2. Data Collection and Ingestion

2.1 Data Sources Ingested

The pipeline ingests data from **multiple heterogeneous sources**, simulating a real-world e-commerce data ecosystem:

1. User Interaction Data (Clickstream)

- **Source Type:** CSV files
- **Origin:** Web and mobile platforms
- **Ingestion Mode:** Batch (daily)

This data captures **implicit user behavior**, such as views, clicks, and add-to-cart events.

---

2. Product Metadata

- **Source Type:** REST API

API Endpoint:

```
https://fakestoreapi.com/products
```

-
- **Ingestion Mode:** Automated API-based ingestion

This API provides product catalog information including price, category, ratings, and descriptions, which supports **content-based feature generation**.

2.2 Ingestion Design

Each ingestion script is designed following modern data engineering best practices.

Automated and Periodic Fetching

- Ingestion scripts are designed to run periodically (e.g., daily)
- Date-based folder partitioning enables incremental ingestion and replay

---

Error Handling and Retry Mechanism

- API ingestion includes retry logic to handle transient failures
- HTTP errors and network issues are captured and retried up to a configurable limit
- Failures do not corrupt previously ingested data

Logging and Audit Trails

- All ingestion activities are logged using a centralized logging mechanism
- Logs capture:
    - Start and end of ingestion
    - Success or failure status
    - Retry attempts and error messages

2.3 Ingestion Implementation

Clickstream Data Ingestion

- Reads interaction data from CSV files

Writes raw data into a timestamp-partitioned data lake structure:

`data/raw/clickstream/YYYY/MM/DD/`

-

Product Metadata Ingestion

- Fetches product data from the Fake Store REST API
- Stores the **raw API response without transformation** to preserve source fidelity

Writes data to:

`data/raw/products/YYYY/MM/DD/products.json`

-

Storing raw API responses ensures reproducibility and allows downstream transformations to be re-applied if needed.

2.4 Raw Data Storage Structure

```
data/raw/
├── clickstream/
│   └── YYYY/MM/DD/clickstream.csv
└── products/
    └── YYYY/MM/DD/products.json
```

This structure mirrors cloud-based data lake designs and supports scalable downstream processing.

2.5 Logs Showing Ingestion Success and Failure

All ingestion runs generate logs stored at:

```
logs/ingestion.log
```

Successful Ingestion Example

```
INFO - Starting product API ingestion
INFO - Product data ingested successfully from API at
data/raw/products/2025/01/01/products.json
```

Failure and Retry Example

```
ERROR - Attempt 1 failed: Connection timeout
ERROR - Attempt 2 failed: Connection timeout
ERROR - Product ingestion failed after retries
```

2.6 Summary

This ingestion layer:

- Integrates batch and API-based data sources
- Ensures fault tolerance through retries and logging
- Preserves raw data for lineage and reproducibility
- Provides a reliable foundation for downstream validation and feature engineering