INNOVATE

DATA EDITION

23 August, 2022

Building scalable data pipelines with Amazon Managed Workflow for Apache Airflow and Amazon Redshift

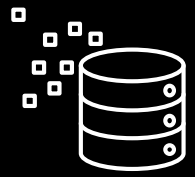
Praveen Kumar
Analytics Solutions Architect
Amazon Web Services

Agenda

- Overview of AWS Modern Data Platform
- Layered architecture and common data pipeline tasks
- Deep dive on Amazon Managed workflow for Apache airflow (MWAA) and Amazon Redshift
- Demo
- Summary and next steps

Customers want more value from their data

Growing Exponentially



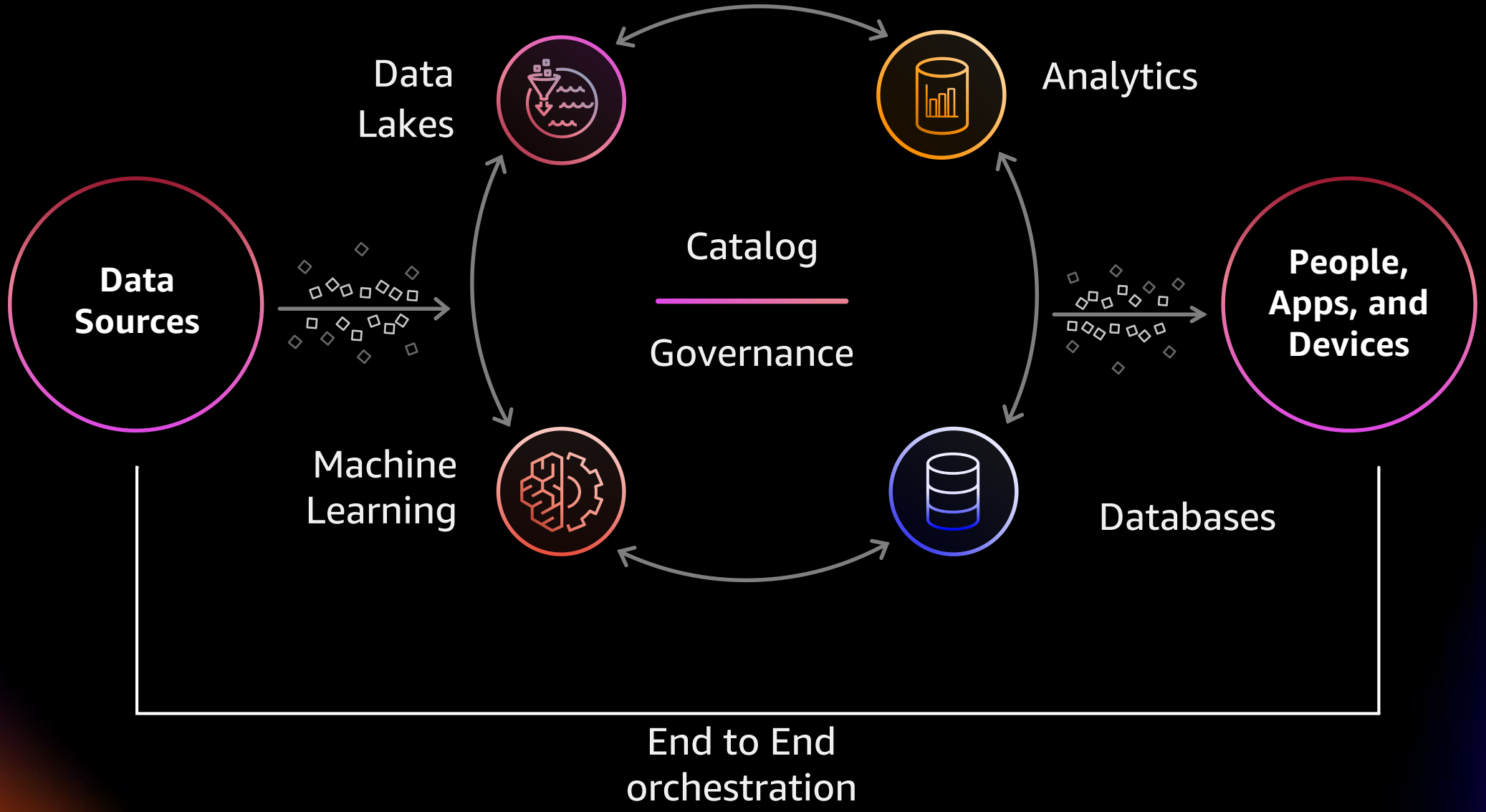
From new sources

Increasingly diverse

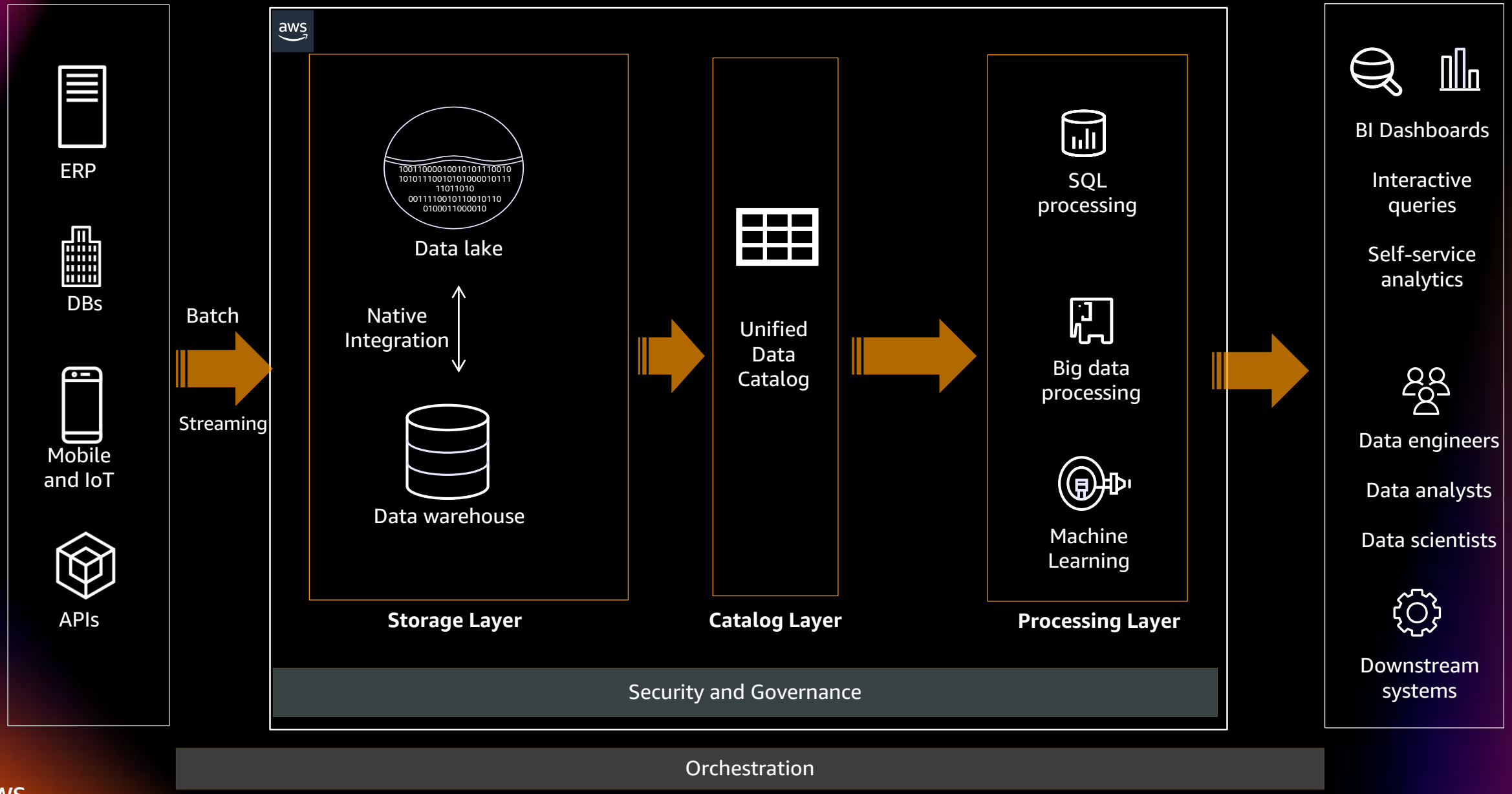
Used by many people

Analyzed by many applications

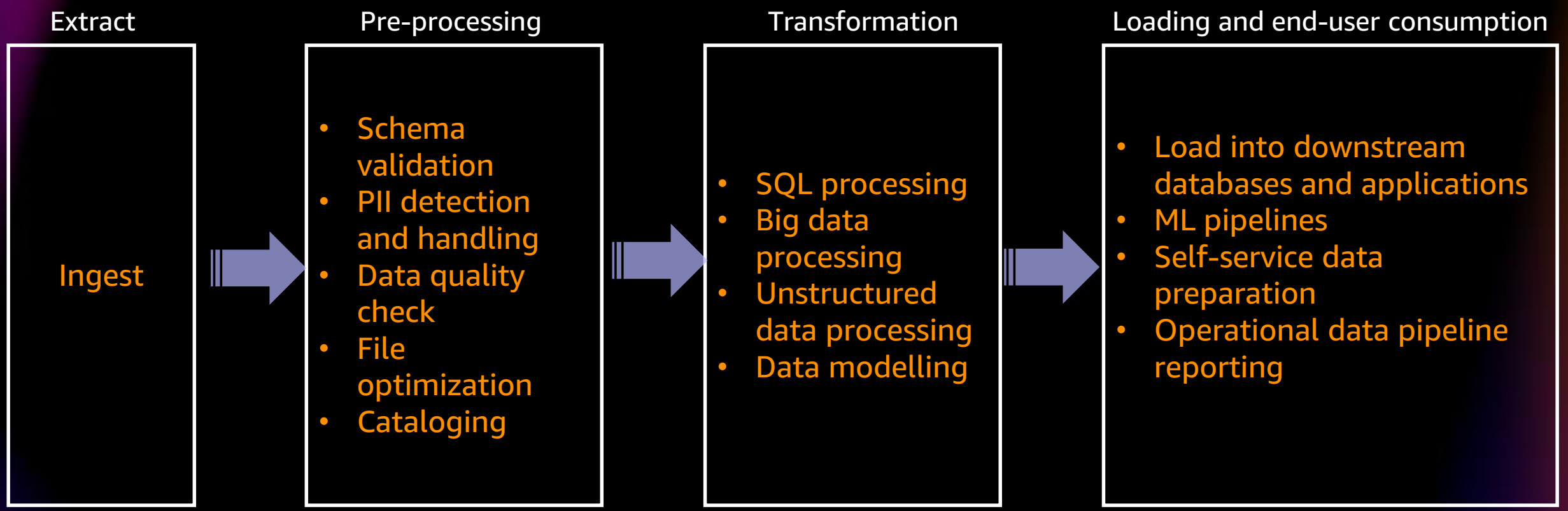
AWS Modern Data Platform



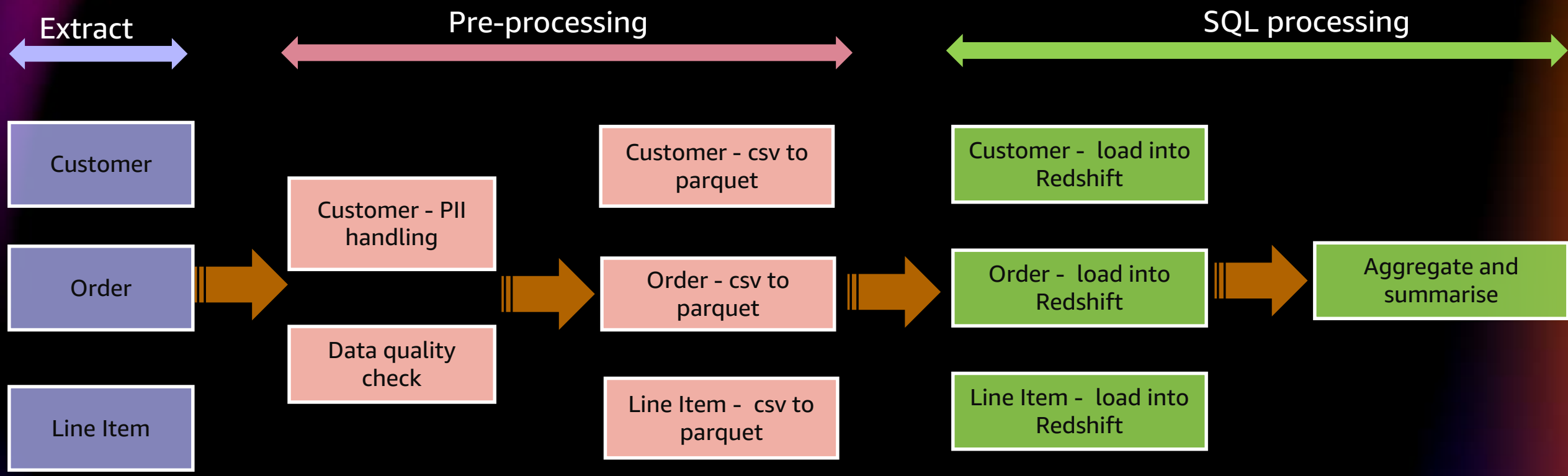
Modern data platform layers



Common tasks in a data pipeline



A sample data pipeline (SQL processing)

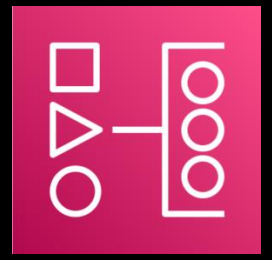


Orchestration

SQL processing with pipeline orchestration

Amazon Redshift

SQL processing
BI and DWH workloads

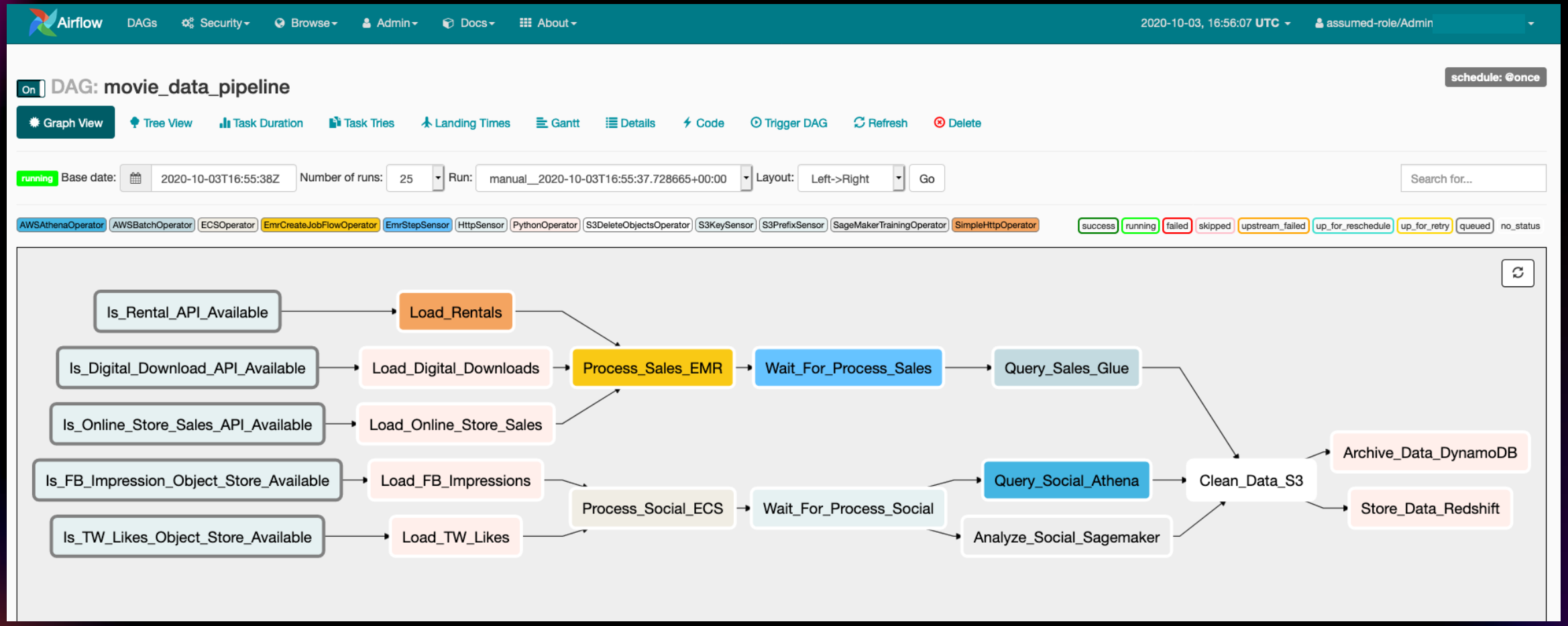


Amazon MWAA

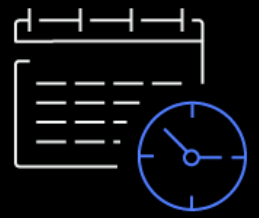
Managed service for Apache Airflow
End-to-end orchestration

Amazon Managed Workflow for Apache Airflow (Amazon MWAA)

What is Apache Airflow?



Apache Airflow components



Scheduler

Worker

Web Server

Meta Database

Apache Airflow key concepts

DAG

Collections of tasks that describes how to run a workflow

Tasks

A task defines a unit of work and is represented as node in the DAG graph

Operators

Atomic components in a DAG describing a single task in the pipeline

Sensors

Type of operator that waits on some external or internal trigger

Hooks

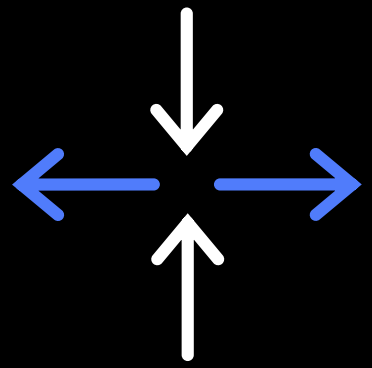
Interface to access external services like Amazon S3, Amazon Redshift, AWS Glue etc

Scheduling

Includes running on demand or at a certain frequency

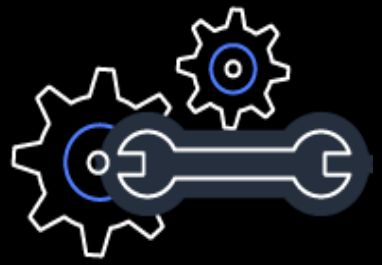
Challenges with self-managed Apache Airflow

Setup



Scaling

Security



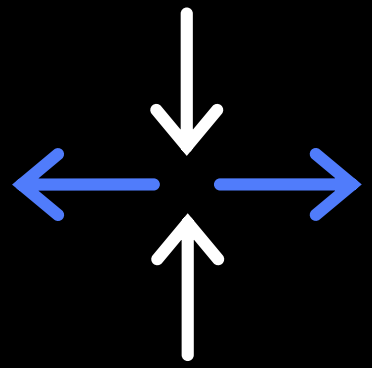
Upgrades

Maintenance

Solution – Amazon MWAA

Setup

- Deploy Airflow rapidly
- Same open-source Airflow

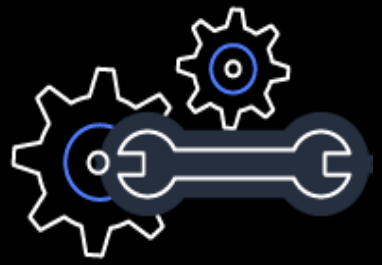


Scaling

- Seamless worker scaling
- Amazon ECS on AWS Fargate

Security

- Integrated with AWS IAM
- VPC only or public Airflow UI



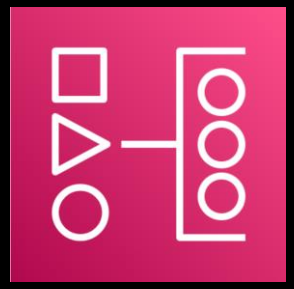
Upgrades

- Maintenance windows
- Rollback in case of failure

Maintenance

- Monitoring with Amazon CloudWatch
- Multi-AZ

How does Amazon MWAA work?



Create an Amazon MWAA Environment

Upload Airflow DAG to Amazon Simplified Storage (Amazon S3)

Access the Airflow UI

The Pokémon Company uses MWAA to simply security controls

"With Amazon MWAA, we can focus on building reliable data pipelines that achieve business goals rather than patching and securing instances." Eric Smith, Data Platform Engineer

The Pokémon Company INTERNATIONAL

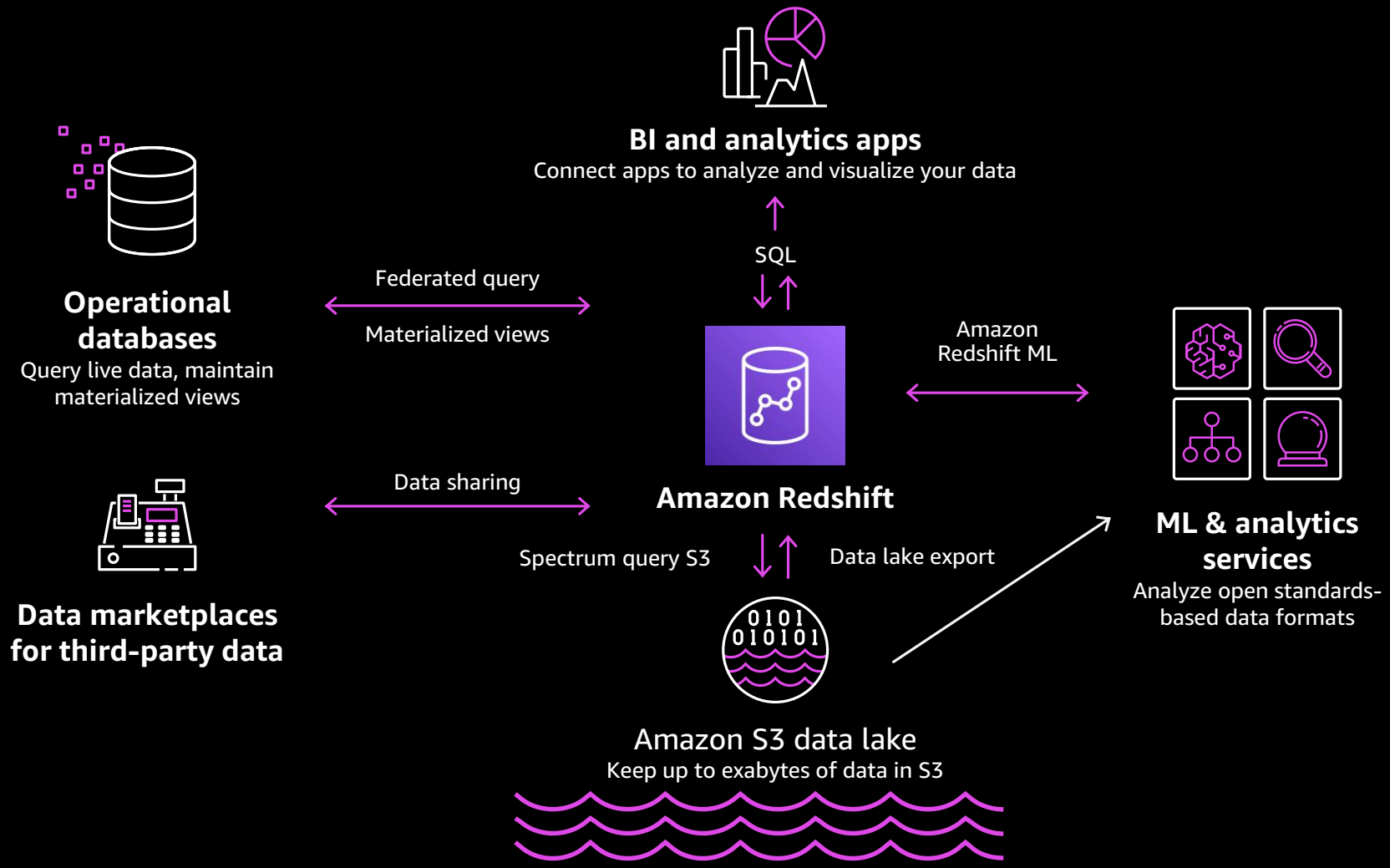
https://press.aboutamazon.com/news-releases/news-release-details/aws-announces-general-availability-amazon-managed-workflows

aws

Amazon Redshift

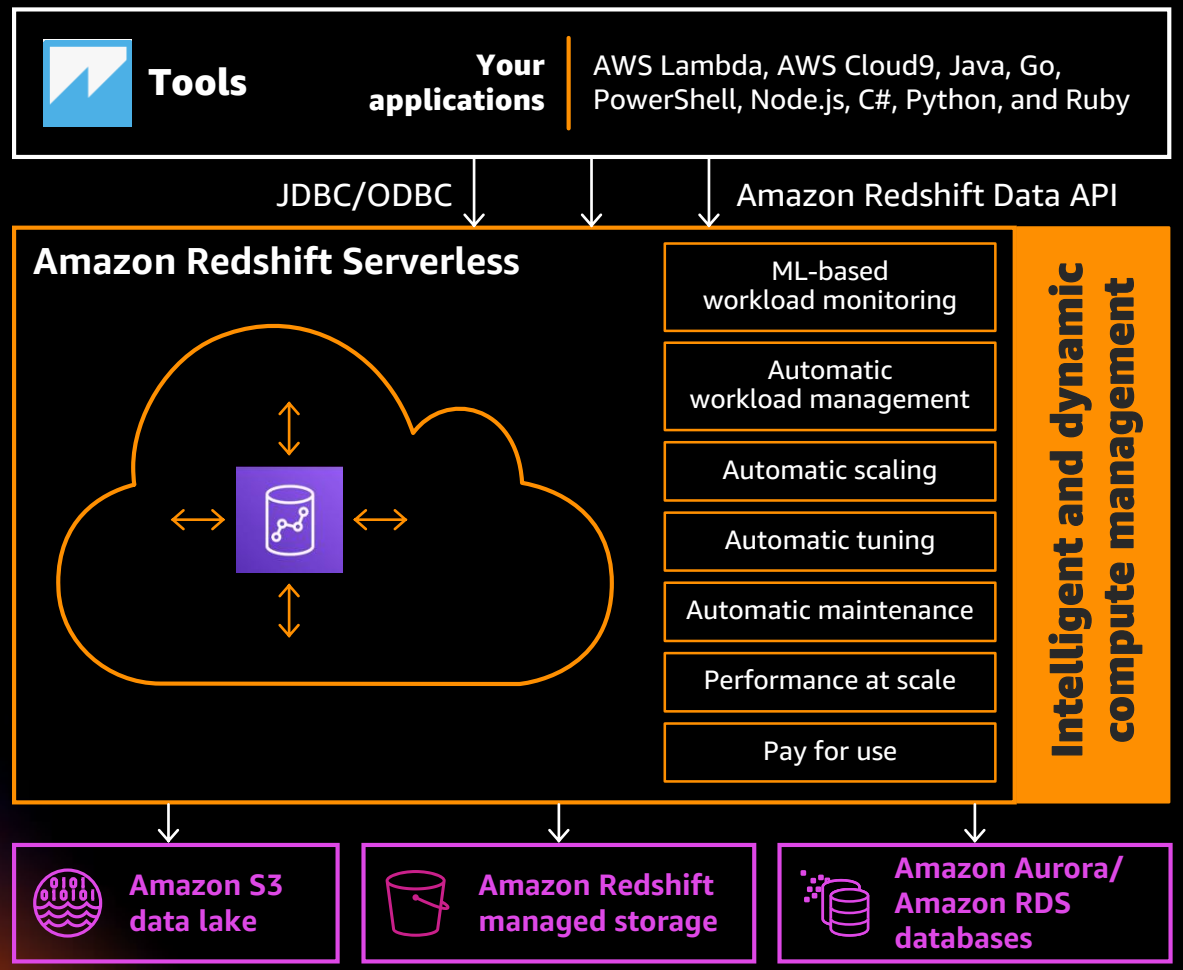
Analyze all your data

WITH AWS SERVICES INTEGRATION



Amazon Redshift Serverless

A NEW SERVERLESS OPTION FOR AMAZON REDSHIFT



Simply point applications to the Amazon Redshift Serverless endpoint and start running

All Amazon Redshift SQL functionality applies

- Security and user management
- Complex joins
- Semi-structured data
- Data sharing
- Machine learning functions
- Data lake queries
- Federated query
- Durability and transactional guarantees
- JDBC/ODBC and Data API
- And more

Amazon Redshift Data API

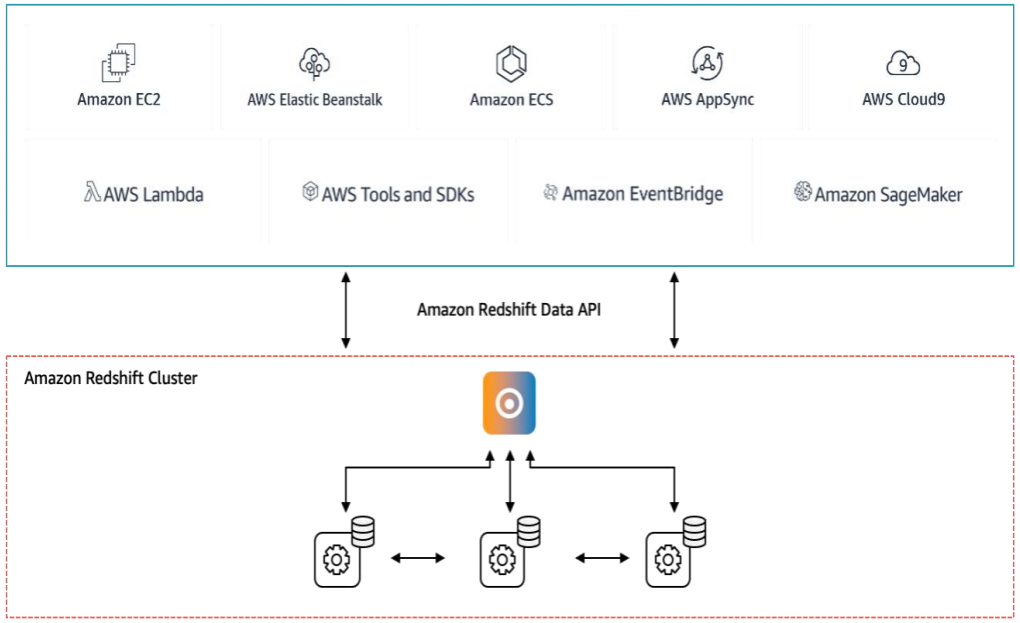
SIMPLIFIES DATA ACCESS FROM WEB SERVICES BASED APPLICATIONS

Simplifies data access from web-services based application without requiring to worry about configuring drivers, connection pools

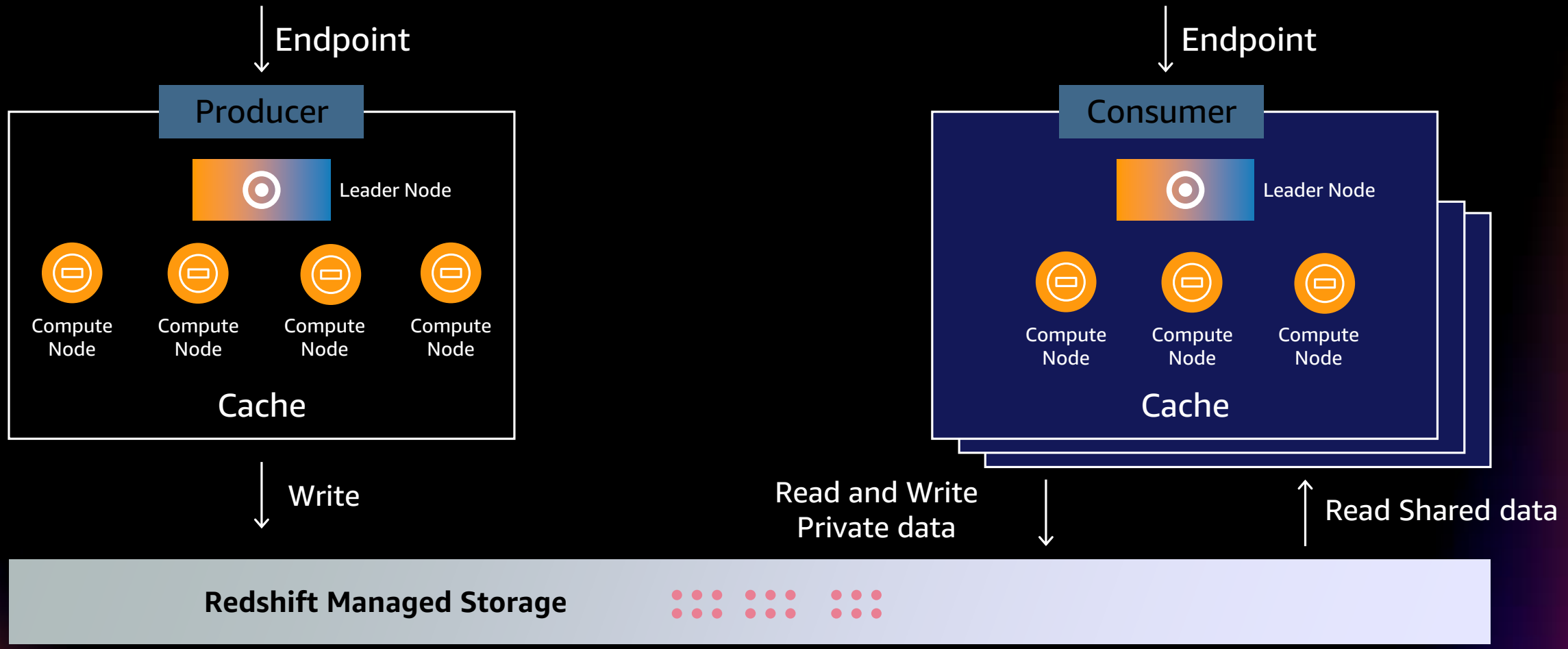
Build your ETL pipeline with AWS Lambda, AWS Step Functions

Build event-driven applications with Amazon EventBridge and AWS Lambda

Simpler access to data from data science tools such as Amazon SageMaker and Jupyter Notebooks



Amazon Redshift Data Sharing



Concurrency scaling

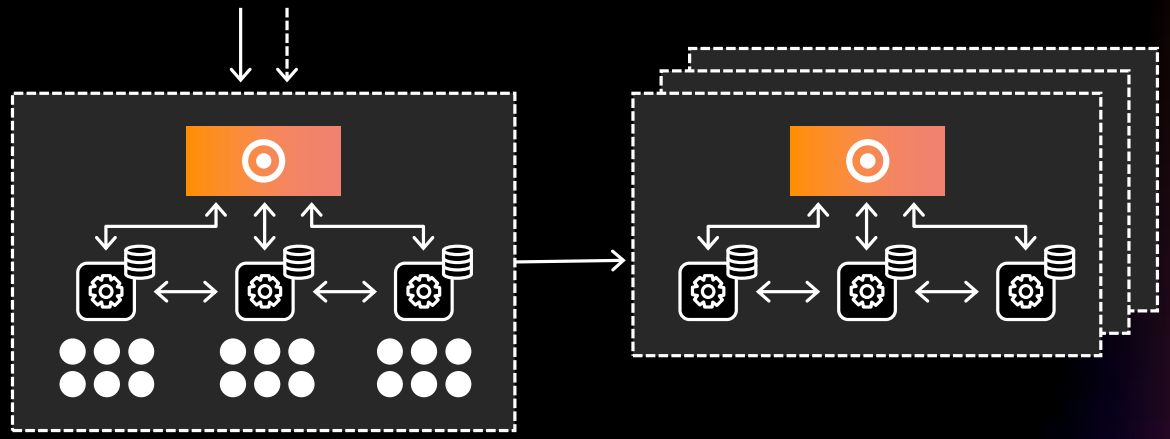
Compute elasticity and scalability to handle unpredictable user demand

Scale out to multiple Amazon Redshift clusters from a single endpoint in seconds

Support virtually unlimited concurrent users and queries while maintaining SLAs

Per-second billing for additional clusters used

Auto-scale write queries, in addition to read queries



Warner Bros. Games

“. . . At Warner Bros. Games, we build and maintain complex data mobility infrastructures to manage data movements across single game clusters and consolidated business function clusters. Using the **Redshift data sharing** feature, we can remove the entire subsystem we built for **data copying, movement, and loading** between Redshift clusters. This will empower all of our business teams to make decisions on the right datasets more **quickly and efficiently**.”

Kurt Lawson
Technical Director
Warner Bros. Analytics

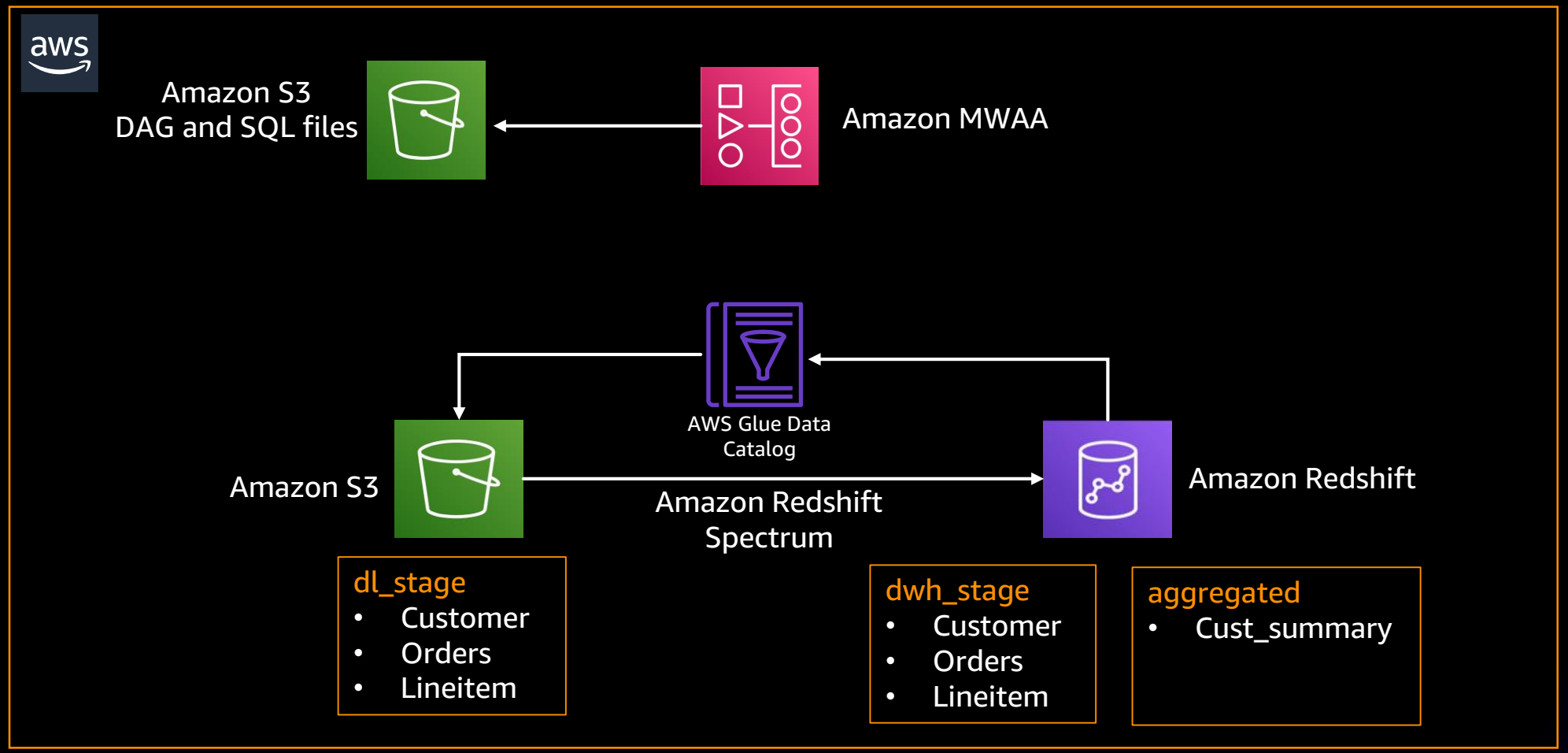
https://aws.amazon.com/redshift/features/data-sharing/

Amazon Redshift Operators for Apache Airflow

RedshiftResumeClusterOperator	Resume a paused Redshift cluster
RedshiftPauseClusterOperator	Pause an available Redshift cluster
RedshiftClusterSensor	Monitors the state of Redshift cluster until it reaches target or terminal state
RedshiftDataOperator	Executes SQL on Redshift cluster using Redshift Data API and waits for query completion
RedshiftSQLOperator	Uses Redshift Postgre connection to execute query

Demo

Demo architecture



Recap

- Build highly extensible and easy to manage pipelines using Amazon MWAA
- Develop SQL based processing and run analytics at scale using Amazon Redshift
- Keep the pipeline build flexible and focus on end-to-end

Visit the AWS Data resource hub

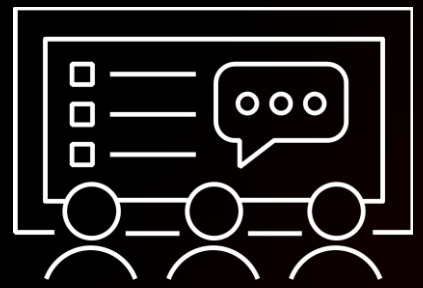
A modern data strategy can help you manage, act on, and react to your data so you can make better decisions, respond faster, and uncover new opportunities. Dive deeper with these resources today.

- Harness data to reinvent your organization
- In unpredictable times, a data strategy is key
- Make data a strategic asset
- Rewiring your culture to be data-driven
- Put your data to work with a modern analytics approach
- … and more!

https://tinyurl.com/data-hub-aws

Visit resource hub

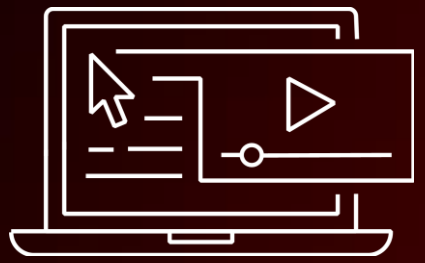
AWS Training and Certification for Data and Analytics



AWS Data & Analytics FREE Training Resources

Discover how to harness data, one of the world's most valuable resources, and innovate at scale.

https://bit.ly/3Ntlhy7



AWS Data Analytics Learning Plan

This learning plan expose you to the fastest way to get answers from all your data to all your users. It can also help prepare you for the AWS Certified Data Analytics - Specialty certification exam.

https://bit.ly/3wBVjD1

AWS Certified Data Analytics - Specialty

Earning AWS Certified Data Analytics – Specialty validates expertise in using AWS data lakes and analytics services.

https://go.aws/3lwF0RR

Thank you for attending AWS Innovate – Data Edition

We hope you found it interesting! A kind reminder to **complete the survey.**
Let us know what you thought of today's event and how we can improve the event experience for you in the future.


aws-apj-marketing@amazon.com


twitter.com/AWSCloud


facebook.com/AmazonWebServices


youtube.com/user/AmazonWebServices


slideshare.net/AmazonWebServices


twitch.tv/aws

Thank you!