
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

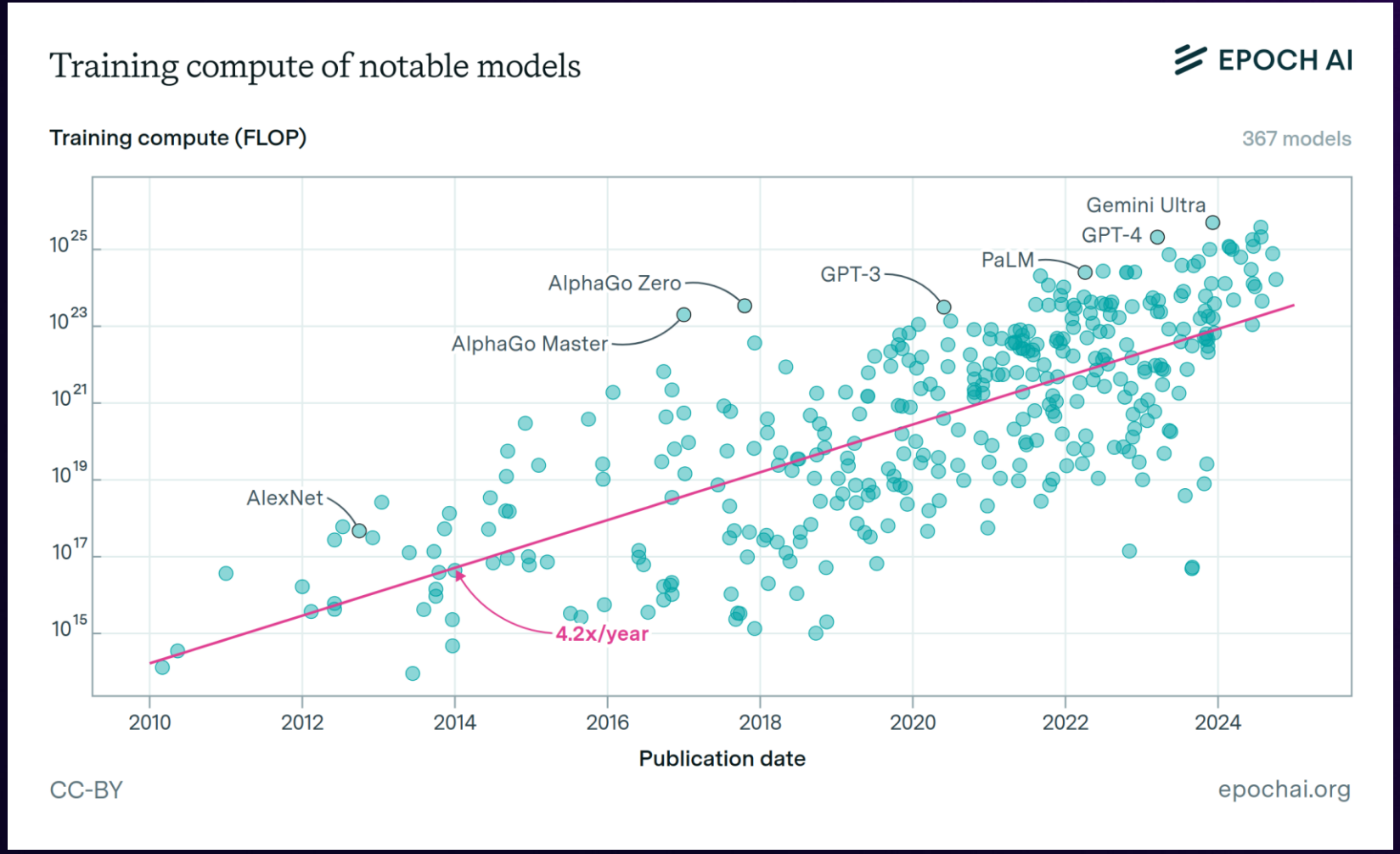
Joy Fan
Sr. Manager,
Software Development
Amazon

Arun Subramaniyan
Founder and CEO
Articul8 AI

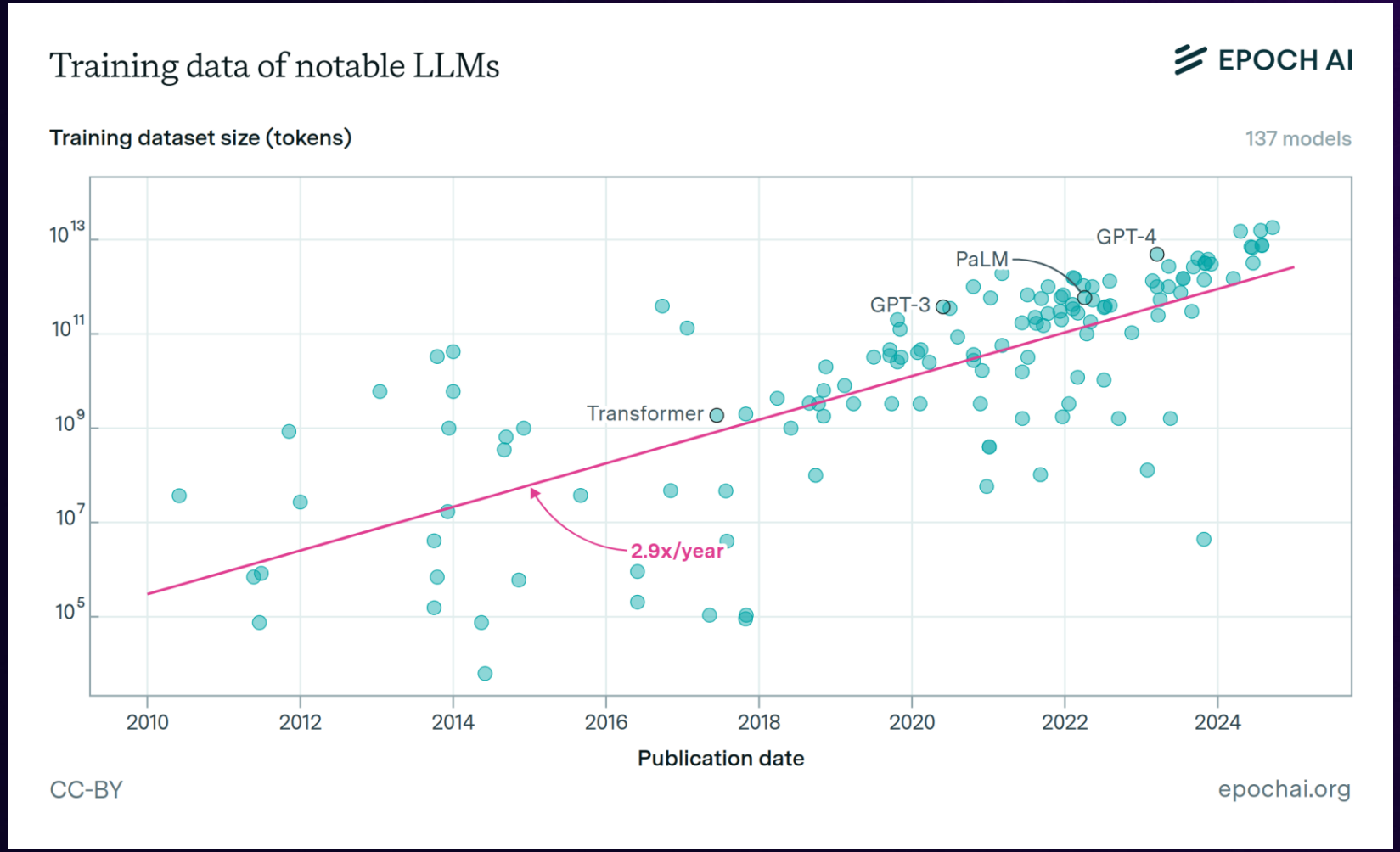
Amazon SageMaker HyperPod

Kareem Syed-Mohammed

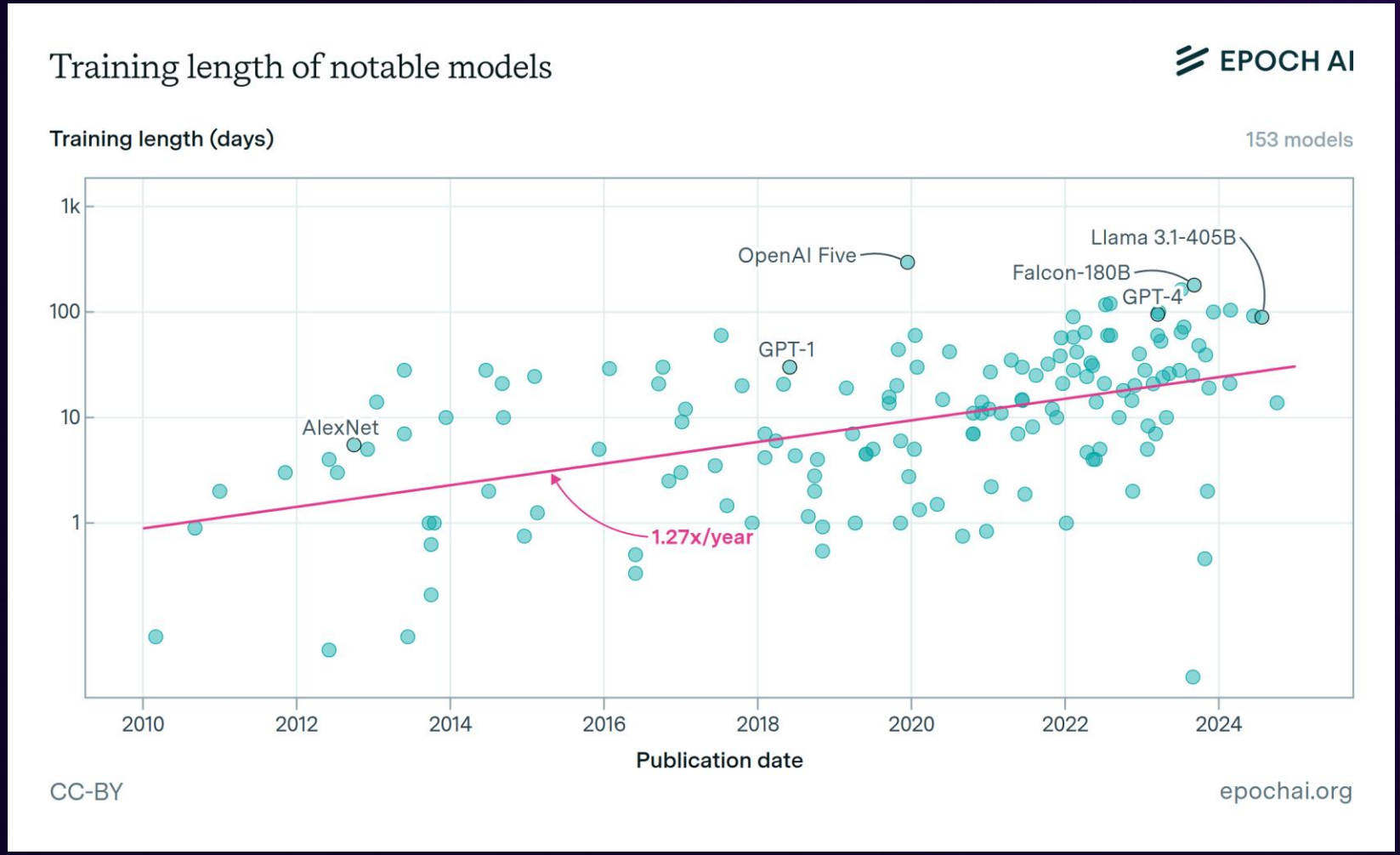
Training compute of frontier models is growing by 4–5x per year, doubling roughly every 6 months



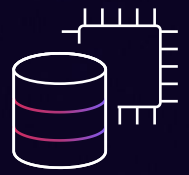
Dataset sizes are doubling every 8 months



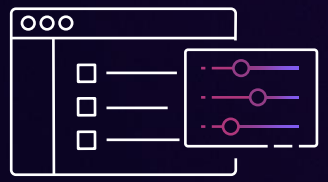
Time to market spans months of investment and continues to increase



Challenges with large-scale gen AI model development



Collect data



Clusters provision & management

Infrastructure stability

Strategies for distributed training

Amazon SageMaker HyperPod

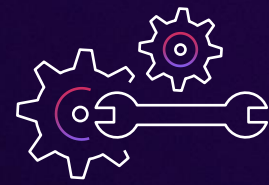
REDUCE TRAINING TIME BY UP TO 40% THROUGH RESILIENCY AND PERFORMANCE OPTIMIZATIONS

Resilient environment

Self-healing clusters reduce training time

Streamline distributed training

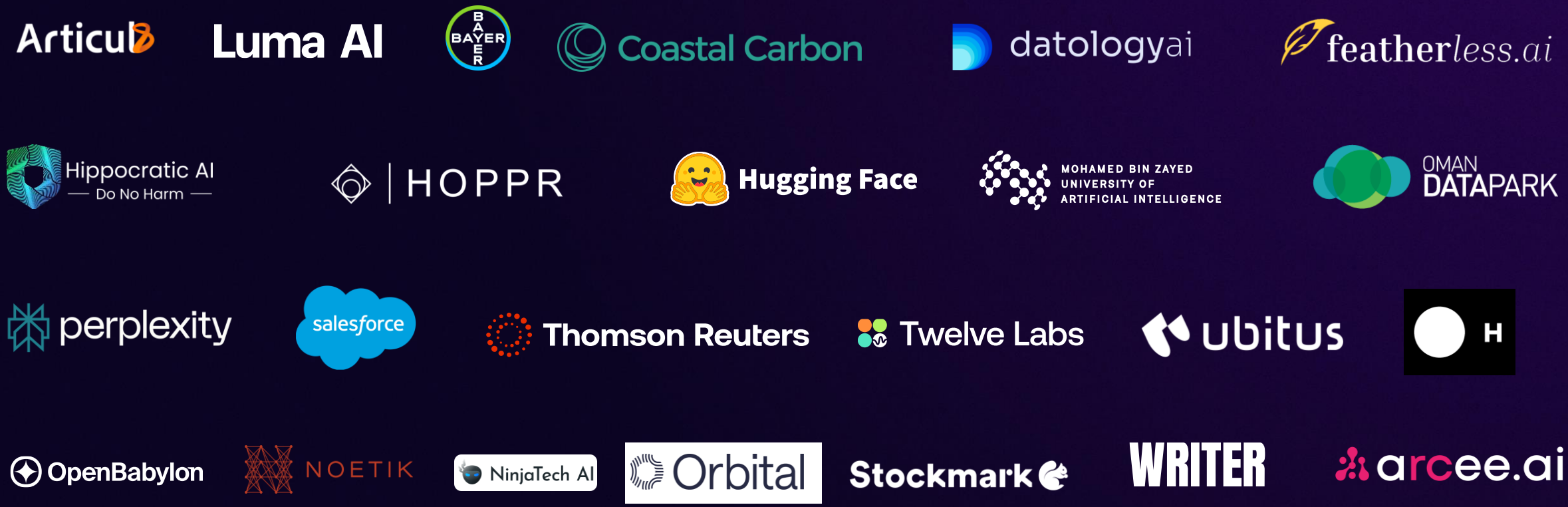
SageMaker distributed training improves performance



Optimized resources utilization

Control over computing environment and workload scheduling

Top AI companies use HyperPod to train and deploy models



RECENTLY LAUNCHED

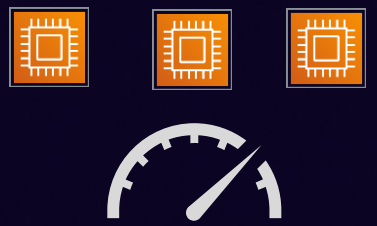
Amazon Elastic Kubernetes Service (Amazon EKS) support in Amazon SageMaker HyperPod

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

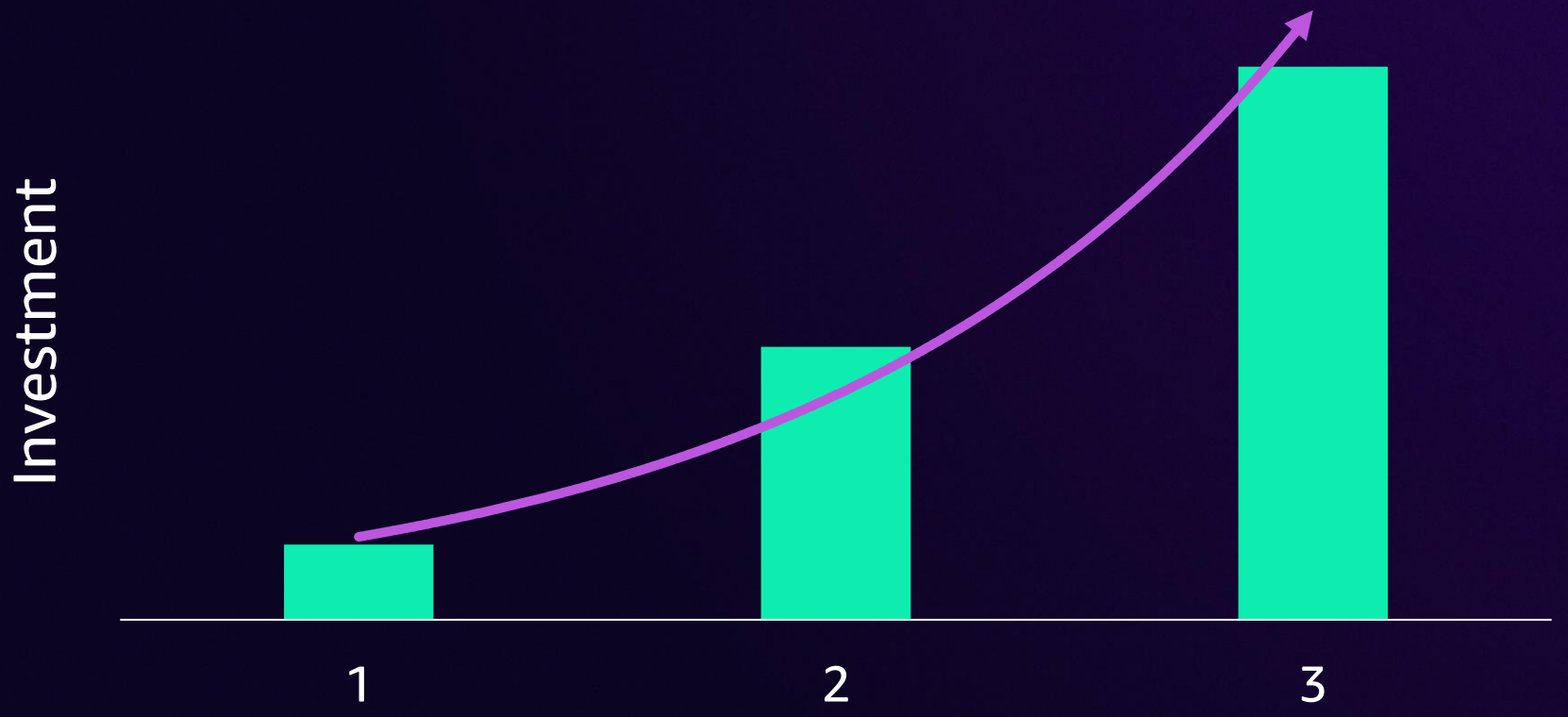
Reduced scientists' productivity

Increased costs

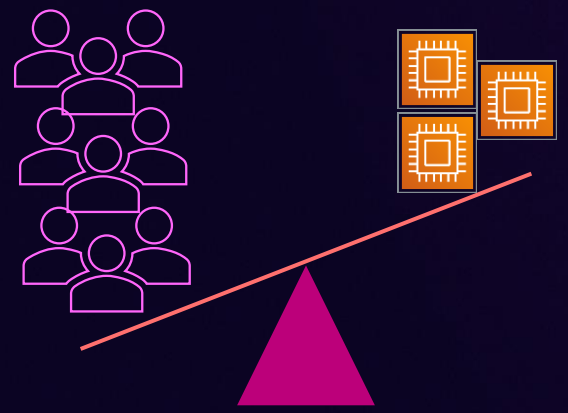
Amazon gen AI challenges and innovation

Joy Fan

Amazon gen AI investment trend



Challenges



High demand

Low supply

Challenges

Static allocation for 100+ teams/projects

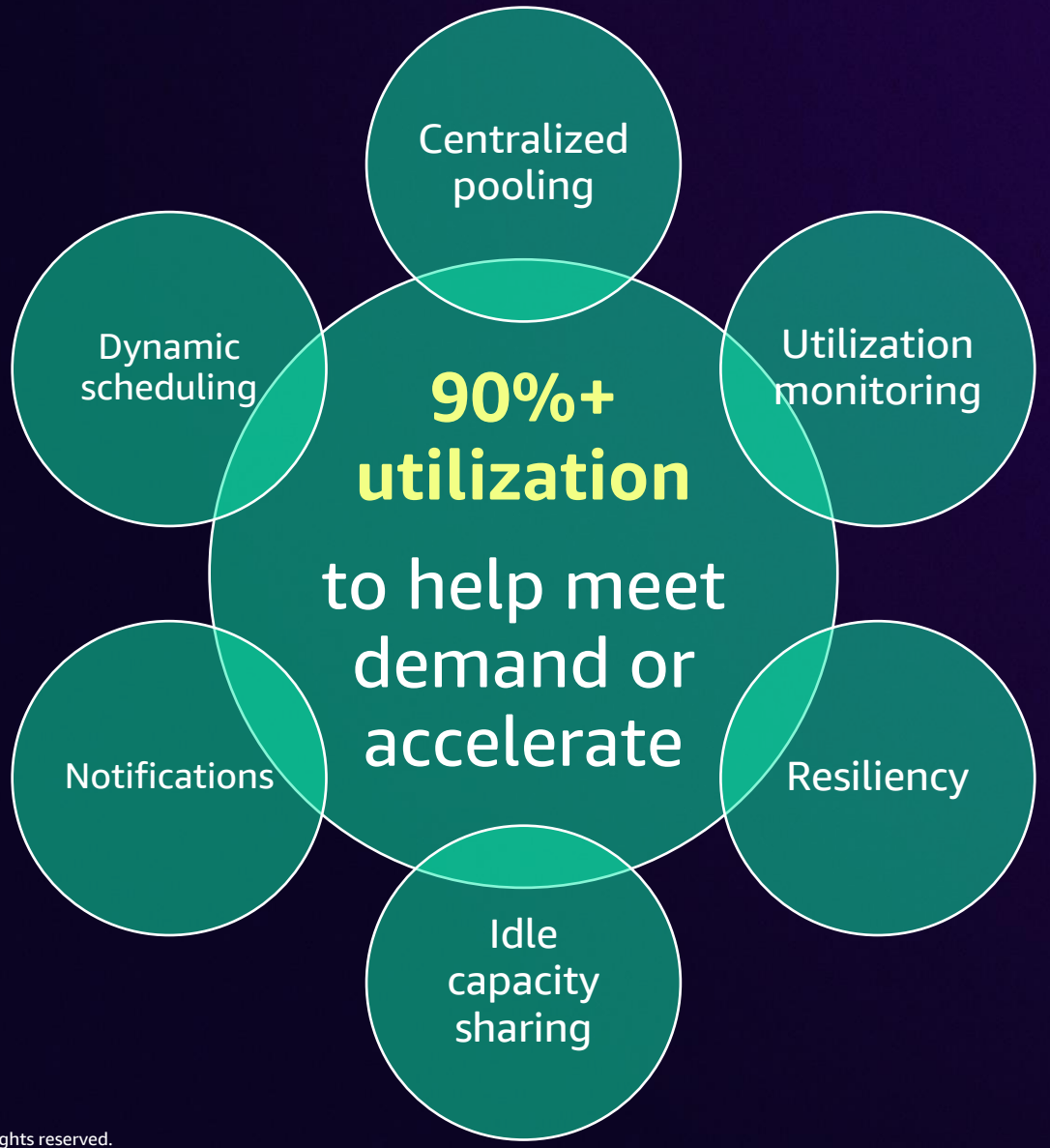
Spikey demand – some wait for allocation while others' resources stay idle

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

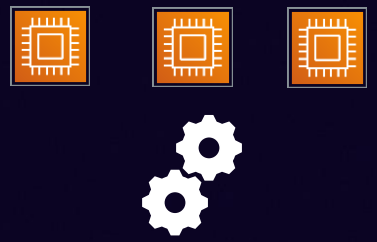
Maximize accelerator utilization and reduce costs for model training, fine-tuning, and inference

- 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%

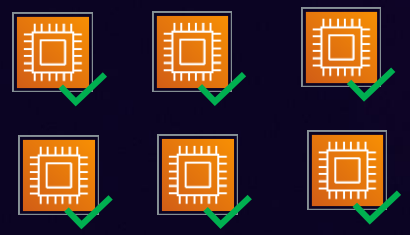
SageMaker HyperPod task governance

PRIORITIZE TASKS, ALLOCATE COMPUTE RESOURCES, AND MAXIMIZE UTILIZATION

A smart scheduler and orchestrator that enables



Dynamic resource allocation



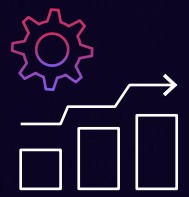
Utilizes idle compute

Reduces task wait times

Real-time task prioritization

Real-time observability

With SageMaker HyperPod task governance, you . . .

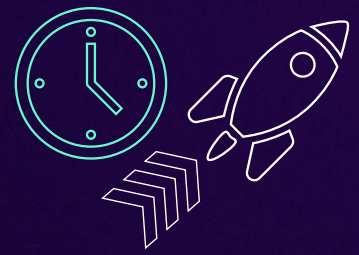


Increase compute utilization

Increases compute utilization by maximizing utilization of idle compute

Reduce costs

Higher utilization can reduce costs up to 40%



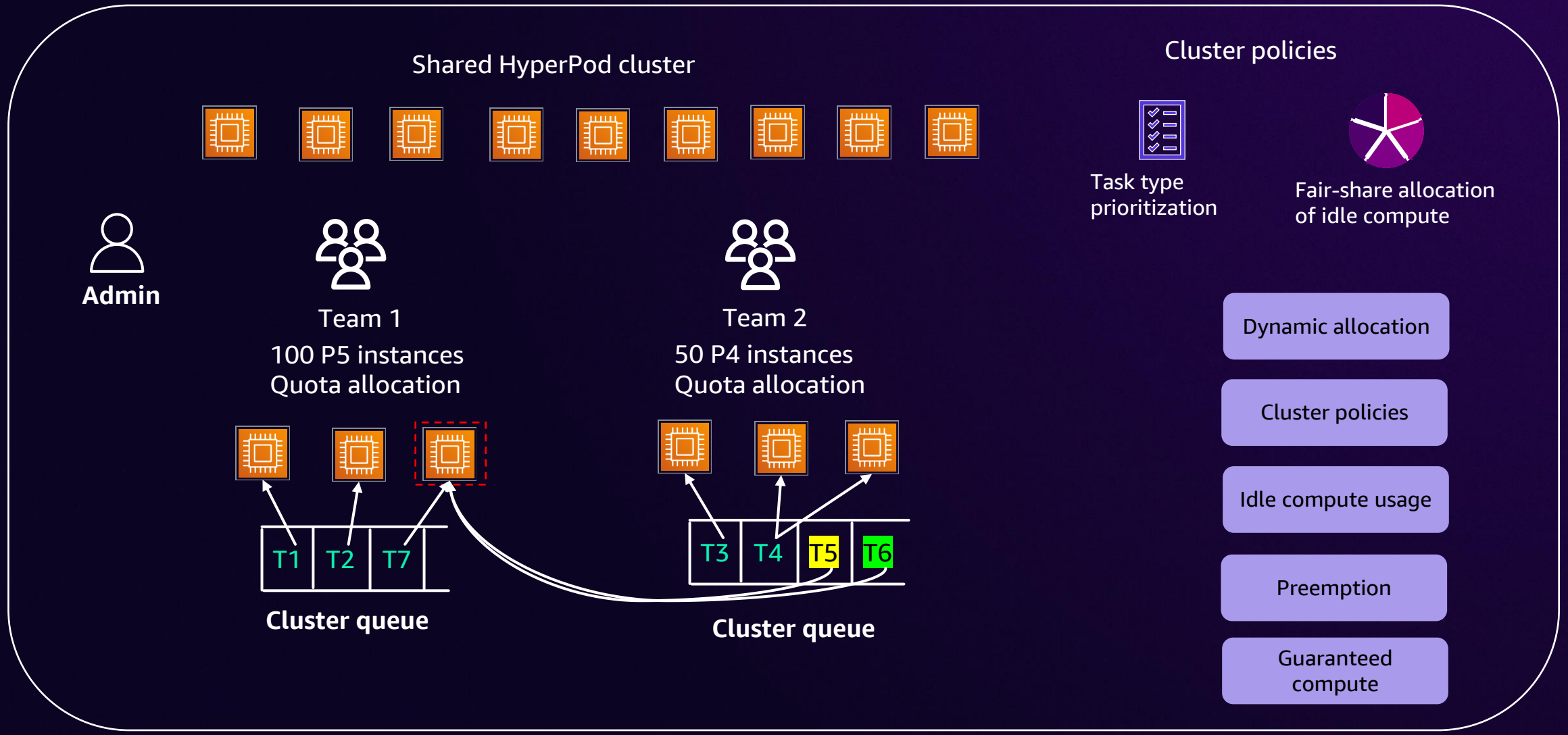
Accelerate time to market

Waiting tasks use idle compute, reducing wait times

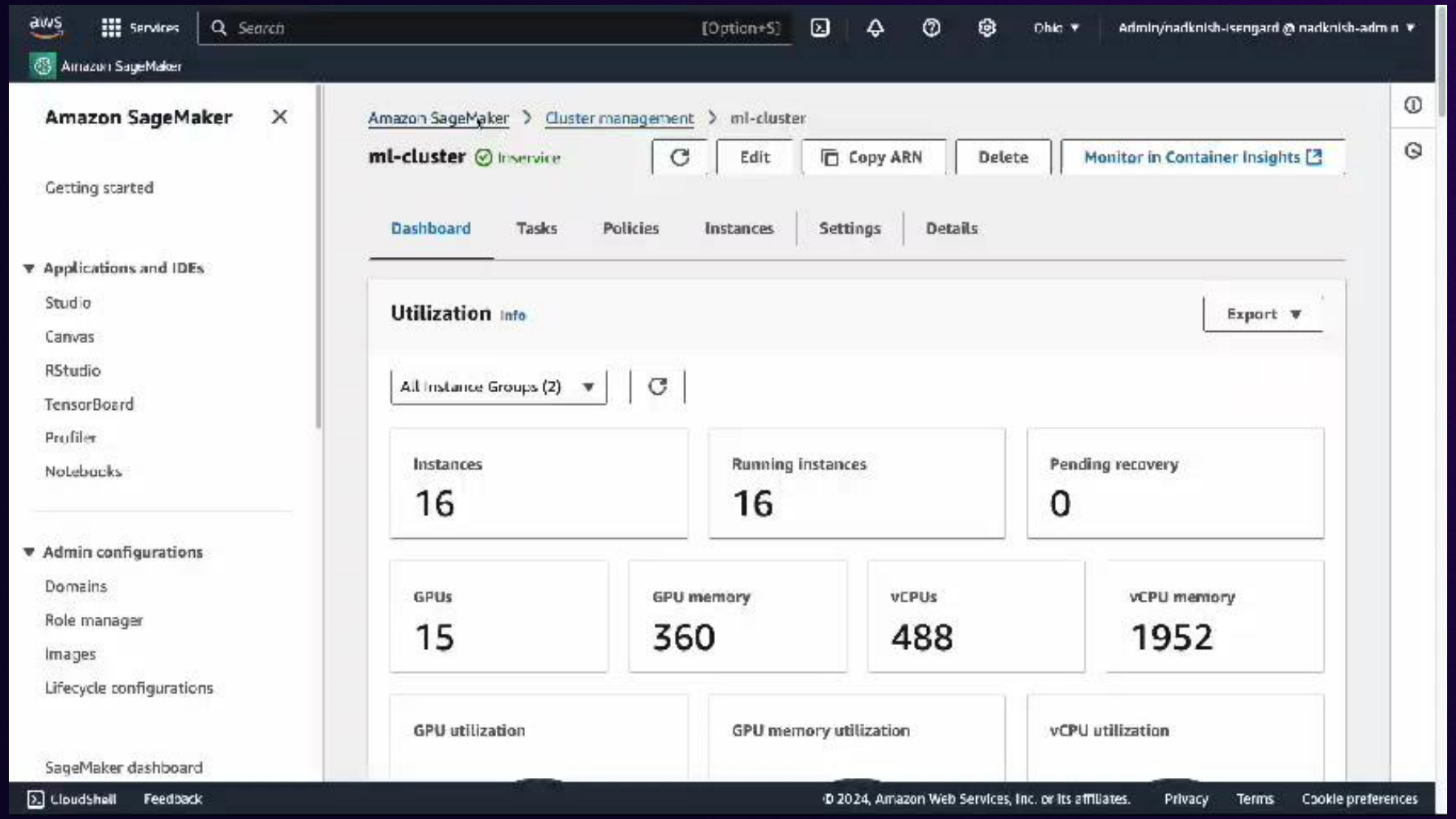
Common use cases

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



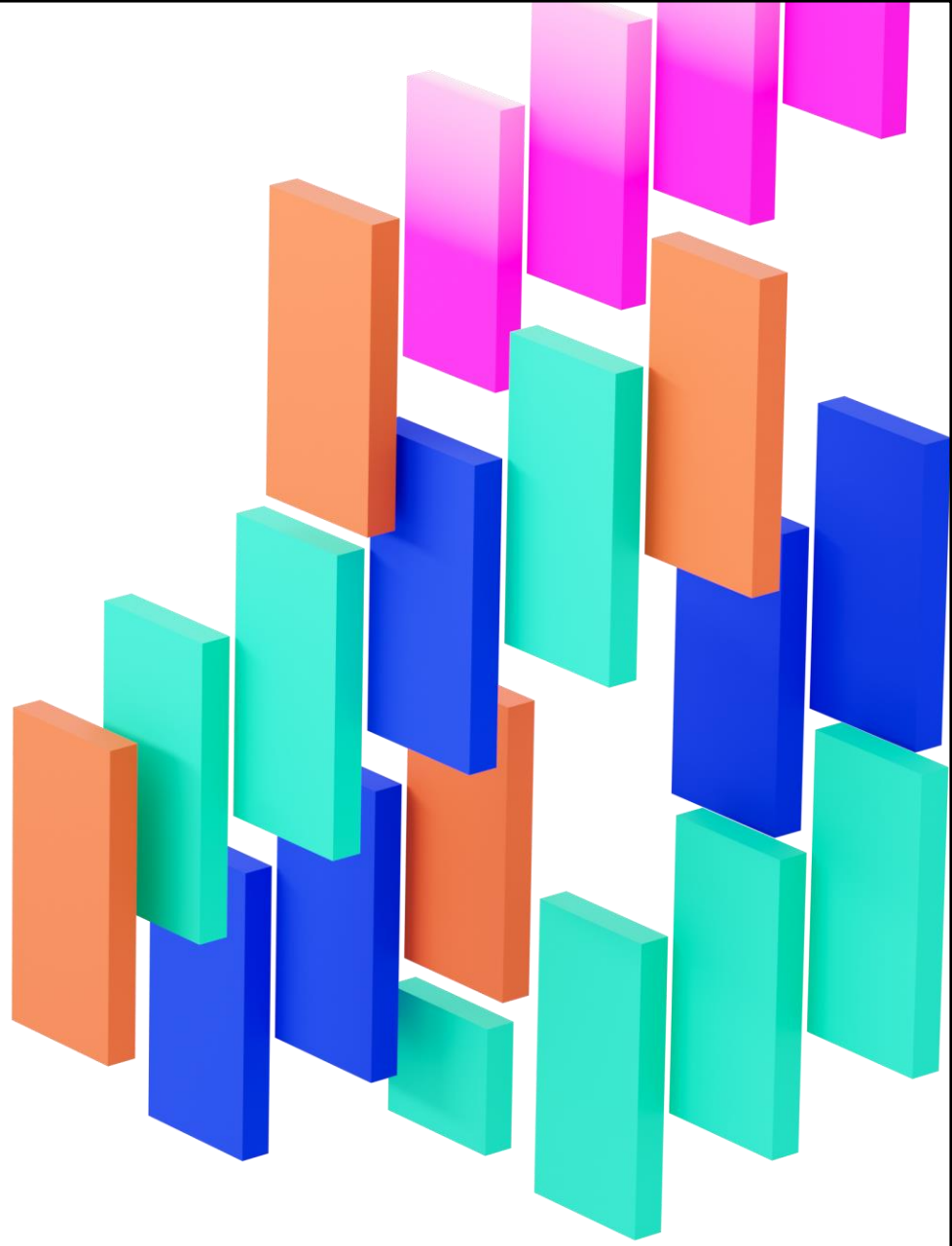
Customer story – Articul8 AI

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

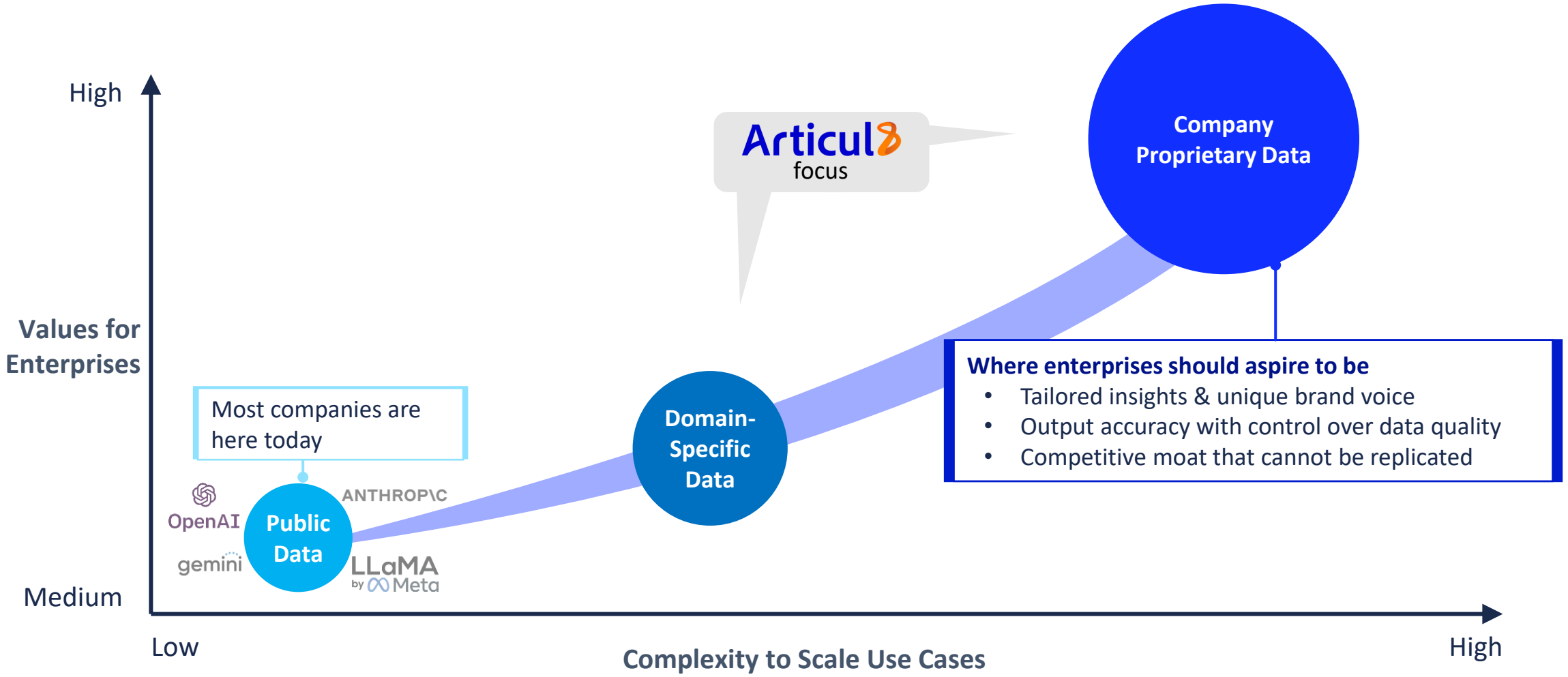
Proven Track Record With Enterprise Customers

FRANKLIN TEMPLETON | NielsenIQ | BCG | intel | Hyperscaler

Articul8 Key Differentiators

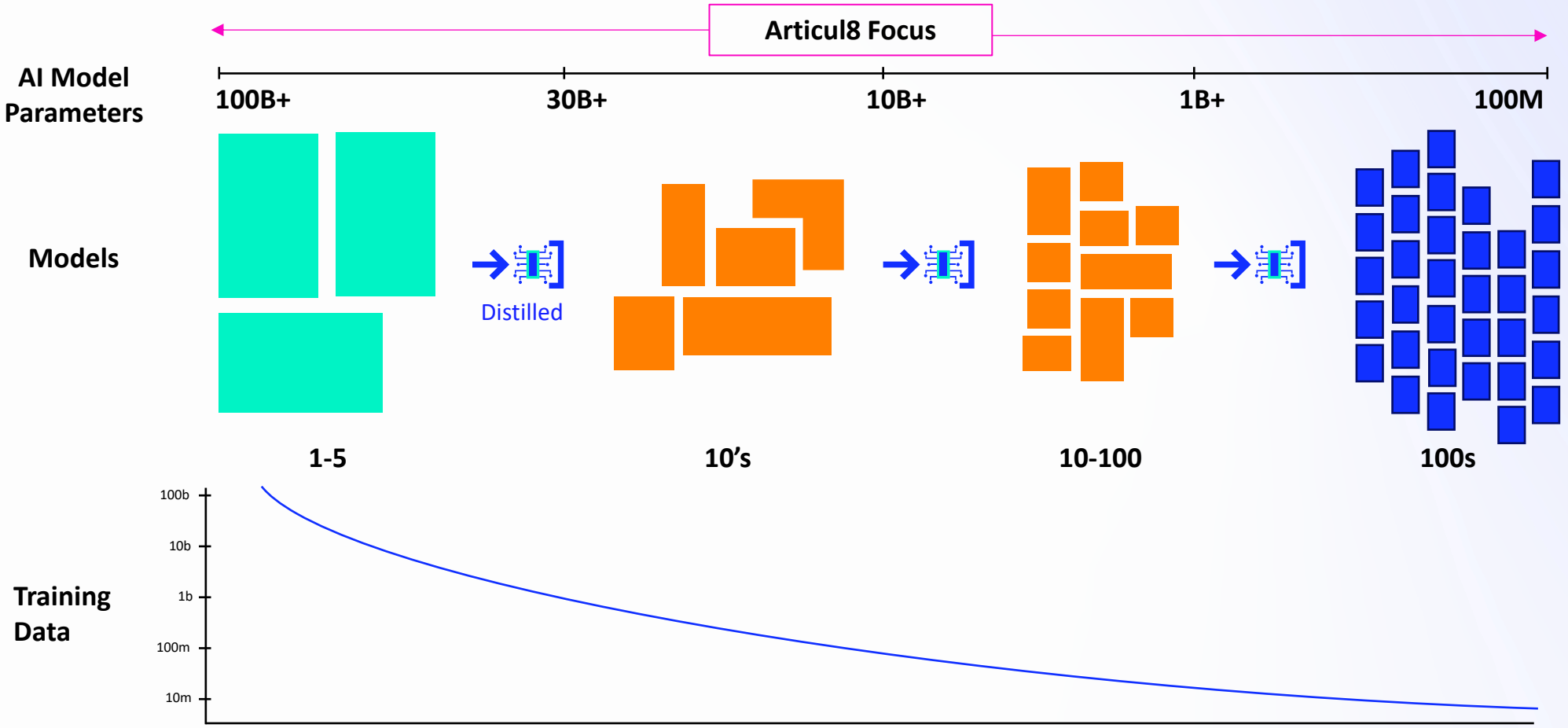
- **Autonomous**: ModelMesh™ selects & manages the right combination of models for autonomous multi-agent decisioning & actioning
- **Domain-specific models**: Build, deploy, and optimize domain-specific models with specialized **data partnerships**
- **4S**: **S**cale, **S**peed, **S**ecurity, & **S**ustainable 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.

01

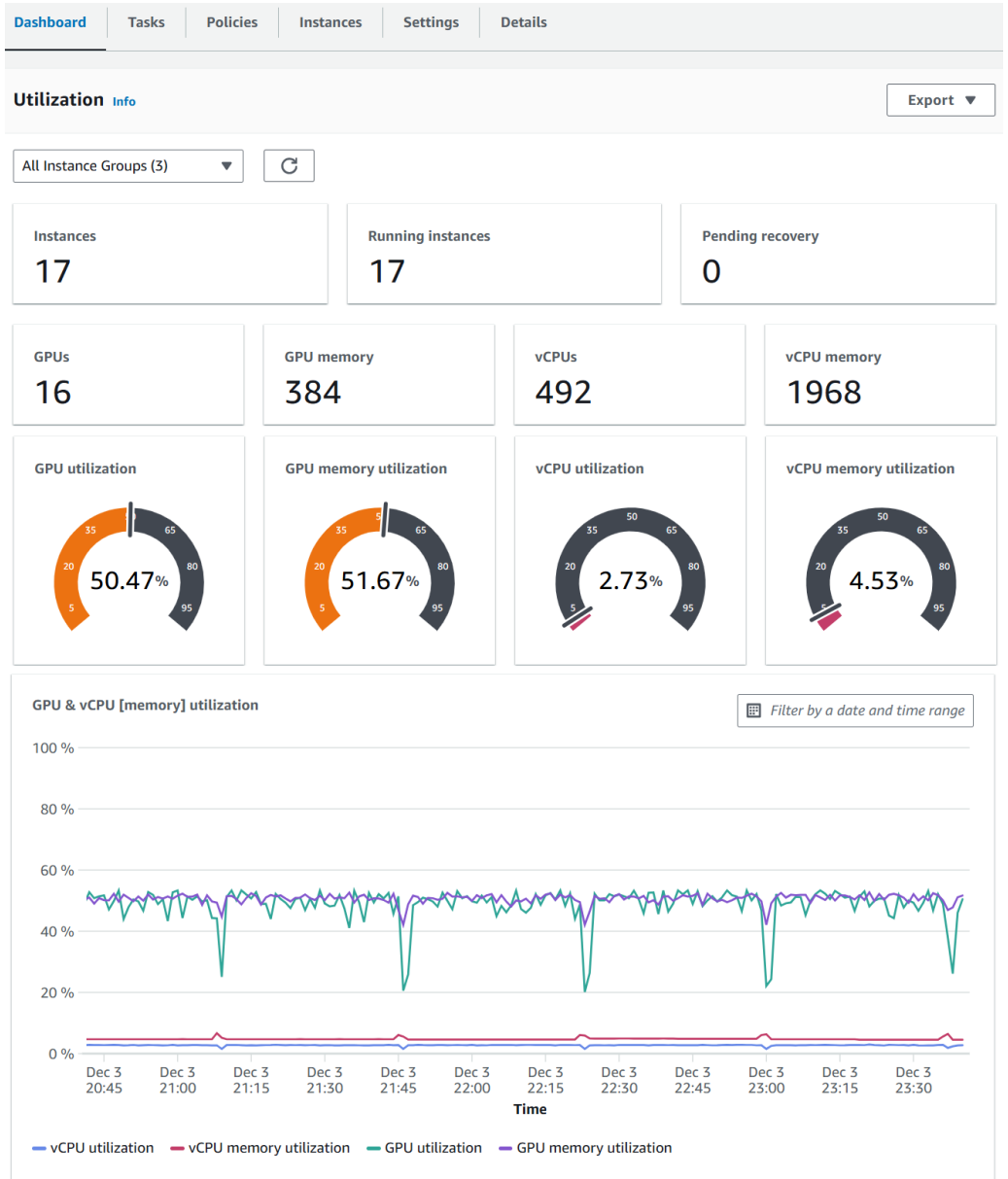
Tracking Interruptions

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

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

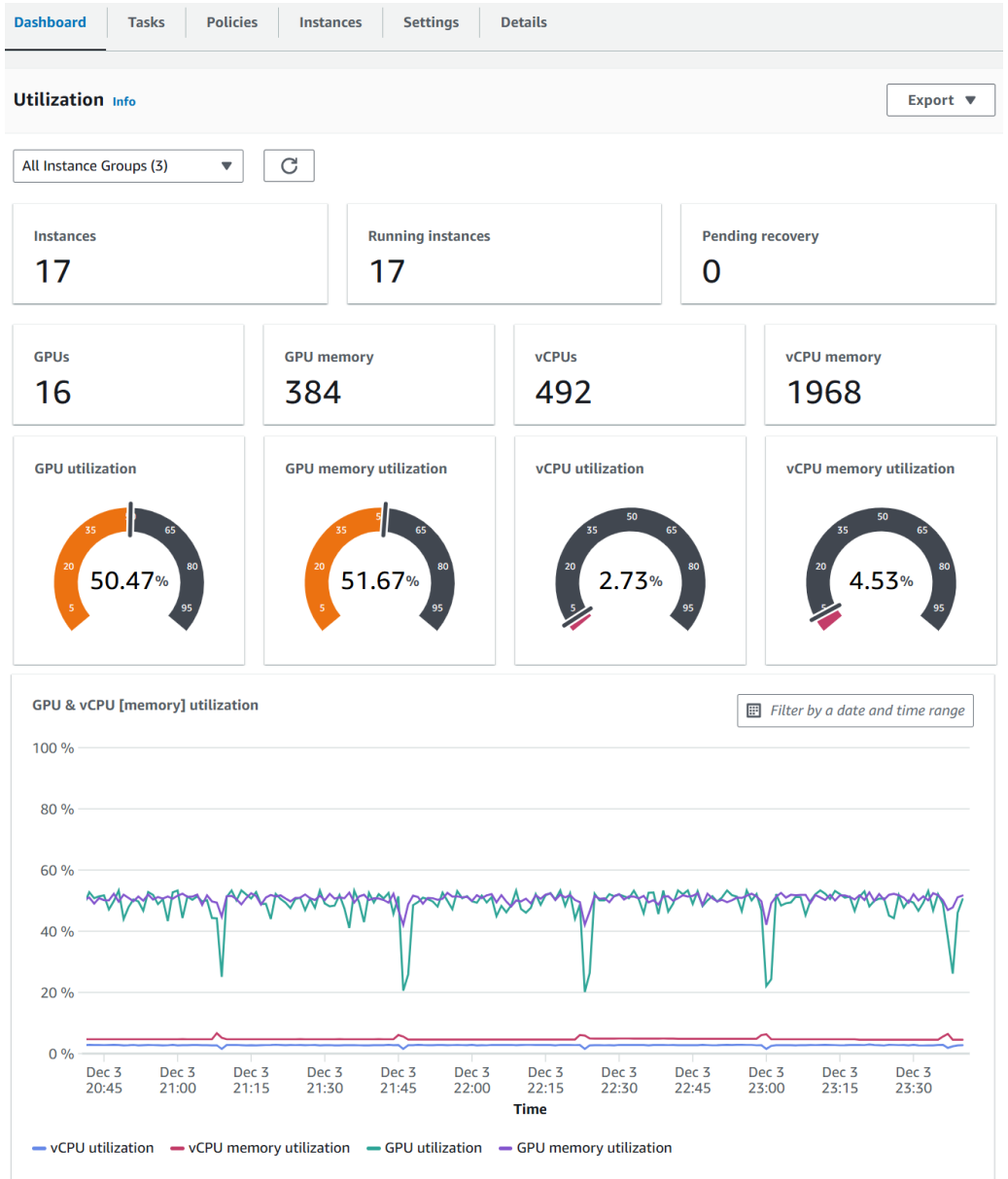
RESOLVED

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

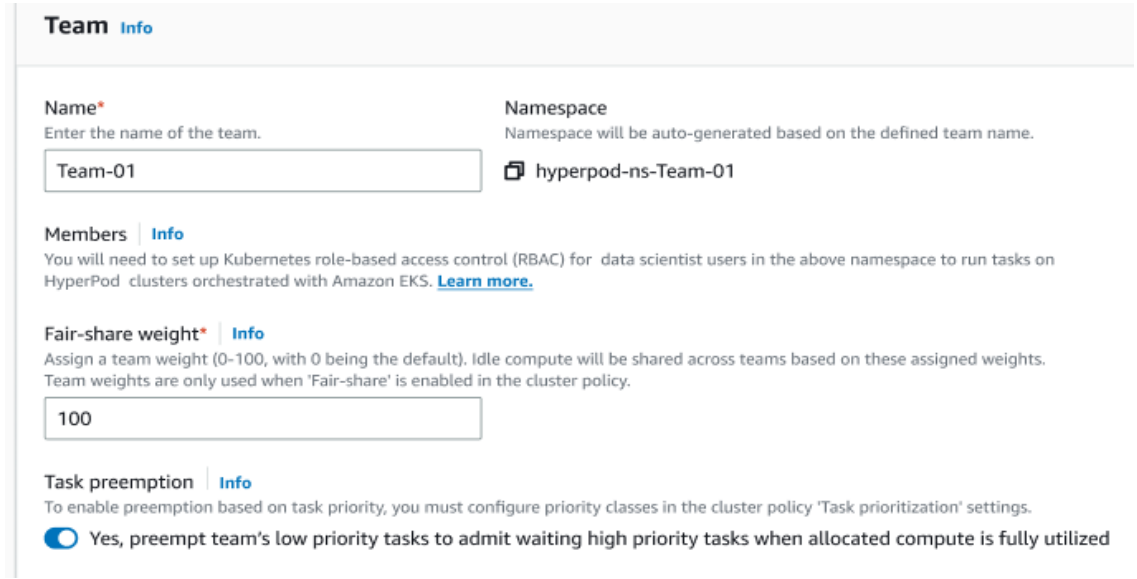
Resource Allocation

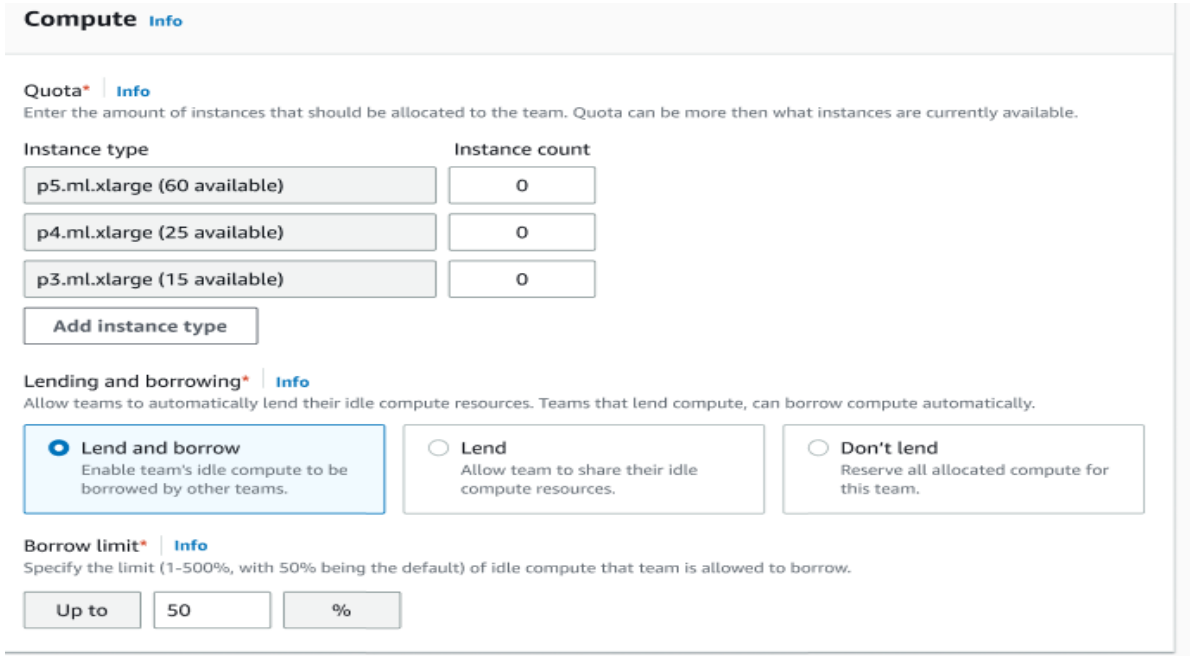
RESOLVED

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





03

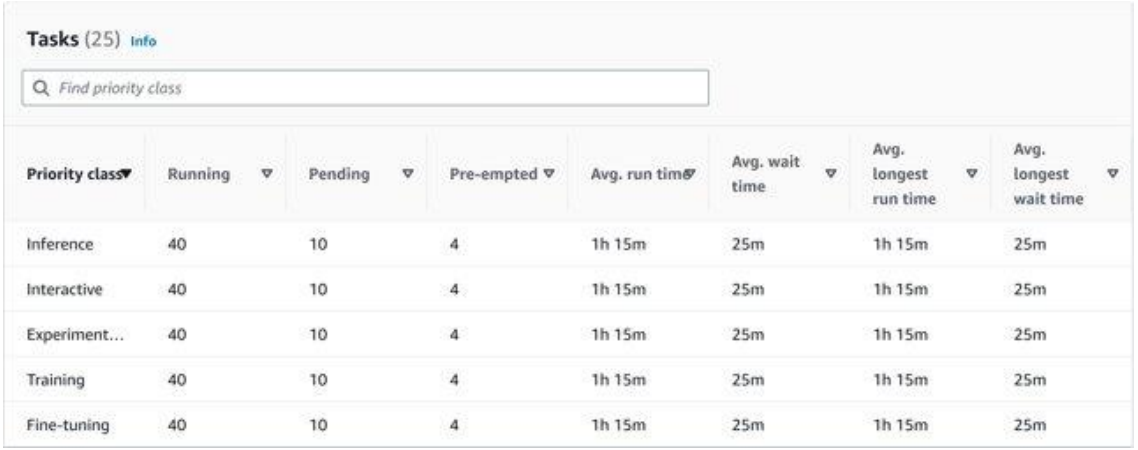
Prioritization

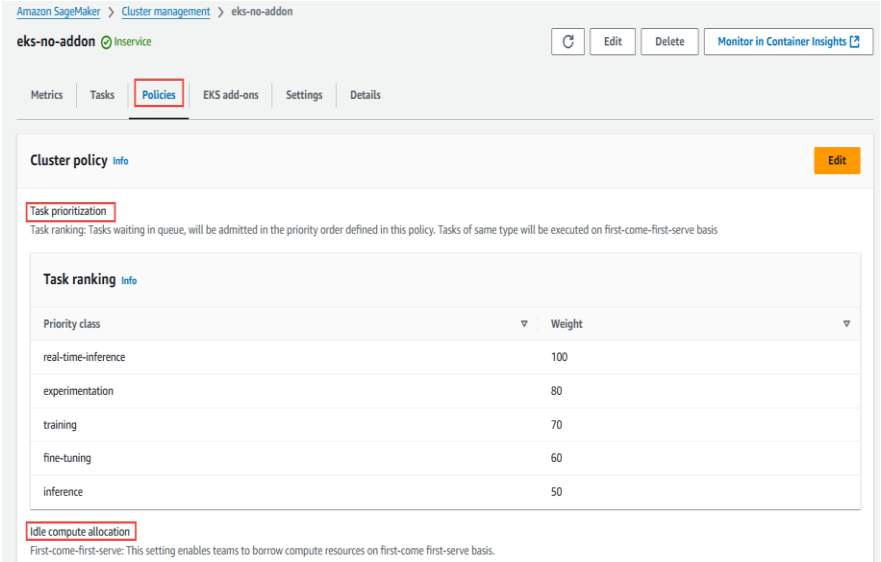
RESOLVED

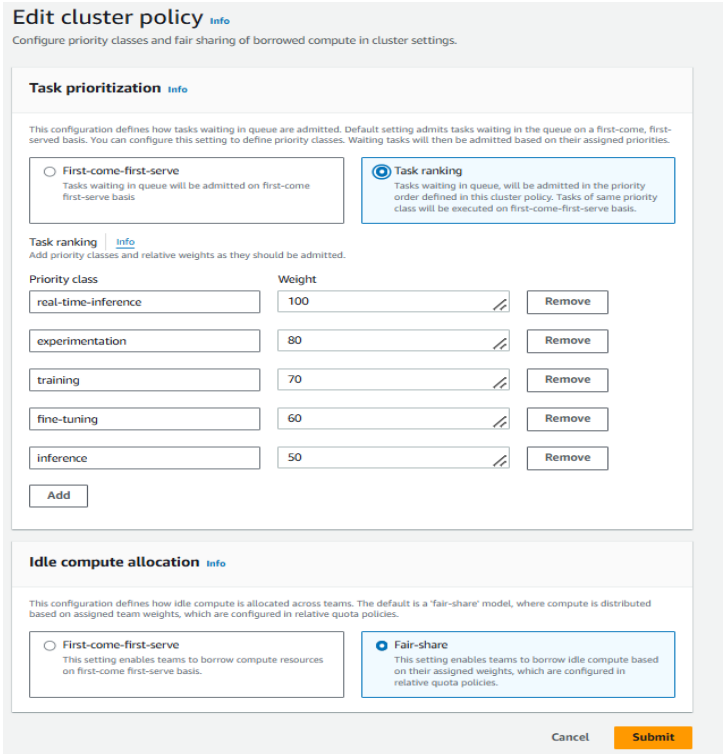
Leveraging the same set of resources for multiple tasks simultaneously requires an effective and efficient means to prioritize one task over another.

Amazon SageMaker HyperPod task governance

Cluster Policy & Prioritization







04

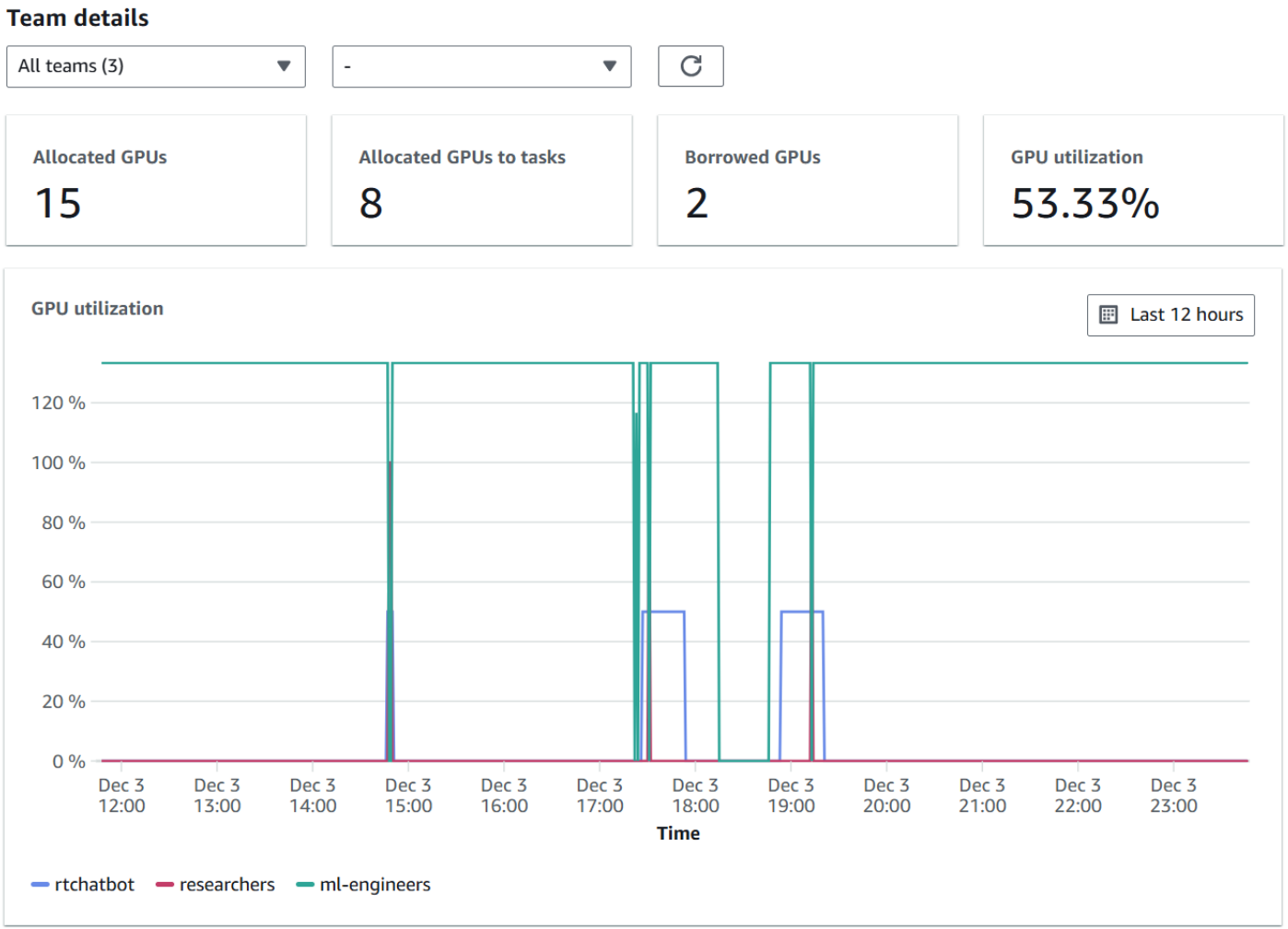
Traceability & Accountability

RESOLVED

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

Amazon SageMaker HyperPod task governance

Team Details



05

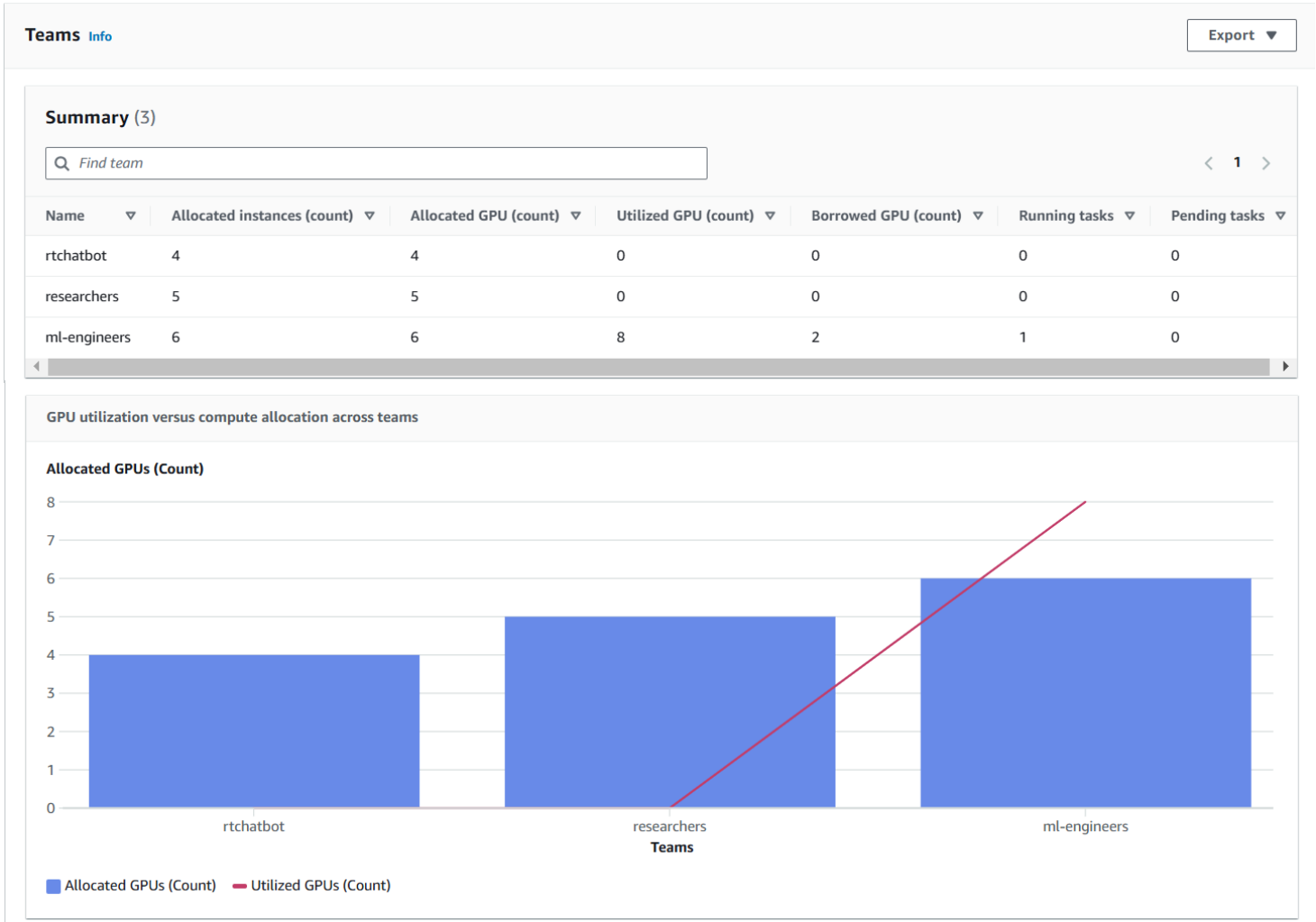
Cost Management & Optimization

RESOLVED

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

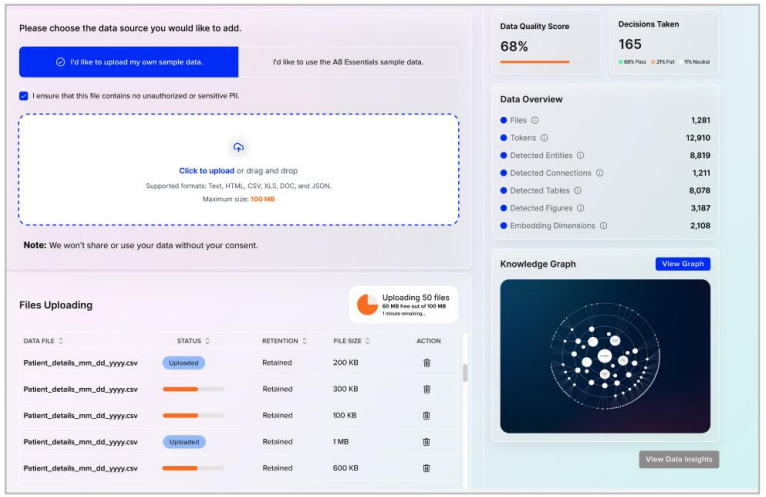
A8 Enterprise

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

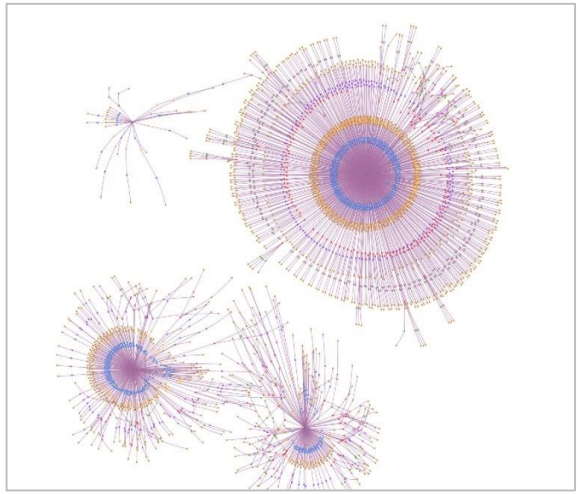
A8 Expert

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

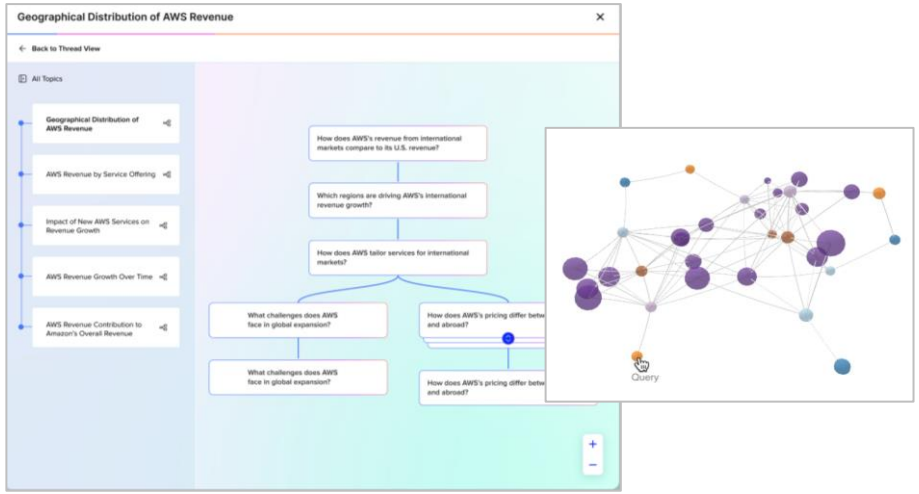
Ingest



Perceive



Act



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