A Principled Technologies report: Hands-on testing. Real-world results.

Dell APEX Cloud Platform for Red Hat OpenShift: An easily deployable and powerful solution to jumpstart your next AI innovation

The 4th Generation Intel Xeon Scalable processor-powered solution deployed in less than two hours and ran a generative AI workload effectively

Deploy a comprehensive AI solution quickly

Less than 2 hours to your Generative AI solution

Generative AI (GenAI) offers myriad benefits, potentially limited only by creativity. The rise of large language models (LLMs), a type of GenAI, has accelerated the desire to explore those bounds in many organizations, but implementation of a solution, in addition to the training and fine-tuning of models, may seem daunting. In light of these concerns, organizations can now bring GenAI and LLMs into their data centers and colocations efficiently and succinctly with the Dell APEX Cloud Platform for Red Hat® OpenShift®, powered by 4th Generation Intel® Xeon® Scalable processors.

How do we know? We followed a combination of official Red Hat OpenShift AI documentation and Dell Validated Design for Red Hat OpenShift AI on APEX Cloud Platform and easily deployed the necessary cloud infrastructure to run a Kubernetes containers-based LLM. In addition to being easy to deploy, the Intel Xeon Scalable processor-powered solution handled the running of the pre-trained LLM, Large Language Model Meta AI (Llama) 2, delivering output quickly.

What do organizations need to run generative AI and LLMs effectively?

People may perceive generative AI, and specifically LLMs, as complex. That is true to some degree, but the Dell APEX Cloud Platform, powered by 4th Generation Intel Xeon Scalable processors, with Red Hat OpenShift AI, can help you bring AI to your data center or colocation so you can begin harnessing its power. Here, we offer an introduction to GenAI, LLMs, and additional information to help you get started.

Overview of generative AI

The popularity of AI reached new heights in 2023, as Yahoo! Finance labeled it "2023's biggest news story."[1] Generative AI was a large driver of that: More than 250 million users tinkered with AI tools,[2] such as generative AI LLMs ChatGPT and others. However, businesses were a major force in pushing the popularity too—McKinsey estimates that "generative AI could add the equivalent of $2.6 trillion to $4.4 trillion annually" to the global economy.[3]

What is it?

Generative AI uses patterns in existing text, audio, video, and other data sets to generate new content in the same forms (though in the case of raw data, the output is synthetic).[4,5] McKinsey goes on to note that "[a]bout 75 percent of the value that generative AI use cases could deliver falls across four areas: Customer operations, marketing and sales, software engineering, and R&D."[6] Specifically, companies in the industries of banking, high tech, and life sciences could see the biggest impact.[7] Generative AI could help design new drugs, develop products, and make supply chains more efficient.[8]

One of the perhaps more recognizable applications of generative AI is customer operations, where it could improve customer experience in a variety of ways. When engaging with customers, for example, generative AI can offer personalized recommendations. For ecommerce operations, that could mean "images and text for personalized product recommendations based on a customer's browsing history and previous purchases."[9] Automating customer service with generative AI chatbots or targeted support could also improve the customer experience due to shorter wait times, more accuracy when answering questions, or proactively addressing customer concerns.[10]

Key terms or elements of AI

The following terms provide deeper context and understanding of our research and testing, which focused on the performance of LLMs; we present those results later in this report.

1. **Generative models** – These algorithms, or networks, can create new synthetic data that looks like a dataset. Examples include autoregressive models, generative adversarial networks (GANs), and variational autoencoders (VAEs).[11]
2. **Autoregressive models** – These models, such as LLMs, "generate data one element at a time, conditioning the generation of each element on previously generated elements."[12] Generative pre-trained transformer (GPT) models, which generate coherent and appropriate text, are popular examples of autoregressive models.[13]
3. **Neural networks** – These building blocks of generative AI take their name from the connections in the human brain. Neural networks use data to "learn patterns and relationships within" the data, which allows the networks to produce clear and understandable output.[14]

4. **LLMs** – LLMs take user-written text and generate tokens based on the model's understanding of the relationships between words and phrases. The models use encoders and decoders with self-attention capabilities to understand basic grammar, languages, and knowledge through self-learning. The encoders and decoders reside in a set of neural networks in a transformer architecture. The flexibility of LLMs enables them to "perform completely different tasks such as answering questions, summarizing documents, translating languages and completing sentences."[15]

What you'll need to run generative AI on-premises effectively

If having an on-premises generative AI solution sounds appealing, consider the basics of what you'll need to run it—the Dell APEX Cloud Platform, powered by 4th Generation Intel Xeon Scalable processors, with Red Hat OpenShift AI provides these basics.

First, you'll need powerful CPUs. Specifically, the CPUs supporting your GenAI solution need to be able to dispatch code to GPUs and handle the necessary I/O and other data processing functions. Processing power is essential for LLMs, and 4th Generation Intel Xeon Scalable processors could serve as an excellent choice. Additionally, the architecture of those Intel Xeon chips includes Intel Advanced Matrix Extensions (AMX), accelerators that "can help you make the most of your CPU to power AI training and inferencing workloads at scale for benefits including improved efficiency, reduced inferencing, training, and deployment costs, and lower total cost of ownership (TCO)."[16]

For many deployments, you'll also want powerful GPUs, which will likely perform much of the machine learning and deep learning training and inferencing. The highly parallel structure of GPUs enables them to perform thousands of operations simultaneously, making them well-suited to handle the tasks required by neural networks.[17]

In addition to powerful CPUs and GPUs, you'll need to support generative AI workloads and LLMs with ample storage to back your large data sets, networking that ensures high-speed data transfers, plus features and architecture that can offer scalability, reliability, security, and more—all things that the hardware of the Dell APEX Cloud Platform offers.

About Llama 2

We used the open source LLM Llama 2 for our testing. The second version of Llama in the Meta family of LLMs, Llama 2 works like other LLMs in that it takes a sequence of input words and generates text by predicting the next word or words. Meta pretrained Llama 2 "on publicly available online data sources."[19] Meta also notes that "Llama 2 models are trained on 2 trillion tokens and have double the context length of Llama 1. Llama Chat models have additionally been trained on over 1 million new human annotations."[20]

What are tokens?

LLMs read or generate chunks of text called tokens. Tokens typically are not words, but they can be. Tokens can be smaller, such as characters or parts of a word, or larger, such as a phrase.[21] Prompt tokens are the words that users send to the LLM, and completion tokens are the words users receive from the LLM.[22] OpenAI notes that "the exact tokenization process varies between models. Newer models like GPT-3.5 and GPT-4 use a different tokenizer than previous models, and will produce different tokens for the same input text."[23]

Retrieval Augmentation Generation (RAG)

According to Amazon, "Retrieval-Augmented Generation (RAG) is the process of optimizing the output of a large language model, so it references an authoritative knowledge base outside of its training data sources before generating a response."[24] RAG enhances the already powerful capabilities of LLMs or an organization's internal knowledge base without requiring the model to undergo retraining. The process introduces an information retrieval component that gathers information from a new data source based on user input. The LLM then receives the user query and the relevant information. The LLM can create better responses by using the new knowledge in addition to its training data.[25]

Vector databases

According to Microsoft, a vector database "is a type of database that stores data as high-dimensional vectors, which are mathematical representations of features or attributes."[26] In the context of LLMs, vector databases store the vector embeddings (the numerical arrays that represent characteristics of an object) that the model's training produces. This enables the database to execute similarity searches, identifying matches between a user's prompt and a specific vector embedding.[27]

Use cases of LLMs

LLMs offer powerful understanding and generation of human-like text, which many organizations could consider valuable for multiple practical applications. LLMs can automate tasks that traditionally require human involvement, which could mean increased efficiency and a possible reduction in operational costs. In addition, many LLMs serve as benchmarks for linguistic AI capabilities, such as Large Language Model Meta AI (Llama) 2, which we used, and can foster the development of improved models, which generally helps promote AI and machine learning research. Other common LLMs include GPT-3.5, GPT-4, Claude, Gemini, and others.[18]

Although relatively nascent, LLMs provide organizations with ample ways to improve their operations. Some of the broader use cases include the following.

Content creation

LLMs can empower writers, marketers, and other creatives by generating initial drafts, suggesting edits, or producing complete articles, reports, and creative works. Using LLMs in these processes could accelerate content creation and allow humans to focus on strategic and creative aspects rather than the mechanical elements of writing.

Customer support

LLMs can provide automated and personalized responses to questions, enabling businesses to offer 24/7 support without extensive human resources while potentially improving customer satisfaction and operational efficiency.

Customer feedback analysis

To better respond to customer concerns and needs, organizations can train LLMs to understand textual and voice sentiment. In this way, LLMs can analyze customer feedback, reviews, and social media mentions at scale to help gain insight into public perception and emerging trends.

Learning and training

LLMs can provide tutoring, generate practice questions, and offer explanations tailored to what students are learning. As such, LLMs can deliver personalized education and training to meet the individual needs of learners. The LLM could help create textbooks and interactive online courses.

Translation

LLMs can translate text, making websites, applications, and digital content more accessible. Additionally, LLMs can learn cultural language nuances to provide localized content that's translated and contextually appropriate, helping organizations remove language barriers and reach a larger, perhaps global, audience.

Dell APEX Cloud Platform with Red Hat OpenShift AI

Overview

Dell APEX Cloud Platform with Red Hat OpenShift AI combines the Dell on-premises cloud hardware with Red Hat container and bare metal orchestration software. The turnkey solution can provide "deep integration and intelligent automation between layers of Dell and OpenShift technology stacks, accelerating time-to-value and eliminating the complexity of management using different tools in disparate portals."[28]

To learn more about the solution, visit https://www.delltechnologies.com/asset/en-us/solutions/apex/briefs-summaries/apex-cloud-platform-for-red-hat-openshift-workload-solution-brief.pdf.

Dell APEX benefits

According to Dell, "APEX is a portfolio of cloud services that is based on a cloud consumption model that delivers IT as-a-service, with no upfront costs and pay-as-you-go subscriptions."[29] The solution measures consumption by using automated tools. Customers pay only for what they use, at a single billing rate that can help accurately predict future costs.[30] The solution we tested reflected what we believe a customer might see: a turnkey offering that includes servers pre-installed with Red Hat OpenShift, pre-configured networking, and other resources available for out-of-the-box use.

Our Dell APEX Cloud Solution used the following hardware:

- 4x Dell APEX MC-760 nodes
- 2x 4th Gen Intel Xeon Scalable processors
- NVIDIA A2 (Note: Though our testing used the A2 GPUs, the server also supports A40 GPUs. The upcoming Apex Cloud Platform release will support the NVIDIA L40S.)

Dell and Red Hat OpenShift benefits

The Dell APEX Cloud Platform for Red Hat OpenShift offers many advantages that can help organizations bring LLMs and other AI services into the data center. Administrators can now manage the infrastructure in the OpenShift Web Console, allowing them to update hardware with the same workflow as updating OpenShift software. One pane of glass for infrastructure management can reduce or even eliminate separate OEM tools, potentially reducing operational costs.

Because Dell ProSupport deploys OpenShift after racking and powering the nodes and setting up their networking, admins could save time by avoiding manual deployment and potentially accelerating the platform and its supported services' times to value. The platform offers advanced DevOps automation tools to help developers focus more on building applications than managing infrastructure. It also supports both container and virtual machine workloads, with configuration options to support GPUs. Overall, Dell APEX Cloud Platform for Red Hat OpenShift delivers a consistent, automated experience for deploying and managing OpenShift across on-premises, public cloud, and hybrid environments.

Red Hat OpenShift and Kubernetes roles in this generative AI solution

Red Hat OpenShift AI is built on the open-source, Kubernetes-based OpenShift Container Platform. The solution we tested was able to leverage the Operators framework, which "automate[s] the creation, configuration, and management of instances of Kubernetes-native applications," according to Red Hat.[31] This framework allowed us to quickly deploy and expand AI and machine learning models to the testbed.

Dell ecosystem

Dell provides MC-760 server nodes for this turnkey solution, but it's also important to note the larger Dell ecosystem surrounding the APEX platform and how Dell seeks to incorporate AI technology throughout its offerings. In two other reports published in 2024, we found that other PowerEdge servers from the Dell portfolio delivered better AI training and inference performance than competing HPE and Supermicro servers.[32, 33]

Dell also offers robust manageability tools that could affect your decision to deploy this solution. Other recent PT studies have shown advantages in remote connectivity, operating system deployment, alert configuration, and firmware updates. As just one example, a 2022 Principled Technologies study showed that:[34]

- iDRAC9, embedded in all new Dell PowerEdge servers, offered 2.5 times the number of HTML5 console features as HPE iLO
- Using Dell OpenManage Enterprise (OME) cut deployment times in half compared to a competitive on-premise solution
- Dell OME enabled easy one-to-many updates
- Dell OME allowed for flexibly alert policy configuration

Using the Dell APEX Cloud Platform for Red Hat OpenShift solution for other natural language processing workloads

The Dell and Red Hat partnership for AI extends beyond the solution we tested. Customers interested in natural language processing (NLP) could choose a Dell APEX Cloud Platform for Red Hat OpenShift solution that runs NVIDIA Riva, which is part of the NVIDIA AI Enterprise platform provides speech models and services. Like the solution we tested, the infrastructure of this solution comprises Dell servers, Red Hat OpenShift, and a management layer integrated into the Red Hat OpenShift console.

This Dell and Red Hat solution offers automated speech recognition (ASR) and text-to-speech (TTS) capabilities because NLP relies on speech recognition and synthesis. Using an NLP model, organizations could develop conversational applications to help sectors such as retail, manufacturing, IT, banking, telecom, and healthcare.

To learn more about this Dell APEX Cloud Platform for Red Hat OpenShift solution for NLP and how to deploy it, visit https://infohub.delltechnologies.com/en-US/t/ai-driven-speech-recognition-and-synthesis-on-dell-apex-cloud-platform-for-red-hat-openshift/.

Considerations of on-site vs. cloud or external API-based LLMs

The case for LLMs in everyday business use is already evident to decision makers. An August 2023 survey by the Cutter Consortium showed that over one-third of organizations indicate they will be deploying LLMs into their applications.[35] A September 2023 Morning Consult study for Dell noted that 76 percent of IT decision makers think GenAI will be "significant if not transformative for their organizations."[36] However, that excitement is tempered by hesitation for 37 percent of respondents.[37] Security, complexity, ethical use of data, and data governance are the top concerns adding to this hesitation. Organizations may be able to mitigate these concerns by using an on-premises solution such as Dell APEX with Red Hat OpenShift AI. In the following sections, we discuss these issues and why you may wish to choose to run your LLM on-premises rather than in the public cloud.

A business may decide to deploy GenAI on-premises because maintaining full control over the solution can help meet the requirements of regulatory agencies or service level agreements (SLAs) around uptime, responsiveness, security, and data location.

Requirements like these were on the minds of IT decision makers that Dell surveyed in August 2023.[38] The survey focused on decision makers who were building or training GenAI models on-premises and asked them why they had chosen on-premises solutions. Fifty percent named cost as their primary consideration. Several of the other top reasons were related to data, including:

- Data governance (30 percent), meaning the maintenance of data to ensure it is current, accurate, and relevant
- Intellectual property (IP) considerations (25 percent), highlighting the value of an organization's data and the importance of keeping that data secure
- Data location (22 percent)[39], of particular concern if governmental regulations or customer SLAs restrict data to certain geographies or require it to stay on-premises.

Survey respondents also described deploying GenAI on-premises for reasons such as performance (55 percent), speed from data to intelligence (30 percent), more control over AI model (30 percent), and legal compliance (21 percent).[40]

A common theme of the survey's questions is maintaining control. Companies chose on-premises GenAI solutions so they could control the AI model, the data they used in the model, the hardware and infrastructure on which it ran, and solution costs. In the following sections, we discuss some of these considerations, specifically around security and compliance, uptime and SLA requirements, data adjacency and latency, and cost.

Security and compliance

Organizations face a multitude of complexities when transitioning to heavy use of generative AI, but perhaps none are as challenging and complex as security. How does an organization protect its confidential data and intellectual property while utilizing AI, if AI tools require employees to input that data? Where do the AI systems transmit that data, and what other organizations may have indirect access to it? For training data, the generative AI solution itself must feed on the organization's content and the conversational responses to improve its functionality. So, when it ingests content, for security purposes it must be able to know exactly who input the content and their relationship to all other users—an incredibly difficult task for a publicly accessible generative AI solution, even a very advanced one. For example, if a model's training used a finance team's data or chat conversations, an organization likely would not want another team's users to gain access to that data by asking the AI solution questions.

In the OpenAI terms of use, we see a few examples of these challenges, as well as other significant issues that you must consider when making the choice between on-premises LLMs versus public ones. Their "Opt Out" clause states: "If you do not want us to use your Content to train our models, you can opt out by following the instructions in this Help Center article. Please note that in some cases this may limit the ability of our Services to better address your specific use case."[41] Forcing the user or organization to take extra steps to opt out means that by default, OpenAI training and content ingest policies are opt-in: Any data a user inputs as a prompt to an OpenAI tool is automatically available for training for that tool. Additionally, in its privacy policy, OpenAI states: "When you use our Services, we collect Personal Information that is included in the input, file uploads, or feedback that you provide to our Services ("Content")".[42]

Cyberhaven, a company that focuses on data security and data loss products, performed a survey in 2023 that showed several findings relevant to this discussion. For this survey, they analyzed OpenAI ChatGPT usage for 1.6 million workers using the Cyberhaven product.[43] Consider these findings from that data analysis:

- 10.8 percent of employees tried using ChatGPT in the workplace
- 8.6 percent of employees pasted data into ChatGPT
- 4.7 percent of employees pasted sensitive data into ChatGPT at least once; the most common incidents included:
  - Internal-only data (319 incidents per 100K employees)
  - Source code (278 incidents per 100K employees)
  - Client data (260 incidents per 100K employees)

With a ChatGPT user base of 100 million weekly users, if just a small fraction of the users are bad actors, this is still a staggering scale of data leakage and privacy issues.

The examples and numbers are not merely academic, however. Well-known public companies have fallen victim to data leaks via ChatGPT, even in its short history. In early 2023, Samsung employees leaked confidential source code and data to ChatGPT.[44] Per one article about the incident, "One Samsung employee entered faulty source code related to the Samsung Electronics facility measurement database download program to find a solution. Another employee entered program code for identifying defective equipment to get code optimization. The third employee converted a smartphone recording of a company meeting to a document file and entered it into ChatGPT to get meeting minutes, according to the report."[45]

In addition to security concerns, there is the issue of compliance. Governmental regulations abound, and they differ by geography. In the United States, every company legally must follow a host of federal and state regulations, which might include the Health Insurance Portability and Accountability Act (HIPAA), the Graham-Leach-Bliley act covering financial data, the Children's Online Privacy Protection Act (COPPA), and the California Consumer Protection Act, among many others.[46,47] In Europe in 2016, the European Union adopted the General Data Protection Regulation (GDPR).[48] Violations are costly, with recent headlines including Meta being charged 1.2 billion Euros just last year.[49]

Considering all of these factors, companies are understandably concerned about protecting the data that their GenAI system may use.

Uptime requirements, SLA, and hardware control

When systems and applications run on the cloud, companies face the risk of failure or downtime. Cloud service outages are not uncommon, and those outages can be extraordinarily costly, with failures in the heavier-used regions being the most impactful and therefore expensive. A 2023 report by Parametrix Insurance estimated that a 24-hour failure in AWS us-east-1, on which over a third of Fortune 500 companies depend in some way, could cost $3.4 billion dollars.[50] Recent examples of cloud service failures just in 2023 include:

- In November 2023, the OpenAI ChatGPT service and APIs were down for more than 90 minutes.[51]
- In January 2023, Microsoft 365 in North America faced an outage of five hours.[52]
- In April 2023, Google Cloud saw outages due to a fire and subsequent water damage in its europe-west9 region.[53]
- In June 2023, AWS Lambda experienced a failure that affected multiple other AWS services. This issue lasted approximately four hours.[54]

This does not mean that on-premises systems are never at risk. But with on-premises systems, the control of the system and its uptime lies with the organization.

Data latency and adjacency

The physical distance between your data and the resource requesting it (e.g., a server or a cloud instance) is important, because the greater the geographical distance or networking complexity between the data and the requesting application, the greater the latencies could be. Naturally, on-premises solutions have an advantage in this area. With on-prem solutions, there is no need for a user prompt and its associated metadata to traverse the internet, be ingested into a potentially large number of API layers, then sent back to the user. The user prompt stays onsite and the on-premises GenAI system delivers the response to the user.

The route that data packets travel can also affect application response times. This routing may change depending on the multiple devices, data centers, and service providers it encounters along its journey, and problems at any of those layers can cascade to others, affecting the user experience. On-premises solutions can have an advantage in these areas due to their proximity and direct control over the networking infrastructure. An organization can dedicate resources to optimizing both software and hardware from networking and latency perspectives in the GenAI application stack.

PromptHub, a site that allows for AI prompt design and testing, tested several different providers and models in October 2023. Specifically, they used token output to benchmark average response times for OpenAI, Azure, and Anthropic models.[55] They found that while response times varied by provider, average response time per output token was approximately 30ms, with the highest latency being 76ms per token. With a 500-word output response (about 660 tokens), this could be 50 seconds. Trimming that response time for users is critical, and if having the GenAI system on-premises helps do so, it could benefit users.

As the PromptHub article points out, different models and providers will deliver varying service levels. We've observed that maintenance, upgrades, or newly released models can change those times. When OpenAI released GPT4 in the spring of 2023, anecdotal data and posts emerged regarding significant increases in GPT4 API response time versus token count.[56, 57] Of course, on-premises solutions could also potentially face changes in latency due to model maintenance or upgrades, but given that an organization would have complete control over those changes and implementation, they could test the issues well ahead of time and take a flexible approach.

In addition to latency considerations, data adjacency could play a significant role in training times. Businesses with extremely large training data sets could choose to train their own LLM on domain-specific knowledge. A SiliconANGLE article noted: "Model training and tuning aren't like conventional IT tasks. Machine learning models must repeatedly churn through large amounts of data – the more, the better. Petabyte-sized data volumes aren't easily shifted between local data centers and cloud platforms. This 'data gravity' issue makes training in-house an attractive option if that's where the data already lives."[58] Bloomberg used this approach when, in 2023, they announced their own BloombergGPT model, using a 700-billion token data set for training.

Cost models of on-prem vs cloud or external API-based solutions

In addition to security and control concerns, leadership of any organization will likely consider costs, both capital and ongoing. An organization could take one of two approaches when deploying an on-premises solution, such as the one we deployed.

The first, covered in depth in another PT report, is to build a similar hardware stack on a cloud provider using cloud services, such as AWS Sagemaker. As we note in that report, depending on the configuration and usage, that approach could be quite costly compared to an on-premises solution, but it does offer some level of control. [Here we will link to the aforementioned PT report once it is published.]

The second approach could be more cost-effective, especially for smaller AI workloads, but comes fraught with the risks outlined above. This approach uses API-based LLM products that generally offer a pricing model based on token usage, with perhaps some different pricing for prompt tokens versus completion tokens. Typically, the LLM vendors express this pricing in $/1K tokens. For example, OpenAI GPT-4 is $0.03/1K tokens for input (prompt) and $0.06/1K tokens for output (completion). Pricing differs based on the LLM that the application requires. For example, our pricing example doubles if the user selects the GPT-4-32K model. Usage would naturally vary, which would likely make costs very hard to predict. And depending on your internal development staff, you may incur costs hiring towards that solution.

This differs from the pricing model of an on-premises solution, which organizations could approach as either a one-time capital expense with annual support costs or pay-as-you-go, where the vendor handles the billing on a per-month basis, perhaps on a 3-to-5-year contract.

Setup of our on-site generative AI LLM infrastructure

Following the Red Hat guide for installing OpenShift AI, we took a base configured Dell APEX cluster with OpenShift and added our own user logins. (This guide is available at https://access.redhat.com/documentation/en-us/red_hat_openshift_ai_self-managed/2.8/html/installing_and_uninstalling_openshift_ai_self-managed/index.) From there, we used the official Red Hat OpenShift documentation to install the prerequisites and OpenShift AI itself.

Next, we referenced the Dell Technologies Validated Design Guide, "Implementing a Digital Assistant with Red Hat OpenShift AI on Dell APEX Cloud Platform" for specific recommendations to add a serving model and download and install the Hugging Face version of the Llama 2 13B pre-trained model. (You can find that guide at https://infohub.delltechnologies.com/t/design-guide-implementing-a-digital-assistant-with-red-hat-openshift-ai-on-dell-apex-cloud-platform-1/).

After we completed that installation, we added the model to our OpenShift AI deployment and surfaced it for other applications to use. We then installed Redis to serve as a document index and uploaded 32 documents to the index for our LLM to use for reference. Finally, we installed Gradio to serve as a GUI for end users and enabled it to incorporate the Redis index with the LLM. At that point, users could ask questions of the LLM.

For more details, see the science behind the report.

What we found

In the constantly evolving AI landscape, organizations may set out to build a model following certain specifications. However, through new innovations or shifting needs, they may choose an option that combines a variety of approaches. In our tests, we used the Dell Technologies Validated Design Guide as a reference, implementing several methods to set up our solution. We found that being open to adjustments while using existing Dell and Red Hat documentation allowed us to completely deploy the Dell APEX Cloud Platform for Red Hat OpenShift solution for our LLM use case in less than two hours. Table 1 shows the steps we took to complete the deployment.

Table 1: The steps we completed to deploy our Dell APEX Cloud Platform for Red Hat OpenShift solution and Llama 2 and the time we needed to complete each step. Source: Principled Technologies.

Step	Time (h:mm:ss)
Create an admin login for OpenShift	0:03:37
Install OpenShift AI	0:03:42
Create Data Science Project	0:01:32
Install both KServe and OpenShift Serverless	0:02:12
Download Llama 2 model	0:45:31
Upload Llama 2 model to cloud storage	0:19:49
Create a workspace with Llama 2 runtime	0:14:34
Deploy Redis	0:13:02
Create index and populate it	0:01:20
Create Gradio deployment	0:05:52
Total	**1:51:11**

To demonstrate the functionality of this proof-of-concept LLM, we also tested the solution's performance by having a single user query the LLM and measuring the latency of the solution's response. We used 10 sample queries and observed that the responsiveness of the system was commendable, with the time for the LLM to respond to our queries being less than a second to begin its response.

Conclusion

The appeal of incorporating GenAI into your organization's operations is likely great. Getting started with an efficient solution for your next LLM workload or application can seem daunting because of the changing hardware and software landscape, but Dell APEX Cloud Platform for Red Hat OpenShift powered by 4th Gen Intel Xeon Scalable processors could provide the solution you need. We started with a Dell Validated Design as a reference, and then went on to modify the deployment as necessary for our Llama 2 workload. The Dell APEX Cloud Platform for Red Hat OpenShift solution worked well for our LLM, and by using this deployment guide in conjunction with numerous Dell documents and some flexibility, you could be well on your way to innovating your next GenAI breakthrough.

1. Yahoo Finance and Luke Carberry Mogan, "AI, 2023's biggest news story," accessed January 22, 2024, https://finance.yahoo.com/video/ai-2023s-biggest-news-story-212425358.html.
2. Statista, "Growth of artificial intelligence (AI) tool users worldwide from 2020-2030," accessed January 25, 2024, https://www.statista.com/forecasts/1425996/ai-tool-user-amount#.
3. McKinsey, "The economic potential of generative AI: The next productivity frontier," accessed January 26, 2024, https://www.mckinsey.com/capabilities/mckinsey-digital/our-insights/the-economic-potential-of-generative-ai-the-next-productivity-frontier#introduction.
4. Generative AI, "All Things Generative AI," accessed January 26, 2024, https://generativeai.net/.
5. George Lawton, "What is generative AI? Everything you need to know," accessed January 26, 2024, https://www.techtarget.com/searchenterpriseai/definition/generative-AI.
6. McKinsey, "The economic potential of generative AI: The next productivity frontier."
7. McKinsey, "The economic potential of generative AI: The next productivity frontier."
8. George Lawton, "What is generative AI? Everything you need to know."
9. Pecan, "The Top 6 Use Cases for GenAI," accessed January 27, 2024, https://www.pecan.ai/blog/top-6-genai-use-cases/.
10. Pecan, "The Top 6 Use Cases for GenAI."
11. Olena H., "10 Essential Terms in Generative AI," accessed January 26, 2024, https://www.linkedin.com/pulse/10-essential-terms-generative-ai-olena-h-.
12. Alexis Porter, "Unveiling 6 Types of Generative AI," accessed January 25, 2024, https://bigid.com/blog/unveiling-6-types-of-generative-ai/.
13. Alexis Porter, "Unveiling 6 Types of Generative AI."
14. Alexis Porter, "Unveiling 6 Types of Generative AI."
15. Olena H., "10 Essential Terms in Generative AI."
16. AWS, "What are Large Language Models (LLM)?," accessed January 26, 2024, https://aws.amazon.com/what-is/large-language-model/.
17. Intel, "What Is Intel® Advanced Matrix Extensions (Intel® AMX)?," accessed January 28, 2024, https://www.intel.com/content/www/us/en/products/docs/accelerator-engines/what-is-intel-amx.html.
18. Ben Lutkevich, 16 of the best large language models," accessed January 29, 2024, https://www.techtarget.com/whatis/feature/12-of-the-best-large-language-models.
19. Meta, "Llama 2: open source, free for research and commercial use," accessed January 29, 2024, https://ai.meta.com/resources/models-and-libraries/llama/.
20. Meta, "Llama 2: open source, free for research and commercial use."
21. Michael Humor, "Understanding "tokens" and tokenization in large language models," accessed January 28, 2024, https://medium.com/@michaelhumor/understanding-tokens-and-tokenization-in-large-language-models-1058cd24b944.
22. Logan Kilpatrick, "What is the difference between prompt tokens and completion tokens?," accessed January 29, 2024, https://help.openai.com/en/articles/7127987-what-is-the-difference-between-prompt-tokens-and-completion-tokens.
23. OpenAI, "Tokenizer," accessed January 29, 2024, https://platform.openai.com/tokenizer.
24. AWS, "What Is RAG?," accessed January 28, 2024, https://aws.amazon.com/what-is/retrieval-augmented-generation/.
25. AWS, "What Is RAG?."
26. Microsoft, "What is a vector database?," accessed January 29, 2024, https://learn.microsoft.com/en-us/semantic-kernel/memories/vector-db.
27. V7, "Vector Databases: Intro, Use Cases, Top 5 Vector DBs," accessed January 26, 2024, https://www.v7labs.com/blog/vector-databases.
28. Dell Technologies, "Red Hat OpenShift AI on Dell APEX Cloud Platform for Red Hat OpenShift," accessed April 12, 2024, https://www.delltechnologies.com/asset/en-us/solutions/apex/briefs-summaries/apex-cloud-platform-for-red-hat-openshift-workload-solution-brief.pdf.
29. Dell Technologies, "Design Guide—Implementing a Digital Assistant with Red Hat OpenShift AI on Dell APEX Cloud Platform > Solution concepts," accessed January 15, 2024. https://infohub.delltechnologies.com/l/design-guide-implementing-a-digital-assistant-with-red-hat-openshift-ai-on-dell-apex-cloud-platform-1/solution-concepts-6/.

30. Dell Technologies, "Dell APEX Flex on Demand," accessed April 29, 2024, https://www.dell.com/en-us/dt/payment-solutions/flexible-consumption/flex-on-demand.htm.

31. Red Hat, "What are Red Hat OpenShift Operators?" accessed April 9, 2024, https://www.redhat.com/en/technologies/cloud-computing/openshift/what-are-openshift-operators.

32. Principled Technologies, "Meeting the challenges of AI workloads with the Dell," accessed April 11, 2024, https://www.principledtechnologies.com/Dell/AI-portfolio-vs-HPE-0124.pdf.

33. Principled Technologies, "Finding the path to AI success with the Dell AI portfolio," accessed April 11, 2024, https://www.principledtechnologies.com/Dell/AI-portfolio-vs-Supermicro-0224.pdf.

34. Principled Technologies, "Simplify administrator tasks and improve security and health monitoring with tools from the Dell management portfolio vs. comparable tools from HPE," accessed January 15, 2024, https://www.principledtechnologies.com/Dell/Management-tools-vs-HPE-1122.pdf.

35. Curt Hall, "Enterprises Keen on Adopting Large Language Models, but Issues Exist," accessed January 9, 2024, https://www.cutter.com/article/enterprises-are-keen-adopting-large-language-models-issues-exist.

36. Dell, "Generative AI Pulse Survey," accessed January 9, 2024, https://www.dell.com/en-us/dt/solutions/artificial-intelligence/index.htm#accordion0&tab0=0&pdf-overlay=//www.delltechnologies.com/asset/en-us/solutions/infrastructure-solutions/templates-forms/dell-technologies-genai-pulse-survey.pdf.

37. Dell, "Generative AI Pulse Survey."

38. Dell, "Generative AI has been 2023's most talked about technology. Organizations are optimistic but also face many challenges as they make their AI dreams a reality," accessed January 17, 2024, https://www.delltechnologies.com/asset/en-us/solutions/infrastructure-solutions/briefs-summaries/gen-ai-research-infographic.pdf.

39. Dell, "Generative AI has been 2023's most talked about technology. Organizations are optimistic but also face many challenges as they make their AI dreams a reality."

40. Dell, "Generative AI has been 2023's most talked about technology. Organizations are optimistic but also face many challenges as they make their AI dreams a reality."

41. OpenAI, "Terms of Use," accessed January 10, 2024, https://openai.com/policies/terms-of-use.

42. OpenAI, "Privacy policy," accessed January 10, 2024, https://openai.com/policies/privacy-policy.

43. Cameron Coles, "11% of data employees paste into ChatGPT is confidential," accessed January 10, 2024, https://www.cyberhaven.com/blog/4-2-of-workers-have-pasted-company-data-into-chatgpt.

44. Mack DeGeurin, "Oops: Samsung Employees Leaked Confidential Data to ChatGPT," accessed January 10, 2024, https://gizmodo.com/chatgpt-ai-samsung-employees-leak-data-1850307376.

45. Lindsey Wilkinson, "Samsung employees leaked corporate data in ChatGPT: report," accessed January 11, 2024, https://www.cybersecuritydive.com/news/Samsung-Electronics-ChatGPT-leak-data-privacy/647219/.

46. F. Paul Pittman, et. al., "Data Protection Laws and Regulations USA 2023-2024," accessed January 11, 2024, https://iclg.com/practice-areas/data-protection-laws-and-regulations/usa

47. Conor Murray, "U.S. Data Privacy Protection Laws: A Comprehensive Guide," accessed January 11, 2024, https://www.forbes.com/sites/conormurray/2023/04/21/us-data-privacy-protection-laws-a-comprehensive-guide/?sh=616915d85f92.

48. European Data Protection Supervisor, "The History of the General Data Protection Regulation," accessed January 11, 2024, https://edps.europa.eu/data-protection/data-protection/legislation/history-general-data-protection-regulation_en#.

49. Data Privacy Manager, "Meta Hit with Record €1.2B GDPR Fine," accessed January 11, 2024, https://dataprivacymanager.net/meta-hit-with-record-e1-2b-gdpr-fine/.

50. Parametrix Insurance, "Cloud Outage and the Fortune 500," accessed January 11, 2024, https://assets-global.website-files.com/64b69422439318309c9f1e44/6554bcf27d66c0c9135d3509_Parametrix%20Insurance-%20Cloud%20Outage%20and%20the%20Fortune%20500%202023.pdf.

51. Tom Warren, "ChatGPT is back online after a 90-minute 'major' OpenAI outage," accessed January 11, 2024, https://www.theverge.com/2023/11/8/23952129/chatgpt-down-openai-api-major-outage.

52. Wade Tyler Millward, "The 15 Biggest Cloud Outages Of 2023," accessed January 11, 2024, https://www.crn.com/news/cloud/the-15-biggest-cloud-outages-of-2023?page=2.

53. Google Cloud, "Google Cloud Service Health > Incidents > Multiple Google Cloud services in the europe-west9-a zone are impacted," accessed January 11, 2024, https://status.cloud.google.com/incidents/dS9ps52MUnxQfyDGPfkY.

54. AWS, "Summary of the AWS Lambda Service Event in Northern Virginia (US-EAST-1) Region," accessed January 11, 2024, https://aws.amazon.com/message/061323/.

[55] PromptHub, “Comparing Latencies: Get Faster Responses From OpenAI, Azure, and Anthropic,” accessed January 12, 2024, https://www.prompthub.us/blog/comparing-latencies-get-faster-responses-from-openai-azure-and-anthropic.

[56] OpenAI community, “GPT-3.5 and GPT-4 API response time measurements - FYI,” accessed January 12, 2024, https://community.openai.com/t/gpt-3-5-and-gpt-4-api-response-time-measurements-fyi/237394.

[57] Anna Dombajova, “Increased latency after switching to GPT-4 version 1106-Preview,” accessed January 12, 2024, https://learn.microsoft.com/en-us/answers/questions/1462200/increased-latency-after-switching-to-gpt-4-version.

[58] Paul Gillin, “AI model training rekindles interest in on-premises infrastructure,” accessed January 12, 2024, https://siliconangle.com/2023/10/16/ai-model-training-rekindles-interest-premises-infrastructure/.

Read the science behind this report ▶

Principled Technologies®

Facts matter.®

This project was commissioned by Dell Technologies.