Getting to Know AIRFLOW

Rosie Hoyem
PyMNtos
04/27/2017

Me. →

- Data Scientist
- Web Developer
- Landlord
- Cyclist
- Traveler

rosiehoyem@gmail.com

rosiehoyem.com

0.

Airflow huh?

???

Airflow In a Nutshell

⊙ Data Engineering tool

⊙ Pimped out Flask app

⊙ Useful for building functional data pipelines and automating workflow

1.

Why do I care?

It’s Popular.

History of Airflow

2014
Maxime Beauchemin began building a tool at Airbnb in October of 2014

2016
Airflow entered incubation as an Apache project

Now
Officially used by dozens of companies large and small

Who Already uses it?

Airbnb [@mistercrunch, @artwr]

Agari [@r39132]

allegro.pl [@kretes]

AltX [@pedromduarte]

Apigee [@btallman]

Astronomer [@schnie]

Auth0 [@sicarul]

BandwidthX [@dineshdsharma]

Bellhops

BlaBlaCar [@puckel & @wmorin]

Bloc [@dpaola2]

BlueApron [@jasonjho & @matthewdavidhauser]

Blue Yonder [@blue-yonder]

Celect [@superdosh & @chadcelect]

Change.org [@change, @vijaykramesh]

Children's Hospital of Philadelphia Division of Genomic Diagnostics [@genomics-geek]

City of San Diego [@MrMaksimize, @andrell81 & @arnaudvedy]

Clairvoyant @shekharv

Clover Health [@gwax & @vansivallab]

Chartboost [@cgelman & @dclubb]

Cotap [@maraca & @richardchew]

Digital First Media [@duffn & @mschmo & @seanmuth]

Easy Taxi [@caique-lima & @WesleyBatista]

FreshBooks [@DinoCow]

Gentner Lab [@neuromusic]

Glassdoor [@syvineckruyk]

HelloFresh [@tammymendt & @davidsbatista & @iuriinedostup]

Holimetrix [@thibault-ketterer]

Hootsuite

IFTTT [@apurvajoshi]

iHeartRadio[@yiwang]

ING

Jampp

Kiwi.com [@underyx]

Kogan.com [@geeknam]

Lemann Foundation [@fernandosjp]

LendUp [@lendup]

liligo [@tromika]

LingoChamp [@haitaoyao]

Lucid [@jbrownlucid & @kkourtchikov]

Lumos Labs [@rfroetscher & @zzztimbo]

Lyft[@SaurabhBajaj]

Madrone [@mbreining & @scotthb]

Markovian [@al-xv, @skogsbaeck, @waltherg]

Mercadoni [@demorenoc]

MiNODES [@dice89, @diazcelsa]

MFG Labs

mytaxi [@mytaxi]

Nerdwallet

OfferUp

OneFineStay [@slangwald]

Open Knowledge International @vitorbaptista

PayPal [@jhsenjaliya]

Postmates [@syeoryn]

Sense360 [@kamilmroczek]

Shopkick [@shopkick]

Sidecar [@getsidecar]

SimilarWeb [@similarweb]

SmartNews [@takus]

Spotify [@znichols]

Stackspace

Stripe [@jbalogh]

Thumbtack [@natekupp]

T2 Systems [@unclaimedpants]

Vente-Exclusive.com [@alexvanboxel]

Vnomics [@lpalum]

WePay [@criccomini & @mtagle]

WeTransfer [@jochem]

Whistle Labs [@ananya77041]

WiseBanyan

Wooga

Xoom [@gepser & @omarvides]

Yahoo!

Zapier [@drknexus & @statwonk]

Zendesk

Zenly [@cerisier & @jbdalido]

99 [@fbenevides, @gustavoamigo & @mmmaia]

GovTech GDS [@chrissng & @datagovsg]

Gusto [@frankhsu]

Handshake [@mhickman]

Handy [@marcintustin / @mtustin-handy]

Qubole [@msumit]

2

What is it?

A Brief Overview

Before Airflow, there was...

Cron Jobs.

(And a hodge-podge of other tools people would duct tape together.)

What’s a Cron Job you say?

Schedule Jobs

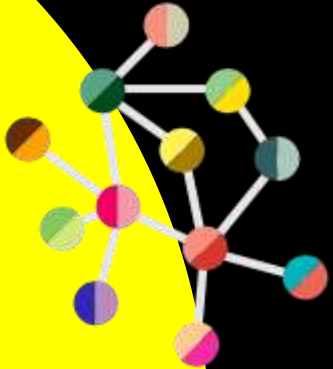
Cron

cron is a Linux utility which schedules a command or script on your server to run automatically at a specified time and date.

Cron Job

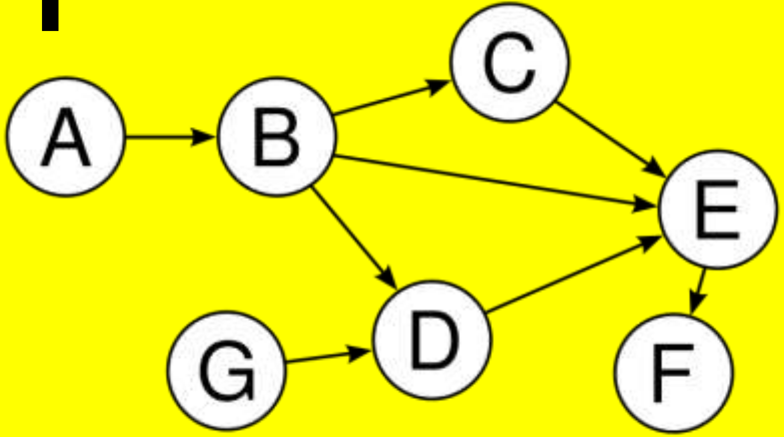
A cron job is the scheduled task itself. Cron jobs can be very useful to automate repetitive tasks.

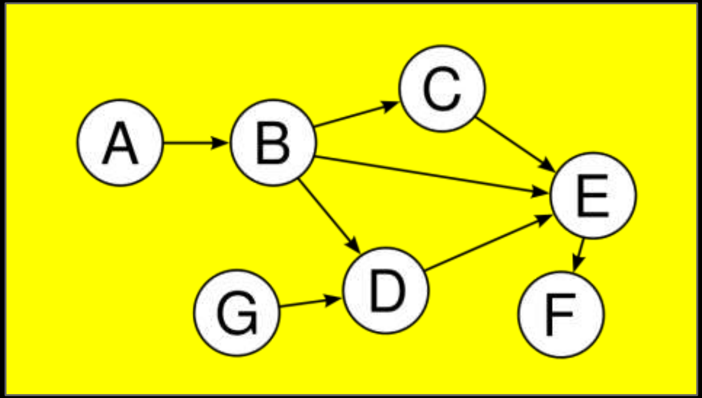
Airflow



DAGs as CODE

I





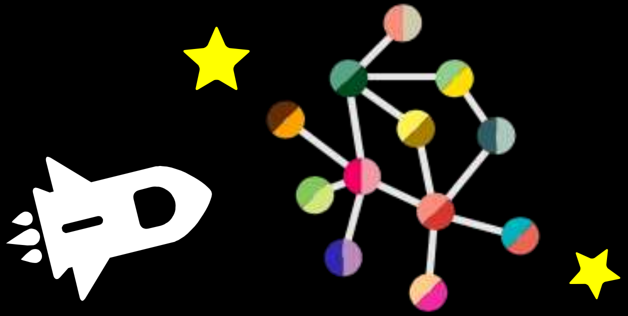
Directed Acyclic Graph

Config file that outlines HOW to carry out a workflow

Contains a collection of tasks

Determines what order tasks will be implemented

Determines when they will be implemented



OPERATORS

Operators are the building blocks of workflows

Action

Performs an **action**, or tell another system to perform an action (i.e., PythonOperator)

Transfer

Move data from one system to another (i.e., RedshiftToS3Transfer

Sensor

Will keep running until a certain criterion is met (i.e., S3KeySensor

Let’s review some concepts

Operators

Classes provided by Airflow. Building blocks of DAGs.

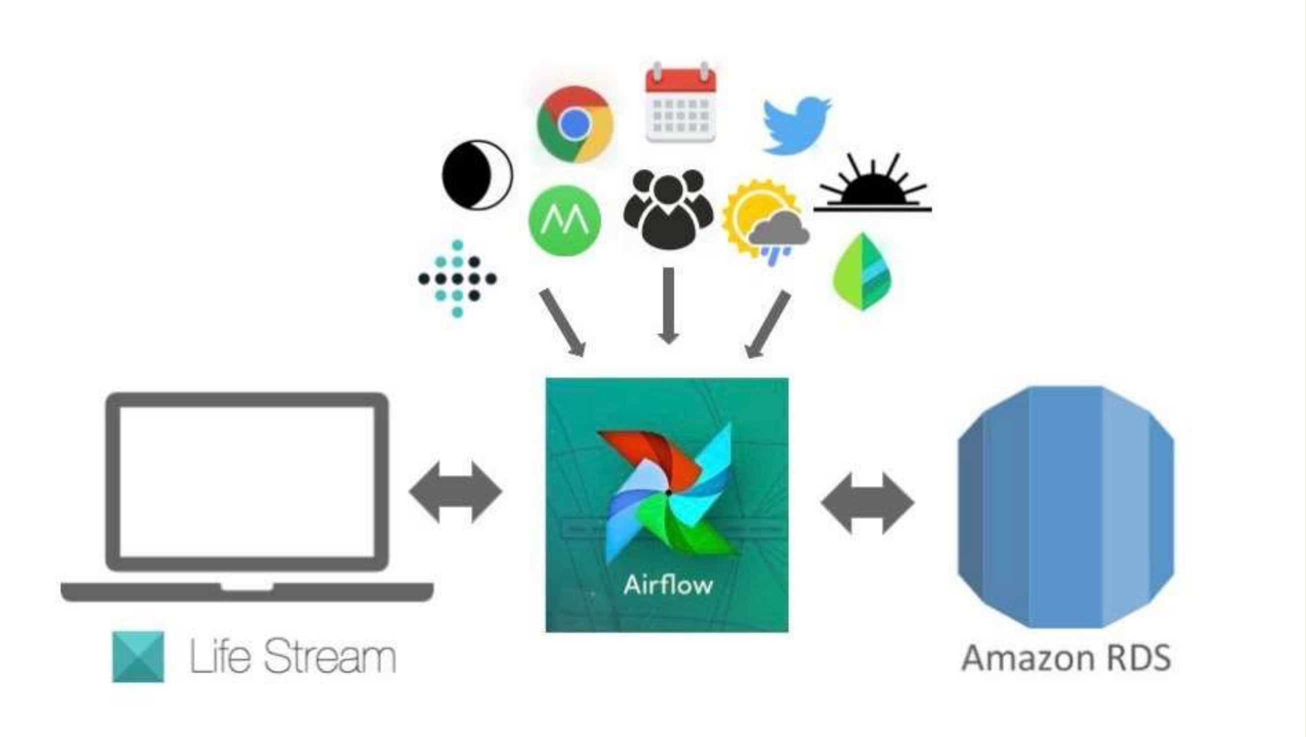
Tasks

Tasks are connected via directed edges that represent an "execute_after" relationship.

DAGS

Directed Acyclic Graphs. Specialized config files for series of tasks.

Life Stream Example



Rails Application

Airflow Process Manager

PostgreSQL Data Store

3

Let’s Try It.

4

Final Thoughts.

What’s It Good For

It Can:

- Schedule complex chains of tasks
- Manage dependencies between tasks
- Define complex relations even in a large distributed environment

It Can’t:

- Store your data
- Clean your house
- Feed your pets while you are gone on vacation (yet)

Competitors

Luigi

Came out of Spotify
Simpler in scope
More object oriented
*Complementary to Airflow?

Azkaban

Created at LinkedIn
Batch workflow job scheduler to run Hadoop jobs

Pachyderm

Containerized data pipeline framework

> Airflow provides a load of functionality, but like any popular, fast-moving project, the documentation gap can be a challenge to adoption.

Thanks!

Any questions?