Getting started with Apache Airflow

Shyarnis Ghising

Dec 12, 2023

Contents

Introduction

- Apache Airflow is a **batch-oriented** workflow for building data pipelines.
- It enables engineers to easily build **scheduled data pipelines** using a flexible Python framework.
- It **orchestrates** the different components responsible for processing data in data pipelines[1].

Introduction

- Apache Airflow is a **batch-oriented** workflow for building data pipelines.
- It enables engineers to easily build **scheduled data pipelines** using a flexible Python framework.
- It **orchestrates** the different components responsible for processing data in data pipelines[1].

Introduction

- Apache Airflow is a **batch-oriented** workflow for building data pipelines.
- It enables engineers to easily build **scheduled data pipelines** using a flexible Python framework.
- It **orchestrates** the different components responsible for processing data in data pipelines[1].

Data Pipelines

- It consists of **several tasks** that needed to be executed.
- Tasks need to be executed in a **specific order**.

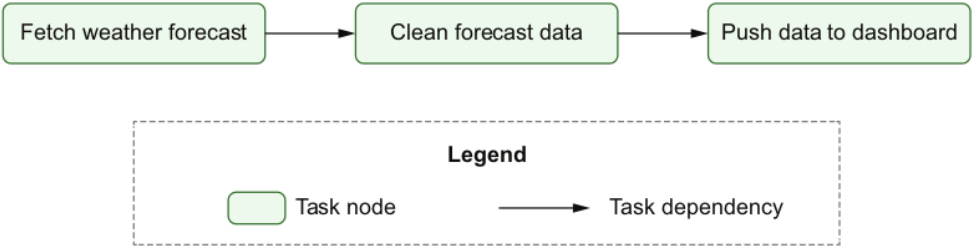


Figure: Data Pipeline for the weather dashboard

Data pipelines as graphs

- **Tasks** are represented by nodes/ vertices.
- **Dependenices between tasks** are represented by directed edges.

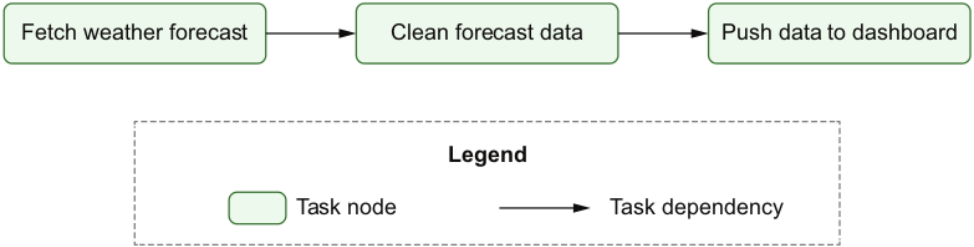
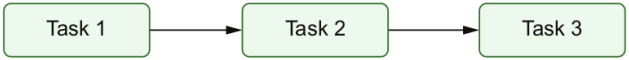


Figure: Data Pipeline represented as DAG

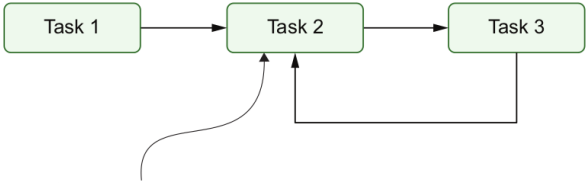
- Such graphs are called *directed acyclic graph* **DAG**.

A directed *acyclic* graph (DAG) of tasks



- Directed cyclic graph leads to **deadlock** situation.

A directed *cyclic* graph of tasks



Task 2 will never be able to execute, due to its dependency on task 3, which in turn depends on task 2.

DAG in Python Code

- Python provide flexibility for building DAGs.

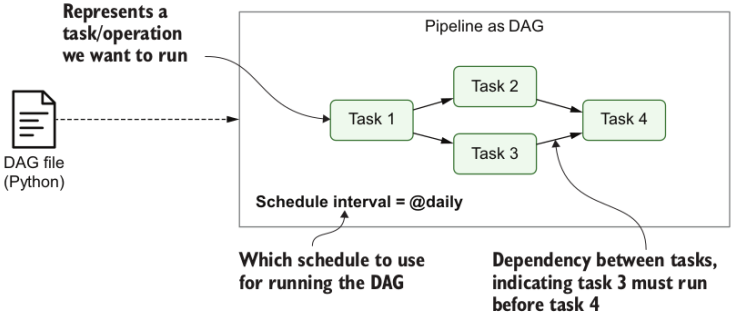


Figure: Pipelines are defined as DAGs using Python code

Installation

Install Airflow on your machine.

Example

```
 - AIRFLOW_VERSION==2.7.2
- PYTHON_VERSION==3.8
- pip install "apache-airflow==2.7.2" --constraint
"https://raw.githubusercontent.com/apache/airflow/
constraints-2.7.2/constraints-no-providers-3.8.txt"
```

constraints

- airflow db migrate
- `airflow users create --username <usernname> --password <password> --firstname <fanme> --lastname <lname> --role Admin --email <email>`
- airflow scheduler
- airflow webserver

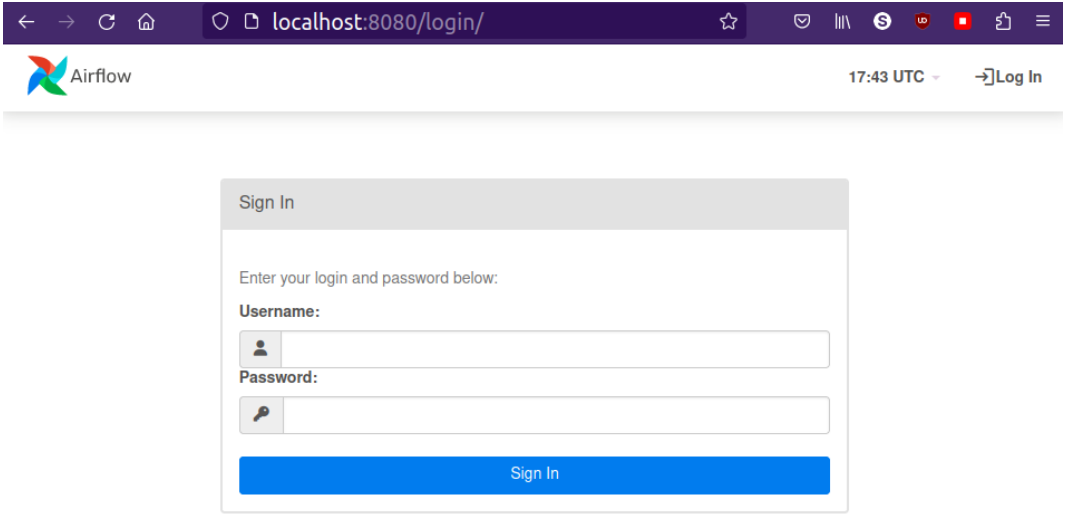


Figure: Airflow login view

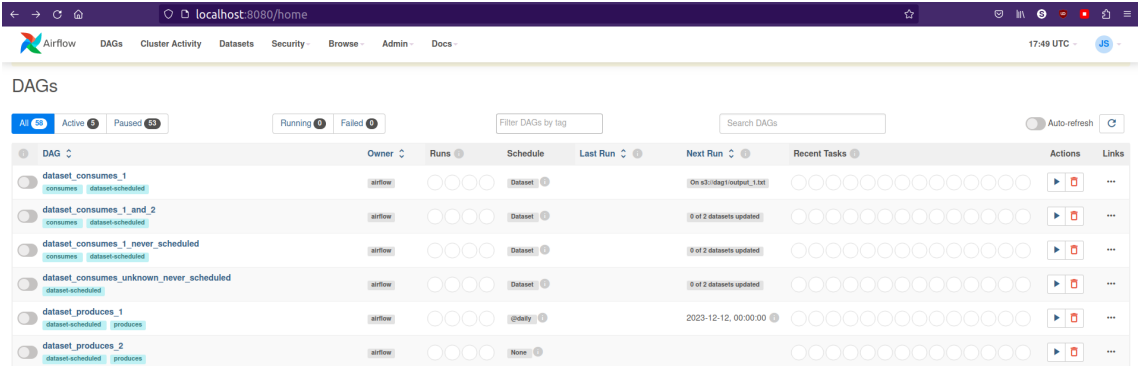


Figure: List of Airflow DAG

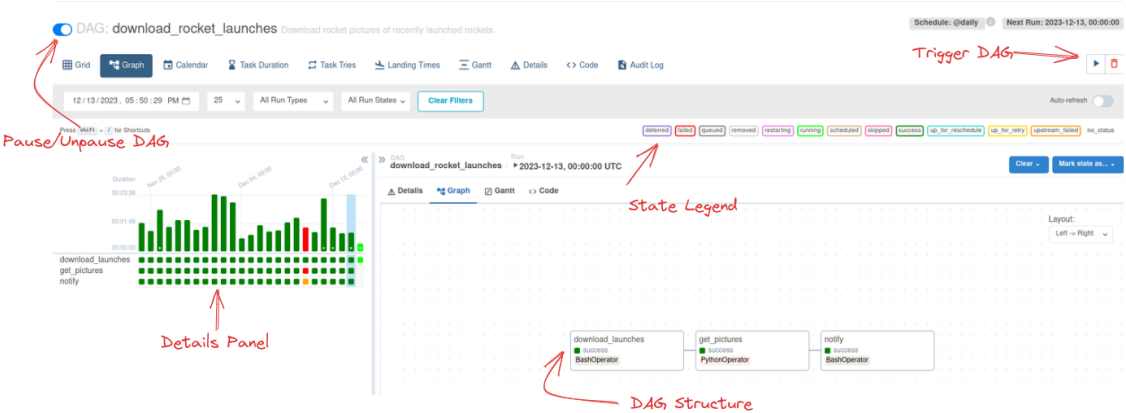


Figure: Airflow DAG in Action

Live Demonstration

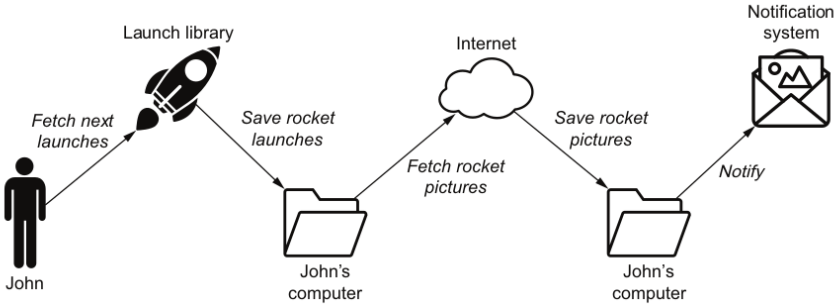


Figure: John's mental model of downloading rocket pictures

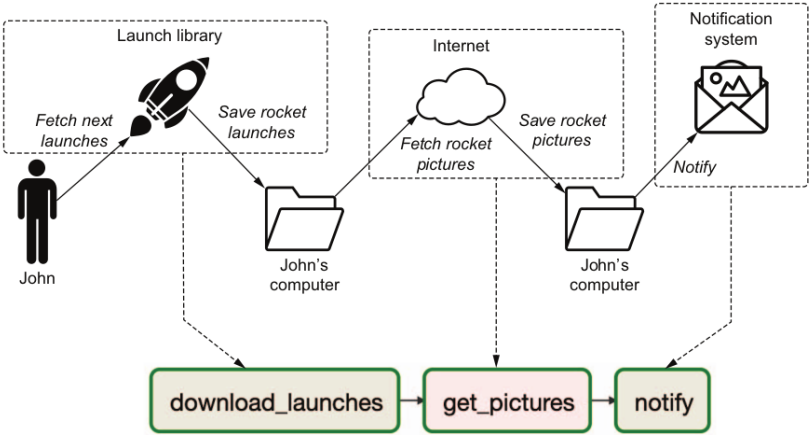


Figure: John's mental model mapped to tasks in Airflow

References

Ruiter Harenslak.
Data Pipelines with Apache Airflow.
Manning, 2020.