Hi,

I've completed an initial analysis of our receipt rewards data system and would like to share some key findings and questions. This will help us ensure we're building a robust and reliable system for our users.

Questions About Our Data:

1. What are our business rules for handling:

* Missing item descriptions in receipts?
* Receipt amounts stored in different formats (numeric vs text)?
* Receipts with future scan dates?
* Receipts that can't be matched to a user?

2. Should we exclude or attempt to fix records with missing critical data?

3. Is there a reason why some receipt totals are stored as text rather than numbers?

4. Which fields are most critical for your reporting and analysis needs?

**How We Found Data Quality Issues:**

Our comprehensive analysis revealed several key areas needing attention:

* Critical fields have missing values (item descriptions, barcodes)
* Receipt totals and points are inconsistently stored (sometimes as text, sometimes as numbers)
* Some receipts show future scan dates
* Over 1,100 receipts can't be matched to any user account
* Nearly 300 users appear to have duplicate accounts
* Many brands are missing category information

**To Fix These Issues, We Need to Know:**

* + Which fields should be mandatory when users submit receipts?
  + What are the valid ranges for receipt totals and item counts?
  + How should we standardize the format of numeric values?
  + What's our process for receipts that get flagged or rejected?
  + Should we implement automatic data format validation?
  + How should we handle records with incomplete information - clean, flag, or remove them?
  + Are there additional user identifiers we could use to reduce orphaned receipts?

**To Optimize Our System, We'd Need:**

* Expected number of daily receipt submissions
* Peak periods (like holidays) when we might see higher submission volumes
* How long we need to keep receipt data
* Which fields are most frequently used in reports and analyses
* Any compliance requirements for storing receipt data
* Details about user activity tracking capabilities
* Information about any additional identifiers that could help match receipts to users

**Performance & Scaling Plans:**

We anticipate these challenges as we grow:

* Higher volume of receipts during peak shopping periods
* Growing storage needs as we accumulate more data
* Need for quick response times when users submit receipts
* Potential slowdown in query performance as data grows

**To address these, we propose:**

* Setting up automatic scaling during high-traffic periods
* Creating a smart archiving system for older data
* Implementing faster processing for receipt submissions
* Building real-time monitoring to catch issues early
* Adding data validation to prevent format inconsistencies
* Creating indexes for frequently queried fields (scan dates, status, user IDs)
* Partitioning receipt data by month to maintain performance
* Implementing batch processing for large data updates

**Implementation Priorities:**

1. Standardize data formats (especially for dates and numeric values)

2. Establish clear rules for handling incomplete data

3. Improve user-receipt matching system

4. Set up performance optimization measures

5. Create automated data quality monitoring

I appreciate your time in reviewing these findings. Please feel free to reach out if you need any clarification or have additional questions.

Best regards,

Rahul Pillarisetty