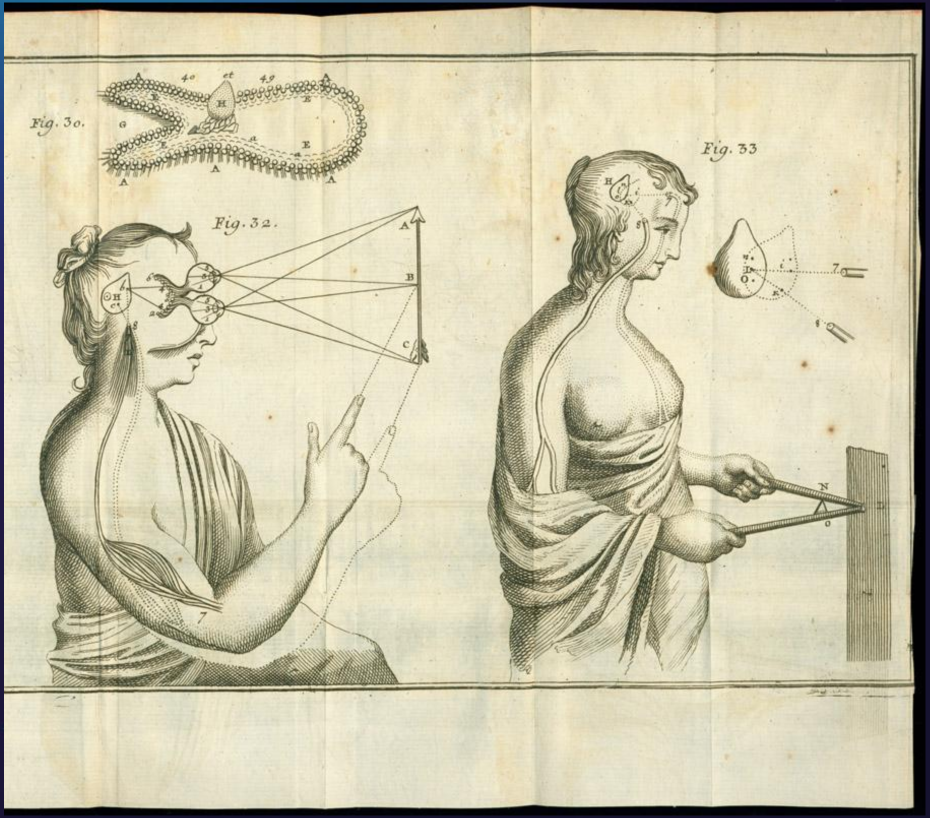
Lessons from the Ecosystem

A little pedantic, don't you think?

Some questionable insights from attempts to measure adoption in the Airflow community and beyond.

Question: what's a PyPI package other than Airflow you're interested in?



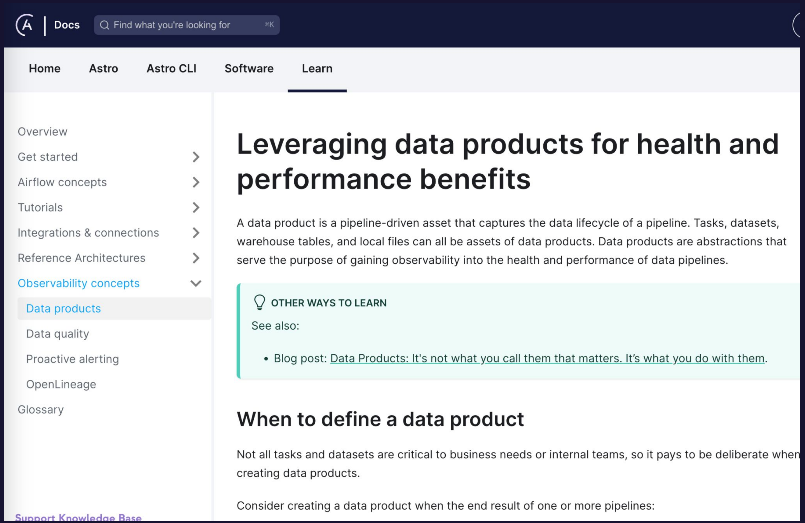
Airflow does not exist

https://thomistica.net/essays/2018/11/9/a-thomistic-critique-of-cartesian-dualism

This talk

1. Frivolous introduction
2. Apache Airflow object metrics
    a. Challenges & an opportunity
3. An example of an Airflow-based solution
    a. Dynamic DAGs
    b. PyPI API
    c. Snowflake
    d. BigQuery
    e. Superset
4. Insights of questionable value!

About me



Leveraging data products for health and performance benefits

A data product is a pipeline-driven asset that captures the data lifecycle of a pipeline. Tasks, datasets, warehouse tables, and local files can all be assets of data products. Data products are abstractions that serve the purpose of gaining observability into the health and performance of data pipelines.

OTHER WAYS TO LEARN

See also:

- Blog post: Data Products: It's not what you call them that matters. It's what you do with them.

When to define a data product

Not all tasks and datasets are critical to business needs or internal teams, so it pays to be deliberate when creating data products.

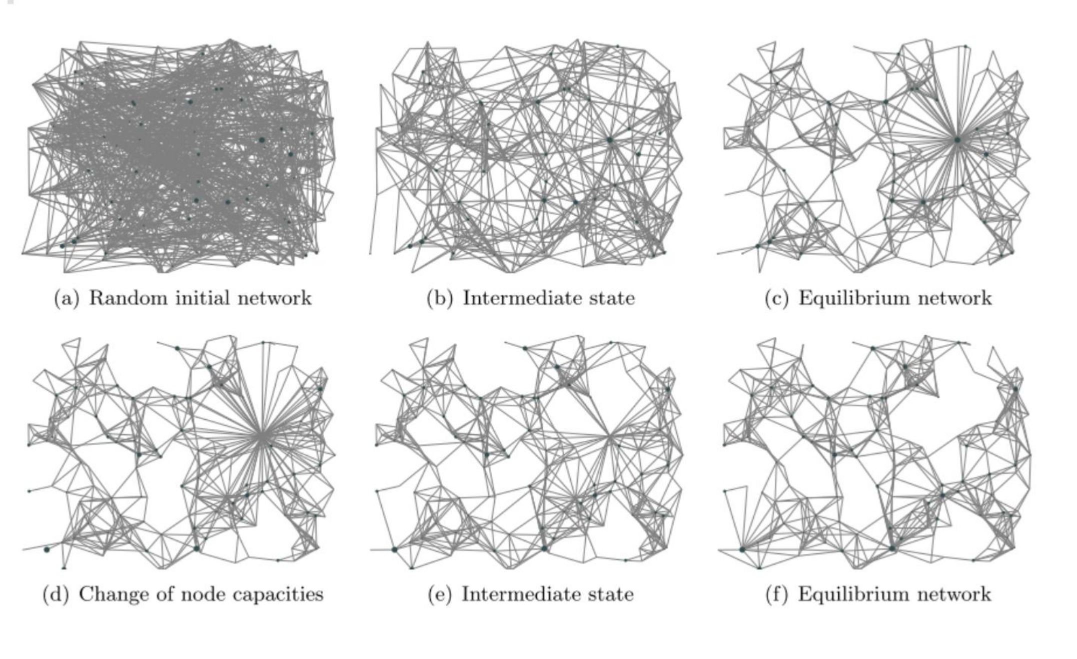
Consider creating a data product when the end result of one or more pipelines:

OpenLineage

Frivolous Introduction

ASTRONOMER

A massively distributed system



(a) Random initial network

(b) Intermediate state

(c) Equilibrium network

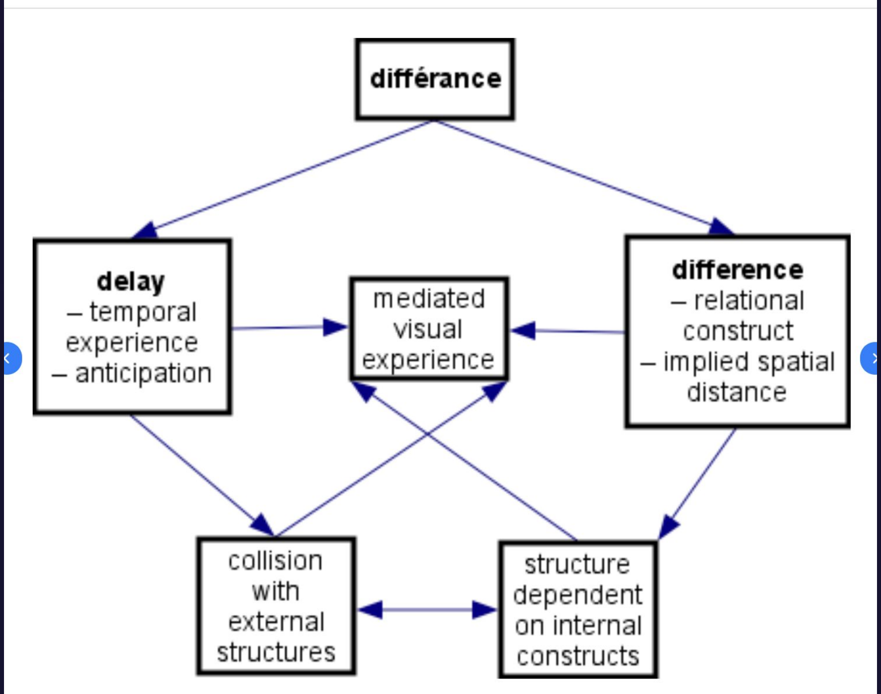
(d) Change of node capacities

(e) Intermediate state

(f) Equilibrium network

https://www.sg.ethz.ch/publications/2012/scholtes2012organic-design-of/

Community ~ "différance"


différance
delay
– temporal experience
– anticipation
mediated visual experience
difference
– relational construct
– implied spatial distance
collision with external structures
structure dependent on internal constructs

https://www.researchgate.net/figure/Visual-interpretation-of-Jacques-Derridas-differance-adapted-from-Malcolm-Richards_fig9_320271353

understanding(perspective):

OCEANUS
MARE
MEDITERRANEUM
SINUS PERSICUS
SINUS GANGETICUS
OCEANUS INDICUS MERIDIONAL

https://images.app.goo.gl/h45vu3dvaxwb1wXTA

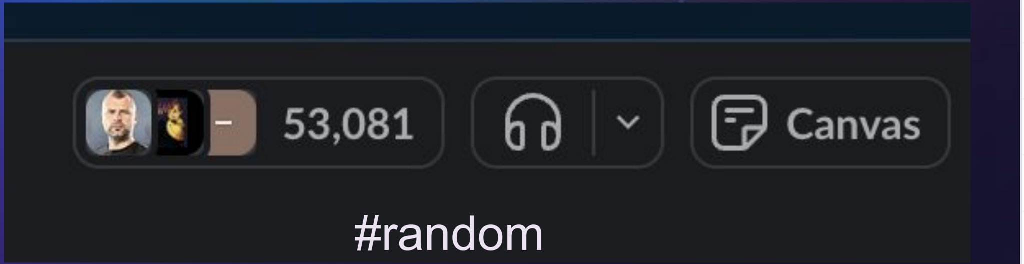
understanding(perspective):

CARTA VNIVERSAL EN QVE SE CONTIENE TODO LO QVE DEL MVNDO SEA DESCVBI
FASTA AORA HIZOLA VN COSMOGRAPHO DE SV MAGES TAD ANNO M D XX VII EN SEVILL
MVNDVS NOVVS
AFRICA
ETHIOPIA
BRASIL
MARE INDICVM

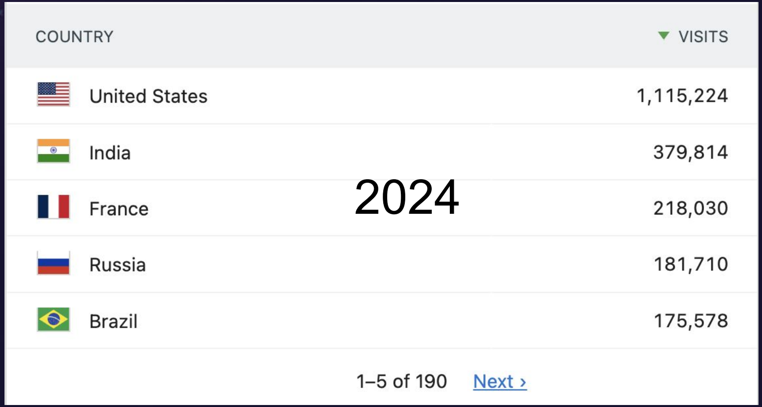
https://images.app.goo.gl/7bjvVz6wLzdXt8kQ7

Airflow “Object” Metrics

Some data points of questionable value



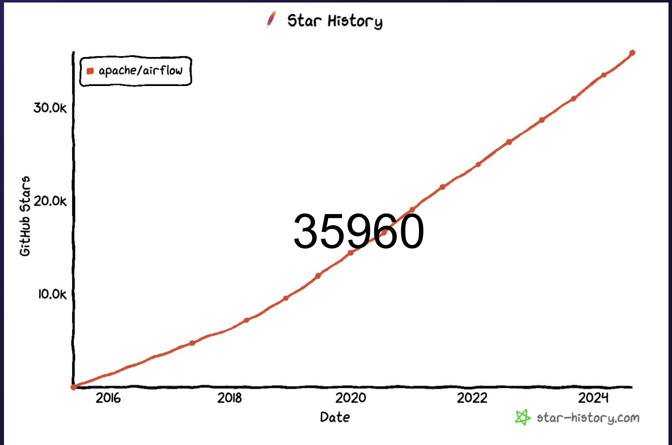
#random



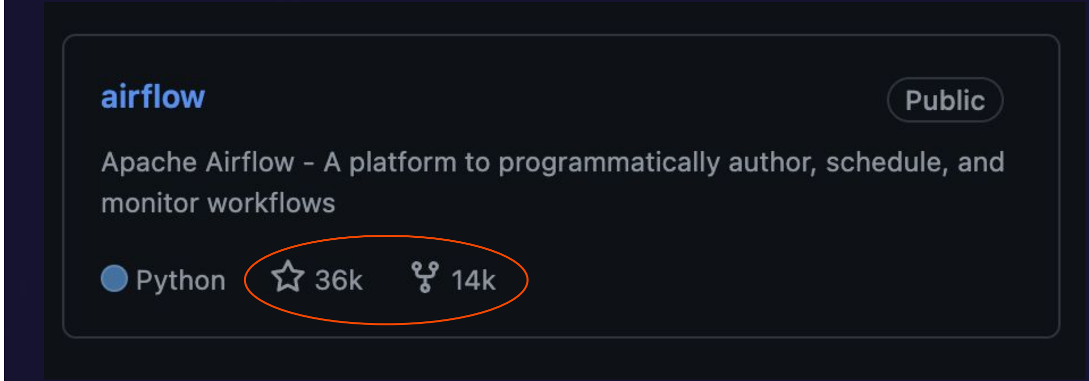
COUNTRY	▼ VISITS
United States	1,115,224
India	379,814
France	218,030
Russia	181,710
Brazil	175,578

1–5 of 190 Next ›

2024



35960



Apache Airflow - A platform to programmatically author, schedule, and monitor workflows

Easily accessed metrics

Some key analytics are supported natively by platforms or by commonly used tools such as Common Room.

GitHub metrics

Issues opened
PRs opened
Contributor #s

Slack metrics

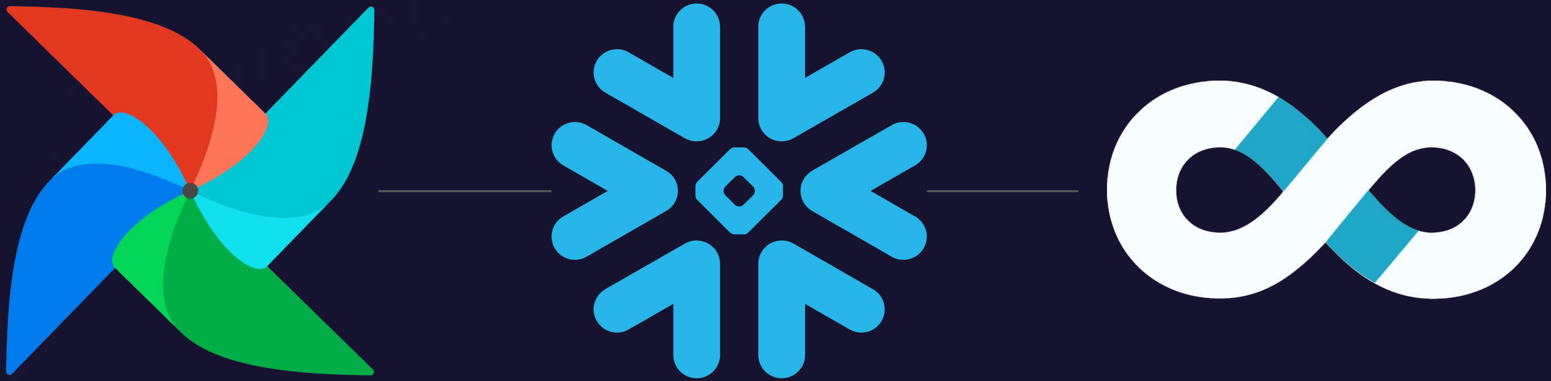
Members
Messages
Organizations

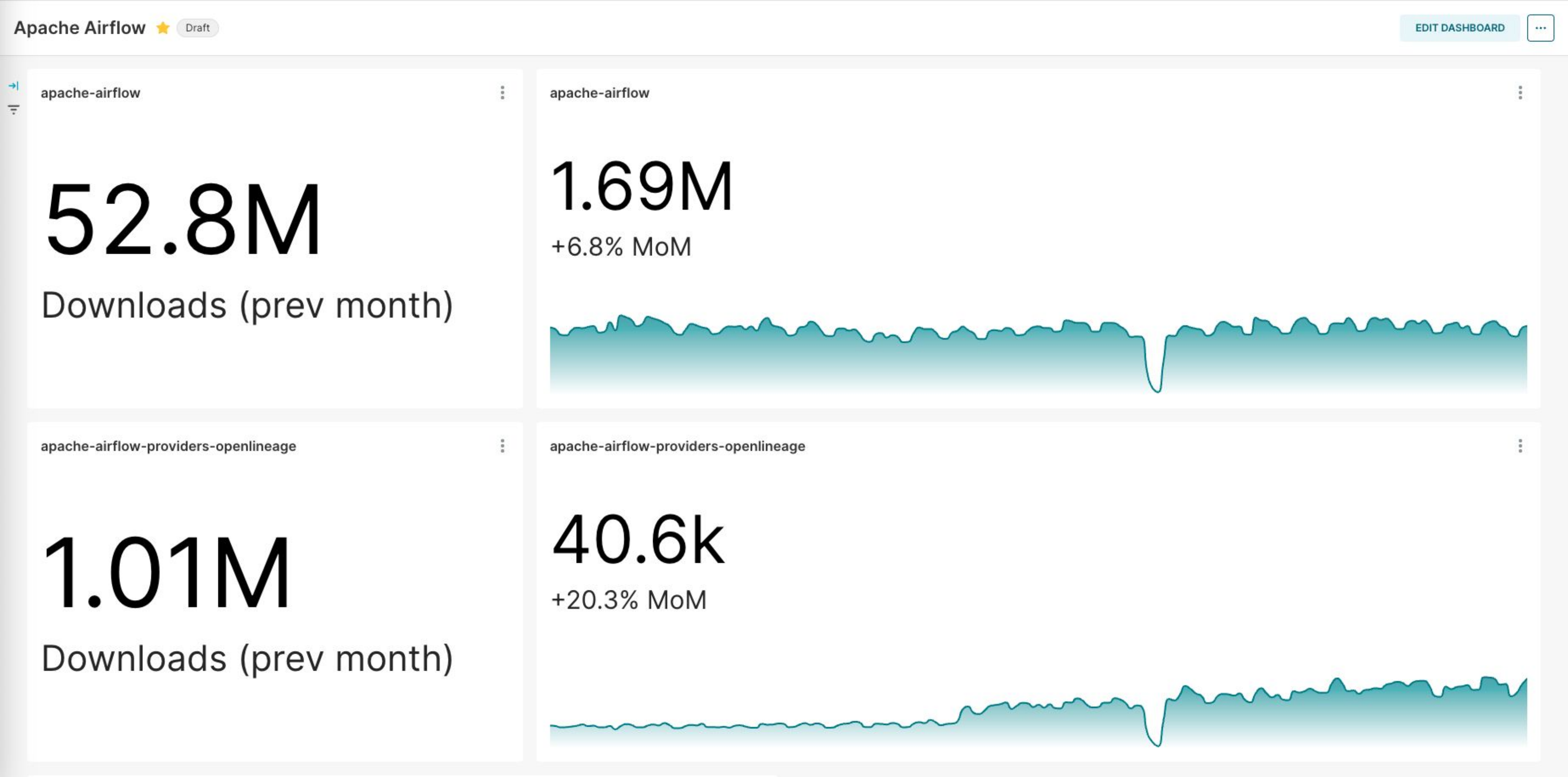
Website & docs analytics

Visitors
Popular pages
Engagement

OSS metrics solution

Dynamic DAGs + PyPI API + Snowflake + BigQuery + Superset





Apache Airflow

apache-airflow

52.8M

Downloads (prev month)

apache-airflow

1.69M

+6.8% MoM

apache-airflow-providers-openlineage

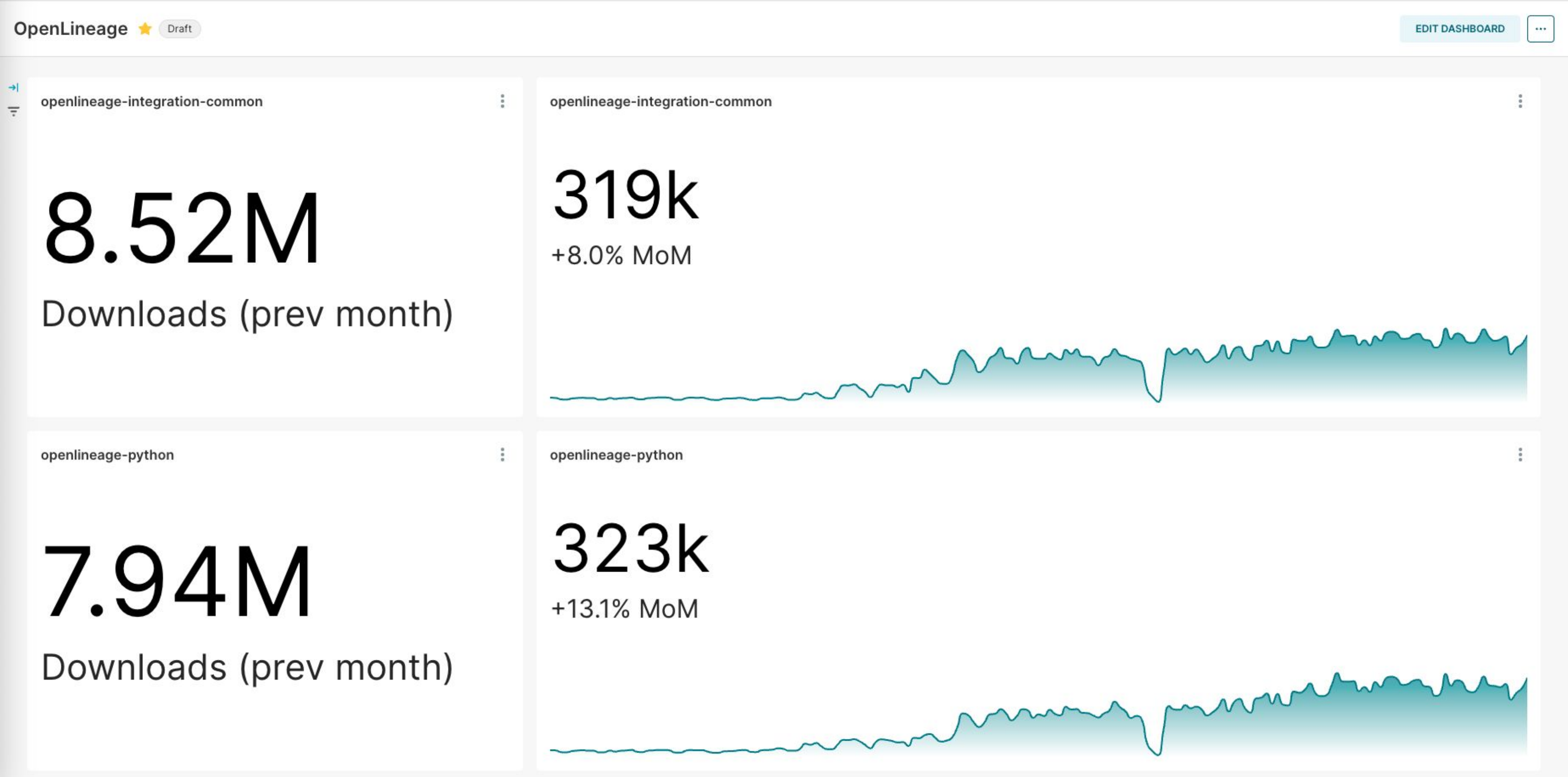
1.01M

Downloads (prev month)

apache-airflow-providers-openlineage

40.6k

+20.3% MoM



openlineage-integration-common

8.52M

Downloads (prev month)

openlineage-integration-common

319k

+8.0% MoM

openlineage-python

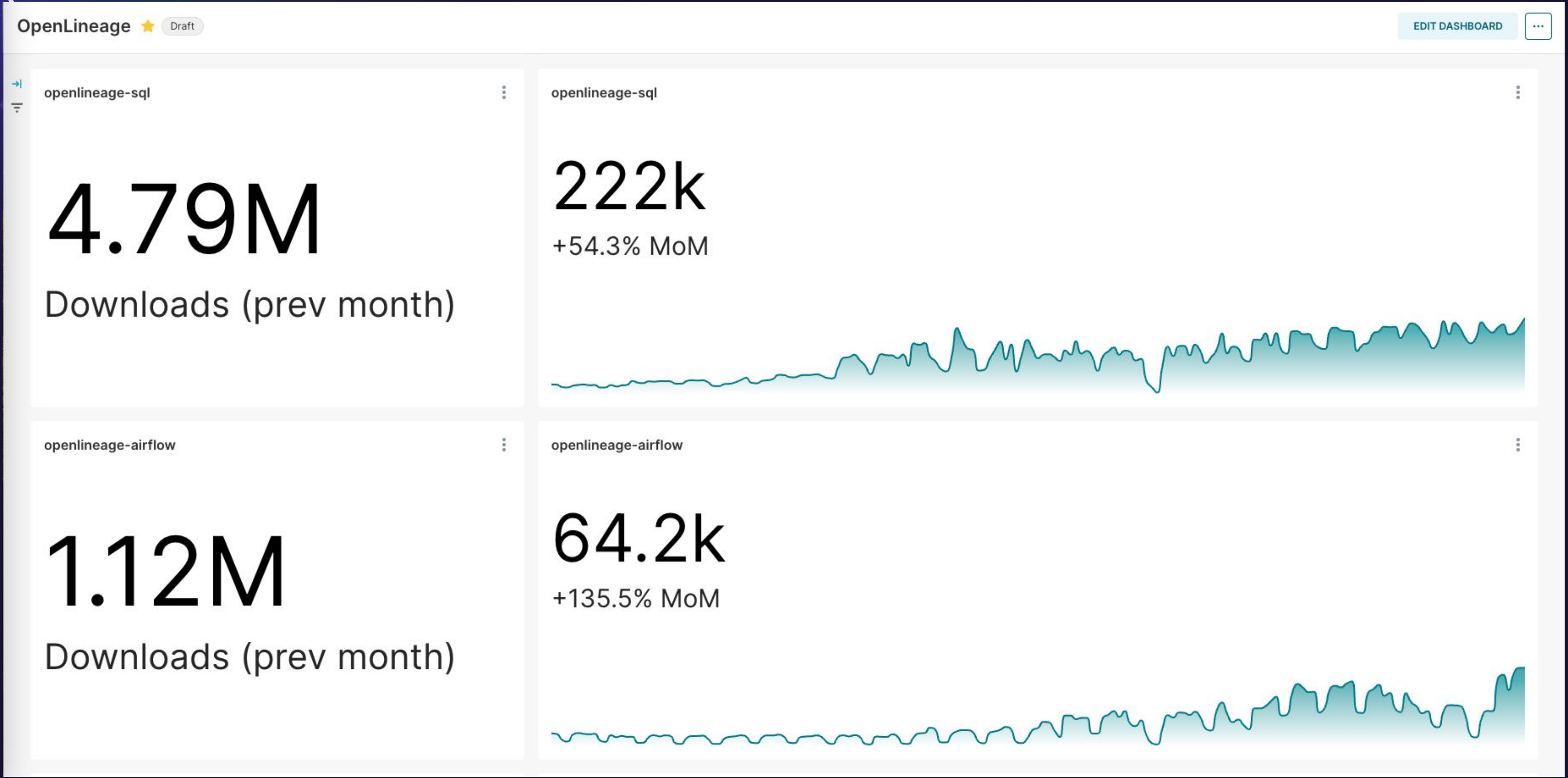
7.94M

Downloads (prev month)

openlineage-python

323k

+13.1% MoM



OpenLineage ★ Draft

openlineage-sql

4.79M

Downloads (prev month)

openlineage-sql

222k

+54.3% MoM

openlineage-airflow

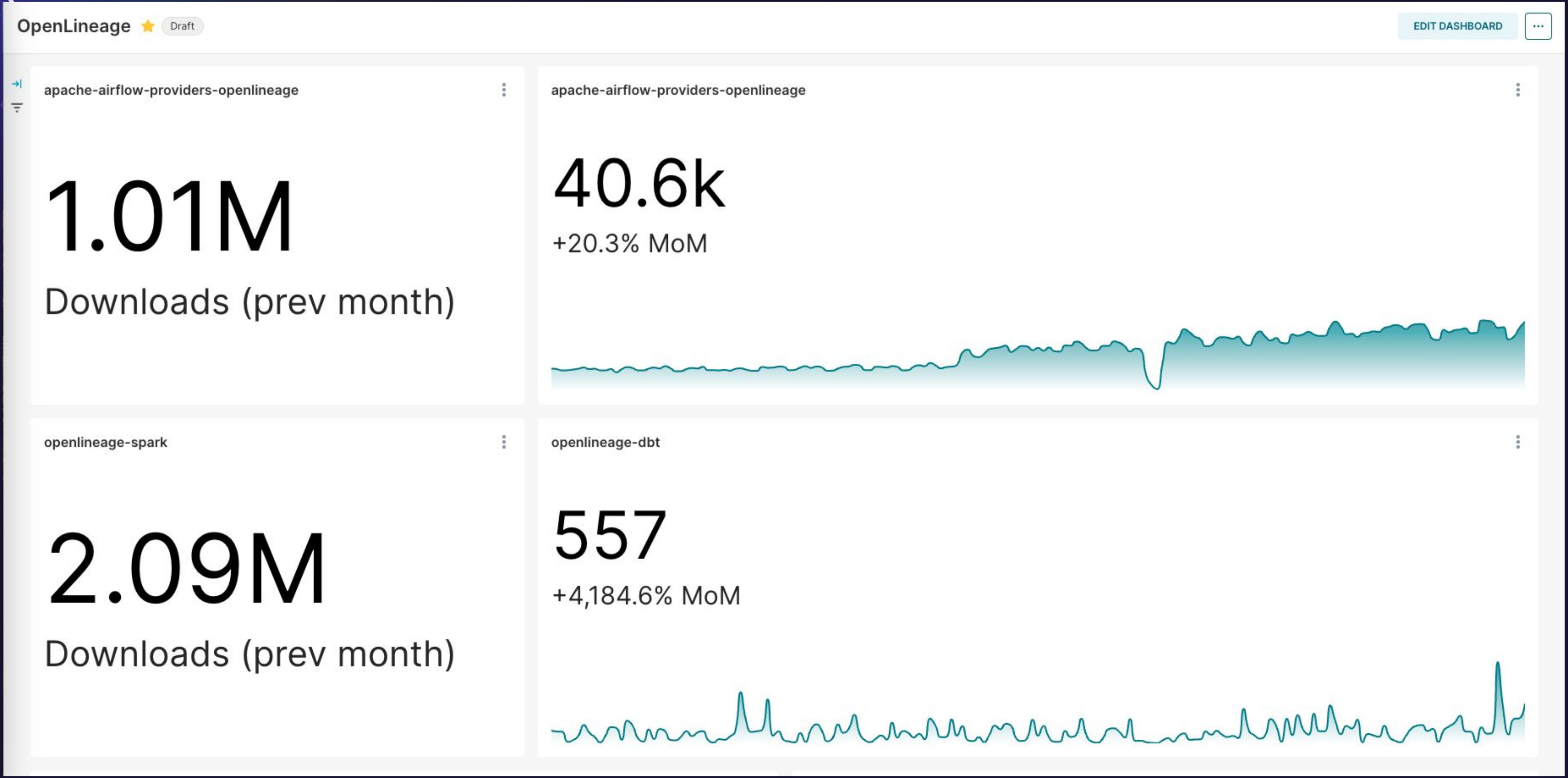
1.12M

Downloads (prev month)

openlineage-airflow

64.2k

+135.5% MoM



OpenLineage

apache-airflow-providers-openlineage

1.01M

Downloads (prev month)

apache-airflow-providers-openlineage

40.6k

+20.3% MoM

openlineage-spark

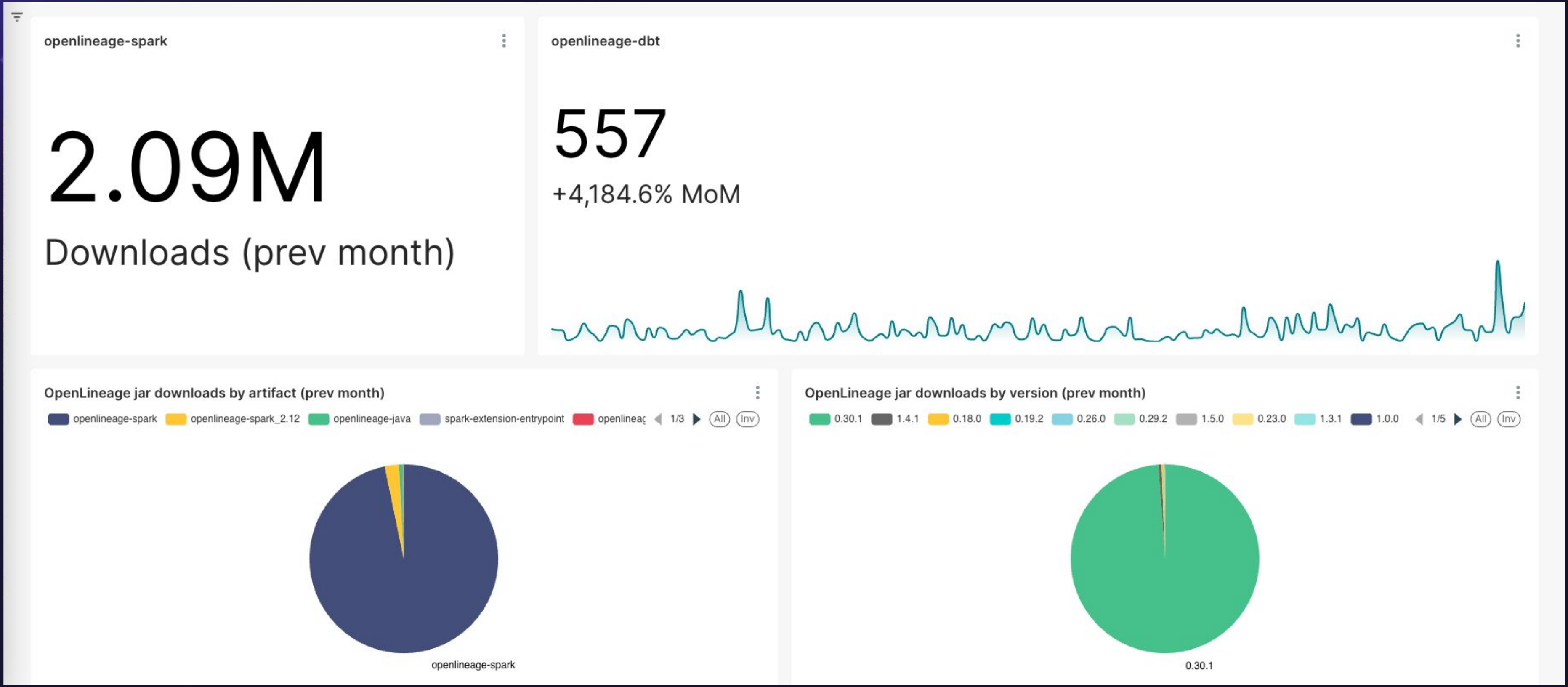
2.09M

Downloads (prev month)

openlineage-dbt

557

+4,184.6% MoM



openlineage-spark

2.09M

Downloads (prev month)

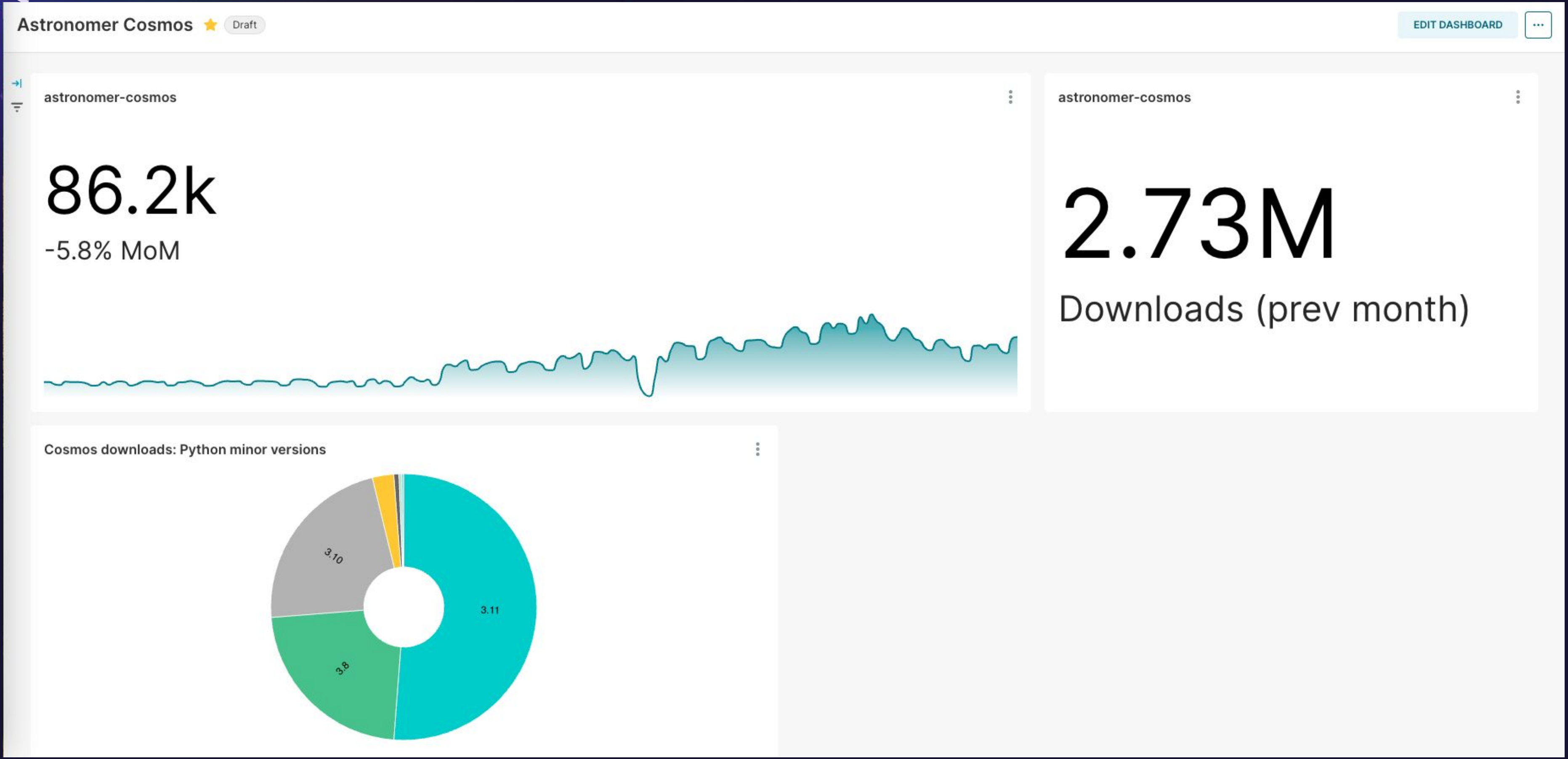
openlineage-dbt

557

+4,184.6% MoM

OpenLineage jar downloads by artifact (prev month)

OpenLineage jar downloads by version (prev month)



Astronomer Cosmos

astronomer-cosmos

86.2k

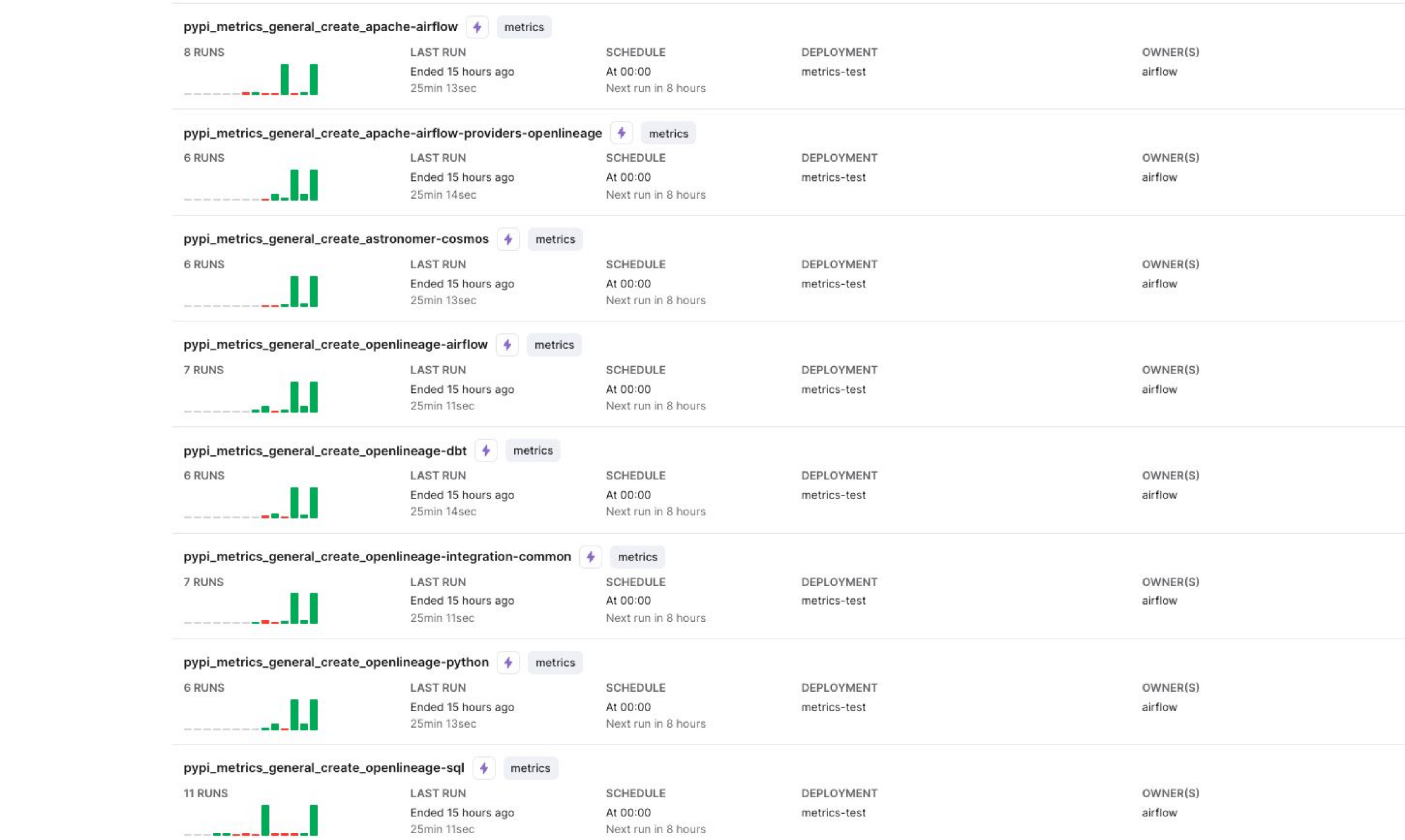
-5.8% MoM

astronomer-cosmos

2.73M

Downloads (prev month)

Cosmos downloads: Python minor versions



pypi_metrics_general_create_apache-airflow metrics

8 RUNS

LAST RUN
Ended 15 hours ago
25min 13sec

SCHEDULE
At 00:00
Next run in 8 hours

DEPLOYMENT
metrics-test

OWNER(S)
airflow

pypi_metrics_general_create_apache-airflow-providers-openlineage metrics

6 RUNS

LAST RUN
Ended 15 hours ago
25min 14sec

SCHEDULE
At 00:00
Next run in 8 hours

DEPLOYMENT
metrics-test

OWNER(S)
airflow

pypi_metrics_general_create_astronomer-cosmos metrics

6 RUNS

LAST RUN
Ended 15 hours ago
25min 13sec

SCHEDULE
At 00:00
Next run in 8 hours

DEPLOYMENT
metrics-test

OWNER(S)
airflow

pypi_metrics_general_create_openlineage-airflow metrics

7 RUNS

LAST RUN
Ended 15 hours ago
25min 11sec

SCHEDULE
At 00:00
Next run in 8 hours

DEPLOYMENT
metrics-test

OWNER(S)
airflow

pypi_metrics_general_create_openlineage-dbt metrics

6 RUNS

LAST RUN
Ended 15 hours ago
25min 14sec

SCHEDULE
At 00:00
Next run in 8 hours

DEPLOYMENT
metrics-test

OWNER(S)
airflow

pypi_metrics_general_create_openlineage-integration-common metrics

7 RUNS

LAST RUN
Ended 15 hours ago
25min 11sec

SCHEDULE
At 00:00
Next run in 8 hours

DEPLOYMENT
metrics-test

OWNER(S)
airflow

pypi_metrics_general_create_openlineage-python metrics

6 RUNS

LAST RUN
Ended 15 hours ago
25min 13sec

SCHEDULE
At 00:00
Next run in 8 hours

DEPLOYMENT
metrics-test

OWNER(S)
airflow

pypi_metrics_general_create_openlineage-sql metrics

11 RUNS

LAST RUN
Ended 15 hours ago
25min 11sec

SCHEDULE
At 00:00
Next run in 8 hours

DEPLOYMENT
metrics-test

OWNER(S)
airflow

AIRFLOW__CORE__PARALLELISM=19

```
projects = [
        "openlineage-sql",
        "openlineage-integration-common",
        "openlineage-dbt",
        "openlineage-airflow",
        "openlineage-python",
        "openlineage-dagster",
        "apache-airflow-providers-openlineage",
        "apache-airflow",
        "astronomer-cosmos",
        ]

def create_dag(project, table_name):
        @dag(
                dag_id=f"pypi_metrics_general_create_{project}",
                start_date=datetime(2024, 1, 1, 16, 30),
                schedule="@daily",
                catchup=False,
                doc_md=__doc__,
                tags=["metrics"],
        )
        def pypi_metrics_general_create():
        /* … */
        overall_downloads_obj >> create_overall_table() >>
        load_overall_data_into_snowflake.expand(downloads=overall_downloads_obj) >> create_new_clean_table()

        pypi_metrics_general_create()

for project in projects:
        create_dag(
                project=project,
                table_name=project.replace('-', '_'),
        )
```

To dos

Improvements and additions planned.

Deployment

- Deploy Superset on an EC2 instance?
- Put the DAGs on GitHub and migrate the Astro deployment to the GH integration

DAGs

- For data, migrate from pypistats.org to public PyPI datasets on BigQuery
- Put the DAGs on GitHub and migrate the Astro deployment to the GH integration

Some Insights of Questionable Value

Insights at this time

Some very dubious gleanings!

Versions

Most openlineage jar downloads are an old version, 0.30.1

- Is one particularly large user tied to this version?

This talk

1. Frivolous introduction
2. Apache Airflow object metrics
    a. Challenges & an opportunity
3. An example of an Airflow-based solution
    a. Dynamic DAGs
    b. PyPI API
    c. Snowflake
    d. BigQuery
    e. Superset
4. Insights of questionable value!

Thank you!
Any questions?

This evening: OpenLineage Meetup

OpenLineage

Please sign up! Get the link at **https://openlineage.io/blog**

Where

Astronomer
8 California St.

When

This evening!
6-9 pm

Join us for in-depth talks & discussion over dinner!

Agenda

- **Unlocking Data Products with OpenLineage at Astronomer**: Julian LaNeve and Jason Ma, Astronomer.
- **OpenLineage: From Operators to Hooks** by Maciej Obuchowski, Astronomer+GetInData/Xebia.
- **Activating Operational Metadata with Airflow, Atlan and Openlineage** by Kacper Muda, GetInData/Xebia.
- **Hamilton, a Scaffold for all Your Python Platform Concerns (and a New OpenLineage Producer)** by Stefan Krawczyk
- **Lightning Talk on New Marquez Features and the Marquez Project Roadmap** by Willy Lulciuc, Marquez Lead, and Peter Hicks, Marquez Committer.