AI in the Public Sector:

Harnessing Generative AI for Enhanced Productivity and Citizen Services

NUTANIX | Red Hat

NUTANIX | Red Hat

The New Era of AI Value Creation for the Public Sector

While artificial intelligence (AI) has been around for decades, the recent explosion of new generative tools like ChatGPT, Bard, Codex, and DALL-E 3 is the result of rapid advances over the past decade in cloud computing, deep learning and neural networks, open-source and user-friendly tools, as well as increased research collaborations. This confluence of new technologies represents an exciting era in business, government, and society in general, but comes with a wide range of views and emotions. Gonzalo Gortázar, CEO of CaixaBank, perhaps best sums up the current mood in the C-suite as follows: "Generative AI models surprise, impress, and scare us, all at the same time."

According to McKinsey, the economics of generative AI is off the charts, suggesting that it could add the equivalent of $2.6 trillion to $4.4 trillion annually to the global economy. The implications of this technology are truly game-changing as it stands poised to unleash the next wave of productivity across all segments of society.

For those working in the government and public services sector, in particular, generative AI will significantly transform how federal, state, and local agencies and citizens interact with each other. This comes on the heels of the recent White House executive order to spur use of artificial intelligence (AI) to find and fix security flaws in U.S. government infrastructure, in the face of growing use of the technology by hackers for malicious purposes. President Joe Biden has called it "the most consequential technology of our time."

NUTANIX | Red Hat

“One thing is clear: To realize the promise of AI and avoid the risk, we need to govern this technology”…calling the order the “most significant action any government anywhere in the world has ever taken on AI safety, security and trust.”

— President Joe Biden

While generative AI presents a host of new innovations that enable predictions, recommendations, or decisions which can benefit society, there are also inherent risks that make it a unique challenge to deploy and manage. For one, this technology can be maliciously used by opportunistic bad actors who wish to inflict harm on individuals, communities, public institutions, or the environment. For example, one only has to consider the socio-psychological impacts of deep fakes as a misuse of AI for exploitive purposes.

According to the Artificial Intelligence Risk Management Framework released by the National Institute of Standards and Technology (NIST) in early 2023, “Without proper controls, AI systems can amplify, perpetuate, or exacerbate inequitable or undesirable outcomes for individuals and communities.” NIST goes on to describe the document as “intended for voluntary

Helping customers tackle the biggest challenges they face in IT is at the core of what we do, from managing increasing multicloud complexity, to data protection challenges, and now adoption of generative AI solutions while keeping control over data privacy and compliance. Nutanix GPT-in-a-Box is an opinionated AI-ready stack that aims to solve the key challenges with generative AI adoption and help jump-start AI innovation.

— Thomas Cornely,
SVP, Product Management at Nutanix

use and to improve the ability to incorporate trustworthiness considerations into the design, development, use, and evaluation of AI products, services, and systems."

Whether from an innovation perspective or security standpoint, the sky's truly the limit with regards to generative AI, and government agencies today must be ready to manage this disruptive and game-changing technology. That's why federal, state, and local agencies need to rely on trusted partners that can accelerate digital transformation and quickly onboard AI solutions while mitigating risk and meeting regulatory requirements.

Nutanix and Red Hat have joined forces to offer an unparalleled solution for the public sector, combining the robust capabilities of the Nutanix Cloud Platform and Red Hat OpenShift AI. This integrated solution empowers federal, state, and local agencies with a secure, centralized hub that simplifies the management of diverse IT environments and ensures seamless delivery of mission-critical apps and data across various platforms (XaaS, on-prem, edge, public cloud, hybrid, or multi-cloud). Nutanix's GPT-in-a-Box™, a turnkey AI solution, addresses the challenges of complexity, scaling, and security in adopting AI/ML technologies, while Red Hat Open Shift AI provides IT operations leaders, data scientists, and developers with a unified solution to streamline the lifecycle of AI/ML models and applications, from experimentation to production. As Sarwar Raza, VP and GM of Cloud services, Red Hat summarizes.

Red Hat is helping organizations across various industries to accelerate business and mission-critical initiatives through the development and deployment of AI-enabled applications in the hybrid cloud. With this launch, joint customers can take advantage of Red Hat OpenShift AI along with the Nutanix AI-ready stack to jump-start innovation on a consistent, scalable open source foundation while maintaining control over their data.

— ***Sarwar Raza, VP and GM of Cloud services, Red Hat***

The collaborative engineering roadmap and joint support agreement between Nutanix and Red Hat ensure product interoperability and robust support, allowing government agencies to quickly and efficiently onboard new capabilities like generative AI while avoiding bottlenecks and challenging integration touchpoints.

Time to value in today's competitive market is a critical factor for adopting innovative features such as generative AI. Let's take a closer look at some of the challenges and opportunities.

NUTANIX | Red Hat

Transforming urban planning through an innovative Gen AI tech partnership

Paul is a seasoned urban planner who has spent 20 years at the city's Urban Development and Zoning Office. With a constant influx of new development projects and limited staff, Paul and his team are struggling to keep up with the demands of creating, updating, and reviewing complex urban development plans, environmental impact statements, zoning proposals, and community outreach documents. There are frequent calls from project developers waiting for approvals as well as citizens complaining about increased traffic congestion and limited parking resulting from a new development or zoning change.

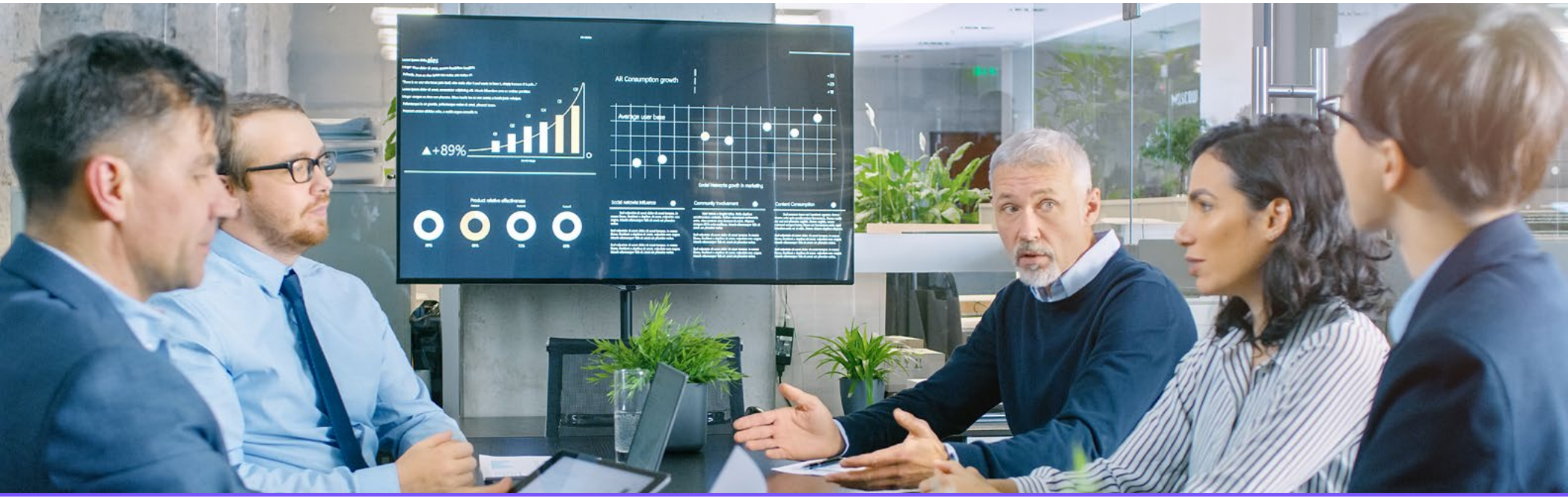
Paul has a rare chance to break away from the office and catch up with a few friends over lunch. Upon hearing about Paul's struggles at work, one of them shares about how her design firm recently started leveraging generative AI to auto-generate portions of their reports, proposals, and outreach documents. It's been a huge time-saver, and she encourages Paul to explore these new technology options.

Intrigued, Paul starts researching solutions and comes across Nutanix and Red Hat. He learns that Nutanix offers a cloud platform that can simplify data management, while Red Hat OpenShift can facilitate the deployment of applications that could automate and enhance certain aspects of the planning process.

Paul sees an opportunity here. He imagines using Nutanix to organize and streamline the vast amounts of data his team works with, from development plans to traffic patterns. At the same time, Red Hat OpenShift could be used to deploy applications that assist in automating routine reviews and analyses, allowing his team to focus on more strategic aspects of urban planning.

Paul proposes this idea to his management team, and they decide to team up with Nutanix and Red Hat to pilot an integration of their solutions. The pilot integration is gradual, following a three-month phased approach. Nutanix Cloud Platform is deployed to declutter and organize their data, making it easier to access and analyze while Red Hat OpenShift automates preliminary reviews of zoning proposals, quickly identifying potential issues.

This pilot partnership doesn't just streamline processes; it revolutionizes the way Paul's team navigates urban planning, turning challenges into opportunities for innovation and community satisfaction.



NUTANIX | Red Hat

Assessing costs of legacy vs. promises of AI-powered government

This Urban Development and Zoning Office example is unfortunately still the exception rather than the norm in the real day-to-day of most state, local, and federal agencies. In most cases, Paul's story ends differently; after returning to the office elated by the possibilities of deploying generative AI, he's soon brought back to earth by all the bureaucratic hurdles, budgetary constraints, and data privacy concerns. This challenge is underscored by a stark reality revealed in a 2023 report: 80% of the government's $100 billion annual IT budget is allocated to operating and maintaining existing systems, leaving limited resources for innovative projects like AI. This isn't to say that governments are not paying attention to AI — they are and quite a few agencies are in fact scaling up test pilot to instances.

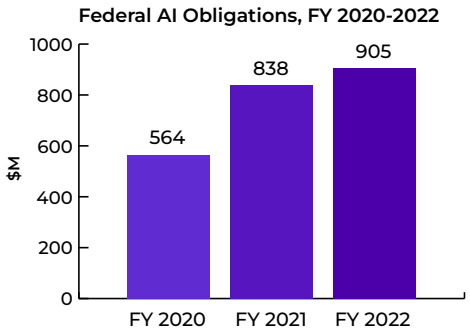
But the majority in government still classify themselves as beginners and full-scale deployments are few and far between. If you're a CTO, CIO, or other government IT manager lingering over the question of AI integration, now is the time to act. Delays only widen the technological gap, while proactive adoption can position your agency at the forefront of Government 3.0.

Imagine if we could reframe Paul's world? What if the Urban Development and Zoning Office began to ingest public data across various projects, community feedback, and environmental impact assessments, and then leveraged a generative AI system to help produce preliminary development plans, zoning proposals, and public outreach documents?

Achieving such a vision may seem far-fetched, but perhaps not when you consider the costs of maintaining legacy systems and waterfall project management approaches. A 2019 study found that 10 critical federal IT legacy systems were in dire need of modernization, which ranged from 8 to 51 years old and collectively cost around $337 million annually of tax payer money to operate and maintain. But aside from the surface costs of legacy IT systems, there are the hidden costs such as employee burnout, frequent downtime, data security, and poor customer experience.

COVID-19 was indeed a tipping point for governments' digital transformation, and indeed in the heightened turmoil caused by the pandemic many agencies had to spin up new technologies on the fly to meet public demands. But the health crisis was also a wake-up call for government leaders to assess just how far they must move away from just "doing digital" to actually "being digital."

That sense of urgency must continue. By 2025, the AI market in the U.S is expected to grow to more than $120B. AI and other cognitive technologies impacting the consumer market have filtered into the government sector, and federal investments in AI have increased 61% between 2020-2022 — making it one of the fastest-growing areas of technology investment. But the reality is that actual government adoption of AI in the U.S has still been slow, which is partly due to a lack of policy direction. A December 2023 report by the U.S. Government Office of Accountability reveals a complex but evolving AI landscape in government: 20 of 23 agencies have reported approximately 1,200 AI use cases, reflecting widespread recognition of AI's potential. However, this also points to an early stage in thorough implementation and policy development with the release of the January 2023 Artificial Intelligence Risk Management Framework (AI RMF 1.0).



While there's a fine balance between applying and regulating AI in government, the momentum is in the direction of adoption, and that's where every conscientious IT government leader should be focused. As Boston Consulting Group well-summarizes:

Implemented effectively, AI can generate benefits for public-sector organizations in three ways: smarter policymaking, reimagined service delivery, and more efficient operations. Thus, the technology can help governments better meet the needs of their citizens while making better use of taxpayer dollars.

Moving to generative AI is the new imperative for federal, state, and local agencies. Beyond saving resources, it's about tapping into AI's transformative power for data-driven decisions, back-office automation, and deeper citizen engagement. Sticking with legacy systems means security risks and wasted employee hours on manual tasks, when the focus could be on groundbreaking innovations using AI.

NUTANIX | Red Hat

Shaping the future: From legacy IT to generative AI in the cloud

While many government IT practitioners may see the benefits of generative AI and the promises of better return on investment through improved automation and enhanced citizen services, the implementation journey can be daunting.

In fact, generative AI might pose challenges in deployment comparable to the intricacies of maintaining legacy systems. Many organizations — governments included — are still trying to assess the risks, understand their data, and navigate the regulations around this new technology.

Yet, there are many opportunities in this space for those willing to move forward decisively despite the many unknowns. Debojyoti (Debo) Dutta, Vice President Of Engineering (AI) at Nutanix makes a strong case for the transformative power of generative AI:

Most people are underestimating the economic impact of generative AI, and AI / ML as a whole. I think this is going to change the way we look at productivity. AI might actually make human beings more intelligent rather than the other way around.

— Debojyoti (Debo) Dutta, Vice President Of Engineering (AI) at Nutanix

The following case study highlights how it's possible for a government entity to quickly overcome the tendency toward analysis paralysis and solve a costly legacy issue.

NUTANIX | Red Hat

How AI operations can fast-track global deployment, saving time and cost

Whenever one U.S. Federal agency needed to deploy to a hot zone halfway around the world, the logistics and costs associated with IT setup were staggering. The traditional approach involved dispatching teams to physically establish IT infrastructure, building out VMs, and handling manual configurations on-site. This not only meant significant financial costs in travel and overtime for an already limited IT staff, but also created extra stress for personnel who had to be away from their families for extended periods of time.

Realizing that this approach was unsustainable, the agency made a business decision. Since it had communications to its Edge sites, staff started focusing on how to automate those deployments. Thanks to a partnership between Nutanix and Red Hat, this agency underwent a major transformation. Instead of complex, manual setups, all that was needed were three Red Hat Ansible scripts to deploy the clusters and set role-based access controls. This method ensured built-in security and micro-segmentation from the start; VMs were categorized to guarantee restricted access and reduce potential attack vectors. The Nutanix Cloud Platform ensured that server integration was quick and easy: plug in a few cables, power on, and the system is recognized on the network. Running the Ansible scripts brought systems to life in a fraction of the previous setup time.

One of the best ways that state and local governments can serve their citizens with AI is to adopt an intelligent search bar on their websites for smart searches. Utilizing a chatbot either on the website or through text messaging is another really popular way for Gen X, millennials, and Gen Z to interact with their government through AI to get timely and relevant information."

— Amelia Gardner, County Commissioner, Utah County

The outcome? A massive reduction in deployment time. Whereas setups previously took two weeks, now they could be done within half a day. Teams would set up, break for lunch, and return to find everything operational. By eliminating IT staff travel time and stress, improving security posture, reducing human error, and achieving 80% faster deployment, this agency reaffirmed the benefits of embracing modern, integrated cloud-based automated solutions for the global theater. And with the power of generative AI, this agency can now explore even more ways to optimize IT deployment strategies, auto-generate system configurations based on the specific needs of a deployment zone, and predict and address potential security issues before they arise.

NUTANIX | Red Hat

Better together: How Nutanix and Red Hat are reenvisioning AI ops from the datacenter to the edge

More than ever today, government agencies must urgently replace IT legacy systems with seamless zero-touch provisioning solutions at the edge that can manage major technology shifts such as Internet of Things, 5G, and generative AI in a secure and resilient manner. That's where Nutanix comes in. Nutanix is a San Jose-based cloud computing company founded in 2009 and now trusted by more than 23,000 leading companies and known for solving some of the world's toughest cloud challenges.

Nutanix has joined forces with Red Hat, the world's leading provider of enterprise open-source software solutions. This strategic partnership brings together the best of both worlds with a pre-integrated full stack cloud platform that simplifies IT operations and enables customers to seamlessly build, scale, and manage containerized and virtualized cloud-native applications — so enterprises can grow and scale with ease.

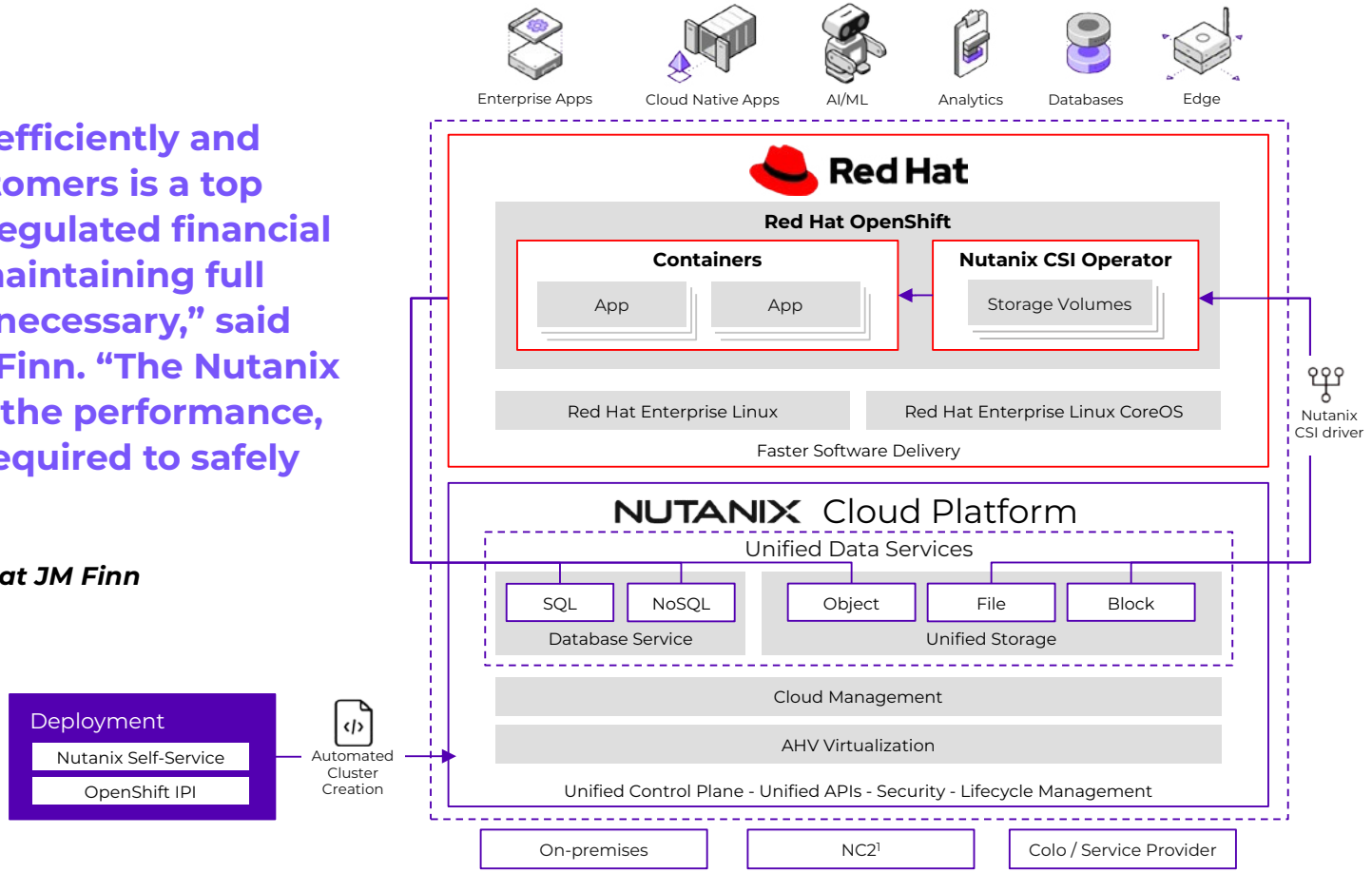
What's more, Nutanix and Red Hat bring together cutting-edge capabilities that make it easier than ever to spin-up AI/ML projects. Generative AI applications require massive amounts of data and computing power. But with Nutanix GPT-in-a-Box, customers get a full-stack, software-defined, AI-ready platform that makes it easy to deploy a curated set of large language models (LLMs) using the leading open-source AI and MLOps frameworks available on the Nutanix Cloud Platform.

And when all this computing power is coupled with Red Hat OpenShift and the Red Hat Ansible Automation Platform, you truly get a powerhouse solution for building, scaling, and managing traditional and cloud-native applications across the edge, datacenter, and cloud.

NUTANIX | Red Hat

“Leveraging AI to more efficiently and effectively help our customers is a top priority for us but, as a regulated financial services organization, maintaining full control over our data is necessary,” said Jon Cosson, CISO at JM Finn. “The Nutanix Cloud Platform delivers the performance, flexibility and security required to safely deploy AI workloads.”

— Jon Cosson, Head of IT & CISO at JM Finn



Imagine a government portal where citizens can interact with ChatGPT to answer questions in real time about tax filings, understanding local zoning laws, or checking on the status of applications (like passports or licenses). For example, you could build a GPT-in-a-Box instance and then integrate Ansible playbook to automate the deployment, scaling, and updating of that instance based on customer high-demand periods (such as during tax season) or scaled down during off-peak hours (to save on costs).

How to run an AI chatbot on Nutanix Cloud Platform

As AI chatbots like ChatGPT continue to revolutionize the way organizations operate, imagine the possibilities of deploying such a tool in a government agency to enhance citizen services by providing real-time information. Nutanix Cloud Platform stands out as an AI-ready solution, offering a consistent and robust environment for developing and deploying AI-driven applications tailored to the unique needs of government operations.

In this section, we walk through high-level steps you can take to set up a Chatbot and train it to run data in your environment. In this example, we are using two Nutanix clusters, one called “Far Edge Cluster (FEC)”, and another called “Near Edge Cluster (NEC)”.

How to Run an AI Chatbot on Nutanix Cloud Platform:
youtube.com/watch?v=WWbeTuzGIW4

1 **Ensure that your Chatbot application is set up on one of the Nutanix clusters powered by NVIDIA GPUs e.g. A100 GPUs. In this case, our application is running on a Kubernetes environment.**

2 **Validate that your clusters are installed correctly along with Prism Central. Prism Central is Nutanix’s centralized interface for managing your virtualized environment across multiple clusters.**

NUTANIX | Red Hat

3 Set up Nutanix object storage on each cluster (i.e. on FEC, NEC), with buckets set up for the models as well as the data, with bucket replication enabled across the two clusters as follows:

a. FEC:

   i. model-bucket – replication FROM model-bucket-src on NEC

   ii. data-bucket-src – replication TO data-bucket on NEC

b. NEC:

   i. model-bucket-src – replication TO model-bucket on FEC

   ii. data-bucket – replication FROM data-bucket on NEC

4 Set up an instance of Kubernetes on each of your Nutanix clusters (i.e. on FEC, NEC) with Kubeflow, an open-source platform for machine learning and MLOps on top of Kubernetes. Instructions for how to get started with Kubeflow can be found here.

5 Once installed, we can see our FEC along with the Kubeflow Pipelines running, which is helping to serve the model powering the chatbot.

6 Confirm the chatbot is running properly and the model is being served with the Kubeflow pipelines. On the Kubeflow dashboard, one can see the chatbot instance running and receiving queries.

NUTANIX | Red Hat

7 **Run test questions on Chatbot.**

When we go back to the Chatbot and ask a question:

"How do I reset the CVM password?"

Note how the response is generic and not helpful:

"To reset the CVM password you will need to contact the Nutanix support team."

Let's try another question:

"Why is Nutanix CVM reporting default password error?"

The answer is partially accurate, but it gets the CVM acronym wrong by calling it a Converged Virtualization Manager instead of a Controller Virtual Machine:

"The Nutanix CVM (Converged Virtualization Manager) is reporting a default password error because the default password has not been changed. The default password for the Nutanix CVM is 'nutanix.'"

---

8 **We can see the Kubernetes Inference Pipeline is running on our inference cluster (running on FEC) using Kubeflow pipelines. We also see that Kubeflow is fetching the new model from the Object store, creating the inference service using KServe to serve that model.**

9 **Then, we upload a new knowledge base article as a JSON file onto the storage bucket. This gets replicated to the cluster running the fine tune Kubeflow pipeline, via Objects replication. Note that we can see the new KB show up in the right bucket via the dashboard in Prism Central.**

NUTANIX | Red Hat

10 **Next, we go back into the Kubeflow dashboard on NEC and verify the updated fine_tune_pipeline is active. As the cluster updates the new KBs, it will trigger the fine tune pipeline to create a new model out of the new KB, this will take a few minutes. The new model will then be inserted into the object store and will get replicated at the edge due to the replication policies of the cluster. Finally, this model will be served at the FEC using KServe.**

---

11 **Go back to Step 7 and re-enter the previous question in the Chatbot app.**

This time the answer is much more accurate based on the new KB data:

"For security reasons, the Nutanix administrators might have changed the nutanix user's default password. Use the below command (run from any CVM) to change the CVM 'nutanix' user password. The change will be automatically replicated to the other CVMs in the cluster. 1. Log in to the CVM using admin user. 2. sudo password nutanix."

Try it out!

If you're a government agency that's interested in saving time, money, and plenty of IT headaches, then let Nutanix help launch your generative AI future. You can even get started today! Go ahead and check out the Nutanix AIOps & Automation Test Drive: https://www.nutanix.com/testdrive?type=tdops

And then email us at Jimmy.Baker@arrow.com to discuss how we can best serve your business needs!