Hi Product Team,

I went through the data files related to Users, Brands and Receipts. I have put into place a basic design to host this data in Snowflake. Upon performing preliminary scan and some data quality checks on the data, I have a couple of things that I would like to get further information on, which are as follows:

1. **Brands data:**
   a. In this data I saw some category codes and brand codes that were empty.
   b. There were many "test" brands. Are these entries valid?
   c. There are some brand codes that are populated with bar codes. Is it safe to assume that the brand code should be based on the "Name"? If so, is there a pattern that I can use to extract the Brand code.
   d. Some barcodes have multiple products associated with them.

BRAND_ID	BARCODE	CATEGORY	CATEGORYCODE	CPG_ID_OID	CPG_REF	NAME
5c45f91b87ff3552f950f027	511111204923	Grocery	null	5c45f8b087ff3552f950f026	Cogs	Brand1
5d6027f46d5f3b23d1bc7906	511111204923	Snacks	null	5332f5fbe4b03c9a25efd0ba	Cogs	CHESTER'S

NAME	TOPBRAND	BRANDCODE
MorningStar	null	null
Calumet	false	CALUMET
Entertainment Weekly	null	511111205012 ←
AUNT JEMIMA Syrup	false	AUNT JEMIMA SYRUP
Molson Canadian	false	MOLSON
Lotrimin®	false	LOTRIMIN
test brand @1597342520277	null →	TEST BRANDCODE @1597342520277
test brand @1605535049181	false	null ←
ST. IVES	false	ST IVES

2. **Receipts data:**
   a. The following observations are similar and need to be addressed. It would be helpful to understand how the status of receipt interacts with the following.
      i. The points earned do not match the total points that were awarded based on the reward receipt item list for a particular receipt.
      ii. The purchased item count do not match the number of items purchased based on the reward receipt item list for a particular receipt.
   b. Most barcodes listed in the rewards receipt item list are not present in the Brands data. Can the missing ones be added to the Brands data based on the Bar codes and any description fields related to it? If so, I would be needing your team's assistance in defining the categories, names and brand codes.
   c. Having all the brands related data in the brands table could help in avoiding repetition of data in the rewards receipt item list table.

RECEIPT_ID	PURCHASEDITEMCOUNT	QUANTITYPURCHASED	QTY_FLAGGED	TOTAL_QTY	DATA_ISSUE
5fa5ad370a720f05ef000089	11	11	10	21	MISMATCH
5ff473b20a720f05230005b7	5	5	0	5	null
5ff5d1fd0a720f05230005de	9	9	0	9	null
5ff473ad0a7214ada10005c3	1	1	0	1	null
5ff4ce690a7214ada10005e2	5	5	5	10	MISMATCH

3. **Users data:**
    a. Some users have duplicate entries. I would like to get your consent before I drop the duplicates. This could help save some storage costs.
    b. Some users who are active don't have a last login time. In these cases, can create date be considered for last login time?
    c. There are some users without information on their sign-up source. Could you provide me with possible options for this? Also, what would be a default value if there is no information provided.
    d. The documentation says that the user role is a constant value of 'CONSUMER' but there are some in the data that say 'FETCH-STAFF'. Can everything be replaced to 'CONSUMER'?

USER_ID	ACTIVE	CREATEDDATE	LASTLOGIN	USER_ROLE	SIGNUPSOURCE	STATE
5fa32b4d898c7a11a6bcebce	true	2020-11-04 22:29:33.309	2021-03-04 07:21:58.047	fetch-staff	Google	AL
5fa41775898c7a11a6bcef3e	true	2020-11-05 15:17:09.396	2021-03-04 16:02:02.026	fetch-staff	Email	null
5fa41775898c7a11a6bcef3e	true	2020-11-05 15:17:09.396	2021-03-04 16:02:02.026	fetch-staff	Email	null
5fa41775898c7a11a6bcef3e	true	2020-11-05 15:17:09.396	2021-03-04 16:02:02.026	fetch-staff	Email	null
5fa41775898c7a11a6bcef3e	true	2020-11-05 15:17:09.396	2021-03-04 16:02:02.026	fetch-staff	Email	null
5fa41775898c7a11a6bcef3e	true	2020-11-05 15:17:09.396	2021-03-04 16:02:02.026	fetch-staff	Email	null
5964eb07e4b03efd0c0f267b	true	2017-07-11 15:13:11.771	2021-03-04 19:07:49.770	fetch-staff	null	IL
5fa41775898c7a11a6bcef3e	true	2020-11-05 15:17:09.396	2021-03-04 16:02:02.026	fetch-staff	Email	null

Performance optimization:

Currently, I have created 4 tables that act as the source of truth for all our analytics. I have created one table each for Brands and Users. I have created two tables for Receipts data, but I feel it can be split into 4 tables with one for higher level transactions, item level transactions, points/rewards and review flagging. This will help us keep the tables narrow and improve query performance for analytics.

I am also, creating tables that houses information that is frequently used for analytics so that we have better query performance and reduce development time (no need for re-writing the logic). An example of this is T3_USER_RECIPT_ITEM table where I have information on receipt id, user id, dates, brand name and prices.

Please let me know of your thoughts. We can setup a meeting to discuss further on this.

Regards

Vidaan Shankar