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

ANT309-R

Build analytics applications using Apache Spark with Amazon EMR Serverless

Damon Cortesi

Principal Developer Advocate
AWS

Matthew Tan

Senior Analytics Specialist SA
AWS

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

Agenda

- What is EMR Serverless and when to use it
- GitHub Actions and automating software workflows
- EMR Serverless and GitHub Actions together
- Workshop
- Demo: VS Code and EMR extension

© 2022, 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

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 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

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

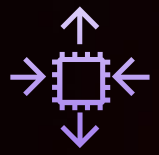
Amazon EMR Serverless

Simple to use

Easily run petabyte-scale data analytics in the cloud without managing, tuning, optimizing, securing, and operating clusters

Fast and highly available

Quickly run large-scale applications that use open-source frameworks of your choice, including Spark, Hive, and Presto



Highly scalable

Automatically scale resources up and down as needed based on the changing requirements of your application

Cost effective

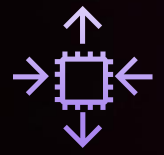
Pay only for the resources you use for data analytics at scale

aws

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

Amazon EMR Serverless

Use cases



Build scalable data pipelines

Extract and transform data at scale from a variety of sources

Accelerate data science and ML

Develop, debug, and visualize big data applications with EMR Studio

Process real-time data streams

Analyze events from streaming data sources in real time

Query any dataset

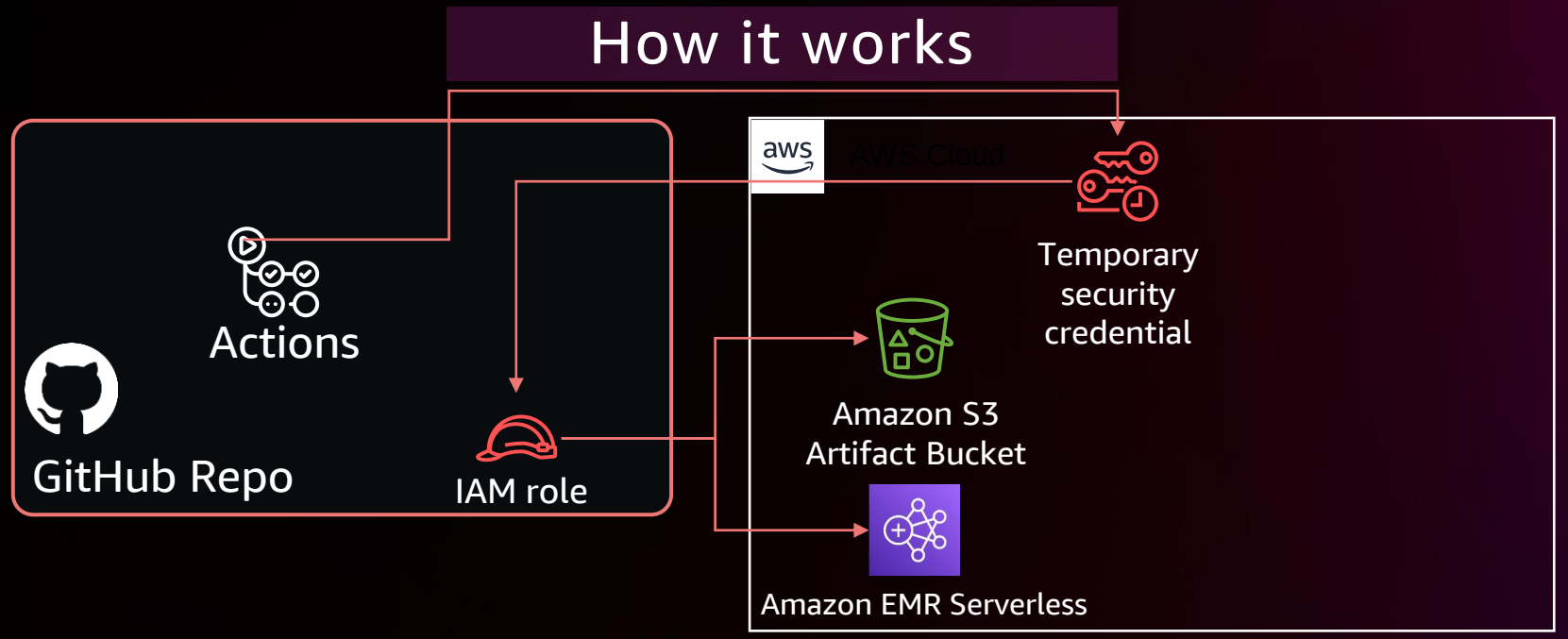
Query datasets interactively using Spark, Hive, and Presto

aws

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

GitHub Actions

GitHub Actions makes it easy to automate all your software workflows, now with world-class CI/CD. Build, test, and deploy your code right from GitHub.



OpenID Connect (OIDC) support in GitHub Actions enables secure cloud deployments using short-lived tokens that are automatically rotated for each deployment.

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

Amazon EMR Serverless and GitHub Actions unified

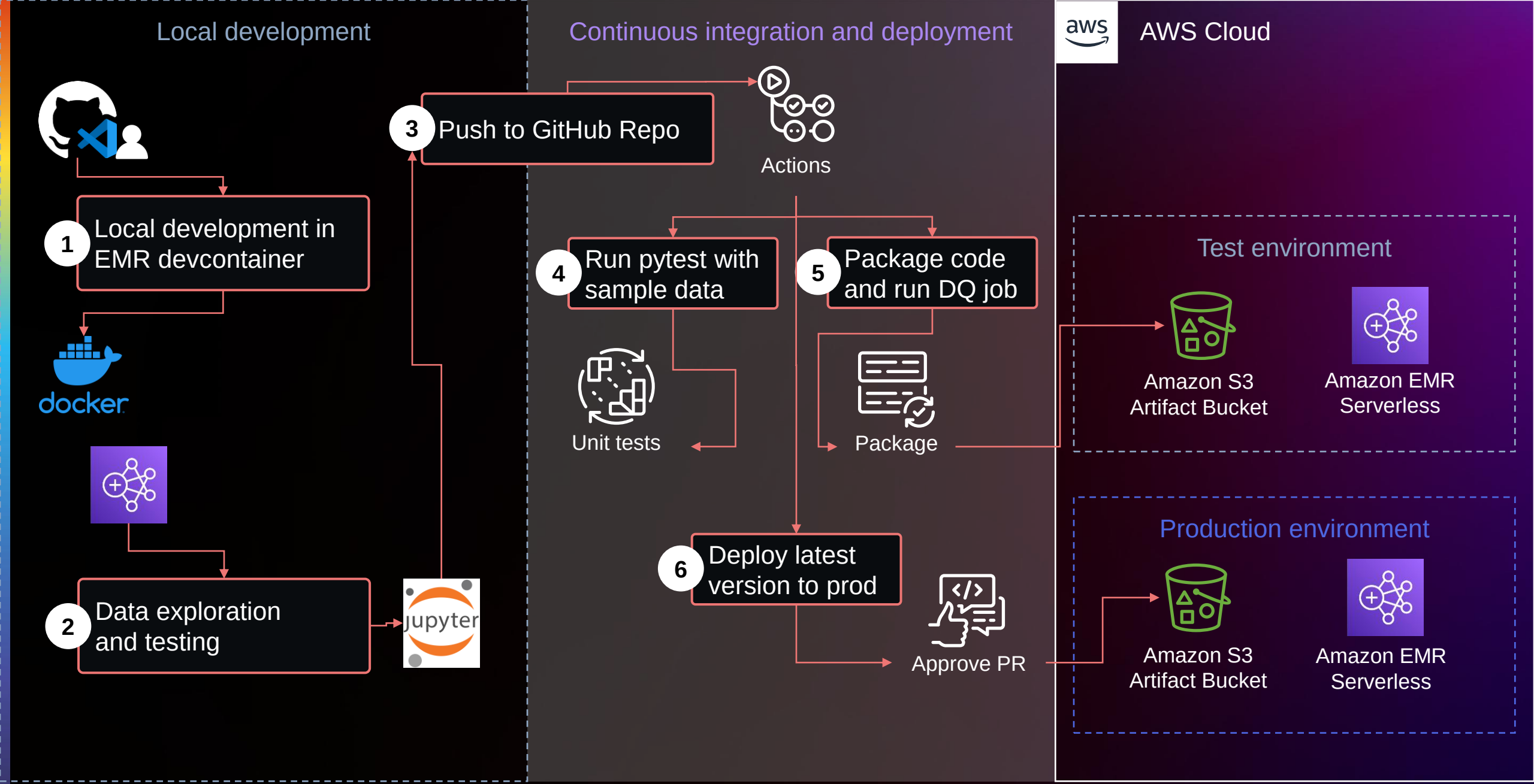
Data scientist

Data engineer

- Data scientists can use GitHub Actions to schedule and automate data quality tests and unit tests.
- Engineers and administrators need to build or manage resources across analytics and ML. They can use GitHub Actions to package and deploy versioned assets to Amazon S3 for use in Serverless Spark jobs. They also can use VS Code development containers to increase the pace of local development for Amazon EMR.

aws

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



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

Workshop details

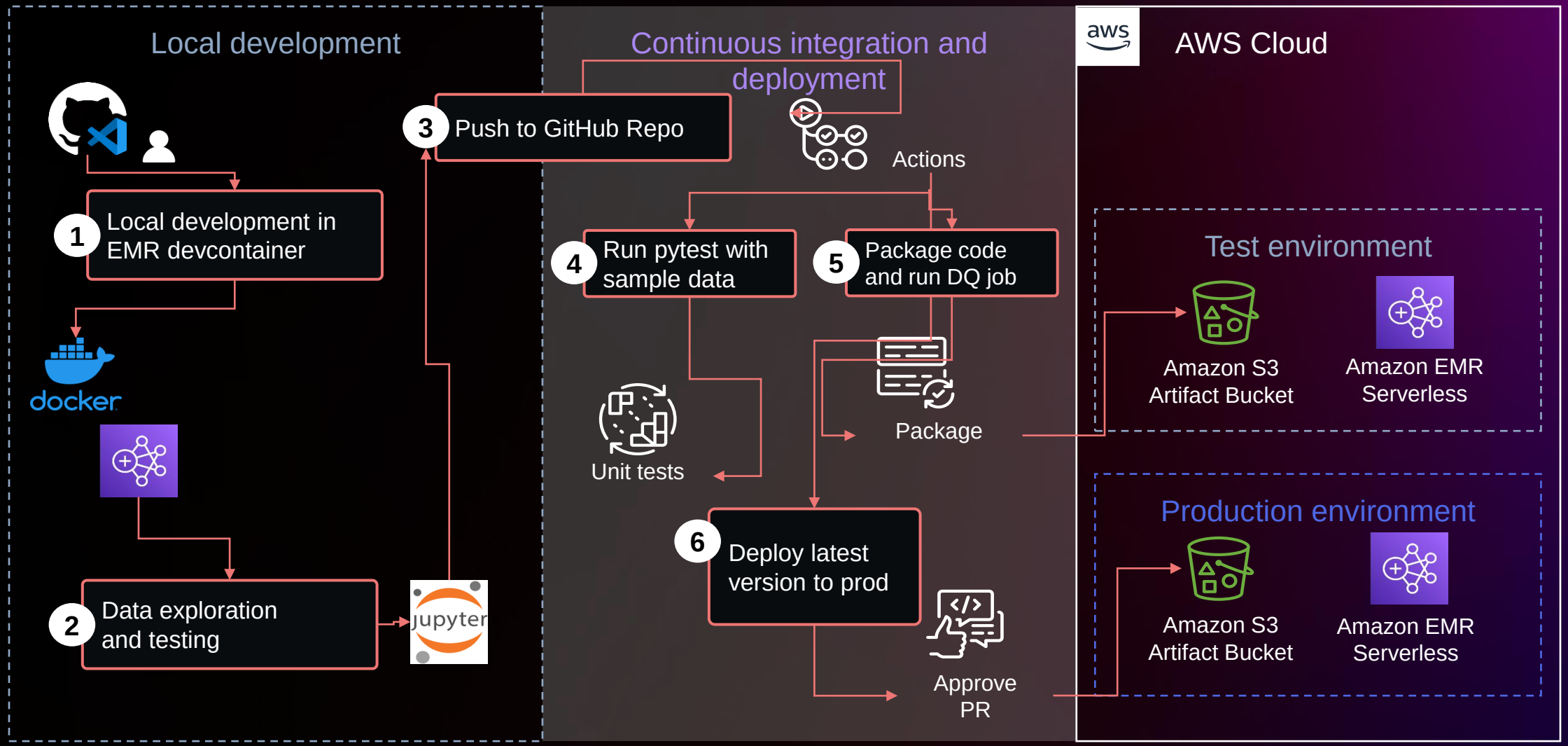
© 2022, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Workshop overview

- In this session, you will learn how to build a Spark-based application and run it using Amazon EMR Serverless.
- You will start with developing your application locally in VS Code, and use GitHub Actions to test and run the application on Amazon EMR Serverless.
- You will debug in-progress applications. You will also see how to configure unit tests when a push is made, run data quality checks when a pull request is opened, and automatically package and deploy new application releases.

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

Workshop architecture



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

Workshop access

Go to:

https://dashboard.eventengine.run

Event hash: To be provided at event

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

Workshop helpers

Name	Session
Vivek Shrivastava	Nov 28, 2022 – MGM Grand 117
Sean Ma	Nov 28, 2022 – MGM Grand 117
Indira Balakrishnan	Nov 28, 2022 – MGM Grand 117
Vivek Shrivastava	Dec 2, 2022 – Venetian Murano 3204
Sean Ma	Dec 2, 2022 – Venetian Murano 3204
Indira Balakrishnan	Dec 2, 2022 – Venetian Murano 3204
Dave Geyer	Dec 2, 2022 – Venetian Murano 3204
Tony Nguyen	Dec 2, 2022 – Venetian Murano 3204

aws
© 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.