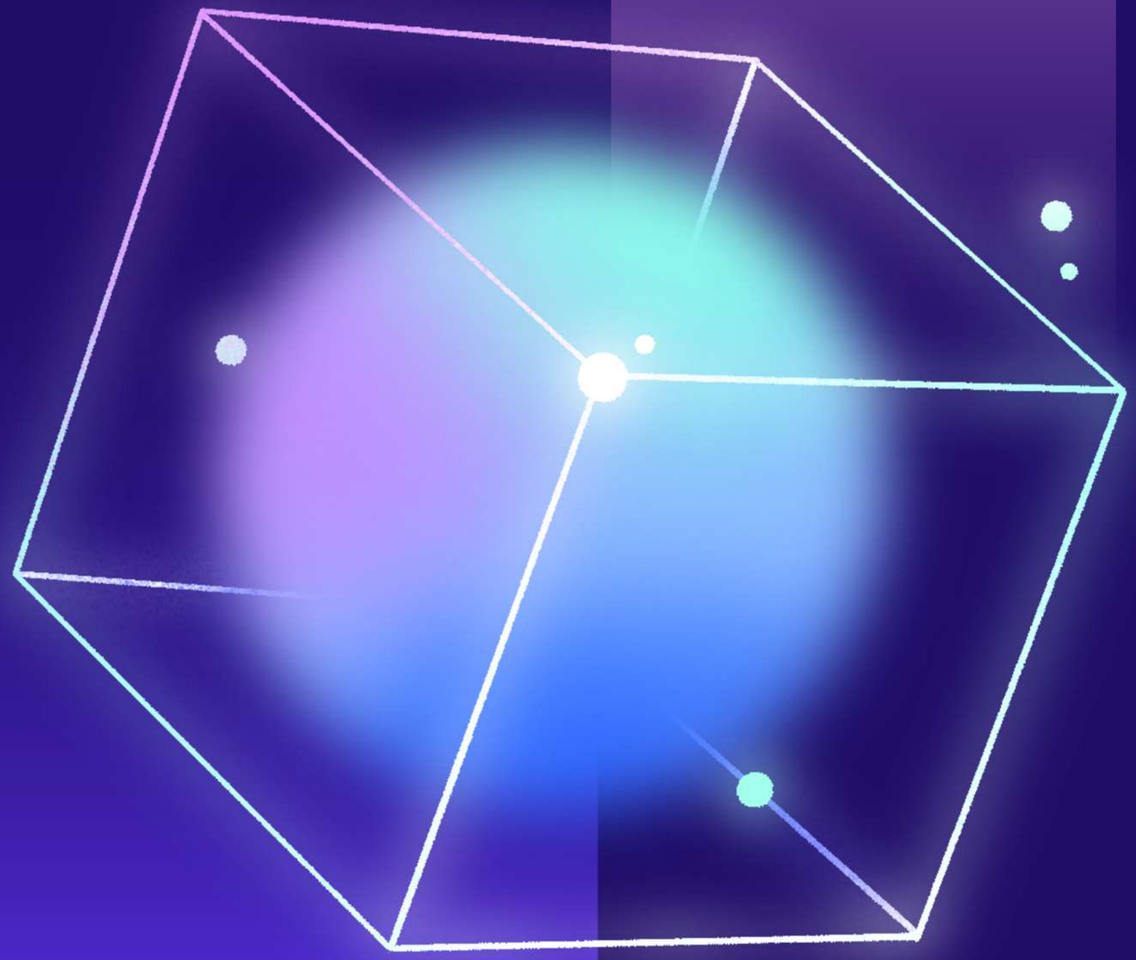
END-TO-END ETL PIPELINES

Pelcharskyi Artur

3 December, 2025



WHAT IS ETL?

ETL stands for **Extract**, **Transform**, **Load**, which is a process used in data management and analytics to move data from various sources into a centralized system, usually a data warehouse.

In essence, ETL ensures that data from multiple sources becomes consistent, accurate, and ready for analysis, enabling better decision-making.

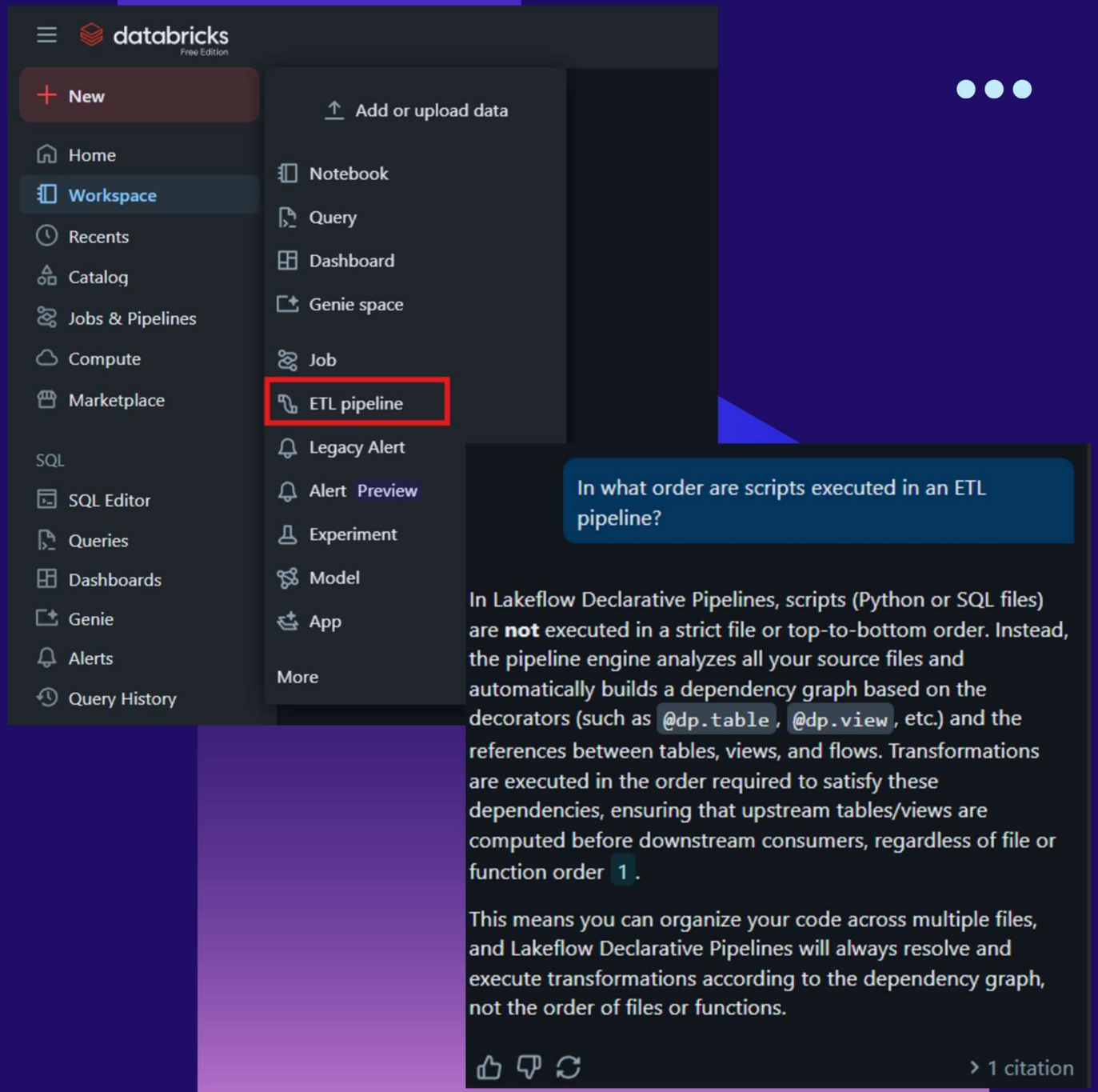
ETL STEPS

- **Extract**: This is the first step, where data is collected from different sources such as databases, APIs, or files. The goal is to gather raw data, even if it’s in different formats.
- **Transform**: Here, the extracted data is cleaned, formatted, and processed to fit the target system. This can include filtering, aggregating, converting data types, or applying business rules.
- **Load**: Finally, the transformed data is loaded into the target system, like a database or data warehouse, where it can be used for reporting, analysis, or machine learning.

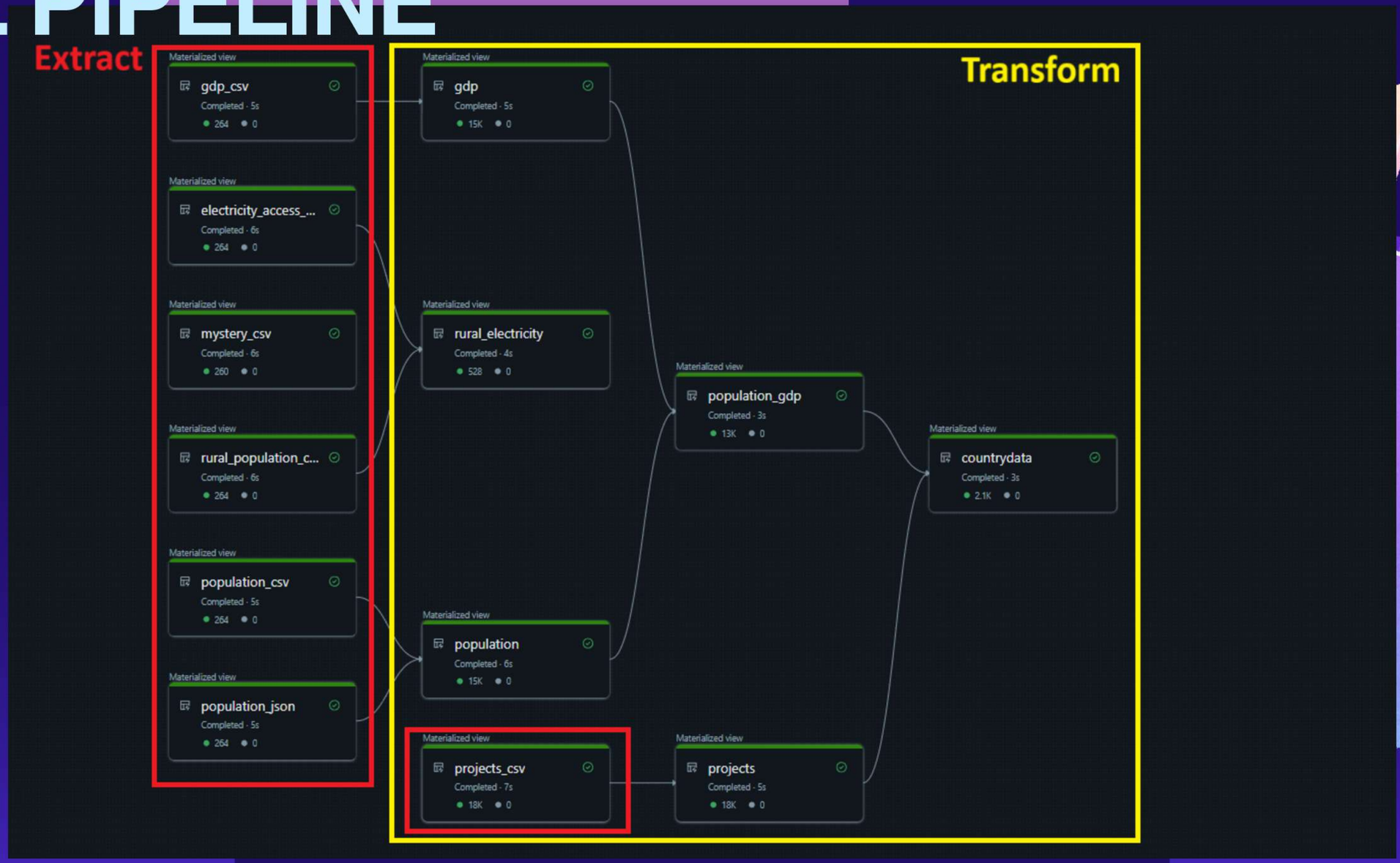
ETL PIPELINES IN DATABRICKS

The ETL Pipeline is part of the Pipelines module in Databricks. This module helps structure code to be modular, scalable, and easy to maintain. Databricks also enables working efficiently with individual tables that are updated within a pipeline.

A **Pipeline** is not the same as a sequence of **Jobs**. In Jobs, we work with notebooks that can be connected in a chain, like **Notebook1 -> Notebook2 -> ….**
In a Pipeline, the focus is on table dependencies, which are automatically resolved. (More details in the demo.)

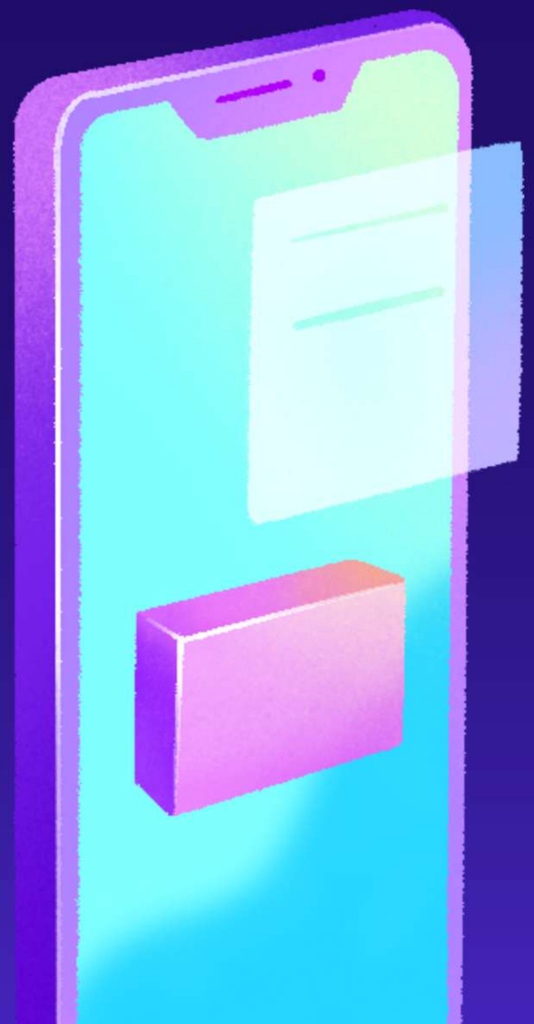


WORLD BANK ETL PIPELINE



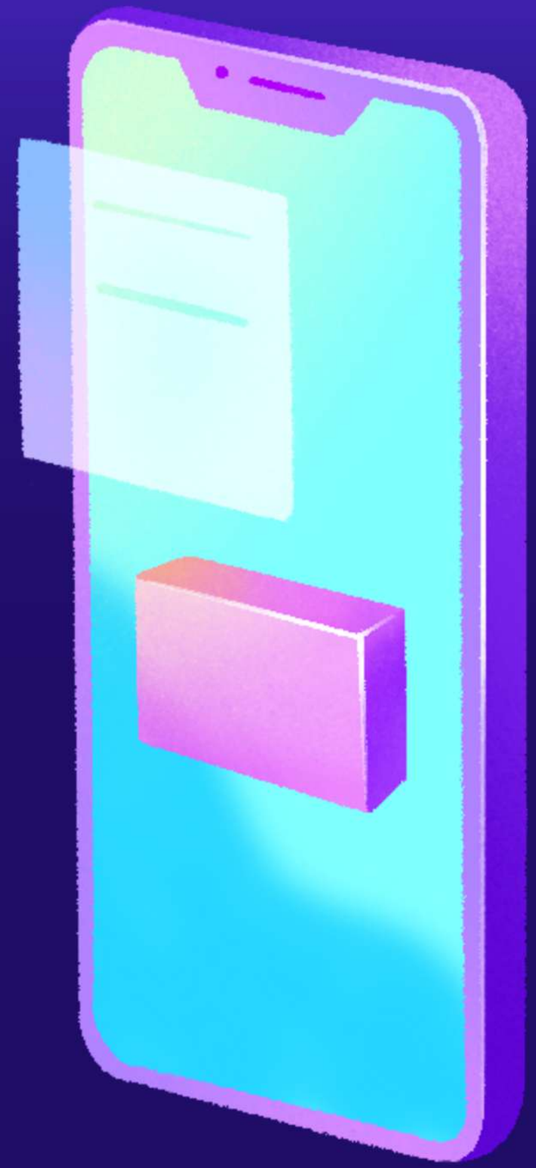
DEMO

REPO LINK



Q/A

...



THANK YOU FOR YOUR ATTENTION