CLOUDERA

Bridging Data & AI: How Cloudera and NVIDIA Drive the Future of Intelligent Enterprises

AI Inference Day with NVIDIA

Speakers : Manick Mehra
Navin Agrawal
Anukrati Saxena

April, 2025

Agenda

What are we going to cover as part of today's presentation ?

1 All About the New Cloudera

2 AI with Cloudera and NVIDIA

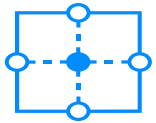
3 Technical Deep dive

4 **Demo** - Remote Code Generation

CLOUDERA

Adoption of AI is a Journey

Identifying AI challenges in an organization

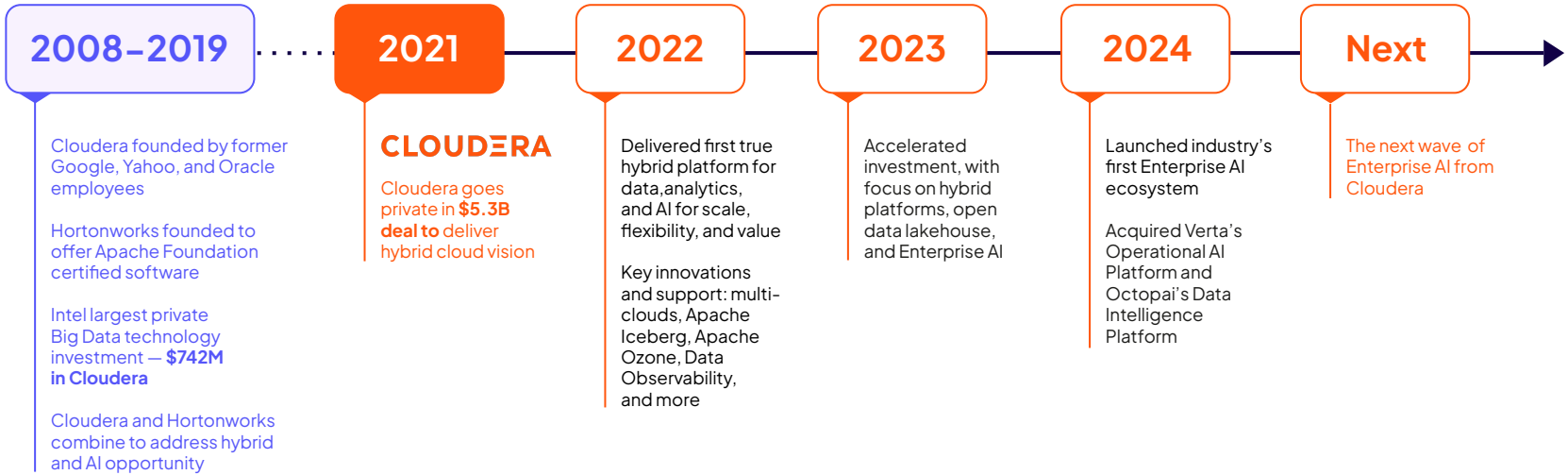
Challenges		What's missing
Data integration barriers		• Streamlined access to enterprise data
Rigid model infrastructure		• Modularity • Flexibility • AI Ops
Lack of security and transparency		• Model control • Built-in security • Visibility & governance

ALL ABOUT THE NEW

CLOUDERA

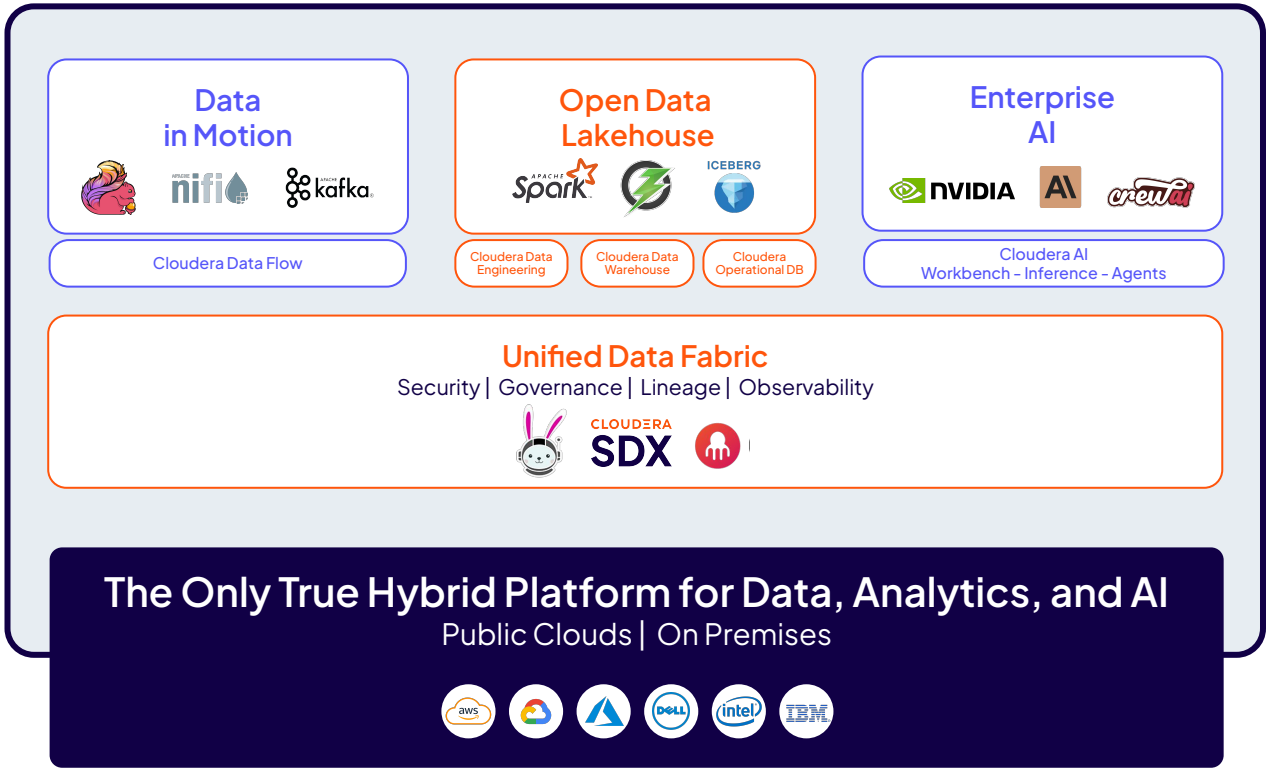
CLOUDERA

Cloudera's Data & AI Evolution



CLOUDERA

Cloudera Key Innovations in Our True Hybrid Platform



CLOUDERA

AI WITH
CLOUDERA & NVIDIA

CLOUDERA

Build, Run, and Infuse AI

Democratizing AI for all enterprises to leverage all their data

Build AI with Cloudera

Create powerful AI applications

Run AI in Cloudera

Run performant AI applications

Infuse AI into Cloudera

Empower with AI Assistants

CLOUDERA

Cloudera AI Ecosystem

AI Applications & Agents

AI Workbench | **AI Inference**

AI Registry

AI Ecosystem

FRAMEWORKS

crewai

mlflow

LangChain

MODELS

Gemini

ANTHROP\C

cohere

Amazon Bedrock

MISTRAL AI_

LLaMA by Meta

VECTOR DATABASES

Pinecone

weaviate

milvus

Solr

Real-time Open Data Lakehouse

Cloud Infrastructure

Google Cloud

Azure

aws

Red Hat

Hardware

NVIDIA

AMD

Google

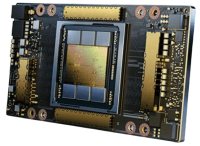
intel

DELL Technologies

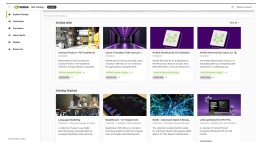
IBM

NVIDIA Capabilities Leveraged by Cloudera

How Cloudera leverages NVIDIA hardware for accelerations ?



NVIDIA GPUs (A100, H100, etc.)
Powering training and inference workloads inside Cloudera's AI/ML pipelines.



NVIDIA NGC (NVIDIA GPU Cloud):
Cloudera can leverage pre-trained models, containers, and optimized AI frameworks from NGC to accelerate development.

NVIDIA Triton Inference Server:
Integrated into the Cloudera ecosystem for model serving at scale

RAPIDS
Cloudera and NVIDIA have collaborated around using RAPIDS to accelerate Apache Spark on GPUs for faster data preparation and feature engineering.

NIM
NIM offers prepackaged, fully optimized containers that include large language models (LLMs) and other AI models.
These containers are designed for plug-and-play usage

CLOUDERA

Build, Run, and Infuse AI

Democratizing AI for all enterprises to leverage all their data

Build AI with Cloudera

Create powerful AI applications

Run AI in Cloudera

Run performant AI applications

Infuse AI into Cloudera

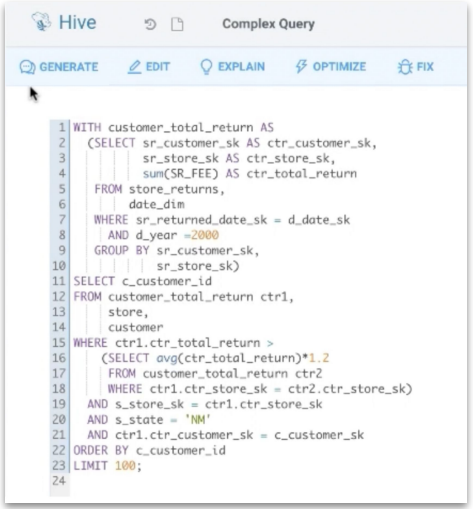
Empower with AI Assistants

CLOUDERA

Infuse AI in Cloudera - Cloudera AI Assistants

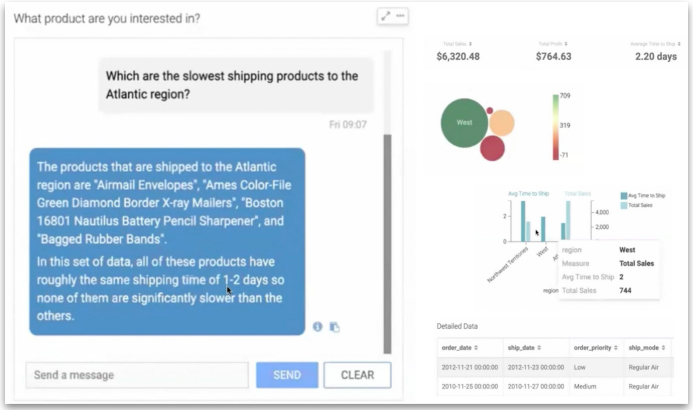
SQL, BI & ML coding AI assistants built-in Cloudera services

SQL AI Assistant



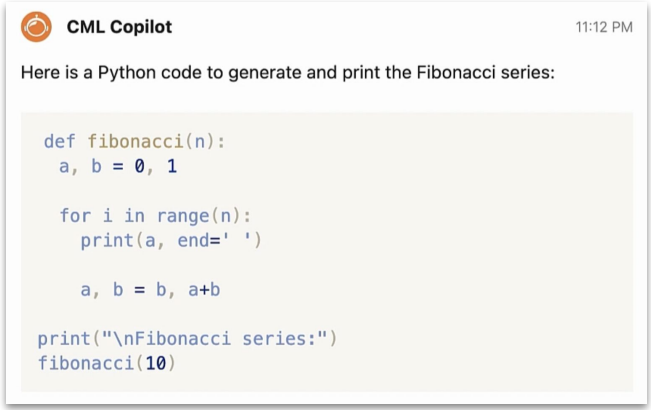
Cloudera Data Warehouse

BI Chatbot



Cloudera Data Visualization

ML Copilot



Cloudera AI

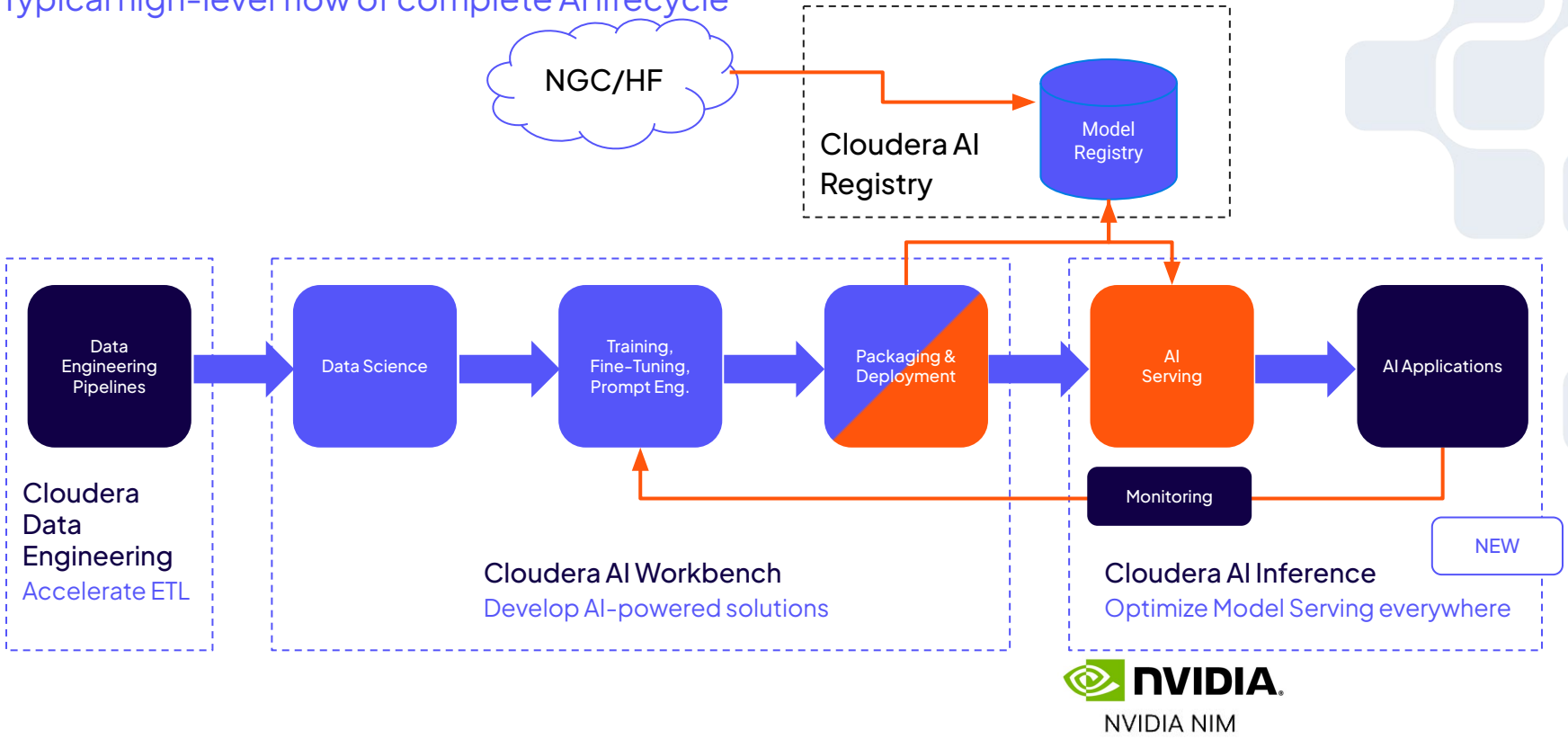
CLOUDERA

Technical Deep Dive

CLOUDERA

MLOPS Lifecycle

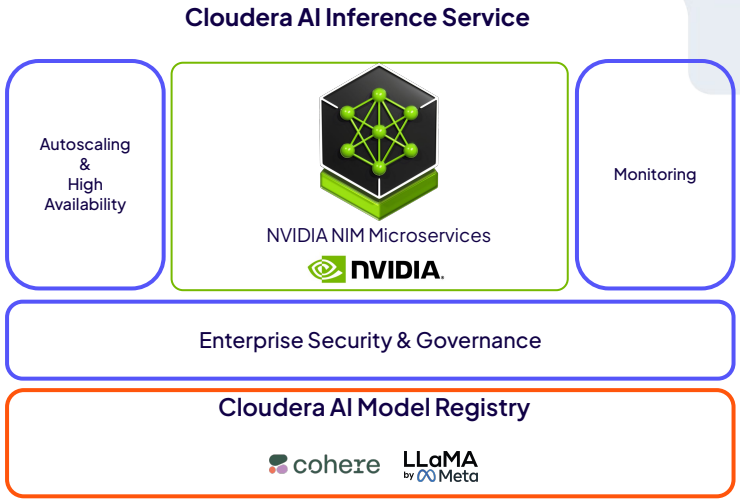
Typical high-level flow of complete AI lifecycle



CLOUDERA

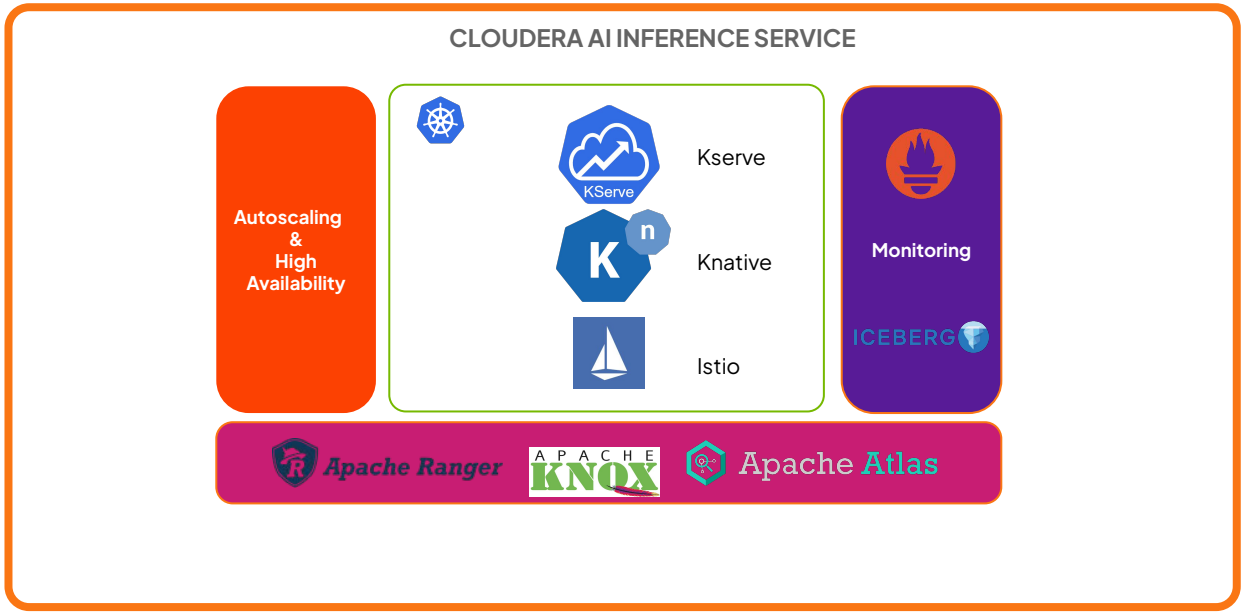
Features of Robust Inference platform

- Support for predictive and generative AI models.
- Use standard inference API standards
  - OpenAI
  - Open Inference Protocol
- Authentication, authorization and auditing.
- HA, fault tolerant, Zero downtime upgrade.
- Autoscaling, Scale to Zero.
- Monitoring for performance and accuracy.
- Run Anywhere (multi cloud, on-premise)
- Private deployment



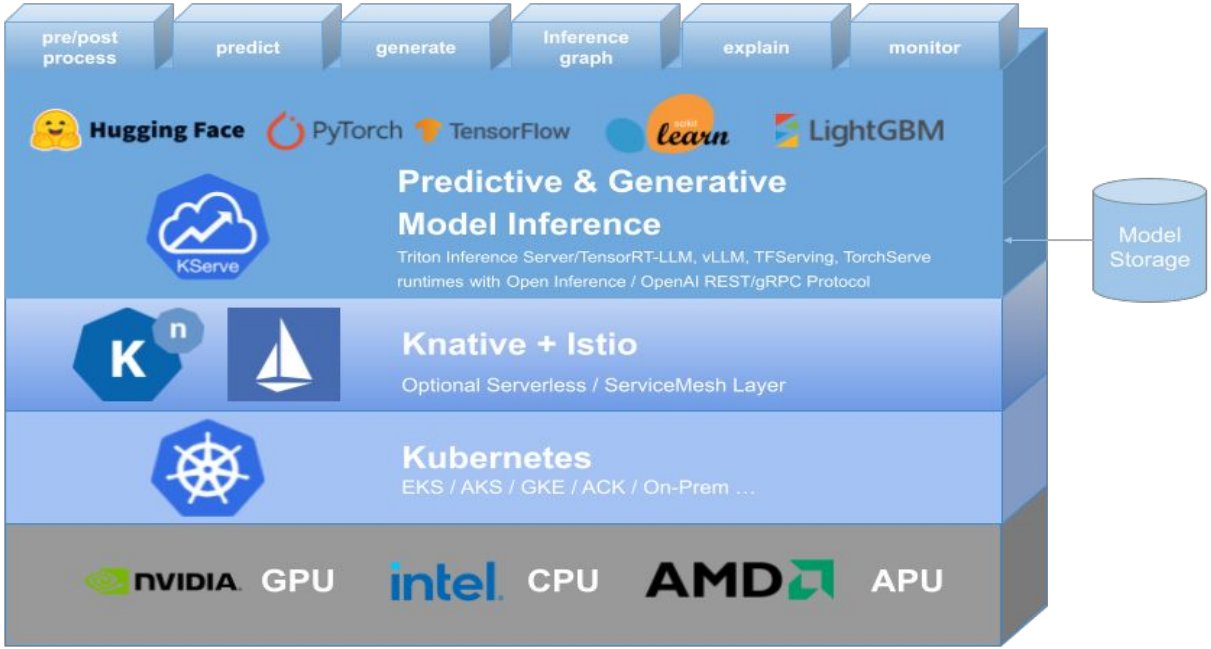
CLOUDERA

Cloudera AI Inference Architecture



CLOUDERA

Under the hood



source: https://github.com/kserve/kserve

CLOUDERA

Supported runtimes

Llama 3.x NIM

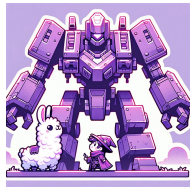
NVIDIA

Mistral NIM

NVIDIA

Mixtral NIM

NVIDIA



vLLM

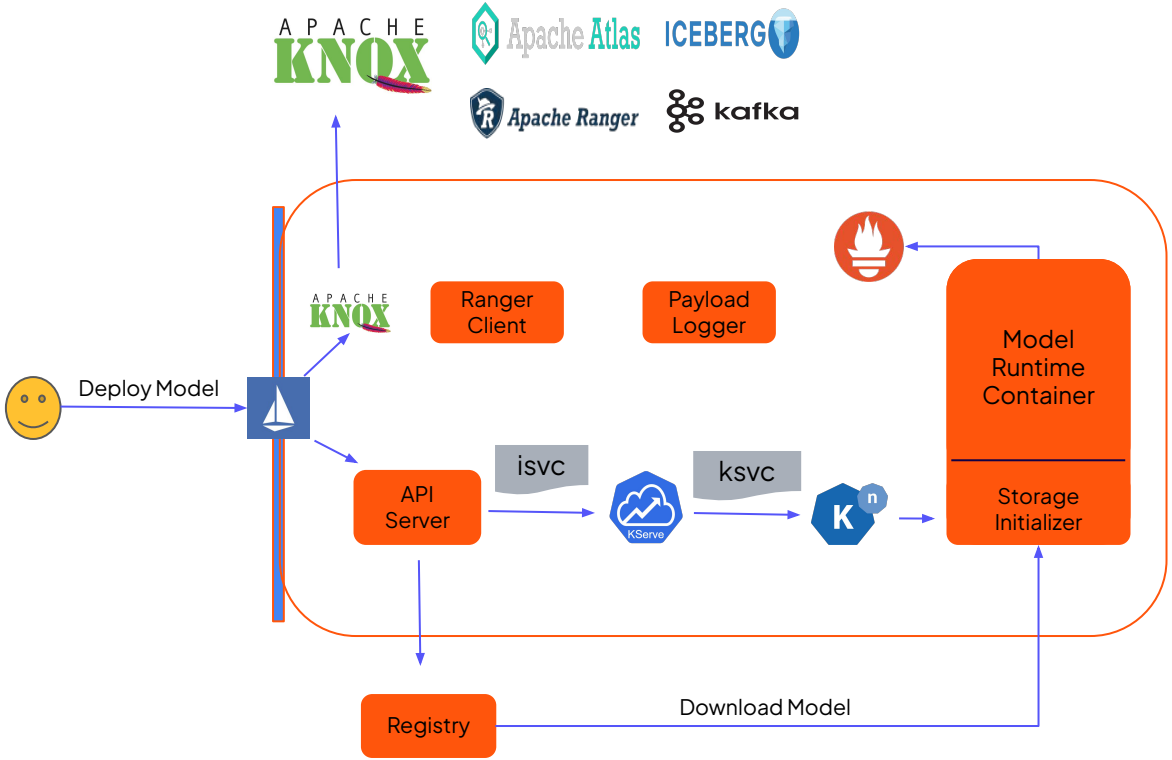
NVIDIA
TRITON INFERENCE SERVER

ONNX

One runtime per NIM

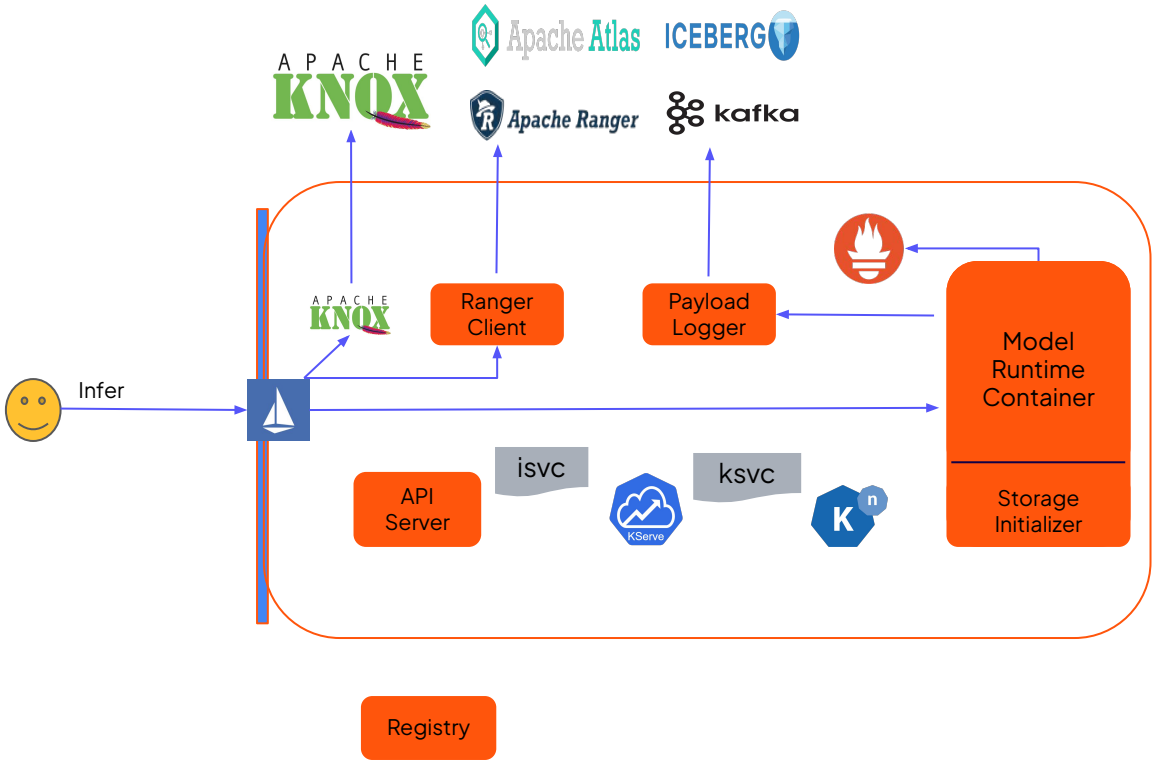
CLOUDERA

Model deployment flow



CLOUDERA

Inference flow



CLOUDERA

DEMO

CLOUDERA

Demo Objectives

An overview of the Cloudera AI Inference Platform and a simple use case.

- Cloudera AI Product Overview
  - Model Hub
  - AI Registry
  - AI Inference
- Use Case
  - Remote code generation via model deployed on AI Inference

CLOUDERA

Key Takeaways

- High Performance - Nvidia NIM & Triton runtime + Nvidia GPUs
- Scalability, HA - KServe & Knative
- Security - Apache Knox, Ranger, Atlas
- Monitoring - Prometheus & Iceberg
- Run Anywhere - K8s
- Private deployment

Note : While applying for a trial use an official email ID as requests from gmail/yahoo are usually rejected

Take the Next Step

Try Cloudera for FREE

Cloudera on Cloud

Cloudera On Premise

CLOUDERA

Thank You

CLOUDERA