


The Road Ahead:
What's Coming in Airflow 3 and Beyond

Vikram Koka
Chief Strategy Officer
Astronomer

Introduction

Vikram Koka

Chief Strategy Officer
Astronomer

Airflow Committer
Contributions include design input for

- Scheduler HA,
- Datasets and Data Driven Scheduling
- Setup and Teardown

vikram@astronomer.io • San Jose, CA

Key Airflow milestones

0.1

Creation
Airflow Started

1.0

1.10.2

Apache Top Level Project

2.0

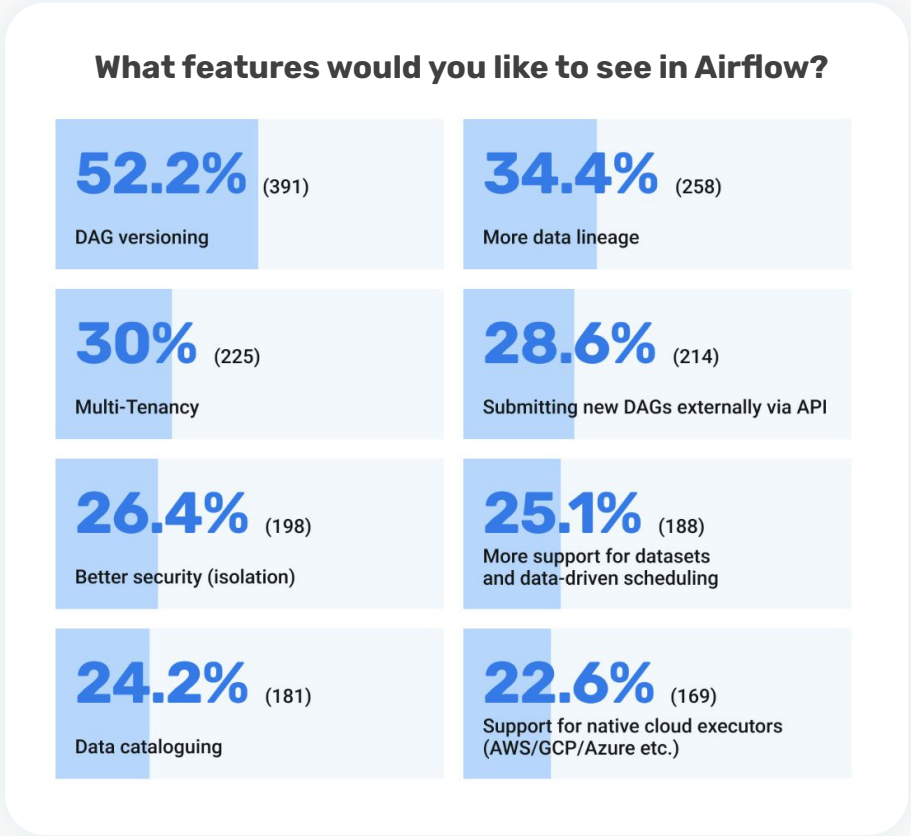
Enterprise Production-Ready
HA Schedulers
Fully specified REST API
TaskFlow API

2.1 - 2.10

Efficiency & Ease of Use
Async Operators
Dynamic Tasks
Setup & Tear-down
Airflow ObjectStore

Data Awareness
Data-aware scheduling
Conditional scheduling
Combined dataset + time scheduling
Dataset Event API

Airflow survey results



DAG Versioning always #1 on the wish list!

More:

- Security
- Execution / invocation
- Datasets
- Observability / Governance

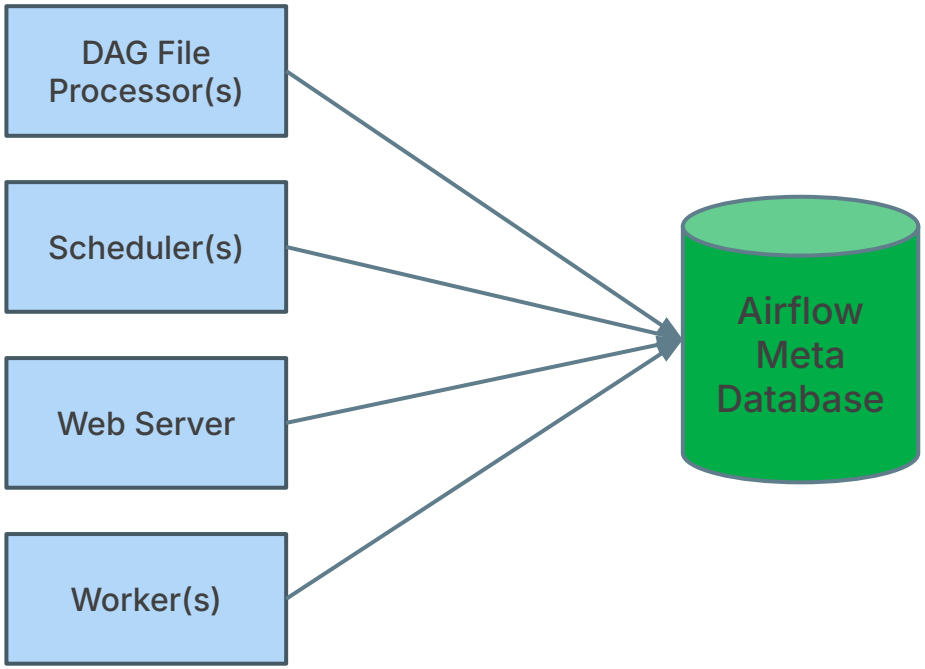
Airflow 3 Themes

Easier to use

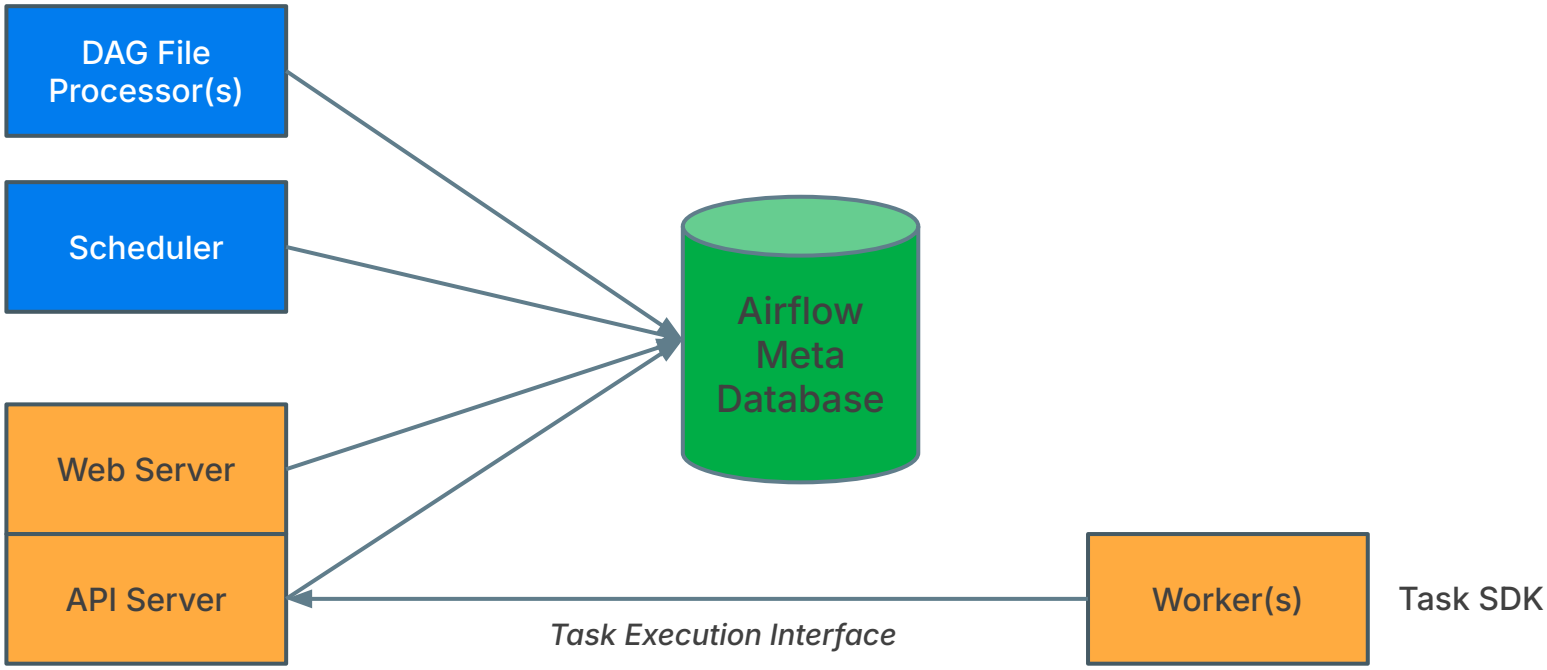
DAG Versioning	View historical versions of DAGs along with all their run information
UI Modernization	React-based UI with support for embedded plug-ins
Backfills at Scale	Extended MLOps support for backfill runs
Stronger Security	Task Isolation: Tasks no longer have direct access to the Airflow metadatabase


Task Isolation needs an architecture change

Current Airflow architecture



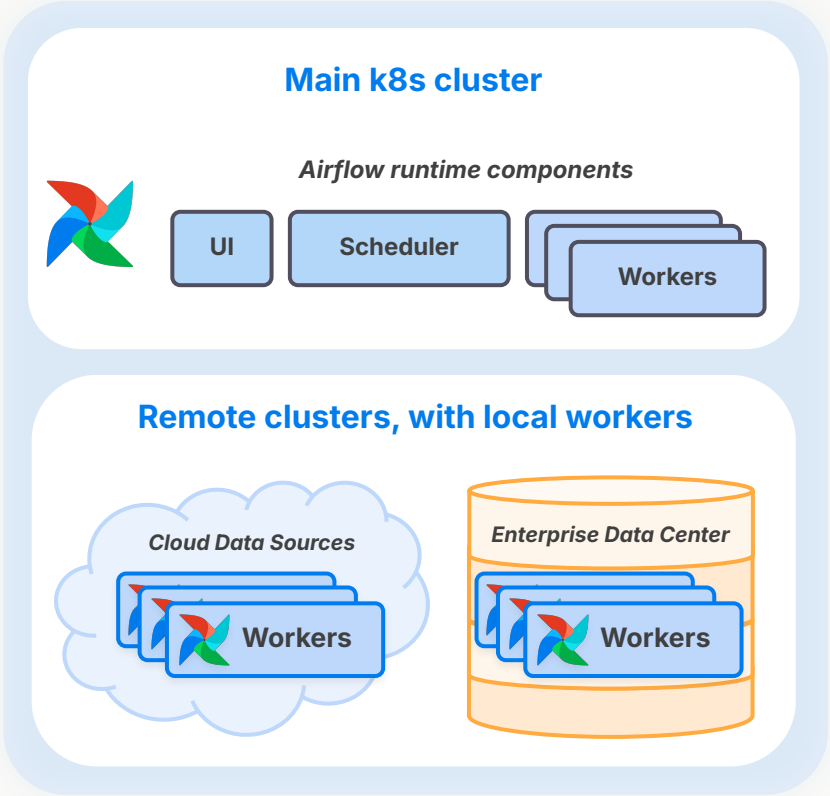
Architectural decoupling:
Task Execution Interface




Airflow 3: Run [tasks]

anywhere,
at any time,
in any language

Run anywhere



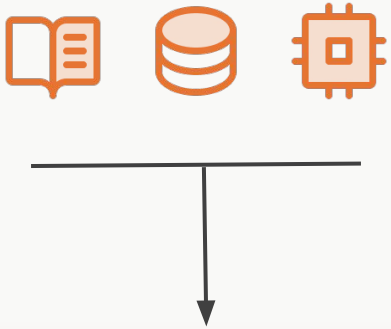
Remote / Edge execution: run tasks on workers in remote clusters

Use cases

- Deployment flexibility with workers on public, hybrid, private cloud, on-prem, edge, GPU clouds
- Higher resilience and scalability
- Improved security isolation
- Easier upgrades, fewer dependencies
- Better meet data locality mandates

Run at any time

- **Event-driven scheduling:** Responsive and real-time pipelines that automatically react to events from external sources. Built on top of data assets
- **Data partitioning:** Time and segment-based partitions. Independently process specific slices of data: improves performance, reduces processing time and resource usage
- **Inference Execution:** Simultaneous execution of the same DAG. Synchronous DAG execution



Data Asset

Collection of logically related data:
i.e., tables, files, models, dashboards

Airflow supports Scheduled (Batch), Event-Driven, and Adhoc Execution

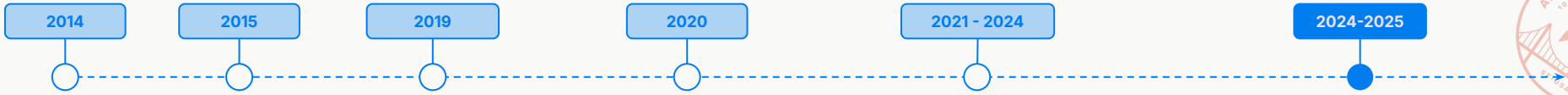
Run in any language

- Airflow 3 is language agnostic
- Software teams building data apps
  - Eg: Typescript
- Airflow 3 is also multi-lingual
  - Extract in Java
  - Transform in Python, SQL
  - Analysis using Scala
  - Feed data into an Go app

Current languages supported

Language support in 3.0+

Airflow 3.0



0.1

Creation
Airflow Started

1.0

1.10.2

Apache Top Level Project

2.0

Enterprise Production-Ready
HA Schedulers
Fully specified REST API
TaskFlow API

2.1 - 2.10

Efficiency & Ease of Use
Async Operators
Dynamic Tasks
Setup & Tear-down
Airflow ObjectStore

Data Awareness
Data-aware scheduling
Conditional scheduling
Combined dataset + time scheduling
Dataset Event API

3.0+

Easier to Use

DAG Versioning
UI modernization
Backfills at Scale
Better Security: Task Isolation

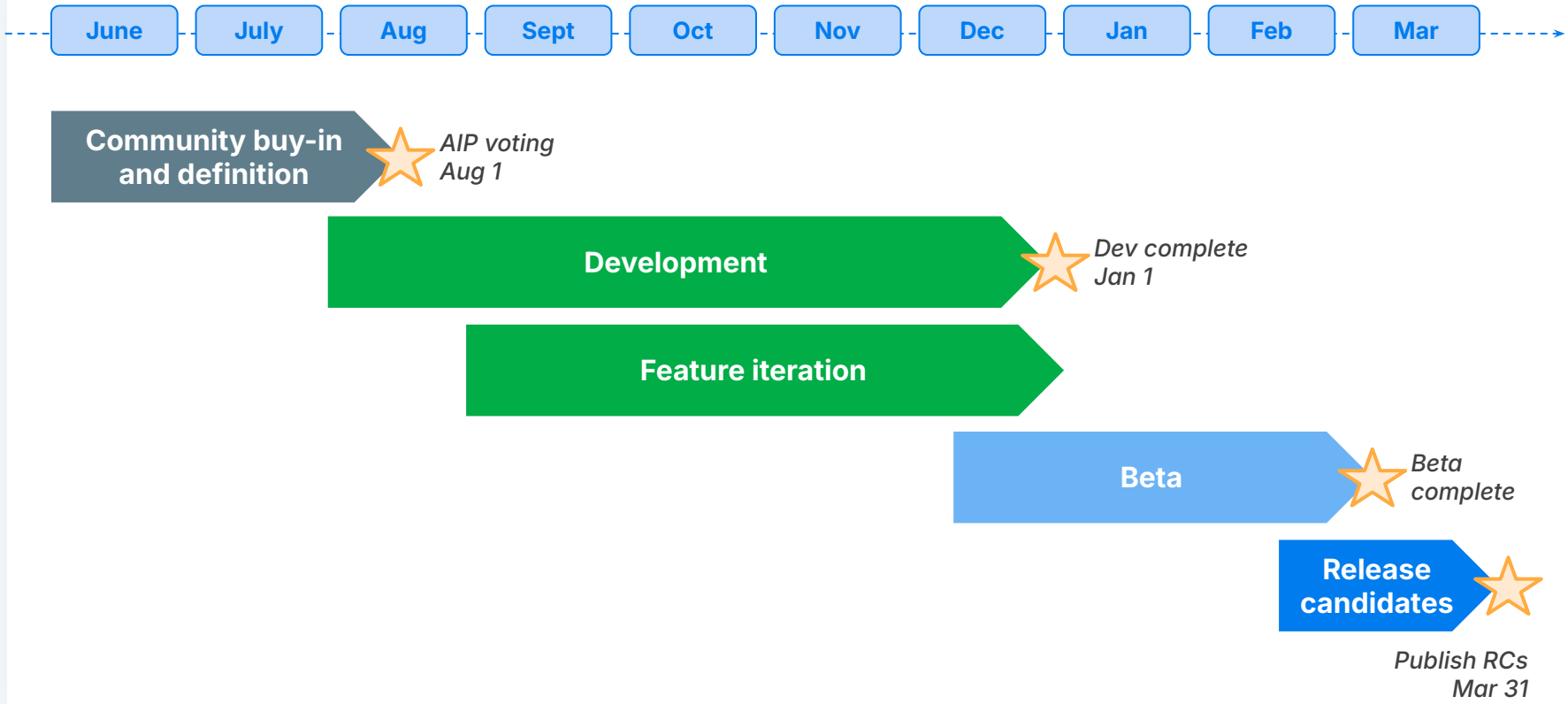
Run anywhere, at any time, in any language

Remote Execution

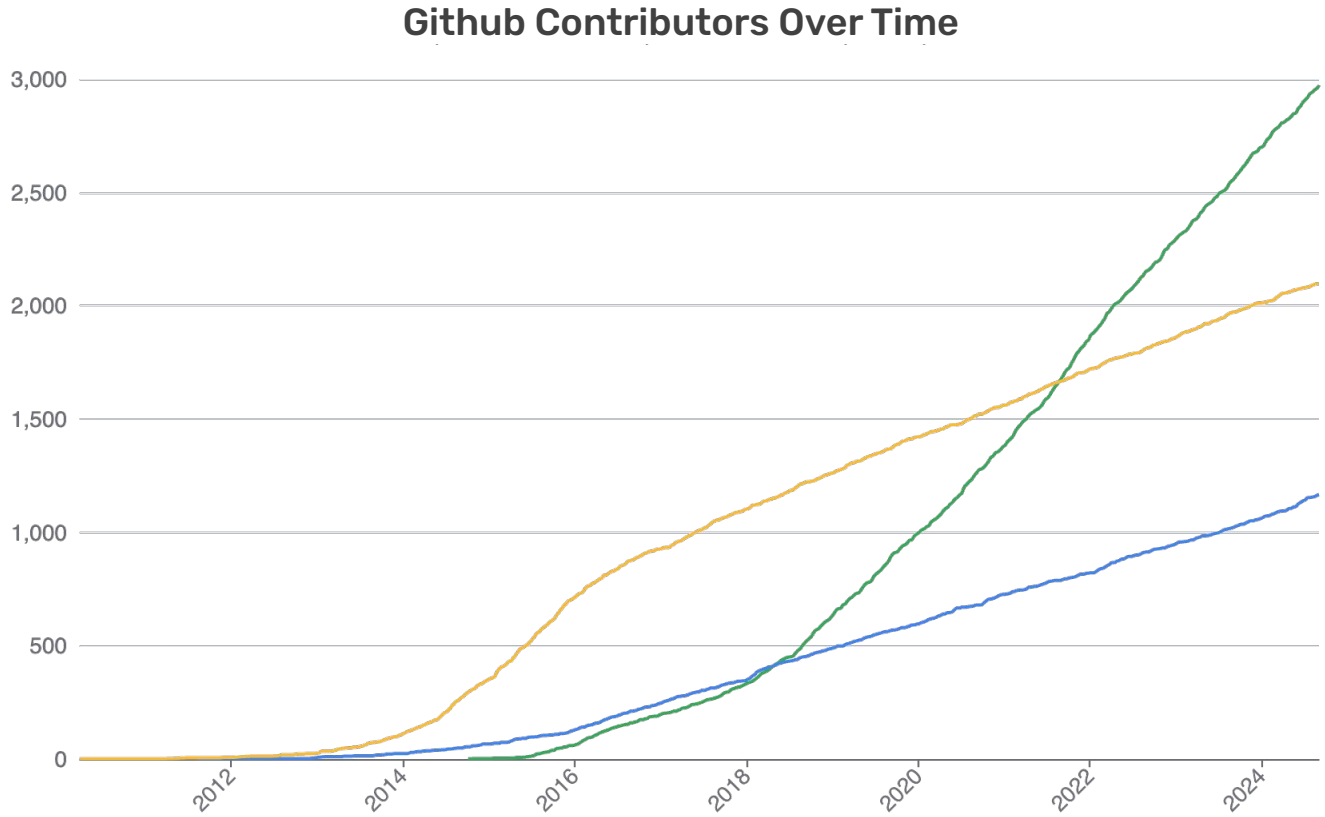
Event Driven Scheduling
Gen AI Inference Execution
Data Assets & Data Partitioning

Task Execution beyond Python: i.e. Java, Typescript, etc.

Airflow 3.0 timeline



Airflow Contributor Growth



Apache Airflow

APACHE Spark™

APACHE kafka®

Shoutout to the community growers!

We need you!

Airflow 3 is a big jump:

Run [Tasks] anywhere, at any time, in any language!

Recruiting beta users:

- ***Anywhere***: Deploying remote execution environments
- ***Any time***: Building Gen AI platforms and use cases

Recruiting contributors:

- ***Any language***: Add Providers for: Typescript, Javascript, Go, Kotlin, ...

Come speak at the next Airflow Summit about your use case on Airflow 3!


Now, over to the panel