- What questions do you have about the data?

Hi John,
I wanted to bring to your attention some observations and concerns regarding our data assets, particularly in the users.json, brands.json, and receipts_fact files. I believe discussing these issues will help us improve the quality and reliability of our data for better decision-making and analysis.

Users.json:

- I noticed several duplicates based on the user_id field in the users.json file. Could you please provide insights into how this data is populated? Are these duplicates expected due to multiple product purchases by the same users, or should we have a separate users database ensuring uniqueness?
- Additionally, there are missing values in certain fields such as last login date. Is this expected behavior, or should we investigate further to ensure data completeness?

Brands.json:

- In the brands.json file, I observed numerous null values for category and category code fields. Does this indicate incomplete data updates for product categories, or are there new products awaiting category updates?
- Furthermore, I noticed duplicate barcodes with different values. Could you clarify if one of these duplicates represents outdated records that haven't been removed from the source yet? We may need to discuss maintaining historical data and adjust our ingestion pipeline accordingly.
- There seem to be data quality issues in the brandCode field, with many null values and some instances where brandcodes are populated as barcodes. I'd appreciate some business context to determine which (barcode or brandcode) will be unique for joining with Fact tables or as a dimension for analysis.

Receipts_fact:

- It appears that the line-level details in the rewardsReceiptItemList field do not sum up to the aggregated values present in the main record. This inconsistency needs investigation to ensure data accuracy.
- Many Nulls in brandcode is resulting in under reporting of business questions asked as can be seen in the queries file, where for the most recent month all the brandcodes are NULL.
- Concerningly, some records in the receipts_fact table have receipt details in the main record but lack details in the inner field. This violates data integrity, indicating missing or inconsistent data.
- Most importantly, some user_ids present in the receipts_fact table do not exist in the users table, highlighting data inconsistency or missing data issues that require resolution.

- How did you discover the data quality issues?

I discovered the data quality issues by conducting an initial analysis using Jupyter Notebook to perform basic checks such as identifying null values, inconsistencies, and potential data quality issues. Subsequently, I further investigated the data using SQL queries to perform aggregation-based and logic-based checks. This multi-step approach allowed me to uncover a variety of data quality issues, including missing values, duplicates, inconsistencies between aggregated and line-level data, and violations of logical constraints.

- What do you need to know to resolve the data quality issues?

Most importantly, we need our team of engineers, along with you and the business team, to come together and gain a thorough understanding of the business context and data sourcing from APIs. This collaborative effort will enable us to address two critical issues: the presence of NULL values in datasets and the determination of unique fields, particularly for brands. I have pointed below the specific areas that require investigation, and acquiring a deeper business context and insights into data sources from Data Engineers will facilitate understanding the logic behind data field population. With your guidance, we can determine the metrics and KPIs necessary for reporting accurately and effectively.

- Insight into Data Population
- Clarification on Missing Values
- Explanation of Duplicates
- Business Context for Brand code: many nulls here resulting in wrong reporting values
- Investigation into Line-Level Details
- Resolution of Null Values and Inconsistencies
- Data Integrity Check

- What other information would you need to help you optimize the data assets you're trying to create?

Based on my analysis of the dataset, I've observed several date fields that have been converted to proper timestamps for flexibility. However, it's essential for us to discuss whether storing timestamps is necessary or if a simpler date format (dd-mm-yyyy) would suffice, depending on the reporting granularity required for analytics.

Additionally, considering the presence of multiple time columns, I propose the implementation of a separate time dimension table. This table would store fields such as calendar year, month, quarter, week, day, and others, depending on granularity requirements. Utilizing a time dimension table can optimize query performance by allowing us to populate date IDs in the main tables instead of storing timestamps and performing joins. However, we need to ensure that the

main files contain continuous dates to avoid unnecessary increases in the size of the time dimension due to sparse date populations.

As for data normalization, we could further optimize the data model by normalizing categories and category codes. Understanding the data size and frequency of reporting (daily/weekly/monthly) is crucial for effective data modeling. This information will enable us to optimize queries for analytical processing and determine the required data size to handle daily. Here getting to know the data growth over time will be an important factor.

Regarding data accuracy, we need clarification on whether stakeholders require 100% data accuracy (non-nulls) or if they are tolerant of slight data mismatches (e.g., 0.01 to 0.1%). Understanding scalability requirements is also essential, as it will inform decisions on storage provisioning and whether to opt for on-premises or cloud services.

Conducting test runs will be necessary to determine the full cycle of data refresh and provide stakeholders with an appropriate range of time required for data refresh in front-end dashboards. The dashboard refresh cycle is a critical parameter that needs consideration in our data optimization efforts as the data grows.

These factors will help us ensure scalability, performance, and data integrity as we optimize our data assets to meet evolving business needs.

- What performance and scaling concerns do you anticipate in production and how do you plan to address them?

In production, I anticipate performance and scaling concerns arising from the growing data volume, complex queries, and increasing user demand. To address these challenges, I plan to implement several strategies. Firstly, I'll focus on optimizing database performance by utilizing appropriate indexing techniques to enhance query execution speed, especially for frequently accessed fields and join operations. Additionally, I'll leverage data partitioning mechanisms. Continuous query optimization and pipeline tuning will also be crucial to minimize resource consumption and enhance efficiency. Furthermore, I'll explore both vertical and horizontal scaling options. Robust monitoring and alerting systems will have to be put in place to proactively identify performance issues and resource constraints, enabling timely intervention and optimization.