INNOVATE

DATA EDITION

23 August, 2022

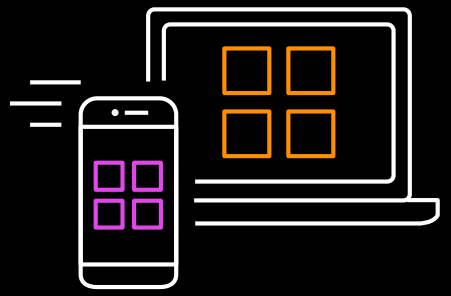
Modernize Spark workloads with Amazon EKS for better price performance

Melody Yang
Senior Big Data Architect, Amazon EMR
Amazon Web Services

Agenda

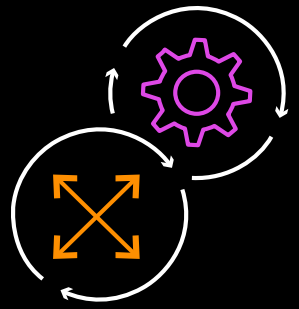
- Customer‘s ask
- Modernization options
- Cost and performance advantages
- Reference architecture
- Demo: Run Hive SQL script with Amazon EMR on Amazon Elastic Kubernetes Service (Amazon EKS)
- Takeaways

What customers ask for

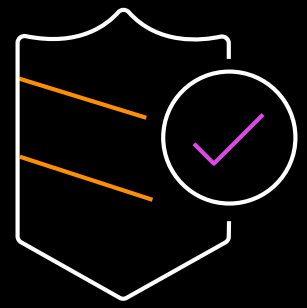


Build applications, not infrastructure

Manage infrastructure to their requirements



Scale quickly and seamlessly



Security and isolation by design

Benefits of adopting containers

Reduced risk		Uniform security across environment, maintained with automation
Operational efficiency		Reduced operational burden by removing undifferentiated heavy lifting
Speed		Consistent environment improves developer velocity
Agility		Automation increases speed and ease of testing and iterating

Some customer feedback

"We run Spark jobs on a common Kubernetes cluster that is provisioned at the minimum capacity. Now our jobs can *start up within 10 seconds* because they can use the capacity already available on the EKS cluster "

"We can now *use any version of Spark or Hive* with our applications, and run *on a single Amazon EKS cluster* that is centrally managed and kept up to date with the latest security updates. *This has simplified operations and reduced our cost* "

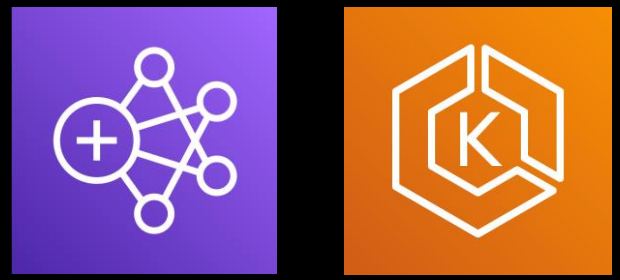
Modernization options

Open flexibility

Amazon Elastic Kubernetes Service (Amazon EKS)

- Gain agility and efficiency with AWS-optimized Kubernetes, and standardize operations everywhere
- Secure, highly available, with observability across all Kubernetes deployments
- Build with choice of solutions from the broader community around Kubernetes

Amazon EMR on Amazon EKS

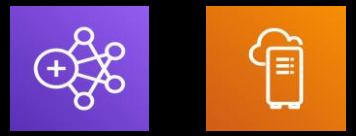


Run Apache Spark jobs using **Amazon EMR on Amazon Elastic Kubernetes Service (Amazon EKS)** - to improve resource utilization and simplify infrastructure management.

Amazon EMR deployment options

Amazon EMR on Amazon EC2

Choose instances that offer the best price performance for your workload

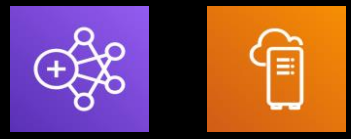


Amazon EMR on AWS Outposts

Set up, manage, and scale 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 EKS



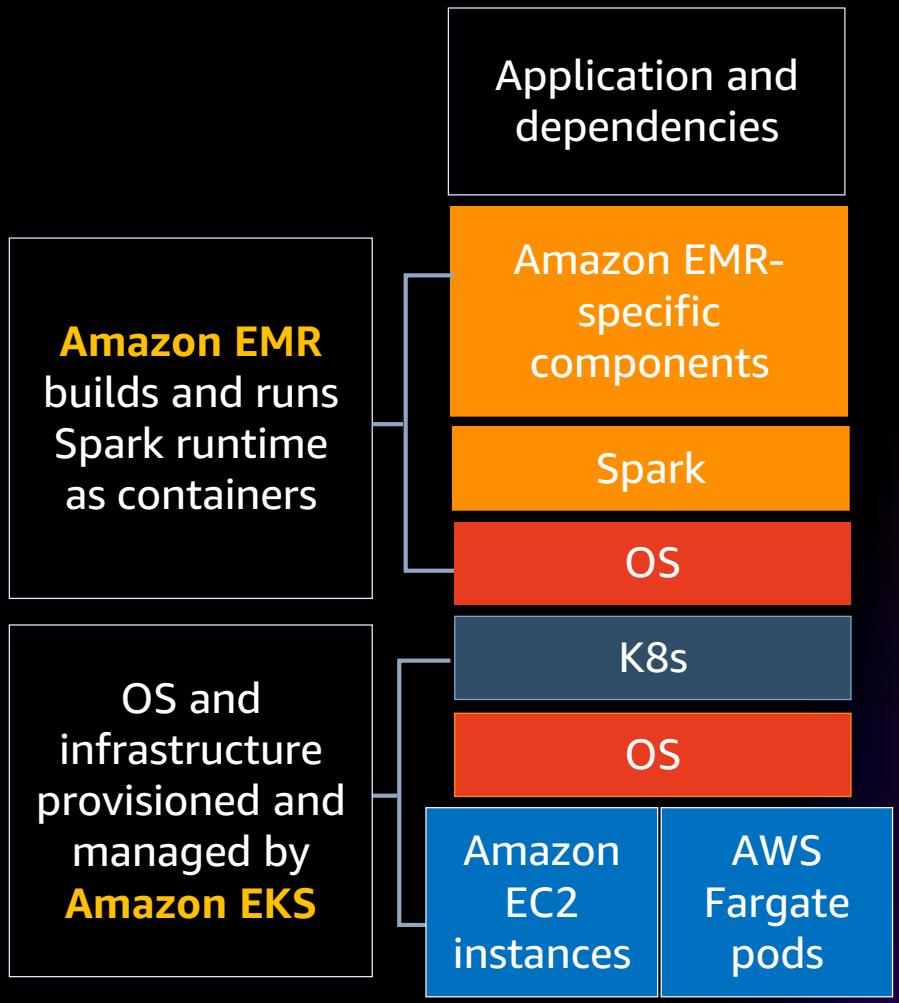
Amazon EMR Serverless

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

Run and scale anywhere

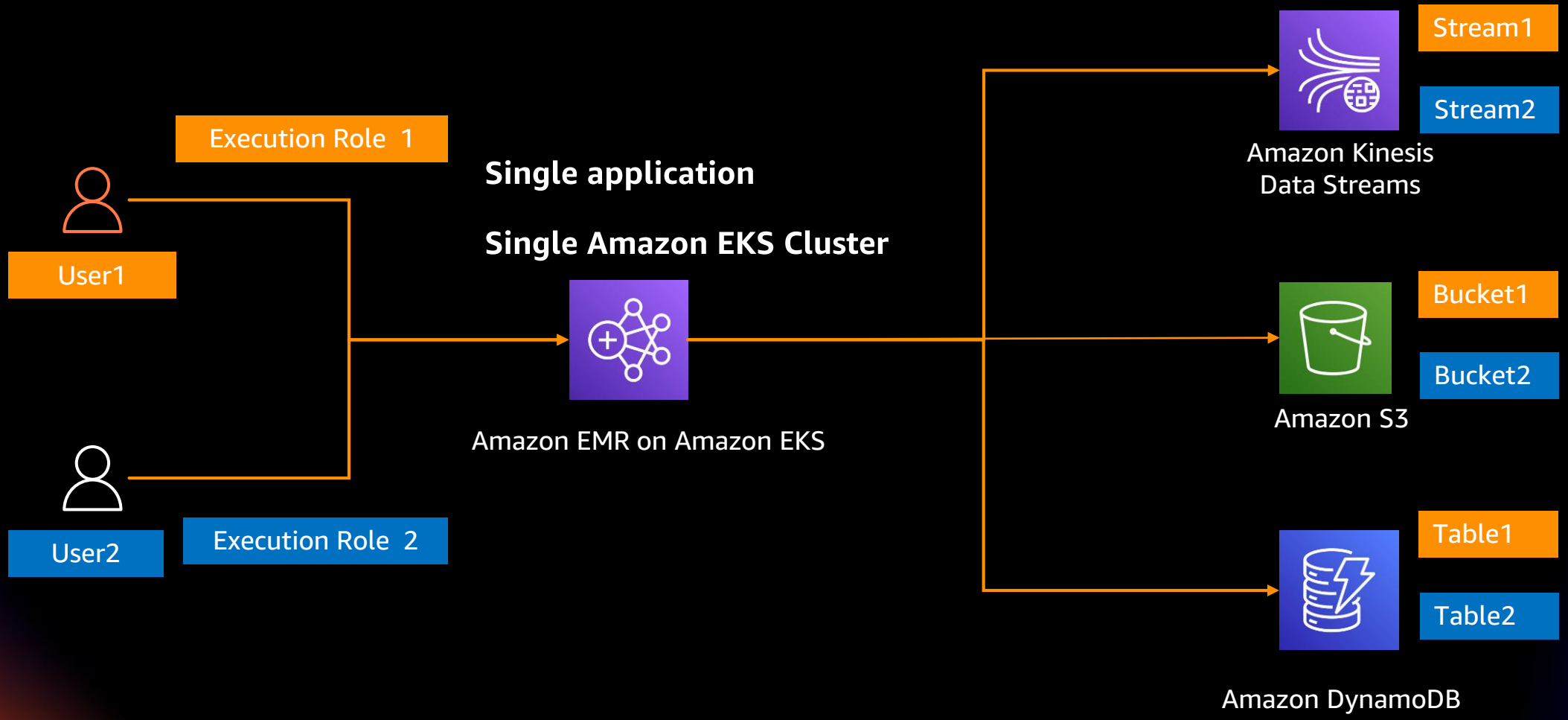
Consolidate analytics workloads with other workloads on Amazon EKS

- 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



Granular security control - Amazon EMR on Amazon EKS

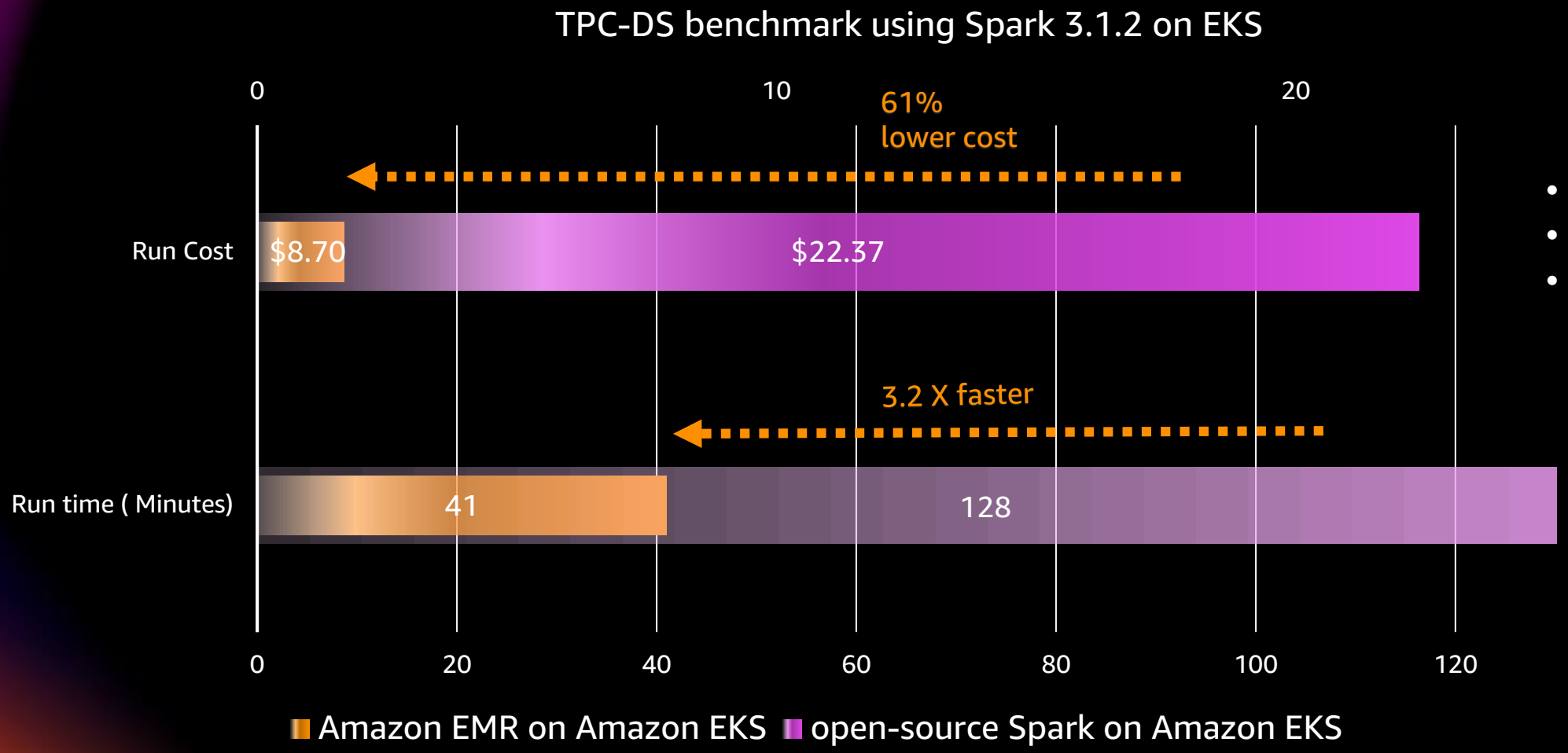
Job scoped execution roles enable fine grained access controls by default



Cost & performance advantages

Performance impacts on cost

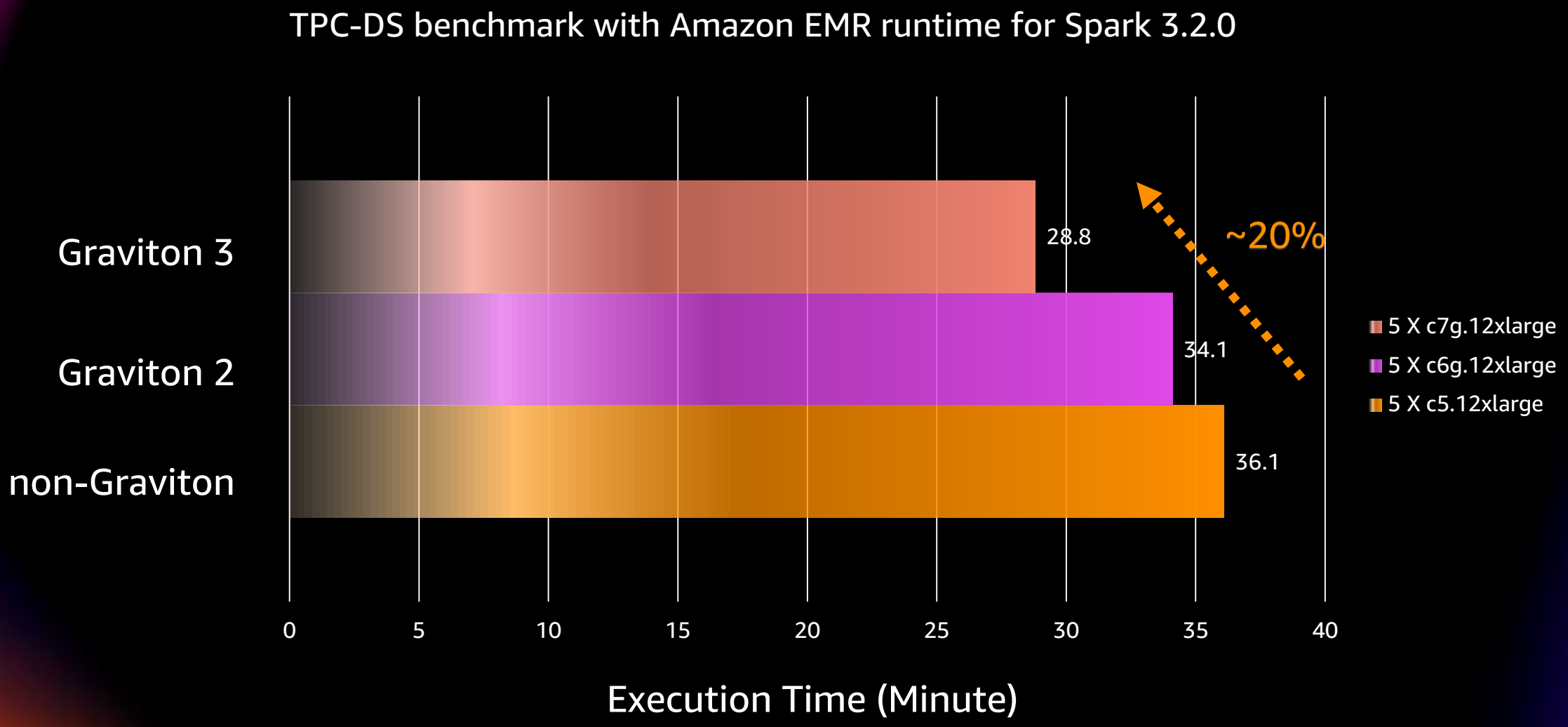
Spark applications finish faster, resulting in lower TCO as well as faster time to insights



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

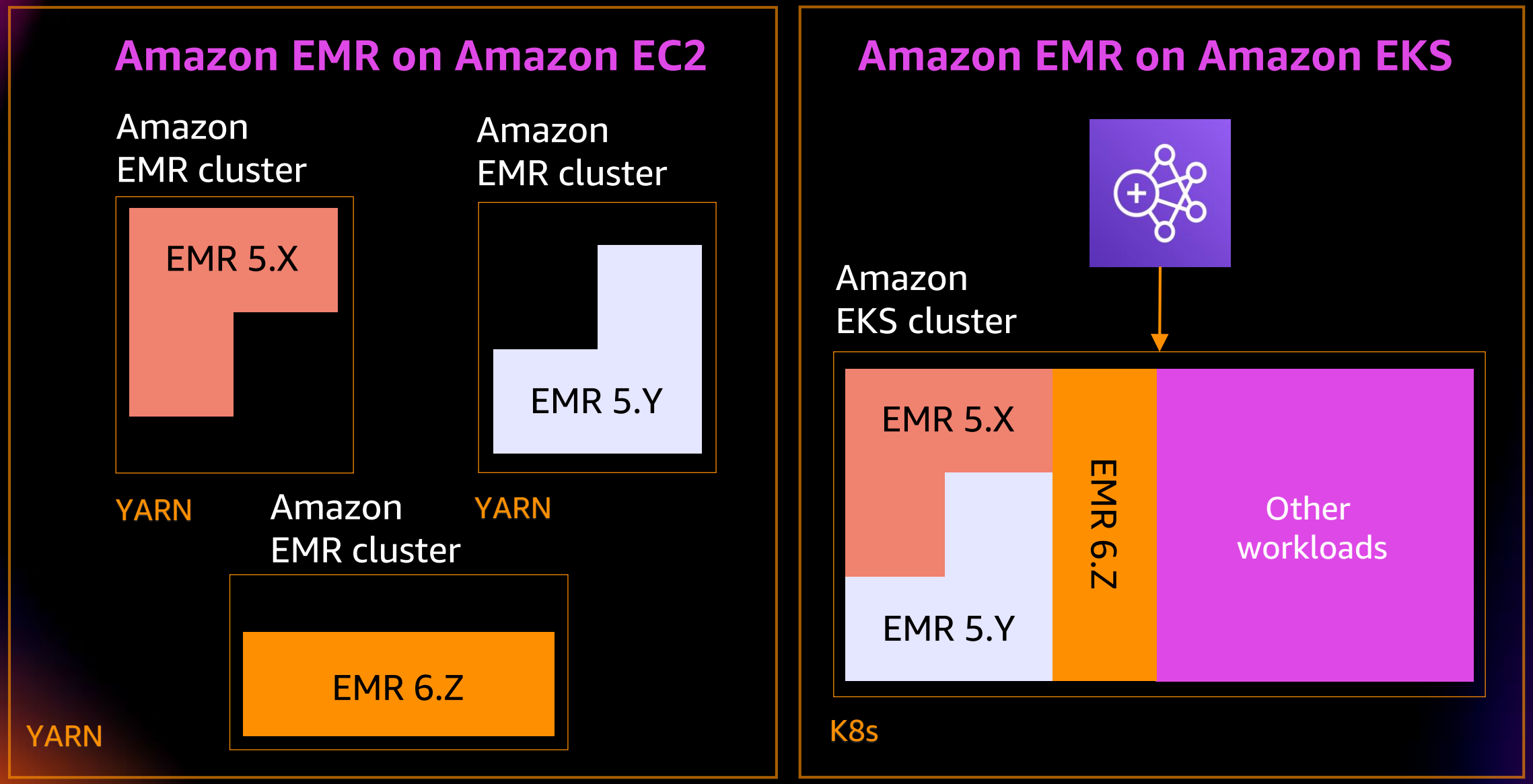
https://aws.amazon.com/blogs/big-data/amazon-emr-on-amazon-eks-provides-up-to-61-lower-costs-and-up-to-68-performance-improvement-for-spark-workloads/

Further performance improvement



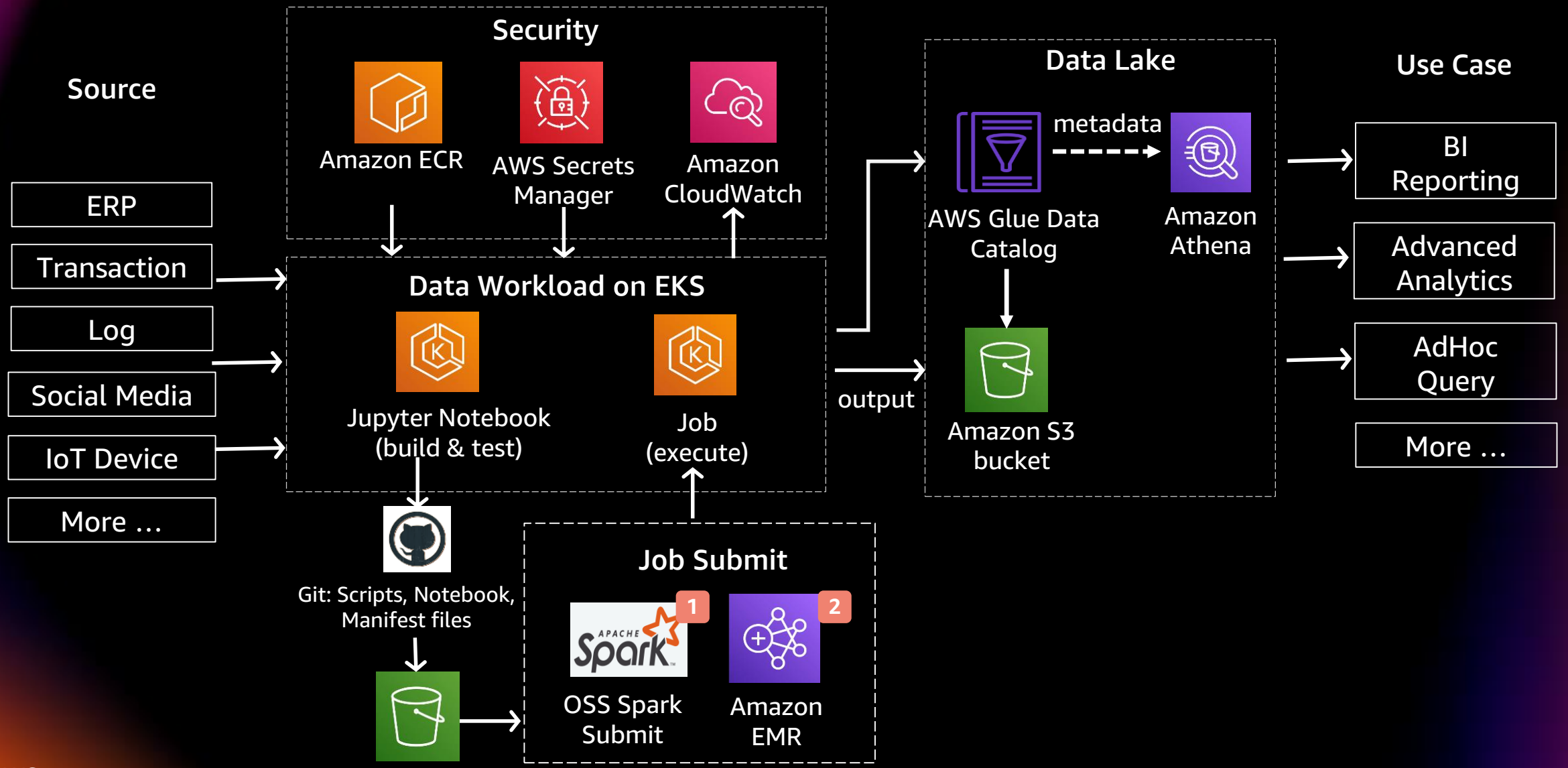
https://aws.amazon.com/blogs/big-data/amazon-on-emr-eks-further-improves-performance-by-20-powered-by-graviton3-instances/

Consolidation saves costs

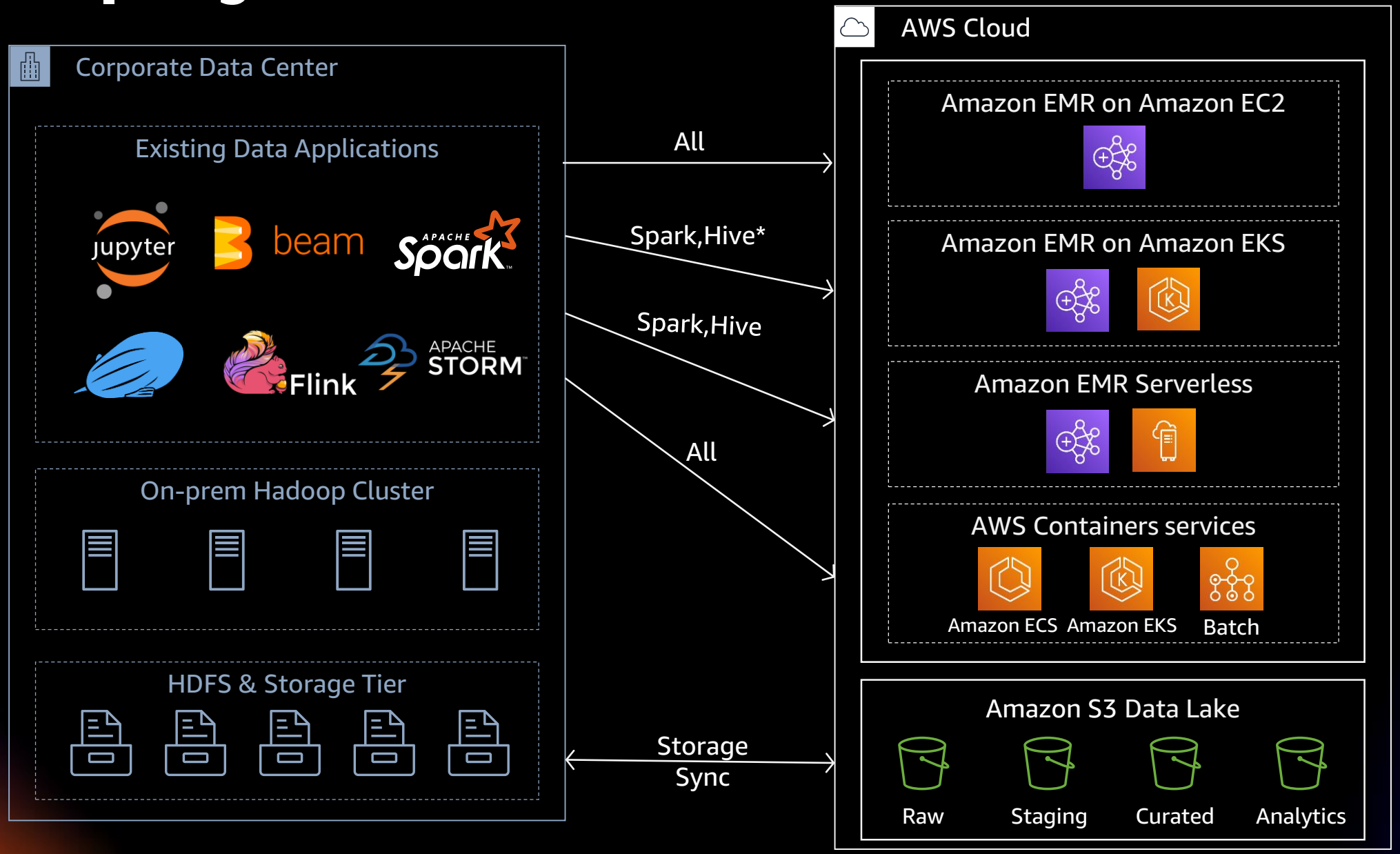


Reference architecture

SQL-based ETL with Spark on Amazon EKS



Hadoop migration with AWS services



* Hive on Spark

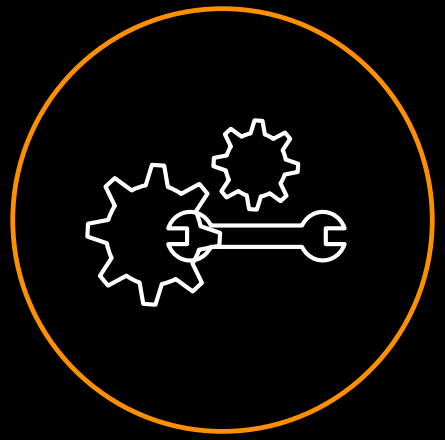
Demo

Run Hive SQL queries with Amazon EMR on Amazon EKS

Takeaways : Containerize once, run anywhere

100% open source compatibility with Amazon EMR on Amazon EKS lets you run your application anywhere

Spark operator on Amazon EKS



Re-install operator as you move between AWS, and on-premises.
Job invocation is different than open-source Spark submit

Spark submit on Amazon EKS

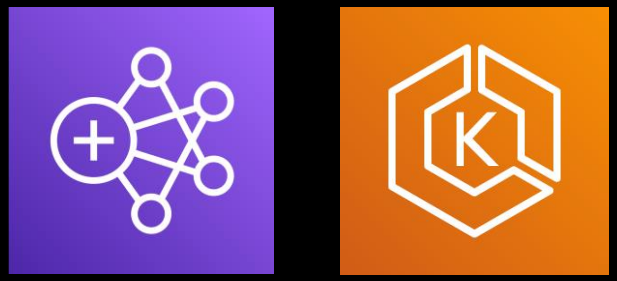
Re-build and maintain an additional API layer to submit jobs, track states, and debug jobs

Amazon EMR on Amazon EKS

Best of both worlds as Spark submit like API allow you to remain platform agnostic without losing functionality

Takeaways: Bring the value of Amazon EMR to Kubernetes

- Build Spark applications, not Infrastructure
- Optimized runtime runs 3.2x faster & 61% cheaper
- Start and scale clusters quickly and easily
- Unify data governance for analytics
- Fully managed service with native integration with AWS platform



Amazon EMR on Amazon EKS

Other resources

- Amazon EMR Containers best practices guides
- Amazon EMR on Amazon EKS self paced lab
- Amazon EKS workshop
- Spark Benchmark Utility: https://github.com/aws-samples/emr-on-eks-benchmark
- Hive metastore for Amazon EMR on Amazon EKS: https://github.com/aws-samples/hive-emr-on-eks
- Streaming with Amazon EMR on Amazon EKS: https://github.com/aws-samples/stream-emr-on-eks

Visit the AWS Data resource hub

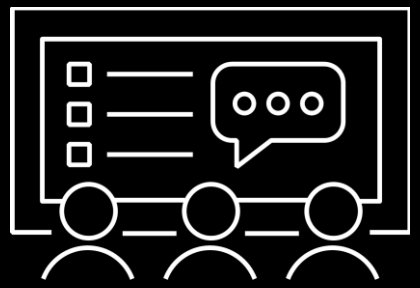
A modern data strategy can help you manage, act on, and react to your data so you can make better decisions, respond faster, and uncover new opportunities. Dive deeper with these resources today.

- Harness data to reinvent your organization
- In unpredictable times, a data strategy is key
- Make data a strategic asset
- Rewiring your culture to be data-driven
- Put your data to work with a modern analytics approach
- … and more!

https://tinyurl.com/data-hub-aws

Visit resource hub

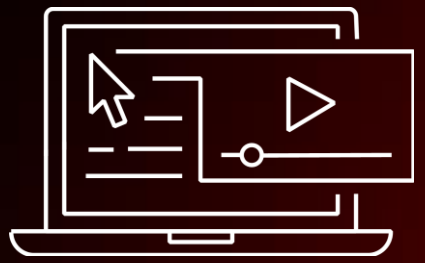
AWS Training and Certification for Data and Analytics



AWS Data & Analytics FREE Training Resources

Discover how to harness data, one of the world's most valuable resources, and innovate at scale.

https://bit.ly/3Ntlhy7



AWS Data Analytics Learning Plan

This learning plan expose you to the fastest way to get answers from all your data to all your users. It can also help prepare you for the AWS Certified Data Analytics - Specialty certification exam.

https://bit.ly/3wBVjD1

AWS Certified Data Analytics - Specialty

Earning AWS Certified Data Analytics – Specialty validates expertise in using AWS data lakes and analytics services.

https://go.aws/3lwF0RR

Thank you for attending AWS Innovate – Data Edition

We hope you found it interesting! A kind reminder to **complete the survey.**
Let us know what you thought of today's event and how we can improve the event experience for you in the future.


aws-apj-marketing@amazon.com


twitter.com/AWSCloud


facebook.com/AmazonWebServices


youtube.com/user/AmazonWebServices


slideshare.net/AmazonWebServices


twitch.tv/aws

Thank you!