OSA CON 24

# Airflow 3 is coming!

November 19-21, 2024

## Jarek Potiuk

Apache Airflow PMC member & committer

Member of Apache Software Foundation

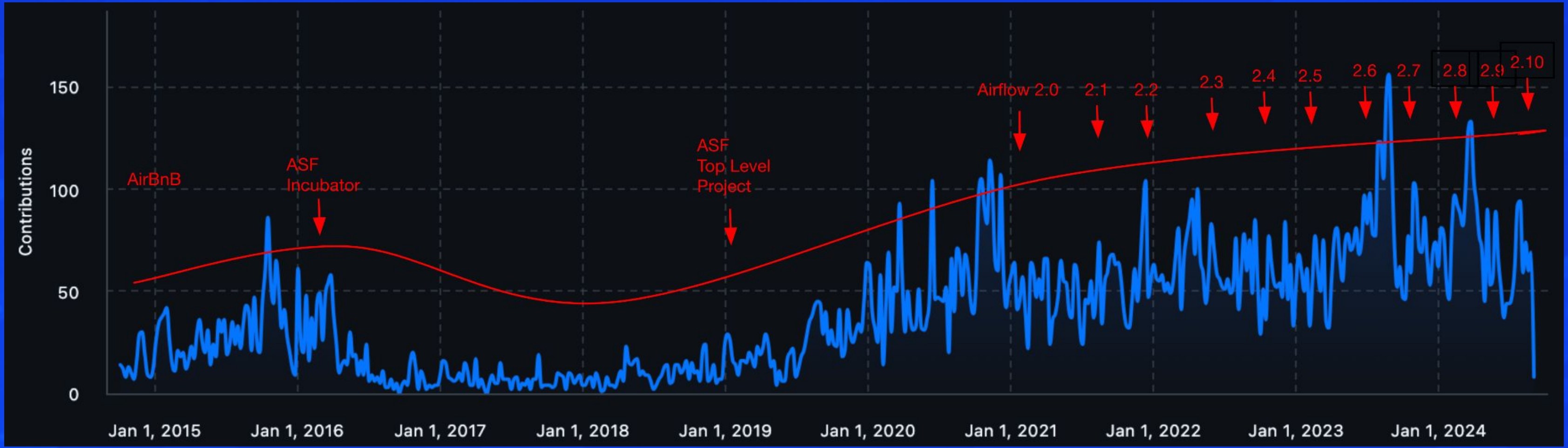
Member of ASF Security Committee

| | |
|---|---|
| 𝕏 | @jarekpotiuk |
| GitHub | @potiuk |
| in | @jarekpotiuk |
| m | https://fosstodon.org/@jarekpotiuk |

# Airflow as Orchestrator

# Why Airflow 3

# Airflow timeline



# Airflow 2.0 strengths

- User base / Community
- Versatility
- Steady stream of new features
- Learning from the industry
  - implement what makes sense
  - when it makes sense

# Airflow 2 weaknesses

- Rooted in ETL
- Architecture UI not fully modernized (limits)
- Weak AI workflow support
- Weak DAG Change Management
- Deployment complexity (dependencies)
- Steep(ish) learning curve

# Why now?

- Focus on ML/AI workflows
- Users have new, different needs
- User skills are proliferating
- Python is winning - but is not the only in Data
- Technical debt piling
- Importance of streaming and inference

# Basic principles

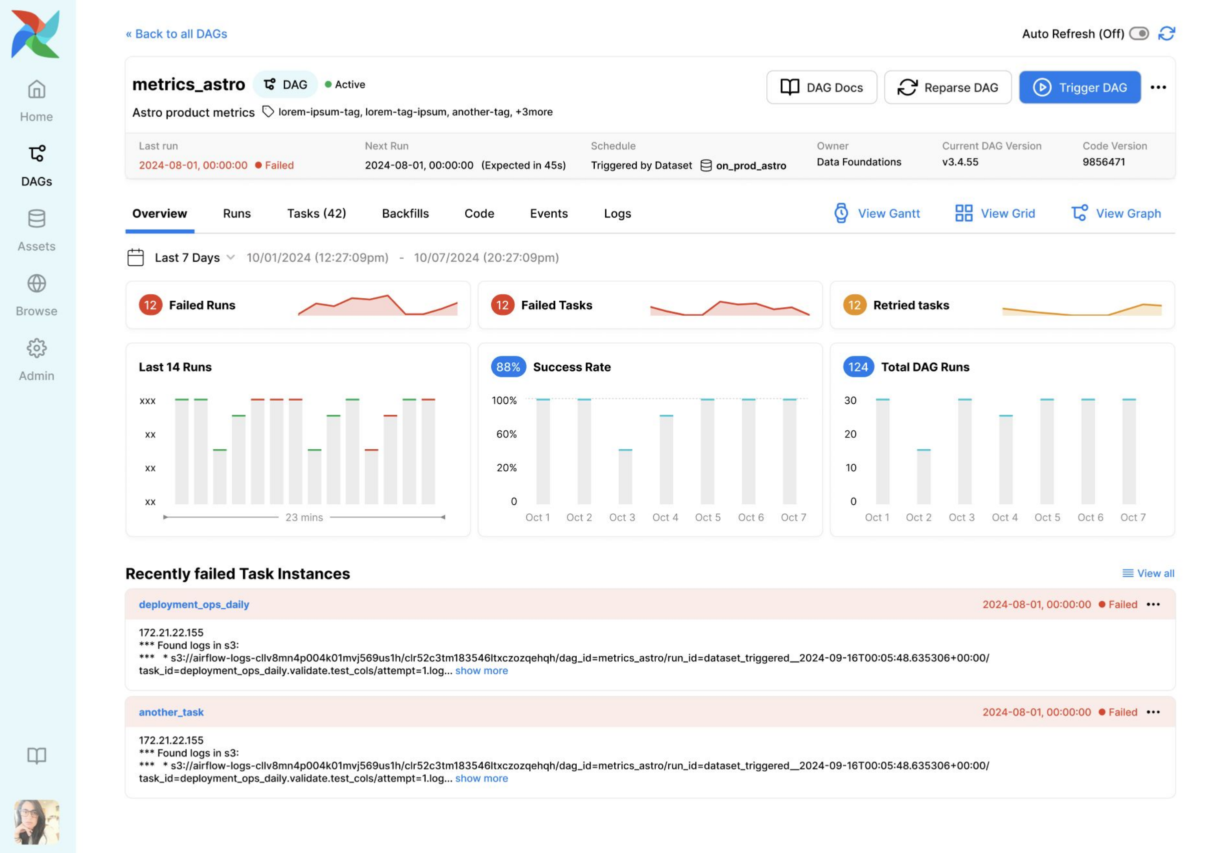
# What to expect?

- Airflow 2 features continue to work (exceptions)
- Smooth migration for existing users is top priority
- AI/ML/Data workflows as first class citizens
- Enhance scalability, performance, enterprise-level security through architectural changes
- Make Airflow Asset and Partition native
- Make Airflow easier to deploy and learn

# UI

# Modern Web Application/API

(AIP-38, AIP-84, AIP-65)



**metrics_astro**

Astro product metrics lorem-ipsum-tag, lorem-tag-ipsum, another-tag, +3more

| Last run | Next Run | Schedule | Owner | Current DAG Version | Code Version |
|---|---|---|---|---|---|
| 2024-08-01, 00:00:00 ● Failed | 2024-08-01, 00:00:00 (Expected in 45s) | Triggered by Dataset on_prod_astro | Data Foundations | v3.4.55 | 9856471 |

Last 7 Days 10/01/2024 (12:27:09pm) - 10/07/2024 (20:27:09pm)

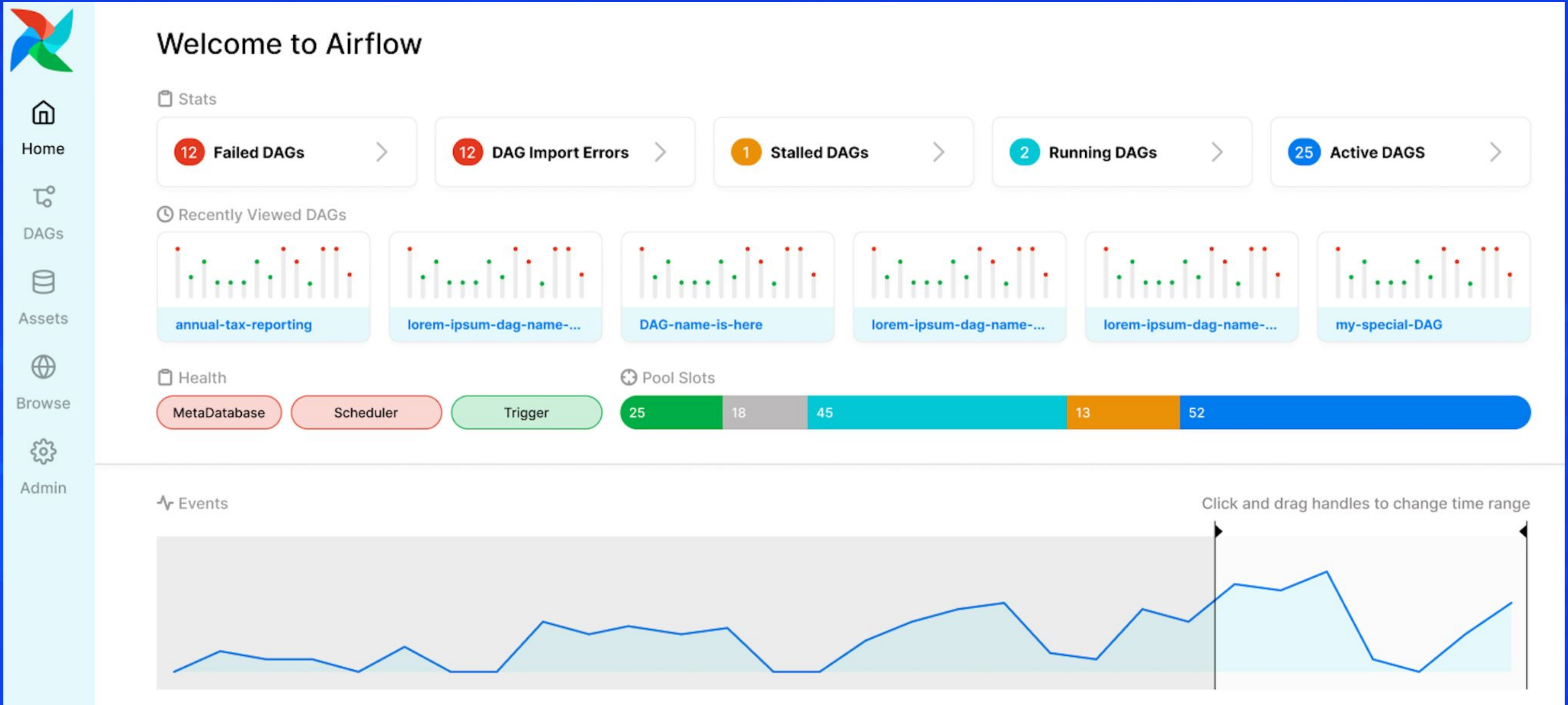
**Recently failed Task Instances**

deployment_ops_daily 2024-08-01, 00:00:00 ● Failed

```
172.21.22.155
*** Found logs in s3:
***   * s3://airflow-logs-cllv8mn4p004k01mvj569us1h/clr52c3tm183546ltxczozqehqh/dag_id=metrics_astro/run_id=dataset_triggered__2024-09-16T00:05:48.635306+00:00/
task_id=deployment_ops_daily.validate.test_cols/attempt=1.log... show more
```

another_task 2024-08-01, 00:00:00 ● Failed

```
172.21.22.155
*** Found logs in s3:
***   * s3://airflow-logs-cllv8mn4p004k01mvj569us1h/clr52c3tm183546ltxczozqehqh/dag_id=metrics_astro/run_id=dataset_triggered__2024-09-16T00:05:48.635306+00:00/
task_id=deployment_ops_daily.validate.test_cols/attempt=1.log... show more
```



Welcome to Airflow

Stats

12 Failed DAGs

12 DAG Import Errors

1 Stalled DAGs

2 Running DAGs

25 Active DAGS

Recently Viewed DAGs

annual-tax-reporting

lorem-ipsum-dag-name-...

DAG-name-is-here

lorem-ipsum-dag-name-...

lorem-ipsum-dag-name-...

my-special-DAG

Health

MetaDatabase

Scheduler

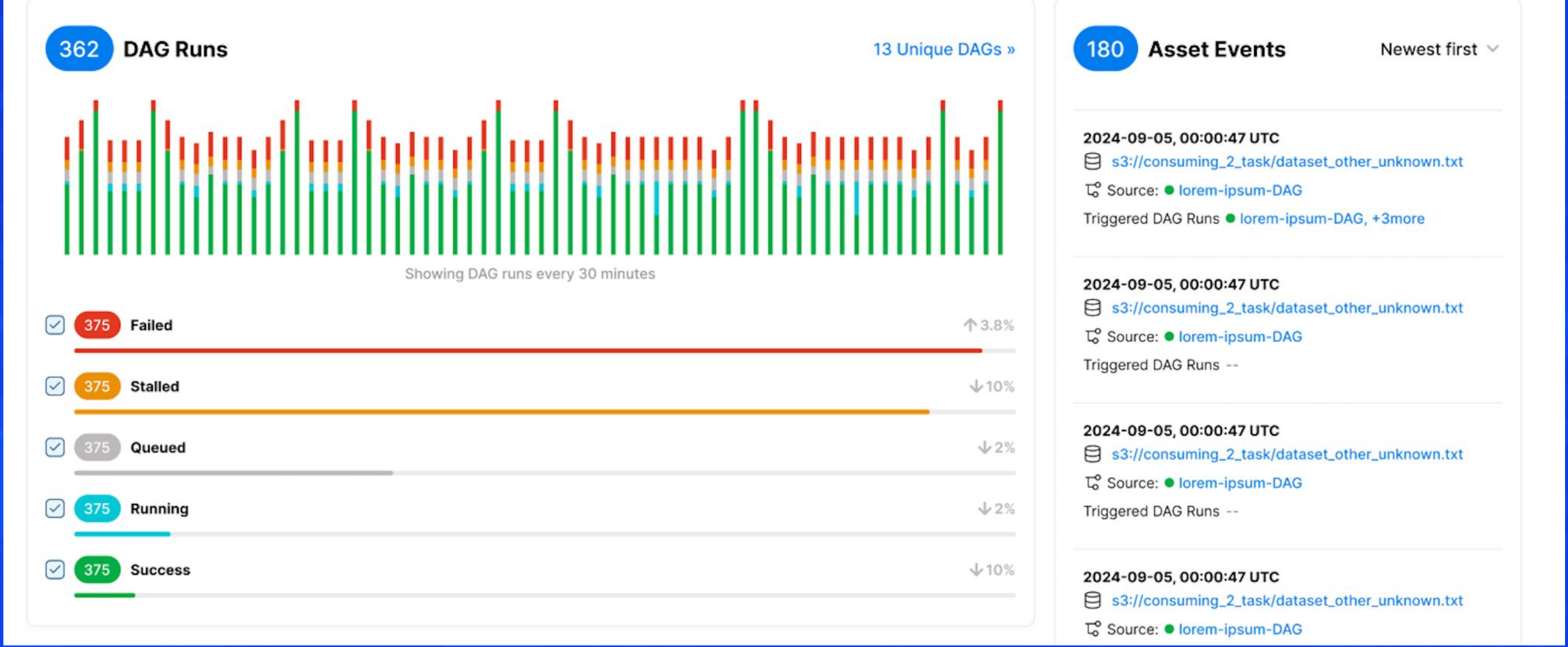
Trigger

Pool Slots

25 18 45 13 52

Events

Click and drag handles to change time range



362 DAG Runs

13 Unique DAGs »

Showing DAG runs every 30 minutes

☑ 375 Failed ↑3.8%

☑ 375 Stalled ↓10%

☑ 375 Queued ↓2%

☑ 375 Running ↓2%

☑ 375 Success ↓10%

180 Asset Events

Newest first

2024-09-05, 00:00:47 UTC
s3://consuming_2_task/dataset_other_unknown.txt
Source: ● lorem-ipsum-DAG
Triggered DAG Runs ● lorem-ipsum-DAG, +3more

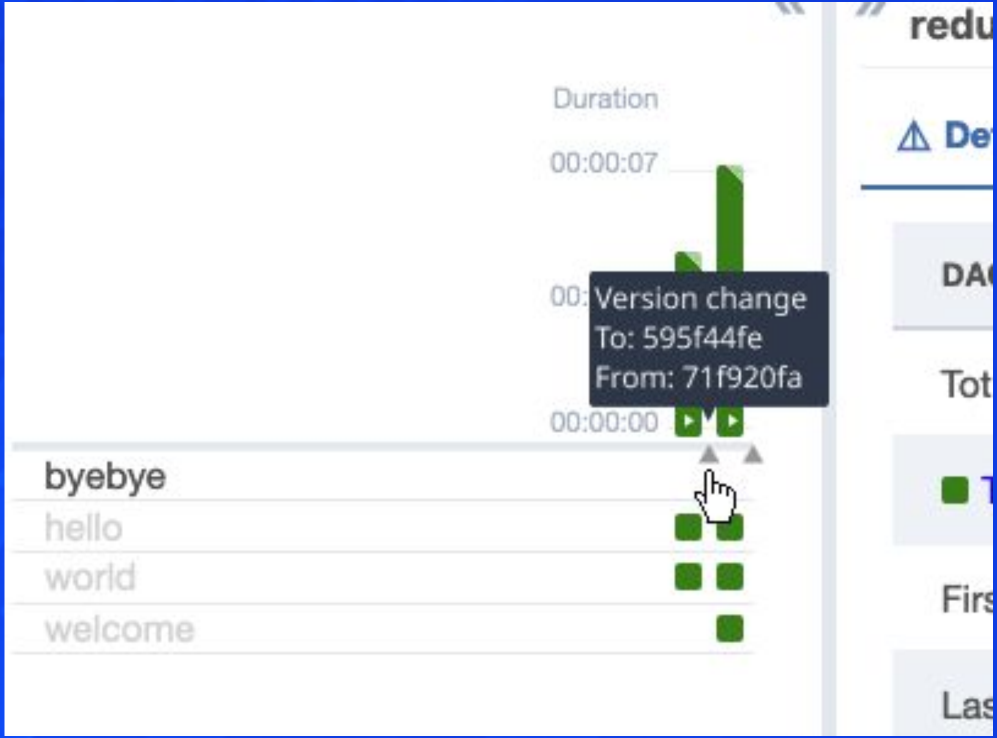
2024-09-05, 00:00:47 UTC
s3://consuming_2_task/dataset_other_unknown.txt
Source: ● lorem-ipsum-DAG
Triggered DAG Runs --

2024-09-05, 00:00:47 UTC
s3://consuming_2_task/dataset_other_unknown.txt
Source: ● lorem-ipsum-DAG
Triggered DAG Runs --

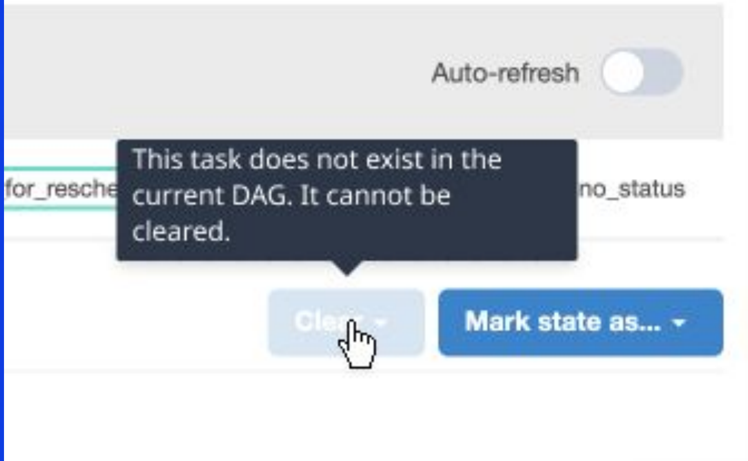
2024-09-05, 00:00:47 UTC
s3://consuming_2_task/dataset_other_unknown.txt
Source: ● lorem-ipsum-DAG

# DAG Versioning

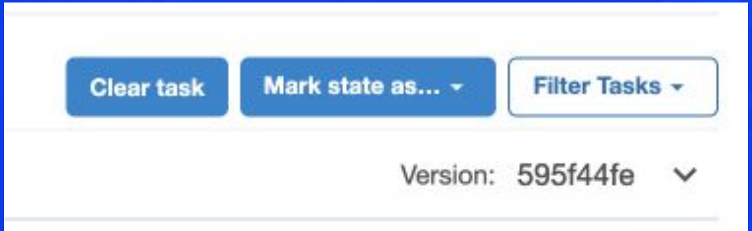
(AIP-63. AIP-65)



Version change
To: 595f44fe
From: 71f920fa



This task does not exist in the current DAG. It cannot be cleared.



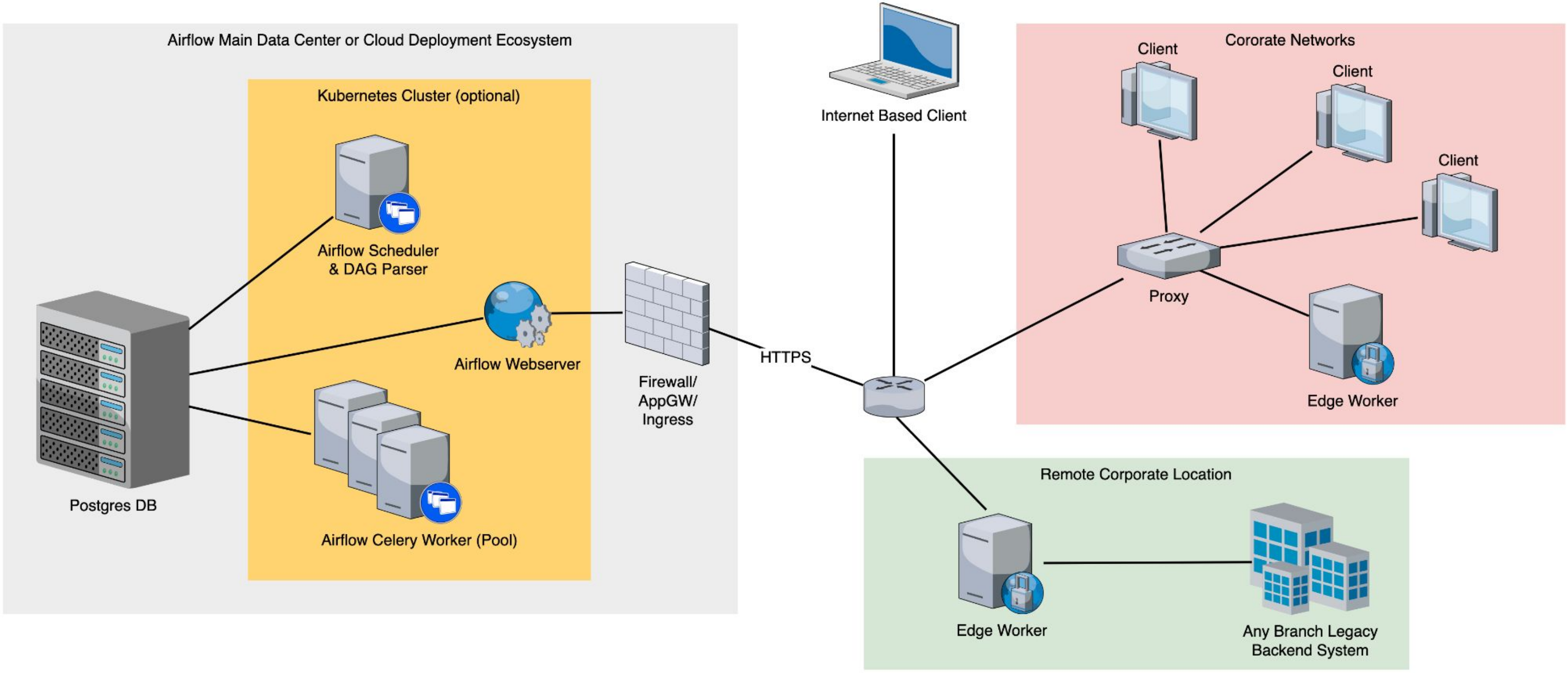
Version: 595f44fe

# Architecture

# Task isolation

(AIP-72, AIP-82, AIP-83, AIP-69)

# Distributed Airflow Architecture


Airflow Main Data Center or Cloud Deployment Ecosystem
Kubernetes Cluster (optional)
Airflow Scheduler & DAG Parser
Airflow Webserver
Postgres DB
Airflow Celery Worker (Pool)
Firewall/ AppGW/ Ingress
HTTPS
Internet Based Client
Cororate Networks
Client
Client
Client
Proxy
Edge Worker
Remote Corporate Location
Edge Worker
Any Branch Legacy Backend System

# Task SDK

- No direct DB access from task
- Small set of dependencies
- Non-Python native support
- Remote task execution

# Ad-hoc / Fast tasks

- No schedule involved
- Low overhead for  tasks
- Inference ready
- “Almost” streaming
- External-event driven scheduling

# Edge execution

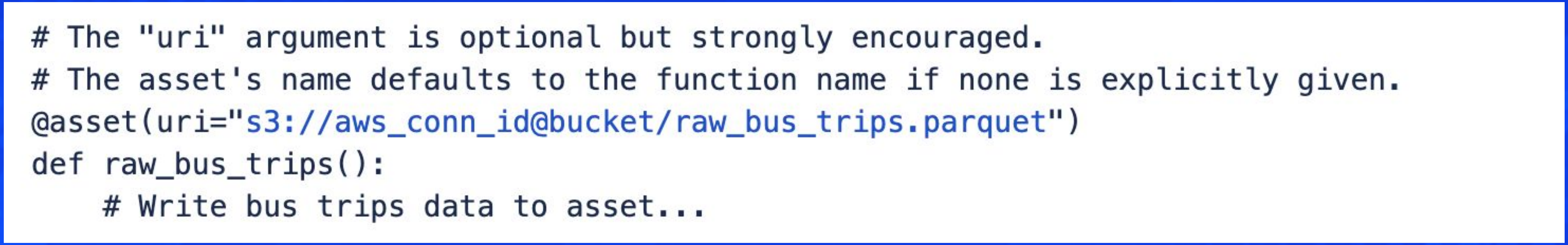
- Agent execution
- Very low overhead
- Enterprise-security
- Multiple Platform support
- Multi-language support

# Dag Authoring

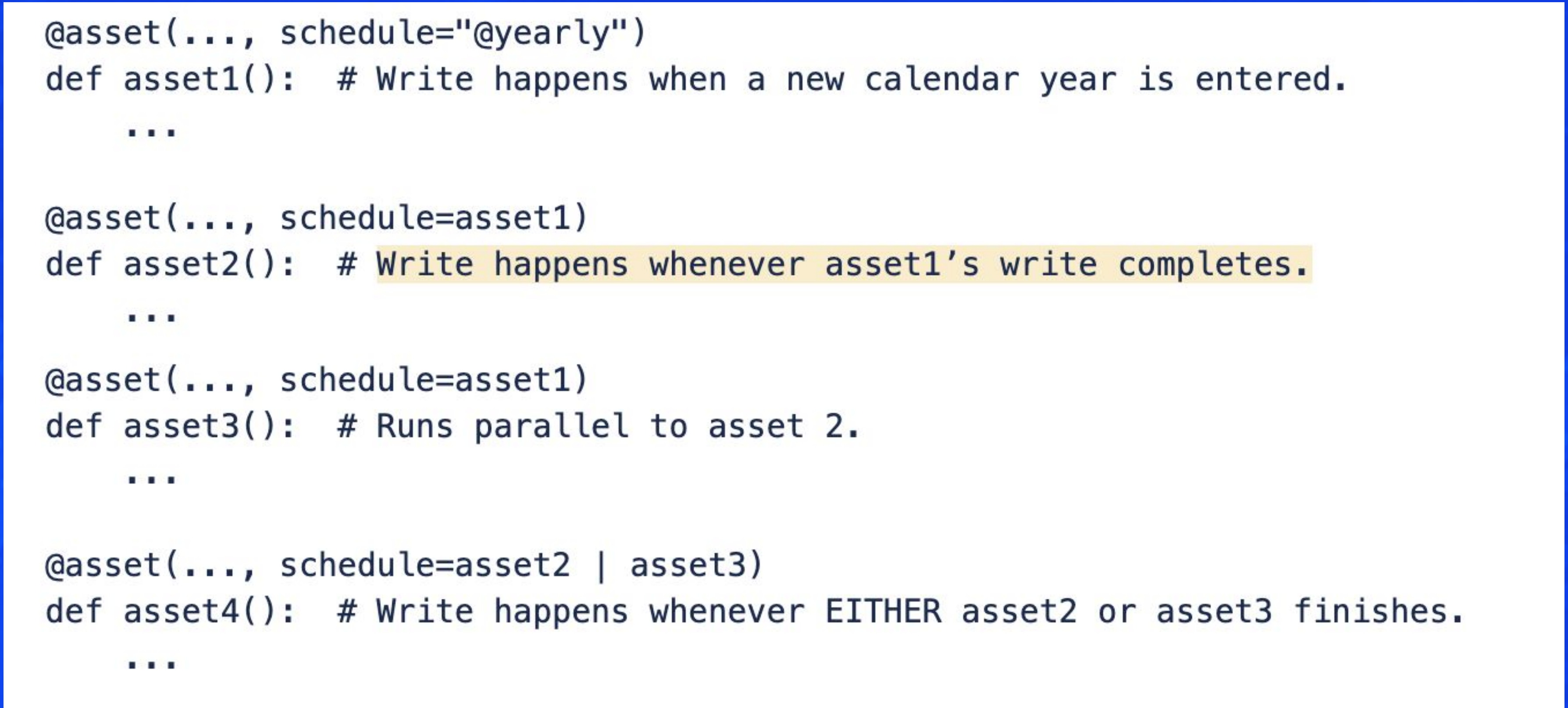
# Data Assets and Partitions

(AIP-74, AIP-75, AIP-76)

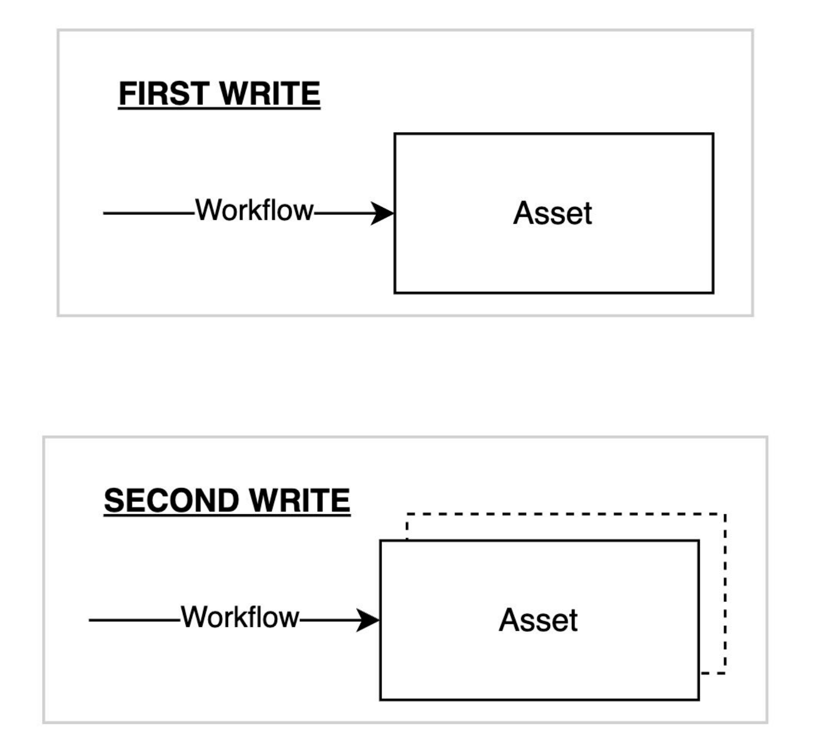
# Dataset changes to Asset

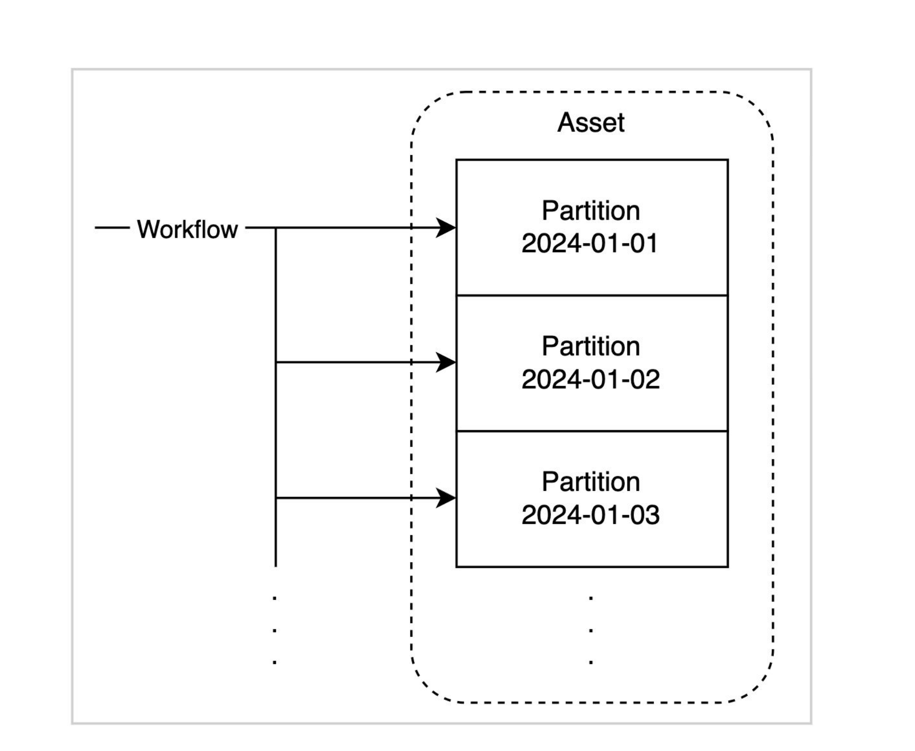


```
# The "uri" argument is optional but strongly encouraged.
# The asset's name defaults to the function name if none is explicitly given.
@asset(uri="s3://aws_conn_id@bucket/raw_bus_trips.parquet")
def raw_bus_trips():
    # Write bus trips data to asset...
```



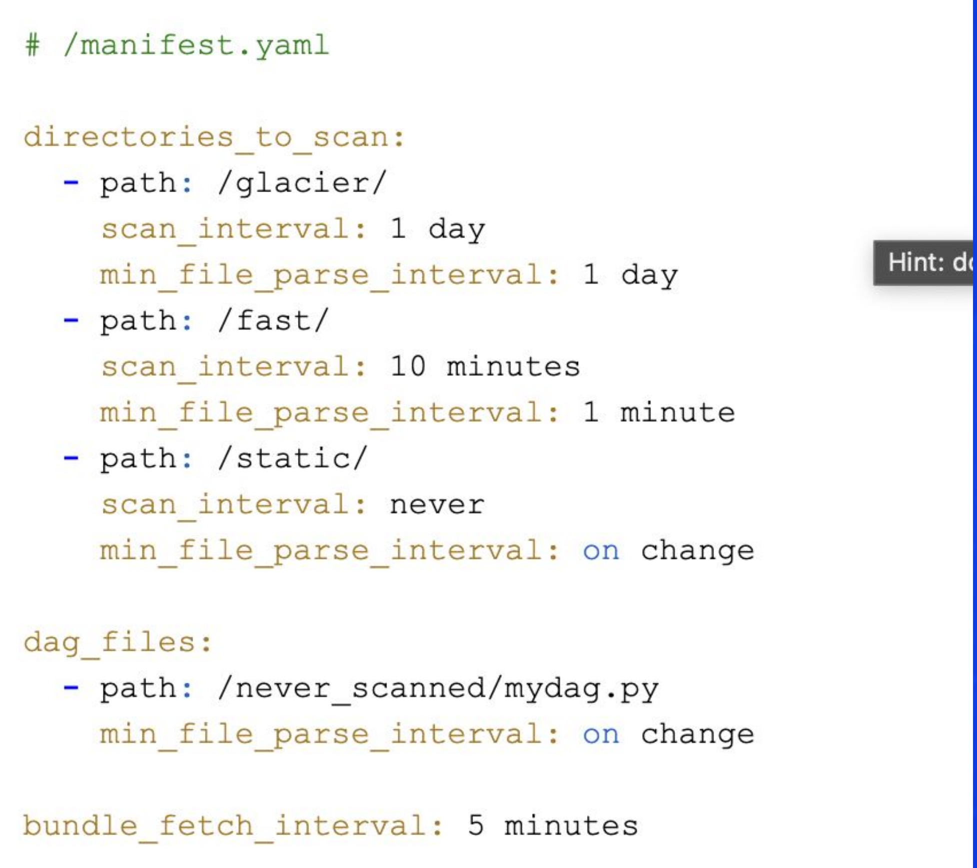
```
@asset(..., schedule="@yearly")
def asset1():  # Write happens when a new calendar year is entered.
    ...

@asset(..., schedule=asset1)
def asset2():  # Write happens whenever asset1’s write completes.
    ...
@asset(..., schedule=asset1)
def asset3():  # Runs parallel to asset 2.
    ...

@asset(..., schedule=asset2 | asset3)
def asset4():  # Write happens whenever EITHER asset2 or asset3 finishes.
    ...
```


FIRST WRITE
Workflow
Asset
SECOND WRITE
Workflow
Asset

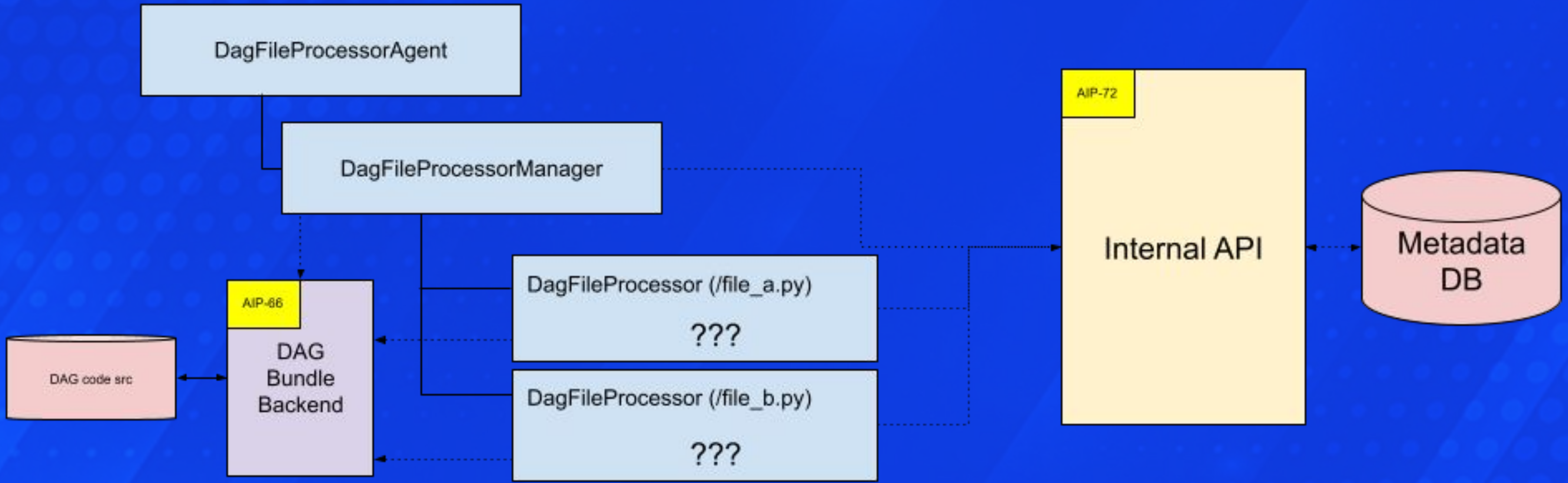

Asset
Workflow
Partition
2024-01-01
Partition
2024-01-02
Partition
2024-01-03

# DAG Bundles / Generalize DAG loading

(AIP-66/AIP-71)



```
# /manifest.yaml

directories_to_scan:
  - path: /glacier/
    scan_interval: 1 day
    min_file_parse_interval: 1 day
  - path: /fast/
    scan_interval: 10 minutes
    min_file_parse_interval: 1 minute
  - path: /static/
    scan_interval: never
    min_file_parse_interval: on change

dag_files:
  - path: /never_scanned/mydag.py
    min_file_parse_interval: on change

bundle_fetch_interval: 5 minutes
```


DagFileProcessorAgent
DagFileProcessorManager
AIP-66
DAG Bundle Backend
DAG code src
DagFileProcessor (/file_a.py)
???
DagFileProcessor (/file_b.py)
???
AIP-72
Internal API
Metadata DB

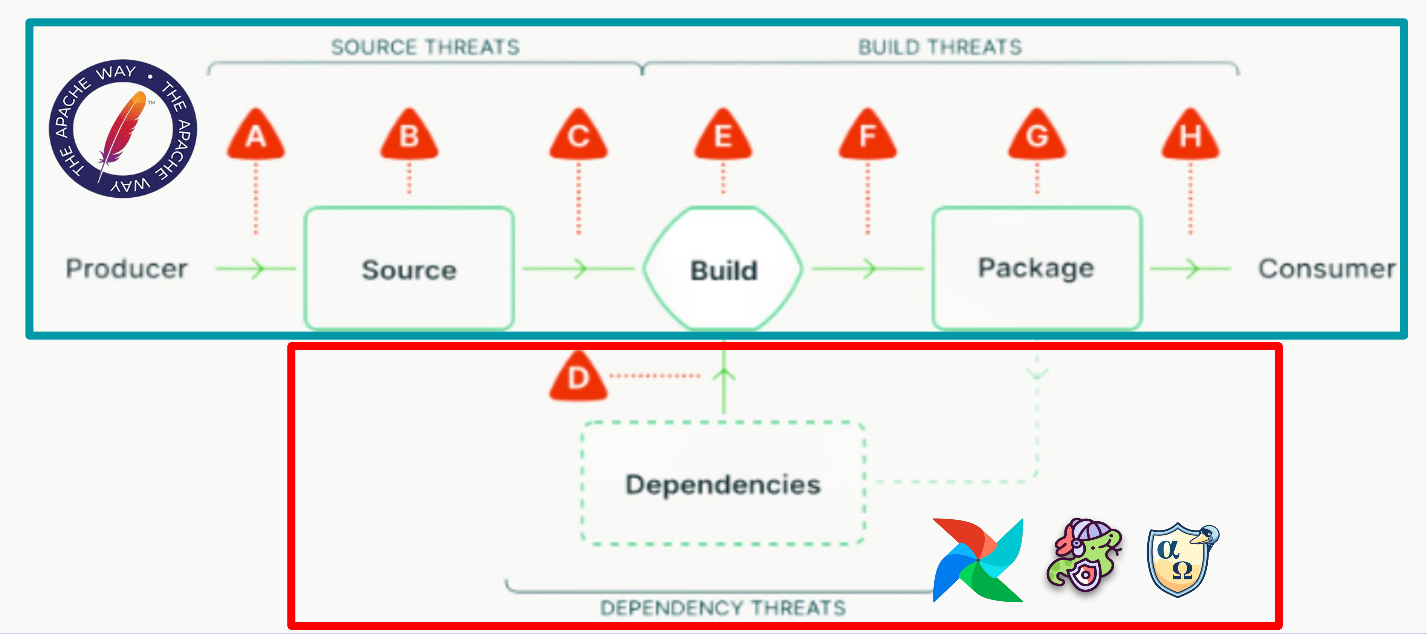
# Easy Deployment / Running

# Easy deployment / running

- Standalone Airflow
- Improve messaging and debugging (Survey !!!)
- Separate dependencies for Core/Providers/SDK/UI
- Decouple Authentication and Authorization
- Enterprise-grade deployment options

# Security

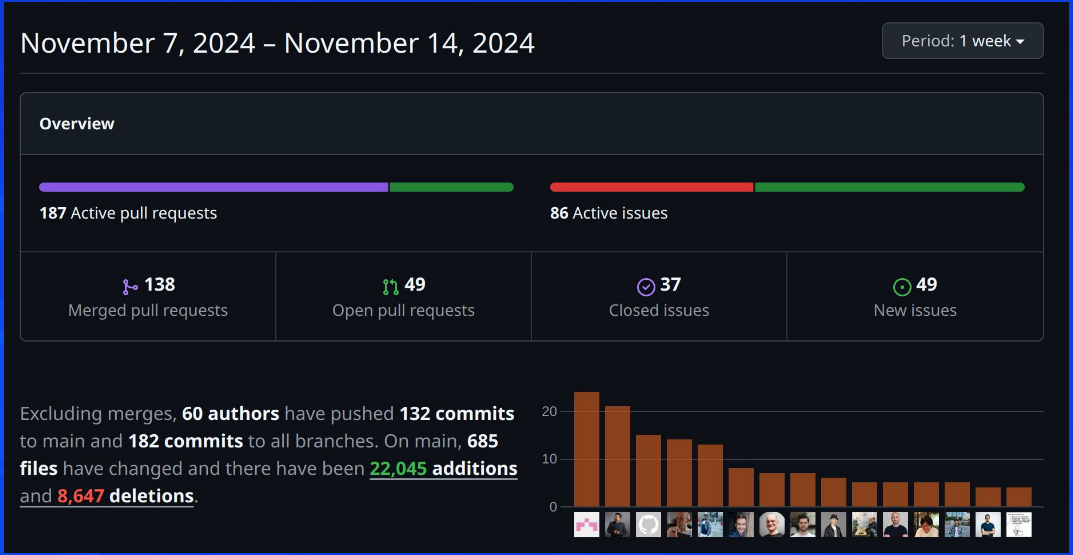
# SLSA Threat model



# More / Future

- React UI plugins
- UI-controlled backfills
- Explicit Template Fields
- Lineage auto-hook instrumentation
- Enhanced CLI Security
- Multi-team
- Asset auto read/write

# Airflow is more active than ever



November 7, 2024 – November 14, 2024

Period: 1 week ▾

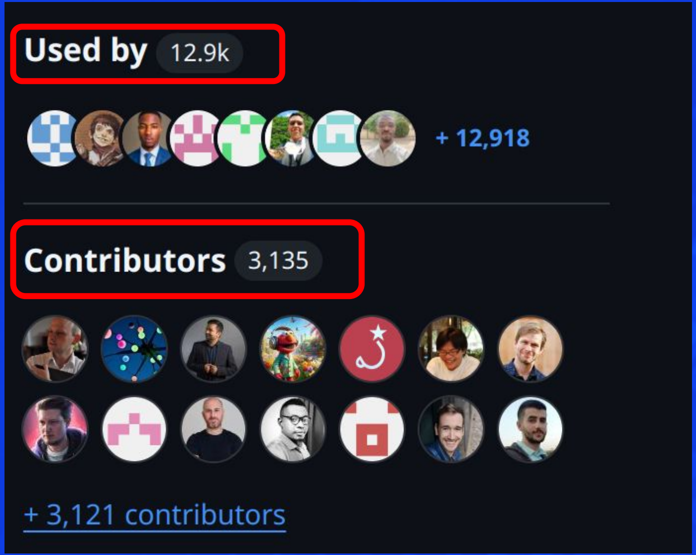
**Overview**

**187** Active pull requests

**86** Active issues

| | | | |
|---|---|---|---|
| **138**<br>Merged pull requests | **49**<br>Open pull requests | **37**<br>Closed issues | **49**<br>New issues |

Excluding merges, **60 authors** have pushed **132 commits** to main and **182 commits** to all branches. On main, **685 files** have changed and there have been **22,045 additions** and **8,647 deletions**.



**Used by** 12.9k

+ 12,918

**Contributors** 3,135

+ 3,121 contributors

# How users are affected?

# When ?

Early

Q2 2025