KubeCon | CloudNativeCon

North America 2024

Engaging the KServe Community

The Impact of Integrating a Solutions with Standardized CNCF Projects

Introductions

Speakers

Andreea Munteanu | Canonical | AI Product Manager

Tessa Pham | Bloomberg | Senior Software Engineer

Johnu George | Nutanix | Technical Director

Adam Tetelman | NVIDIA | Principal Product Architect

Taneem Ibrahim | Red Hat | Senior Engineering Manager

Current State of OSS: AI, Inference, Engagement, Recent Successes

Future AI Project Roadmap: Challenges, Insights, Community Support

WG-Serve & KServe Talks

Engaging the KServe Community, The Impact of Integrating a Solutions with Standardized CNCF Projects | **Thursday** 5:25pm - 6:00pm

Unlocking Potential of Large Models in Production | **Thursday** 2:30pm - 3:05pm

WG Serving: Accelerating AI/ML Inference Workloads on Kubernetes | **Friday** 11:55am - 12:30pm

Optimizing Load Balancing and Autoscaling for Large Language Model (LLM) Inference on Kubernetes | **Wednesday** 3:25pm - 4:00pm

Best Practices for Deploying LLM Inference, RAG and Fine Tuning Pipelines on K8s | **Friday** 4:00pm - 4:35pm