DELL Technologies

Dell Technologies Partner Webinar Series:

Powering the AI

Dell Technologies

Artificial Intelligence

Powering the AI

Stephen Tuomey
Datacenter Account Executive – Federal Distribution
stephen_tuomey@federal.dell.com

DELL Technologies

Dell Technologies

- Dell Technologies' Advantage
- Infrastructure – Modernizing
- What is Artificial Intelligence (AI)
- AI & Partner(s)
- AI in the Data Center
  - Compute
- Dell's Approach to AI
- Questions & Answers

Dell Technologies

Accelerating IT from Ideas to Innovation

Strategic Partnerships highlighting technology leaders and innovators

DELL DELL EMC

Tracewell

Hugging Face

RSA

Thales

Open Source

VMware

Microsoft

Artificial Intelligence

Cloud Native Applications

DevSecOps

Containers

Agile

Modern Data Center

Modernization

Private Cloud

AI Ready

Agile

Virtualization

Multicloud

SaaS

FedRamp

Cloud Portability

Hybrid Cloud

Edge

RealTime Analysis

Edge AI

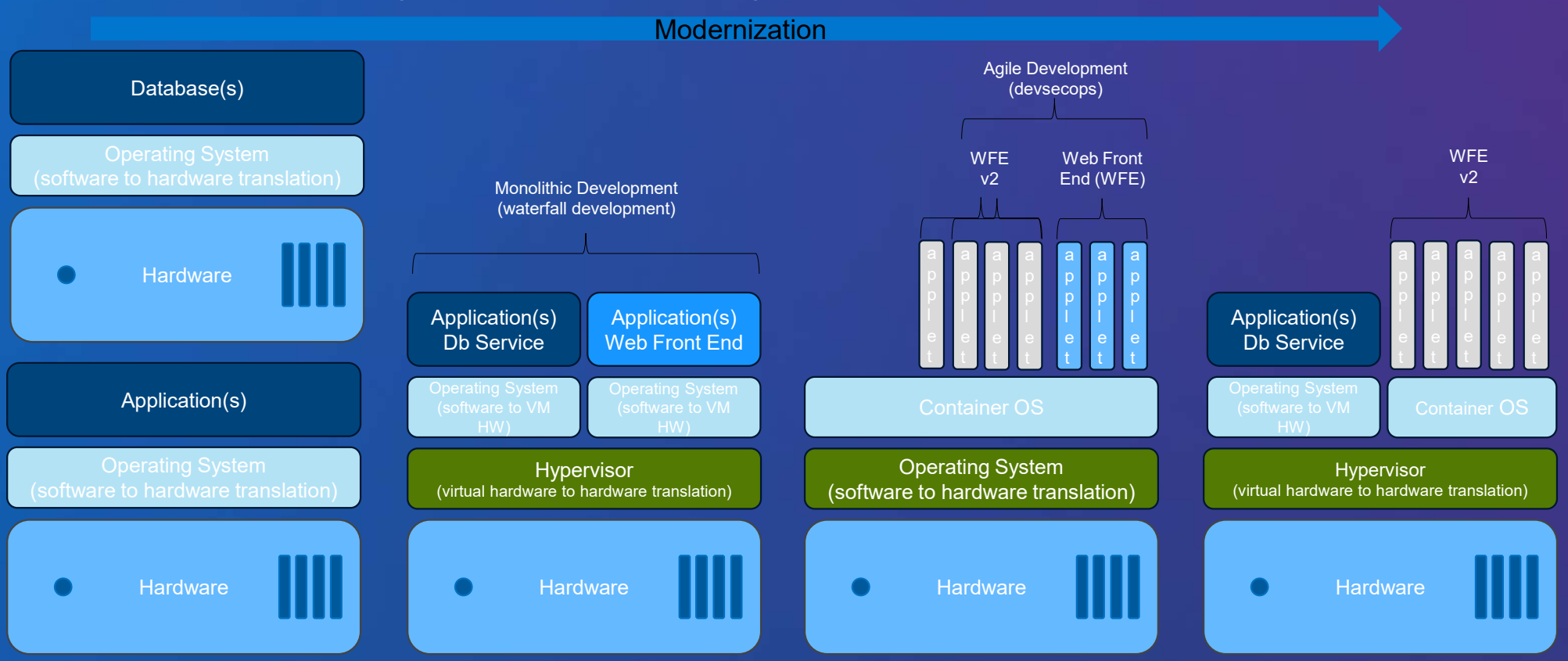
IoT

Telemetry

Zero Trust | Authentication | Security | Audit | Logging

DELL Technologies

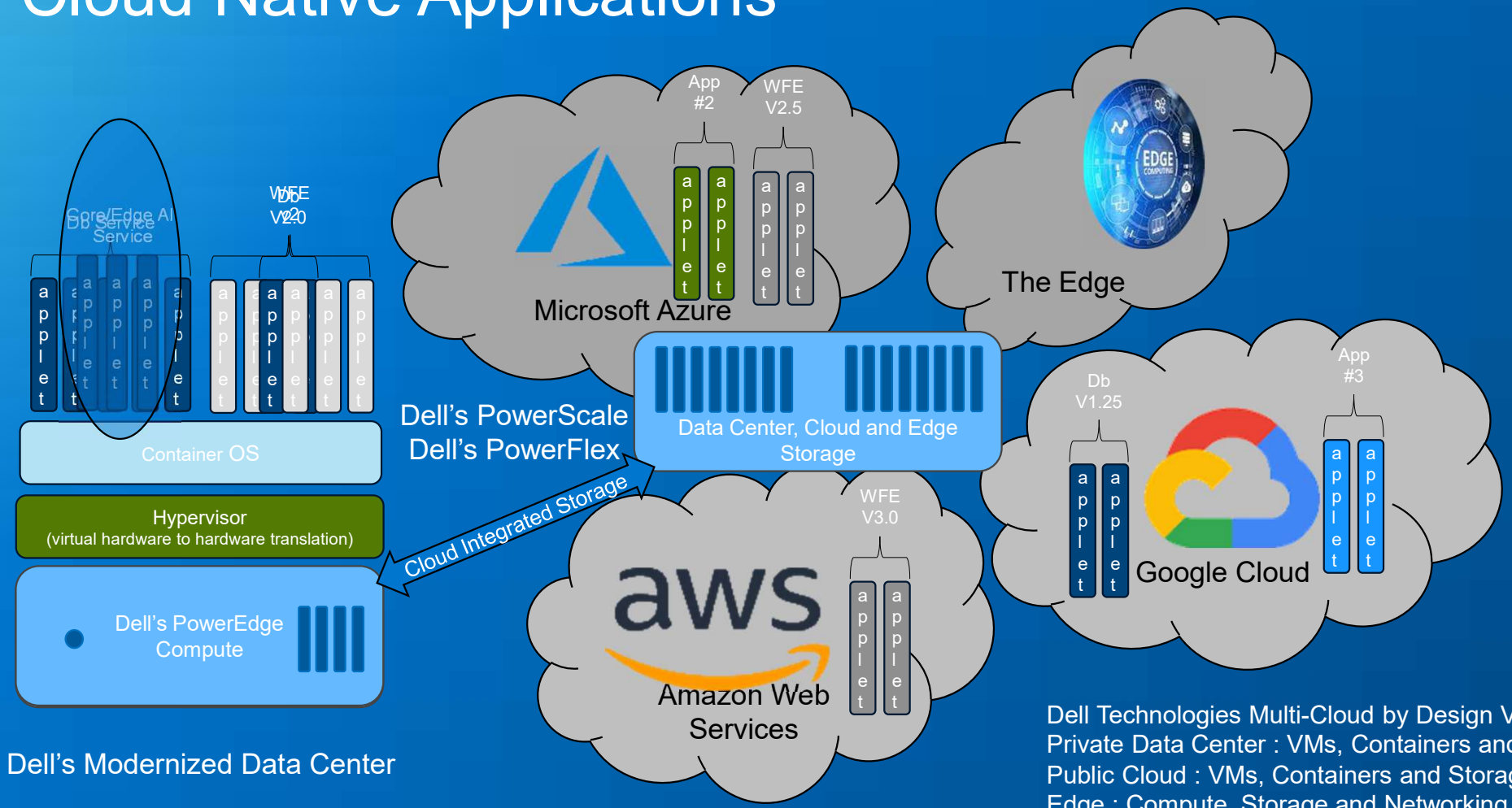
Information Systems – Physical to Virtual to Cloud



Hardware Based Systems | Virtual Machine Based Systems | Container Based Systems | Dell's HCI Systems

DELL Technologies FEDERAL PARTNER PROGRAM

Cloud Native Applications



Dell's Modernized Data Center

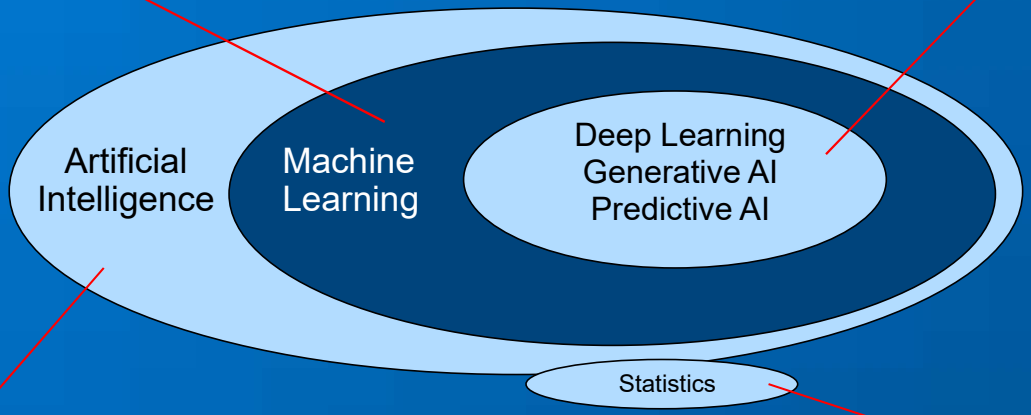
Dell Technologies Multi-Cloud by Design Vision:
Private Data Center : VMs, Containers and Storage
Public Cloud : VMs, Containers and Storage
Edge : Compute, Storage and Networking

DELL Technologies

Artificial Intelligence

Discerning patterns in the data without the need for explicit rules. Learning to generalize and extend the pattern recognition.

Best fit "formulas" based on related data and near related data "self generated formulas" with forecasting



Techniques patterned after the way multiple layers of neurons work in our brain

General "formulas" Advanced self generated Relationships based on related data and near related data

Imitating how humans think, exhibiting the ability to generalize.

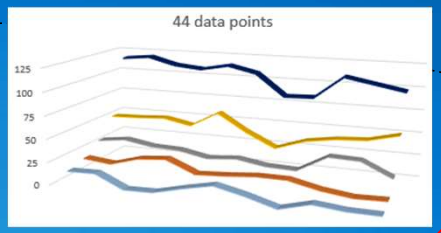
Best fit "formulas" with forecasting

Descriptive
Inferential

Best fit line with forecast

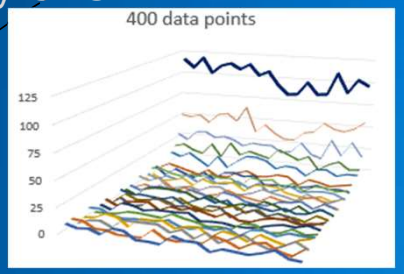
DELL Technologies

Artificial Intelligence – Data & Compute requirements



Multiple Gigabytes
Multi CPU
One+ GPU

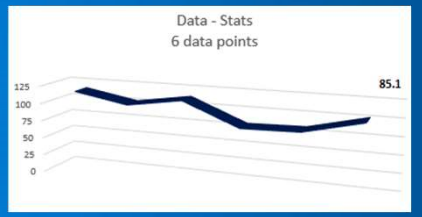
Multiple Terabytes
Many Multi CPU Servers
Many GPUs



Machine Learning

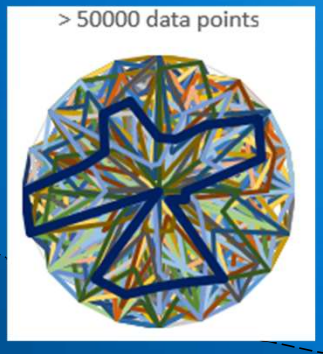
Artificial Intelligence

Statistics



Multiple Megabytes
Single CPU

Deep Learning
Generative AI
Predictive AI



Multiple Exabytes
Datacenter scale compute
Large quantity GPUs

DELL Technologies

Chat GPT – Generative AI

About 2nd / 3rd grade

About 5th grade

Focus areas	GPT 1	GPT 2	GPT 3	GPT 3.5	GPT 4
Launched	Jun'2019	Feb'2020	May'2021	Jan'2023	Nov'2023
Parameters	117 mm	1.5 bn	175 bn	3 variants: 1.3 bn, 6 bn, 175 bn	100 tn
Decoded layers	12	48	96	12	XXXXX
Hidden layer	768	1600	12288	XXXX	XXXXX
Content token size	512	1024	2048	4,096	8,192 (8k) 32,768 (32k)
No. of English words	XXXX	768	1,536-3,000	3,000- 8,000	24,000- 64,000
Feature	Add text here	Generate human-like text	Improved answers	State of art with lesser toxic outputs	Highly accurate output with data variants such as image, text, speech and numerical data
Model type	Uni-model	Uni-model	Uni-model	Multi-model	Multi-model

If creation of the "Algorithm" is the hard part.

- Why re-invent the "basic" LLM ?

Human - Education

Reading
Writing
Arithmetic
History
Sciences
…

i.e. Schooling

21 Years

2+ Years

Statistician

Some more Math
Linear Models
Programming
…

Programmer

Lots of Math ☺
Modeling
Programming
& examples
…

Help Desk

Help Desk Transcripts
Model Technical Info
Software Info
…

Basic schooling → Specialized training → Skilled employee

Large Language Models (LLMs)

Reading
Writing
Arithmetic
History
Sciences
…

i.e. Internet

~1 Year

Purchased Copy

Statistician

Some more Math
Linear Models
Programming
…

Programmer

Lots of Math ☺
Modeling
Programming
& examples
…

Help Desk

Help Desk Transcripts
Model Technical Info
Software Info
…

LLM creation and initial training → Specialized training → Skilled employee assistant

Base Model

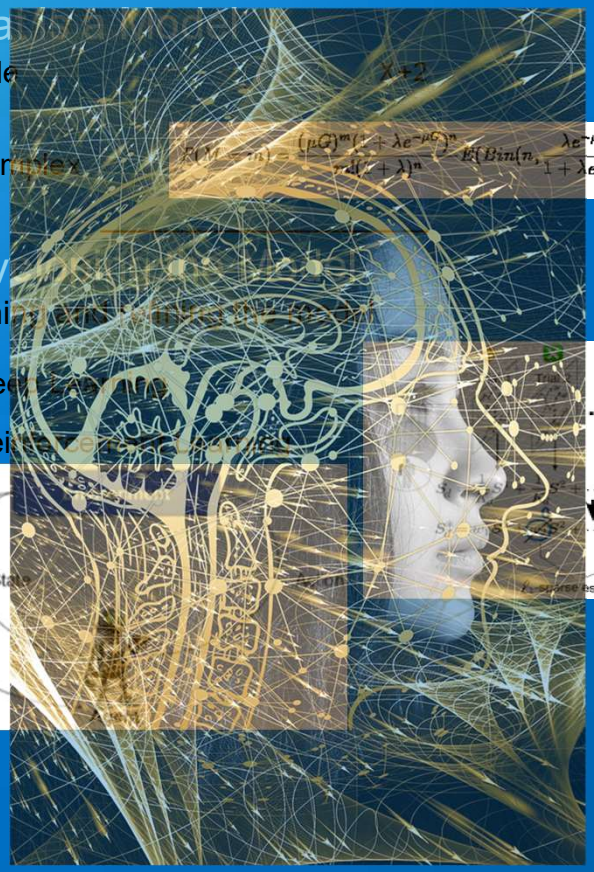
Wha
simple

or com

Dev
Traini

– De

– Rei



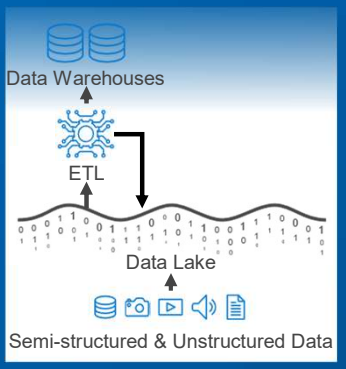
Purchased AI

VS

Purchased AI

+

Customer data



=

Specialized Model

DELL Technologies

AI – The LLM example – “fictious Dell help desk”

Integrating custom data into the model

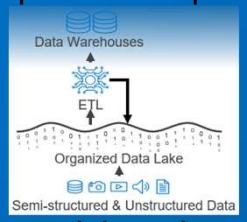
Basic Model (in the cloud)

Specialized Model (in the data center)

Refining the model

+

Dell’s previous help desk logs



Dell’s models and parts info

Continually ‘re-tuning’ the model

New Help Desk Tickets

Prompt> AI I have a Dell 7740 laptop and need to get a replacement power supply.

AI> Laptops generally require between 40 watts and 75 watts of power.

Prompt> AI I have a Dell 7740 laptop and need to get a replacement power supply.

AI> The Dell 7740 laptop requires a 65 watt power supply, model number RDX-4545 is the recommended Dell part number.

DELL Technologies

Selecting the Large Language Model

Integrating the LLM into the Data Center

Selection Criteria

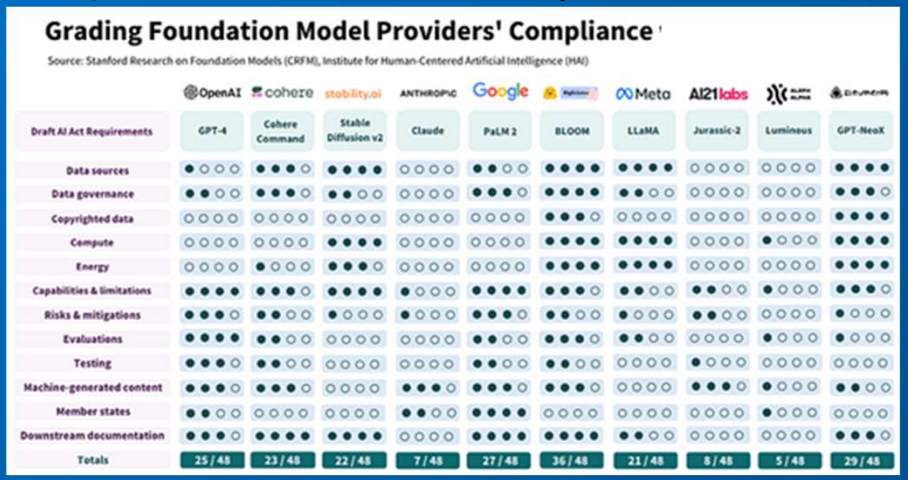
Performance / Quality

Model	Provider	Open-Source	Speed	Quality	Params	Fine-tuneability
gpt-4	OpenAI	No	★☆☆	★★★★	-	No
gpt-3.5-turbo	OpenAI	No	★★☆	★★★★	175B	No
gpt-3	OpenAI	No	★☆☆	★★★★	175B	No
ada, babbage, curie	OpenAI	No	★★★	★☆☆☆	350M - 7B	Yes
claude	Anthropic	Yes	★★☆	★★★★	52B	no
claude-instant	Anthropic	Yes	★★★	★★☆☆	52B	No
command-xlarge	Cohere	No	★★☆	★★☆☆	50B	Yes
command-medium	Cohere	No	★★★	★☆☆☆	6B	Yes
BERT	Google	Yes	★★★	★☆☆☆	345M	Yes
T5	Google	Yes	★★☆	★☆☆☆	11B	Yes
PaLM	Google	Yes	★☆☆	★★☆☆	540B	Yes
LLaMA	Meta AI	Yes	★★☆	★★☆☆	65B	Yes
CTRL	Salesforce	Yes	★★★	★☆☆☆	1.6B	Yes
Dolly 2.0	Databricks	Yes	★★☆	★★☆☆	12B	Yes

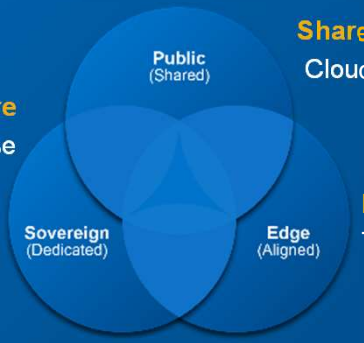
Customer Criteria / Output Requirement(s)

	Reasoning	Knowledge	Conversation	Creativity	Personality	Storytelling	Empathy
LaMDA	0.84	0.69	1.0	0.53	0.85	0.58	0.94
ChatGPT	0.74	0.82	0.92	0.77	0.72	0.74	0.7
GPT-3	0.87	0.86	0.72	0.75	0.66	0.72	0.49
T5	0.7	0.6	0.19	0.51	0.1	0.36	0.04
PaLM	0.76	0.56	0.21	0.24	0.21	0.18	0.17
BLOOM	0.48	0.35	0.29	0.36	0.15	0.18	0.24
Turing-NLG	0.56	0.42	0.29	0.07	0.16	0.07	0.0

Compliance, Governance and Security



Deployment



Shared infrastructure
Cloud Hosted

Dedicated infrastructure
On Premise

Distributed infrastructure
Tactical and disperse

DELL Technologies

Running AI – Partner Opportunity

Integrating the LLM into the Data Center

Purchase the desired model

OpenAI

Chatsonic

Hugging Face

Jasper AI

Grok

Gemini

Testing Workstations

1x CPU
256GB RAM
GPU

Specializing the Model

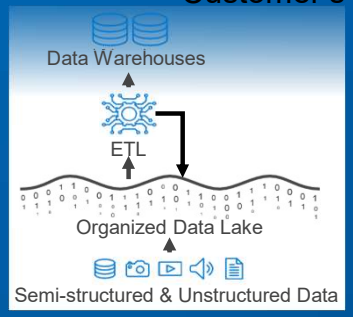
Refining the model

Customer has 1+ PB of data

5 Servers

2x CPU
1024GB RAM
4x A100 GPU

Customer's data



Data Classification Services

PowerScale
500+ TB
+ Cloud Archive

1TB data / hr / server

Running the basic Model (in the data center)

1,000 queries / min / server

Customer requirement: 20,000 queries / min

20 Servers

2x CPU
1024GB RAM
2x A100 GPU

Running the 'customized' model(s)

925 queries / min / server

+

1+TB data / hr / server

Edge Servers

DELL Technologies

AI Logistics:
Large Language Model or Lots of Small Language Models
Integrating the LLM into the Data Center

https://www.scientificamerican.com/article/when-it-comes-to-ai-models-bigger-isnt-always-better/
https://www.boldbusiness.com/digital/what-kind-ai-do-you-want-in-your-future/

As models have gotten bigger, they've also become more unwieldy, energy-hungry and difficult to run and build. Smaller models and datasets could help solve this issue.

A chatbot inside a smart fridge might need to understand common food terms and compose lists but not need to write code or perform complex calculations.

Some experts suggest that parameters of Big AI models could be reduced by 60% without losing significant performance. Not only does this favor Small AI versus Big AI based on complexity and power. But it also means Small AI would be less expensive to develop and use when compared to LLMs. This explains why some expect Small AI to have an advantage in the battle to dominate AI in the years to come.

PRECISION PORTFOLIO

POWER AS BIG AS YOUR IDEAS

Intelligent Performance | Immersive User Experience | Mission Critical Reliability

Mobiles 3000, 5000, 7000 Series

Towers 7000, 5000, 3000 Series

Dell has been a leader in workstation technology for over 25 years. Driven by customer insights and experiences, we offer intelligent performance and mission critical reliability for your key applications. Our groundbreaking innovations are a direct result of working with our customers to understand their pain points and how we can deliver the best solutions.

DELL Technologies

PRECISION TENENTS

- Intelligent Performance
  - Dell Precision has the most powerful workstations on Earth
    - Realize high-end performance with the latest Intel Core and Xeon processors and professional graphics from AMD and NVIDIA
    - Optimize application performance with Optimizer for Precision
    - Keep applications running smoothly with extensive, high-performance memory
    - Experience super-fast data transfers with Optane DC Persistent memory/storage

- Immersive User Experience
  - Providing what the customers desire
    - Tackle immerging technologies with Ready for VR and AI options
    - Get amazing picture quality with 4K, HDR and InfinityEdge displays
    - Experience true color with Dell Premium Color Displays that feature 100% Adobe color gamut
    - Leverage award-winning designs that provide amazing acoustics, heat reducing thermals and Hot-plug capabilities

- Mission Critical Reliability
  - Reliable performance and manageability
    - Rigorous tested to earn ISV certifications
    - Reduce memory errors with ECC Memory & RMT Pro
    - Ensure data availability with RAID capable systems
    - Manage systems smoothly with Dell Client Command Suite

PowerEdge 16G Server Portfolio

Purpose-built | Intelligent | Cyber Resilient | Sustainable

Purpose-built

AI, for the Edge and Scale Anywhere

Intelligent

Accomplish more with Automation & Improve Operational Efficiencies

Cyber Resilient

Zero Trust & Verifiable Confidence

Sustainable

Maximize power efficient performance

DELL Technologies

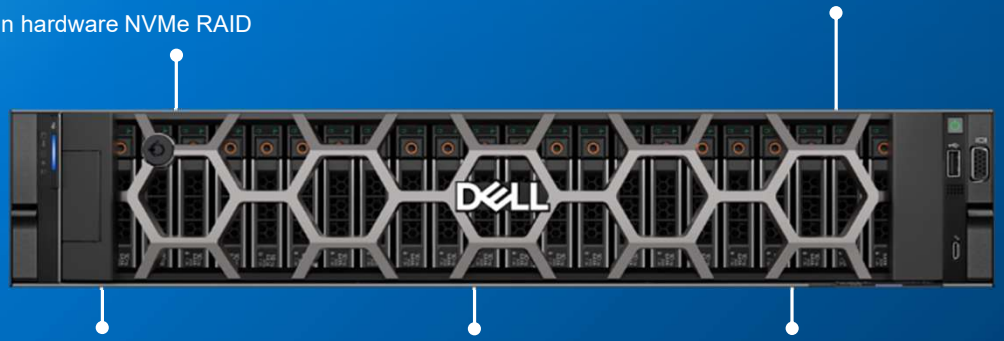
PowerEdge R760

Support for up to 28 Drives

- 24 NVMe direct-attached drives
- Gen5 NVMe* & SAS4 support
- Rear Hot-Plug BOSS-N1 (2 x M.2 NVMe) for boot
- Next-gen hardware NVMe RAID

Support for high-speed and memory capacity

- Up to 32 DDR5 DIMMs
- Up to 4800 MT/s (1DPC) or 4400 MT/s (2DPC)



2 Socket Capable

- Up to two 4th Generation Intel® Xeon® Scalable processors with up to 56 cores per processor
- High-bandwidth memory CPUs

Support for GPU

- 2 x 300W (DW) or 6 x 75W (SW)

Flexible I/O

- Up to 8 x PCIe Slots
- Optional 2 x 1GbE LOM + 1 x OCP 3.0 slot

MAINSTREAM

- Smart Cooling
- Direct Liquid Cooling
- Designed for growing scale-out solutions and air-cooled support
- Industry-leading manageability and security

Data Center (private cloud)

TARGET WORKLOADS

High Performance Scale-Out Databases

Architect for growth and scalability using high core count CPUs with the latest DDR5 memory technology, high-bandwidth networking and Gen5 based NVMe storage.

Next Level of Virtualization

8TB of memory combined with 112 cores of the latest generation Intel CPU enables high-density virtualization in a 2S server..

AI Training

With the latest Gen5 PCIe enabled NVIDIA GPUs and NVMe drives designed to offer the highest throughput on the largest datasets, customers benefit from reduced training cycles and faster AI deployments.

DELL Technologies

PowerEdge XE9680 / XE9640

Industry Technology

8-way H100 SXM or
8-way A100 SXM

Sapphire Rapids

PCIe Gen5

DDR5

NVME Gen5

Dell's first 8-way SXM GPU server that provides leading AI performance in a dense air-cooled design

AI / ML & HPC

Purpose-built for the most demanding AI – ML/DL training, modeling and simulation workloads

6U air-cooled design chassis supports the highest wattage next-gen technologies in up to 35C ambient

Massive flexibility with 8-way H100 SXM or A100 SXM, 10 Gen5 x16 PCIe slots, and up to 16 drives

Data Center (private cloud)

TARGET WORKLOADS

High Performance Computing

Architect for growth and scalability using high core count, 8x CPU Capable, CPUs with the latest DDR5 memory technology, high-bandwidth networking and Gen5 based NVMe storage.

Accelerated AI Training

With the latest Gen5 PCIe enabled NVIDIA GPUs and NVMe drives designed to offer the highest throughput on the largest datasets, customers benefit from reduced training cycles and faster AI deployments.

DELL Technologies

Dell Technologies – 17th Generation in the Data Center (AI)

Next Generation AI Capability

PCIe Gen5

DDR5

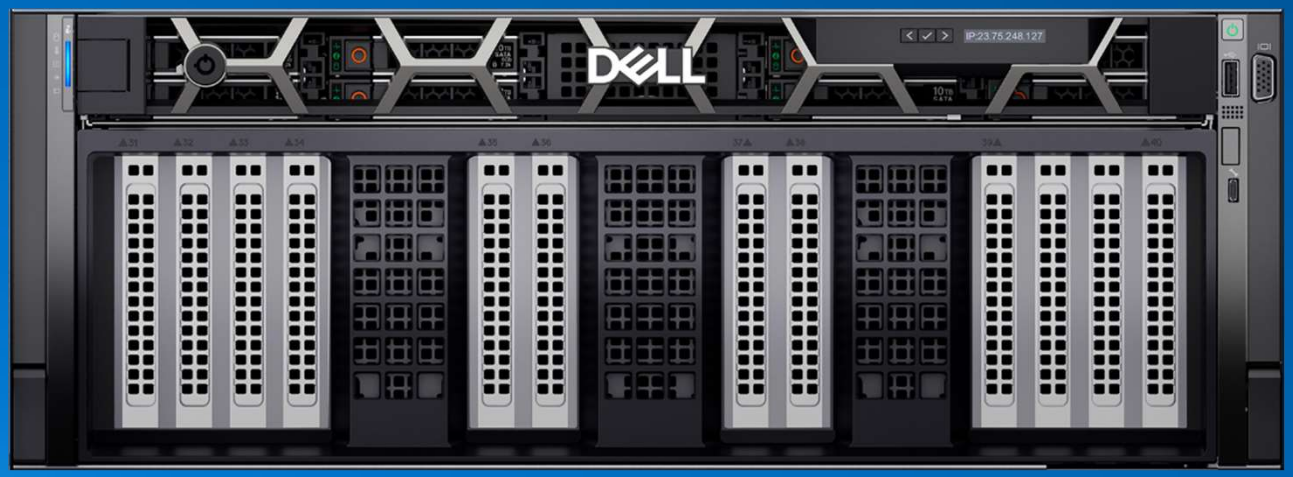
NVME Gen5

PowerEdge XE9680L

Liquid cooling

2.5x energy efficiency

Highest industry throughput (12 PCIE)



NVIDIA B200 GPU

Highest density (16GPU / 4U)

16-way H100 SXM or
8-way B200 SXM

NVLINK

Dell Technologies – Validated Designs for AI

Dell Technologies Validated Designs

Validated Designs for Generative AI

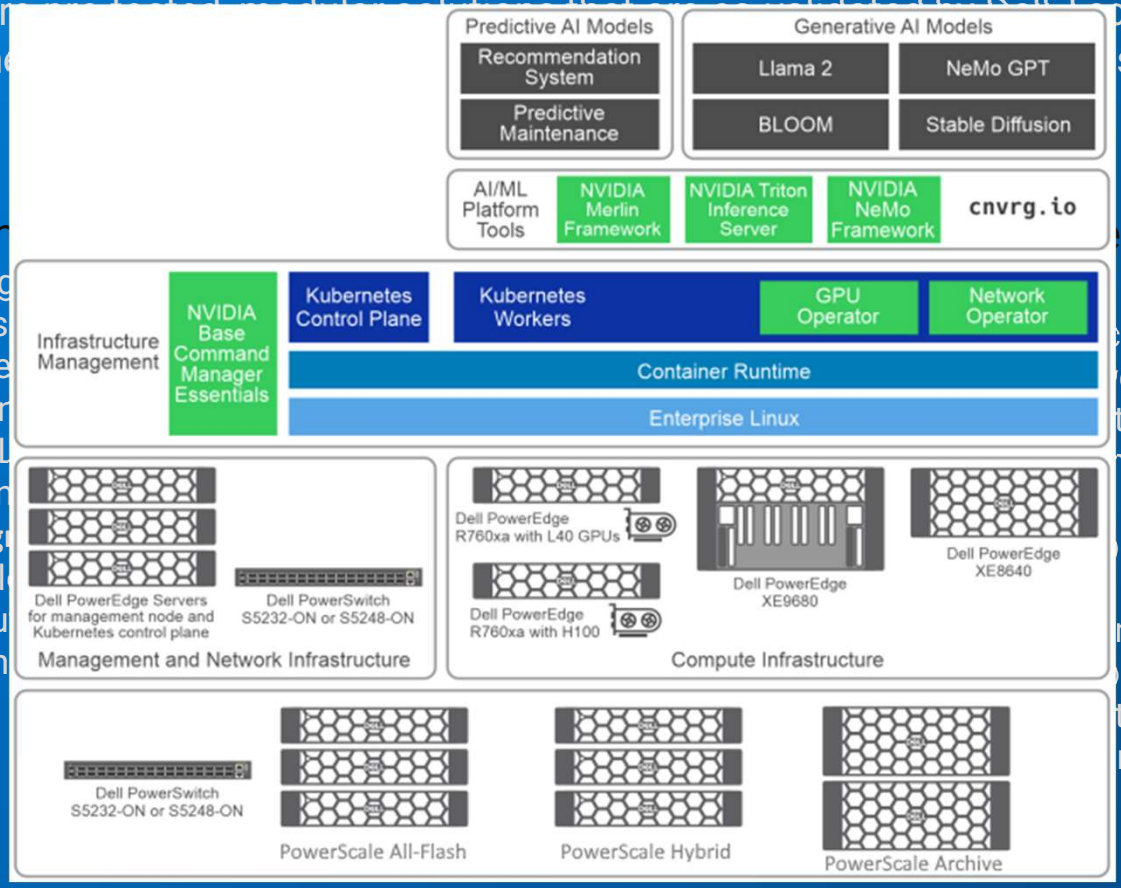
Dell Validated Designs ar chnologies experts and partners. They streamlin se cases with scalable

Validated Design for In l Customization and

Inferencing involves utilizing predictions, make decisions h businesses can add input data. This process is e ertical into a pre-existing large implementation of GenAI, er t to perform deeper generation and responses. L n pertinent to the business. help deliver results faster an

Generative AI Validated Desig del Customization and Tuning

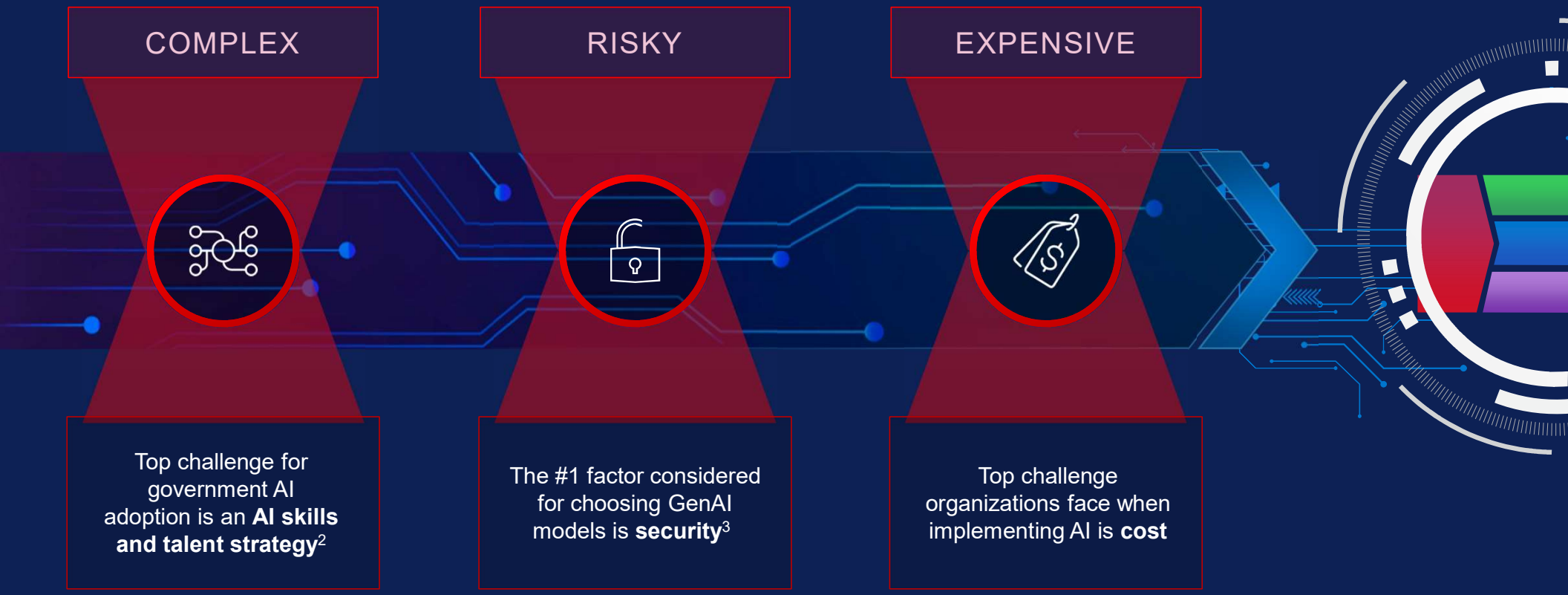
Addresses inferencing chall responsiveness, and compu n with proven guidance on re- transform enterprise data in r your use cases and tion techniques such as ng.



DELL Technologies

AI Headwinds Slowing Down AI Adoption

76% of IT and government leaders believe GenAI will deliver transformative value for their organization[1]



Top challenge for government AI adoption is an **AI skills and talent strategy**[2]

The #1 factor considered for choosing GenAI models is **security**[3]

Top challenge organizations face when implementing AI is **cost**

[1] Dell Technologies Innovation Catalyst Study, February 2024
[2] IDC Report: From Breakthrough Innovation to Impact: Monetizing the AI Moment Philip Carter, Directions 2024
[3] Dell Technologies Generative AI Pulse Survey, August and September 2023, www.dell.com/GenAIPulse
[4] IDC, Global AI Buyer Sentiment, Adoption and Business Value Survey, October 2023.

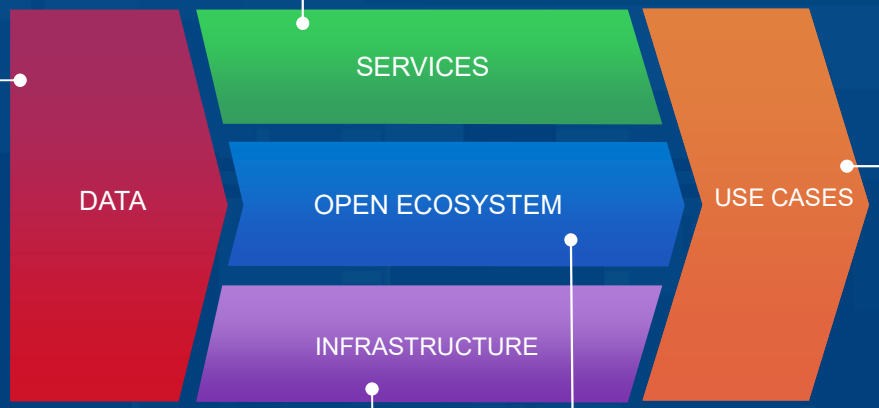
The Dell AI Factory

Dell's approach to help customers embrace and implement AI

EXPERT AI SERVICES

An effective AI Factory needs a skilled team to succeed, but AI-ready skills are in short supply and the ecosystem is diverse

Dell has extensive experience guiding customers through their AI journeys, accelerating AI outcomes aligned to business objectives while utilizing the right technical solutions at scale



DATA: FUELING THE AI FACTORY

Data is the raw material that powers the AI Factory

Most valuable data is on-premises and at the edge

Dell helps you bring AI to your most valuable data and is a leader in storing, protecting and managing that data

OUTCOMES POWERED BY USE CASES

The AI Factory produces business outcomes powered by your highest priority use cases

Dell simplifies the deployment of your most important AI uses cases with validated solutions and tailored services

AI OPTIMIZED INFRASTRUCTURE

Infrastructure is the foundation of the AI Factory

AI use cases are diverse and AI technologies are rapidly evolving

Dell has the broadest AI portfolio from client to cloud allowing you to right size your AI investment and giving you the flexibility to run AI anywhere

AI OPEN ECOSYSTEM

No single vendor can meet all your AI needs

Dell has a proven history of nurturing open ecosystems and is committed to building a broad AI ecosystem giving customers greater access to innovation and flexibility

Dell AI Factory Accelerates AI Adoption

SIMPLE	SECURE	ECONOMICAL
Comprehensive AI services and validated, optimized solutions make AI easier, augmenting skills gaps and addressing data readiness	Trusted infrastructure, full-stack solutions and services helps ensure valuable data is protected, managed and accessible A0	Industry's broadest AI solutions portfolio[1] addresses "no one-size-fits all" approach A1

*1 Based on Dell analysis, August 2023. Dell Technologies offers hardware solutions engineered to support **AI** workloads from Workstations PCs (mobile and fixed) to Servers for High-performance Computing, Data Storage, Cloud Native Software-Defined Infrastructure*

Slide 25

A0 [Mention was removed] - revised to "helps ensure" to avoid putting forth a guarantee
Author, 2024-05-30T17:20:58.200

A0 0 thanks. Adjusted the colors
Author, 2024-06-03T20:42:17.726

A1 [Mention was removed] - please provide support for this claim
Author, 2024-05-30T17:22:12.716

A1 0 Attribution added!
Author, 2024-06-03T20:43:44.051

A modern data strategy brings AI to data

AI Lifecycle

Data Management & Preparation

Training

Model Customization and Tuning

Inferencing

Discover, process data across the organization with modern scale-out lakehouse strategies

Reduce data learning curve with expertise and education

Deploy managed GenAI services to simplify operations

Get ahead with optimized data management and prep for fastest ROI

Bring AI models to customer data and customize based on use cases

DELL Technologies

Accelerate GenAI outcomes and ensure long-term success with help at every stage

Establish strategy

Outcome

Consensus on Roadmap

- Accelerator Workshop
- ProConsult Advisory Services
- Advisory Services for GenAI Process Optimization
- Workforce Persona Services

Prepare data

Outcome

Validated data for model

- Advisory Services for GenAI Data Security
- Implementation Services for GenAI Data Preparation
- Services for Dell Data Lakehouse

GenAI Platform

Outcome

Deployed GenAI platform

- Deployment Services
- Platform Implementation Services
- Red Hat OpenShift AI Implementation Services
- NEW Platform Implementation Services for Digital Assistants

Deploy & test model

Outcome

Tuned Model

- Use Case Implementation Services
- NEW Use Case Implementation Services for Digital Assistants
- NEW Accelerator Services for RAG + on Precision workstations
- NEW Accelerator Services for Dell Enterprise Hub
- NEW Services for Microsoft Copilot

Operate & Scale

Outcome

Simplified GenAI operations

- Residency Services
- Managed Services for GenAI
- Training and Certification
- Support Services
- Adoption & Change Management

Simplify your AI journey with the Dell AI Factory

Services

DELL Technologies

Dell's AI Advisory Approach

Inform the journey towards data-driven and value-focused use of AI capabilities

Assess Current State (As-Is)

- Conduct Mission and technical interviews to identify Mission objectives and requirements, relevant technologies and systems, and data sources and capabilities
- Collect previously defined uses cases, current AI capabilities, and other supporting documentation to inform requirements
- Assess AI Readiness

Project Kickoff
(presentation & project plan)

Interview Schedule
(roles & stakeholders)

AI Readiness Assessment
(6-dimensional view)

Define Future State (To-Be)

- Customize and refine use case prioritization framework and criteria weights
- Prioritize AI Mission use cases based on Mission value and feasibility
- Define AI archetypes, identify repeating use case patterns, and organize use cases accordingly

Prioritization Framework
(Mission value & feasibility)

Use Case Priorities
(Mission value & feasibility)

AI Archetypes and Patterns
(shared value through reusability)

Plan Transition (Roadmap)

- Finalized AI and Data Management guiding principles and logical architecture
- Create and validate transformation roadmap based on current state gaps
- Develop recommendations and next steps and present to executive stakeholders

AI Data Management
(guiding principles & arch)

Transformation Roadmap
(activities & dependencies)

Recommendations
(executive summary)

Enact Use Case (Execute)

- Select and analyze data sources for top use case
- Deploy pre-trained model
- Conduct unsupervised fine-tuning and provide recommendations for retrieval augmented generation
- Conduct functional demo to prove concept

Deploy Model
(pre-trained)

Tune Model
(fine-tuning, RAG)

Functional Demo
(proof of concept)

AI use cases and tech continue to expand and evolve – there is no one-size-fits-all approach

Rapid evolution of models, AI tools and AI-led use-cases across all industries

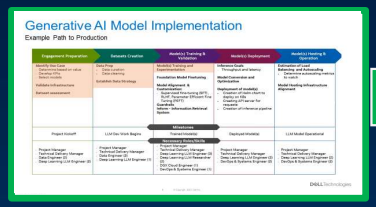
Many approaches for creating expert systems based on private data

AI workloads running on-prem, in the cloud, and at the edge

On-prem AI workloads will be accelerated by AI servers and AI-ready PCs using a diverse range of silicon

Gen AI Use Cases & Services Step by Step

Accelerate customer Gen AI adoption with a comprehensive portfolio



Step #0: Engage your **Dell Services** sales counterpart to initiate a fee-waived Workshop

Fee-waived Accelerator Workshops

Services Step by Step

Step #1 Strategy — Create customer Gen AI transformation roadmap (As-Is / To Be)

Step #2 Implement — Validated data for Implementation and enable the Gen AI environment

Step #3 Adopt — Gen AI use case implementation

Step #4 Scale — Operationalized use case

Gen AI Use Cases

Financial Services
- Financial analytics
- High frequency trading
- Fraud detection

Healthcare & Life Sciences
- Drug design
- Genomics
- Bio informatics

Government & Academic Research
- Particle physics
- Life sciences
- Humanities
- Climate modeling

Government Classified
- Homeland security
- Defense
- Nuclear safety

Manufacturing
- Structural analysis
- Fluid dynamics
- Packaging
- Impact modeling

Geosciences Oil & Gas
- Seismic
- Reservoir modeling

Accelerator workshops are the first step

What is an Accelerator Workshop?

Interactive, strategic session with business and IT leadership

Align cross-organizational priorities and gain consensus

Deliver actionable roadmap with recommendations to achieve desired results

Dell Technologies Differentiators

As-Is/ To-Be methodology rapidly assess environment and implements plan

IT expertise to architect solutions and operationalize technology

Completed over 10,000 transformation projects and 1,000+ strategic workshops

DELL Technologies

Thank you for viewing this Dell Technologies presentation! Carahsoft is the distributor for Dell Technologies public sector solutions available via GSA, ITES-SW, MHEC, and other contract vehicles.

To learn how to take the next step toward acquiring Dell Technologies' solutions, please check out the following resources and information:

For additional resources:
carah.io/Dell-Resources

For additional Dell Technologies solutions:
carah.io/Dell-Solutions

To purchase, check out the contract vehicles available for procurement:
carah.io/Dell-Contracts

For upcoming events:
carah.io/Dell-Events

For additional public sector services:
carah.io/Dell-Solutions

To set up a meeting:
DellQuotes@carahsoft.com 888-662-2724

carahsoft

For more information, contact Carahsoft or our reseller partners:
DellQuotes@carahsoft.com | 888-662-2724