**ETL PIPELINE PROJECT REPORT**

**Nashra Ghaffar (CT-20032)**

**Muhammad Abdul Rafay (CT-20012)**

**Find the code here:** [**https://github.com/sngwebs786/Pipeline-ETL-using-MERN**](https://github.com/sngwebs786/Pipeline-ETL-using-MERN)

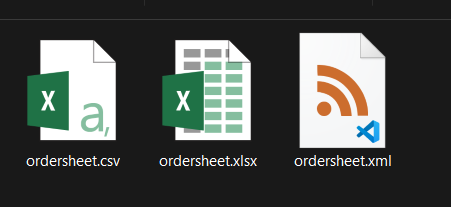
**Data Set:**

In this project we are going to work on the dataset of Sales available in different formats, comprises of the following columns:

* Order ID
* Order Date
* Ship Date
* Ship Mode
* Customer ID
* Customer Name
* Segment
* Country
* City
* State
* Postal Code
* Region
* Product ID
* Category
* Sub-Category
* Product Name
* Sales
* Quantity
* Discount
* Profit

**Data Structure**

The data in the files exhibits unique rows across the three file types: CSV, XLSX, and XML. While the data differs between files, the structure remains consistent.



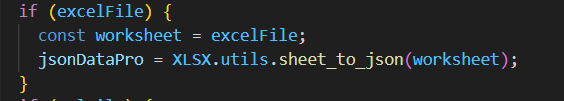
**Technologies used:**

1. Node JS (Backend)
2. React JS (Frontend)
3. MongoDB (Database)

**Extract:**

Data extraction involves uploading three distinct file types: CSV, XLSX, and XML. The data from these files will be extracted and converted into JSON format for further processing.

* 1. **Excel File:**

****

* 1. **Xml File:**

****

* 1. **Csv File:**

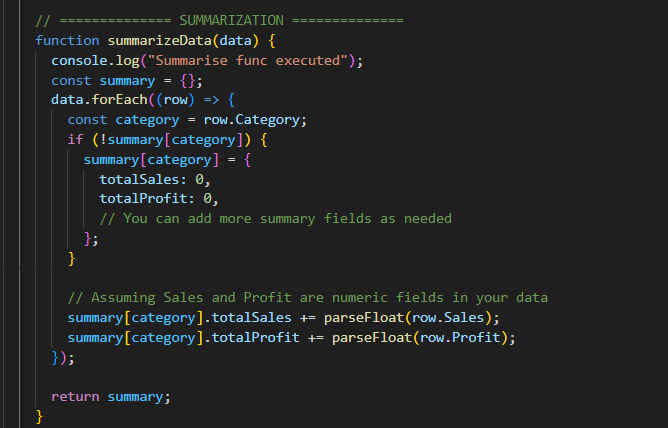
****

**Transform**

The transformation phase comprises three main functions:

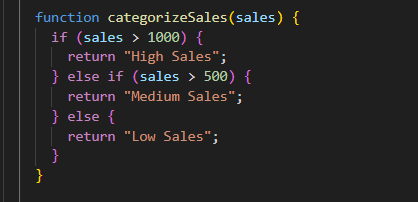
**Summarize:** Aggregates sales and profit based on product categories.

* + Total sales and profit for the "Technology" category.
  + Total sales and profit for the "Office Supplies" category.

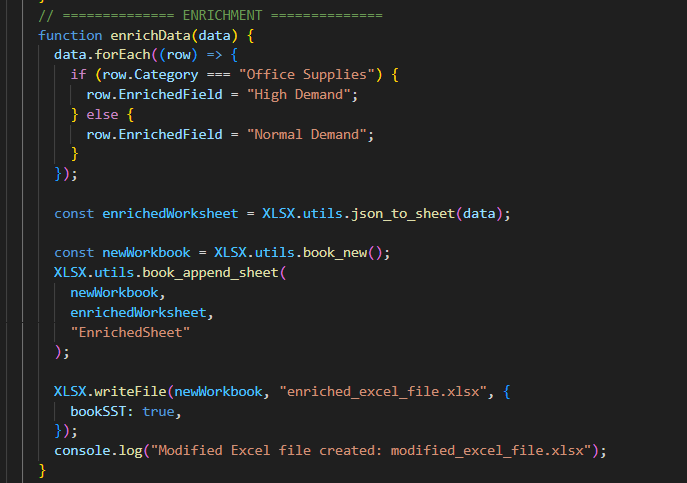


**Categorize:** Classifies products into three categories based on sales:

* + Low Sale: Sales < 500
  + Medium Sale: 500 ≤ Sales < 1000
  + High Sale: Sales ≥ 1000



**Enrichment:** Determines demand for a specific category. If the product belongs to the "Office Supplies" category, it is marked as high demand; otherwise, it is labeled as medium demand.

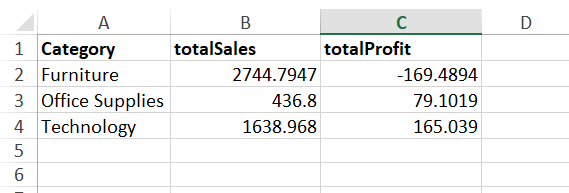


**New Columns added:**

After transformation, new columns are introduced:

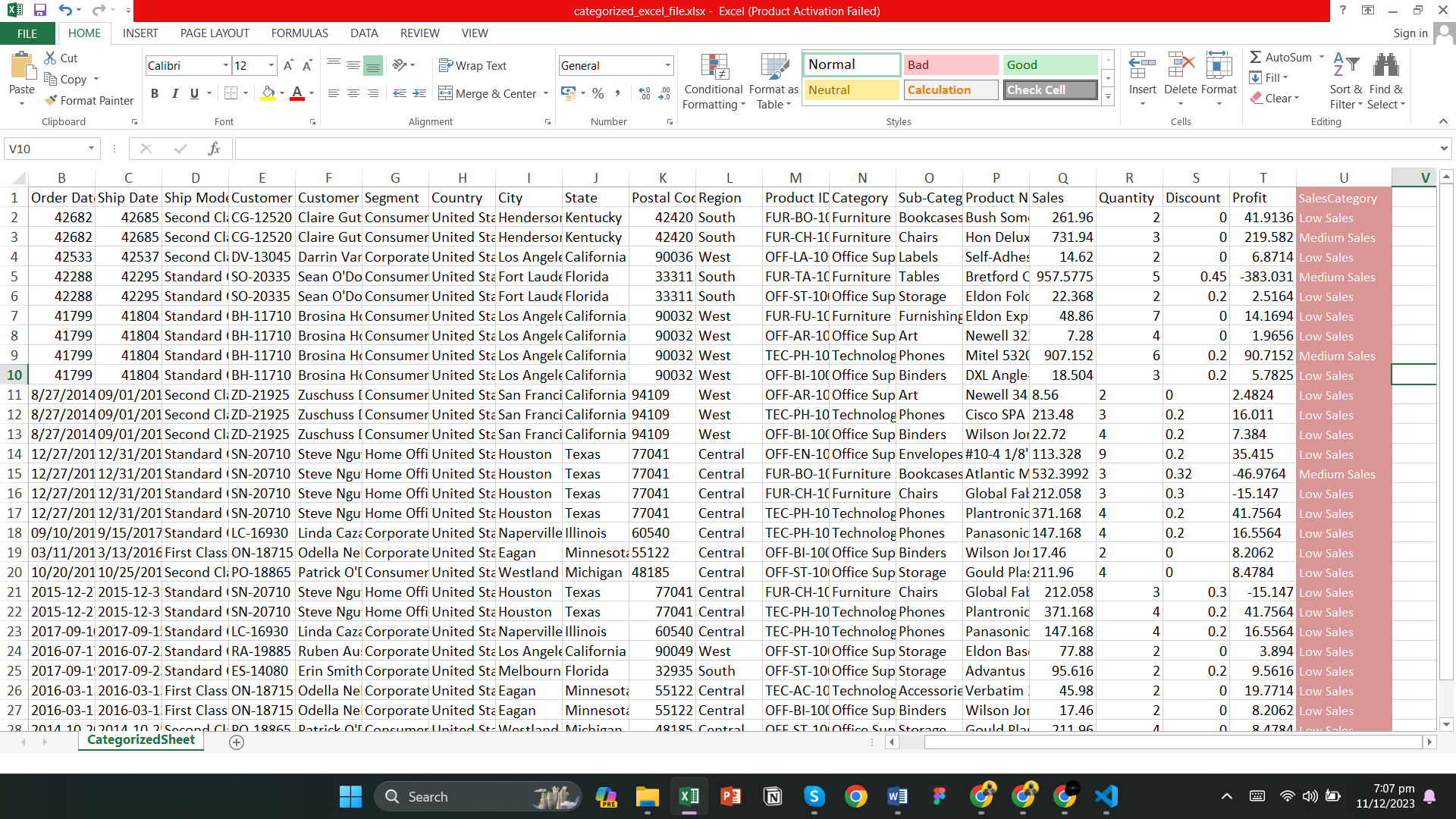
**Summarize:**

* + Total Sales
  + Total Profit



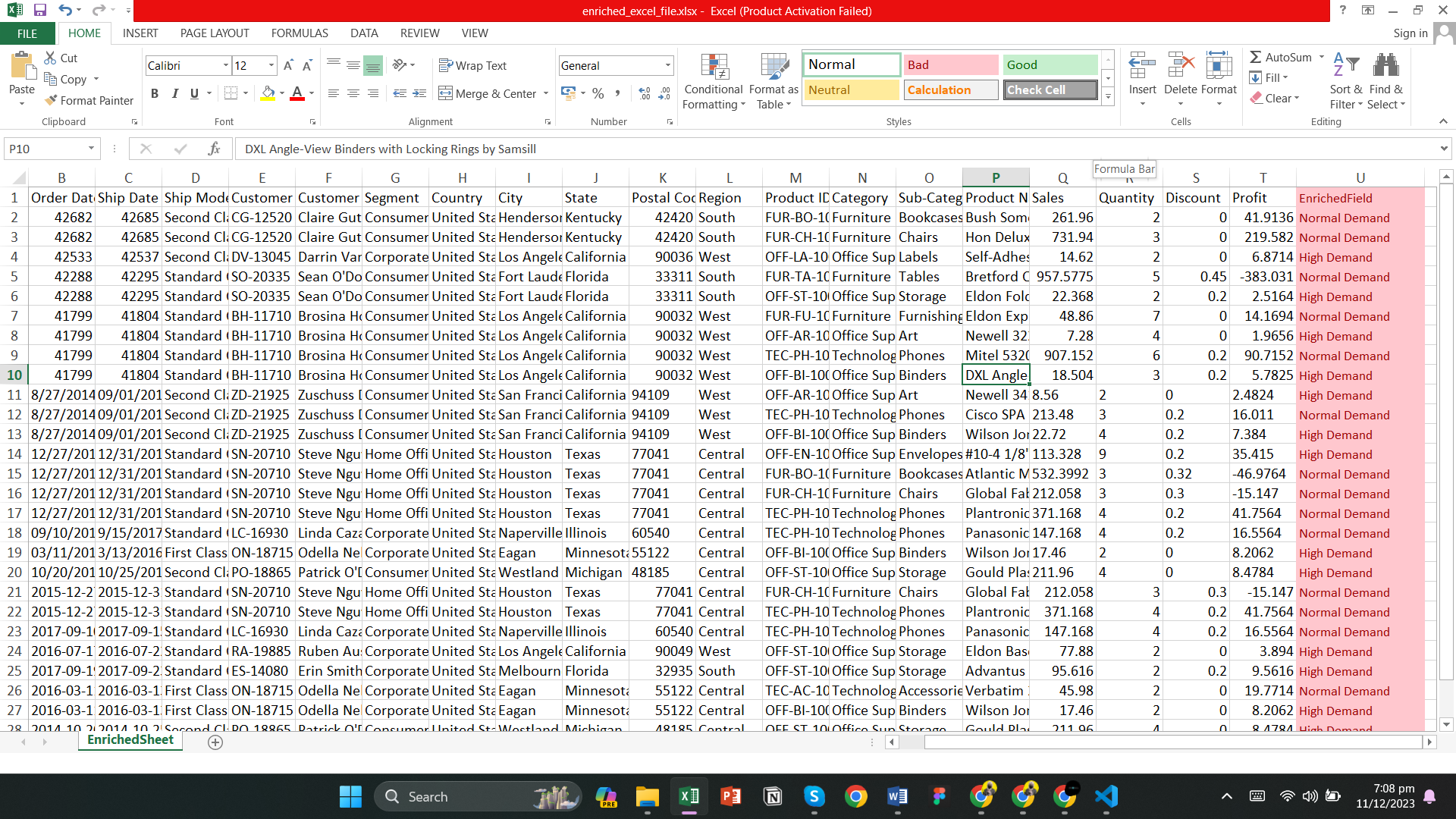
**Categorize:**

* + Sales Category



**Enrichment:**

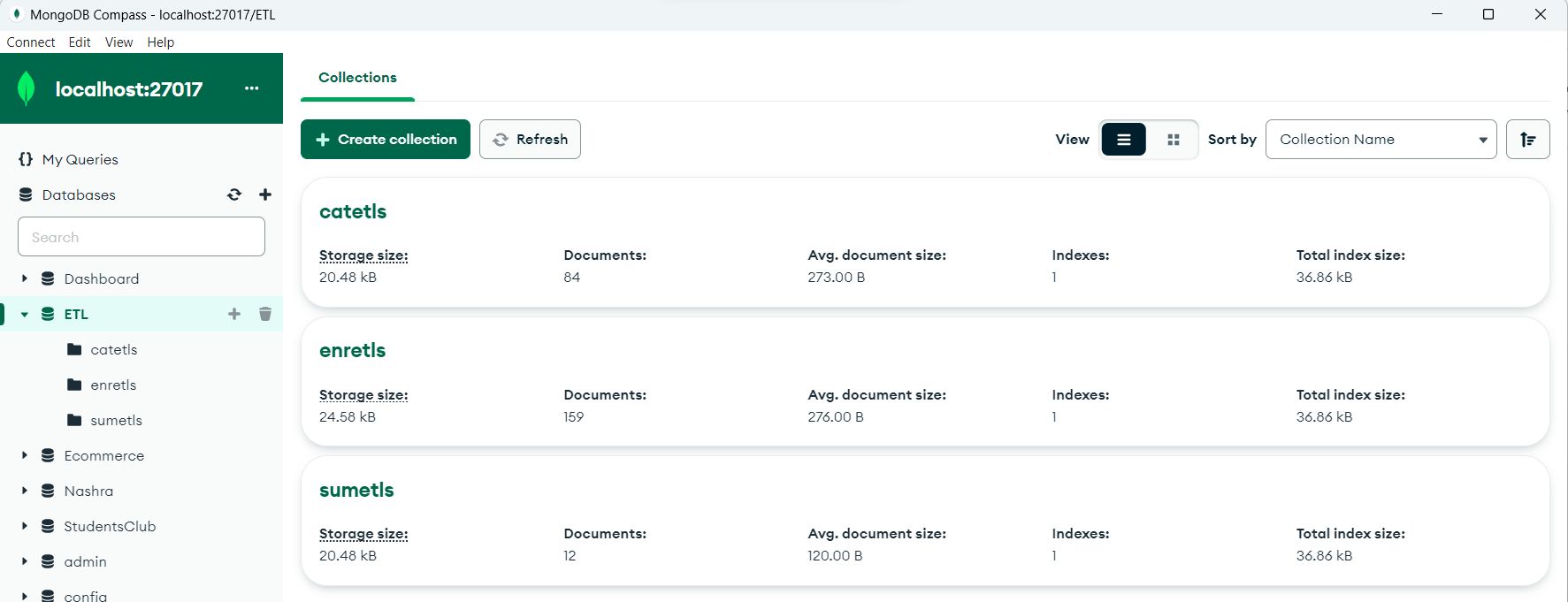
* + Enriched Field



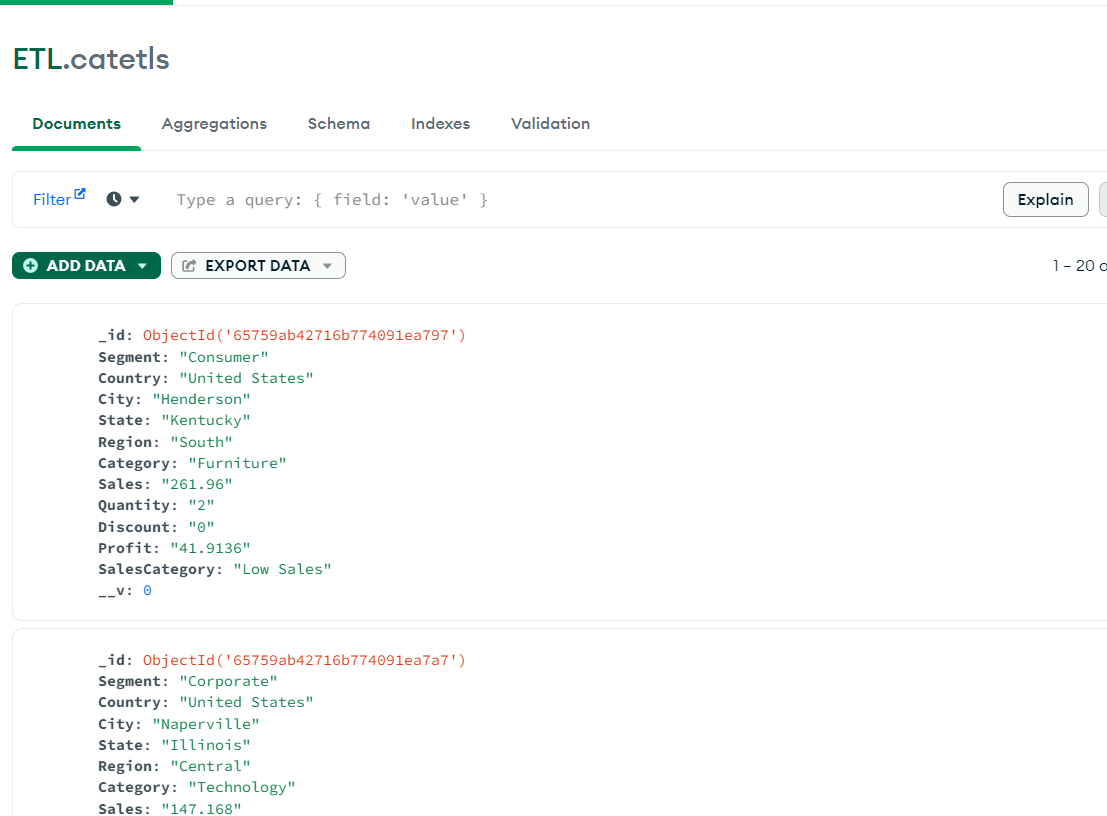
These columns enhance the dataset's analytical capabilities and provide insights into sales performance and demand patterns.

**Load**

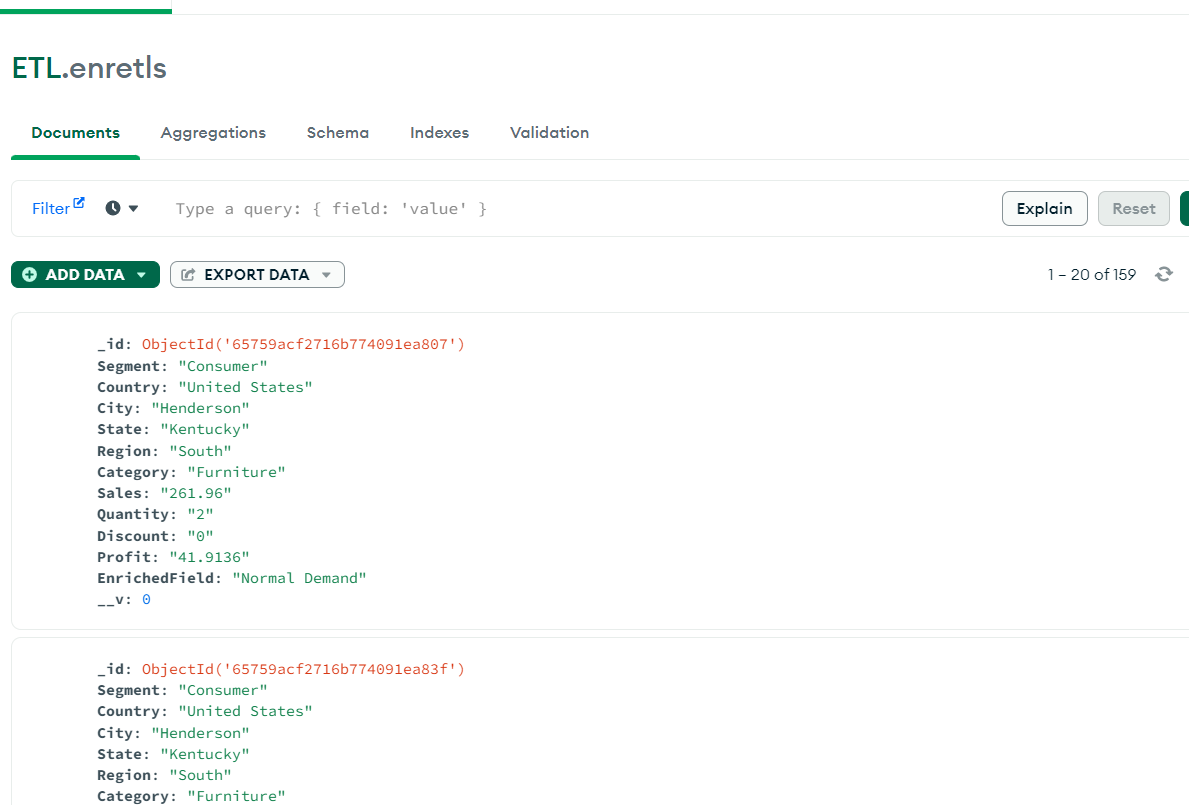
The final step involves loading the transformed data into the desired destination for analysis, reporting, or further integration with other systems. We have used Mongo DB for the storage of data.



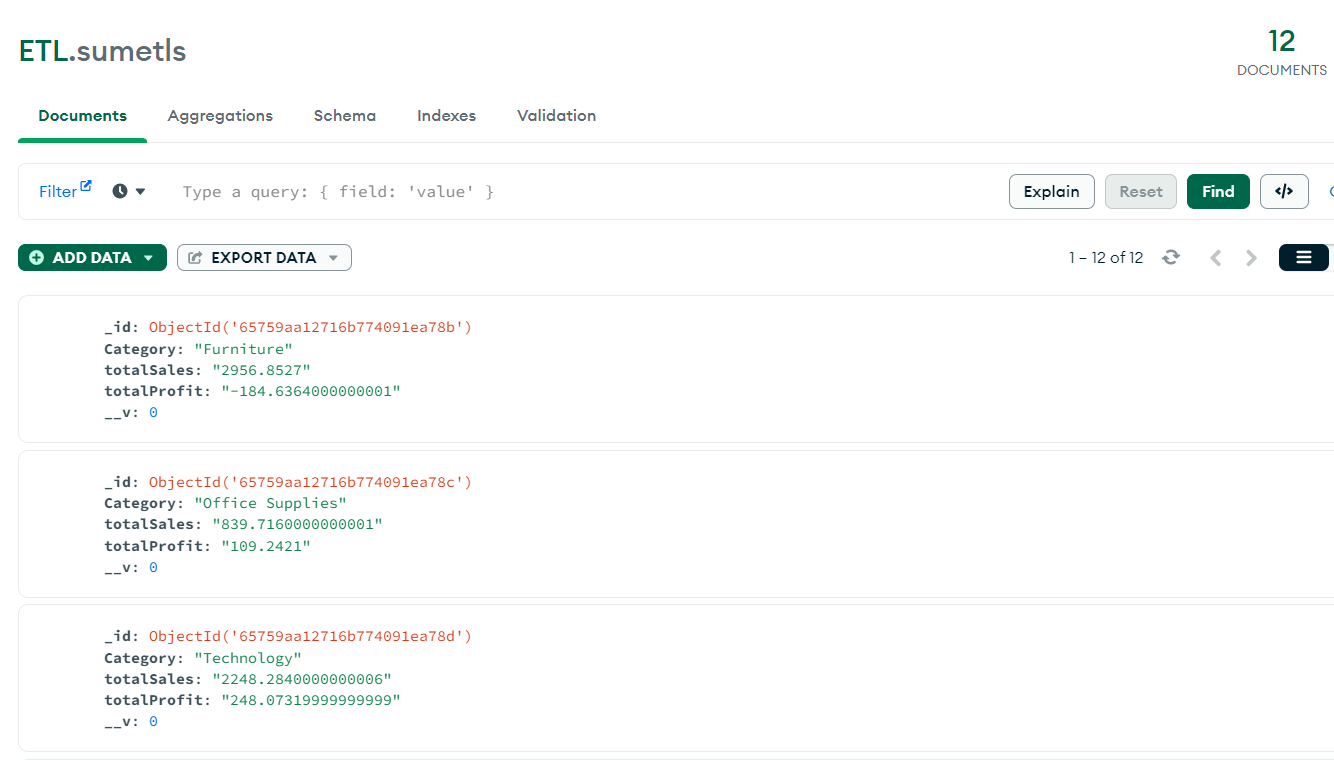
1. **Categorized**:



1. **Enrichment**:



1. **Summarized**:



**Conclusion**

This documentation provides an overview of the ETL project, detailing the data structure, extraction process, transformation steps, and the resulting dataset's new columns. The project aims to facilitate meaningful analysis and decision-making based on the enriched and categorized data.