Understanding and Addressing Data Quality Issues

Disclaimer: The following breakdown is structured to highlight key data quality challenges, how they were discovered, and what is needed to resolve them. Since data engineering can get pretty technical, I have framed the discussion in a way that aligns with business goals that are more relatable to a business leader —ensuring accurate insights, reliable performance, and scalability as Fetch grows.

We are primarily looking at **four key entities**—**Receipts, Brands, Items, and Users**—since they form the foundation of our data ecosystem.

What questions do you have about the data?

Before taking action, we need to confirm a few assumptions to ensure we align with business expectations.

Receipts (Transaction Data)

- Should I track the **full history** of a receipt's status, or is the latest status enough? This impacts how receipt validation trends are analyzed - in the sample data currently, it seems we are tracking the last status of a receipt
- Sometimes, receipts are missing item details—should those be considered valid, or are they incomplete transactions
- When bonus points aren't awarded, there's often no reason recorded. Should we infer a default reason, or is that a problem we need to solve upstream?

Brands (Product Attribution)

- Are brand codes supposed to be **unique**, or can two brands technically share the same code?
- I noticed test brands (e.g., names starting with "@test")—are those to be intentionally included in reporting, I assume not?
- How should the missing brand assignments be handled? If an item has no associated brand, should it be excluded from analysis or attempted to assign one?

Items (Products in Transactions)

- Should **barcodes be globally unique** across all brands, or do they sometimes get reused by different brands? This changes how we match items to brands correctly.
- What is the **source of truth for pricing**? I see different price fields (finalPrice, discountedPrice, itemPrice), and they don't always match.
- What should be done when an item appears in a receipt but is not found in our product database? Should it be flagged, or is that expected behavior?

Users (Spenders and Account Management)

- Can a single person have **multiple account IDs**, or is that a data issue? There are cases where this happens, possibly due to data entry errors.
- What should happen to **receipts from inactive users**? Do they still count toward rewards and reporting, or should they be filtered out?
- If accounts get merged, should I **reassign past transactions** to the new account, or leave them as they were?
- Should **fetch-staff** and **consumer** both be considered valid for reporting?

How did you discover the data quality issues?

I discovered the data quality issues during a detailed audit of the data, where I closely examined different datasets and their relationships.
Here's what stood out:

Receipts (Transaction Data) - performed data checks to compare the receipt data with the product database. By cross-referencing barcodes and looking for missing information (like brand assignments or reasons for zero bonus points), I identified inconsistencies

- A huge mismatch between barcodes on receipts and our product database—only 82 out of 6,941 barcodes matched.
- Brand assignments were missing for over 6,300 receipts, meaning I couldn't correctly attribute purchases.
- Nearly 50% of receipts with zero bonus points had no recorded reason—this makes it difficult to explain why points are not awarded.

Brands (Product Attribution) - reviewed the brand codes and categories in the dataset, which involved checking for duplicates and test data. By looking for discrepancies in how brands were categorized and assigned, I spotted issues like brands with multiple codes or missing category assignments

- Some brands were incorrectly categorized under the wrong parent brands, affecting the accuracy of reporting at the parent brand level.
- A few brands had outdated data, with historical information not being properly updated, leading to discrepancies in trend analysis.

Items (Products in Transactions) - performed a validation against the item master database and found items listed in receipts that didn't exist in the master list. Additionally, I cross-checked prices and quantities to uncover mismatches, including duplicated barcodes.

- Certain items had mismatched unit prices, raising concerns about the accuracy of cost and revenue reporting.
- Some items were linked to incorrect product categories, which made it difficult to analyze sales trends by the correct product types.

Users (Spenders and Account Management) - analyzed user accounts and transaction histories. By reviewing account IDs and transaction logs, I found issues such as multiple accounts linked to a single spender and inactive users still showing up in reward calculations.

- Several users had no recorded transaction history, though their accounts were active, which could be indicative of missing or incorrectly processed data.
- Some users had outdated contact details in the system, affecting customer communication and potential marketing outreach

What do you need to know to resolve the data quality issues?

To fix these issues, we need more clarity on business rules and expected behavior. Some of these are also covered in the first question.

Receipts (Transaction Data)

- Some receipts had no recorded date, making it impossible to track when purchases occurred and potentially affecting reporting timelines.
- There were receipts with missing item quantities, which led to incomplete transaction records and inaccurate sales calculations.
- If a barcode or brand is missing from a receipt, should we estimate it based on other data, or leave it blank?
- Do we need to store the **full history** of receipt status changes for auditing, or just keep the latest update?

Brands (Product Attribution)

- Should test brands be **completely excluded**, or do they serve a purpose for analytics?
- Can we rely on a **standardized brand mapping** to correct inconsistencies?

Items (Products in Transactions)

- Should we **enforce barcode uniqueness**, or is it normal for the same barcode to appear under different brands?
- Which price field should be considered the most accurate for reporting?

Users (Spenders and Account Management)

- Should we merge multiple account IDs for a single person, or keep them separate?
- How should we handle transactions from inactive users—should they be removed, or just flagged?

What other information would help optimize the data assets?

- **Receipts:** A reference table linking valid barcodes to items and brands.
- **Brands:** A **canonical brand list** that clearly maps brands to categories.
- **Items:** A **master item table** that standardizes barcode-brand relationships.
- **Users:** A clear **policy on handling duplicate accounts** and inactive users and employee/consumers

What performance and scaling concerns do you anticipate in production, and how do you plan to address them?

If we don't proactively address these challenges, we could face **slow reporting, incorrect insights, and difficulties scaling as data volume grows**.
Here's how we plan to mitigate risks:

Receipts (Transaction Data)

- Since receipt volume will grow rapidly, I suggest to **partition data by date** to speed up queries.
- Barcode lookups are expensive—we can improve performance by **caching product data** instead of reloading it every time.
- Instead of rewriting the receipt table each time, we can implement **incremental updates** to track changes more efficiently.

Brands (Product Attribution)

- Brand mappings **change frequently**, so a **versioned mapping table** (SCD Type 2) must be maintained for historical tracking.
- Since brand-category relationships are high-cardinality, we can optimize lookups with **indexed joins** to speed up reporting.

Items (Products in Transactions)

- To prevent mismatches, we need to enforce **barcode uniqueness constraints** or implement a deduplication process.
- Price discrepancies will be flagged via a **reconciliation process** that detects and corrects inconsistencies.

Users (Spenders and Account Management)

- If we merge accounts, we will need a **controlled process to consolidate user history** while maintaining accuracy.
- Inactive users can be filtered via a **materialized view**, so queries don't slow down when processing large datasets.