unisys

Client Connect México

12 de noviembre, Ciudad de México



Patrocinio

facephi



Israel Erick Pacheco Caballero

Director de Arquitectura de Soluciones, LATAM

DELL Technologies

© 2026 UNISYS CORPORATION. ALL RIGHTS RESERVED. CONFIDENTIAL AND PROPRIETARY INFORMATION.

Inteligencia Artificial Híbrida

Next Best Action

Acelere su jornada de IA



Israel Erick Pacheco Caballero, M.Sc.
Director Arquitectura de Soluciones LATAM
Dell Technologies

Israel.pachecocaballero@dell.com

CTOAmbassadors

"Esto no es un piloto. Así opera nuestra empresa"

— Jeff Clarke, Vicepresidente y COO, Dell Technologies · Dell Technologies World 2026

DELL Technologies

Copyright © Dell Inc. All Rights Reserved.

Dell Customer Communication - Confidential

México exige crecer con más disciplina: eficiencia antes que expansión

La eficiencia convierte presión de crecimiento en capacidad de ejecución.

0.6% 2025 › 1.5% 2026 › 1.8% 2027

Proyección de crecimiento del PIB de México (FMI)

El foco directivo cambia:
Hacer más con menos, decidir más rápido y liberar recursos para innovar.

Presión de crecimiento
Menos margen para absorber complejidad

➡

Eficiencia operativa
Menos fricción, mejor control y datos conectados

➡

Capacidad de innovación
Más recursos para innovar con disciplina y seguridad

La tecnología deja de ser solo un costo cuando reduce fricción, conecta información crítica y refuerza control y seguridad.

Fuente: World Economic Outlook, October 2026: Resilience Under Strain, Urgent Choices

DELL Technologies

La eficiencia ya es una capacidad de negocio

Simplificar operaciones, ordenar los datos y automatizar decisiones permite ejecutar con velocidad y control.

Estrategia

01

Proceso

Reducir complejidad, eliminar tareas repetitivas y acelerar decisiones.

02

Datos

Elevar calidad, disponibilidad y confianza en la información crítica.

03

Gobernanza

Proteger, controlar y operar con seguridad desde el diseño.

CAPACIDAD DE EJECUCIÓN | INTELIGENCIA ARTIFICIAL

INNOVACIÓN | PERSONAS Y CULTURA | ENFOQUE AL CLIENTE

DELL Technologies

y 6,500 empresas ya están en su camino hacia la IA on-premises | híbrida con Dell Technologies



Con 6.500+ Clientes en Dell AI Factory

Derechos de autor © Dell Inc. Todos los derechos reservados.

Por qué las organizaciones no logran generar un impacto medible con AI

Soluciones integrales

Proyectos piloto estancados sin objetivos claros	Desafíos de cargas de trabajo, datos y seguridad	La infraestructura existente obstaculiza la innovación	La falta de herramientas de AI ralentiza el despliegue	Brechas de habilidades que impiden la adopción y el éxito de AI	Aumento del costo de AI y del consumo de recursos a escala agéntica
RESULTADOS DE NEGOCIO DESEADOS	**PLATAFORMA DE DATOS PARA AI**	**PLATAFORMA AI FABRIC**	**ECOSISTEMA ABIERTO DE AI**	**PERSONAS Y CULTURA**	**ROI MEDIBLE Y CONTROL DE COSTOS**
20%	**TOP 2**	**95%**	**12%**	**25%**	**60%**
Logra aumentar ingresos con IA vs 74% que esperaba lograrlo[1]	**Barrera de implementación: seguridad y preparación de datos[3]**	**De los pilotos de IA no logran escalar a producción[2]**	**De los proyectos de IA se han desplegado por completo por falta de flexibilidad del ecosistema 4**	**De los especialistas dicen estar preparados para implementar proyectos de IA4**	**de los casos de uso digitales impulsados por AI no alcanzarán los objetivos de ROI en 2026[3]**

1. Deloitte, *El estado de la IA en la empresa*, enero 2026.
2. MIT NANDA, *La brecha GenAI: Estado de la IA en los negocios 2025*, julio 2025.
3. Gartner, encuesta sobre madurez en IA, junio 2025.
4. Riverbed. (2025, 23 de septiembre). Encuesta global revela una brecha de preparación para AI: pese a inversiones récord, solo uno de cada diez proyectos de AI se ha desplegado por completo.

Copyright © Dell Inc. All Rights Reserved.

Era del razonamiento (2022) – Carga de trabajo agéntica (2026 – actualidad)

Mantener el control total de costos | Precio basado en tokens = Unidad de costo de inferencia ($$$)

Era del razonamiento (2022) – Carga de trabajo agéntica (2026 – actualidad)

Mantener el control total de costos | Precio basado en tokens = Unidad de costo de inferencia ($$$)



Reasoning-era
(6 nov 2023)
Single-turn. Human in the loop.

- User asks → model answers

Early enterprise AI deployments have largely relied on cloud-hosted models due to their accessibility, scalability, and consumption-based pricing. Cloud providers offer on-demand access to both open-weight models, such as those that can be run on-premises, and several other popular leading frontier models—some of which can also be run on-premises, all at varying token-based pricing. While well suited for bursty

Prompt determinista: Explícame esta receta de pastel

Razonamiento: *El usuario solicita explicar la receta del pastel basándose estrictamente en el contexto inyectado, sin alucinar*

Instrucciones del sistema	Razonamiento del sistema	Respuesta del sistema
10 tokens	50 tokens	180 tokens

Era del razonamiento (2022) – Carga de trabajo agéntica (2026 – actualidad)

Mantener el control total de costos | Precio basado en tokens = Unidad de costo de inferencia ($$$)



Agentic workload
(2026 – today)
Multi-step. Model in the loop.

- Plan → act → observe → repeat

significantly increasing long-term token consumption. Compared to standard chat interactions, agentic workloads consume orders of magnitude more tokens with agents easily utilizing 4x to 15x more tokens. As agentic workloads continue to evolve, token growth is expected to grow even further, with autonomous agents driving up to 1000x more inference demand than reasoning AI.

Meta-Prompt: Actúa como un Consultor Autónomo de Pastelería e Inteligencia de Producto . Tu objetivo es diseñar e implementar el pastel de chocolate óptima resolviendo la misión de extremo a extremo

Planificar – actuar – observar – repetir: Sé un consultor de pastelería: busca recetas, compara chocolate, valora ingredientes, cuestiónate y luego escribe la respuesta final

Planifica la misión de tarta perfecta	Buscar candidatos a recetas	Compara las opciones de chocolate	Ponle precio a la cesta de la compra	Autocrítica del plan	Crear la salida final	Añadir notas de servicio y compra
800 tokens	1,300 tokens	1,500 tokens	1,700 tokens	1,600 tokens	2,200 tokens	1,500 tokens

Copyright © Dell Inc. All Rights Reserved.

Era del razonamiento (2022) – Carga de trabajo agéntica (2026 – actualidad)

Mantener el control total de costos | Precio basado en tokens = Unidad de costo de inferencia ($$$)



Reasoning-era
(6 nov 2023)
Single-turn. Human in the loop.

- User asks → model answers

Modelo ligero ~240 tokens

~$0.50 USD / MTokens

Prompt determinista: Explícame esta receta de pastel

Razonamiento: *El usuario solicita explicar la receta del pastel basándose estrictamente en el contexto inyectado, sin alucinar*

Instrucciones del sistema	Razonamiento del sistema	Respuesta del sistema
10 tokens	50 tokens	180 tokens



Agentic workload
(2026 – today)
Multi-step. Model in the loop.

- Plan → act → observe → repeat

significantly increasing long-term token consumption. Compared to standard chat interactions, agentic workloads consume orders of magnitude more tokens with agents easily utilizing 4x to 15x more tokens. As agentic workloads continue to evolve, token growth is expected to grow even further, with autonomous agents driving up to 1000x more inference demand than reasoning AI.

Meta-Prompt: Actúa como un Consultor Autónomo de Pastelería e Inteligencia de Producto . Tu objetivo es diseñar e implementar el pastel de chocolate óptima resolviendo la misión de extremo a extremo

Planificar – actuar – observar – repetir: Sé un consultor de pastelería: busca recetas, compara chocolate, valora ingredientes, cuestiónate y luego escribe la respuesta final

Planifica la misión de tarta perfecta	Buscar candidatos a recetas	Compara las opciones de chocolate	Ponle precio a la cesta de la compra	Autocrítica del plan	Crear la salida final	Añadir notas de servicio y compra
800 tokens	1,300 tokens	1,500 tokens	1,700 tokens	1,600 tokens	2,200 tokens	1,500 tokens

DELL Technologies

Copyright © Dell Inc. All Rights Reserved.

Era del razonamiento (2022) – Carga de trabajo agéntica (2026 – actualidad)

Mantener el control total de costos | Precio basado en tokens = Unidad de costo de inferencia ($$$)



Reasoning-era
(6 nov 2023)
Single-turn. Human in the loop.

- User asks → model answers

Modelo ligero ~240 tokens

~$0.50 USD / MTokens

Prompt determinista: Explícame esta receta de pastel

Razonamiento: *El usuario solicita explicar la receta del pastel basándose estrictamente en el contexto inyectado, sin alucinar*

Instrucciones del sistema	Razonamiento del sistema	Respuesta del sistema
10 tokens	50 tokens	180 tokens



Agentic workload
(2026 – today)
Multi-step. Model in the loop.

- Plan → act → observe → repeat

Modelo medio ~10,600 tokens

44X | 4,316% más costo en inferencia

Meta-Prompt: Actúa como un Consultor Autónomo de Pastelería e Inteligencia de Producto . Tu objetivo es diseñar e implementar el pastel de chocolate óptima resolviendo la misión de extremo a extremo

Planificar – actuar – observar – repetir: Sé un consultor de pastelería: busca recetas, compara chocolate, valora ingredientes, cuestiónate y luego escribe la respuesta final

Planifica la misión de tarta perfecta	Buscar candidatos a recetas	Compara las opciones de chocolate	Ponle precio a la cesta de la compra	Autocrítica del plan	Crear la salida final	Añadir notas de servicio y compra
800 tokens	1,300 tokens	1,500 tokens	1,700 tokens	1,600 tokens	2,200 tokens	1,500 tokens

Copyright © Dell Inc. All Rights Reserved.

Era del razonamiento (2022) – Carga de trabajo agéntica (2026 – actualidad)

Mantener el control total de costos | Precio basado en tokens = Unidad de costo de inferencia ($$$)



Reasoning-era
(6 nov 2023)
Single-turn. Human in the loop.

- User asks → model answers

Modelo ligero ~240 tokens

~$0.50 USD / MTokens

Prompt determinista: Explícame esta receta de pastel

Razonamiento: *El usuario solicita explicar la receta del pastel basándose estrictamente en el contexto inyectado, sin alucinar*

Instrucciones del sistema	Razonamiento del sistema	Respuesta del sistema
10 tokens	50 tokens	180 tokens



Agentic workload
(2026 – today)
Multi-step. Model in the loop.

- Plan → act → observe → repeat

Modelo medio ~10,600 tokens

44X | 4,316% más costo en inferencia

Modelo complejo ~28,500 tokens

119x más | 11,775% más en costo de inferencia

~$15.00 USD / MTokens

Meta-Prompt: Actúa como un Consultor Autónomo de Pastelería e Inteligencia de Producto . Tu objetivo es diseñar e implementar el pastel de chocolate óptima resolviendo la misión de extremo a extremo

Planificar – actuar – observar – repetir: Sé un consultor de pastelería: busca recetas, compara chocolate, valora ingredientes, cuestiónate y luego escribe la respuesta final

Planifica la misión de tarta perfecta	Buscar candidatos a recetas	Compara las opciones de chocolate	Ponle precio a la cesta de la compra	Autocrítica del plan	Crear la salida final	Añadir notas de servicio y compra
800 tokens	1,300 tokens	1,500 tokens	1,700 tokens	1,600 tokens	2,200 tokens	1,500 tokens

DELL Technologies

Copyright © Dell Inc. All Rights Reserved.

Economía sostenible de AI (Tokenomics)



EDITORS' PICK | INNOVATION > CLOUD

Uber Burns Its 2026 AI Budget In Four Months On Claude Code

By Janakiram MSV, Senior Contributor. ⓘ I cover emerging technologies with... Follow Author

Published May 17, 2026, 11:18pm EDT, Updated May 27, 2026, 12:17pm EDT



AI Basics

Microsoft Cancels Claude Code Licenses: The Real Reason (2026)

James Swierczewski · May 15, 2026

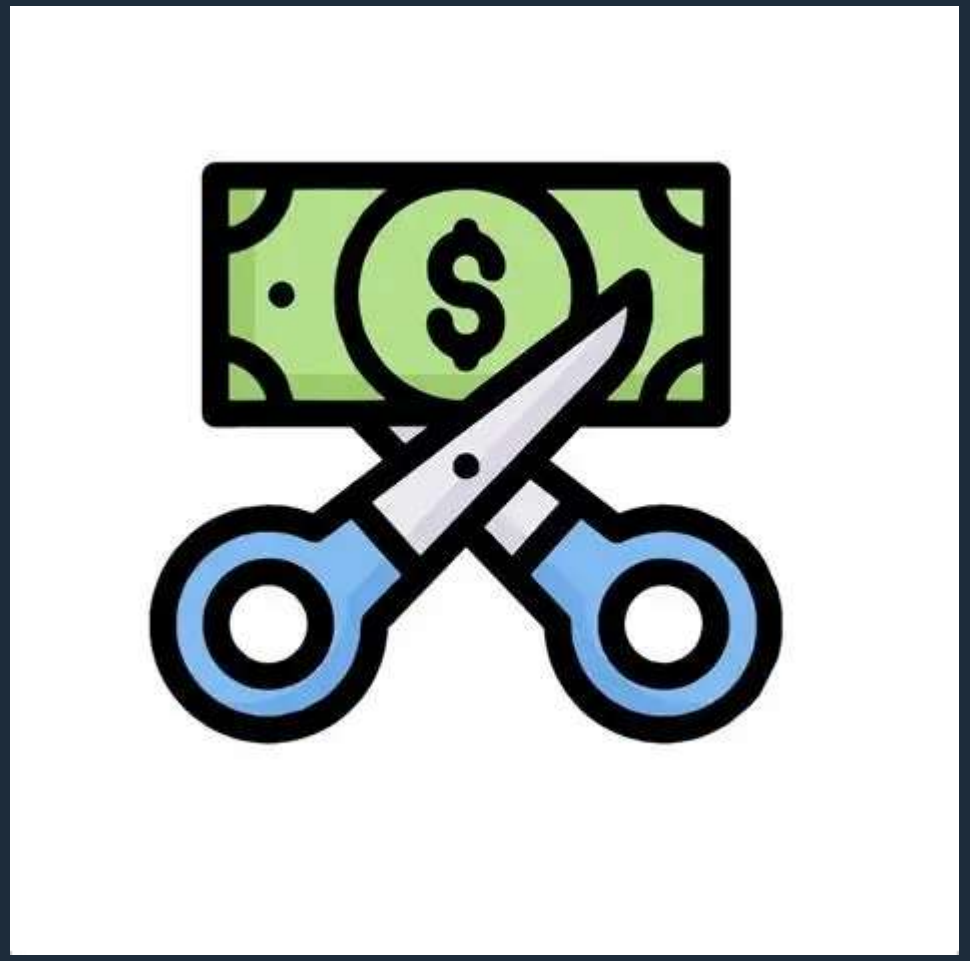
One: cost, and the timing of cost.

Two: telemetry and product feedback.

El costo no depende solo de los tokens, sino de la ampliación de contexto, la recuperación, las llamadas a herramientas, los reintentos, la orquestación, el egreso y la observabilidad.

Los informes públicos ya muestran que el problema afecta a grandes organizaciones tecnológicas.

La AI híbrida preserva la elasticidad de la cloud y añade capacidad privada sin sacrificar el LLM ni la transparencia para el usuario. Además, da a las organizaciones control directo de sus cargas de trabajo de AI, independientemente de cambios de precios, límites de velocidad, descontinuaciones de modelos o disponibilidad del proveedor.




accenture
EVOLVING AI
ACCENTURE LIMITS EMPLOYEE AI SPENDING ON CLAUDE TO $29 PER MONTH AS SOME WORKERS REPORTEDLY SPEND UP TO $1,000 MONTHLY

c. All Rights

Protege la libertad para innovar

El valor oculto de AI on-premises

Innovación

Sin límites para casos de uso de AI en I+D

Cada empleado puede mejorar y experimentar con nuevos casos de uso sin costos asociados

Sin exposición de datos

Acceso seguro a datos sensibles y propiedad industrial para explorar y mejorar datos y modelos

Retención de talento y entrenamiento de modelos

Las personas más talentosas prefieren empresas de tecnología avanzada

El uso continuo y la retroalimentación mejoran sus propios modelos



Copyright © Dell Inc. All Rights

Capa de IA híbrida

1. Prompt de negocio

Nube Privada | Co-hosted

Datos sensibles y procesamiento de baja latencia

92 modelos validados para despliegue

Dell AI Factory

Red Hat OpenShift | Linux AI | SUSE | Ubuntu

AI Data Platform | Security and Resilience Platform

Llama | Mistral AI | DeepSeek | Qwen | Dell Enterprise Frontier

AMD | NVIDIA | intel.

Orquestación y seguridad

Gobierno, cifrado y movimiento seguro de datos

Nube pública

Cómputo GPU escalable para entrenamiento e inferencia

Modelos Frontier | OpenAI | Anthropic | Google Gemini

Google Cloud | aws | Microsoft Azure

Resultados de negocio: analítica, productividad y automatización

Prompt



Mejores decisiones y mayor agilidad



Equipos más productivos



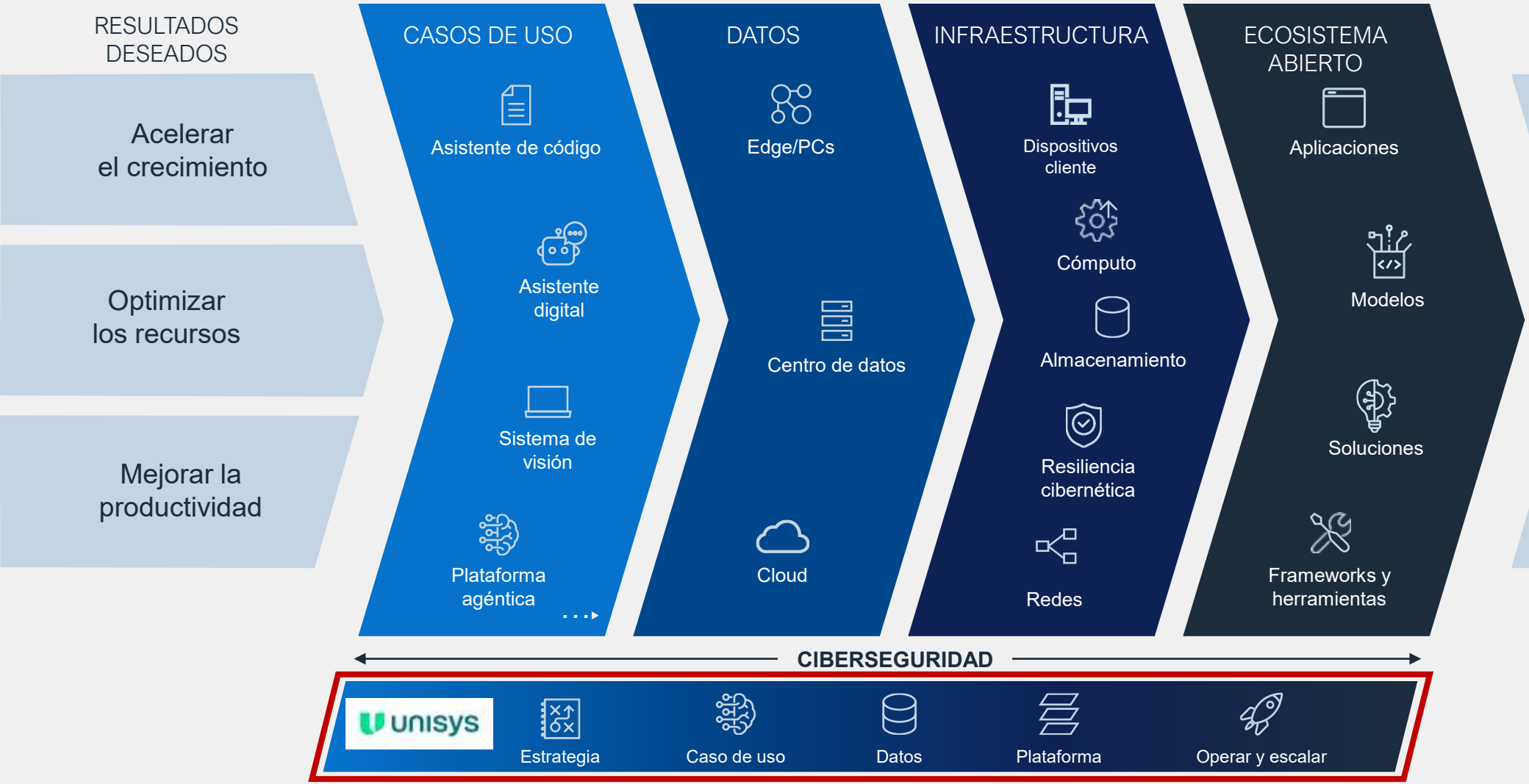
Mayor eficiencia operativa

Costo-beneficio:
mantenga cargas críticas localmente y use la nube bajo demanda para optimizar inversión, rendimiento y escalabilidad.

DELL Technologies

THE DELL AI FACTORY

Ofrece el portafolio de AI más amplio[1], elegido por más de 6,500 clientes



[1] Prowess Consulting, "¿Qué proveedor ofrece el portafolio de AI más amplio para infraestructura y servicios?", octubre de 2025. La investigación de Prowess Consulting muestra que Dell Technologies ofrece el portafolio de AI end-to-end más amplio del mundo de un solo proveedor.\

Copyright © Dell Inc. All Rights Reserved.

Creando acceso a un ecosistema abierto amplio* para la innovación en IA



NEW!

cohere

Concéntrate de forma segura implementando búsquedas inteligentes y agentes personalizables

NEW!

Google Cloud

Trae el poder y las capacidades de Google Gemini a las instalaciones

NEW!

Red Hat

Ofrece opciones y mejora la interoperabilidad trayendo la pila Red Hat AI a Dell AI Factory

NEW!

Meta

Prototipar rápidamente y construir aplicaciones basadas en agentes con los modelos más recientes de Llama 4 y Llama Stack

NEW!

glean

Ofrecer gestión intuitiva del conocimiento y capacidades avanzadas de búsqueda

** representativo, no integral , así como proveedores de hosting y de soluciones globales sujetos a disponibilidad por país*

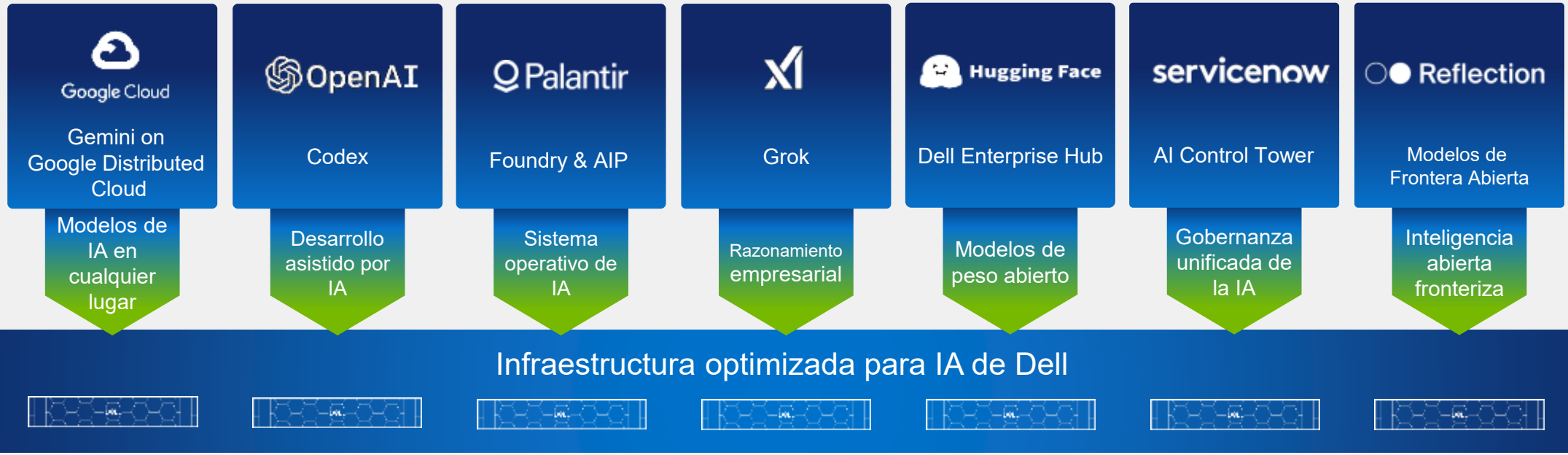
Copyright © Dell Inc. All Rights Reserved.

Trayendo líderes de IA a la empresa | IA híbrida | on-premises

Mantener el control total sobre los datos, modelos y operaciones

La elección no es por etiqueta: es por resultado de negocio, TCO, datos, continuidad y control operativo.

~50 familias de LLMs y sus variants preconfiguradas con scripts: tan solo en Dell Enterprise Hub en Hugging Face



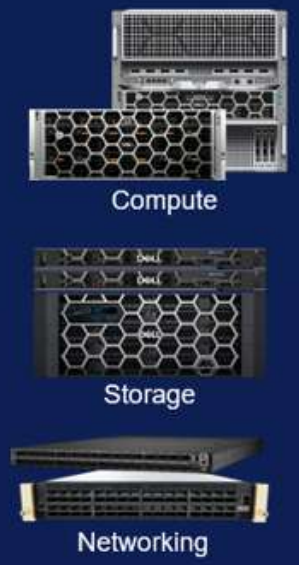
Copyright © Dell Inc. All Rights Reserved.

Dell IA FULL-STACK, DISEÑADA PARA INTELIGENCIA

Solo Dell entrega la pila completa, desde escritorio hasta centro de datos, así que las organizaciones convierten la ambición de IA en resultados más rápido, con menos riesgo y un socio responsable.



Dell Deskside Agentic AI



Plataformas Dell de primer nivel



Plataforma Dell de Datos de IA con NVIDIA



Dell PowerRack



Dell AI Factory con NVIDIA | AMD | INTEL Modular, validado

DELL Technologies

Copyright © Dell Inc. All Rights Reserved.

Dell AI Factory | Ejemplo: 8 casos de uso comunes de IA

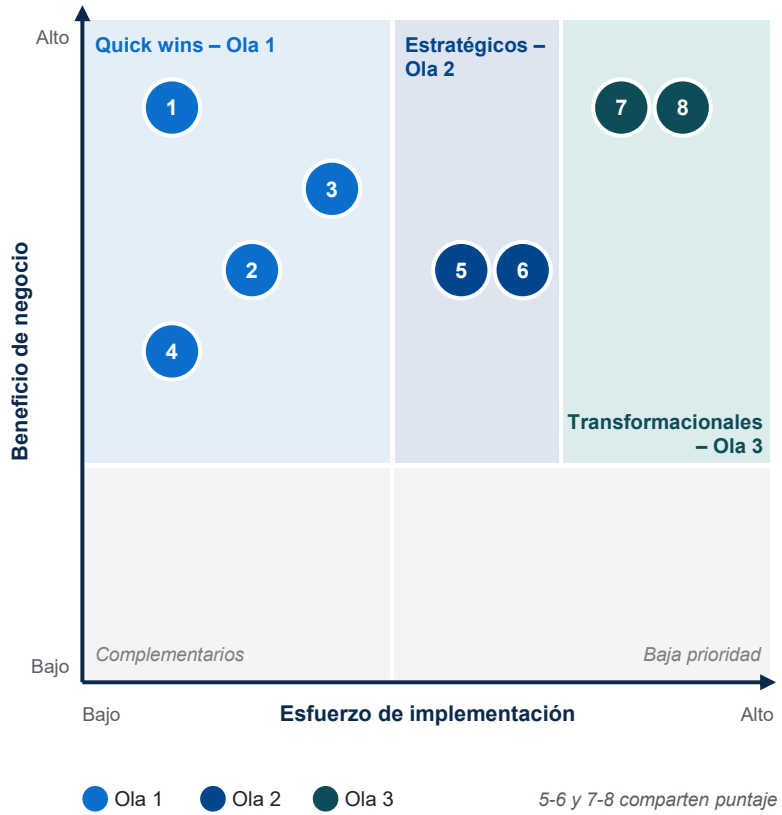
Capturar eficiencia primero y escalar a nuevos ingresos



Priorización preliminar por impacto de negocio, factibilidad técnica y time-to-value (metodología Dell AI Accelerator Workshop)

Ejemplos ilustrativos para cualquier industria (manufactura, retail, servicios financieros, salud, gobierno, educación): empezar en el back office, construir la plataforma de datos y escalar a IA especializada que genera nuevos ingresos.

Matriz de priorización: Beneficio vs. Esfuerzo



Ola 1 Ola 2 Ola 3 *5-6 y 7-8 comparten puntaje*

#	Caso de uso	Valor de negocio	Modelo potencial · arquitectura probable	B/E	Ola
			EDGE · DELL AI FACTORY · NUBE PÚBLICA · APIs FRONTIER		
1	**Asistente de conocimiento (RAG sobre políticas, manuales y procesos)**	Respuestas consistentes; menos tiempo buscando información interna	— · ● · ◐ · ○ — *RAG privado con LLM open-weight on-prem; nube para picos de demanda; frontier solo para casos complejos*	5/2	Ola 1
2	**Asistente de ventas (propuestas, CRM, preparación de cuentas)**	Mayor productividad comercial; propuestas más rápidas	◐ · ● · ◐ · ○ — *RAG sobre CRM y catálogo en AI Factory; AI PC para tareas locales; integración con la suite de productividad*	4/2.5	Ola 1
3	**Asistente de código (desarrollo, pruebas, modernización)**	Ciclos de desarrollo más cortos; protege la propiedad intelectual	◐ · ● · ○ · ◐ — *Modelos de código on-prem para proteger IP; workstations / AI PC; frontier para tareas no sensibles*	4.5/3	Ola 1
4	**Asistente de marketing (contenido, campañas, personalización)**	Más contenido en menos tiempo con guardrails de marca	◐ · ◐ · ● · ○ — *Cloud-first para generación creativa; RAG de marca y datos de clientes en AI Factory; AI PC para edición local*	3.5/2	Ola 1
5	**Predicción de demanda y siguiente mejor oferta (NBO)**	Más ingresos por cliente; inventario y capacidad optimizados	— · ● · ◐ · ○ — *ML predictivo sobre Dell AI Data Platform on-prem; BI en nube; GenAI para personalizar la oferta*	4/4	Ola 2
6	**Agentes de automatización de procesos (facturas, contratos, reclamos)**	Menor costo por transacción con humano en el loop	— · ● · ○ · ◐ — *Agentes on-prem por sensibilidad fiscal/contractual; frontier solo para excepciones anonimizadas*	4/4	Ola 2
7	**Visión por computadora en el edge (calidad, seguridad, inventario)**	Menos defectos, mermas e incidentes en planta, tienda o sitio	● · ◐ · ○ · — — *Entrena y reentrena en AI Factory; infiere en planta, tienda, hospital o sitio remoto con baja latencia*	5/5	Ola 3
8	**Nuevos productos / servicios de IA para clientes (modelo propio)**	Nuevas líneas de ingreso y diferenciación competitiva	◐ · ● · ◐ · ○ — *Modelo afinado con datos propios en AI Factory; soberano y multi-tenant; entrega en sitio vía edge*	5/5	Ola 3

● principal ◐ complemento ○ opcional — no aplica · Router de modelos decide por costo, latencia y sensibilidad del dato

Ola 1 · 0–6 meses
Quick wins en back office: asistentes de conocimiento, ventas, código y marketing; PoV con criterios de aceptación (adopción, horas ahorradas, costo por millón de tokens).

Ola 2 · 6–12 meses
Plataforma de datos: Dell AI Data Platform, analítica predictiva y agentes de procesos; gobierno, seguridad y ciberresiliencia.

Ola 3 · 12–24 meses
Diferenciación y monetización: IA en el edge y nuevos productos / servicios de IA para clientes.

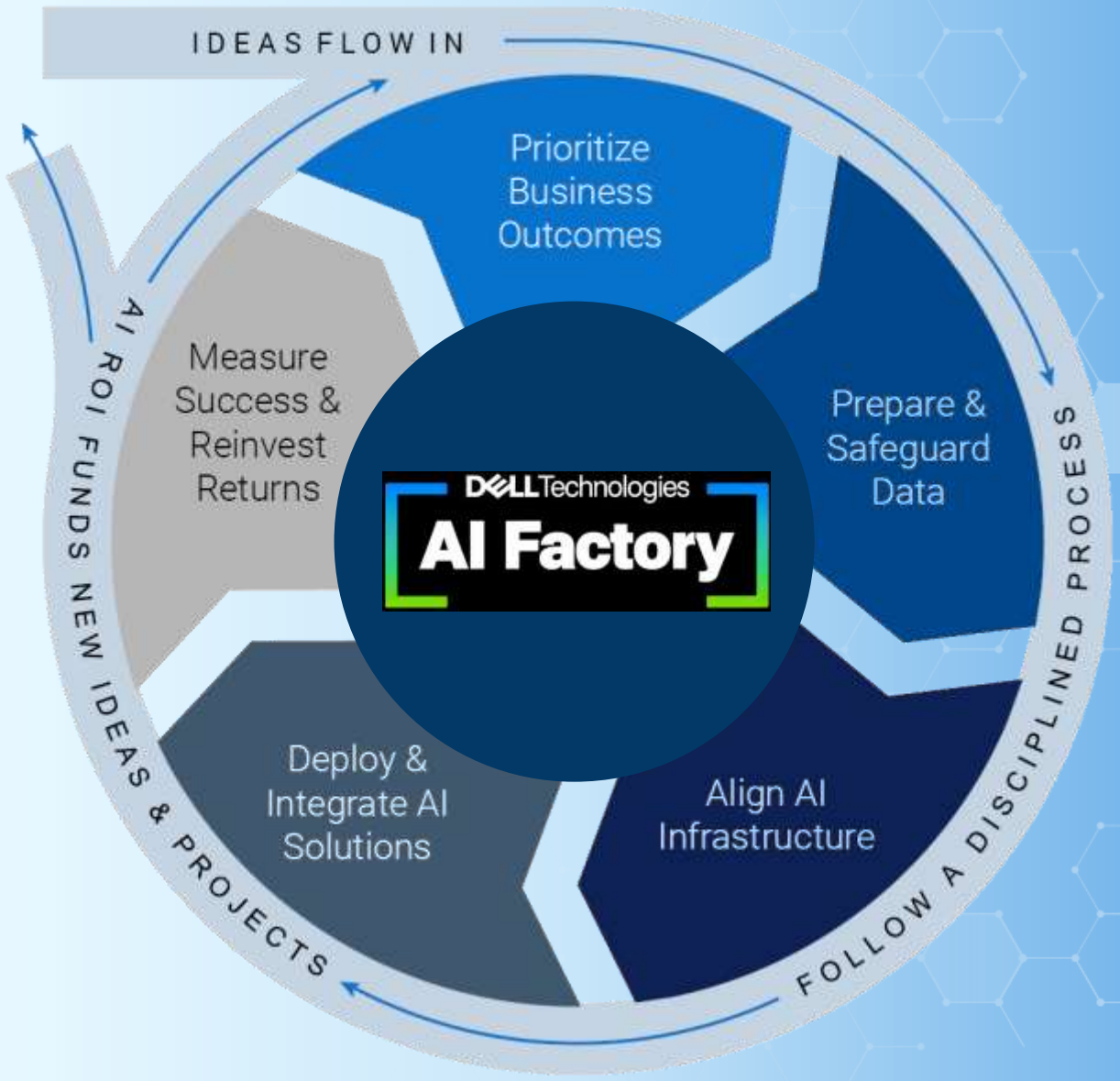
Casos y puntajes B/E ilustrativos: adaptar por industria y validar en Dell AI Accelerator Workshop con datos del cliente (volúmenes, tokens, SLA, data readiness). Sizing oficial vía herramientas Dell. Referencias: Dell AI Factory; Dell AI Data Platform.

Copyright © Dell Inc. All Rights Reserved.

No te vendemos una teoría: así opera Dell hoy

Dell Consulting Services
hace que la IA sea repetible

- Gen AI Accelerator Workshop
- Data Management
- Dell AI Data Platform
- AI Strategy
- AI PCs
- Dell Cyber Resilience
- AI Infra Deployment Services
- Data Center for AI
- AI for Industry
- AI Lab | Dell Pro Max w/GB10



Copyright © Dell Inc. All Rights Reserved.

IA Híbrida: Next Best Action

Acelere su jornada de IA con un socio tecnológico estratégico.

Comienza con el área que corresponde a tu cuello de botella actual y luego amplía el modelo

1

Evaluación de control

Evalúa la preparación de gobernanza, seguridad, riesgo y soberanía de datos antes de escalar AI.

2

Revisión de optimización

Evalúa la ubicación de cargas de trabajo y la economía de AI en el escritorio, edge, núcleo y cloud para mejorar el costo por resultado.

3

Taller de escalamiento

Diseña el despliegue y la adopción en producción para que los casos de uso comprobados se conviertan en una capacidad para toda la empresa

DELL Technologies AI Factory

Dell Technologies es el socio estratégico para poner en acción, dar seguimiento o retomar un plan de adopción de IA con valor


Experiencia de suscripción


Experiencia de consumo


Experiencia de propiedad

INNOVACIÓN PERSONAS Y CULTURA ENFOQUE AL CLIENTE

Copyright © Dell Inc. All Rights Reserved.

Dell Pro Max GB10 | características clave



1000 TFLOPS FP4 de Poder de cómputo*

Soporta Modelos hasta **200Bn** de parámetros*

Stack **2x** GB10 a través de CX7 a modelos de parámetros de admisión de 400 mil millones (nivel Llama 4 Maverick)*

*Desempeño preliminar, sujeto a cambios

2025 COMMERCIAL DESIGN

NVIDIA. DELL Technologies



Israel Erick Pacheco Caballero, M.Sc.
Director Arquitectura de Soluciones LATAM
Dell Technologies

Israel.pachecocaballero@dell.com

Aplica la IA en los procesos más impactantes dentro de sus áreas clave para acelerar su negocio

Next Best Action with

DELL Technologies & unisys

CTOAmbassadors

Copyright © Dell Inc. All Rights Reserved.