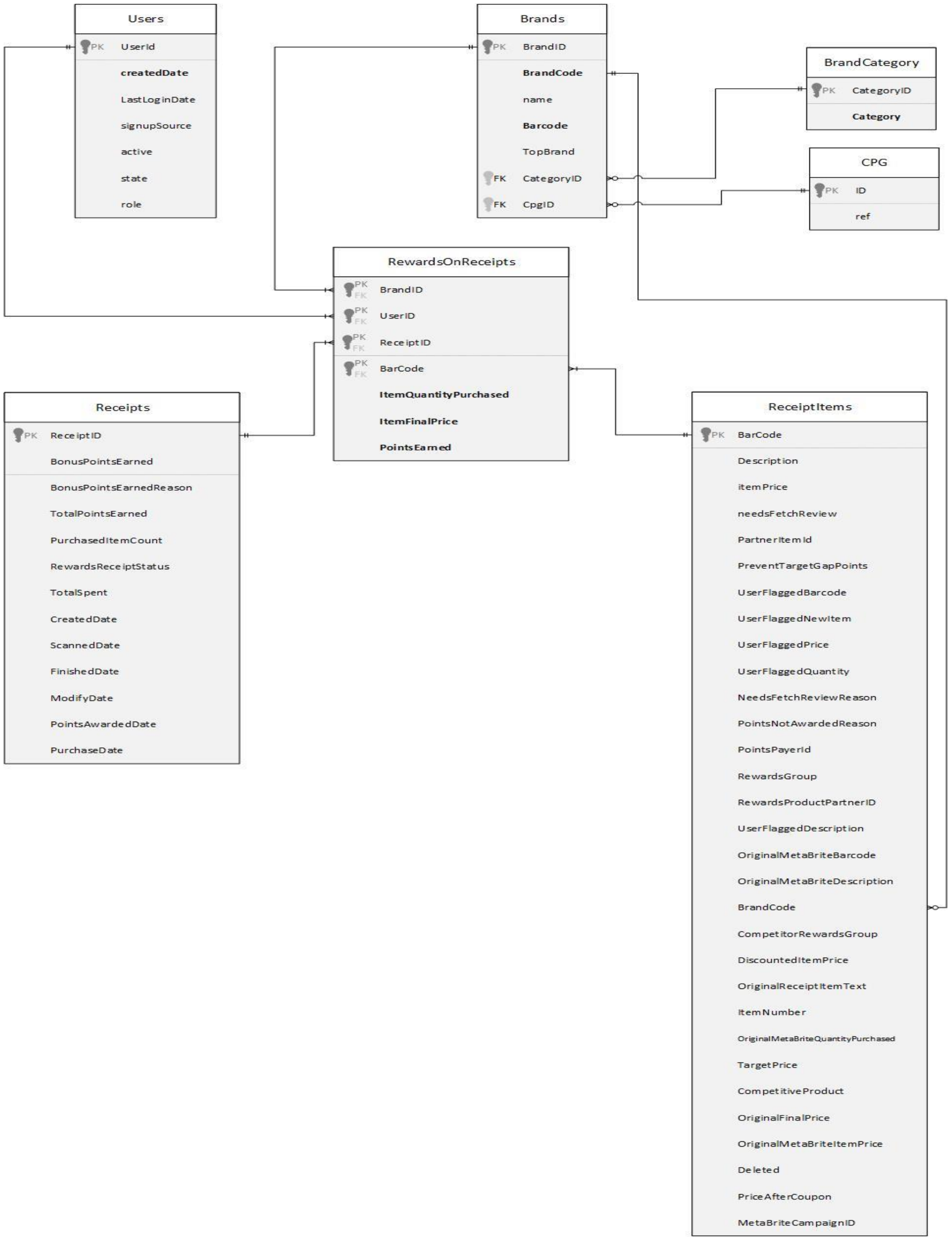
Fetch Rewards DA Challenge

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



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

I have used Oracle SQL dialect for querying the database.

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

   ```
   WITH CTE AS (SELECT rw.brandid, sum(rw.itemfinalprice) as total_amount
   FROM rewardsonreceipts rw
   JOIN receipts r
   ON rw.receiptsid = r.receiptid
   where r.scanneddate BETWEEN trunc (sysdate, 'mm') AND sysdate
   GROUP BY rw.brandid
   ORDER BY total_amount DESC)
   SELECT cte.brandid, b.name
   FROM CTE
   JOIN BRANDS_FR b
   ON cte.brandid = b.brandid
   FETCH FIRST 5 ROWS ONLY
   ```

2. **How does the ranking of the top 5 brands by receipts scanned for the most recent month compare to the ranking for the previous month?**

   ```
   WITH CTE AS (SELECT rw.brandid, sum(rw.itemfinalprice) as total_amount
   FROM rewardsonreceipts rw
   JOIN receipts r
   ON rw.receiptsid = r.receiptid
   WHERE TRUNC(r.scanneddate,'mm') =  trunc (sysdate, 'mm')
   GROUP BY rw.brandid
   ORDER BY total_amount DESC),
   CURRENT_TOP_BRANDS AS (
   SELECT cte.brandid, DENSE_RANK(cte.total_amount) OVER (ORDER BY cte.total_amount DESC)
   AS 'CURRENT_RANK'
   FROM cte
   ),
   CTE1 AS (SELECT rw.brandid, sum(rw.itemfinalprice) as total_amount
   FROM rewardsonreceipts rw
   JOIN receipts r
   ON rw.receiptsid = r.receiptid
   WHERE TRUNC(r.scanneddate,'mm') =  add_months(trunc (sysdate, 'mm'),-1)
   GROUP BY rw.brandid
   ORDER BY total_amount DESC),
   ```

```
PREVIOUS_TOP_BRANDS AS (
SELECT cte1.brandid, DENSE_RANK(cte1.total_amount) OVER (ORDER BY cte1.total_amount
DESC) AS 'PREVIOUS_RANK'
FROM cte1
)
SELECT CURRENT_TOP_BRANDS.BRANDID, CURRENT_TOP_BRANDS.CURRENT_RANK,
PREVIOUS_TOP_BRANDS.PREVIOUS_RANK
FROM CURRENT_TOP_BRANDS
JOIN PREVIOUS_TOP_BRANDS
ON CURRENT_TOP_BRANDS. BRANDID = PREVIOUS_TOP_BRANDS.BRANDID
WHERE CURRENT_TOP_BRANDS.CURRENT_RANK <= 5
```

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

```
WITH accept_cte as (SELECT 'Accepted' AS STATUS , ROUND(AVG(TOTALSPENT),2) as
AVG_SPENT
FROM RECEIPTS
WHERE UPPER(rewardsreceiptstatus) = 'ACCEPTED'),
reject_cte as (SELECT 'Rejected', ROUND(AVG(TOTALSPENT),2) as AVG_SPENT
FROM RECEIPTS
WHERE UPPER(rewardsreceiptstatus) = 'REJECTED')
SELECT * FROM accept_cte
UNION
SELECT * FROM reject_cte
```

The above query gives the average spent for both the accepted and rejected rewards status which can be used to find out the greater average spent status.

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

```
WITH accept_cte as (SELECT 'Accepted' AS STATUS , SUM(PURCHASEDITEMCOUNT) as
TOTAL_ITEMS
FROM RECEIPTS
WHERE upper(rewardsreceiptstatus) = ‘ACCEPTED’),
reject_cte as (SELECT 'Rejected', SUM(PURCHASEDITEMCOUNT) as TOTAL_ITEMS
FROM RECEIPTS
WHERE upper(rewardsreceiptstatus) = 'REJECTED')
SELECT * FROM accept_cte
UNION
SELECT * FROM reject_cte
```

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

```
SELECT BRANDID, SUM(ITEMFINALPRICE) as TOTAL_SPEND
FROM REWARDSONRECEIPTS
WHERE BRANDID IS NOT NULL AND USERID IN (SELECT USERID
FROM USER
WHERE CREATEDDATE >= add_months(TRUNC(sysdate,'mm'),-6))
GROUP BY BRANDID
ORDER BY TOTAL_SPEND DESC
FETCH FIRST ROW ONLY
```

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

```
SELECT BRANDID, COUNT(BARCODE) AS TOTAL_COUNT
FROM REWARDSONRECEIPTS
WHERE BRANDID IS NOT NULL AND USERID IN (SELECT USERID
FROM USER
WHERE CREATEDDATE >= add_months(TRUNC(sysdate,'mm'),-6))
GROUP BY BRANDID
ORDER BY TOTAL_COUNT DESC
FETCH FIRST ROW ONLY
```

Third: Evaluate Data Quality Issues in the Data Provided

I have analyzed the data from Users, Brands and Receipts datasets using Python and listed the major data quality issues.

Some of the major data quality issues I found during the analysis are:

1. More than 50% user records are duplicate.
2. pointsEarned are missing in approximately 15% of records.
3. purchaseDate, purchasedItemCount is missing in about 6% of receipts records. It could pose an to validate if the rewards were applicable when the item was purchased.
4. There are about 186 products in receipts of brands which are not included in the brands dataset.
5. There are about 117 users listed in receipts who don’t exist in users dataset.

Please refer the file FetchRewards.ipynb file for detailed analysis and insights.

Fourth: Communicate with Stakeholders

Subject: Regarding data Quality Issues with Receipts, Brands and Users data

Hello Business Analyst,

As discussed in our last meeting, I worked on exploring and analyzing the data on users, receipts, and brands. After importing and cleaning the data using Python, I did a preliminary data quality check. Please find the issues with regards to data quality listed below.

1. There are multiple records for a single user in the dataset. In fact, more than 50% of the data is duplicated.
2. The category for more than 50% of brands is missing. So, we cannot find the category of a significant number of products. This will lead to a problem if we need to analyze which category of products/brands have more sales or if we plan on any category based brands promotional offers.
3. There is a lot of missing values in purchasedDate, purchasedItemCount, pointsEarned in receipts data. The missing data in purchaseDate and purchasedItemCount would make it difficult to validate if the rewards were being offered for the products on the date of purchase.
   Additionally, I would like to know if the missing values in pointsEarned indicate system error or that the user did not earn any points for the particular purchase. It would be better to set zero for points earned if the user doesn't earn any points which could help in points based analysis and also help identify system errors.
4. There are about 186 products listed in receipts which belong to brands that do not exist in brands data. Again, this would lead to a problem when we need to do any brand based aggregations.
5. Similarly, about 117 receipts belong to users that do not exist in the user's data. This is highly concerning since we cannot keep a track of user history of purchase and rewards. Also, we would need to identify if there is any issue in the application for user registration since there is duplicate and missing data for users.

I also observed that certain data especially dates is not in right format. As listed above, there are quite a few issues related to the quality of data available and I would like to discuss with you in detail to understand the data better and address these concerns. Please let me know your availability and we can schedule a meeting accordingly.

Best,

Daya Nayak