DELL Technologies

Powering Federal AI: Compute Platforms

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
- AI & Partner(s)
- AI at the user
- AI in the Data Center
  - Compute
- Dell's Approach to AI
- Questions & Answers

DELL Technologies

Dell Technologies Advantage

Modern Workplace

- World's smallest,[1] thinnest, most powerful commercial AI PCs[2]
- World's #1 monitors[3]
- Most manageable and secure commercial PCs[4]
- Leading in use of sustainable materials

Modern Data Center & Multi-cloud

- #1 servers [5]
- #1 external storage [6]
- #1 storage software [7]
- #1 Purpose-Built Backup Appliances [8]

Artificial Intelligence

- The world's broadest AI solutions [9]
- Open ecosystem that continues to grow with proven, validated solutions
- World's best management consulting firms – Forbes 2024

Security / Resilience

Industries most secure supply chain

Sustainability

An industry leader in sustainable materials and packaging

EDGE | CORE | CLOUD

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

GB10

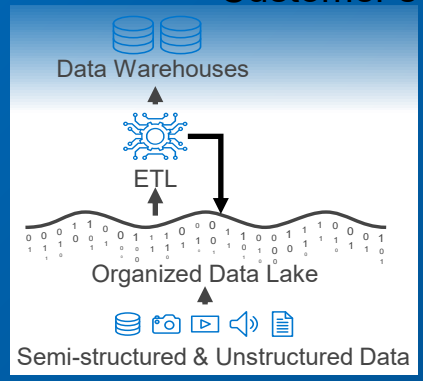
Specializing the Model

Refining the model

PowerStore

Additional Servers

Customer's data



Data Classification Services

PowerScale

Running the basic Model (in the data center)

Servers

Running the 'customized' model(s)

AI Client Systems

Constant Stream of NEW data

Edge Servers

DELL Technologies

Selecting the correct Large Language Model

Integrating the LLM into the Data Center

Selection Criteria

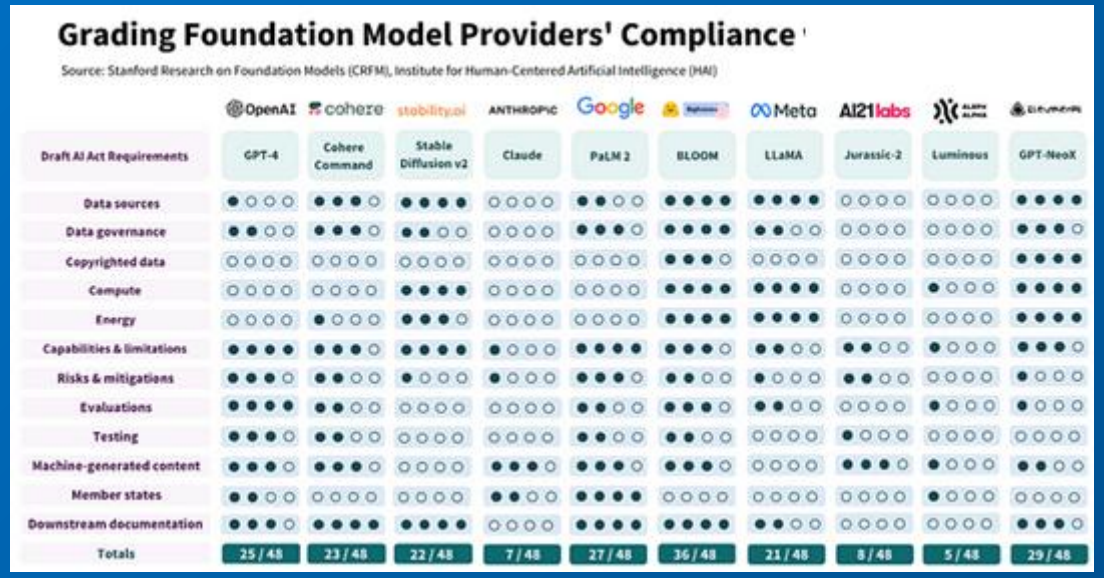
Performance / Quality

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

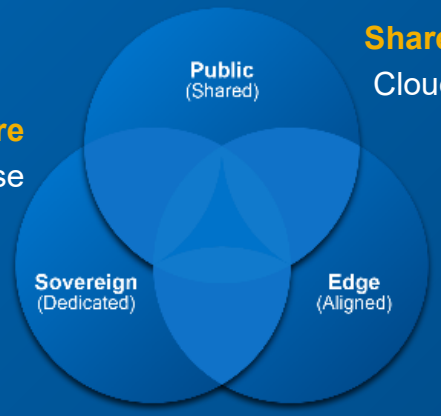
Customer Criteria / Output Requirement(s)

	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

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.

How does RAG work ?

(Retrieval Augmented Generation)

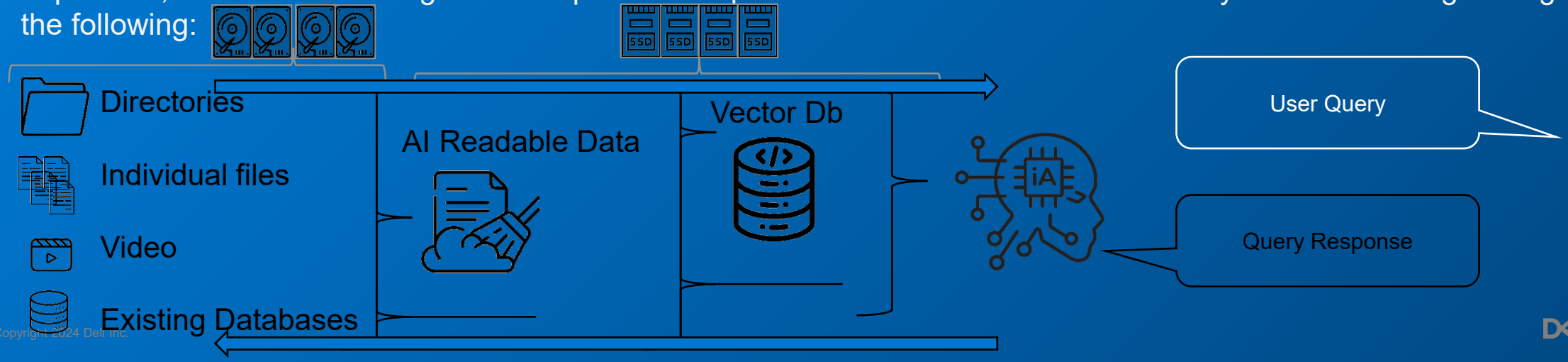
Storage for AI processing and User based workloads

As we look to the process outlines in the previous example:

- Cleaning and Processing and Vectorizing the data, is usually a one time event. Though intensive at that time it can be thought of as a batch process that takes as long as needed. Very much like an Archive event.
- Having the AI engine query the vector Db however is something that every query will undergo. Every user, every query will cause the Db to be queried.
- When the Db query returns a valid sample of existing (cleaned and processed) documents to include, those documents will also need to be “read” into the AI engine.

As we examine the processes, the initial step can be completed on large capacity, lower performance disk subsystem. But as we progress, the Db and in usage the cleaned and processed documents need to be quickly accessible to the AI engine.

In process, for RAG to be integrated and provide the performance needed for multi-user systems the storage design follows the following:



Modern Solutions for the Modern Workplace

Mobile & Fixed PCs

Dell
Laptops & Desktops

Dell Pro
Laptops & Desktops

Dell Pro Max
Laptops & Desktops

Dell Pro Rugged

Dell Thin Clients

Silicon Diversity

intel. AMD NVIDIA Qualcomm

AI Built In

- AI PCs, Copilot+ PCs for Business Workers
- High Performance AI PCs for Developers

Displays & Accessories

Monitors

Docking Stations

Keyboard, Mice & Active Pens

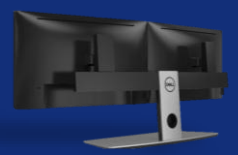
Audio

Webcams

Bags

Power Adapters

Cables & Adapters



Monitor Arms & Stands

IT Solutions

Dell Endpoint Security
Dell Trusted Workspace

Dell Manageability Solutions
Dell Management Portal
Dell Device Management Console

IT Services

- APEX PCaaS
- Employee Experience Measurement Services
- Lifecycle Hub Services
- Implementation and Migration Services for Windows 11
- Dell ProDeploy Suite for PCs
- Dell ProSupport Suite for PCs
- Asset Recovery Services

End-to-end Sustainability

Embedded across our products, packaging, and supply chain

- Reducing Product Carbon Footprints (PCFs)
- Building Durable & Reliable PCs
- Designing with Recycled & Renewable Materials
- Improving Energy Efficiency

DELL Technologies

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

Dell Pro Max with GB10 – Standalone vs. Connected

Fine-tuning

Prototyping

Inferencing

Large model hosting

RAG

Edge simulations

AI development
& deployment Accelerator
(Linux-based)

Offload AI tasks with secure remote-in connectivity via OpenSSH

General-purpose workstation
(Windows-based)

Dell Pro Max with GB10 as a Connected device for AI acceleration (Linux-based)

Design, coding, CAD, 3D, Video editing, ISV apps

Faster, parallel workflows + system responsiveness

AI-specific tasks, fine-tuning, RAG, inferencing, Edge simulations

Works best as a standalone device when…

- You want a ready-to-code AI development system with NVIDIA's DGX OS, CUDA libraries, and software stack pre-configured.
- You prefer Linux for AI development and deployment.
- You need to run large models locally without GPU or cloud dependency.
- You need a portable AI environment for research, fieldwork, or edge projects.
- You want a self-contained system that's easy to redeploy to different projects.

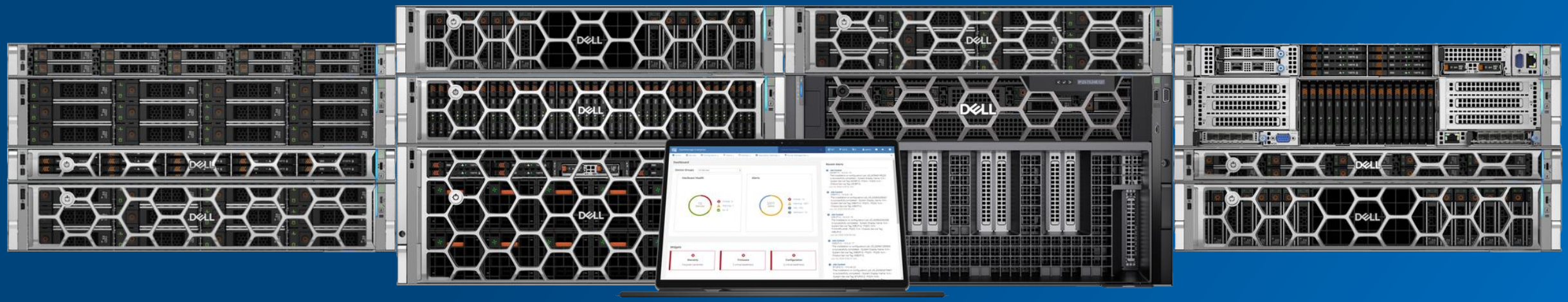
Works best as a connected device when…

- You have a workstation for design, coding, or content creation and don't want AI jobs slowing it down.
- You want to offload AI workloads to free up your main system.
- You need parallel workflows — development on one machine, inference on another.
- You want dual OS flexibility — Windows on your workstation, Linux on the GB10.
- You want a cost-efficient alternative to high-end GPU upgrades.

DELL Technologies

PowerEdge Servers

Purpose-built | Intelligent | Cyber Resilient



Purpose-built

Scale AI, Edge & Performance Anywhere

Intelligent

Accomplish more with Automation & Improve Operational Efficiencies

Cyber Resilient

Accelerate Zero Trust Adoption

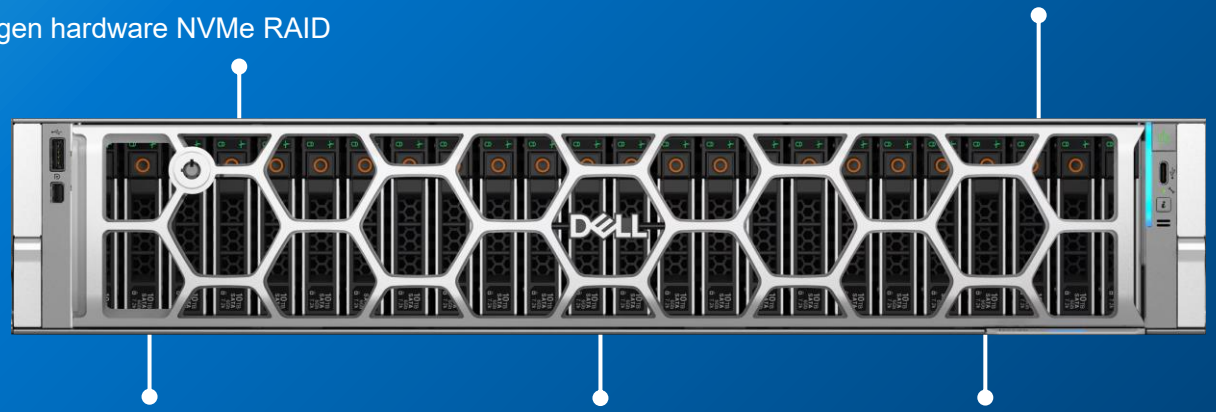
PowerEdge R770 / R7725

Support for up to 28 Drives

- Up to 40 x EDSFF E3.S Gen5 NVMe
- 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 6TB



2 Socket Capable

- Intel / AMD processors
- High-bandwidth memory CPUs

Support for GPU

- 2 x 450W (DW) or 6 x 75W (SW)

Flexible I/O

- Up to 8 x PCIe Slots
- Optional 2x OCP 3.0 slot

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

MAINSTREAM

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

6TB of memory combined with >192 cores of the latest generation Intel/AMD CPUs enables high-density virtualization in a 2S server..

AI Training

With the latest Gen5 PCIe enabled AMD 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 XE9780 / XE9785

8-way H100 SXM or
8-way B300 SXM

Intel / AMD
CPU

PCIe
Gen5

DDR5

NVME
Gen5

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

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

7U air-cooled design chassis supports the highest wattage next-gen technologies in up to 35C ambient

Massive flexibility with 8-way B300 SXM or A100 SXM, 10 Gen5 x16 PCIe slots, and up to 16 drives

AI / ML & HPC

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

The Dell AI Factory

Dell's approach to standardizing and accelerating AI use cases

People – Processes – Technology
Use Case – Data – Outcomes – Technology

USE CASE: DIRECTING THE AI

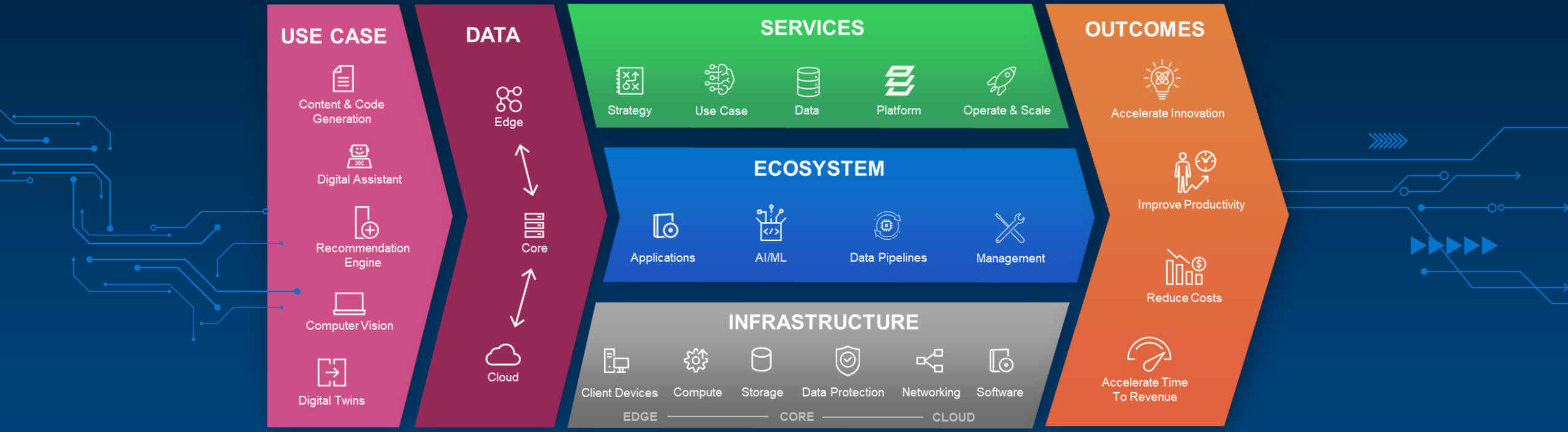
How the AI is being utilized is an important step in creation of the AI, and how that AI is going to enhance the human experience.

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

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



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

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

DELL Technologies

Dell Technologies Ecosystem

- NVIDIA – Core to the Dell AI Factory with NVIDIA, used extensively in federal AI and HPC initiatives as the integrated development and runtime environment for LLMs, RAG, and agentic workloads in secure data centers and labs.
- Google (Gemini) – Positioned specifically for on-prem Gemini LLM deployments on Dell AI Factory for regulated and public sector environments that require tight data control and sovereignty.
- Meta (Llama / Llama Stack) – Used with Dell AI Solutions with Llama and Llama Stack on PowerEdge for sovereign and public-sector-friendly GenAI, where open models and on-prem control are important.
- Cohere – Sovereign / regulated enterprise GenAI and agentic AI partner (Cohere North on Dell AI Factory) with explicit positioning for secure, no-code agents and RAG in highly regulated sectors including public sector and government.
- H2O.ai – Highlighted as a GenAI and agentic AI platform for air-gapped, on-prem, data-sovereign deployments, with “Federal” called out as one of its top three segments on Dell AI Factory reference material.
- Red Hat (OpenShift / OpenShift AI) – Core AI platform partner for federal-grade containerized AI, often combined with Dell Private Cloud Platform and AI Factory for Sovereign AI / Private AI-in-a-Box offers for government and NATO-style environments.

Dell Technologies Data Protection Solutions

Comprehensive portfolio built to protect proven & modern workloads

Data Center Backup

Backup Targets

Integrated Appliances

Direct Cloud Backup

In Cloud Protection

In-Cloud
Protect Cloud Born Workloads

SaaS Apps
O365, G-Suite, Salesforce, Slack

Protect any cloud-born workload

Disaster-Recovery

Cloud DR
Replicate VMs to the Cloud

On-Prem
Replicate to DR Site

DRaaS
On-Prem/Cloud DRaaS Offering

Ability to recover from disasters

Cloud Archive

ECS Tier
Long-Term Retention

Cloud Tier
Long-Term Retention

Tier rarely accessed data to free space

Cyber-Recovery

Air-Gap
Recover From Cyber-Attacks

Ability to recover from cyber-attacks

CLOUDS: VMC, aws, Microsoft Azure, Google Cloud

PLATFORMS: vSphere 7, kubernetes, Tanzu, VCF, OpenShift, Rancher, AKS, GKE, EKS

APPS: RDS, cassandra, MySQL, ORACLE, mongoDB, Microsoft SQL Server, PostgreSQL

Dell Technologies Services

Accelerate the power of **AI** for your data

Dell Accelerator Workshop for Generative AI is the first step in achieving a successful journey into GenAI

Experience an easy point of entry from Dell Accelerator Workshop

Half-day, fee-waived facilitated workshop brings IT and business stakeholders together to begin developing a point of view on important AI and data questions.

Dell experts work with your team to create a vision for your future state, utilizing our "AS-IS / TO-BE" methodology to align on priorities and next steps. We will conduct interview and review the existing environment, identifying challenges and driving consensus that is synthesized into an Executive Overview.

Assessment Dimensions

The Accelerator Workshop is a great starting point to determining how your business will achieve maximized value from AI and data. We begin to consider the needs for getting your organization ready across key priority areas, defining a high-level transformation plan aligned to your strategic vision and architectural principles.

Strategy & Governance	Data Preparation & Onboarding	Data Virtualization & Serving	AI Models
Platform	People	Adoption & Scaling	Security & Resilience

What is an Accelerator Workshop?

Interactive, strategic session with mission personnel and IT leadership

Align cross-organizational priorities and gain consensus

Deliver actionable roadmap with recommendations to achieve desired results

Next Steps and Q & A

Architect your AI solutions to assist your customer's mission

- Data Center Infrastructure
  - Compute
  - Storage
- Platforms
- Applications
- Multi-Clouds

Higher Revenue & Higher Profit

Next Steps

- Schedule a Test Drive
- Contact your Distributor or Dell team

DELL Technologies

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/DellResources

For additional Dell Technologies solutions:
carah.io/DellSolutions

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

For upcoming events:
carah.io/DellEvents

For additional public sector solutions:
carah.io/DellSolutions

To set up a meeting:
Dellgroup@carahsoft.com or 866-Dell-2-Go

carahsoft

For more information, contact Carahsoft or our reseller partners:
Dellgroup@carahsoft.com | 866-Dell-2-Go