
AWS
re:Invent
NOV. 28 – DEC. 2, 2022 | LAS VEGAS, NV

BOA312

Orchestrating hybrid workflows with Apache Airflow

Ricardo Sueiras (he/him)

Developer Advocate
Amazon Web Services

Fernando Gamero (he/him)

Sr. Solutions Architect
Amazon Web Services

© 2022, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Apache Airflow

© 2022, Amazon Web Services, Inc. or its affiliates. All rights reserved.



Regulation and compliance



Heritage systems

© 2022, Amazon Web Services, Inc. or its affiliates. All rights reserved.

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

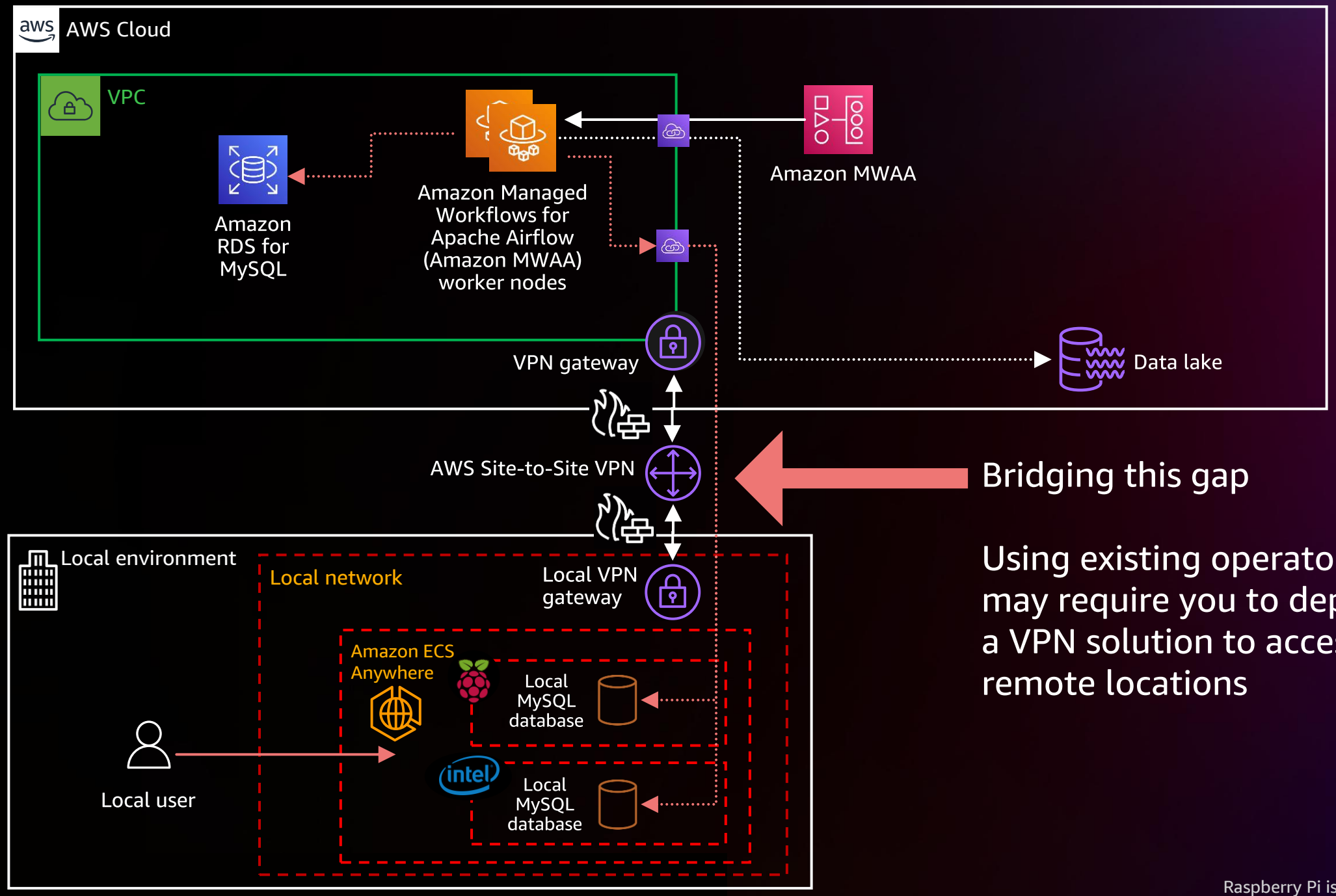
PythonOperator

- 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 (Amazon Athena)
- Create Athena federated queries (running as AWS Lambda functions)
- Additional work required to enable connectivity (inbound/controlled networks)

aws
© 2022, Amazon Web Services, Inc. or its affiliates. All rights reserved.



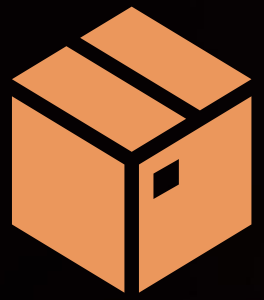
Bridging this gap

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

Raspberry Pi is a trademark of Raspberry Pi Ltd.



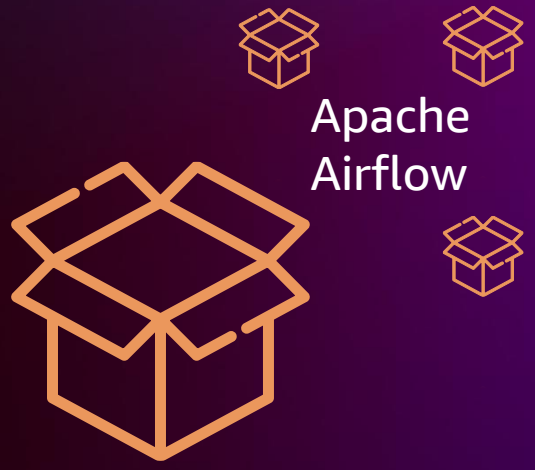
Develop ETL application



Package as container



Push to container repository



Orchestrate container

aws

© 2022, Amazon Web Services, Inc. or its affiliates. All rights reserved.

KubernetesOperator

- Need to integrate networks between cloud and remote (via VPN, for example)
- Need to build your ETL container image
- Good reusability
- Need to provision and manage Kubernetes clusters locally

ECSOperator

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

aws
© 2022, Amazon Web Services, Inc. or its affiliates. All rights reserved.



Run container on Amazon ECS cluster



Run containers in the cloud

Run containers locally



AWS Outposts



Amazon ECS Anywhere

- Outbound connectivity
- Local installation requirements
  - Docker engine
  - AWS Systems Manager Agent
  - Amazon ECS container agent
- Supported Linux distributions
  - Ubuntu 18.04, 20.04
  - RHEL/CentOS 7/8
  - Fedora, Debian, openSUSE
- Windows

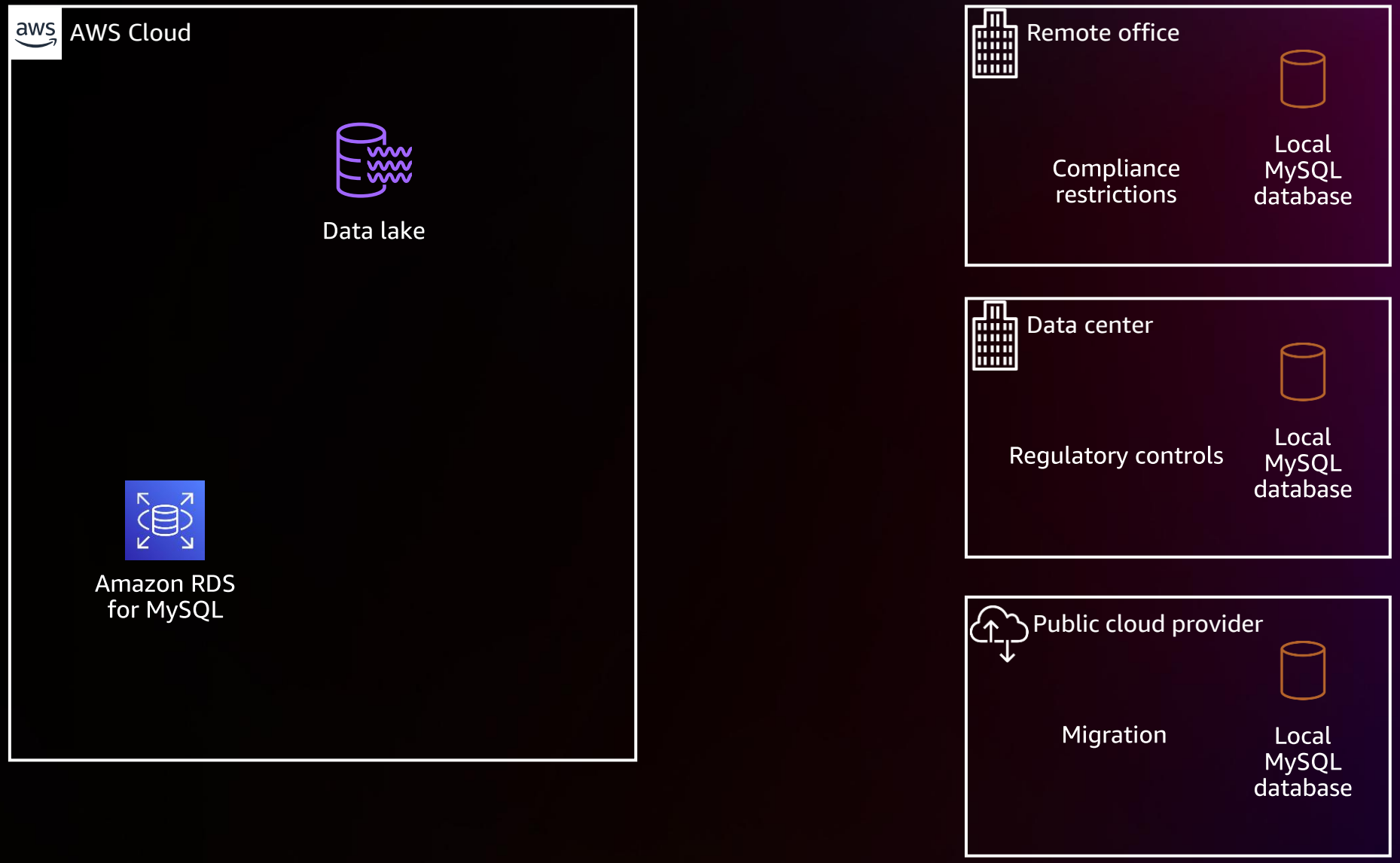
New!

Outbound network connectivity

aws

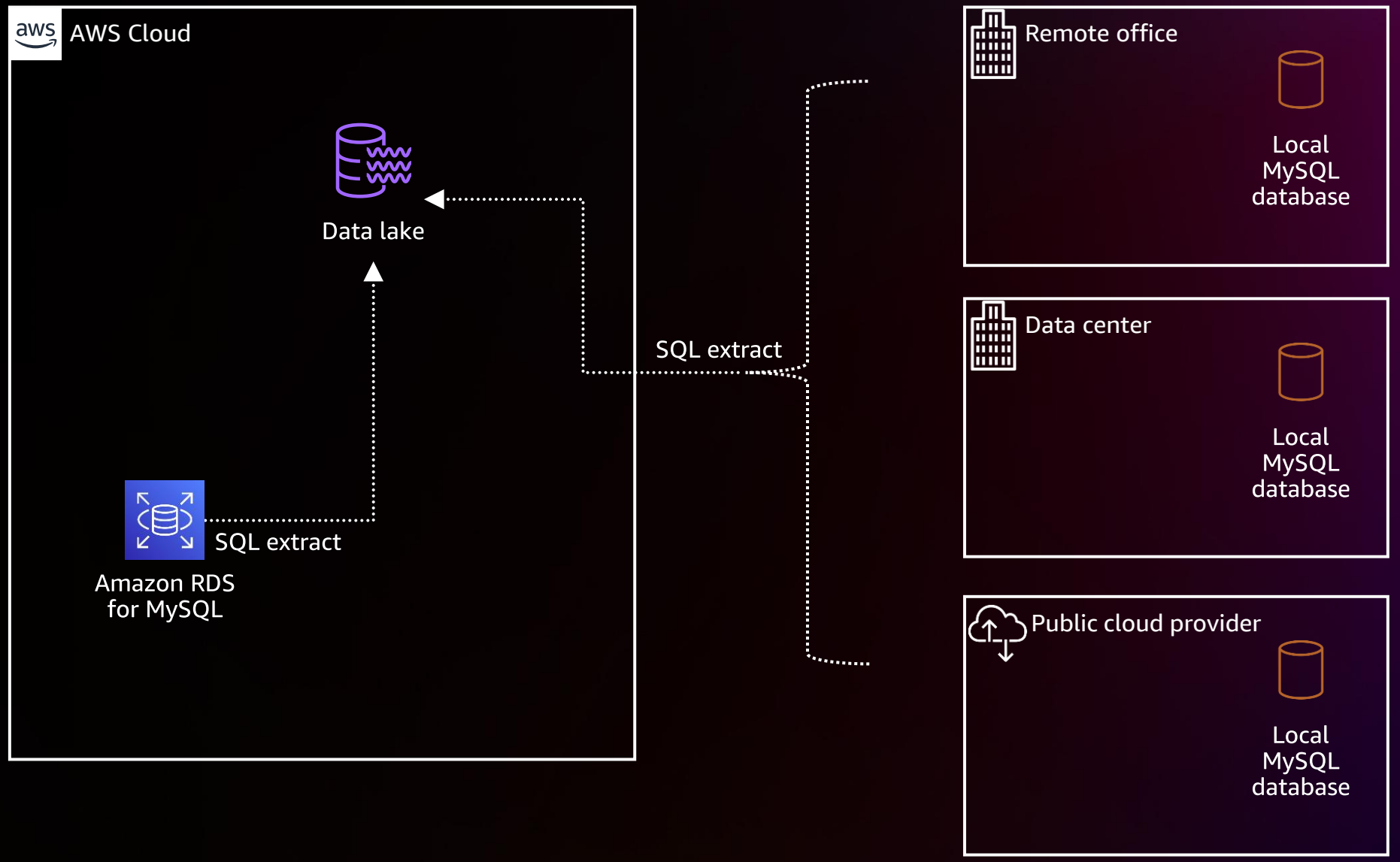
© 2022, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Hybrid data pipeline



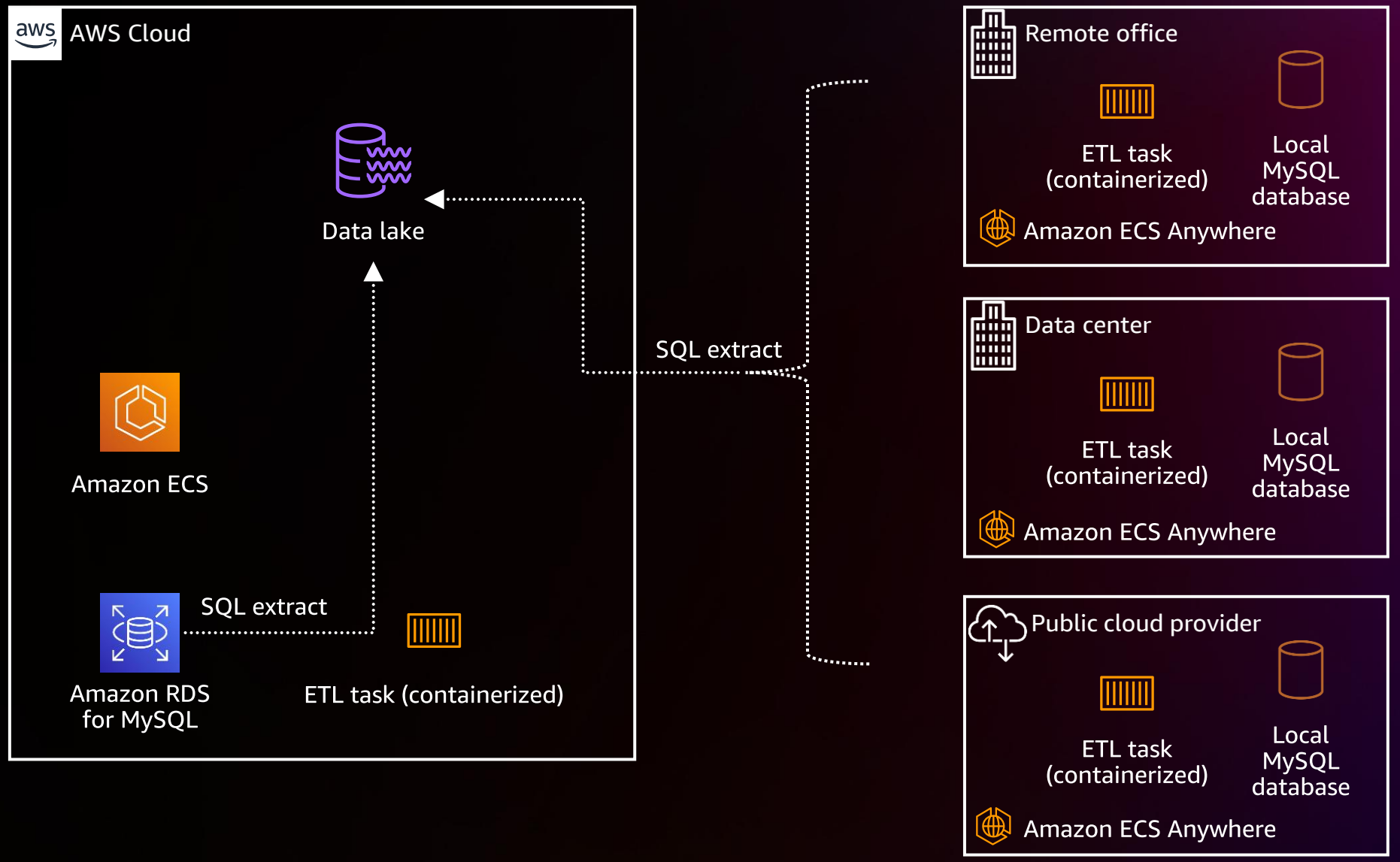
© 2022, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Hybrid data pipeline



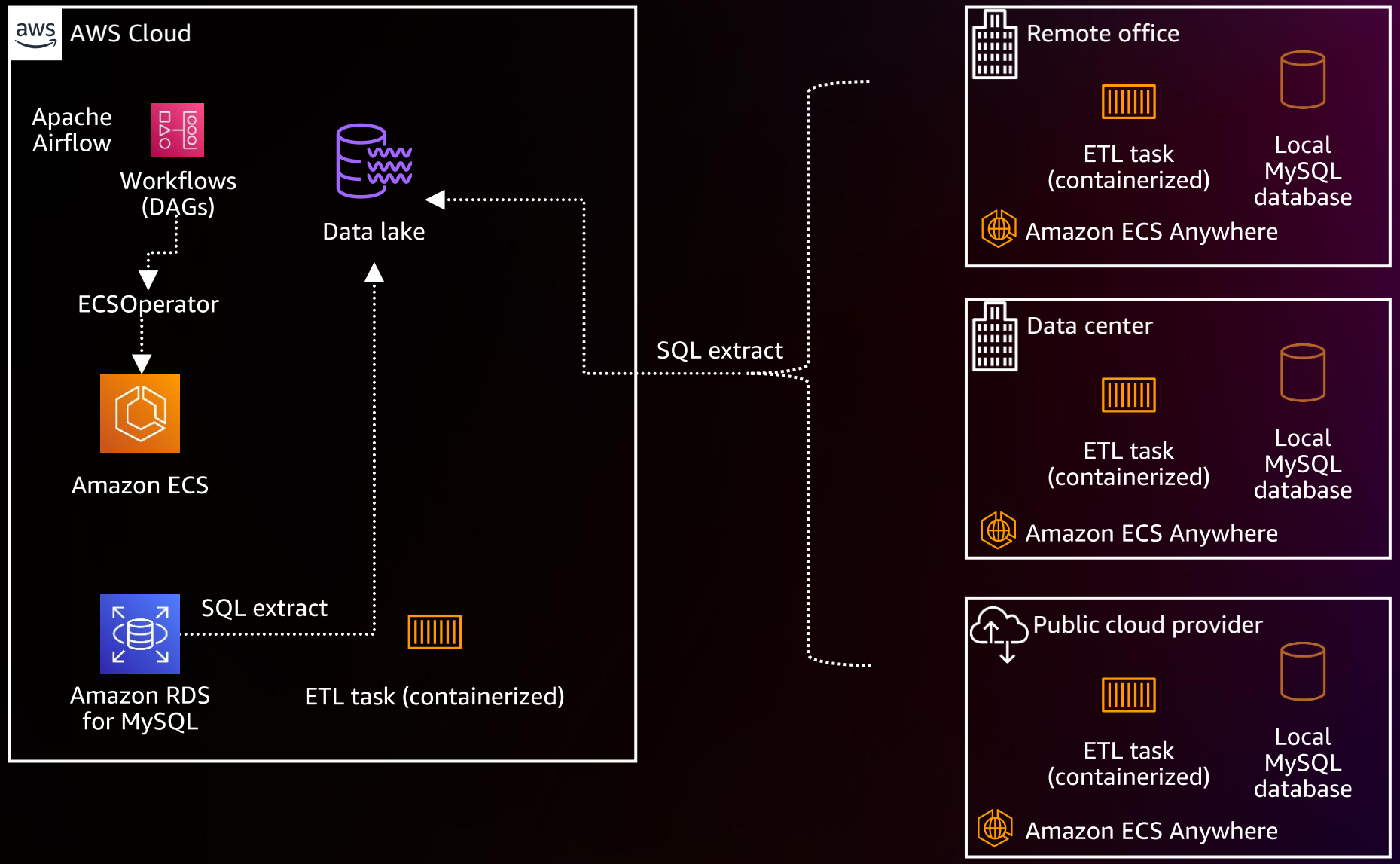
© 2022, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Hybrid data pipeline

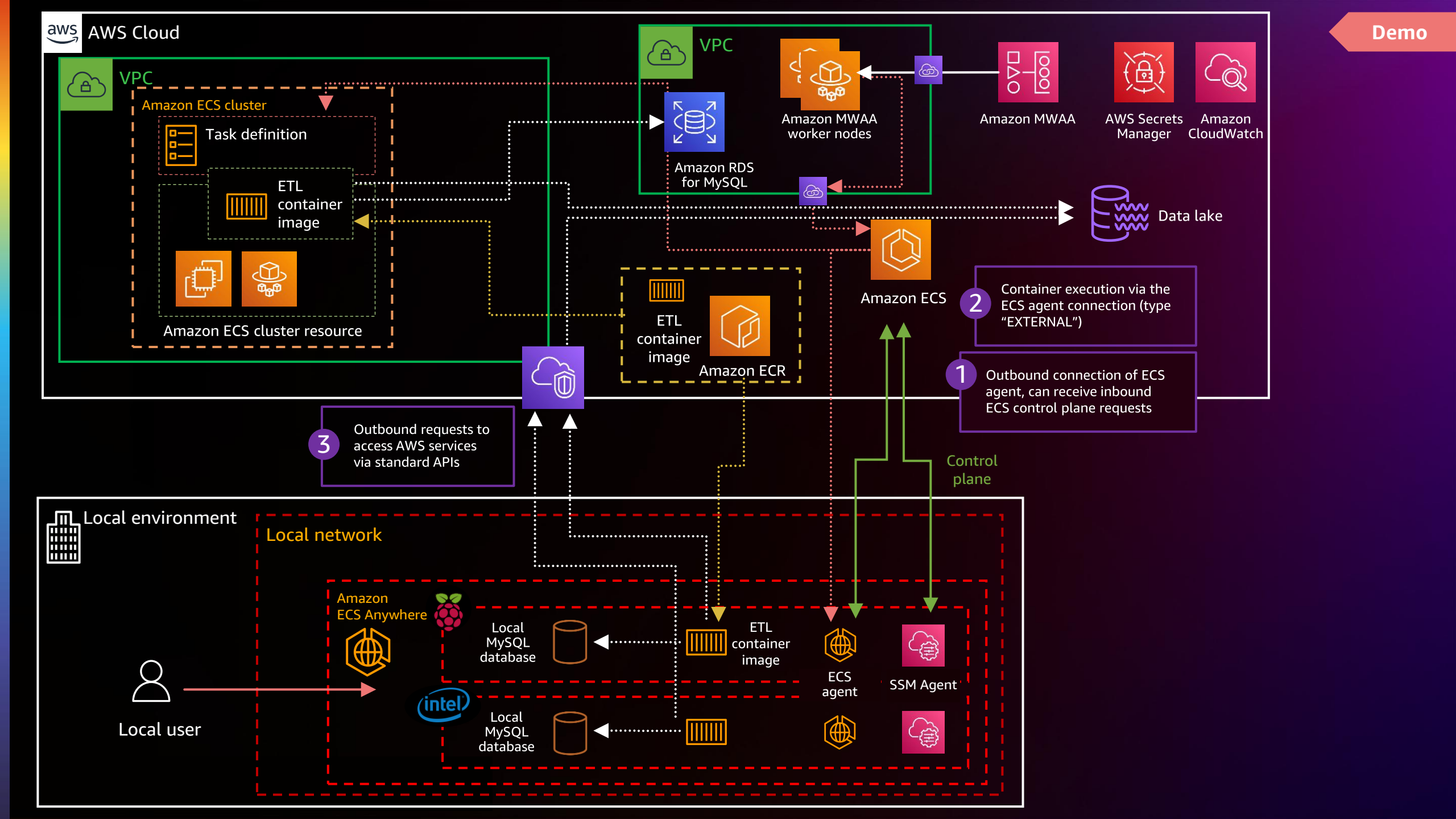


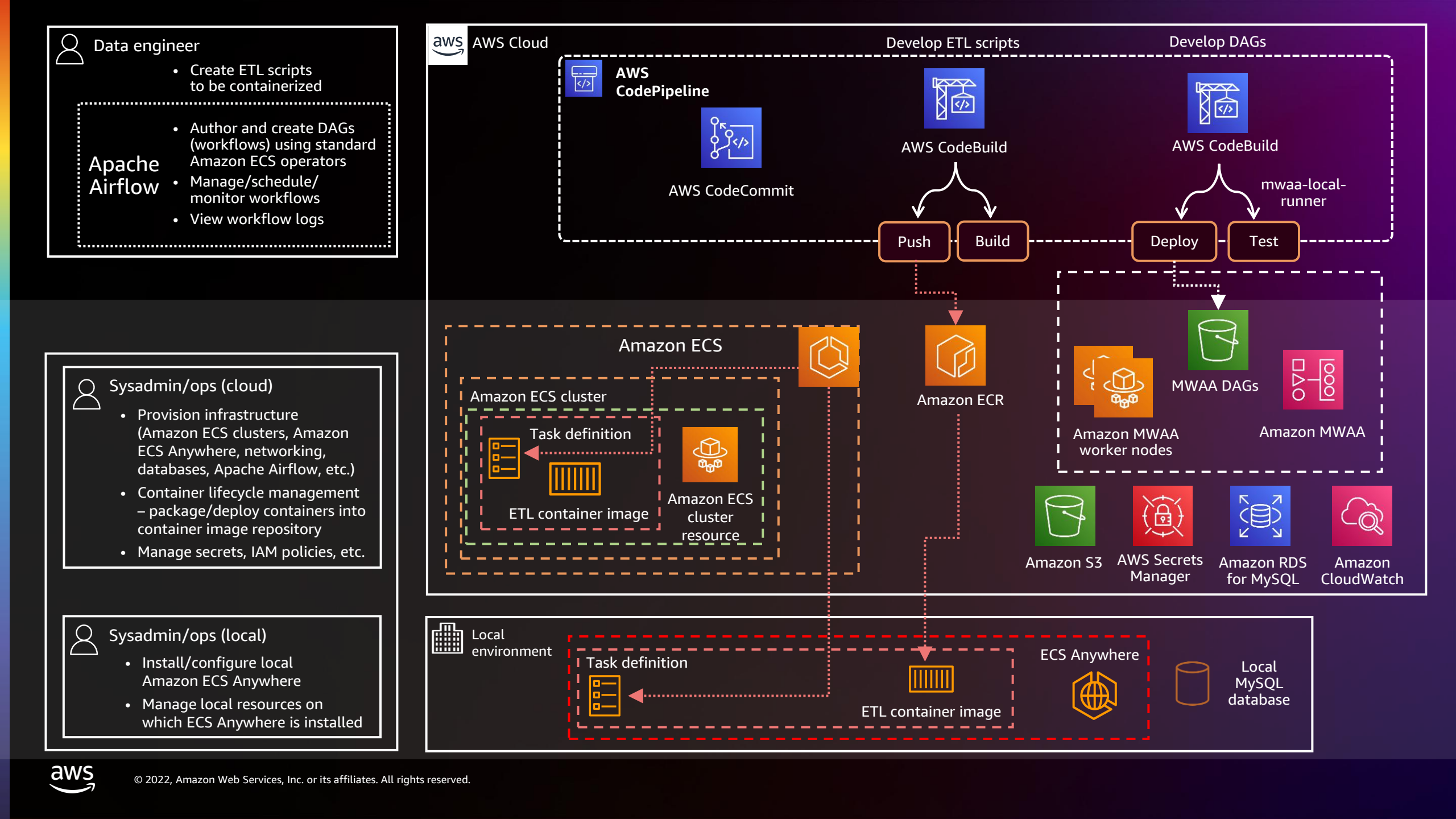
© 2022, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Hybrid data pipeline



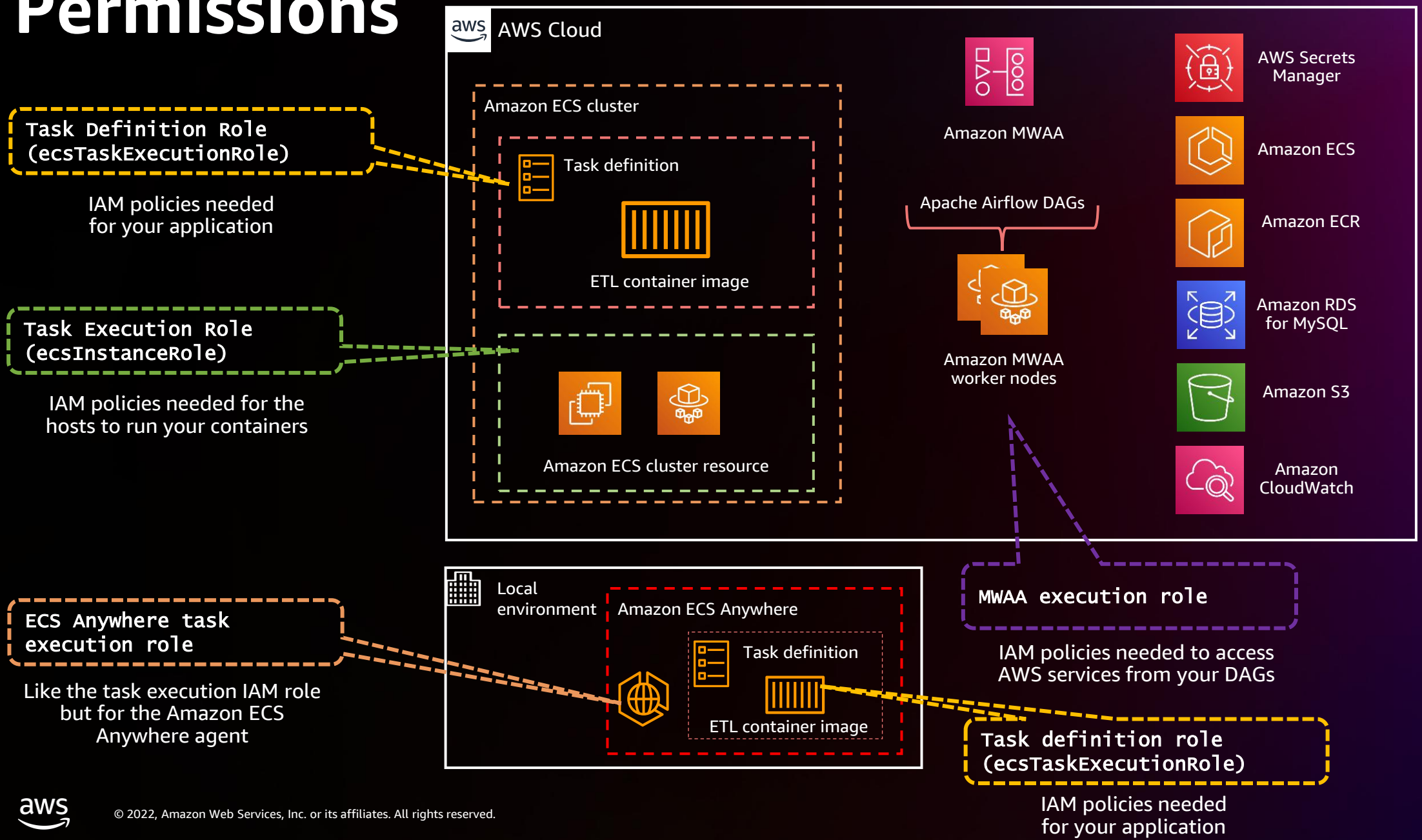
© 2022, Amazon Web Services, Inc. or its affiliates. All rights reserved.


Demo
AWS Cloud
VPC
Amazon ECS cluster
Task definition
ETL container image
Amazon ECS cluster resource
VPC
Amazon RDS for MySQL
Amazon MWAA worker nodes
Amazon MWAA
AWS Secrets Manager
Amazon CloudWatch
Data lake
Amazon ECS
ETL container image
Amazon ECR
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
Amazon ECS Anywhere
Local MySQL database
ETL container image
ECS agent
SSM Agent
Local MySQL database
Local user



© 2022, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Permissions



Task Definition Role (ecsTaskExecutionRole)

IAM policies needed for your application

Task Execution Role (ecsInstanceRole)

IAM policies needed for the hosts to run your containers

ECS Anywhere task execution role

Like the task execution IAM role but for the Amazon ECS Anywhere agent

MWAA execution role

IAM policies needed to access AWS services from your DAGs

Task definition role (ecsTaskExecutionRole)

IAM policies needed for your application

© 2022, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Thank you!



Please complete the session survey in the **mobile app**

© 2022, Amazon Web Services, Inc. or its affiliates. All rights reserved.