1. Problem Formulation

1.1 Business Problem Definition

RecoMart is an e-commerce platform that aims to improve **customer engagement and conversion rate** by providing **personalized product recommendations** to its users.

Currently, users browse products without personalized guidance, which results in:

- Lower click-through rates
- Missed cross-selling opportunities
- Reduced average order value

Business Goal:
Design a data-driven recommendation system that suggests relevant products to users based on their past behavior and product characteristics, thereby improving:

- Conversion rate
- User engagement
- Cross-selling effectiveness

The recommendation system must be continuously updated using fresh user interaction and transaction data.

1.2 Key Data Sources and Attributes

RecoMart collects data from multiple sources. The pipeline integrates the following key datasets:

1. User Interaction Data (Clickstream Logs)

Captured from web and mobile platforms.

Attributes:

- `user_id` – Unique identifier for users
- `product_id` – Identifier of interacted product
- `event_type` – View, click, add-to-cart
- `timestamp` – Time of interaction
- `device` – Web or mobile

This data reflects **implicit user preferences**.

---

2. Transactional Purchase Data

Records confirmed purchases made by users.

Attributes:

- `transaction_id`
- `user_id`
- `product_id`
- `quantity`
- `price`
- `timestamp`

This data represents **explicit user intent and value-based interactions**.

---

3. Product Metadata (Catalog / API)

Fetched from internal or external product services.

Attributes:

- `product_id`
- `category`
- `brand`
- `price`
- `popularity_score`

This data supports **content-based recommendations** and feature enrichment.

1.3 Expected Outputs from the Pipeline

The end-to-end data pipeline is expected to generate the following outputs:

1. Cleaned and Validated Datasets

- Structured and validated datasets for exploratory data analysis (EDA)
- Removal of duplicates, handling missing values, and schema consistency

2. Engineered Feature Sets

Features suitable for recommendation algorithms, such as:

- User activity frequency
- Product popularity
- Average user-item interaction strength
- Aggregated behavioral statistics

These features support:

- Collaborative filtering

- Content-based recommendation models

3. Deployable Recommendation Model

- A trained recommendation model capable of generating personalized product suggestions
- A simple inference interface to retrieve top-N product recommendations for a user

1.4 Evaluation Metrics

The recommendation model will be evaluated using **ranking-based metrics**, which are standard for recommendation systems:

- **Precision@K**
  Measures the proportion of relevant items among the top-K recommended products.
- **Recall@K**
  Measures the proportion of relevant items successfully retrieved in the top-K recommendations.
- **Normalized Discounted Cumulative Gain (NDCG)**
  Evaluates ranking quality by assigning higher importance to correctly ranked items at higher positions.

These metrics align directly with the business objective of improving recommendation relevance and user engagement.

1.5 Expected Pipeline Outcomes

By implementing this pipeline, RecoMart will achieve:

- A scalable and automated data ingestion and processing system
- High-quality, versioned datasets for machine learning

- Reproducible feature generation for training and inference

- A recommendation model that learns continuously from fresh data