APACHE AIRFLOW

Introduction:

Apache Airflow is an open-source **workflow orchestration platform** developed by Airbnb in 2014 and later donated to the **Apache Software Foundation**. It is designed to **author, schedule, and monitor complex workflows** (known as DAGs – Directed Acyclic Graphs).

Airflow allows data engineers and developers to build workflows as **Python code**, making them dynamic, testable, and maintainable. It is widely used in **data engineering, ETL pipelines, machine learning workflows, and cloud automation**.

Features of Apache Airflow

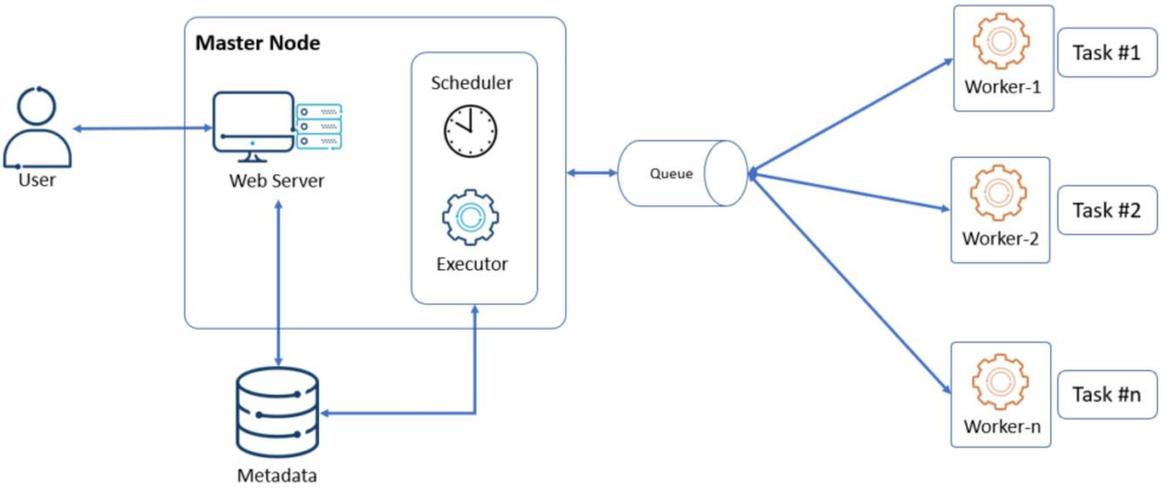
- **Open-source & Scalable** → Highly extensible, supports distributed execution.
- **Python-based** → DAGs are defined using Python code.
- **UI Monitoring** → Rich web interface to monitor, trigger, and retry workflows.
- **Dynamic Workflows** → Supports loops, conditionals, and parametrization.
- **Scheduler** → Handles job execution based on defined schedules or triggers.
- **Plugins & Integrations** → Works with databases, cloud providers (AWS, GCP, Azure), and messaging systems.
- **Task Dependencies** – Easily manage task order using operators (>>, <<).
- **Retries & Alerts** – Auto-retry failed tasks and send alerts (Email/Slack).
- **Extensible Plugins** – Add custom operators and integrations.

Apache Airflow

Airflow Architecture

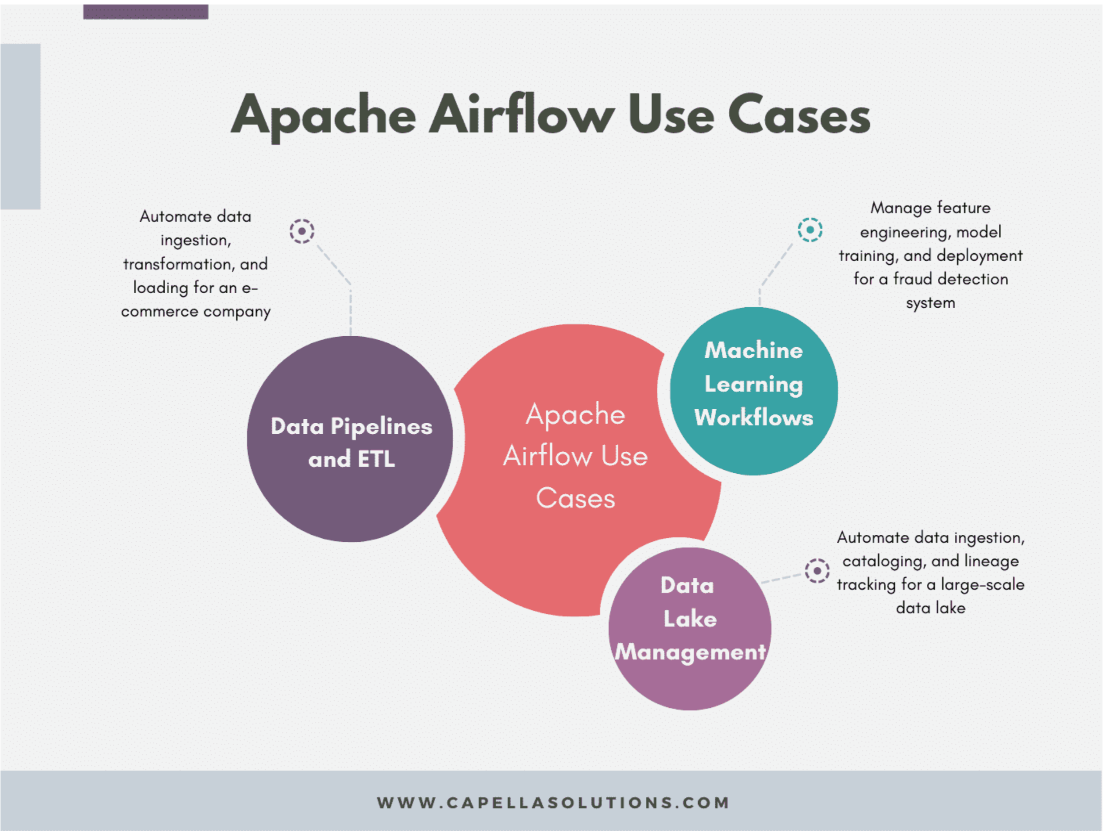
Airflow consists of several key components:

- **Web Server** → Provides a UI to monitor and manage workflows.
- **Scheduler** → Decides what tasks need to run and when.
- **Executor** → Runs the actual tasks (e.g., LocalExecutor, CeleryExecutor, KubernetesExecutor).
- **Metadata Database** → Stores DAGs, task status, logs, and user info.
- **Workers** → Execute the tasks assigned by the scheduler.



Use Cases of Airflow

- **ETL Pipelines** → Extracting, transforming, and loading data into data warehouses.
- **Machine Learning** → Training and deploying ML models.
- **Data Processing** → Managing Spark/Hadoop jobs.
- **Cloud Automation** → Orchestrating cloud resources and APIs.
- **Business Workflows** → Report generation, monitoring, and alerting.



Apache Airflow Use Cases

Automate data ingestion, transformation, and loading for an e-commerce company

Manage feature engineering, model training, and deployment for a fraud detection system

Automate data ingestion, cataloging, and lineage tracking for a large-scale data lake

WWW.CAPELLASOLUTIONS.COM