

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](#)



Airflow

DAGs

Datasets

Security

Browse

Admin

Docs

03:40 UTC

AU

DAGs

All 11 Active 10 Paused 1

Filter DAGs by tag

Search DAGs

Auto-refresh

DAG	Owner	Runs	Schedule	Last Run	Next Run	Recent Tasks	Actions
<input checked="" type="checkbox"/> copy_data_to_prod	airflow	<div><div>70</div></div>	0 1 ***	2023-06-26, 01:00:00	2023-06-27, 01:00:00	<div><div>2</div></div>	▶ 🗑️
<input checked="" type="checkbox"/> postgres_update_time	airflow	<div><div>11</div><div>8</div></div>	@once	2023-03-23, 03:56:35		<div><div>1</div></div>	▶ 🗑️
<input checked="" type="checkbox"/> SRS_ETL_LANDING_PAGE emr	airflow	<div><div>1</div></div>	@once	2023-04-12, 00:00:00		<div><div>13</div></div>	▶ 🗑️
<input checked="" type="checkbox"/> SRS_ETL_MAIN_STATS_FILTERS emr	airflow	<div><div>1</div></div>	@once	2023-04-12, 00:00:00		<div><div>1</div><div>12</div></div>	▶ 🗑️
<input checked="" type="checkbox"/> SRS_ETL_PROCESS_INPUT_FILES emr	airflow	<div><div>1</div><div>3</div></div>	@once	2023-04-10, 09:07:46		<div><div>9</div></div>	▶ 🗑️
<input checked="" type="checkbox"/> SRS_ETL_SCENARIOS1 emr	airflow	<div><div>1</div><div>2</div></div>	@once	2023-04-11, 06:52:13		<div><div>9</div></div>	▶ 🗑️
<input checked="" type="checkbox"/> SRS_ETL_SCENARIOS2 emr	airflow	<div><div>4</div></div>	@once	2023-04-13, 08:19:54		<div><div>8</div><div>3</div></div>	▶ 🗑️
<input checked="" type="checkbox"/> SRS_ETL_SCENARIOS3 emr	airflow	<div><div>1</div><div>1</div></div>	@once			<div><div>1</div><div>1</div></div>	▶ 🗑️

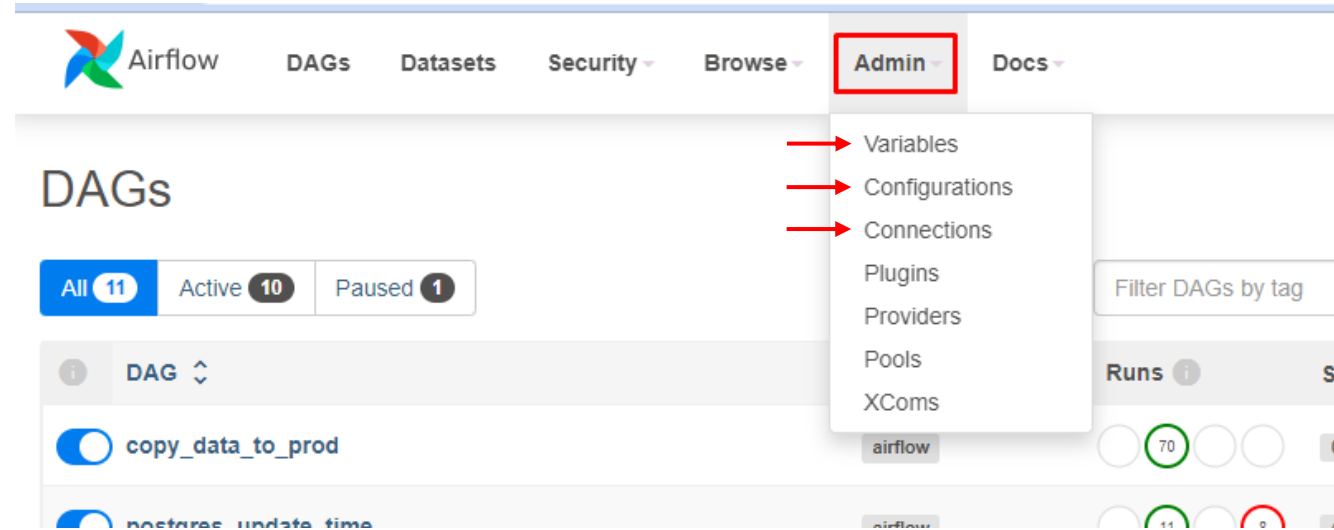
Key components

- **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:

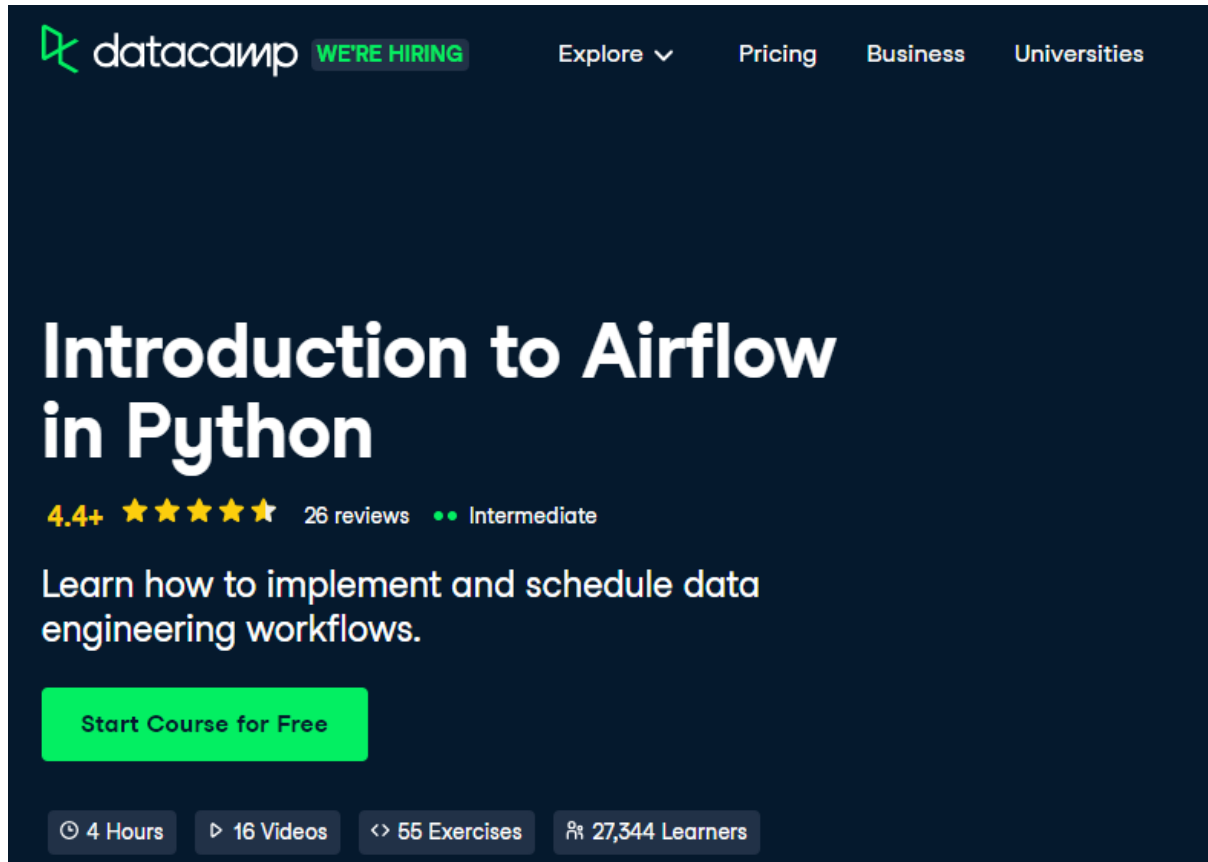
- to enable WSL2 (Windows Subsystem for Linux 2)
- to install WSL2

Command Prompt → Run as administrator

```
> wsl --update
```

```
> wsl --install -d ubuntu
```

DataCamp: [Introduction to Airflow](#)



Course content:

[https://github.com/JNYH/DataCamp Introduction to Airflow/blob/master/Course notes solutions answers
_Introduction to Airflow.pdf](https://github.com/JNYH/DataCamp%20Introduction%20to%20Airflow/blob/master/Course%20notes%20solutions%20answers_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/>



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