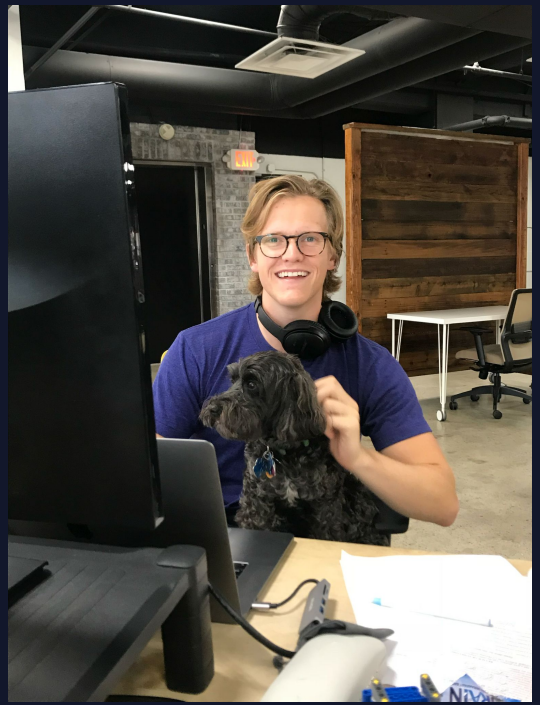
Things to Consider When Building an Airflow Service

About Us

Pete Dejoy
SVP Product, CoFounder

Viraj Parekh
Field CTO, CoFounder

About Us

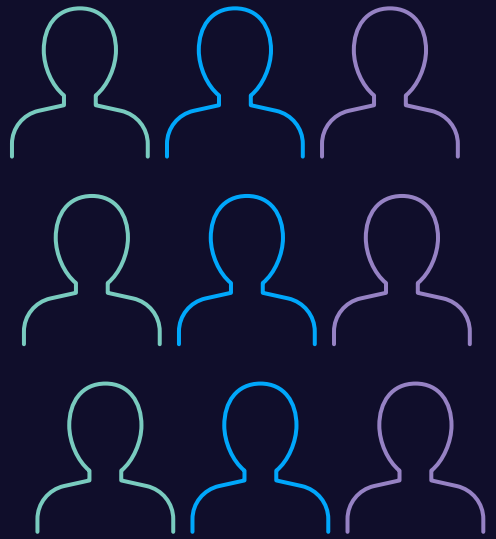


About Us

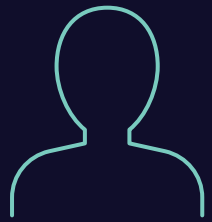
A data engineering team uses Airflow...

Data Engineers

But then more users come along...



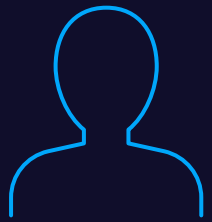
...some of whom are different personas



Analytics Engineer

SQL queries, dbt jobs

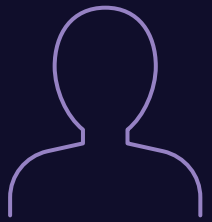
"I need this report on monthly active users to be refreshed daily!"



Data Engineers

Data ingestion, ETL, Schema Maintenance

"I need to enrich our MAU tables with data from our legacy system!"

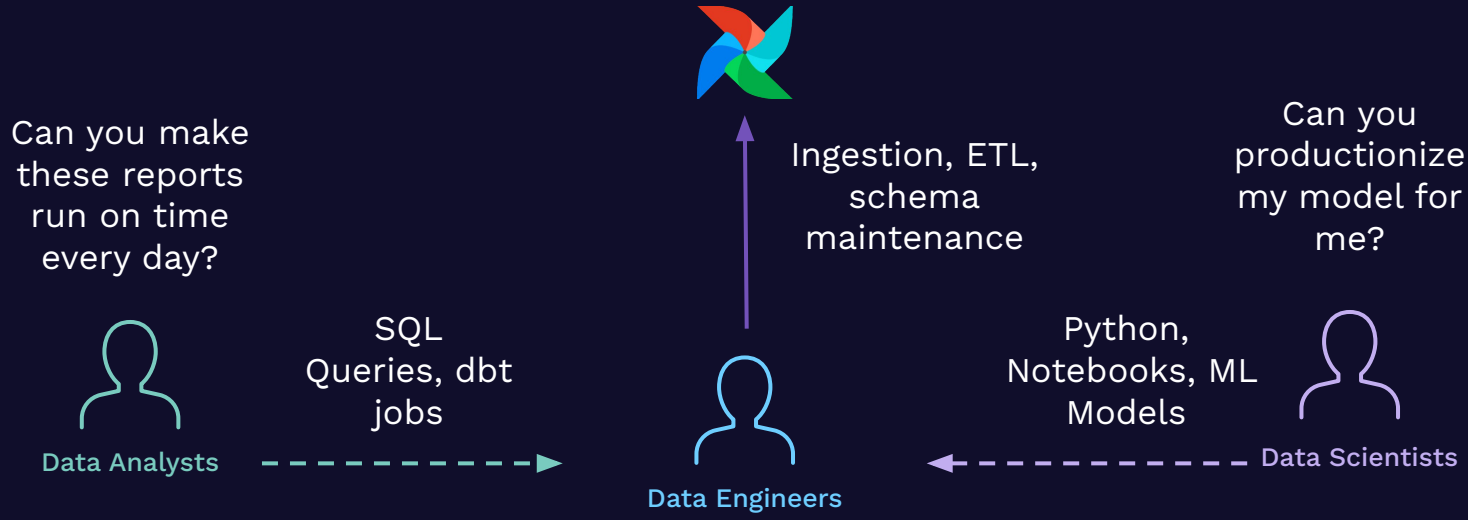


Data Scientists

Notebooks, models, R, Python

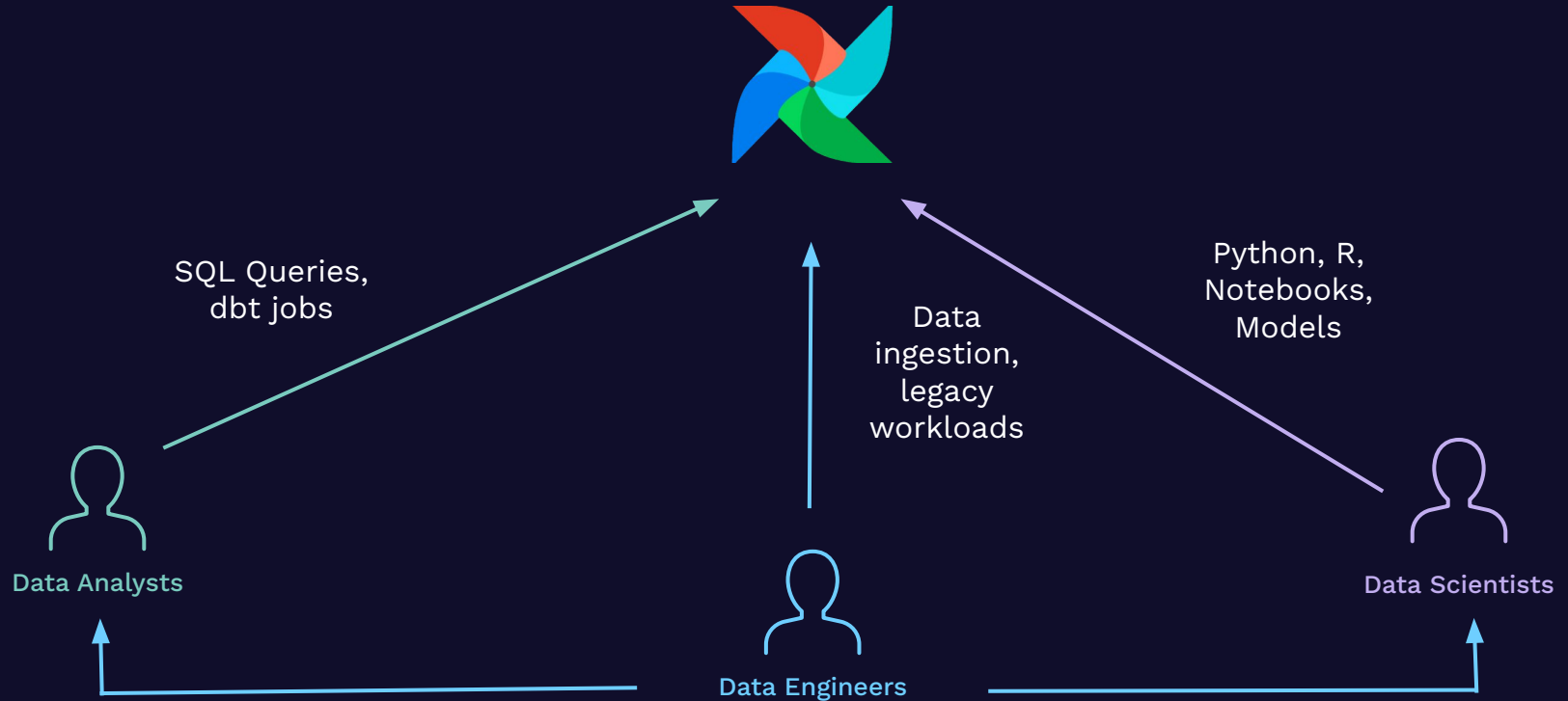
"I need my model to be trained on the latest data!"

You want to avoid this...



"Can you all help me out?"

...and build towards this.



To build a robust Airflow service...

Three biggest challenges to enabling Airflow for all.

Infra Mgmt

How do you manage the infrastructure abstractions to deliver a reliable service?

Workload Compute

How do you give end users access to compute specialized for their use case?

DevEx

How do you deliver a world-class developer experience for different kinds of end users?

Infrastructure Management

There are a lot of Airflow services to build your service around...

CLOUDERA

...but they all differ in:

- **Tenancy Abstractions** - how do you scale across teams?
- **Feature set** - what Airflow features are exposed? What can you modify?
- **Cost** - What's the pricing model? How do you justify ROI?

Many Approaches to Multitenancy

Astronomer @ Fanduel

- Leverages Astronomer to run multiple Airflows
- Teams leverage their own workload patterns
- Cost savings resulted from optimizing workloads

Cloud Composer @ BMG

- Monolithic architecture
- More opinionated data processing patterns
- Tracks cost with GKE native features

Workload Compute Needs

Practitioners and Infra Needs

Practitioner	Skillset	**Use Case**	Infra Requirements
 **Analytics Engineers**	SQL, dbt, BI Tools	“I need this ***report*** on MAU to be refreshed daily.”	• Transformation-oriented workflows • Compute is generally offloaded to a DWH
 **Data Engineers**	Python, SQL, Schema design, git	“I need to **ingest some data** from my app db to my data warehouse.”	• Need for ephemeral storage and disk space • Need to manage python dependencies
 **Data Scientists**	Pandas, Notebooks, R	“I need to write & deploy a **predictive model** on MAUs.”	• Every compute need imaginable

Airflow has rich abstractions for customizing workload compute...

Containers

Operators for Kubernetes, Docker, ECS, ACI, and more for running pre-existing images

Executors

Celery, Kubernetes, and hybrid style executors to configure interface between user code and underlying infrastructure

https://shopify.engineering/lessons-learned-apache-airflow-scale

ASTRONOMER

Airflow has Abstractions for workload compute:

Containers

Run all their tasks in KubernetesPodOperators to handle dependencies + infra specifics

Links (1) (2)

Executors

Executor optionality allows teams to choose the right tool for their workload profile.

Links (1) (2)

ML Launcher – to containerize task execution. ML Launcher integrates compute backends like Sagemaker, Databricks, and Snowflake to perform container runs **and meet the unique hardware requirements for ML such as GPUs, instances with large memory, and disks with high IO throughput.** This design choice enables MLEs to develop and deploy pipelines without worrying about Airflow runtime and allows us to scale easily to hundreds of DAGs (Directed Acyclic Graphs) with thousands of tasks in a short period

Airflow @ Instacart

Compute Abstractions → Interfaces + DevEx

Build your devex based on the needs of your end user!

Practitioners and Interface

Practitioner	Skillset	Use Case	Interface of choice
 **Analytics Engineers**	SQL, dbt, BI Tools	"I need this ***report*** on MAU to be refreshed daily."	• BI Tools, SQL tools, dbt
 **Data Engineers**	Python, SQL, Schema design, git	"I need to **ingest some data** from my app db to my data warehouse."	• IDEs, Object oriented Python, SQL tools , some Terraform
 **Data Scientists**	Pandas, Notebooks, R	"I need to write & deploy a **predictive model** on MAUs."	• Notebooks,Python, Pandas, scripting

Analytics Engineer

Updater

[dbt + Airflow @ Updater]

DevotedHealth

[dbt @ Devoted Health]

Data Engineers

[BrickFlow @ Nike]

lyft

[Airflow + Flyte @ Lyft]

Data Scientists

Walmart

[ML @ Walmart]

ck

[Vega @ Credit Karma]

There are a ton in the community that exist:

https://airflow.apache.org/ecosystem/

TL;DR:

- Use a managed service, but know they’re not all built the same
- Different personas are going to have different compute & interface requirements
- Focus on developer experience, but if you’re building your own Airflow DSL, make sure you know what you’re getting into.

ASTRONOMER

(After) Party Under the Stars

Wednesday, September 20th
6:30pm-10:00pm

The Sheraton Centre
123 Queen St W
(7 min walk)

RSVP Now

Airflow Summit
September 19 - 21, 2023
Toronto, Canada
airflowsummit.org

Let's flow together

Workshop

Get Airflow Certified

Thursday, September 21st
12:00 pm in Trinity 4

Marc Lamberti
Head of Customer Education
at Astronomer