An On-Demand Airflow Service for Internet Scale Gameplay Pipelines

Apache Airflow

Yuanmeng Zeng & Nitish Victor

Electronic Arts
EADP Data & AI

Yuanmeng Zeng
Software Engineer II

EA Digital Platform(EADP) DATA & AI

yzeng@ea.com


APEX
LEGENDS


PARIS
PARIS EST
FIFA 21
ACCOR LIVE LIMITLESS

EA Digital Platform(EADP)

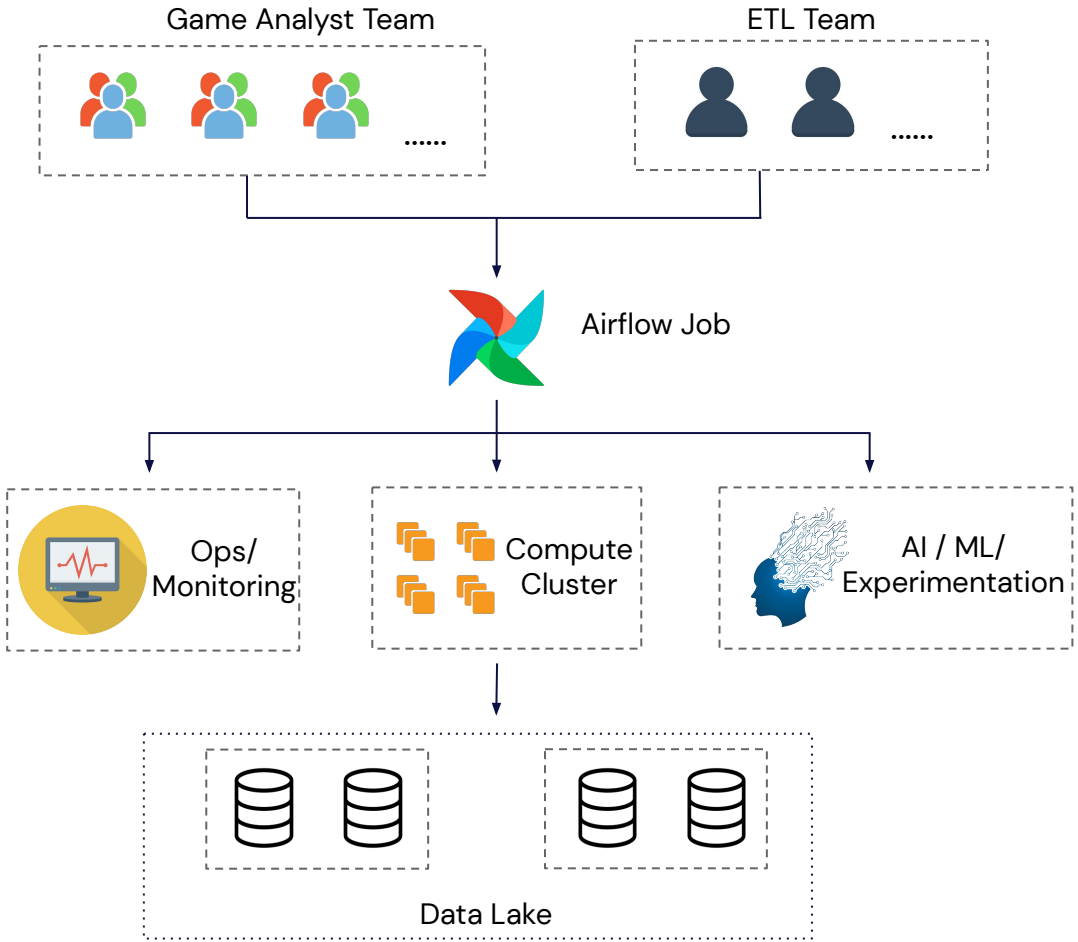
- **EADP Data & AI**
  - Data capture
  - Data storage and processing
  - Data science and analysis
- **Game Telemetry adhering to a common taxonomy**
- **Centralized data platform with unified access to the data**

Scale

- About 40 game studios under EA's umbrella
- Ever-changing data landscape with new acquisitions such as Codemasters and Glu Mobile
- Terabyte-scale data generated daily and petabyte-scale data access
- Thousands of ETL jobs
- Different needs in terms of data from different studios

Airflow at EADP

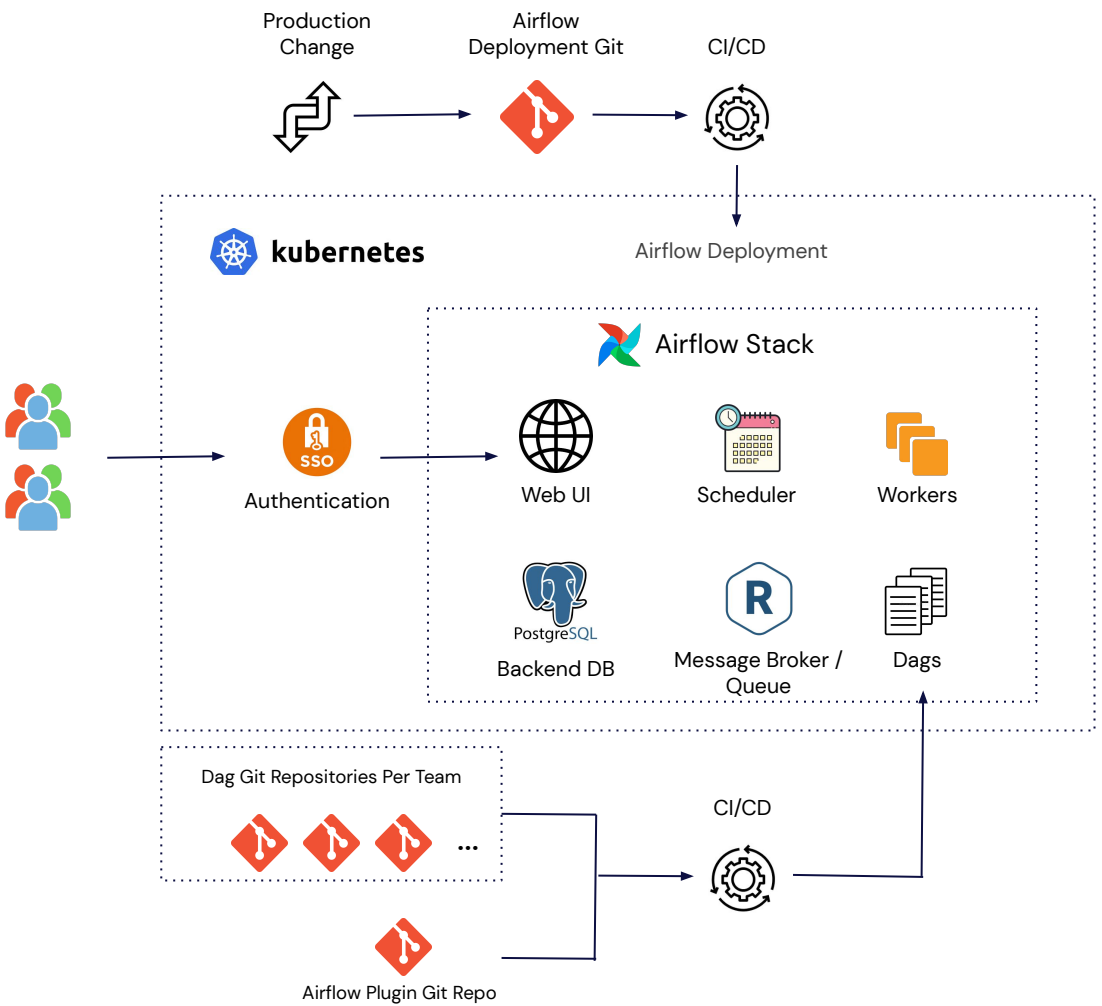
- Airflow as a Service
- Support multiple game analyst teams and EADP ETL team
- DevOps job for monitoring and scaling
- Experimentation job for AI and machine learning
- Data processing job using our compute cluster



Before

Single Airflow Cluster

- Monolithic design
- Multi-tenant shared environment
- Variables and connections managed through single Airflow instance.



Motivation for On-Demand Airflow Service

Lack of Isolation between teams

Need Of team-level ACL on sensitive datasets, variables and connections

Implementation is monolithic

Any platform level change impacts all production workloads

Difficult to support diverse use-cases across teams

Teams have very different use cases and ways they use Airflow

Lack of self-serve management capabilities

Teams are unable to manage user access on their own and scale up/down according to their needs

Ever increasing DAG count and workload

High stress on a single scheduler

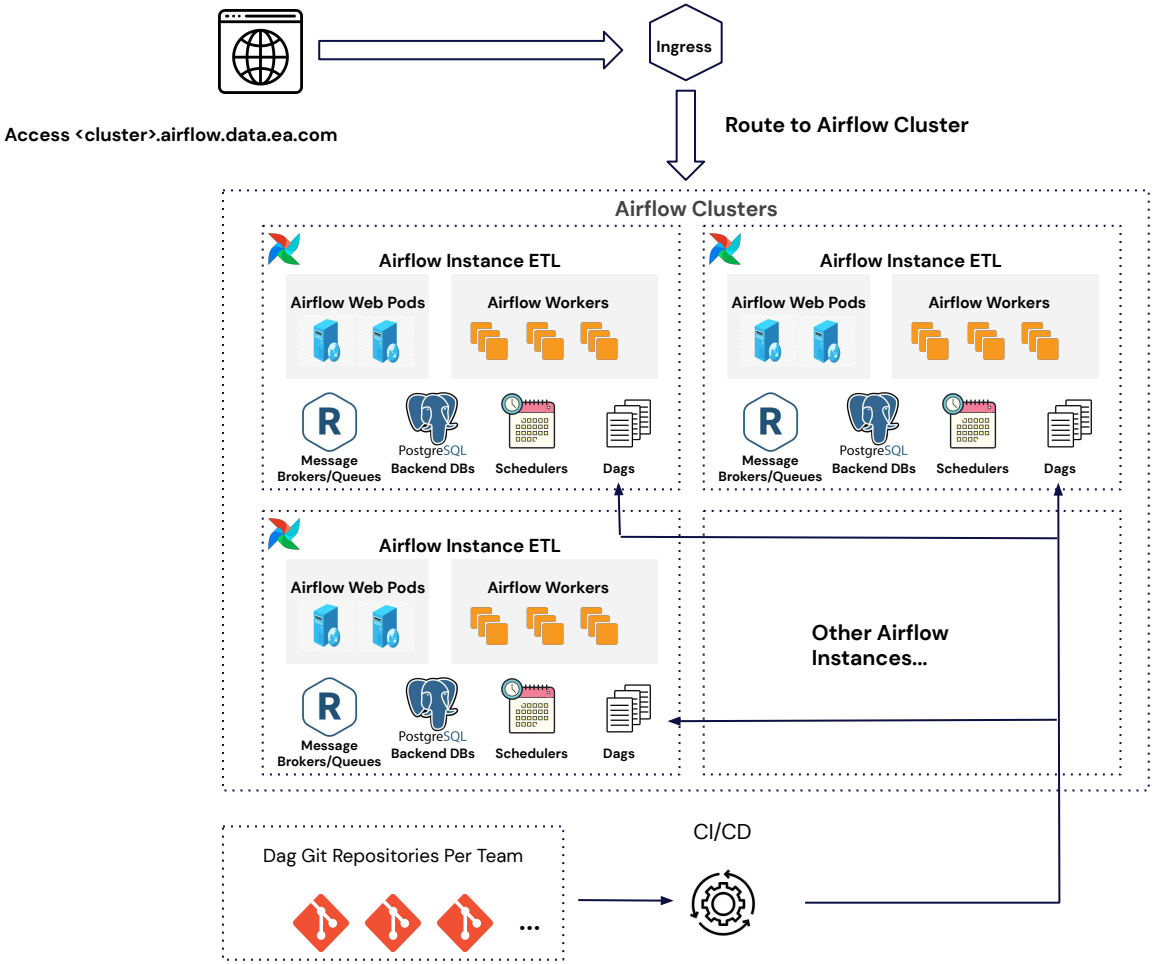
Variance resource requirement of dags

Workload of different dags has different requirements for hardware resources

On-Demand Airflow Design

Multiple Airflow clusters

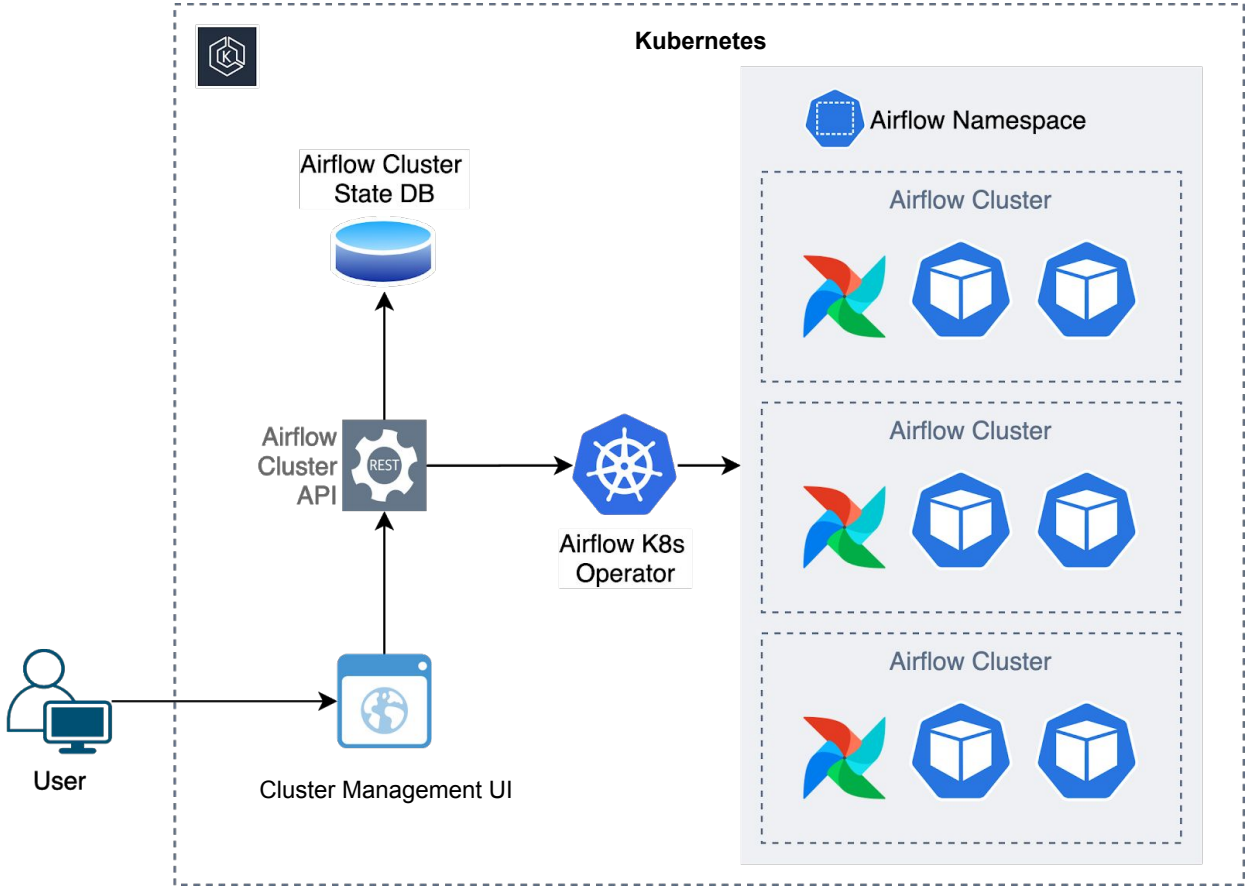
- Each “team” or entity gets their own isolated Airflow cluster(s)
- Route to their cluster using subdomains under *airflow.data.ea.com*
- For example, FIFA team could have a cluster at *fifa.airflow.data.ea.com*
- Deploy new dags through CI/CD without redeployment of airflow



On-Demand Airflow Design

Airflow using K8s Operator

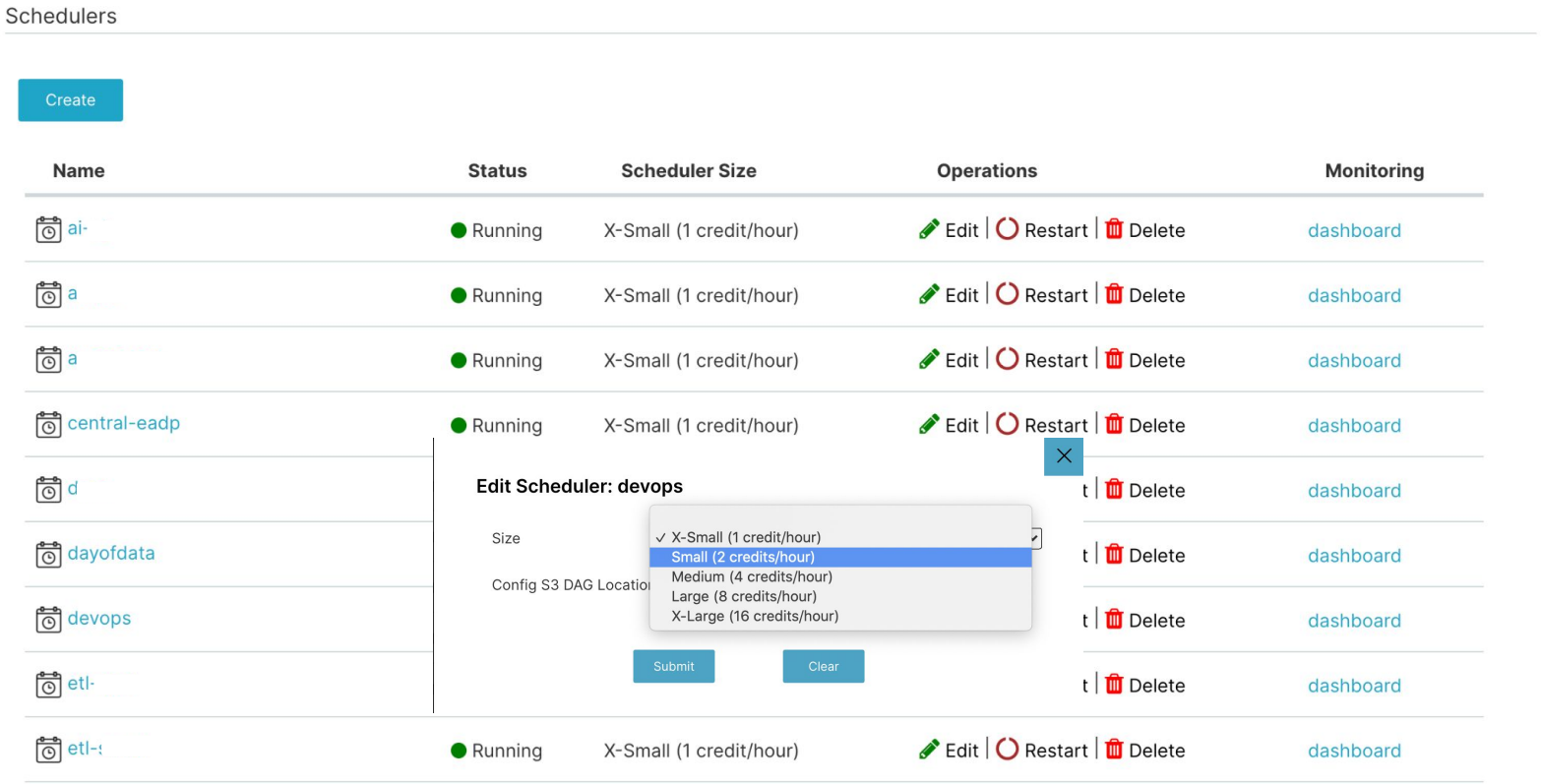
- Use cloud-native Kubernetes operator to deploy/scale/shutdown clusters
- Customize or create an operator for EADP use-case and expose functionality via an API.
- REST API layer in front of K8s Operator to allow external calls



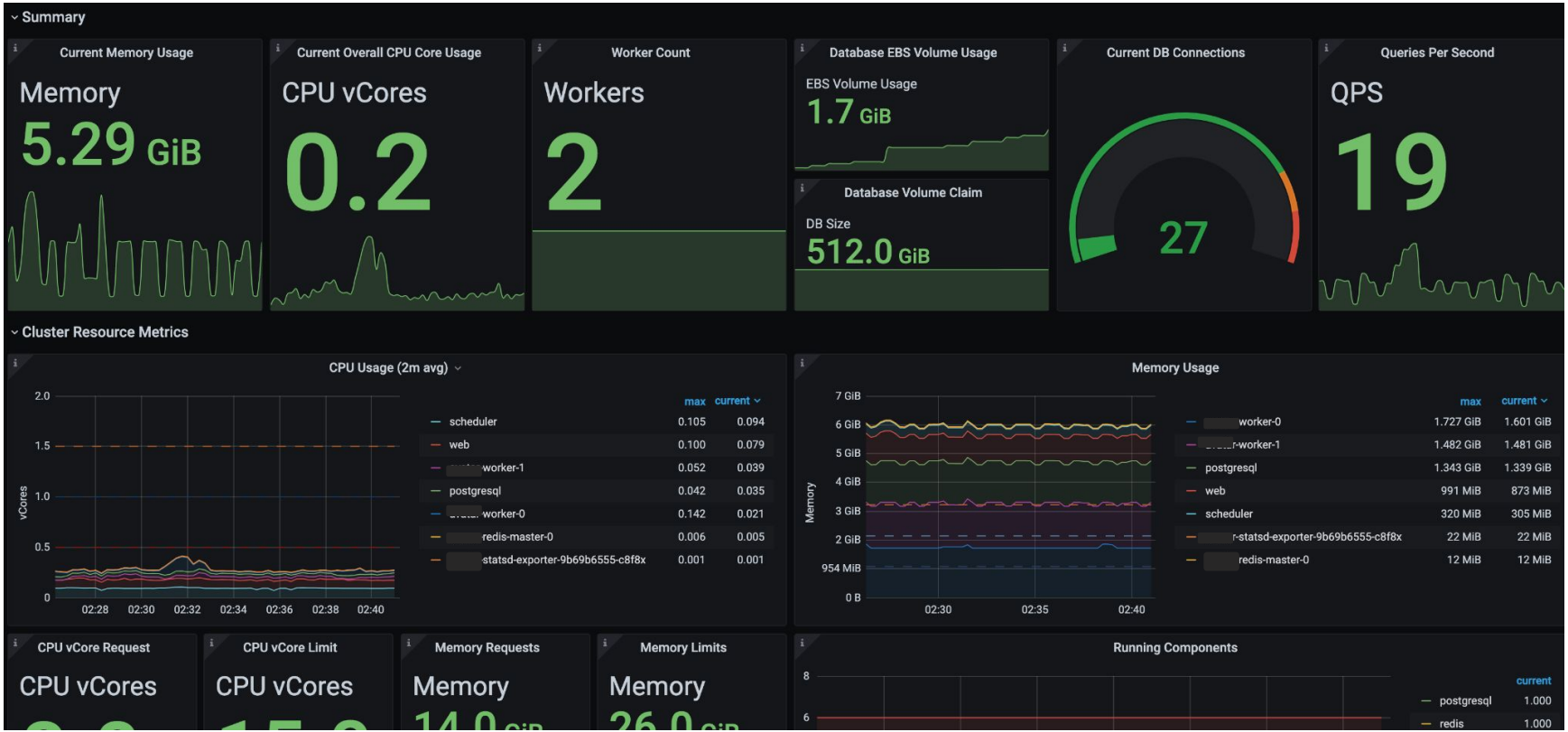
Cluster Sizes

Cluster Size	XS	S	M	L	XL
Task Concurrency	4	16	32	64	128
Worker Concurrency	2	4	4	4	8
Worker Count	2	4	8	12	16

Self-Serve Management UI



Monitoring Dashboard



Improvements

Fully self-serve

Minimum impact of downtime

Cluster flexibility & Diversity

Homogeneous jobs in the same cluster

Easy troubleshooting

Cost saving

Future work

- AI-based cluster scaling
- Various hardware support
- More granular access control on cluster operation

Q&A

Apache Airflow