Dell AI Factory

Como a Mágica Acontece

With NVIDIA & INTEL

Rogério Chola
HPC&AI Solution Architect

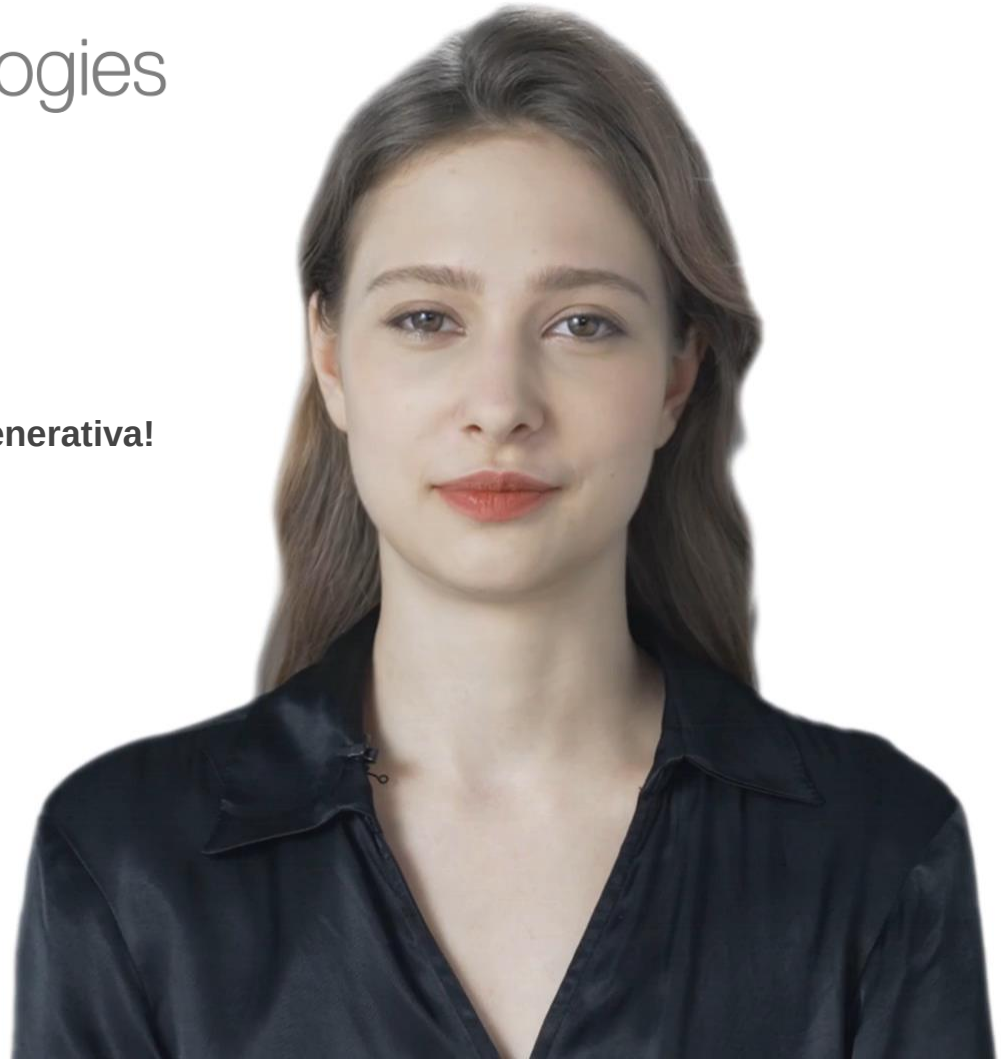
DELL Technologies

DELL Technologies

NVIDIA

intel

Bem-vindos a Era da IA Generativa!



HeyGen

Pilares da Tecnologia: Prontidão para o que vem a seguir

Principais tecnologias emergentes no mundo atual de dados em qualquer lugar

5G

Segurança Intrinseca

Digital Twin Industria 4.0

IA Generativa IA Cognitiva

Computação Quântica

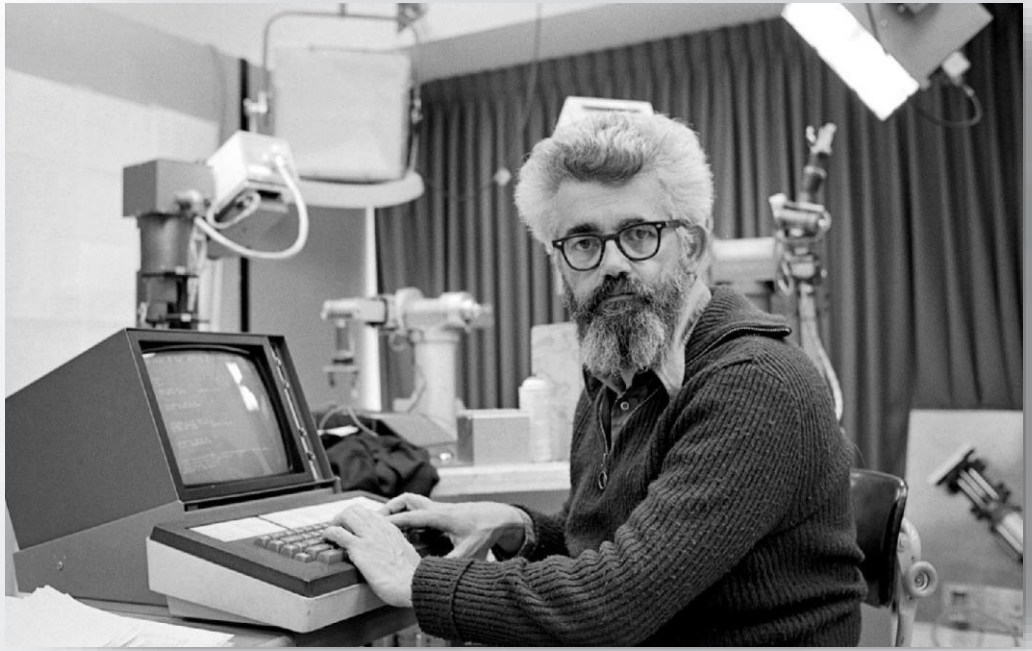
Edge Multicloud

Futuro do Trabalho

Generative AI

DELL Technologies

Introdução à IA - Nascimento e Crescimento



John McCarthy @ Dartmouth College - Hanover, NH - 1956

Generative AI

Generative AI refere-se a técnicas de inteligência artificial que aprendem uma representação de artefatos a partir de dados e os usam para gerar artefatos novos, completamente originais em escala, que preservam uma semelhança com os dados originais.

Generative AI permite que os computadores gerem variações de conteúdo totalmente originais (incluindo imagens, vídeo, música, fala e texto). Ele pode melhorar ou alterar o conteúdo existente, e pode criar novos elementos de dados e novos modelos de objetos do mundo real, como edifícios, peças, medicamentos e materiais.

IA Tradicional vs IA Generativa

A IA Tradicional visa resolver problemas específicos analisando dados existentes e fazendo previsões, enquanto a IA Generativa cria conteúdo novo e original com base em padrões aprendidos. A IA Tradicional é adequada para tarefas como classificação, recomendação e tomada de decisão, enquanto a IA Generativa libera o potencial para exploração criativa e geração de conteúdo.

AI TIMELINE

Desde a metado do século 20 estamos avançando no uso de IA.

Rule Based Systems

- A primeira fase do desenvolvimento da IA abordou a criação de sistemas baseados em regras
- Os especialistas codificaram seus conhecimentos em um conjunto de regras para o computador seguir.
- Esses sistemas eram limitados em sua capacidade de aprender com novos dados ou se adaptar a novas situações.

Machine Learning

- A próxima fase do desenvolvimento de IA abordava o uso de algoritmos de aprendizado de máquina para treinar computadores a reconhecer padrões em dados e fazer previsões ou decisões com base nesses padrões.
- Esta fase viu o desenvolvimento de algoritmos como árvores de decisão, regressão logística e redes neurais.

Deep Learning

- Esta fase da IA aborda o aprendizado profundo.
- O aprendizado profundo é um subconjunto do aprendizado de máquina que usa redes neurais com várias camadas para reconhecer padrões complexos nos dados.
- Esta fase tem sido eficaz no processamento de imagens, vídeos e dados de linguagem natural.

Generative AI

- A atual fase aborda a IA generativa.
- Gerador de IA usa algoritmos de aprendizagem profunda para gerar conteúdo, como imagens, vídeos,música e até texto que se assemelha muito aos padrões dos dados originais.
- Esta fase tem enorme potencial para criar novos tipos de conteúdo e gerarnovos insights e previsões com base em grandes quantidades de dados.

AI TIMELINE

Uma série de avanços cada vez mais frequentes que dão sentido à linguagem natural

2010 — Near-Perfect Translation of Natural Language

Por volta de 2010, pesquisadores de IA que trabalham com tradução de linguagem natural descobriram que modelos expostos a grandes quantidades de texto produziam resultados muito melhores do que modelos usando regras gramaticais de cima para baixo.

2014 — Mastering the Meaning of Words

Em 2014, os modelos de linguagem começaram a dar sentido ao significado das palavras em uma língua natural, analisando o contexto em que a palavra apareceu.

2017-2022 — Large Language Foundation Models

Os avanços feitos de 2017 para 2022 resultaram em modelos de linguagem que podem servir de base para a personalização.

Criar modelos básicos tem um custo ALTO, mas, uma vez criados, eles **podem ser personalizados usando uma pequena quantidade de dados adicionais** para obter desempenho de última geração em novas tarefas sem investimento significativo.

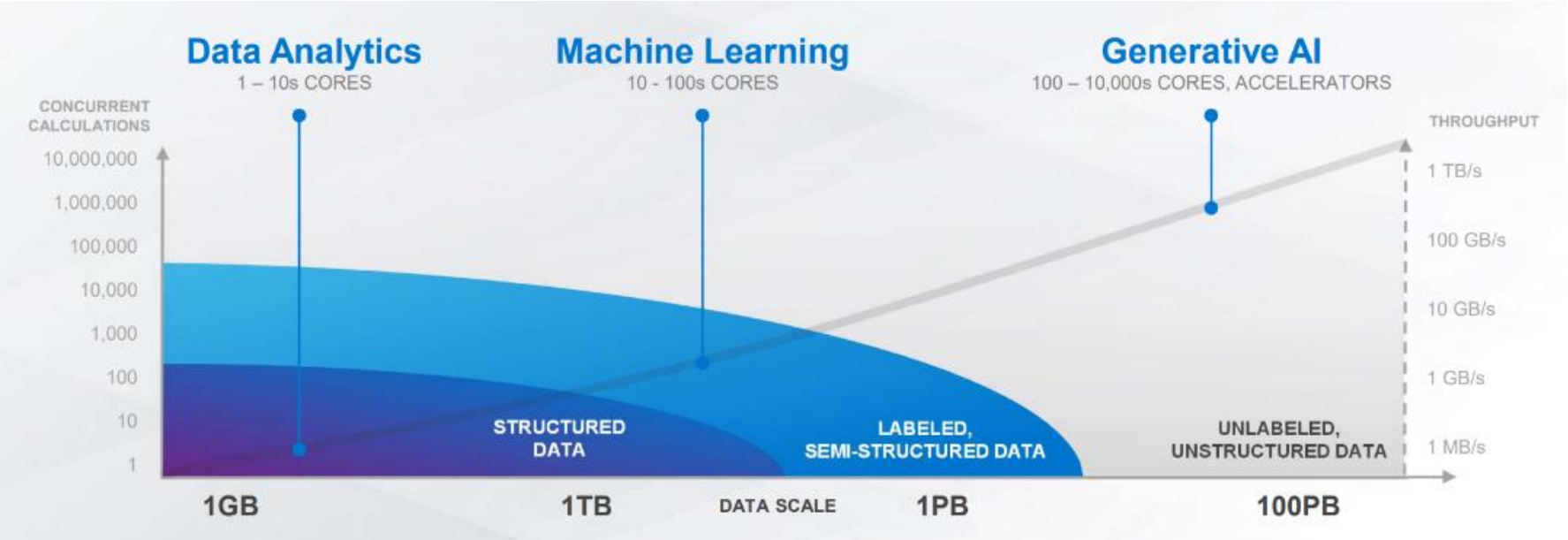
2022 — Conversational AI Large Language Models

2022 marcou a chegada do **ChatGPT**, que deu aos usuários uma maneira simples de acessar um modelo básico de linguagem grande.

O brilho do ChatGPT não está apenas no modelo incrivelmente avançado em sua essência
Igualmente, é a capacidade de explorar esse modelo conversando com ele em linguagem natural.

Generative AI Evolution

Dell Technologies & Nvidia provides expertise and guidance to build, train, and model-tune Generative AI.



Absolutamente todas as pessoas e negócios do mundo serão profundamente impactados pelo o uso da Inteligência Artificial.

Agência d
modelos v

A Deep Agency
utilizada para tira

tar à
ificial

Inteligên
predição
câncer

Estudo na Faculdade de Saúde Pública da USP revela benefícios na utilização de inteligência artificial para auxiliar tomadas de decisões médicas

riar
naval

InCor incorpora tecnologia de imagem da Abbott que usa Inteligência Artificial para angioplastias

Inovação

Audi usa inteligência artificial para produzir rodas de carro; entenda como funciona

Programa de IA é combinado por dois sistemas que competem para criar uma imagem do zero; montadora quer levar tecnologia para outras áreas

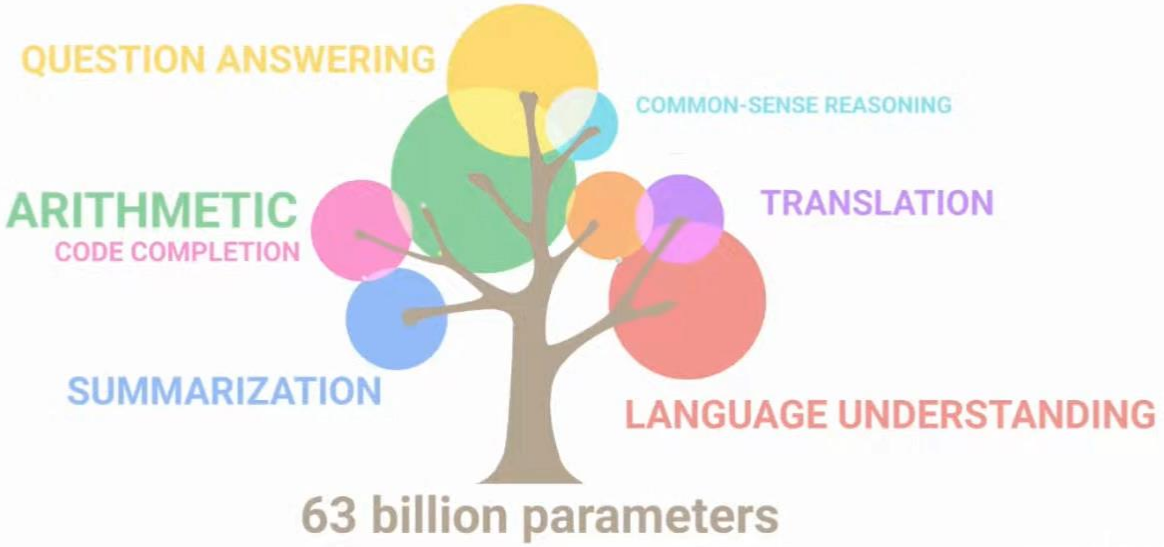
Extensão para Chrome usa ChatGPT para responder e-mails do Gmail

"Artificial intelligence is the most important advancement of technology in decades. That will change the way how people work learn, travel, get medical assistance and communicate with each other others" Bill Gates

"iPhone moment for AI" Jensen Huang, Nvidia´s CEO about rise of Generative AI

DELL Technologies

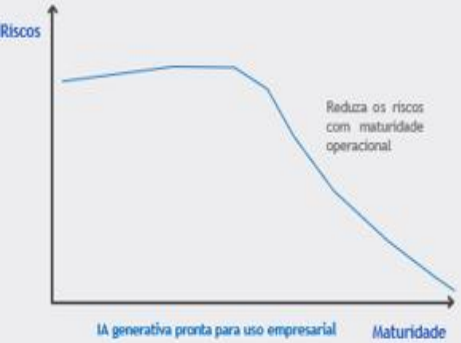
DELL Technologies



A Função Vital dos Dados na IA Generativa

Gerenciando os riscos e aprimorando o valor

Ao iniciar sua jornada, você precisa fazer investimentos correspondentes em tecnologia e treinamento para reduzir os riscos e oferecer mais controle e valor crescente à sua organização.



Bring GenAI to Your Data

Os Dados são o Grande Diferencial !

Dados e a qualidade desses dados são elementos cruciais para projetos de IA Generativa e quanto mais se utiliza os próprios dados, maior será o valor dos seus projetos para a organização

Noções básicas sobre os riscos de dados na IA Generativa

- Perda de propriedade intelectual
- Vazamento de dados
- Problemas de privacidade
- Violações de conformidade
- Perda de credibilidade e integridade
- Viés
- Violação de IP

Problemas e Desafios comuns relacionados a adoção de IA

- ✓ Confusão sobre arquitetura
- ✓ Preparação de Dados, Qualidade e Organização
- ✓ Infraestrutura Compartilhada
- ✓ Ferramentas e Modelos
- ✓ Gestão de Recursos e Conhecimento
- ✓ Acumulação de Silos e Recursos

Problemas e Desafios comuns relacionados a adoção de IA

- ✓ Comportamento e Viés Tóxico
- ✓ Alucinação e Respostas Incoerentes
- ✓ Conteúdo Congelado
- ✓ Conhecimento Específico
- ✓ Complexidade dos Algoritmos
- ✓ Infraestrutura de Larga Escala – Treinamento/Inferência

DELL Technologies

Coesão – Coerência – Contexto

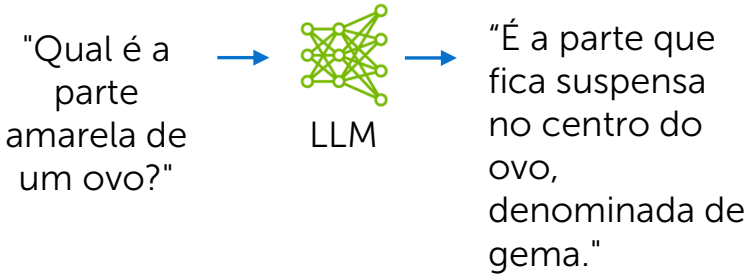
Os chatbots de IA Generativa, como o ChatGPT, utilizam um modelo de linguagem de larga escala (LLM) com capacidade de processamento de linguagem natural (PLN) para compreender e gerar textos coesos e contextualmente relevantes.

Para realizar essa tarefa, os modelos precisam ser treinados em grandes volumes de dados com o objetivo de aprender relações e padrões linguísticos.

A partir dessa base de informações, as IA´s geram respostas em um processo de predição que busca determinar a continuação mais provável de um texto. Isso envolve analisar o contexto e o que foi dito anteriormente para elaborar uma resposta coerente e natural — sempre de acordo com os dados em que o modelo foi treinado.

LLM´s – Exemplo de Resposta sem Contexto

Resposta Sem Contexto (Coesão/Coerência)



A personalização é necessária para lidar com tarefas específicas do negócio - Contexto

Resposta Sem Conetxto (com Coesão/coerência)

"Qual é a parte amarela de um ovo?"

LLM

"Essa é a parte que fica suspensa no centro do ovo."

P-Tuned Response

"Qual é a parte amarela de um ovo?

+

Prompts treinados (contexto)

LLM

Chatbot de Nutrição
"A parte amarela de um ovo é a gema. Contém gordura, colesterol e proteínas."

Chatbot Pré-natal
"A parte amarela de um ovo é rica em colina, que é importante para o desenvolvimento do cérebro fetal"

Chatbot culinário
"A parte amarela de um ovo é usada para fortificar molhos e molhos para salada, e para emulsionar ingredientes ricos e gordurosos, como óleo e manteiga"

Casos de uso corporativos exigem conhecimento específico do domínio

Codifique e incorpore sua IA com as informações em tempo real da sua empresa para fornecer as respostas mais recentes

Resposta de base/modelo personalizado

"Qual era a pressão do tanque às 23h de ontem à noite?"

"Fui treinado há 2 meses e não tenho os dados atuais"

Modelo com técnicas RAG/C-RAG

"Qual era a pressão do tanque às 23h de ontem à noite?"

Dataset (RAG/CRAG)

"A pressão às 23h da noite de ontem era de 345 psi"

70%

Dos dados corporativos é inexplorado
Desbloqueie muitas novas oportunidades para maior inteligência

Retreinamento menos frequente
Economia significativa de tempo e custo a longo prazo para manter LLMs

Modelos de Implementação de IA: avaliando as vantagens e desvantagens de custo e valor

Os três primeiros tipos de modelos de implementação abaixo são o que 90% das organizações estão implementando agora. O modelo de IA que você escolher dependerá do nível de preparo da ciência de dados de sua organização, dos padrões de implementação e das implicações de cada um deles. O aumento de modelos é um ótimo ponto de partida, e a maioria das organizações migra dessa fase para modelos de ajuste fino.

Inferência simples com um modelo pré-treinado

Conhecido como "prompt engineering", que faz uma pergunta a um modelo pré-treinado e recebe um resultado. O uso do ChatGPT é um exemplo de inferência simples.

Aumento de modelo

Inferência mais acesso a seus dados, para o qual o RAG é um caso de uso comum. É um acesso fácil para tornar um modelo de IA generativa mais inteligente com base em seus dados.

Modelos de ajuste fino

Envolve alterar a ponderação de um modelo e informá-lo com seus dados. Aproveitado à medida que as organizações começam a dimensionar seus modelos, ele oferece melhores resultados, mas também exige mais esforço para ser estabelecido.

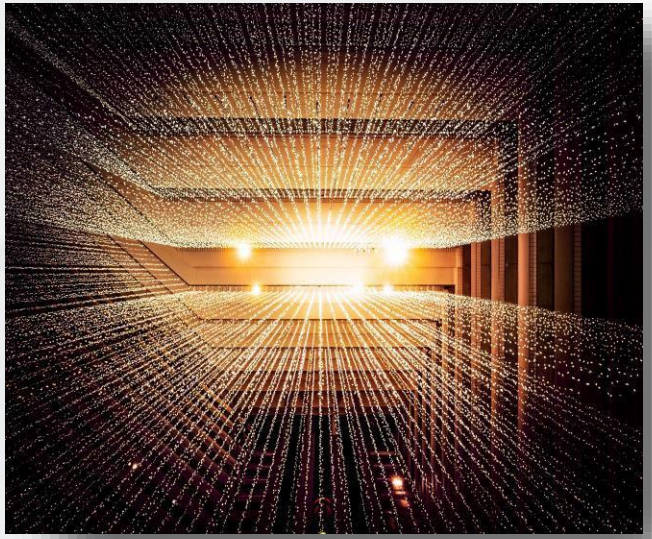
Treinamento de modelos

Inclui a criação de um modelo muito específico e o treinamento com um conjunto de dados. Geralmente, isso requer a maioria dos recursos e trabalho e normalmente é reservado para resolver os problemas mais complexos.

	Modelo pré-treinado	Aumento do modelo (RAG, PAL)	Modelos de ajuste fino	Treinamento de modelos
Definição do modelo	Modelo inicialmente treinado ou pré-treinado  Modelo do transformador			
Esforço				
Custo				
Valor e diferenciação				
Integração de dados				
Infraestrutura	Client – servidor	Client – servidor	Otimizado para a GPU	Implementação de GPU grande
Habilidades	Operações de TI	Desenvolvedor	Cientista(s) de dados	Cientista(s) de dados

Fundamentos para Construção - IA Generativa

- ✓ Bons Modelos de Dados (LLM´s) – Qualidade dos Dados
- ✓ Acesso a Grandes Quantidades de Dados Brutos (ETL)
- ✓ Infraestrutura de Alto Desempenho
- ✓ Arquitetura – Robusta – Escalável – Orientado ao Caso de Uso
- ✓ Custo Acessível para Processamento Rápido e Eficiente
- ✓ Software de Ciência de Dados - Framework
- ✓ Segurança – Guardrails
- ✓ Expertise

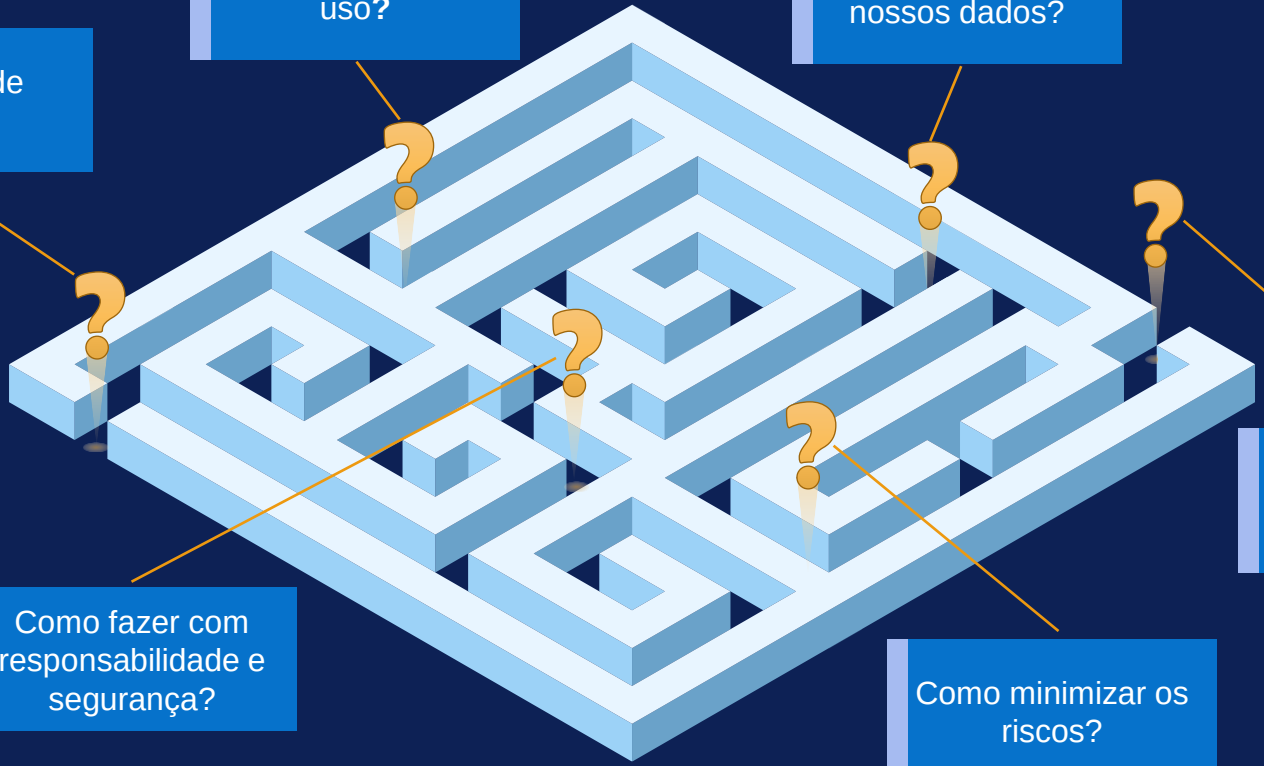


A jornada para o GenAI tem muitas perguntas que precisam ser respondidas

Quais os casos de uso**?**

Como usar melhor nossos dados?

Como e onde começar?



Como envolver a todos?

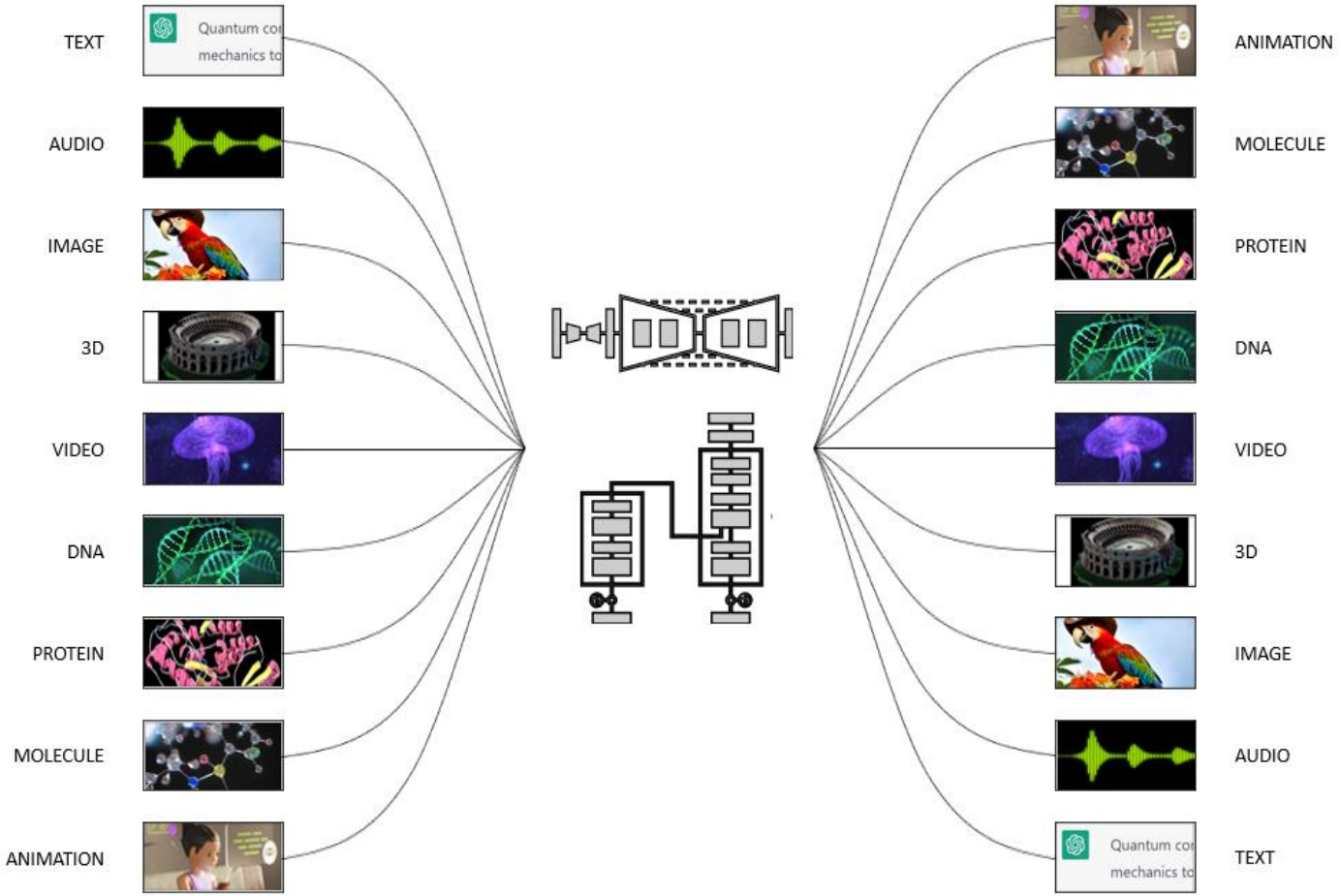
Como fazer com responsabilidade e segurança?

Como minimizar os riscos?

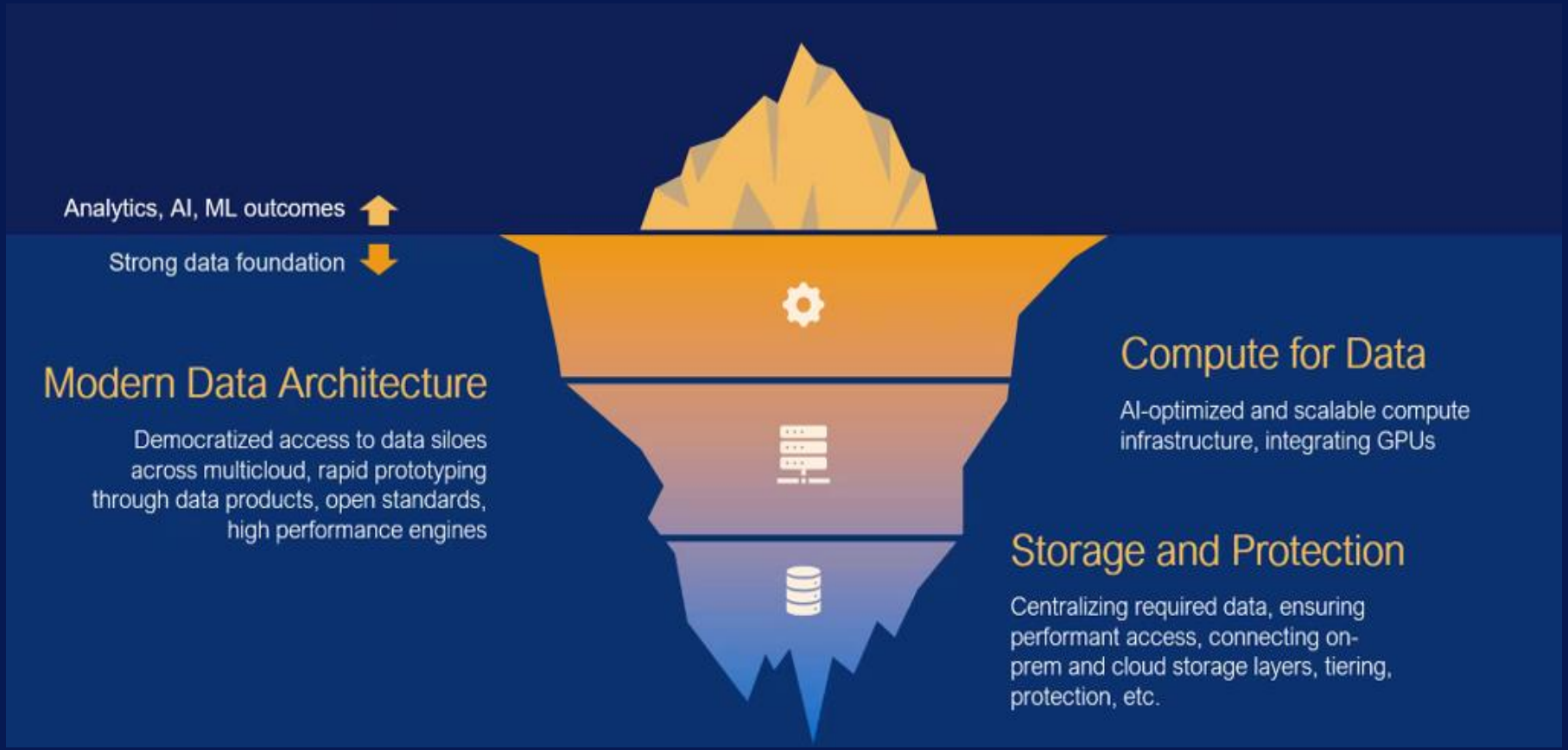
What´s Generative AI

Bring Your Own ChatGPT

Bring GenAI to Your Data

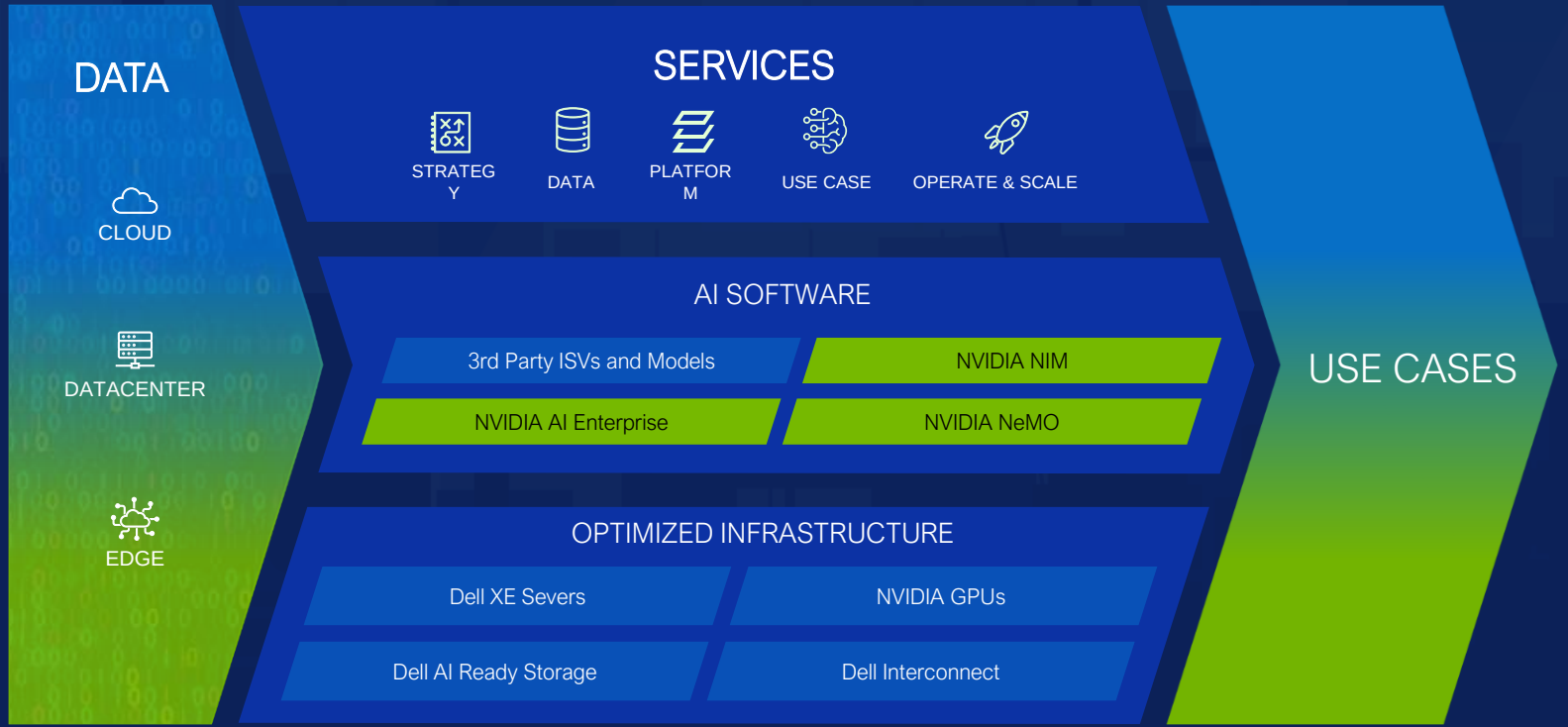


Uma base sólida e estruturada é Fundamental para IA e Analytics



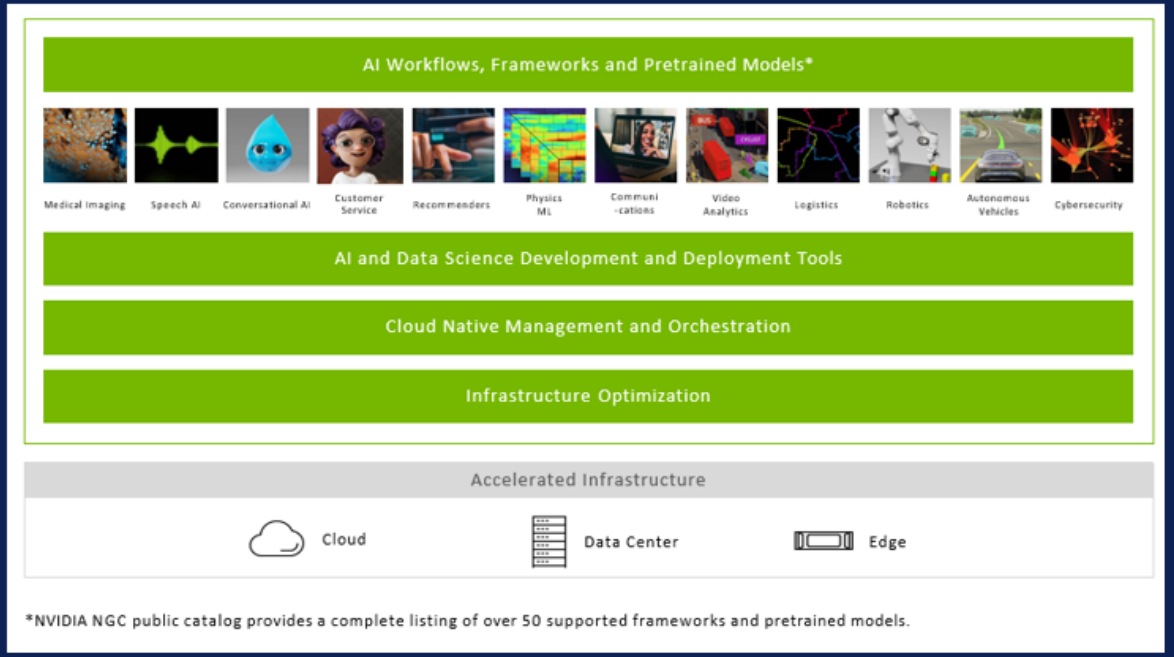
The Dell AI Factory with NVIDIA

INDUSTRY'S FIRST END-TO-END OPERATIONAL AI PLATFORM



NVIDIA AI ENTERPRISE SOFTWARE

The Proven Standard for Enterprise AI



*NVIDIA NGC public catalog provides a complete listing of over 50 supported frameworks and pretrained models.

NVIDIA Base Command

Provides AI workload and cluster management, libraries that accelerate compute, storage, and networking infrastructure, and an operating system optimized for AI workloads.

Optimize Performance

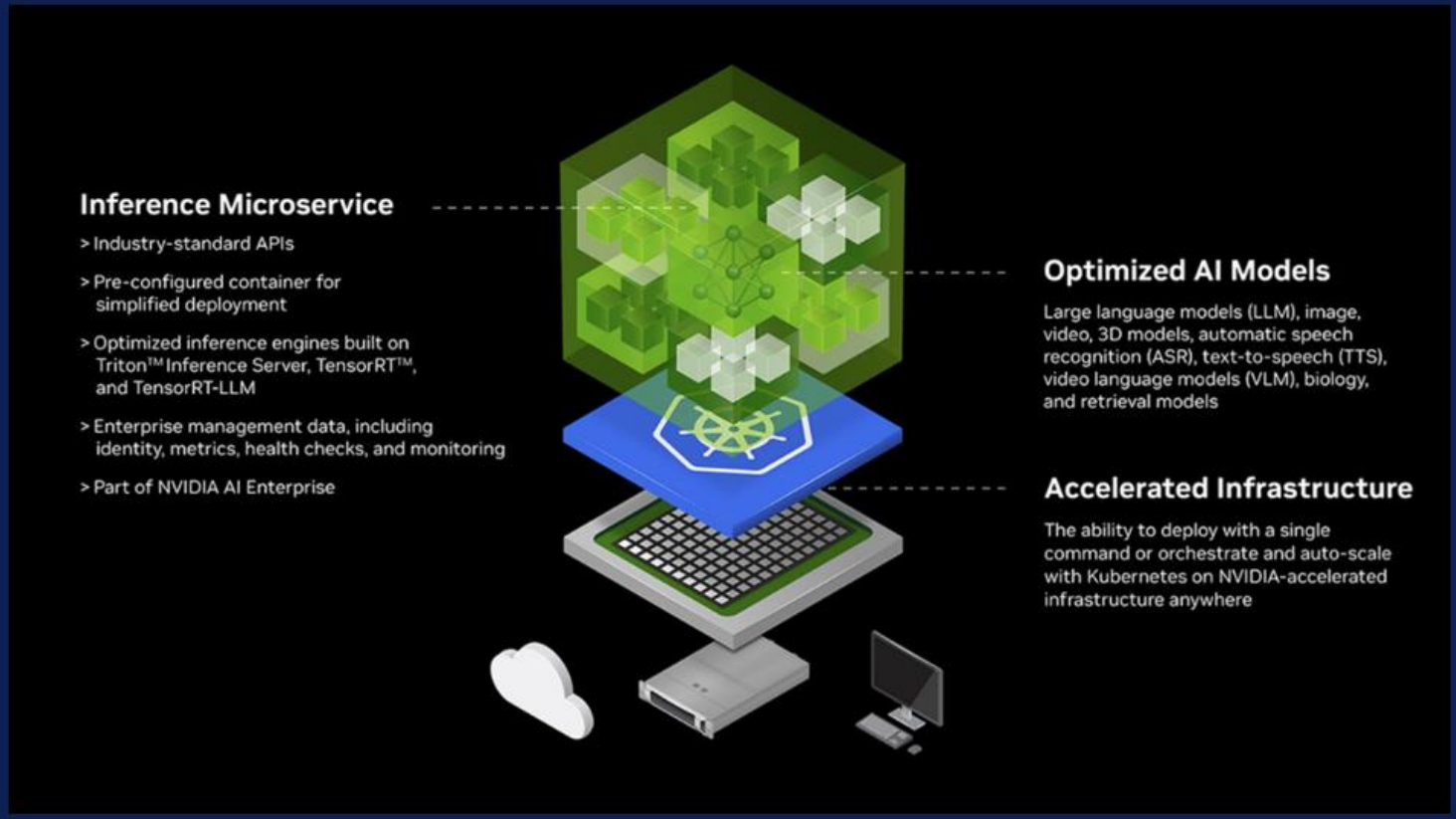
NVIDIA NIM and CUDA-X microservices provided an optimized runtime and easy to use building blocks to streamline generative AI development

NVIDIA API Catalog

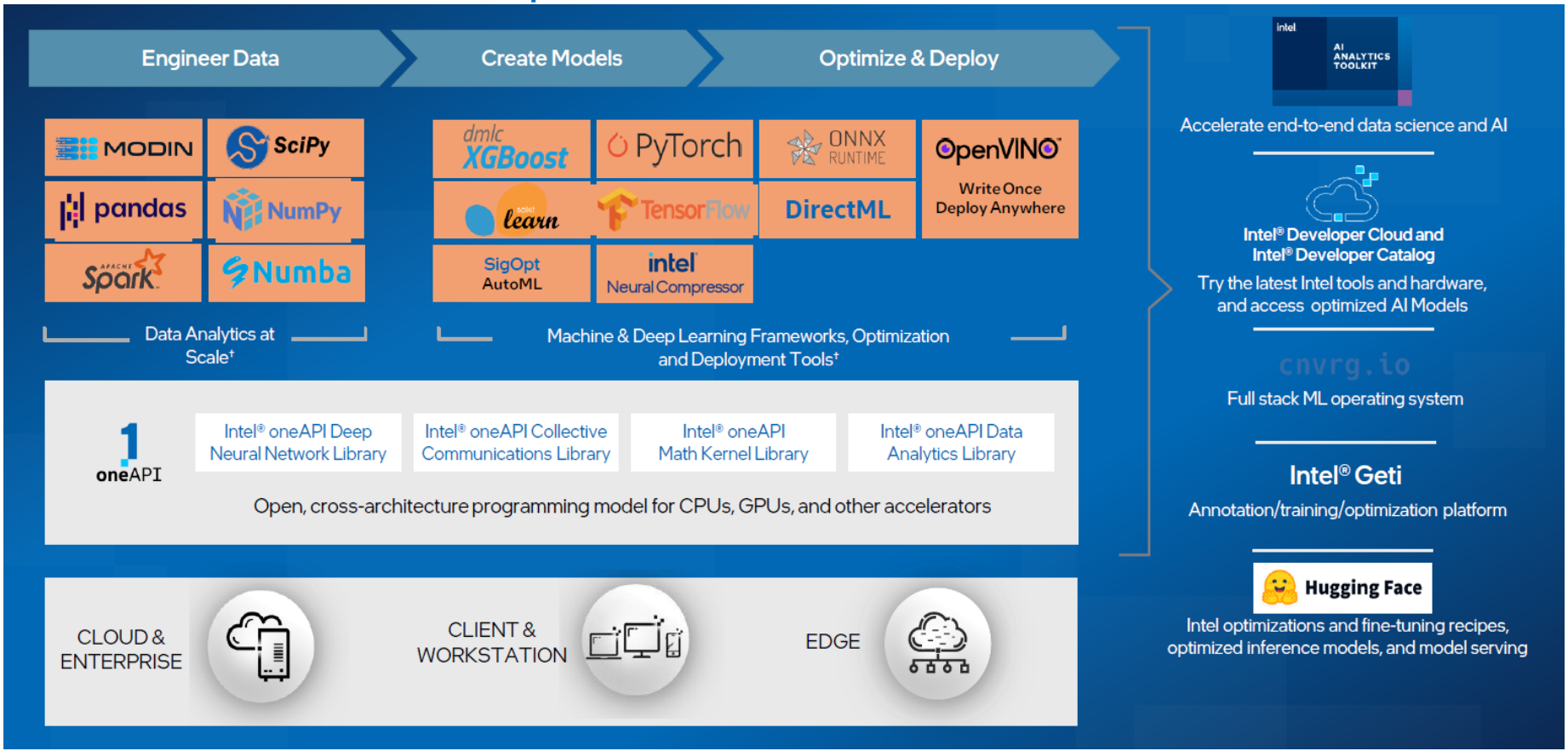
Pre-trained models help deliver results faster, more cost-efficiently

NVIDIA NIM - Inference Microservices for Generative AI

The best way to deploy AI models on Dell AI Factory



O Portfólio INTEL para IA



Construindo Pipelines de Dados de IA

Establish Strategy

- Define Vision / Solution Arch
- Identify Use Cases / Models

Prepare Data

- Data Management
- Data Engineering

Gen AI Platform

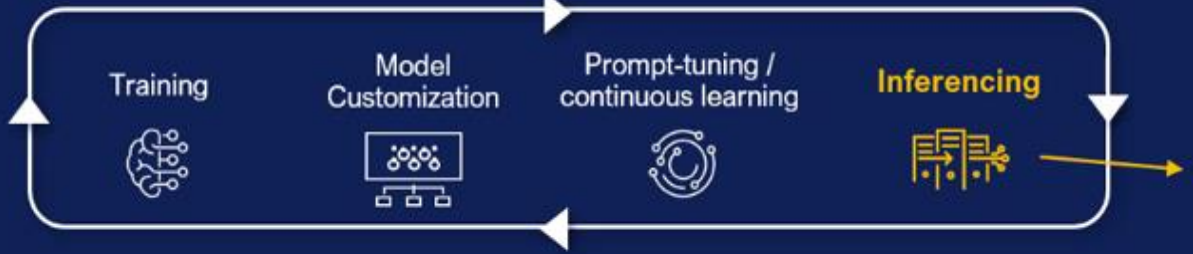
- Gen AI Software
- Gen AI Hardware

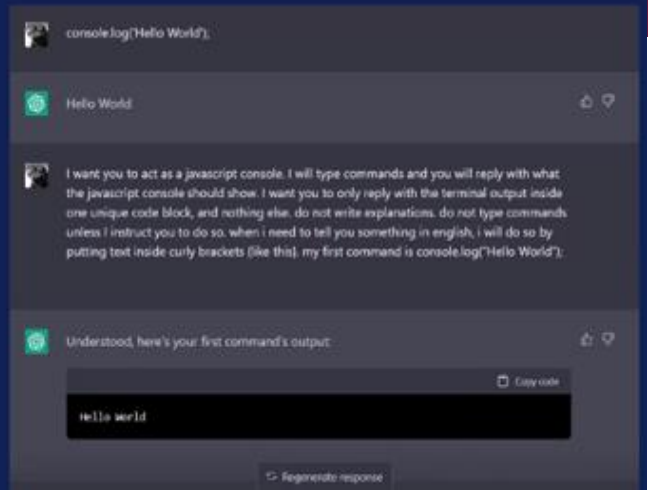
Deploy & Test Model

- Language Modeling
- Implement AI Use Cases

Operate & Scale

- AI/Data Governance
- Gen AI Operate / Optimize





DELL Technologies

Dell AI Factory Accelerates AI Adoption

SIMPLE

Comprehensive AI optimized solutions make AI easier, augmenting skills gaps and addressing data readiness

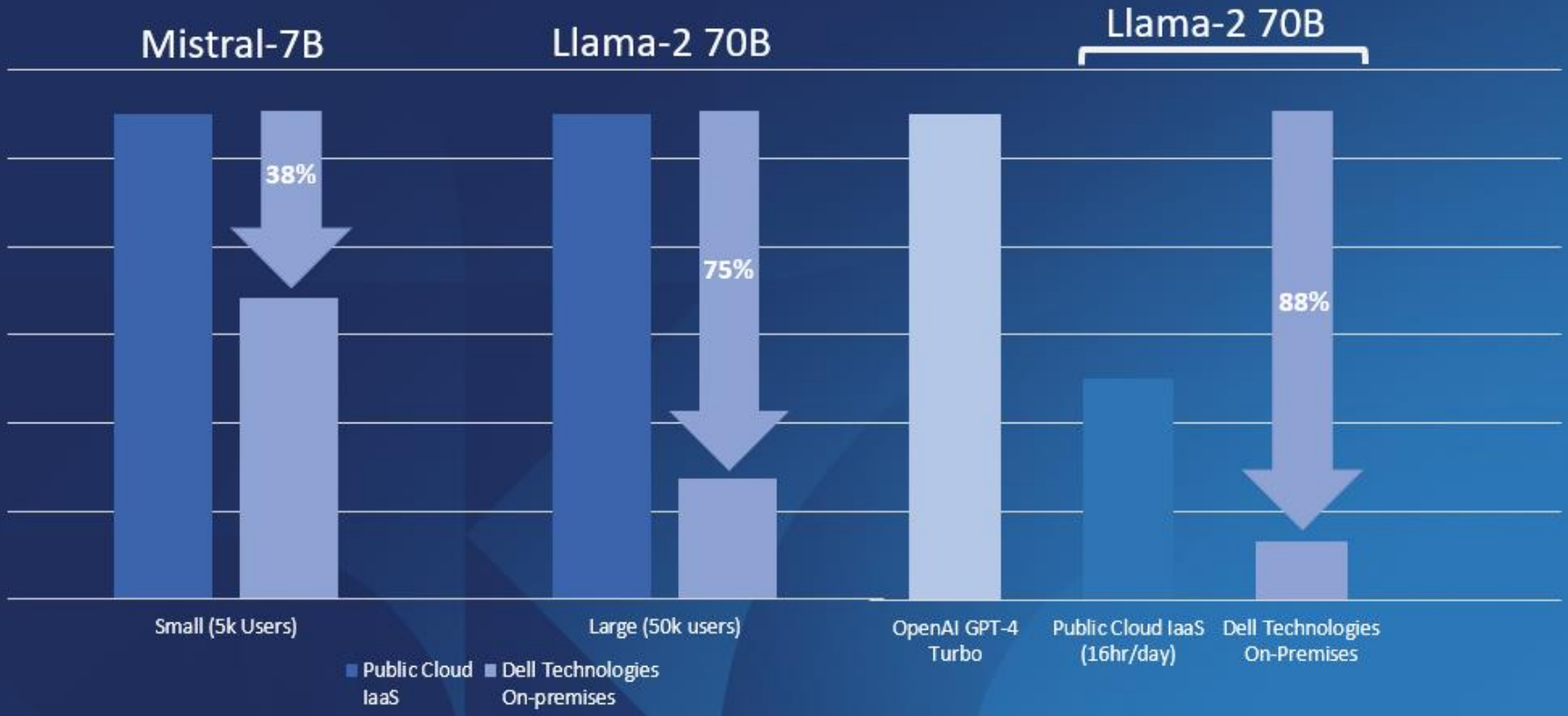
SECURE

Trusted infrastructure, full-stack solutions and services ensure valuable data is protected, managed and accessible

ECONOMICAL

Right sizing AI investment to fit their use cases requirements with Dell extensive AI portfolio to lower costs.

Cost-effectiveness of Dell Technologies Solutions for Gen AI



One-stop shop for AI

3

Dell Professional Services

Forbes & Forrester: Awarded consulting organization
Proven methodology
Accelerate AI / Gen adoption at scale
Flexible services bundles

2

AI / GenAI toolkit bundle

Complete set of tools
Opensource ecosystem
Intuitive IDEs
Customizable sample use cases

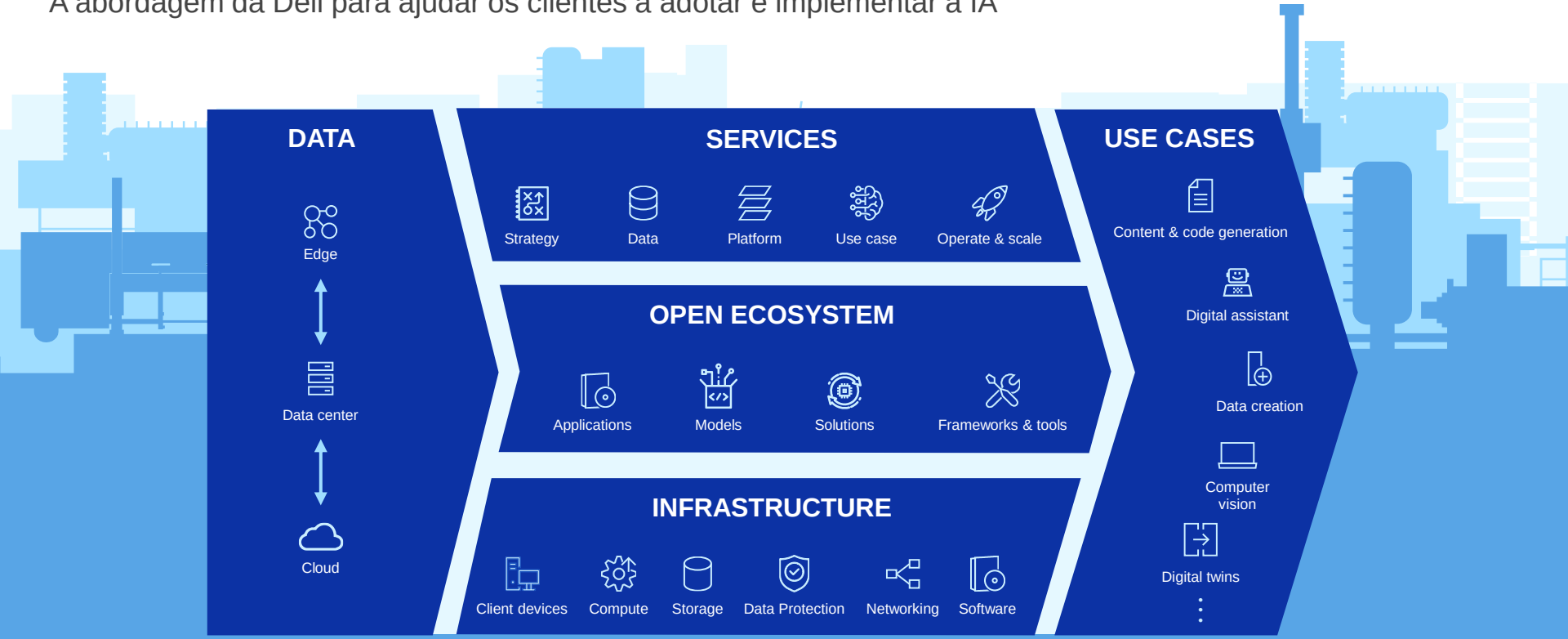
1

Dell Validated Design

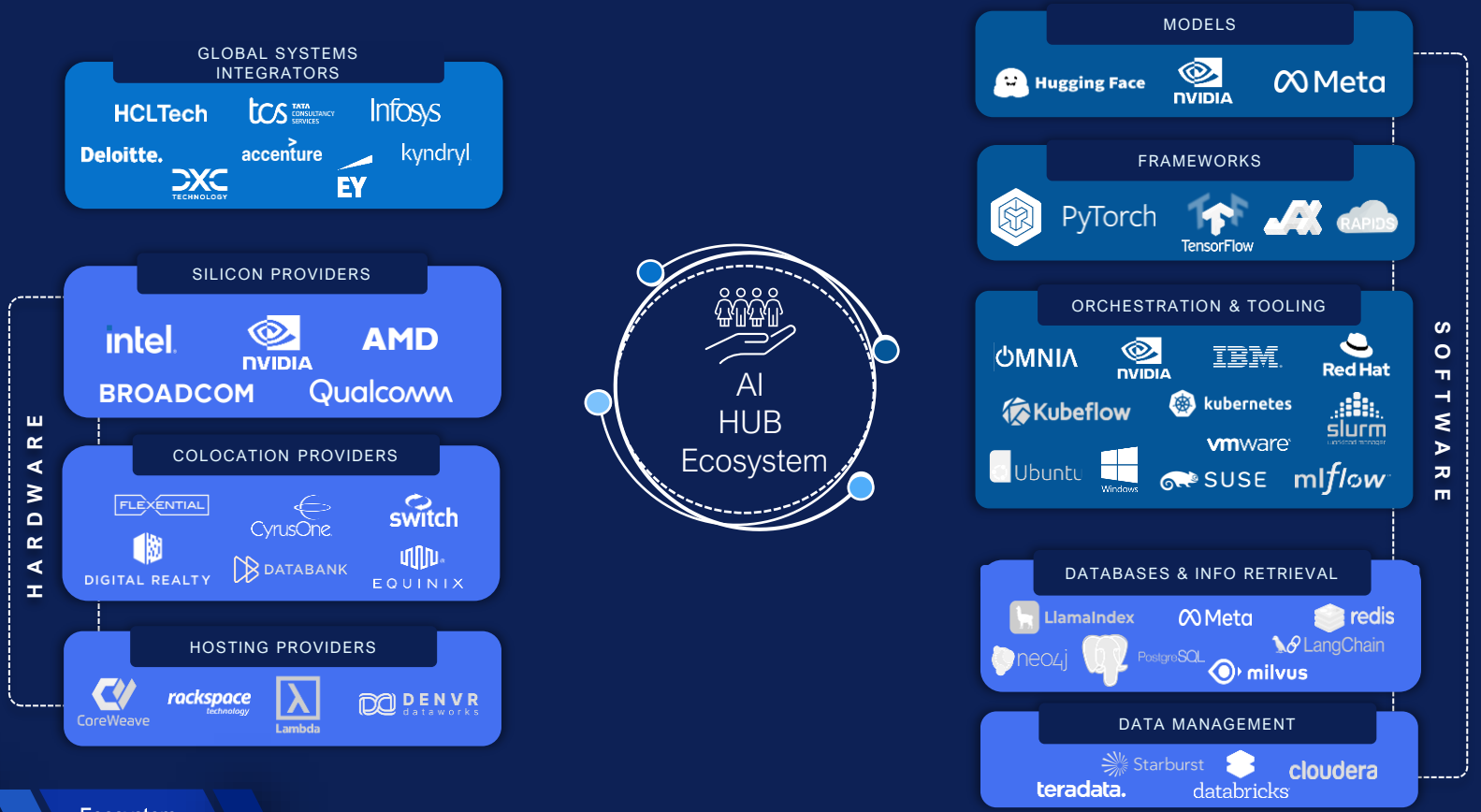
Validated architecture
Accelerated deployment
On-prem & APEX

DELL AI FACTORY E USE CASES

A abordagem da Dell para ajudar os clientes a adotar e implementar a IA



Open Ecosystem HUB for AI Innovation



Ecosystem

DELL Technologies

Right-sizing your AI investment

Focus Area

Business strategy and operations management | Product innovation and R&D | Manufacturing and supply chain | Marketing, sales, and customer service | IT, HR ,and Finance | Users | Location | Dataset | •••

Use Case

- Content creation
- Digital assistant
- Natural language search
- Design and data creation
- Code generation
- Document automation

Patterns

- Pre-trained model/ inferencing
- Model augmentation
- Fine-tuning
- Model training

AI Models

Open Ecosystem HUB for AI

Infrastructure

Laptops

Workstations

Dell Validated Designs

Compute

Network

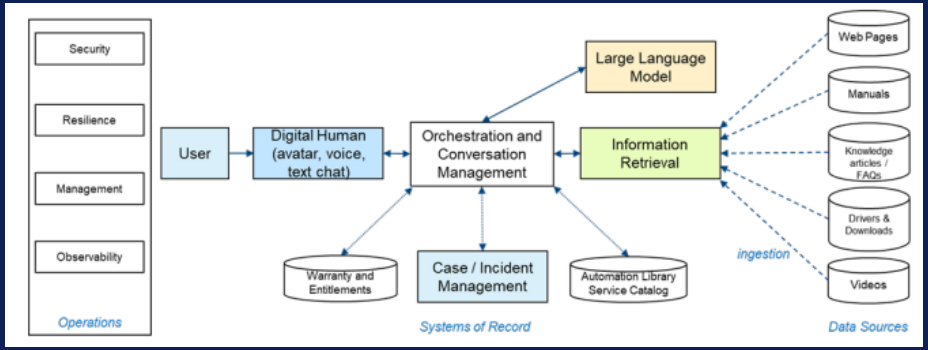
Storage

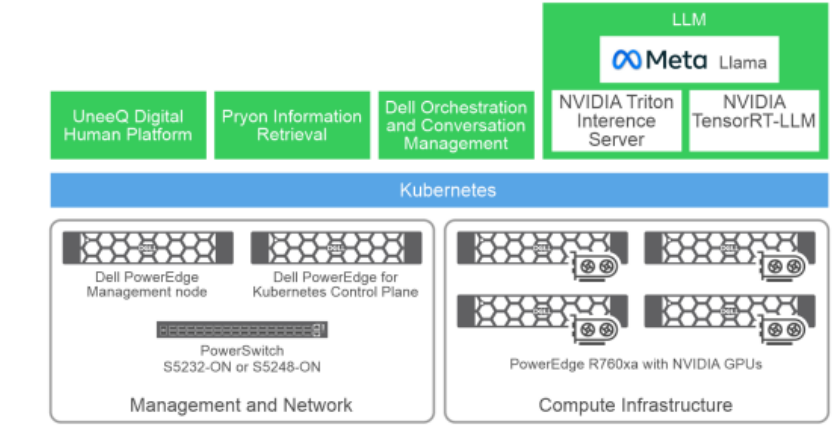
Professional Services

Strategy | Data | Deployment | Implementation | Residency

Infrastructure

Digital Assistant Conceptual Architecture





Orchestration and Conversation Management

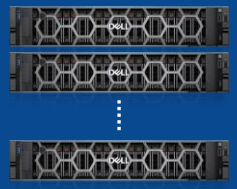
- **Speech to text (STT)—SST** is a technology that converts spoken language into written text. It is commonly used in voice recognition systems, transcription services, and various applications where it is more convenient or efficient for the user to speak rather than type.
- **Text to Speech (TTS)—TTS** is the opposite of speech to text. It converts written text into spoken words. It is often used in assistive technologies for visually impaired users, in navigation systems, or in applications for which audio content is more suitable. The text-to-speech component is only able to accept complete sentences (compared to streamed tokens) to achieve proper pronunciation. Our LLM observability must add a 'time to first sentence' metric to the standard set of LLM metrics.
- **Natural Language Processing (NLP)—**NLP is the interaction between computers and humans through natural language. The goal is to enable computers to understand, interpret, and generate human language in a valuable way. NLP performs several tasks such as language understanding, language generation, translation, and sentiment analysis. Generative LLMs perform most NLP tasks well, however our orchestration layer maintains the ability to perform classical NLP or keyword recognition tasks if needed.
- **Prompt engineering**—Prompt engineering is the process of designing and optimizing prompts to guide an AI model's responses effectively. It is a crucial aspect of interacting with AI models, especially in language models where the quality and structure of the prompt can significantly influence the model's output.
- **Retrieval Augmented Generation (RAG)—**RAG is a method used in AI language models in which the model retrieves relevant documents or information from a database before generating a response. This method allows the model to retrieve real-world facts and details, enhancing the accuracy and richness of its responses.
- **Translation—**The digital assistant solution can be equipped with the ability to translate text from one language to another. The task involves understanding the semantics and context of the source language and accurately reproducing the meaning in the target language. This process is crucially important in multinational environments in which original content is created in a single language and translated on demand by a digital assistant.
- **Integration hub—**An integration hub facilitates the seamless integration of various systems of record (see Systems of record) into the overall solution. It acts as a bridge, connecting disparate systems and enabling them to communicate and share data

Design Guide—Generative AI Digital Assistants in the Enterprise | Dell Technologies Info Hub

Optimized Enterprise-Scale Architecture

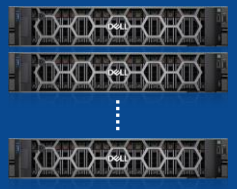
Orchestration & LLM Mgmt

Digital Assistant Renderer & Information Retrieval



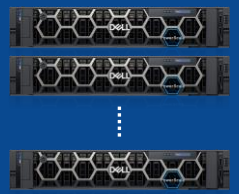
PowerEdge with NVIDIA GPUs

Large Language Model (LLM)



PowerEdge with NVIDIA GPUs

Data Sources



Dell Storage

Network fabric

PowerSwitch 25/100gbps

Management

Cluster Management

PowerEdge Servers

K8s Control Plane

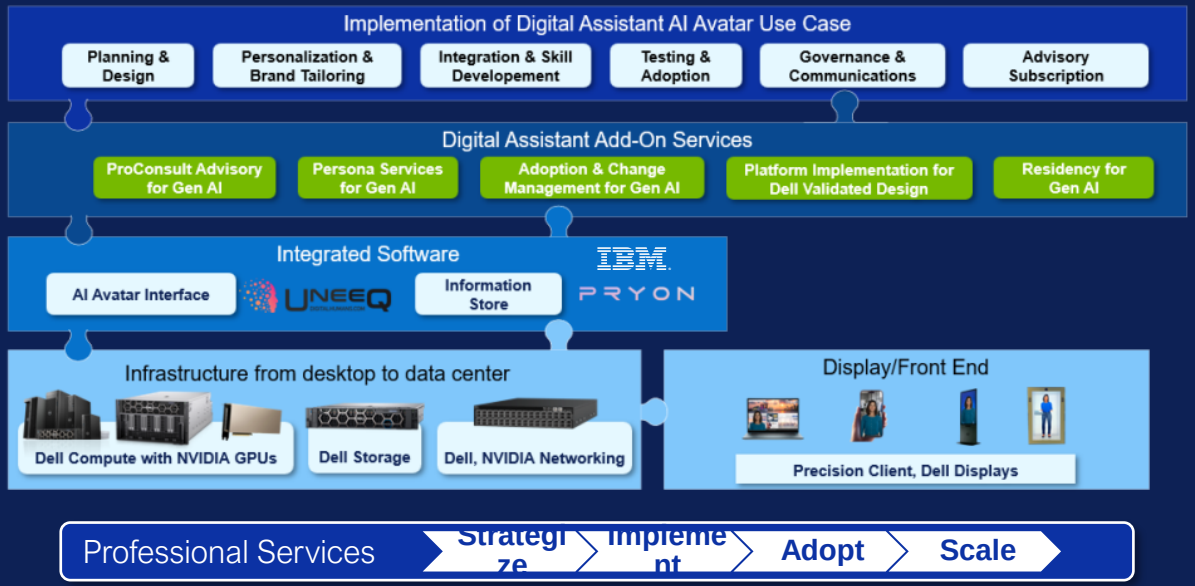
Application Integration

(EHR, ERP, EDI, etc. + APIs)

DELL Technologies

Dell Professional Services

Bringing the complete solution together, Dell Professional Services for Digital Assistants provide strategic guidance for customer use cases, personalized AI Avatar development, platform implementation with data source integrations and ongoing support to drive continued success, upskilling and optimization.



Implementation Services for Digital Assistants - AI Avatar Use Case:
Create a custom Digital Assistant featuring a humanistic AI Avatar interface that is tailored to customer use cases, integrated with their data, and personalized for their brand and audience.

Implementation Services for Digital Assistants - AI Avatar Platform:
Deploy the Digital Assistant on Dell hardware and integrated with their monitoring ecosystem, leveraging the Dell Validated Design for Digital Assistants.

Serving Generative AI future use cases, today

Manufacturing

- Structural analysis
- Fluid dynamics
- Packaging
- Impact modeling

Geosciences Oil & Gas

- Seismic
- Reservoir modeling

Government Classified

- Homeland security
- Defense
- Nuclear safety

Government & Academic Research

- Particle physics
- Life sciences
- Humanities
- Climate modeling

Financial Services

- Financial analytics
- High frequency trading
- Fraud detection

Electronics Design

- Electrical design
- Circuit verification
- Board layout

Life Sciences Pharma/Biotech

- Drug design
- Genomics
- Bio informatics

Media & Entertainment

- Rendering
- Gaming

Enterprise HPC

- Analytics – HPDA
- Marketing
- Simulations

DELL Technologies

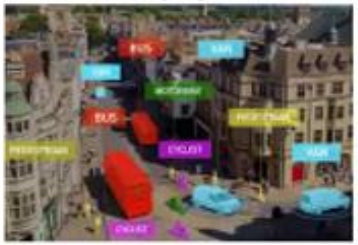
AI já está Transformando cada Segmento e Industria

CREDIT CARD FRAUD
1.1B Credit Transactions / Day

CONTACT CENTER AI
500M Calls / Day

MEETING TRANSCRIPTION
15B Meeting Minutes / Day

PUBLIC SAFETY
> 1B Smart City Cameras Deployed



PRODUCT RECOMMENDATIONS
300M E-commerce Visitors / Day

RETAIL ASSET PROTECTION
$275M Inventory Loss / Day

MEDICAL IMAGING
10M Diagnostic Scans / Day

INDUSTRIAL INSPECTION
94M Vision Sensors Installed by 2025

Why Dell for Generative AI?

Dell's unique approach leverages simplified, tailored and trusted solutions that combine your data with the power of Generative AI to drive innovation and tangible business value.

Dell is the #1 worldwide provider in AI server plus storage infrastructure[1]

We offer full-stack scalable Generative AI solutions, in collaboration with NVIDIA

Dell Services provide deep AI expertise at every stage to accelerate tangible time-to-value

[1] Source: IDC Semiannual AI Tracker: Worldwide Server and Storage Revenue, 2021 and 2022 H1