



H2O GenAI WORLD

TRAINING • NEW YORK

**AI Agents, Combining Predictive and
Generative AI and Kernel Internal HAIC
Training**

21st November 2024



Table of Contents

- **AI Agents 101**
- **AI for Business Stakeholders**
- **Kernel Internal and HAIC Deployments**



H2O
GenAI WORLD
NEW YORK

AI Agents 101 & Hands-on Workshop with h2oGPTe



Agenda

- **Theory:** AI Agents, LLM Chains, and Prompt Engineering
- **Workshop:** Building AI Agents and Mastering Effective Prompt Engineering with h2oGPTe: A Hands-On Workshop with Real-World Data Science Applications

AI Agents, LLM Chains, and Prompt Engineering

What Are LLM Chains?

LLM Chains are pipelines or a sequence of actions linking LLMs with tools and functions, other LLMs, services, parsers together. Chains can be chained together in a sequential, recursive, or parallel manner and have different role.

For example an LLM might be chained with:

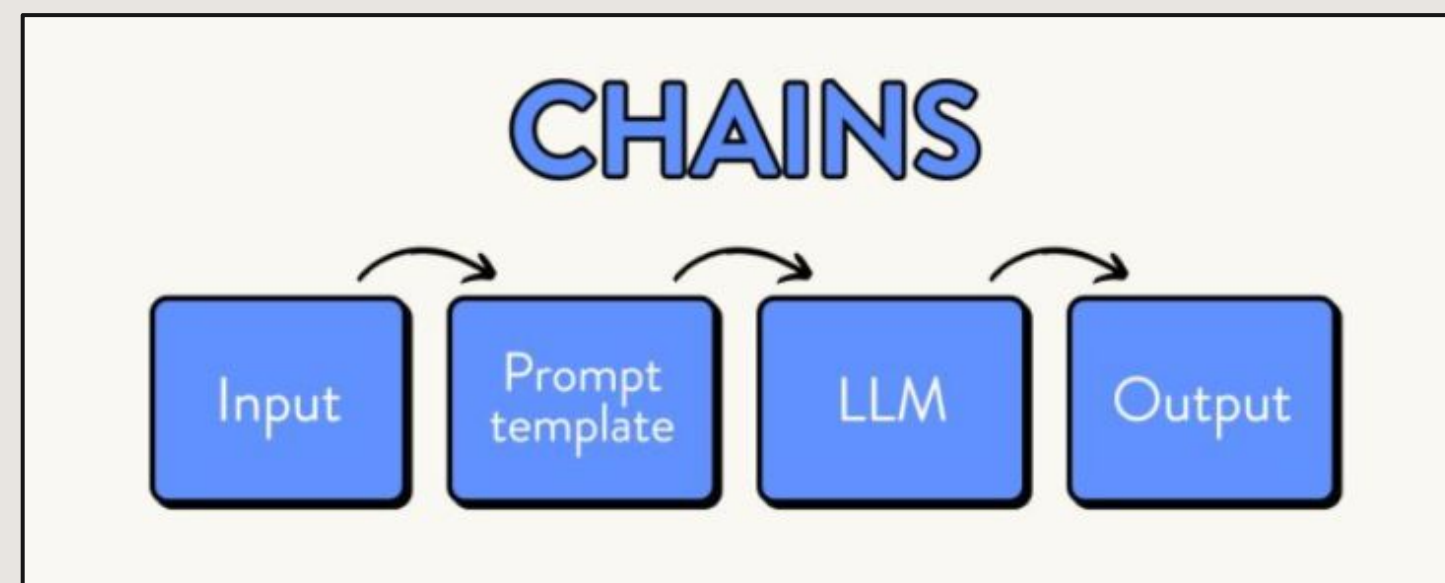
- an API to fetch real-time information on a specific stock
- Perform real time prediction on a deployed ML model
- Format the input to an LLM and/or the output of an LLM.
 - the simplest component of LLM Chain? the Prompt Template: it formats the user input and provides a consistent and standardized way to present the prompt to the LLM and generates a response accordingly.

Example of several Open Source LLM Chaining frameworks: ***AutoGen, LangChain, LLaIndex, Haystack***

What Are LLM Chains?

Large Language Model Chaining is the process of integrating one or multiple large language models with other applications, tools, and services to:

- leverage the strengths of other tools and services to create a more powerful and versatile AI system (such as personal assistants) that can provide accurate and useful outputs to a various range of inputs.
- to overcome LLM own limitations (using RAG for example to mitigate hallucination)
- format the input to an LLM and/or the output of an LLM.

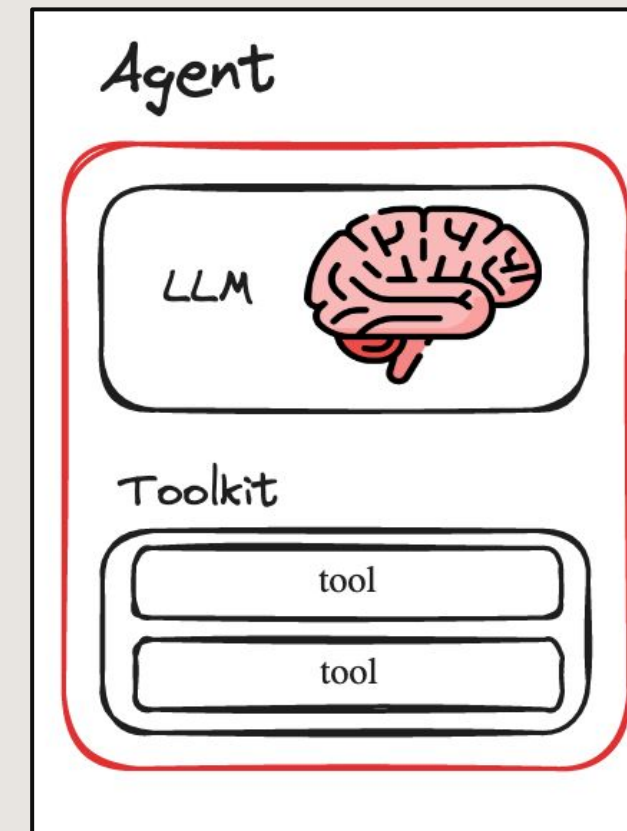


What Are Tools?

Tools or Functions or "Functions calling" are specialized functionalities that can accomplish a specific task given a set of inputs (function parameters) and are interfaces that an agent (using the LLM as brain power) is "aware of" and can access and use to interact with.

For example:

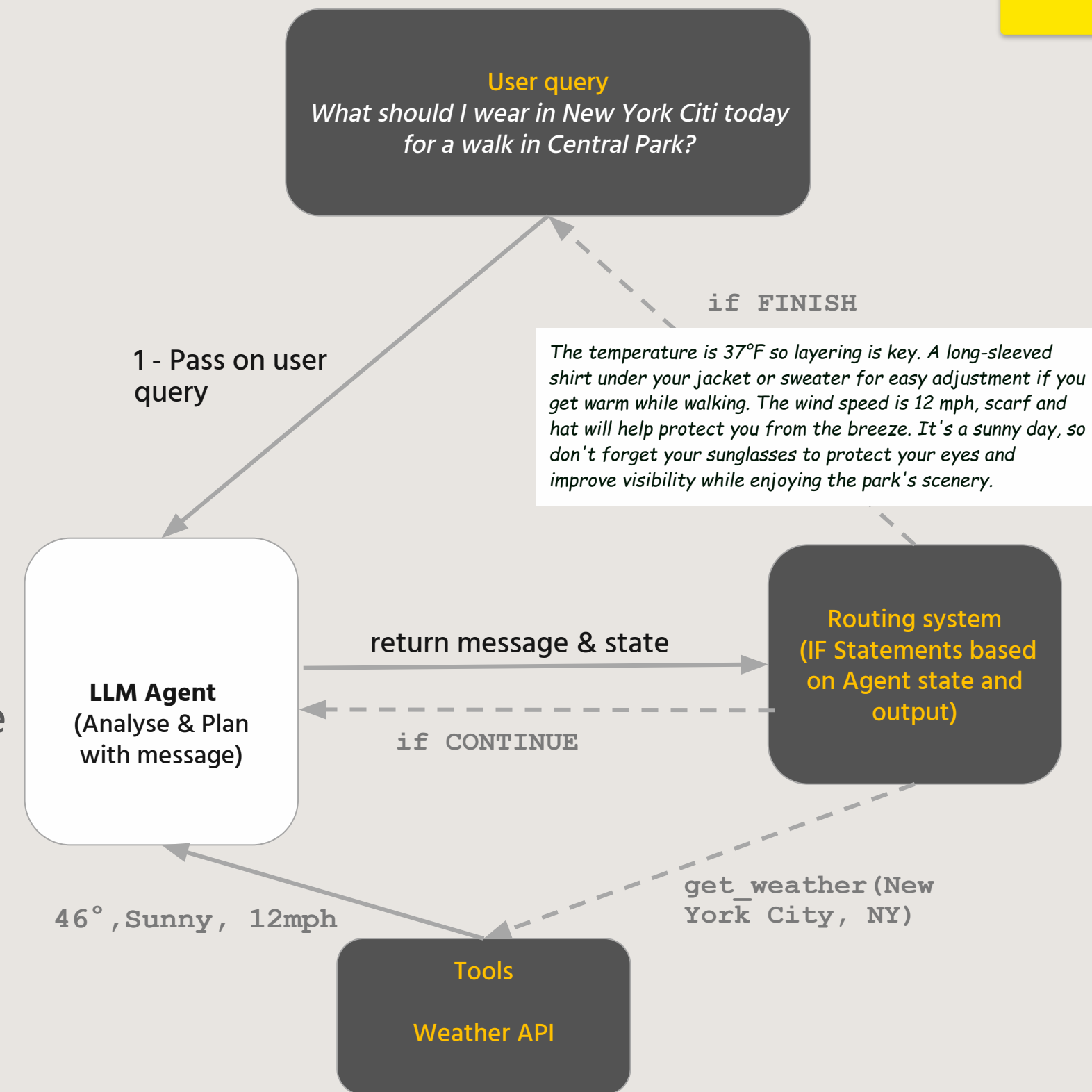
- News API
- Web Search Tool
- H2O MLOps Prediction endpoint
- AutoML tool API (like H2O DriverlessAI)
- Audio Transcription
- Retrieval Augmented Generation (RAG) pipeline
- Code Interpreter



```
{
  "name": "get_country_information",
  "description": "the function get_country_information can be used to Get information about a country such as",
  "parameters": {
    "type": "object",
    "properties": {
      "country": {
        "type": "string",
        "description": "The country of interest, for example Italy",
      },
      "field_to_extract": {"type": "string", "enum": ["capital", "currency", "population", "maps"]},
    },
    "required": ["country", "field_to_extract"],
  },
}
```


What Is Routing?

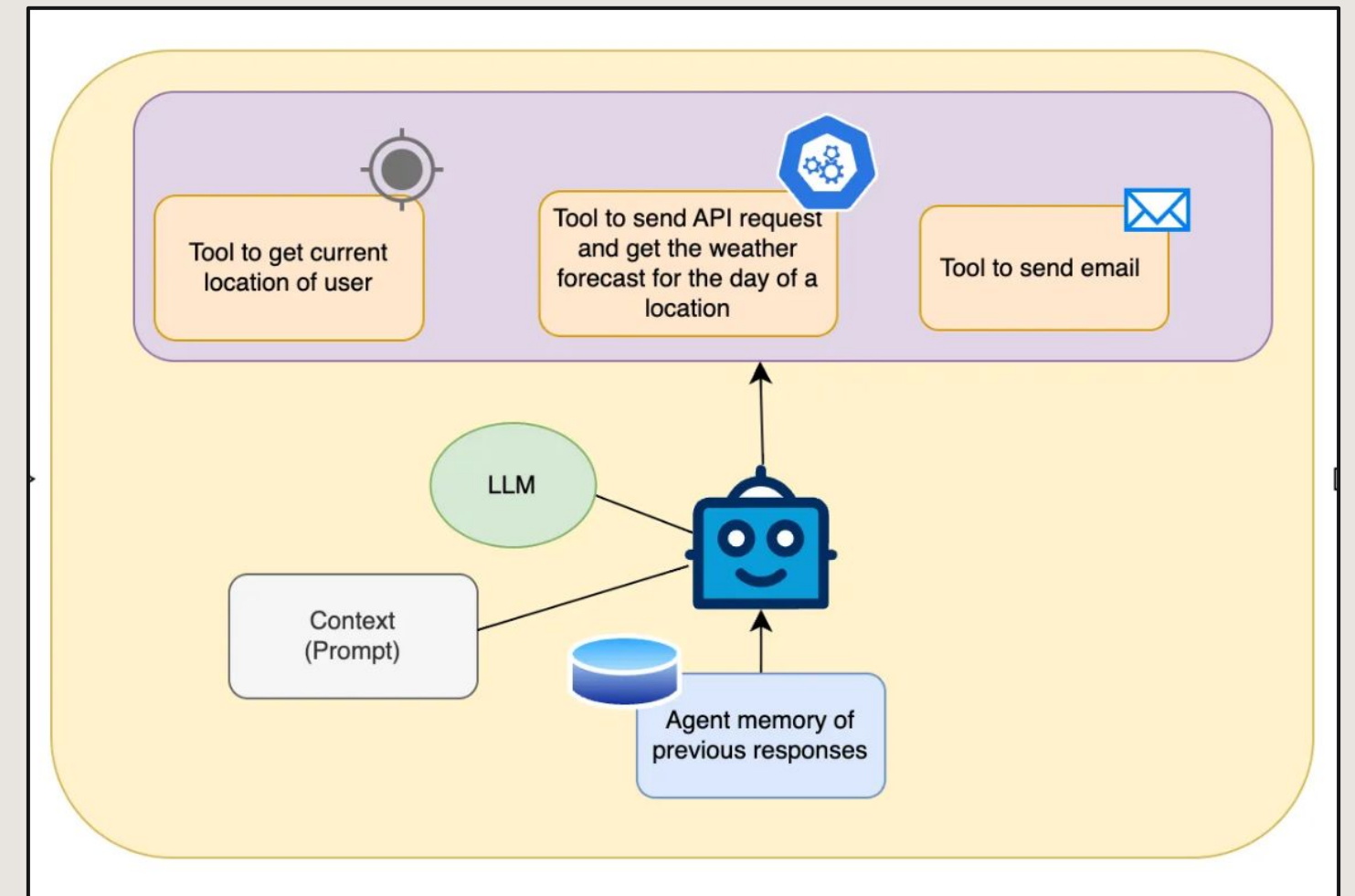
- The Agent is really the chain in charge for deciding what path or chain to take next based on original user query, intermediary information available and current state.
- For this, we can define a route function or routing system, which will check if the agent has reached a final status and return the final answer or will route the query towards the appropriate tools from the available tool and execute it with the function call arguments.
- The routing system can be very complex and includes retries, routing towards specialized agents, fallbacks, stop rules etc.



After executing an action (like getting temperature, wind speed etc.) the results can be fed back into the LLM Agent to assess if further actions are required or if the process can be concluded.

What Are LLM Agents?

- Agents are systems that use LLMs as reasoning and planning engines (thanks to LLMs Natural Language Understanding and Generation capabilities).
- Their role is to decide which actions and steps to take, tools to use and the inputs to pass them based on original query or context, the previous responses and intermediary steps and function calls outputs kept in memory.
- An agent can be good at specific or general actions (RAG Agent) and should understand and execute tasks and can even collaborate with another specialised agent to achieve more sophisticated outcomes.



Should You Use LLM Agents?

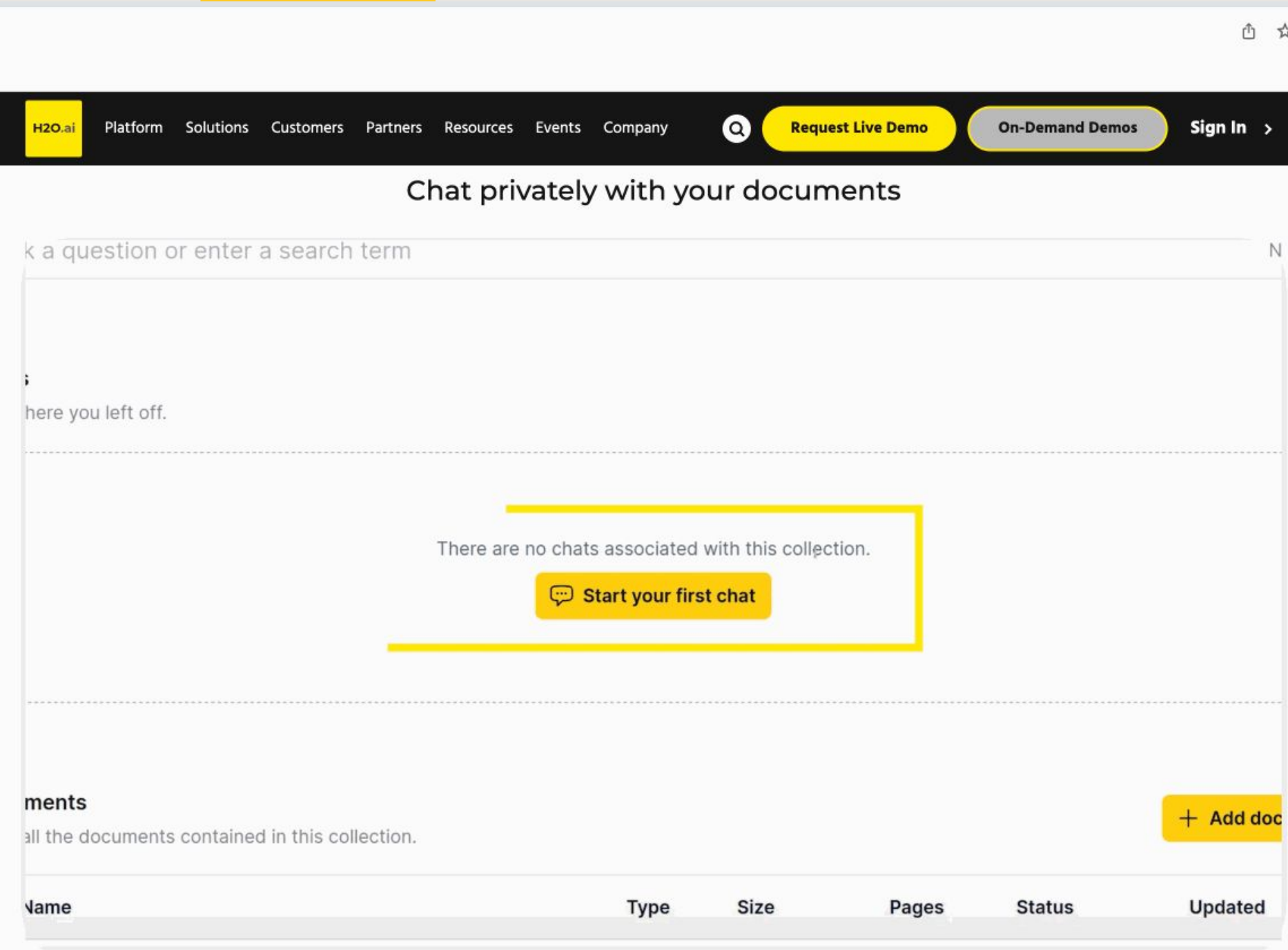
- **When To Use LLM Agents?**

Examples of agents and uses cases are role-playing conversational Agent, Customer Service Support, Healthcare Assistant, Programming and Coding assistant, Analytical Agent, Legal & Compliance Assistant etc.

- **When NOT to use LLM Agents?**

Typically, whatever could be resolved by a rule-based system do not require an LLM agent. Do not replace your existing well-established and inexpensive workflow with agents if flexibility and versatility is not going to add substantial value to your existing solution.

Enterprise h2oGPTe



Transparency, Compliance & Governance

Our platform utilizes open-source LLM's and provides transparency on usage cost, prompts, and any grounding evidence utilized by LLM's. EvalStudio offers the evaluation frameworks to ensure industry standard and customized benchmarks and validation tests.

Secure

On-premise, air gapped and cloud VPC

Only H2O.ai provides an end-to-end GenAI platform where you own every part of the stack. Built for air-gapped, on-premises or cloud VPC deployments. Own your data, own your prompts.

Flexible & Safe

No vendor lock-in. You choose which LLMs are the best fit for your use case. Guardrails and frameworks for evaluations to prevent hallucinations.

Does Enterprise h2oGPTe provide its own LLM Agents to use?

YES! Let's now find out more how we can use it!

Workshop:

Building AI Agents and Mastering Effective Prompt Engineering with h2oGPTe:
A Hands-On Workshop with Real-World Data Science Applications

Go to: <https://h2ogpte.h2oworld.h2o.ai/>

Prompt 1: Generate key insights, charts, and figures using this [climate change report](#). Provide recommendations on how we can reduce the impact of climate change.

Prompt 2: Download Dell, Oracle, MSFT, and Google stock prices for the past two years and plot, compare, and analyze.

Additional Exercises

- Design a futuristic city that is both eco-friendly and technologically advanced. Generate images to showcase this city.
- Have a conversation with h2oGPTe. Make sure you modify the prompt template such that every response it outputs is in a funny tone.
- Pick any company of your choice (i.e. Dell) and ask the following: Analyze and provide key findings of [insert company] stock, and forecast the next year's stock price. Plot relevant charts and figures demonstrating any key trends.
- Using any synthetic data available online, build a simple AI algorithm that can predict fraudulent transactions.

Key Takeaways from this training

AI Agents & their applications: AI Agents are systems that utilize Large Language Models (LLMs) as their core reasoning and planning engines. The applications include conversational agents for *customer service support systems, healthcare assistants, programming and coding aids, analytical agents, and assistants for legal and compliance tasks.*

LLM Chains: Pipelines that connect LLMs with various components such as tools, other LLMs, services, and parsers. LLM Chains can be structured sequentially, recursively, or in parallel, allowing for the creation of powerful and versatile AI systems.

Tools and Functions: Specialized tools that perform specific tasks. These include APIs for news and weather, web search tools, MLOps prediction endpoints, and image analysis tools etc.

Routing: A rule based system to determine the next steps in the process based on the Agent decisions and state. The routing logic can include sophisticated elements like retries, specialized agent routing, fallback mechanisms, and termination rules.

Enterprise h2oGPTe: Enterprise h2oGPTe, a versatile AI assistant, capable of answering questions about various types of content, including documents, websites, and workplace information. It utilizes advanced Retrieval-Augmented Generation (RAG) approaches and comes with its own set of LLM Agents.



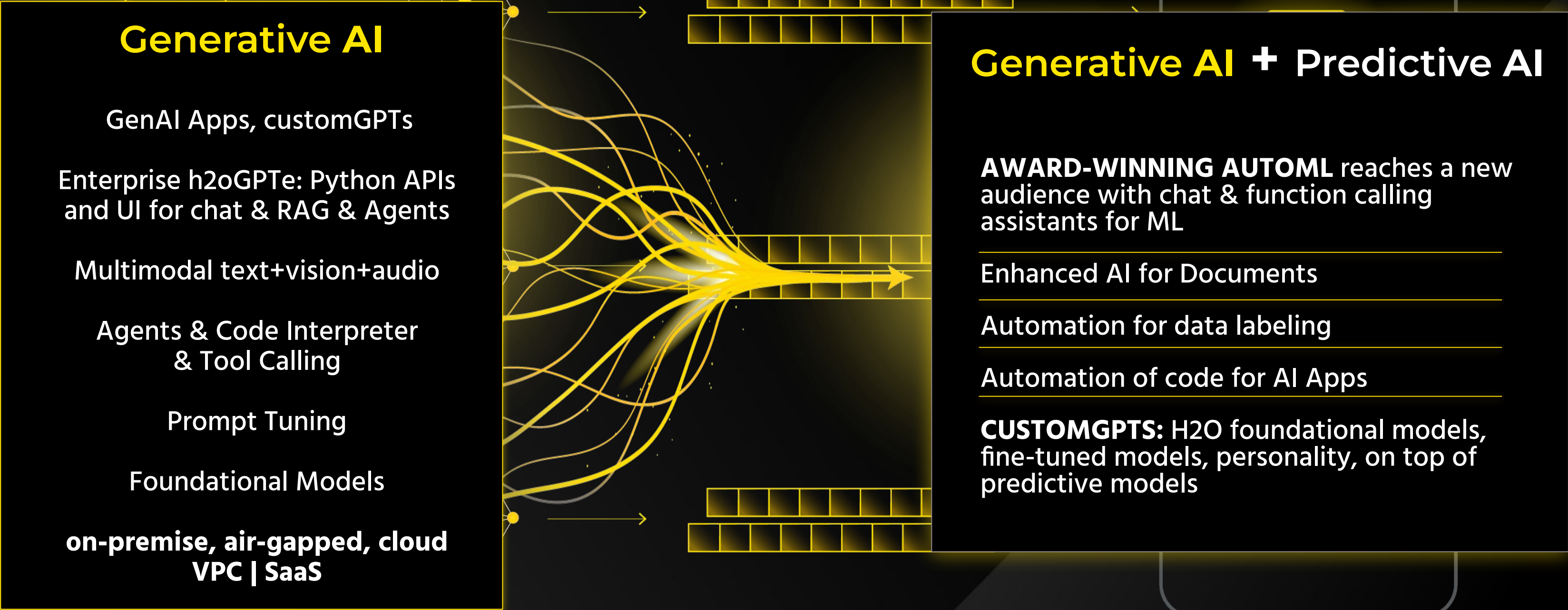
H2O
GenAI WORLD
TRAINING • **NEW YORK**

**AI for Business Stakeholders: Combining Predictive +
GenerativeAI**



Convergence of the World's Best Predictive and Generative AI for Private, Protected Data

Own your data, own your prompts – on-premise, airgapped and cloud VPC



Supercharge human productivity

Enterprise h2oGPTe synthesizes **diverse data** and lets you process **thousands of documents** or web pages—any way you want: inspect, extract, translate, transform, find differences, summarize. Your imagination is the limit.

H2O Generative AI Platform

Convergence of H2O Generative and Predictive AI



Build, deploy, and productionalize custom processes, solutions, and applications.



Customized vertical solution Document AI to analyze contracts/legal documents.



A powerful search assistant on your private data.



Run detailed benchmarks to evaluation LLM & RAG pipelines for high accuracy and security.



Extract data from unstructured.



Automatically create labeled data for LLM fine-tuning.



Developed for OCR and Document AI use cases.



Create your own on-device SLM and fine-tune domain specific datasets for offline use cases.



Fine-tune your LLM for your use cases.

H2O.ai

Our Products

First GenAI platform to unify Generative and Predictive AI for total data privacy—featuring intelligent agents, adaptive workflows, and robust safeguards.

H2O Predictive AI Platform



Deploy and share ML and LLM apps and chatbots to help business users get the most out of the work of data scientists.



Build AI Apps with python for your models and chatbots.



An AutoML platform to streamline model building, explainability, and deployment.



Easily build ML models at scale

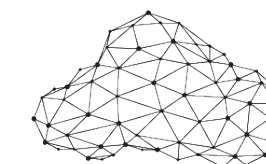


A no-code UI for deep learning use cases built by the world's top Kaggle Grandmasters.



H₂O MLOps

Deploy, monitor, and manage H2O and third-party models at scale efficiency and reliability.



H2O AI Cloud

A comprehensive platform for scalable AI solutions, including machine learning, data visualization, and no-code tools for diverse applications.



Hands-on Workshop

<https://h2oworld.h2o.ai/home>

Hands on lab 1: Using GenerativeAI to extract useful information from unstructured data (eg. Free text)



Gartner estimates that unstructured data represents 80 to 90% of all new enterprise data.

<https://www.smithhanley.com/2023/06/08/structured-versus-unstructured-data-sets/>

Use Case

Large Language Models, as well as Small Language Models can unlock insights from unstructured data, that can then be combined with structured data for predictive analytics use cases.

Why H2O?

Small language models from H2O.ai (Eg. Danube-3) are small enough to fit on and run on an on-premise machine – allowing users to unlock value from their unstructured data in a secure manner!

Hands on lab 2: Data analysis and model building with GenerativeAI + PredictiveAI, no code required!



While code is still the lingua franca for computers, language models have now democratized the ability to translate intent to code – making the “what to do” piece at least as important as the “how to translate what to do into code” piece

Use Case

Data Analysis & model building with GenerativeAI + PredictiveAI.

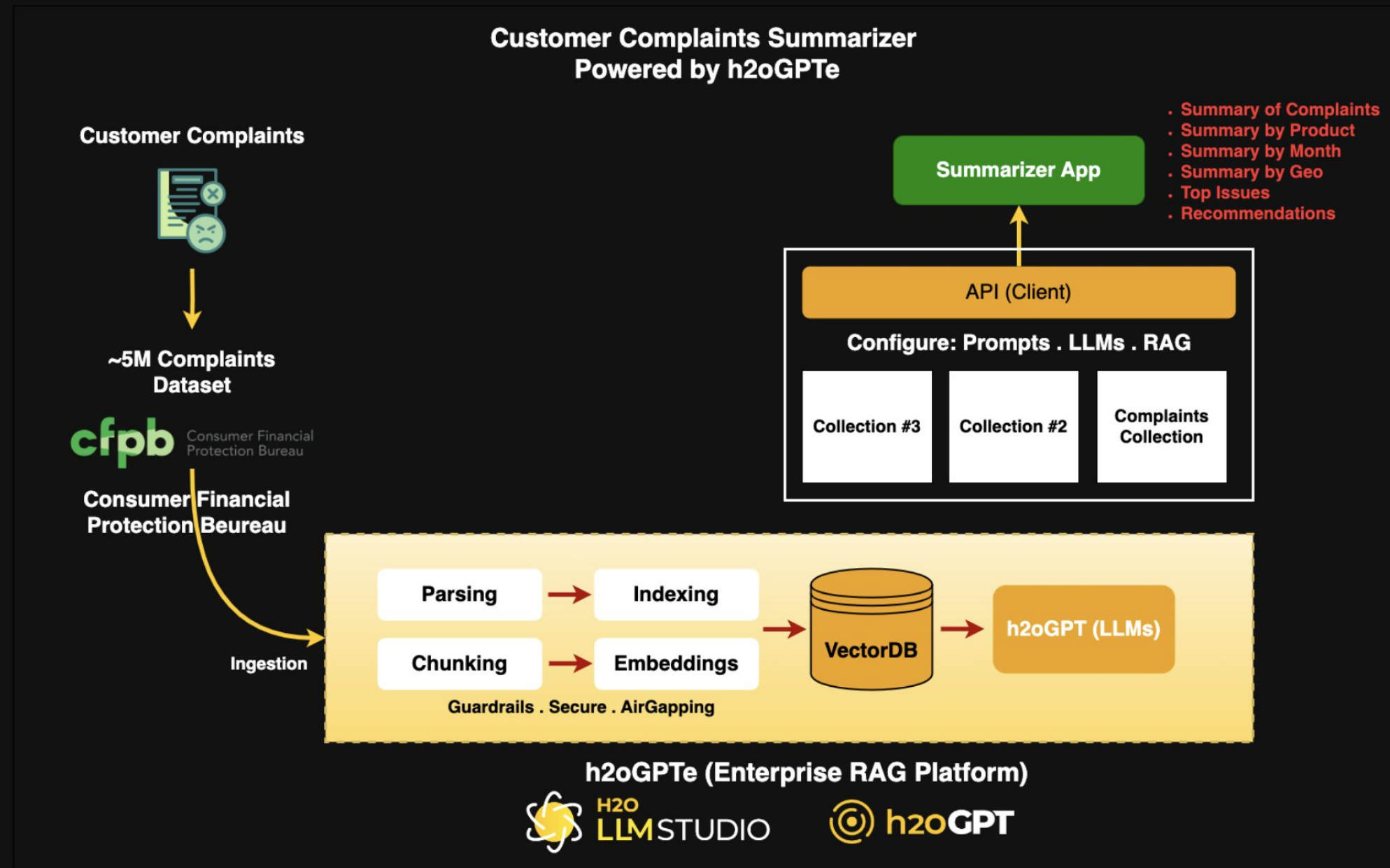
Required a lot of code earlier, today it really requires expressing intent accurately, plus the ability to verify results.

Why H2O?

H2ogpte and the agent framework have extensive support for coding agents that simplify data analysis, modeling and associated tasks!

One can even utilize it in conjunction with other powerful tools such as H2O.ai’s award winning AutoML, DriverlessAI.

Workflows combining Predictive + GenerativeAI : Complaint Summarizer



< ANALYSIS >
✕

e. Annex E - Accepted Notice of Award (NOA); and
f. Annex F - Performance Security.

2. Goods/Services Supplied

2.1. Compliance with Specifications and Pricing The Supplier agrees to supply the

may have, to defer payment of part or all of the Price until the Supplier has completed, to the satisfaction of IOM, the delivery of the Goods and the Incidental Services to which those payments relate.

3.2. Payment.

a. The Supplier shall invoice IOM [upon delivery of all Goods / upon each delivery] in accordance with this Agreement and payment shall become due 30 (thirty) calendar

of 0.1% of the price of the undelivered goods for every day of breach of the delivery schedule by the Supplier.

4.3. Product Certification Requirements.
For all product deliveries, the vendors/suppliers/manufacturers must

REQUIREMENTS

- ✔ a. The Supplier shall invoice IOM [upon delivery of all Goods / upon each delivery] in accordance with this Agreement and payment shall become due 30 (thirty) calendar days after acceptance by IOM of the Goods. 👍 🗨
- ✔ b. The invoice will be accompanied by the following documents: air waybill number, shipping invoice, packing list, certificate of origin [add or delete as required] 👍 🗨
- ✔ c. Payments shall be made in [currency] (currency code) by bank transfer to the following bank account of the Supplier: [bank account details] 👍 🗨
- ✘ d. The amount outlined in Article 3.1 represents the maximum fee payable by IOM. The Supplier will not be liable for any taxes, duties, levies, or charges imposed on IOM related to this Agreement. 👍 🗨
Requirement not met because [reason].
- ✘ e. IOM shall be entitled, without derogating from any other right it 👍 🗨
Requirement not met because [reason].

RELATED SECTIONS
✕

and to charge the Supplier any loss incurred as a result of the Supplier's failure to make the delivery within the time specified; or

b. For any Quality issues, potential disputes and inconsistency of the items, the applicable penalty charging (equivalent percentage of the rejected quality attribute) will be

of company, and in the format acceptable to IOM.

5.2. Purpose and Duration. The Performance Security shall serve as the guarantee for the Supplier's faithful performance and compliance with the terms and conditions of this Agreement. The amount of the Performance Security shall not be construed

days from the completion of Supplier's obligations] following which it will be discharged by IOM.

6. Inspection and Acceptance

Page 3 (0%)

failure to make the delivery within the time specified; or

b. For any Quality issues, potential disputes and inconsistency of the items, the applicable penalty charging (equivalent percentage of the rejected quality attribute) will be applied and any additional inspection and/or laboratory costs,

5.2. Purpose and Duration. The Performance Security shall serve as the guarantee for the Supplier's faithful performance and compliance with the terms and conditions of this Agreement. The amount of the Performance Security shall not be construed as the limit of the Supplier's liability to IOM, in the event of breach of this Agreement by

and charges assessed on it in connection with this Agreement.

e. IOM shall be entitled, without derogating from any other right it may have, to defer payment of part

account. Sample of Penalty computation can be found in Annex D

c. Liquidated damages will also be applied if the Supplier fails to deliver

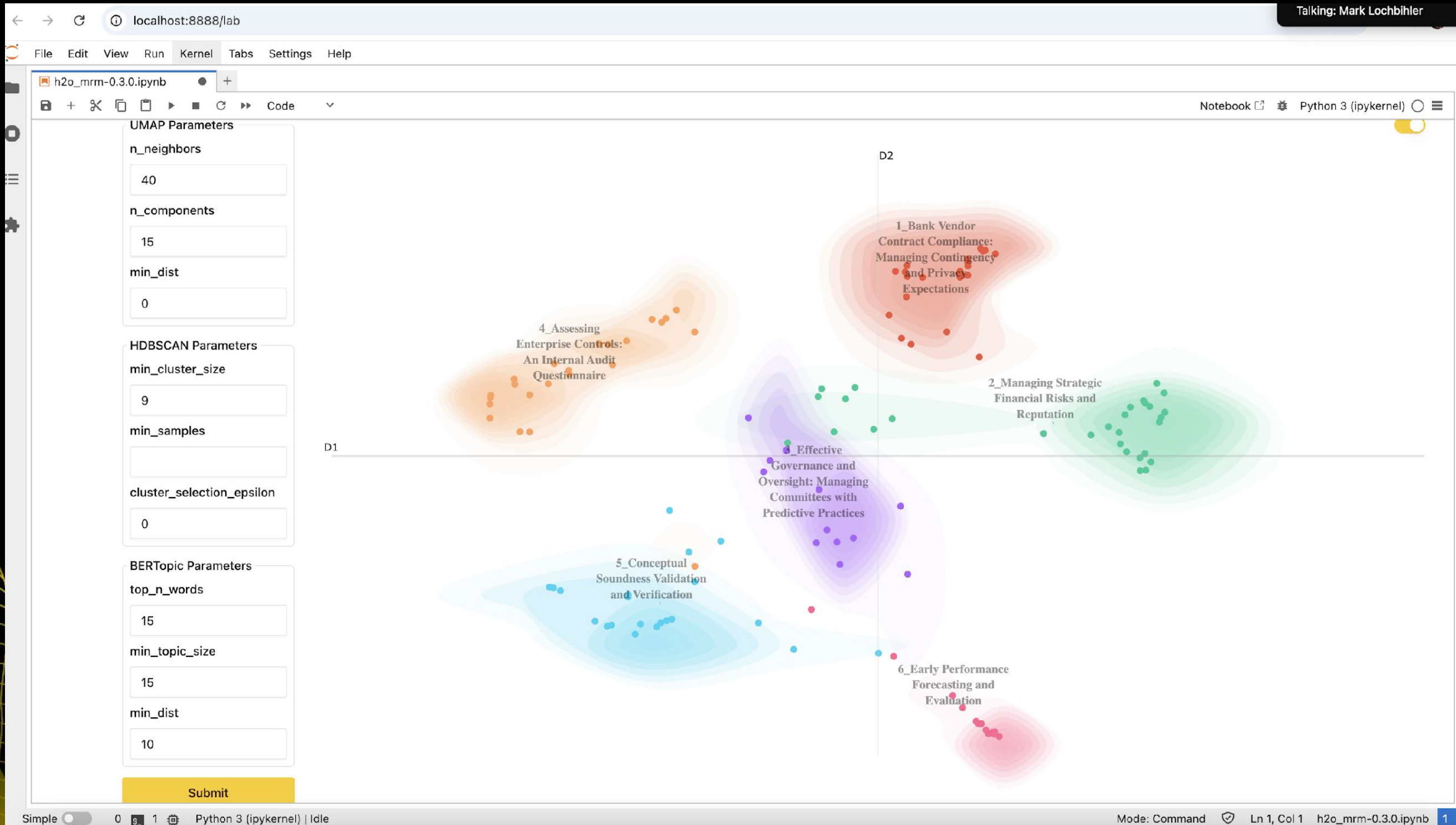
Page 7 (0%)

indemnity, and hold harmless IOM, its officers, employees, and agents from and against all losses, costs, damages and

be refunded on or before the date of termination.

b. If IOM terminates this Agreement in

Workflows combining Predictive + GenerativeAI : Model Risk Management and semi-automatic validation of LLM applications!



AutoML & Enterprise H2oGPTe powered Use-Cases Quick Showcase

Assistant

[Demo #1](#)

RFI Assistant

Get assistance for your Request For Information

Customer Service

[Demo #2](#)

Sentiment Analysis

What are your customers talking about?
Identify topics and sentiment

Customer Service

[Demo #3](#)

Complaint Summarizer

Identify key & actionable feedbacks

Fraud

[Demo #4](#)

Scenario Fraud Analysis

Assessing fraud risk given risk profiles
(medical industry)

Credit Risk

[Demo #5](#)

Loan Prequalification

Credit Risk Simulation and advices

Operations

[Demo #6](#)

GPTeller

AI-powered, efficient customer service automation.



H2O
GenAI WORLD
TRAINING • **NEW YORK**

**Kernel Internal and
HAIC Deployments**

21st November 2024

Mark Lochbihler, mark.lochbihler@h2o.ai





Kernel Internal and HAIC Deployments

Introduction to "Kernel Internal"
capabilities powering [H2O.ai](https://h2o.ai) Cloud



H2O AI Cloud: Managed & Hybrid

H2O AI Cloud

Democratize and Accelerate AI Results with Trust and Confidence



Make.

Make Models with Accuracy, Speed and Transparency



H2O-3
Open Source Distributed Machine Learning



H2O Driverless AI
Award-Winning AutoML



H2O Hydrogen Torch
No-Code Deep Learning



H2O Document AI
Intelligent Document Models



LLM Studio
Open Source LLM Fine-Tuning

Operate.

Streamline Performance Monitoring and Rapidly Adapt to Changing Conditions



H2O AI Feature Store
High-Scale and Intelligent Feature Store



H2O Label Genie
AI labeling because no one wants to label data



H2O MLOps
Model Hosting, Monitoring and Deployment



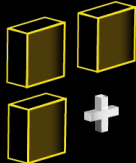
Hybrid Cloud
Deploy in your VPC, on prem, airgapped, where you need



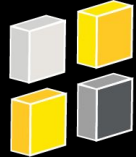
Managed Cloud
Hosted and managed for you, by H2O

Innovate.

Easily Deliver Innovative Solutions to End Users with an Intuitive AI App Store



H2O AI App Store
Find, Share, and Discover AI Apps



Industry Apps
Pre-built AI Apps



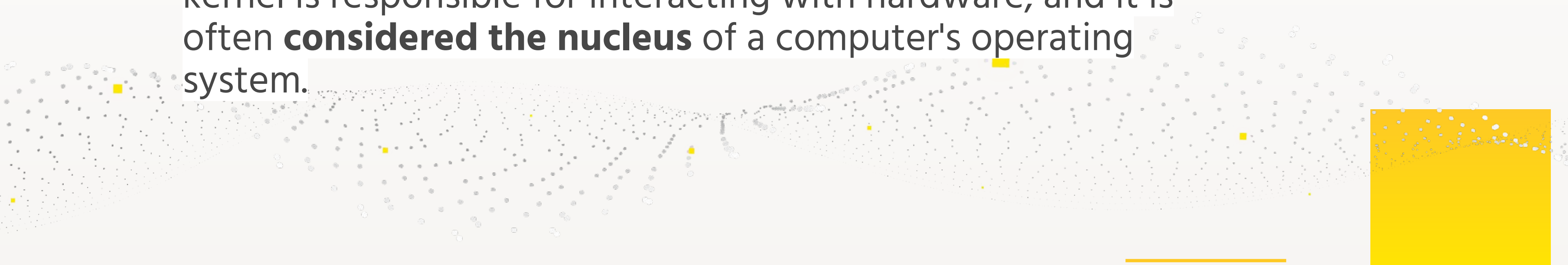
H2O Wave
Low-Code AI AppDev Framework



Enterprise h2oGPT
ChatBots and more with H2O LLMs

Kernel Definition

A kernel, in the context of computing, is essentially **the core of an operating system**. It is the fundamental layer that exists between the computer hardware and the software. The kernel is responsible for interacting with hardware, and it is often **considered the nucleus** of a computer's operating system.



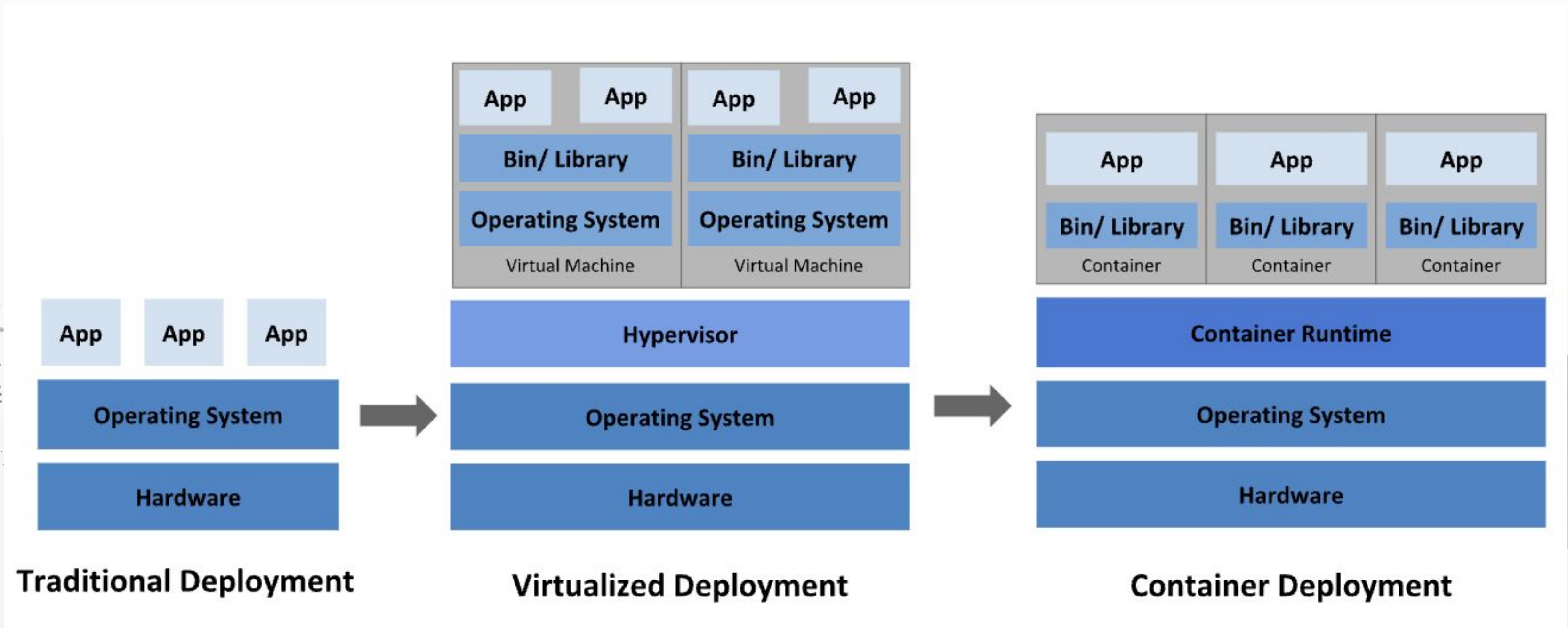
What is inside H2O.ai Cloud (HAIC)

HAIC's "Kernel Internal" is Kubernetes (K8s), a portable, extensible, open source platform for managing containerized workloads and services, that facilitates both declarative configuration and automation. It has a large, rapidly growing ecosystem. Kubernetes services, support, and tools are widely available.

Reference:

<https://kubernetes.io/docs/concepts/>

History



Reference:

<https://kubernetes.io/docs/concepts/>

What is K8s Good For?

Kubernetes provides you with:

- Service discovery and load balancing
- Storage orchestration
- Automated rollouts and rollbacks.
- Automatic bin packing
- Self-healing
- Secret and configuration management
- Batch execution
- Horizontal scaling
- IPv4/IPv6 dual-stack
- Designed for extensibility

Reference:

<https://kubernetes.io/docs/concepts/>

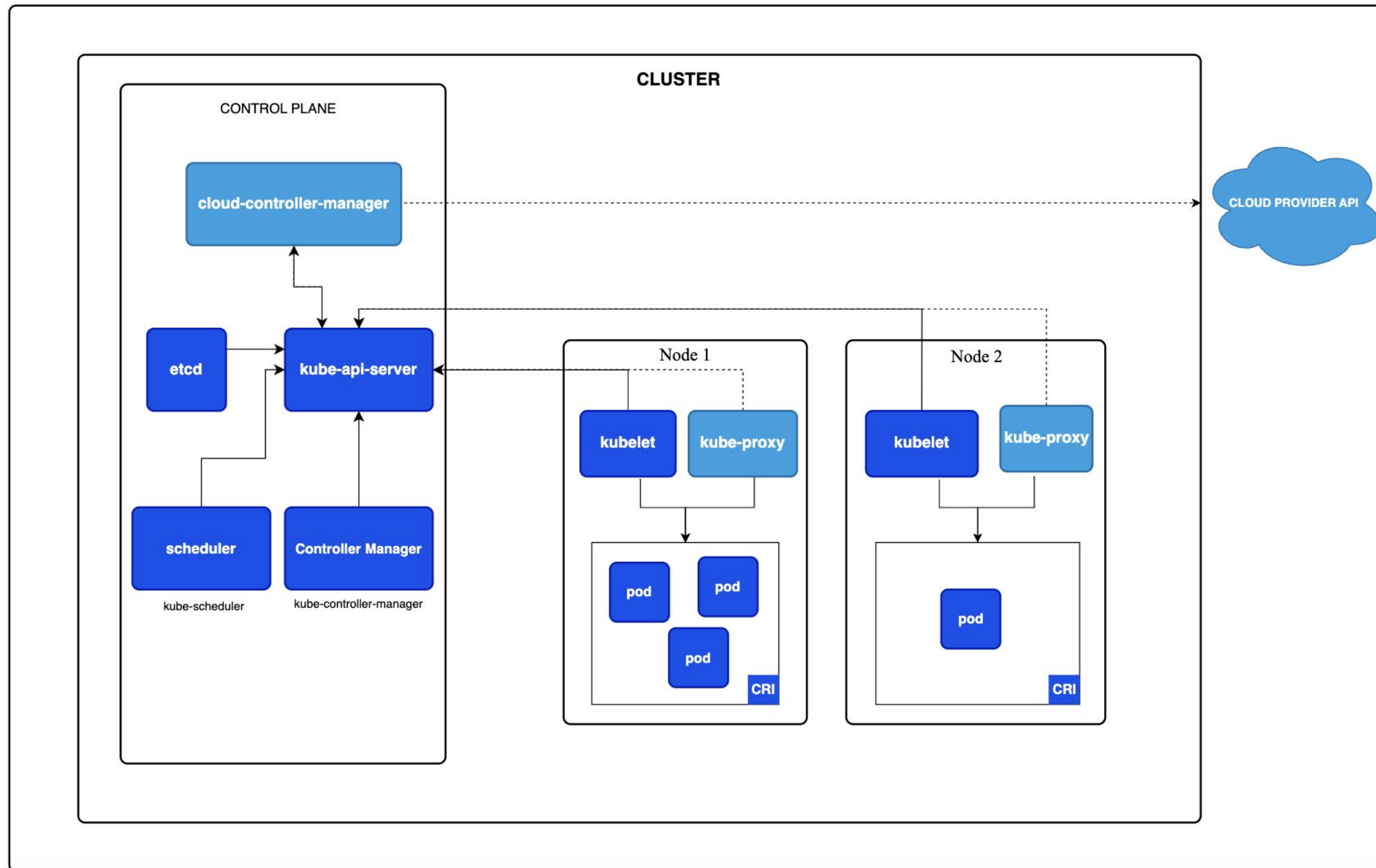
K8s is not merely an orchestration system

The technical definition of orchestration is execution of a defined workflow: first do A, then B, then C. In contrast, **Kubernetes comprises a set of independent, composable control processes that continuously drive the current state towards the provided desired state.**

- It shouldn't matter how you get from A to C. Centralized control is also not required. This results in a system that is easier to use and more powerful, robust, resilient, and extensible.

Reference:

<https://kubernetes.io/docs/concepts/>



Pods are the smallest deployable units of computing that you can create and manage in Kubernetes.

A Pod models an application-specific "logical host": it contains one or more application containers which are relatively tightly coupled.

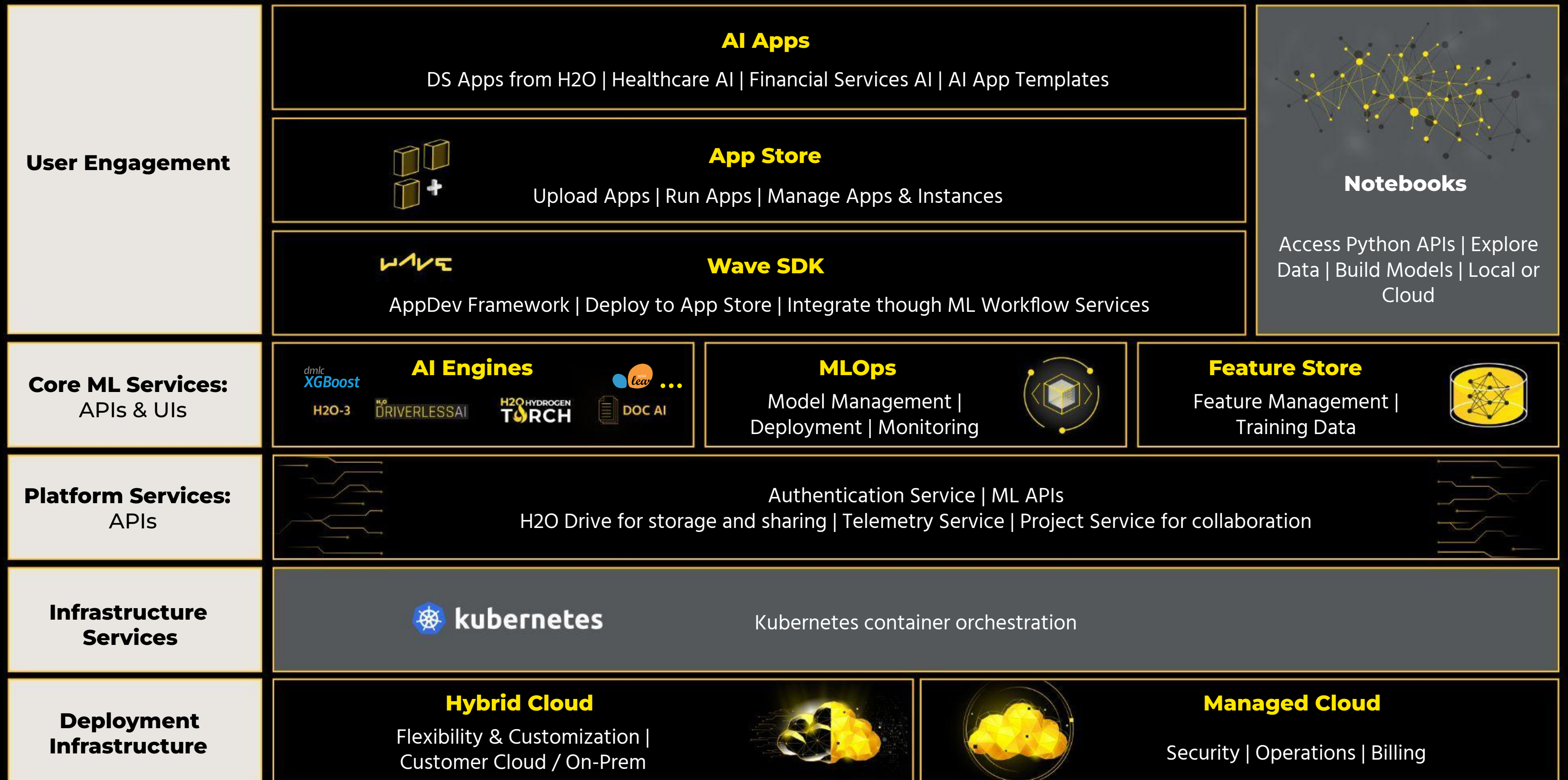
HAIC runs on Top K8s Distributions - On Premise and Public Cloud



kubernetes



What is the H2O AI Cloud?



AI Platform Requirements and the H2O AI Cloud

H2O AI Cloud is the most comprehensive AI Cloud, that meets the demanding requirements of the largest enterprises and the fastest growing startups.

AI Clouds deliver faster time-to-value, optimize AI's impact, and make it much simpler and lower risk to manage and govern AI across an organization.

H2O AI Cloud has a set of products and capabilities, that help it stand out against every key AI Cloud platform requirement.

	H2O AI Cloud
Support all AI Use Cases	<ul style="list-style-type: none"> • Big Data >1TB • Structured Data • Time-Series Data • Text, Image and Audio Data • Document Data
Deliver Rapid Time-to-Value	<ul style="list-style-type: none"> • Advanced AutoML • No Code Deep Learning and Document AI • 1-Click Deployment to Production • Apps for Business Explainability • Low Code AI App Development • AI Apps for Users
Support Multiple Users / Democratize AI	<ul style="list-style-type: none"> • Code Interfaces • No Code Interfaces for all Data Types • #1 No Code AutoML • Snowflake SQL Interface Integration • AI AppStore
Easy to Explain, Monitor, and Govern AI	<ul style="list-style-type: none"> • Most Robust Explainable AI Capabilities • Model Registry with Hosting, Scoring, and Monitoring • Single Pane of Glass for models in DBs, apps, H2O MLOps, existing tools
Provide the Highest Accuracy	<ul style="list-style-type: none"> • AutoML produces the highest accuracy results with the fewest iterations • 100s of Kaggle Grandmaster Optimized Recipes for Use cases • Deep Learning and Document AI Optimized by Kaggle GMs
Integrate into Existing Data, Apps and AI Tools	<ul style="list-style-type: none"> • No infrastructure management, eliminates undifferentiated heavy lifting • Any Cloud, Multi-Cloud, or Hybrid Environment • Integrates with all existing data, AI tools and apps

Key Takeaways from this training

What is H2O AI Cloud?

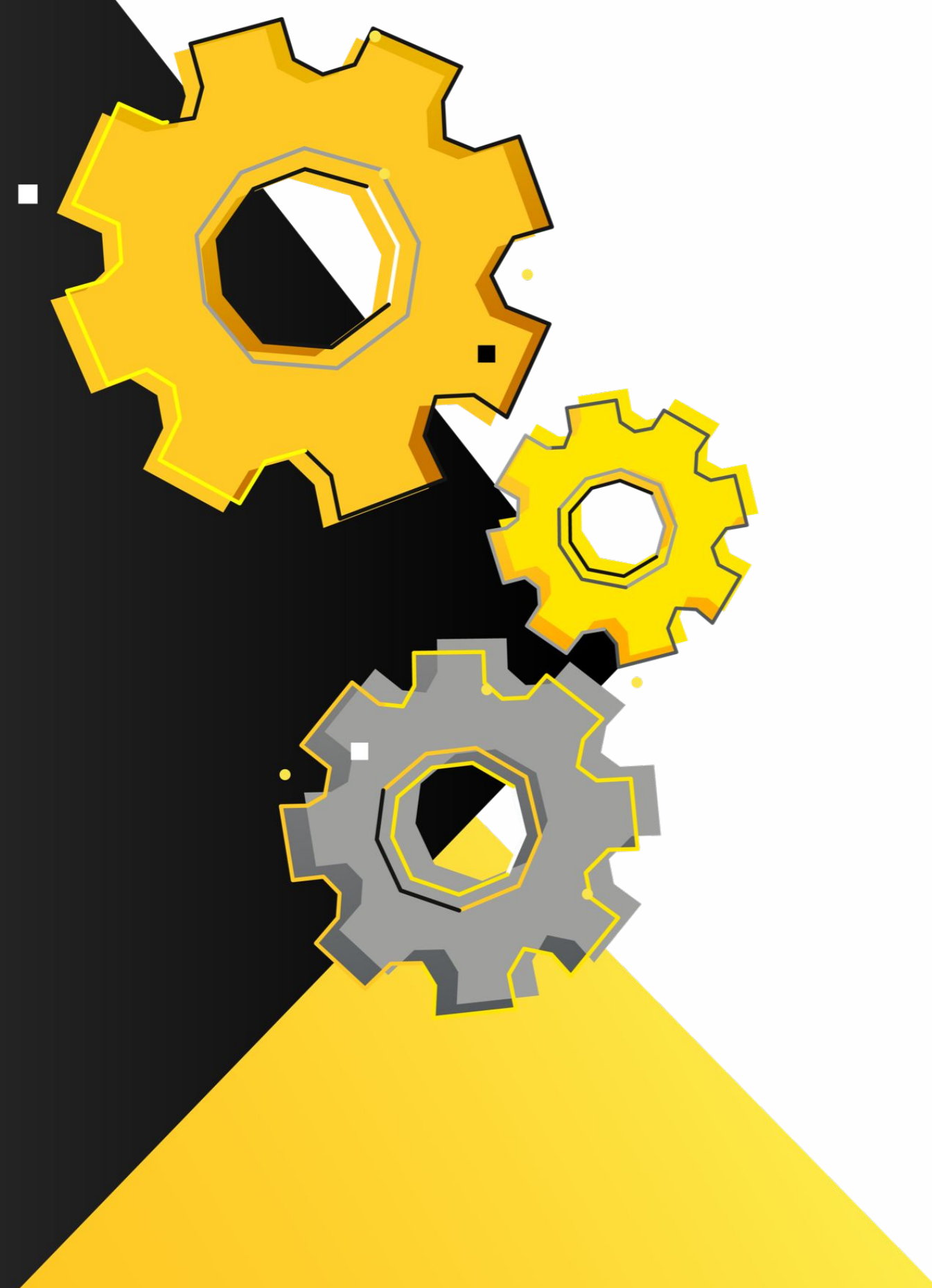
It is the most comprehensive AI Cloud, that meets the demanding requirements of the largest enterprises and the fastest growing startups.

HAIC Pillars:

- **Make.** Make Models with Accuracy, Speed and Transparency.
- **Operate.** Streamline Performance Monitoring and Rapidly Adapt to Changing Conditions.
- **Innovate.** Easily Deliver Innovative Solutions to End Users with an Intuitive AI App Store.

HAIC's Kernel Internal: Is Kubernetes (K8s), a portable, extensible, open source platform for managing containerized workloads and services.

HAIC Runs on Top K8s: On Premise and Public cloud, including OpenShift, Rancher, AKS, EKS, GKS and native K8s.



Configuring and Deploying [H2O.ai](https://h2o.ai) on Dell AI Factory with NVIDIA infrastructure

H2O Generative AI Platform

Convergence of H2O Generative and Predictive AI



Build, deploy, and productionalize custom processes, solutions, and applications.



Customized vertical solution Document AI to analyze contracts/legal documents.



A powerful search assistant on your private data.



Run detailed benchmarks to evaluation LLM & RAG pipelines for high accuracy and security.



Extract data from unstructured.



Automatically create labeled data for LLM fine-tuning.



Developed for OCR and Document AI use cases.



Create your own on-device SLM and fine-tune domain specific datasets for offline use cases.



Fine-tune your LLM for your use cases.

H2O.ai

Our Products

First GenAI platform to unify Generative and Predictive AI for total data privacy—featuring intelligent agents, adaptive workflows, and robust safeguards.

H2O Predictive AI Platform



Deploy and share ML and LLM apps and chatbots to help business users get the most out of the work of data scientists.



Build AI Apps with python for your models and chatbots.



An AutoML platform to streamline model building, explainability, and deployment.



Easily build ML models at scale

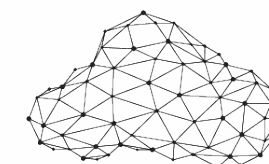


A no-code UI for deep learning use cases built by the world's top Kaggle Grandmasters.



H₂O MLOps

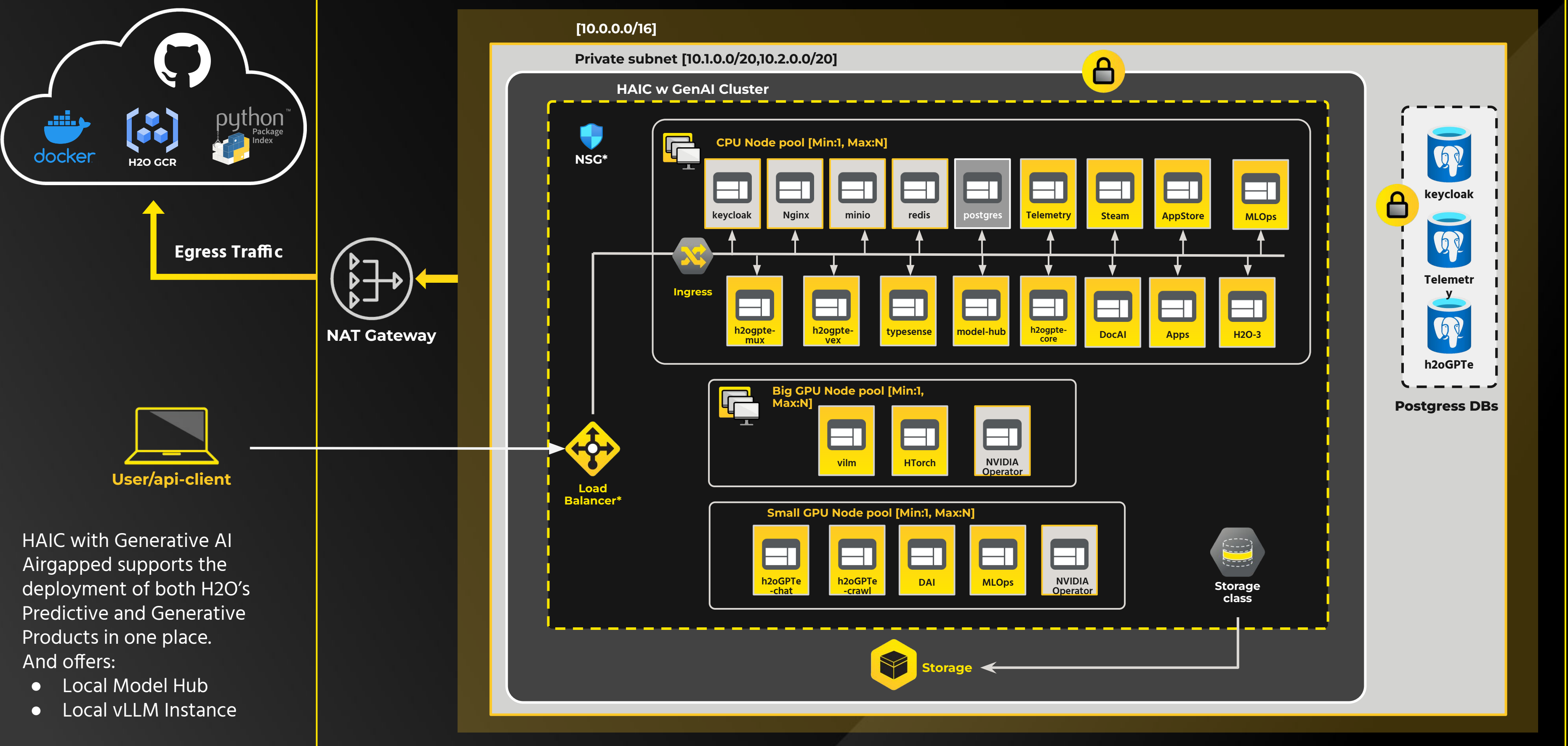
Deploy, monitor, and manage H2O and third-party models at scale efficiency and reliability.



H2O AI Cloud

A comprehensive platform for scalable AI solutions, including machine learning, data visualization, and no-code tools for diverse applications.

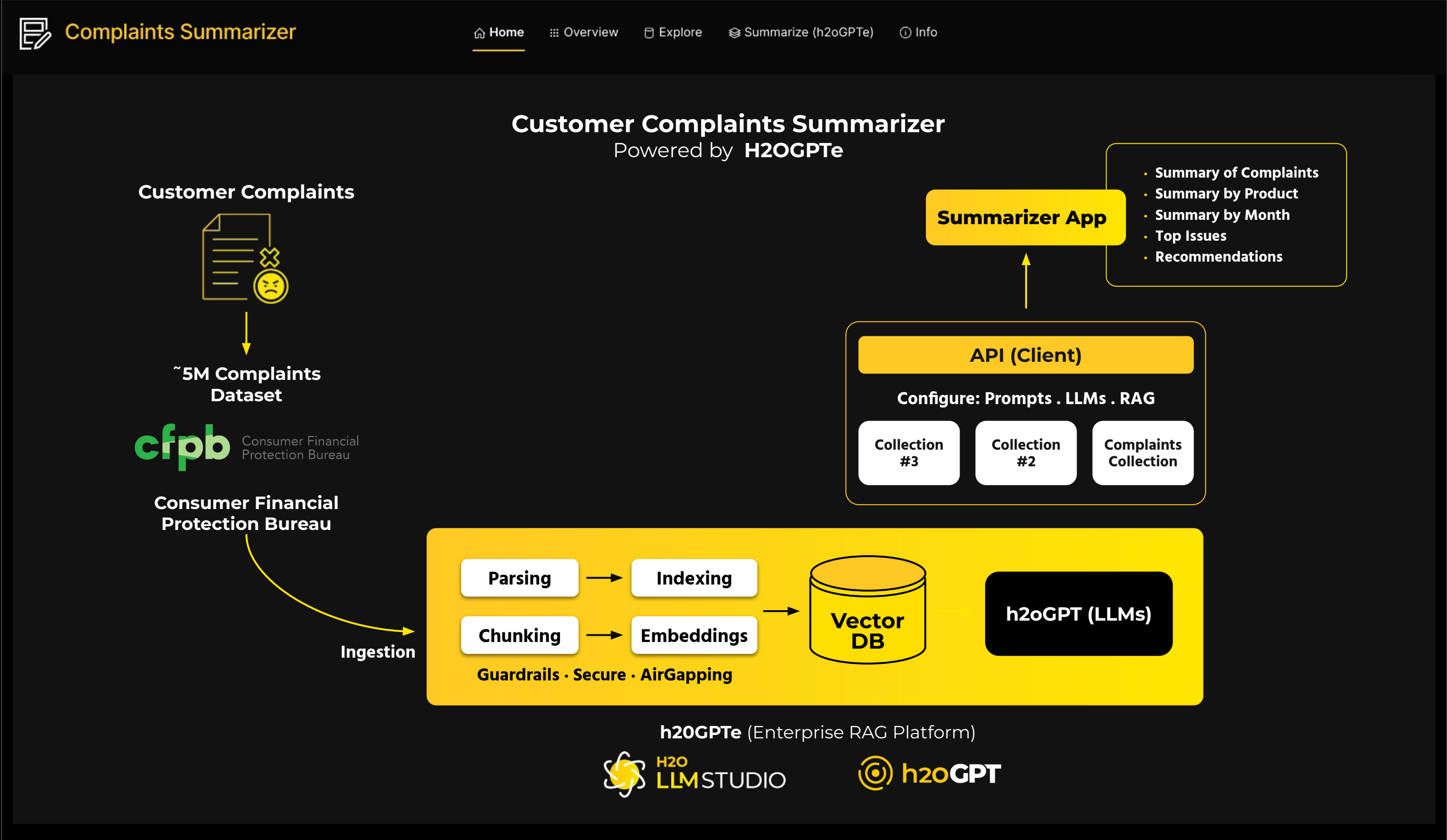
HAIC with Generative AI Airgapped Deployment



HAIC with Generative AI Airgapped supports the deployment of both H2O's Predictive and Generative Products in one place.

- And offers:
- Local Model Hub
 - Local vLLM Instance

Deployment Based on Complaint Summarizer Application



Dell Reference Design Published

Home > AI Solutions > Gen AI > White Papers > Accelerated Customer Issue Resolution with H2O.ai Complaints Summarizer on Dell AI Factory with NVIDIA

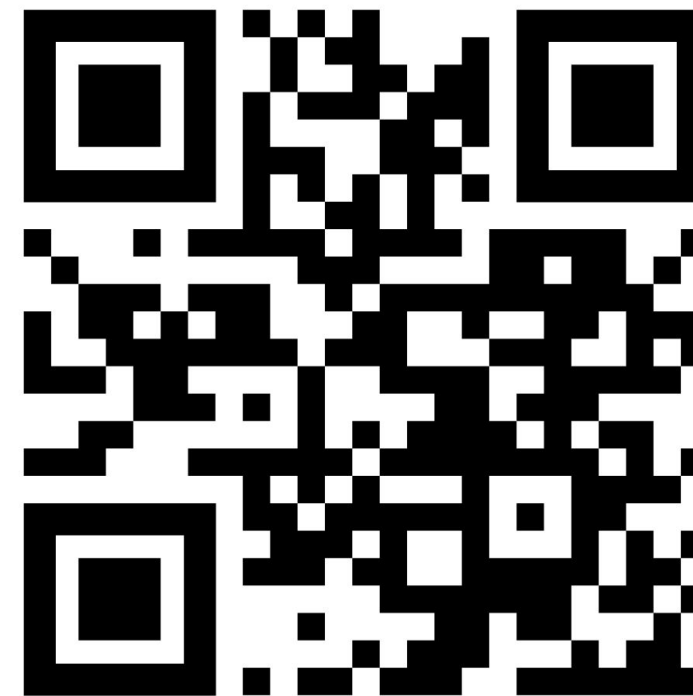
Accelerated Customer Issue Resolution with H2O.ai Complaints Summarizer on Dell AI Factory with NVIDIA

This technical white paper describes a solution for on-premises customer service issue resolution using a Dell Reference Design for H2O.ai Enterprise h2oGPTe on Dell AI Factory with NVIDIA.

★★★★★



Download PDF



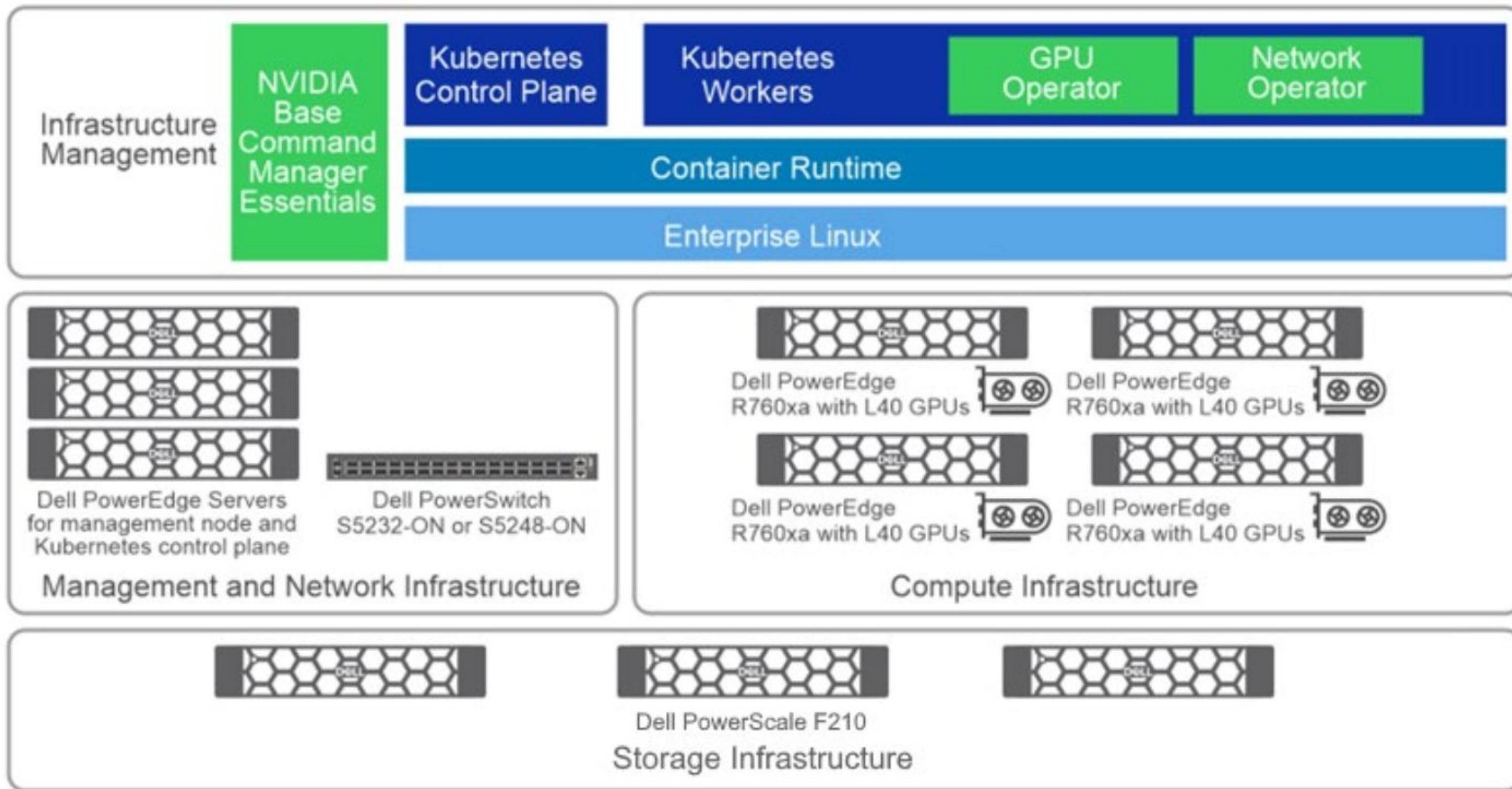


Figure 1. Dell AI Factory with NVIDIA hardware and infrastructure management



RackScale 1 (~14kWs)

Hardware

- 3x Dell PowerEdge R660
- 4x Dell PowerEdge R760XA
 - 2x Intel Gold CPU (32 cores)
 - 512GB RAM
 - 2x 3.84TB Storage
 - 16x NVIDIA L40S GPU
- NVIDIA Bluefield-3 (1x3220/2x3140H)
- 3x NVIDIA Spectrum4 SN5600
- 2x PowerSwitch S5232F-ON Storage Cluster BE
- 1x NVIDIA SN2201 ToR
- 3x PowerScale F210
- 1x APC 750x1200 Rack (A7096292)
- 2x PDU (AC021024)

Software

- 5x Ubuntu OS
- Upstream Kubernetes

|

Services

- ProSupport+ or ProSupport One
- ProDeploy, ProConsult

For Complaint Summarizer, we started with the following HAIC deployment pods

- **H2O.ai Cloud (HAIC) pods**
 - **Enterprise h2oGPTe** - AI powered search and agents
 - o h2ogpte-mux - User interface and API endpoint service
 - o h2ogpte-core - the main enabling service for h2oGPTe
 - o h2ogpte-chat - a scalable chat service
 - o h2ogpte-crawl - a scalable ingestion service
 - o h2ogpte-vex - a local, private embedded VectorDB service
 - o vllm - enables local LLM hosting and inference
 - o model-hub - a local, private model hub
 - o Typesense – typo-tolerant search engine
 - **AppStore** – H2O.ai application deployment service
 - **Operational Services**
 - o Keycloak – Provides user federation, strong authentication, user management, fine-grained authorization and is extensible to customer security frameworks
 - o Nginx – web server for reverse proxy, load balancing and caching.
 - o Minio – Local S3 compatible object store
 - o Redis – Local in-memory storage
 - o Postgres – Local relational database management system
 - o NVIDIA Operator – from Kubernetes, manages GPU resources
 - o Telemetry - create and manage telemetry data
 - **Other** – During discovery, it may be determined that additional H2O.ai Predictive and Generative AI capabilities are needed, like Driverless AI, Document AI and MLOps. These Kubernetes based services can easily be added to this cluster to meet the customer needs.

Complaint Summarizer - Initial Persistent Storage Requirements

K8s Pod PVC Requests were satisfied:

- appstore - 128GB
- h2ogpt - 254GBi
- vllm - 512GBi
- chat, core, crawl and mux - each w 100GB
- typesense - 64GB
- vex - 1000GB
- postgresql - 20GBi
- minio - 512GB
- redis - 60GB

For Complaint Summarizer, Initial GPU Considerations

The Enterprise h2oGPTe platform supporting this solution leverages GPUs and here are some considerations to keep in mind:

- **LLM Inferencing** - We have deployed the open source **Llama 3.1 8B Instruct** LLM and it is **leveraging a single GPU** and approximately **16GB of GPU RAM**. As workloads increase due to more demand, it is possible in the current deployment to increase parallelism for inferencing to leverage 2 or even 4 GPUs as needed to maintain acceptable latency for responses. It is important to note that you are able to leverage a different LLM with this solution, and this can lead to increased GPU and GPU RAM requirements..
- **Chat** - For initial Prompt and Response workloads, we have enabled the Chat service with a single replicate, **leveraging a single GPU**. It is possible to increase the number of replicates to scale Chat throughput and response times as the need arises.
- **Crawl** - Crawl is used for the Document Ingestion process and is initially set to a single replicate **leveraging a single GPU**. This can be increased as well as demands in this area of the solution increase over time.

For Complaint Summarizer, Security Considerations

Security considerations

Note that **HAIC offers a completely private deployment** of Generative and Predictive AI capabilities, with all AI services and customer data deployed 100% within a customer's firewall.

This means that data used, as an example, for Generative AI RAG never leaves your network, **as a private LLM inference server and vector database** are included as **part of the Enterprise h2oGPTe platform**.

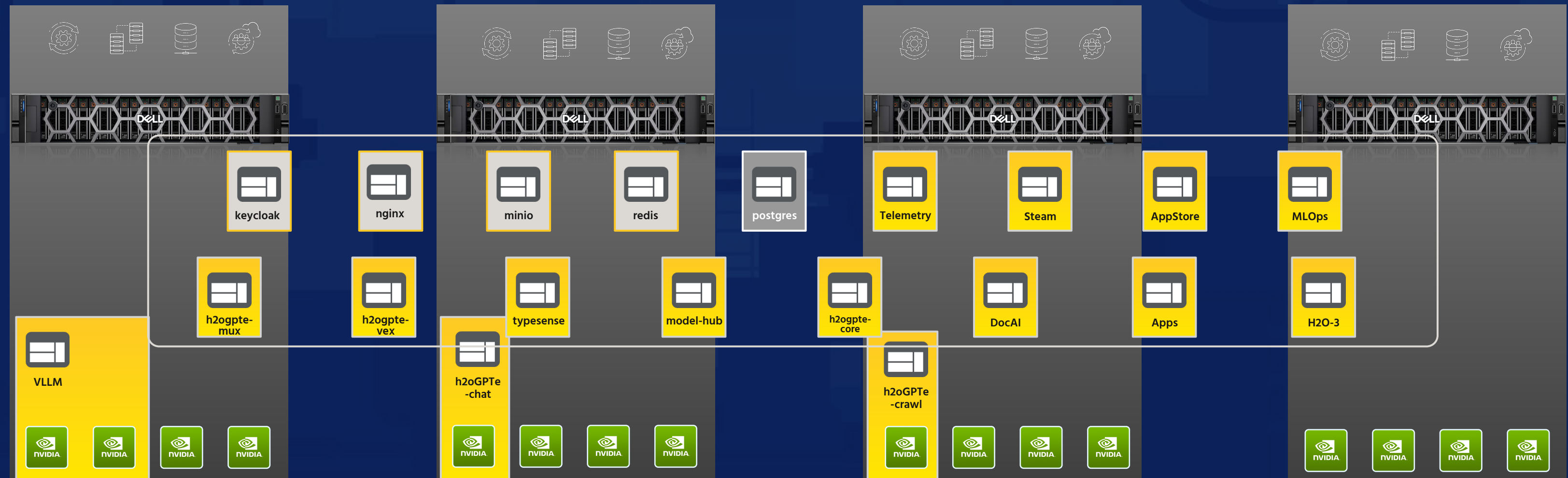
And for **identity management**, HAIC offers **Keycloak**, which **enables integration into a customer's existing Active Directory or LDAP, or other third party identity provider using OIDC**.

H2O.ai on Dell AI Factory with NVIDIA

The world's broadest AI solutions portfolio from desktop to data center to cloud

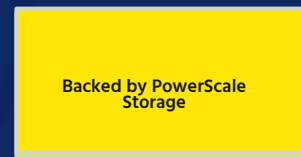
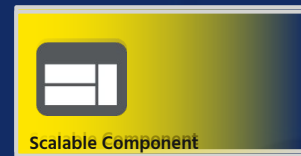
H2O.ai Cloud INFRASTRUCTURE

4x R760xa 16x NVIDIA L40s GPU
3x R660



H2O.ai on Dell AI Factory with NVIDIA

4x R760xa w/16x NVIDIA L40s GPU, 3x R660



Key Takeaways from this Session

HAIC Airgapped Deployment supports:

The deployment of one or both of **H2O's Predictive and Generative Product Stacks** in one place. And **offers both a local Model Hub and local vLLM Instance**.

HAIC is Validated on Dell AI Factory with NVIDIA

Working with the Dell Customer Solution Center, H2O validated HAIC with the Complaint Summarizer accelerator application running on deployed DAIFWN.

Enterprise h2oGPTe pods which can be configured to leverage one or more GPUs

- **vLLM** - For local, private LLM Inferencing
- **Chat** - Supports LLM and RAG Chat Sessions
- **Crawl** - Supports Document ingestion, chunking and embedding

HAIC Dell Server Configuration

4x R760xa 16x NVIDIA L40s GPU, 3x R660

HAIC Persistent Storage

For this deployment, they are backed by Dell PowerScale Storage Servers

Certification Exam

Introduction to Enterprise h2oGPTe, Agents and Combining Predictive and

Generative AI - Training Technical Track

- **Data Science**
- **AI for Business Stakeholders**
- **Kernel Internal and HAIC Deployments**



<https://tinyurl.com/h2oaitraining>



Thank you

