Amazon EMR Studio

Accelerate Apache Spark and Other Big Data Application Development with EMR Studio

Damon Cortesi

Amazon EMR

Easily Run Spark, Hive, Presto, HBase, Flink, and more big data apps on AWS

Latest versions

Updated with latest open source frameworks within 30 days

Support for popular OSS like Flink, Hudi

Best Performance at Lowest cost

Spark workloads run 2.4x faster compared to Open Source

50–80% reduction in costs with EC2 Spot and Reserved Instances

Per-second billing for flexibility

Use S3 storage

Process data in S3 securely with high performance using the EMRFS connector

Scale Compute and Storage independent of each other

Easy & Scalable

Fully managed, no cluster setup, node provisioning or cluster tuning

Vertical and Horizontal Auto-Scaling to suit workload demands

aws

Amazon EMR flexible deployment options

Amazon EMR on Amazon EC2

AWS offers more instance options than any other cloud provider. Choose the instance that gives the best performance or cost for your workload, including Graviton2. Take advantage of On-Demand, Reserved, and Spot Instances to optimize costs.

Amazon EMR on Amazon EKS

NEW!

Use EMR to automate the provisioning, management, and scaling of Apache Spark jobs on Amazon Elastic Kubernetes Service (EKS), and take advantage of the optimized EMR runtime.

Amazon EMR on AWS Outposts

Set up, deploy, manage, and scale EMR in your on-premises environments, just as you would in the cloud. AWS Outposts brings AWS services, infrastructure, and operating models to virtually any data center, co-location space, or on-premises facility.

aws

EMR Studio integrated development environment

Easily build and deploy data science code without logging in to AWS console

Start notebooks in seconds, run jobs later

Build production pipelines simply and flexibly

Save debugging time with native application UIs in one place

aws

Use existing identity provider with AWS SSO

Easily enable single sign-on in the AWS SSO console

- Integrated with AWS organizations
- Choose identity source like Okta, Azure AD, OneLogin, Ping Identity, etc.
- Import users and groups from identity source to AWS SSO

aws

Abstract infrastructure details using templates

Create cluster templates in AWS Service Catalog to simplify running jobs for your data scientists and data engineers –

- Control Spark version, Amazon EMR version, etc.
- Restrict number of instances and instance types
- Preconfigure libraries and packages
- Add security settings

aws

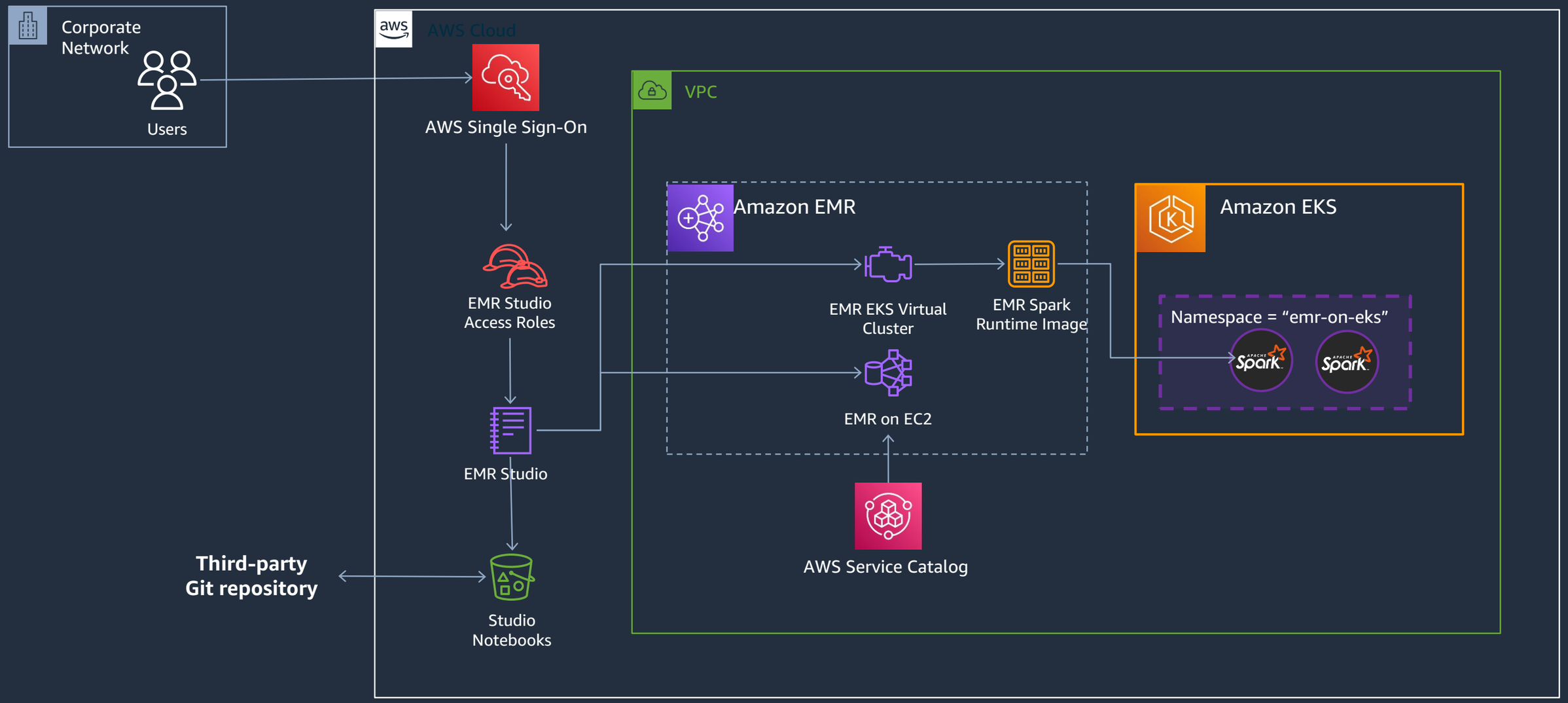
Restrict who can provision a cluster or template

Control who can submit jobs by spinning up clusters on demand instead of keeping long-running clusters for all

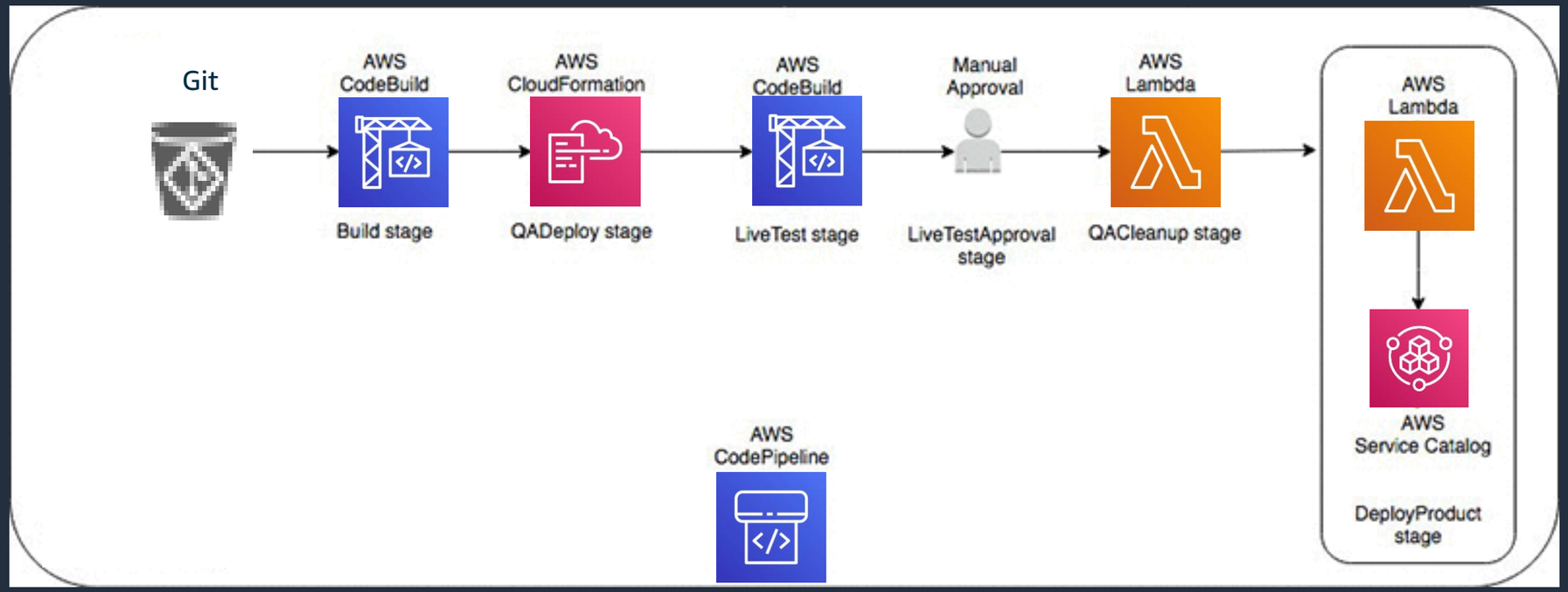
- Allow users and groups assigned to a Studio to access a set of cluster templates
- Authorize users and groups in a Studio to create clusters using configurations
- Limit users and groups to provision permitted templates to run jobs for cost savings

aws

Example Architecture



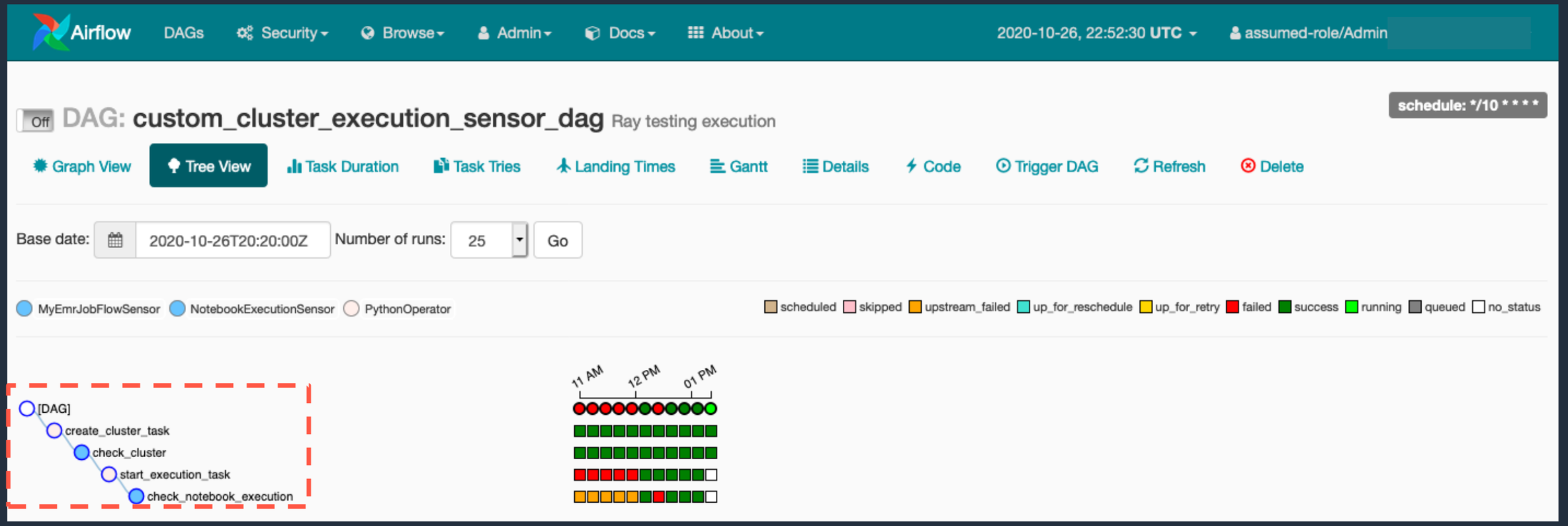
Trigger pipelines from code repositories

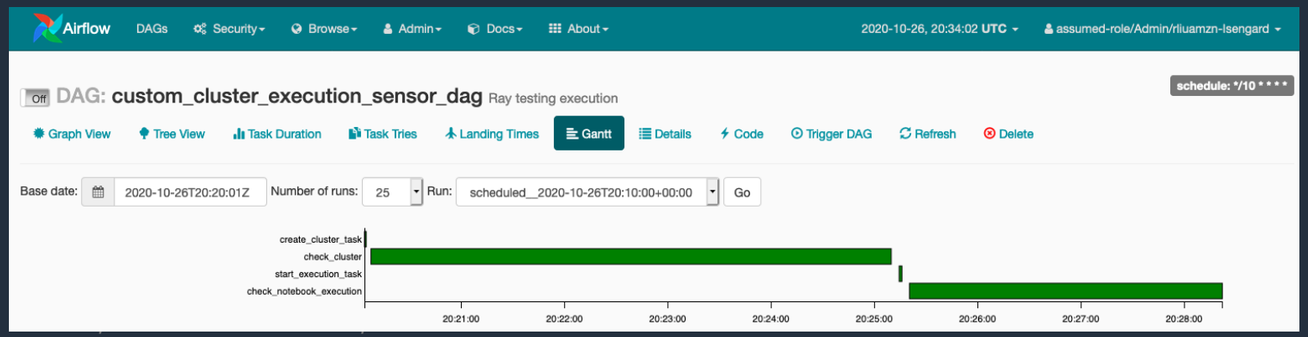


Commit code to Git and trigger a continuous integration and deployment pipeline supported by AWS services

aws

Simplify building pipelines from notebooks





Run notebooks as pipelines via Amazon Managed Workflows for Apache Airflow (MWAA) NEW!

Parameterize and chain notebooks that can be run as pipelines

aws

Resources

- EMR Studio Workshop: https://emr-developer-experience.workshop.aws/
- EMR Studio Templates: https://github.com/aws-samples/emr-studio-samples

aws

Thank you!