**** Data Process Implementation Observations & Data Quality Callouts ****

Hi Team,

As part of the preliminary review of the datasets for the current project, I've created a High-Level Design (HLD) with an Entity-Relation Diagram. Below are some key observations and considerations based on the initial data checks:

Key Observations

1. Data Quality Issues:

- **Duplicates in Users Dataset:**
  Duplicate entries were observed in the **`userId`** field after extracting the data from JSON files. This violates primary key uniqueness, which is ***critical to accurately identify and distinguish users***.
- **Null Values in Users Data:**
  Key fields like **`LastLogin`** and **`SignUpSource`** contain null values. This may impact customer-level analysis, particularly for ***source attribution and user activity-related metrics***.
- **Data Recency Checks:**
  Ensuring ***up-to-date data*** ingestion on a defined cadence is critical. The preliminary data file provided did not meet this requirement. Implementing a recency check mechanism will help maintain data integrity over time.

2. Data Transformation:

- We currently receive three datasets: **`Users`**, **`Brands`**, and **`Receipts`**. As part of the design, we propose creating an additional table, **`ReceiptsItemList`**, to catalog item-level details with information such as volume and spend. This will enable detailed brand performance analysis and help refine strategies.

Caveat:
A significant challenge is the poor match rate between **`brandCode`** in the **`Brands`** dataset and items scanned in the **`Receipts`** dataset. Many **`brandCode`** fields are being recorded as null or blank. This issue highlights the need to collaborate with the engineering team to improve receipt scanning accuracy and enhance brand-item matching.

Questions for Discussion

1. **Duplicates in Users Data:**
   - What is the upstream strategy for capturing user data?

    - Should duplicate records be pruned based on a specific business rule (e.g., removing rows that are identical across all columns) or by selecting the latest record based on a predefined sort order?

2. **Data Cadence:**
    - What cadence should we adopt for data ingestion and processing?
    - This will impact the incremental processing strategy and needs to align with the expected user volume and business needs.
3. **Defining Metrics:**
    - Are there additional business metrics or flags that need to be incorporated into the datasets? For example, should we define categorical codes for certain fields like **`Description`**, **`Groups`**, or **`Reason`** to facilitate more granular analysis?
4. **Scalability and Growth Projections:**
    - The current design is adequate for the current volume of users, brand collaborations, and transactions. However, if there are growth projections available, we can work with the infrastructure team to ensure scalability and plan resource allocation and cost distribution accordingly.

Let me know if there’s a good time to discuss these points further or if additional details are required.

Best regards,
Aditya Roy Choudhary