Enhance performance and efficiency with generative AI

Reduce incident resolution time with an all-in-one solution from HPE and Wipro

Too much complexity, too many security concerns

Most enterprises will agree that their IT landscape — and the accompanying IT operations support — are growing increasingly complex. As complexity escalates, IT administrators require more time to resolve infrastructure problems.

Traditionally, administrators have referred to infrastructure documents to help them identify and resolve issues. During major incidents today, however, administrators have no time for research; they are under tremendous pressure to fix issues quickly and efficiently and keep systems performing within agreed-upon SLAs.

Coupled with the need for faster problem resolution is the growing concern about data security and control. Organizations that have adopted a pure cloud-based operating model face the constant challenge of ensuring data security and compliance during data transfers.

Rather than continue to use yesterday's IT operations solutions to tackle today's challenges, you can choose a smart new solution from HPE and Wipro — leveraging the power of generative AI[1] and LLMs to rescue your IT administrators, secure your data, and protect your IT infrastructure.

What is a large language model?

"In the AI world, a language model provides a basis to communicate and generate new concepts. A large language model (LLM) is a type of artificial intelligence (AI) algorithm that uses deep learning techniques and massively large data sets to understand, summarize, generate, and predict new content. The term generative AI is also closely connected with LLMs, which are, in fact, a type of generative AI that has been specifically architected to help generate text-based content."[2]

All-in-one generative AI appliance, all on-premises

To ensure data security, while also slashing incident resolution time and increasing control, HPE and Wipro created an on-premises generative AI appliance based on Wipro's SmartOps software running on the HPE Machine Learning Development Environment (MLDE) Software. With this complete hardware/software appliance, your organization benefits from a proven AI / machine learning (ML) solution, delivered via HPE Cray Supercomputing XD670 or the HPE hybrid cloud platform. HPE GreenLake is HPE's hybrid cloud platform. It is designed to help organizations unlock the power of data and AI and connect across hybrid landscapes, where data and apps are distributed from edge to cloud. HPE GreenLake delivers a single, consistent cloud experience for all IT through a portfolio of platform-enabled services and an ecosystem that helps customers.

This next-generation appliance delivers the performance and efficiency your IT administrators need to keep your systems and operations running, together with targeted support for a variety of use cases and technologies. Here's how the solution works:

- Constantly monitors the IT service management (ITSM) queue
- Detects new incidents
- Compares each incident with past similar incidents
- Applies deep analytics to the incident
- Reports a plan of action for resolution
- Updates the incident with standard operating procedures and real-time metrics

[1] "What is generative AI? Everything you need to know," TechTarget, accessed February 2024.

[2] "Large language models (LLMs)," TechTarget, accessed February 2024.

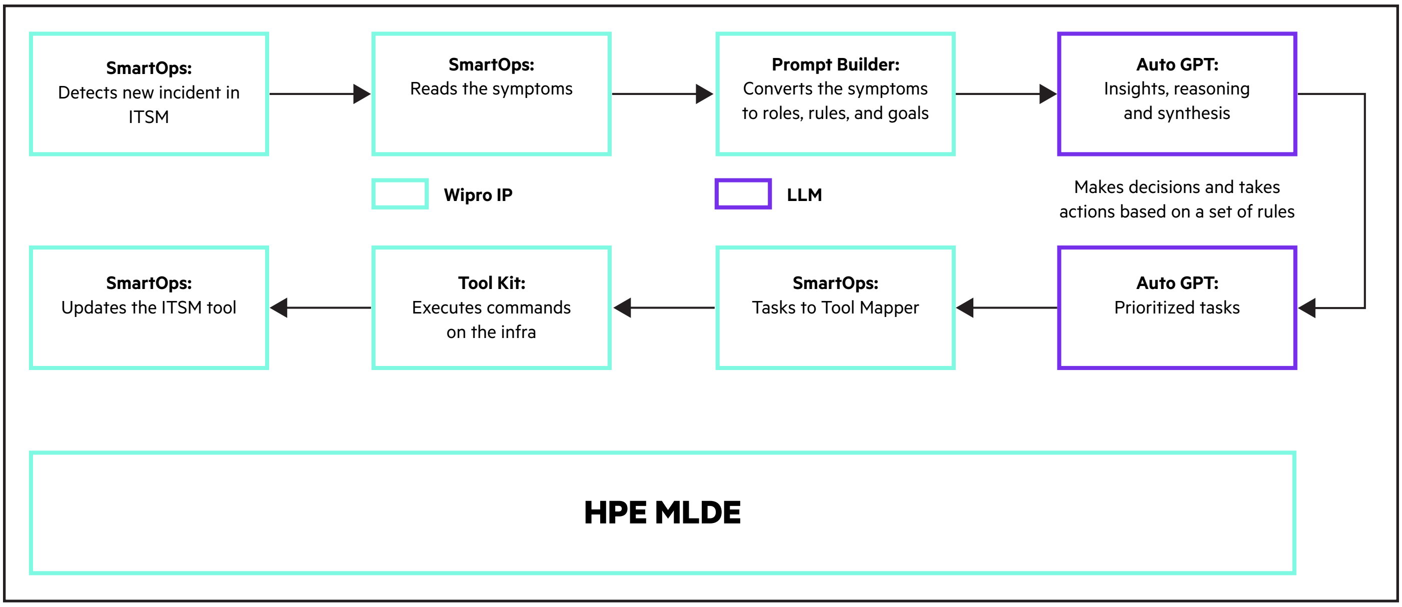

Figure 1. Wipro SmartOps and HPE MLDE workflow

With this HPE and Wipro on-premises generative AI solution, your organization will receive:

- Your choice of LLM
  - Select an LLM model from an available set of tested models.[3] Choose the model that best suits your intended purposes and meets your organization requirements.
  - Use domain-specific LLMs, fine-tuned for ITOps.
  - Use the Retrieval Augmentation Generation method to embed enterprise-specific confidential information in the LLM.
- Enhanced data security and compliance
  - Retain complete ownership and control of your data.
  - Reduce data exposure to external threats and vulnerabilities.
  - Comply with industry regulations.
  - Privately host LLMs to help secure the prompts not being used to train cloud-hosted LLMs.[4]
  - Deploy guardrails to protect sensitive information.
  - Eliminate incorrect LLM training data.
- Higher performance, lower latency
  - Accurately predict resource allocation; gain complete control of your LLMs to ensure consistent performance and avoid potential bottlenecks during peak usage.
  - Minimize network latency.
- Reduce the cost of running the LLMs (build vs buy)
  - LLM-based inference has token-based billing associated with every prompt and its response. For organizations handling large volumes of users/queries, it is more economical to deploy a privately hosted generative AI solution, compared to a cloud-hosted solution using the software-as-a-service (SaaS) model.

[3] Some locally hosted LLMs tested with Wipro SmartOps include Mistral and Llama 2.

[4] For example, prompts not remaining in temporary memory in the cloud for 30 days for audit purposes.

Enterprise-class machine learning framework

With data security being an ever-present threat, many organizations prefer to deploy a private generative AI/LLM solution. Doing so eliminates external dependencies on cloud-based services or third-party providers, which in turn reduces the potential attack surface and minimizes the risk of data breaches or unauthorized access to data through external systems.

Optimized to reduce the time and resources needed for AI/ML development, HPE MLDE is a complete enterprise-class ML framework, designed to be hosted on-premises with HPE Cray Supercomputing XD670 or delivered via HPE GreenLake. To deliver a cohesive experience, HPE MLDE logically groups all the pieces of the ML workflow together, running on high-performance GPU-powered HPE infrastructure.

The HPE MLDE is characterized by:

- Enterprise-grade capabilities — Offering role-based access control
- Scalability — Supporting LLM training across environments
- Pay-as-you-go* consumption-based model — Delivered via HPE GreenLake for HPC
- Choice — Supporting either AMD or NVIDIA® accelerators/graphical processing units (GPUs)
- Flexibility — Offering support for distributed training
- Reduced cost — Enabling you to run both AI and high-performance computing (HPC) workloads on a shared cluster

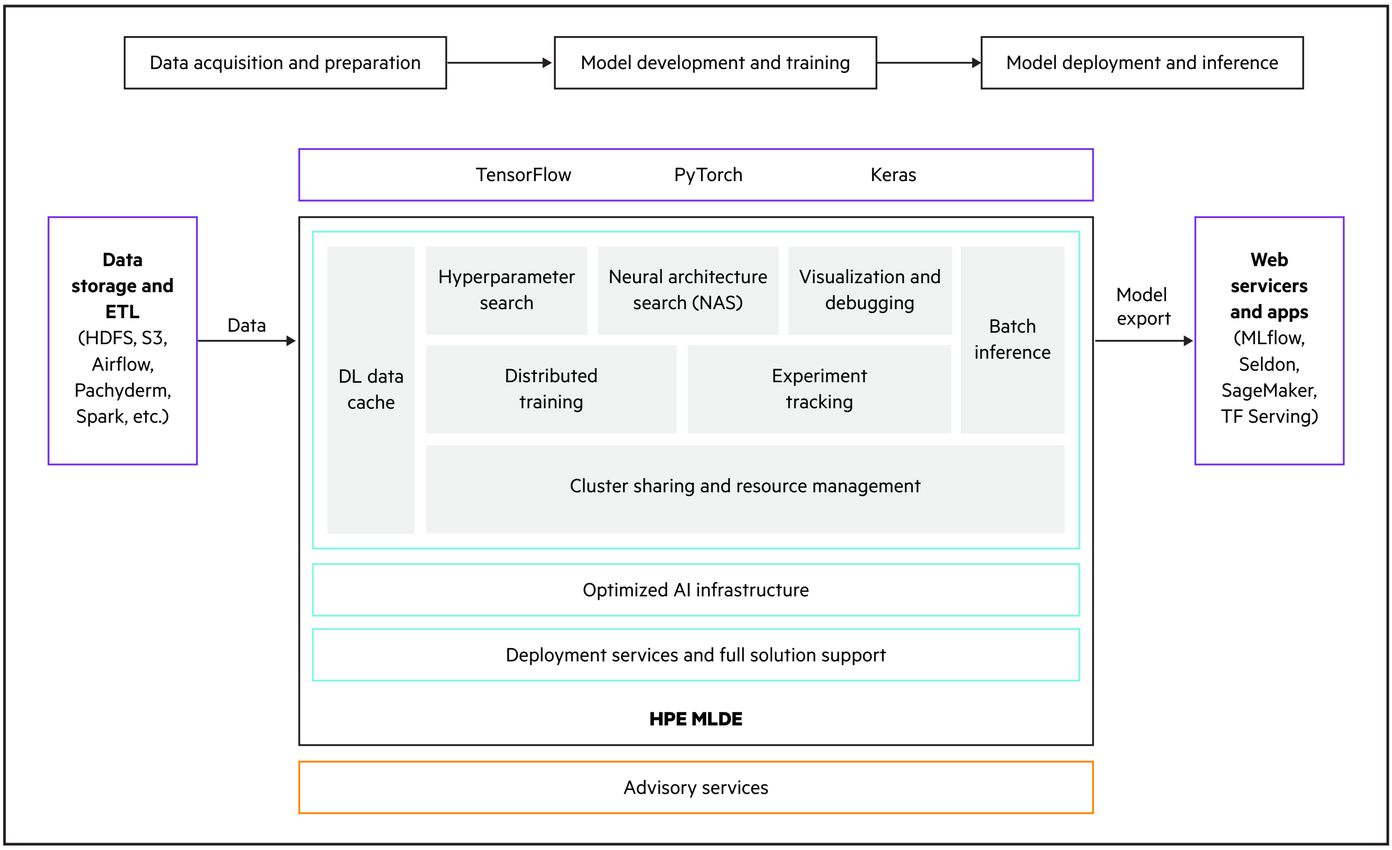

Figure 2. HPE Machine Learning Development Environment

* May be subject to minimums or reserve capacity may apply

Solution in action

Securing the IT infrastructure

With SmartOps and HPE MLDE, L2 and L3 IT administrators have the tools they need to solve problems typically reserved for L4 IT administrators. The solution applies deep analytics to search an organization's database for similar kinds of problems that happened previously and then offers a viable solution to the L2 or L3 IT engineer directly.

This on-premises solution meets security challenges by:

- Making the deployment of LLMs more secure across the enterprise. Data remains within the enterprise, which resolves the primary problem of ensuring data security.
- Reducing dependencies on other third-party vendors by using a flexible solution (SmartOps) that can be customized and trained to support a variety of use cases.
- Deploying the LLM with real-time IT operations data and enterprise-specific data. With accurate, up-to-date infrastructure status, L2 and L3 IT administrators can resolve problems more efficiently.

Identifying application and infrastructure hot spots

Leveraging generative AI to analyze historical incident and resolution data, Smart Causal Analysis identifies application and infrastructure hotspots in the IT environment that are causing service outages, performance degradations, and frequently recurring issues.

First, Smart Causal Analysis preprocesses data with grammar corrections, language translation, image-to-text translations, and email/document attachment processing to understand the intent of the data. Then, LLMs are used to segregate issues into topics, sub-topics, and similarities to create a more refined view of the issues' root causes and better understand how frequently the issues recur.

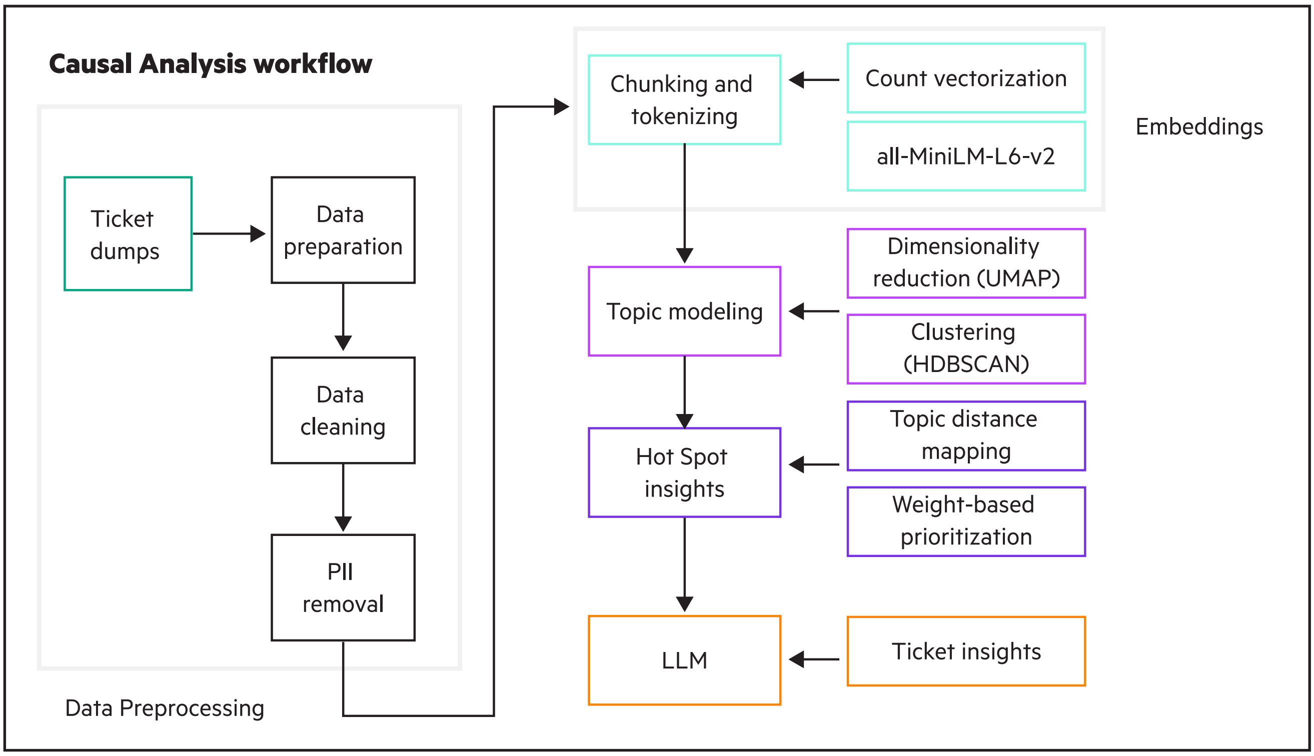

Figure 3. Smart Causal Analysis workflow

Speeding problem resolution time[5]

Most enterprises with large, complex IT infrastructures experience similar operational challenges including:

- Increased mean time to resolution (MTTR), resulting in longer system downtime after an incident
  - Average MTTR is 30 hours based on a survey of 200 organizations that are not using AI.
  - An AI-driven solution can reduce the MTTR by 50%.
    - MTTR at the top 25% of companies that are leveraging AI it is about 15 hours.
    - MTTR can be improved by optimizing ticket routing, leveraging automation, enhancing self-service options, and completing regular performance analyses using AI-driven tools.
  - On average, companies have seen a 35% increase in ticket volumes over the past year.
- High cost for ticket resolution
- Overworked staff
  - Administrators are finding it increasingly difficult to identify answers to infrastructure issues using traditional manual methods.

Wipro SmartOps running on HPE MLDE can help IT administrators resolve incidents faster by providing context-specific AI-driven information along with incident logs — resulting in 25% reduction in incident resolution time.

Tackle today's IT operations challenges — and tomorrow's — with confidence

Regardless of the size of your organization or your industry focus, you need a smart IT operations solution designed to flex and adapt to an ever-changing marketplace. You want a complete solution that leverages the power of generative AI to speed your time to problem resolution. You need an on-premises appliance that protects your data while also supporting your modern hybrid cloud environment. And you require professional services to help you with smart strategy, deployment, management, and more.

Get what you need — with a generative AI solution built on Wipro SmartOps and HPE MLDE. This solution is designed to help you overcome the challenges enterprises typically face when setting up a generative AI environment:

- Managing the hardware that supports AI model development
- Manually implementing ML features through piecemeal code
- Meeting the evolving requirements of the ML infrastructure
- Understanding the changing business requirements that are driving the need for AI solutions
- Setting up lighthouse wins for AI to drive future success within the organization
- Creating a road map for AI expansion
- Ensuring data security

[5] Based on Wipro's experience and internal solution deployments

Learn more at

HPE.com/partners/Wipro

HPE.com/GreenLake

HPE.com/us/en/hpe-machine-learning-development-environment

Chat now (sales)

Hewlett Packard Enterprise

a50010316ENW