tech | ICCON
infra cloud conference

Agenda:

2:00 – 4:30 PT

Welcome to ICCON – Target's Infra Cloud Conference

Scalable high-throughput data cache for machine learning training

Vinayak Kamath, Lead Data Engineer, Target

High-performance data caches are essential for efficiently loading large datasets to GPUs. It increases GPU utilization when training machine learning models and reduces idle times. We'll discuss our extensive analysis of high-performance distributed file systems and provide insights into how we built an accelerated data caching system to improve the speed and efficiency of ML training. Our exploration is helpful to anyone seeking to optimize computational performance in the dynamic realm of machine learning.

Live Q&A with Vinayak

Short Break

Infra Dev – ML CI/CD at Scale

Damon Allison, Principal Machine Learning Engineer, Shipt

Deploying ML solutions to production quickly and safely requires CI/CD tooling and processes specifically geared around ML's primary artifact: the model. In this talk, Damon will cover how Shipt has implemented tooling and processes for ML engineers and data scientists to build, deploy, monitor, and audit machine learning models in production systems at scale. We'll discuss how models are tagged, versioned, deployed, scaled, and monitored with tooling like mlflow, seldon core, airflow, and drone. We'll also cover advanced deployment scenarios like A/B deployment, shadow versions, as well patterns for integrating models into both batch and real time production systems.

Live Q&A with Damon

Short Break

Scaling training and deployment of LLMs for retail applications

Aastha Jhunjhunwala, Solution Architect, NVIDIA

Large Language Models (LLMs) are revolutionizing the retail industry by leveraging advanced natural language processing to enhance customer experience, accelerate time to business insights, generate digital assets, and much more. However, with great power comes great challenges. Join us as we unfold the intricacies and challenges encountered in the LLM lifecycle and how to overcome these and scale efficiently while maximizing GPU utilization. We'll touch upon considerations such as when to fine-tune vs. pretrain, 3D parallelism techniques for efficient training, information retrieval, and techniques for optimized inference such as in-flight batching. This talk will be of value to anyone seeking to plan, build, and optimize their LLM applications.

Live Q&A with Aastha

4:30 – 6:00 PT

In-Person: Happy Hour

Stick around and meet the presenters and other attendees