Hello,

While doing a preliminary exploration of the receipts, users and brands json files, following data quality issues were found in the datasets

- About 50% of records are duplicated in the users table
- About 50% of values in categoryCode column is missing
- Non-homogeneity in values for brandCode column
- There are duplicate barcodes found in the brands dataframe which pose a data quality issue.
- there is no information about some items and their corresponding barcodes. If the receipts had any partner purchases, that information is lost as well and consequently so is the partner revenue.

As receipts grow into billions, the team anticipates performance and scaling concerns which can be addressed by writing efficient SQL queries to fetch data, indexing the tables as well as adding nodes to scale as demand for storage grows.

Regards