

Amazon Reviews DWH:

Assignment

Design and implement Data Warehouse (DWH) with corresponding Extract/Transform/Load (ETL) scripts to load the DWH.

Data

The description of the data can be found [here](#).

You can find the data [here](#). You may use the smaller dataset for categories (Clothing, Shoes and Jewelry).

- reviews_Clothing_Shoes_and_Jewelry_5.csv
- metadata_category_clothing_shoes_and_jewelry_only.csv

Business Requirements

Solution must be able to query the following metrics:

- Average review rating per category per month.
- Analysis of review rating per brand per month.

Technical Requirements

Must have:

- A DWH Model (ER Diagram) for showing relationships between tables. Preferred data modelling technique is Kimball Dimensional Modelling (i.e. [dim and fact tables](#)).
- Data Ingestion and Loading Scripts for the designed tables.
- Description of Data Quality Checks which can be used. (hint: data cleaning steps)
- If a product has multiple categories, you can use the first category.

Good to have:

- DDL statements for the tables.
- Implementation of Data Quality Checks
- Loads data incrementally and on a high-volume basis (10x bigger) more than once in a day.
- Remarks / Comments of those parts (optimization) where you could speed up execution and / or save network traffic.

You can use whatever tools/languages (preferably open source) in order to solve the questions. Although the above assignment may feel specific and restrictive, please feel free to use your creativity and flexibility.

Feel free to ask any questions you may have (It's appreciated).