CLOUDERA

TRAINING
DATASHEET

DSCI-276

Cloudera AI Inference: From Model Deployment to Intelligent Applications

Course Overview

Course Type
Instructor-led training

Level
Intermediate

Duration
1 day

Platform
Cloudera AI

Topics Covered

- Introduction to Enterprise AI Inference
- Cloudera AI Inference Architecture and Core Components
- AI Registry and Model Lifecycle Management
- Model Registration and Scalable Endpoint Deployment
- Performance Optimization: Model Size, Quantization, Latency, Throughput, and GPU Selection
- Cloudera Copilot Integration with AI Inference
- AI-Assisted Development in JupyterLab
- SQL AI Assistant in Cloudera Data Warehouse
- Integrating Cloudera AI Inference with Agent Studio for Multi-Agent Workflows

About This Training

This 1-day instructor-led course provides a hands-on introduction to Cloudera AI Inference for deploying and managing LLMs models using AI Registry and scalable endpoints.

Participants will learn how AI Inference powers real-world use cases such as Cloudera Copilot, SQL AI Assistant, and multi-agent workflows with Agent Studio.

In addition, learners will explore key concepts such as model size and quantization, latency vs. throughput trade-offs, GPU selection, and workload sizing, enabling them to design cost-efficient and high-performance AI systems.

The training focuses on deploying and integrating scalable, secure, enterprise-ready AI capabilities using Cloudera.

What Skills You Will Gain

Participants will develop the following skills:

- Understand the architecture and core components of Cloudera AI Inference and AI Registry
- Register, manage, and deploy ML models and LLMs as scalable inference endpoints
- Configure CPU/GPU resources, and optimize performance using quantization and workload sizing
- Understand key concepts such as latency, throughput, concurrency, and I/O ratios
- Invoke and integrate inference endpoints using APIs and enterprise workflows
- Enable AI-assisted development using Cloudera Copilot in JupyterLab
- Use SQL AI Assistant to generate, optimize, and debug queries using natural language
- Build and integrate AI-powered agents using Agent Studio for end-to-end workflows

Who Should Take This Course?

This course is ideal for AI/ML Engineers, Data Scientists, MLOps Engineers, Platform Engineers, Data Engineers, DevOps Engineers supporting AI workloads, Solution Architects, and Technical Innovation Teams responsible for deploying and operating AI models in enterprise environments.

DSCI-276

Cloudera AI Inference: From Model Deployment to Intelligent Applications

Use Case: AI-Powered Retail Sales Intelligence

Overview

This training simulates a retail business scenario where organizations leverage AI to analyze sales data, understand customer behavior, and enable faster, data-driven decision-making.

The dataset represents a simplified retail environment, including customer profiles and sales transactions across products, categories, and regions.

Business Problem

Retail organizations generate large volumes of transactional data but face key challenges:

- Slow insight generation due to manual analysis
- High dependency on technical teams for querying data
- Limited adoption of AI for predictions and automation

These challenges lead to delayed decisions and missed revenue opportunities.

Objective

This use case demonstrates how AI can be applied across the data lifecycle to:

- Analyze structured sales and customer data
- Enable natural language-based data access (Text-to-SQL)
- Accelerate data science workflows using AI-assisted development
- Deploy pre-trained models for real-time inference
- Enable conversational analytics through AI agents

Mapping to Cloudera AI Capabilities

- SQL AI Assistant (Hue)
  Enables business users to query retail data using natural language without writing SQL
- Cloudera Copilot (Jupyter)
  Enhances data science productivity through AI-assisted code generation and analysis
- Cloudera AI Inference
  Demonstrates how pre-trained models from Model Hub (e.g., NVIDIA, Hugging Face) can be imported and deployed for real-time inference
- Cloudera Agent Studio (Talk to Your Data)
  Enables conversational interaction with enterprise data for business insights

Business Value: Faster time to insight and self-service analytics for business users, reducing dependency on specialized technical skills.Enables rapid AI adoption using pre-trained models and improves decision-making through real-time and conversational intelligence.

VERSION:
2.0.0

DSCI-276

Cloudera AI Inference: From Model Deployment to Intelligent Applications

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 Exercise

- Explore the AI Inference interface
- Import an LLM into AI Registry
- Deploy a Model Endpoint via the CAII 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 Exercise

- Enable Copilot within AI Workbench
- Connect Copilot to an AI Inference endpoint
- Execute AI-assisted code generation
- Loading the extension, listing models, registering aliases
- Code generation with --format code for Python, PySpark
- Output formats: markdown, html, math
- Debugging SQL and Python errors with %ai error
- Explaining complex code in plain language
- Multi-turn iterative development with conversation context
- Chat panel commands: /ask, /fix, /generate
- Real-world challenges: ETL, ML templates, documentation
- Alias management: register, use, update, delete

VERSION:
2.0.0

DSCI-276

Cloudera AI Inference: From Model Deployment to Intelligent Applications

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 Exercise

- Enable AI Assist in Hue
- Use GENERATE to create SQL queries from business questions
- AI-powered query optimization and performance improvement techniques
- Use FIX to debug and correct SQL errors
- Use EXPLAIN to understand complex SQL queries in plain language
- Multi-turn conversational interface using Assistant actions (Generate, Edit, Optimize, Fix, Comment)

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 Exercise

- Build and explore a Database Agent in Agent Studio
- Connect the agent to Cloudera Data Warehouse (CDW) using Hive tools
- Understand how the SQL Execution Agent interacts with database tools
- Ask business questions in plain English using the agent interface
- Automatically generate and execute Hive SQL queries on CDW
- Observe multi-agent collaboration (SQL + Analysis Agent) for enhanced insights
- Execute an end-to-end workflow from user query → SQL generation → data retrieval → response
- Build confidence in using conversational AI for real-world data exploration

VERSION:
2.0.0