3. Raw Data Storage

3.1 Storage Approach

The ingested data is stored in a **local filesystem-based data lake**, which simulates cloud object storage systems such as AWS S3.
Using the local filesystem allows the pipeline to be executed and validated without external dependencies while still following modern data lake design principles.

All storage paths are **centrally managed using configuration files**, ensuring that data locations are not hardcoded in the ingestion logic.

3.2 Data Lake Folder Structure

Raw data is organized using a structured, hierarchical folder layout based on:

- **Data source** (clickstream, products)
- **Ingestion date** (YYYY/MM/DD)

Raw Data Layout

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

This layout mirrors industry-standard data lake structures used in large-scale data platforms.

3.3 Partitioning Strategy

A **date-based partitioning strategy** is used for raw data storage:

```
/<data_source>/<YYYY>/<MM>/<DD>/
```

Benefits:

- Enables incremental data processing
- Supports historical reprocessing and backfills
- Simplifies debugging and auditability
- Preserves ingestion-time data lineage

3.4 Storage Configuration

All storage locations are defined in a centralized configuration file:

📄 `config/paths.yaml`

```
raw: data/raw
validated: data/validated
processed: data/processed
features: data/features
logs: logs
source_files: data/source_files
```

Ingestion scripts dynamically resolve storage paths using this configuration, allowing the pipeline to remain flexible and environment-independent.

3.5 Data Upload Mechanism

During execution, ingestion scripts:

- Read source data from configured upstream locations
- Create date-partitioned directories automatically
- Write raw data files to the data lake
- Log storage paths and execution status

No manual upload steps are required, as data is written programmatically during ingestion.

3.6 Traceability and Reproducibility

- Raw data files are stored as **immutable artifacts**
- Each ingestion run creates a new, timestamp-partitioned directory
- Ingestion logs capture file paths and execution timestamps

This design ensures full traceability and allows downstream pipeline stages to be re-run reliably using historical raw data.