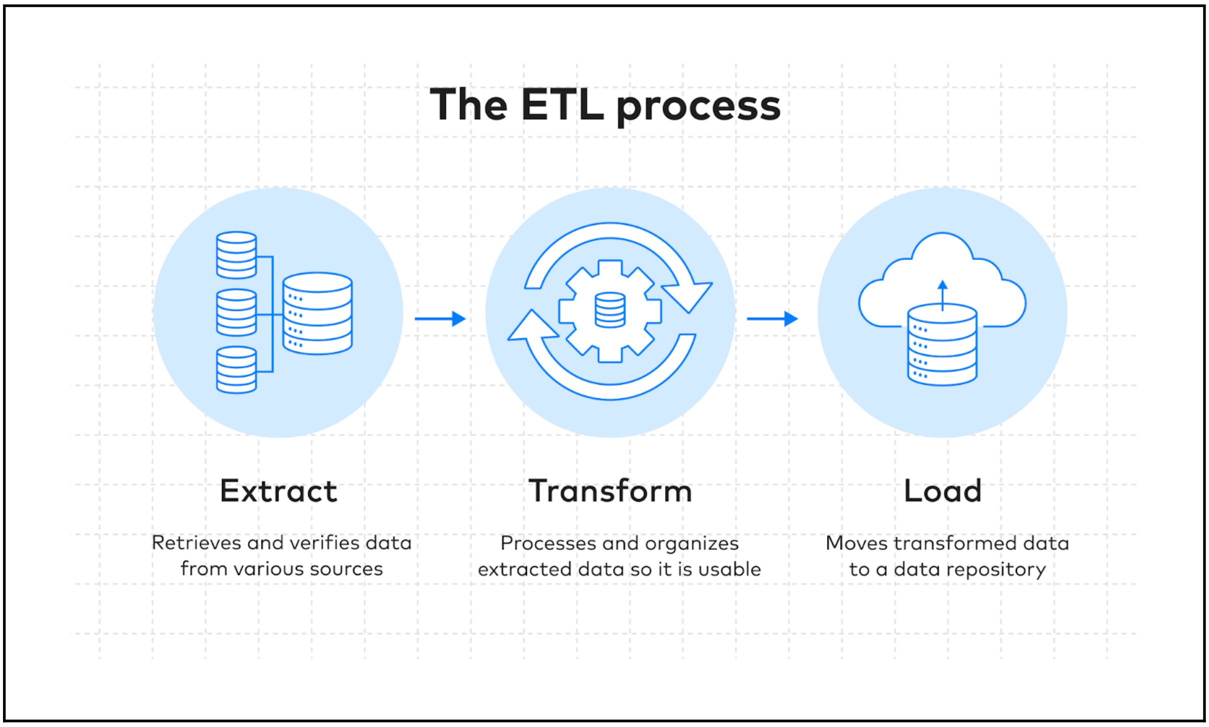
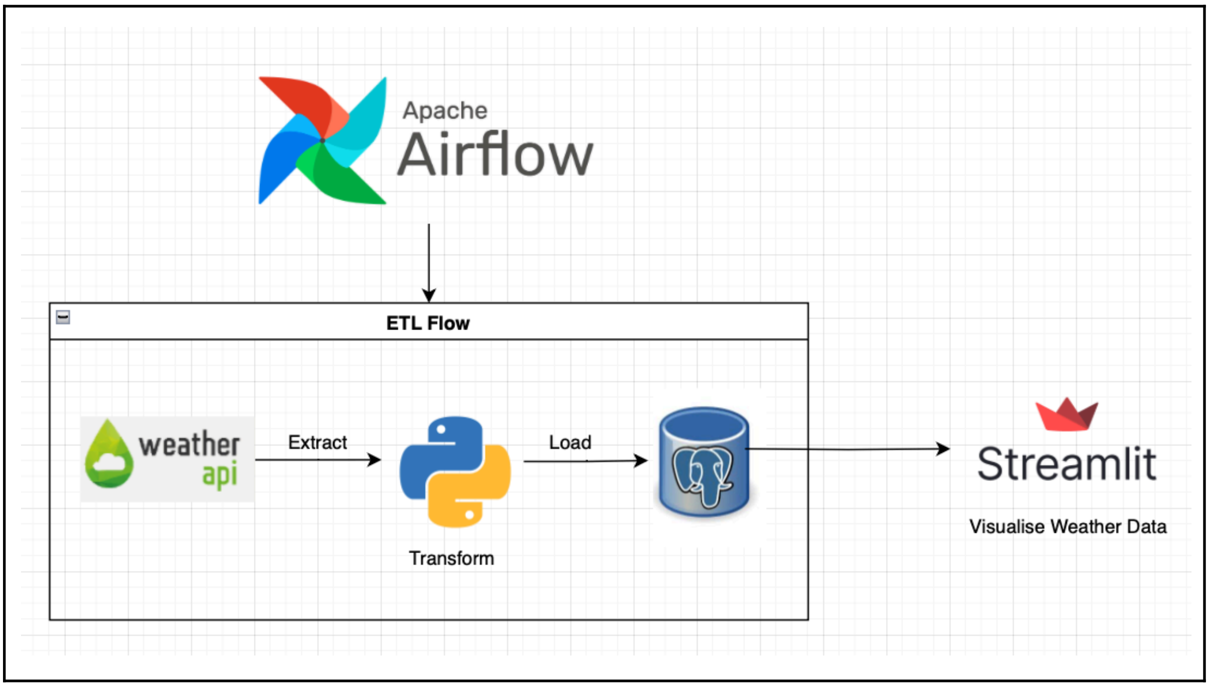
ETL Pipeline

In a data science project, once the required data or data source (API's Internal Database or IOT devices) is identified to solve the problem, then they give the data acquisition task to big data or data engineering team where they are going to have to create an entire data pipeline. This data pipeline follows a process known as ETL pipeline process. This ETL pipeline is only just a part of the data pipeline.

The main task of this pipeline is to integrate or combine all these data sources together, then do some preprocessing or transformation and finally load the transformed and preprocessed data to a single source like (SQL, Mongodb or Postgres), which will have the entire data together. Basically the task is to extract, transform and load the data that's why it's called ETL (Extract, Transform & Load).


The ETL process
Extract
Retrieves and verifies data from various sources
Transform
Processes and organizes extracted data so it is usable
Load
Moves transformed data to a data repository

Apache Airflow is a widely used orchestration tool for building, scheduling, and monitoring ETL (Extract, Transform, Load) and ELT data pipelines.



With airflow, we can schedule the entire ETL Process.

.