Airflow

Name: Priyeshwar

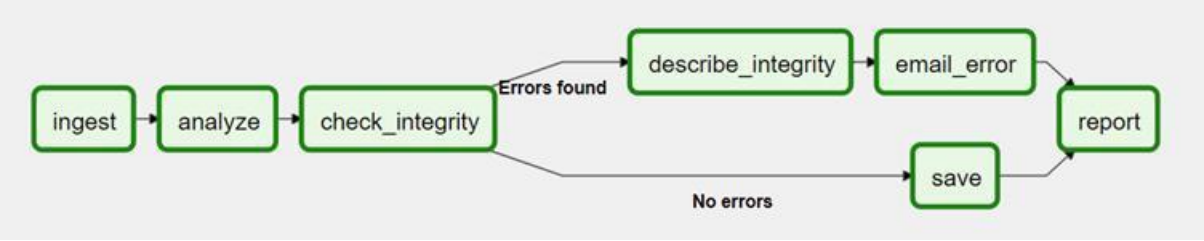
Mail: priyesh2664@gmail.com

What is Airflow?

Apache Airflow is an open-source workflow orchestration tool used by data engineers to programmatically author, schedule, and monitor workflows (pipelines). Workflows are defined in Python code and represented as Directed Acyclic Graphs (DAGs).

Airflow Architecture

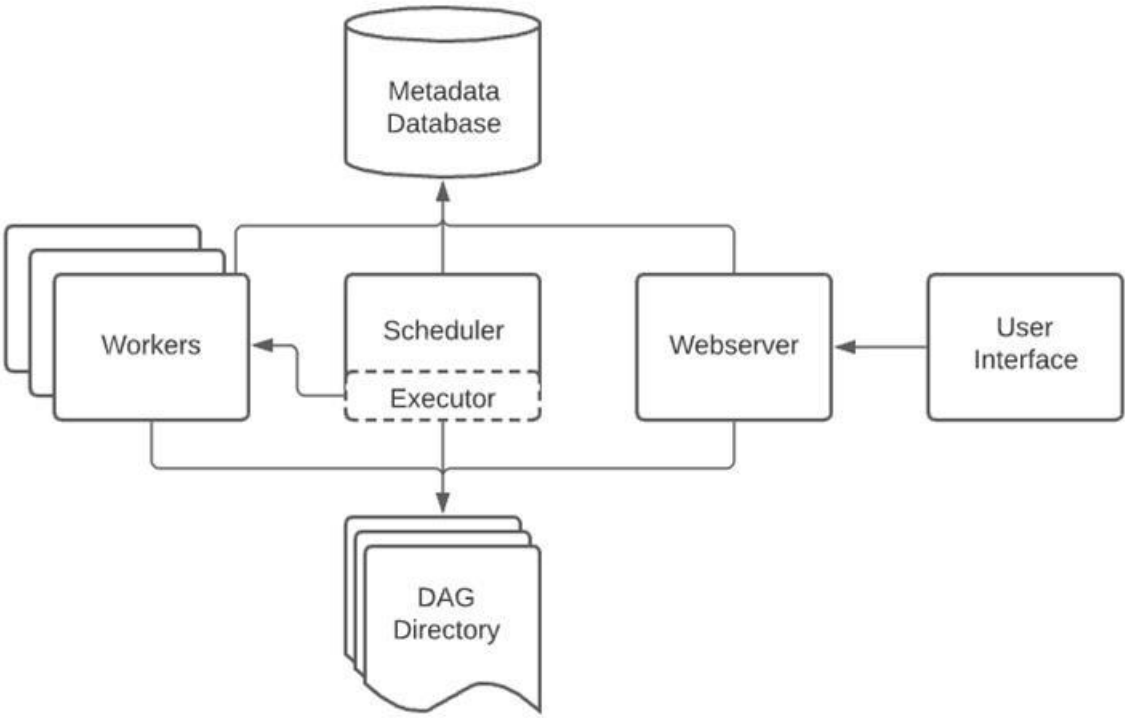
The Airflow platform lets you build and run workflows, which are represented as Directed Acyclic Graphs (DAGs). A sample DAG is shown in the diagram below.



DAG

A DAG contains Tasks (action items) and specifies the dependencies between them and the order in which they are executed. A Scheduler handles scheduled workflows and submits Tasks to the Executor, which runs them. The Executor pushes tasks to workers.

Other typical components of an Airflow architecture include a database to store state metadata, a web server used to inspect and debug Tasks and DAGs, and a folder containing the DAG files.



Control Flow

In Apache Airflow, control flow defines how tasks within a DAG are executed and related. A DAG can run multiple times, and multiple DAG runs can also happen in parallel. Every DAG must include the mandatory parameter execution_date, along with other optional parameters that define its behavior.

Dependencies between tasks are represented using the symbols >> (downstream) and << (upstream), for example:

first_task >> [second_task, third_task] or third_task << fourth_task.

By default, tasks wait for all their upstream tasks to succeed before running, but this behavior can be customized using features like LatestOnly (to run tasks only in the most recent DAG run), Branching (to choose different execution paths dynamically), and Trigger Rules (to allow tasks to run under different conditions, such as when some upstream tasks fail). To manage complexity, Airflow also provides SubDAGs, which embed reusable DAGs inside others, and TaskGroups, which allow visually grouping related tasks in the user interface.

Apache Airflow Setup with Docker & VS Code

1. Install Required Tools

- **Docker Desktop** → Download here
- **Visual Studio Code** → Download here

Make sure Docker Desktop is **running** before continuing.

---

2. Prepare Materials Folder

1. Create a directory:
2. C:\Users\<your_username>\materials
3. Save a file named **docker-compose.yaml** in that folder.
4. Open **Visual Studio Code** → File → Open Folder → select the materials folder.

---

3. Create Environment File

Inside materials, create a file named **.env** with these contents:

AIRFLOW_IMAGE_NAME=apache/airflow:2.4.2

AIRFLOW_UID=50000

---

4. Start Airflow

In **VS Code terminal**, run:

docker-compose up -d

This will download Airflow images and start containers (webserver, scheduler, worker, etc.).

---

5. Create Admin User

Once containers are running, create an **admin user**:

docker-compose run airflow-worker airflow users create \

--role Admin \

--username admin \

--email admin@example.com \

--firstname admin \

--lastname admin \

--password admin

---

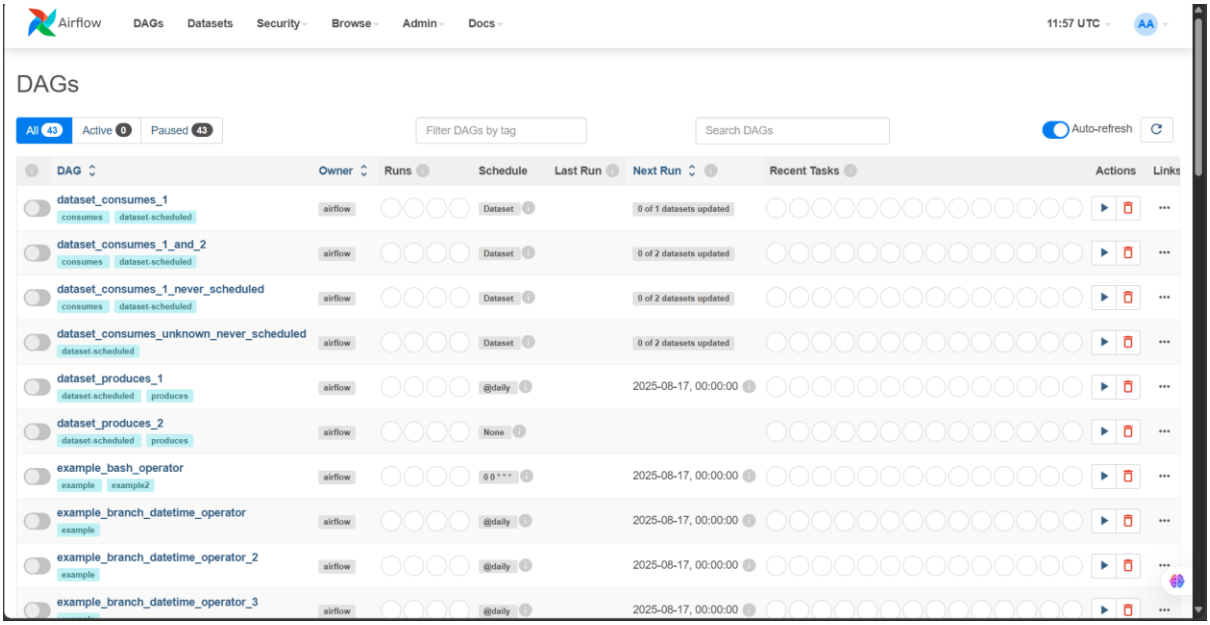
6. Access the Airflow UI

Open your browser and go to:

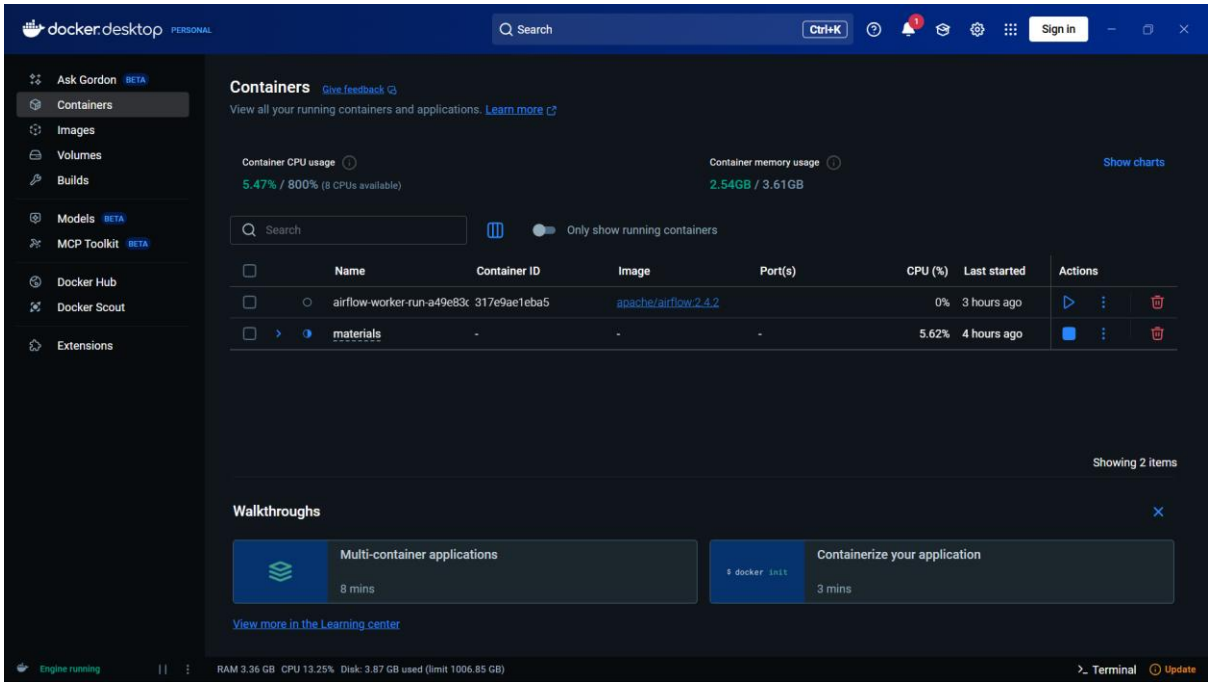
http://localhost:8080

Log in with:

- **Username:** admin
- **Password:** admin



At this point, Airflow should be up and running inside Docker.


docker desktop PERSONAL
Search
Ctrl+K
Sign in
Ask Gordon BETA
Containers
Images
Volumes
Builds
Models BETA
MCP Toolkit BETA
Docker Hub
Docker Scout
Extensions
Containers
Give feedback
View all your running containers and applications. Learn more
Container CPU usage
5.47% / 800% (8 CPUs available)
Container memory usage
2.54GB / 3.61GB
Show charts
Search
Only show running containers
Name
Container ID
Image
Port(s)
CPU (%)
Last started
Actions
airflow-worker-run-a49e83c
317e9ae1eba5
apache/airflow:2.4.2
0%
3 hours ago
materials
-
-
-
5.62%
4 hours ago
Showing 2 items
Walkthroughs
Multi-container applications
8 mins
$ docker init
Containerize your application
3 mins
View more in the Learning center
Engine running
RAM 3.36 GB CPU 13.25% Disk: 3.87 GB used (limit 1006.85 GB)
Terminal
Update