Enabling Data Mesh by Moving from a Monolithic Airflow to Several Smaller Environments

Stanislav Repka
Filip Kunčar
About us

Stanislav Repka
Data Engineer
@ Kiwi.com

Filip Kunčar
Data Platform Lead
@ Kiwi.com
About us

Jan Petřík
Data Engineer
@ Kiwi.com

Eliška Povolná
Data Engineer
@ Kiwi.com

Marian Špilka
Data Engineer
@ Kiwi.com

Filip Kunčar
Data Platform Lead
@ Kiwi.com

Stanislav Repka
Data Engineer
@ Kiwi.com

Igor Orság
Data Engineer
@ Kiwi.com
We find unique flight options and prices other search engines can’t see.

**100M**  Daily Searches

**70K**  Seats sold per day

**BILLIONS**  Daily price checks
History of Airflow in Kiwi.com

First deployment.
*Airflow as such dates back only to October 2014.

June 2016
History of Airflow in Kiwi.com

First deployment.
*Airflow as such dates back only to October 2014.

June 2016

Decision to shift data paradigm to Data Mesh concept.

2021
History of Airflow in Kiwi.com

First deployment. *Airflow as such dates back only to October 2014.*

2016

- June 2016: First deployment.

2021

- 2021: 25 teams running more than 500 active DAGs.
- 2021: Infrastructure spending mounts to $20k/mo for the single instance.
- 2021: Scheduler consuming more than 128vCPUs.

2022

- 2022: Decision to shift data paradigm to Data Mesh concept.
Unmaintainable Monolith that Didn't Fit The Concept

Skyflow
Unmaintainable Monolith that Didn't Fit The Concept

Skyflow → Security
Unmaintainable Monolith that Didn't Fit The Concept

Skyflow

Security
Stability
Unmaintainable Monolith that Didn't Fit The Concept

Skyflow

Security

Stability

Scalability
Data Mesh Principles

- Decentralized Data Ownership
- Self-Serve Data Infrastructure
- Data as a Product
- Federated Governance
Migration Riddles

**Non-technical**

- Persuade stakeholders
- Plan migration strategy

**Technical**

- Change sensors
- Prepare integration design
- Prepare cloud infrastructure
Implementation IAM
Implementation IAM

- Secret Manager
- BigQuery
- Cloud SQL
Implementation IAM

- Terraform
- IAM
- Secret Manager
- BigQuery
- Cloud SQL
Workload identity

Airflow Workload

Airflow Instance Service Account

Team Service Account

Cloud SQL

BigQuery

Secret Manager
Workload identity

Team Service Account

Service account A

Service account X

Cloud SQL

BigQuery

Secret Manager
Example of Representative DAG

PostgreSQL
Authorise to Cloud
PostgreSQL database using
Workload identity
Example of Representative DAG

PostgreSQL
Authorise to Cloud PostgreSQL database using Workload identity

Google Cloud Storage
Load data to GCS bucket by Workload identity authorisation
Example of Representative DAG

PostgreSQL
Authorise to Cloud PostgreSQL database using Workload identity

Google Cloud Storage
Load data to GCS bucket by Workload identity authorisation

BigQuery
Load data into BigQuery using workload identity authorisation
```python
dst_hook = BigQueryStorageHook.get_hook(
    impersonation_chain="data-platform@bi-sandbox-f64c11b3.iam.gserviceaccount.com",
    bq_project="bi-sandbox-f64c11b3",
    location="EU",
)
bq_client = dst_hook.bq_client
```
HTTP sensors

HTTP request

HTTP response

Team X

Team Y
Challenges

- Infrastructure costs on the hybrid plan.
- Central management of the decentralized instances.
Future

Streaming

Batch

BI

= 

Heart
Conclusion

Decentralization
(security, scalability, stability)

Engineers and Data Analysts
Cooperation Enhanced

Cloud Native Data
Thank you