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

CMP317-R

AWS Batch ask me anything

Naina Thangaraj (she/her)
Sr. Product Manager, Technical
AWS

Jo Adegbola (she/her)
GM, AWS Batch
AWS

Tom Burggraf (he/him)
Principal Product Manager, AWS Batch
AWS

Matt Vaughn (he/him)
Principal Developer Advocate
AWS

Vinoth Narasimhan (he/him)
Sr. Manager, Product Management Technical
AWS

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

Agenda

What is AWS Batch?

Why AWS Batch?

AWS Batch workloads

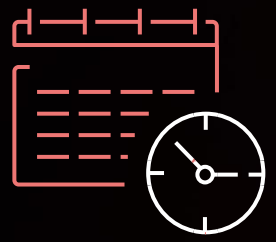
AWS Batch concepts

New AWS Batch features

Customers using AWS Batch

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

What is AWS Batch?



Job scheduler



Orchestrator

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

Why AWS Batch?



Fully managed



Integrated with AWS services



Optimized resource provisioning



Cost-efficient

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

Who uses AWS Batch?



Autonomous vehicle
ML and simulation



Gene sequencing
and drug discovery



Big data



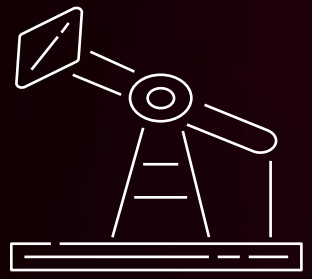
Machine learning



Financial
risk analysis



EDA, CAD, FDC



Renewable energy,
oil and gas exploration



Weather
simulation

aws

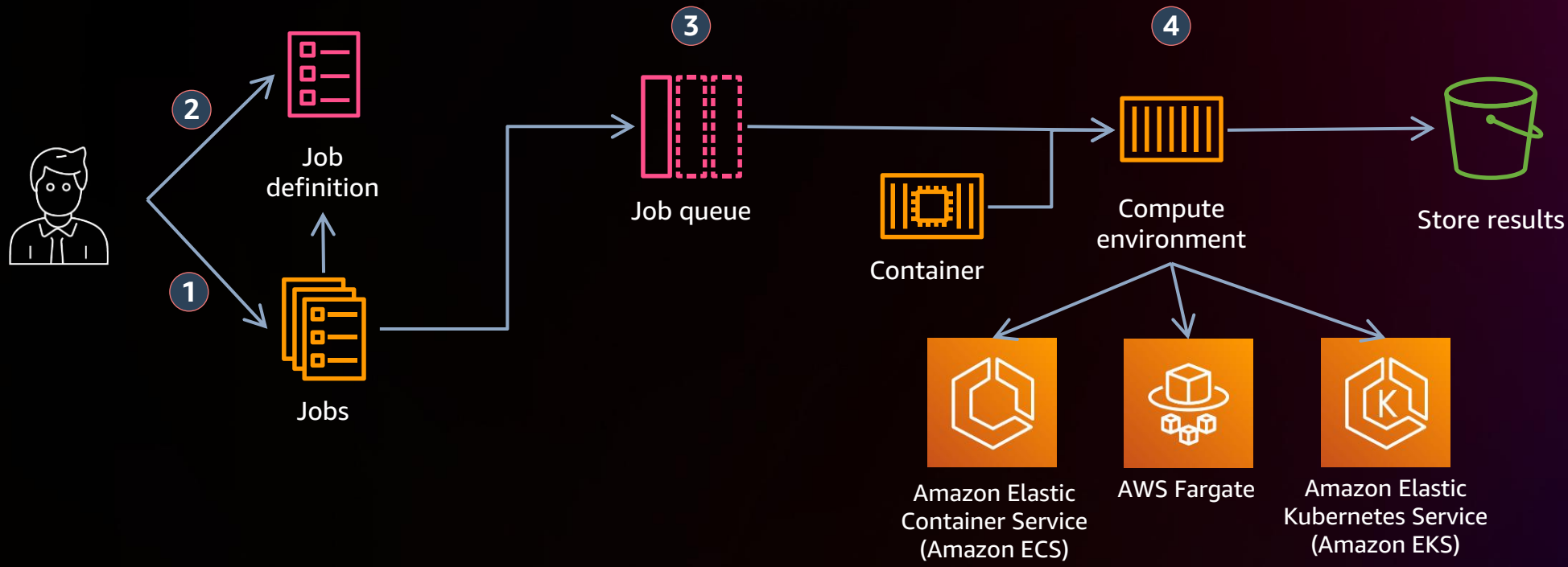
© 2022, Amazon Web Services, Inc. or its affiliates. All rights reserved.

AWS Batch concepts

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

AWS Batch overview

1. **Job**
   Each job must reference a job definition, but many parameters may be overridden when submitted
2. **Job definition**
   Template that has common attributes (such as container image, IAM role, vCPU and memory requirements)
3. **Job queue (JQ)**
   Queue determines priorities. Each JQ is connected to one or more CE
4. **Compute environment (CE)**
   Resource mix (defines On-Demand vs. Spot and other instance types); CE can be connected to more than one JQ

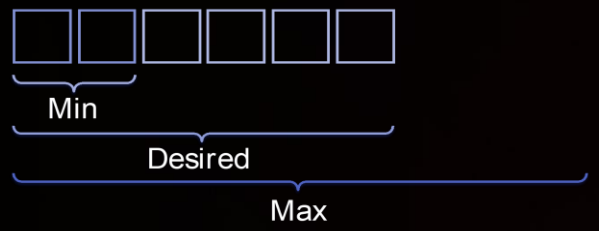


© 2022, Amazon Web Services, Inc. or its affiliates. All rights reserved.

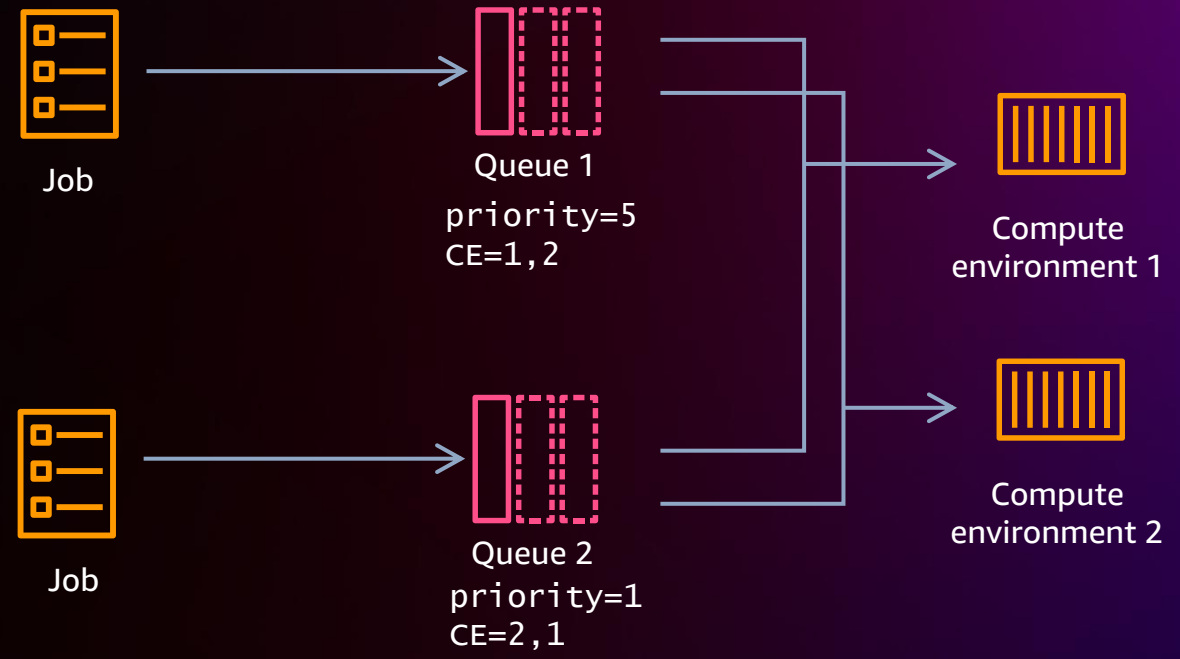
Compute environments (CEs)

Compute environments contain the underlying resources that are used to run jobs

- Types
  - **Managed:** AWS scales and configures underlying instances (recommended)
  - **Unmanaged:** Customers control and manage instance configuration, provisioning, and scaling
- Parameters
  - **Scaling:**



  - **Instance types:** Instance families or specific instances on which jobs will be running

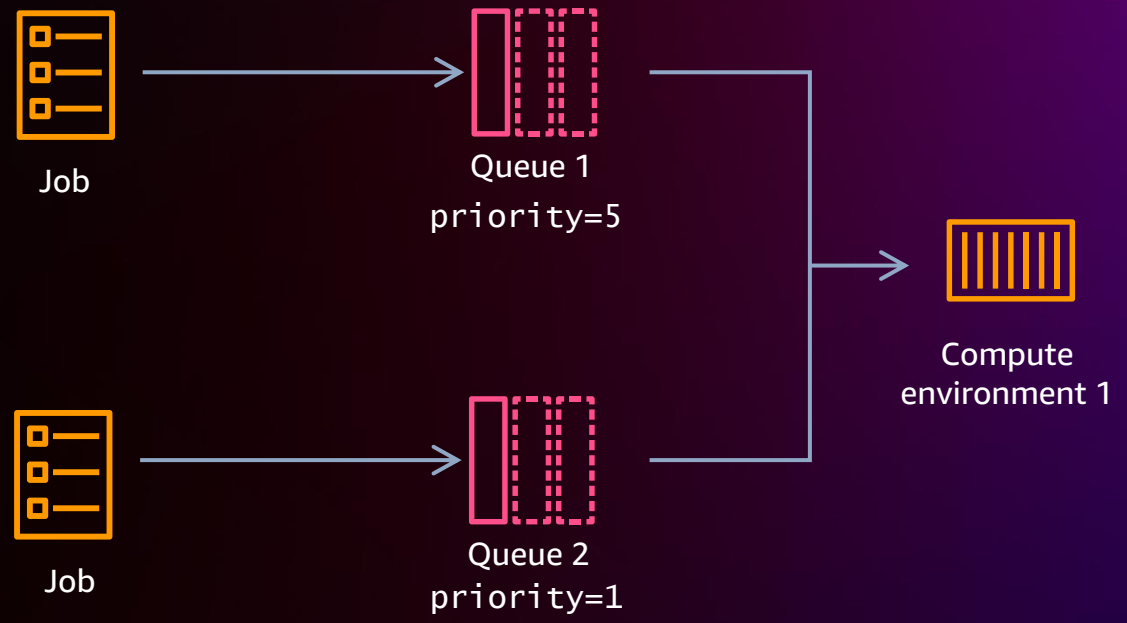


© 2022, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Job queues (JQ)

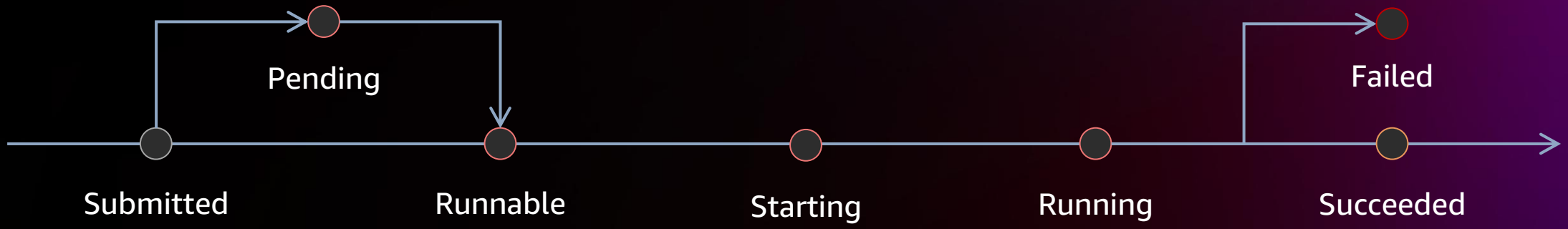
Job queues are where jobs are submitted and reside throughout their lifetime

- A single queue can connect up to three compute environments (CE) and can share a CE with other queues
- JQ parameters
  - **Priority:** Scheduling priority to assign a job to a CE shared with multiple JQs in descending order
  - **CE order:** Job placement in order of listing



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

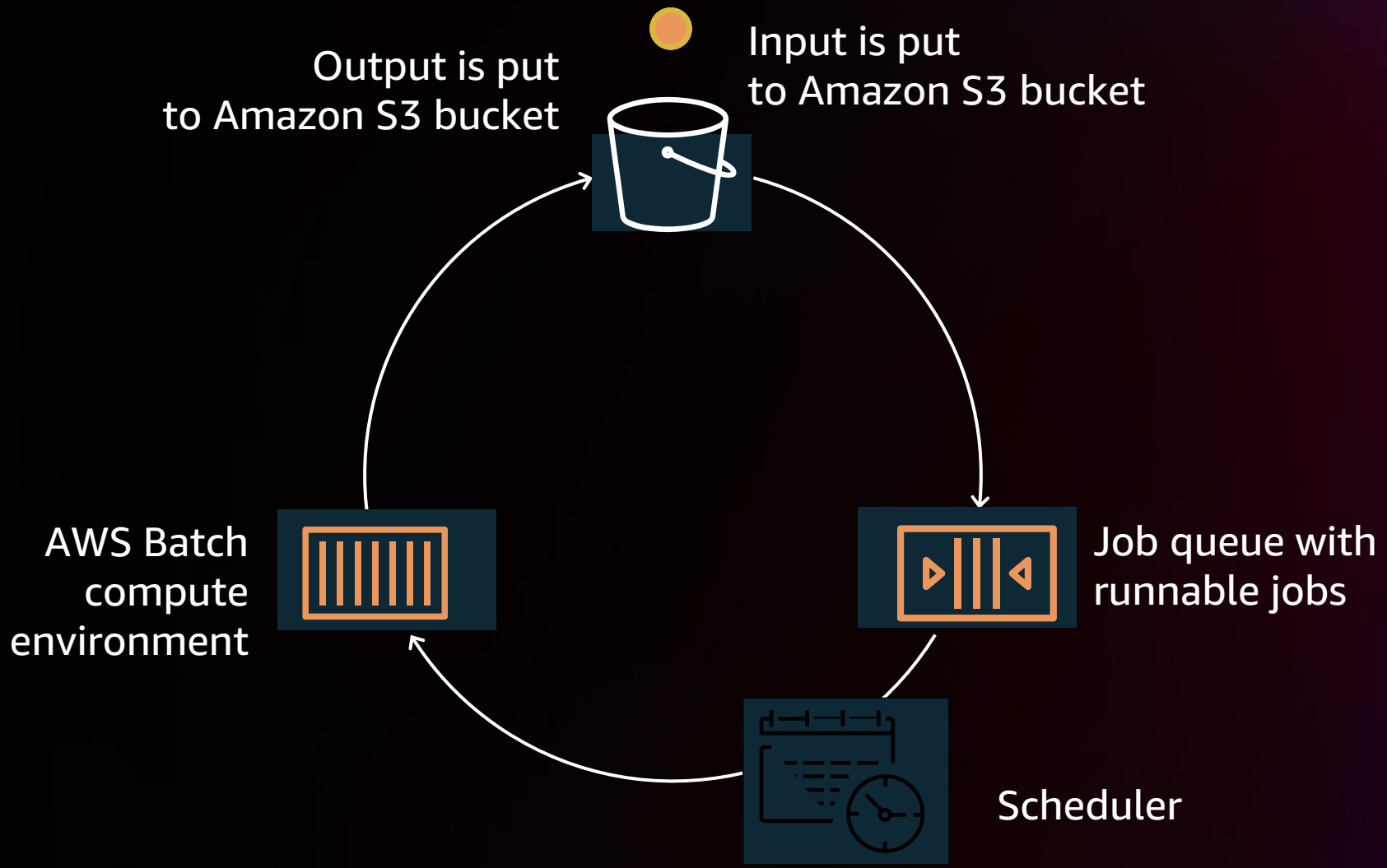
Jobs and their states



- **SUBMITTED:** Accepted into the queue, but not yet evaluated by the scheduler for execution
- **PENDING:** The job has dependencies on other jobs which have not yet completed
- **RUNNABLE:** The job is evaluated by the scheduler and is ready to run
- **STARTING:** The job is in the process of being scheduled to a compute resource
- **RUNNING:** The job is currently running
- **SUCCEEDED:** The job has finished with exit code 0
- **FAILED:** The job finished with a non-zero exit code, was cancelled or terminated

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

Typical AWS Batch job architecture



Job definition

Application
image + config

IAM role

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



SubmitJob

JQ1

Instance: Optimal

Max vCPU: 100

CE1 (On-Demand)

Allocation Strategy: Best Fit Progressive

Instance: Optimal

Max vCPU: 2,000

CE2 (Spot)

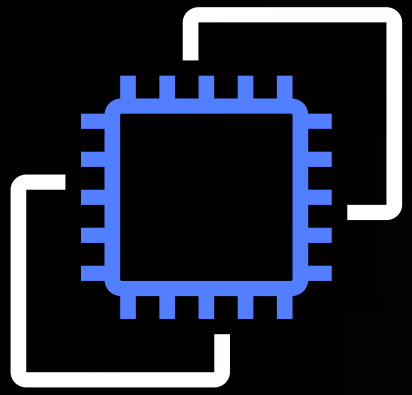
Allocation Strategy: Spot Capacity Optimized

Best Fit Progressive – When capacity is reached a capacity error (insufficient capacity, reclaim, EC2 limit), Batch will progressively sort through the list and pick the next best fit instance type

Spot Capacity Optimized – Allow AWS to find the most available Spot capacity pools, and launch instances accordingly. Available in Spot only

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

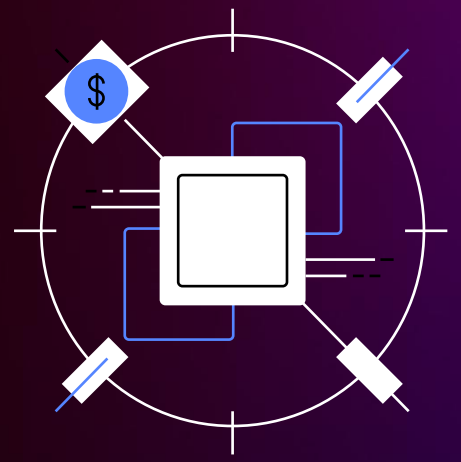
Allocation strategies for AWS Batch



Best Fit



Best Fit Progressive



Spot Capacity Optimized

Make capacity/throughput/cost tradeoffs

aws

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

EC2 Spot Instances are how you can do more for less

Up to 90% off On-Demand



Same infrastructure as On-Demand

89+ security standards

300+ instance options

24+ Regions

Spot is more than cost savings...

More compute + Faster results = Accelerated innovation



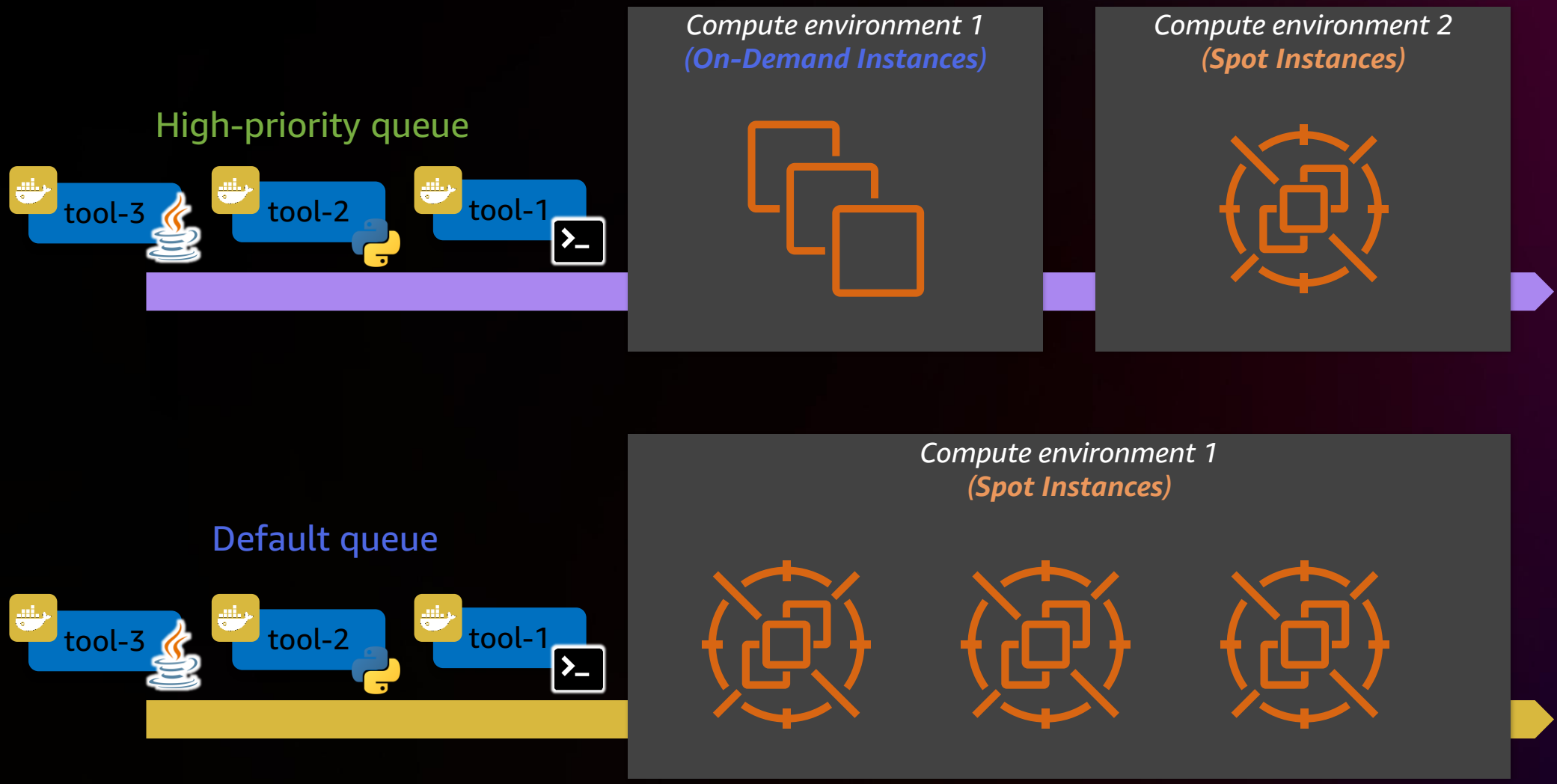




aws

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

Cost-efficient AWS Batch queues and compute environments



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

AWS Batch in numbers

500K – The record number of simultaneous jobs (and container images)

1B+ jobs
In a month across all Regions

5.6M vCPUs – Largest single run

We have thousands of concurrent batch customers

Some of whom run some of the largest workloads in the world.

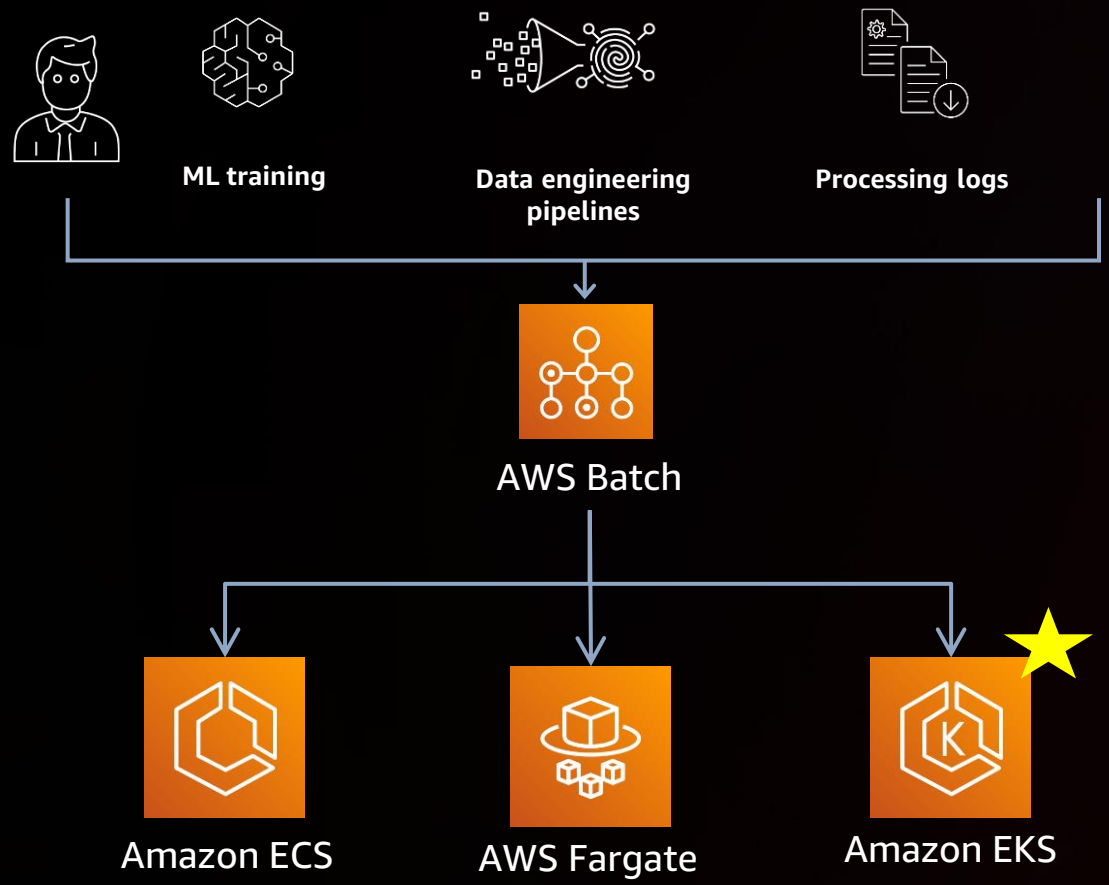
Oh, and it's free.

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

AWS Batch features

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

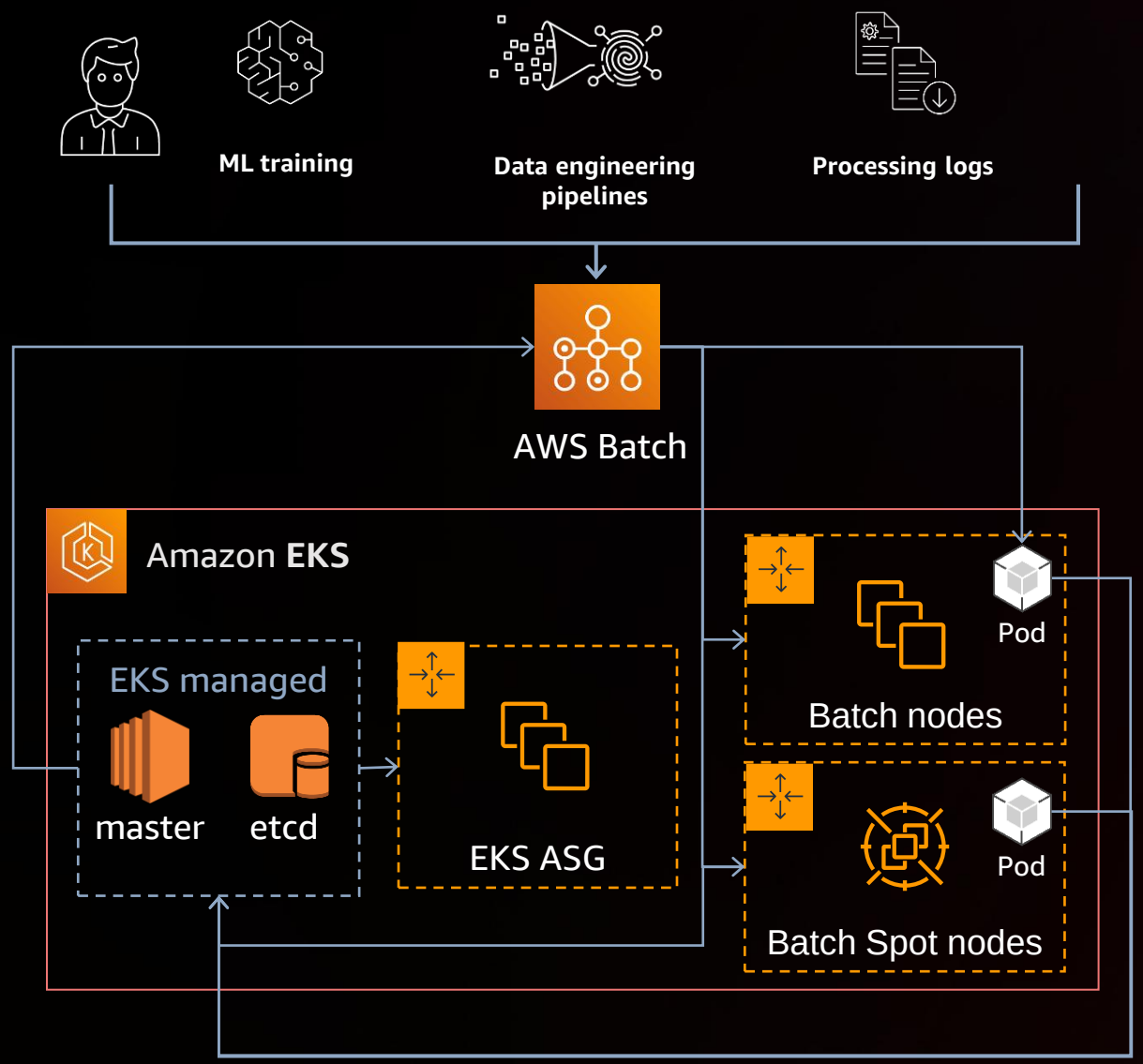
AWS Batch support for Amazon EKS



- **Fully managed** service for batch workloads
- Optimized for throughput, scalability, and cost
- Workload-aware cluster scaling, job scheduling and full lifecycle management
- Works with your existing Amazon EKS clusters
- Does not conflict with your other Amazon EKS workloads
- Works well with your current monitoring and other cluster governance daemon sets

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

AWS Batch for Amazon EKS – Under the hood

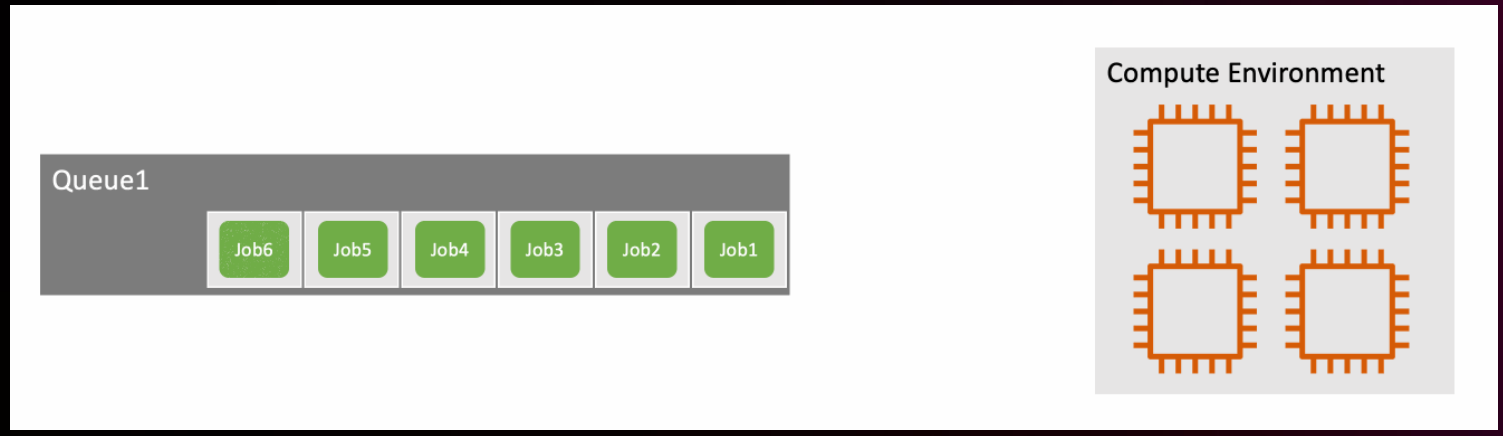


- Batch is the main entry point for batch workload requests
- Scales the nodes based on the jobs queued
- Scales nodes separate from other Amazon EKS workloads using Batch's scaling knowledge
  - Capacity pools and Spot interruption likelihood
- Taints prevent placement of other pods on Batch-managed nodes
- Batch handles the pod placements

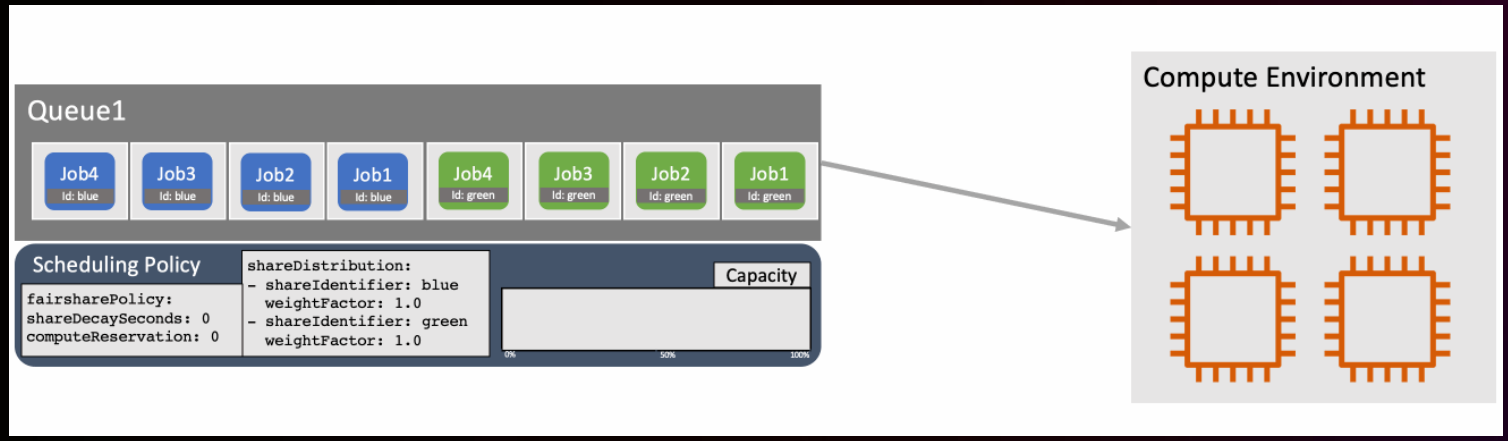
© 2022, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Job queue: Fair-share scheduling

FIFO



Fair-share



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

Customers using AWS Batch

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

Simulation workloads using AWS Batch and Spot Instances

Analyzing autonomous vehicle data to drive new vehicle capabilities

Mobileye will ingest, process, and analyze hundreds of petabytes of sensor, image, and video data gathered from Mobileye's autonomous vehicle fleets driving worldwide

Mobileye uses AWS Batch to simulate workloads on >200K cores for an extended periods of time using Spot Instances, testing compute, and network performance

"A job that took weeks in our data center due to limited resources **took hours on Spot thanks to the great parallelism**, in a very cost-efficient price."

—Shay Asoolin, Sr. Director Development Infrastructure, Mobileye

MOBILEYE®
Our Vision. Your Safety.™

Company:	Mobileye
Industry:	Software and internet
Country:	United States
Employees:	750
Website:	www.mobileye.com

About Mobileye

Mobileye is the leading supplier of software that enables Advanced Driver Assist Systems (ADAS). Our technology has rapidly evolved to also support the three pillars of autonomous driving – sensing, mapping, and driving policy

aws

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

Fred Hutch microbiome researchers use AWS to perform 7 years of compute time in 7 days

Challenge

At Seattle's Fred Hutch Microbiome Research Initiative (MRI), a team of researchers are engaged in analysis of the microbiome, which is the collection of microbes on and inside the human body. But these researchers aren't just studying the microbiome – they're striving to manipulate microbiomes to make therapeutic cancer drugs more effective. To support their efforts, researchers must analyze and process an immense number of whole genome datasets.

Solution

Fred Hutch chose AWS to power the high-performance computing (HPC) platform that runs microbiome analysis. These researchers execute their computational analysis using the Nextflow framework to orchestrate **AWS Batch** processes and scale the HPC platform to accelerate processing time. The organization runs its HPC workloads on Amazon EC2 instances, powered by Intel Xeon Platinum 8000 series processors, and it stores research data in Amazon S3 buckets.

Benefits

- Processes data from more than 15,000 biological samples
- Reduced 7 years of compute time to 7 days, so researchers can get results faster
- Increases resolution on microbiome samples to find links to improve health outcomes
- Enables collaboration with other scientific researchers

Company: Fred Hutchinson Cancer Research Center

Country: US

Employees: 3,500

Website: FredHutch.org

The Fred Hutch Microbiome Research Initiative, funded by Seattle's Fred Hutchinson Cancer Research Center, includes microbiome investigators with expertise in study design, laboratory methods, animal models, human intervention studies, data analysis, and visualization. These researchers are working to predict health outcomes, understand the pathogenesis of disease, and manipulate the microbiota to promote health.

"Our goal is to accelerate our research processes on AWS so we can get closer to developing therapeutics to fight cancer."

-Sam Minot, PhD and Staff Scientist

aws

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

AstraZeneca is raising the bar by running its genome sequencing pipeline on AWS

Challenge

AstraZeneca's Centre for Genomic Research (CGR) has a bold target to analyze 2 million genomes by 2026. The organization faced 3 key challenges:

On-premise compute resources, which limit the performance capacity, were difficult and costly to scale to this demand.

Dependency upon third-party informatics providers reduced flexibility to adapt and improve future iterations of the analysis.

Orchestration of the bioinformatics pipeline was time-intensive, therefore costly and hard to scale.

Solution

With support from AWS Professional Services, AstraZeneca built a highly scalable and high-performance sequence data processing pipeline on AWS.

The solution leveraged FPGA instances for compute and extensive use of AWS Step Functions, AWS Lambda, Amazon SQS and **AWS Batch**.

The output is stored in scalable and highly secure AWS-managed databases and Amazon S3 storage.

Benefit

AstraZeneca has been able to build, at the time of writing, possibly the fastest and most operationally efficient sequence data processing pipeline in the industry.

The bespoke pipeline was able to decrease processing time by 2400%, from 20 days at end of 2016 to 20 hours for 20,000 samples in June 2019, with expectations to accelerate further.

The results from this early batch have been used to provide scientists with advanced access to the clinical effects of natural mutations in humans that mimic drug inhibition/suppression for the majority of human genes.

AstraZeneca

Company:	AstraZeneca
Industry:	Pharmaceuticals
Country:	Global, UK HQ
Employees:	Circa 61K
Website:	www.astrazeneca.com

AstraZeneca PLC is a leading global pharmaceutical organization that engages in the research, development, and manufacturing of pharmaceutical products. The company produces and commercializes medicines for gastrointestinal, cardiovascular, neuroscience, respiratory and inflammation, oncology, and infectious diseases.

"Building our bespoke analysis pipeline on AWS with the support of AWS Professional Services has enabled us to achieve acceleration of key objectives towards the analysis of up to 2 million genomes (AstraZeneca's Genomics Initiative)."

-Slavé Petrovski, PhD, Vice President and Head of Genome Analytics and Bioinformatics, Discovery Sciences, R&D

aws

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

New AWS Batch features

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

Major launches in 2022

- CE and JQ insights in the AWS console
- Dynamic updates to CE properties
- AWS PrivateLink
- 7-day retention period
- 5000 instances per cluster
- Instance support - G5, Hpc6a, C6i/R6i/M6i, DL1 (with standard ECS AMI), I4I, C7g,TRN1, C6a[d]/R6a[d]/M6a[d]
- AWS Batch on Amazon EKS

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

Thank you!

Naina Thangaraj
tnaina@amazon.com

Vinoth Narasimhan
vkumarn@amazon.com

Matt Vaughn
mwvaughn@amazon.com

Jo Adegbola
joadegbo@amazon.com

Tom Burggraf
tomburg@amazon.com



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

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