Hello team,

I was recently given an assignment focused on the receipts, brands, and users data that we are collecting. I have several questions and concerns regarding the data provided.

Linking Receipts to Brands

One main question I have is surrounding the best way to link receipts to brands. I ended up using the brand code field, but there was significant data loss since many receipt records do not have a brand code. There are also many records in the brands data that do not have a brand code. I used this field as it was the only clear common field in the two tables. If there is another field that I am not aware of, please let me know and so I can update my joining criteria.

User Data

I also have concerns regarding the user data. There are several user IDs in the receipt data which do not occur in the users data. We will not be able to access user data for those records. There are also users which have duplicate records in the users data. It is my understanding that the _id field in the users table should be unique. There may be an issue with how the user data is being collected.

Brand Data

Regarding brands, I found two brand codes (GOODNITES and HUGGIES) which are duplicated. There are also many brands with no brand code in the brand table, which is part of the joining issue I mentioned above. I would like to find out if the data should have blanks in that field or if it is required. It would benefit me to know more about how the data is being collected.

Once the above concerns are addressed, I will be able to more effectively design the tables, for example by adding indices. This will allow for enhancing the performance of the data model. Thank you.

Regards,

William Davis

Analytics Engineer