Data Pipelines and Workflow Orchestration

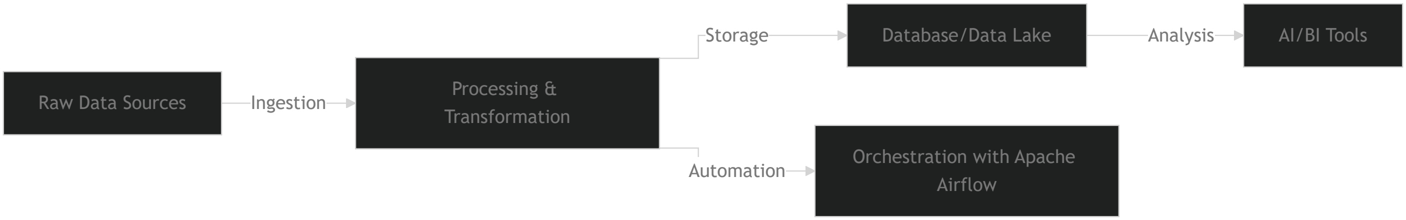
Introduction to Data Pipelines and Workflow Orchestration

What is a Data Pipeline?

A **data pipeline** is a series of steps that move and transform data from one system to another. In modern AI and data-driven applications, pipelines ensure that data is collected, cleaned, transformed, and stored efficiently.

Key Components of a Data Pipeline:

1. **Data Ingestion** - Collecting raw data from multiple sources (APIs, databases, files, etc.).
2. **Processing & Transformation** - Cleaning, filtering, aggregating, and preparing data for use.
3. **Storage** - Saving processed data in a database, data warehouse, or lake.
4. **Analysis & Visualization** - Using data for machine learning, reporting, or dashboards.
5. **Automation & Orchestration** - Managing dependencies, scheduling, and ensuring smooth execution.



Why Are Data Pipelines Important?

- **Efficiency:** Automates data flow, reducing manual work.
- **Scalability:** Handles large volumes of data reliably.
- **Consistency:** Ensures accurate, well-structured data for analysis.
- **Integration:** Connects different data sources seamlessly.
- **AI & Machine Learning:** Provides high-quality data for predictive modeling.

Workflow Orchestration: Managing Complexity

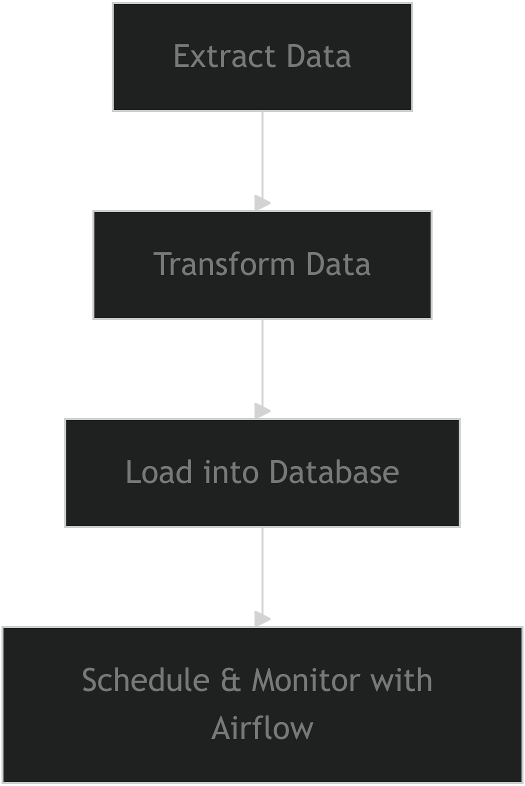
Workflow orchestration tools help automate and monitor data pipelines, ensuring that tasks run in the correct order.

Popular Orchestration Tools:

- **Apache Airflow** - Python-based task scheduler for complex workflows.
- **Prefect** - Dataflow automation with built-in fault tolerance.
- **Luigi** - Task dependency management for ETL pipelines.

How Orchestration Works:

1. **Task Scheduling** - Defining execution order & dependencies.
2. **Error Handling** - Automatic retries, alerts, and logging.
3. **Monitoring & Logging** - Tracking pipeline performance and failures.



How This Ties Into Your Learning Path

This introduction lays the foundation for the rest of this lesson where we will:

- Run a **hands-on ETL pipeline** using the NYC Taxi dataset.
- Learn how to **optimize & troubleshoot pipelines**.
- Prepare for **building your own pipeline with Apache Airflow** in the upcoming lab.

< Previous

Attributions

Next >