

The background features a dark blue gradient with abstract, glowing shapes in shades of purple and pink. Two thin, light blue lines cross the scene diagonally. The text is positioned on the left side.

AWS re:Invent

DECEMBER 2 - 6, 2024 | LAS VEGAS, NV

AIM388 - NEW

Amazon SageMaker HyperPod: Reduce costs with new governance capability

Kareem Syed-Mohammed

Sr. Product Manager,
Amazon SageMaker
Amazon Web Services

Joy Fan

Sr. Manager,
Software Development
Amazon

Arun Subramaniyan

Founder and CEO
Articul8 AI

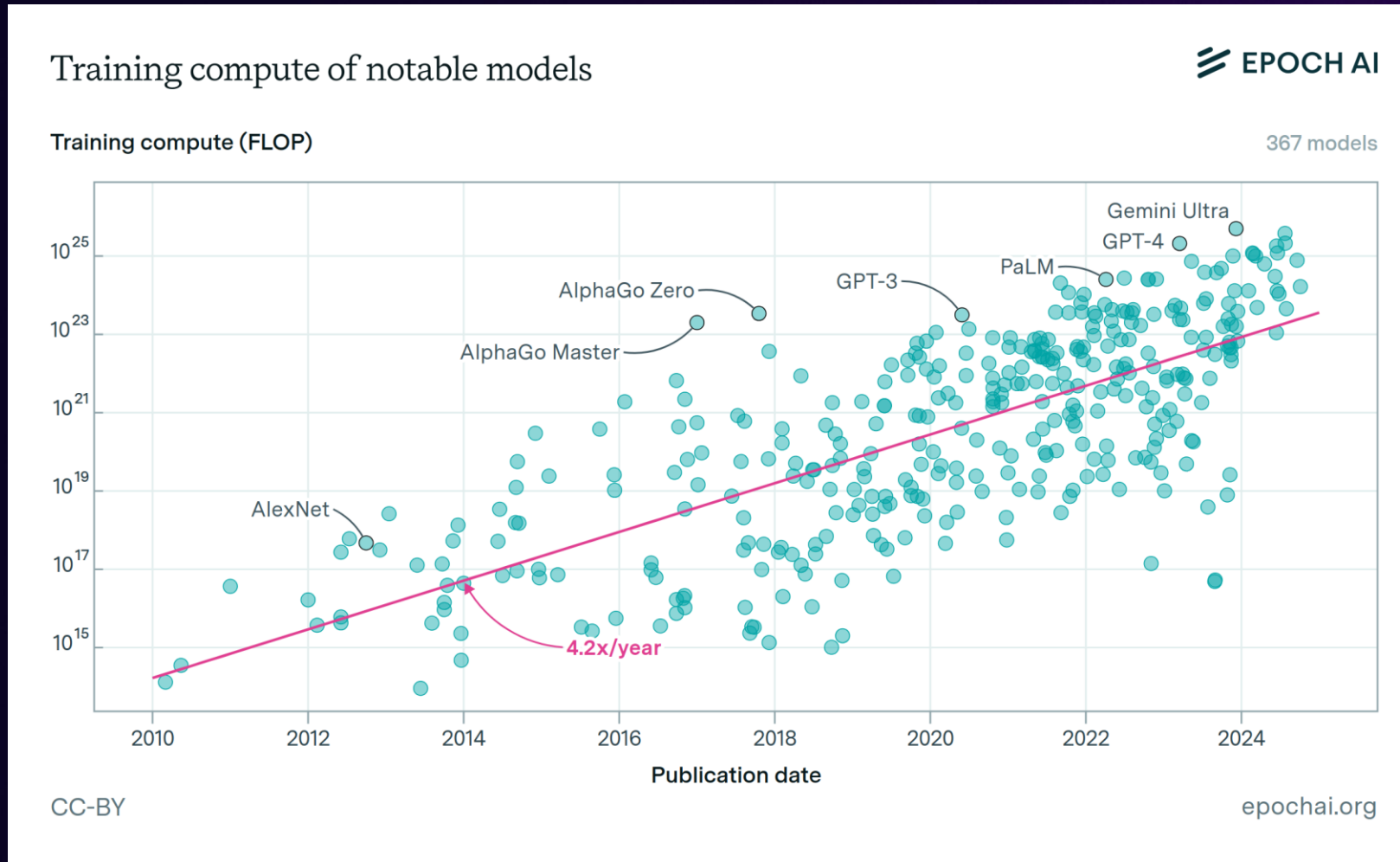


Amazon SageMaker HyperPod

Kareem Syed-Mohammed



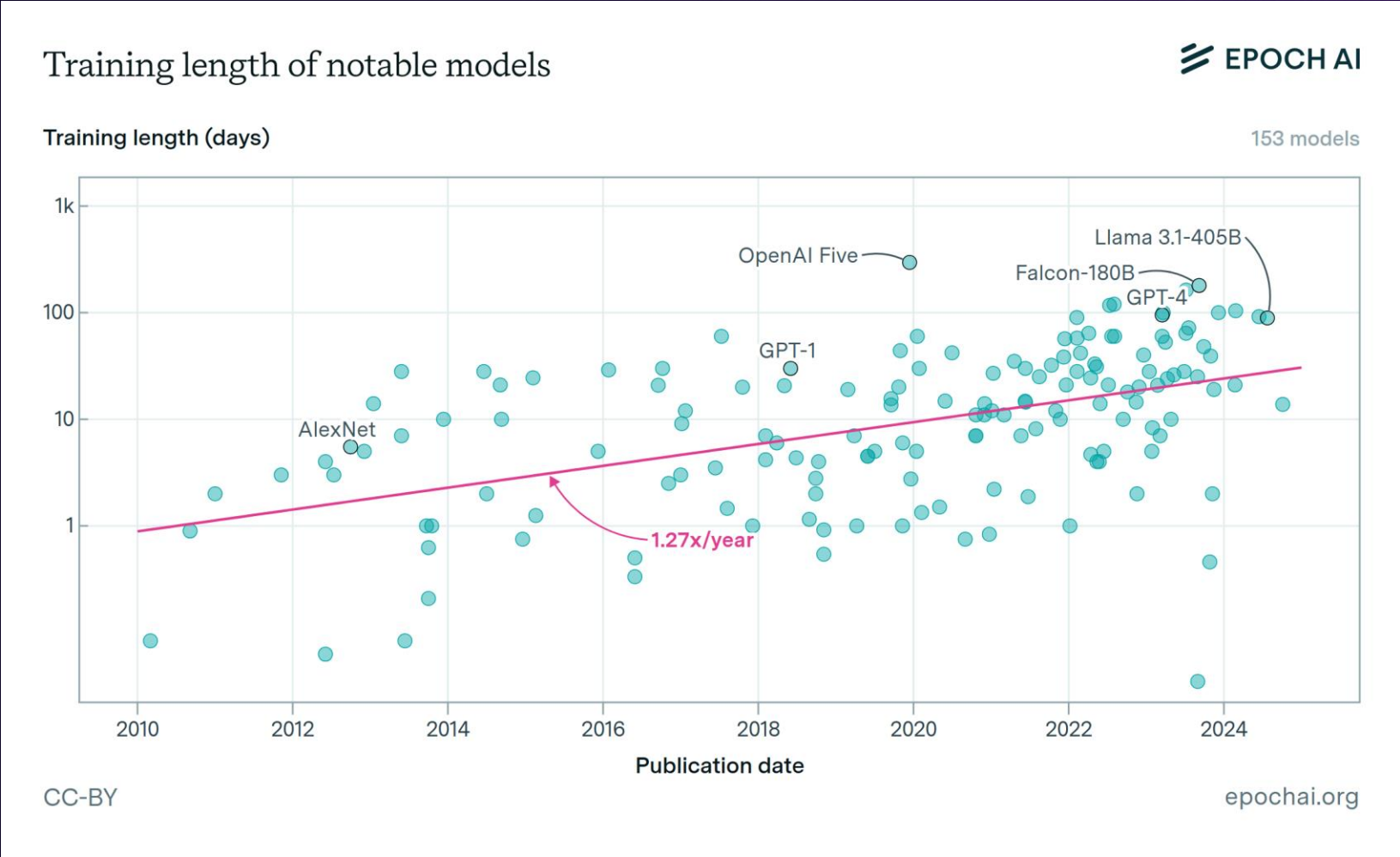
Training compute of frontier models is growing by 4–5x per year, doubling roughly every 6 months



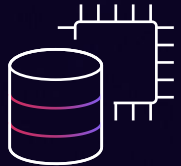
Dataset sizes are doubling every 8 months



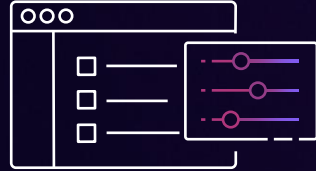
Time to market spans months of investment and continues to increase



Challenges with large-scale gen AI model development



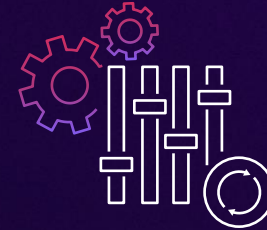
Collect data



Clusters provision & management



Infrastructure stability



Strategies for distributed training

Amazon SageMaker HyperPod

REDUCE TRAINING TIME BY UP TO 40% THROUGH RESILIENCY AND PERFORMANCE OPTIMIZATIONS



Resilient environment

Self-healing clusters reduce training time



Streamline distributed training

SageMaker distributed training improves performance



Optimized resources utilization

Control over computing environment and workload scheduling

Top AI companies use HyperPod to train and deploy models

Articub

Luma AI



Coastal Carbon

datologyai

featherless.ai

Hippocratic AI
— Do No Harm —

HOPPR

Hugging Face

MOHAMED BIN ZAYED
UNIVERSITY OF
ARTIFICIAL INTELLIGENCE

OMAN
DATAPARK

perplexity

salesforce

Thomson Reuters

Twelve Labs

ubitus



OpenBabylon

NOETIK

NinjaTech AI

Orbital

Stockmark

WRITER

arcee.ai

RECENTLY
LAUNCHED

Amazon Elastic Kubernetes Service (Amazon EKS) support in Amazon SageMaker HyperPod

Remove the heavy-lifting to scale across more than a thousand AI accelerators

A fully resilient infrastructure purpose-built for gen AI model development

Optimize utilization of cluster's compute, memory, and network resources between training and inference workloads



Customer pain points



Under- or over-allocation



Idle compute



Waiting tasks

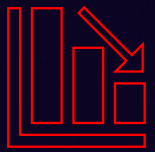


Lack of task prioritization



Lack of compute utilization observability

Outcome



Low utilization of compute



Reduced scientists' productivity



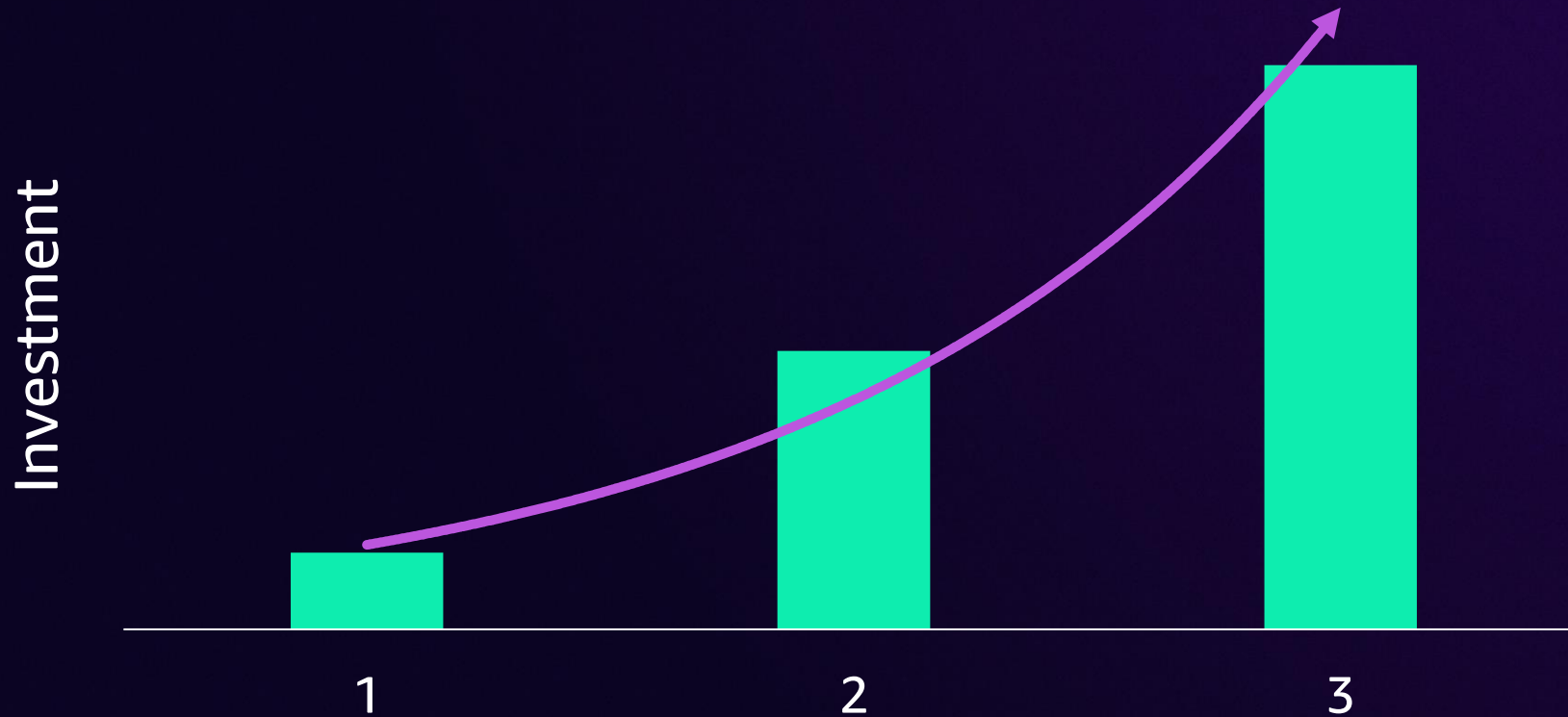
Increased costs

Amazon gen AI challenges and innovation

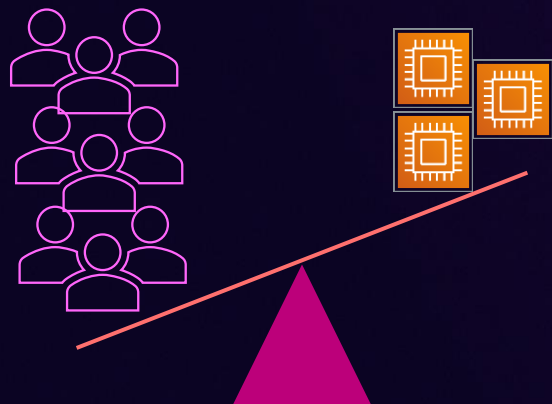
Joy Fan



Amazon gen AI investment trend



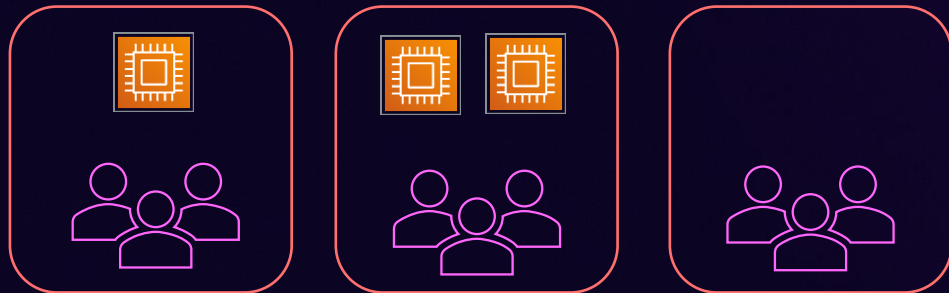
Challenges



High demand

Low supply

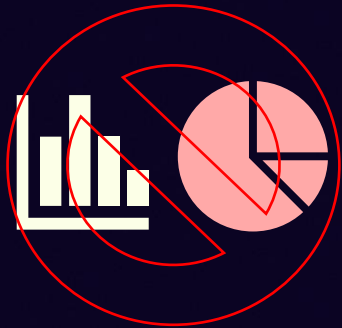
Challenges



**Static allocation for
100+ teams/projects**

**Spikey demand – some
wait for allocation while
others' resources stay idle**

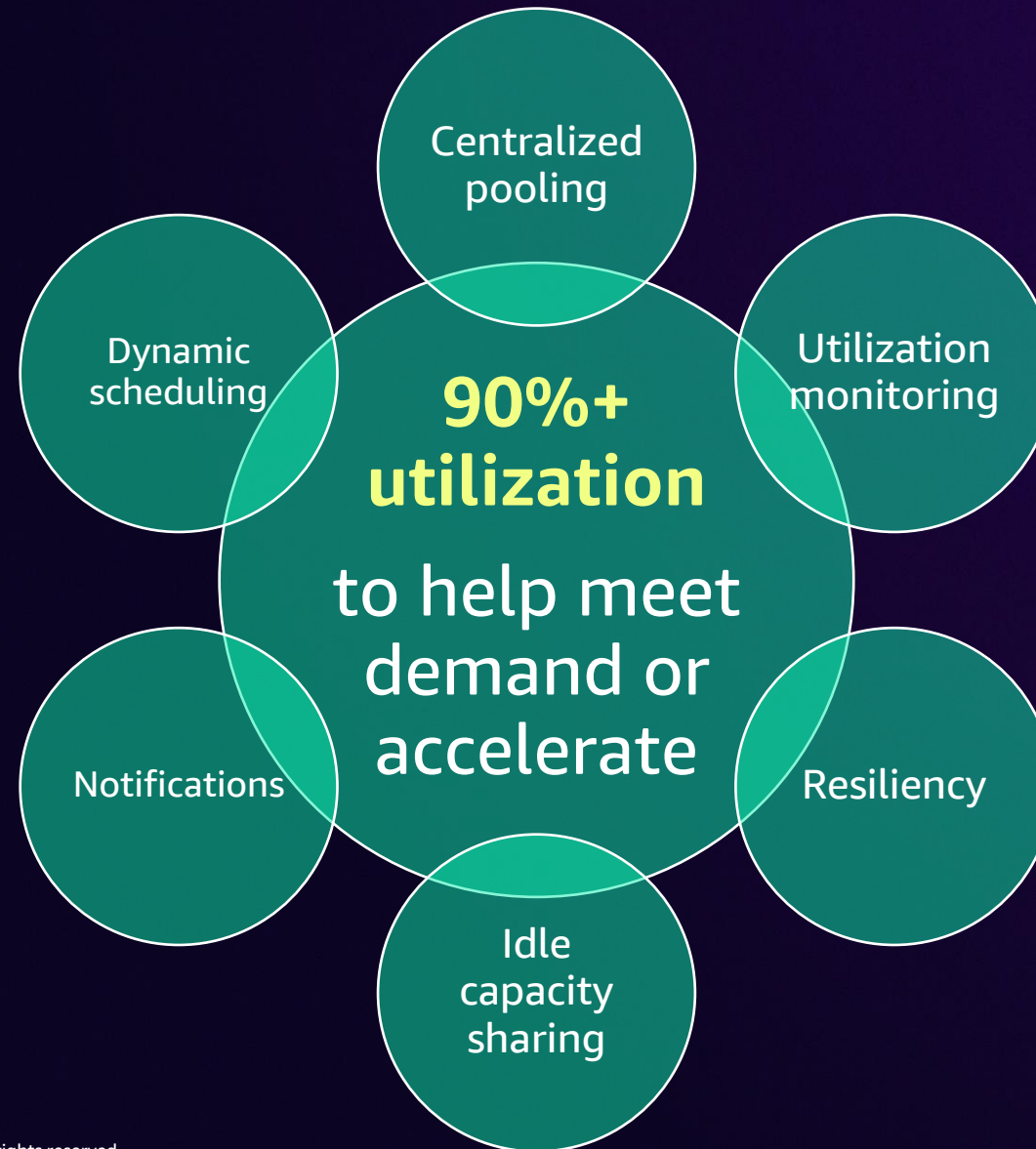
Challenges



Lack of utilization metrics

Lack of standard centralized monitoring

Amazon innovation – AI orchestration platform



Introducing Amazon SageMaker HyperPod task governance

PRIORITIZE TASKS, ALLOCATE COMPUTE RESOURCES, AND MAXIMIZE UTILIZATION

Kareem Syed-Mohammed



NEW

Amazon SageMaker HyperPod task governance

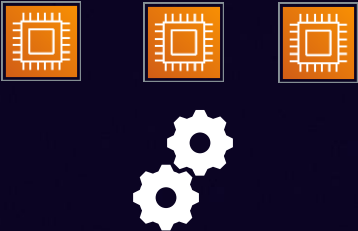
Maximize accelerator utilization and reduce costs
for model training, fine-tuning, and inference

- Dynamically allocate compute resources across tasks
- Ensure high-priority tasks are completed on time
- Monitor and audit compute allocation in real-time
- Maximize compute resource utilization and reduce costs by up to 40%

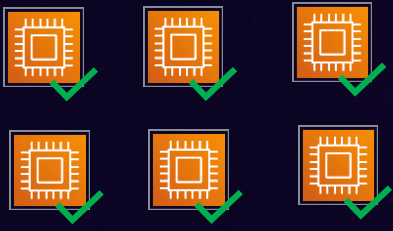
SageMaker HyperPod task governance

PRIORITIZE TASKS, ALLOCATE COMPUTE RESOURCES, AND MAXIMIZE UTILIZATION

A smart scheduler and orchestrator that enables



Dynamic resource allocation



Utilizes idle compute



Reduces task wait times

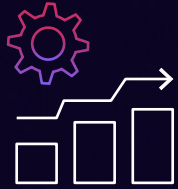


Real-time task prioritization



Real-time observability

With SageMaker HyperPod task governance, you . . .



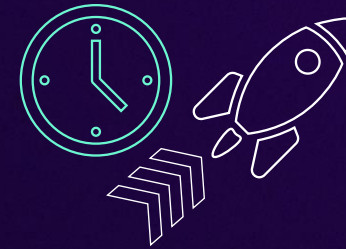
Increase compute utilization

Increases compute utilization by maximizing utilization of idle compute



Reduce costs

Higher utilization can reduce costs up to 40%



Accelerate time to market

Waiting tasks use idle compute, reducing wait times

Common use cases

Resource management

Efficiently allocating GPUs across different AI/ML projects and teams

Dynamic allocation

Automatically allocating idle resources to accelerate waiting tasks

Task orchestration

Automating the scheduling, prioritizing, and preempting of AI/ML workloads

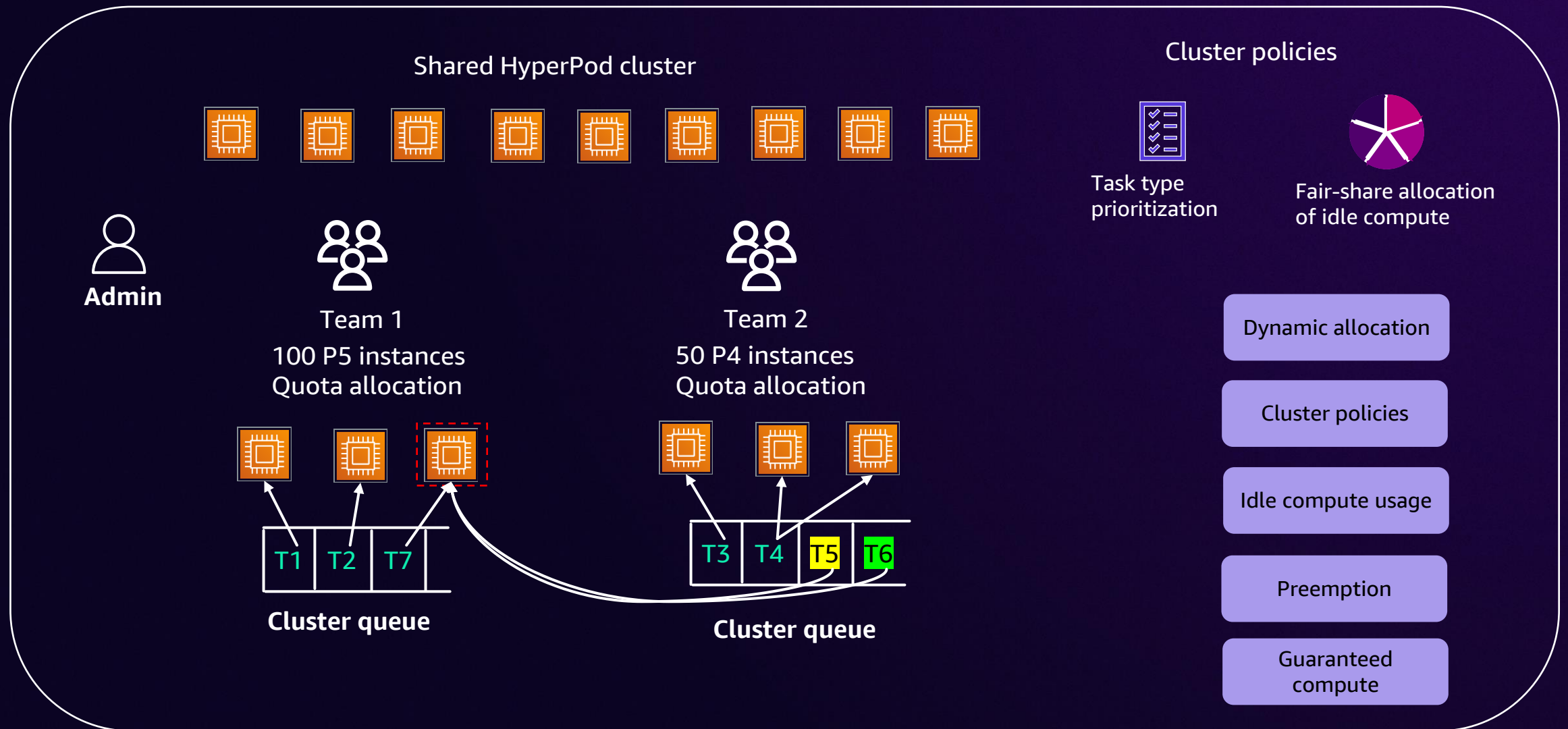
Monitoring and analytics

Providing insights into resource utilization, job performance, and overall system health

Cost optimization

Helping organizations minimize cloud computing costs for AI/ML workloads

How SageMaker HyperPod task governance works



Demo

The screenshot displays the Amazon SageMaker console interface. The top navigation bar includes the AWS logo, a search bar, and the user's name 'Admin/nadknish-isengard'. The main content area is titled 'Amazon SageMaker' and shows the path 'Cluster management > ml-cluster'. The cluster name 'ml-cluster' is displayed with a green 'In service' status. Action buttons for 'Edit', 'Copy ARN', 'Delete', and 'Monitor in Container Insights' are visible. Below this, a navigation bar includes 'Dashboard', 'Tasks', 'Policies', 'Instances', 'Settings', and 'Details'. The 'Utilization' section is active, showing a dropdown for 'All Instance Groups (2)' and an 'Export' button. The utilization metrics are as follows:

Metric	Value
Instances	16
Running instances	16
Pending recovery	0
GPUs	15
GPU memory	360
vCPUs	488
vCPU memory	1952

Below the metrics, there are three sections for utilization: GPU utilization, GPU memory utilization, and vCPU utilization.

Customer story – Articul8 AI

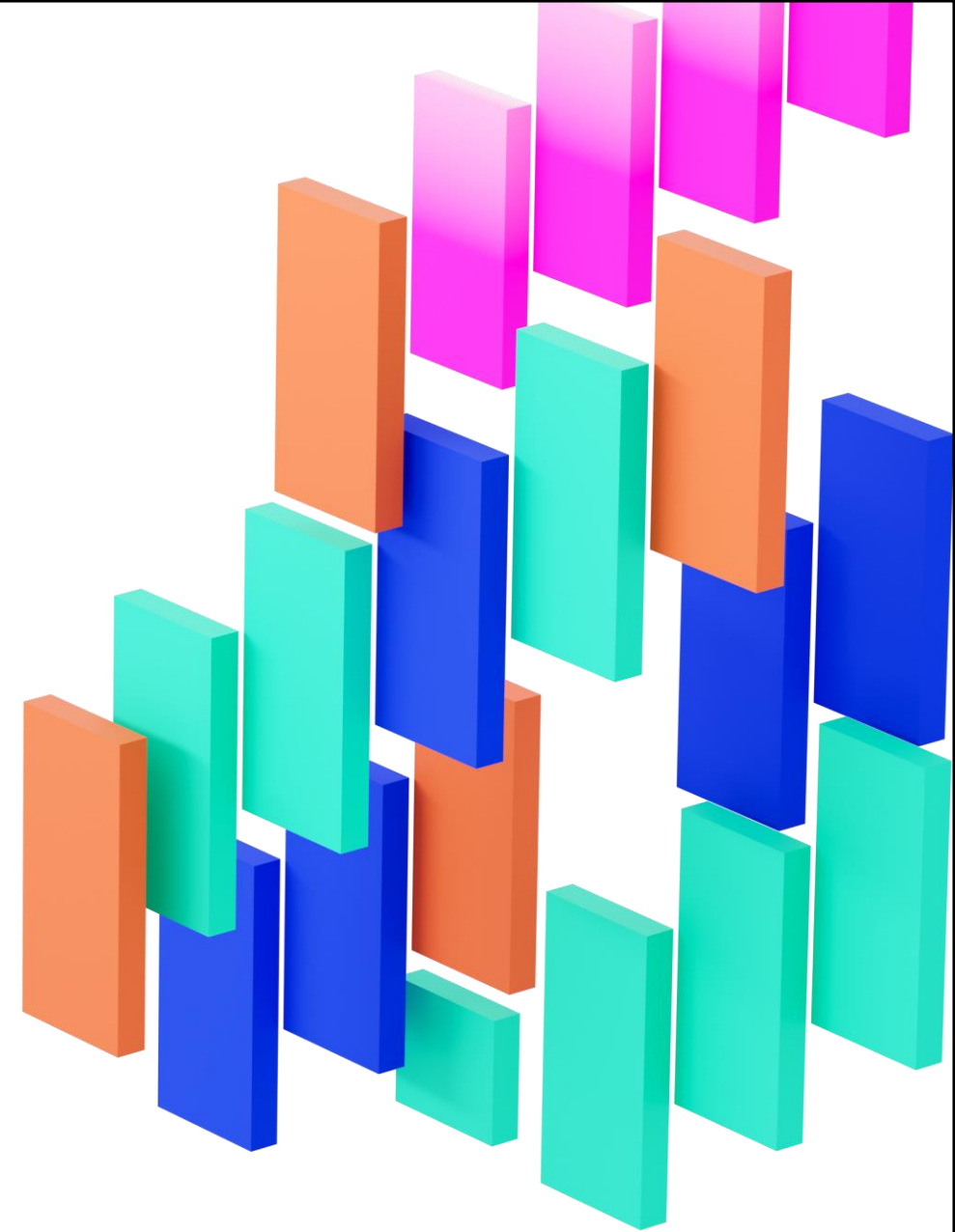


Articul8

Generative AI (GenAI) Platform for Enterprises

Arun Subramaniyan

November 2024



Articul8

The fastest way to build **sophisticated enterprise gen AI applications** with your data & expertise

Proven Track Record With Enterprise Customers



NielsenIQ

BCG

intel.

Hyperscaler

Articul8 Key Differentiators



Autonomous: ModelMesh™ selects & manages the right combination of models for autonomous multi-agent decisioning & actioning

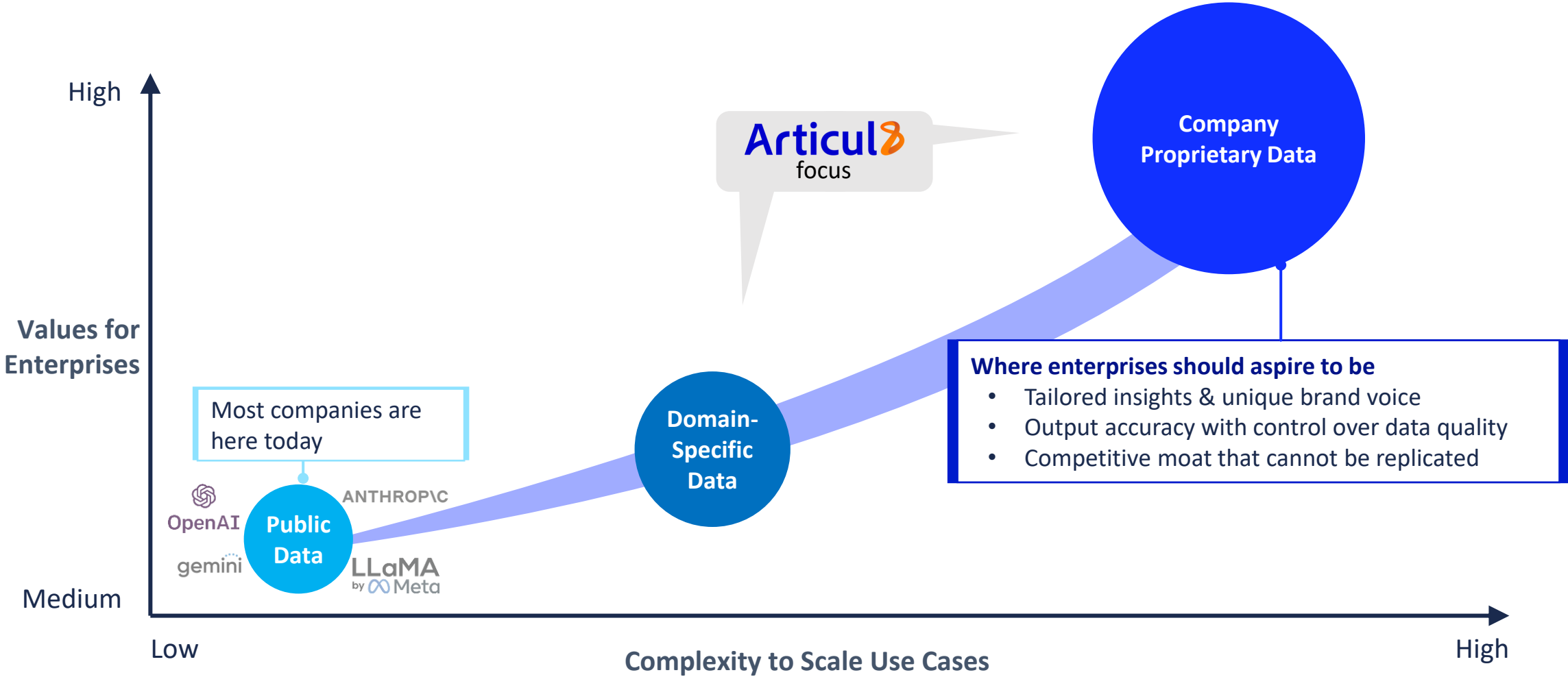


Domain-specific models: Build, deploy, and optimize domain-specific models with specialized **data partnerships**



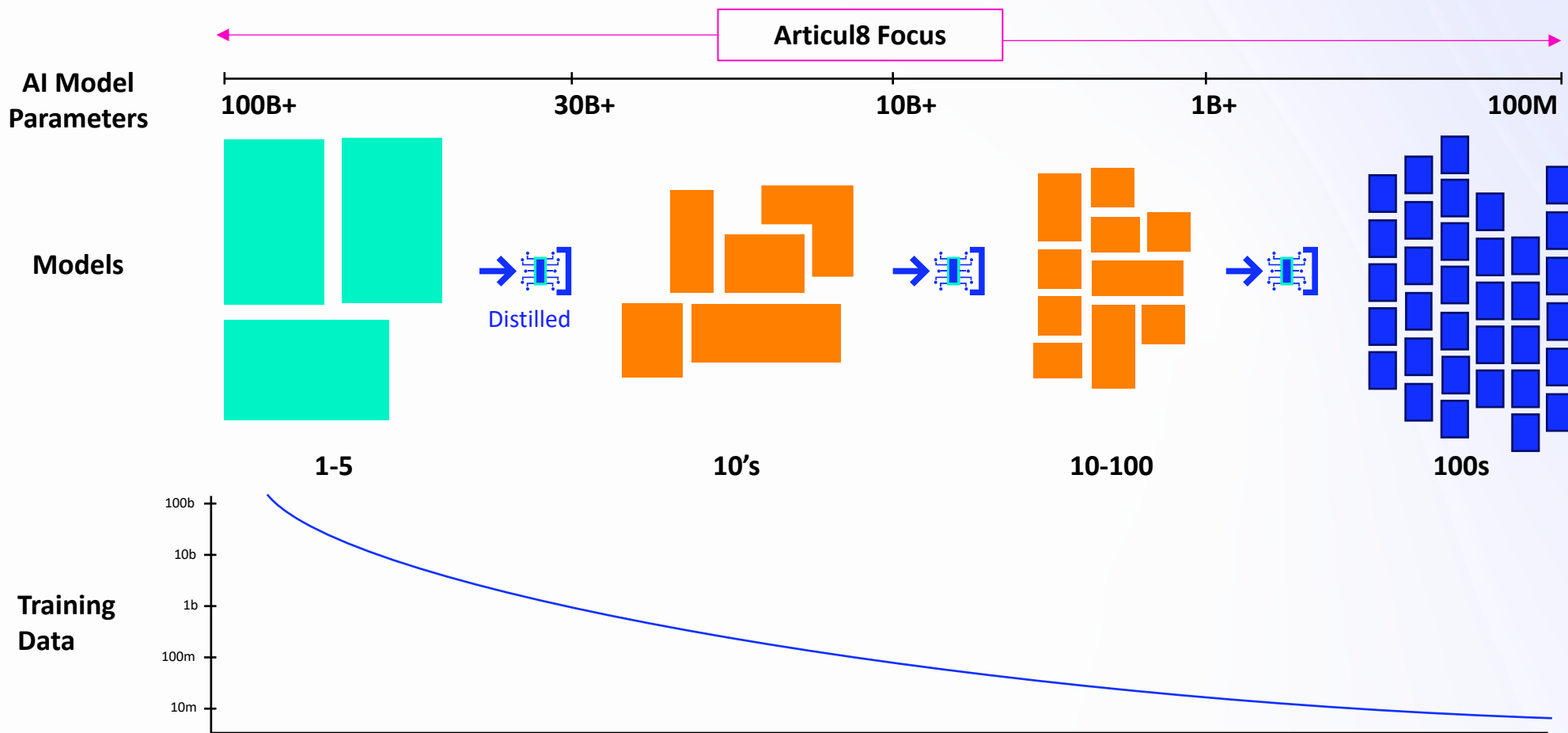
4S: Scale, Speed, Security, & Sustainable Cost

Highest Value Gen AI Use Cases Require Proprietary Data & Are Complex to Scale



Articul8 Approach to Domain-Specific Models (DSMs)

Articul8 approach is to build data efficient, task-specific DSMs and autonomously orchestrate them at scale with Articul8's ModelMesh™ technology



Task-specific DSMs allow for an iterative and incremental approach to training the appropriate # of DSMs for the task on hand

Expert reinforced, continuous learning and attestation

Articul8's Challenges with Model Training, Fine-Tuning, and Inference

01

Tracking Interruptions

Hardware failures, among a number of other factors, cause frequent interruptions. Troubleshooting and restarting/replacing faulty instances manually is tedious & time-consuming.

02

Resource Allocation

GPU resources are expensive and to minimize wastage of unused resources, grouping jobs by teams and allocating resources accordingly, is not available today.

03

Prioritization

Leveraging the same set of resources for multiple tasks simultaneously requires an effective and efficient means to prioritize one task over another.

04

Traceability & Accountability

Explainability, traceability, accountability, and auditability are critical, especially in regulated industries.

05

Cost Management & Optimization

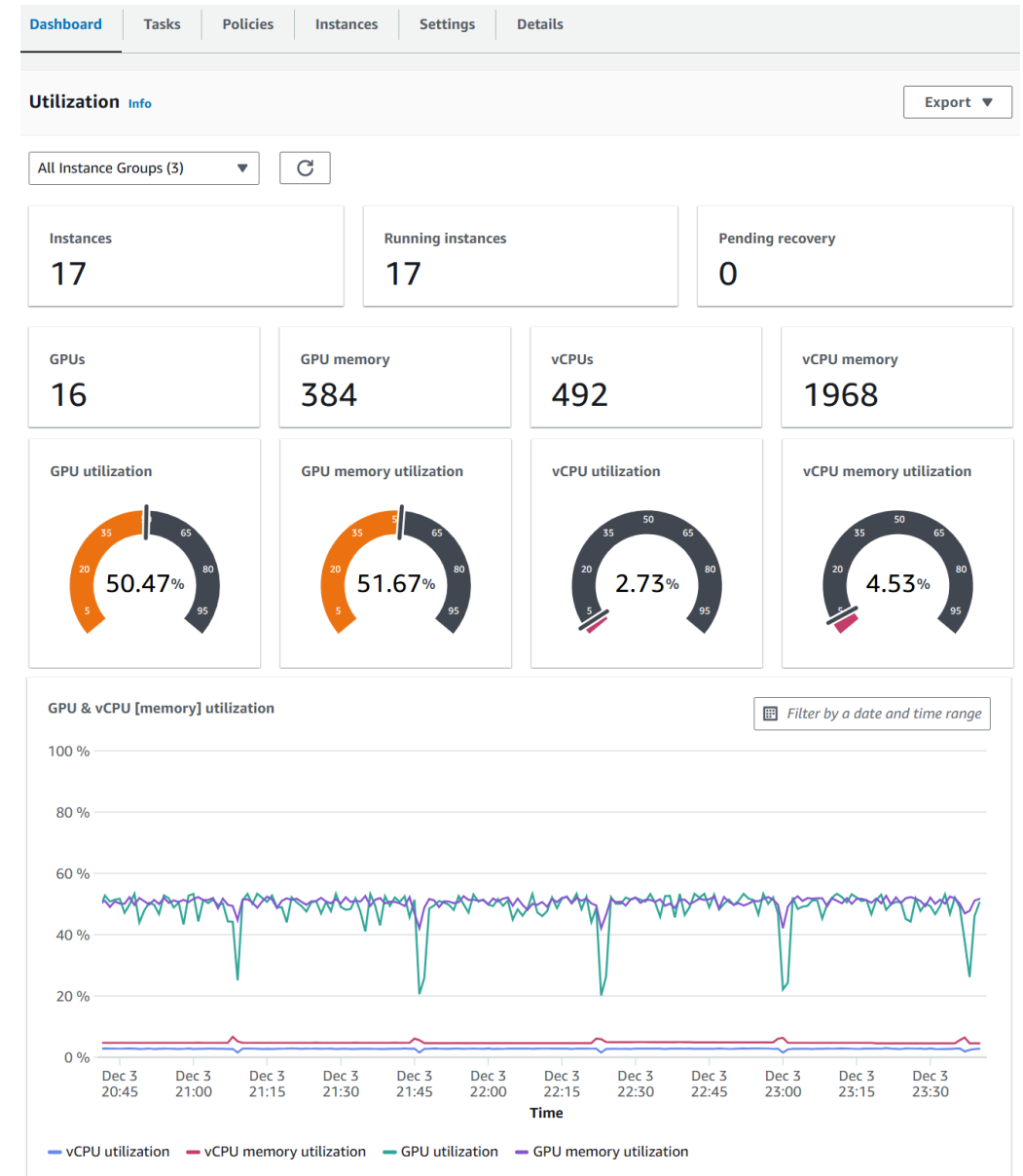
Resource sharing across multiple customers in our SaaS model (for Inference) has its own set of cost-tracking, management, and optimization challenges.

Tracking Interruptions

Hardware failures, among a number of other factors, cause frequent interruptions. Troubleshooting and restarting/replacing faulty instances manually is tedious & time-consuming.

Amazon SageMaker HyperPod task governance Cluster Metrics

- Single pane of glass for insight into health of the cluster
- Provides information available and unavailable/faulty resources



01

Tracking interruptions

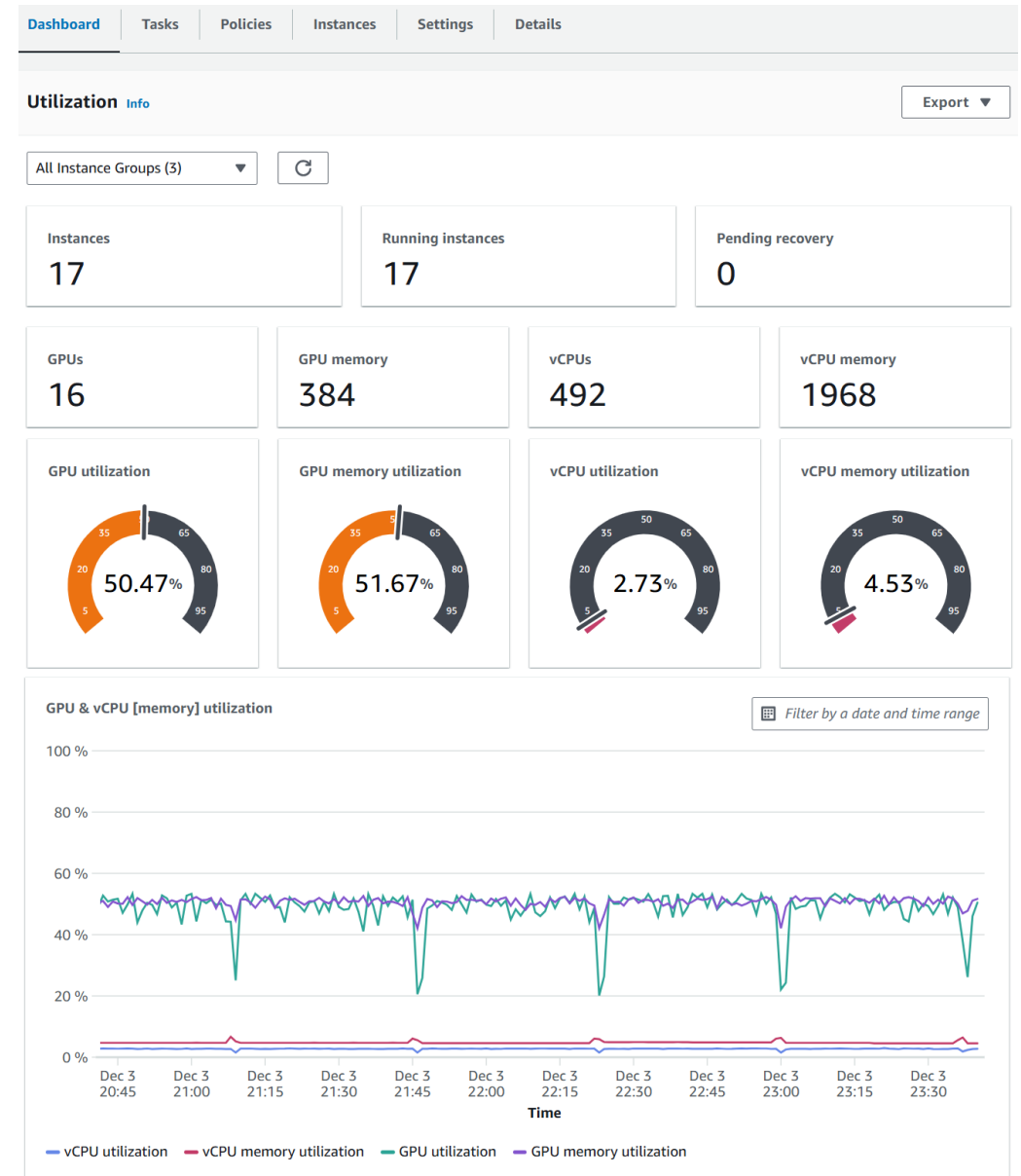
RESOLVED

Hardware failures, among a number of other factors, cause frequent interruptions. Troubleshooting and restarting/replacing faulty instances manually is tedious & time-consuming.

Amazon SageMaker HyperPod task governance

Cluster Metrics

- Single pane of glass for insight into health of the cluster
- Provides information available and unavailable/faulty resources



02

Resource Allocation

RESOLVED

GPU resources are expensive and to minimize wastage of unused resources, grouping jobs by teams and allocating resources accordingly, is not available today.

Amazon SageMaker HyperPod task governance

Compute Allocation & Preemption

Team [Info](#)

Name* [Info](#)
Enter the name of the team.

Namespace
Namespace will be auto-generated based on the defined team name.

Members [Info](#)
You will need to set up Kubernetes role-based access control (RBAC) for data scientist users in the above namespace to run tasks on HyperPod clusters orchestrated with Amazon EKS. [Learn more.](#)

Fair-share weight* [Info](#)
Assign a team weight (0-100, with 0 being the default). Idle compute will be shared across teams based on these assigned weights. Team weights are only used when 'Fair-share' is enabled in the cluster policy.

Task preemption [Info](#)
To enable preemption based on task priority, you must configure priority classes in the cluster policy 'Task prioritization' settings.
 Yes, preempt team's low priority tasks to admit waiting high priority tasks when allocated compute is fully utilized

Compute [Info](#)

Quota* [Info](#)
Enter the amount of instances that should be allocated to the team. Quota can be more than what instances are currently available.

Instance type	Instance count
<input type="text" value="p5.ml.xlarge (60 available)"/>	<input type="text" value="0"/>
<input type="text" value="p4.ml.xlarge (25 available)"/>	<input type="text" value="0"/>
<input type="text" value="p3.ml.xlarge (15 available)"/>	<input type="text" value="0"/>

Lending and borrowing* [Info](#)
Allow teams to automatically lend their idle compute resources. Teams that lend compute, can borrow compute automatically.

Lend and borrow
Enable team's idle compute to be borrowed by other teams.

Lend
Allow team to share their idle compute resources.

Don't lend
Reserve all allocated compute for this team.

Borrow limit* [Info](#)
Specify the limit (1-500%, with 50% being the default) of idle compute that team is allowed to borrow.

03

Prioritization

RESOLVED

Leveraging the same set of resources for multiple tasks simultaneously requires an effective and efficient means to prioritize one task over another.

Amazon SageMaker HyperPod task governance

Cluster Policy & Prioritization

Tasks (25) [Info](#)

Find priority class

Priority class	Running	Pending	Pre-empted	Avg. run time	Avg. wait time	Avg. longest run time	Avg. longest wait time
Inference	40	10	4	1h 15m	25m	1h 15m	25m
Interactive	40	10	4	1h 15m	25m	1h 15m	25m
Experiment...	40	10	4	1h 15m	25m	1h 15m	25m
Training	40	10	4	1h 15m	25m	1h 15m	25m
Fine-tuning	40	10	4	1h 15m	25m	1h 15m	25m

Amazon SageMaker > Cluster management > eks-no-addon

eks-no-addon [Inservice](#) [Edit](#) [Delete](#) [Monitor in Container Insights](#)

Metrics | Tasks | **Policies** | EKS add-ons | Settings | Details

Cluster policy info [Info](#) [Edit](#)

Task prioritization
Task ranking: Tasks waiting in queue, will be admitted in the priority order defined in this policy. Tasks of same type will be executed on first-come-first-serve basis.

Task ranking info

Priority class	Weight
real-time-inference	100
experimentation	80
training	70
fine-tuning	60
inference	50

Idle compute allocation
First-come-first-serve: This setting enables teams to borrow compute resources on first-come first-serve basis.

Edit cluster policy info [Info](#)

Configure priority classes and fair sharing of borrowed compute in cluster settings.

Task prioritization info

This configuration defines how tasks waiting in queue are admitted. Default setting admits tasks waiting in the queue on a first-come, first-served basis. You can configure this setting to define priority classes. Waiting tasks will then be admitted based on their assigned priorities.

First-come-first-serve
Tasks waiting in queue will be admitted on first-come first-serve basis.

Task ranking
Tasks waiting in queue, will be admitted in the priority order defined in this cluster policy. Tasks of same priority class will be executed on first-come-first-serve basis.

Task ranking info
Add priority classes and relative weights as they should be admitted.

Priority class	Weight	
real-time-inference	100	Remove
experimentation	80	Remove
training	70	Remove
fine-tuning	60	Remove
inference	50	Remove
Add		

Idle compute allocation info

This configuration defines how idle compute is allocated across teams. The default is a 'fair-share' model, where compute is distributed based on assigned team weights, which are configured in relative quota policies.

First-come-first-serve
This setting enables teams to borrow compute resources on first-come first-serve basis.

Fair-share
This setting enables teams to borrow idle compute based on their assigned weights, which are configured in relative quota policies.

[Cancel](#) [Submit](#)

Traceability & Accountability

RESOLVED

Explainability, traceability, accountability, and auditability are critical, especially in regulated industries.

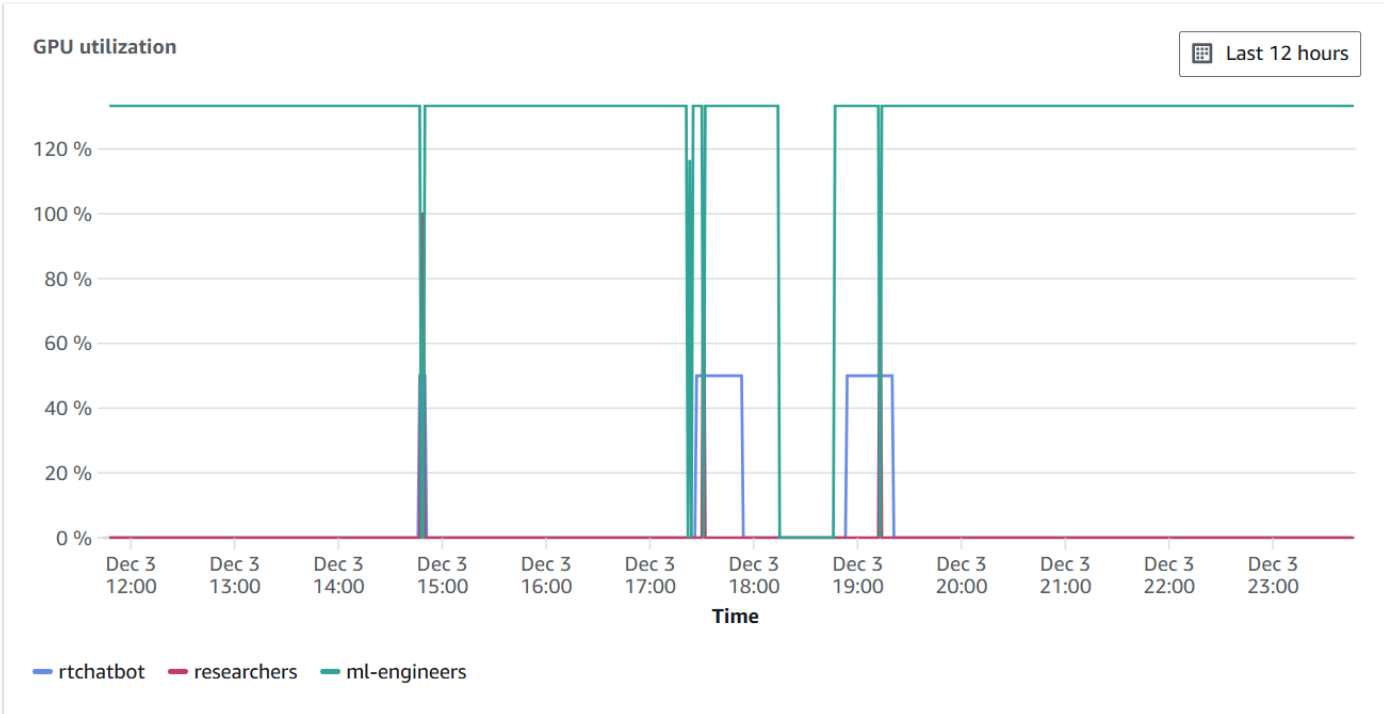
Amazon SageMaker HyperPod task governance

Team Details

Team details

All teams (3) [dropdown] [dropdown] [refresh]

Allocated GPUs 15	Allocated GPUs to tasks 8	Borrowed GPUs 2	GPU utilization 53.33%
-----------------------------	-------------------------------------	---------------------------	----------------------------------



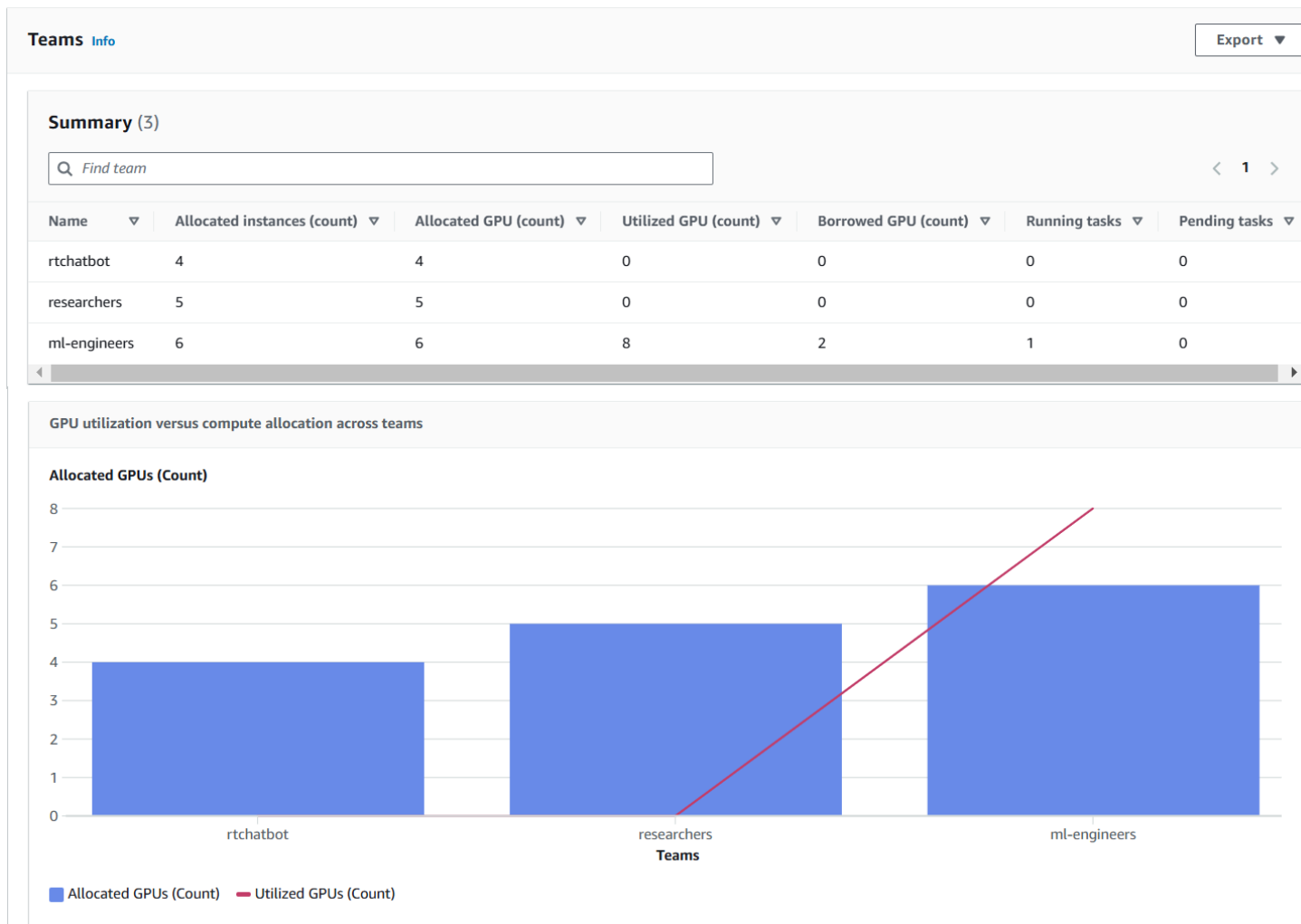
RESOLVED

Cost Management & Optimization

Resource sharing across multiple customers in our SaaS model (for Inference) has its own set of cost-tracking, management, and optimization challenges.

Amazon SageMaker HyperPod task governance

Team Metrics



Articul8 + AWS => Faster Time to Market

A8 Essential

The essential gen AI experience with your data, delivering outcomes from Day 1

A8 Enterprise

Build resilient enterprise gen AI applications and derive ROI within 6 weeks

A8 Expert

Build expert-level gen AI applications that encode your enterprise domain expertise

Ingest

Please choose the data source you would like to add.

I'd like to upload my own sample data. I'd like to use the A8 Essentials sample data.

I ensure that this file contains no unauthorized or sensitive PII.

Click to upload or drag and drop
Supported formats: Text, HTML, CSV, XLS, DOC, and JSON.
Maximum size: 100 MB

Note: We won't share or use your data without your consent.

Files Uploading

DATA FILE	STATUS	RETENTION	FILE SIZE	ACTION
Patient_details_mm_dd_yyyy.csv	Uploaded	Retained	200 KB	
Patient_details_mm_dd_yyyy.csv		Retained	300 KB	
Patient_details_mm_dd_yyyy.csv		Retained	100 KB	
Patient_details_mm_dd_yyyy.csv	Uploaded	Retained	1 MB	
Patient_details_mm_dd_yyyy.csv		Retained	600 KB	

Data Quality Score: 68%

Decisions Taken: 165

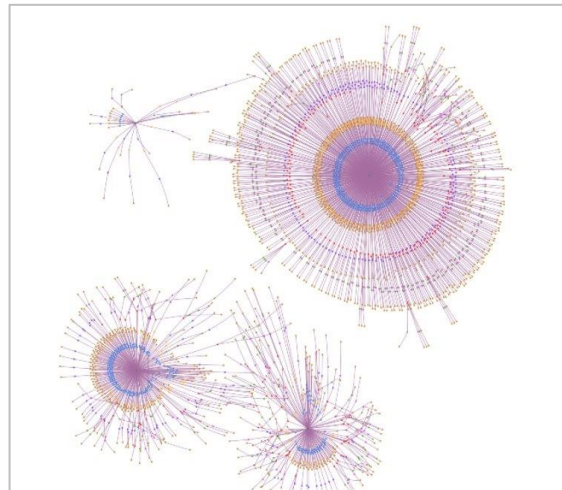
Data Overview

- Files: 1,281
- Tokens: 12,910
- Detected Entities: 8,819
- Detected Connections: 1,211
- Detected Tables: 6,078
- Detected Figures: 3,187
- Embedding Dimensions: 2,108

Knowledge Graph

View Data Insights

Perceive



Act

Geographical Distribution of AWS Revenue

Back to Thread View

All Topics

- Geographical Distribution of AWS Revenue
- AWS Revenue by Service Offering
- Impact of New AWS Services on Revenue Growth
- AWS Revenue Growth Over Time
- AWS Revenue Contribution to Amazon's Overall Revenue

How does AWS's revenue from international markets compare to its U.S. revenue?

Which regions are driving AWS's international revenue growth?

How does AWS tailor services for international markets?

What challenges does AWS face in global expansion?

How does AWS's pricing differ between and abroad?

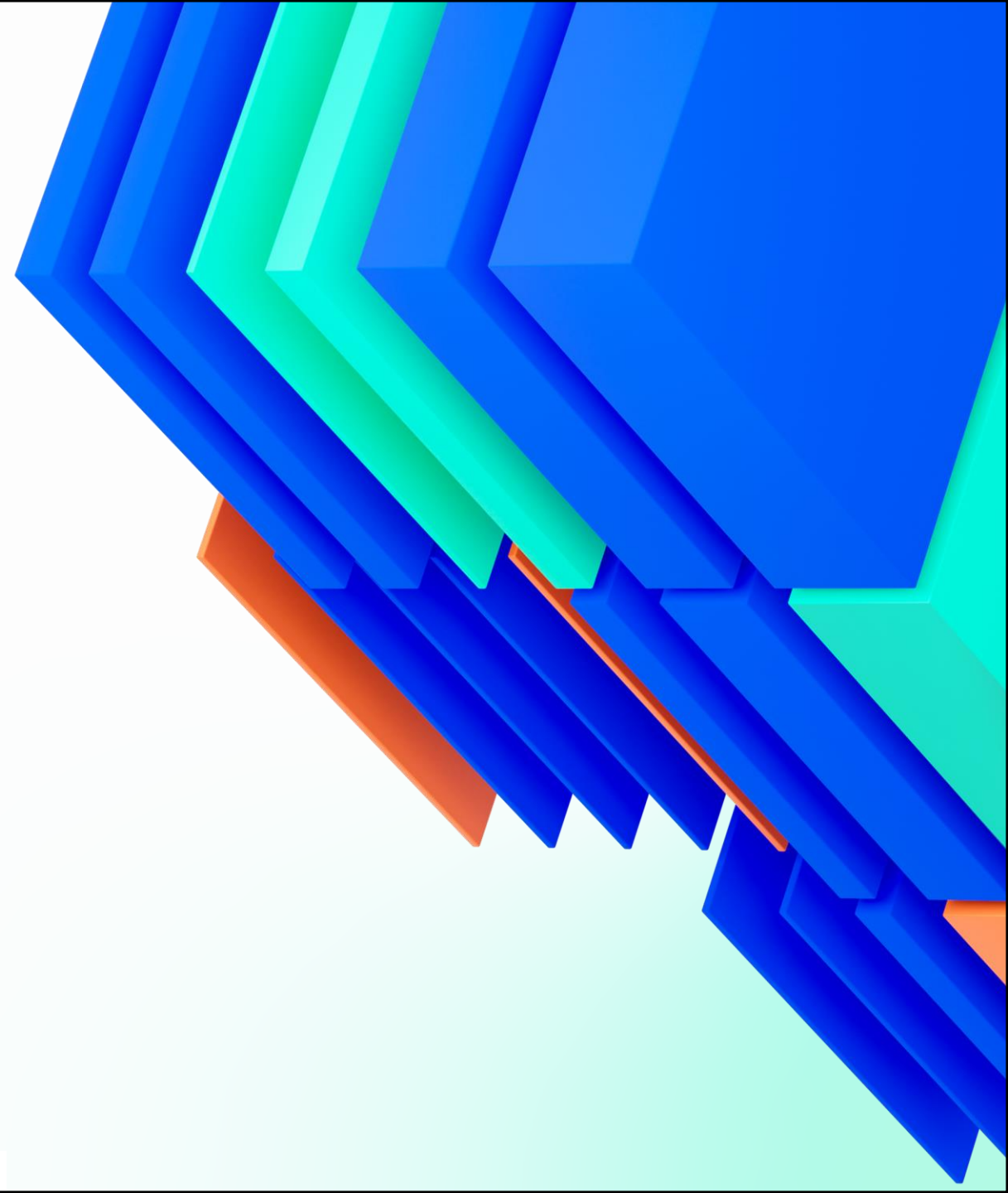
How does AWS's pricing differ between and abroad?

Query

Ingest > Perceive > Decide > Act > Outcomes...all in a few hours.

Articul8

Thank you!



Conclusion



NEW

Amazon SageMaker HyperPod task governance

Maximize accelerator utilization and reduce costs
for model training, fine-tuning, and inference

- Dynamically allocate compute resources across tasks
- Ensure high-priority tasks are completed on time
- Monitor and audit compute allocation in real-time
- Maximize compute resource utilization and reduce costs by up to 40%



Getting started



Announcement blog



SageMaker HyperPod
webpage



Documentation

Thank you!



Please complete the session survey in the mobile app