Proposition: Modify existing pipeline to include new cleaning steps and drop unnecessary fields.

Process

1. Python file executed on EC2 reads JSONL files from S3 to the local machine
2. On the local machine, desired columns are selected and uploaded to RDS
    a. Check if primary key already exists when adding
3. When initially added to the table, modify the recently added rows (using an added time field??):
    a. Drop the oldest products, which are not required
    b. Other cleaning here, such as negative values for price (in that case, the price would be set to null)

Modified schemas

Review

Old

asin	user_id	timestamp	verified purchase	helpful_votes	text	title	rating

New

ID	parent_asin	text	title	rating

Explanation: Asin, user_id, and timestamp do not provide any meaningful information to the user. Verified purchase and helpful votes could be added at a later time, but for now we have not included them because we believe the risk of overcomplexity they present is not worth the benefit they might offer.

Product

Old

parent_asin	main_category	title	average_rating	rating_number	description	price	store	features

New

parent_asin	main_category	title	rating_number	description	features	price

Explanation: Dropped average_rating: it can be calculated using a SELECT query on the review table. Dropped store: we do not consider this useful information to our app's target user (we are providing them an Amazon link, not directions to a store).

Child

Old

parent_asin	asin

New
**REMOVED

Explanation: asin most likely does not add value to our dataset, as the most common subproduct descriptors are now present in details (color, size), and considering others would potentially overcomplicate our pipeline.

Categories

Old

parent_asin	category

New
**REMOVED

Explanation: categories is already represented in the Product table

Bought Together

Unchanged

parent_asin_1	parent_asin_2

Explanation: *still a self-relation*

Details

Old
**DID NOT EXIST

New

parent_asin	color	size	material	brand	style	dimensions	upc	manufacturer

Explanation: *Need details to provide index #3 for Pinecone vector embedding (done by row of this table)*