ETL and Warehousing

GROUP 6

Anna Mika , Hugo Leduc, Camil Nitel and Elisa Leclerc

Our Brand

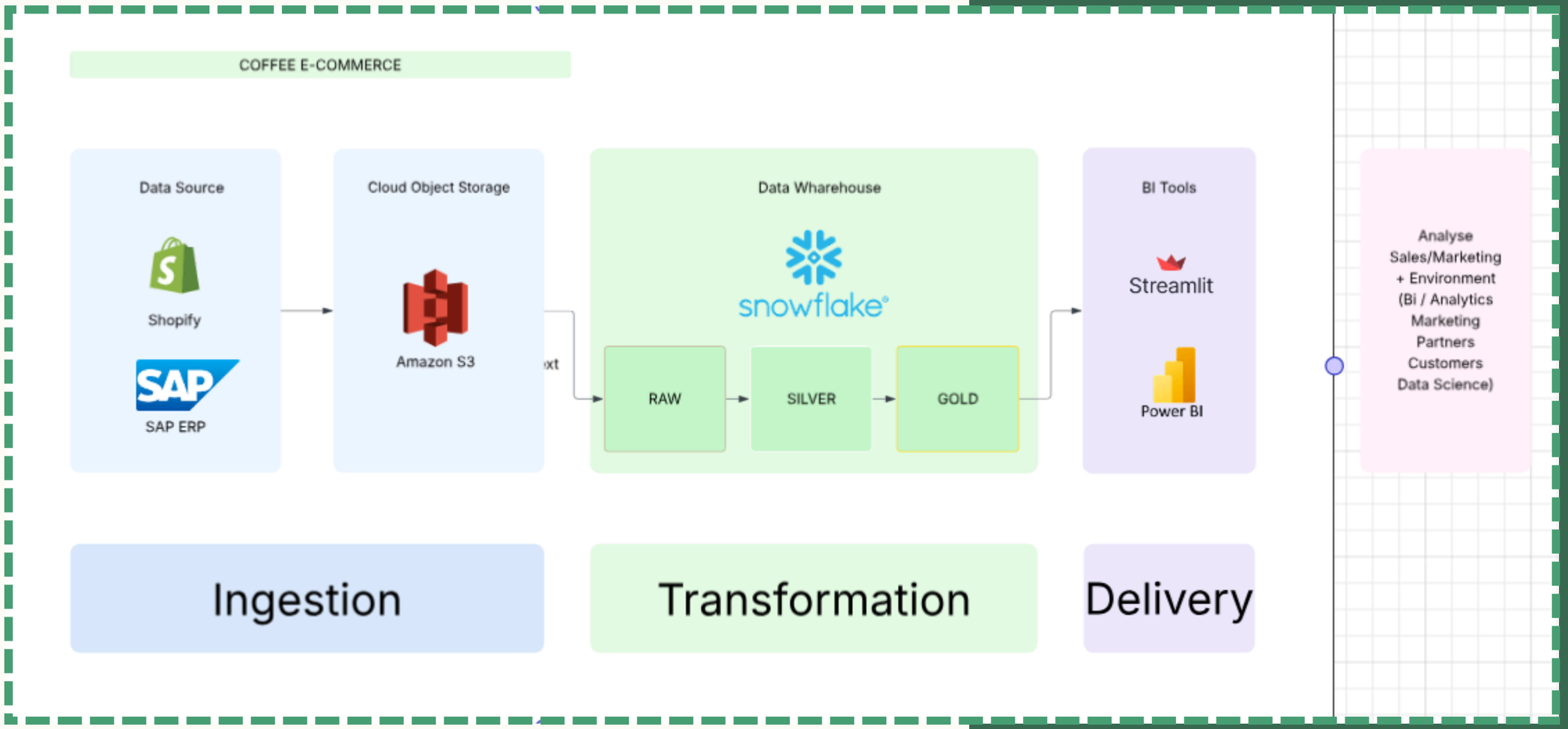
Ecocoffee offers organic, specialty-grade coffee and matcha products designed to make high-quality drinks accessible and enjoyable for everyone. We focus on sustainability, using eco-friendly packaging and responsibly sourced ingredients. With a fun and modern identity, our goal is to bring people together through a coffee experience that feels both premium and approachable.

Snowflake Data Pipeline Architecture

This project implements an automated data ingestion and transformation pipeline in Snowflake, designed to handle client orders and carbon emissions data for sustainability analytics.

The pipeline uses multi-layer architecture (RAW → SILVER → GOLD → STREAMLIT_APPS) combined with streams and tasks to ensure data moves seamlessly and cleanly through each stage.

End-to-End Flow Summary



BRONZE Layer

RAW Data Ingestion

Store unprocessed raw data exactly as it is ingested from the source

Tables

- RAW_CLIENT_SUPPORT_ORDERS_PY_SNOWPIPE
- RAW_CARBON_EMISSIONS_PY_SNOWPIPE

Data is loaded automatically via Snowpipe. A Python script triggers Snowpipe when new source files are available in cloud storage.

Streams

- RAW_CLIENT_SUPPORT_ORDERS_STREAM
- RAW_CARBON_EMISSIONS_STREAM

These streams track changes (new rows) in the RAW tables, allowing tasks to automatically process new data.

SILVER Layer

Data Cleaning

Clean and standardize the raw data for analytical use

- Removing duplicate transactions (TXID)
- Ensuring no NULL TXID
- Converting item names to proper case (INITCAP)
- Date validation (no future dates)
- Standardizing bag sizes to uppercase
- Nulling invalid or future timestamps
- Handling negative or null carbon emissions values

These tasks run immediately after new data ingestion, ensuring that SILVER tables are always up to date.

GOLD Layer

Curated Data for Analytics

Store finalized, cleaned data ready for reporting and dashboards

Tables

- GOLD_CLIENT_SUPPORT_ORDERS
- GOLD_CARBON_EMISSIONS

Tasks

- task_gold_orders
- task_gold_emissions

These tasks are triggered after the corresponding SILVER streams detect changes, ensuring that GOLD tables always mirror the cleaned SILVER tables

STREAMLIT_APPS Database

Data Serving Layer

Provide read-optimized copies of the GOLD tables for Streamlit dashboards and visualization tools

TABLES

- CLIENT_SUPPORT_ORDERS
- CARBON_EMISSIONS

Stored under: STREAMLIT_APPS.GOLD_COPY

TASKS

- task_copy_gold_orders_to_streamlit
- task_copy_gold_emissions_to_streamlit

These copy tasks ensure that the STREAMLIT_APPS database always reflects the latest GOLD-layer data, ready to be queried by Streamlit applications.

Snowpipe Trigger Script

The snowpipe.py script is used to automate ingestion into the RAW tables.

Purpose

- Detect new source files (from S3, local upload, etc.)
- Call Snowflake's REST API to trigger Snowpipe ingestion

Typical Workflow

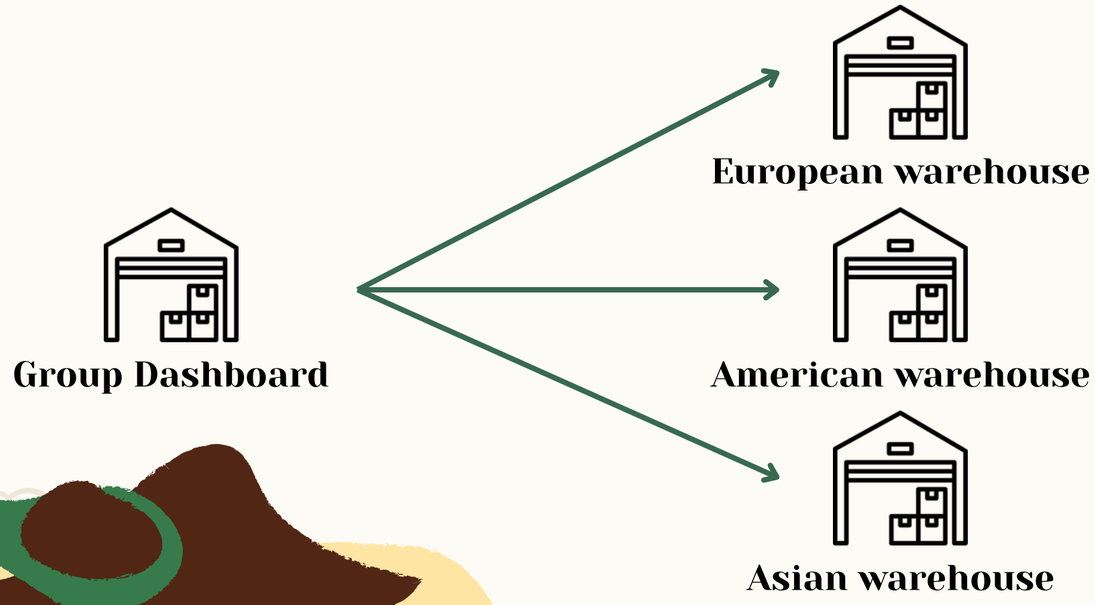
- The script authenticates to Snowflake using credentials (user/role/key).
- It identifies new data files to load.
- It sends a REST API call to the Snowpipe endpoint:
  https://<your_account>.snowflakecomputing.com/v1/data/pipes/<pipe_name>/insertFiles
- Snowpipe loads the data into the corresponding RAW (Bronze layer) table.
- The RAW → SILVER → GOLD → STREAMLIT_APPS pipeline runs automatically via streams and tasks.

Pipeline Benefits

- Fully automated with scheduled tasks (Every minute).
- Modular multi-layer architecture (RAW → SILVER → GOLD → STREAMLIT_APPS)
- Data quality guaranteed through cleaning logic
- Near real-time updates for analytics and dashboards
- Minimal manual intervention — once snowpipe.py runs, everything else flows automatically

Next steps & Upgrades

Dashboarding



Anlytical usage

Cost of the pipeline structure

Size of databases

Change to an event based pipeline

Thank you for your attention !