TRAINING 2

Machine Learning Model Deployment and Serving with Ray Serve

RAY SUMMIT 2022

Ray Serve is a framework-agnostic and Python-first machine learning model serving library built on Ray. This training will cover how Ray Serve makes it easy to deploy, operate, and scale a machine learning model using Ray Serve APIs.

Key takeaways:

- Understand Ray Serve architecture, components, and flow of requests across replicas
- Learn how to use Ray Serve APIs to create, access, and deploy your models and mechanisms to access model deployments via Python APIs and HTTPs endpoints
- Implement common model deployment patterns for serving ML models using the inference graph API as a directed acyclic graph (DAG)
- Scale up/down individual components of an inference graph node, utilizing appropriate hardware resources (GPUs/CPUs) and replicas
- Use operational-friendly APIs to integrate with your custom CI/CD
- Inspect load and deployments in a Ray dashboard

Level:

Beginners or intermediate ML/DS/MLOps practitioners