(Note: I'm assuming that this is an email directed to a number of people asking questions about the data)

Subject: Questions about the Inbound Data

Hello Everyone,

First off, I appreciate everyone's support working through this data. All of your input will be extremely helpful in understanding how we can best work with this customer data.

I performed a large amount of analysis on the datasets you gave me, here's my notes, findings, and questions for each:

- Overview questions:
    - Where are we getting these datasets from and how are they being sent to us? If we know the source of the datasets, maybe we can work directly with them to handle some of the data and formatting difficulties together.
    - What is the date format being sent and can we work with the source development team to change to a more standard MM-DD-YYYY format? If not, can they send us a legend on how to read it so we can format it on our end?
        - EX: Here's a createdDate value from Users: 1418998882381. I don't know what this means.

- Receipts:
    - Formatting Difficulties:
        - First off, this dataset was difficult to properly analyze due to how the individual items scanned show on a receipt. The column 'rewardsReceiptItemList' could be its own dataset. Is there any way we could get that sent in a better format? That would make it so we don't feel as if we're fighting the formatting to get at the data.
        - Comparing the possible columns in 'rewardsReceiptItemList' to the columns in the Receipts dataset acutely shows the difficulty:
            - Receipts: 15
            - RewardsReceiptItemList: 34
        - That's a lot of data embedded in a single column of a dataset.
    - Performance:
        - My only issue with performance is the embedded data in rewardsReceiptItemList. Processing the large amount of data within data will be slow.

        - Having this data sent in an easier-to-read format will help with both difficulty (slower to develop) and processing time (slower to run).
        - If we can't have it done on their end, we could possibly store receipts in something like an AWS S3 bucket and have it be processed there before moving it into our data warehouse. We would have to pay for the S3 storage and processing, but it would be a possible workaround.
    - Missing Data:
        - There's a large amount of missing data here. The dataset contains 1119 rows of data, and the below screenshot shows the columns with missing data and their amounts.

```
_id                       0
bonusPointsEarned       575
bonusPointsEarnedReason 575
createDate                0
dateScanned               0
finishedDate            551
modifyDate                0
pointsAwardedDate       582
pointsEarned            510
purchaseDate            448
purchasedItemCount      484
rewardsReceiptItemList  440
rewardsReceiptStatus      0
totalSpent              435
userId                    0
```

        - From my findings, the missing data seems to highly correlate to receipts that are earning bonus points - most often when a certain number of receipts have been scanned.
        - In order to avoid creating extra receipts, could we combine regular receipts with bonus points receipts? The source seems to recognize when a bonus point receipt should get created, so I assume modifying their end to tack on the bonus points to the receipt that created the bonus points event is possible.
- Brands:
    - The biggest issue I have with the Brands dataset is between the category & categoryCode columns.
        - The current usage of these columns seem redundant, is there a need to have categoryCode be a repeat of each category?
            - EX:

Category	CategoryCode
Baking	BAKING
Candy & Sweets	CANDY_AND_SWEETS

        - The majority of the missing data in this dataset comes from categoryCode, why is categoryCode missing from so many records when category isn't? They're repeated, so it's strange.
        - My suggestion is turn categoryCode into a short identifier, and then make a smaller table that maps from categoryCode to category. This will avoid redundancy, and can cut down on some storage.
            - EX:

CategoryCode	Category
BAKE	BAKING
CAND	CANDY_AND_SWEETS

- Users:
    - Can someone check if there was a problem with the dataset we were sent for Users? The data we were sent has a large number of duplicates in the _id column. That column should have a unique value tied to every record, and we cannot create more than 1 record in the warehouse with a _id sequence.
    - I could be wrong, but I believe we were sent test data. I believe this because:
        - The majority of the records with _id values were tied to 'fetch-staff' roles.
        - Same correlation for the records with missing data in other fields, roughly ~80% of records with missing data were 'fetch-staff' roles.
    - Even if this was test data, we should expect that each record has a unique _id value, please have the source development team look at the reason why we are getting any repeated _id values. We can easily work with them to figure out the issue.

Let's have a meeting with the key figures sometime in the next week so we can discuss each item. It'll be easier to get everyone on the same page that way. In the meantime, please let me know if you have any questions, comments, notes, etc. about anything I listed above.

Thank you again for all your support,
Gregory Schamberger