
AWS
re:Invent

ANT323

Introducing a new notebook-first IDE experience with Amazon EMR

Shuang Li

Senior Product Manager

AWS

AWS re:Invent

© 2020, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Agenda

Data science and engineering workflow

What is Amazon EMR Studio

Customer benefits

How Amazon EMR Studio works

Demo

Administrator capabilities



© 2020, Amazon Web Services, Inc. or its affiliates. All rights reserved.

What does a data science and data engineering workflow look like?



Data science and engineering



Code is written in Scala, R, Python

Run a notebook on a compute resource



Data science and engineering



Code is written in Scala, R, Python



In the cloud



On-premises clusters

Share notebook as code



Data science and engineering





In the cloud



On-premises clusters



Share notebook as code via code repositories

Run notebooks as pipelines



Data science and engineering







In the cloud



On-premises clusters



Share notebooks as code via code repositories

What Amazon EMR customers told us

1. Use corporate identity to log directly in to notebooks and not the AWS console



Data science and engineering



Apache Airflow



Attach to new, existing, or predefined clusters



Share notebook as code via code repositories

1. Use corporate identity to log directly in to notebooks and not the AWS console

Apache Airflow



Data science and engineering

Attach to new, existing, or predefined clusters

2. Fully managed notebooks with no management

Share notebook as code via code repositories

1. Use corporate identity to log directly in to notebooks and not the AWS console



Data science and engineering



Apache Airflow



Attach to new, existing, or predefined clusters

2. Fully managed notebooks with no management



Simple way to connect to repositories to share code

3. Share notebook as code via code repositories

1.

Use corporate identity to log directly in to notebooks and not the AWS console

Data science and engineering



4.

Simple way to run notebooks as jobs / pipelines from airflow / step functions

Apache Airflow

Attach to new, existing, or predefined clusters

2.

Fully managed notebooks with no management

Simple way to connect to repositories to share code

3.

Share notebook as code via code repositories

1. Use corporate identity to log directly in to notebooks and not the AWS console

Data science and engineering



2. Fully managed notebooks with no management



3. Simple way to connect to repositories to share code

Share notebook as code via code repositories

Apache Airflow

4. Simple way to run notebooks as jobs / pipelines from airflow / step functions

5. Cluster templates so that there is no cluster configuration; admins sets limits

Attach to new, existing, or predefined clusters

1. Use corporate identity to log directly in to notebooks and not the AWS console

Data science and engineering



2. Fully managed notebooks with no management



3. Simple way to connect to repositories to share code

Share notebook as code via code repositories

4. Simple way to run notebooks as jobs / pipelines from airflow / step functions

Apache Airflow

5. Cluster templates so that there is no cluster configuration; admins sets limits

Attach to new, existing, or predefined clusters

6. Multitenancy: many notebooks attached to a single cluster to save costs

1. Use corporate identity to log directly in to notebooks and not the AWS console

Data science and engineering



Apache Airflow

4. Simple way to run notebooks as jobs / pipelines from airflow / step functions

5. Cluster templates so that there is no cluster configuration; admins sets limits

Attach to new, existing, or predefined clusters

2. Fully managed notebooks with no management

Simple way to connect to repositories to share code

Simple way to debug jobs

3.

7.

6. Multitenancy: many notebooks attached to a single cluster to save costs

Share notebook as code via code repositories

Introducing Amazon EMR Studio (preview)



AWS re:Invent

© 2020, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Amazon EMR Studio: A notebook-first IDE for data scientists and data engineers

Easy to develop, visualize, and debug data engineering and data science applications written in R, Python, Scala, and PySpark

1. Use corporate identity to log directly into notebooks instead of the AWS console
2. Install custom kernels and libraries directly from notebook
3. Collaborate with peers by sharing notebooks via GitHub and other repositories
4. Build pipelines using orchestration services like Apache Airflow

Customer benefits



Simple to use



Fully managed Jupyter Notebooks



Easy to build production pipelines



Simplified debugging

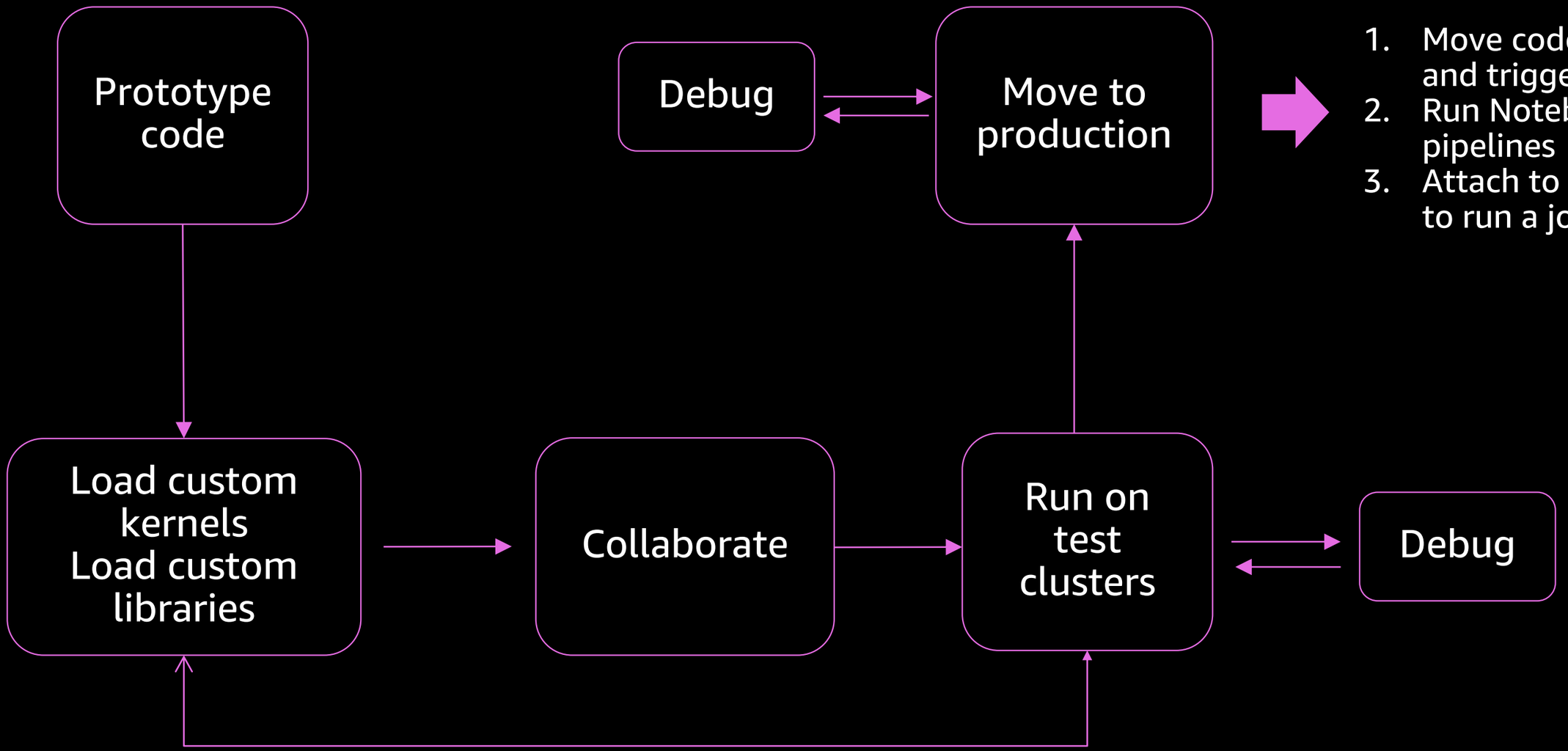
Make data scientists and data engineers more productive



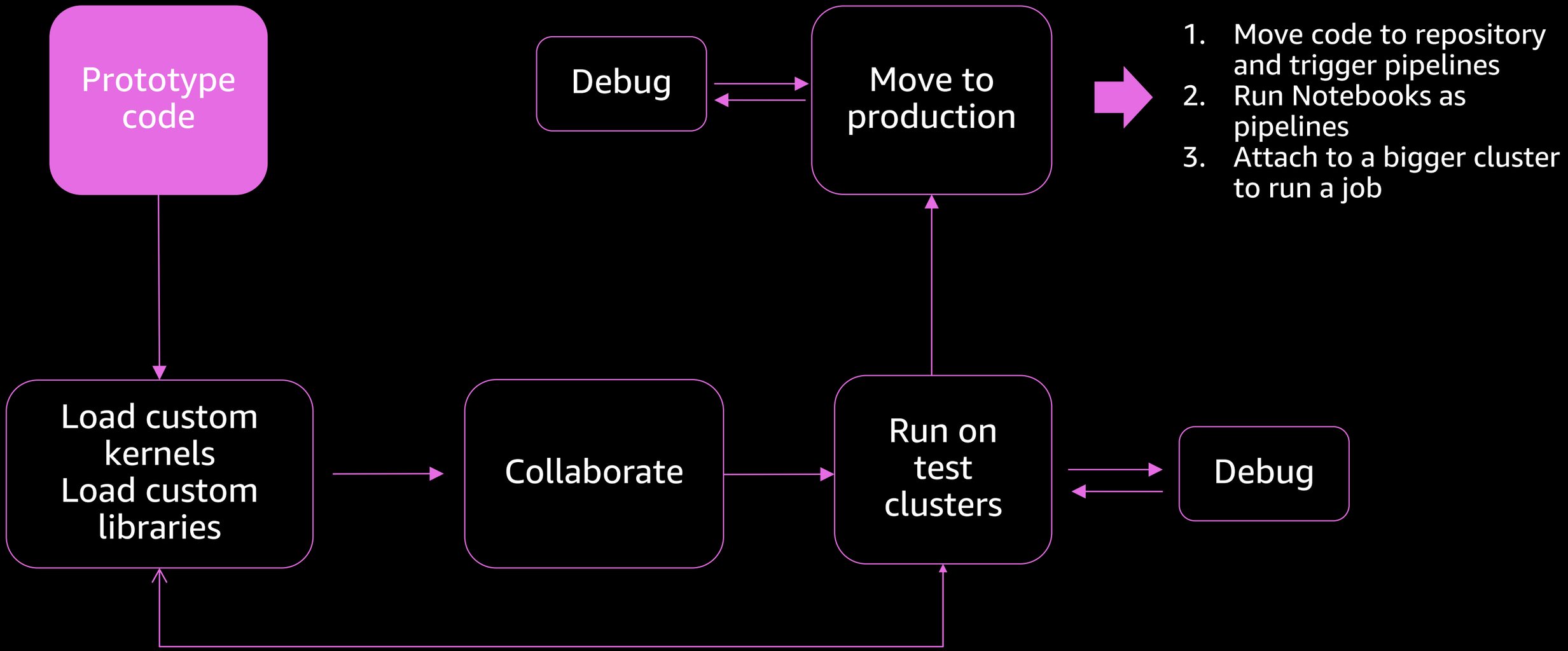
AWS re:Invent

© 2020, Amazon Web Services, Inc. or its affiliates. All rights reserved.

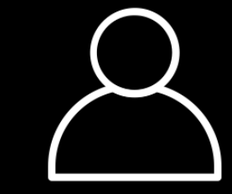
Data science and engineering workflows



Data science and engineering workflows



Log into Amazon EMR Studio without logging into the AWS Console

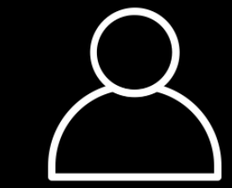


Data scientist

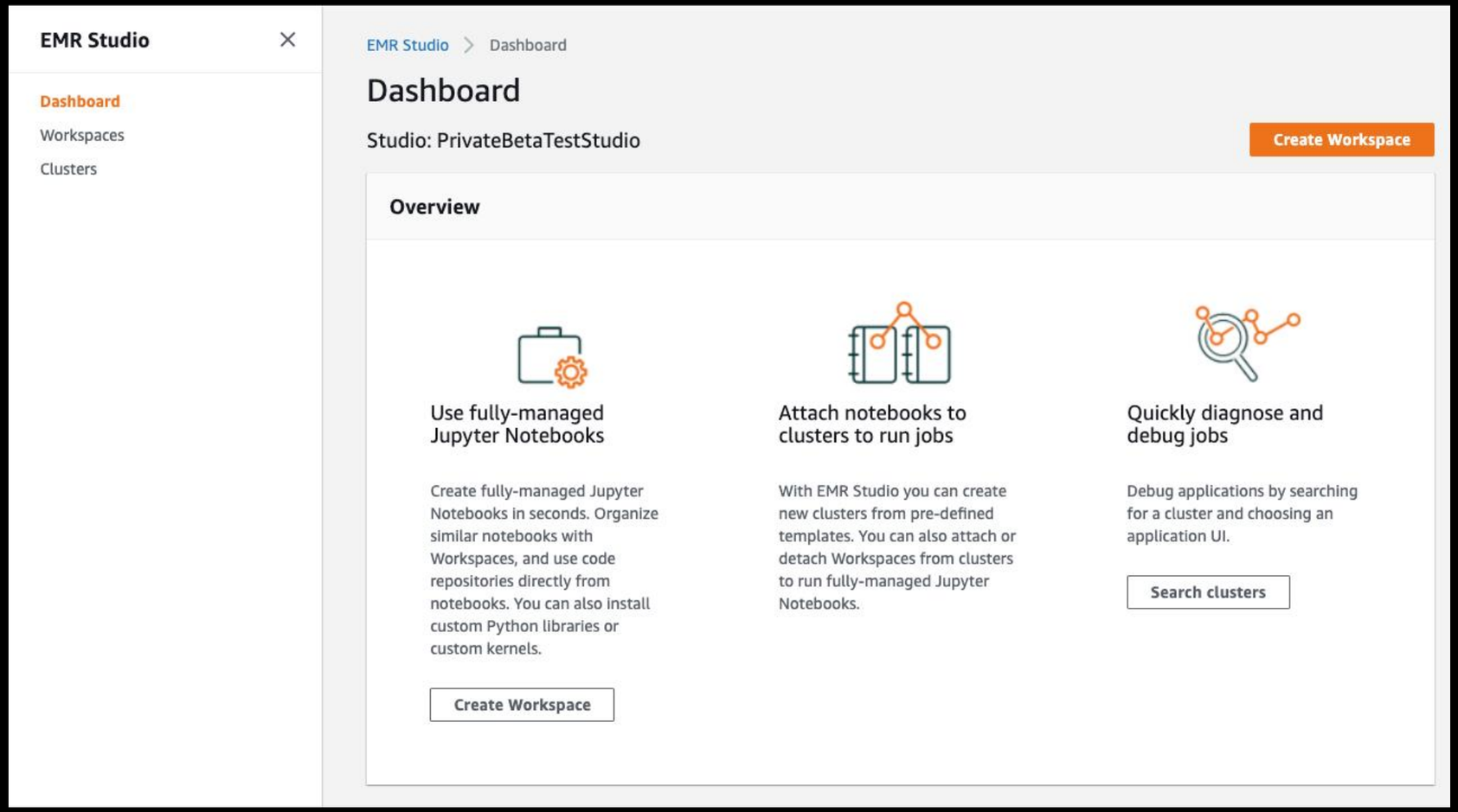


Log in with corporate identity using AWS Single Sign-On

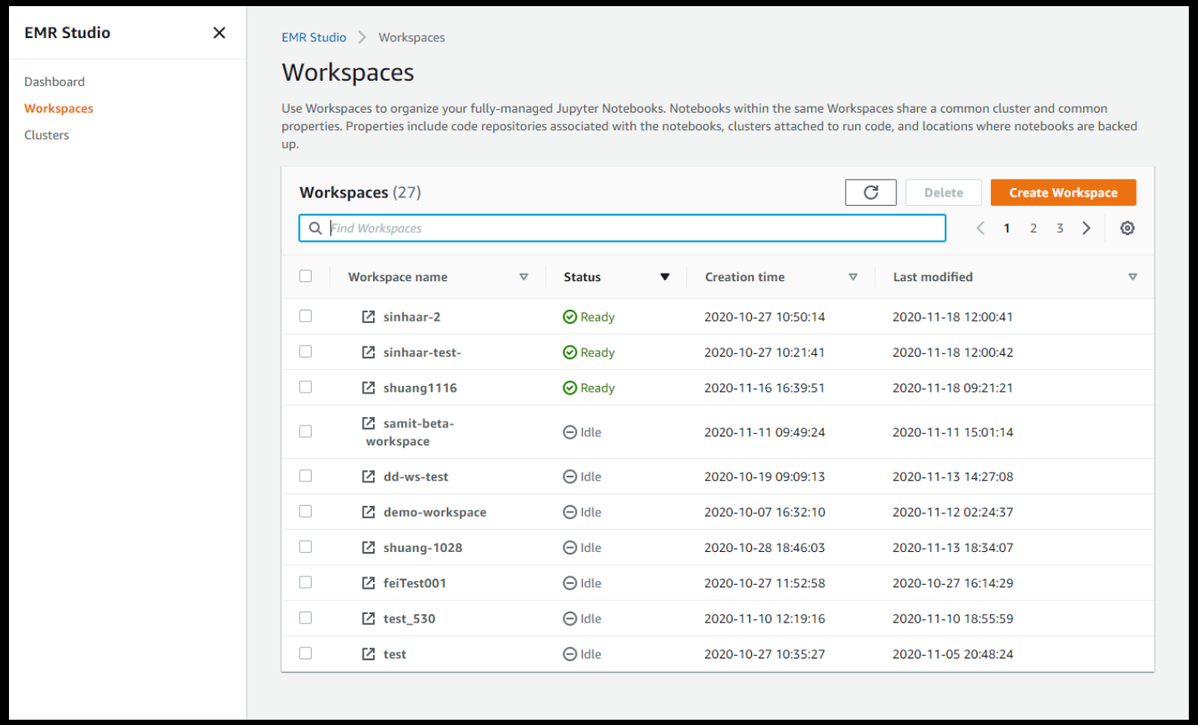


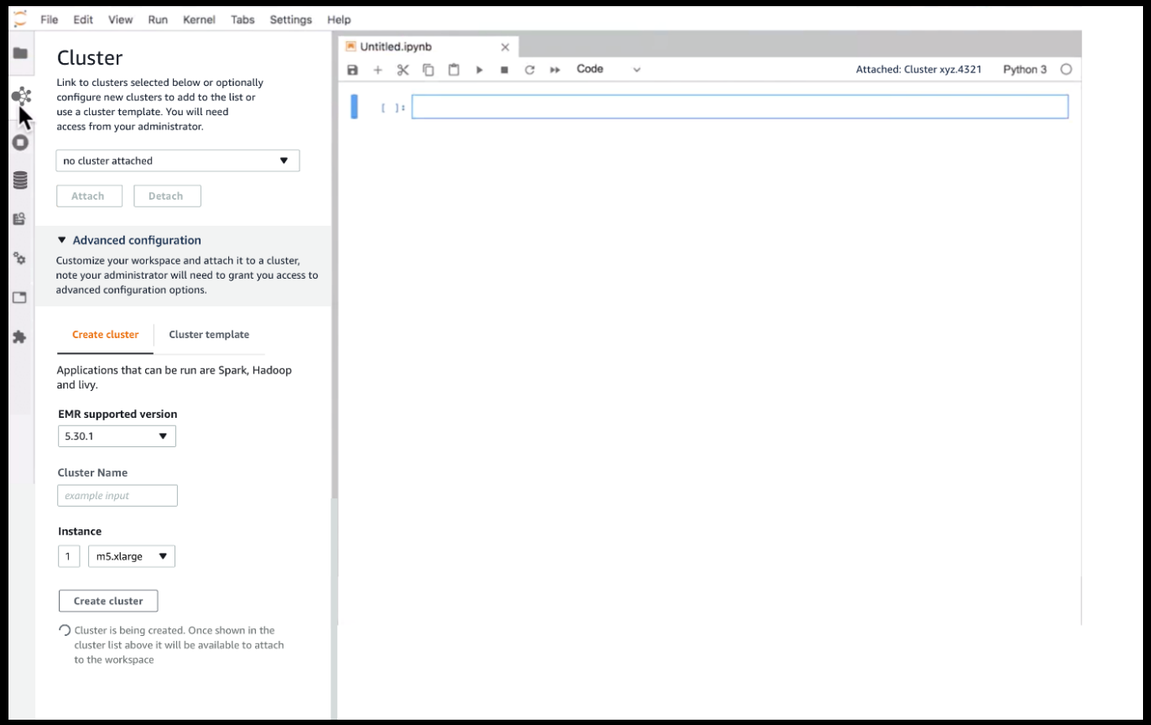


Data engineer



Amazon EMR Studio gives you a fully-managed Notebook





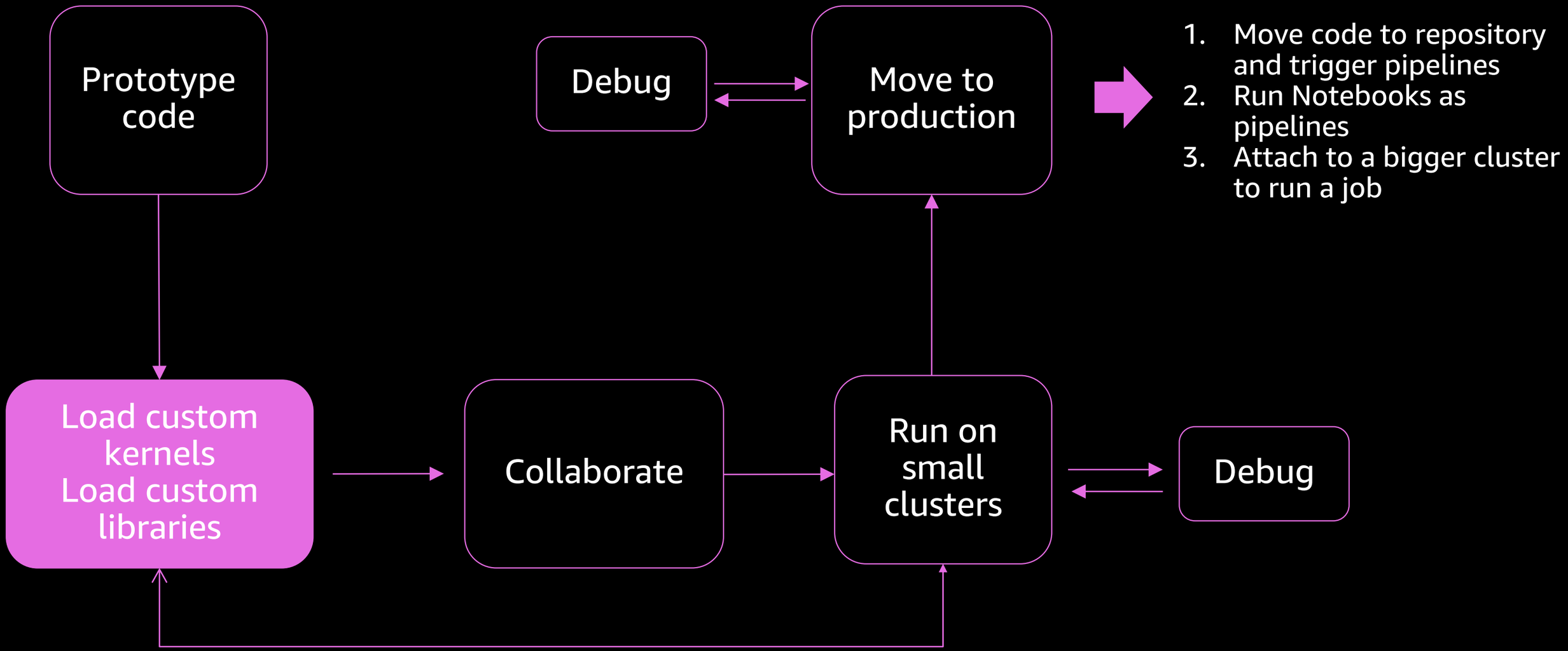
Workspaces help organize notebooks

Workspaces share similar properties

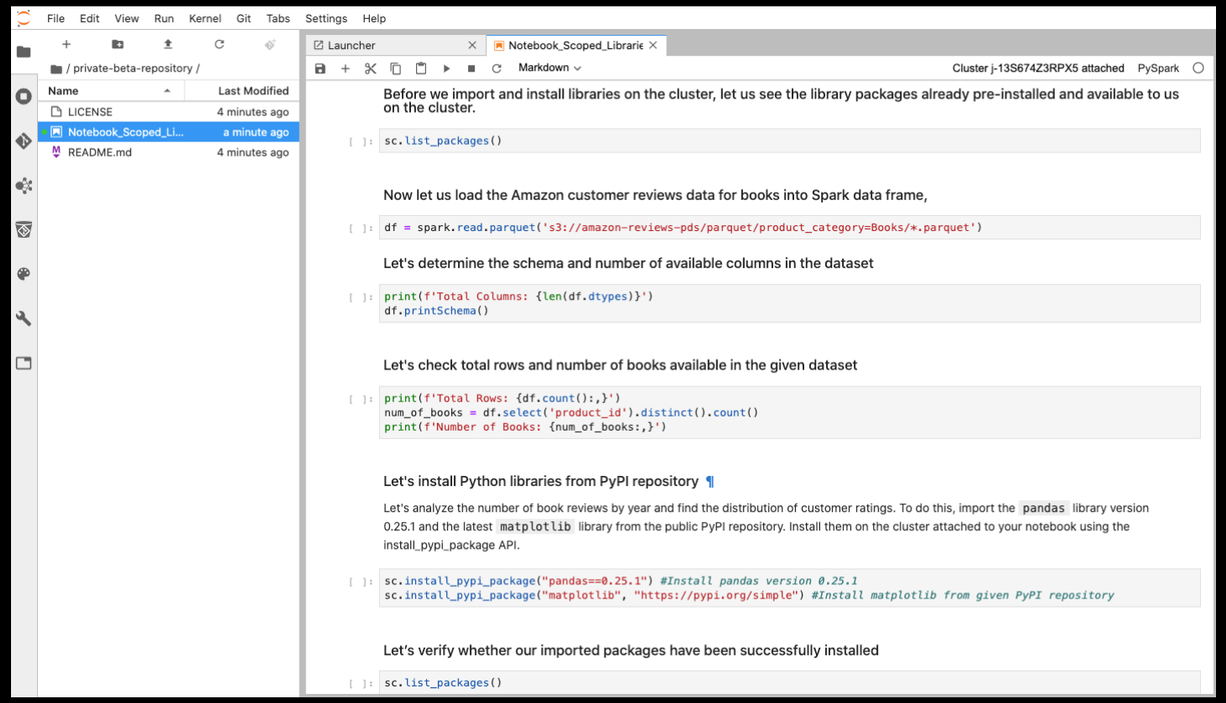
Fully-managed Jupyter Notebooks

Write Python, R, PySpark, Scala

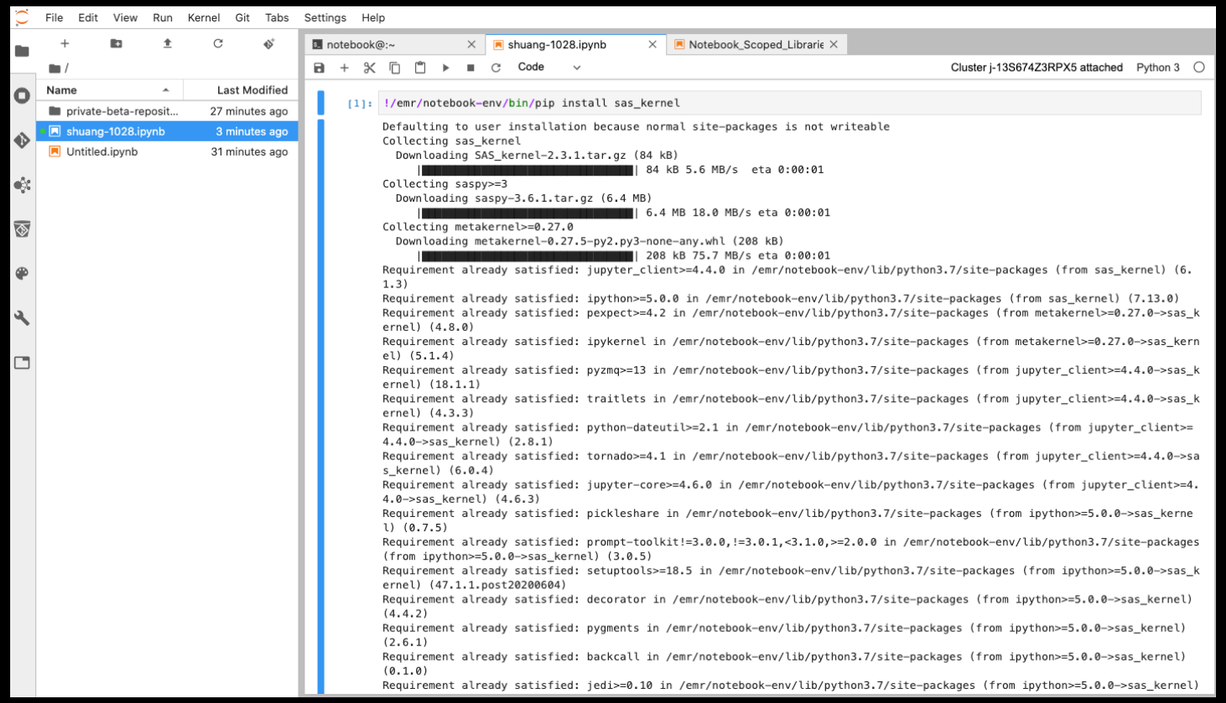
Data science and engineering workflows



Simple to load custom libraries and kernels

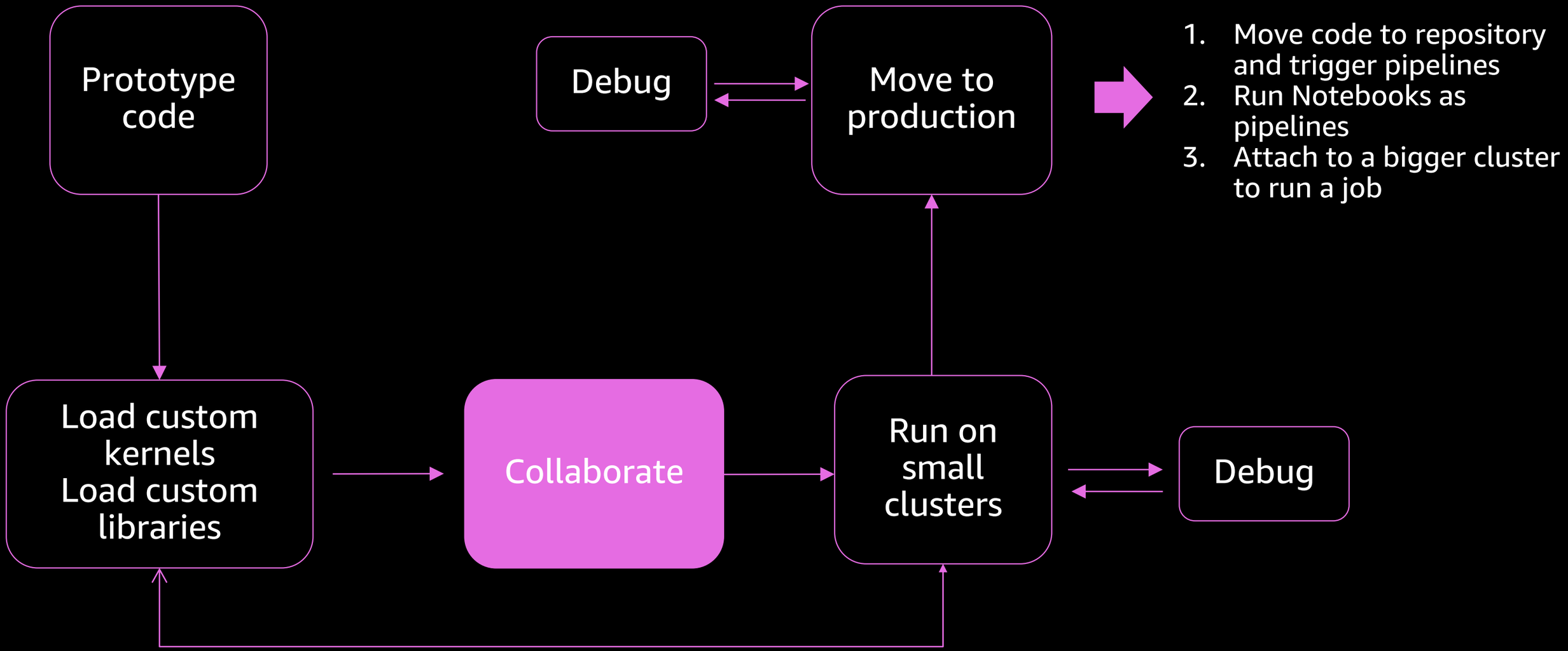


Install notebook-scoped libraries with PySpark kernel

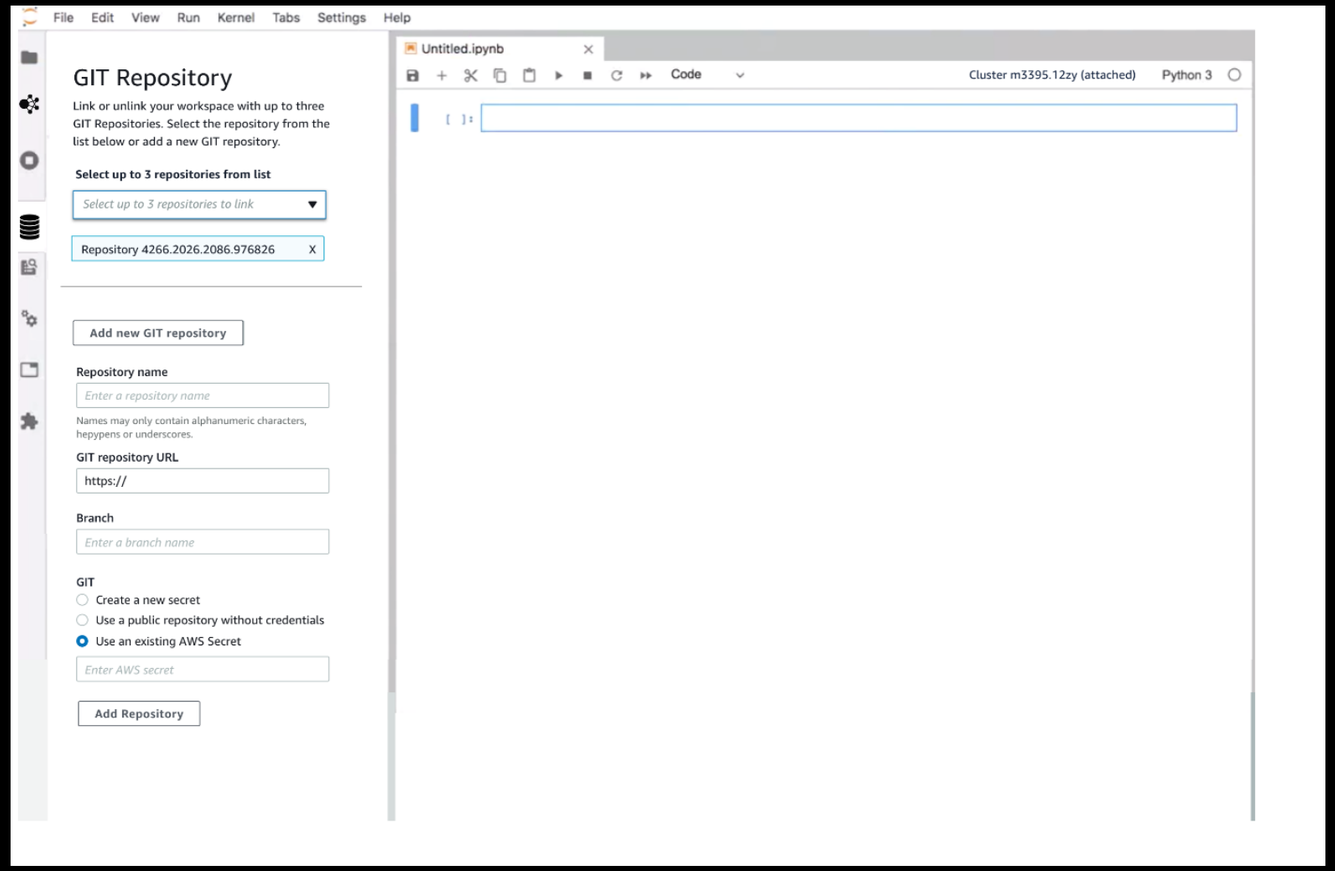


Install additional Python libraries and kernels on the master node of the cluster

Data science and engineering workflows



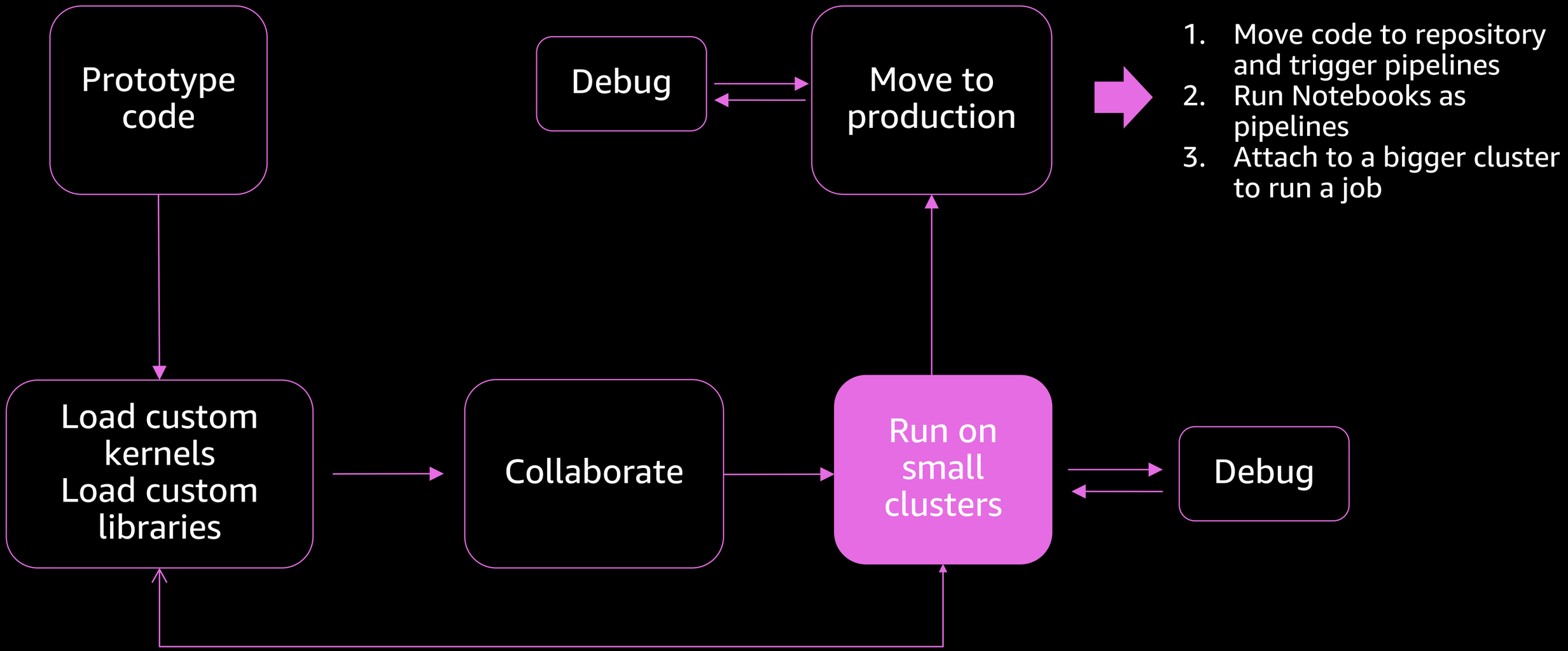
Simple to connect to Code repositories



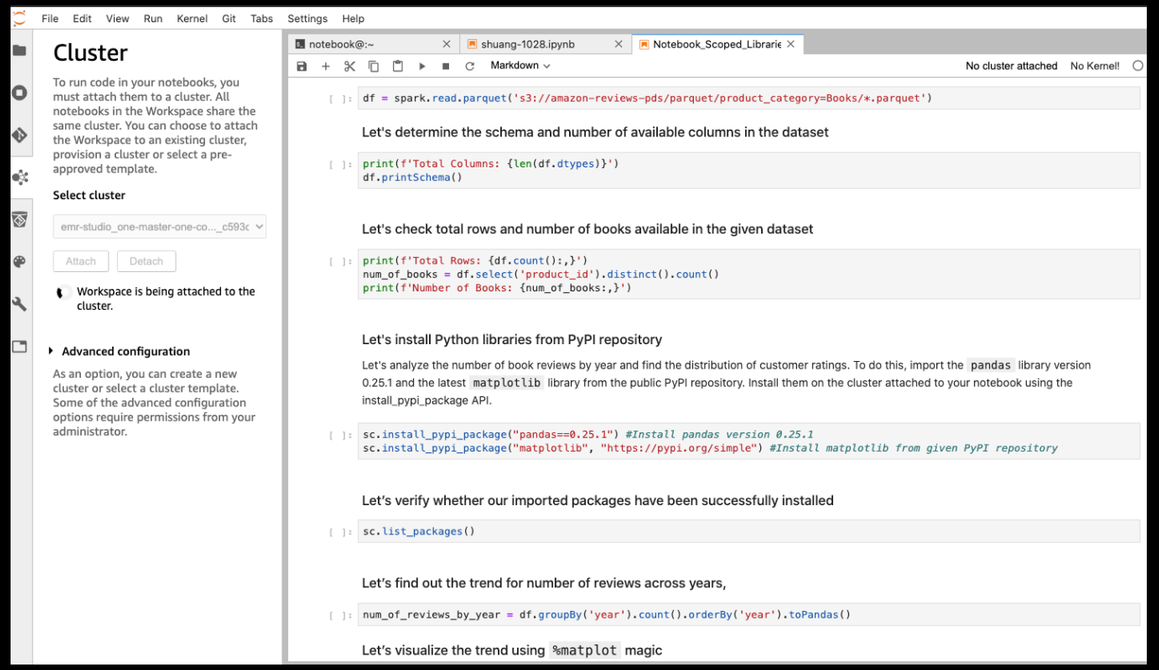
Connect to AWS CodeCommit, GitHub, and Bitbucket

Select existing or add new Git repositories

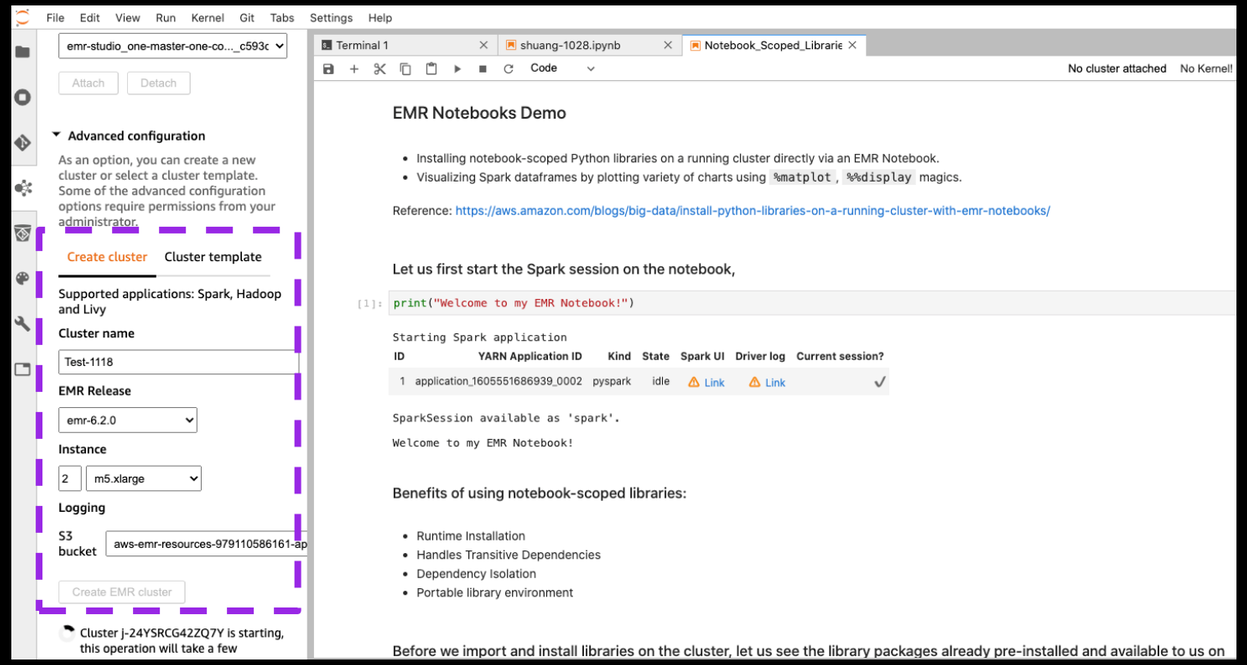
Data science and engineering workflows



Single-click attach to clusters to run jobs

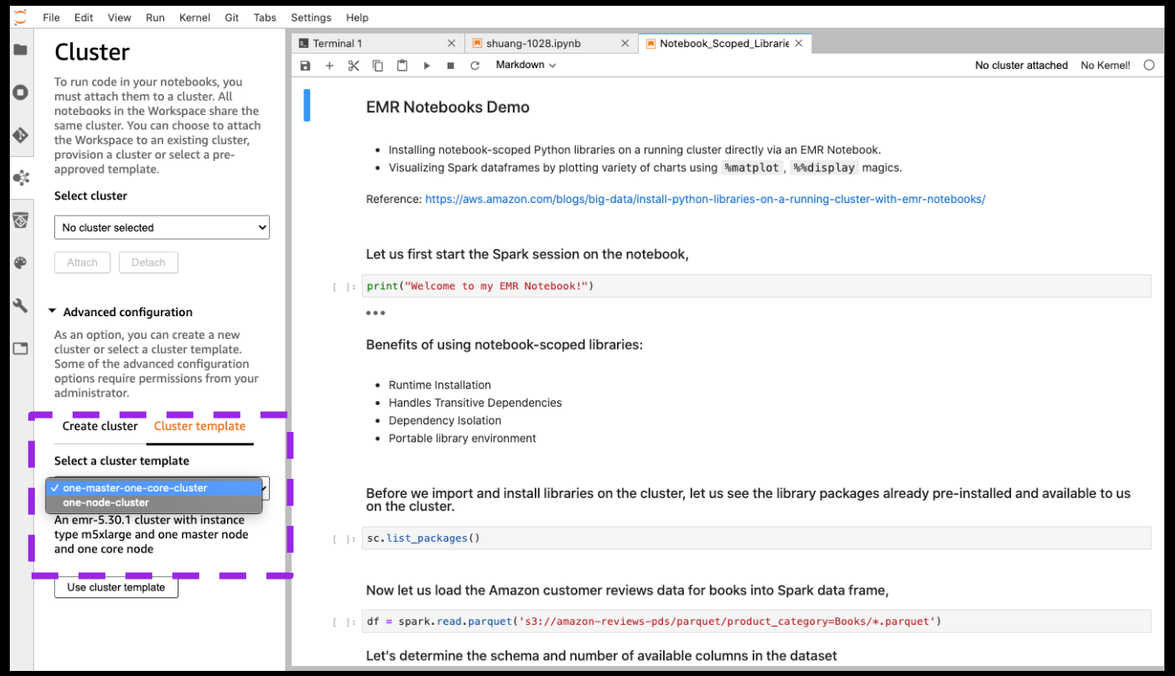


Attach workspace to an existing Amazon EMR cluster

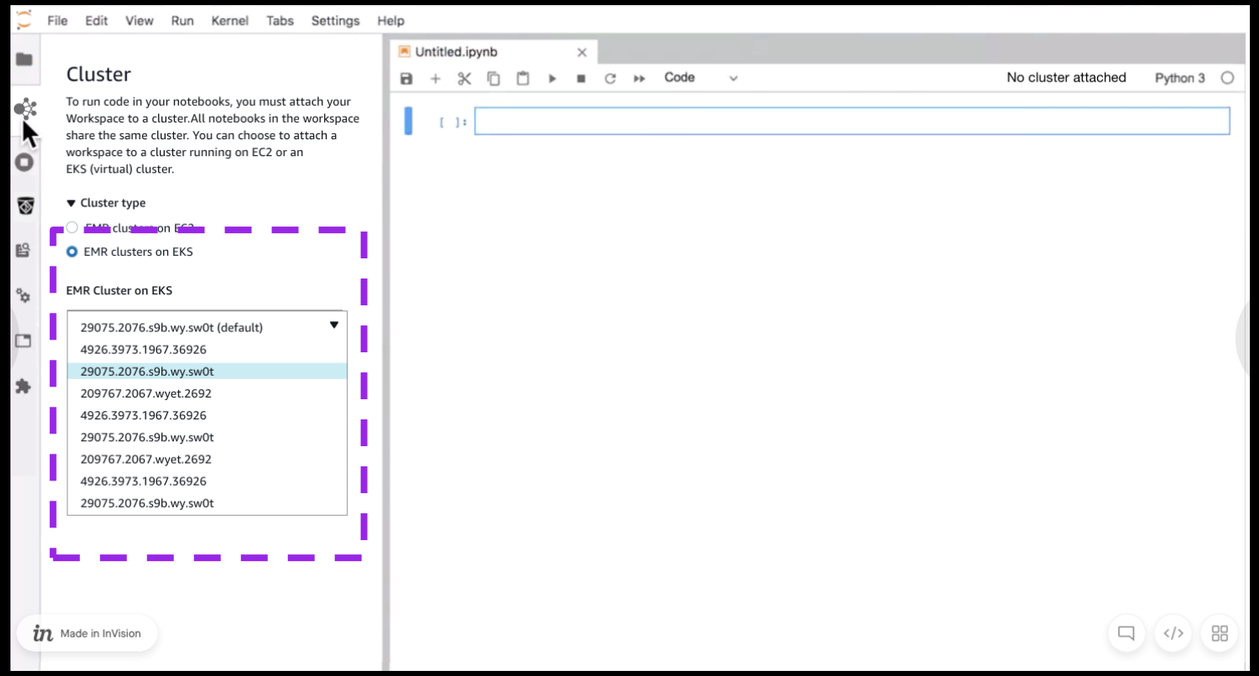


Provision Amazon EMR clusters using simple configurations

Single-click attach to clusters to run jobs

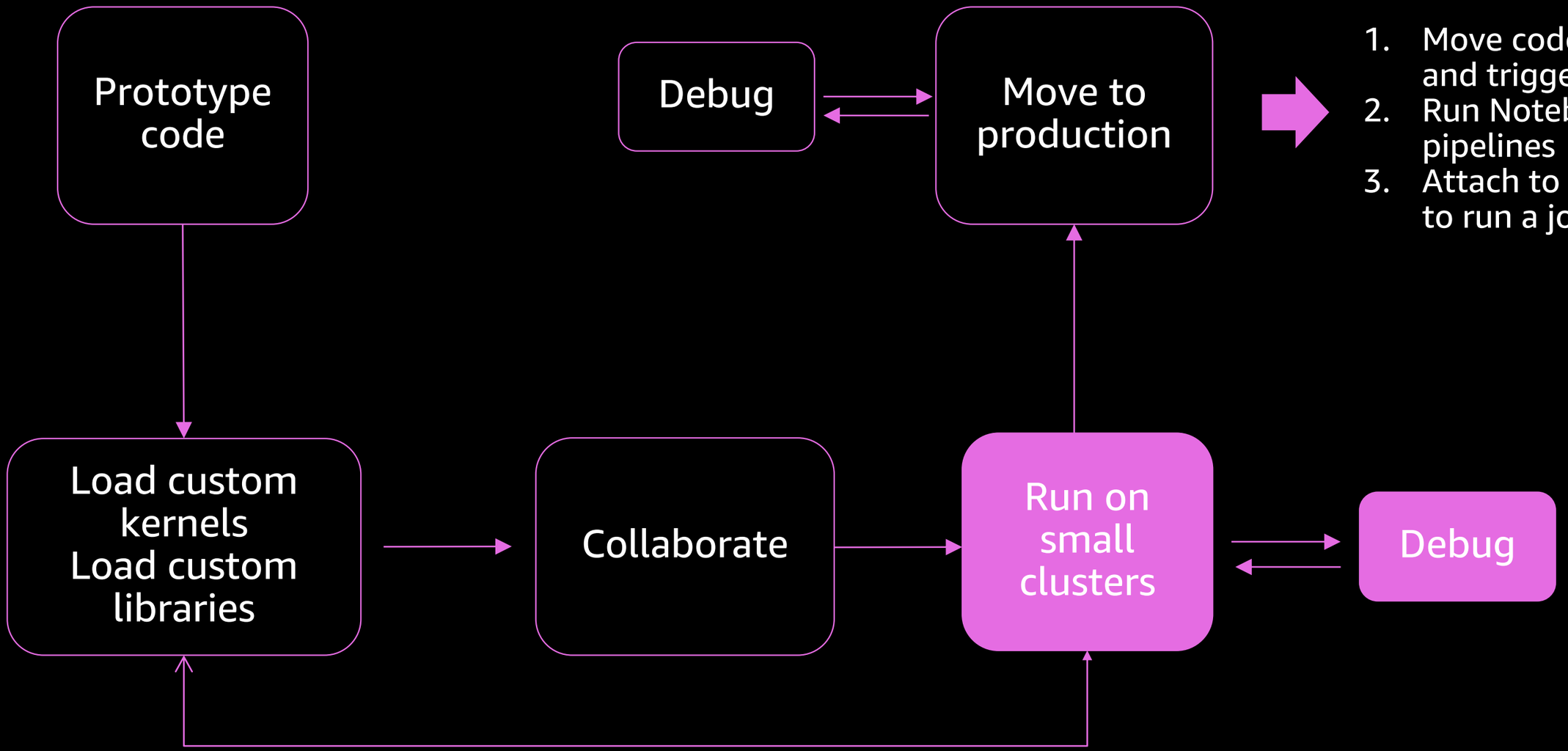


Provision Amazon EMR clusters using preconfigured cluster templates via AWS Service Catalog

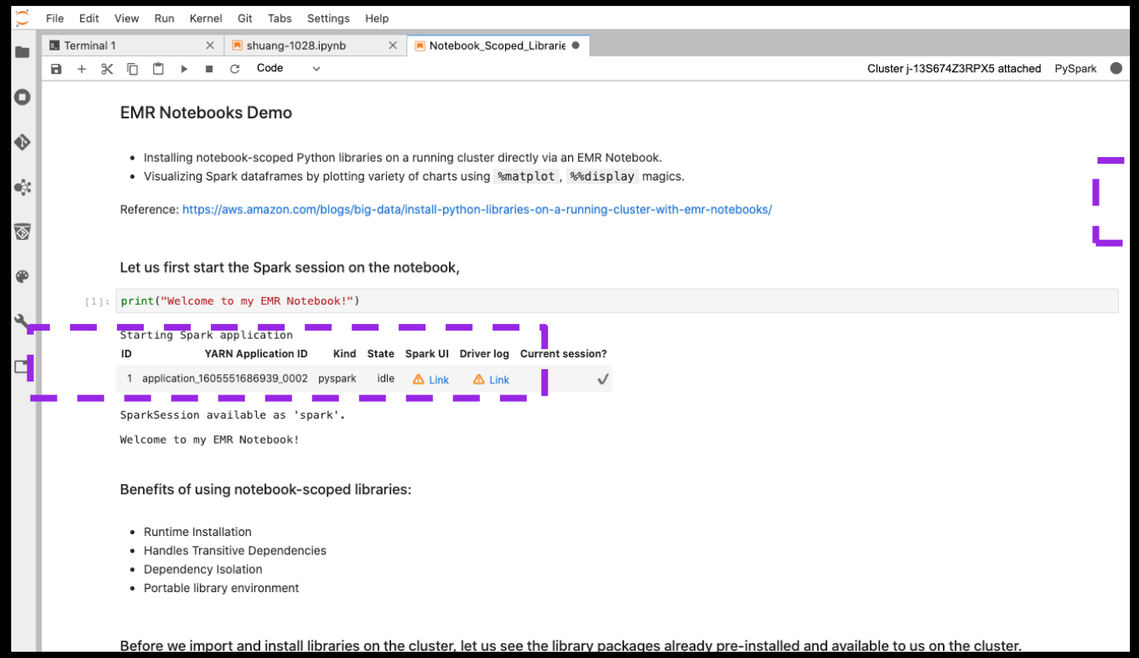


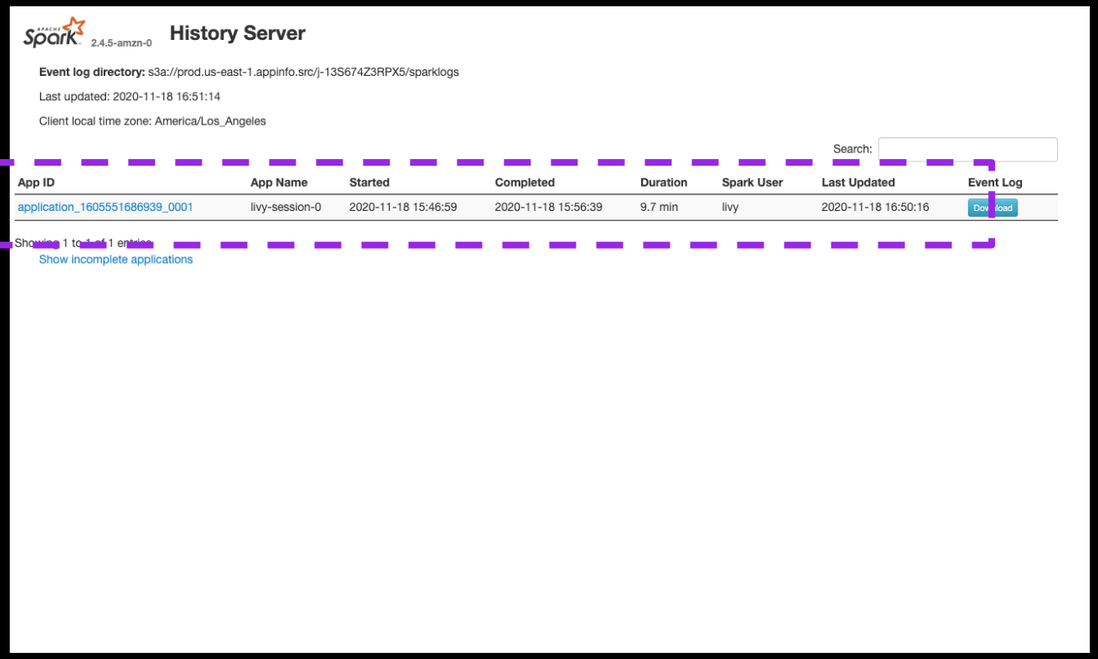
Provision clusters from Amazon EKS

Data science and engineering workflows



Debugging directly from Notebooks is simple

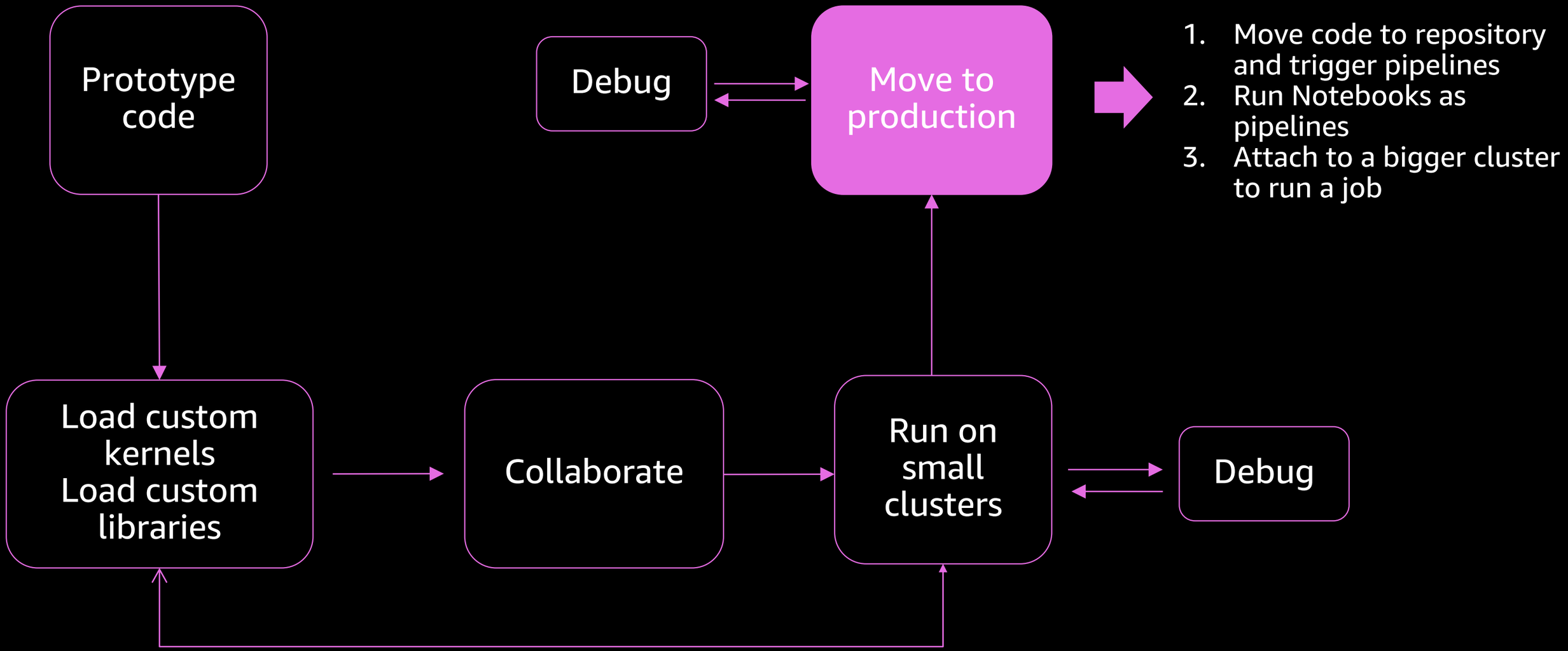




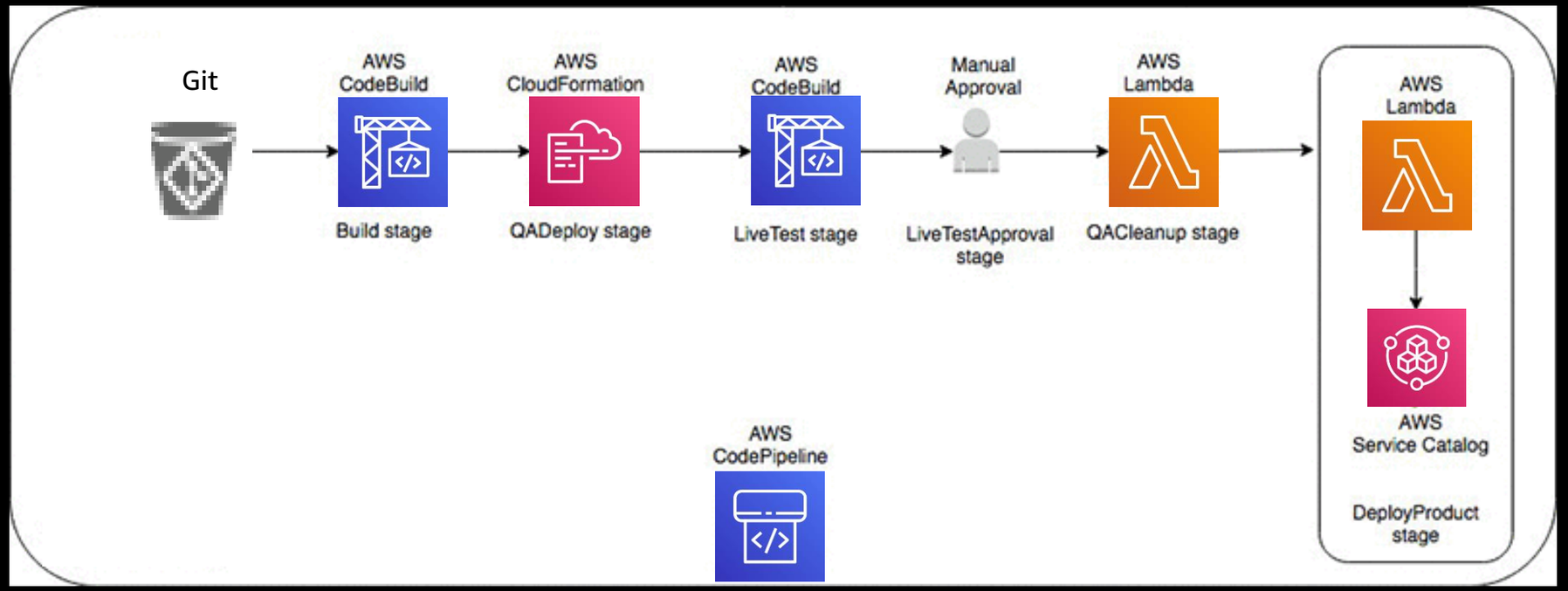
Debug by clicking the “Spark UI” link in notebook

View debugging information in Spark History Server for the application in a separate browser tab

Data science and engineering workflows

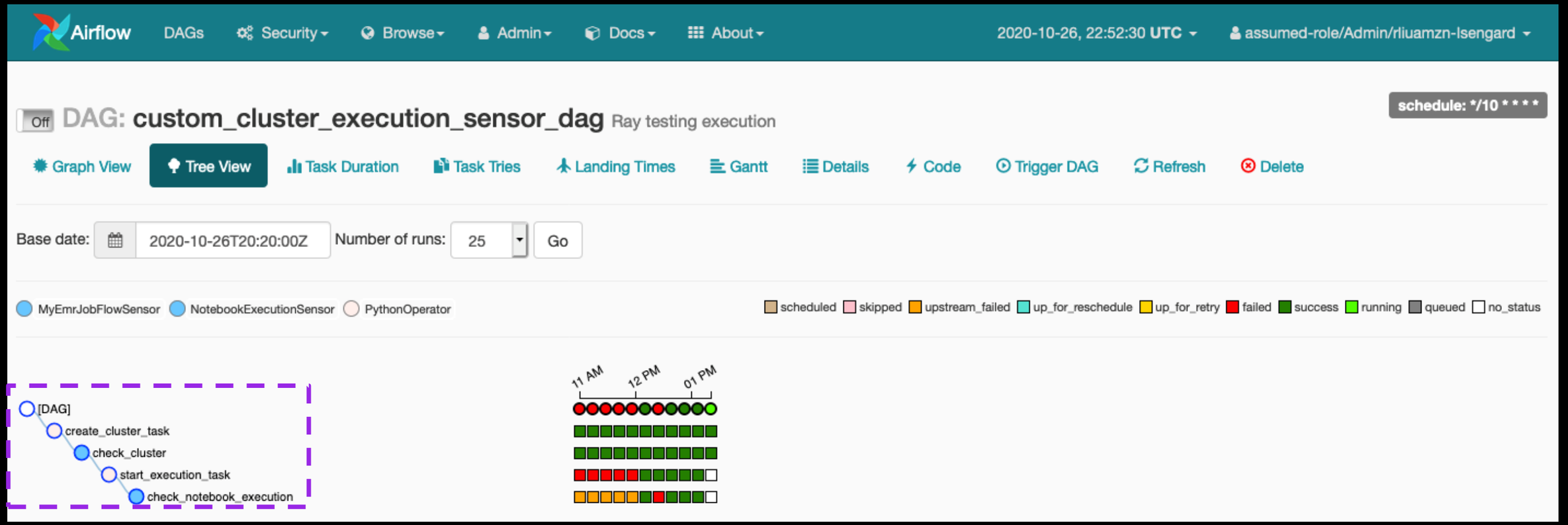


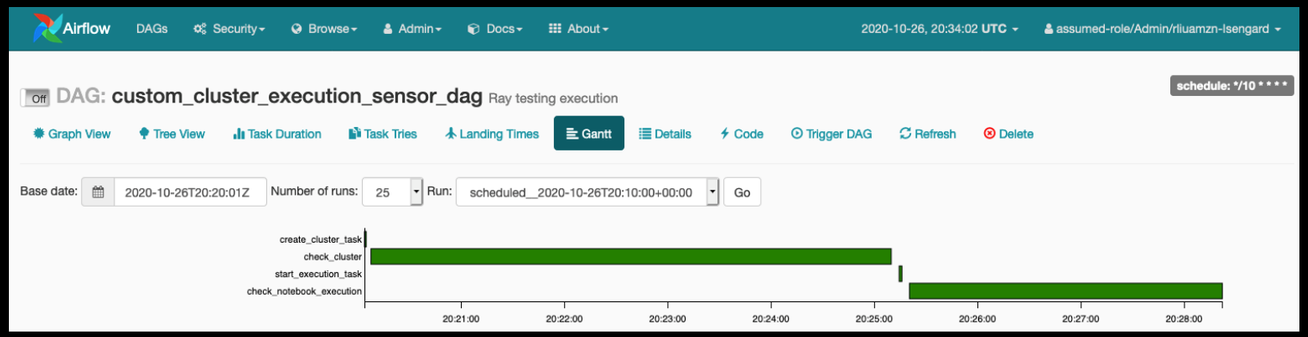
Trigger pipelines from code repositories



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

Simplify building pipelines from notebooks

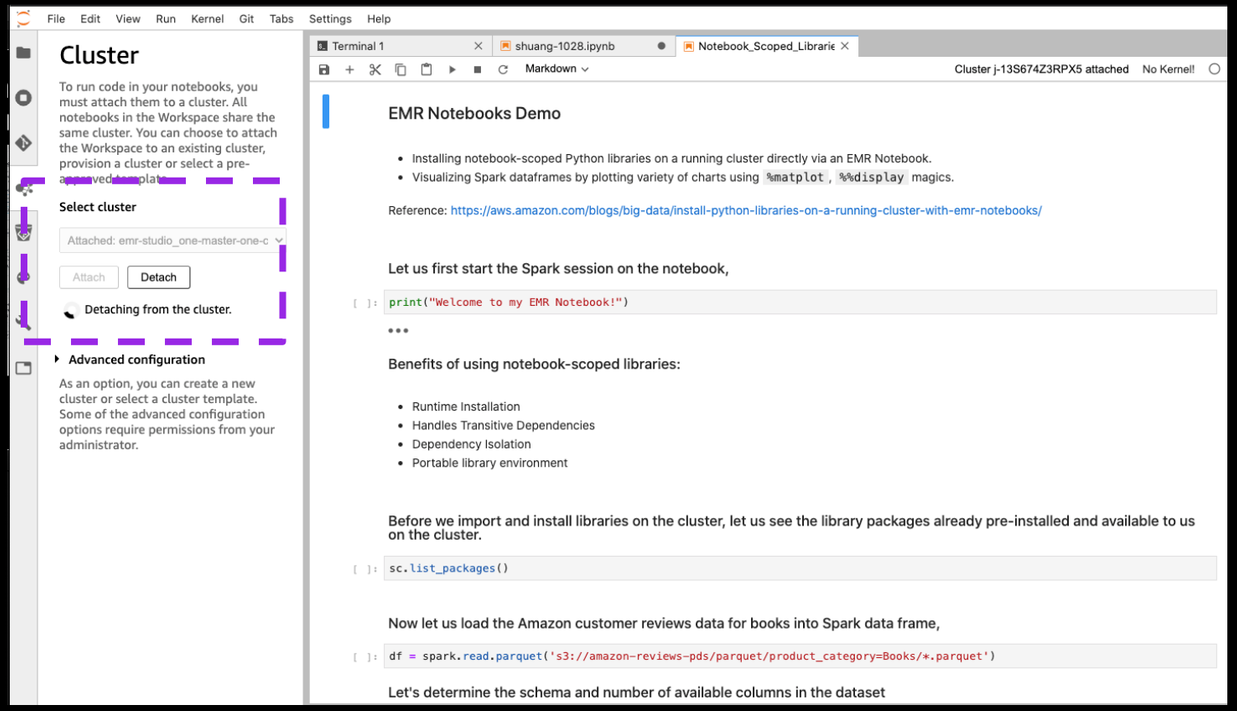




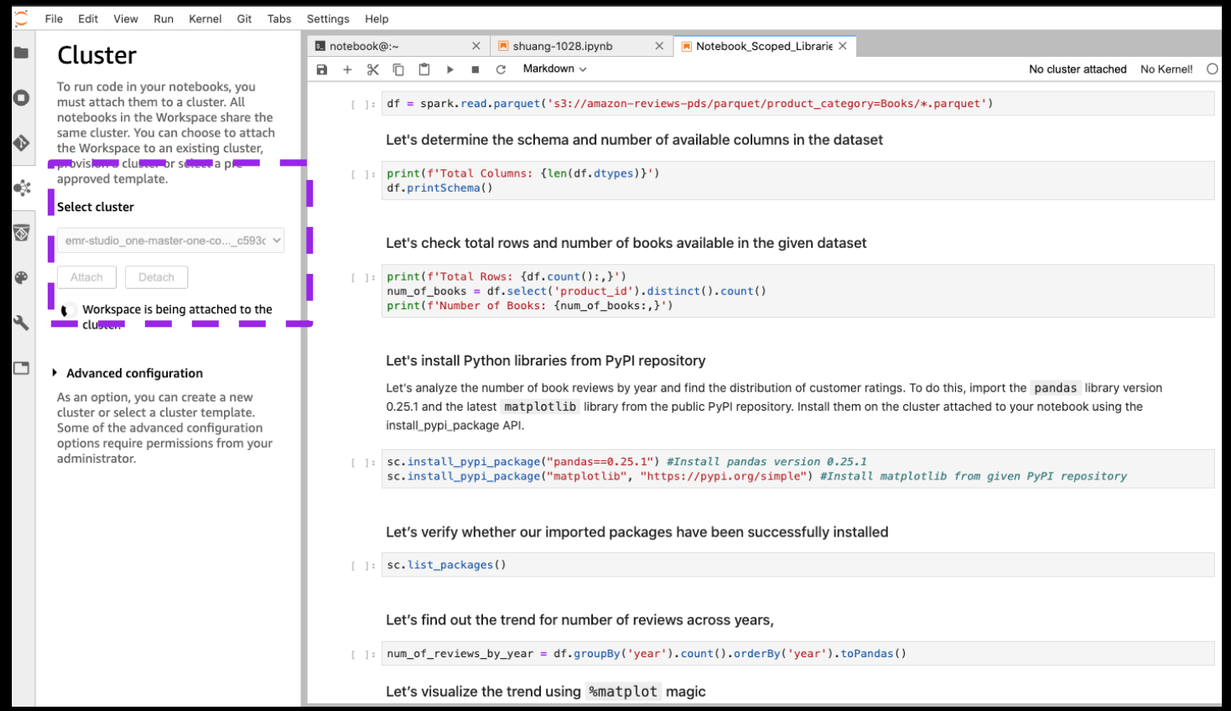
Run notebooks as pipelines via Amazon Managed Workflows for Apache Airflow (MWAA) NEW!

Parameterize and chain notebooks that can be run as pipelines

Detach from and re-attach to clusters

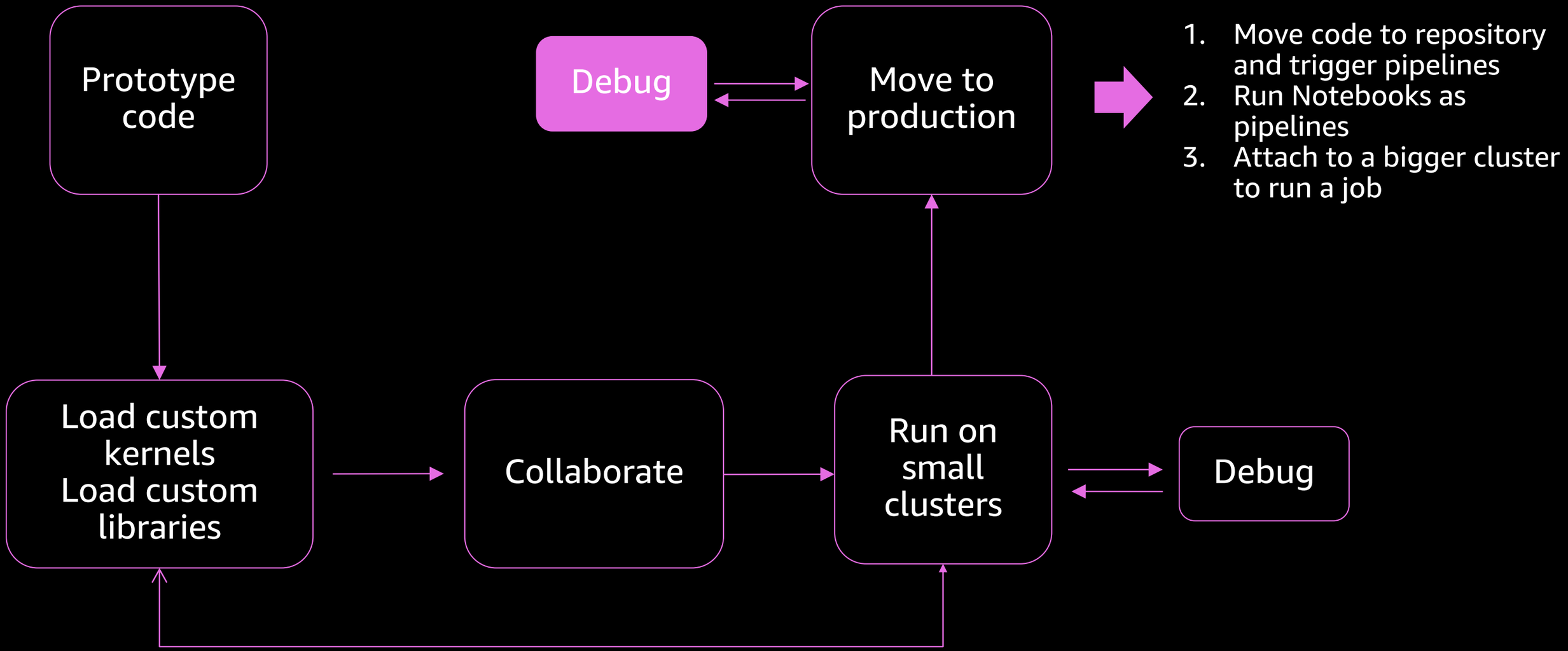


Detach workspace from a cluster

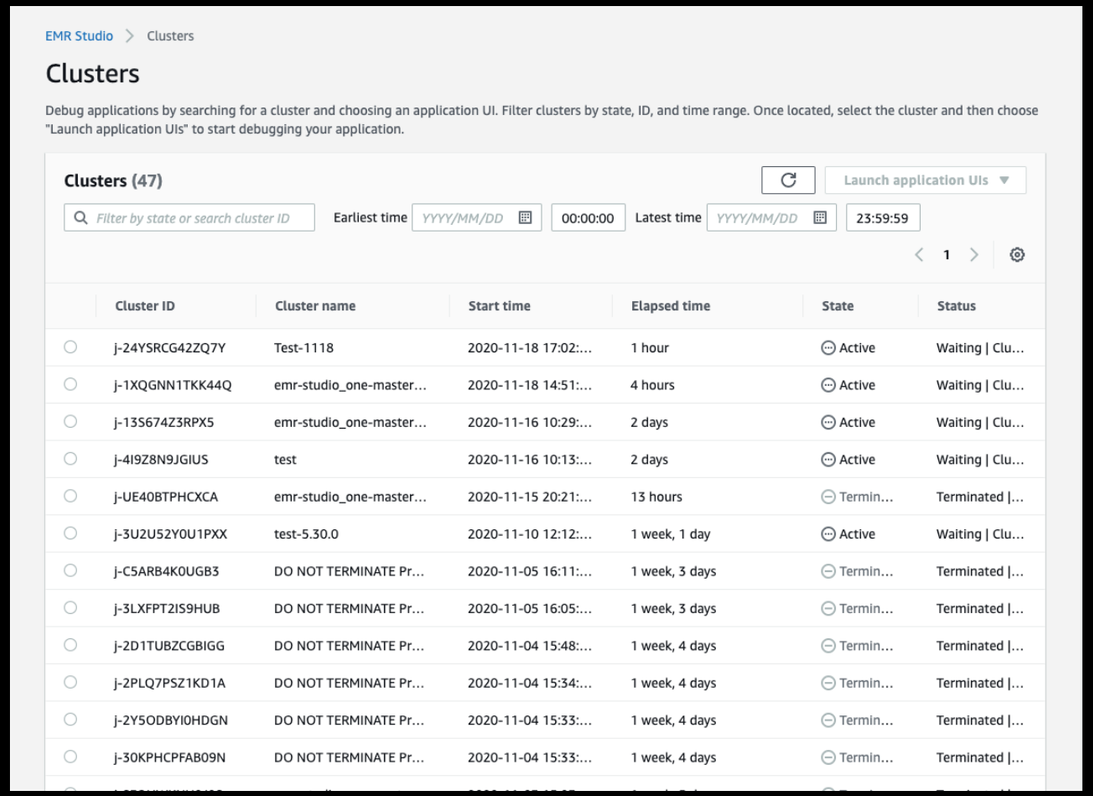


Re-attach workspace to another cluster

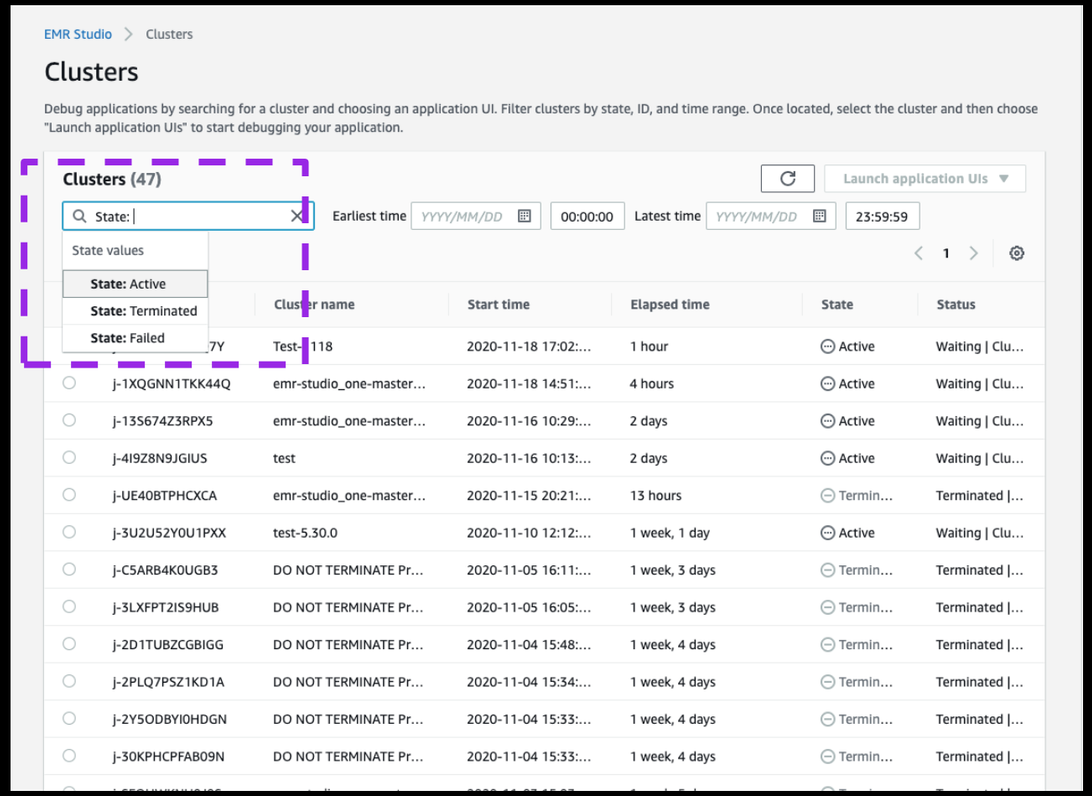
Data science and engineering workflows



Debugging production pipelines is easy

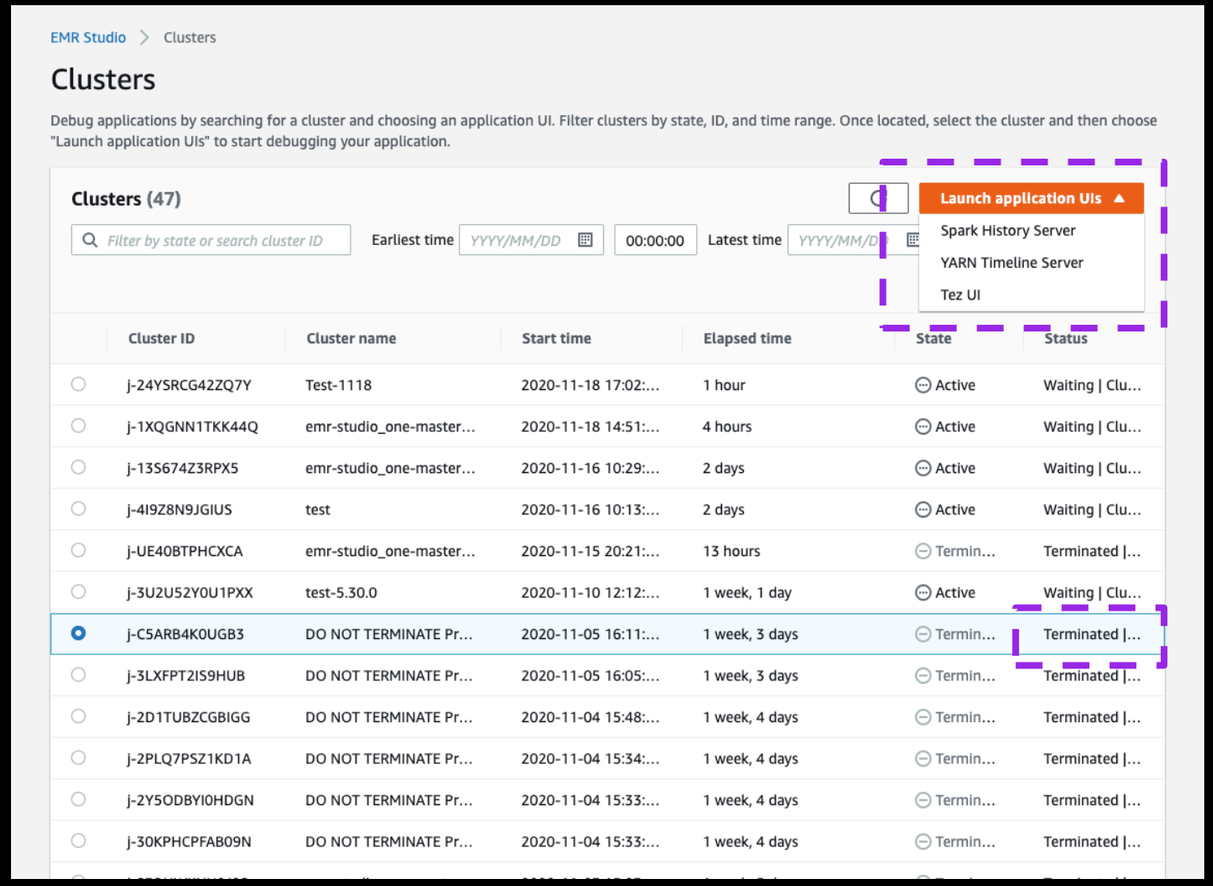


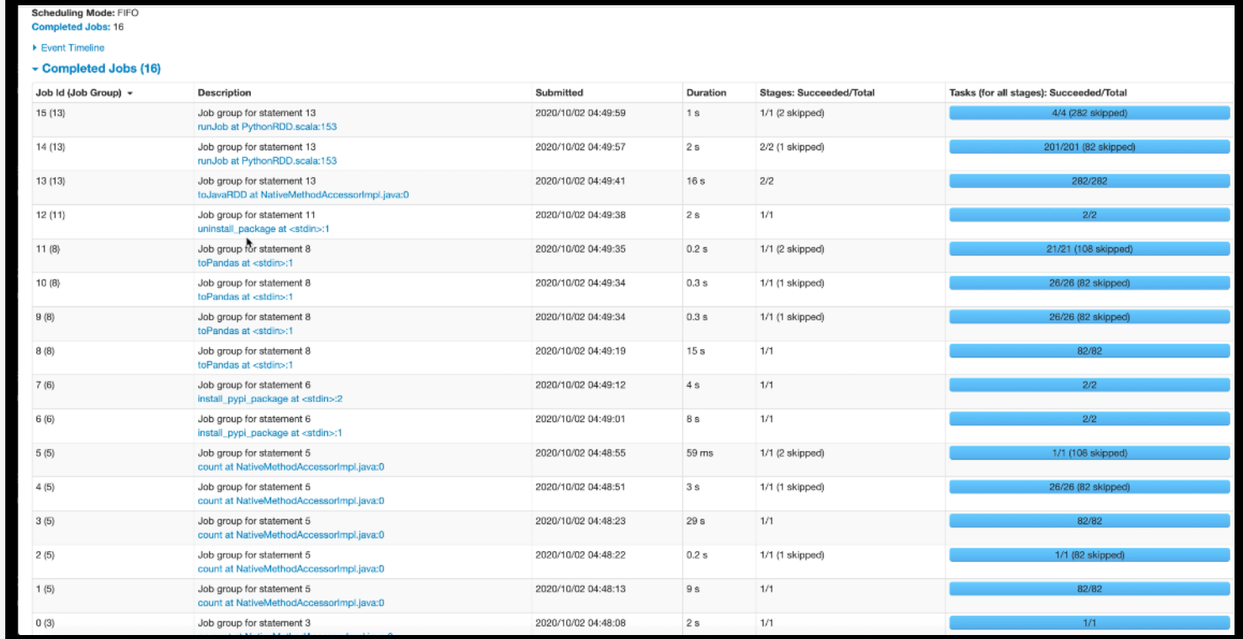
Browse all clusters in one place



Narrow down clusters for investigation using filters such as cluster state

Debugging production pipelines is easy





Diagnose jobs on both active and terminated clusters using Spark UI, Tez UI, and Yarn Timeline Service

Overlay execution context on jobs, even for terminated clusters and jobs

Demo



AWS re:Invent

© 2020, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Amazon EMR Studio provides simplicity with control



AWS re:Invent

© 2020, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Use your existing identity provider with AWS SSO

Easily enable AWS SSO in the AWS SSO console

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

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

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

Use roles and permission policies to scope down access for users and groups

Restrict resource access using roles and permission policies

- Enable permissions like create/edit workspaces, use cluster templates, create/link GIT repositories, create clusters, access to job diagnostics, in user role for users and groups assigned to a Studio
- Scope down access for users and groups in a Studio by assigning permission policies; for example, restrict some users to running jobs only on existing clusters
- Simply select a predefined permission policy, an existing IAM policy in your AWS account, or create your own policy in AWS IAM console with more flexibility and control

Use tags to enforce authorization and charge back

Tag clusters to restrict access for security and usage tracking

- Assign one division or one team to a Studio
- Authorize users and groups in a Studio to access clusters with a division tag so no other divisions can use those clusters for security
- Scope down permissions for users who can view cluster attributes but not launch application UIs to access logs with sensitive data, using tags
- Charge a division for running jobs on tagged clusters in a Studio

Thank you!



AWS re:Invent

© 2020, Amazon Web Services, Inc. or its affiliates. All rights reserved.



Please complete the session survey

AWS re:Invent

© 2020, Amazon Web Services, Inc. or its affiliates. All rights reserved.