Generative AI Digital Assistants in the Enterprise

A scalable and modular production infrastructure for artificial intelligence-powered digital assistants

October 2024

H20028.1

Design Guide

Abstract

This guide describes the architecture and design of the Dell Validated Design for Generative AI Digital Assistants, also known as AI avatars, virtual humans, and digital humans. This document describes the solution and specifies the architecture and key hardware and software components. The solution is delivered along with Dell professional services for a complete implementation.

Dell Technologies Solutions

Dell
Validated Design

DELL Technologies

Copyright

Contents

Chapter 1 Introduction 4
Overview 4
Document purpose 5
Audience 6
Revision history 6

Chapter 2 Digital Assistant Solution Overview 7
What are digital assistants? 7
Use cases for digital assistants 9
Requirements for digital assistants 11
Workload characterization of digital assistants 12
Tenets of an architecture for AI-powered digital assistants 12

Chapter 3 Software Architecture 14
Conceptual architecture 14
Solution software components 18
NVIDIA AI Enterprise and LLM operations 22

Chapter 4 Physical Architecture 25
Architecture overview 25
Validation and performance data 27
Deployment scenarios 27
Dell Precision AI-Ready workstations 28

Chapter 5 Professional Services and Consulting 29
Professional Services 29
Customer Solution Centers (CSC) 30
Security Considerations 31

Chapter 6 Conclusion 33
Summary 33
We value your feedback 33

Chapter 7 References 34
Dell Technologies documentation 34
Partner documentation 34

Chapter 1 Introduction

Overview

As Artificial Intelligence (AI) and generative AI technologies evolve, human interaction with machine intelligence systems will change. In the last decade, we observed the transition from keyboard interfaces to natural human communication platforms using speech. Examples include speech-to-text (STT) for text messages and information systems such as Siri, Google, and Alexa. With large language models (LLMs) making significant leaps in natural language understanding, we are witnessing another jump forward in the useability of AI for both technical and nontechnical users. The frontier of a scenario where AI can comprehend nuances in human language has arrived. Generative AI, the branch of AI designed to generate new data, images, code, or other types of content that humans do not explicitly program, is rapidly becoming pervasive across nearly all facets of business and technology.

Based on the previously published Dell Validated Designs for Inferencing and Model Customization, this solution is the first to address a full-stack use case, from the server, storage, and networking hardware to a cross-industry solution that implements a digital assistant solution based on generative AI.

A digital assistant is a 2D or 3D representation of a persona that acts as a “co-pilot” or assistant to human users. This AI-powered avatar extends the traditional chatbot and voice experience because it can mimic various human body languages. It interprets clients’ input and provides factual responses and appropriate nonverbal reactions. Digital assistants can connect to any domain or company-specific dataset to share knowledge. They interact using verbal and nonverbal cues such as tone of voice and facial expressions. Digital assistants are accessible 24/7, making it possible to re-create natural human interaction at a scale. For customer service, digital assistants can provide client support and coaching while maintaining a human-like emotional connection. They are designed to offer the best of both AI and human conversation, enhancing the customer experience by providing personalized, real-time, and data-driven interactions.

An LLM is at the core of a digital assistant. It is an advanced type of AI model that has been trained on an extensive dataset, typically using deep learning techniques that can understand, process, and generate natural language text, voice, and images. However, AI built solely on public or generic models is not well suited for enterprises to use in their business as they have been trained in general knowledge and do not have the domain-specific knowledge needed for the customer scenario. Enterprise use cases require domain-specific knowledge to train, customize, and operate their LLMs. However, training LLMs from scratch is difficult, both from a skill and a compute capacity perspective.

Retrieval Augmented Generation (RAG) is an advanced AI technique that optimizes the output of an LLM by referencing an authoritative knowledge base outside of its training

data before generating a response. It is an architectural approach that improves the efficacy of LLM applications by retrieving relevant data or documents pertinent to a task and providing them as context for the LLM. This process helps to mitigate the unpredictability of LLM responses and ensures that the generated content remains relevant, accurate, and useful in various contexts.

Dell Technologies has designed a scalable, modular, and high-performance architecture that enables enterprises to create a range of generative AI-based digital assistant solutions that apply specifically to their businesses, reinvent their industries, and give them competitive advantages.

This guide describes the Dell Validated Design for Generative AI Digital Assistants for an on-premises enterprise environment. It is based on high-performance Dell and NVIDIA infrastructure and uses innovative technologies, such as RAG for secure information retrieval based on the Pryon Information Retrieval Platform and the 2D/3D rendering capabilities of the Digital Human Platform by UneeQ.

This Dell Validated Design for on-premises digital assistants is the first application-level solution in a series of designs for generative AI that focus on all facets of the generative AI life cycle, based on designs for model training, customization, and inferencing. This solution, which is implemented along with a Dell Professional Services engagement, enables the widespread industrialized rollout and adoption of next-generation workforce automation capabilities.

Document purpose

A **Dell Validated Design (DVD)** is a strategic approach that streamlines the process of designing, testing, and integrating components for specific use cases. It is built on these key tenets:

1. **Reliability**—DVDs provide proven and documented solutions through design guides and white papers, which help to reduce risk and improve operational efficiency.
2. **Efficiency**—DVDs are fully integrated and have been pretested and validated by Dell Technologies experts and their partners so that customers spend less time planning, deploying, and testing.
3. **Flexibility**—DVDs offer a choice of consumption models, including as-a-service through Dell Technologies APEX solutions and implementations led by Dell Professional Services, which provides better alignment with application, IT, and business strategies.

The purpose of a DVD is to provide reliable, efficient, and flexible solutions that are designed to reduce the time, effort, and risk required to implement infrastructures and use cases.

This guide explains the concept of digital assistants, also known as AI avatars, virtual humans, and digital humans, as well as many of the use cases. It describes the architecture and design of the Dell Validated Design for Generative AI Digital Assistants.

Note: This guide uses the term digital assistant, whereas UneeQ uses the term “digital human.” The terms are interchangeable; however, digital human is used when referring specifically to the UneeQ platform.

Because this solution for digital assistants is accompanied by a Dell professional service engagement, we present only a portion of the full design and configuration information, as well as the validation and performance data. All information is available as part of an implementation.

This design guide can be read with the Generative AI in the Enterprise white paper. The white paper provides an overview of generative AI, including its underlying principles, benefits, architectures, and techniques; the various types of generative AI models and how they are used in real-world applications; the challenges and limitations of generative AI; and descriptions of the various Dell and NVIDIA hardware and software components to be used in the series of validated designs to be released.

Audience

This design guide is intended for experts interested in the implementation of solutions and infrastructure for customer-facing generative AI, including professionals and stakeholders involved in the development, deployment, and management of generative AI systems for customer care and customer service.

Key roles include AI architects and IT infrastructure architects and designers. Other audience members might include system administrators and IT operations personnel, AI engineers and developers, data scientists, and AI researchers. Some knowledge of generative AI principles and terminology is assumed, including familiarity with the referenced Dell documentation.

Revision history

Table 1. Revision history

Date	Version	Change summary
May 2024	1.0	Initial release
October 2024	1.1	Updated for hardware configurations and Pryon version

Chapter 2 Digital Assistant Solution Overview

This chapter provides an overview of digital assistants and describes the use of generative AI to support them, including how the core components of the solution fit into the general workflow of generative AI development and operations. It also presents several use cases for digital assistants.

What are digital assistants?

Our objective is to build an ethical and secure digital assistant interface platform that uses concepts originally developed for gamification applied to enterprise systems. By integrating Dell's generative AI solution with an ecosystem of advanced AI software platforms, we can harness technology that was traditionally focused on gaming use cases and apply it to real-world business applications. The fully integrated solution designed by Dell Technologies brings digital assistants to the forefront of human interactions and drives operational efficiencies in both internal use cases and externally facing customer interactions.

By building this ecosystem and evolving the next generation of technologies, we aspire to bring the intelligence grid of the future to every customer and drive practical uses in retail, healthcare, and beyond.

The following figure depicts an example of a digital assistant that has been programmed to act as a technical support agent by supporting multiple modalities:

- The visual avatar interface (the digital assistant persona)
- The voice and language
- The chat functionality

Figure 1. Example of digital assistant implementation as a technical support agent

Depending on a user's preference, the buttons at the bottom of the screen allow the system to turn on or turn off individual modalities.

UneeQ

Dell Technologies has partnered with UneeQ for our digital assistant frontend. UneeQ has differentiated itself from the competition by not only focusing on generative AI for life-like animations, but by also addressing the visual components associated with what people find to be unnatural. The nuances in the extent of lip movement along with emotion driven from the eyes, ears, and forehead are important. These nuances move people beyond the uneasy reaction that often occurs (also known as uncanny valley responses) at the first interaction. UneeQ is sensitive to cultural differences and has considered which animations are acceptable in one region compared to another.

The need for translation and speech in any language also requires that the synthetic generation dynamically adjusts not only to sentiment but also to movements that account for the user's background. As we focus on pushing the boundaries of human acceptance, Dell Technologies partners with UneeQ to bring our high-performance infrastructure and AI solutions expertise into the mix. This migration to an on-premises solution enables us to achieve response times in line with a truly interactive conversation: for each question asked, the latency of the answer is less than a few seconds.

Pryon

For implementing a RAG approach, we have partnered with Pryon, which provides a commercial off-the-shelf AI-powered data retrieval product. The Pryon Retrieval Engine is a versatile AI tool that securely extracts answers from all types of content including audio,

images, text, and video. Pryon ensures high-precision ingestion by emulating human-like document understanding, employing proprietary OCR technology to extract text in reading order from images, graphics, schematics, and handwritten notes. Pryon's retrieval process uses a proprietary combination of query expansions, out-of-domain detection, and query embedding to comprehensively understand natural language queries for ingested content matching. Then, Pryon uses three proprietary models to find and rank the best matched content quickly. All answers are presented with attribution to the source content.

Pryon seamlessly integrates with major enterprise knowledge base systems, enabling users to construct a unified retrieval knowledge base effortlessly without content consolidation. Pryon Retrieval Engine maintains document-level access control lists (ACL), and Pryon AI models do not train on enterprise data. This combination of features allows Pryon Retrieval Engine to offer industry-leading accuracy, at enterprise scale, with security controls, and fast time-to-value.

Use cases for digital assistants

Digital assistants are next-generation conversational AI systems that go beyond chatbots and voice processing. While the Generative AI in the Enterprise white paper discusses multiple use cases for generative AI across various industries, some examples of use cases specifically for digital assistants include:

- **Customer service**—Customer service and support activities use conversational agents, chatbots, and digital assistants extensively by generating natural language responses based on user queries or instructions. While the 2010s focused on automating tasks and business processes “behind the service desk,” AI-powered digital assistants enable enterprises to automate the front-end aspects of customer service. In an age in which staffing customer service agent positions is a challenge, digital assistants represent a welcome opportunity to scale up the virtual work force through an additional conversational channel.
- **Education and training**—Digital assistants provide personalized responses, recommendations, and content creation. While generative AI can be used to generate education and training materials, a digital assistant can deliver the training itself interactively and allow for questions. In contrast to a human trainer, a digital assistant is always patient and never runs out of time.
- **Marketing and sales**—Digital assistants can be used in marketing and sales to automate and enhance communication with customers. They provide personalized responses, product recommendations, and content creation. For instance, a digital assistant interprets customer queries and generates relevant responses, improving customer support, which leads to higher customer engagement and conversion rates.
- **Digital kiosks**—Digital assistant-enabled kiosks can revolutionize customer service by providing intelligent, personalized, and efficient service. They use AI technologies such as machine learning, natural language processing, and computer vision to interact with customers in a more human-like manner. Also, they can understand and respond to customer queries, provide personalized recommendations based on customer preferences, and potentially recognize customers through facial recognition technology. Equipping these kiosks with 3D displays provides a seamless, personalized customer experience that goes far

beyond what traditional kiosks can offer. Moreover, digital kiosks can learn from each interaction, continuously improving their performance and accuracy over time. By integrating such AI workforce-enabled kiosks into their customer service strategy, enterprises can provide superior service, enhance customer satisfaction, and gain a competitive edge in the market.

In addition to these use cases, there are multiple other applications for digital assistants, including applications such as:

- **Self-service knowledge bases**—Generative AI can automatically generate and update knowledge base articles, FAQs, and troubleshooting guides. When customers encounter issues, they can search the knowledge base to find relevant self-help resources that provide step-by-step instructions or solutions.
- **Contextual responses, problem solving, and proactive troubleshooting**—Digital assistants using generative AI can analyze customer queries or problem descriptions and generate contextually relevant responses or troubleshooting suggestions. By understanding the context, the solution can offer tailored recommendations, guiding customers through the troubleshooting process.
- I**nteractive diagnostics**—Digital assistants can conduct interactive diagnostic conversations to identify potential issues and guide customers towards resolution. Through a series of questions and responses, the system can identify the problem, offer suggestions, or provide the next steps for troubleshooting.
- **Intelligent routing and escalation**—Generative AI models supporting digital assistants can intelligently route customer inquiries or troubleshoot specific issues based on their complexity or severity. They can determine when a query must be escalated to human support agents, ensuring efficient use of resources and timely resolution.
- **Automating contact center operations**—Contact centers typically operate in a voice-only mode. A first step to introduce digital assistants is to use the voice feature of this conversational AI system to replace human labor by starting with Level 1 and Level 2 customer support.

While the validation work that we performed in this design is primarily centered on an avatar in front of a static background, there are also logical extensions to the use cases of generative AI that include virtual worlds and environments. Digital assistants can be placed in virtual worlds, landscapes, or architectural designs for use in augmented reality (AR), virtual reality (VR), the Metaverse, or simulations.

These examples show how generative AI-powered digital assistants are applied across various domains. The versatility of this cross-industry solution allows for creative and innovative applications in customer service, content and code generation, creative arts, virtual environments, personalization, and more. The use cases continue to expand as digital assistants powered by generative AI technologies advance.

Requirements for digital assistants

There are several important requirements for a digital assistant solution. Key functional requirements include:

- **Accuracy, explainability, and traceability**—Understanding and explaining the reasoning behind the model's predictions or responses is crucial for digital assistants because they disseminate authoritative information. Therefore, accountability, transparency, and ethical considerations are of paramount importance. Any digital assistant architecture must include an authoritative knowledge base in an information retrieval system. The knowledge base allows the establishment of data lineage for any of the model's decisions during inferencing by identifying the document from which the information has been gathered.
- **Quality control and error handling**—Detecting and handling errors or inaccuracies during inferencing is important to maintain the quality and reliability of the system. Implementing effective error handling, monitoring, and quality control mechanisms to identify and rectify issues is essential. The typical mechanism for achieving these results is to provide for a human in the loop and Reinforcement Learning with Human Feedback (RLHF).
- **Latency and responsiveness**—Achieving low-latency and highly responsive inferencing is essential for an interactive application such as a digital assistant. The overall round-trip delay throughout the system (consisting of the steps of request → intent identification → information retrieval → LLM → response) must be considered for the experience to be truly interactive. Balancing the computational demands of the model with the need for fast responses can be challenging, particularly when dealing with high volumes of concurrent user requests.
- **Computational resources**—High-quality digital assistant implementations are computationally intensive, especially for large and complex applications. Generating predictions or responses in real time requires significant processing power, memory, network, and storage resources.
- **Deployment scalability**—To handle increasing user demand, scaling up the deployment of a model is needed. Ensuring that the system can handle the required number of concurrent inferencing requests with acceptable latency and can dynamically allocate resources to meet the workload can be complex, requiring careful architecture design and optimization.

These challenges highlight the need for careful consideration and optimization in various aspects of inferencing, ranging from computational efficiency and scalability to model optimization, interpretability, and quality control. Addressing these challenges contributes to the development of robust and reliable AI systems for powering digital assistants.

Dell Technologies and its partners help solve these challenges by collaborating to deliver a validated and integrated hardware and software solution. This solution is built on Dell high-performance best-in-class infrastructure and uses the award-winning software stack and the industry-leading accelerator technology and AI enterprise software stack of NVIDIA.

Workload characterization of digital assistants

Dell Technologies provides a diverse selection of acceleration-optimized servers with an extensive portfolio of accelerators. Due to the specific workload characteristics of a digital assistant system and its overall sizing and scaling needs, we chose a particular set of PowerEdge servers, the PowerEdge R760xa and R660 servers (see Chapter 4 for more details).

To understand this decision better, consider the characteristics of a digital assistant workload:

- From a customer perspective, the most important characteristics are the number of tenants, the number of simultaneous conversations, and the response time – which most directly affects the user experience.
- The workload exhibits a high degree of parallelism. Incoming requests are entirely independent from another as they reflect the scope and type of requests from users. This case holds true for all components in the system.
- The LLM workload of a digital assistant solution is characterized as an inference workload, as opposed to a training workload, so we designed an inference-optimized infrastructure.
- Cost is a key concern for any AI workload. To offer a high-performing digital assistant solution while addressing cost concerns, it is imperative to find the right GPU/server mix. In this design, we answer the following questions with an experimental validation of the documented configurations to avoid costly (and lengthy) trial-and-error runs when the solution is deployed at customer sites:
  - How many servers are required?
  - How many GPUs are required in each server?
  - What type of GPUs are best suited for the workload processed?
  - The key variable part of the system, assuming that latency is held constant, is essentially the number of simultaneous conversations that the system must process. When properly designed, the system can easily scale up and down by adding and removing servers in the respective tiers, as needed.

A cloud-native design based on Kubernetes containers addresses the operations and management requirements. The Kubernetes control plane itself consumes few compute resources. Kubernetes best practices dictate implementing the control plane on three independent control nodes. For the Kubernetes control nodes, we chose Dell PowerEdge R660 servers, which offer immense value for limited-size management workloads, while meeting the Kubernetes quorum requirement.

Tenets of an architecture for AI-powered digital assistants

The use cases and requirements introduced in this chapter are guidelines for developing the architecture for digital assistants. This section describes the key tenets of the architecture that influence the design decisions that were made to assemble this solution from Dell Technologies products, from open-source implementations, and from products stemming from the Independent Software Vendors (ISVs) Pryon and UneeQ.

Open-source software

To facilitate integration into existing customer environments, the solution uses a combination of leading-edge commercial components and widely adopted open-source software. This approach applies to all layers of the system. Starting from operating systems, middleware, and databases, this integration extends further to control planes, observability tools, and dashboards. Including open-source components provides for easier integration into existing environments and lowers the learning curve when the solution is handed over to the operations team.

Cost-effectiveness

Where possible, the most cost-effective alternatives are chosen for hardware infrastructure to find the most cost-effective server/CPU/ GPU/storage combination. The key goal of performance benchmarking and sizing efforts is to allow an implementation team to make the right tradeoffs between different alternatives. For instance, a GPU accelerator that offers a better ratio of performance to cost is favored over an alternative setup.

Cloud native

Every component of the architecture must be automatically deployable and managed by a cloud-native, declarative approach. This approach is achieved by using Kubernetes, a cloud-native open-source technology for orchestrating and managing the deployment of containerized environments. Ensuring that each component of the solution can be deployed (using Helm charts) and managed by Kubernetes from a single point of control allows for one common way of managing the solution.

High availability

Overall deployment must be designed to provide High Availability (HA) or Disaster Recovery (DR). This requirement includes, but is not limited to, the Kubernetes control plane, log aggregation, telemetry, alerting, image repository, Kafka, Redis, or other shared components.

Observability

The solution uses standard observability frameworks such as Prometheus (for metric collection) and Grafana (for dashboards). These frameworks establish a single point of control of all hardware and software components in the overall solution.

Managed service readiness

The deployment of generative AI solutions might suffer from the lack of skills to run and evolve the solutions as part of a continuous improvement process. To address this challenge, enterprises frequently procure such solutions as managed services, whether the solution is delivered on-premises or hosted by a managed service provider (MSP). The present solution enables this option by delivering a single control point, for which pre-existing integration into a managed services environment (such as the Dell Managed Services Platform (DMSP)) can be used.

Chapter 3 Software Architecture

This chapter describes the overall architecture and the primary software components used for the digital assistant solution, beginning with the conceptual architecture and then describing the specific software components used in the design.

Conceptual architecture

Similar to most distributed applications, our digital assistant solution follows the Model-View-Controller (MVC) design pattern. In addition to the LLM and information retrieval components, multiple components of the digital assistant solution are based on generative AI.

The following figure shows the key components of the digital assistant solution. They match the MVC pattern as follows:

- Model—LLM and information retrieval
- View—Digital assistant rendering platform
- Controller—Orchestration and conversation management

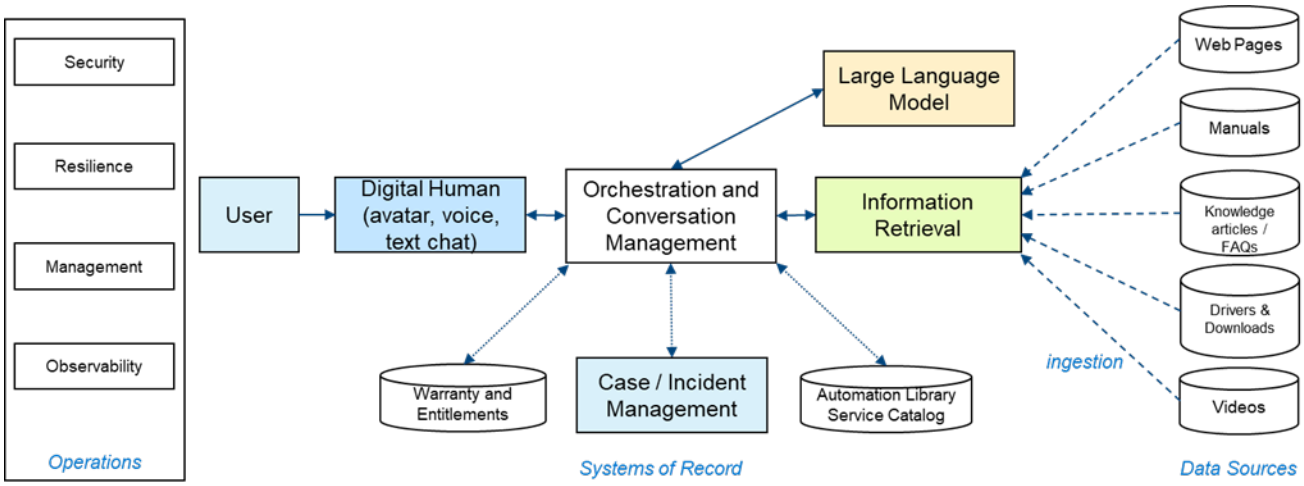

Figure 2. Digital assistant conceptual architecture

While the solution consists of four core components (digital assistant rendering platform, orchestration and conversation management, information retrieval, and large language model), you draw the most benefit from the system if high-quality data sources are available, and the existing systems of record are integrated. These components are all described in the following sections.

Digital assistant rendering platform

A **digital assistant rendering platform** (also known as a digital human platform) provides the user-facing frontend of the overall solution, specifically the avatar, voice, and chat capabilities. While the avatar with voice (and additional text chat) capabilities provides the most immersive experience, customers can turn on and turn off different modalities, if needed. A text-only experience – effectively emulating a chatbot – might be desirable if the available network bandwidth between the client device and the digital assistant solution is low.

This component interacts with the orchestration and conversation management component using a speech-to-text component to submit requests provided by the user to the overall solution. Conversely, a text-to-speech component delivers the solution responses back to the user.

Orchestration and conversation management

Orchestration and conversation management functionality is at the core of the solution and ties the different software subsystems together. It consists of the following key components:

- **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.

Information retrieval

Information retrieval (IR) systems, sometimes referred to as knowledge bases, are designed to manage, process, and retrieve information from a vast array of data sources. They ingest data, maintain either a copy or reference to the original documents, and store this information for efficient and accurate retrieval. These systems are fundamental to many applications, including search engines, recommendation systems, and digital libraries, among others.

The primary goal of an IR system is to provide authoritative information, minimizing or eliminating hallucination. Hallucination is the generation of information that is not grounded in the source data. By ensuring that the retrieved information is accurate, relevant, and based on the original documents, IR systems enhance the reliability and credibility of the results.

IR systems employ sophisticated algorithms and techniques to index, search, and retrieve data, ensuring that the digital assistant solution can find the most relevant information that it needs from the vast digital landscape. By providing quick and easy access to this information, IR systems help find the most relevant information for a given query from thousands of documents that are ingested from various data sources.

Data sources

Data sources represent the authoritative, domain-specific data that an enterprise owns and that must be ingested into the IR system. Typical examples of data sources are:

- **Web pages**—Documents that are accessible on the Internet, and that typically represent the nucleus of data that forms the body of knowledge on which a digital assistant solution is built.
- **Product manuals**—Detailed documents provided by product manufacturers. They contain essential information about the product, including how to use it, safety instructions, and troubleshooting tips.
- **Knowledge articles and FAQs**—Technical resources often found on company websites. They provide answers to frequent questions (FAQs) and offer in-depth information (knowledge articles) about specific topics related to the company's products or services.
- **Rich media, pictures, and videos**—Pictures and videos that can provide abundant information in an easily digestible format. They are often used for tutorials, product demonstrations, and illustrations of troubleshooting procedures.
- **Drivers and downloads**—Drivers that you can download from the manufacturer's website and are essential for the proper functioning of the device.

The original content continues to exist in addition to the digital assistant solution. For example, a website continues to exist even if the digital assistant solution makes this content more accessible to a broad user base.

Large language models

Large language models (LLMs) are advanced natural language processing models that use deep learning techniques to understand and generate human language. LLMs can include a range of architectures and approaches, such as recurrent neural networks and transformers. In particular, Generative Pre-trained Transformer (GPT) is a popular and influential example of an LLM that is based on the transformer architecture, which is a deep neural network architecture designed to handle sequential data efficiently. Transformers use self-attention mechanisms to process input sequences and learn contextual relationships between words, enabling them to generate coherent and contextually relevant language.

In the digital assistant solution, we follow a RAG approach to improve accuracy and relevance. Hence, the LLM plays a slightly reduced role (in comparison to pure LLM-based approaches) as the LLM is given several prompts along with factual information that has been obtained from the IR system. The purpose of the LLM is to provide succinct summaries of the obtained information so that:

- The information can be displayed on text panels or buttons in the chat canvas.
- A high-level summary of the content as the ‘talk track’ is provided for the digital assistant. Merely reading the detailed information about the panels causes a detrimental user experience.

There are multiple LLMs used across the solution, some embedded in other components of the overall solution, and another to provide for the dialog management. The following considerations apply to each LLM.

Foundation models

A foundation or pretrained model is a machine learning model that has been trained on a large dataset for a specific task before it is fine-tuned or adapted for a more specialized task. Foundation models are crucial because they provide a starting point that already understands a broad range of concepts and language patterns. Beginning with a pretrained foundation model makes the process of customizing and fine-tuning for specific tasks more effective and efficient.

Parameters

Parameters in LLMs refer to the learnable components or weights of the neural network that make up the model. These parameters determine how the model processes input data and makes predictions or generates output. Typically, GPTs are measured in billions of parameters. These parameters are learned during the training process, in which the model is exposed to vast amounts of data and adjusts its parameters to generate language. Assuming that the model architecture and training data are comparable, generally the higher the parameters in the model, the greater the accuracy and capability of the models. Sometimes a smaller model that is trained to be specific to a particular outcome might be more accurate. Models with higher parameters also require more compute resources, especially GPU resources.

Accuracy

The accuracy of LLMs is typically measured based on their performance on specific NLP tasks. The evaluation metrics used depend on the nature of the task. The publicly available ChatGPT and Llama 2 models are foundation models. While these models offer

a strong starting point with general capabilities, they are often customized for specific use. This design guide does not address model customization as we use the LLM with RAG.

Systems of record

Systems of record play a pivotal role in shaping the user experience and streamlining operations. These systems are domain-specific backend systems that establish context based on user-supplied information, facilitating personalized and efficient service delivery. They serve as the backbone of many business operations, providing critical data and functionality. Some examples of these systems include:

- **Customer repositories**—These databases store comprehensive information about customers, including their contact details, preferences, transaction history, and more. They enable businesses to understand their customers better and tailor their services accordingly.
- **Entitlement databases**—These databases are crucial in customer service scenarios. They identify the warranty and level of support to which a customer is entitled, based on a user-supplied device ID or serial number. This identification helps businesses provide appropriate support and manage expectations.
- **Case management systems**—These systems manage and track customer support cases or tickets. They allow the system to reduce the potential scope of inquiries, for instance, if support cases and tickets are open for a given device. This process leads to quicker resolution times and better customer satisfaction.
- **Automation libraries and service catalogs**—These backend systems supply additional information and carry out actions identified during the conversation that were implicitly or explicitly approved by the requester. They automate routine tasks, increasing efficiency and reducing the scope for human error.

Integrating systems of records into the digital assistant experience represents a significant stride towards a more data-driven and customer-centric approach to business operations.

Solution software components

Whereas the previous section described the conceptual architecture and software, this section describes the specific software modules and components used in the design.

UneeQ Digital Human Platform

The UneeQ Digital Human Platform is used in the digital assistant solution and implemented with an industry-leading software stack. UneeQ is an ISV specializing in immersive customer experiences. Their platform offers a simple interface and a reliable digital assistant that looks and sounds natural. Enterprises use the UneeQ portal and tools to create instances of digital assistants and conversation streams.

The UneeQ Digital Human Platform is built on Unreal Engine by Epic Games, a 3D graphics game engine, and adds the following distinct components, all delivered as images that can be deployed and managed using Kubernetes:

- **Digital Human renderer**—The renderer is the component that displays the details of the avatar representing the digital assistant. This component requires GPUs to achieve high quality.

- **Digital Human variant options**—The platform offers multiple digital assistant variant options, allowing users to choose a digital assistant that best suits their needs.
- **Virtual background design capabilities**—The platform provides virtual background design capabilities, enabling users to customize the digital environment in which the digital assistant operates.
- **Persona creation**—The platform allows persona creation, enabling users to define the personality and behavior of the digital assistant.
- **Voice and language customization**—The platform offers voice and language customization options, enabling the digital assistant to communicate in a way that aligns with the user's preferences.
- **Integrated NLP of choice**—The platform integrates with the NLP system (see the description of the Conversational Pod below), enabling the digital assistant to understand and respond to user queries.
- **Desktop and mobile browser support**—The platform supports users interacting with digital assistants through the web browser on their desktop or mobile phone.

A core component of the UneeQ Digital Human Platform is the NVIDIA Audio2Face microservice, which animates 3D character facial characteristics to match any audio input and is part of NVIDIA AI Enterprise.

Dell Orchestration and Conversation Management

The Dell Orchestration and Conversation Management component is a subsystem of multiple software components that enables the secure definition and management of multiple users simultaneously. It is necessary because different tenants, or types of users, might be defined in the system, and each use case implemented by the solution might require access to a different combination of backend systems such as LLMs and IR systems. For example, an enterprise might choose to provide help desk support to its employees using a digital assistant while offering a digital assistant-enabled storefront to its customers. While the digital assistant avatar might be identical, different user experience (UX) elements might be enabled for each use case, and different backend systems might need to be accessed.

It is necessary to integrate the digital assistant solution into an existing customer environment seamlessly and securely. In particular, the entry point of the solution is provided by this subsystem, which resides in its own Kubernetes namespace and includes multiple capabilities, including and ingress control, user interface management, transaction orchestration, conversation management, and others.

Pryon Retrieval Engine

The Pryon Retrieval Engine is a commercial off-the-shelf information retrieval system. It uses several proprietary LLMs to represent ingested content to find the most relevant content rapidly. Content is topically grouped as "collections," which are accessed one-by-one. Key Pryon characteristics include:

- **High-precision ingestion is the starting point for accuracy**. Pryon emulates human-like document analysis, employing proprietary OCR technology to extract text in reading order from images, graphics, schematics, and handwritten notes. It uses vision segmentation to identify and label key components, performs content

normalization and filtering to remove unnecessary objects, and employs visual semantic segmentation to assemble smarter document chunks.

- **Accurate retrieval begins with understanding the question**. Pryon's retrieval process uses a proprietary combination of query expansions, out-of-domain detection, and query embedding to comprehensively understand natural language queries for ingested content matching. Then, Pryon uses three proprietary models to quickly find and rank the best matched content.
- **Enterprise scale is key to sustained value**. Pryon seamlessly integrates with major enterprise systems such as Microsoft SharePoint, Confluence, AWS S3, Google Drive, Zendesk, ServiceNow, Salesforce, and Documentum, enabling users to effortlessly construct a unified retrieval knowledge base without content consolidation.
- **Maximum security is a prerequisite to AI adoption**. Pryon Retrieval Engine maintains a document-level ACL and can be implemented in on-premises and air-gapped environments. Pryon AI models do not train on enterprise data.
- **Rapid time to value is necessary to get business buy-in**. Pryon's prebuilt Ingestion, Retrieval, and Generative Engines make it possible to implement the digital assistant solution efficiently.
- **A comprehensive admin console is required**. The admin console provides the ability to carry out standard administration tasks, launch content ingestion, and define synonyms and validated answers to improve the guidance for the search. Query history, volume, and other metrics can also be viewed at a collection or system level.

Llama 3 LLM

True to its purpose of an on-premises solution, the digital assistant solution uses the Llama 3 model from Meta running on a Dell PowerEdge R760xa server with NVIDIA GPUs.

AI models are typically optimized for performance before deployment to production. Optimized models offer faster inference speed, improved resource efficiency, and reduced latency, resulting in cost savings and better scalability. They can handle increased workloads, require fewer computational resources, and provide a better user experience with quick responses.

To test the quality of LLM-generated responses, multiple methods are used:

- **Using another LLM to check the response quality**—Responses from GPT 3.5 turbo and Llama 3 are ranked using ChatGPT.
- **Answer relevancy**—The answer relevancy metric measures the generated response's quality by evaluating how relevant the output of the LLM is compared to the provided input by considering the retrieval context.
- **Summarization**—The summarization metric uses LLMs to determine whether the LLM is generating factually correct summaries while including the necessary details from the original text.
- **Faithfulness**—The faithfulness metric measures the quality of the LLM by evaluating whether the output aligns factually with the contents of the retrieved context.

- **Contextual precision**—The contextual precision metric measures the LLM by evaluating whether nodes in the retrieval context that are relevant to the given input are ranked higher than the irrelevant ones.
- **Contextual relevancy**—The contextual relevancy metric measures the output of the LLM by evaluating the overall relevance of the information in the retrieval context for a specific input.
- **Contextual recall**—The contextual recall metric measures the quality of the LLM by evaluating the extent to which the context aligns with the output.
- **Hallucination**—The hallucination metric determines whether the LLM generates factually correct information by comparing the output to the provided context.
- **Bias**—The bias metric determines whether the LLM output contains gender, racial, or political bias.
- **Token cost**—The token cost metric calculates the token cost of the LLM.

Other software components

Due to the solution's considerable complexity, it is imperative to provide a single point of control, both from a security operation and from an administration and management perspective.

In this validated design, the following components facilitate solution operations:

- **Kubernetes** is an open-source system for automating deployment, scaling, and management of containerized applications. It groups the containers that make up an application into logical units for easy management and discovery.
- **Prometheus** is an open-source systems monitoring and alerting toolkit. It collects and stores metrics as time series data, that is, metric information is stored with the timestamp at which it was recorded, alongside optional key-value pairs called labels. Prometheus excels at collecting time series data from various sources. It can scrape metrics from applications, services, and infrastructure components.
- **Grafana** is a multiplatform open-source analytics and interactive visualization web application. It provides charts, graphs, and alerts for the web when connected to supported data sources. Grafana provides the tools to transform raw data into meaningful insights through rich, interactive dashboards.

Together, these three systems deliver full observability. Kubernetes provides a platform for running and managing applications. Prometheus, tightly integrated with Kubernetes, collects metrics, and provides real-time monitoring. Grafana, in turn, uses the data collected by Prometheus to create visualizations, providing a user-friendly way to monitor the performance, reliability, and overall health of the Kubernetes-managed applications.

This validated design also uses Keycloak, which is an open-source Identity and Access Management (IAM) tool. It offers features like user federation, identity brokering, and social login. Keycloak simplifies the authentication process and eliminates the need to handle user storage or authentication. It also supports single sign-on and single sign-out. Keycloak can integrate with existing IAM solutions such as Microsoft Active Directory, Ping Identity, and Okta.

NVIDIA AI Enterprise and LLM operations

Large language model operations (LLMOps) refer to the specialized practices and workflows that accelerate development, deployment, and management of LLM models throughout their complete life cycle. These operations include data preprocessing, language model training, monitoring, fine-tuning, and deployment. Similar to machine learning operations (MLOps), LLMOps enables collaboration among data scientists, AI developers, DevOps engineers, and IT professionals.

NVIDIA AI Enterprise provides enterprise support for various software frameworks, toolkits, workflows, and models that support inferencing. For the scope of this design, we used several NVIDIA capabilities in support of LLMOps, as described below.

As well as providing enterprise support for components such as Triton Inference Server and TensorRT-LLM, NVIDIA AI Enterprise is also required for Audio2Face, an NVIDIA Inference Microservice (NIM) that is part of the UneeQ software implementation and used to animate 3D character facial movements to match the audio input.

Triton Inference Server

NVIDIA Triton Inference Server (also known as Triton) is an inference serving software that standardizes AI model deployment and execution and delivers fast and scalable AI in production. Enterprise support for Triton is available through NVIDIA AI Enterprise. Alternatively, Triton is available as open-source software.

Triton streamlines and standardizes AI inference by enabling teams to deploy, run, and scale trained machine learning or deep learning models from any framework on any GPU- or CPU-based infrastructure. It provides AI researchers and data scientists the freedom to choose the appropriate framework for their projects without impacting production deployment. It also helps developers deliver high-performance inference across cloud, on-premises, edge, and embedded devices.

Benefits

The benefits of Triton for AI inferencing include:

- **Support for multiple frameworks**—Triton supports all major training and inference frameworks, such as TensorFlow, NVIDIA TensorRT, NVIDIA FasterTransformer, PyTorch, Python, ONNX, RAPIDS cuML, XGBoost, scikit-learn RandomForest, OpenVINO, custom C++, and more.
- **High-performance AI inference**—Triton supports all NVIDIA GPU-, x86-, ARM CPU-, and AWS Inferentia-based inferencing. It offers dynamic batching, concurrent execution, optimal model configuration, model ensemble, and streaming audio/video inputs to maximize throughput and use.
- **Designed for LLMOps**—Triton integrates with Kubernetes for orchestration and scaling, exports Prometheus metrics for monitoring, supports live model updates, and can be used in all major public cloud AI and Kubernetes platforms.
- **Support for model ensembles**—Because most modern inference requires multiple models with preprocessing and postprocessing to be run for a single query, Triton supports model ensembles and pipelines. Triton can run the parts of the ensemble on CPUs or GPUs and allows multiple frameworks inside the ensemble.

- **Enterprise-grade security and API stability**—NVIDIA AI Enterprise includes NVIDIA Triton for production inference, accelerating enterprises to the leading edge of AI with enterprise support, security, and API stability while mitigating the potential risks of open-source software.

The Triton Inference Server is software that hosts generative AI models, which in our case is Llama 3. Triton Inference Server, along with its integration with TensorRT-LLM and the NeMo framework provides the infrastructure for deploying generative AI models.

TensorRT-LLM

NVIDIA recently unveiled TensorRT-LLM as a new backend for hosting LLMs. It encompasses the TensorRT deep learning compiler, featuring optimized kernels, preprocessing and postprocessing workflows, and specialized multi-GPU and multi-node communication components, all of which contribute to remarkable performance gains on NVIDIA GPUs. TensorRT-LLM empowers developers to explore novel LLMs, providing both peak performance and rapid customization options. For additional insights and in-depth information, see NVIDIA's Technical Blog.

NVIDIA offers the TensorRT-LLM API to define LLMs and build TensorRT engines to perform inference on NVIDIA GPUs. It includes a backend for integration with the NVIDIA Triton Inference Server to serve the LLMs. Models built with TensorRT-LLM can be run on a wide range of configurations going from a single GPU to multiple nodes with multiple GPUs using Tensor Parallelism and Pipeline Parallelism.

The Python API of TensorRT-LLM is designed to be similar to the PyTorch API. It provides both a functional module and a layer's module to assemble LLMs. To maximize performance and reduce the footprint, models can run using different pretraining and post training quantization using the NVIDIA Algorithmic Model Optimization (AMMO) toolkit. The first step to create an inference solution is to either define your own model or select a predefined network architecture from the list of models supported by TensorRT-LLM. Then, the model must be trained using a training framework like NVIDIA NeMo or PyTorch because TensorRT-LLM does not support model training. For predefined models, checkpoints can be downloaded from various providers. TensorRT-LLM can use model weights obtained from HuggingFace Hub.

Equipped with the model definition and the weights, TensorRT-LLM's Python API is used to re-create the model in a format that TensorRT can compile into an efficient engine. TensorRT-LLM provides components to create a runtime that starts the TensorRT engine. The Triton backend for TensorRT-LLM serves TensorRT-LLM models by using the Triton.

The final optimized model is ready for production deployment on a PowerEdge R760xa server equipped with NVIDIA GPUs using Triton Inference Server. It can be accessed through an API endpoint or using HTTPS/gPRC protocols. The Triton Inference Server also offers health and performance metrics of the model in production that can be consumed and visualized through Prometheus and Grafana (see Other software components). These monitoring tools provide valuable insights into the model's performance and overall system health during deployment.

GPU Operator

The NVIDIA GPU Operator automates the life cycle management of the software required to expose GPUs on Kubernetes. It enables advanced functionality, including better GPU performance, utilization, and telemetry. The NVIDIA GPU Operator uses the Kubernetes

operator framework to automate the management of all NVIDIA software components needed to provision the GPU.

These components include the drivers to enable NVIDIA's Compute Unified Device Architecture (CUDA), the Kubernetes device plug-in for GPUs, the NVIDIA Container Toolkit, automatic node labeling using GPU Feature Discovery (GFD), and Data Center GPU Manager (DCGM) monitoring, among others.

Chapter 4 Physical Architecture

This chapter describes physical architecture, including the relationship to the main software components, and the primary infrastructure components for the digital assistant solution.

Because this solution for digital assistants is designed to be accompanied by a Dell professional service engagement, we present only a portion of the full design and configuration information. All information is available as part of an implementation.

Architecture overview

The following figure shows an overview of the physical architecture, along with the main software components:

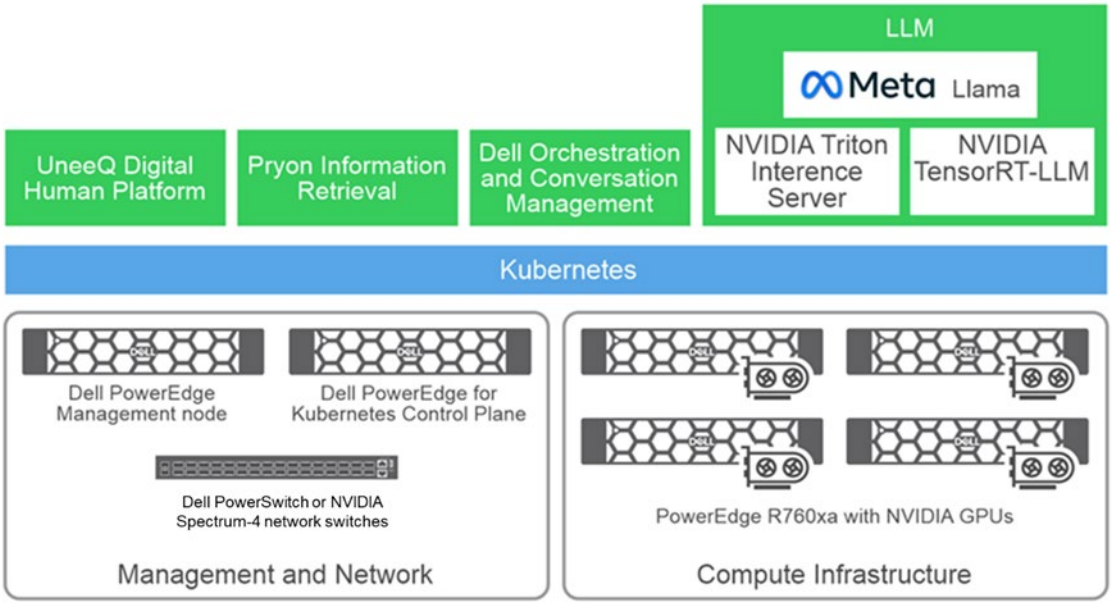

Figure 3. Digital assistant solution architecture with physical infrastructure

Dell servers and NVIDIA GPU options

Dell Technologies provides a diverse selection of acceleration-optimized servers with an extensive portfolio of accelerators featuring NVIDIA GPUs. As shown in Figure 3, we highlight **Dell PowerEdge R760xa** servers, which are purpose-built for generative AI for the primary workloads, plus Dell PowerEdge R660 servers for the control plane nodes.

PowerEdge R760xa servers support the following GPU accelerator configurations:

- Up to four NVIDIA L40s GPUS per server

- Up to four NVIDIA H100 GPUs per server

The selection and sizing of the proper server and GPU configurations will be done as part of the service engagement for each customer.

Network infrastructure

Organizations can choose from Dell PowerSwitch or NVIDIA Spectrum-networking, with performance up to 800 Gb, based on their specific requirements.

Dell PowerSwitch S5232F-ON or PowerSwitch S5248F-ON switches can be used as the network switches. PowerSwitch S5232F-ON supports both 25 Gb and 100 Gb Ethernet, while PowerSwitch S5248F-ON is a 25 Gb Ethernet switch. ConnectX-6 Network adapters are used for network connectivity, available in both 25 Gb and 100 Gb options.

NVIDIA Spectrum SN5600, part of NVIDIA Spectrum-X Networking Platform can also be used as the network switches. Spectrum SN5600 supports various connection speeds from 200 Gb to 800 Gb Ethernet. NVIDIA B3220 BlueField-3 SuperNIC is used as the network adapter. Connectivity is available from 10 Gb to 200 Gb. The NVIDIA Spectrum SN2201 is a 1 Gb Ethernet switch that can be used for the OOB management network.

Storage infrastructure

Local storage available in PowerEdge servers is used for operating system and container storage. Kubernetes deploys the Rancher local path Storage Class that makes local storage available for provisioning pods.

The need for external storage for the digital assistant solution depends on the specific requirements and characteristics of the AI model and the deployment environment. Often, external file storage is required for storing the graphic elements used to render the digital assistant itself. LLMs, on the other hand, reside in GPU memory after having been loaded from either local storage, external file, or block storage.

Dell PowerScale

Dell PowerScale offers massive AI performance with the ultimate density. It accelerates all phases of the AI pipeline, from model training to inferencing and fine-tuning. Boasting up to 24 NVMe SSD drives per node and 300 PBs of storage per cluster, it ensures GPU utilization for large-scale model training and drives faster time to AI insights with up to 127 percent improved throughput[1].

PowerScale is no stranger to AI-optimized infrastructure, being one of the first storage products to offer GPUDirect support with NVIDIA, low latency storage access with Network File System over Remote Direct Memory Access (NFSoRDMA), multitenant capabilities, simultaneous multiprotocol support, and 6x9s availability and resiliency to ensure uninterrupted uptime.

Building on that foundation and using continuous software and hardware innovation, our next-generation all-flash systems form a key component of Dell's AI-ready data platform. They offer multicloud agility with our Dell APEX File Storage for public cloud portfolio, federal-grade security features to safeguard the AI process from attacks such as data

[1] **Disclosure**: Based on internal testing, comparing streaming write of F910 on OneFS 9.8 to streaming write of F900 on OneFS 9.5. Results might vary. April 2024

poisoning and model inversion, and exceptional efficiency with the world's most efficient scale-out NAS[2] to manage the AI data growth while controlling storage costs. PowerScale is the first ethernet-based NVIDIA DGX SuperPOD-certified storage platform and is enabling seamless AI Adoption for enterprise customers everywhere.

Validation and performance data

Dell Technologies, along with Pryon and UneeQ have performed extensive validation and performance characterization of the solution, at both the subsystem level and in end-to-end implementations.

While this information is not published in this design guide, such information is available or can be used as part of an implementation engagement.

Deployment scenarios

There are different options for implementing the digital assistant solution. While a digital assistant solution by nature may imply the use of a human-like visual interface, in practice there are valid reasons to consider different implementations in some scenarios.

The two different options for deploying a digital assistant solution include:

- **"Full" digital assistant**—As the name suggests, the full solution allows the implementation of a digital assistant in support of its three key modalities: avatar, voice, and text chat. Each end user can turn on or turn off those options, as well as closed captioning. The full option uses all four core components of the solution:
  - Digital Human Platform
  - Orchestration and Conversation Management
  - Information Retrieval
  - Large Language Model
- **"Headless" digital assistant**—Although not necessarily intuitive, there are cases when a digital assistant solution may be deployed without the 2D/3D human-like interface and the Digital Human Platform.

  This scenario might be required either as a primary or an on-demand approach, with a limited set of modalities, notably as a text chatbot with the voice option. For example, the compute and storage in the environment may be capacity-constrained (for example, under heavier use than planned), or it has been dynamically determined that the bandwidth of the network through which an end user connects to the solution is insufficient for the full experience.

  Under these circumstances, the enterprise providing the solution might not offer the digital assistant avatar frontend, and offer a reduced, traditional chatbot and voice

[2] **Disclosure**: Based on Dell analysis comparing efficiency-related features—data reduction, storage capacity, data protection, hardware, space, life cycle management efficiency, and ENERGY STAR certified configurations, June 2023.

experience without compromising the quality of the responses. Yet the same entire knowledge base is still available.

The choice of which approach depends on the intended use cases of the organization.

Dell Precision AI-Ready workstations

There are certain scenarios in which deploying a digital assistant on a Precision AI-ready workstation, powered by NVIDIA RTX Ada generation GPUs, is appropriate and offers significant benefits. While not part of this design, it is important for customers to be aware of the full range of deployment options that are available.

Dell Precision AI-ready workstations can be customized and configured to meet generative AI workload requirements. The breadth of the portfolio now supports up to four NVIDIA RTX 6000 Ada generation GPUs in a single fixed workstation. This scalability ensures that the workstation remains a viable solution for AI development and can accommodate evolving workloads in preparation for data centers and the cloud. This scalability fosters seamless innovation and experimentation without limiting production resources.

The use case becomes the primary decision factor when deploying a digital assistant, whether at the edge or the data center. For instance, if low latency, enhanced privacy and security, or offline functionality are wanted, consider deploying on a Precision workstation platform in either mobile or tower form factors. Furthermore, a Precision workstation can serve as an excellent sandbox for testing your digital assistant experience prototypes before full-scale data center deployment.

Deploying a digital assistant on a Precision workstation might offer superior performance, privacy, and control, while data center deployment can provide scalability, redundancy, and centralized management. Organizations must evaluate these factors based on their specific use cases and requirements.

Chapter 5 Chapter 5 Professional Services and Consulting

Dell Professional Services for Generative AI harness the power of this rapidly evolving technology in a meaningful and secure way to drive the outcomes that your business expects. By partnering with Dell Technologies, your business can confidently advance generative AI initiatives. You can rely on us every step of the way, with services for strategy, implementation, adoption, and scaling generative AI solutions across your organization.

This Dell Validated Design for Generative AI Digital Assistants in the Enterprise is intended to be deployed as part of a Dell Professional Services engagement, to ensure the best application to the customer's use cases as well as the proper infrastructure design for the right sizing and performance. A brief description of this and other services for generative AI is shown below. For more information, see Dell Professional Services for Generative AI.

Professional Services

The following services and more are available from Dell Technologies.

Digital Assistant Services

Bridging the gap between low interactivity chatbots and human interfaces, create a digital assistant to allow for broader conversations encompassing search capabilities and other system integrations. Dell experts will:

- Facilitate custom digital assistant design and development
- Develop knowledge fabric definitions, content ingestion, and fine-tuning
- Design, develop, and integrate additional custom digital assistant skills such as opening a help desk request on the user's behalf, or check an order status on the customer's behalf, and so on

Advisory Services

Create a strategy and road map to achieve your generative AI vision:

- Dell experts assess your current environment, including generative AI business drivers, goals, challenges, and constraints.
- Prioritizing your business use cases, our professionals define your ideal future state, design a new generative AI solution architecture using this design, and identify the skills your IT organization needs to succeed.
- We then create and validate a generative AI road map for your business to realize the value of generative AI, define your qualitative business rationale, develop recommendations and next steps, and present the road map to your executives.

Implementation Services

Establish your generative AI inferencing platform using the Dell Validated Design for Generative AI Inferencing with NVIDIA:

- Dell Technologies holds a workshop with your project stakeholders, reviewing the approach required to implement platform architecture using Dell Validated Design for Generative AI with NVIDIA
- We then deploy the necessary tools and frameworks to establish an operational generative AI platform, guided by this validated design.
- Before handing over operations to your team, we conduct generative AI platform knowledge transfer that is focused on deployed solutions to enable your team's generative AI success.
- Prepare your data for Generative AI inferencing and model customization with data preparation consulting services.

Adoption Services

Achieve your unique inferencing use cases using a pretrained model with the deployed generative AI platform using the Dell Validated Design for Generative AI with NVIDIA:

- Through multiple workshops, Dell professionals align with project stakeholders to review your use cases and determine the best model to meet your needs.
- With your unique use cases in mind, Dell generative AI experts deploy and configure the pretrained model for your business.
- We then conduct knowledge transfer sessions covering software stack use, architecture, and best practices for adoption of your new inferencing model.

Scaling Services

Optimize processes and advance a generative AI mindset throughout your organization:

- Address key IT skills gaps with Education Services for Generative AI; we work directly with your team and ensure that you are up to speed.
- Simplify your generative AI operations with Managed Services for Generative AI, with Dell managing your NVIDIA solutions stack so that you can focus on your generative AI model development and use cases.
- Dell experts provide the expertise that is needed to drive your generative AI initiative and keep your generative AI infrastructure using the Dell Validated Design for Generative AI with NVIDIA running at its peak.

Customer Solution Centers (CSC)

Our global Dell Technologies Customer Solution Centers are trusted environments where world-class IT experts help customers to design, validate, and build innovative solutions across technology domains including AI, modern data center, multicloud, edge, and the modern workplace. Through engagements focused on real-world use cases and outcomes, we enable customers to see solutions in action and empower them to advance their innovation priorities.

Wherever you are on your journey, we can help you move from ideas to implementation, including proofs of concept (POCs). Though not part of Dell professional services, CSC services are available to Dell Technologies customers at no charge, and you can engage

with our experts in person or remotely from any location. Contact your account team today to submit an engagement request.

Security considerations

Deploy and use the hardware and software recommended in this solution securely. Follow recommendations from Dell Technologies and the other vendors that are cited in this validated design.

For additional information, go to the Dell Technologies Security and Trust Center.

Dell Engineering performed a threat analysis using the Microsoft Threat Modeling Tool, which revealed several factors to consider when deploying this solution in your lab.

Our threat modeling findings are grouped into two categories: external threats and inherent threats.

External threats

The number of attacks and the potential consequences of any external threats must receive priority. Our modeling revealed three distinct types of external threats: unauthorized acquisition of user credentials, subversion of the authentication mechanism, and denial of service (DoS):

Threat actors steal credentials—The unauthorized acquisition of user credentials includes usernames, passwords, or two-factor authentication (2FA) codes transmitted over insecure channels. Defaults are not changed, or secrets are predictable. When stolen, credentials can be exploited for impersonation, privilege escalation, or other malicious activities.

Threat actors exploit insecure authentication and logging mechanisms—Attackers can compromise authentication mechanisms using various techniques, including credential stuffing, subverting 2FA system and common threat vectors such as default credentials, predictable secrets, password recovery, or change vulnerabilities, online password cracking, and offline password cracking.

Threat actors perform a Denial-of-Service attack—Attackers aim to disrupt an application or system's normal functioning by overwhelming it in ways that render it unavailable to legitimate users. Attackers use different techniques to exhaust system resources such as bandwidth, memory, or CPU or exploit vulnerabilities in the system to degrade performance.

Inherent threats

Inherent threats in software application and system design arise from design choices, architecture decisions, and the integration of various components. These threats can leave applications vulnerable to attacks or failures:

Insecure Authentication Mechanisms—Eliminating vulnerabilities in the authentication and session management processes is critical for maintaining a secure application environment. Weaknesses in these areas can result in unauthorized access, privilege escalation, and compromise of user accounts. Attackers often exploit these flaws to bypass authentication, impersonate legitimate users, and gain unauthorized access to sensitive data or functionality.

Insecure Logging—Audit logs capture the security-relevant sequence of events for any operation. When a system has insecure logging, an attacker can tamper with the integrity of this data by various methods to mislead or cover other elevated attacks. An attacker can exploit software vulnerabilities to perform log injection, log forging, and log poisoning. An attacker can also cause a denial of service to the target system or downstream consumers of logs through vulnerabilities in log parsers.

Chapter 6 Conclusion

Summary

The Dell Validated Design for Generative AI Digital Assistants has been developed to address the needs of enterprises that want to develop and run AI-powered digital assistants for using domain-specific data that is relevant to their own organization.

Dell Technologies, along with its partners, have designed a scalable, modular, and high-performance architecture. It enables enterprises to design and deploy an inferencing solution that has been validated and performance-tested to accelerate the time to value and to reduce the risk and uncertainty by using a proven design quickly.

This guide provides design guidance and a fully validated reference architecture for digital assistants in an on-premises environment, which is designed to be customizable and scalable with a Dell Professional Services engagement.

With this validated design, Dell Technologies enables organizations to deliver full-stack generative AI digital assistant solutions built on the best of Dell infrastructure and software, combined with the latest NVIDIA accelerators and software, and the advanced capabilities of UneeQ and Pryon. This combination of components enables enterprises to use purpose-built generative AI on-premises to solve their business challenges. Together, we are leading the way in driving the next wave of innovation in the enterprise AI landscape.

We value your feedback

Dell Technologies and the authors of this document welcome your feedback on the solution and the solution documentation. Contact the Dell Technologies Solutions team by email.

Chapter 7 References

The following links provide additional information about the solution design and components in this guide.

Dell Technologies documentation

The following Dell Technologies documentation provides additional and relevant information:

- Dell Generative AI Solutions
- Dell Info Hub for AI
- White Paper – Generative AI in the Enterprise
- Design Guide – Generative AI in the Enterprise – Inferencing
- Technical White Paper – Generative AI in the Enterprise – Model Customization
- Dell Professional Services for Generative AI
- Dell Technologies Security and Trust Center

Partner documentation

The following documentation from partners in this design provides additional and relevant information about components in the solution:

- NVIDIA AI Enterprise documentation
- UneeQ Digital Humans
- Pryon Retrieval Engine
- Meta Llama 3
- Meta's Responsible Use Guide