AI Factories & Agentic AI

Acelerando la adopción y escalabilidad de la IA

Enzo Fedrizzi
Director, AI & Hybrid Cloud Digital Transformation Practice - Latin America & Southern Europe

Septiembre 2025

SERVICIO AL CLIENTE DE NUEVA GENERACIÓN

AYUDA 7X24

ESPERA CERO

CON LA MISMA VOZ

SERVICIO QUE YA ME CONOCE

RÁPIDO, CONFIABLE E HIPERPERSONALIZADO

HPE | intel

SALUD PROACTIVA E
INTERVENCIÓN TEMPRANA

ANALIZA MIS DATOS GENÓMICOS PARA PARA PREDECIR E IDENTIFICAR RIESGOS DE SALUD Y DISEÑAR TRATAMIENTOS PERSONALIZADOS

HPE | intel

OPTIMIZACIÓN DE INVENTARIO,
VENTAS Y LOGÍSTICA

ANTICIPA LA DEMANDA, OPTIMIZA INVENTARIO, PRECIOS Y DESPACHOS CON PREDICCIONES INTELIGENTES QUE MAXIMIZAN MÁRGENES Y REDUCEN QUIEBRES DE STOCK

Industrias y Casos de Uso de AI con HPE

Agricultura

Arquitectura e Ingeniería

Fuerzas Armadas

Educación

Medios y comunicaciones

Gobierno y Sector Público

Energía

Servicios Financieros

Servicios de Salud

Manufactura

Transporte y Logística

Multi industria

Retail

Ciudades y Espacios Inteligentes

Deportes

Telco

HPE | intel

Casos de Uso y Valor de Negocio

EXPERIENCIA

96%

de las organizaciones planean aumentar su uso de agentes de IA en el próximo año*

DIGITALIZACIÓN

20%

de mejora de rendimiento de planta, optimizando procesos industriales con IA*

Crecimiento del Negocio → Retorno para el futuro

Valor

Productividad → Crear beneficios → Retorno para el empleado

Valor

Reducción de Costos → Retorno de inversión

HPE | intel

Casos de Uso - Cómo los Ayuda HPE

Diagnosticar Madurez Organizacional y de los Datos

Definir y Priorizar un Portfolio de Casos de Uso Alineado a Objetivos del Negocio

Definir Estrategias, Plan de Ruta y Quick Wins

Sólo 1 de cada 10 pilotos de IA llegan a producción

HPE | intel

IDC Capgemini

Solo el 5 % de los pilotos de IA que alcanzan producción poseen un impacto medible

Adicionalmente no pasan de ser pilotos por que no escalan

Assess Your Data Needs Depending on the AI Technique

AI techniques heat map

AI techniques stability

- Low (L)
- Medium (M)
- High (H)

	Common AI techniques					
Use-case families	Generative models	Nongenerative machine learning	Optimization	Simulation	Rules/ heuristics	Graphs
Prediction/ forecasting	L	H	L	H	M	L
Planning	L	L	H	M	M	H
Decision intelligence	L	M	H	H	H	M
Autonomous systems	L	M	H	M	M	L
Segmentation/ classification	M	H	L	L	H	H
Recommendation systems	M	H	M	L	M	H
Perception	M	H	L	L	L	L
Intelligent automation	M	H	L	L	H	M
Anomaly detection/ monitoring	M	H	L	M	M	H
Content generation	H	L	L	H	L	L
Conversational user interfaces	H	H	L	L	M	H
Knowledge discovery	H	M	L	L	M	H

Source: Gartner

Gartner

Los LLM son una base poderosa, pero no resuelven por sí solos los retos empresariales;

es necesario un enfoque de **Agentic AI** que combine:

modelos, datos y agentes inteligentes para **orquestar acciones** y **generar verdadero valor**

FinOps Foundation

FinOps for AI

On-Demand (Shared) Capacity

- Cost is based on actual tokens processed through the API. Pricing is typically per 1,000 tokens and is specific to token type (prompt and completion). Shared capacity can have inconsistent latency and availability, especially at peak usage times.
- Shared capacity can be used for one time or non-prod/LT applications

HPE | intel

https://www.finops.org/wg/how-to-forecast-ai-services-costs-in-cloud/

FinOps Foundation

On-Demand (Shared) Capacity

- Cost is based on actual tokens processed through the API. Pricing is typically per 1,000 tokens and is specific to token type (prompt and completion). Shared capacity can have inconsistent latency and availability, especially at peak usage times.
- Shared capacity can be used for one time or non-prod/LT applications

Provisioned (dedicated) Capacity

- Provisioned capacity is guaranteed to be available and not shared with other customers.
- Includes a guaranteed amount of throughput (prompt tokens + completion tokens). Provisioned capacity is ideal for production applications sensitive to API latency and availability.

HPE | intel

https://www.finops.org/wg/how-to-forecast-ai-services-costs-in-cloud/

Las organizaciones pueden subestimar los costos de los proyectos de GenAI en nubes públicas entre un

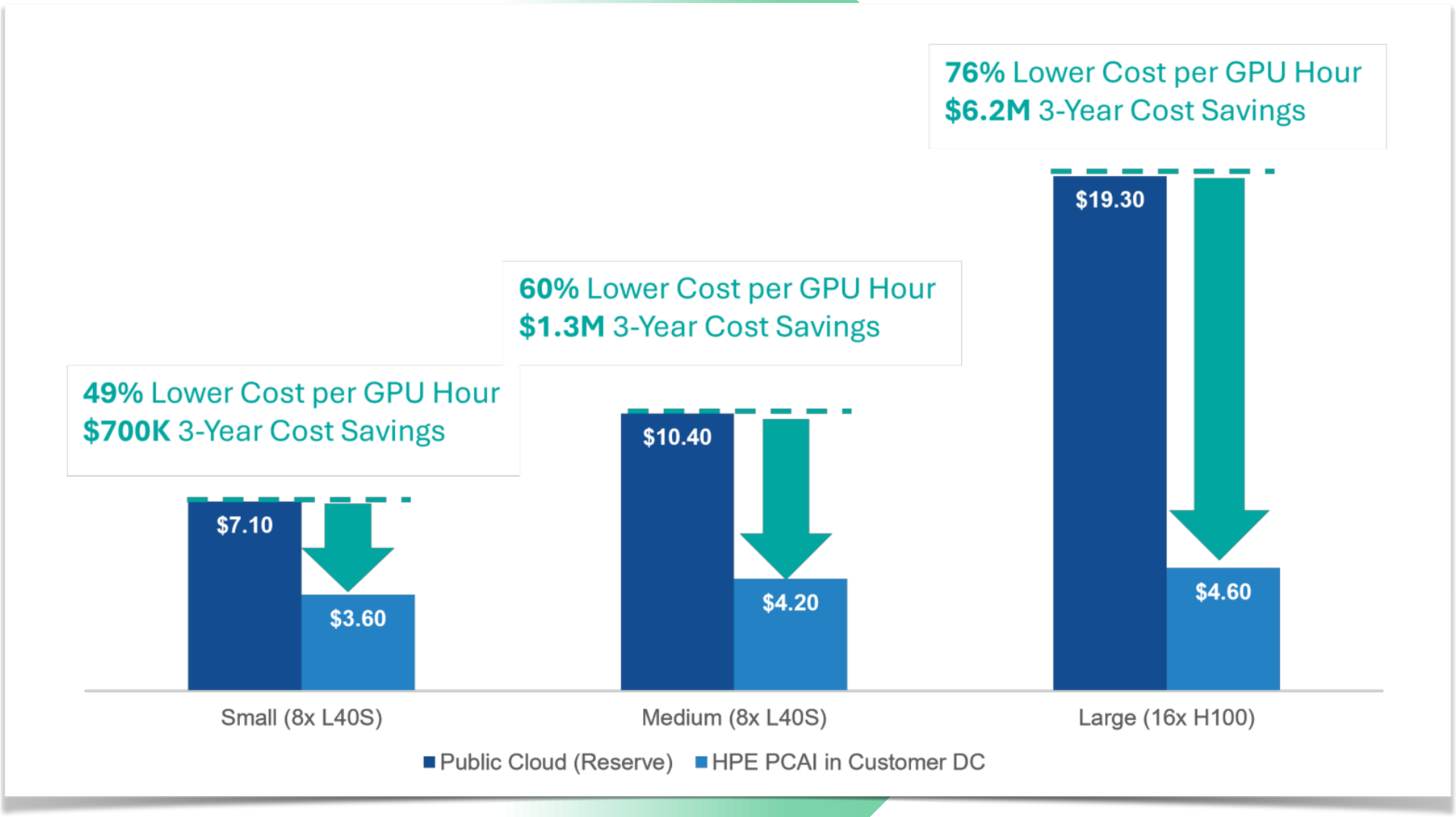
500% a 1000%

si no tienen una comprensión clara de cómo escalan los gastos del stack requerido para IA

HPE | intel

https://www.gartner.com/en/newsroom/press-releases/2024-10-21-gartner-identifies-four-emerging-challenges-to-delivering-value-from-ai-safely-and-at-scale

Gartner®


76% Lower Cost per GPU Hour
$6.2M 3-Year Cost Savings
60% Lower Cost per GPU Hour
$1.3M 3-Year Cost Savings
49% Lower Cost per GPU Hour
$700K 3-Year Cost Savings
$7.10
$3.60
$10.40
$4.20
$19.30
$4.60
Small (8x L40S)
Medium (8x L40S)
Large (16x H100)
Public Cloud (Reserve)
HPE PCAI in Customer DC

HPE | intel

Technology Sandwich for AI in Public Clouds

Category	Azure	AWS	GCP
Infrastructure	Azure AI Infrastructure: GPUs/TPUs, Virtual Machines, AI-optimized hardware	Amazon EC2, AWS Inferentia, Trainium, Elastic Inference	Google Cloud GPUs/TPUs, Compute Engine, Vertex AI infrastructure
Data	Azure Synapse Analytics, Azure Data Lake, Azure Data Factory	AWS Glue, Amazon S3, AWS Lake Formation	BigQuery, Cloud Storage, Dataflow, Pub/Sub
AI App Platform	Azure Machine Learning, Cognitive Services, Bot Service	SageMaker, AI-specific AWS Lambda integrations	Vertex AI, AutoML, AI Hub
Developer Tools	Azure DevOps, Visual Studio Code, ML Pipelines	AWS CodePipeline, SageMaker Studio, AWS Cloud9	AI Platform Notebooks, Cloud Code, Vertex AI Workbench
Inference	Azure AI Inference with ONNX Runtime, Azure Kubernetes Service (AKS)	AWS Elastic Inference, SageMaker Endpoints	Vertex AI Prediction, TensorFlow Serving, Kubernetes Engine
Marketplace	Azure Marketplace: AI solutions and pre-built apps	AWS Marketplace: AI and ML algorithms and models	Google Cloud Marketplace: Pre-trained models and solutions
AI Models	OpenAI models (GPT), Custom Vision, Text Analytics, Speech API	AWS AI Models (Text-to-Speech, Rekognition, Comprehend)	PaLM, Bard, Vision AI, Natural Language API, Translation API

HPE | intel

El uso de IA en nubes públicas plantea importantes desafíos;

en **control y variabilidad de costos**,
riesgos de **lock-in** y **gobierno de datos**,

dañando la obtención de valor por la
baja predictibilidad de costos,
performance y ROI incierto

HPE | intel

Decision Intelligence Platforms — Putting Everything Together

Decision support

Decision augmentation

Decision automation

Decision intelligence platform

Decision-centric UI | decision flow models and orchestration | monitoring and adaptation

Business intelligence

Diagnostic Analytics

Data science & machine learning

Optimi-zation

Knowledge graphs

Simulation

Generative AI

AI Agents

Rules & heuristics

Data ecosystem (fabric)

Internal and external | batch and streaming | analytical and operational | cross-silo | Multistructured

Gartner

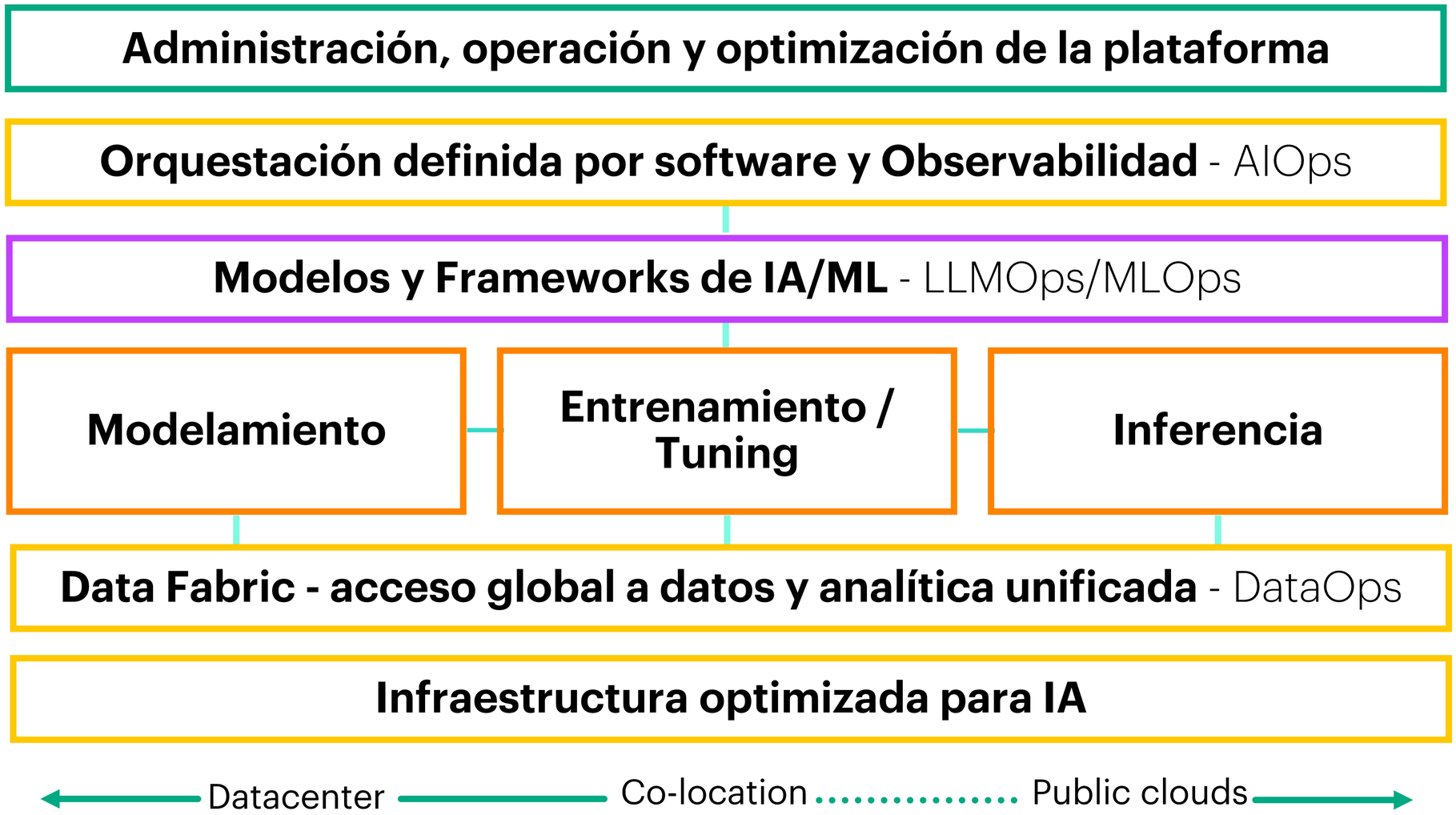
Un AI Factory on-premise

es un stack **listo para usar** que permite operar con **costos predecibles,** garantiza una **operación eficiente** y **gobernanza total de los datos**,

acelerando la adopción de AI mientras se **reducen los riesgos** y se **aumenta el control** estratégico de las iniciativas

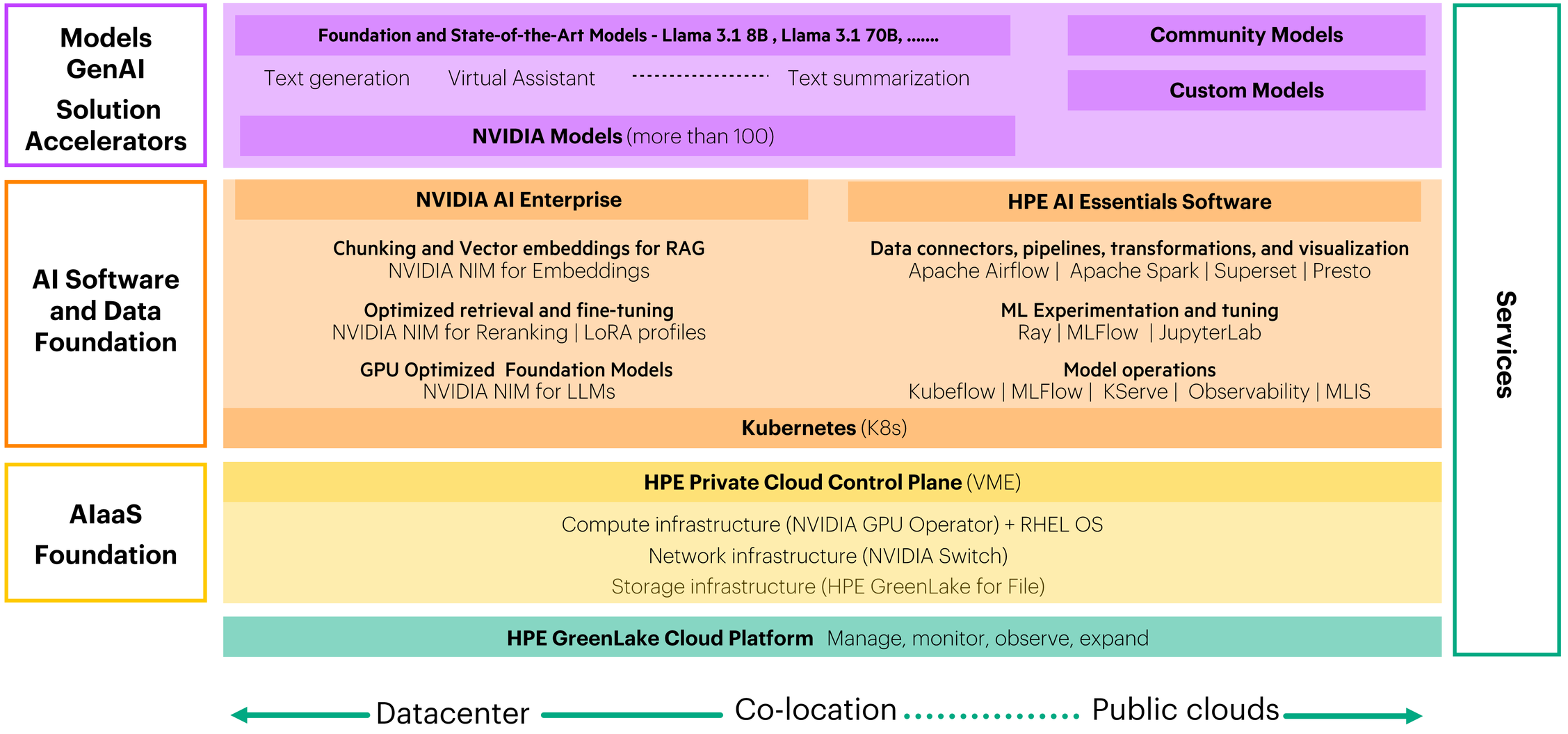
HPE | intel

HPE PRIVATE CLOUD AI



HPE | intel

HPE PRIVATE CLOUD AI



HPE | intel

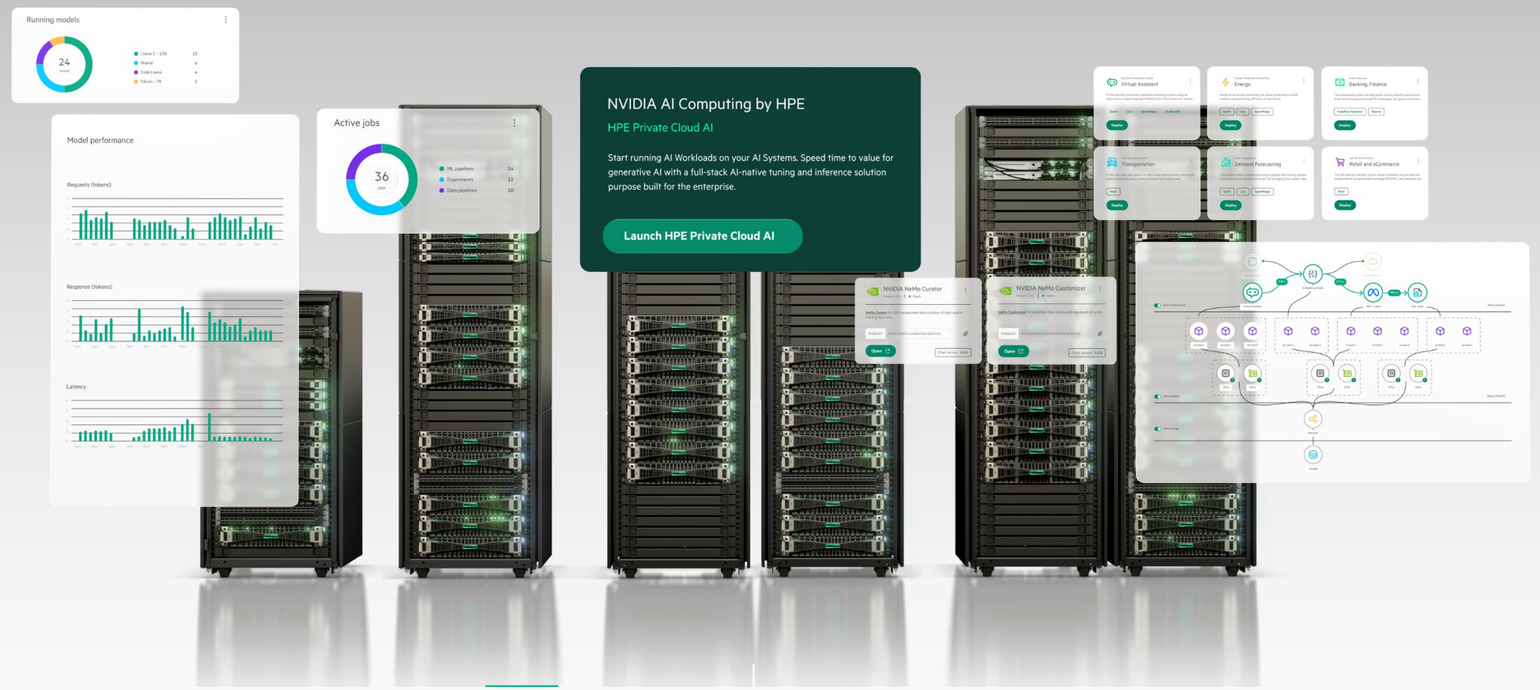
HPE PRIVATE CLOUD AI

Start Fast	Small	Medium	Large	Scale Fast
				
Developer System	**Small**	**Medium**	**Large**	**Extra Large**
2 NVIDIA H100 NVL GPU's	4 or 8 NVIDIA L40S GPUs	8 or 16 NVIDIA L40S GPUs	16 or 32 NVIDIA H100 NVL GPUs	Future Blackwell Support
32 TB Integrated	109 TB	217 TB	670 TB	TBD
Customer Network	100GbE NVIDIA Networking	200GbE NVIDIA Networking	400GbE NVIDIA Networking	800GbE NVIDIA Networking
Up to 2.2 kW	up to 8 kW rack	up to 17.7 kW	up to 16.5 kW x 2	TBD

Datacenter ← Co-location ..... Public clouds →

HPE | intel

HPE PRIVATE CLOUD AI



Zora AI™

by Deloitte

Accenture AI Refinery™

Accelerated by NVIDIA AI

HPE | intel

HPE

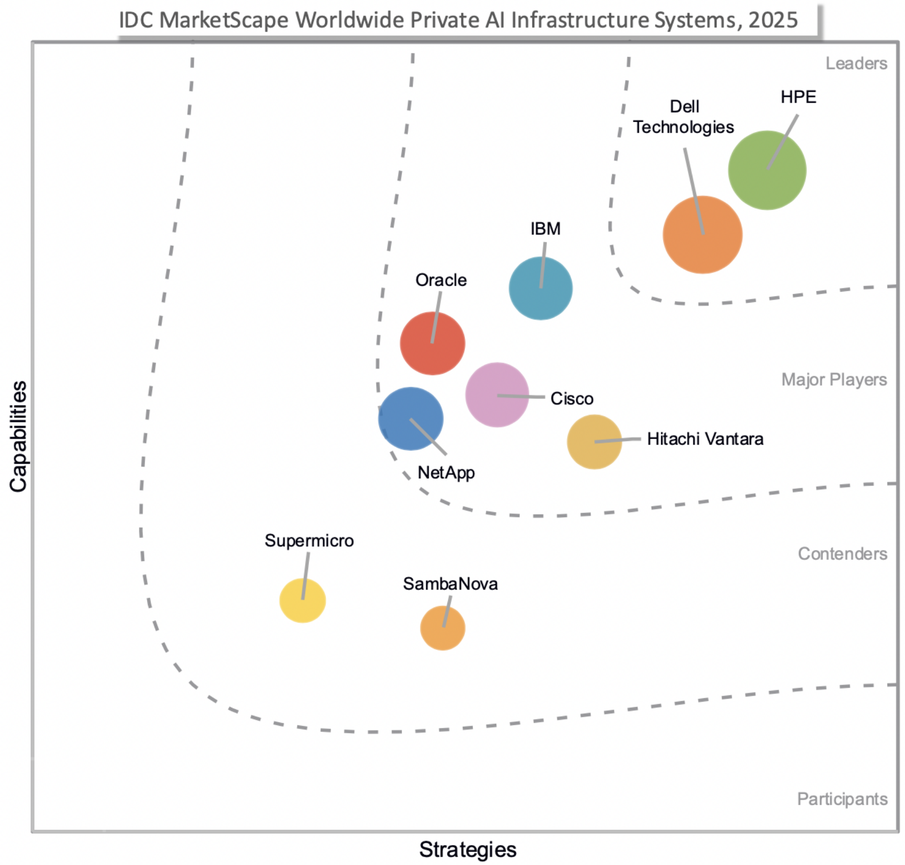
PRIVATE CLOUD AI

HPE | intel

IDC MarketScape: Worldwide Private AI Infrastructure Systems 2025 Vendor Assessment

Mary Johnston Turner

IDC MARKETSCAPE FIGURE



Source: IDC, 2025

Próximos Pasos Posibles de Trabajo Conjunto

Everything as a service

¿Cuáles de sus casos de uso prioritarios de IA y sus implicaciones deberíamos explorar con mayor profundidad?

¿Qué casos de uso recomendarías que sigamos con una sesión de exploración de soluciones más detallada?

Experiencia

- Online Meeting Summary
- Conversational Search
- Real-time Translation
- First Draft Content Creation
- Image Generation for Presentations
- Intelligent Code Compilation

Digitalización

- Quality Control
- Supply Chain Optimization
- Predictive Maintenance
- Robotics and Cobots
- Energy Management
- Digital Twins

Plataformas

- Public and Private Balance
- Sovereignty / IP Leakage
- Data Gravity / Entanglement
- Data Access / Egress Costs
- Use Case Support and Roadmap
- AI Accelerated Infrastructure

Ingeniería

- Data Infrastructure and Architecture
- Data Ingestion and Integration
- Data Cleansing and Preprocessing
- Data Governance and Security
- Data Quality and Validation
- Metadata Management and Documentation

Everything as a service

HPE | intel

GRACIAS