# AWS re:Invent

NOV. 29 – DEC. 3, 2021 | LAS VEGAS, NV

ANT305

# Data science and DataOps workflows with Amazon EMR Studio

Vincent Gromakowski
Principal analytics solution architect
AWS

© 2021, Amazon Web Services, Inc. or its affiliates. All rights reserved.

# Agenda

- Introduction to Amazon EMR Studio
- Introduction to DataOps
- Creating your Studio
- Connecting to Git
- Deploying to “production”

© 2021, Amazon Web Services, Inc. or its affiliates. All rights reserved.

# 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

## Fast performance at low cost



Spark workloads run 2.4x faster compared to open source

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

Per-second billing for flexibility

## Use Amazon S3 storage



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

Scale compute and storage independent of each other

## Easy and scalable



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

Vertical and horizontal Auto Scaling to suit workload demands

aws

© 2021, Amazon Web Services, Inc. or its affiliates. All rights reserved.

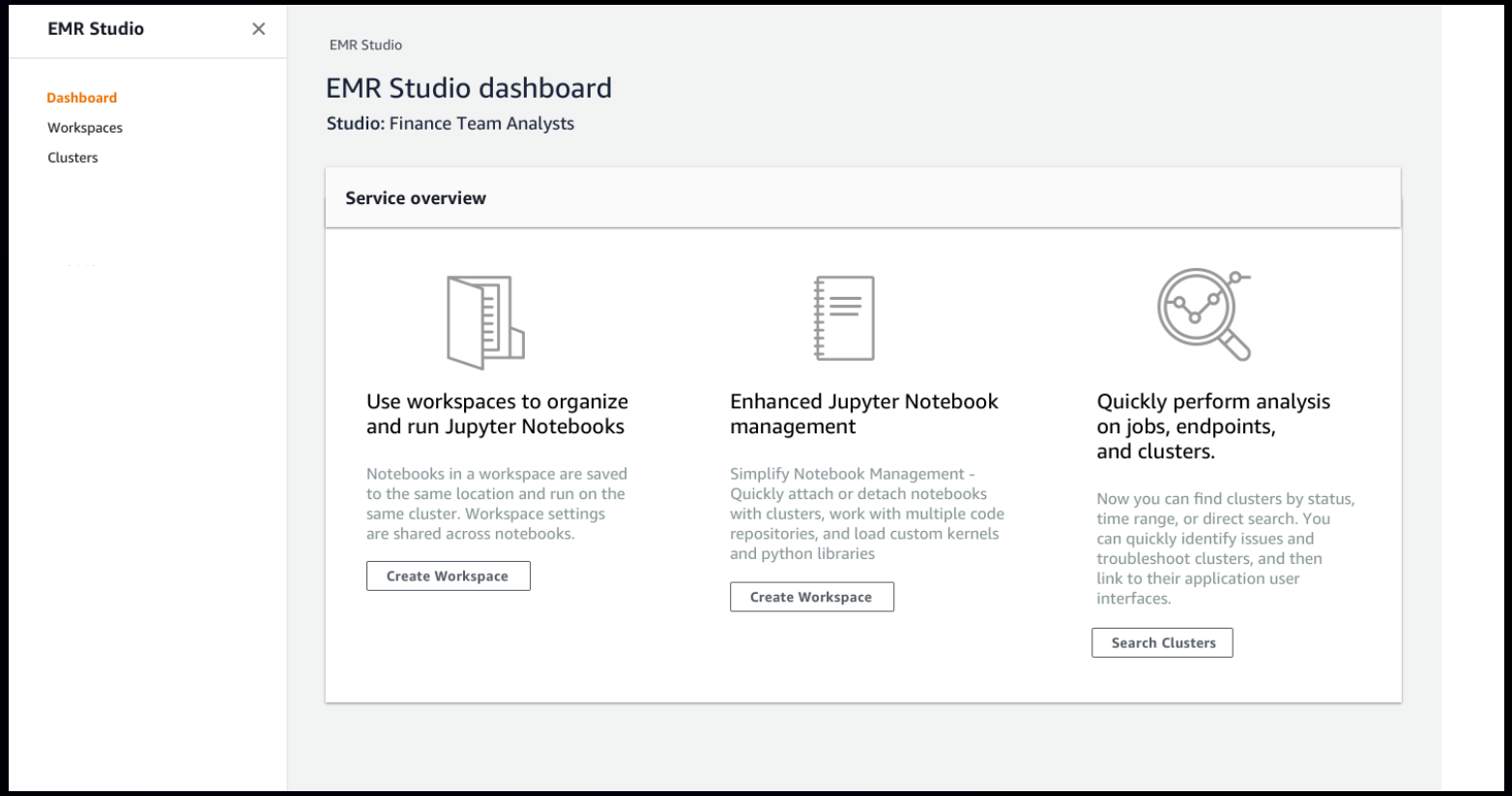
# Amazon EMR Studio

Fully managed notebook-first IDE experience for Amazon EMR with single sign-on, Jupyter notebooks, automated infrastructure provisioning, and job diagnosis

© 2021, Amazon Web Services, Inc. or its affiliates. All rights reserved.

# EMR Studio – Easy access and management

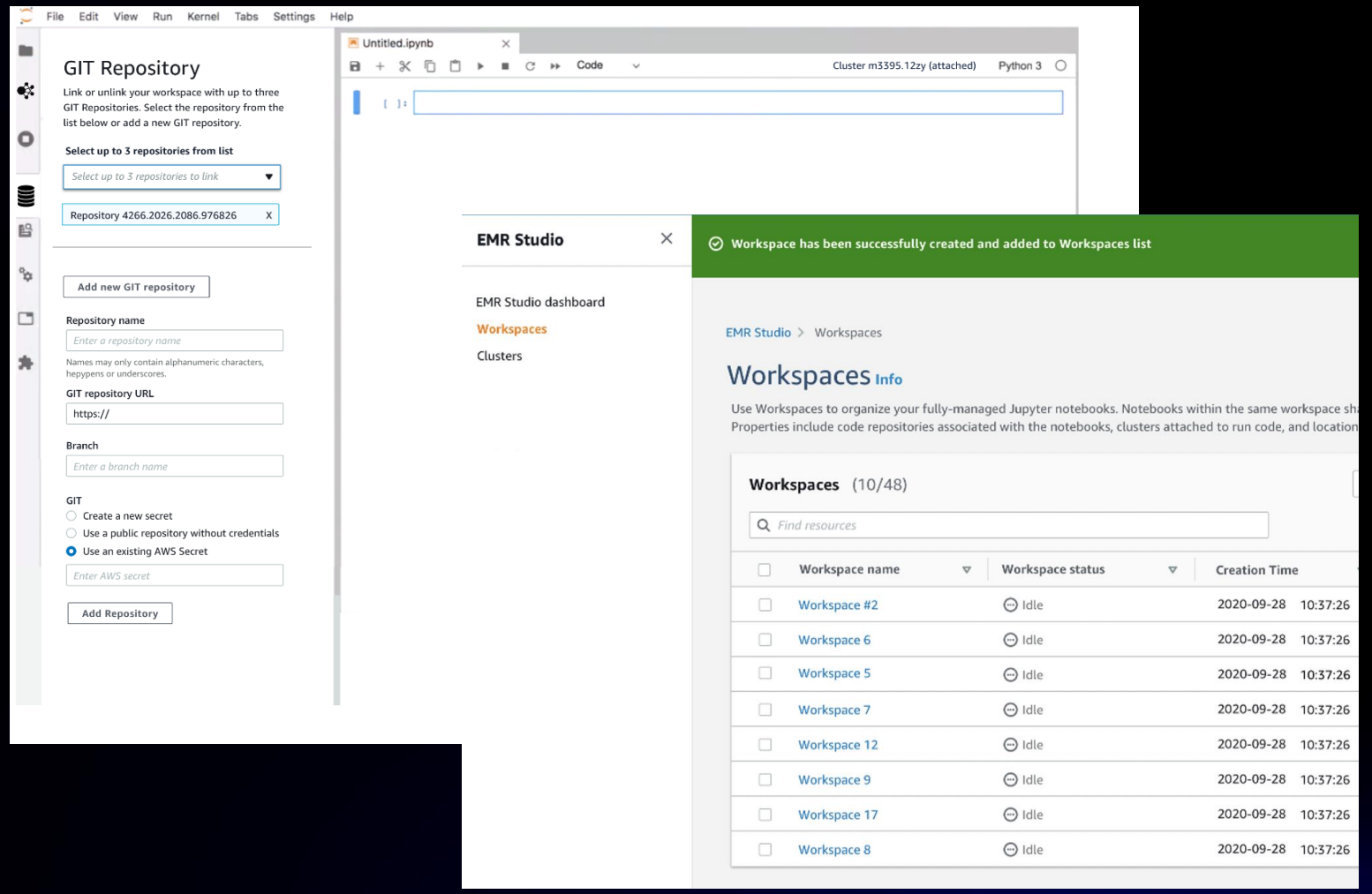
- Single sign-on with AWS SSO or use existing IAM credentials
- Create different instances of EMR Studio for different teams
- Web application that can be accessed outside the console
- Use prebuilt cluster templates that shut down automatically
- View and manage existing Amazon EMR on EC2 and Amazon EMR on EKS clusters




© 2021, Amazon Web Services, Inc. or its affiliates. All rights reserved.

# EMR Studio – Improve productivity

- Connect to GitHub and other git repositories
- Add custom kernels and Python libraries
- Run Notebooks as pipelines via Airflow
- Parameterize and chain notebooks
- Create workspaces with one or more notebooks




© 2021, Amazon Web Services, Inc. or its affiliates. All rights reserved.

# 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
- Add optional parameters

© 2021, Amazon Web Services, Inc. or its affiliates. All rights reserved.

# DataOps

“DataOps is an automated, process-oriented methodology, used by analytic and data teams, to improve the quality and reduce the cycle time of data analytics.”*

- Integrated continuous integration/continuous deployment (CI/CD)
- Revision-controlled data pipelines
- Automated testing and deployment
- "DataOps is a data management method that emphasizes communication, collaboration, integration, automation and measurement of cooperation between data engineers, data scientists and other data professionals.”

Andy Palmer, CEO and co-founder, Tamr Inc.

Source: *Wikipedia contributors, "DataOps," *Wikipedia, The Free Encyclopedia,* https://en.wikipedia.org/w/index.php?title=DataOps&oldid=1045562971

© 2021, Amazon Web Services, Inc. or its affiliates. All rights reserved.

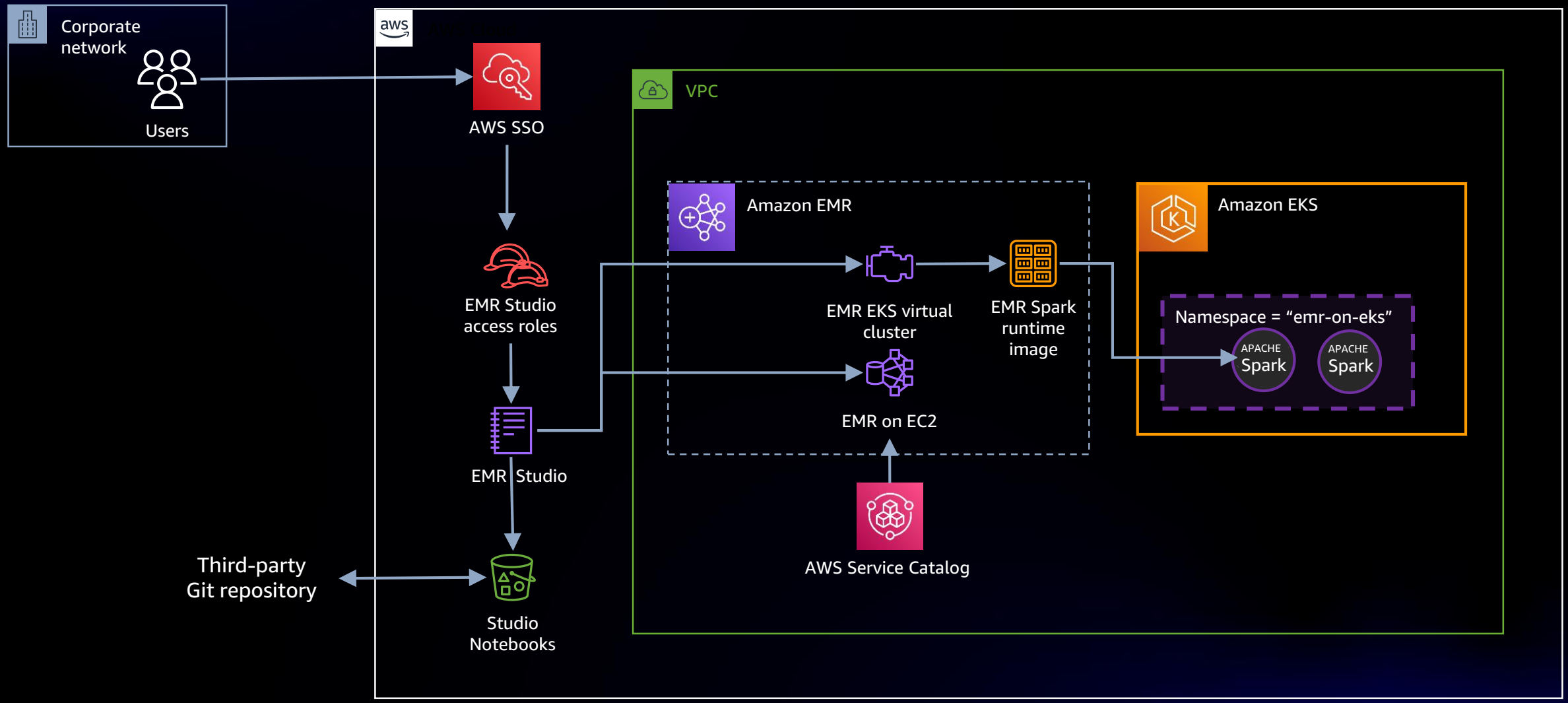
# Today’s goal

I am a data professional, and I want to be able to easily explore, process, and visualize data at scale

I also want to be able to easily deploy the results of my analysis in an automated fashion so I can move from data to insights and share with my team

© 2021, Amazon Web Services, Inc. or its affiliates. All rights reserved.

# Example architecture




© 2021, Amazon Web Services, Inc. or its affiliates. All rights reserved.

# Workshop instructions

Go to dashboard.eventengine.run

Use the AWS event hash: **669a-1711def104-a9**

Follow the **Content Guide**

© 2021, Amazon Web Services, Inc. or its affiliates. All rights reserved.

# Thank you!

Vincent Gromakowski
gromav@amazon.com

© 2021, Amazon Web Services, Inc. or its affiliates. All rights reserved.