AIRFLOW SUMMIT

3.0

Allegro's Airflow Journey: From On-Prem to Cloud Orchestration at Scale

Marek Gawiński & Piotr Dziuba

TL;DR

How to get to over **250** active Airflow environments and **survive**

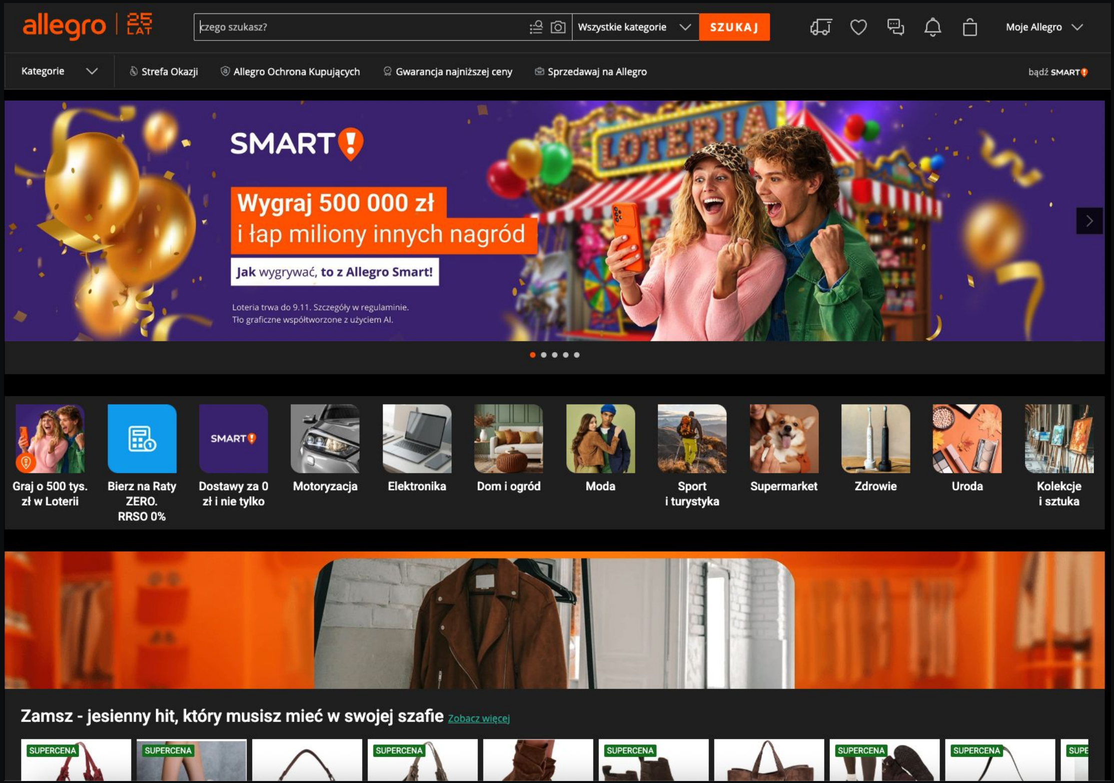
Operate data platform for almost **1000 users** with over **16,000 DAGs**

How to save **$1,147,233** on airflow orchestration costs and **FAR MORE** in human time

AIRFLOW SUMMIT

About Allegro

- Allegro.pl + CZ, SK, HU
- 25 years on the market
- ~21,1 million active buyers
- ~20 millions users per month
- > 160k merchants
- ~7000 people across CEE



SMART!

Wygraj 500 000 zł i łap miliony innych nagród

Jak wygrywać, to z Allegro Smart!

Loteria trwa do 9.11. Szczegóły w regulaminie.
Tło graficzne współtworzone z użyciem AI.

Zamsz - jesienny hit, który musisz mieć w swojej szafie Zobacz więcej

From on-prem to cloud scale

3.0

History points

2012-07 - Hadoop Cluster in our DC (CHD3)

2015-12 - Oozie + Tez support

2017-08 - Airflow As A Service (AaaS) in Allegro (Airflow **1.8.1**)

2022-04 - Migrating Data Platform to the Cloud: AaaS -> Cloud Composer (Google)

2023-03:

- Hadoop Outro
- AaaS - At the peak 249 dedicated instances for Dev, Testing & Production envs (**1.9.0/1.10.12**)

2023-06 - End of Support for Airflow as a Service

Organization

- Teams organization
  - Over 100 teams
  - Over 900 internal users
  - Different competence levels
- Projects groups on GCP - **360**
  - 3 environments each (dev/test/prod)
- Composers - **175** instances (DEV - **48**, TEST - **43**, PROD - **84**)

Conway's law!

AIRFLOW SUMMIT

Current state

- Main use cases
  - Data processings
  - MLOps
  - Governance
  - Utils

- Infrastructure
  - Composer environments: 175
  - # DAGs: 15.6k
  - # Tasks: 166k
  - Over 300 types of operators

Orchestrated tasks

- 38k BQ processings
- 35k BQ Sensors
- 8k Spark processings
- 11k Snowflake processings
- 5k DBT processings
- 15.5k PythonOperators

.... and 50k other

mongoDB
salesforce
Jira
tableau

~~Problems~~
Solutions

3.0

AIRFLOW SUMMIT

Data Platform

- DAG authoring and deployment process
- Cloud resource management
  - Datasets & Tables
  - Also Cloud Composer environment
- Governance
  - Access management
  - GDPR
  - Ownership attribution
  - Auditability
- Documentation and support

Gen AI - Gemini

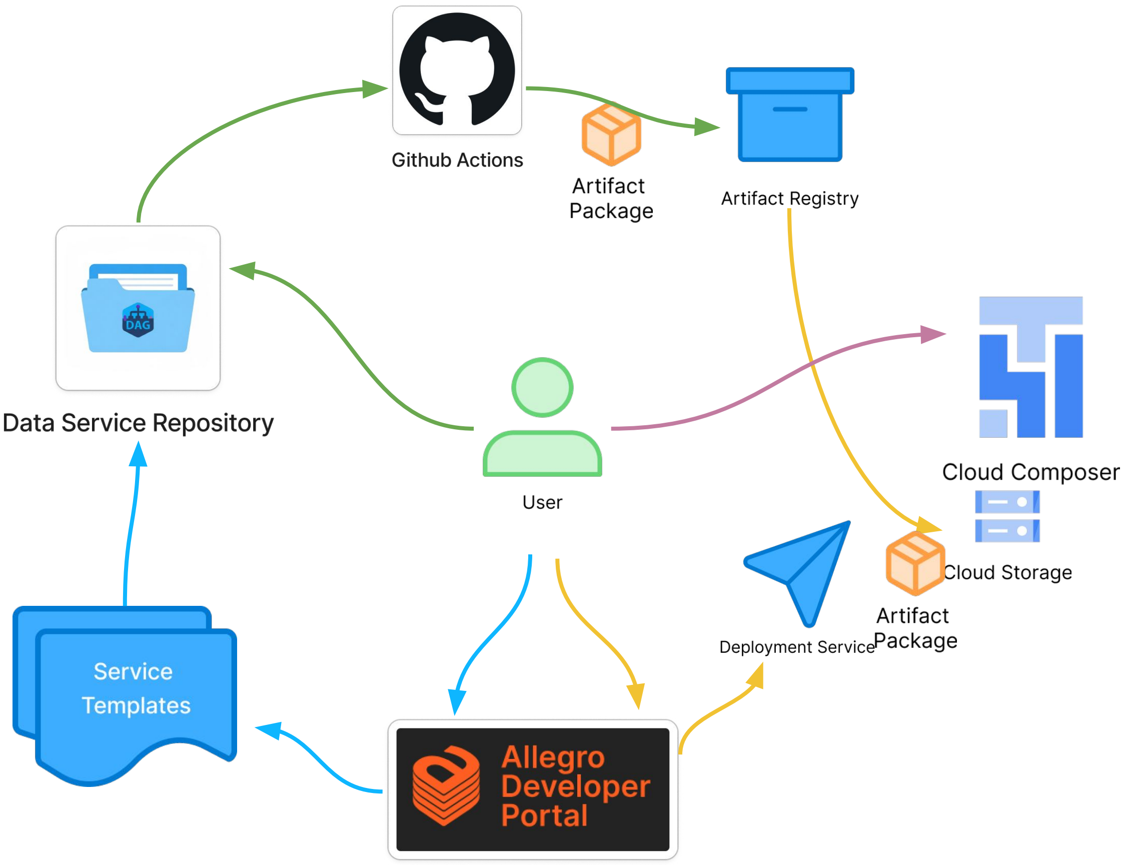
User

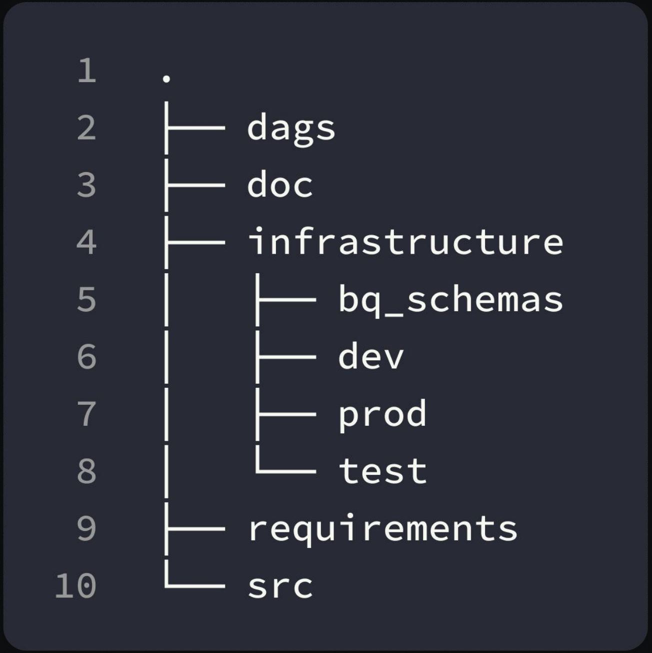
AIRFLOW SUMMIT

DAG authoring flow

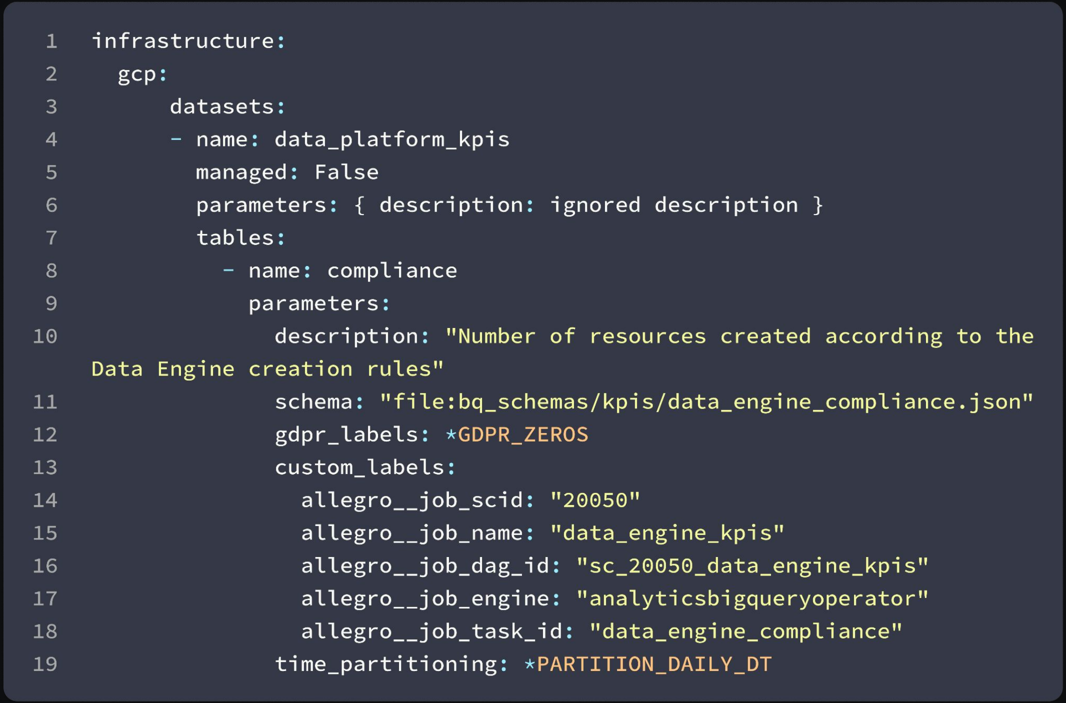
Flow elements:

- Service generation based on predefined template
- Processing defining and build automation
- Target environment deployment





```
.
├── dags
├── doc
├── infrastructure
│   ├── bq_schemas
│   ├── dev
│   ├── prod
│   └── test
├── requirements
└── src
```

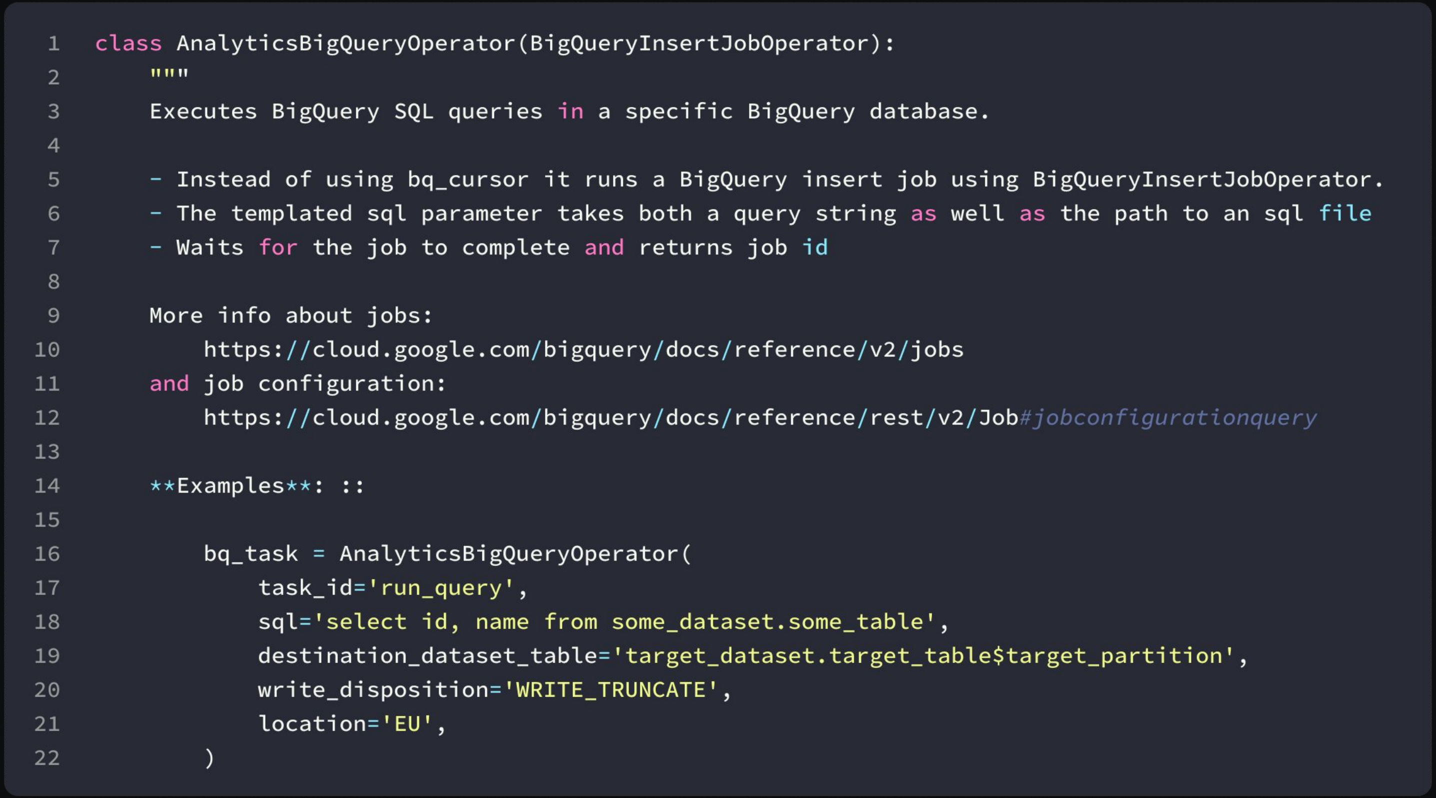


```yaml
infrastructure:
  gcp:
      datasets:
      - name: data_platform_kpis
        managed: False
        parameters: { description: ignored description }
        tables:
          - name: compliance
            parameters:
              description: "Number of resources created according to the Data Engine creation rules"
              schema: "file:bq_schemas/kpis/data_engine_compliance.json"
              gdpr_labels: *GDPR_ZEROS
              custom_labels:
                allegro__job_scid: "20050"
                allegro__job_name: "data_engine_kpis"
                allegro__job_dag_id: "sc_20050_data_engine_kpis"
                allegro__job_engine: "analyticsbigqueryoperator"
                allegro__job_task_id: "data_engine_compliance"
              time_partitioning: *PARTITION_DAILY_DT
```

AIRFLOW SUMMIT

Data Platform - operators

- Core provider operators are great but…
  - are of general purpose
- data-engine-composer-extras - allegro airflow extension library
  - Common utils
  - Curated fine-tuned operator set
    - Artifact-structure-oriented
    - Curated default values
    - Governance labels
  - Pre-installed on each environment via dedicated terraform module



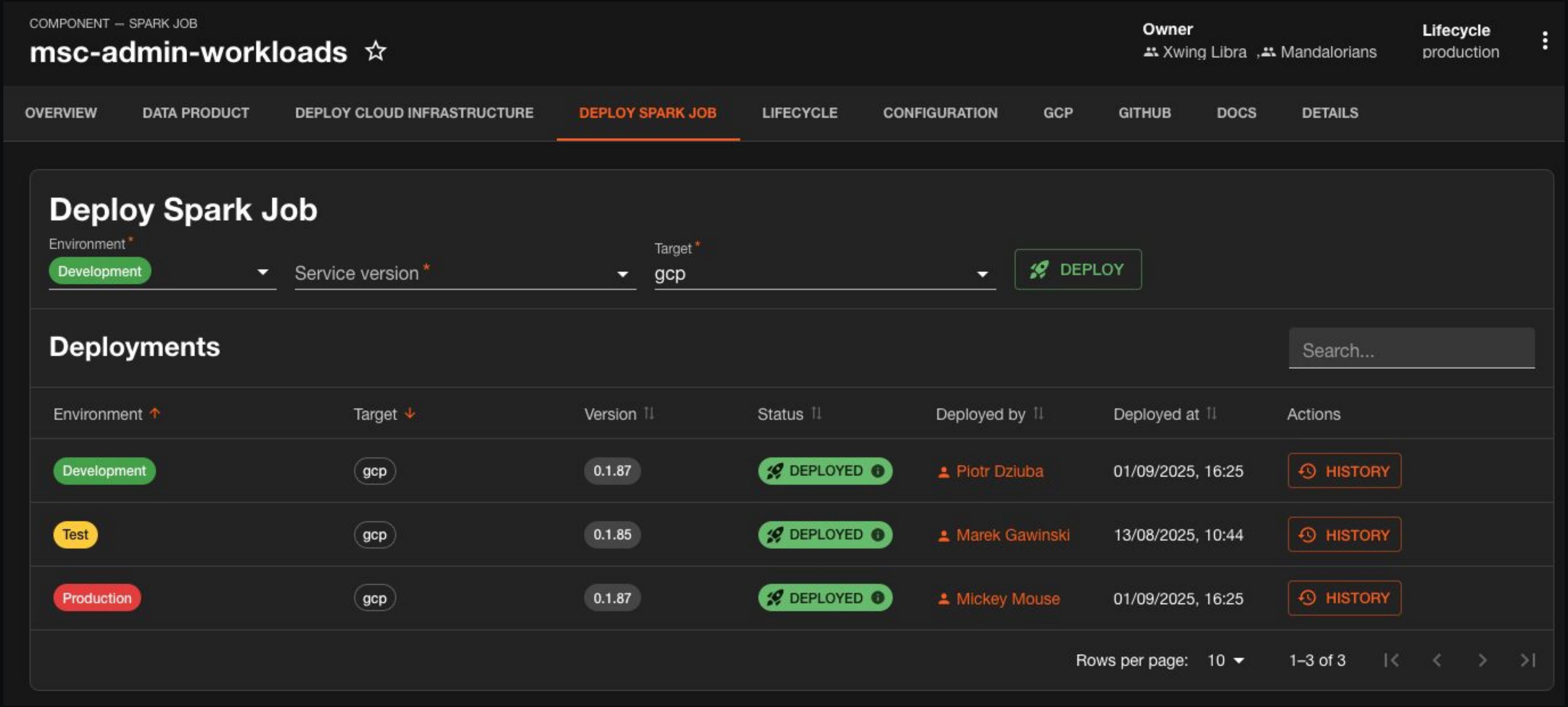
```python
class AnalyticsBigQueryOperator(BigQueryInsertJobOperator):
    """
    Executes BigQuery SQL queries in a specific BigQuery database.

    - Instead of using bq_cursor it runs a BigQuery insert job using BigQueryInsertJobOperator.
    - The templated sql parameter takes both a query string as well as the path to an sql file
    - Waits for the job to complete and returns job id

    More info about jobs:
        https://cloud.google.com/bigquery/docs/reference/v2/jobs
    and job configuration:
        https://cloud.google.com/bigquery/docs/reference/rest/v2/Job#jobconfigurationquery

    **Examples**: ::

        bq_task = AnalyticsBigQueryOperator(
            task_id='run_query',
            sql='select id, name from some_dataset.some_table',
            destination_dataset_table='target_dataset.target_table$target_partition',
            write_disposition='WRITE_TRUNCATE',
            location='EU',
        )
```



COMPONENT — SPARK JOB

msc-admin-workloads ☆

Owner
Xwing Libra , Mandalorians

Lifecycle
production

Deploy Spark Job

Environment *: Development

Service version *

Target *: gcp

Deployments

Environment ↑	Target ↓	Version	Status	Deployed by	Deployed at	Actions
Development	gcp	0.1.87	DEPLOYED	Piotr Dziuba	01/09/2025, 16:25	HISTORY
Test	gcp	0.1.85	DEPLOYED	Marek Gawinski	13/08/2025, 10:44	HISTORY
Production	gcp	0.1.87	DEPLOYED	Mickey Mouse	01/09/2025, 16:25	HISTORY

Rows per page: 10 1–3 of 3

AIRFLOW SUMMIT

Data Platform - is the story complete?

- Self-service data platform
  - It’s easy to get started with DAG authoring
  - It’s easy to get your own airflow environment
- Airflow as first go-to solution for any regular task handling
- Numerous teams with different level of expertise

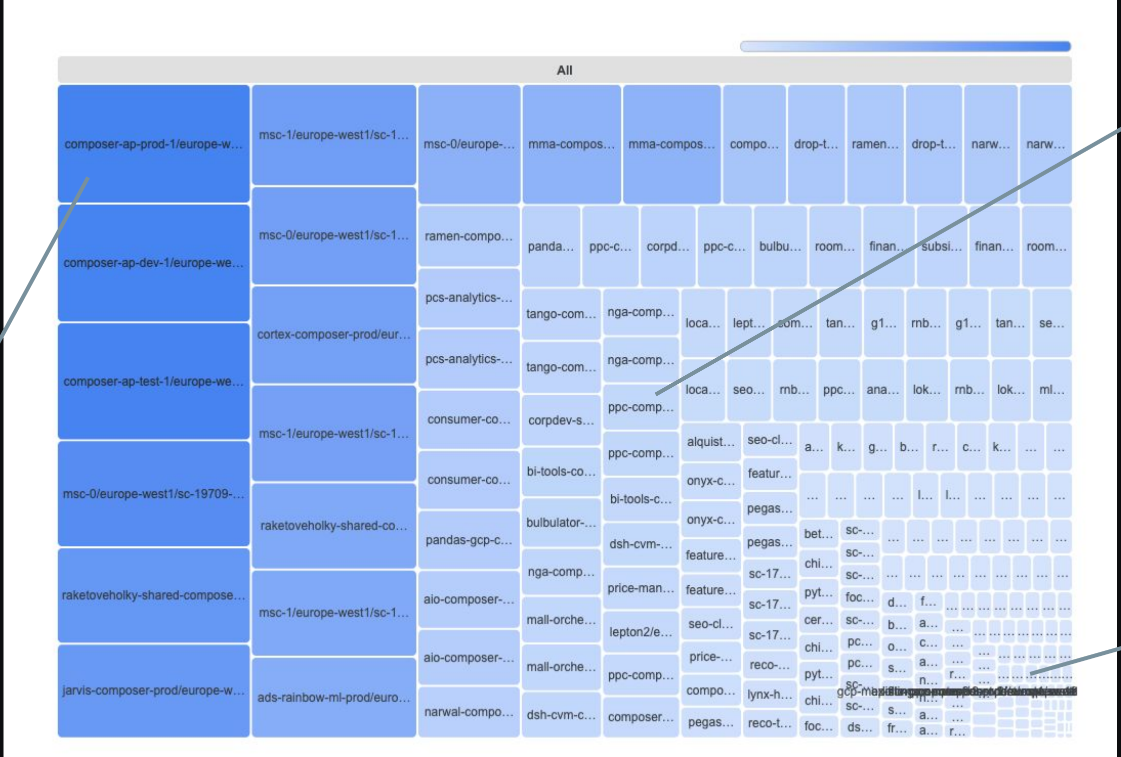
- Can there be any downsides?

DAG distribution

180 cloud composer instances

16k DAGs

100 teams



63 DAGs

7 DAGs

524 DAGs

Fragmentation consequences

Cost

$$$ * 180 = $$$$$

Maintenance time

4h*/ month x 180 = 720h / month = 4 FTE

* Estimation based on surveyed average of 4h maint time per month per team

Managed Shared Composer

3.0

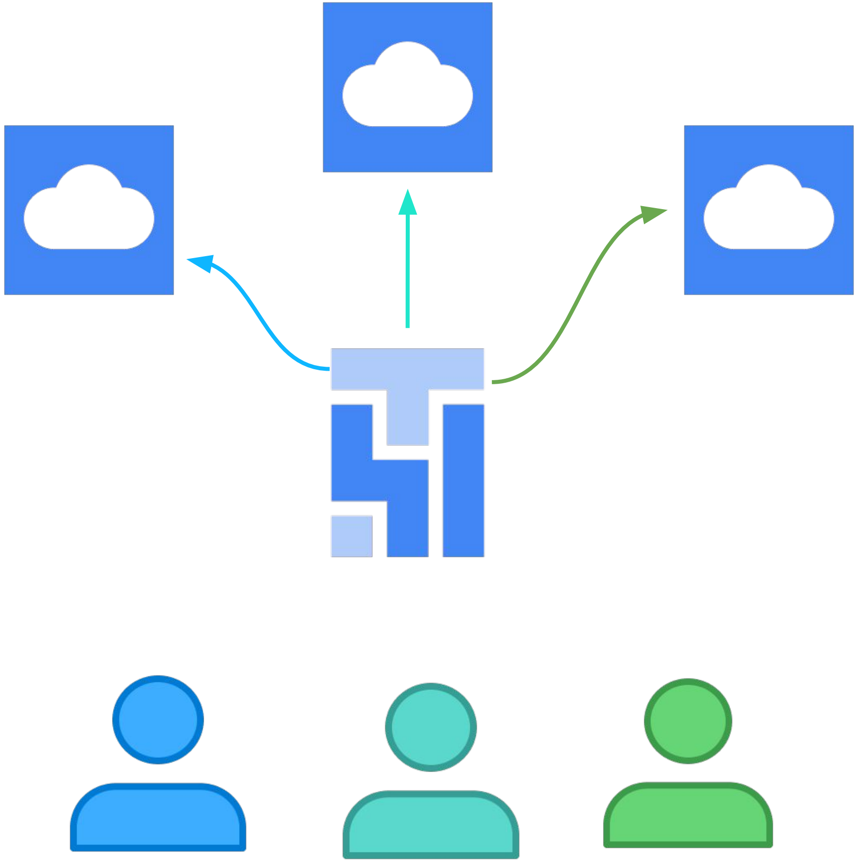
AIRFLOW SUMMIT

Shared Environment

Multiple teams

Each with one or more GCP projects

Ready-to-use Cloud Composer environment



Shared environment ingredients

User project

SA impersonation binding

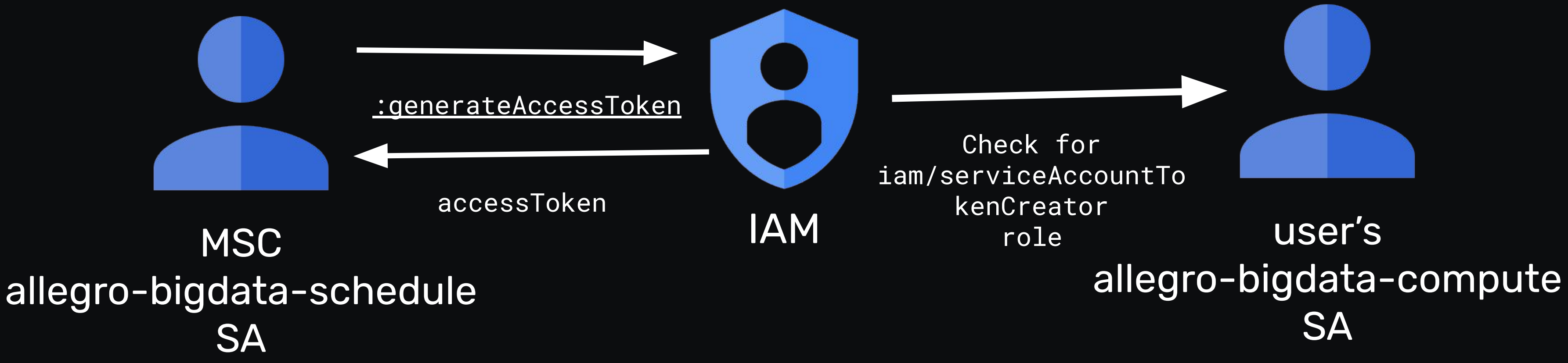
Composer

Composer bucket access

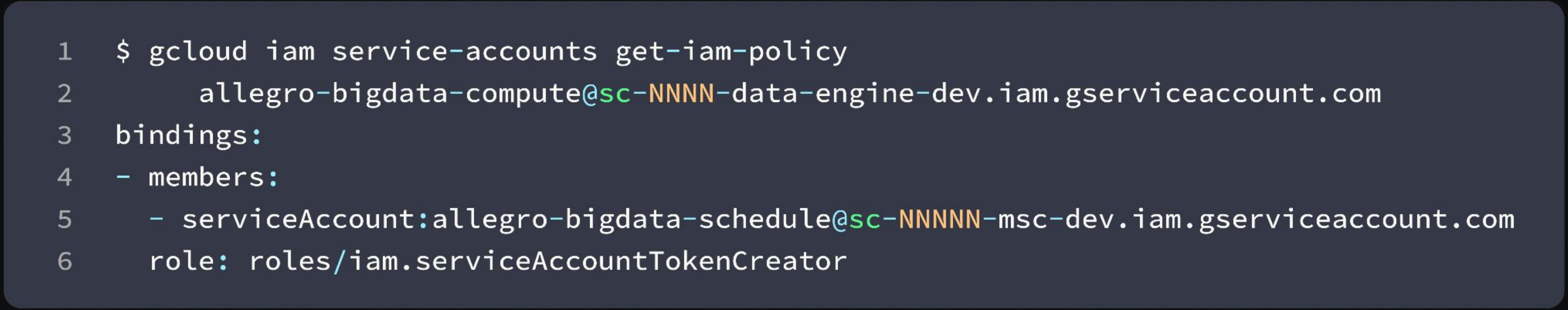
Composer User role

AIRFLOW SUMMIT

Service Account impersonation



Impersonation - IAM config



```
$ gcloud iam service-accounts get-iam-policy
     allegro-bigdata-compute@sc-NNNN-data-engine-dev.iam.gserviceaccount.com
bindings:
- members:
  - serviceAccount:allegro-bigdata-schedule@sc-NNNNN-msc-dev.iam.gserviceaccount.com
  role: roles/iam.serviceAccountTokenCreator
```

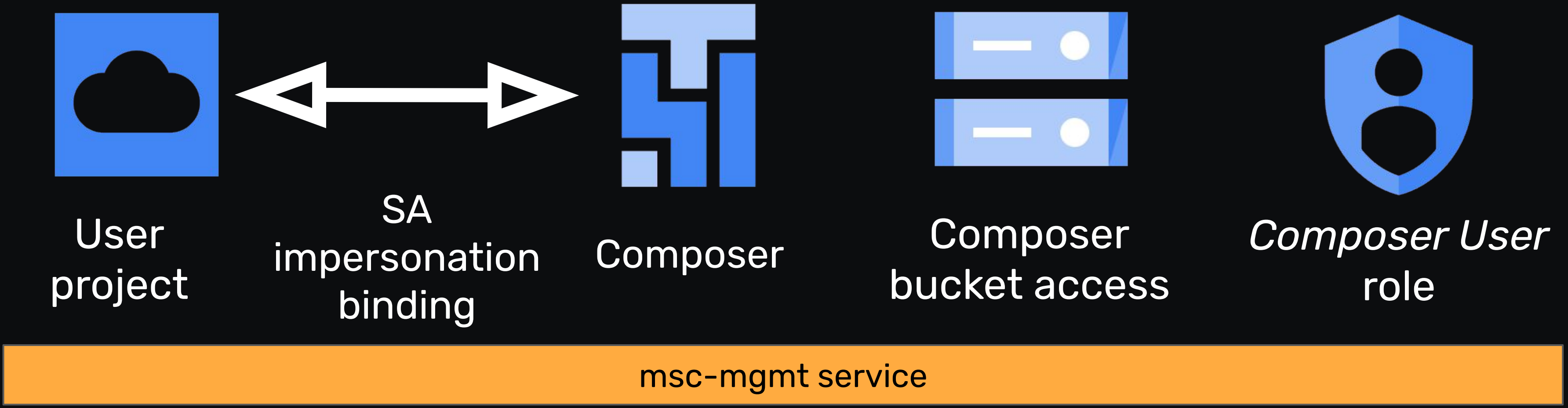
AIRFLOW SUMMIT

Impersonation - DAG code

```
gcp_project = f'sc-NNNN-data-engine-{env}'
gcp_service_account = f'allegro-bigdata-compute@{gcp_project}.iam.gserviceaccount.com'
IMPERSONATION_CHAIN = [gcp_service_account]

DEFAULT_DAG_ARGS = {
  …
    'project_id': gcp_project,
    'impersonation_chain': IMPERSONATION_CHAIN
}
```

Shared environment ingredients



User project

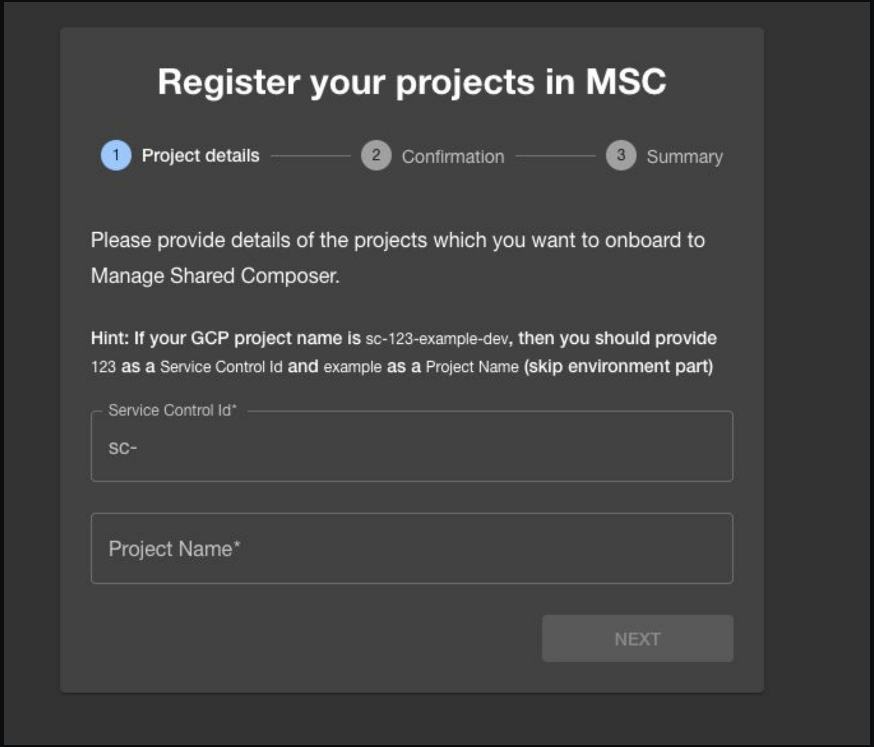
SA impersonation binding

Composer

Composer bucket access

Composer User role

msc-mgmt service



Register your projects in MSC

Please provide details of the projects which you want to onboard to Manage Shared Composer.

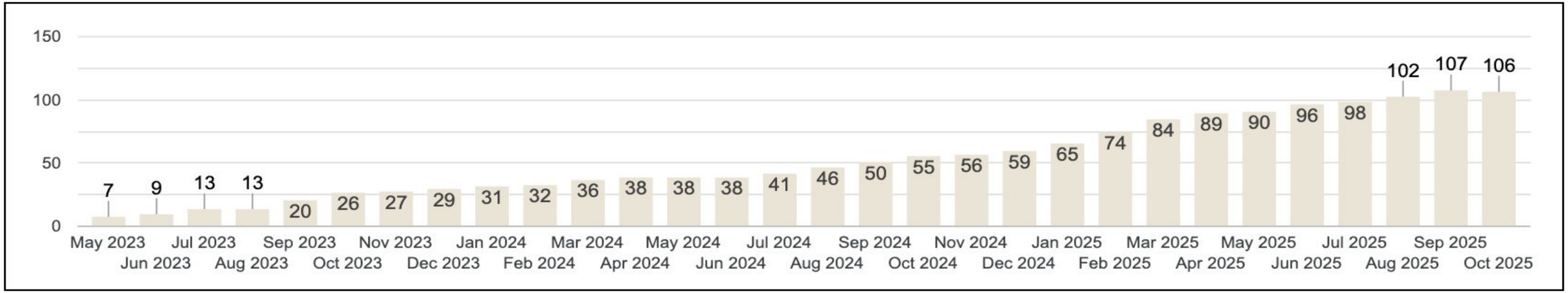
Hint: If your GCP project name is sc-123-example-dev**, then you should provide** 123 **as a** Service Control Id **and** example **as a** Project Name **(skip environment part)**

Service Control Id*: sc-

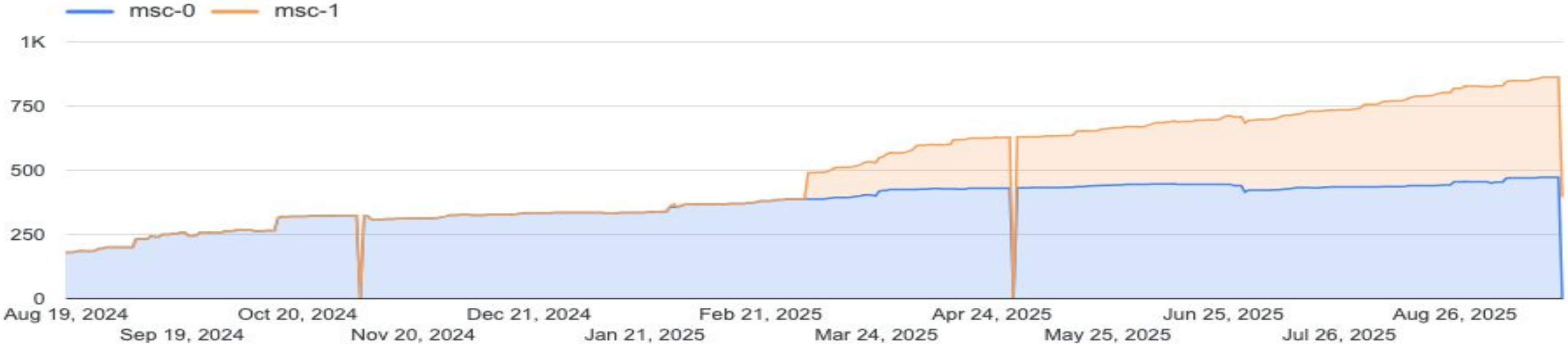
Project Name*:

Adoption

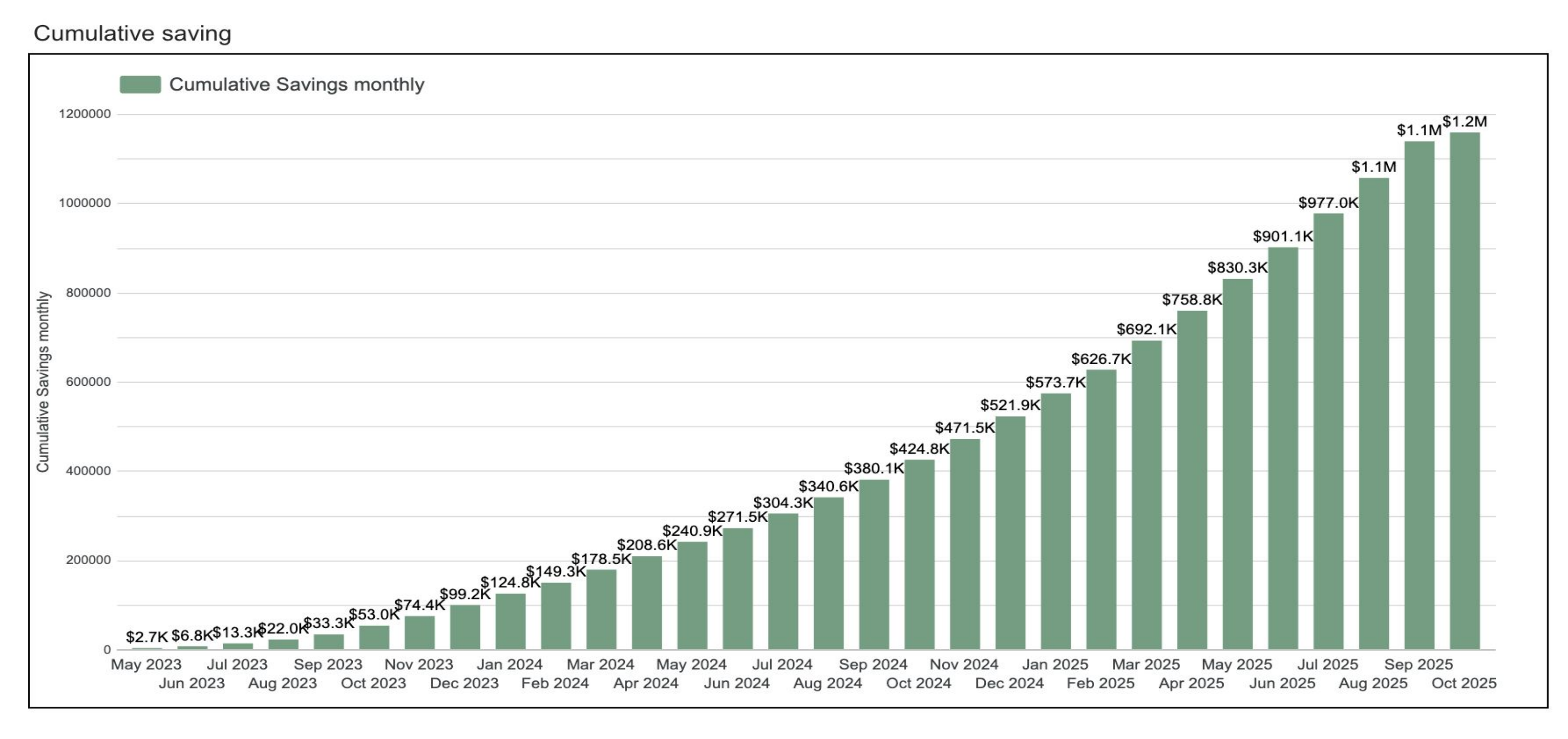
MSC Members Count



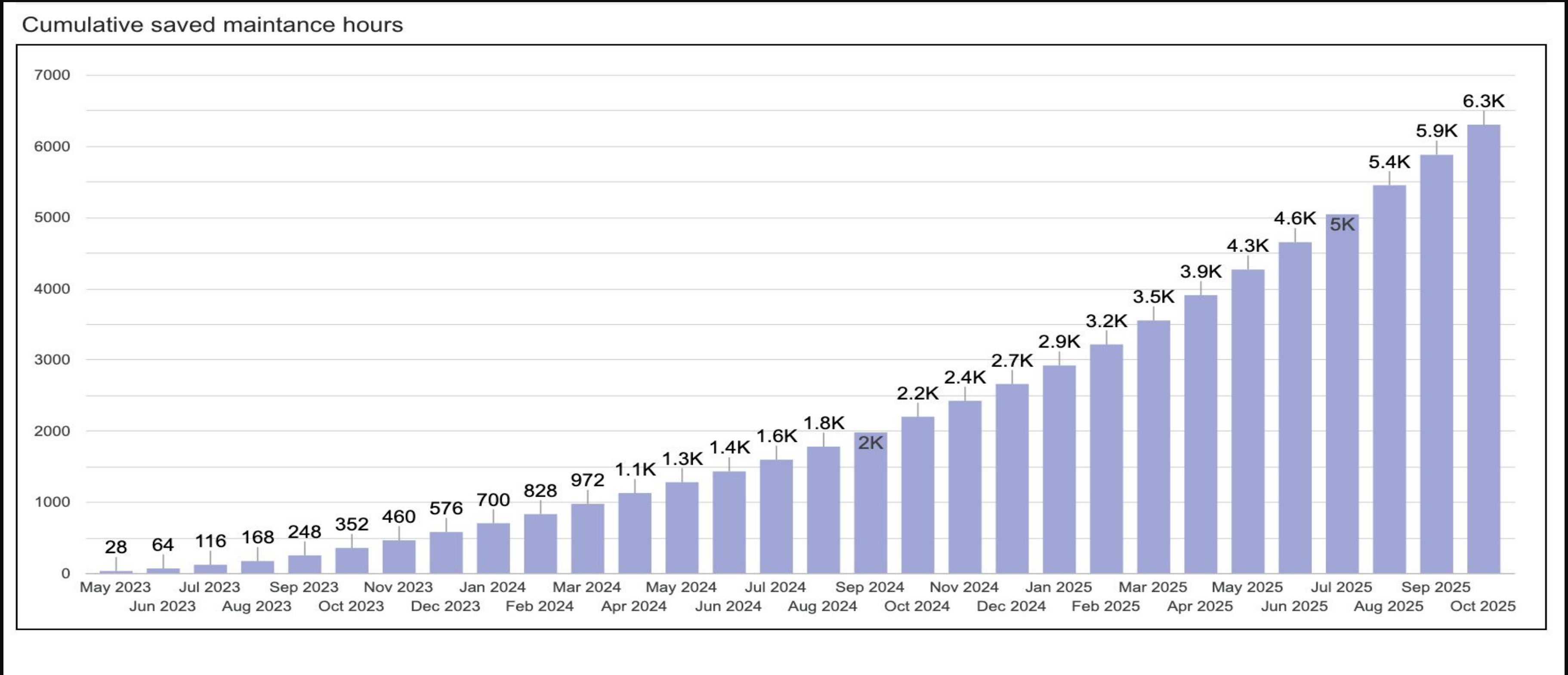
# Production DAGs



Cloud costs savings



Maintenance time saved



*Estimation based on surveyed average of 4h maint time per month per team

Shared environment challenges

- No place for special handling
- DAGs maintenance is still a member duty
- Noisy neighbours
- Limited access restrictions
- Single point of failure

Conclusions

3.0

Conclusions - opinionated ;)

- Airflow can be a backbone of wide variety of solutions
- Planning for maintenance even more than for start
- Self-service, conventions and processes pay off
- Shared environment(s) - worth considering but definitely not a silver bullet

Thank you!

Questions?

https://allegro.tech/