



HPE Private Cloud AI Live Demo



HPE Geneva
Customer
Innovation Center

Introduction & Demonstration platform

HPE Private Cloud AI is a turnkey AI solution built on NVIDIA AI computing. It includes the hardware, software, and services required to deploy AI applications quickly and efficiently while keeping private data safe and accessible.

Powered by HPE AI Essentials, HPE Private Cloud AI provides a single interface for managing one or more solutions across an organization regardless of system location

The purpose of this demonstration is to propose a Live Tour of the HPE Private Cloud AI solution, management from HPE GreenLake and the main features enables through HPE AI Essentials.

The demonstration platform is a Small HPE Private Cloud AI, its architecture is detailed on the right side.



HPE GreenLake Platform

NVIDIA Networking

Nvidia SN4600M Switches (100GbE Networking)

3 x Control Nodes (PCAI and K8 Control Plane)

HPE DL325 servers

HPE GreenLake for File Storage

Powered by HPE Alletra MP

2 x AI-optimized Nodes

HPE DL380a servers



Shared PCAI environment and automatic cleanup

The HPE Private Cloud AI environment is a shared demonstration platform.

Automatic cleanup scripts are run on a daily basis to remove users and cleanup resources.

If there is a need to secure something you have created, please contact the Demo Portal team:



contacthpedemoportal@hpe.com

Demonstration environments

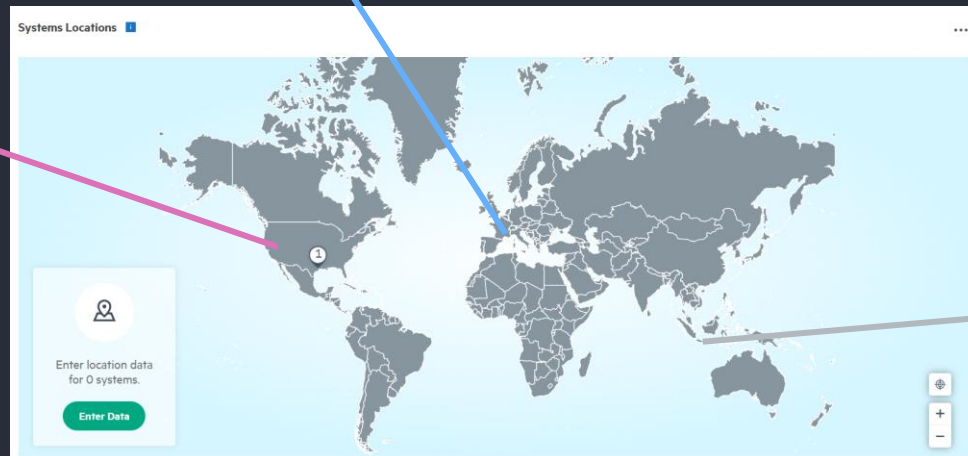
Status of the demonstration platforms September 2025:

A new Private Cloud AI instance has been deployed in Houston (US WEST), in version 1.8.0. It is the preferred platform to demonstrate as it is in the latest version.

The Geneva PCAI instance (EU Central) is deployed in version 1.6, and is still available as a backup platform in case there are issues with the Houston one.

**HPE Private Cloud AI
in the Houston CIC
Region US West**

**HPE Private Cloud AI
in the Geneva CIC
Region EU Central**



A 3rd HPE Private Cloud AI instance is available in the Singapore CIC. It can be booked in a separate demo [here](#)

Demonstration menu

Status of the demonstration platforms September 2025:

- 1. Manage the HPE Private Cloud AI from the HPE GreenLake Platform
- 2. Access HPE AI Essentials to deploy and manage your AI workloads
- 3. Run an imported app :
 - Open WebUI with Test Llama model
 - Open WebUI with RAG (HPE QuickSpecs)
- 4. Deploy an app through Solution Accelerators
- 5. Visual Language Model (VLM) Imported solution / Traffic Analysis demo
- 6. HPE Machine Learning Inference Software MLIS
- 7. PCAI Monitoring in OpsRamp

Geneva EU Central	Houston US West
✓	✓
✓	✓
✓	✓
✓	⊘ Deprecatd in v1.8
✓	✓
✓	✓
✓	✓

1. HPE GreenLake Platform

Access for IT Administrators



1. Accessing the HPE GreenLake Platform

The demonstration is very easy to access directly from your own browser. Once the demonstration booked, you are granted access to the main HPE GreenLake tenant for all demonstrations, the workspace « **DreamCompany** »

You will have access for 2 hours, if you need a longer access, don't hesitate to book the demo again.

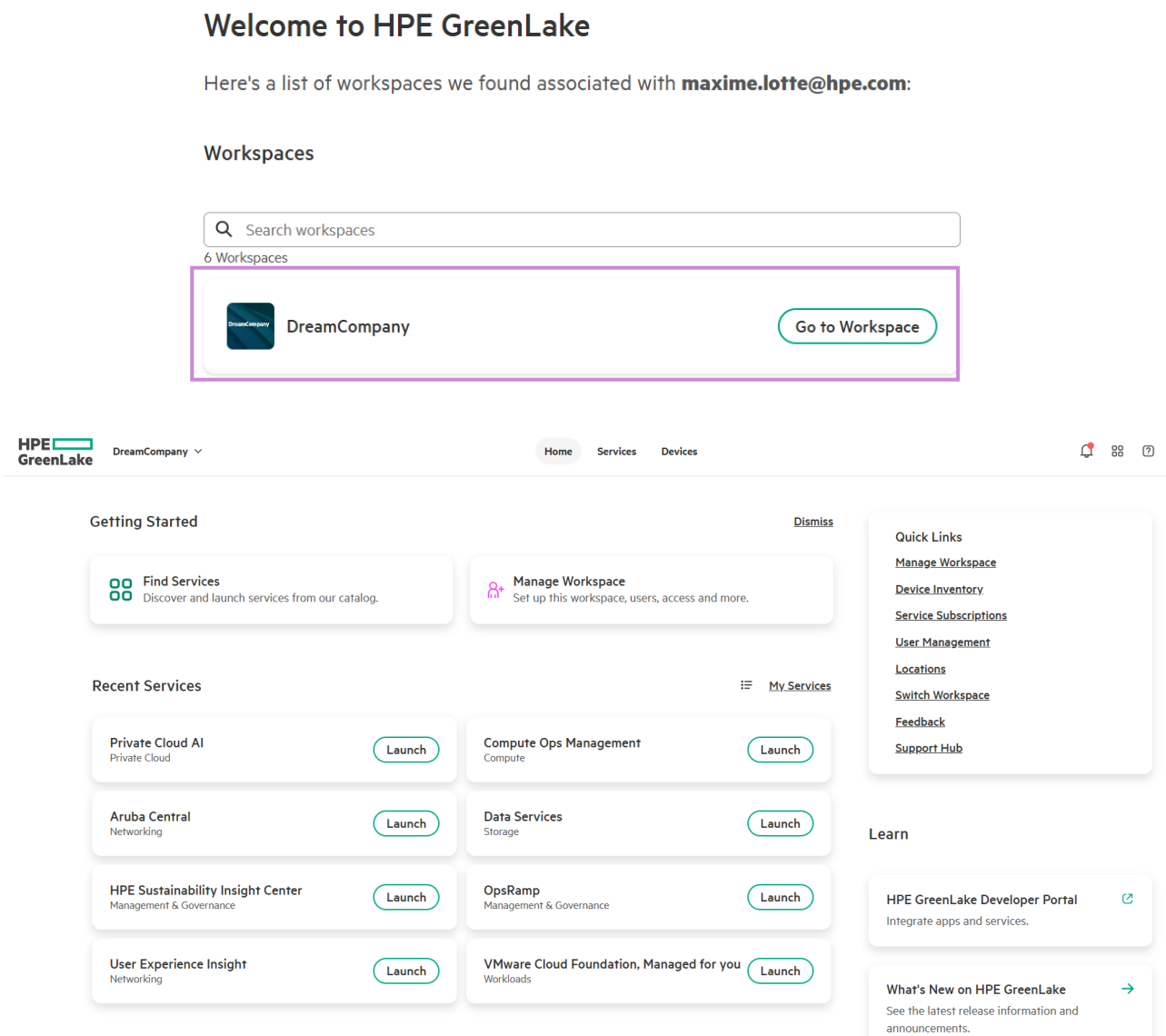
Open your preferred browser to <https://common.cloud.hpe.com/>

Log in with your own email address.

Select the Workspace **DreamCompany**

You are connected to the HPE GreenLake Platform.

Note: this demonstration also gives you access to all other services if you need to demonstrate other features.

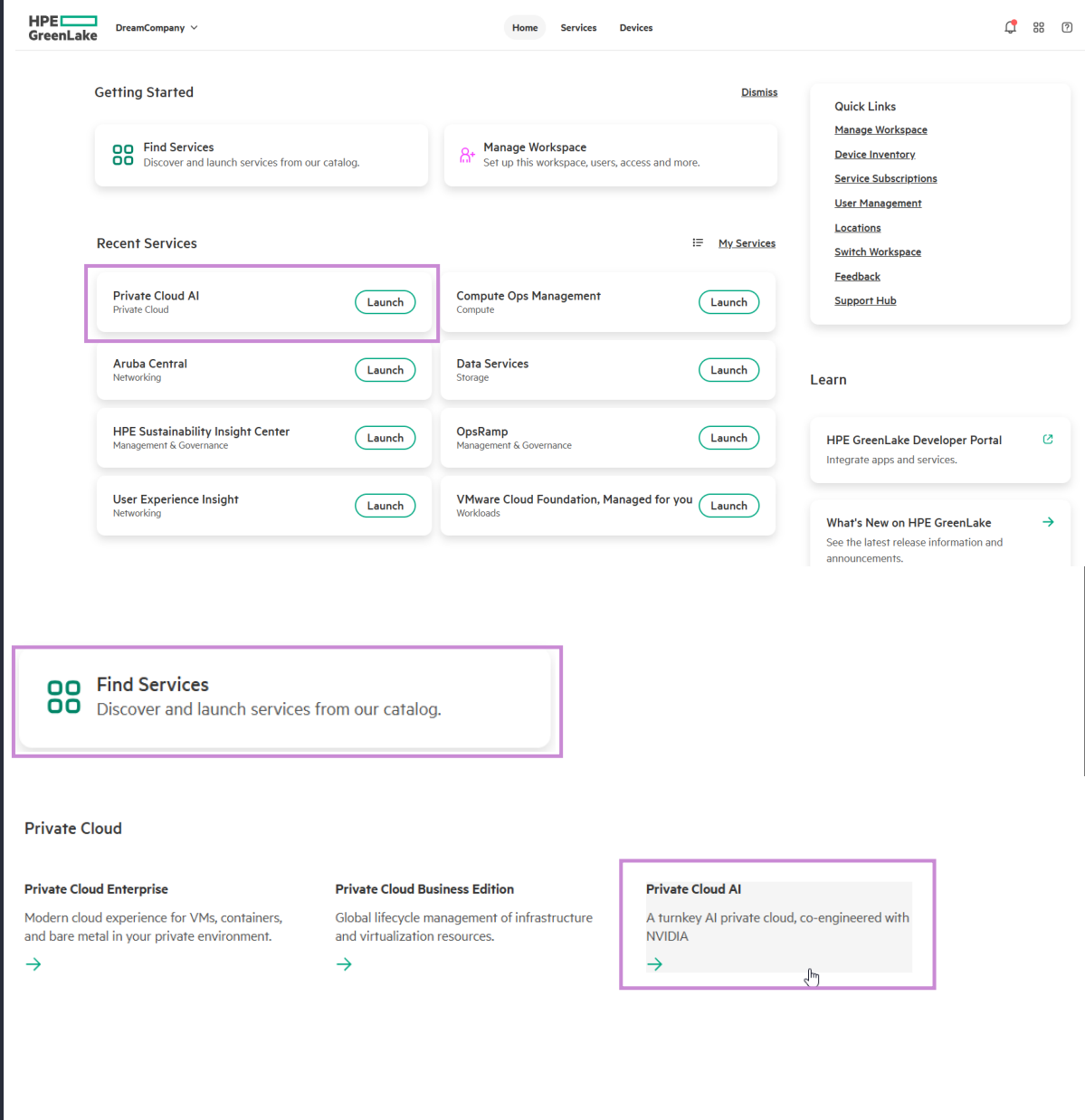


1. Accessing PCAI in the HPE GreenLake Platform

Once there, you can open the **Private Cloud AI Service**.

You can find it:

- In the Recent Services if it was recently used:
Select region US West and click Launch. The Private Cloud AI instance in region EU is the instance always available, with all use-cases deployed and documented.
- If not, you can find it by clicking **Find Services**. Scroll down to the Private Cloud Category, select **Private Cloud AI**, Click Launch, **select region US West** and click Launch again



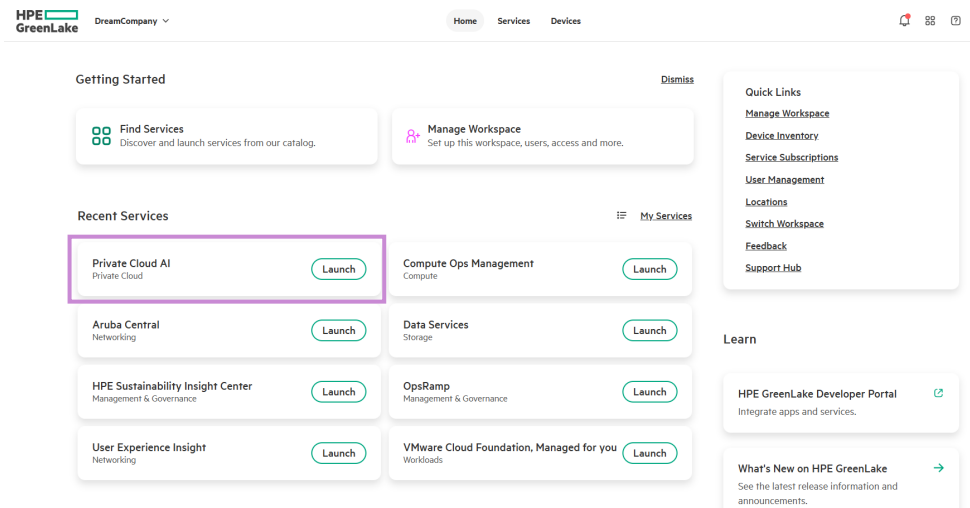
1. Accessing PCAI: Main demo environment / Backup demo environment

Important: In case of an issue with the HPE Private Cloud AI instance in **West US** region (in the HPE CIC in Houston):

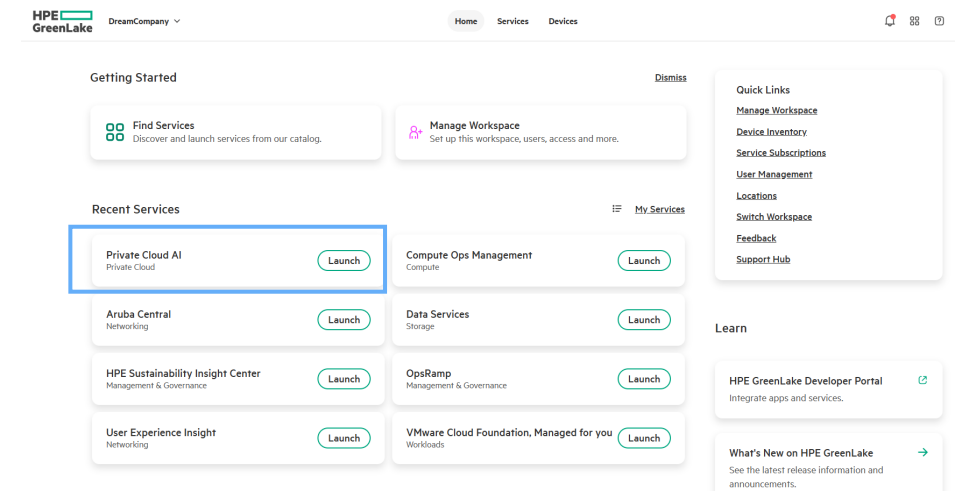
Come back to the HPE GreenLake homepage, and open **Private Cloud AI** in the **EU region** to demonstrate on the HPE Private Cloud AI in Geneva, as a backup demo platform.

The HPE Private Cloud AI in the US West region is deployed in the latest version, and has all use-cases implemented, it should be your preferred platform to demonstrate.

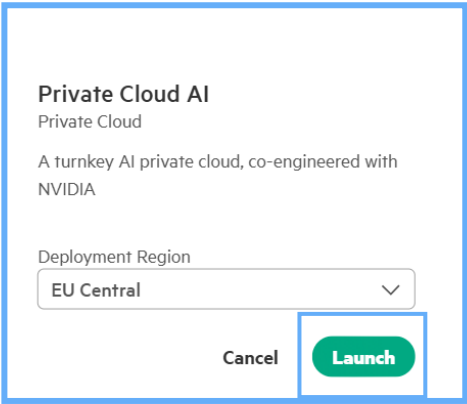
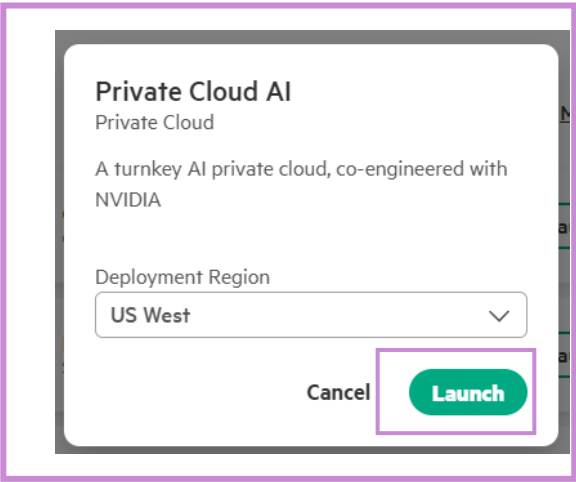
Use the EU region Private Cloud AI instance as a backup in case there is an issue with the US instance.



Main Demo Region



Backup Demo Region



1. Managing PCAI in the HPE GreenLake Platform

Here you have a global view of the different Private Cloud AI platforms you have deployed in the US West region.

For this demo there is 1 Private Cloud AI available, in Houston. You can see it on the map.

Click **1 System** to show this PCAI.

Click the System **hou-pcai.hpecic.net** to open the system.

Data Services Cloud Console | Private Cloud AI

Summary

1 Systems

2 Nodes

8 GPUs

Converged Storage Capacity

1% used959.94 GiB of 98.30 TiB

Issues

No issues!

Performance

GPU

150 MHz0% GPU utilization

GPU Memory

289 MHz0% GPU memory utilization

Systems Locations

Announcements

HPE Private Cloud AI: New features and updates July 21, 2025.

HPE Private Cloud AI: New features and updates May 16, 2025.

Systems

+

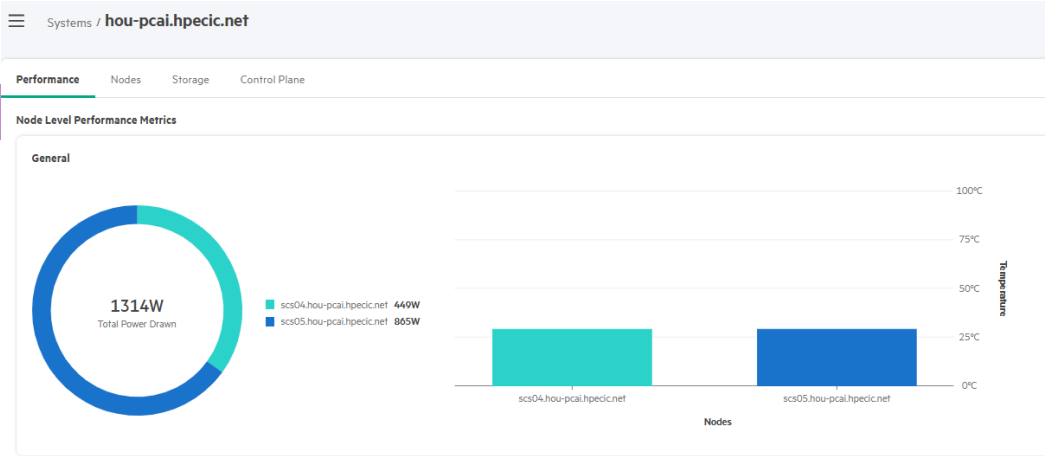
Manage Software Updates

Name	State	Capability	Health	Converged Storage Capacity	Private Cloud AI Instance
☐ hou-pcai.hpecic.net	Online	Baremetal with GPUs	Ok	1%960.15 GiB of 98.30 TiB	Launch

1. Managing PCAI in the HPE GreenLake Platform

Here you will find information about the different parts of the system:

- Performance of the nodes: GPU, CPU usage and power used
- Information and status of the Nodes
- Information and status of the Storage Alletra MP
- Information and status of the Control Plane (3 ProLiant Servers in the PCAI rack)



Systems / hou-pcai.hpecic.net

Performance Nodes Storage Control Plane

Name	Node Health	ILO IP Address	GPUs	OS Image	Credentials
10.5.11.113 ProLiant DL380a Gen11, Hypervisor Host	Ok	10.5.11.88	4	RHEL 8.10 (Outpa)	OS: 7b479a32-4340-11f0-904f-0100a0b330f ILO: bf4e056f-61f7-11f0-bbb3-0100a02854dc
10.5.11.114 ProLiant DL380a Gen11, Hypervisor Host	Ok	10.5.11.89	4	RHEL 8.10 (Outpa)	OS: 7b479a32-4340-11f0-904f-0100a0b330f ILO: bf4e056f-61f7-11f0-bbb3-0100a02854dc

Performance Nodes Storage Control Plane

Details

Name: PCAI

Model: Alletra MP File

Server Serial Number: CZD1W0CAX

Health: Ok

Storage Usage: 1% 960.42 GiB of 98.30 TiB

Performance Nodes Storage Control Plane

Servers

Name	ILO IP Address	Indicator LED	Server Health	Credentials
10.5.11.12 ProLiant DL325 Gen11, Hpe Virtualization	10.5.11.86	Off	Ok	ILO: CZD1W0795-ILO
10.5.11.13 ProLiant DL325 Gen11, Hpe Virtualization	10.5.11.87	Off	Ok	ILO: CZD1W0796-ILO
10.5.11.11 ProLiant DL325 Gen11, Hpe Virtualization	10.5.11.85	Off	Ok	ILO: CZD1W0794-ILO

1. Managing PCAI in the HPE GreenLake Platform

You also have an Actions button in the top right corner, allowing you to check for updates to the Software stack.

Here the platform is up to date. Hovering the version will show more details. It mentions on this one 'Non-Compliant' as it is a demonstration instance of PC AI.

Do not trigger updates on the PC AI environment as this is a shared demonstration platform.

Systems / hou-pcai.hpecic.net

PerformanceNodesStorageControl Plane

Node Level Performance Metrics

General

100°C

Actions

Manage Software Updates

Manage Software Updates

Manage Software Updates: hou-pcai.hpecic.net

Use the table below to maintain the software for hou-pcai.hpecic.net.

Refresh Software VersionsPrecheckDownloadUpdate

Name	Current Version	Latest Version	Update Status
hou-pcai.hpecic.net	Non-Compliant	-	Up To Date

Version Non-Compliant

HPE AI Essentials 1.8.0

HPE VME Host 1.0.0.31

HPE VME Manager 8.0.6.3

HPE GreenLake for File Storage OS 3.2.0-154

Data Services Connector 102.0.0.3

View Details

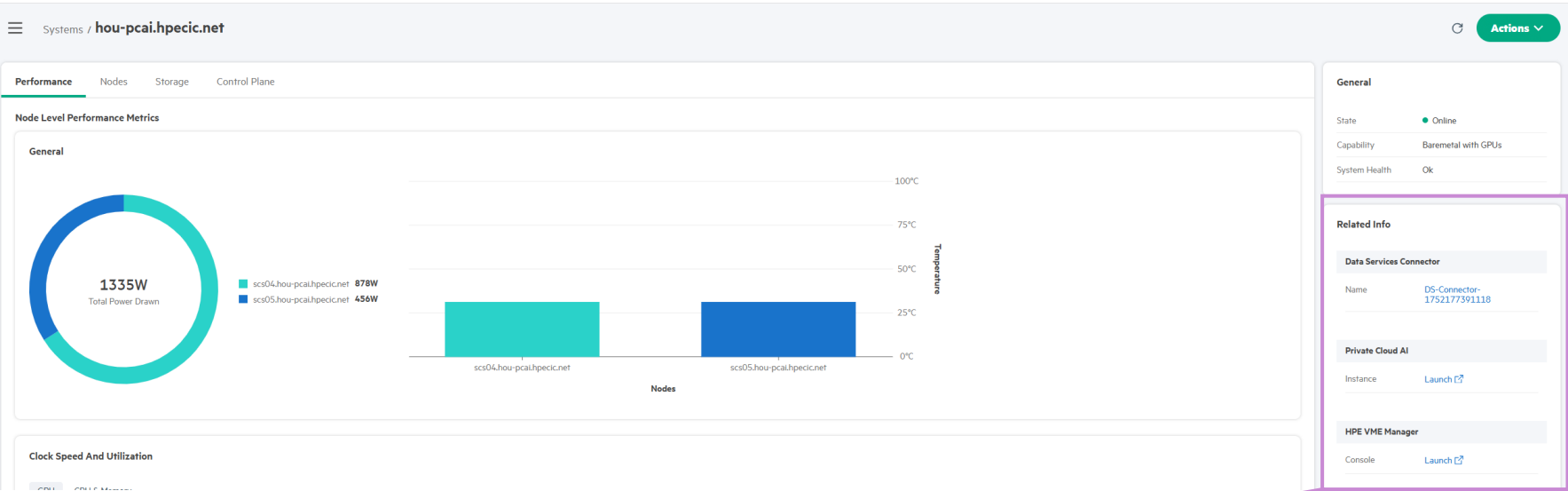
Non-Compliant

Do not trigger Updates !

1. Accessing the PCAI endpoints from the HPE GreenLake Platform

Finally, the HPE GreenLake Platform shows you the main endpoints needed for the PC AI Platform. You can find them on the right side:

- **Private Cloud AI:** Will launch the **HPE AI Essentials** dashboard, to start deploying AI workloads on the platform. This is described in details in the next part of the demo guide. Click this link to continue the demonstration.
- **HPE VME Manager:** This will open the HPE VME console to manage the Control Plane. This is not covered in this version of the demonstration, as it is not central to the PC AI Solution. But it is interesting to see that HPE VM Essentials is running the Private Cloud AI Control Plane.



Related Info

Data Services Connector

Name `DS-Connector-1752177391118`

Private Cloud AI

Instance `Launch` (with an external link icon)

HPE VME Manager

Console `Launch` (with an external link icon)

2. HPE AI Essentials

Access for IT Operators and Data Scientists



2. Accessing HPE AI Essentials

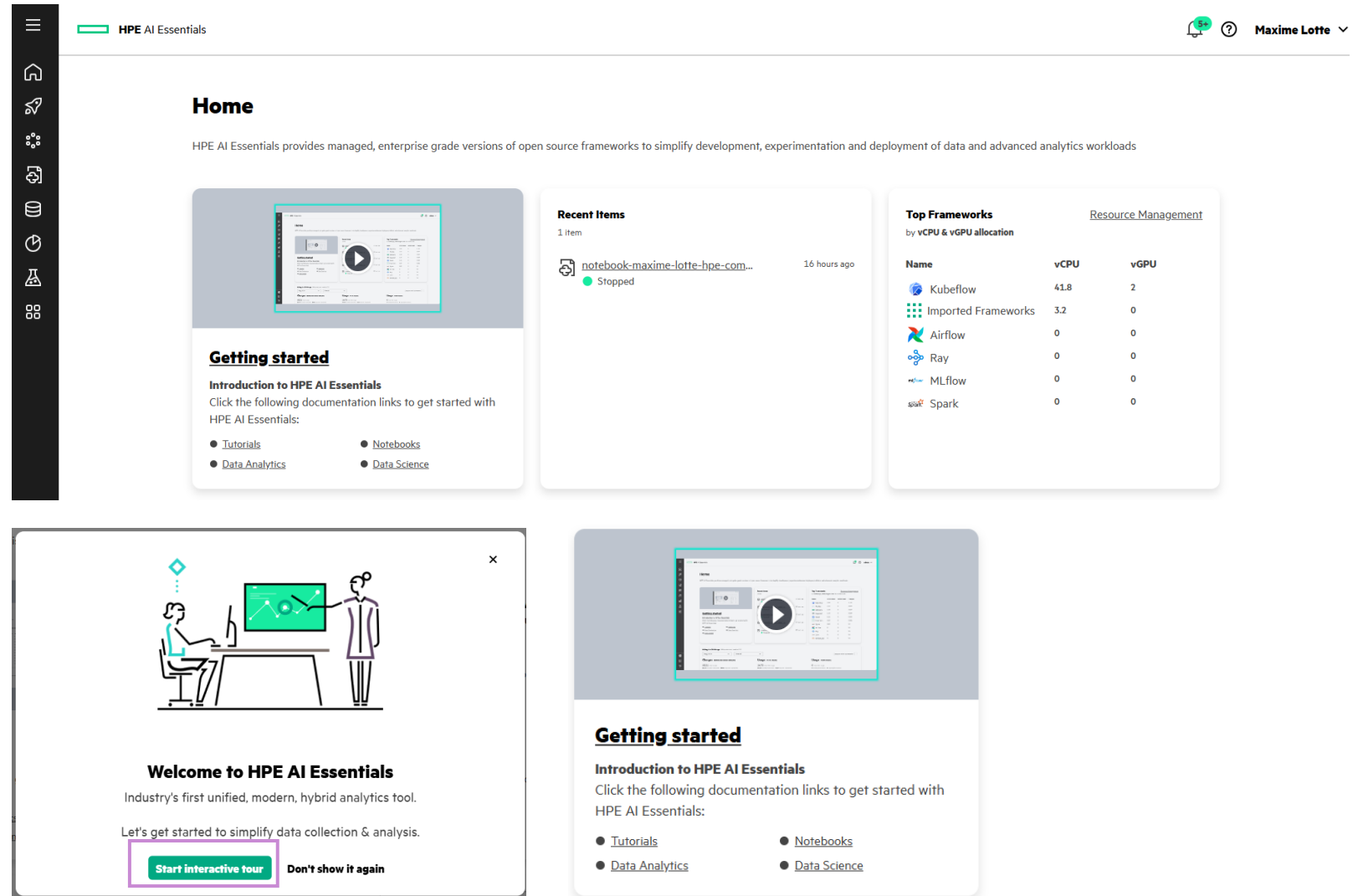
Click the **Private Cloud AI** link in HPE GreenLake. Another tab opens, and you are connected with SSO to HPE AI Essentials part of HPE Private Cloud AI.

Note: You may experience a sporadic SSO issue when connecting to HPE AI Essentials. Just close the browser tab and try again if this issue happens.

The Interactive Tour will help you navigate the different menus, and the video **Introduction to HPE AI Essentials** provides a quick explanation of all features proposed by the platform.

If it is your first time connecting to HPE PC AI, the video can be very useful.

We will also have a tour of these features in the next parts of this demonstration guide.



2. Quick tour of HPE AI Essentials

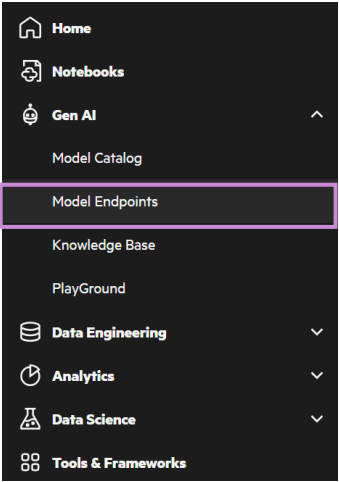
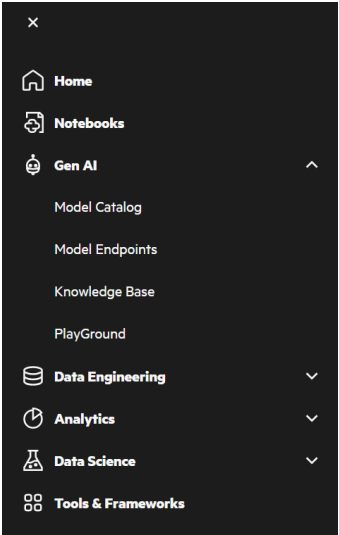
Let’s start the tour of HPE AI Essentials, to explain all capabilities enabled by the platform.

To start, go to **Gen AI, Model Catalog**.

You can see there the pre-installed models that can be quickly deployed on Private Cloud AI to server AI applications. You can see for example the llama-3.1 model.

Please do not stop or delete models that start with « do-not-delete »

Next you can go to **Model Endpoints**, where you can see the deployed models, serving the AI Applications.



Model Catalog

All Models

5 items

NVIDIA AI Enterprise (5)

[See All](#)

nv-rerankqa-mistral-4b-v3
Version 1 | Staged

Registry AI-Essentials-NGC-Catalog

Model Category reranker

text-embedding information-retrieval +3

Deploy

llama-3.1-8b-instruct
Version 1 | Staged

Registry AI-Essentials-NGC-Catalog

Model Category llm

text-generation natural-language-processing +3

Deploy

llama-3.1-70b-instruct
Version 1 | Staged**Registry** AI-Essentials-NGC-Catalog**Model Category** llm

natural-language-processing +4

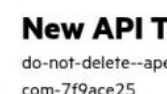
Deploy

Model Endpoints

8 items

Name	Namespace	Status	Model Catalog	Endpoint	Modified On	Actions
do-not-delete--apertus	hugo-hpe-com-7f9ace25	Ready	DO-NOT-DELETE--Apertus.v1	https://do-not-delete--apertus-predictor-hugo-hpe-com-7f9ace25.ai-application.hou-pcai.hpecic.net	09/09/2025 02:57:33 PM	
do-not-delete--gpt-oss-20b	hugo-hpe-com-7f9ace25	Ready	DO-NOT-DELETE--openai-gpt-oss-20bv4	https://do-not-delete--gpt-oss-20b-predictor-hugo-hpe-com-7f9ace25.ai-application.hou-pcai.hpecic.net	08/19/2025 02:25:32 PM	
do-not-delete--llama-3-1-8b	hugo-hpe-com-7f9ace25	Ready	llama-3.1-8b-instruct.v2	https://do-not-delete--llama-3-1-8b-predictor-hugo-hpe-com-7f9ace25.ai-application.hou-pcai.hpecic.net	08/19/2025 02:19:49 PM	
do-not-delete--embed	hugo-hpe-com-7f9ace25	Ready	llama-3.1-8b-instruct.v2	https://do-not-delete--embed-predictor-hugo-hpe-com-7f9ace25.ai-application.hou-pcai.hpecic.net	09/08/2025 11:19:53 AM	
embed	karson-lin-hpe--fe48973e	Paused	nv-embedqa-e5-v5		09/10/2025 08:16:59 AM	
embedder-hpe-explore	dennis-geisse-h-0249149f	Paused	nv-embedqa-e5-v5		09/09/2025 08:55:32 PM	
hpeexplore	dennis-geisse-h-0249149f	Paused	HPE-Explore		09/09/2025 08:55:23 PM	
taide	karson-lin-hpe--fe48973e	Paused	taide		09/10/2025 08:54:21 AM	

Generate API Token: Token needed for applications to identify when accessing the model.



New API Token

do-not-delete--apertus-predictor-hugo-hpe-com-7f9ace25

Token Lifetime*

Enter the number of days to set token lifetime.

30

Close Generate API Token

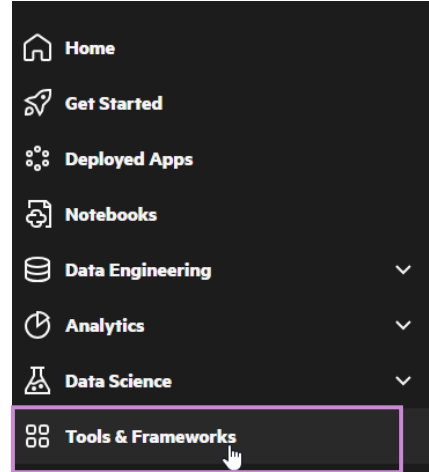
2. Quick tour of HPE AI Essentials

HPE AI Essentials also includes a curated set of Open Source applications to enable enterprises to build, deploy & operationalize analytics in GenAI applications.

These Tools can be found in **Tools & Frameworks**, in the different tabs.

Data engineers have tools to design and implement data pipelines for ML and AI workloads, including

- **Apache Airflow** to connect a data set and automate data pipelines.
- **Apache Spark** to run large scale analytics workloads.



Tools & Frameworks

[Import Framework](#)

Data Engineering

Analytics

Data Science



Airflow

Version 2.10.3 | ● Ready

Workflow engine for scheduled batch ingestion of data from a wide variety of external data sources

Endpoint <https://airflow.ai-application.hou-pcai.hpecic.net>

Chart Version 1.6.2

Open



EzPresto

Version 1.8.0-prestodb-0.289 | ● Ready

Distributed query engine designed for analytic queries on data of any size

Endpoint <https://ezprestodb.ai-application.hou-pcai.hpecic.net>

Chart Version 1.8.0

Open

Tools & Frameworks

[Import Framework](#)

Data Engineering

Analytics

Data Science



Spark History Server

Version 3.5.2.3 | ● Ready

Monitoring service for spark applications

Endpoint <https://sparkhs.ai-application.hou-pcai.hpecic.net>

Chart Version 0.6.19

Open



Spark Operator

Version 1.3.8-202503251142 | ● Ready

Controller for non-interactive spark applications

Chart Version 1.3.33

Open

2. Quick tour of HPE AI Essentials

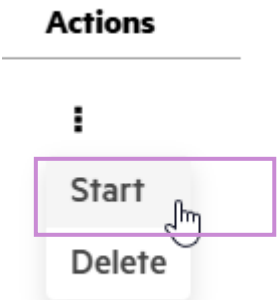
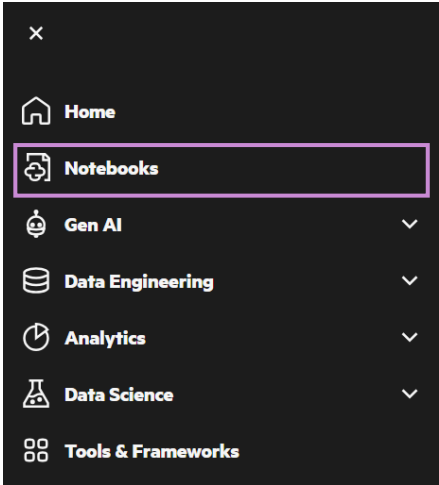
Notebooks:

Every PC AI User is allocated a secure JupyterLab environment to create notebooks, develop scripts, run experiments, and deploy applications. All within their own private workspace

Notebooks automatically connect to the full suite of integrated tools and frameworks, and come pre-packaged with additional popular frameworks.

Notebooks also include pre-packaged tutorials, with predictive analysis examples to use as references when developing models.

You may need to start your Notebook to be able to open it, you can do this in **Actions**, **Start**



Notebook Servers

New Notebook Server

Search

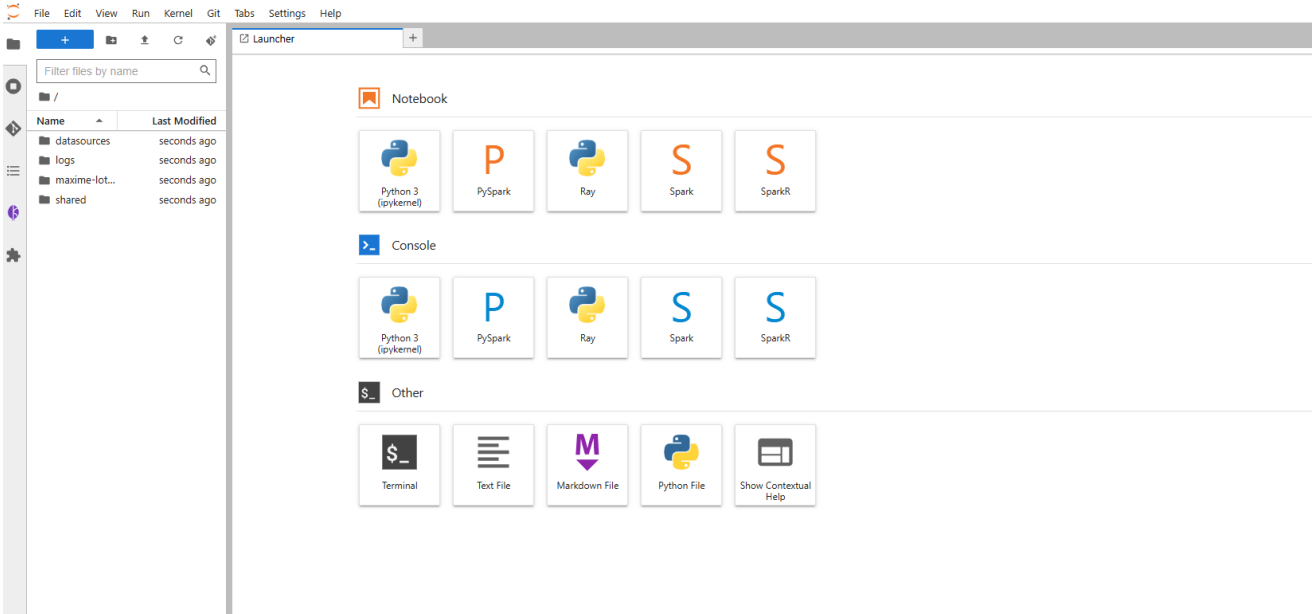
Filter

Grid

Delete

1 item

☐	Name	Type	Status	Image	CPUs	Memory	GPUs	Actions
☐	notebook-maxime-lotte-hpe-com-66e986b2	jupyter	Running	10.6.127.47/ezmeral-common/gcr.io/mapr-2527111/kubeflow/notebooks/jupyter-tensorflow-fullzeua-1.5.2-175c1e34	1	2Gi	N/A	



2. Quick tour of HPE AI Essentials

Data Sources:

HPE AI Essentials utilizes HPE GreenLake for File Storage as the default File & Object storage.

However, PC AI Administrators can also connect external Data Sources, including relational Data bases, S3 object Storage, and Data Volumes in File Storage.

Once Data Sources are connected, an administrator can grant permissions on data through Identity & Access Management

Home

Notebooks

Gen AI

Data Engineering

Data Sources

Data Catalog

Query Editor

Cached Assets

Airflow Pipelines

Analytics

Data Science

Tools & Frameworks

Administration

Settings

Resource Management

Identity & Access Management

Audit Logs

Contact Us

Guided Tour

Data Sources

Object Store DataData Volumes

Search existing data sources

Add New Data Source

1 Data Source

local-s3

HPE GreenLake for File Storag...

Ready

Seamlessly run any app without compromise from edge to cloud with a cloud experience for every workload.

http://local-s3-service.ezdata-system.svc.cluster.local:30000

Add New Data Source

Cancel

Objects Data Sources

HPE Ezmeral Data Fabric S3

HPE Ezmeral Data Fabric Software is SaaS-based data foundation for analytics and AI across hybrid cloud to seamlessly access, analyze, and govern data globally.

Add HPE Ezmeral Data Fabric S3

Amazon S3

Amazon Simple Storage Service (Amazon S3) is an object storage service offering industry-leading scalability, data availability, security, and performance from Amazon.

Add Amazon S3

MinIO S3

MinIO is a high-performance, S3 compatible object store. It is built for large scale AI/ML, data lake and database workloads.

Add MinIO S3

HPE GreenLake for File Storage S3

Seamlessly run any app without compromise from edge to cloud with a cloud experience for every workload.

Add HPE GreenLake for File Storage S3

HPE AI Essentials

Maxime Lotte

Identity and Access Management

Users

Search

9 items

Delete

Add Privileges

☐	Name	Username	Email	Roles	Status	Actions
☐	Hugo Carigny	hugo-hpe.com	hugo@hpe.com	admin, member	Ready	
☐	Jean-Marc ANDRE	jean-marc.andre@hpe.com	jean-marc.andre@hpe.com	admin, member	Ready	
☐	Jordan Nanos	jordan.nanos@hpe.com	jordan.nanos@hpe.com	admin, member	Ready	
☐	Karim Wackenheim	karim.wackenheim@hpe.com	karim.wackenheim@hpe.com	admin, member	Ready	
☐	Laurent Wojcik	laurent.wojciak@hpe.com	laurent.wojciak@hpe.com	admin, member	Ready	
☐	Maxime Lotte	maxime.lotte@hpe.com	maxime.lotte@hpe.com	admin, member	Ready	
☐	Philippe Garnier	garnier-hpe.com	garnier@hpe.com	admin, member	Ready	

2. Quick tour of HPE AI Essentials

Administration:

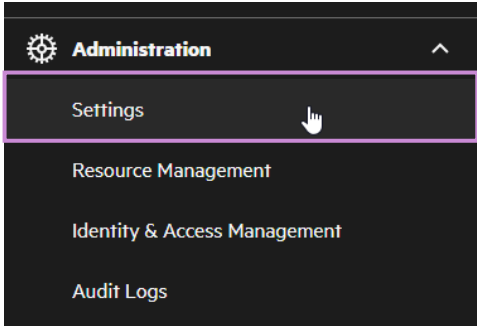
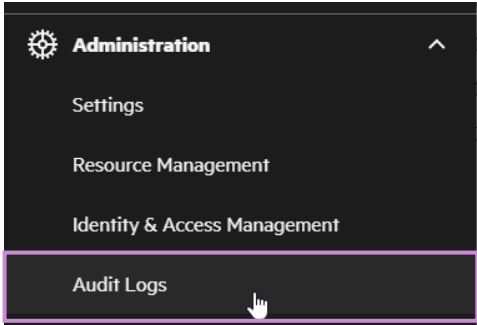
Administrators can audit user operations and manage GPU Resources.

On the Settings page, Admins can view cluster config, add endpoints, and modify timeouts.

On the **Tools & Frameworks** tab, administrators can configure the installed tools & frameworks, or install the included tools and frameworks that are not installed by default, such as Superset and Feast.

This concludes the HPE AI Essentials overview.

Let's move to the next step of the demonstration.



Cluster Configurations

Status	● Ready
Installation Name	houpcaihpe
Domain Name	ai-application.hou-pcai.hpecic.net
vCPU	256
vGPU	8
GPU Configuration	Whole
High Availability	Enabled

Audit Logs

Search

1 hour

Download Logs

9 items

Time	Action	Message	User	Status
11/08/2024 11:49:29 AM	User login	user maxime.lotte-hpe.com logged in	maxime.lotte-hpe.com	● Success
11/08/2024 11:44:48 AM	User login	user garnier-hpe.com logged in	garnier-hpe.com	● Success
11/08/2024 11:42:56 AM	User login	user laurent.wojkowiak-hpe.com logged in	laurent.wojkowiak-hpe.com	● Success
11/08/2024 11:41:52 AM	User login	user karim.wackenheim-hpe.com logged in	karim.wackenheim-hpe.com	● Success
11/08/2024 11:29:26 AM	User login	user maxime.lotte-hpe.com logged in	maxime.lotte-hpe.com	● Success
11/08/2024 11:21:38 AM	User login	user karim.wackenheim-hpe.com logged in	karim.wackenheim-hpe.com	● Success
11/08/2024 11:19:36 AM	User login	user garnier-hpe.com logged in	garnier-hpe.com	● Success
11/08/2024 11:09:09 AM	User login	user maxime.lotte-hpe.com logged in	maxime.lotte-hpe.com	● Success
11/08/2024 10:59:23 AM	User login	user garnier-hpe.com logged in	garnier-hpe.com	● Success

Settings

Cluster	Activation Key	Configurations	Tools & Frameworks
Search			
11 items			
	Livy	Version 0.8.0.9 ● Initialized	
Analytics			
	Spark History Server	Version 3.5.2.3 ● Ready	
Analytics			
	Spark Operator	Version 1.3.8-202503251142 ● Ready	
Analytics			
	Airflow	Version 2.10.3 ● Ready	
Data Engineering			
	EzPresto	Version 1.8.0-prestodb-0.289 ● Ready	
Data Engineering			
	Superset	Version 4.1.1 ● Initialized	
Data Engineering			
	Feast	Version 0.41.3 ● Initialized	
Data Science			
	HPE MLIS	Version 1.4.0 ● Ready	
Data Science			

3. Run an imported application
to start interacting with models



3. Run an imported application

It is also possible in PC AI to import **your own developed application**, to run on the platform.

An application has been already imported, to demonstrate this feature.

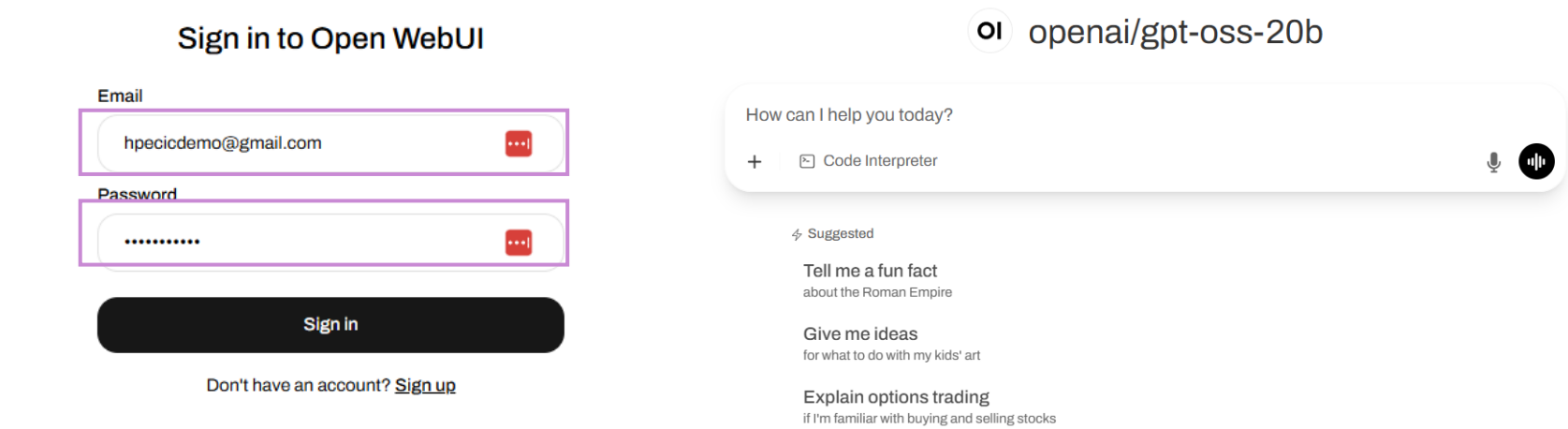
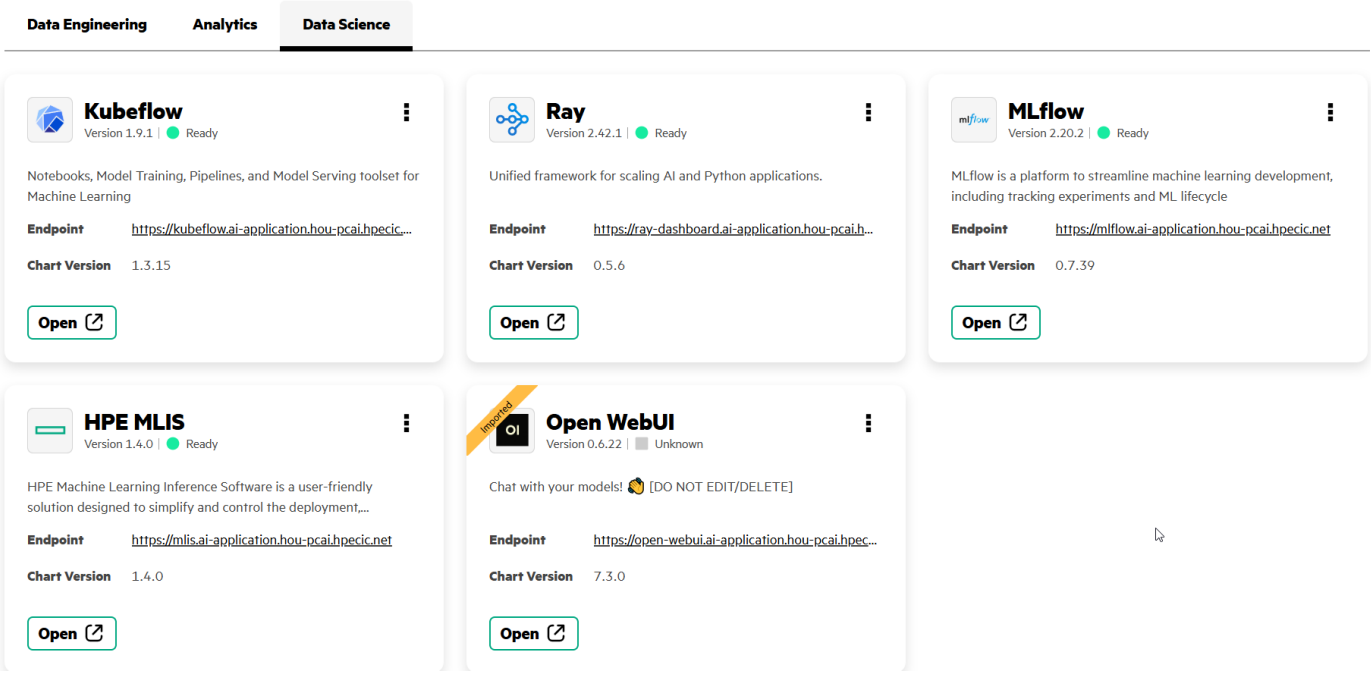
Go back to Tools & Frameworks, go to **Data Science**. You will find there the application **Open WebUI**.

Click Open, and log in the application with:

hpecicdemo@gmail.com

ITAdmin1234

You are logged in the application, that is a Chat-GPT like engine, used to test the different LLM Models.



3. Run an imported application

In the top left corner, you can select which model you would like to interact with. Here you can see multiple options available.

Apertus is the model pushed by Swiss AI, trained on the Alps Supercomputer provided by HPE to Swiss AI.

GPT-oss: is the Chat-GPT model hosted on-premises. This model has reasoning capabilities and is unique in that regard.

Llama-3.2-Vision: Is a multi-modal model and supports text and image as prompts.

Clicking the + button will allow you to add a second model, and to see the responses that 2 models provide for the same prompt:

Here we see the result for 2 models (OpenAI and Llama) for the same prompt:

openai/gpt-oss-20b +

Search a model

All External

swiss-ai/Apertus-8B-Instruct-2509

openai/gpt-oss-20b ✓

meta/llama-3.1-8b-instruct

HPE HPE Quickspecs

Temporary Chat

openai/gpt-oss-20b +

meta/llama-3.1-8b-instruct -

Explain options trading in simple terms if I'm familiar with buying and selling stocks.

openai/gpt-oss-20b

Thought for less than a second

Options trading in plain language

TERM	WHAT YOU ALREADY KNOW (STOCKS)	WHAT AN OPTION IS
Stock	You own a piece of a company that you can buy or sell anytime.	The price of the company's share is the "underlying" that an option references.
Buy a share	You pay the current price and own the stock.	You <i>buy</i> a call option = right (but not the obligation) to buy the stock at a set price later.
Sell a share	You get the money at the current price.	You <i>sell</i> a put option = right (but not the obligation) for someone to sell you the stock at a set price later.
Premium	You pay nothing extra when buying a share.	The person who writes (sells) the option gets an upfront fee, the premium. That's the cost of the option.
Strike price	No equivalent.	The price at which the option holder can exercise their right.

meta/llama-3.1-8b-instruct

Let's break down options trading in simple terms if you're already familiar with buying and selling stocks:

Options Trading

Options trading is like buying or selling a contract that gives you the right, but not the obligation, to buy or sell a stock at a predetermined price (strike price) at a specific time (expiration date). Think of it like a game of leverage, where you can bet on the price movement of a stock without actually owning it.

Key Players

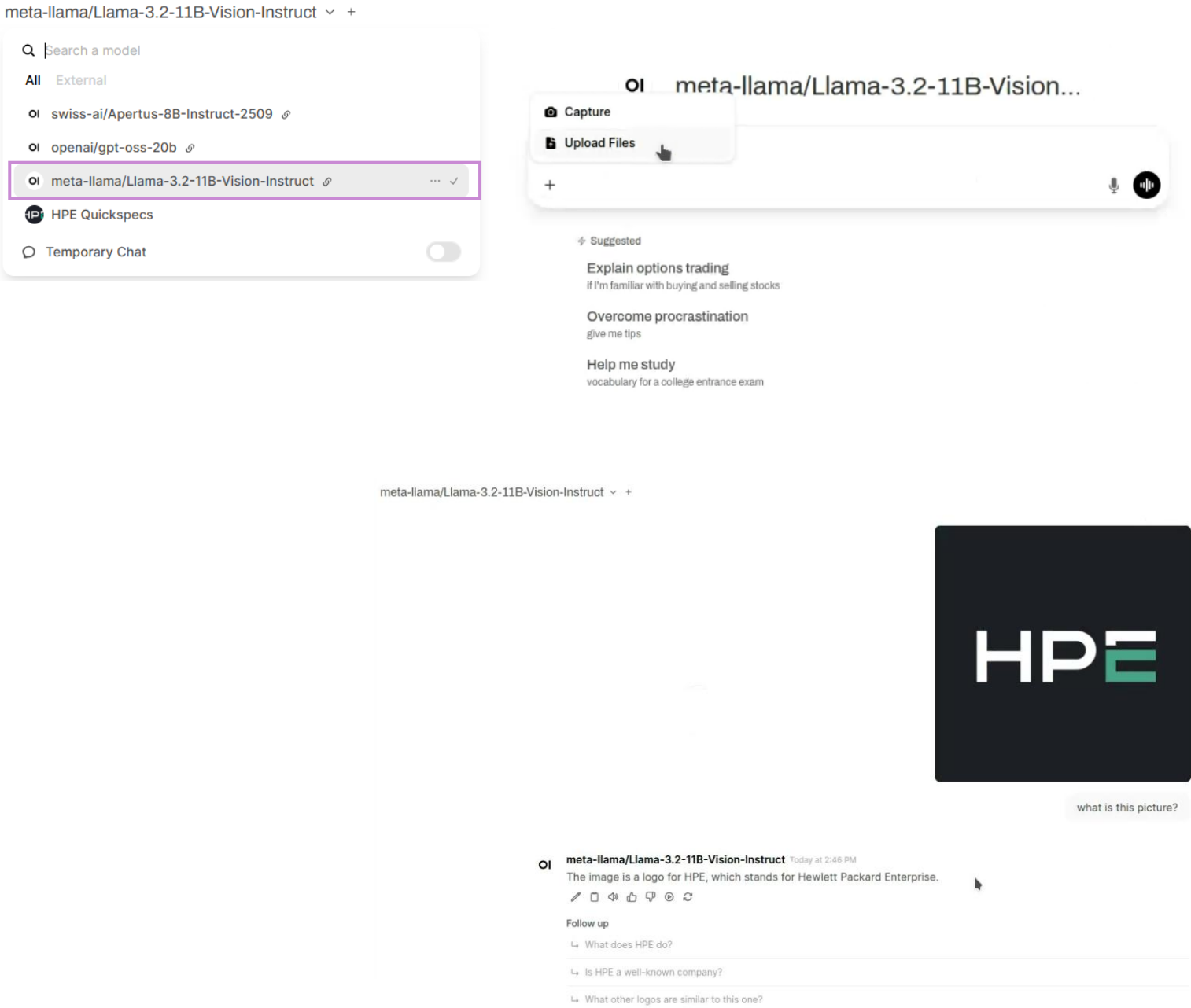
- Underlying Asset:** The stock you're trading options on. For example, Apple Inc. (AAPL).
- Option Buyer (Long):** The person who buys the option contract, hoping the stock price will move in their favor.
- Option Seller (Short):** The person who sells the

Send a Message

3. Run an imported application

Another test you can run is the Llama-3.2 model with Vision. This model will allow you to upload a file to analyze it.

Upload a file of your choice, here for example the HPE logo, and you can start a conversation from here as you can see in the screenshot.



3. Run an imported application

Finally, you can the HPE Quickspecs, to interact with a model enhanced with RAG, with specific information on HPE products & technology.

Here the answers will leverage the additional context provided, and the model will mention the source for the answer that was formed.

HPE Quickspecs

Set as default



HPE Quickspecs

Ask questions about HPE products!

How can I help you today?

+



what is HPE Slingshot ?



HPE Quickspecs

HPE Slingshot is a high-performance interconnect solution from Hewlett Packard Enterprise (HPE). It's designed for High-Performance Computing (HPC) and Artificial Intelligence/Machine Learning (AI/ML) clusters. According to the provided context, HPE Slingshot enhances realized performance on tightly coupled HPC and AI workloads by innovative congestion management, reducing tail latency and run-to-run execution time variability.

The solution includes standalone HPE Slingshot components, systems, and software for HPC and AI/ML clusters based on HPE Apollo, HPE ProLiant, and HPE Cray XD server systems.

- 1 HPE Slingshot QuickSpecs.pdf 2 HPE Slingshot Monitoring Software QuickSpecs.pdf



Follow up

- ↳ How does HPE Slingshot improve performance over traditional HPC networking solutions?
- ↳ Can HPE Slingshot be integrated with existing HPC or AI/ML infrastructure, or is it a replacement?
- ↳ Are there any specific configurations or use cases where HPE Slingshot is particularly beneficial?

4. Additional demonstration use-cases



4.1 Traffic Analysis

A new application has been deployed to demonstrate AI in action on HPE PC AI.

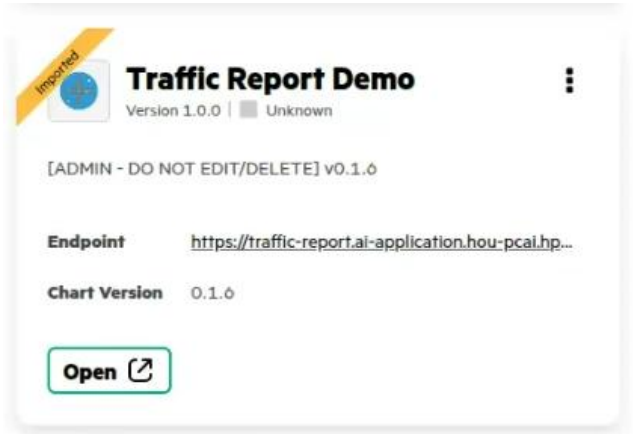
This demo is the **Traffic Report Demo**, that you can open from the **Tools & Frameworks, Data Science** :

Here we leverage 2 models to achieve the objects identification in the image, and then the analysis of the traffic.

Model YOLO v8: defines bounding boxes around identified elements, with a confidence index on what has been identified.

Llama-Multimodal: we send to this model the source image and the output from the YOLO model. And we ask the Llama model to analyse and describe the model.

Let's see this in action.



Smart Traffic Report

Powered by HPE AI Solutions

Upload an image or video to analyze traffic conditions using YOLO detection and Qwen2.5-VL analysis.

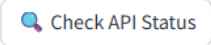


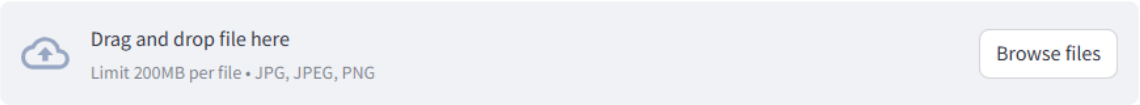
Image Analysis Video Analysis Settings

Image Analysis

Upload Traffic Image

Results

Choose an image...



4.1 Traffic Analysis

You will need pictures to upload to the model.

You can easily download 2 images that work well for this demo, here:

<https://hdpग्vaanimatedemos.hpecic.net/images/659362.jpeg>

https://hdpग्vaanimatedemos.hpecic.net/images/14951887_061424-wls-i355-rollover-crash-10p-vo-vid.jpg

Save the images to your PC.

Then you can upload the image.

First you will see the result of the Image objects recognition by the YOLO model, green rectangles with the confidence index for each object detected.

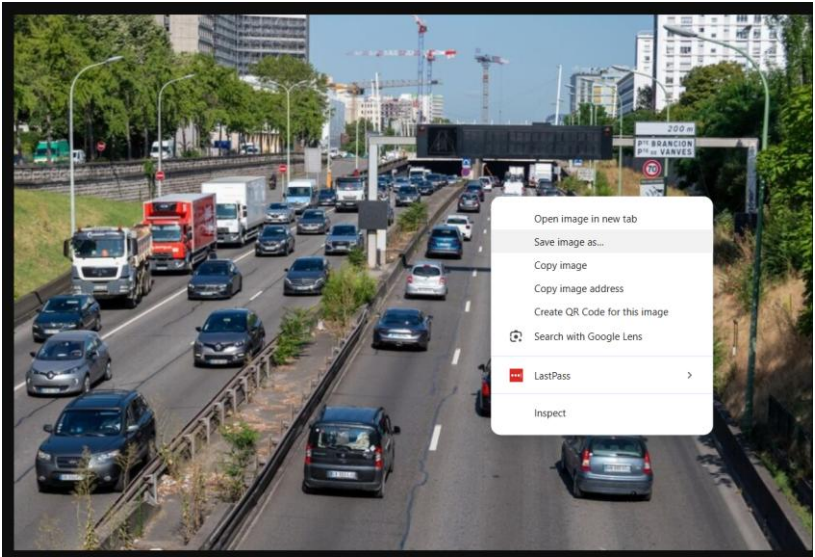


Image AnalysisVideo AnalysisSettings

Image Analysis

Upload Traffic Image

Choose an image...

Drag and drop file here
Limit 200MB per file • JPG, JPEG, PNG

Browse files

659362.jpeg201.2KB

×

Analyze Image

Results

4.1 Traffic Analysis

Below the picture, you will see the analysis of the picture, the AI report, formed by the second model, the Llama model.

You can test as well with the second picture, the results are very interesting as well, with a Safety Assessment and Recommendations.

Detected Vehicles

- Car: 21
- Truck: 5

AI Report

- Traffic Status: Normal
- Traffic Flow: Moderate
- Incident Analysis: No incidents or unusual situations observed.
- Safety Assessment: No safety concerns or violations observed.
- Recommendations: Maintain normal traffic flow.

Image Analysis

Upload Traffic Image

Choose an image...

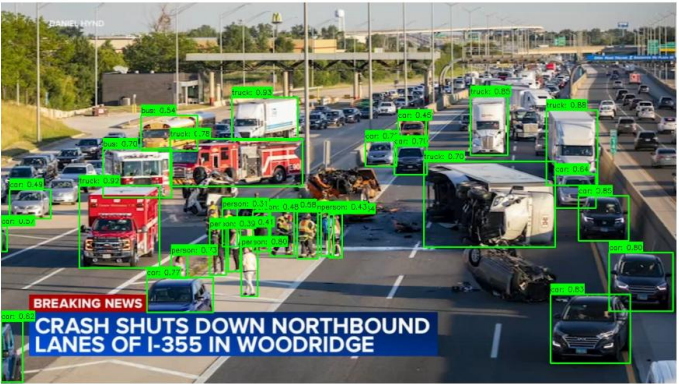
 Drag and drop file here
Limit 200MB per file • JPG, JPEG, PNG

Browse files

 14951887_061424-wls-i355-rollover-crash-10p-vo-vid.jpg 0.6MB

Analyze Image

Results



Detected Vehicles

- Truck: 6
- Car: 11
- Person: 10
- Bus: 2

AI Report

- Based on the provided image, here is a concise analysis of the traffic scene:
- Traffic Status: ABNORMAL
- Traffic Flow: HEAVY
- Incident Analysis: A multi-vehicle crash has occurred in the northbound lanes of I-355 in Woodridge, involving 6 trucks, 11 cars, and 2 buses. The incident has caused significant disruptions to traffic flow.
- Safety Assessment: There is a high risk of injury or fatality due to the severity of the crash and the number of vehicles involved.
- Recommendations: Emergency services should be dispatched immediately to assist with the incident, and traffic authorities should consider diverting traffic around the affected area to minimize further congestion and ensure public safety.

4. HPE PC AI monitoring in OpsRamp



4. Monitor PCAI with OpsRamp

Accessible from HPE GreenLake Platform, the OpsRamp service provides IT Operations and Monitoring capabilities.

Select OpsRamp and launch

You are connected to OpsRamp as a Partner.

A partner can manage multiple clients.

Go on the top menu

Select DreamCompany client

Recent Services

My Services

Private Cloud AI Private Cloud	Launch	OpsRamp Management & Governance	Launch
Compute Ops Management Compute	Launch	Consumption Analytics Management & Governance	Launch
HPE Sustainability Insight Center Management & Governance	Launch	HPE Aruba Networking Central Networking	Launch
Data Services Storage	Launch		

eu-pod1.opsramp.com/dashboard.do?action=loadDashboard

Partner GLP - DreamCompany

Dashboards Infrastructure Automation Configuration Management Command Center Reports Setup

PartnerDefaultDashboard

Partner dashboard

Global Assets

309 Total

195 Other
39 Server
29 Storage
16 Server >> Windows
14 Server >> Linux
16 All Other Resource ...

Device Availability

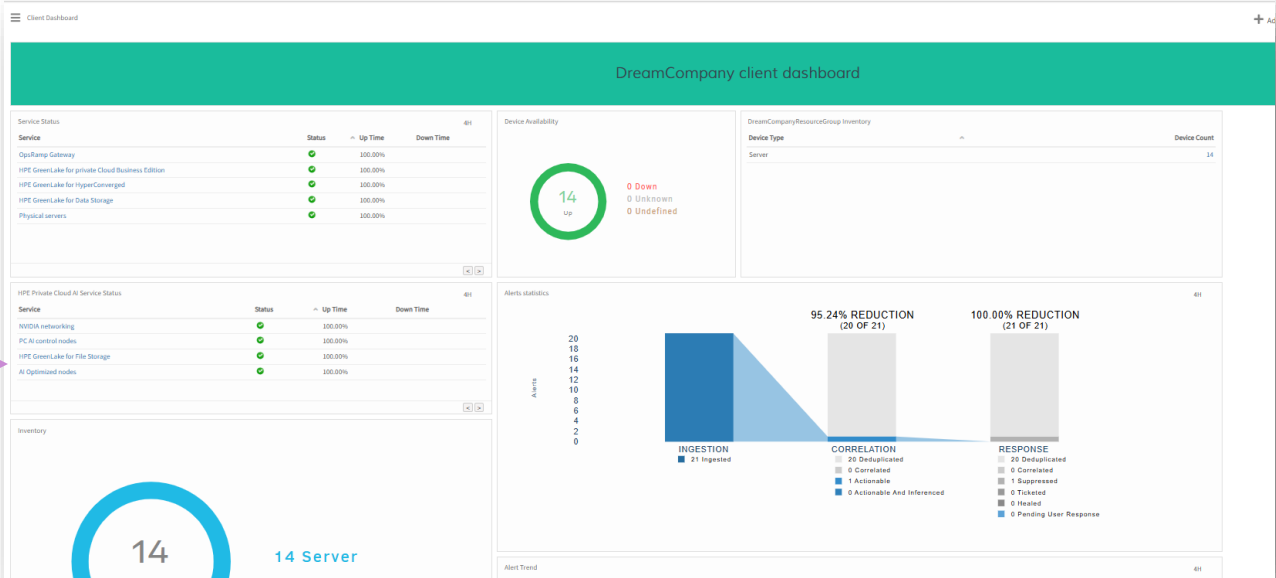
185 Up

6 Down
72 Unknown
40 Undefined

This Partner can see multiple clients. Go on top menu, Clients / DreamCompany dashboard...

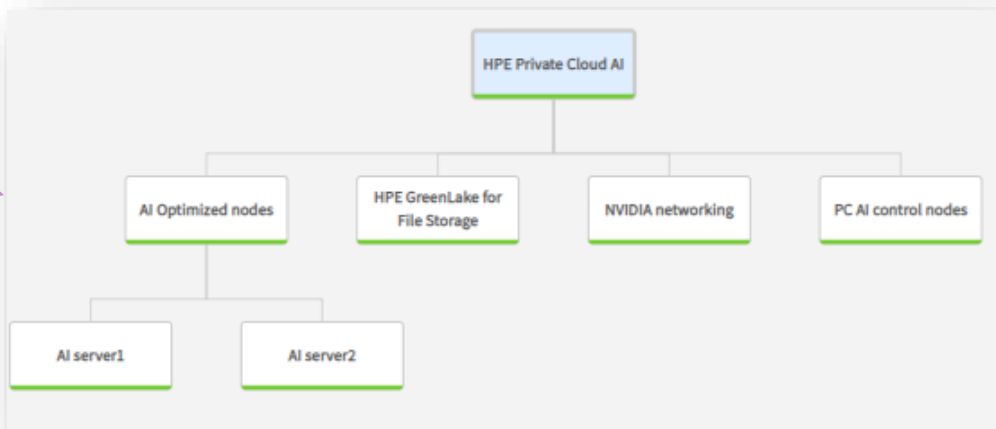
4. Monitor PCAI with OpsRamp

DreamCompany client dashboard shows various services monitored: including HPE private Cloud AI



Go in Infrastructure / Service Maps

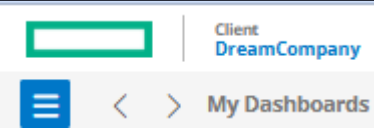
to view HPE private Cloud AI service and sub services monitored in OpsRamp



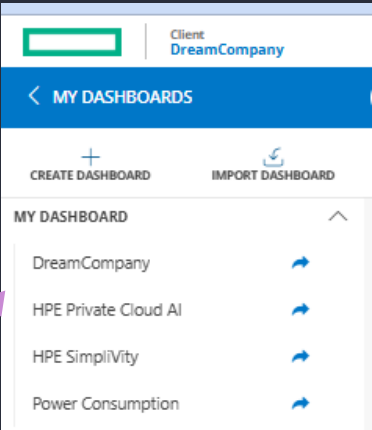
4. Monitor PCAI with OpsRamp

Go to the Dashboard menu

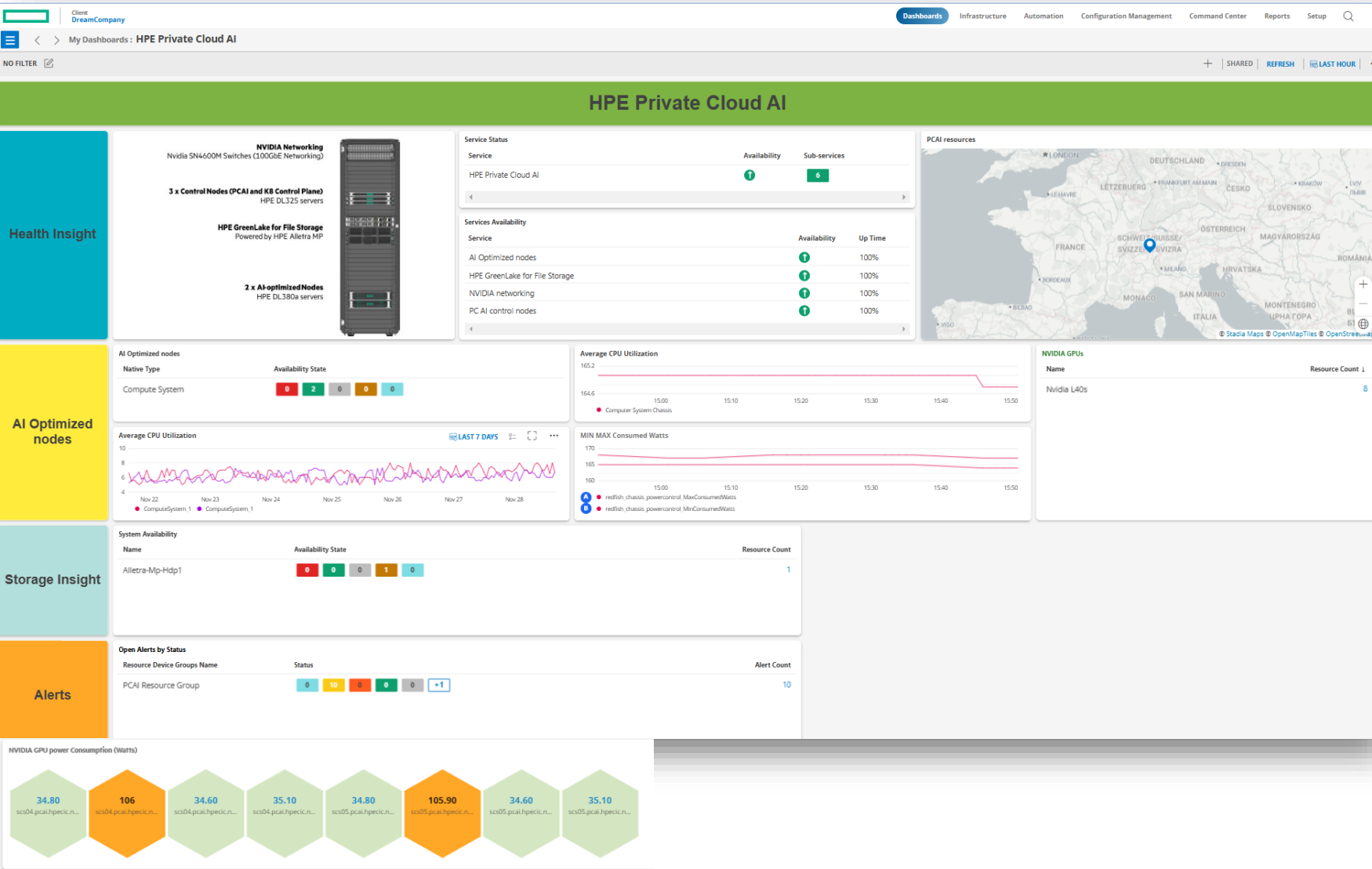
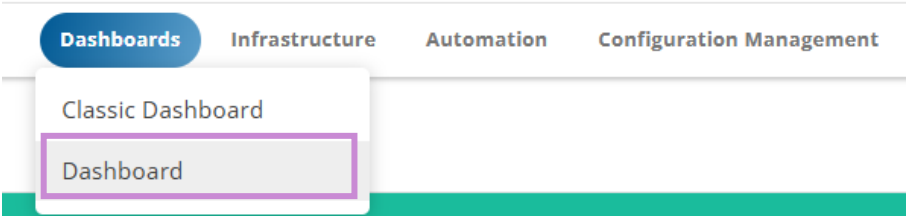
Click on

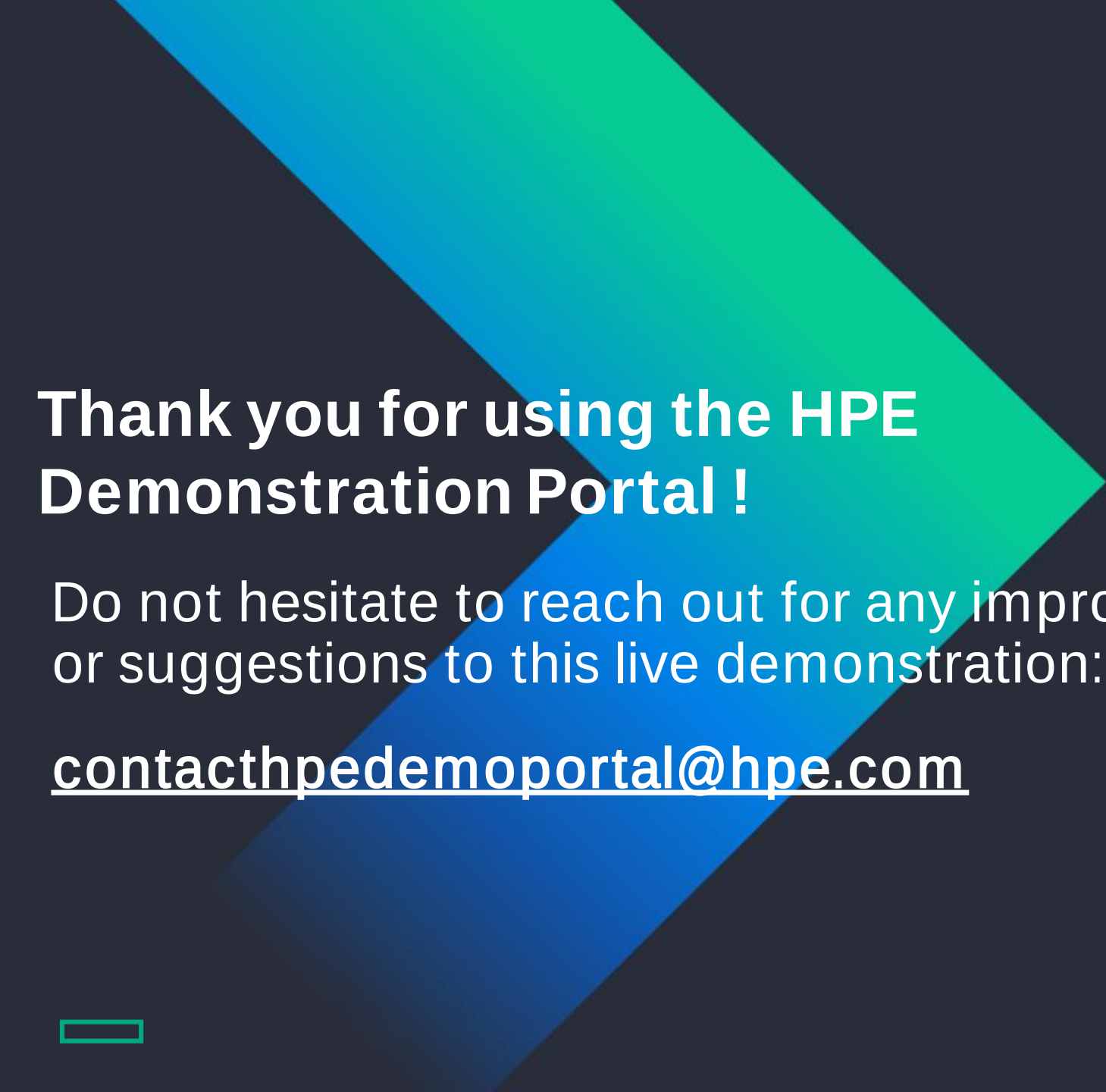


to see predefined dashboards
Open HPE Private Cloud AI



PCAI Heath, metrics and alerts are shown





Thank you for using the HPE Demonstration Portal !

Do not hesitate to reach out for any improvements
or suggestions to this live demonstration:

contacthpedemopoportal@hpe.com



Thank You

