Communication with stakeholders

Subject: Data Quality Findings
Hi Product/Business Leader,

I have been running data quality checks by developing a python script that runs multiple SQL queries to check for data in-consistencies. As part of that exercise, I identified key data quality issues that may impact analysis and reporting. Your input is needed to determine the next steps.

Key Issues:

1. Missing brandCode values – Critical for mapping receipt items with brands. Should we enforce brandCode as a not-null field to ensure accuracy?
2. Duplicate barcodes – Can the same barcode be linked to multiple products, or should it be unique? Identified duplicate barcode values in our dataset.
3. Receipts with totalSpent = 0.0 – Are these valid cases (e.g., refunds/cancelled transactions), or do they indicate any missing data?
4. Missing itemPrice and finalPrice – Could this impact the pointsearned/bonuspointsearned calculations. Should these fields be not null?

Recommendations to address the issues:

1. Implement data validation rules (NOT NULL, UNIQUE) to the required data fields to prevent future inconsistencies.
2. Confirm if the brandCode is critical for mapping or alternately we can use brandId.
3. Clarification on barcode uniqueness to determine if duplicates are expected.
4. Validate if the fields like totalSpend, ItemPrice, FinalPrice can be null or 0.

Production scaling concerns:
As the production data gets scaled, having too many null or missing values may slow down the queries. We need to make sure to select proper secondary indexes to optimize query performance.
Check for possibilities to implement data validation rules at the source to filter out the missing, invalid and duplicate values.

Please let me know for any questions and thoughts.
I can schedule a quick sync up call if required.
Thank you,
Deepeka