Gmail

Sashanth Embakula <embakulas@gmail.com>

Update on Data Quality Assessment and Data Model Creation for Fetch Rewards Project

Sashanth Embakula <embakulas@gmail.com> Tue, May 28, 2024 at 4:15 PM
To: Sashanth Embakula <embakulas@gmail.com>

Hi,

I hope you're doing well. I wanted to provide you with a detailed update on our recent efforts concerning the Fetch Rewards analytics project. We've made significant progress in assessing the data quality and structuring a relational data model. Below are the specifics of our findings and the steps we've taken:

Data Quality Issues Identified:
Missing or Null Values:
Users Table: We identified missing values in several columns, including state, createdDate, and lastLogin.
Receipts Table: Fields such as bonusPointsEarnedReason and rewardsReceiptStatus had null values.
ReceiptItems Table: Several entries had missing values in columns like description and userFlaggedBarcode.

Inconsistent Data Types:
We encountered inconsistencies in date columns across different tables. For example, createdDate and lastLogin in the Users table, and various date-related columns in the Receipts table were not in a consistent format.

Foreign Key Violations:
Some userId entries in the Receipts table did not have corresponding records in the Users table.

Invalid Date Formats: Ensured all date fields were correctly formatted.
Logical Date Consistency: Verified that createDate was not later than modifyDate, pointsAwardedDate, or finishedDate in the Receipts table.
Negative Values: Checked that numerical fields such as pointsEarned, totalSpent, itemPrice, and finalPrice were non-negative.
Boolean Consistency: Ensured boolean fields contained only True or False values.
Natural Key Uniqueness: Verified uniqueness in natural keys such as state in Users, barcode in Brands, and partnerItemId in ReceiptItems.

Steps Taken to Address These Issues:
Data Cleaning and Validation:
We filled in missing values where appropriate and ensured that all fields adhered to the expected data types.
For the Brands table, we resolved the issue of non-unique barcodes by creating a separate mapping table (BarcodeBrandMapping) to handle duplicates.

Default Entries:
To handle missing foreign key references, we created default user entries in the Users table and default brand entries in the Brands table.

Findings:
Data Integrity Improvements: The structured model and data cleaning efforts have significantly improved data integrity and consistency.
Foreign Key Consistency: The introduction of default entries and careful validation ensured that all foreign key relationships are now intact and we did not lose any data.
Enhanced Data Usability: The cleaned and validated data is now more reliable and ready for further analysis and reporting.
Next Steps:

We will continue to monitor and address any remaining data quality issues. Further, refine the data model based on ongoing insights and requirements.

Please let me know if you have any questions or need further details on any of the points mentioned above.

--
Sashanth Embakula
+1 8136062755
embakulas@gmail.com
https://www.linkedin.com/in/sashanth-embakula-b1b032143/
https://github.com/embakulas