Dell AI Strategy & Overview

HPC Large Language Model (LLM) Solutions

Stephen Sofhauser
HPC/AI Federal Team Lead
stephen.sofhauser@federal.dell.com

DELL Technologies

A history of customer-inspired innovation

16+ generations of Dell Technologies in HPC and AI

1st HPC CLUSTER	TUNGSTEN - #4	THUNDERBIRD - #6	STAMPEDE	PITZER	FRONTERA	Tri-Labs CTS-2	Denvr Data Works
Cornell University	NCSA	Sandia National Laboratories	TACC	Ohio Supercomputer Center	TACC	CTS-2	

1999 ★ **24 YEARS OF LEADERSHIP** ★ **2023**

TOP 500 The List.

#1 in XSEDE HPC systems for US open science

Frontera: Fastest HPC in higher education

Great Lakes: 1st Mellanox HDR InfiniBand system

The GREEN 500

T-Gen: Collaborative development for clinical needs

Orion: Weather system with MSU and NOAA

Commodity Technology Systems 2: Large scale deployment of 1.5PF scalable units

Texas A&M ACES: Composable HPC & AI

Wrangler: High performance petascale data analytics

Dept. of Energy: Commodity HPC systems

Denvr Data Works: Generative AI at scale

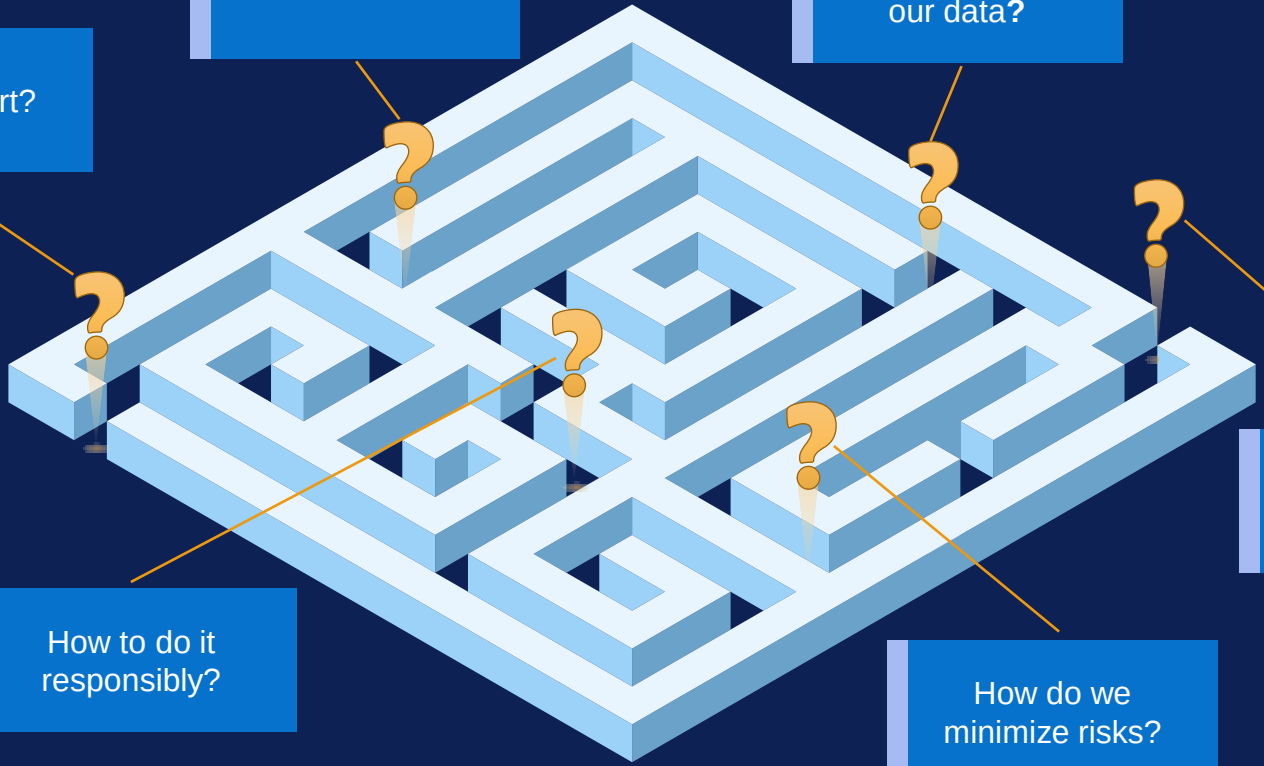
A sampling of ML, DL and Generative AI use cases

Leaders have many questions they need to answer

Where to start?

Which use cases**?**

How best to use our data**?**



How do we activate everyone?

How to do it responsibly?

How do we minimize risks?

Risks

- Late to market
- Slow adoption of key technologies
- Over or under investment in resources
- Too late getting infrastructure allocated
- Overreaching on your first projects
- Insufficient resources to deliver projects on time
- Getting entangled in legal issues / infringement

How Dell can help?

DELL Technologies

Dell Technologies HPC AI and Large Language Model Strategy

ADVANCING

Provide expertise, innovation, and partnerships

Inform and develop a strategy on opportunities to accelerate your business with LLMs

DEMOCRATIZING

Making LLMs and AI accessible for everyone

Optimized Dell Validated Design platforms for running your own LLMs, providing performance, efficiency, and security

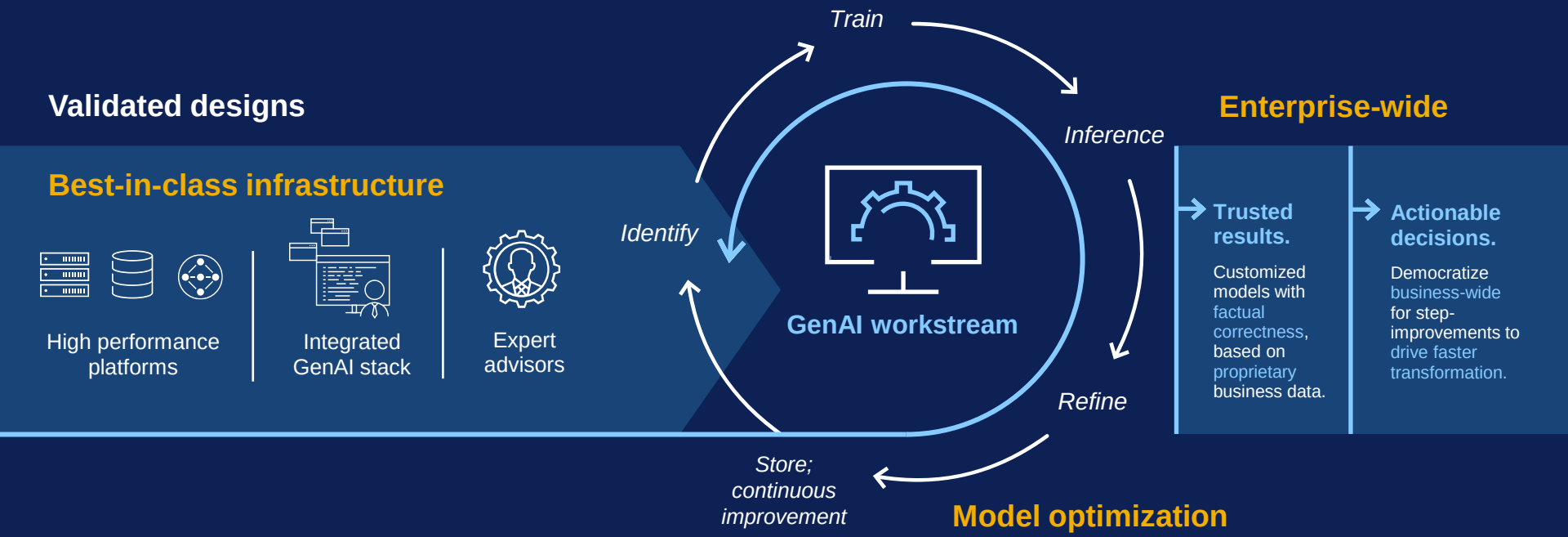
OPTIMIZING

Guidance on fine-tuning for your business applications

Empower your organization with LLMs developed and fine-tuned on your data, for your business, and secured within your infrastructure

Dell Validated Design for Generative AI

Deploy Generative AI with Dell Technologies and NVIDIA expertise



Validated Designs for AI

AI SIMPLIFIED

- 18-20% faster configuration and integration
- Save 12 employee hours a week with automated reconciliation feeds
- Reduce support requirements by 25%

Faster AI insights

- 44% faster deployment
- 18X faster AI models

Proven AI expertise

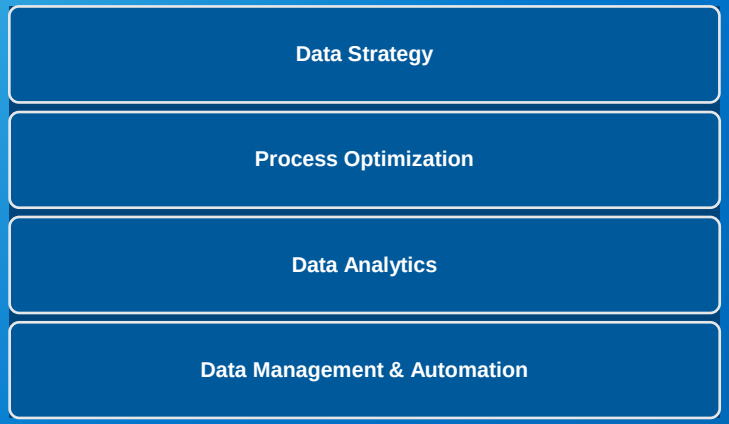
- AI Experience Zones
- Dell AI Specialists

- Accelerate AI by optimizing Compute Resources (run.ai)
- AI in Virtualized Environments (vmware)
- Automatic Machine Learning (AutomML with H2O)
- AI Machine Learning Operations (MLOps with cnvrg.io)
- Conversational AI (with Kore.AI)
- Computer Vision for Smart Transport (Genetec/Ipsotek/Briefcam) & (Milestone /Ipsotek/Briefcam)
- Datarobot
- Domino Data Science Platform
- Deep Learning with nvidia
- HPC for AI and Data Analytics
- Iguazio
- Intelligent Video Analytics with Epic.IO and Ipsotek (Formerly Intellisite/Deep Vision)
  Kubeflow
- NVIDIA Fleet Command – Manage and Scale AI at the edge
- On-premises Training, Prompt Tuning and Inference of LLM's at Scale (Project Helix)
- Retail Loss Prevention

Dell AI Services for customers

- Advice on model trends, best practices, systems
  - Access to data scientists, subject matter experts working on cutting edge GenAI
  - Gap analysis, planning, strategy, data center design, multi-cloud strategy
- Consult on business strategy, workflow process efficiency
- Training seminars on workflow, model development, tools
- Design and supply of LLM HPC infrastructure
  - On prem, multi-cloud, scalable HPC
  - Accurate sizing to best fit system budget, load, environment
  - Sustainability designs
  - Access to solution architects familiar with GenAI design
  - Modular data centers for large scale LLM model deployment
- Turnkey development packages
- Proof of concept testing to reduce risk
- Flexible financial models

Enabling AI with Dell Services



DATA ANALYTICS STRATEGY

Define how to use data to deliver enhanced business outcomes:
- SKU: 893-9943
- Name: ProConsult Core: Advisory Data Analytics
- Engagement length: 3 weeks
- As-Is → To-Be (Data/Workload focused)
- Vision workshop – AI/ML use cases and business priorities
- High-Level architecture
- Transformation roadmap

PROCESS OPTIMIZATION

Define how to operate more efficiently and effectively:
- SKU: 875-6022
- Name: Consulting Residency For Applications And Data
- Engagement length: 6 weeks
- As-Is → To-Be (Process focused)
- Value Stream Mapping
- Identify automation/AI/ML opportunities
- Transformation roadmap

AI/ML USE CASE

Design and build an AI/ML model:
- SKU: 893-9944
- Name: ProConsult Plus: Advisory Data Analytics
- Engagement length: 6 weeks
- Prove out a proposed AI/ML model
- Discovery, Data Assessment, First Model Iteration, Define Path to Production

DATA MANAGEMENT & AUTOMATION

Design, build, and operationalize repeatable, flexible, on-demand data services :
- SKU: 875-6022
- Name: Consulting Residency for Applications and Data
- Engagement length: 12 weeks
- Delivering platforms for analytics/AI/ML
- Data Ingestion, Persistence, Integration and Consumption
- Data Management, Governance, and Security
- Process Automation / DevOps / DataOps / MLOps

Sampling of Dell Services AI projects

- Pump manufacturer: Quality control for bolt alignment, scratches, tag placement
- Railroad: Track inspection
- USPS: Potential dangerous package detection
- Large retailer: Shrinkage detection at self-checkout
- Utility company: Monitoring sub-stations
- Healthcare: CT and MRI scan reading with over 90% accuracy (better than human)
- Planes: Download data to detect anomalies
- Airport: Manage airplane turnaround at the gate
- Dell: Manufacturing, sales, finance and more

How can we help you with your AI opportunities?

JOIN THE COMMUNITY

NEWS:
insidehpc.com/dell

DELL TECHNOLOGIES HPC COMMUNITY:
dellhpc.org

HPC & AI ENGINEERING:
hpcatdell.com and https://infohub.delltechnologies.com/t/high-performance-computing/

HPC & AI INNOVATION LAB:
delltechnologies.com/innovationlab

HPC / AI CENTERS OF EXCELLENCE:
delltechnologies.com/coe

To learn more, visit delltechnologies.com/hpc

DELL Technologies


DELL Technologies

Dell Infrastructure For GenAI LLM

Dell Optimized GPU Server Portfolio for GenAI

PCIe Optimized	4-way SXM	4-way Dense	8-way SXM
			
R760xa	**XE8640**	**XE9640**	**XE9680**
Available Now *Up to 160GB model size memory, bridged* • **2U** monolithic • **2-socket** Sapphire Rapids CPU • **Up to 4 x** double-wide GPUs • **Up to 12 x** single-wide GPUs • **Full PCIe GPU portfolio** supported • **Air cooled** with optional liquid cooling for CPU	**Available Now** *Up to 320GB model size memory, NVLink* • **4U** monolithic • **2-socket** Sapphire Rapids CPU • **4 x Nvidia H100 SXM** NVLink GPUs; • **Air cooled**	**Available Soon** *Up to 320GB model size memory, NVLink* • **2U** monolithic • **2-socket** Sapphire Rapids CPU • **4 x Nvidia H100 SXM** NVLink GPUs -or- • **4 x Intel Data Center Max 1550 OAM** XeLink GPUs • **Direct liquid cooled** CPUs and GPUs	**Available Now** *Up to 640GB model size memory, NVLink* • **6U** monolithic • **2-socket** Sapphire Rapids CPU • **8 x Nvidia H100 SXM** NVLink GPUs -or- • **8 x Nvidia A100 SXM** NVLink GPUs • **Air cooled**
High performance 2U server purpose built for dense PCIe GPU acceleration. Maximize AI, HPC, VDI and performance graphics supporting multiple GPU choices. Use cases: • AI/ML Inferencing • AI/ML Training • Rendering/Perf. Gfx • VDI	Accelerate and automate analysis into insights. Maximize AI initiatives performance in a 4-way GPU, 4U server. Use cases: • AI/ML Training • HPC Modeling & Simulation	Push performance boundaries with a dense form-factor, liquid cooled approach to AI initiatives. Smallest form factor 4-way GPU, dense 2U AI/ML/DL & HPC server. Use cases: • AI/ML Training • HPC Modeling & Simulation	Modernize operations and infrastructure to drive new AI initiatives. Optimized for demanding AI/Machine Learning & Deep Learning applications Use cases: • Large AI/ML/DL Training

DELL Technologies