

Cloudera AI Inference: From Model Deployment to Intelligent Applications

OEM: Cloudera • Duration: 1 Day (8 hrs) • Code: DSCI-276

COURSE MODULES & TOPICS

Product Overview & Architecture

- Overview of Cloudera AI Inference
- Control plane and runtime architecture
- Integration within the Cloudera AI ecosystem
- CPU and GPU deployment models
- Core enterprise capabilities
- Overview of AI Registry
- Cloudera Model Hub
- Cloudera AI Registry — Register & Govern Models
- Theory: Model Size & Precision
- Model Size & Precision / Quantization
- Latency vs. Throughput — The Core Trade-off
- Theory: I/O Ratio & Concurrency
- I/O Ratio & Concurrency — Sizing Your Workload
- NVIDIA GPU Reference Guide for AI Inference
- Theory: GPU Selection for AI Inference
- GPU Selection for Cost Optimisation
- NVIDIA NIMs & Benchmarking — Complexity to Confidence
- Hands-On: Explore the AI Inference interface, import an LLM into AI Registry, deploy a Model Endpoint via the CAI UI, authenticate & call the endpoint, import model via Hugging Face

Cloudera Copilot & AI-Assisted Development

- Overview of Cloudera Copilot
- Integration with JupyterLab editor
- Foundation on JupyterAI extension
- Connecting Copilot to models deployed in Cloudera AI Inference
- Hands-On: Enable Copilot within AI Workbench, connect to an AI Inference endpoint, execute AI-assisted code generation, alias management

SQL AI Assistant in Cloudera Data Warehouse

- Introduction to SQL AI Assistance in Cloudera Data Warehouse and its key capabilities

- Natural language to SQL generation with support for Hive
- Foundation on JupyterAI extension
- Connecting Copilot to models deployed in Cloudera AI Inference
- Hands-On: Enable AI Assist in Hue, use GENERATE/FIX/EXPLAIN, multi-turn conversational interface

Integrating Cloudera AI Inference with Agent Studio for Multi-Agent Workflows

- Overview of AI Agents in enterprise automation and decision-making
- Cloudera Agent Studio for building and managing multi-agent workflows
- Integration with Cloudera AI Inference for real-time agent intelligence
- Architecture of agent systems using LLMs, tools, and orchestration
- Hands-On: Build a Database Agent in Agent Studio, connect to CDW using Hive tools, execute an end-to-end workflow from user query to response