

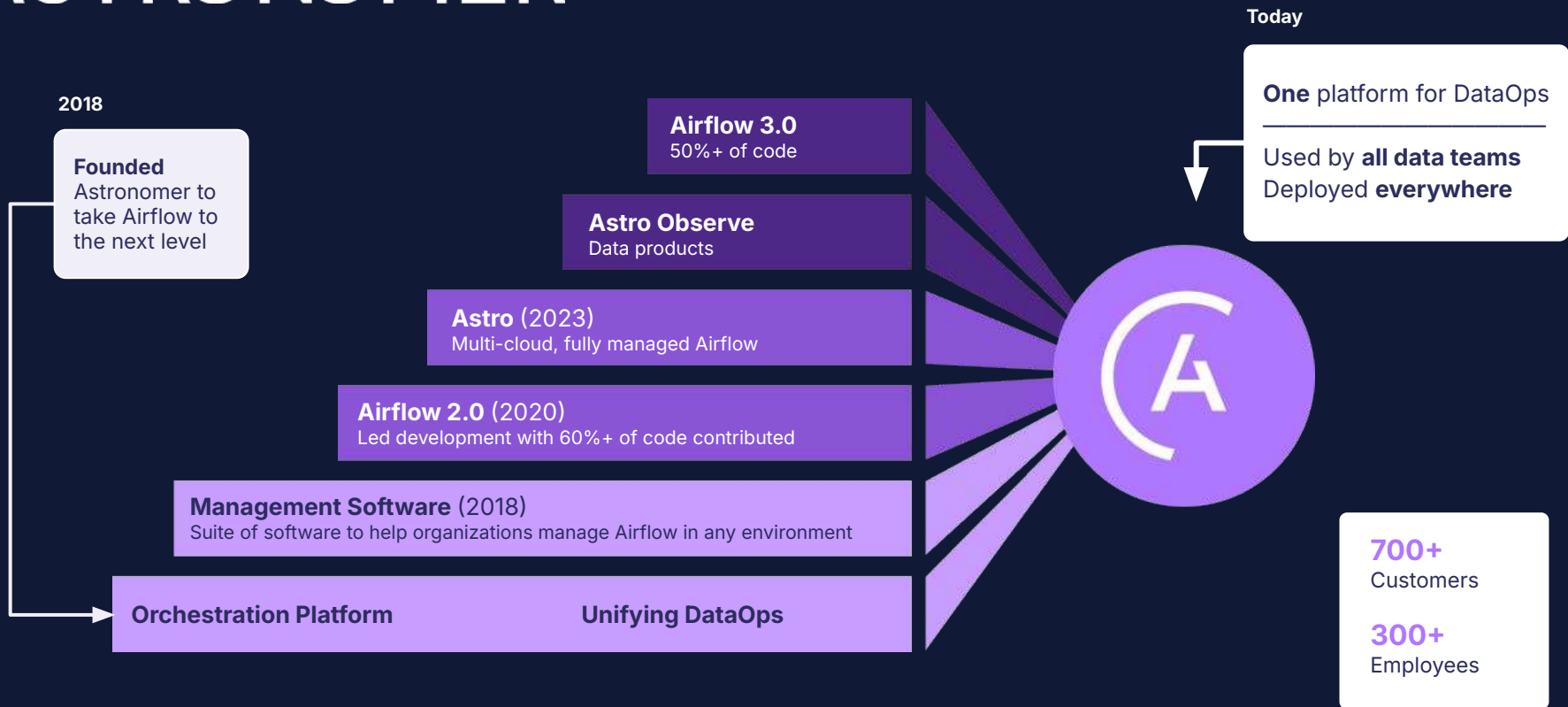


# Orchestrating Data Quality - Quality Data Brought To You By Airflow

Maggie Stark

# 3.0

# ASTRONOMER





- ⚗ Centralized Data Team
- 🔗 Build critical operational and analytical pipelines with Airflow
- 👥 Product Influencers
  - 📊 Providing a Data-First Perspective

# Configuration as Code

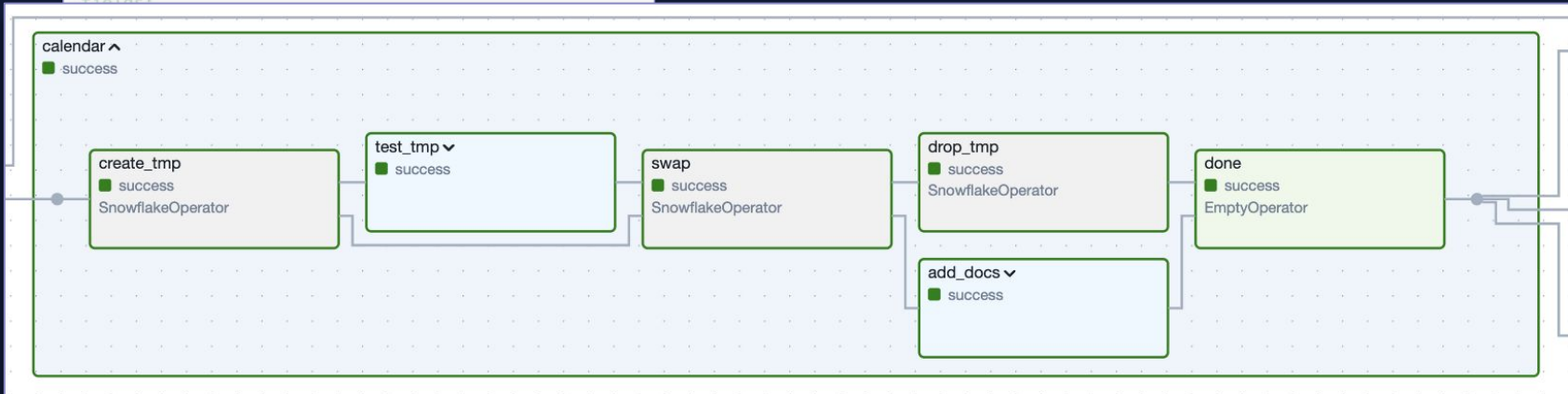
```
/*
operator: include.task_groups.transform.CreateTable
description: >-
  Creates standard calendar, where each row
  represents a date with common date operations.
fields:
  date: Date (iso format yyyy-mm-dd).
  is_weekend: Date is Sat. or Sun.
  is_weekday: Date is weekday (MTWTF).
schema: !switch_value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

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

# Configuration as Code

```
/*  
operator: include.task_groups.transform.CreateTable  
description: >--  
  Creates standard calendar, where each row  
  represents a date with common date operations.  
fields:
```



```
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

# Configuration as Code

```
/*  
operator: include.task_groups.transform.CreateTable  
description: >--  
  Creates standard calendar, where each row  
  represents a date with common date operations.  
fields:
```

**1 file = n tasks!**

Create a table; get  
documentation and  
testing come free.

calendar ^

■ success

create\_tmp

■ success

SnowflakeOperator

test\_tmp v

■ success

swap

■ success

SnowflakeOperator

drop\_tmp

■ success

SnowflakeOperator

done

■ success

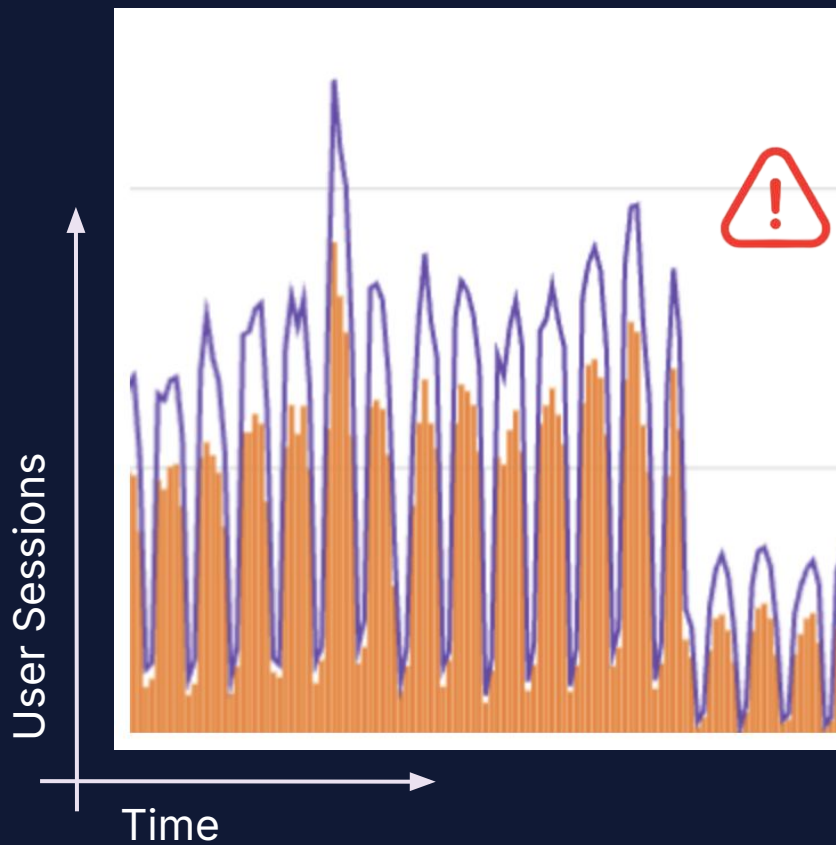
EmptyOperator

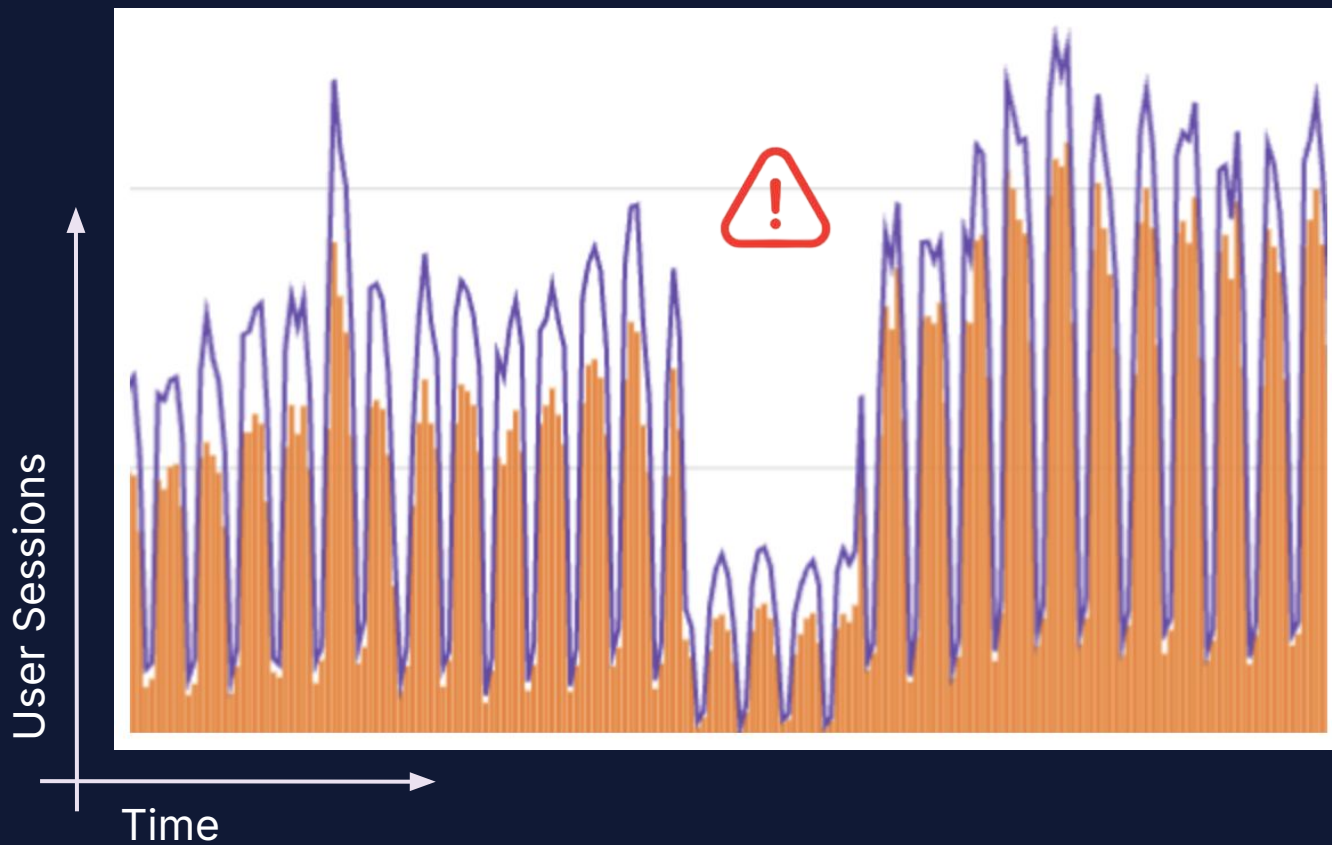
add\_docs v

■ success

```
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





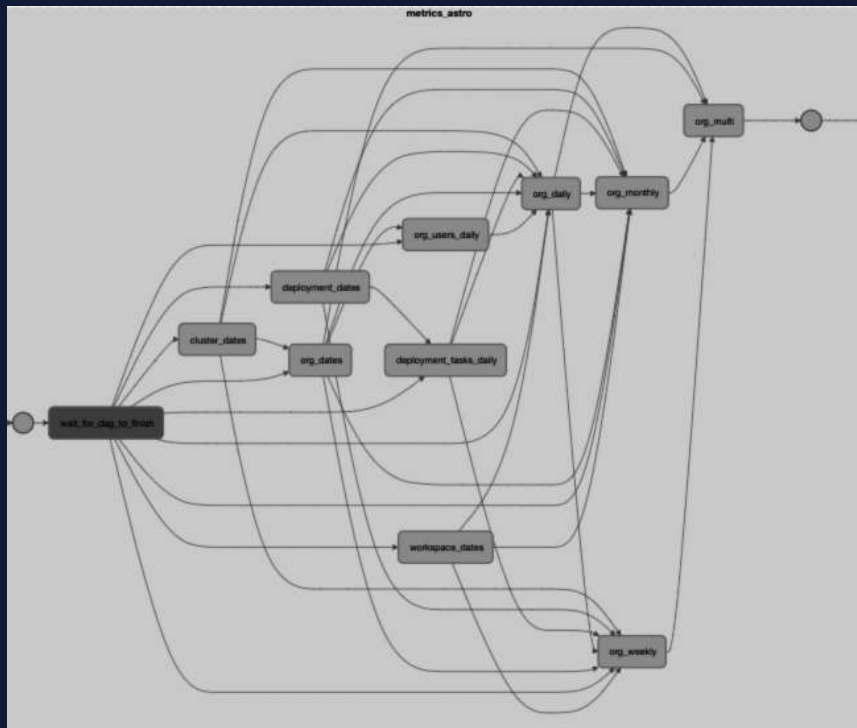


## ❏ Standalone Data-Quality DAG

📖 Hard to Maintain

❏ Duplicative

📉 Poor Visibility and Adoption



# Configuration as Code

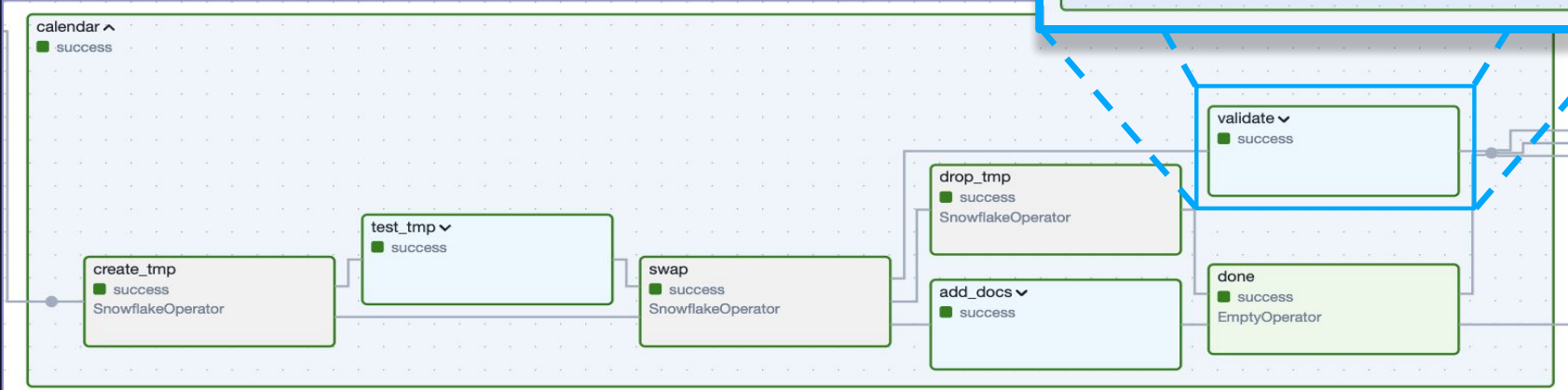
```
/*
operator: include.task_groups.transform.CreateTable
description: >-
  Creates standard calendar, where each row
  represents a date with common date operations.
fields:
  date: Date (iso format yyyy-mm-dd).
  is_weekend: Date is Sat. or Sun.
  is_weekday: Date is weekday (MTWTF).
schema: !switch_value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

# Configuration as Code

```
/*  
operator: include.task_groups.transform.CreateTable  
description: >--  
  Creates standard calendar, where each row  
  represents a date with common date operations.  
fields:
```



```
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

astro\_org\_dates ^

■ failed

test\_tmp ^

■ success

test\_table

■ success

SQLTableCheckOperator

test\_cols

■ success

SQLColumnCheckOperator

create\_tmp

■ success

SnowflakeOperator

swap

■ success

SnowflakeOperator

drop\_tmp

■ success

SnowflakeOperator

add\_docs v

■ success

validate ^

■ failed

test\_table

■ failed

SQLTableCheckOperator

test\_cols

■ success

SQLColumnCheckOperator

sql\_check\_done

■ success

EmptyOperator

done

■ success

EmptyOperator

## **Data Contracts**

Expected attributes exist, NULLs, field values, conditionals based on understanding of data source.

## **Data Integrity**

PKs, FKs, NULLs, unique field composites, field values.

## **Data Accuracy**

Metrics align between related grains of aggregations, business-process conditionals which have external impact.

## **Non-Blocking Validations**

Business-process conditionals with no external impact, field values, parity between dev./prod.

# Where are data quality checks performed?

	Ingest	Model	Maps	Metrics	Marts	Reporting
Data Contracts						
Data Integrity						
Data Accuracy						
Non-Blocking Validations						

astro\_org\_dates ^

■ failed

test\_tmp ^

■ success

test\_table

■ success

SQLTableCheckOperator

test\_cols

■ success

SQLColumnCheckOperator

create\_tmp

■ success

SnowflakeOperator

swap

■ success

SnowflakeOperator

drop\_tmp

■ success

SnowflakeOperator

add\_docs v

■ success

validate ^

■ failed

test\_table

■ failed

SQLTableCheckOperator

test\_cols

■ success

SQLColumnCheckOperator

sql\_check\_done

■ success

EmptyOperator

done

■ success

EmptyOperator

```

/*
operator: include.task_group.transform.CreateTable
description: >-
  Creates standard calendar table
  represents a date with fields:
  fields:
    date: Date (ISO format)
    is_weekend: Date is weekend
    is_weekday: Date is weekday
  schema: !switch_value
    sandbox: env__sandbox
    default: commons
  primary_key:
    - date
  tests:
    check_null:
      - is_weekend

```

## validations:

### check\_condition:

#### external:

- date = eom\_date

#### internal:

- is\_weekend OR is\_weekday
- COUNT(\*) = 365

```

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




# Logging

	<u>A</u> DAG_ID	<u>A</u> TASK_ID	 LOGICAL_TS
1	model_crm	sf_accounts.validate.test_table	2025-09-18 06:29

[ ] RESULTS

```
[
  {
    "failed_row_count": "50",
    "is_success": "False",
    "owner": "external",
    "sql": "zd_org_id IS NOT NULL OR
is_deleted OR NOT is_current_cust",
    "test_name": "check_condition_1"
  },
  {
    "failed_row_count": "10",
    "is_success": "False",
    "owner": "external",
    "sql": "NOT is_current_cust OR
support_level IS NOT NULL OR product IS NULL",
    "test_name": "check_condition_2"
  }
]
```

	<u>A</u> DAG_ID	<u>A</u> TASK_ID	 LOGICAL_TS
1	test_mart_cust	test_current_astro_custs	2025-09-17 17:






A RESULTS

```
{ "owner": "internal", "test_severity": "validation", "failed_row_count": 2, "is_success":
false, "failed_rows":
"D566485B7B22414343545F4944223A223031347830010124627653357A414145222C091F4C4E414D45223A2
6E647265645820496E632E111CC454414753223A2250726F6A6563746564207573616765206F76657220313235
F6E73756D7074696F6E2067726F777468091C4035302520284C61737420372064617973298A2A000433300D2B5
772075736572732061646465643B4163746976650D1324696
57211174C55524C223A2268747470733A2F2F617374726F68
2E636F6D2F1514282F722F4163636F756E742F463601102F7
33044223A395A1A00185F4C4147223A32361E000450430510
```

[ ] HUMAN\_READABLE

```
[
  {
    "ACCT_ID": "0014x0000bvS5zAAE",
  }
]
```

# Recapping

-  Use Airflow-native operators to perform DQ checks.
-  Write-audit-publish approach.
-  Checks are mainly configured in task definition
  -  New/different standalone DQ pipelines
-  Surface some validation results in dashboards to take action.

# The 2025 Apache Airflow<sup>®</sup> Survey is here!

Fill it out to for a free Airflow 3  
Fundamentals or DAG Authoring in  
Airflow 3 certification code



# Questions?