Day 14

hamza - Rehman

Apache AirFlow is an open-source platform used to programmatically author, schedule, and monitor complex workflows. It is designed to help you create and manage data pipelines, which are sequence of Tasks that process and transform data from one state to another.

purpose of Apache AirFlow:

The main purpose of AirFlow is to provide a flexible and scalable way to manage workflows and data pipelines. Here is how it achieves that:

① workflow Orchestration:

AirFlow allows you to define workflows as Directed Acyclic Graphs (DAGs). A DAG is a collection of tasks organized in a way that specifies their dependencies and execution order.

② Task management:

You can break down complex workflows into smaller, manageable tasks. Each task can be independently retried, scheduled, and monitored.

③ Scheduling:

Airflow provides robust scheduling capabilities, allowing you to run tasks at specific intervals, dates, or based on complex rules.

④ Monitoring and Logging:

Airflow offers built-in tools to monitor the progress of your workflow, track failure and view logs for debugging.

⑤ Extensibility:

With a modular Architecture Airflow can be extended through custom operators, hooks and executors to interact with different systems and data sources.

Benefits of Apache Airflow:

① Flexibility:

Airflows ability to define workflows as a code makes it highly flexible, allowing you to customize and adjust workflows easily.

② open-source:

As an open source tool, Airflow has a large and active community.

③ Integrations:

Airflow supports integrations with third-party services, databases and cloud providers, making it a versatile choice for various data engineering task.

④ Dynamic pipelines:

You can create dynamic pipelines that change based on external conditions, such as checking for the availability of data or the outcome of previous tasks.

⑤ Visualization:

AirFlow web based user interface offers clear visualization of DAGs, allowing users to track progress and analyze task dependencies easily.

History:

Apache AirFlow was originally developed by Airbnb in October 2014 to manage the company's complex workflows and data pipelines.

It was released as an open-source project under the Apache License in 2015.

Features of Apache AirFlow

① DAGs Based Architecture:

② Extensible operators and Hooks.

③ Scheduling and Triggers

④ Task dependencies

⑤ Web Based User Interface

⑥ Task Retry and alerting
⑦ pluggable Executors.
⑧ configuration as code

Use Cases of AirFlow

① Data Engineering and ETL Pipelines:

AirFlow is commonly used to build ETL pipelines for processing and moving data between systems.

② Machine learning workflows:

it can orchestrate machine learning workflows from data processing to model training and deployment.

③ Batch processing:

AirFlow suitable