Communicate with Stakeholders

Subject: Comprehensive Report on Data Quality Issues and Next Steps

Hi [Product Leader],

I wanted to share some insights and findings from our recent data quality evaluation of the Fetch Rewards relational database. As we analyze the data, we've identified a few important data quality issues that we need to address to ensure data accuracy and reliability.

Key Questions Regarding the Data:

a. What are the data quality issues?

- Date Inconsistencies in Receipts:
    - We discovered **13 instances** where the purchaseDate (the date the purchase occurred) is later than the dateScanned (the date the receipt was scanned). This is logically incorrect, as the purchase should always precede or coincide with the scan.
    - Additionally, we identified **7 cases** where pointsAwardedDate (the date when rewards points were credited) occurs before the purchaseDate, which violates the expected sequence of events in our system.
- Date Inconsistencies in Brands:
    - We discovered 14 instances where the same barcode is linked to multiple products, which violates the principle that barcodes should uniquely identify a single product.
    - For example, barcode 511111504788 is associated with both a test product labeled as "test", and a real product called The Pioneer Woman. This inconsistency can lead to issues in product identification, inventory management, and sales analytics.

b. Questions

- Why are there instances where the purchaseDate is later than the dateScanned?
- What causes points to be awarded before the purchase is completed (pointsAwardedDate before purchaseDate)?
- Why are there missing or invalid categoryCode and topBrand values in the brand data?

- Are there system delays or synchronization issues impacting the logging of real-time data?
- What process can ensure more accurate and complete data entry across all fields?

Discovery of Data Quality Issues

- By comparing the various date fields within the *Receipts* table, we analyzed the relationships between the purchaseDate, dateScanned, and pointsAwardedDate. This analysis highlighted instances where the sequence of events didn't align with business logic.
- In the *Brands* table, we examined the product information associated with each barcode. This examination revealed cases where a single barcode was linked to different product names, identifying inconsistencies in product identification.

What does this mean for our data?

These inconsistencies suggest potential data entry or synchronization issues, either during the data logging process or due to system delays in capturing real-time events. This can undermine the reliability of our data and, if not corrected, may lead to inaccurate analytics or customer dissatisfaction if reward points are not processed correctly.

Requirements for Resolving Data Issues

To effectively address these data quality concerns, we need:

1. Cross-functional Collaboration:
    - We will need to work closely with the engineering team to investigate potential data sync or logging issues that are causing these date-related inconsistencies. A review of how our systems handle real-time data across different services might help pinpoint the root cause.
2. Validation and Error Handling Logic:
    - We should consider implementing stronger validation checks in the data ingestion pipeline to catch and flag these discrepancies as they occur. For instance, an automated alert when a purchase date is after the scan date would allow us to address the issue in real-time.
3. Additional Metadata and Process Clarification:
    - Understanding the end-to-end flow of how receipts are scanned, processed, and rewarded will help us isolate where the breakdowns might be happening.

For example, knowing the exact timing of when rewards points are processed after receipt scanning could clarify if this is a system lag issue or something more fundamental.

4. Data Enrichment and Cleansing:
    - For brand-related data, we may need to enrich the Brands table by cross-referencing external sources to fill in missing categoryCode values and correct invalid entries. An automated data cleansing process might also help ensure more consistency going forward.

Additional Information for Data Asset Optimization

- Operational Context:
  It would be helpful to know if there are any system performance issues or downtimes that may have impacted data logging during specific periods, which could explain some of the discrepancies.
- User Behavior Data:
  Understanding patterns in how users interact with our system (e.g., timing between purchases and receipt scanning) could provide additional context for why these issues are happening and how widespread they might be.

Anticipating Performance and Scaling Challenges

As we address these issues, we also need to keep performance and scaling in mind, particularly because Fetch Rewards is growing rapidly. The following concerns and strategies are worth noting:

- Increased Volume of Data:
    - With a larger user base, we anticipate higher volumes of receipts, purchases, and rewards data being processed. To handle this, we need to ensure that our database queries remain optimized. For example, indexing fields like purchaseDate, dateScanned, and user_id can improve query performance during large-scale analysis.
- Real-Time Data Validation:
    - To prevent the recurrence of these data quality issues in production, we plan to implement real-time validation at the data entry point. This will catch inconsistencies (e.g., incorrect date sequences) immediately, reducing the chances of erroneous data entering the system.

- Error Monitoring and Alerts:
    - As we optimize our processes, setting up automated monitoring systems that track the integrity of data flows will help. For instance, we can establish an alert mechanism to flag any future instances of dates being recorded out of sequence or missing key fields (like brand category).

Next Steps:

Our immediate plan is to:

- Collaborate with the engineering team to investigate system-related causes for these inconsistencies.
- Propose enhancements to our data ingestion pipeline to prevent further data quality issues.
- Initiate a data enrichment process to clean and complete brand-related information.

We will keep you updated as we make progress on resolving these issues.

Please let me know if you have any questions or need additional details.

Best regards,
Surbhi Dilip Bhor
Analytics Engineer, Fetch Rewards