# Monthly Returns & Shipments ETL

---

## Objective

The goal of this assignment is to build a **monthly data pipeline** for ecommerce **returns** and **shipments**, using a structured **bronze → silver → gold** approach in **PySpark**.

You will:

- Ingest raw monthly data.
- Apply cleaning and transformation logic.
- Generate **business-enriched metrics** for analytics.
- Ensure **incremental monthly processing** without duplicating previous data.

## 1. Bronze Layer (Raw Monthly Data)

**Purpose:** Ingest raw monthly data from ADLS into bronze tables.

**Instructions:**

- Load **monthly returns** and **shipments** data from ADLS. (using the same approach as daily data ingestion)
- Push the raw data **as-is** to bronze tables.
- No transformations are required in this layer; maintain the schema exactly as provided.
- Use Autoloader functionality for ingestion.
- Ensure incremental monthly loads are appended without affecting existing bronze data.

## 2. Silver Layer (Cleaned Monthly Data)

**Purpose:** Clean, standardize, and prepare monthly data for analysis.

1. **Returns Table Transformations**
    - Remove duplicates based on `order_id`, `order_dt`, `return_ts`.
    - Convert `order_dt` to **DateType**.
    - Convert `return_ts` to **TimestampType**.
    - Convert `reason` column to **uppercase** and trim whitespace.
    - Add `processed_time` column with the current timestamp.
2. **Shipments Table Transformations**
    - Convert `order_dt` to **DateType**.
    - Convert `carrier` column to **uppercase** and trim whitespace.
    - Add `processed_time` column with the current timestamp.

## 3. Gold Layer (Business-Enriched Monthly Data)

**Purpose:** Enrich silver data with business-relevant columns for analytics.

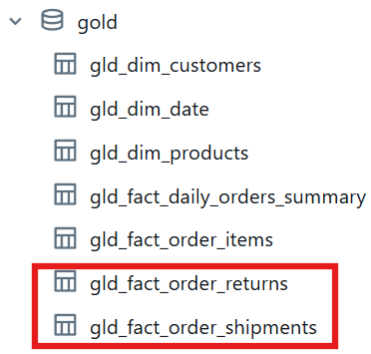
1. **Returns Table Gold Transformations**
    - Add `date_id` column: yyyyMMdd format from order_dt.
    - Calculate `return_days` = difference in days between return_ts and order_dt.
    - Create policy compliance flags:
        - `within_policy` → 1 if `return_days` <= 15, else 0.

- `is_late_return` → 1 if `return_days` > 15, else 0.

**B) Shipments Table Gold Transformations**

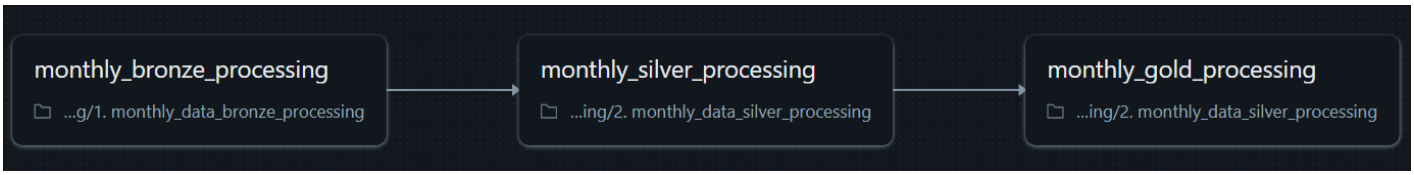
- Add `carrier_group`:
  - Domestic: ECOMEXPRESS, DELHIVERY, XPRESSBEES, BLUEDART.
  - International: All other carriers.
- Add `is_weekend_shipment` flag:
  - True if `order_dt` is Saturday or Sunday, else False.

**Gold Layer – Final Tables**




## 4. Orchestration: Monthly Job

- **Create Job:** Name it `monthly_job` to run the full bronze → silver → gold pipeline every month.
- **Schedule:** Run the job once every month (e.g., 1st day of the month at 02:00 AM).
- **Dependencies:** Ensure bronze → silver → gold executes in order.
- **Incremental Data:** Job should process only new monthly data without duplicating previous months.


monthly_bronze_processing
...g/1. monthly_data_bronze_processing
monthly_silver_processing
...ing/2. monthly_data_silver_processing
monthly_gold_processing
...ing/2. monthly_data_silver_processing

---