

2. Data Collection and Ingestion

2.1 Data Sources

The data ingestion layer is designed to collect data from multiple heterogeneous sources required for building a recommendation system. The following data sources are ingested:

1. User Interaction Data (Clickstream Logs)

This data captures user behavior on the RecoMart platform and is ingested from CSV files generated by web and mobile applications.

Attributes include:

- `user_id`
- `product_id`
- `event_type` (view, click, add_to_cart)
- `timestamp`
- `device`

This dataset represents **implicit feedback**, which is essential for learning user preferences.

2. Product Metadata

Product-related information is ingested from a REST API (simulated using a mock JSON source).

Attributes include:

- `product_id`
- `category`

- `brand`
- `price`
- `popularity_score`

This dataset supports **content-based recommendation** and feature enrichment.

2.2 Ingestion Strategy

The ingestion process is designed with the following objectives:

Automated and Periodic Ingestion

- Ingestion scripts are designed to run on a scheduled basis (daily).
- Data is partitioned by ingestion date (`year/month/day`) to support automation, historical tracking, and reprocessing.

Error Handling and Fault Tolerance

- Try–except blocks are implemented to handle runtime errors.
- Retry mechanisms are included for API-based ingestion to handle transient failures.
- Ingestion failures do not corrupt existing data.

Logging and Auditability

- A centralized logging mechanism records ingestion events.
 - Logs capture execution start time, success messages, and error details.
 - These logs enable monitoring, debugging, and audit trails.
-

2.3 Raw Data Storage Layout

All ingested data is stored in a **raw data lake** using a structured and time-partitioned directory layout:

```
data/raw/
  └── clickstream/
      |   └── yyyy/mm/dd/
  └── products/
      └── yyyy/mm/dd/
```

This storage design ensures:

- Preservation of original source data
 - Traceability of ingestion events
 - Support for backfilling and data versioning
-

2.4 Outcomes of the Ingestion Layer

The data collection and ingestion layer produces:

- Immutable raw datasets stored in a structured data lake
- Logged records of ingestion success and failure
- A reliable foundation for downstream data validation, preparation, and feature engineering