synpulse

solve · evolve

Airflow Introduction

James Ng

2023-06-27

Agenda

01 | **What is Airflow?**

02 | **Python scripts**

03 | **Resources**

What is Airflow?

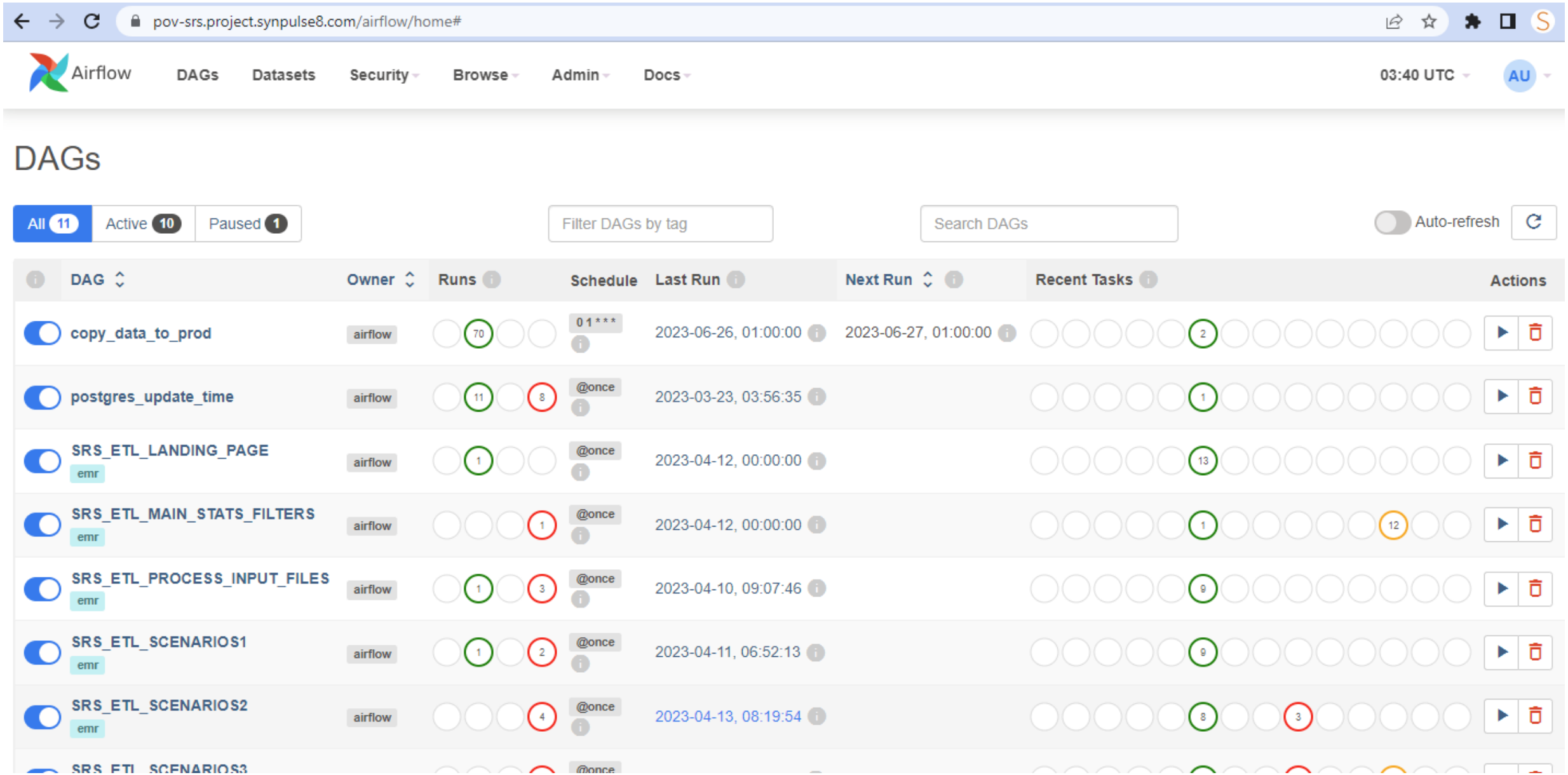
01

What is Airflow?

Apache Airflow
- orchestrate/automate workflows
- schedule workflows
- execute workflows, including (Extract, Transform, Load) ETL pipelines

History
- October 2014: developed by Airbnb
- January 2019: open-sourced under the Apache Software Foundation
  https://github.com/apache/airflow

Airflow allows you to define, schedule, and monitor complex workflows as Directed Acyclic Graphs (DAGs).
Airflow workflows are created via Python scripts, made easy by importing libraries and classes.
More info: Wikipedia



Key components

- **Scheduler**
  To manage the execution of tasks based on their dependencies (workflow) and the defined schedule.
  It determines which tasks are ready to run and triggers their execution.

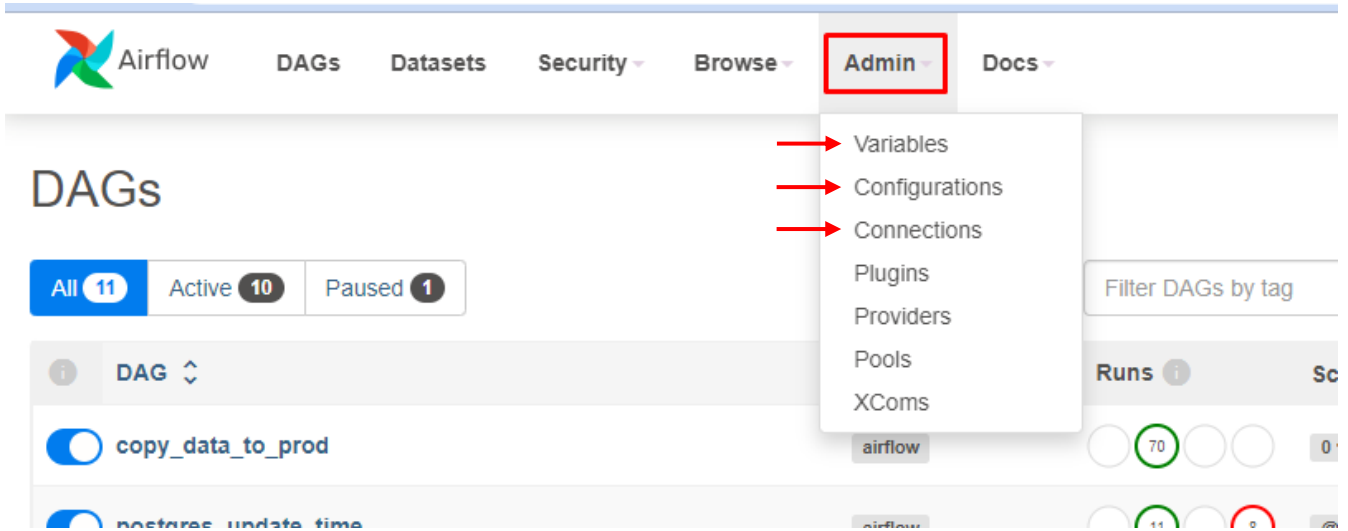
- **Web Server**
  To provide a web-based user interface (UI) for interacting with Airflow.
  It allows users to monitor workflows, view logs, and manage DAGs, tasks, and connections.

- **Metadata Database**
  To store the state of workflows, task instances, and other operational metadata.
  This database keeps track of the execution history, task status, and other relevant information.

Python scripts

02

Configuration



- Airflow → Admin → **Configurations** (view only)
  Airflow's configuration is managed through a configuration file: `airflow.cfg` which is usually located in the AIRFLOW_HOME directory (default is ~/airflow)

- Airflow → Admin → **Variables**
  To store common variables used by any DAG

- Airflow → Admin → **Connections**
  To securely store username and password

Python code

3 main script sections

- import libraries (Airflow operators, sensors, etc)
- set DAG variables

- define DAG functions
    - DAG comprises 1 or more task(s)
    - Each task in the DAG is implemented as an operator, which defines what the task does

- instantiate DAG objects
- set task dependencies (workflow)

Resources

03

Installing Airflow (Windows, with WSL2)

Step-by-step instructions:
https://www.freecodecamp.org/news/install-apache-airflow-on-windows-without-docker/

Requirements:

- to enable WSL2 (Windows Subsystem for Linux 2)

- to install WSL2

Command Prompt → Run as administrator

```
> wsl --update

> wsl --install -d ubuntu
```

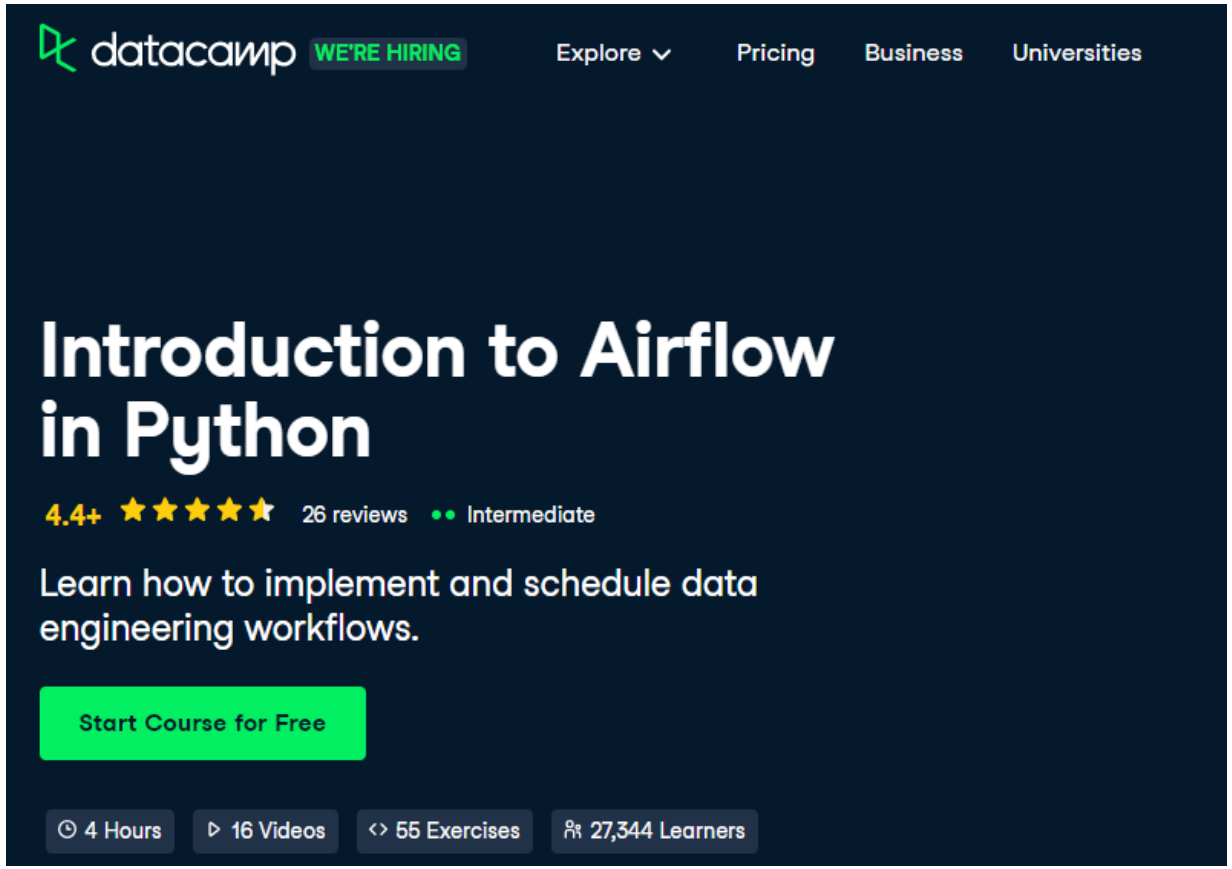
Installing Airflow with Docker (recommended)

Step-by-step instructions:
https://airflow.apache.org/docs/apache-airflow/stable/howto/docker-compose/index.html

Requirements:

- Install Docker Community Edition (CE) on your workstation

DataCamp: Introduction to Airflow



Course content:
https://github.com/JNYH/DataCamp_Introduction_to_Airflow/blob/master/Course_notes_solutions_answers_Introduction_to_Airflow.pdf

Thank You.

James Ng
Assistant Vice President

Synpulse Singapore Pte. Ltd.
Management Consulting
80 Amoy Street
Level 3
Singapore 069899
Singapore

+65 97677746

james.ng@synpulse.com

https://www.linkedin.com/in/jnyh/

synpulse
solve · evolve

Americas
New York
Toronto

Asia
Bangkok
Hong Kong
Hyderabad
Jakarta
Manila
Pune
Shenzhen
Singapore
Taipei

Europe
Zurich (Headquarters)
Bratislava
Dusseldorf
Geneva
London
Luxembourg

Oceania
Sydney