
May 23–27, 2022
AIRFLOW
SUMMIT

Airflow Summit 2022

Orchestrating hybrid workflows with Apache Airflow

Ricardo Sueiras
Principal Developer Advocate, AWS

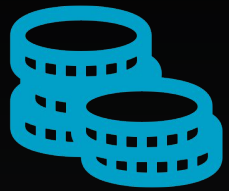
Apache

Airflow

Regulation &
Compliance

Heritage
systems

Reduce
complexity



Cost
effective

SqlToS3Operator

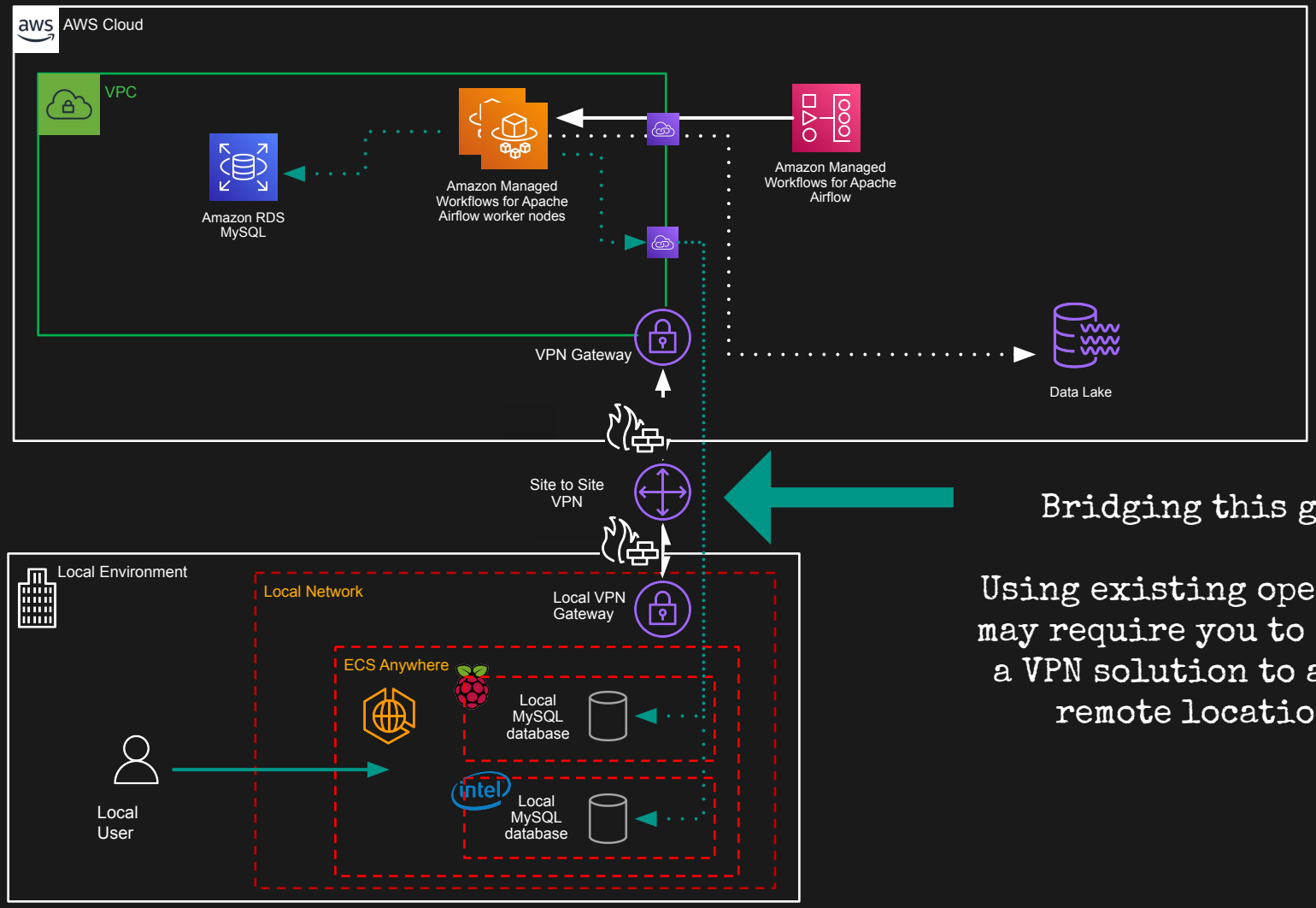
- Need to integrate networks between Cloud and remote (via VPN for example)
- Additional work required to enable connectivity (inbound/controlled networks)
- Processing of data in Airflow Worker

PythonOperator

- Need to integrate networks between Cloud and remote (via VPN for example)
- More complex DAGs
- Processing of data in Airflow Worker
- Potential Apache Airflow anti-pattern

AthenaOperator

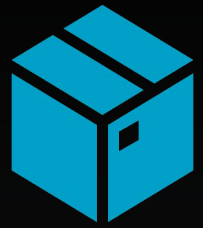
- Need to integrate networks between Cloud and remote (via VPN for example)
- Processing of data in Cloud (Athena)
- Create Athena Federated Queries (running as Lambda functions)
- Additional work required to enable connectivity (inbound/controlled networks)



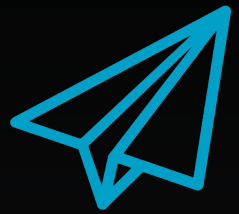
Bridging this gap.

Using existing operators may require you to deploy a VPN solution to access remote locations

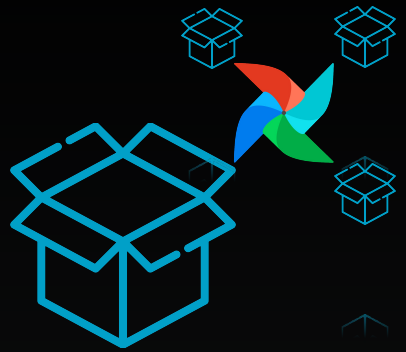
Develop ETL application



Package as container



Push to Container Repository



Orchestrate running container

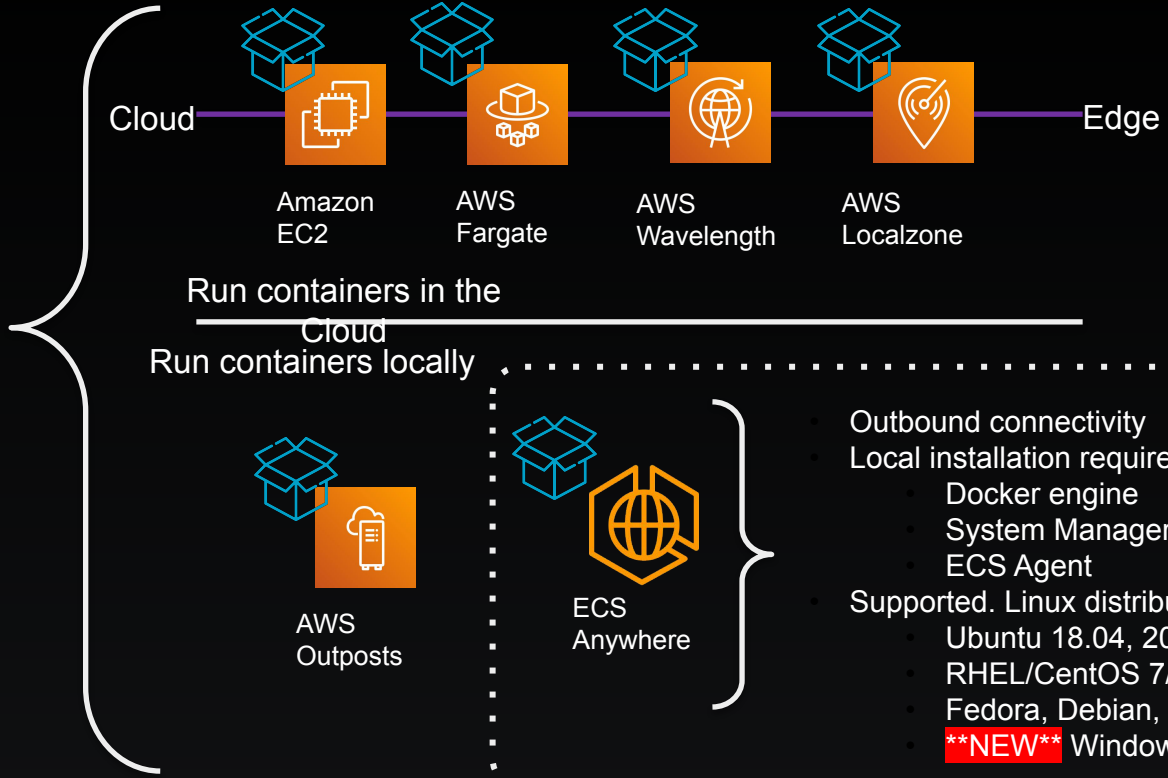
KubernetesOperator

- Need to integrate networks between Cloud and remote (via VPN for example)
- Need to build your ETL container image
- Good re-usability
- Need to provision and manage K8s clusters locally

ECSOperator

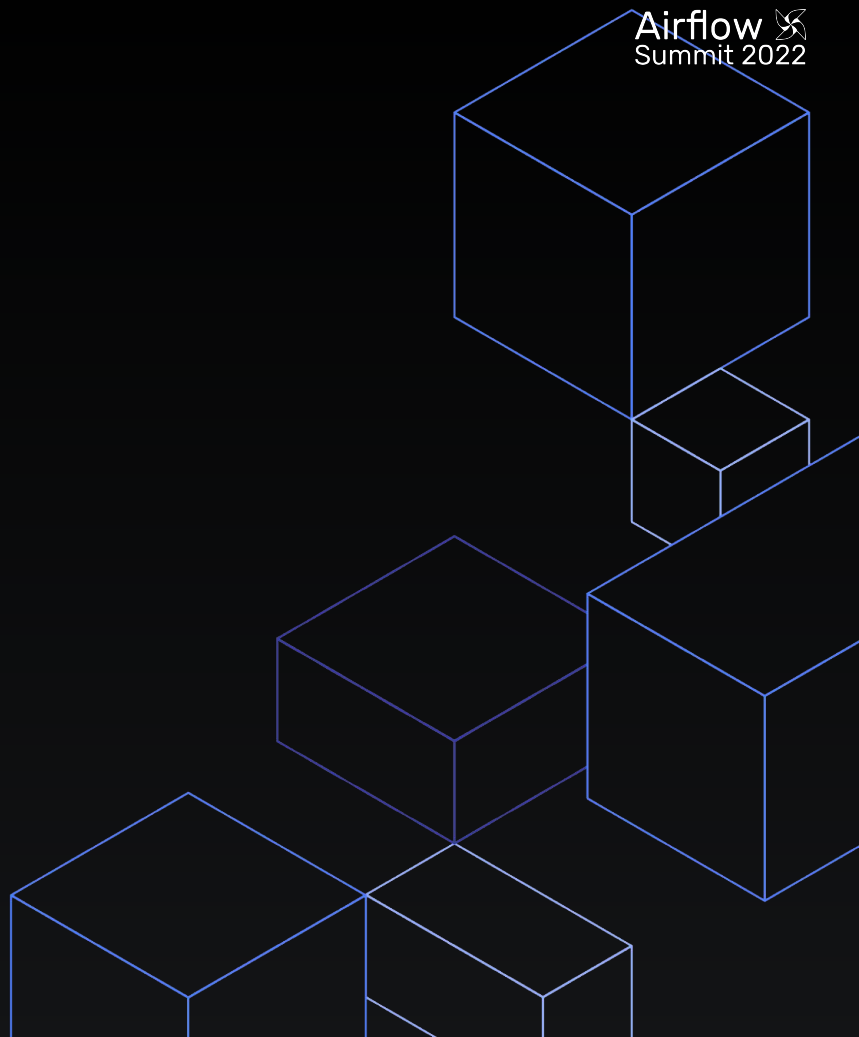
- Need to install local ECS Anywhere agent and create Amazon ECS clusters
- Not open source
- Need to build your ETL container image
- Good re-usability
- Processing of data locally

Run container on ECS Cluster

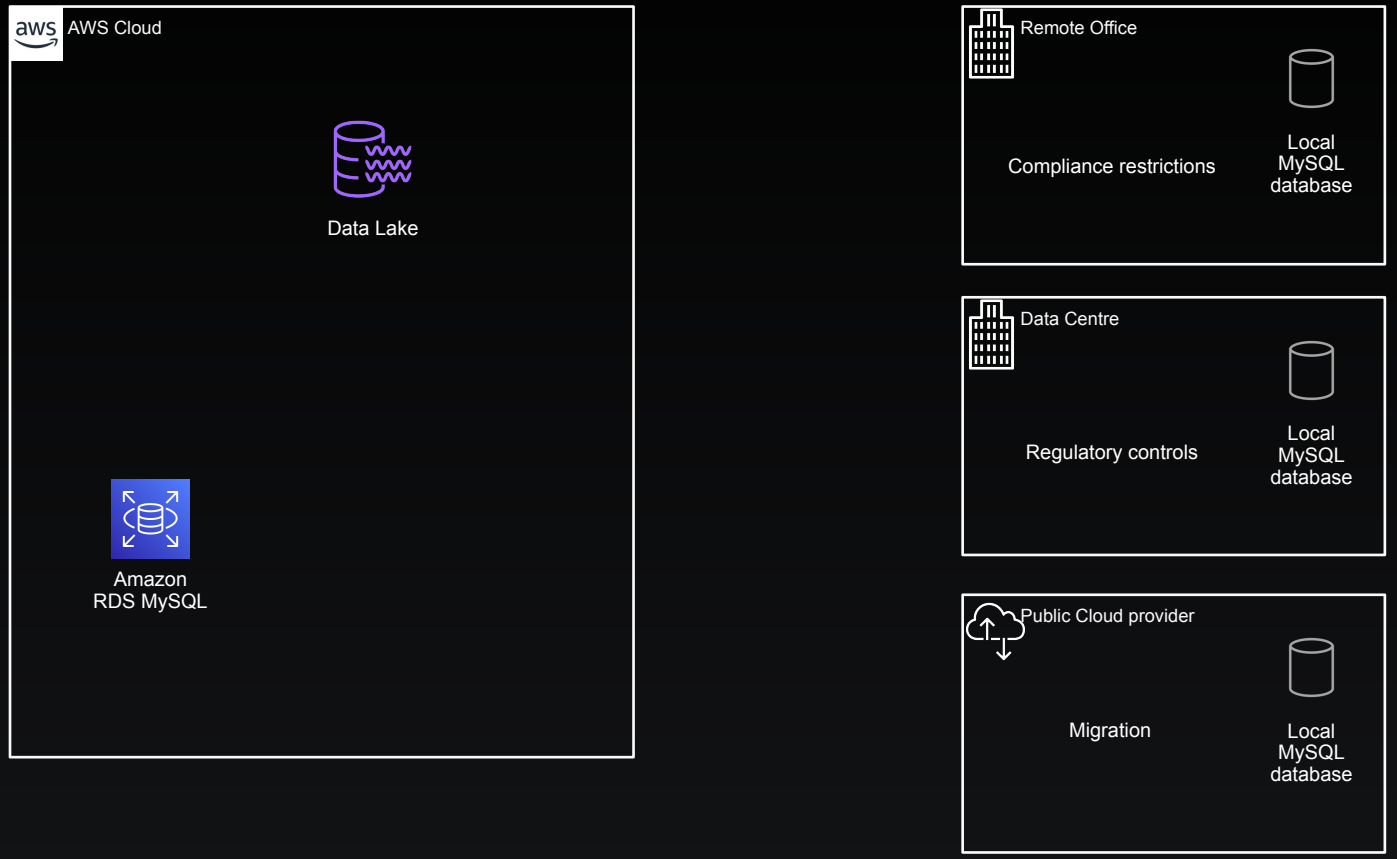


Demo

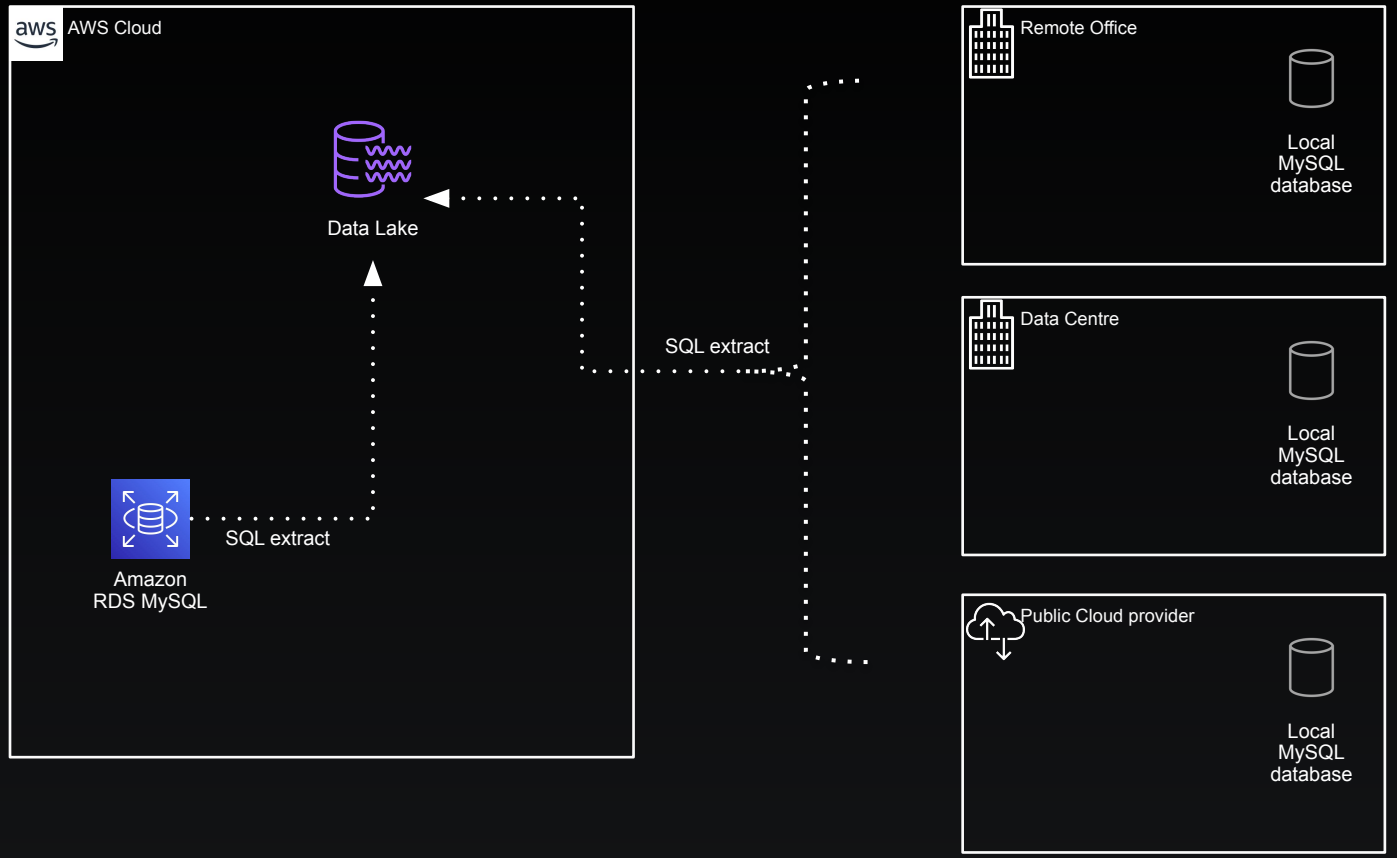
(live, pray for me!)



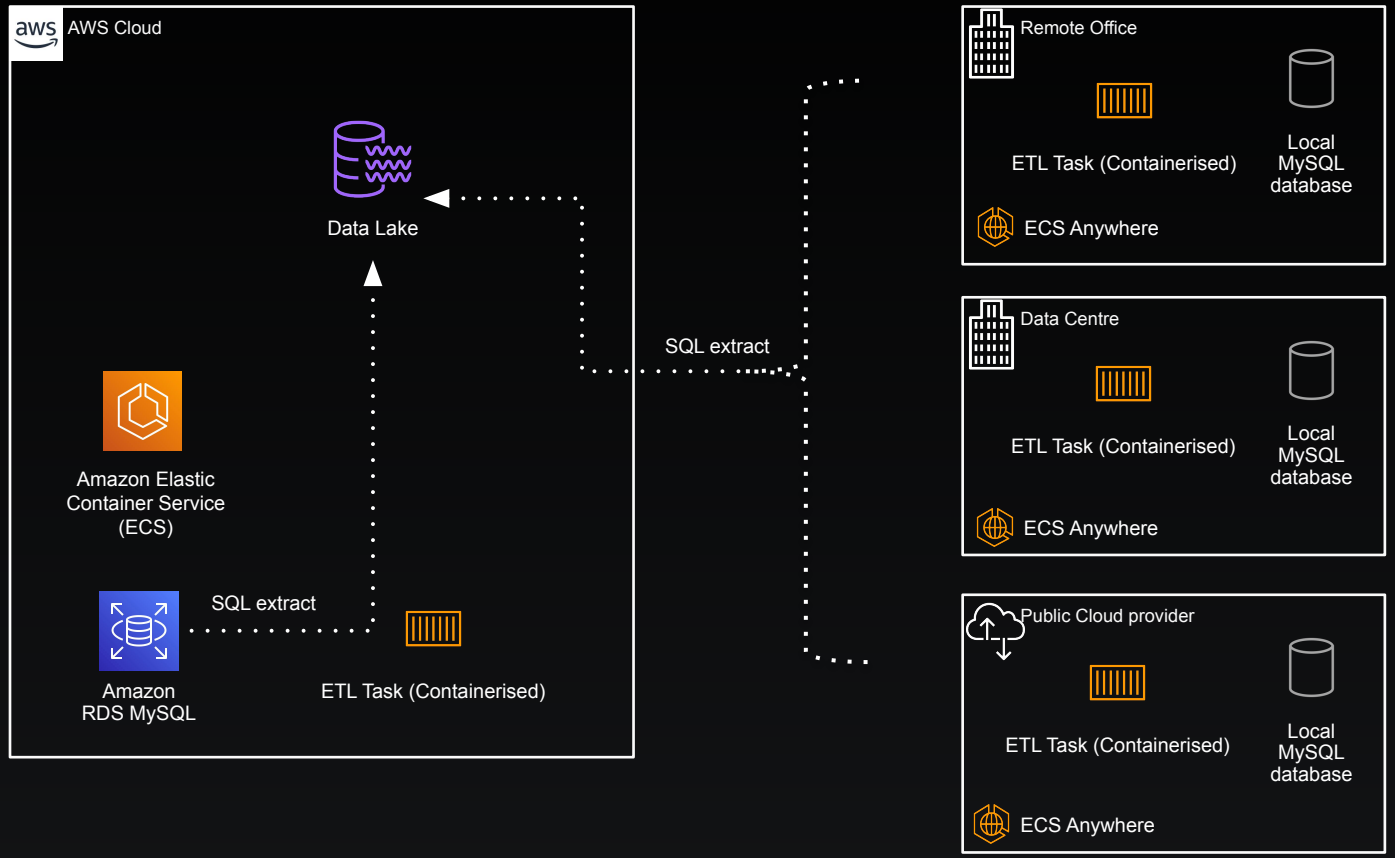
Hybrid data pipeline



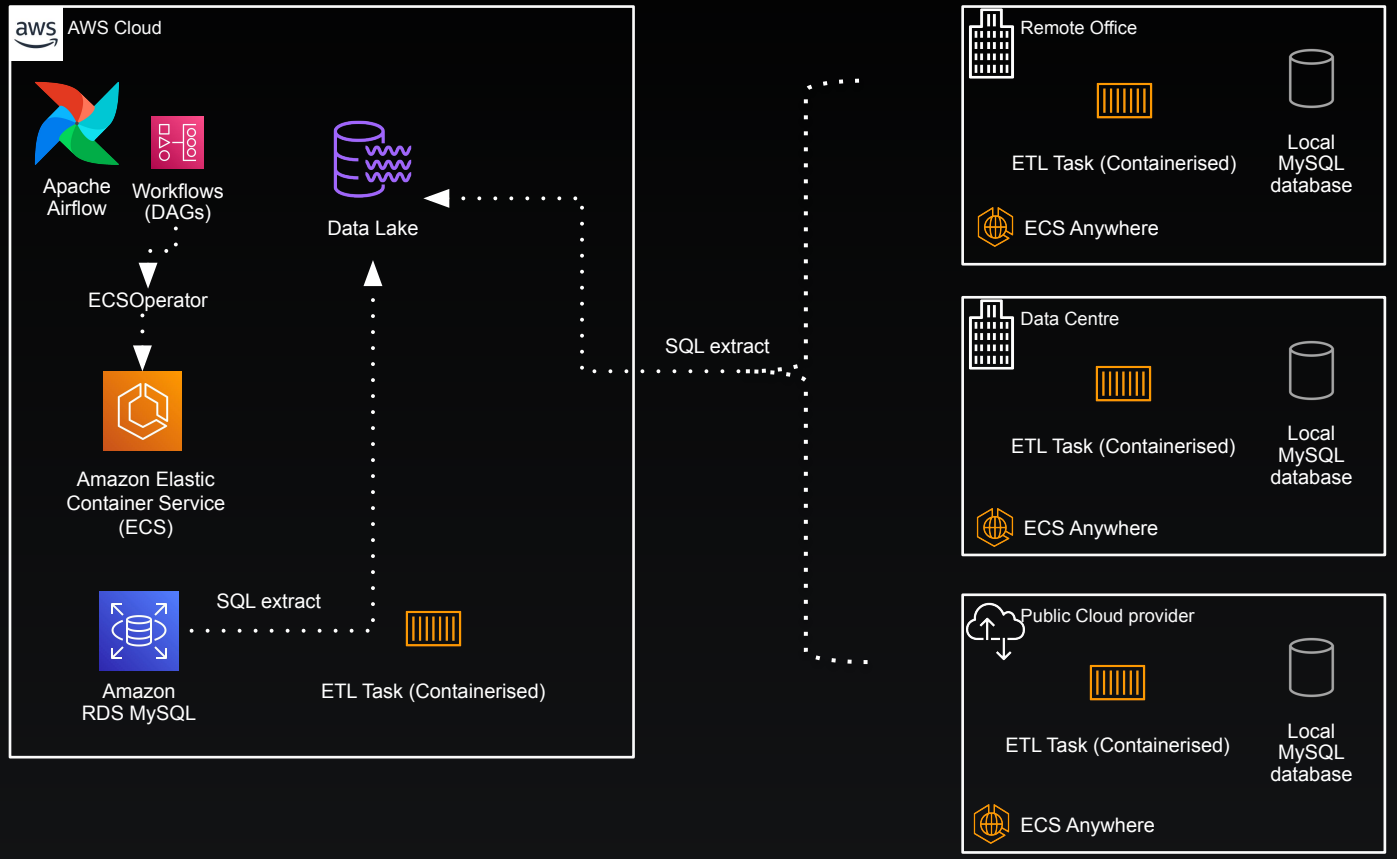
Hybrid data pipeline



Hybrid data pipeline

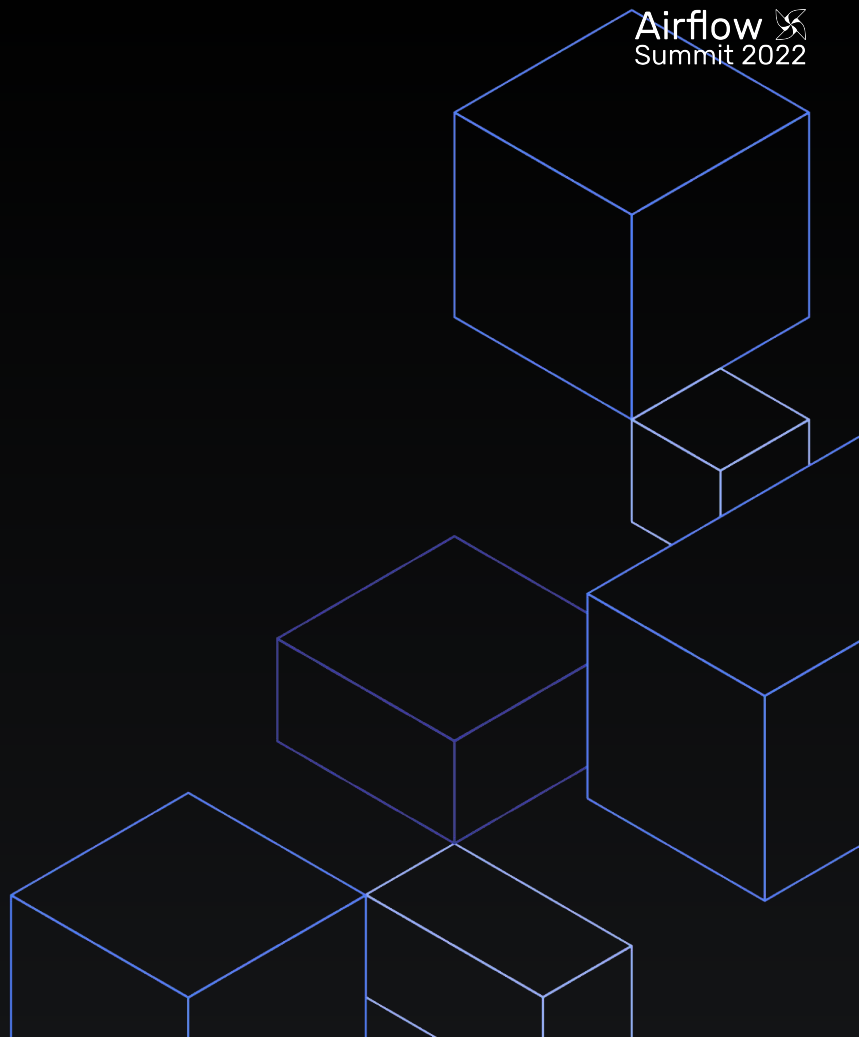


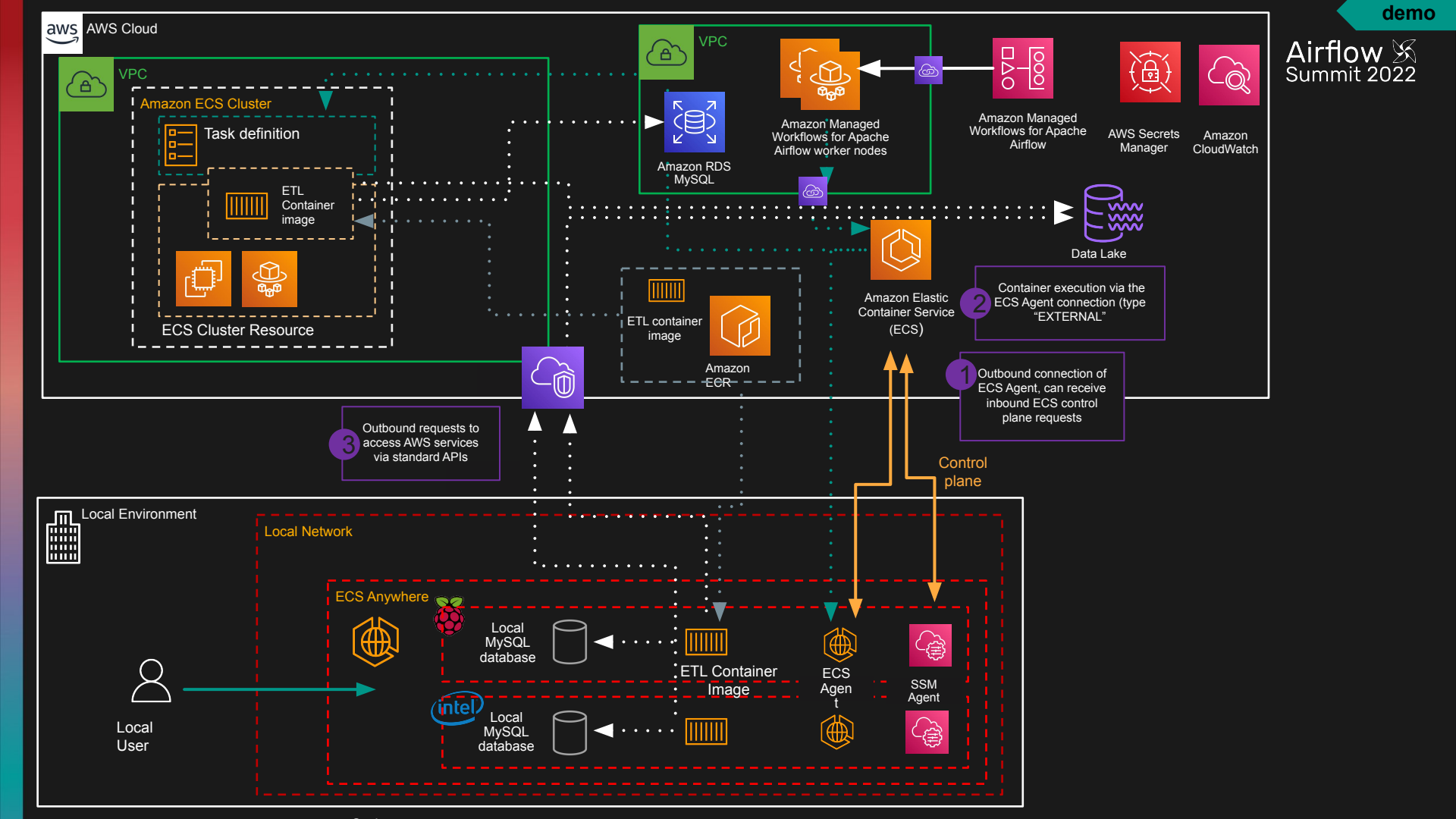
Hybrid data pipeline

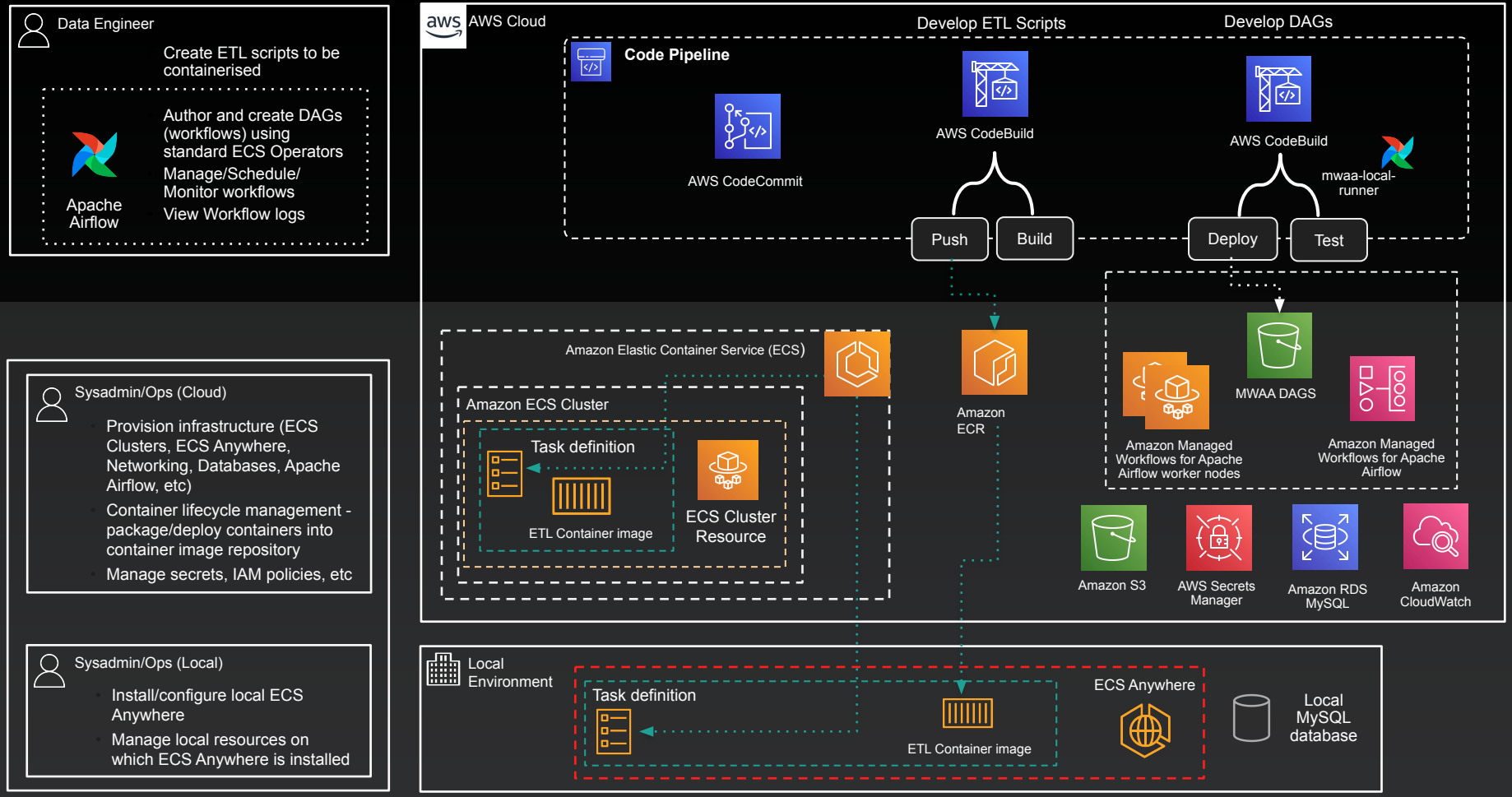


Airflow Summit 2022

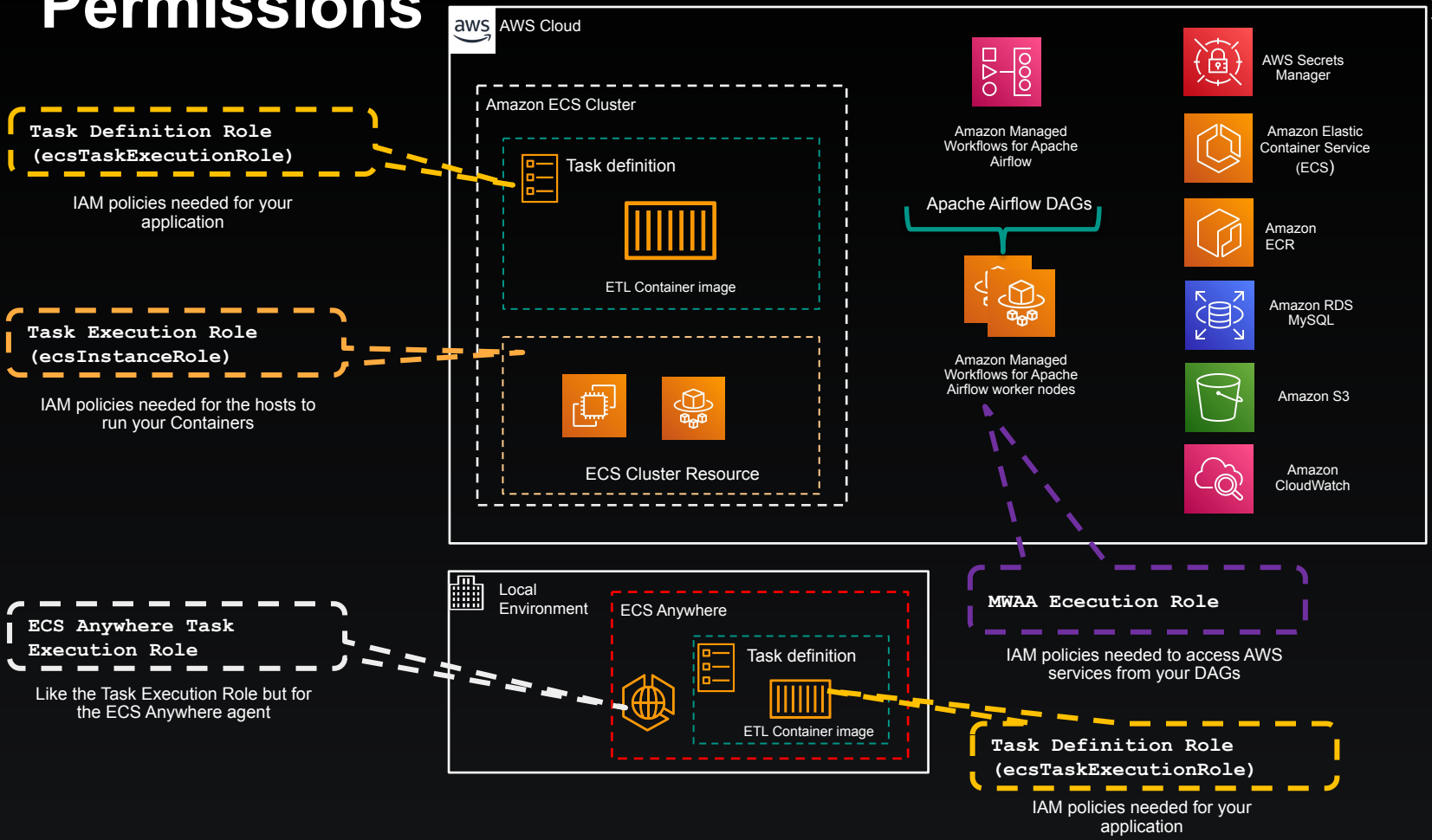
Demo




demo
Airflow Summit 2022
AWS Cloud
VPC
Amazon ECS Cluster
Task definition
ETL Container image
ECS Cluster Resource
VPC
Amazon RDS MySQL
Amazon Managed Workflows for Apache Airflow worker nodes
Amazon Managed Workflows for Apache Airflow
AWS Secrets Manager
Amazon CloudWatch
Data Lake
ETL container image
Amazon ECR
Amazon Elastic Container Service (ECS)
2 Container execution via the ECS Agent connection (type "EXTERNAL"
1 Outbound connection of ECS Agent, can receive inbound ECS control plane requests
3 Outbound requests to access AWS services via standard APIs
Control plane
Local Environment
Local Network
ECS Anywhere
Local MySQL database
Local MySQL database
ETL Container Image
ECS Agent
SSM Agent
Local User


Data Engineer
Create ETL scripts to be containerised
Author and create DAGs (workflows) using standard ECS Operators
Manage/Schedule/ Monitor workflows
View Workflow logs
Apache Airflow
Sysadmin/Ops (Cloud)
Provision infrastructure (ECS Clusters, ECS Anywhere, Networking, Databases, Apache Airflow, etc)
Container lifecycle management - package/deploy containers into container image repository
Manage secrets, IAM policies, etc
Sysadmin/Ops (Local)
Install/configure local ECS Anywhere
Manage local resources on which ECS Anywhere is installed
aws
AWS Cloud
Develop ETL Scripts
Develop DAGs
Code Pipeline
AWS CodeCommit
AWS CodeBuild
AWS CodeBuild
mwaa-local-runner
Push
Build
Deploy
Test
Amazon Elastic Container Service (ECS)
Amazon ECS Cluster
Task definition
ETL Container image
ECS Cluster Resource
Amazon ECR
MWAA DAGS
Amazon Managed Workflows for Apache Airflow worker nodes
Amazon Managed Workflows for Apache Airflow
Amazon S3
AWS Secrets Manager
Amazon RDS MySQL
Amazon CloudWatch
Local Environment
Task definition
ETL Container image
ECS Anywhere
Local MySQL database

Permissions



https://github.com/094459/blogpost-airflow-hybrid

Thank you!

Ricardo Sueiras

https://www.linkedin.com/in/ricardosueiras

@094459