Hi Team,

After diving deep into our datasets, I've uncovered several data quality issues and optimization opportunities that we should address to improve the reliability and scalability of our analytics platform. Below is a detailed overview of my findings, along with some questions and recommendations to guide our next steps.

Key Findings

1. I found several **data quality gaps** in our datasets that need to be addressed:

- **Users Data**:
    - **lastLogin**: 12.5% of records are **missing** this field.
    - **signUpSource**: ~10% of records are **missing** this field; this could limit how effectively we segment our users.
    - 283 **duplicate user records,** potentially causing inflated user counts in metrics.
- **Receipts Data**:
    - Key fields like **purchaseDate** (15% missing) and **totalSpent** (13% missing) have **significant gaps**.
    - I detected **outliers** in key metrics such as:
        - **pointsEarned** (max value is 10,199.8 compared to a median of just 150)
        - **purchasedItemCount** (max value is 689 items, which is around ~45x higher than the 75th percentile)
- **Brands Data**:
    - Missing values in **topBrand** (~50%) and **brandCode** (~20%); this may affect brand-level reporting.

2. I also noticed **relationship integrity risks** between **Receipts** and **Users**:
    - Over 1,119 **userId** entries in the Receipts dataset have no corresponding/matching user records in the Users dataset.
    - This accounts for approximately 25% of the Receipts dataset and raises questions about data completeness during ingestion.

3. There are some field which could be better organized. Below are my **structural observations**:

    - **Nested fields** such as rewardsReceiptItemList require **normalization** to improve querying efficiency.

    - **topBrand**, currently stored as a float, would be more suitable as a **Boolean** for clarity and consistency.

In order to uncover these Data quality issues, I did the following analysis and validation checks:

- Data Type Validation
- Missing Values Analysis
- Duplicate Records Check
- Numerical Outlier Analysis
- Relationship Integrity Validation

To move forward effectively, I also have a few **questions** that I would like to align on.

1. **Handling Missing Values and Duplicates**:

    • For duplicates in Users, should we prioritize the most recent record or define a merge rule?

    • How should we impute missing values in topBrand and signUpSource—is a default value or machine-learning-based imputation preferred?

2. **Validation Logic**:

    • Are there defined thresholds for metrics like pointsEarned and purchasedItemCount to distinguish between valid outliers and data anomalies?

3. **Integration and Mapping**:

    • What processes can ensure all userId fields in Receipts correspond to valid users? Should these records be excluded or enriched during ETL processes?

To address the issues that I outlined above, I have a few **recommendations**:

1. **Data Type and Structure Enhancements**:

    • Enforce strict typing (datetime, Boolean) to minimize type-related inconsistencies.

    • Normalize nested fields like rewardsReceiptItemList into separate tables for faster and more accurate joins.

2. **Automated Data Quality Checks**:

- Implement real-time validation pipelines to flag outliers and nulls during ingestion.
- Build periodic deduplication workflows for user records.

3. **Scaling and Performance Optimization**:

- Propose partitioning large datasets (e.g., Receipts by purchaseDate) to enhance query performance.
- Anticipate growth by validating storage efficiency for nested JSON fields and consider flattening where appropriate.

Lastly, I have some feedback regarding **next steps** and support needed:

1. Can we establish clear business rules for acceptable ranges in numerical metrics and the handling of unlinked references?

2. It would be helpful to validate our scaling assumptions, such as anticipated data growth and query concurrency levels, to ensure the infrastructure can support future workloads.

3. I suggest scheduling a strategy session to align on priorities and finalize these approaches.

Let me know if we can schedule a sync to discuss these findings further and align on next steps. I'd love to hear your thoughts and input!

Best regards,
Raghav Mehta