SENG8081-Case Studies

Conestoga College

Team Members:

Jyoti Maske  
Hiral Parekh  
Yash Khamar  
Neel Patel

Team 5

ONLINE RETAIL SALES ANALYSIS

System Flow Diagram

A diagram of a company

AI-generated content may be incorrect.

1. **Start**

* This marks the **initiation of the data pipeline**—the point at which data requirements are defined and the flow begins.

1. **Data Collection *(from Kaggle + FakeAPI)***

* You are collecting data from **two sources**:
  + **Kaggle** (CSV datasets—usually batch/historical data)
  + **FakeAPI** (simulated API—represents real-time or dynamic input)
* This step brings **raw, unstructured or semi-structured data** into your system.

1. **Data Cleaning**

* This involves removing:
  + **Missing values**
  + **Duplicates**
  + **Inconsistent formats**
* Ensures the dataset is **reliable and standardized** before use.

1. **Preprocessing**

* You perform key transformations here, such as:
  + **Feature selection** – choosing the most relevant fields
  + **Renaming columns** for consistency
  + **Changing data types** (e.g., string to datetime)
  + **Handling nulls** or encoding values
* This stage **prepares the data for storage and analysis**.

1. **Export to CSV**

* After cleaning and preprocessing, data is **saved into CSV files**.
* This acts as an **intermediate stage** before loading into databases.

1. **Store in SQL Server (SSMS)**

* The structured CSV data is **imported into SQL Server**, inside a specific database (e.g., SalesAnalysis).
* SQL Server acts as your **central data warehouse** where data is:
  + Indexed
  + Queryable using SQL
  + Securely stored

1. **Data Visualization (Tableau)**

* Data from SQL Server is **connected to Tableau**.
* You build **interactive dashboards** to analyze:
  + Sales trends
  + Product performance
  + Ratings and categories
* Users can gain **insights from visual data representations**.

1. **Future Prediction (Forecasting)**

* Based on the **analyzed trends**, you implement models or logic to:
  + **Predict future sales**
  + **Forecast demand**
  + **Estimate performance**
* This helps in **strategic planning and business decisions**.

**Summary:**

Our project flow captures a **complete data lifecycle**:  
**Collect → Clean → Prepare → Store → Visualize → Predict**

It combines both **batch and semi-real-time data** and is aligned with how modern big data pipelines work in analytics-focused projects.