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

ANT330-R

Run Apache Spark on Kubernetes with Amazon EMR on Amazon EKS

Melody Yang

Sr. Big Data Architect, EMR
AWS

Lorenzo Ripani

Big Data Architect, EMR
AWS

aws

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

Agenda

Amazon EMR on EKS overview

Getting started and demo

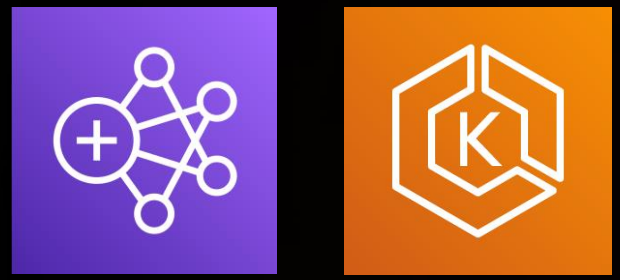
Cost and performance benefit, demo

Ease of use and demo

Q&A

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

Amazon EMR on Amazon EKS



Run Apache Spark jobs using **Amazon EMR on Amazon Elastic Kubernetes Service (Amazon EKS)** to improve resource utilization and simplify infrastructure management

aws

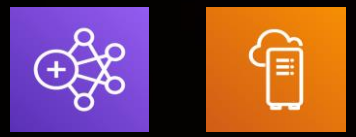
© 2022, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Amazon EMR deployment options



Amazon EMR on Amazon EC2

Choose instances that offer the best price performance for your workload



Amazon EMR on AWS Outposts

Set up, manage, and scale Amazon EMR in your on-premises environments, just as you would in the cloud



Amazon EMR on Amazon EKS

Automate provisioning, management, and scaling of Apache Spark jobs on Amazon EKS



Amazon EMR Serverless

Run petabyte-scale data analytics in the cloud without managing and operating clusters

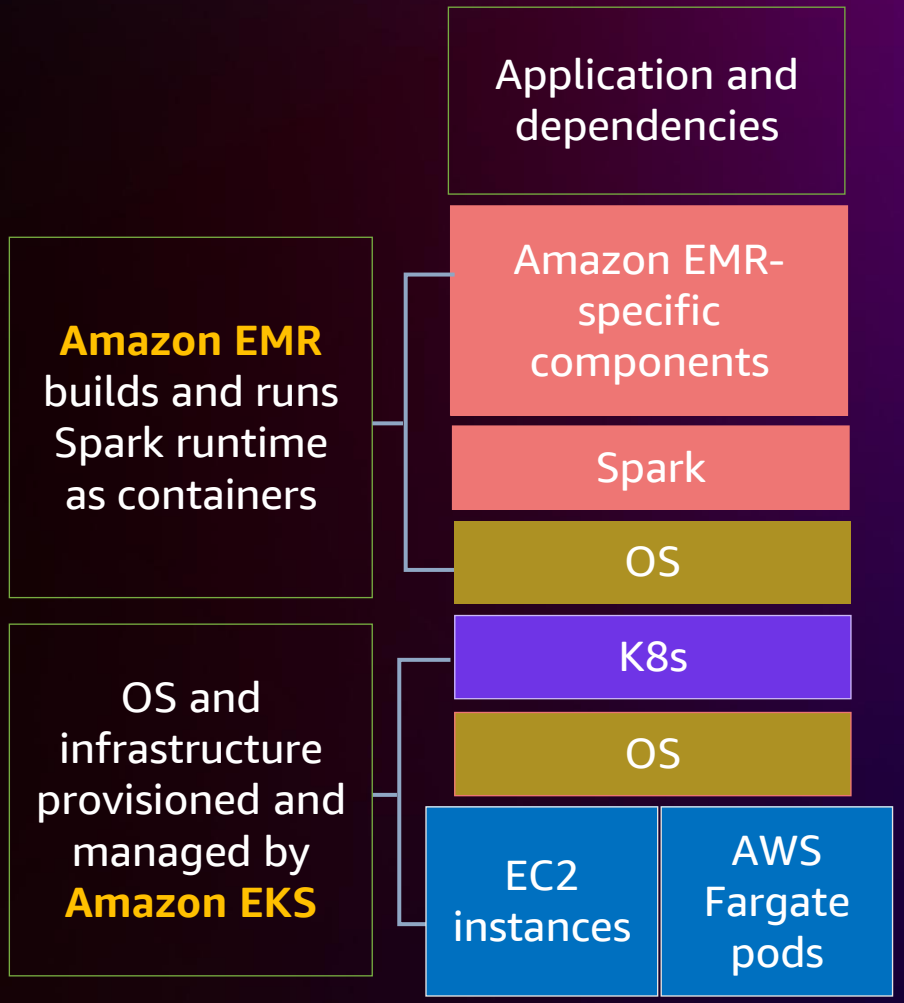
aws

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

Simplified operation

CONSOLIDATE ANALYTICS WORKLOADS WITH OTHER WORKLOADS ON AMAZON EKS

- Simplify infrastructure management
- Consolidate multiple versions of Spark on same Amazon EKS cluster
- Simplify Spark application upgrades
- Add multi-AZ resiliency by Amazon EKS with worker nodes across multiple AZs



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

Getting started and demo

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

Two core concepts



Virtual cluster

- Registered handle to a K8s namespace
  - Does not create any resources
- Can register multiple Amazon EKS cluster



Job run

- Represents work to be submitted
  - Spark JAR, PySpark, Spark SQL
- Integration with Apache Airflow

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

Getting started is easy

VIRTUAL CLUSTER = K8S NAMESPACE ON AMAZON EKS

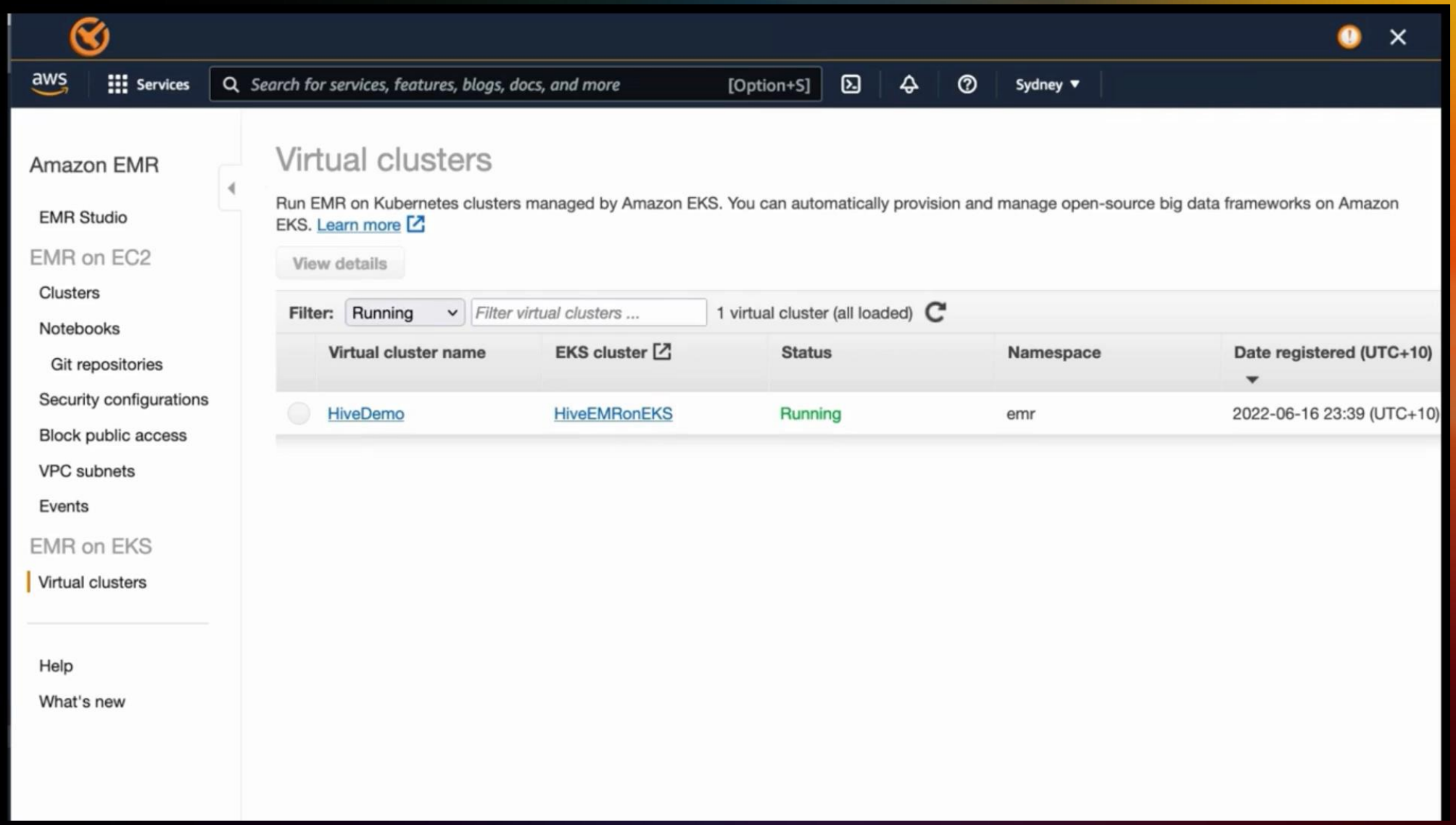


Create virtual cluster

```
$ aws emr-containers create-virtual-
cluster \
--name <virtual_cluster_name> \
--container-provider '{
  "id": "<eks_cluster_name>",
  "type": "EKS",
  "info": {
    "eksInfo": {
        "namespace": "<namespace_name>“
    }}
}'
```

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

Demo 1: Create a virtual cluster



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

Getting started is easy

New

Submit Spark job

```
$ aws emr-containers start-job-run \
--name <job_name> \
--virtual-cluster-id <cluster_id> \
--execution-role-arn <IAM_role_arn> \
--release-label <emr_release_label> \
--job-driver '{
  "sparkSubmitJobDriver": {
    "entryPoint": <entry_point_location>,
    "entryPointArguments": ["<arguments_list>"],
    "sparkSubmitParameters": <spark_parameters>
  }
}'
```

Submit Spark SQL script

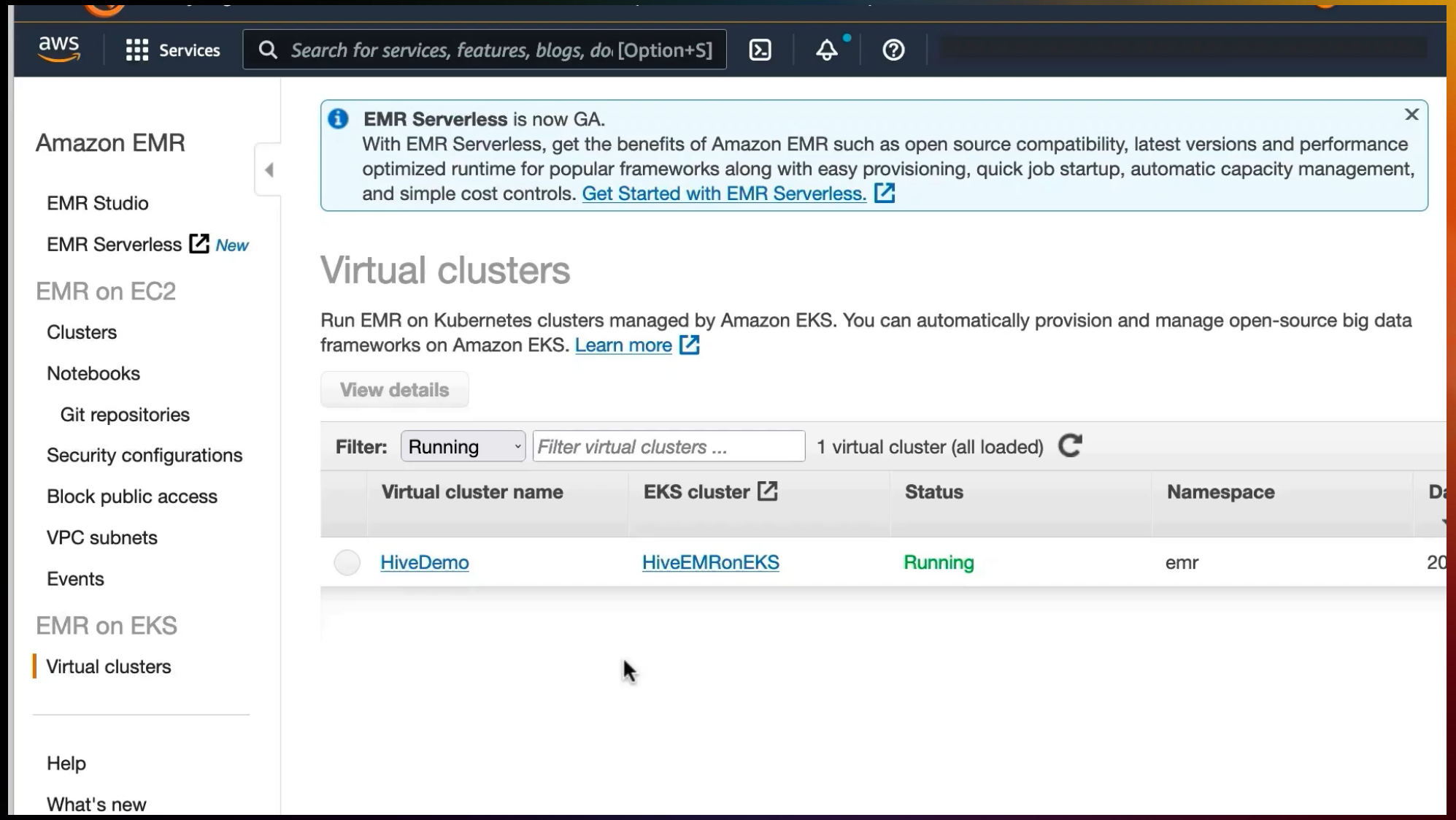
```
$ aws emr-containers start-job-run \
--name <job_name> \
--virtual-cluster-id <cluster_id> \
--execution-role-arn <IAM_role_arn> \
--release-label <emr_release_label> \
--job-driver '{
  "sparkSqlJobDriver": {
    "entryPoint": <entry_point_location>,
    "sparkSqlParameters": <spark_parameters>
  }
}'
```



aws

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

Demo 2: Submit a Spark SQL job



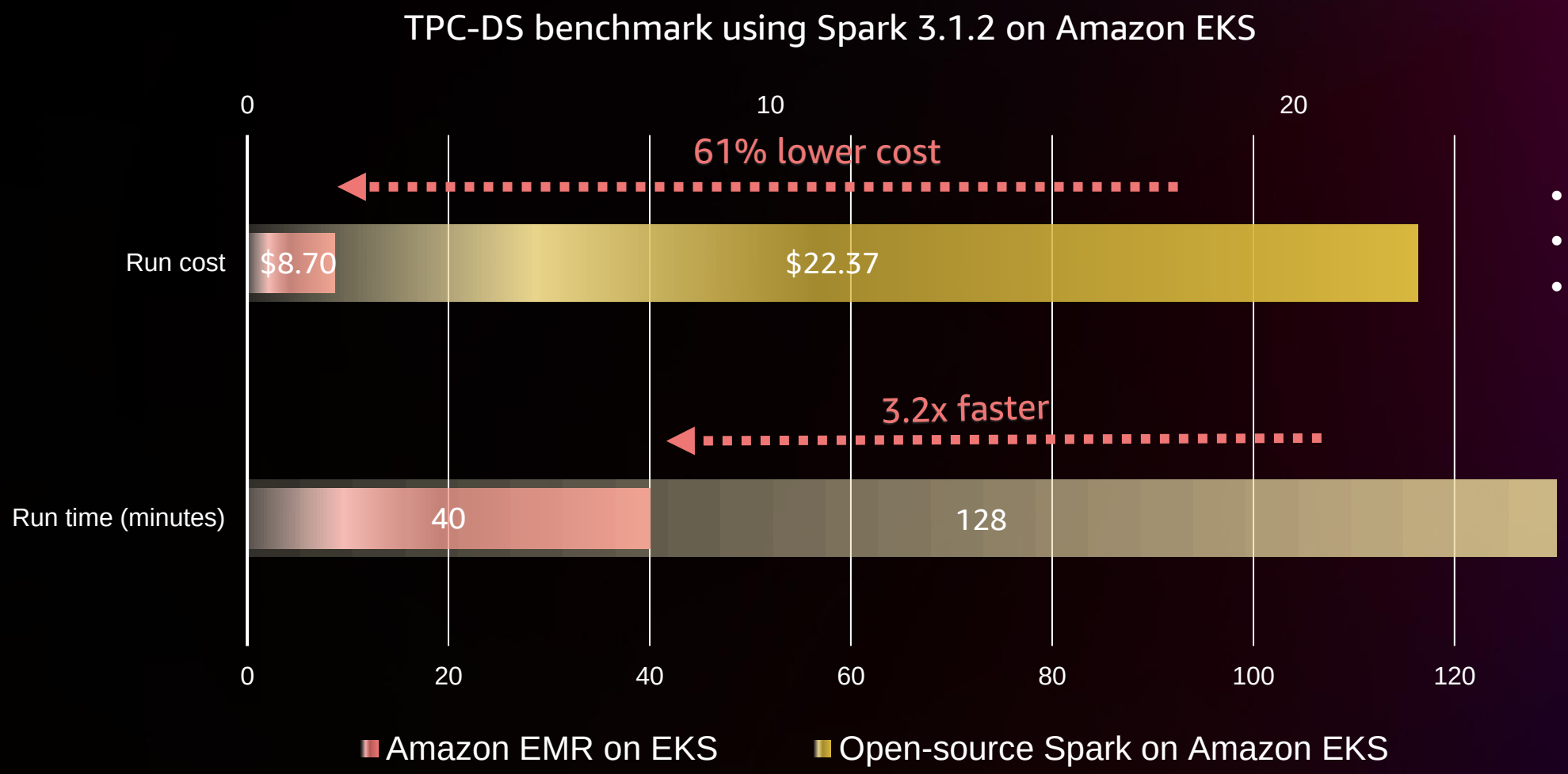
© 2022, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Cost and performance benefits, demo

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

Performance impacts cost: Amazon EMR vs. OSS

SPARK APPLICATIONS FINISH FASTER, RESULTING IN LOWER TCO AS WELL AS FASTER TIME TO INSIGHTS

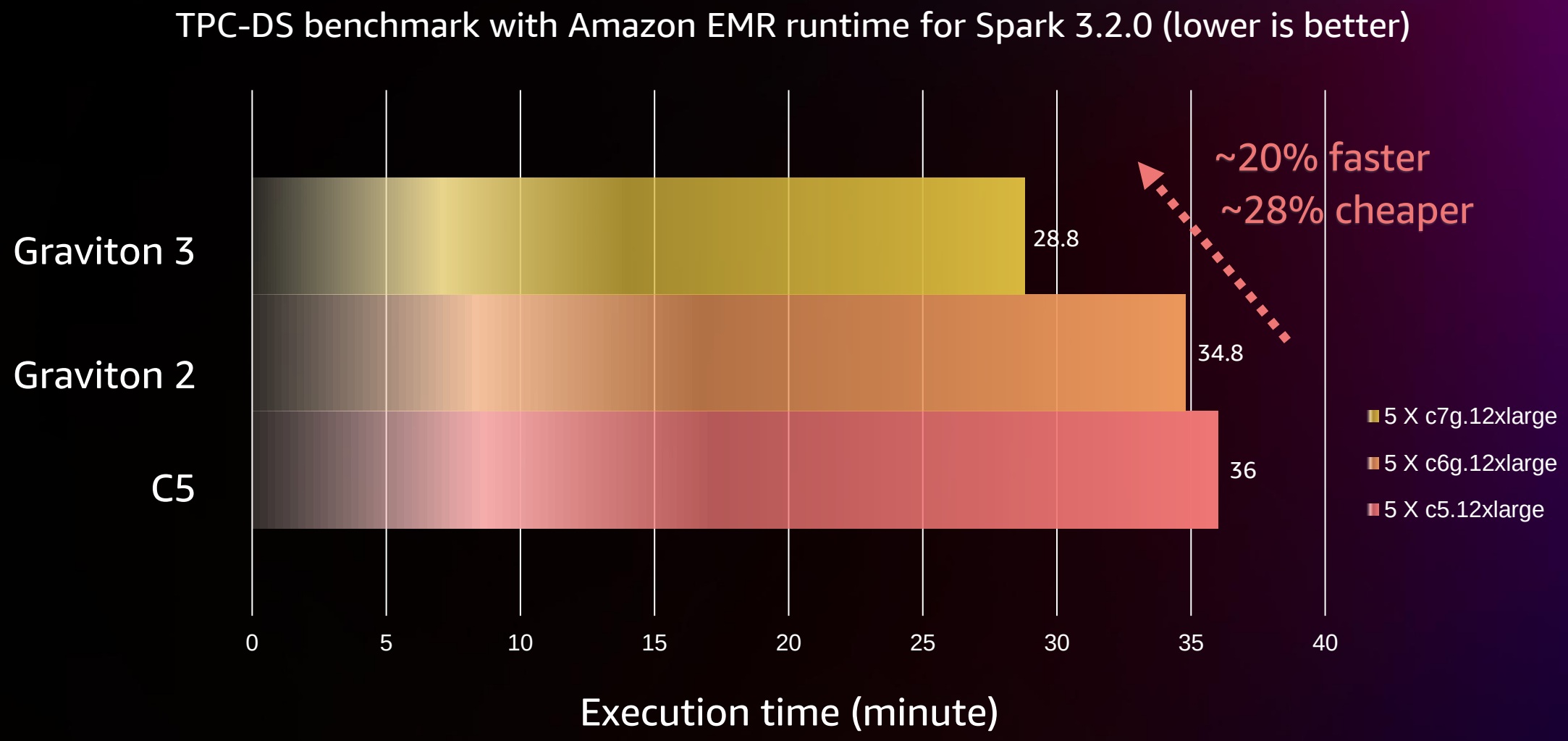


- 3 TB source data
- 104 test queries
- 6 X c5d.9xlarge

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

Source: AWS blog, https://aws.amazon.com/blogs/big-data/amazon-emr-on-amazon-eks-provides-up-to-61-lower-costs-and-up-to-68-performance-improvement-for-spark-workloads/

Further performance improvement



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

Source: AWS blog, https://aws.amazon.com/blogs/big-data/amazon-emr-on-eks-gets-up-to-19-performance-boost-running-on-aws-graviton3-processors-vs-graviton2/

Pricing

EMR uplift price is based on resources used by the job

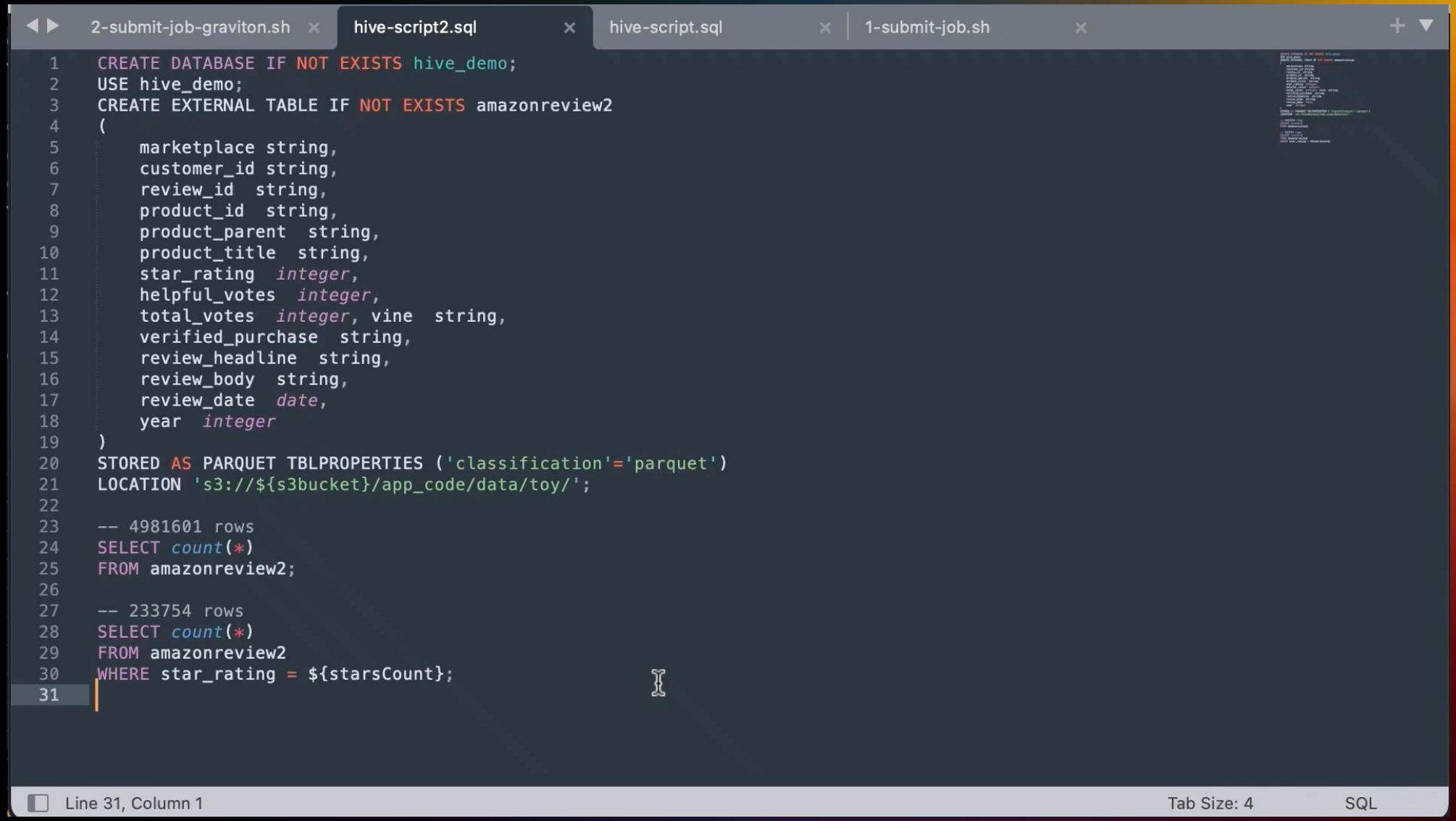
- VCPU-hour = $0.01 USD per VCPU-hour
- Memory-hour = $0.001 USD per GB-hour
- Per-second billing
- Minimum 1 minute

Additional costs

- Amazon EKS cluster: $0.10 USD per hour per EKS cluster
- EC2 instances (on demand/spot)
- AWS Fargate (vCPU and memory used), minimum 1 minute
- Per-second billing

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

Demo 3: Run a job on AWS Graviton

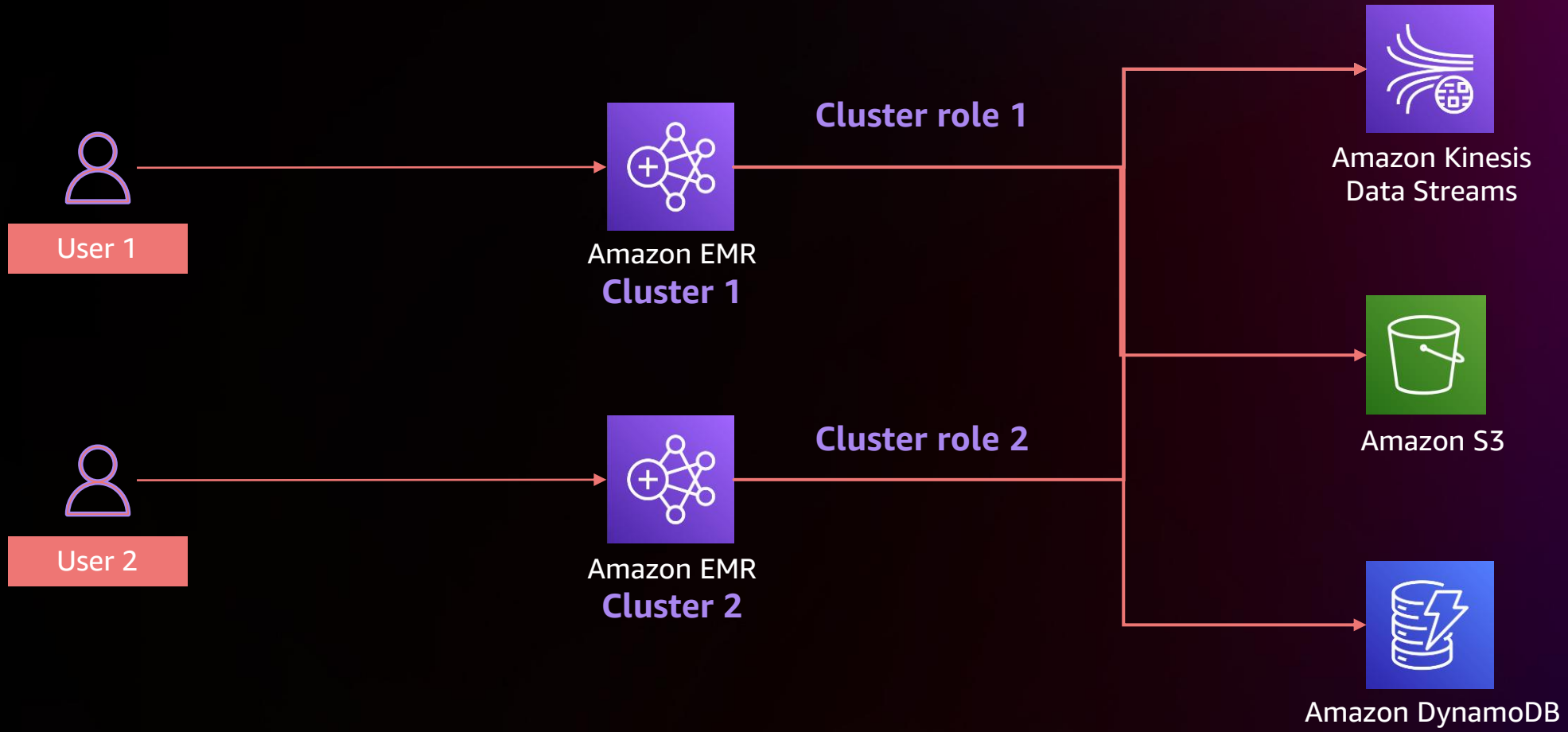


© 2022, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Ease of use and demo

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

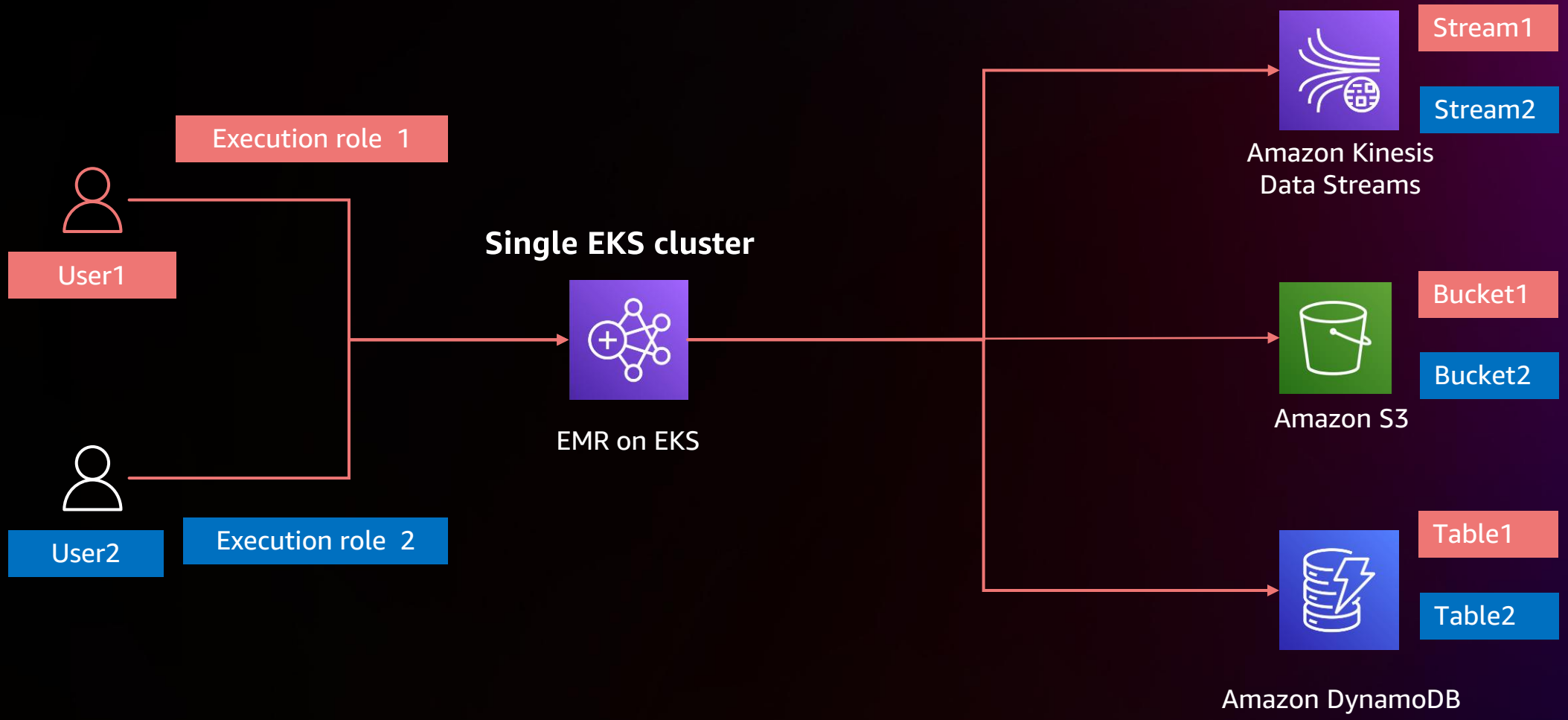
Spark on Yarn: One cluster, one role



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

Granular security control: Amazon EMR on EKS

JOB SCOPED EXECUTION ROLES ENABLE FINE-GRAINED ACCESS CONTROLS BY DEFAULT



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

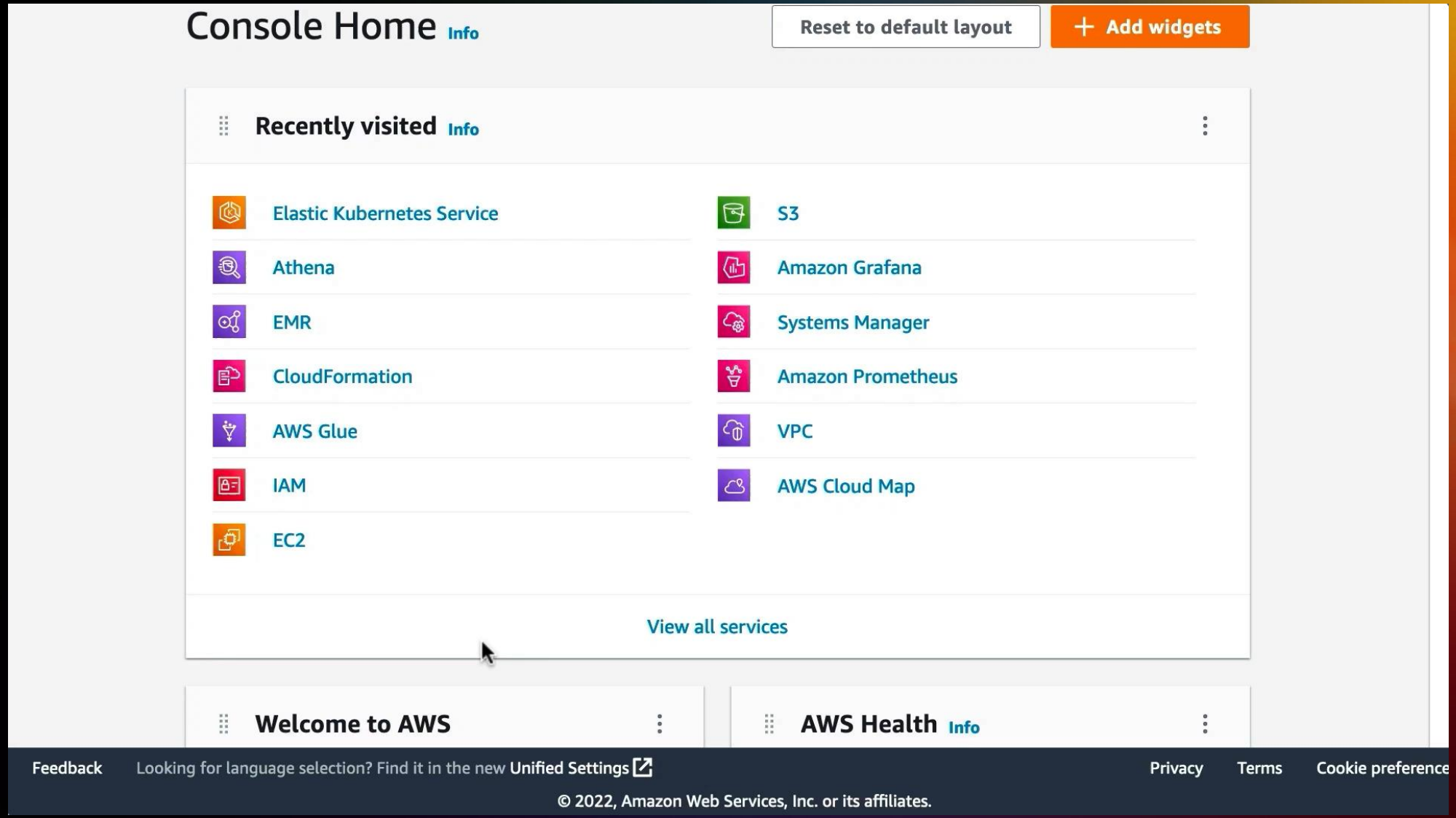
Auto scaling: Infrastructure

Cluster Auto Scaling

- Kubernetes cluster auto scaler
  - Use Amazon EC2 Auto Scaling group to scale in AWS
- Karpenter
  - Open-source high-performance Kubernetes cluster auto scaler
  - Removes Amazon EC2 Auto Scaling group
  - Evaluates pending pods and chooses optimal instance type
- AWS Fargate (serverless)

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

Demo 4: Switch to Fargate (serverless)



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

AWS Jam: Data and analytics

Hands-on lab: Spark on Amazon EKS made easy

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

“With Amazon EMR on EKS and the ARM-based AWS Graviton 2 instances, we improve the overall performance of big data operations by 30% and reduce the cost by 20%.”

Li Rui

Vice Technology President of Technology, Mobiuspace

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

Thank you!

Melody Yang
meloyang@amazon.com

Lorenzo Ripani
ripani@amazon.com



Please complete the session survey in the **mobile app**

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