EMR on EKS

Run Apache Spark on Kubernetes with Amazon EMR on Amazon EKS

Damon Cortesi, Principal Developer Advocate

March 22, 2021

aws

Agenda

- Amazon EMR overview
- Amazon EKS overview
- Apache Spark on EMR on EKS
- Demos
  - Running Spark jobs on EMR on EKS
  - How to monitor your jobs
  - Customizing your job performance
- Q&A

aws

Customers are running Big Data / Analytics workloads on Kubernetes

Highly Scalable

Better resource utilization

Increasingly diverse

Accessed by many teams

Connected by many applications

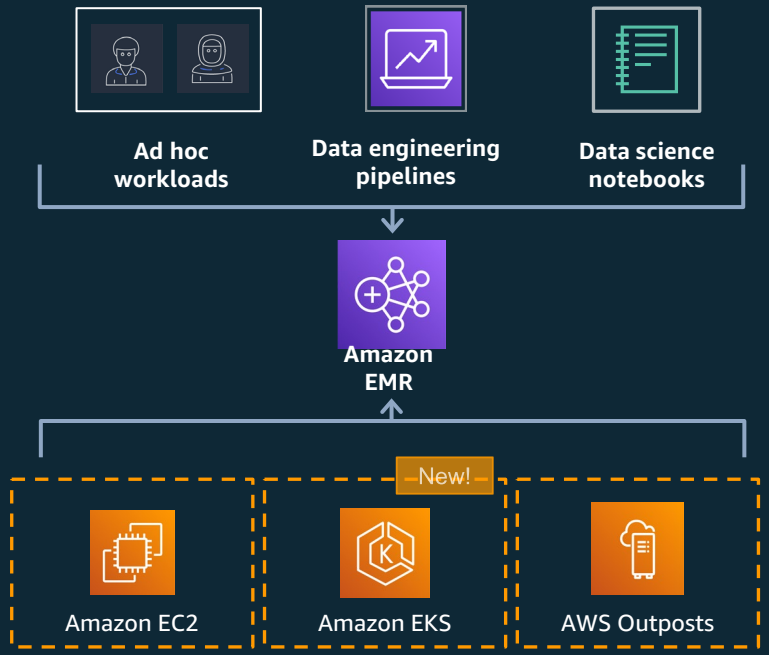
Open source Community

aws

However, managing and maintaining open source workloads is challenging

- Container build and deployment
- Job submission
- Spark UI accessibility
- Performance optimization
- Autoscaling considerations
- Security and patch management

A new deployment model for Amazon EMR



Run Amazon EMR on Amazon EKS

In addition to existing deployment modes

Simplifies running Spark on Kubernetes

Amazon EMR on Amazon EC2

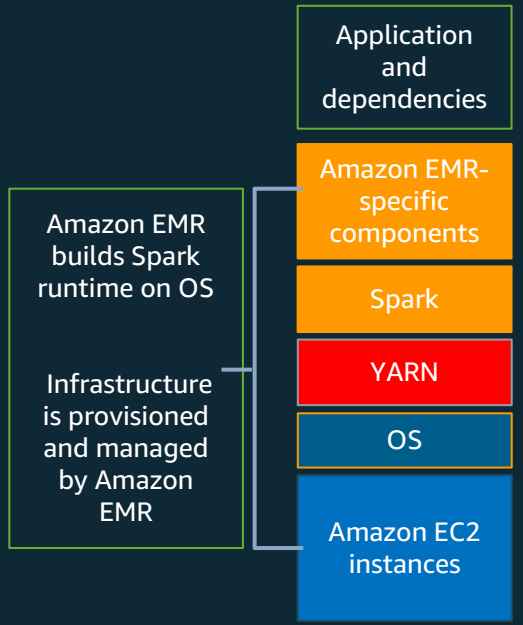
Control over instances drives cluster-centric model

Run single version of Spark per cluster

Shared execution role

Great for jobs with cluster-scoped dependencies

Great for clusters running at high utilization



aws

Amazon EMR on Amazon EKS

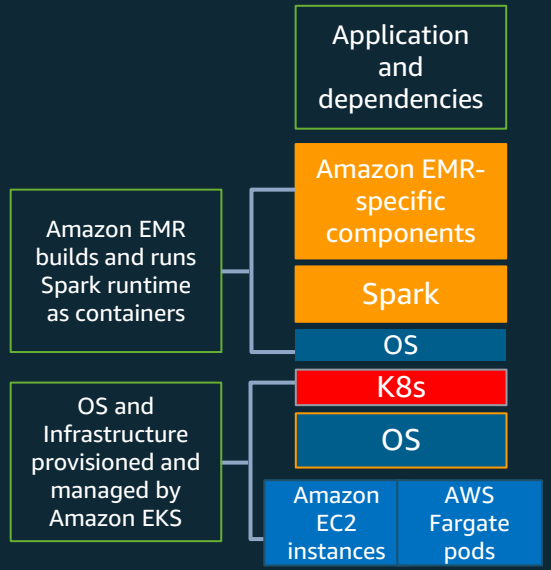
Containerization drives job-centric model

Run multiple versions of Spark per cluster

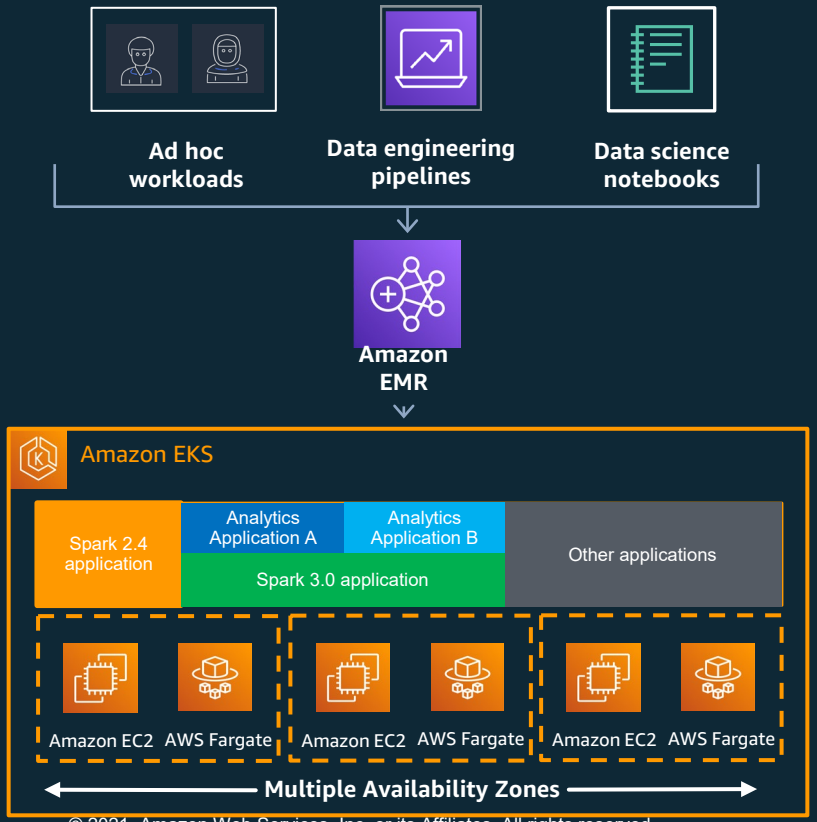
Per job execution role

Great for quick test and upgrade cycles

Great for consolidating resource utilization



Run Spark on Amazon EKS clusters



Register Amazon EKS clusters with Amazon EMR

Consolidate infrastructure across organization

Manage resource limits by teams and workload

Start jobs quickly, no cluster provisioning delays

Run application on single AZ or across multiple AZs

Choose serverless with AWS Fargate on Amazon EKS

aws

Amazon EMR helps accelerate move to EKS

Provide managed distribution Spark on Kubernetes (2.4 and 3.0)

Manage job execution your behalf

Simplify secure execution using granular access control

Native integration with Amazon S3, AWS Glue Data Catalog, and more…

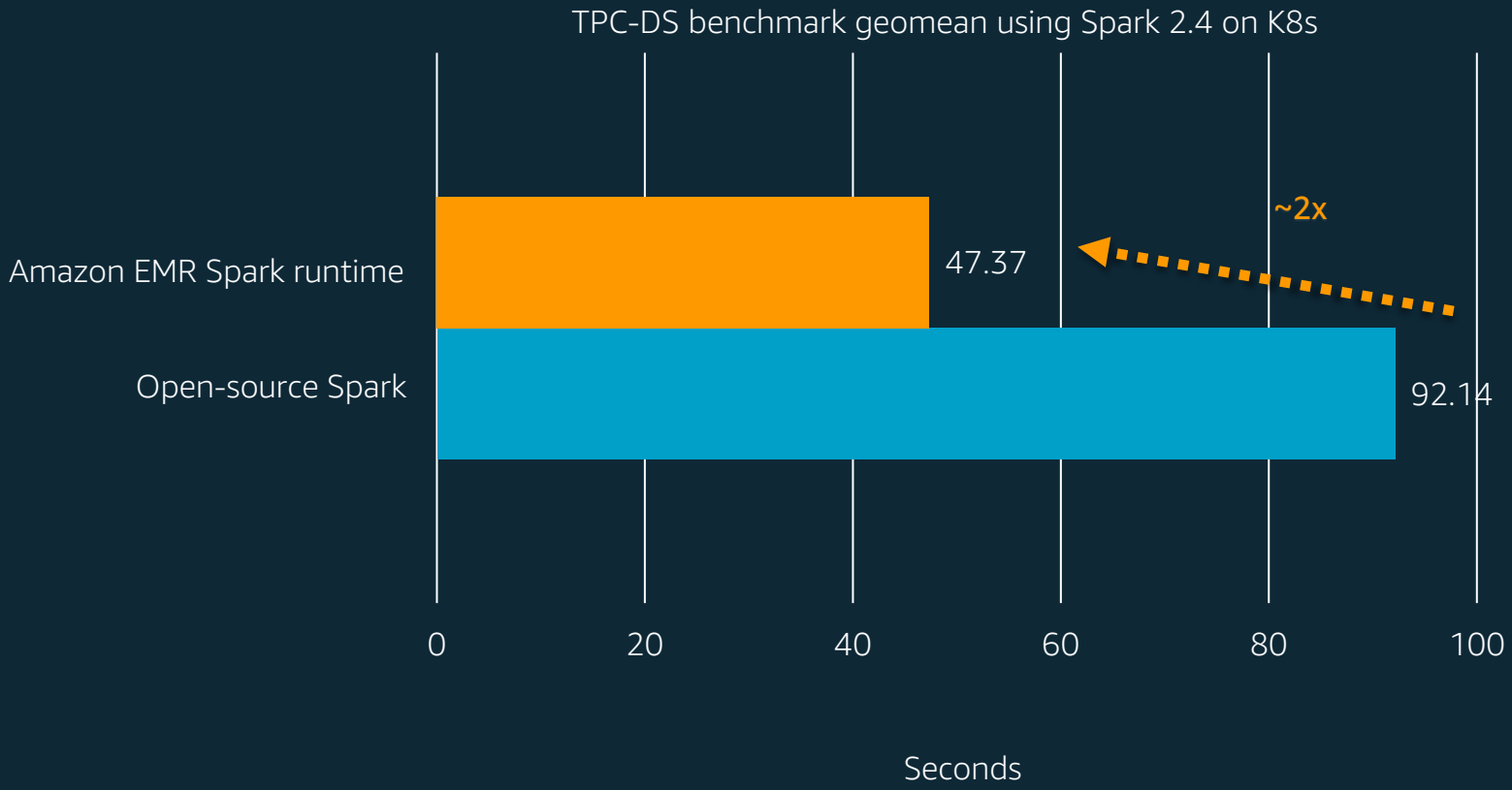
Simplify debugging with Spark History Server

Support integration with Apache Airflow

Differentiated performance with Amazon EMR runtime for Apache Spark

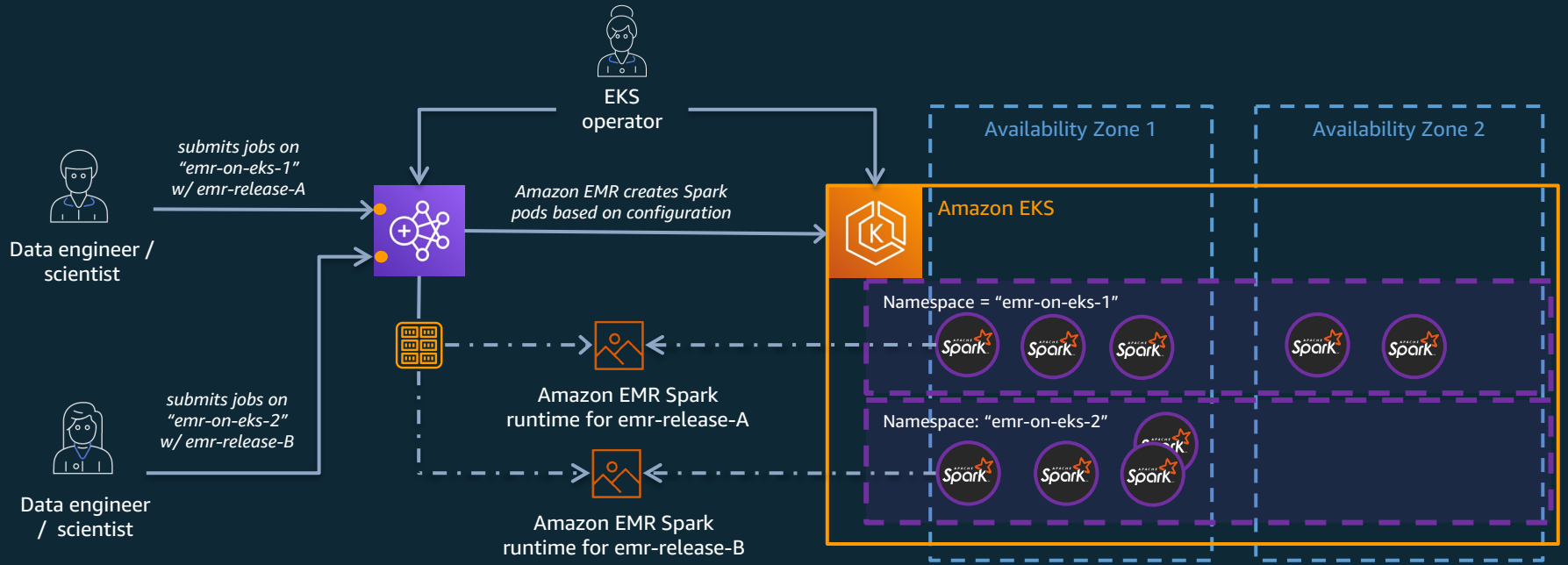
aws

Spark performance with Amazon EMR runtime



Example Architecture

Job-centric workflow



Job submission options

AWS Command Line Interface

AWS Tools and AWS SDK

Apache Airflow

AWS Step Functions (coming soon)

aws

Job debugging options

Amazon CloudWatch

Amazon Simple Storage Service (Amazon S3)

AWS Management Console

Amazon EMR Studio (Preview)

Demo – Job Creation and App UI

Demo – S3 and CloudWatch Logs

Demo – Job Optimization

Thank you!

Join us for the 2021 AWS Summit Online
https://aws.amazon.com/events/summits/online/americas/