White Paper

Creating an End-to-End Infrastructure for AI Success

Sponsored by: Dell Technologies

Peter Rutten Eric Burgener

September 2021

SITUATION OVERVIEW

IDC expects that within the next few years, artificial intelligence (AI) will start to permeate business processes for most enterprises. In general, more data will drive better products and services, improved customer experience, and more relevant business insights. There will be a proliferation of data capture points as enterprises collect data from edge devices, products and services, employees, supply chain partners, and customers. These data sets are generated by the many operations that together constitute a business but that themselves are not yet integrated or understood in relation to each other.

Big data analytics applications leveraging AI hold the promise of driving better business insights, fueled by the data collected from these sources. These applications increasingly operate with a real-time orientation and require performance, availability, and scalability that legacy information technology (IT) infrastructure cannot meet.

To make the most effective use of AI-driven big data analytics, enterprises will need to create an "end to end" AI strategy that is well integrated across three different deployment models — from edge to datacenter to cloud. Compute and storage platforms that offer edge, core, and cloud-based deployment options can enable enterprises to build a more cohesive, well-integrated infrastructure that supports a common management paradigm across locations.

Infrastructure requirements for AI can be considered from three angles:

- **Scale.** The scale dimension describes the scale at which the workload operates. Hardware — compute, networking, and data persistence (storage) — plays a crucial role, but software, such as virtualization and orchestration, is becoming just as important for managing ever larger and more complex AI models.
- **Portability.** Portability refers to the ability of the workload to be moved across edge, core, and cloud locations. As AI workloads grow, the best deployment model may change, and enterprises may choose to move them to a different one.
- **Time.** This relates to the time continuity of the workload itself. AI workloads are increasingly designed to analyze streaming data in a real-time or near-real-time manner.

There is an emerging opportunity to not only develop and deploy AI near the data but also connect locations with low-latency, high-bandwidth networks and only move data that is used for reads/writes by an AI model across these networks. This way, the bulk of the data can always remain in place, regardless of where it is processed. The benefits of leaving data in place are substantial. Data movement is minimized, reducing infrastructure, administrative, network bandwidth, and potential egress costs. The time to move through the stages of the AI data pipeline to get to better business decisions is shorter. And the data security and/or governance implications of moving the data are minimized or entirely avoided.

DELL TECHNOLOGIES' AI SOLUTIONS

To help its customers succeed with AI, Dell Technologies has put together Dell Technologies Validated Designs for AI. These engineering-validated stacks make it easy for enterprises to buy, deploy, and manage successful AI projects, offering not only the underlying IT infrastructure but also the expertise to create optimized solutions that drive business value. Dell Technologies markets a range of systems for every AI scenario, allowing businesses to grow their capabilities at their own pace as their needs shift and as their data sets grow.

Challenges

The key challenge for enterprises, as businesses become more data driven, will be to identify and capture the data they need to improve their offerings and then use that data effectively to drive value for the business and its customers and partners. As they craft their strategies, enterprises will need to ensure that they respect the privacy of their constituents, effectively safeguard the data they do collect, and stay within various governments' regulatory requirements. A key consideration will be to build a cost-effective IT infrastructure that can continue to meet business performance, availability, and capacity requirements as their AI workloads and data continue to scale over time. Successful AI applications tend to grow very rapidly, and businesses will need to ensure they don't outgrow their supporting IT infrastructure.

Infrastructure requirements are evolving. Compute, networking, and storage infrastructure must provide higher performance and availability for workloads that will depend on data sets in the tens of petabytes range and beyond. In addition, the preferred deployment model for IT infrastructure will be hybrid multicloud, and technology providers need to support this strategy by offering products and services that can be deployed effectively and simply, either on premises or in public cloud environments. They will also need to support the ecosystem that is developing around AI workloads with a focus on providing end-to-end AI solutions that are easy to buy and deploy and offer the features needed across the spectrum of AI environments.

CONCLUSION

AI workloads will become increasingly common in enterprises over the next several years, and they will require many new capabilities from the underlying IT infrastructure. Enterprises should consider the infrastructure requirements for AI from three angles — scale, portability, and time — as they modernize their IT infrastructure for the data-centric digital era. Getting the underlying infrastructure right is a key determinant of success as enterprises look to AI to help drive better business decisions.

Enterprises successfully deploying AI will have their AI infrastructure distributed across edge, core, and cloud deployment locations, each of which will exhibit different workload profiles. Rather than thinking about AI infrastructure as a series of point deployments in different locations, enterprises should strive to craft a well-integrated end-to-end AI infrastructure strategy.

MESSAGE FROM THE SPONSOR

Read the paper "End-to-End AI is Within Reach."

Learn more about Dell Technologies AI solutions at delltechnologies.com/ai

About IDC

International Data Corporation (IDC) is the premier global provider of market intelligence, advisory services, and events for the information technology, telecommunications and consumer technology markets. IDC helps IT professionals, business executives, and the investment community make fact-based decisions on technology purchases and business strategy. More than 1,100 IDC analysts provide global, regional, and local expertise on technology and industry opportunities and trends in over 110 countries worldwide. For 50 years, IDC has provided strategic insights to help our clients achieve their key business objectives. IDC is a subsidiary of IDG, the world's leading technology media, research, and events company.

Global Headquarters

140 Kendrick Street
Building B
Needham, MA 02494
USA
508.872.8200
Twitter: @IDC
blogs.idc.com
www.idc.com