San Francisco
12 September 2024

A New DAG Paradigm

Less Airflow, More DAGs

Maggie Stark Senior Data Scientist
Marion Azoulai Senior Data Scientist

ASTRONOMER

Astronomer

The driving force behind Apache Airflow

5 offices | 249 employees | 24×7 worldwide support

100%

Drives 100% of Airflow releases

55%

Of Airflow code contributed

18 of 25

18 of the top 25 committers on board, 8 PMC members

30K+

30K+ Airflow students in Academy ecosystem

Data Team

- Centralized Data Team
- Building critical operational and analytical pipelines with Airflow
- Product Influencers
  - Providing a Data-First Perspective

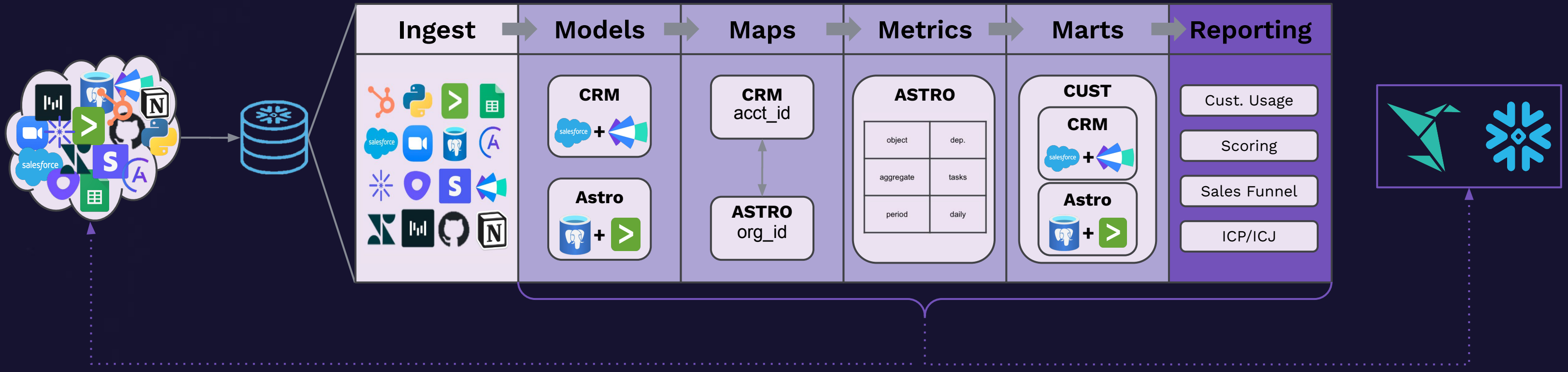
Data Team Ecosystem

Our day-to-day work is standard ELT

Ingest External Sources

Transform and Propagate

Deliver Data and Insights



Ingest

Models

Maps

Metrics

Marts

Reporting

CRM

Astro

CRM
acct_id

ASTRO
org_id

ASTRO

object	dep.
aggregate	tasks
period	daily

CUST

CRM

Astro

Cust. Usage

Scoring

Sales Funnel

ICP/ICJ

Flow

Exposure

What we do

- Application Integrations
- Outbound Marketing
- Account Scoring
- Centralized Analytics
- Billing
- Embedded Dashboards
- Sales Alerting
- Warehouse Governance
- rETL
- Cost Controls
- Ad Hoc Analysis
- Product Testing

To name a few...

Our Initial Architecture

Focused on governance and onboarding.

A DAG Factory

Quickly stand up pipelines.

- Abstracts Airflow
- Remain in a familiar context
  - Generate tasks from SQL, R, YAML, Notebooks, etc.

Custom TaskGroups

Standardizes pipeline operations.

- Prioritizes code reusability
- Contract mechanism for production suitability

Sensors

Manage cross DAG dependencies.

- Asynchronous when possible

Configuration as Code

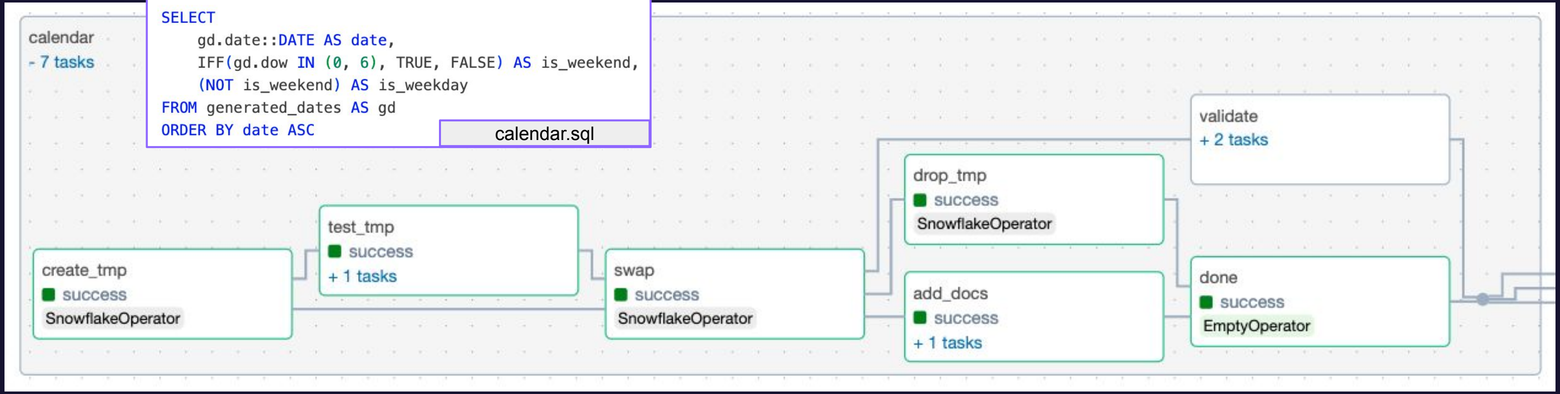
```
/*
operator: include.task_groups.transform.CreateTable
description: >-
    Creates standard calendar, where each row
    represents a date with common date operations.
fields:
    date: Date (iso format yyyy-mm-dd).
    is_weekend: Date is Sat. or Sun.
    is_weekday: Date is weekday (MTWTF).
schema: !switch_value
    sandbox: env__sandbox_schema
    default: commons
primary_key:
    - date
tests:
    check_null:
        - is_weekend
validations:
    check_condition:
        - is_weekend != is_weekday
*/

SELECT
    gd.date::DATE AS date,
    IFF(gd.dow IN (0, 6), TRUE, FALSE) AS is_weekend,
    (NOT is_weekend) AS is_weekday
FROM generated_dates AS gd
ORDER BY date ASC
```

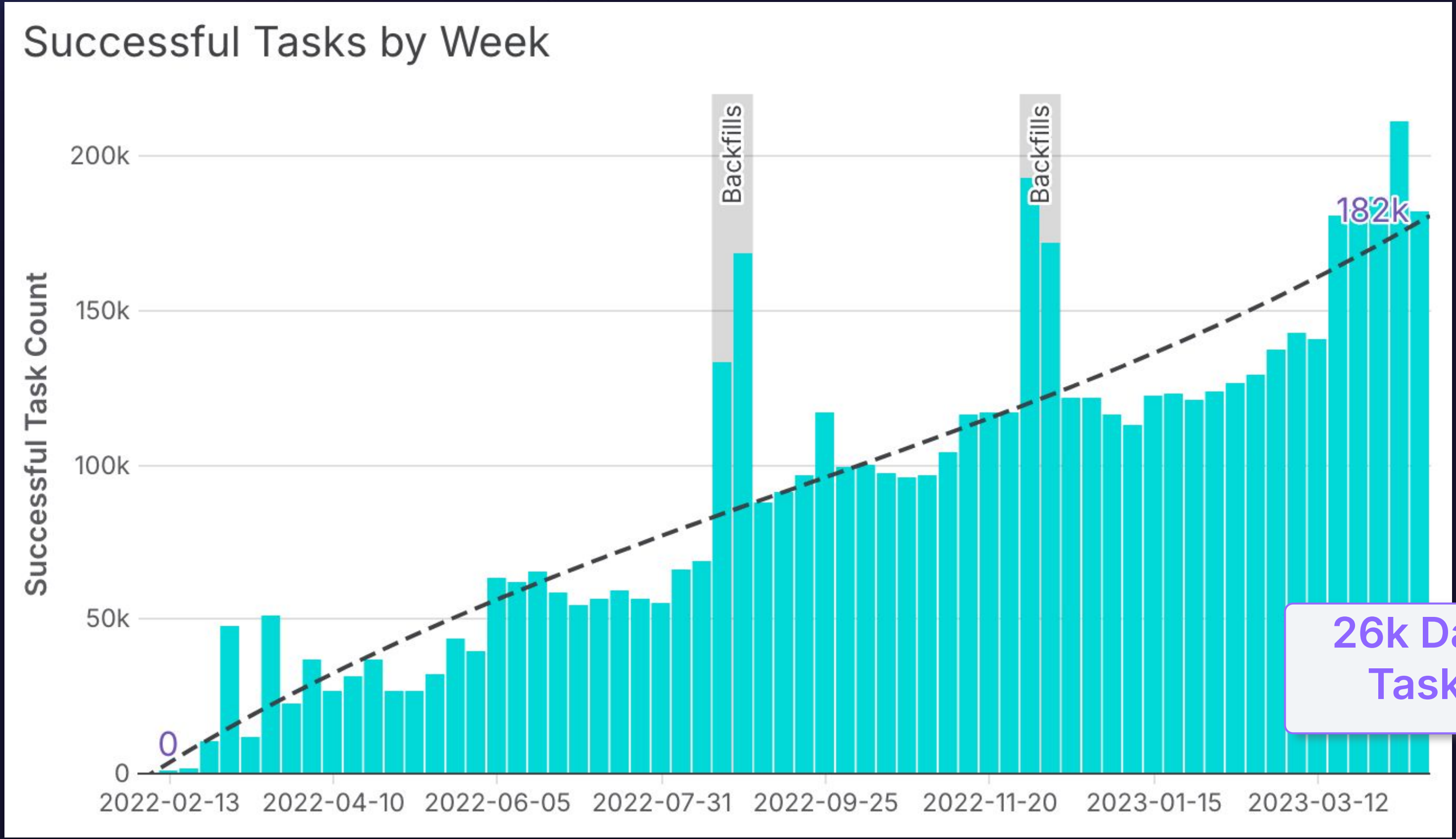
calendar.sql

1 file = n tasks!

Create a table; get documentation and testing come free.

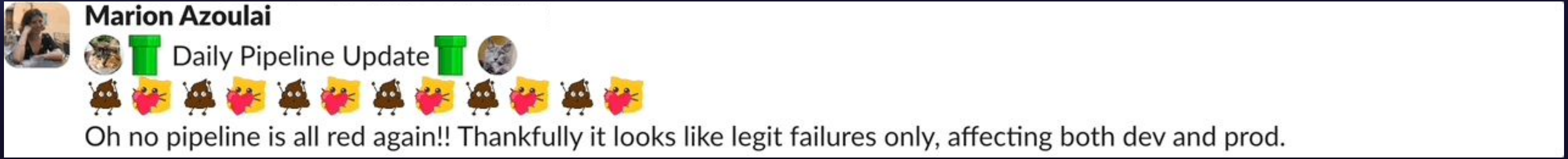


And this was great!



26k Daily Tasks

Until it wasn't...

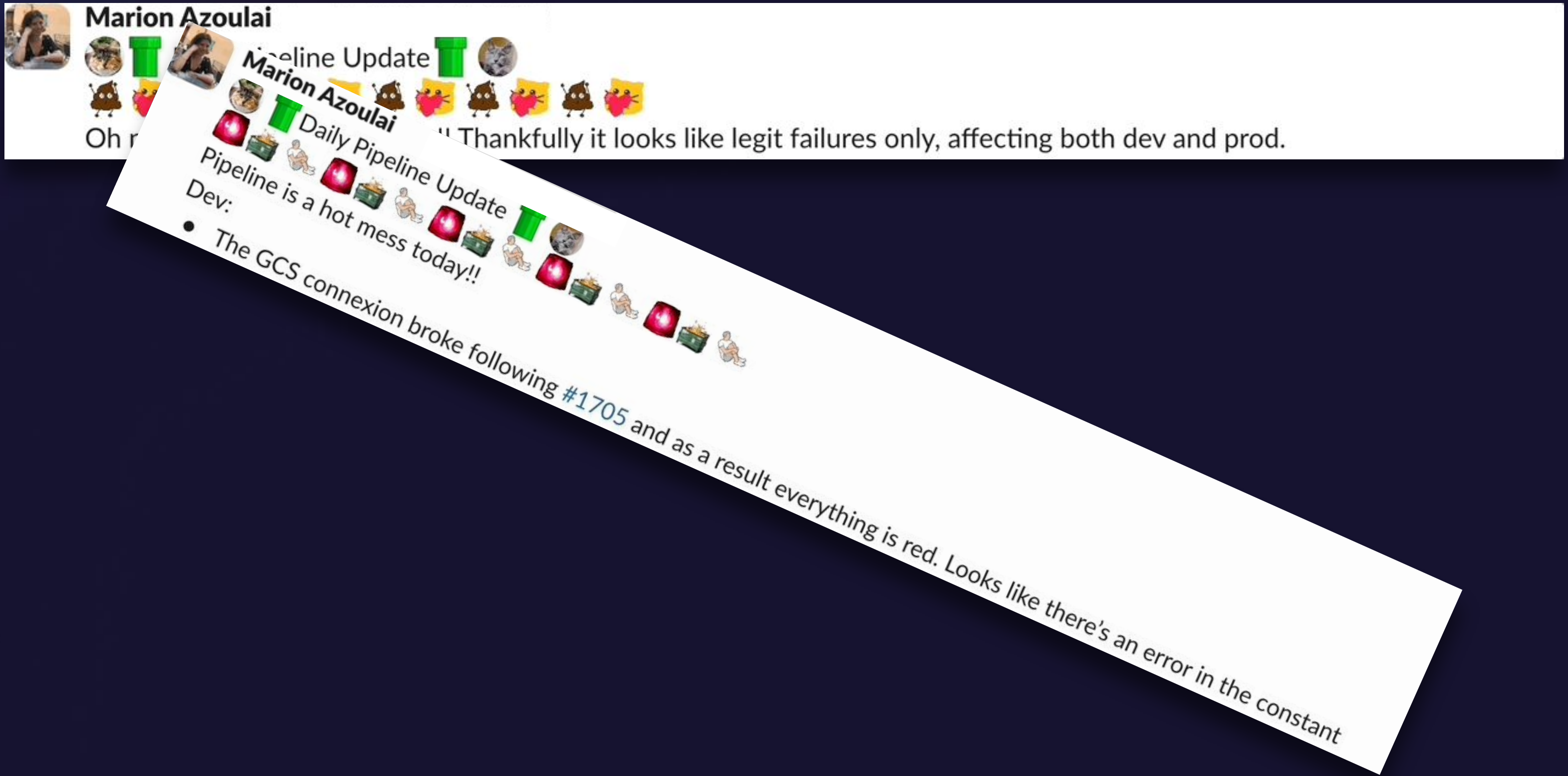


Marion Azoulai

Daily Pipeline Update

Oh no pipeline is all red again!! Thankfully it looks like legit failures only, affecting both dev and prod.

Until it wasn't...



Marion Azoulai

…eline Update

Oh r… !! Thankfully it looks like legit failures only, affecting both dev and prod.

Marion Azoulai

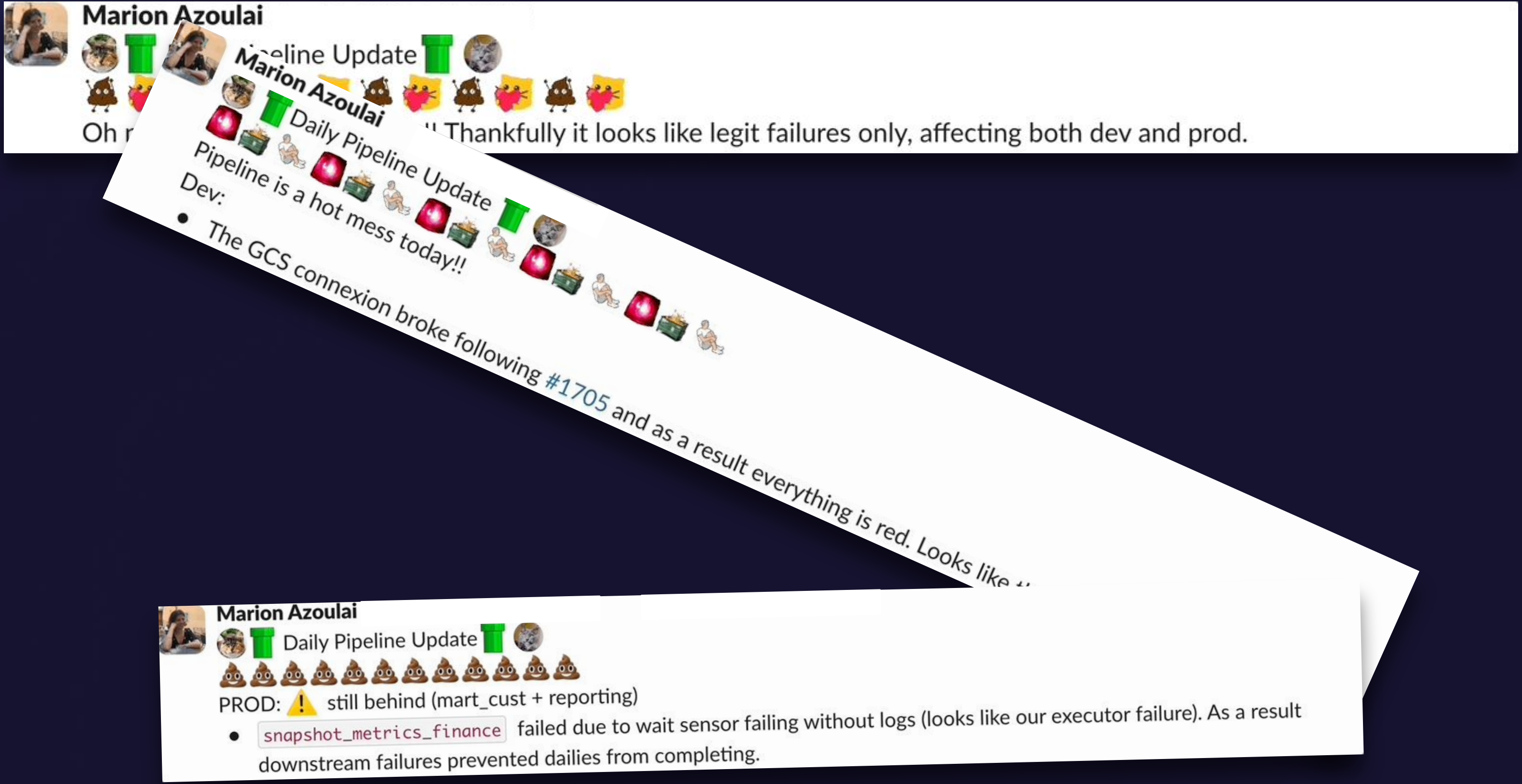
Daily Pipeline Update

Pipeline is a hot mess today!!

Dev:

- The GCS connexion broke following #1705 and as a result everything is red. Looks like there's an error in the constant

Until it wasn't...



Marion Azoulai

…eline Update

Oh r… !! Thankfully it looks like legit failures only, affecting both dev and prod.

Marion Azoulai

Daily Pipeline Update

Pipeline is a hot mess today!!

Dev:

- The GCS connexion broke following #1705 and as a result everything is red. Looks like t…

Marion Azoulai

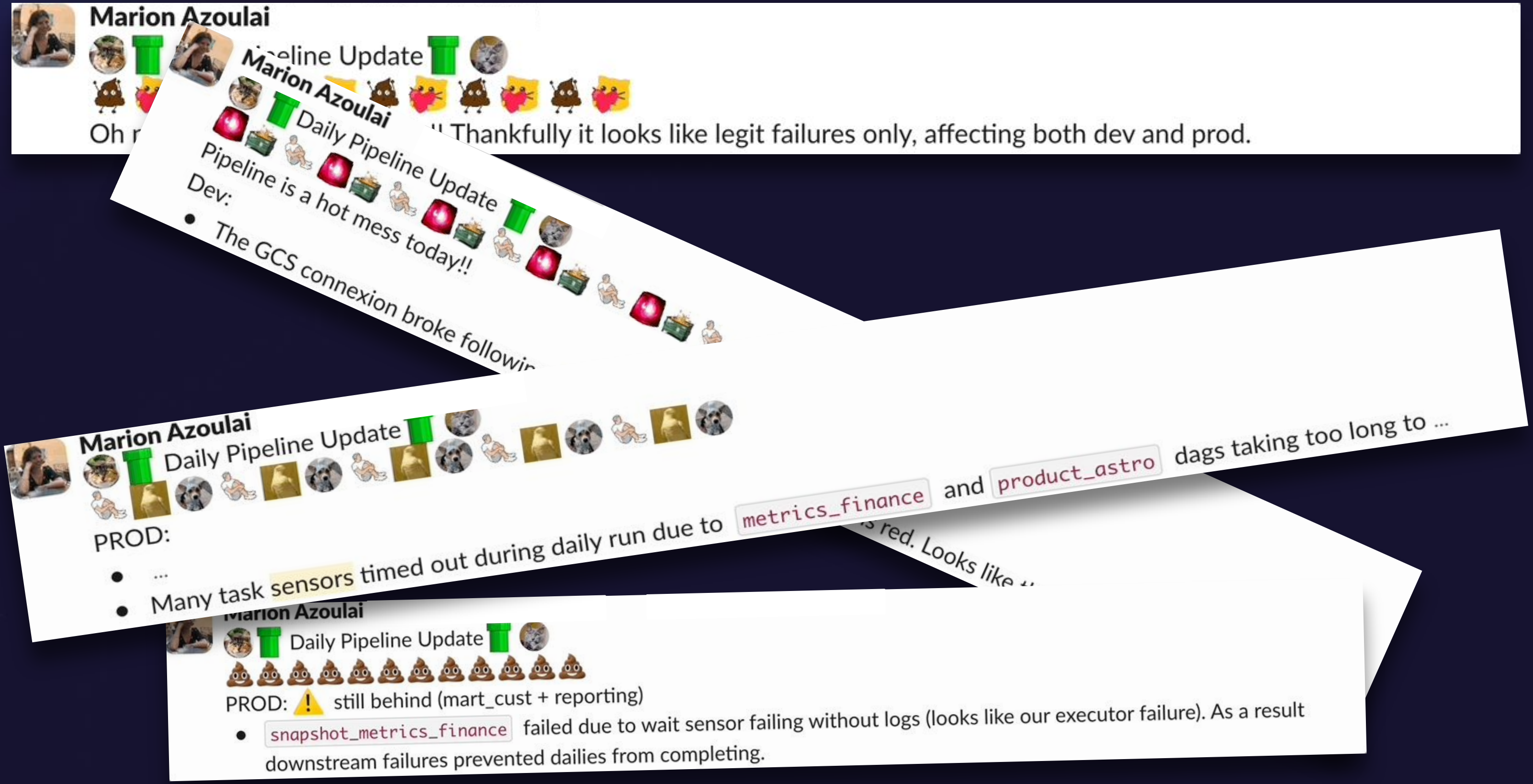
Daily Pipeline Update

💩💩💩💩💩💩💩💩💩💩💩💩

PROD: ⚠️ still behind (mart_cust + reporting)

- `snapshot_metrics_finance` failed due to wait sensor failing without logs (looks like our executor failure). As a result downstream failures prevented dailies from completing.

Until it wasn't...



Marion Azoulai
…eline Update

Oh r… !! Thankfully it looks like legit failures only, affecting both dev and prod.

Marion Azoulai
Daily Pipeline Update

Pipeline is a hot mess today!!

Dev:

- The GCS connexion broke followin…

…s red. Looks like …

Marion Azoulai
Daily Pipeline Update

PROD:

- ...
- Many task sensors timed out during daily run due to `metrics_finance` and `product_astro` dags taking too long to ...

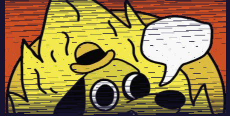
Marion Azoulai
Daily Pipeline Update

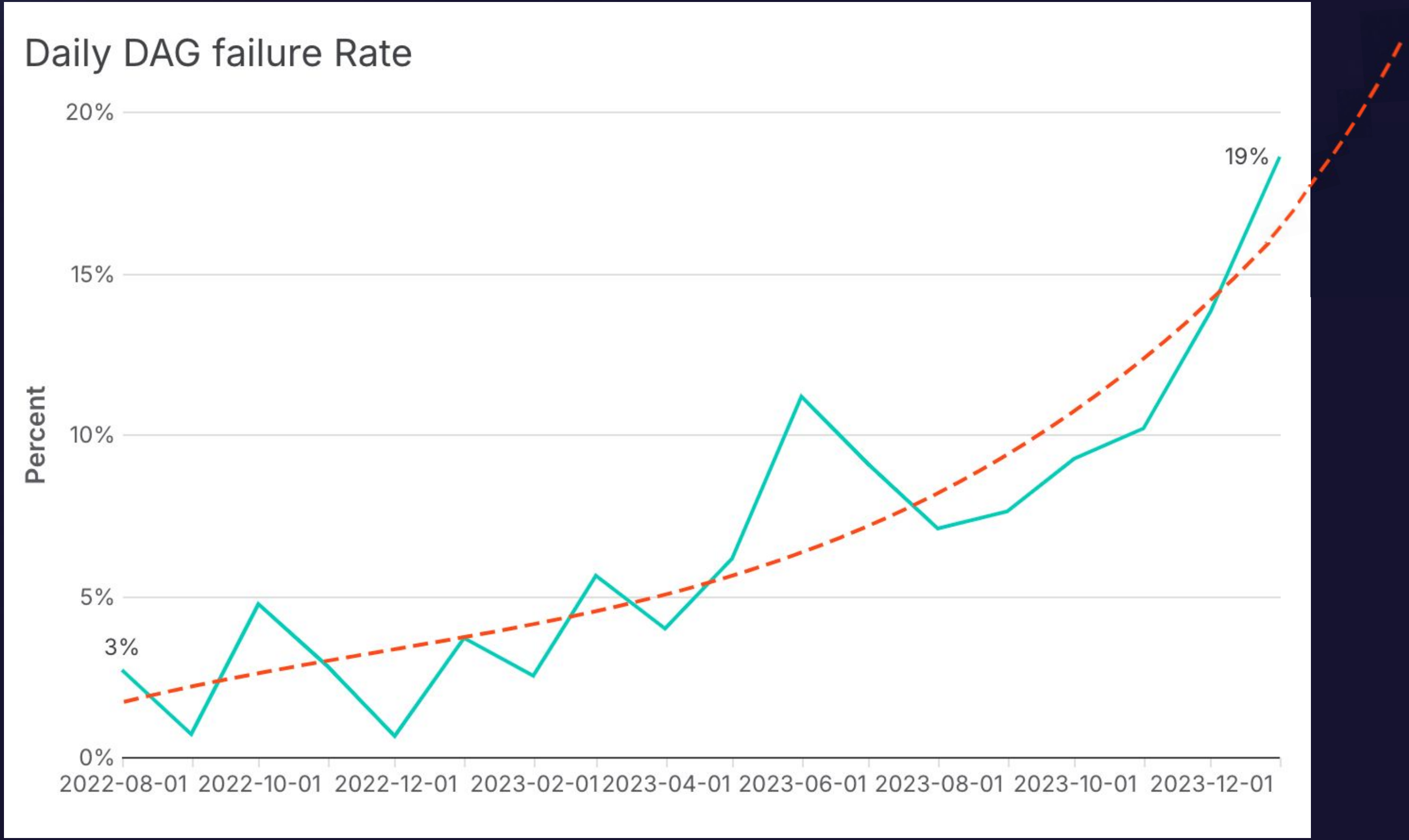
💩💩💩💩💩💩💩💩💩💩💩💩

PROD: ⚠️ still behind (mart_cust + reporting)

- `snapshot_metrics_finance` failed due to wait sensor failing without logs (looks like our executor failure). As a result downstream failures prevented dailies from completing.

Until it wasn't...





Pain Points

The challenges of a sensor-dependent setup

Over-reliance on Task and DAG Sensors

Pain Points

The challenges of a sensor-dependent setup

Over-reliance on Task and DAG Sensors

Failure Cascades

Pain Points

The challenges of a sensor-dependent setup

Over-reliance on Task and DAG Sensors

Failure Cascades

Complex Error Diagnosis

Pain Points

The challenges of a sensor-dependent setup

- Over-reliance on Task and DAG Sensors
- Failure Cascades
- Complex Error Diagnosis
- Recovery Delays

Pain Points

The challenges of a sensor-dependent setup

- Over-reliance on Task and DAG Sensors
- Failure Cascades
- Complex Error Diagnosis
- Recovery Delays
- Hitting System Limits

Pain Points

The challenges of a sensor-dependent setup

- Over-reliance on Task and DAG Sensors
- Failure Cascades
- Complex Error Diagnosis
- Recovery Delays
- Hitting System Limits
- Excessive Compute

Pain Points

The challenges of a sensor-dependent setup

- Over-reliance on Task and DAG Sensors
- Failure Cascades
- Complex Error Diagnosis
- Recovery Delays
- Hitting System Limits
- Excessive Compute
- Scaling Challenges

Pain Points

The challenges of a sensor-dependent setup

- Over-reliance on Task and DAG Sensors
- Failure Cascades
- Complex Error Diagnosis
- Recovery Delays
- Hitting System Limits
- Excessive Compute
- Scaling Challenges
- Development Challenges

Brainstorming the Solution

Focused on scalability and reliability

How do we solve our sensor problem?

Brainstorming the Solution

Focused on scalability and reliability

How do we solve our sensor problem?

Datasets

Brainstorming the Solution

Focused on scalability and reliability

How do we solve our sensor problem?

Datasets

How do we solve the visibility issue of datasets?

Brainstorming the Solution

Focused on scalability and reliability

How do we solve our sensor problem?

Datasets

How do we solve the visibility issue of datasets?

A Control DAG

Brainstorming the Solution

Focused on scalability and reliability

How do we solve our sensor problem?

Datasets

How do we solve the visibility issue of datasets?

A Control DAG

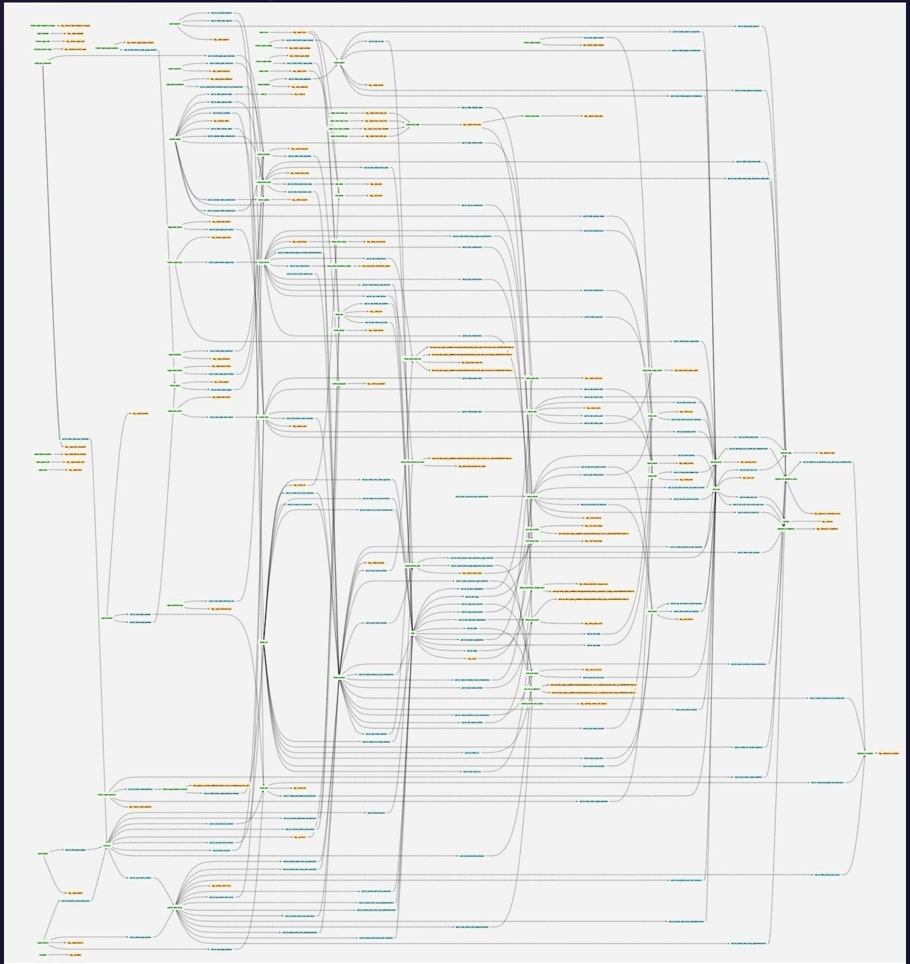
How do we build a Control DAG that's scalable?

Brainstorming the Solution

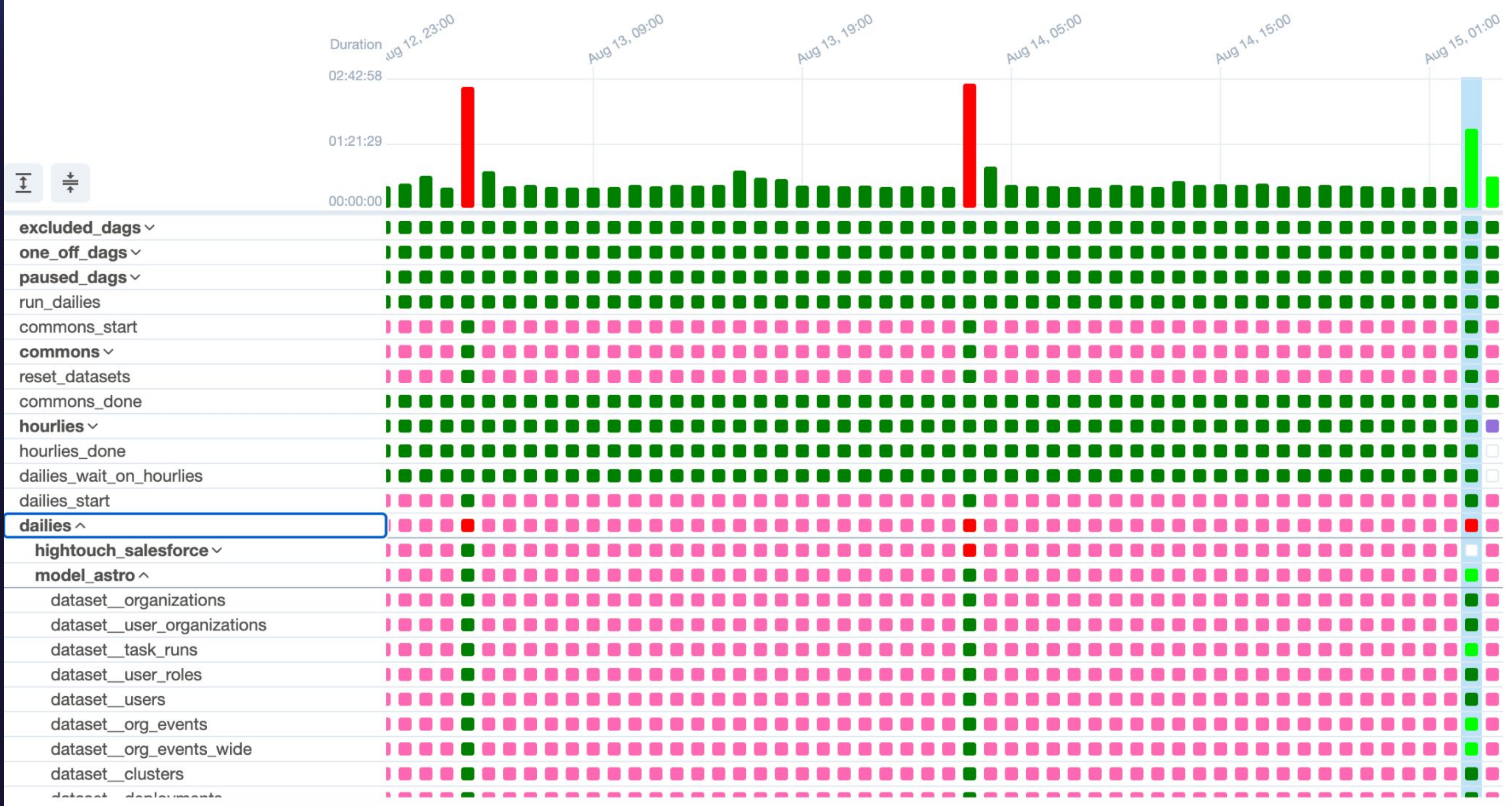
Focused on scalability and reliability

How do we solve our sensor problem?	How do we solve the visibility issue of datasets?	How do we build a Control DAG that's scalable?
Datasets	A Control DAG	Airflow

Our DAG Dependency Graph



But… make it functional



Pain Points

The challenges of a sensor-dependent setup

- Over-reliance on Task and DAG Sensors
- Failure Cascades
- Complex Error Diagnosis
- Recovery Delays
- Hitting System Limits
- Excessive Compute
- Scaling Challenges
- Development Challenges

Pain Points

The challenges of a sensor-dependent setup

- ✓ Over-reliance on Task and DAG Sensors
- ✓ Failure Cascades
- Complex Error Diagnosis
- Recovery Delays
- ✓ Hitting System Limits
- ✓ Excessive Compute
- Scaling Challenges
- Development Challenges

Our Re-Architecture

Focused on scalability and reliability

Dataset Scheduling

=

Increased Reliability

Pain Points

The challenges of a sensor-dependent setup

✓ Over-reliance on Task and DAG Sensors	✓ Failure Cascades	✓ Complex Error Diagnosis	Recovery Delays
✓ Hitting System Limits	✓ Excessive Compute	✓ Scaling Challenges	Development Challenges

Our Re-Architecture

Focused on scalability and reliability

Dataset Scheduling

=

Increased Reliability

End to End Visibility

=

Increased Confidence

Pain Points

The challenges of a sensor-dependent setup

- Over-reliance on Task and DAG Sensors
- Failure Cascades
- Complex Error Diagnosis
- Recovery Delays
- Hitting System Limits
- Excessive Compute
- Scaling Challenges
- Development Challenges

Our Re-Architecture

Focused on scalability and reliability

Dataset Scheduling

=

Increased Reliability

End to End Visibility

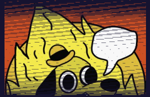
=

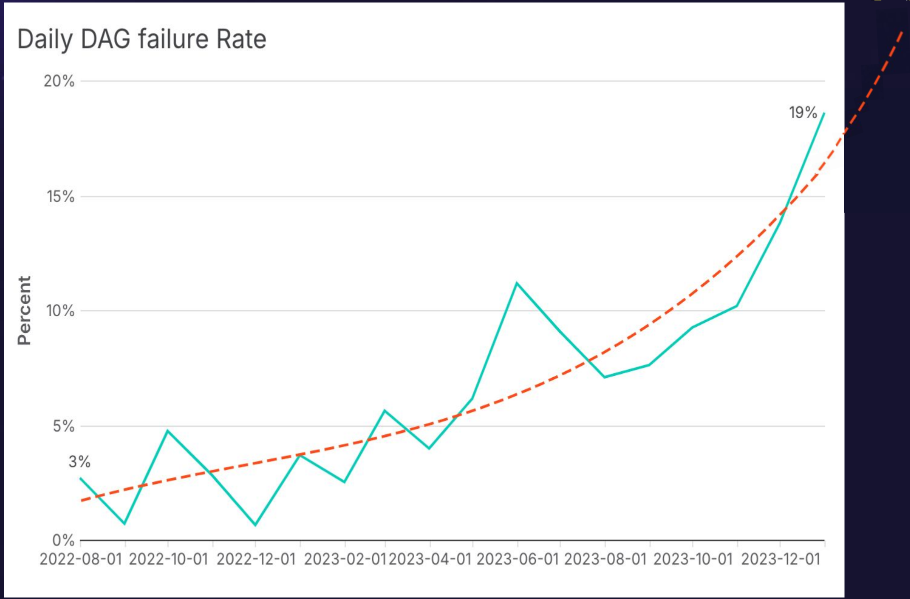
Increased Confidence

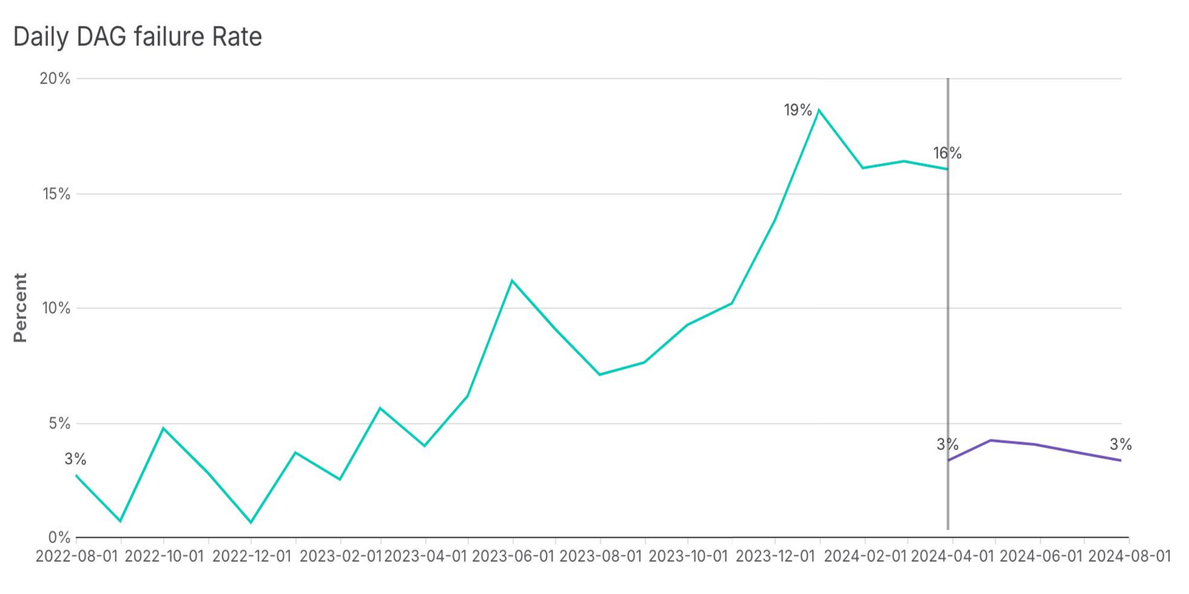
Micro-Pipelines

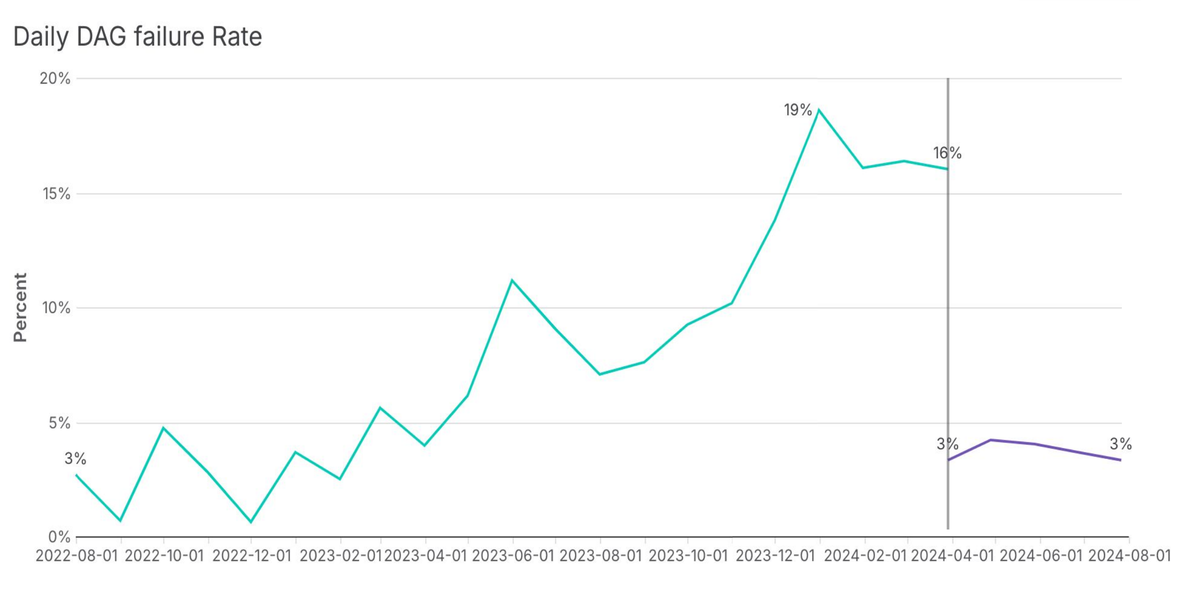
=

Failure Minimization




Daily DAG failure Rate
Percent
20%
15%
10%
5%
0%
19%
3%
2022-08-01
2022-10-01
2022-12-01
2023-02-01
2023-04-01
2023-06-01
2023-08-01
2023-10-01
2023-12-01

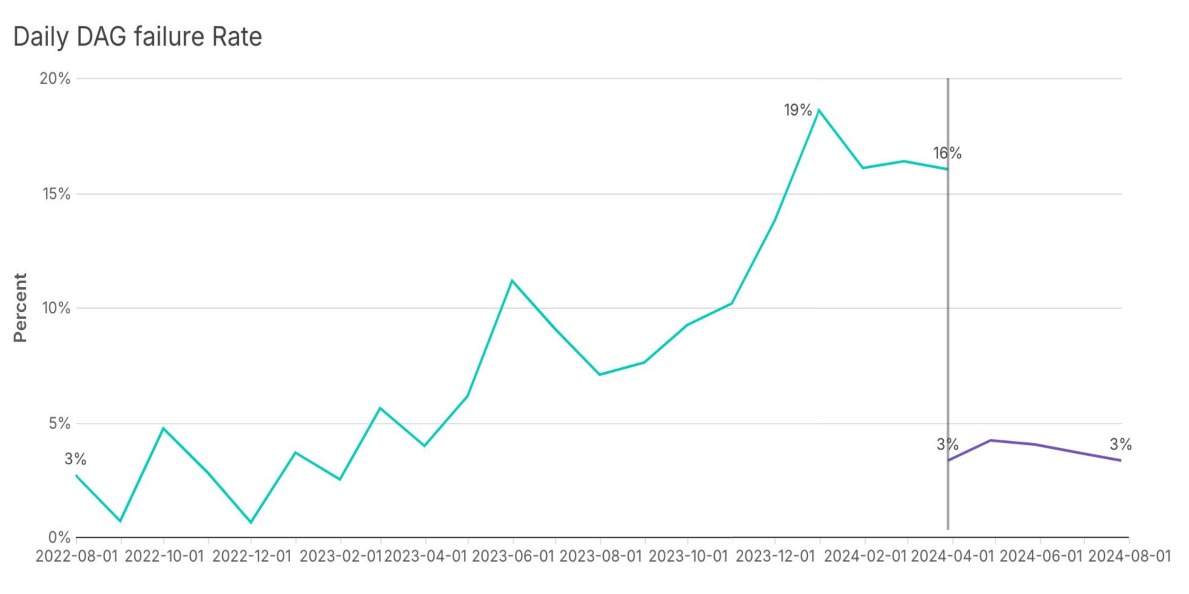

Daily DAG failure Rate
Percent
20%
15%
10%
5%
0%
3%
19%
16%
3%
3%
2022-08-01
2022-10-01
2022-12-01
2023-02-01
2023-04-01
2023-06-01
2023-08-01
2023-10-01
2023-12-01
2024-02-01
2024-04-01
2024-06-01
2024-08-01



↓ 37.9%

Avg. Hourly Run Duration

8.4 min improvement



↓ 37.9%

Avg. Hourly Run Duration

8.4 min improvement

↓ 33.0%

Avg. Daily Run Duration

1.7 hr improvement

Pop Quiz

WHO WANTS TO BE A ASTRONOMER

What is the most commonly used operator?

A: BashOperator

B: SnowflakeOperator

C: PythonOperator

D: LazyAutomationOperator

WHO WANTS TO BE A ASTRONOMER

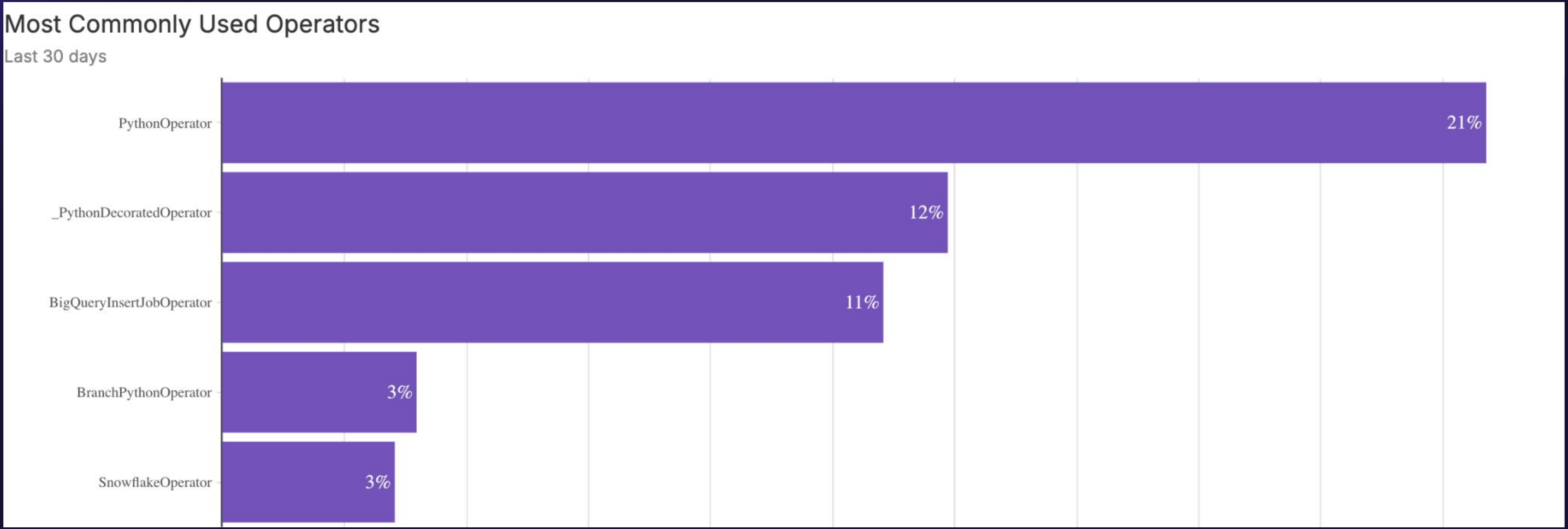
What is the most commonly used operator?

A: BashOperator

B: SnowflakeOperator

C: PythonOperator

D: LazyAutomationOperator



Most Commonly Used Operators

Last 30 days

WHO WANTS TO BE A

ASTRONOMER

What is the most failure prone operator?

A: BigQueryCheckOperator

B: EmptyOperator

C: SSHOperator

D: PythonSensor

WHO WANTS TO BE ASTRONOMER

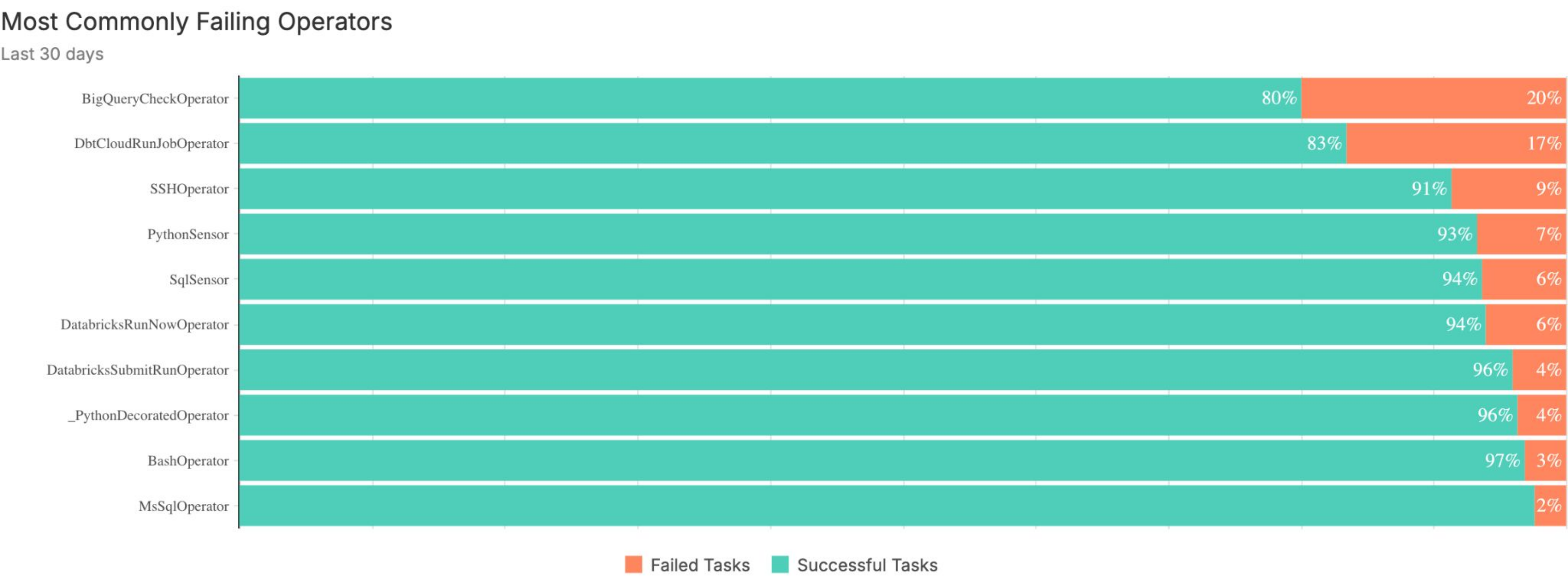
What is the most failure prone operator?

A: BigQueryCheckOperator

B: EmptyOperator

C: SSHOperator

D: PythonSensor



Most Commonly Failing Operators

Last 30 days

WHO WANTS TO BE A

ASTRONOMER

What are the longest-running operator types?

A: Branch Operators

B: DBT Operators

C: Batch Operators

D: Databricks Operators

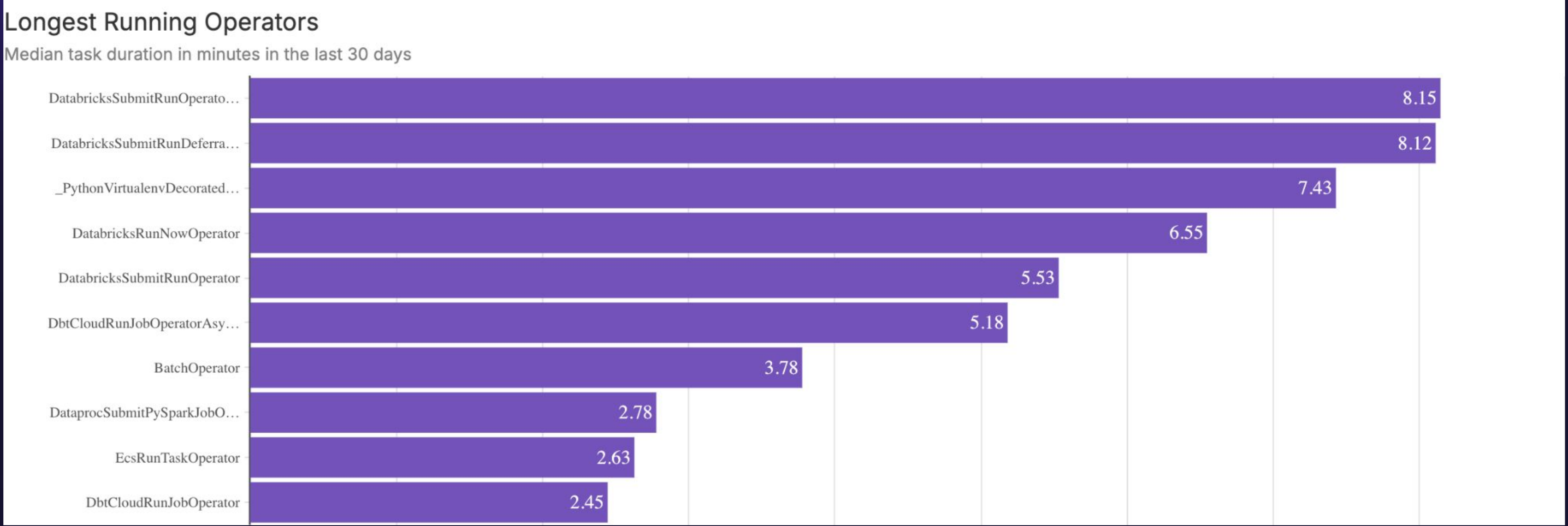
What are the longest-running operator types?

A: Branch Operators

B: DBT Operators

C: Batch Operators

[✓ D: Databricks Operators]



Longest Running Operators

Median task duration in minutes in the last 30 days

How many tasks in a DAG?

How many tasks does a DAG typically have?

A: 1

B: 5

C: 10

D: 20

WHO WANTS TO BE A ASTRONOMER

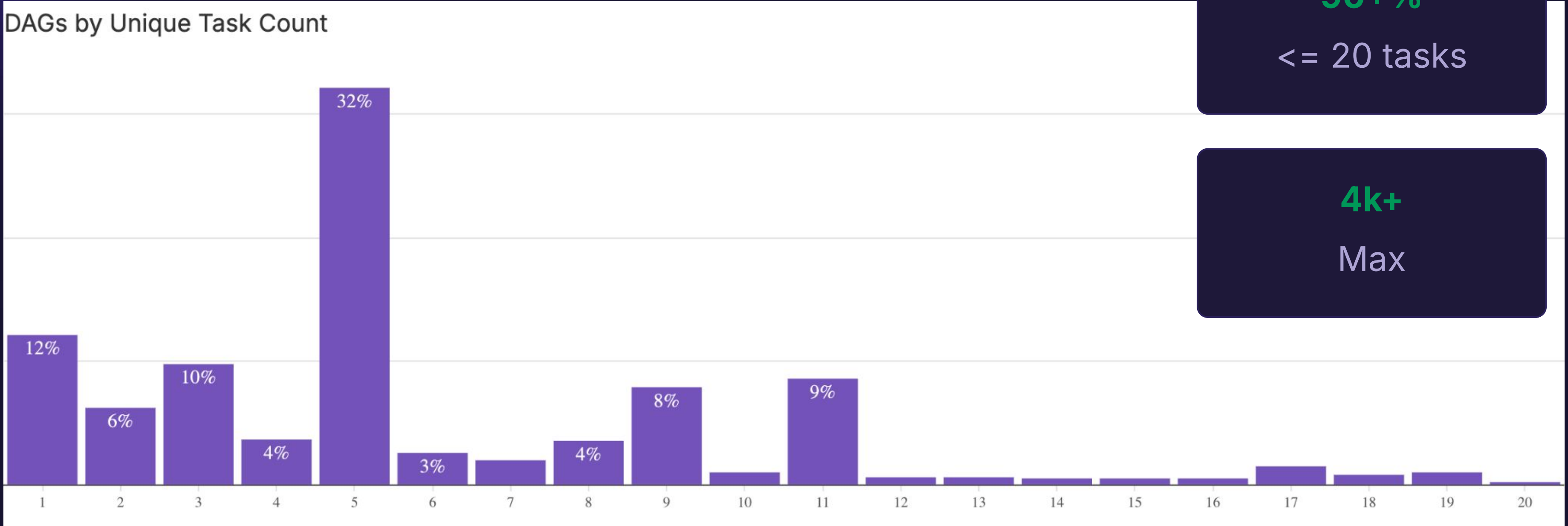
How many tasks does a DAG typically have?

A: 1

[✓ B: 5]

C: 10

D: 20



DAGs by Unique Task Count

90+%

<= 20 tasks

4k+

Max

Which hour (UTC) are daily DAGs most often scheduled?

A: 0

B: 8

C: 12

D: 20

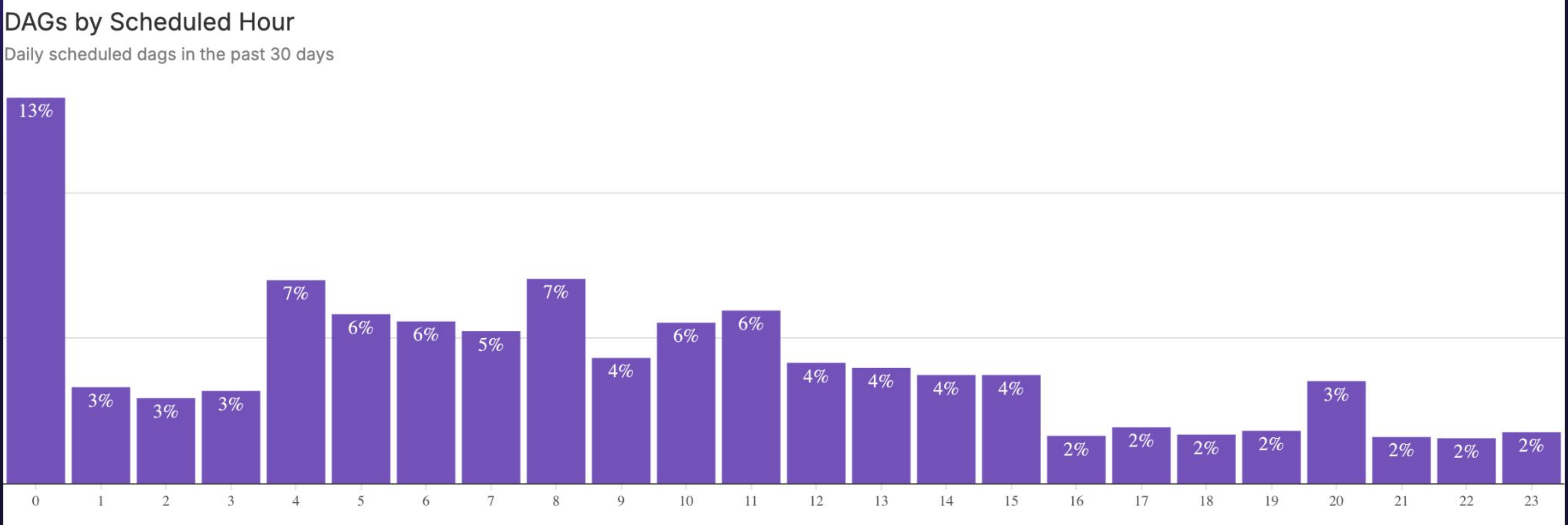
Which hour (UTC) are daily DAGs most often scheduled?

A: 0

B:8

C: 12

D: 20



DAGs by Scheduled Hour

Daily scheduled dags in the past 30 days

Which minute are DAGs most often scheduled?

A: 0

B: 15

C: 30

D: 45

WHO WANTS TO BE A

ASTRONOMER

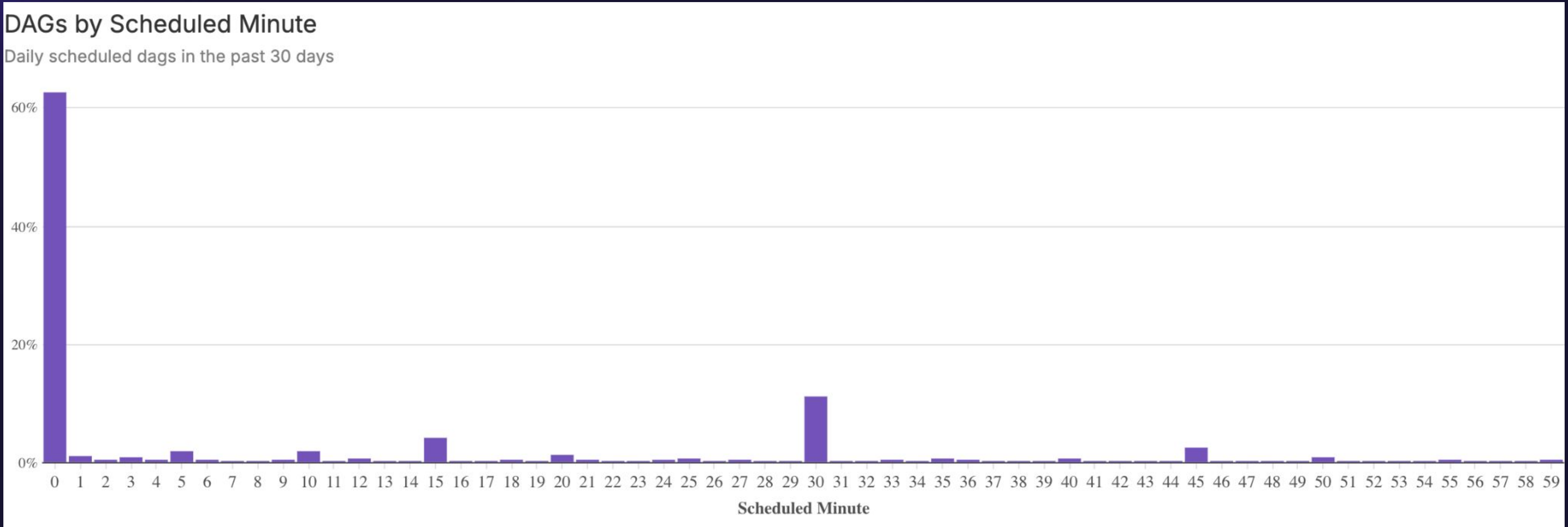
Which minute are DAGs most often scheduled?

A: 0

B: 15

C: 30

D: 45



DAGs by Scheduled Minute

Daily scheduled dags in the past 30 days

Thank you!
Any questions?

ASTRONOMER