



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

AI Inference Day with NVIDIA

April, 2025

Speakers : Manick Mehra
Navin Agrawal
Anukrati Saxena

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

Adoption of AI is a Journey

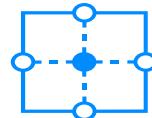
Identifying AI challenges in an organization

Challenges

Data integration
barriers



Rigid model
infrastructure



Lack of security
and
transparency

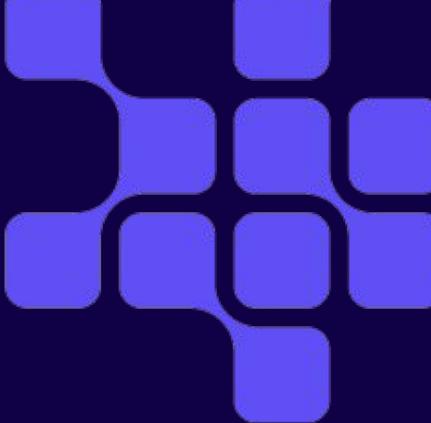


What's missing

- Streamlined access to enterprise data

- Modularity
- Flexibility
- AI Ops

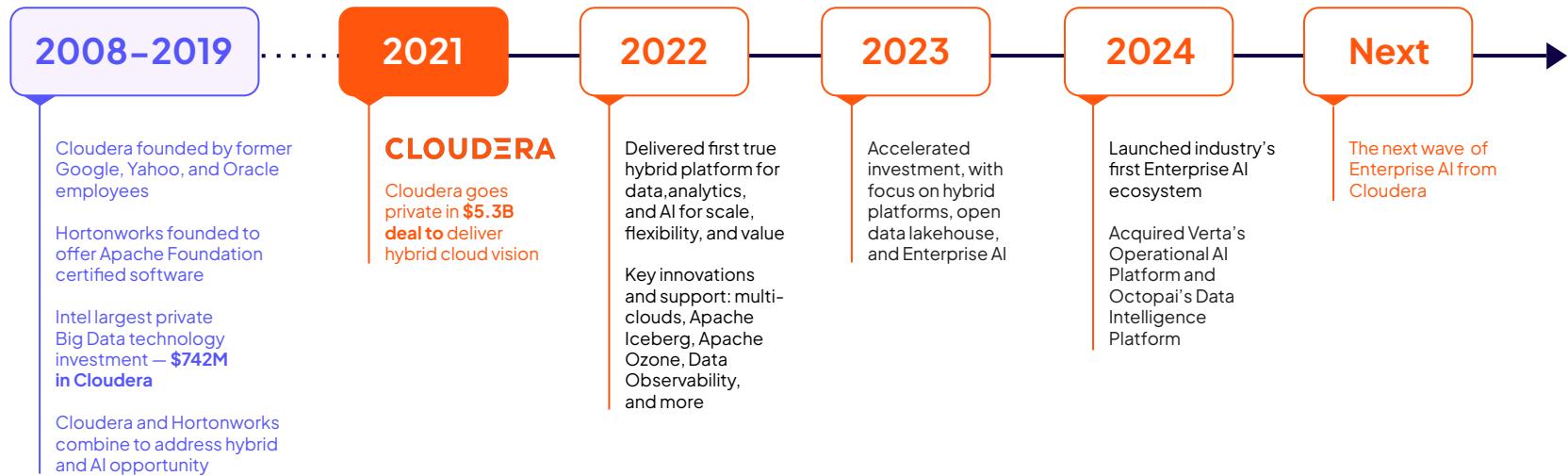
- Model control
- Built-in security
- Visibility & governance



ALL ABOUT THE NEW CLOUDERA

CLOUDERA

Cloudera's Data & AI Evolution



Cloudera Key Innovations in Our True Hybrid Platform

Data in Motion

Apache nifi Apache kafka

Cloudera Data Flow

Open Data Lakehouse

APACHE Spark ICEBERG

Cloudera Data Engineering Cloudera Data Warehouse Cloudera Operational DB

Enterprise AI

NVIDIA AI crew.ai

Cloudera AI Workbench - Inference - Agents

Unified Data Fabric

Security | Governance | Lineage | Observability

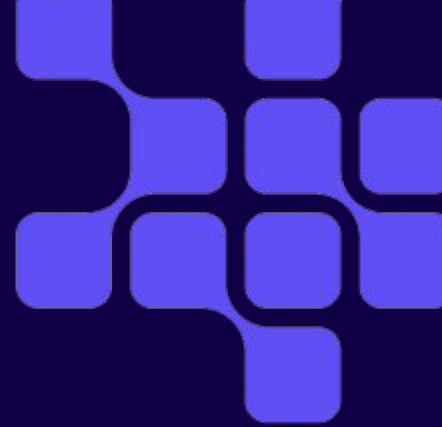
CLOUDERA
SDX

The Only True Hybrid Platform for Data, Analytics, and AI

Public Clouds | On Premises

aws Microsoft Azure Google Cloud DELL intel IBM

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 AI Ecosystem

AI Applications & Agents

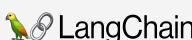
AI Workbench

AI Inference

AI Registry

AI Ecosystem

FRAMEWORKS



MODELS



VECTOR DATABASES



Real-time Open Data Lakehouse

Cloud Infrastructure

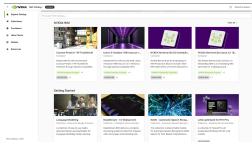
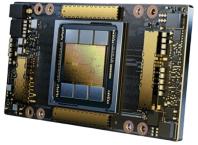


Hardware



NVIDIA Capabilities Leveraged by Cloudera

How Cloudera leverages NVIDIA hardware for accelerations ?



RAPIDS



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

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



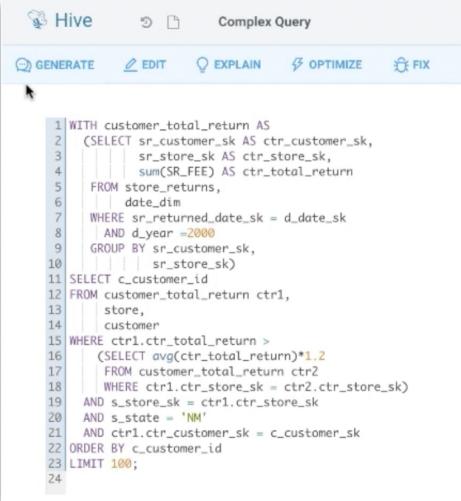
Infuse AI into Cloudera

Empower with AI Assistants

Infuse AI in Cloudera - Cloudera AI Assistants

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

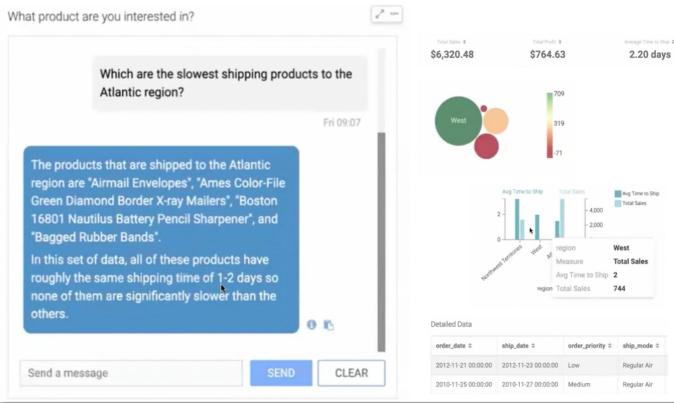
SQL AI Assistant



```
1 WITH customer_total_return AS
2   (SELECT sr_customer_sk AS ctr_customer_sk,
3    sr_store_sk AS ctr_store_sk,
4    sum(SR_FEE) AS ctr_total_return
5   FROM store_returns,
6    date_dim
7   WHERE sr_returned_date_sk = d_date_sk
8    AND d_year >=2000
9   GROUP BY sr_customer_sk,
10    sr_store_sk)
11 SELECT c_customer_id
12 FROM customer_total_return ctr1,
13  store,
14  customer
15 WHERE ctr1 ctr_total_return >
16   (SELECT avg(ctr_total_return)*1.2
17    FROM customer_total_return ctr2
18    WHERE ctr1 ctr_store_sk = ctr2 ctr_store_sk)
19  AND s_store_sk = ctr1 ctr_store_sk
20  AND s_state = 'NM'
21  AND ctr1 ctr_customer_sk = c_customer_sk
22 ORDER BY c_customer_id
23 LIMIT 100;
```

Cloudera Data Warehouse

BI Chatbot



What product are you interested in?

Which are the slowest shipping products to the Atlantic region?

Fri 09:07

The products that are shipped to the Atlantic region are "Airmail Envelopes", "Ames Color-File Green Diamond Border X-ray Mailers", "Boston 16801 Nautilus Battery Pencil Sharpener", and "Bagged Rubber Bands".

In this set of data, all of these products have roughly the same shipping time of 1-2 days so none of them are significantly slower than the others.

Send a message SEND CLEAR

Visualizations:

- Total Sales: \$6,320.48
- Total Profit: \$764.63
- Average Time to Ship: 2.20 days
- Heatmap: West (green), East (orange), South (red)
- Bar chart: Avg Time to Ship vs Region Measure
- Table: Detailed Data

Cloudera Data Visualization

ML Copilot

CML Copilot

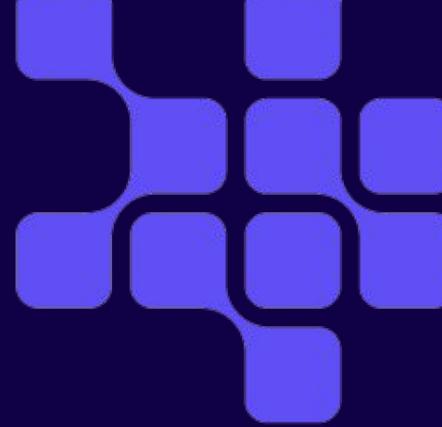
Here is a Python code to generate and print the Fibonacci series:

```
def fibonacci(n):
    a, b = 0, 1

    for i in range(n):
        print(a, end=' ')
        a, b = b, a+b

print("\nFibonacci series:")
fibonacci(10)
```

Cloudera AI

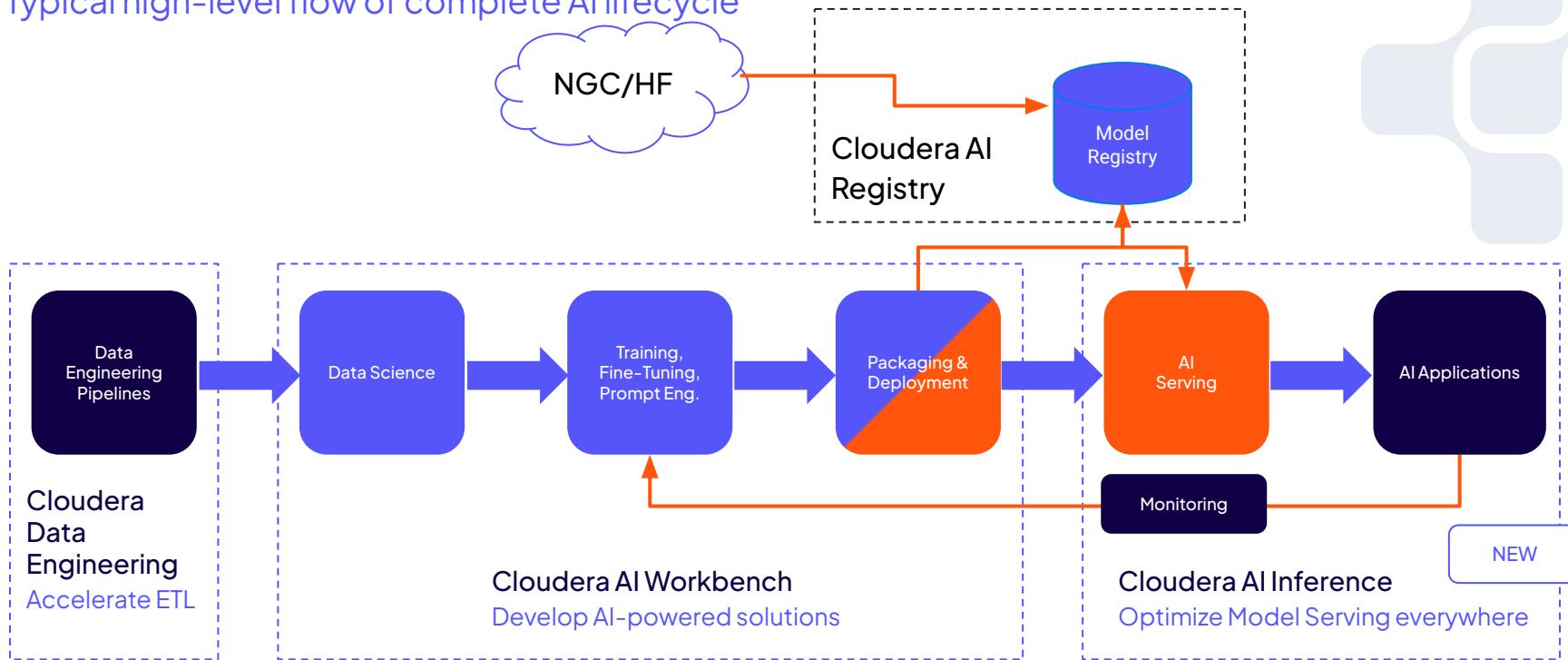


Technical Deep Dive

CLOUDERA

MLOPS Lifecycle

Typical high-level flow of complete AI lifecycle

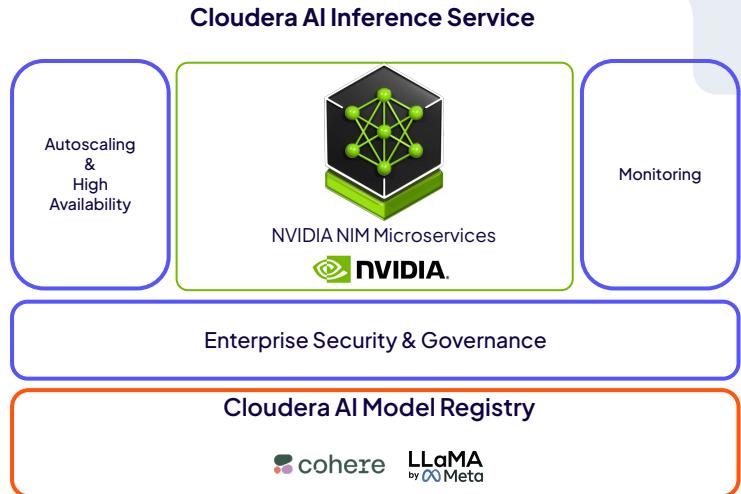


NVIDIA NIM

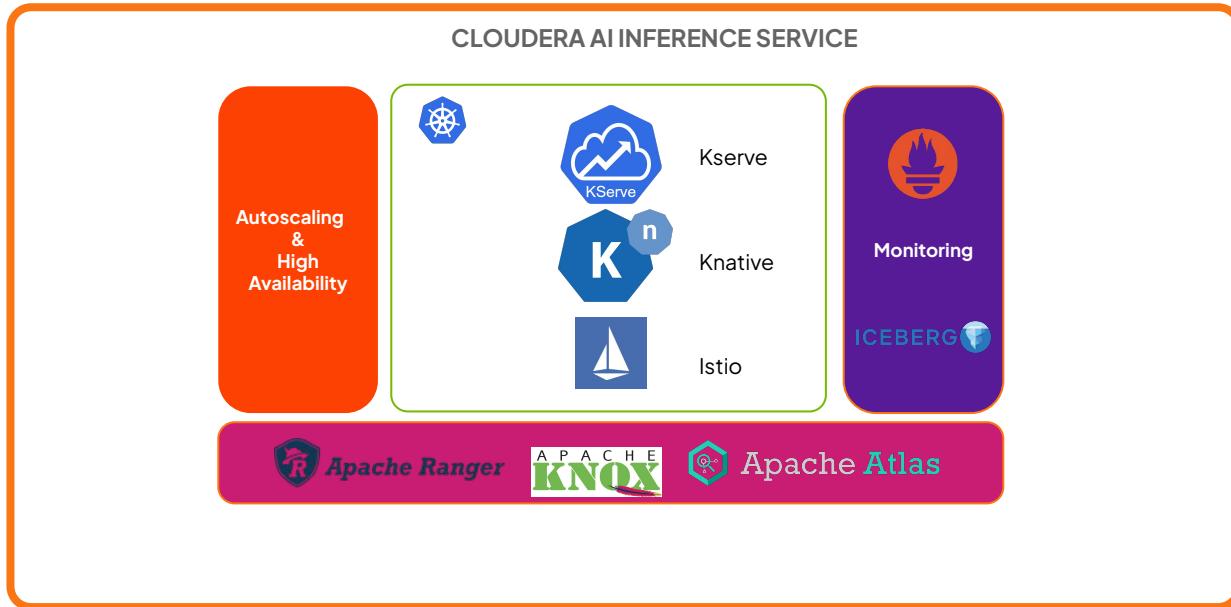
CLOUDERA

Features of Robust Inference platform

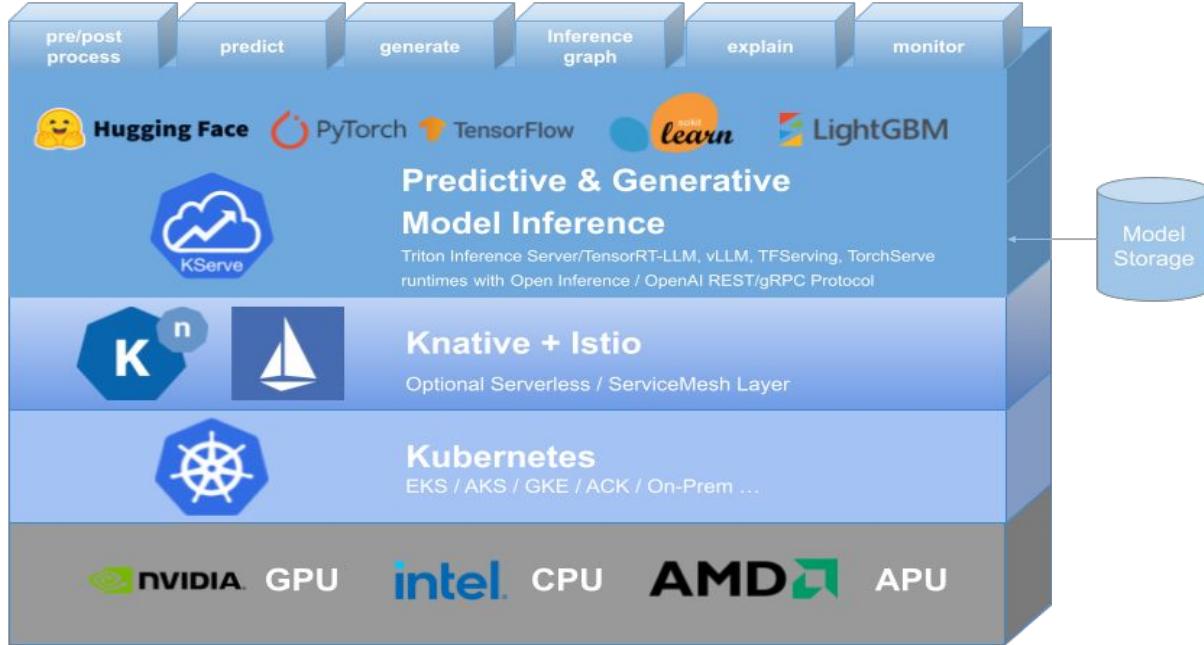
- Support for predictive and generative AI models.
- Use standard inference API standards
 - OpenAI
 - Open Inference Protocol
- Authentication, authorization and auditing.
- HA, fault tolerant, Zero downtime upgrade.
- Autoscaling, Scale to Zero.
- Monitoring for performance and accuracy.
- Run Anywhere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



Cloudera AI Inference Architecture

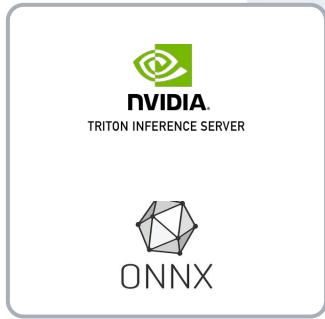


Under the hood



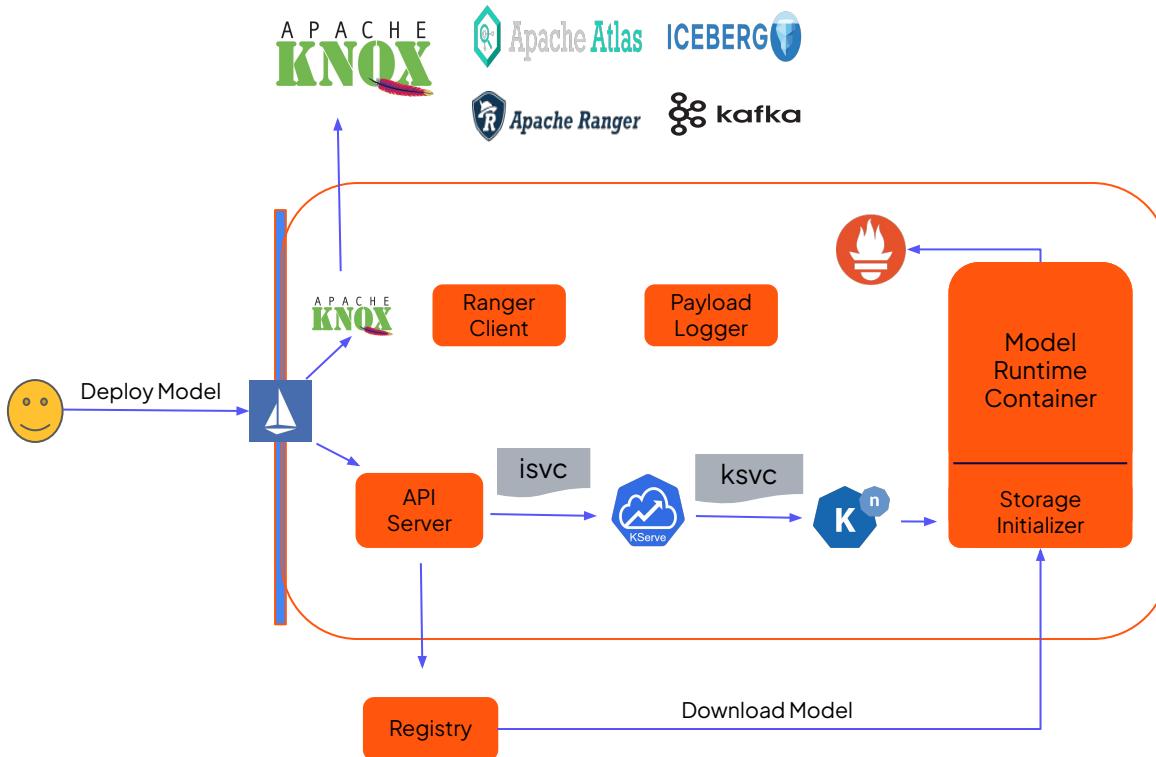
source: <https://github.com/kserve/kserve>

Supported runtimes

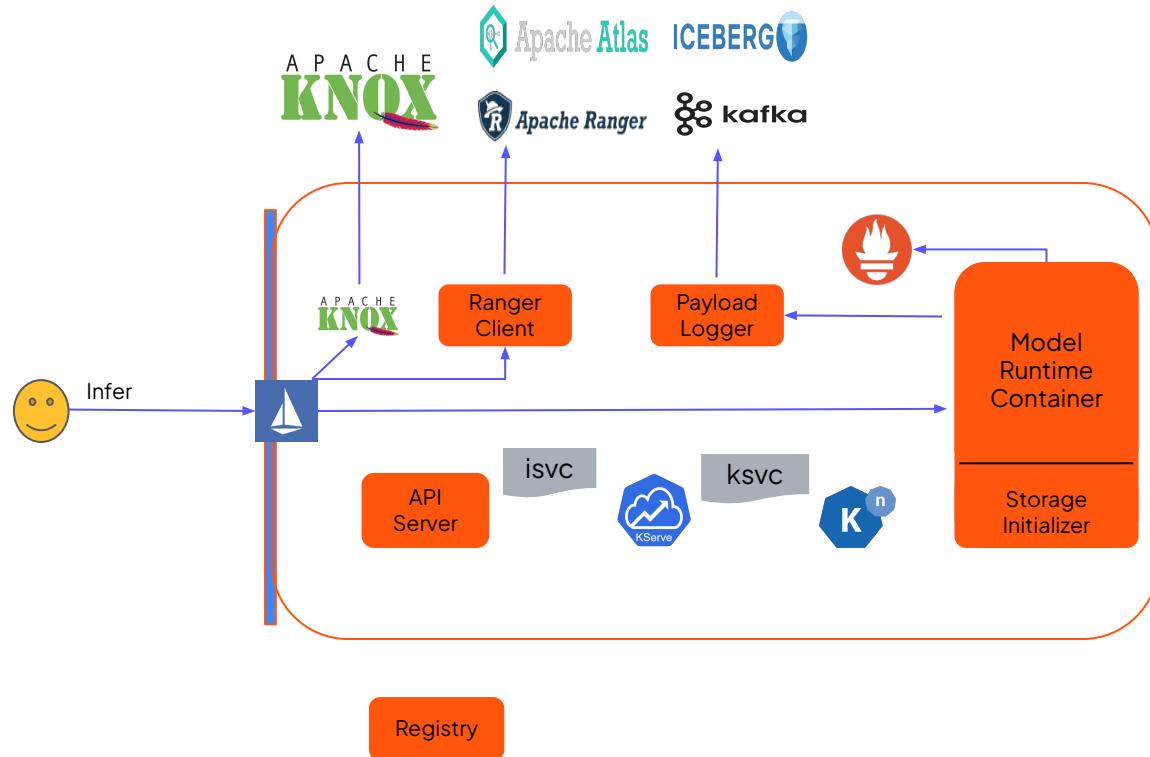


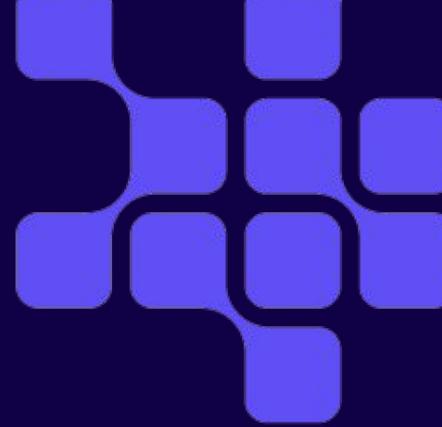
One runtime per NIM

Model deployment flow



Inference flow





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

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

Take the Next Step

Try Cloudera for FREE

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

Cloudera on Cloud



Cloudera On Premise





Thank You