

Dear Stakeholders,

Wanted to take a moment to touch base on the latest data you provided for ingestion to our data warehouse. I noticed a few things that seemed off;

1 The users file has a lot of duplicate records. We can handle the deduplication on our side if needed but I just wanted to bring this to your attention in case it's a larger issue with the json file creation process on your side.

2. The receipts file you provided does not properly reference the brands file when associating items to brands. We are provided a brandcode on the receipt, however this brandcode is not unique in the brands file. As a result, there are cases where it is unclear from the data on the receipts file exactly which brand should be associated with the item purchased. An example of this is Huggies which appears twice in the brands file under different brandids

3. There are many brandcodes in the brands file that are null or blank. Do you know why this is? We will need to have this information filled in correctly if we want to see a clear picture of brand performance. Please let me know if I can be of help with this process.

4. I see that most receipts have no brandcode information associated with them at all, and those that do are mostly from purchases made in January 2021. Did something change during this time? It is critical that we capture the brand information on every item purchased so that we can paint a complete picture of brand performance for the quarterly meeting next Thursday. Please let me know if records missing this tag can be reprocessed with this tag included so I can update the data warehouse on my side with this informations

As we continue to scale up our operations, it would also be helpful to know what receipt list values are guaranteed to be on a receipt and which are optional. This will allow me to further normalize required fields on the receipts table which will free up disk space and improve data load speeds. Fields 'description and 'rewardsGroup at first glance might be a good candidate for this.

Finally. it would be great to run incremental refreshes on the receipts data to minimize the amount of strain on the database during daily loads. Please let me know if any of this data can be restated, and if so what field I can use to key off of to indicate a recent restatement. I see the field 'modifyDate' which perhaps is used for this purpose. If records can be restated, I will adjust my daily load process accordingly to account for a merging of updated and new records with records that have already been pulled in the past. If records cannot be restated, then I will only pull in new records every day and keep what has already been pulled as is

Please let me know if you have any questions or concerns about the above points.

Thanks you for interest and collaboration!

Best Regards,

Ismet Yavas, Data Analyst