Modern Data Pipelines with Apache Airflow

Andy Cooper & Taylor Edmiston @ Astronomer.io

Momentum Dev Con 2018

About Us

Andy Cooper

- Data Engineer
- 6 years of experience developing software and data pipelines
- Began career developing traditional data warehouses with Microsoft stack
- Using Airflow since 1.7

Taylor Edmiston

- Backend software engineer building the Airflow platform at Astronomer.io
- 9 years with Python, 6 years as a professional developer
- Top 20% all time on Stack Overflow with a reach of 750k developers
- Enjoys travel - 9 countries / 4 continents

What is Astronomer?

- **Astronomer** is a data engineering platform built on Apache Airflow and clickstream analytics
- Building tools that make data engineers lives easier
- Seed-stage startup, founded ~3 years ago, located in Cincinnati (OTR)
- AngelPad #9 batch
- https://www.astronomer.io
- https://www.crunchbase.com/organization/astronomer

What do we do?

Airflow

- **Astronomer Cloud (Managed Airflow)**
  - Get up and running with Airflow quickly
- **Astronomer Enterprise** (docs)
  - Keep your data and workflows in your private cloud
  - Astronomer Spacecamp - Enterprise support & training available (https://www.astronomer.io/blog/announcing-astronomer-spacecamp/)
- **Astronomer Open** (docs)
  - The core of our platform is open source — try our Docker images on your machine

Clickstream

- A clickstream analytics pipeline and router for user events
- Client-side (web, native mobile) or server-side
- Not an analytics service! We integrate with 50+
- Free tier
- astronomer.io/clickstream
- 2-min demo video - https://www.youtube.com/watch?v=ru7VMe5MXZk

(~40 min) Outline

- (5 min) Intro
- (10 min) Part I - Airflow overview & concepts
- (10 min) Part II - Example DAGs
- Midpoint Q&A?
- (10 min) Part III - Getting started with Airflow + Astro CLI demo
- (5 min) Summary / Outro
- Q&A

What We'll Cover

- Airflow Concepts
- Getting Started with Airflow
- Astro CLI
- Preview and Discussion Of Airflow UI
- Q&A

What is Apache Airflow?

- “Airflow is a platform to programmatically author, schedule and monitor workflows.”
- Open Source currently in the Apache Incubator phase
  - 7,500 stars
  - 4,000 commits
  - 400 contributors
- Written in Python
- Leverages Flask web framework

Airflow Concepts

What is a DAG?

Directed Acyclic Graph

Define Your Pipelines in Code

A Centralized Web App for All Workflows

Web App Features

- A quick look into DAG and task progress
- Error Logging
- Connections & Variables
- Connection Pooling

Hooks and Operators

Hooks

- An interface to an external system
- Often a wrapper for an API client
- Examples
  - DbApiHook
  - S3Hook
  - SlackHook

Operators

- Sensor Operators
  - S3KeySensor
  - S3PrefixSensor
  - HTTPSensor
- Action Operators
  - BashOperator
  - PythonOperator
  - EmailOperator
- Transfer Operators
  - SalesforceToRedshiftSchemaSync
  - SalesforceToS3

DAG Runs & Task Instances

Airflow DAGs Data Profiling Browse Admin Docs About 17:26 UTC

Dag Runs

List (170242) Create Add Filter With selected Search

		State	Dag Id	Execution Date	Run Id	External Trigger
		success	clickstream_v2_to_redshift__597247068b386500015db396	04-19T15:15:00	scheduled__2018-04-19T15:15:00	
		success	clickstream_v2_to_redshift__59834120a942890001096936	04-19T15:15:00	scheduled__2018-04-19T15:15:00	
		failed	clickstream_v2_to_redshift__5acf6731ba1fa926db24be81	04-19T15:15:00	scheduled__2018-04-19T15:15:00	
		failed	clickstream_v2_to_redshift__5ab1c89048837d774362365d	04-19T15:15:00	scheduled__2018-04-19T15:15:00	
		success	clickstream_v2_to_redshift__59f1f814e57fda0001becd72	04-19T15:15:00	scheduled__2018-04-19T15:15:00	
		success	clickstream_v2_to_redshift__5981e6f7a942890001096911	04-19T15:15:00	scheduled__2018-04-19T15:15:00	
		failed	clickstream_v2_to_redshift__5abcb07f5da097d233451817	04-19T15:15:00	scheduled__2018-04-19T15:15:00	
		success	clickstream_v2_to_redshift__597247068b386500015db384	04-19T15:15:00	scheduled__2018-04-19T15:15:00	
		running	clickstream_v2_to_redshift__59f1f8cae57fda0001becd73	04-19T15:15:00	scheduled__2018-04-19T15:15:00	

enu

Airflow DAGs Data Profiling Browse Admin Docs About 17:27 UTC

Task Instances

List (4931169) Add Filter With selected Search

× State not equal skipped Reset Filters

		State	Dag Id	Task Id	Execution Date	Operator	Start Date
		success	clickstream_v2_to_redshift__5978db9928d2dc0001549c75	s3_sensor_search_ads	2017-10-20T01:15:00		2017-10-24T23:57:41.58
		success	clickstream_v2_to_redshift__5978db9928d2dc0001549c75	s3_sensor_search	2017-10-20T02:15:00		2017-10-24T23:57:42.37
		success	clickstream_v2_to_redshift__5978db9928d2dc0001549c75	s3_sensor_search	2017-10-20T01:15:00		2017-10-24T23:57:41.57
		success	clickstream_v2_to_redshift__5978db9928d2dc0001549c75	s3_sensor_push_notification_received	2017-10-20T01:15:00		2017-10-24T23:57:41.56
		success	clickstream_v2_to_redshift__5978db9928d2dc0001549c75	s3_sensor_group	2017-10-20T01:15:00		2017-10-24T23:57:41.49
		success	clickstream_v2_to_redshift__5978db9928d2dc0001549c75	s3_sensor_screen	2017-10-20T02:15:00		2017-10-24T23:57:42.34
		success	clickstream_v2_to_redshift__5978db9928d2dc0001549c75	s3_sensor_instant_shipping_quote	2017-10-20T02:15:00		2017-10-24T23:57:42.32
		success	clickstream_v2_to_redshift__5978db9928d2dc0001549c75	s3_sensor_identify	2017-10-20T01:15:00		2017-10-24T23:57:41.50
		success	clickstream_v2_to_redshift__5978db9928d2dc0001549c75	s3_sensor_page	2017-10-20T01:15:00		2017-10-24T23:57:41.53
		success	clickstream_v2_to_redshift__5978db9928d2dc0001549c75	s3_sensor_identify	2017-10-20T02:15:00		2017-10-24T23:57:42.31
		success	clickstream_v2_to_redshift__5978db9928d2dc0001549c75	s3_sensor_follow_item	2017-10-20T02:15:00		2017-10-24T23:57:42.28
		success	clickstream_v2_to_redshift__5978db9928d2dc0001549c75	s3_sensor_deep_link_opened	2017-10-20T02:15:00		2017-10-24T23:57:42.25

Dynamic DAGs

Executors & Scaling

Executors

- SequentialExecutor
- LocalExecutor
  - No additional dependencies
  - Multi-threaded out of the box
- CeleryExecutor
- MesosExecutor
- KubernetesExecutor (future)

Plugins

What can a plugin do?

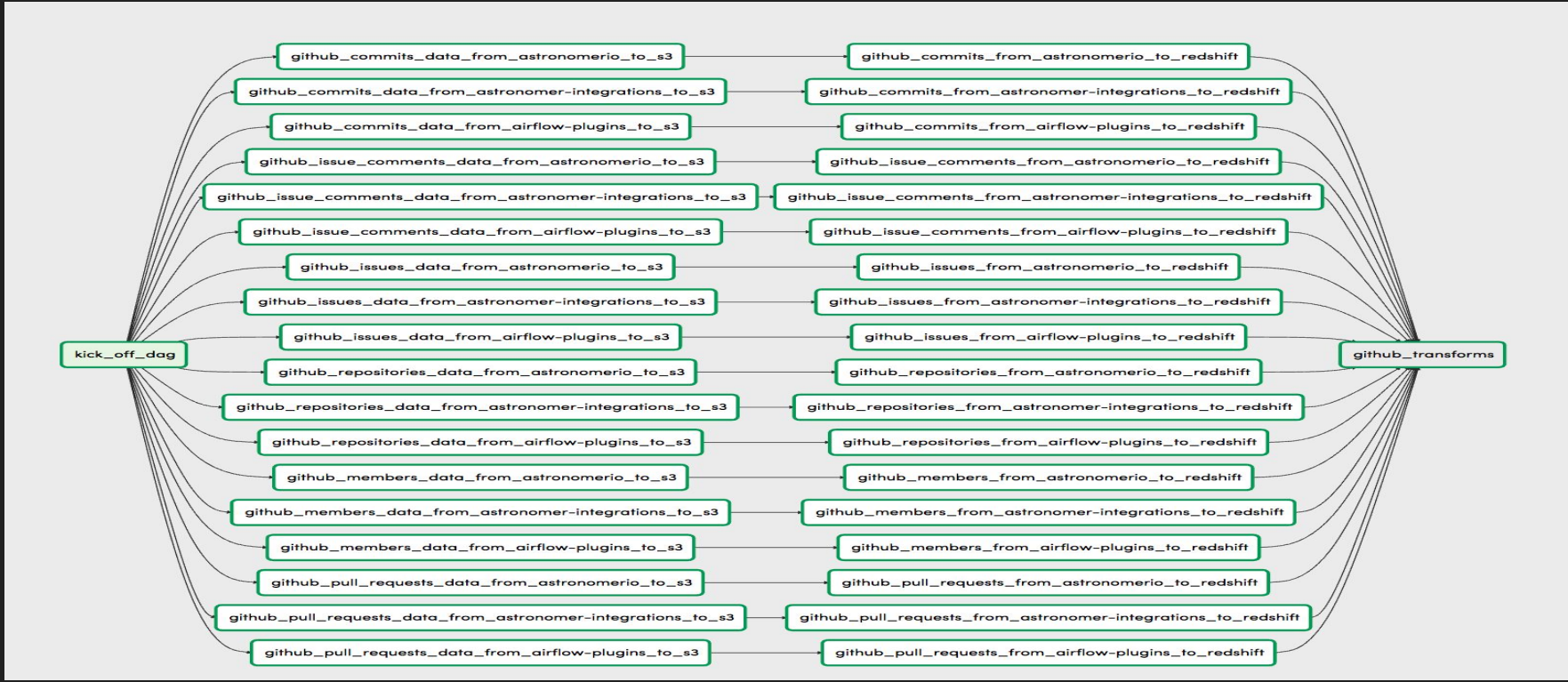
- Extend the Airflow API
- Build new dashboards
- Create custom Hooks and Operators
- Astronomer maintains the most comprehensive collection of Airflow Plugins
  - github.com/airflow-plugins
- Code reuse, composition, good software engineering practices, etc
- Examples
  - Salesforce To Redshift Plugin
  - airflow-api-plugin
  - Airflow DAG Creation Manager Plugin

Example DAGs

DAG Examples

- GitHub stats DAG
- Clickstream Redshift loader DAG
  - ~200 million events per month from customer apps
  - ~2 million Airflow task instances per month
- https://github.com/airflow-plugins/Example-Airflow-DAGs

Github Issue and Commit Tracking Ex.



Clickstream Redshift DAG

Clickstream Redshift DAG

- Your Website → Astronomer Clickstream → S3 → [S3 sensor → Redshift copy via Apache Spark]
- Dynamic DAGs configured via API → Scheduler (cached) → Variable

Airflow DAGs Data Profiling Browse Admin Docs About 17:14 UTC

On DAG: clickstream_v2_to_redshift__5978db9928d2dc0001549c6c

schedule: 15 * * * *

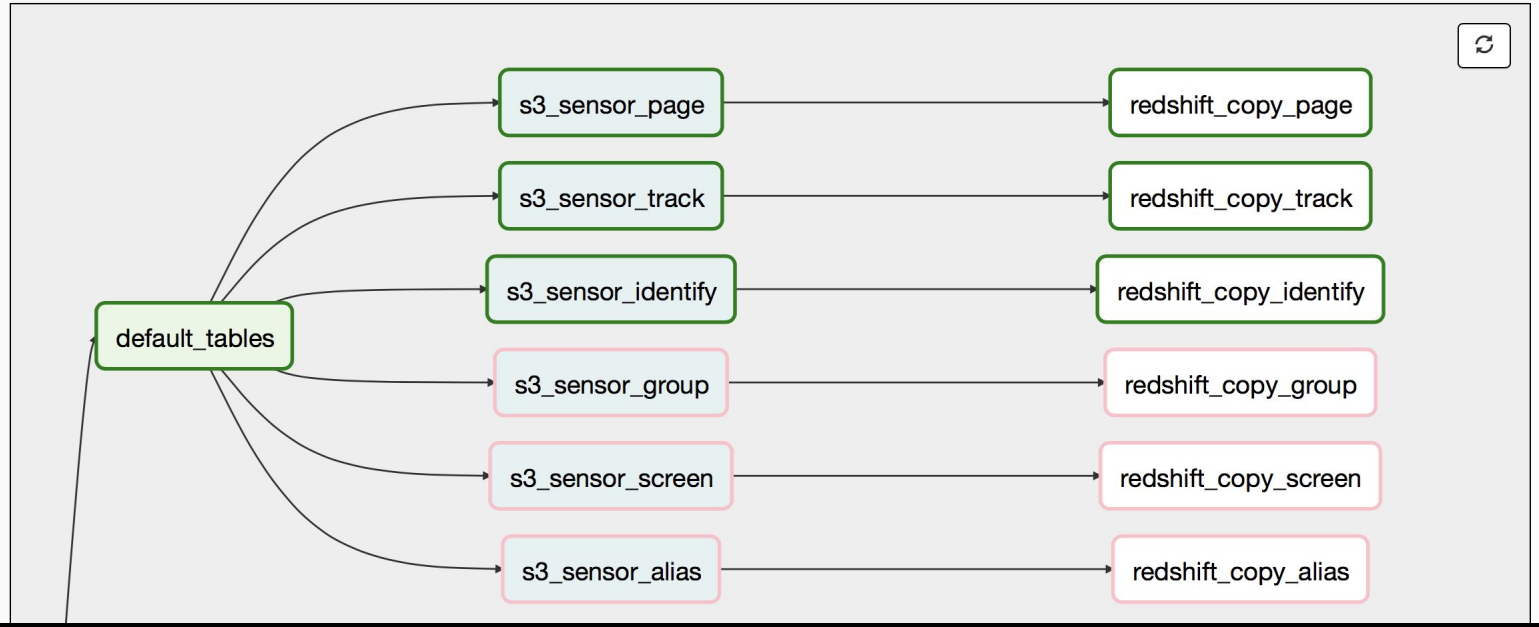
Graph View | Tree View | Task Duration | Task Tries | Landing Times | Gantt | Details | Code | Refresh

success Run: scheduled__2018-04-19T07:15:00 Layout: Left->Right Go

Search for...

DockerOperator DummyOperator PythonOperator S3ClickstreamKeySensor

success running failed skipped retry queued no status



Airflow DAGs Data Profiling Browse Admin Docs About 17:16 UTC

On DAG: clickstream_v2_to_redshift__5978db9928d2dc0001549c6c

schedule: 15 * * * *

Graph View | Tree View | Task Duration | Task Tries | Landing Times | Gantt | Details | Code | Refresh

Base date: 2018-04-19 08:15:00 Number of runs: 50 Go

DockerOperator DummyOperator PythonOperator S3ClickstreamKeySensor

success running failed skipped retry queued no status

Airflow DAGs Data Profiling Browse Admin Docs About 17:21 UTC

List [redacted] Create With selected

		Pool	Slots	Used Slots	Queued Slots
		redshift_loader_597247068b386500015db382	30	5	21
		redshift_loader_597247068b386500015db396	30	0	0
		redshift_loader_598342dca942890001096952	50	0	1

Astro CLI

The fastest way to get started with Airflow

How can I get started with Airflow?

- Source Code
  - https://github.com/astronomerio/astro-cli
- Install CLI
  - $ curl -sL https://install.astronomer.io | sudo bash
- Start a Project
  - $ mkdir test-project && cd test-project
  - $ astro airflow init
  - $ astro airflow start

Takeaway

- Part I - Airflow overview & concepts
- Part II - Example DAGs
- Part III - Getting started with Airflow + Astro CLI demo

Resources

- Official
  - https://github.com/apache/incubator-airflow
  - https://airflow.apache.org
  - Airflow Dev Mailing List
  - Apache Airflow meetups
- Community
  - https://github.com/airflow-plugins
  - https://soundcloud.com/the-airflow-podcast
  - https://github.com/jghoman/awesome-apache-airflow
- Related Talks
  - https://blog.tedmiston.com/talks/

Contact Info

- Andy
  - https://twitter.com/andscoop
  - https://www.linkedin.com/in/andscoop/
  - https://andscoop.com/
  - andy.cooper@astronomer.io
- Taylor
  - https://twitter.com/kicksopenminds
  - https://www.linkedin.com/in/tedmiston/
  - https://blog.tedmiston.com
  - taylor@astronomer.io