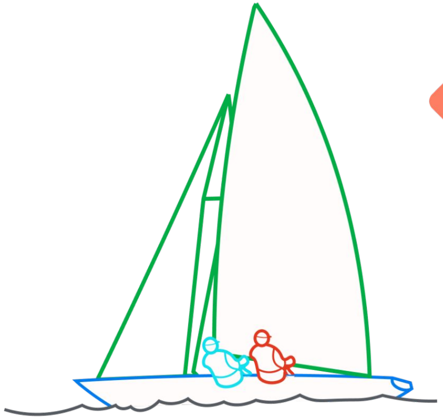
Enabling Data Mesh by Moving from a Monolithic Airflow to Several Smaller Environments

Stanislav Repka
Filip Kunčar



Airflow Summit
Let's flow together

September 19-21, 2023,
Toronto, Canada

About us

Stanislav Repka
Data Engineer
@ Kiwi.com

Filip Kunčar
Data Platform Lead
@ Kiwi.com

About us

Eliška Povolná
Data Engineer
@ Kiwi.com

Jan Petřík
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

2021

Decision to **shift** data paradigm **to Data Mesh concept**.

History of Airflow in Kiwi.com

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

June 2016

2021

Decision to **shift** data paradigm **to Data Mesh concept**.

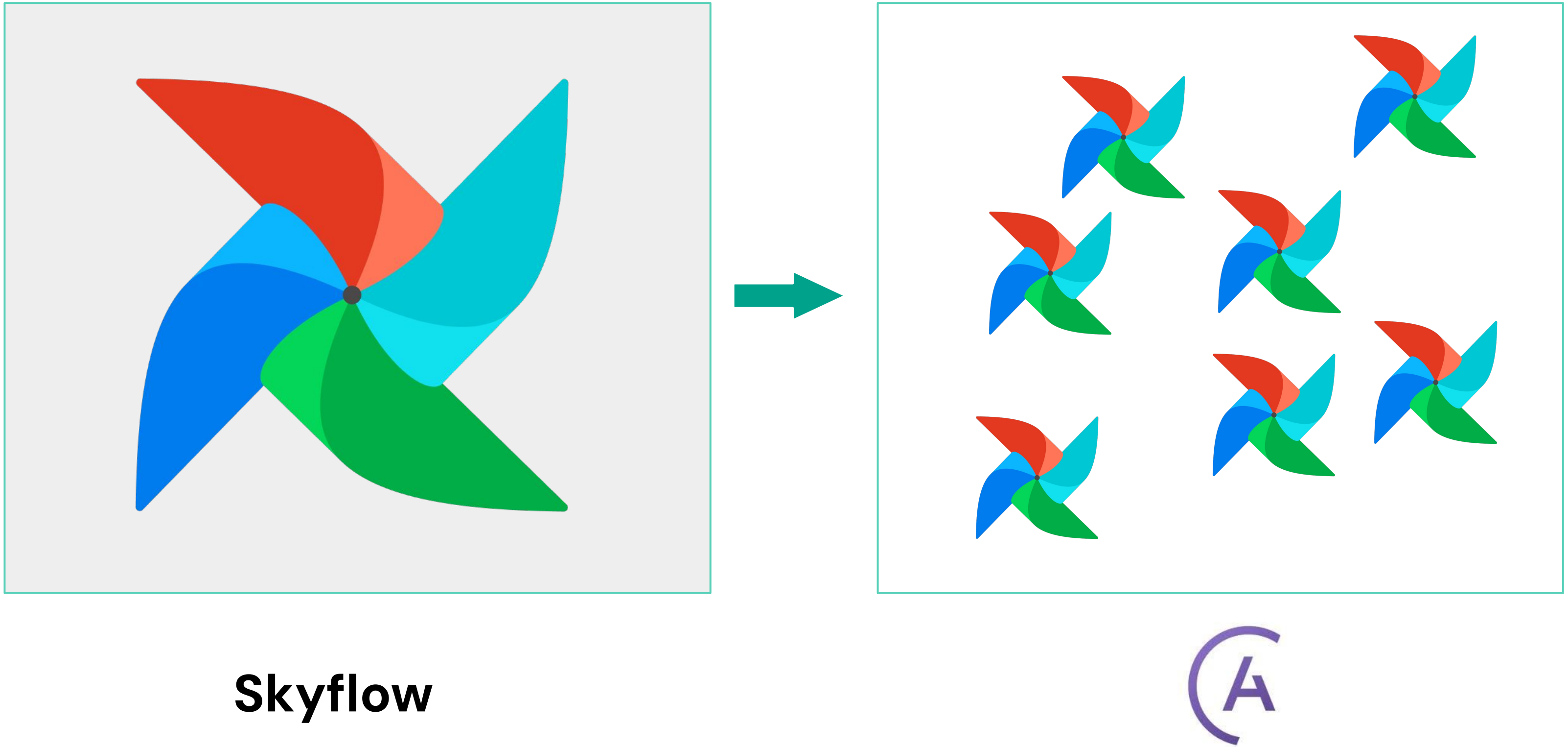
2022

25 teams running more than **500 active DAGs.**

Infrastructure spending mounts to **$20k/mo for the single instance.**

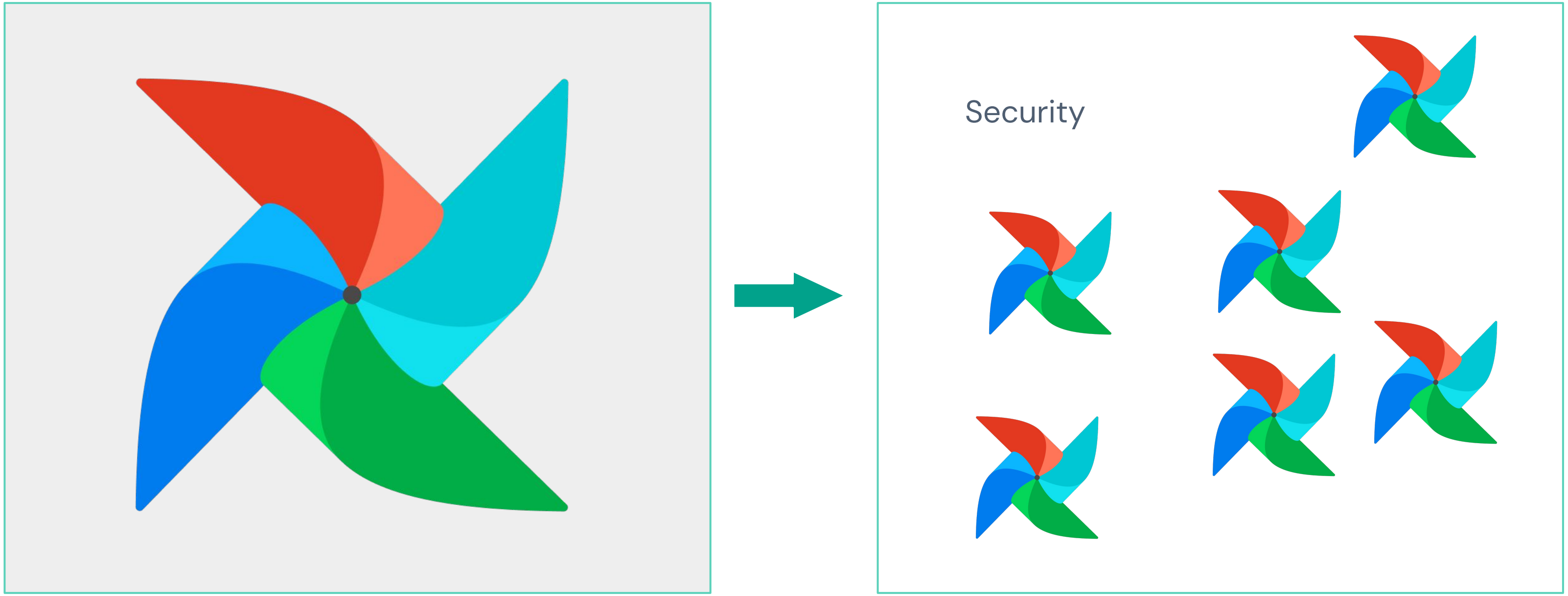
Scheduler consuming **more than 128vCPUs.**

Unmaintainable Monolith that Didn't Fit The Concept



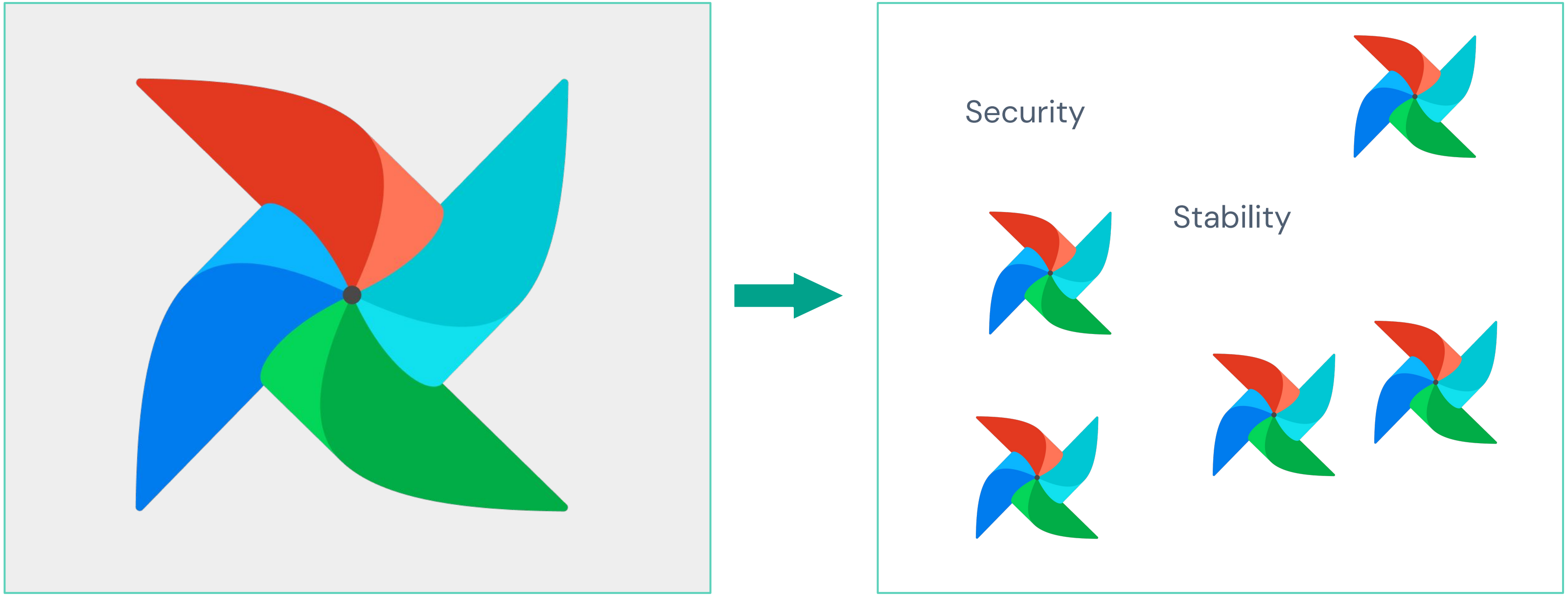
Skyflow

Unmaintainable Monolith that Didn't Fit The Concept



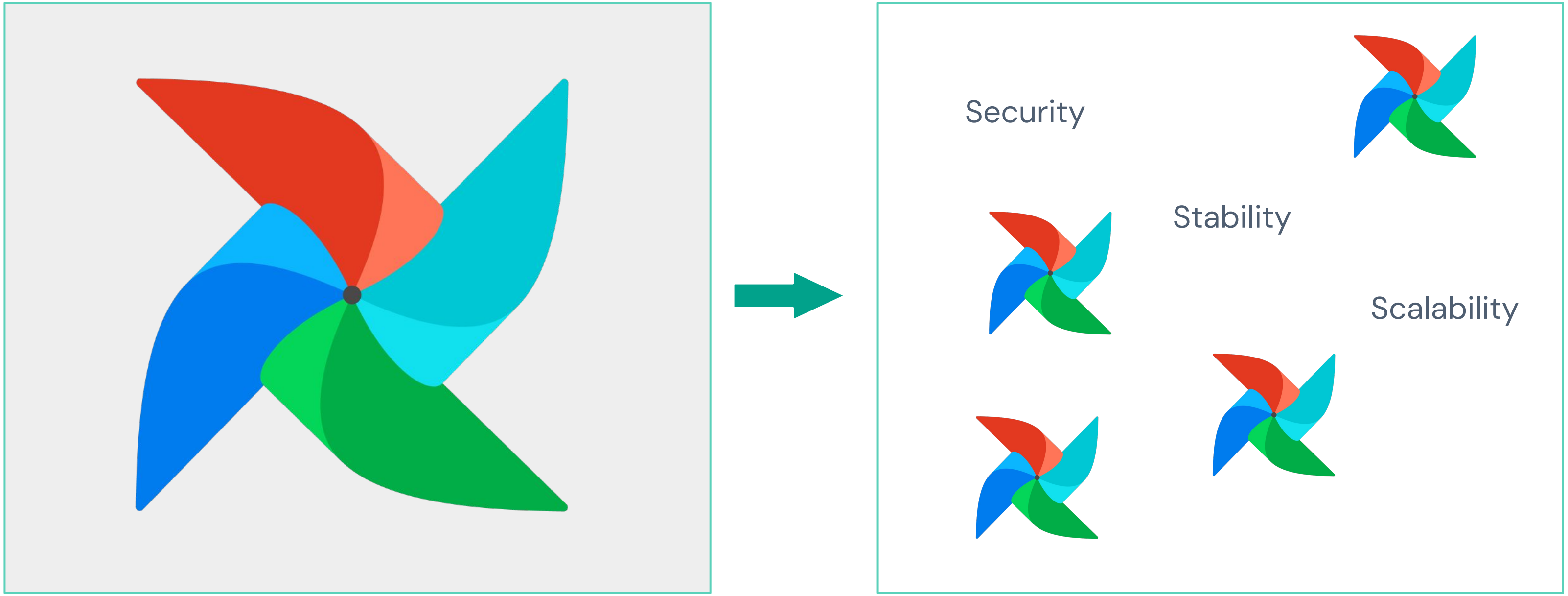
Skyflow

Unmaintainable Monolith that Didn't Fit The Concept



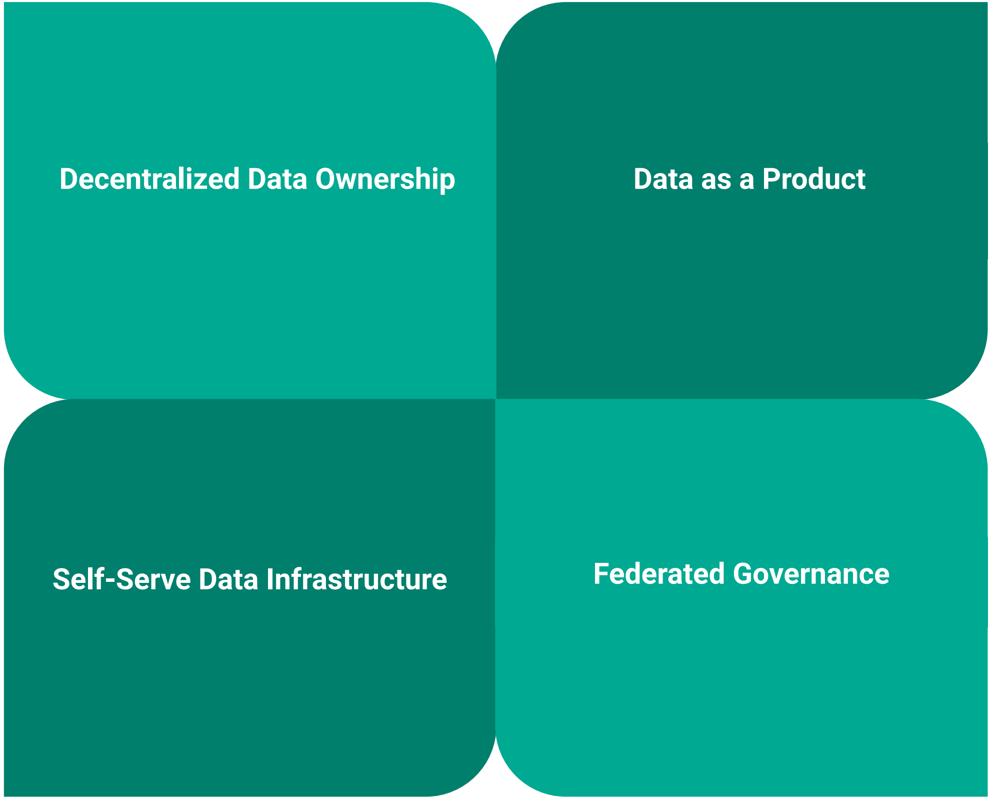
Skyflow

Unmaintainable Monolith that Didn't Fit The Concept



Skyflow

Data Mesh Principles



Decentralized Data Ownership

Data as a Product

Self-Serve Data Infrastructure

Federated Governance

Migration Riddles

Non-technical

Persuade stakeholders

Plan migration strategy

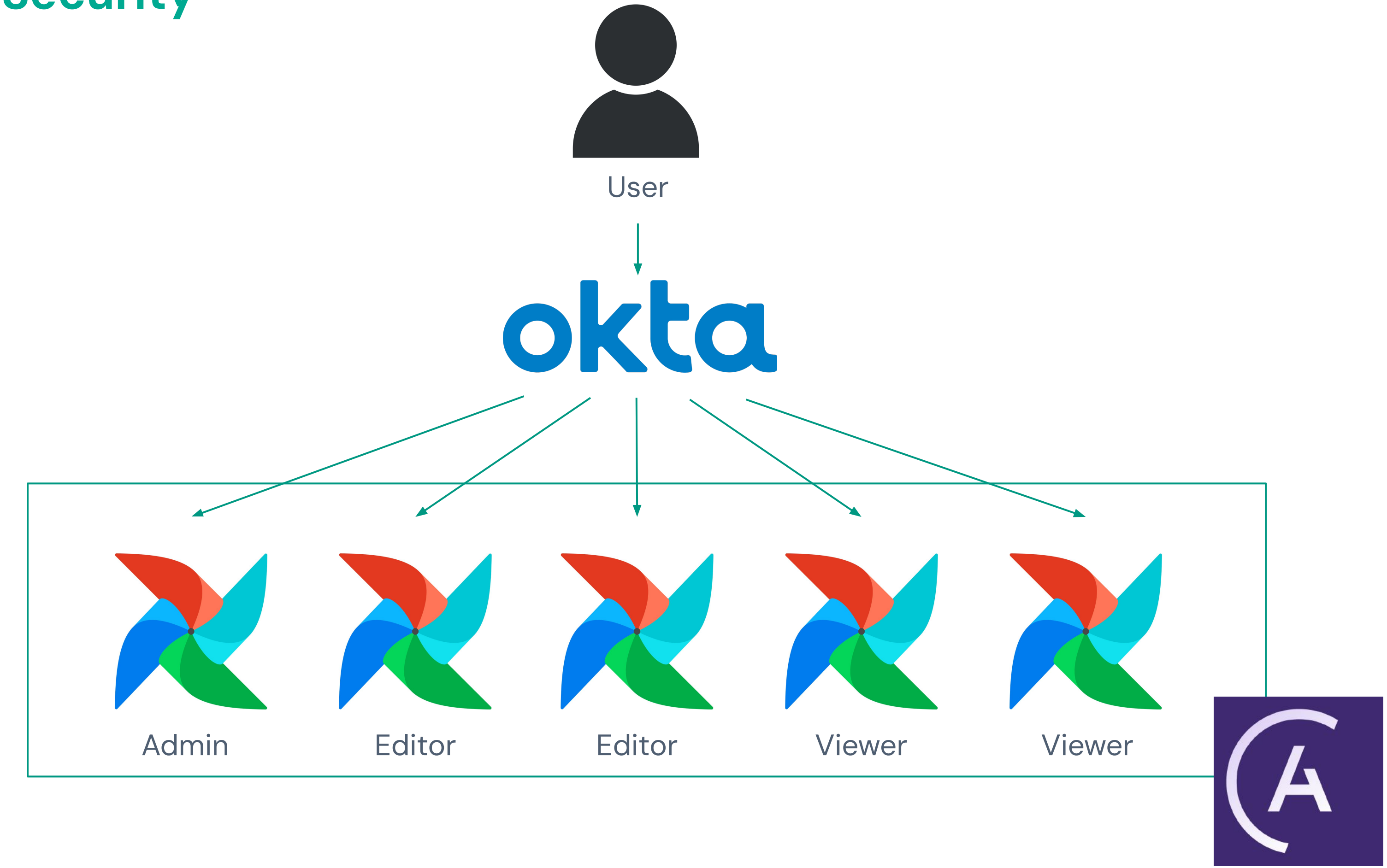
Technical

Change sensors

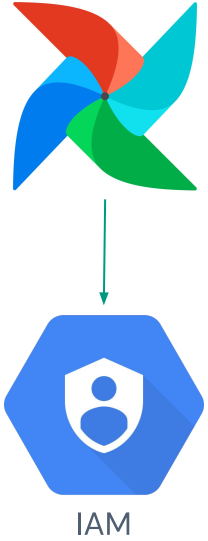
Prepare integration design

Prepare cloud infrastructure

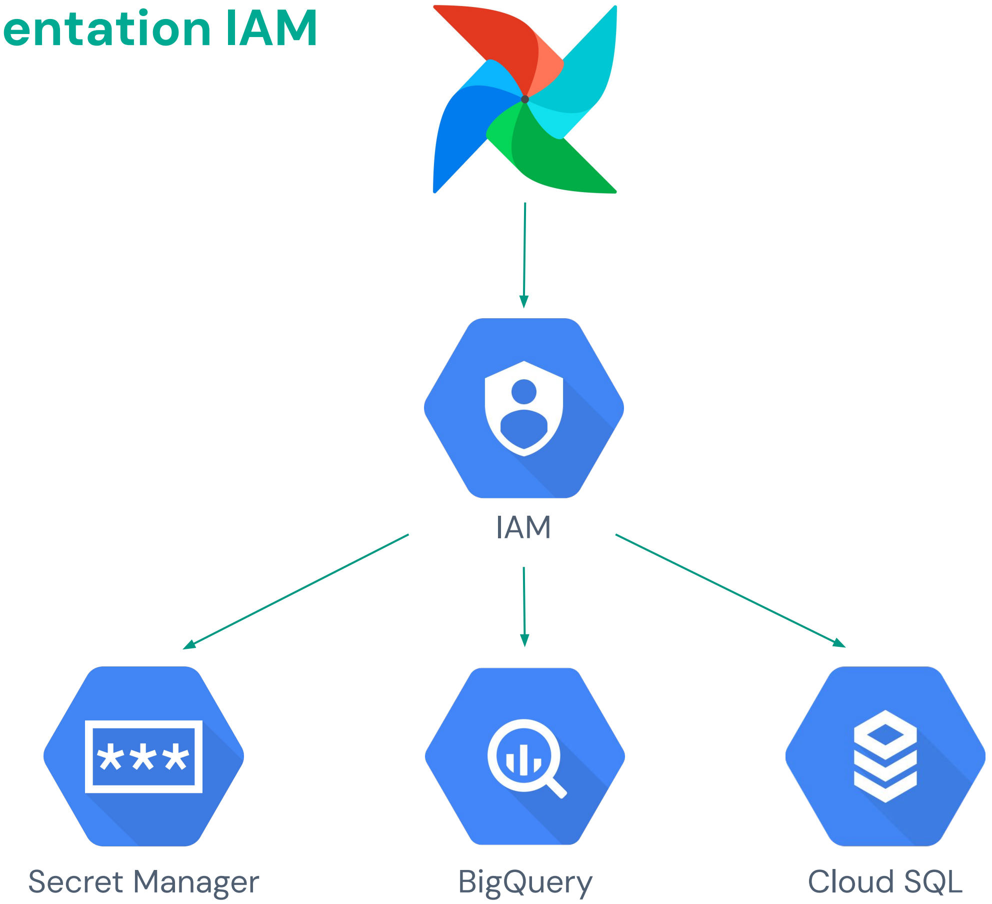
Security



Implementation IAM



Implementation IAM



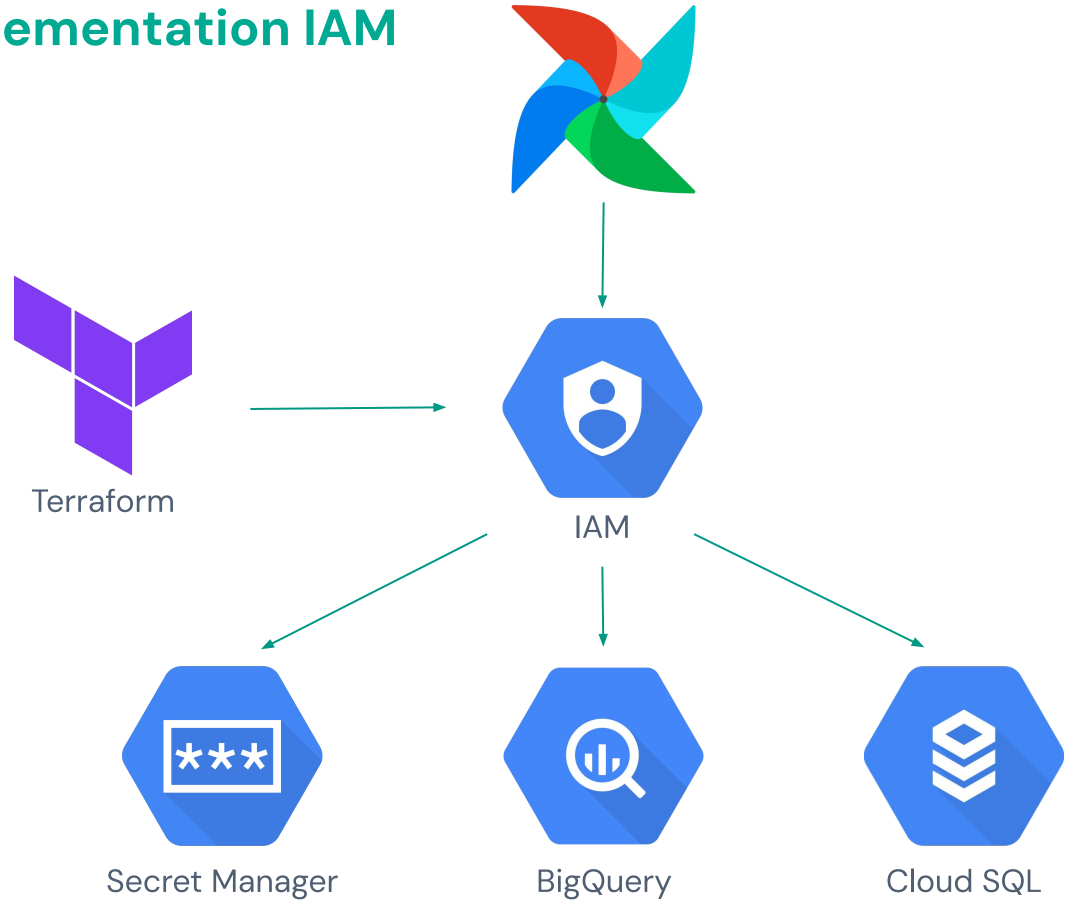
IAM

Secret Manager

BigQuery

Cloud SQL

Implementation IAM



Terraform

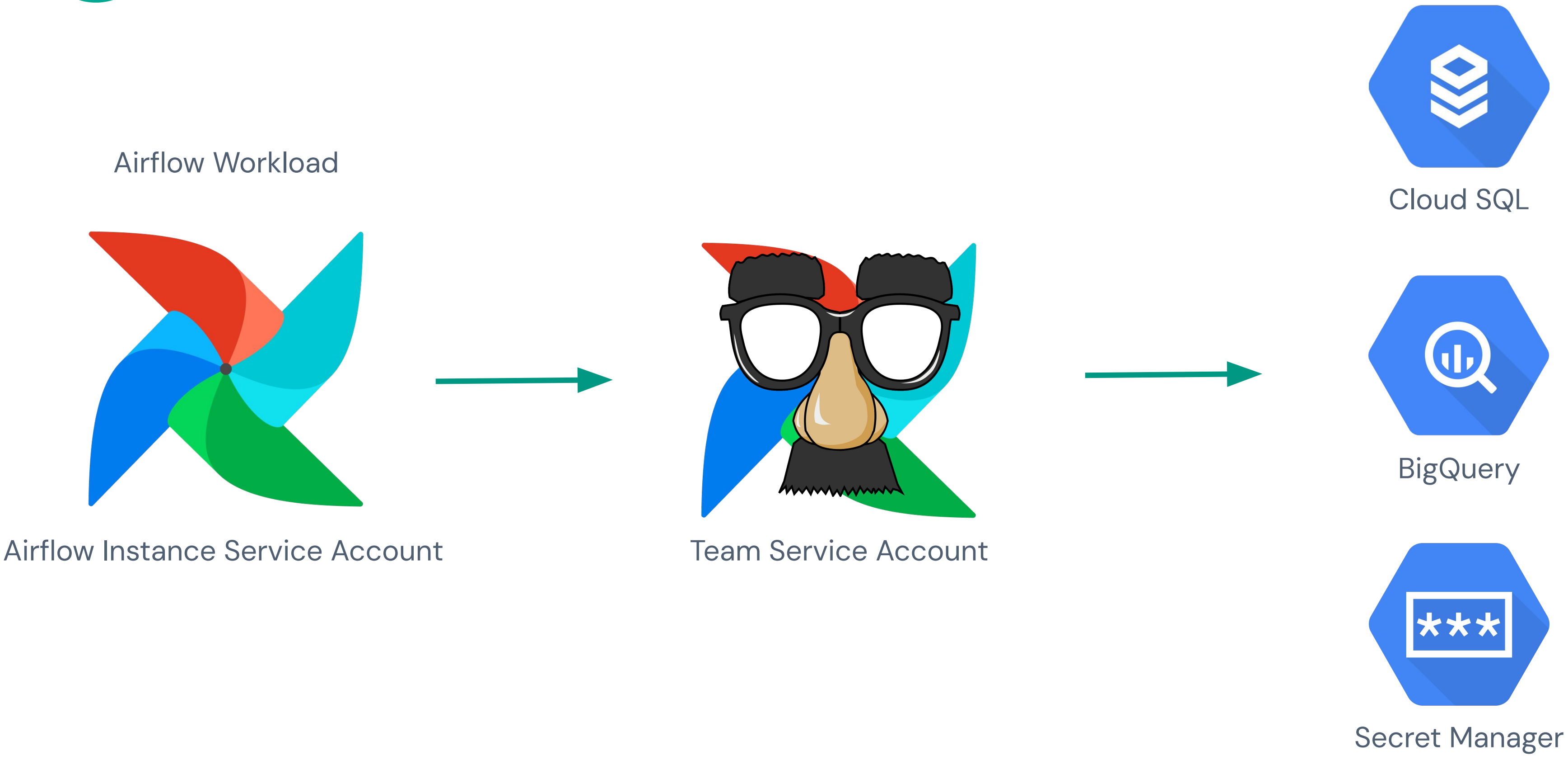
IAM

Secret Manager

BigQuery

Cloud SQL

Workload identity



Airflow Workload

Airflow Instance Service Account

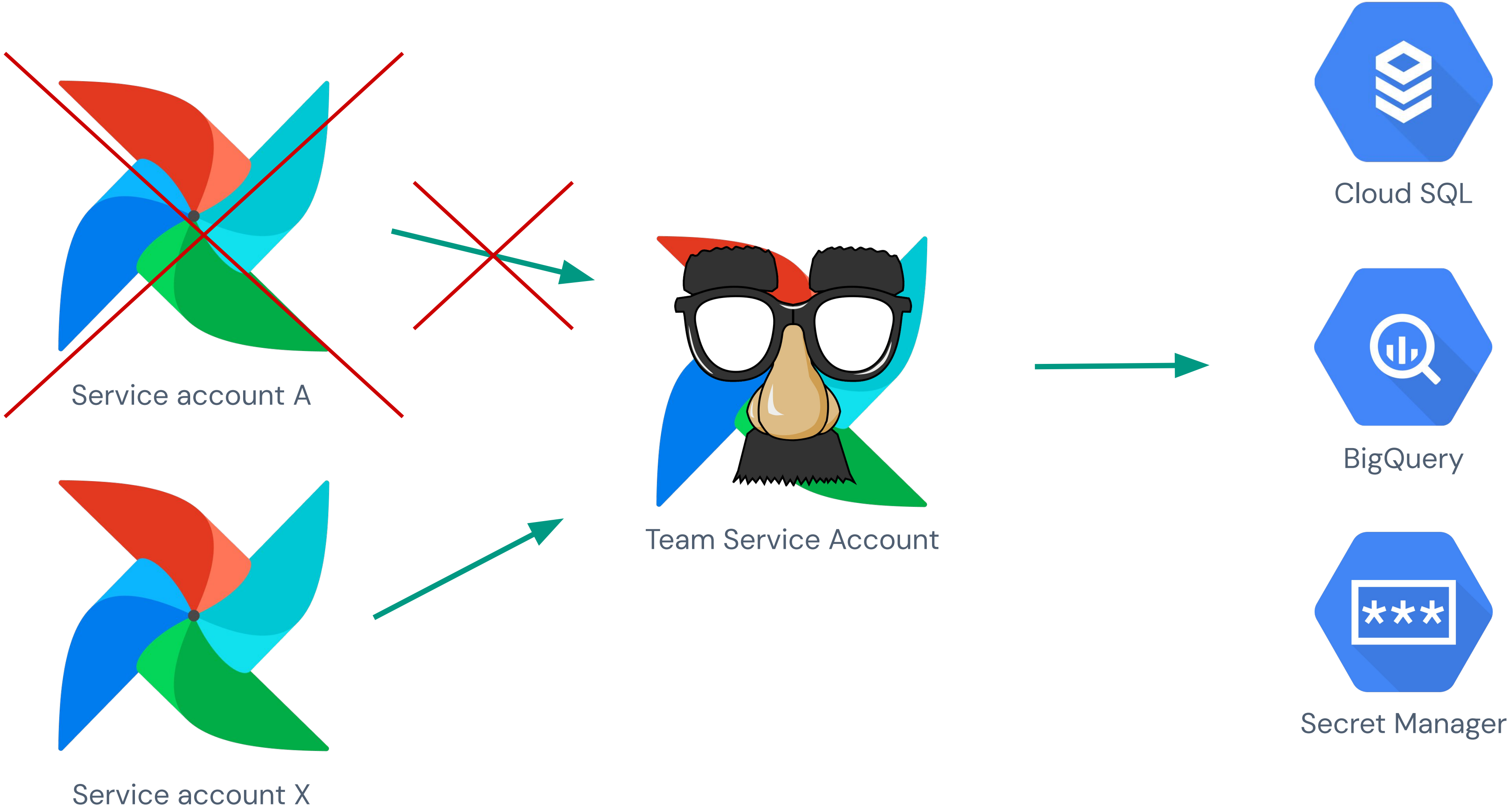
Team Service Account

Cloud SQL

BigQuery

Secret Manager

Workload identity



Service account A

Service account X

Team Service Account

Cloud SQL

BigQuery

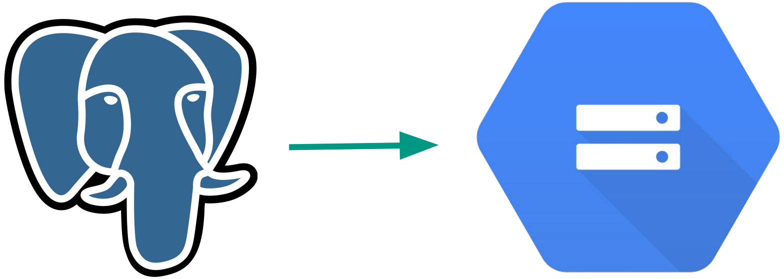
Secret Manager

Example of Representative DAG

PostgreSQL

Authorise to Cloud
PostgreSQL database using
Workload identity

Example of Representative DAG



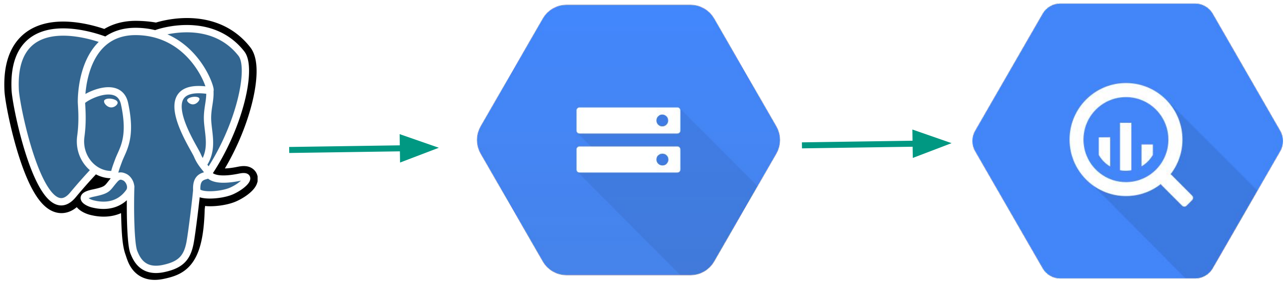
PostgreSQL

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

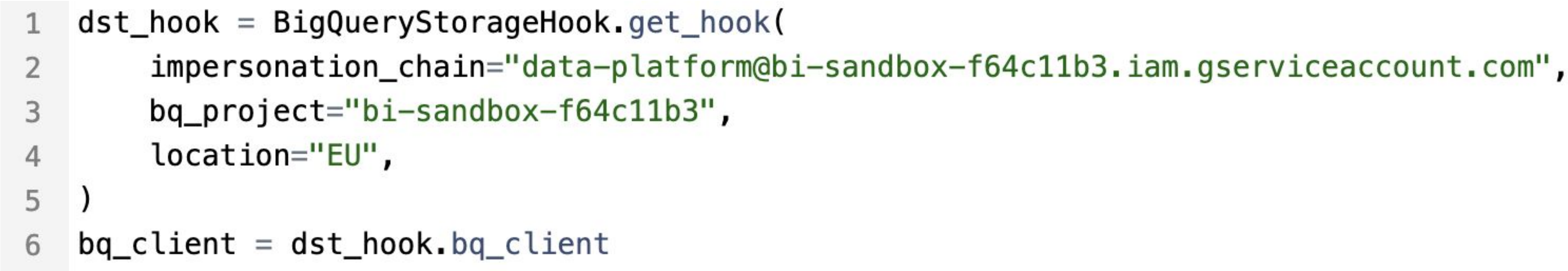
Google Cloud Storage

Load data to GCS bucket by Workload identity authorisation

BigQuery

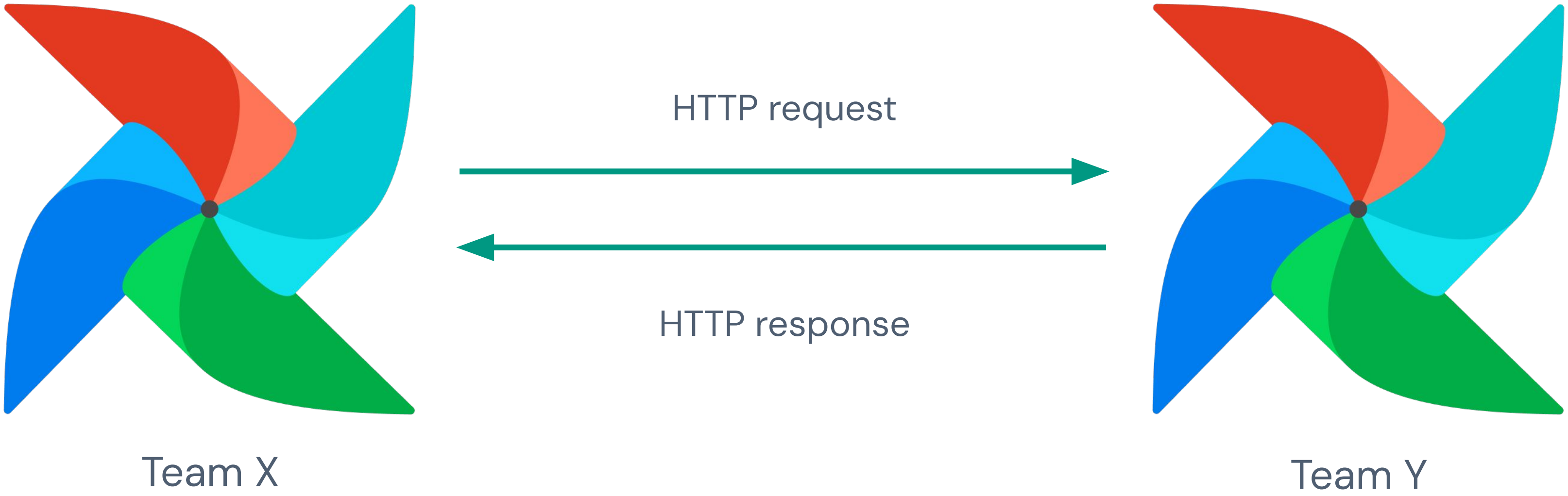
Load data into BigQuery using workload identity authorisation

Code Example



```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

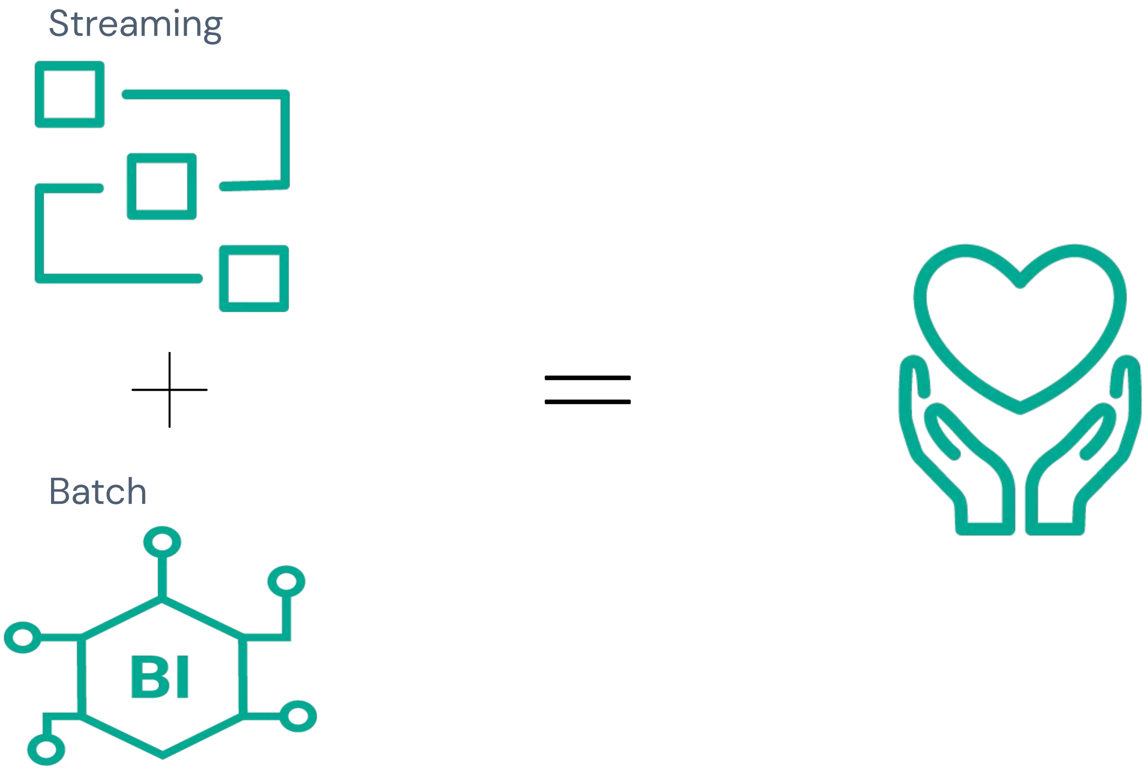
Team Y

Challenges

Infrastructure costs on the hybrid plan.

Central management of the decentralized instances.

Future



Streaming

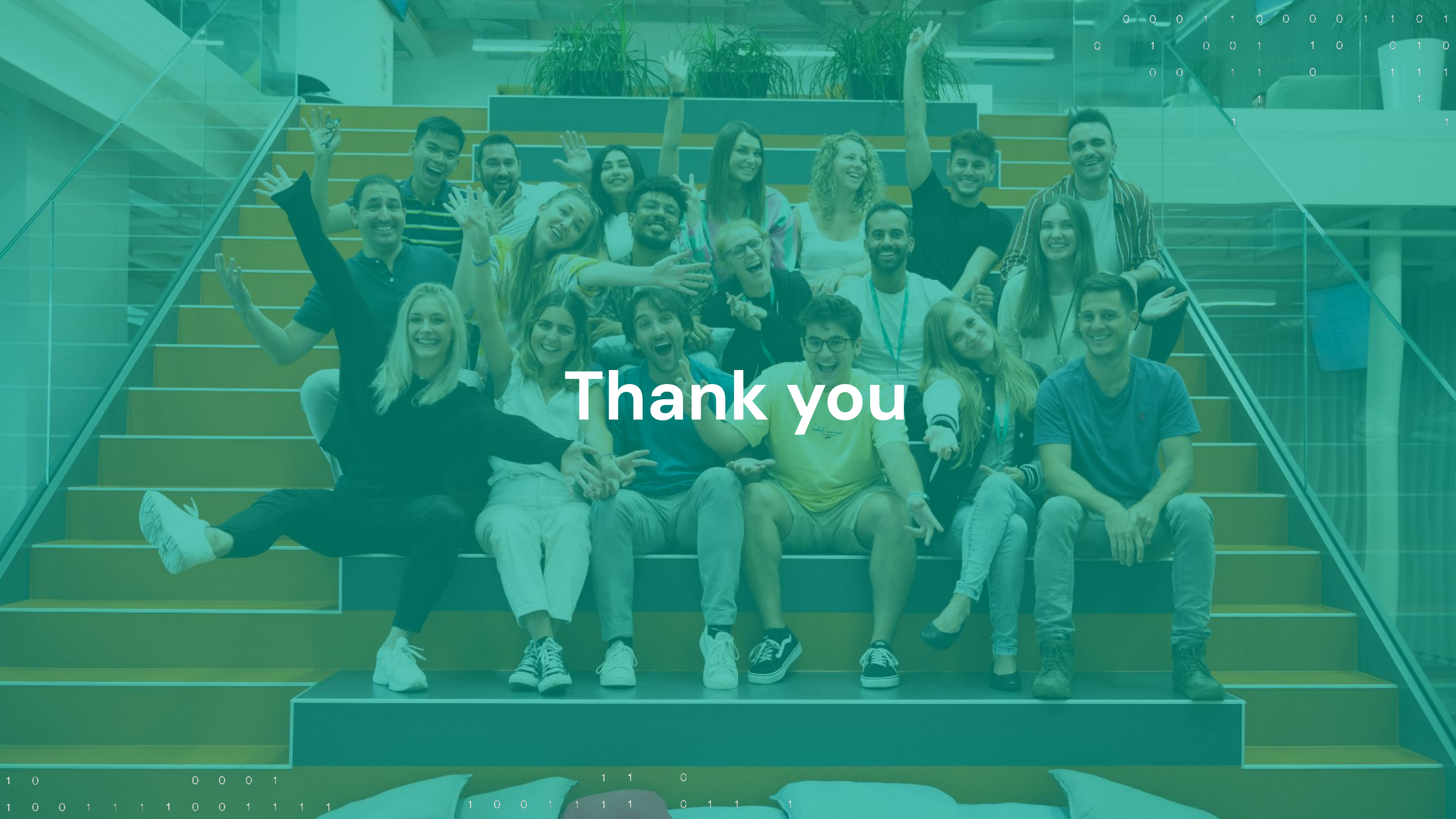
Batch

Conclusion

Decentralization (security, scalability, stability)

Engineers and Data Analysts Cooperation Enhanced

Cloud Native Data



Thank you

Q&A

Stanislav Repka
Data Engineer @ Kiwi.com

Filip Kunćar
Data platform lead @ Kiwi.com