Clear your schedule: Running Data Science with python and airflow

•••

Jonathan Stott
mago

@namelessjon

Production?

@namelessjon

Photo by Birmingham Museums Trust on Unsplash

Schedule?

```
* * * * *
1 * * * *
5 8 * * *
```

@namelessjon

Photo by Amanda Jones on Unsplash

More Schedules

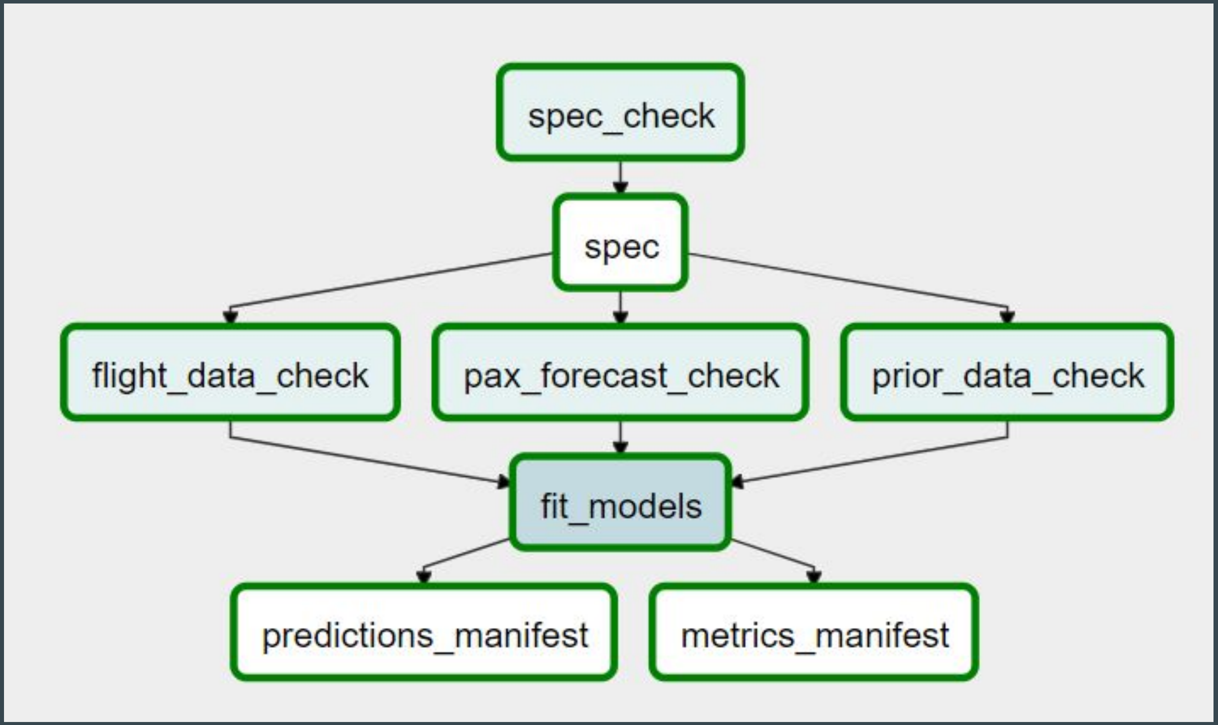
@namelessjon

Photo by Ahmad Ossayli on Unsplash

Enter: Airflow

```
$ pip install apache-airflow
```

Directed Acyclic Graphs



Operators

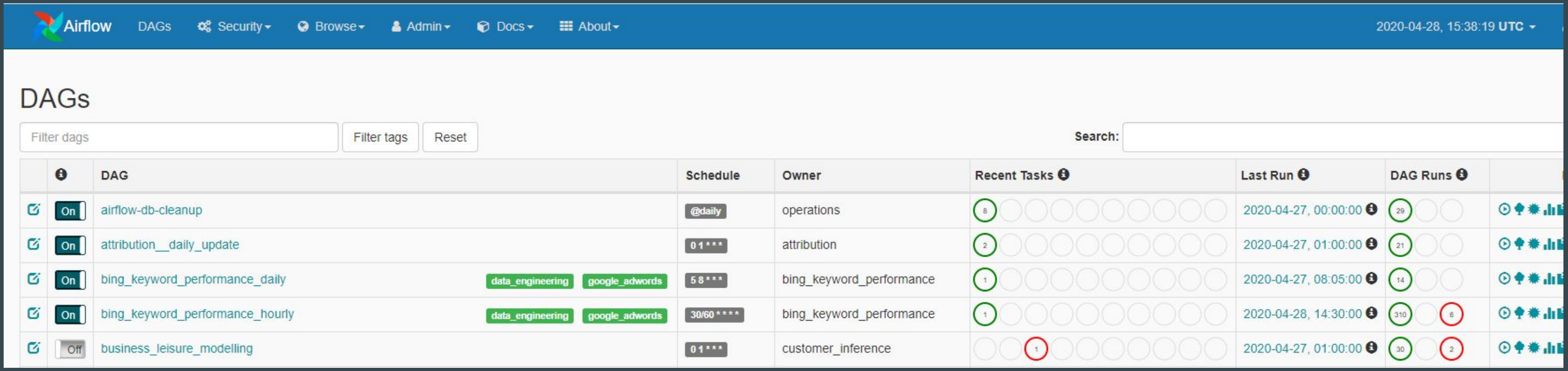
Photo by Science in HD on Unsplash

Sensors

@namelessjon

Photo by eberhard grossgasteiger on Unsplash

UI / Dashboard

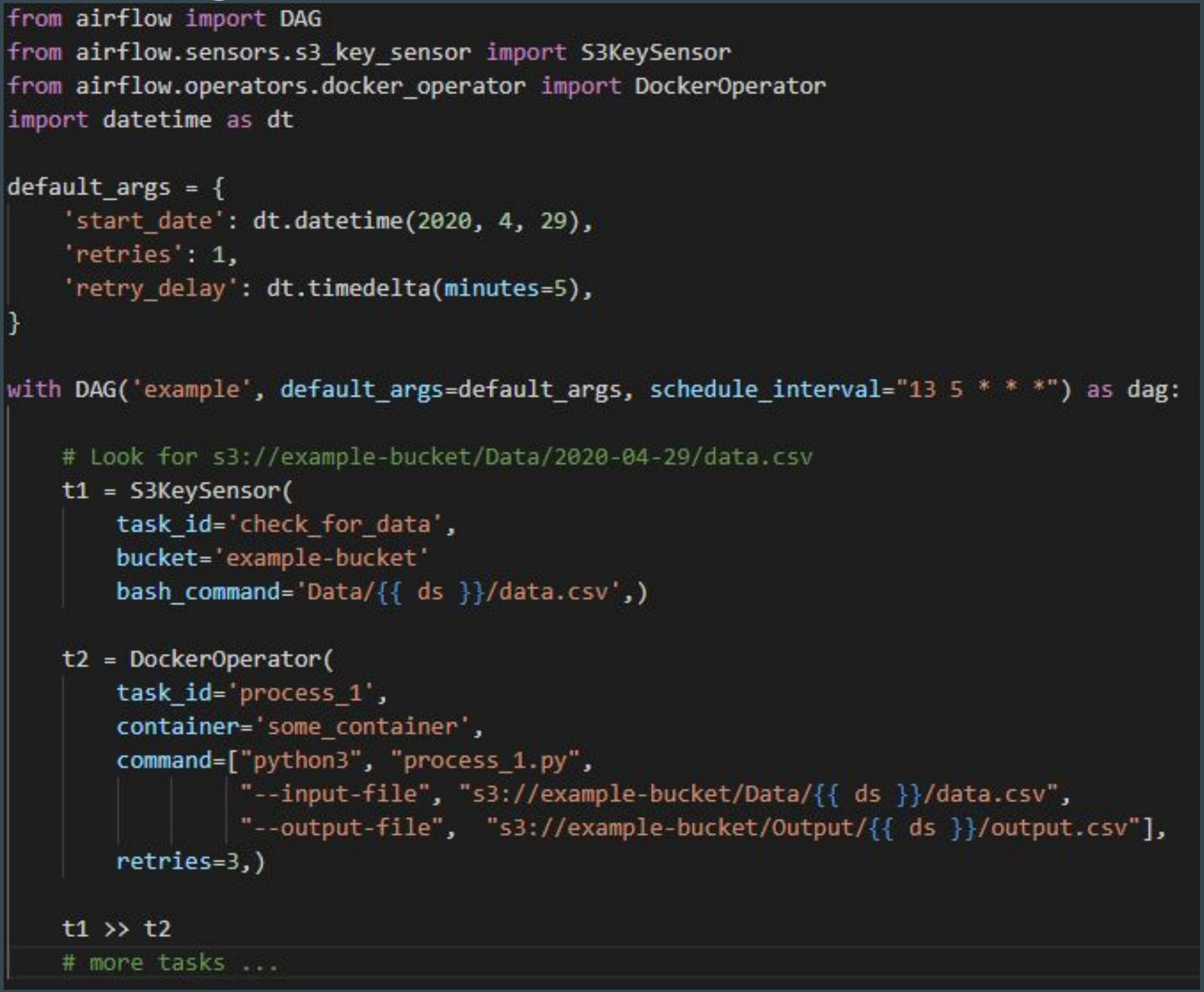


DAGs

		DAG		Schedule	Owner	Recent Tasks	Last Run	DAG Runs
	On	airflow-db-cleanup		@daily	operations	8	2020-04-27, 00:00:00	29
	On	attribution__daily_update		0 1 * * *	attribution	2	2020-04-27, 01:00:00	21
	On	bing_keyword_performance_daily	data_engineering google_adwords	5 8 * * *	bing_keyword_performance	1	2020-04-27, 08:05:00	14
	On	bing_keyword_performance_hourly	data_engineering google_adwords	30/60 * * * *	bing_keyword_performance	1	2020-04-28, 14:30:00	310 6
	Off	business_leisure_modelling		0 1 * * *	customer_inference	1	2020-04-27, 01:00:00	30 2

@namelessjon

Building a workflow



```python
from airflow import DAG
from airflow.sensors.s3_key_sensor import S3KeySensor
from airflow.operators.docker_operator import DockerOperator
import datetime as dt

default_args = {
    'start_date': dt.datetime(2020, 4, 29),
    'retries': 1,
    'retry_delay': dt.timedelta(minutes=5),
}

with DAG('example', default_args=default_args, schedule_interval="13 5 * * *") as dag:

    # Look for s3://example-bucket/Data/2020-04-29/data.csv
    t1 = S3KeySensor(
        task_id='check_for_data',
        bucket='example-bucket'
        bash_command='Data/{{ ds }}/data.csv',)

    t2 = DockerOperator(
        task_id='process_1',
        container='some_container',
        command=["python3", "process_1.py",
                 "--input-file", "s3://example-bucket/Data/{{ ds }}/data.csv",
                 "--output-file",  "s3://example-bucket/Output/{{ ds }}/output.csv"],
        retries=3,)

    t1 >> t2
    # more tasks ...
```

@namelessjon

Other helpful features

- Task logging
- Connections
- Alerting
- Celery
- SLAs
- ... and more!

Use operators where you can

Use containers for scripts

@namelessjon

Photo by Guillaume Bolduc on Unsplash

Keep business logic out of the DAG file

@namelessjon

Photo by Sarah Shaffer on Unsplash

Remember DAGs are python

Summary

- Use airflow to manage workflows with dependencies between tasks
- Docker is great for packing up tasks in workflow
- Start small and scale up!

Thank you

Links

https://airflow.apache.org/

https://towardsdatascience.com/best-practices-for-airflow-developers-990c8a04f7c6

https://medium.com/wbaa/datas-inferno-7-circles-of-data-testing-hell-with-airflow-cef4adff58d8