University Of St. Thomas

A purple shield with white letters

Description automatically generated

# Course 732 - Data Stores and Feature Design

# Project – Amazon Product Database

Project by – Mugdha Dixit

1. Using the marketplace locate one or more free data sets that interest you. (5 pts)

* **TRAJECT\_DATA\_AMAZON\_PRODUCT\_RESULTS\_DATA**

1. Worksheet #1. Build a curation layer. (20 pts)

# Worksheet 1

## Created new schema named Amazon Schema

A computer code with blue and black text

Description automatically generated

## Creating tables and adding data into it.

### Created table named ‘CUR\_Search\_Product’ and added data from

A screenshot of a computer code

Description automatically generated

### Pulling data from ‘Amazon\_Product’ database to the ‘CUR\_Search\_Product’ table.

A close-up of a computer screen

Description automatically generated

### Creating table ‘Category’ and adding data to the table.

A table with numbers and letters

Description automatically generated with medium confidence

A close-up of a computer screen

Description automatically generated

### Creating table ‘Buyer’ and adding data to the table.

A close-up of a table

Description automatically generated

A close-up of a computer code

Description automatically generated

## Enhancing the data with additional fields.

### Adding new column to the ‘Product’ table and inserting values, 'High', 'Medium' and 'Low' based on ‘P\_rating’.

A black text on a white background

Description automatically generated

A close-up of a computer screen

Description automatically generated

### Identifying missing values in Product table



### Adding new column 'has\_coupon' to ‘Product’ table and adding Yes/No indicators.

A close up of words

Description automatically generated

### Calculating the average rating from non-null values in the ‘B\_ratings’ column.



### Identifying the duplicate values.

A close-up of a computer screen

Description automatically generated

### Finding number of null values in ‘CUR\_Buyer’ table.



### Updating null values in ‘B\_ratings’ with the average rating

A screenshot of a computer program

Description automatically generated

### Joining CUR\_Search\_Product with CUR\_Category to get details from both tables.

A black text on a white background

Description automatically generated

# Worksheet 2

## Build an aggregation layer.

### Creating new schema named 'AGG\_Schema'.

A close up of words

Description automatically generated

## Creating different views using different types of aggregation.

### Counting number of products with Coupons from CUR\_Search\_Product table

A close-up of a computer screen

Description automatically generated

### Calculating the total purchase amount for each customer from ‘CUR\_Buyer’ table.

A black text on a white background

Description automatically generated

### Counting products in each category from CUR\_Category table

A close up of text

Description automatically generated

### Finding the maximum rating among products and the corresponding product title using CUR\_Search\_Product table.

A close-up of a computer code

Description automatically generated

# Worksheet 3

### Creating a table function from aggregated view.

A screen shot of a computer code

Description automatically generated

# Worksheet 4

### Creating a stored procedure.

A screen shot of a computer

Description automatically generated

## Dashboard

A screenshot of a computer

Description automatically generated

A screenshot of a computer

Description automatically generated

# Worksheet 5

1. **Provide the name and description of the data set you chose.**

**Name of the Dataset:** Traject Data: Amazon Product Results Data

**Description:**

The Traject Data: Amazon Product Results dataset provides real-time information on Amazon product listings, offering detailed insights such as brand details, specifications, seller information, imagery, videos, deals, and more. This dataset enables accurate monitoring of product listings, price trends, market research, and competitor intelligence, supporting various business functions across industries.

1. **Explain in 1-2 sentences what your naming convention is and what your schemas are so the instructor can locate them quickly.**

I created three different tables in the curation layer as Product, Category and Buyer.

All the tables names are named starting with ‘CUR’.

1. **Briefly explain the logic/formulas used for any custom fields you create in your curation.**

**layer (This is a mini data catalog)**

* I used the Amazon product dataset to establish three distinct tables aimed at analyzing product sales based on various factors.
* Firstly, I structured a schema and populated data from the Amazon database to newly created tables.
* Initially the database I selected had a single table with multiple columns. I derived three separate tables.
* While addressing null values within the dataset, I removed irrelevant entries and incorporated average values into certain columns using the 'AVG' aggregation function.
* In next worksheet, using SUM, COUNT, AVG, and MAX, I crafted views to showcase key insights:
* Total products popularized through coupons.
* Total purchase amounts per customer.
* Product categorization.
* Maximum product rating with its title.
* From these views, I developed a function in the third worksheet to retrieve category names and product counts based on search criteria.
* In worksheet four, I crafted a stored procedure and created dashboard. The dashboard featured four tiles,
* Following two visualizations were shown in the dashboard:
* Product sales based on Product Category
* Product Ratings based on product category.