eBook

AI in Your Enterprise

Best Practices for Deploying and Running AI Applications

NUTANIX

Introduction

Artificial intelligence (AI) is transforming the business world in unprecedented ways, helping enterprises improve their efficiency, productivity, innovation, and customer satisfaction. AI can also introduce new opportunities, complexity, challenges, and risks for organizations. In this eBook, we will explore the importance of AI in the enterprise, its current state and trends, and best practices and strategies for implementing and managing AI projects in your company.

Table of Contents

The Evolving World of AI

AI has entrenched itself in the enterprise. It is no longer a futuristic concept, but a reality that is transforming the way businesses operate. AI has become an essential part of many industries including manufacturing, healthcare, finance, education, public sector, and more.

Thanks to advances in data, algorithms, and computing power, AI is continuing to evolve at a rapid pace. Traditional AI is designed to perform specific tasks based on pre-programmed rules and data, for purposes such as process automation, supply chain optimization, natural language processing, data analytics, cybersecurity, and much more. However, a new form of AI called generative AI has burst onto the scene, with enterprises sprinting to derive immediate value from it.

Generative AI rose to popularity with the release of GPT-3, a large language model (LLM) from OpenAI, in late 2022. It differs from traditional AI by training on a vast data set with trillions of data points to generate new content derived from learned patterns. Generative AI is also known for its ability to mimic general human responses, a far cry from the task-specific capabilities of traditional AI.

LLMs are foundational models such as GPT-3 and GPT-4, Llama 2, PaLM 2, and Falcon, that have been trained on enormous quantities of unlabeled data. Building these LLMs requires sizable investments beyond most organizations' reach. The alternative is to create a compact LLM, which involves fine-tuning an LLM, typically with proprietary organizational data, to generate specific answers for a wide variety of use cases.

Though they are quite discrete, both traditional AI and generative AI possess the ability to profoundly transform the enterprise of today and tomorrow.

55%

Amount of respondents from a McKinsey Global Survey reporting that their organizations have adopted AI in early 2023.

33%

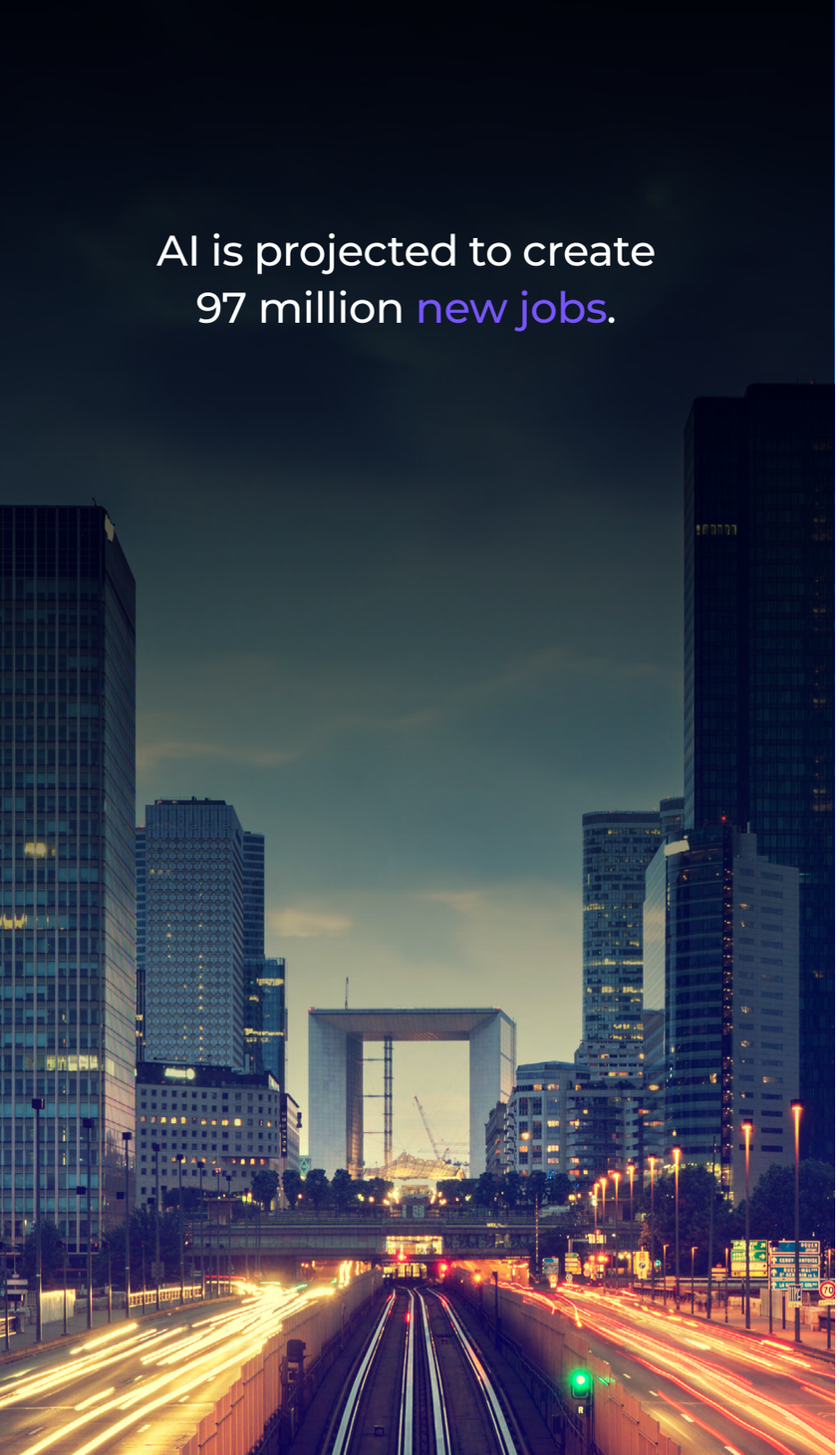
Amount of organizations already regularly using generative AI in at least one function.

84%

Technology executives expecting to modernize their IT infrastructure to support AI, according to the 2023 Nutanix State of Enterprise AI report.

$407B

Projected AI market size by 2027, a $320B increase from 2022.



Challenges to AI Implementation

Despite AI's tremendous benefits, organizations face key challenges when implementing and growing their AI practices. To ensure success with AI initiatives, they need to plan carefully, involve all the necessary stakeholders, and continuously monitor their AI applications and data.

AI Model Management and Monitoring

Model management and monitoring in AI are essential for ensuring the quality, reliability, and performance of AI systems. Continual fine-tuning of models is necessary to produce timely, relevant results. Monitoring AI models involves collecting, analyzing, and reporting on metrics and feedback of all AI models in production.

These functions form the core of organizational machine-language operations (MLOps), a set of practices designed to deploy and maintain machine-learning models. Each model and the training and fine-tuning of its associated data creates a new version of a generative AI app. The models must be monitored and controlled to ensure that each version is of successively higher quality, is protected, and has full data lifecycle management like any other enterprise application.

Workforce Skill Gaps

As AI is a relatively new tool in the enterprise, organizations face a skillset gap in their workforces. A lack of formal AI training exists in many enterprises, turning reskilling into a critical imperative. The skills gap remains the biggest barrier to AI adoption.

There is also enormous incentive for workers to upskill: 82% of executives believe workers who are skilled using AI should be paid more, 74% believe they should be promoted more often, and 72% believe their company should increase their investment in AI learning and development programs.

Alignment of the Business and IT

As with most newer technologies, the business side sees almost unlimited potential with AI in terms of resource savings, employee productivity gains, and enhancements to products and services. AI helps the business to make faster and better decisions, improve customer experience and satisfaction, automate tasks, enhance security, and so much more.

IT shoulders vast responsibilities in planning, executing, and monitoring AI projects in the enterprise, including designing and developing AI infrastructure, deploying and maintaining the AI system, ensuring the AI practice meets privacy and governance standards, and much more. At peak efficiency, optimal AI environments help to reduce complexity, enhance predictability, and completely protect enterprise AI assets.

With vastly different responsibilities on both sides, harmony between the business and IT is essential for AI to flourish in any organization.

Data Consistency

You can leverage your organizational data to fine-tune your models. Once your models are trained and updated, they can be deployed and inferenced at the edge to reduce latency. High-quality data makes your models more efficient and yields the most optimal results.

Structured data is organized and searchable in databases, with a predefined format and structure. In the world of AI, it is referred to in terms of parameters and embeddings. Parameters and embeddings are tagged and categorized and fed into traditional AI models, while untagged parameters (unstructured, or raw data) are fed into the LLMs. The model then needs to be verified to ascertain whether the data quality represents an increase in accuracy. More data is not always good; it is the quality of the data that matters.

Data cleansing is the process of detecting and removing inaccurate records from the parameters and embeddings, such as errors, duplicates, inconsistencies, or irrelevant parts. Data also needs to be sanitized—this is the highly important process of removing personally identifiable information (PII) and other sensitive data including social security numbers, customer names, passport numbers, etc. from your datasets.

Cost

Cost considerations for AI practices depend on numerous factors such as the type of AI model required, current and future infrastructure, existing AI workforce skillsets, and much more. Regardless of how you choose to set up AI projects in your organization, consider the following cost areas:

Implementation: On-premises instances of LLMs and pay-as-you-go cloud services are among the options for those wanting to start quickly.

Infrastructure: Upgrades can involve newer servers, GPUs, and other hardware.

Energy: Large-scale AI systems can require significant power.

Staffing and Training: Upskilling the workforce and acquiring relevant talent can be costly.

Data Acquisition and Management: Acquiring raw data from sensors and other means can be a costly setup. Managing data includes structuring, cleansing, and maintenance measures.

Training and Upgrading Your AI Models: Costs can vary on model size and type, hardware software requirements, and more.

Performance: Monitoring data usage and other features can be costly. Using an Application Performance Monitoring tool can help.

Interoperability Issues

Operating an enterprise AI platform often involves multiple clouds, starting with your on-premises datacenter and extending to any number of cloud providers. Leveraging data, applications, and services between clouds is complex, requiring specific knowledge of each cloud provider's toolsets. Vendor services can unite private and public clouds, leveraging the strengths of each while imparting simplicity and system-wide security to ensure business continuity for AI operations throughout your cloud estate.

AI Use Cases are as Boundless as the Imagination

Cross-industry

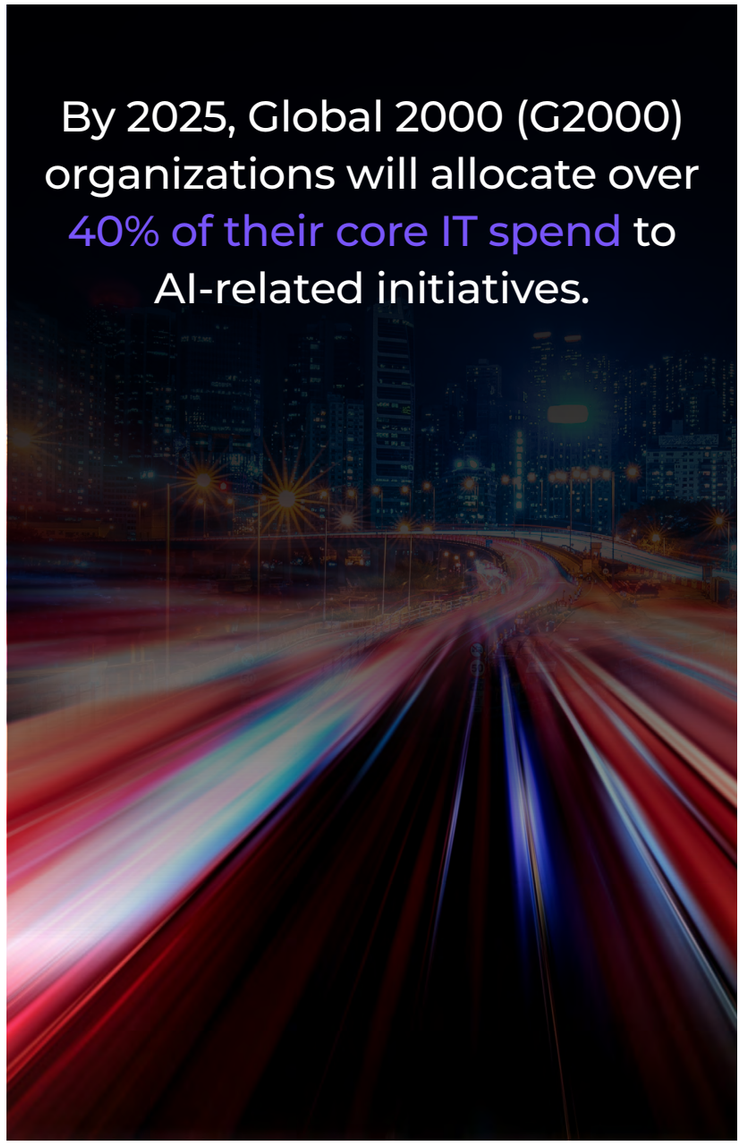
- Support and technical knowledge-base chatbots
- Document automation
- Fraud detection
- Video inferencing
- Content creation

Industry-specific

- **Healthcare:** Patient data mining and analysis to identify risk factors
- **Financial Services:** Detecting money laundering and other unlawful activity
- **Manufacturing:** Image recognition for quality checks and inspections
- **Logistics:** Forecasting demand patterns to enhance supply-chain operations
- **Public Sector:** Automated budgeting for smarter resource allocation

Generative AI-specific

- Process automation
- Code co-pilot development
- Review summarization for e-commerce apps
- Customer-specific marketing campaigns
- Predictive modeling in research and development



Key AI Strategies

As you create and take your first steps along the pathway of enterprise AI, it is important to follow a set of essential strategies to help form the goals, scope, and methods for your practice. The following will help you to enable effective management of processes, resources, and outcomes to ensure the highest quality, accountability, and reliability of your AI operations.

Determine Where to Build and Deploy

Enterprises enjoy the aspect of predictability when they run AI models locally in their data centers, fine-tune them for efficiency, and then deploy at the edge, cycling through this process repeatedly to fine-tune their models. The predictability arises from the familiarity and reliability of leveraging known resources; for example, using their own data versus data from cloud services. Examples of this practice include:

- Video inferencing for loss prevention at retail self-checkouts, where small AI systems reside at the edge in self-checkout kiosks.
- Security and surveillance systems with face recognition capabilities deployed at multiple locations.
- Smart traffic lights that change in real-time to coordinate efficient routes for emergency-services vehicles.

Cloud Costs

Pay-as-you-go cloud services can grow out of control quickly if not continually monitored. Ensure that you encounter no surprises in this regard. One way that expenses can escalate is when you fine-tune your own data with a large LLM in the cloud (as opposed to on-premises LLMs, which leverage large LLMs but process the data locally). Not only can this be incredibly cost-prohibitive, but it also risks leaking sensitive data.

Latency

Latency—the time it takes for a system to respond to a request or perform an action—affects the performance and efficiency of AI apps. Real-time or near-real-time processing of data, as well as the expected frequency and timing of results, depend on an efficient, low-latency AI system. Numerous factors impact latency in AI, including network conditions, the speed between the data source and the system running the AI model, AI model complexity and size, the architecture of the AI system, and where data and the AI system are located. GPU processing power also comes into play. Sometimes fast GPUs are needed for low latency; other times, cheaper and smaller GPUs can provide similar latency but be better optimized for cost.

Data Gravity

As AI systems grow, they leverage more and more data, which in turn requires larger and more extensive applications, models, and services to support it. This concept is known as data gravity, and it plays an important role in how effectively an AI practice can evolve at the enterprise level. Data gravity affects performance, security, and scalability of AI systems and creates additional challenges involving bandwidth and movement of data when not properly managed. As you grow your AI platform, try to anticipate how much “mass” your data will have and plan ahead.

Flexible AI Software Stack

Additionally, consider deploying AI with an open-source software stack. To maintain the most flexibility for your current and future AI projects, it is wise not to get locked into proprietary technology. An opinionated AI stack, which is a bundled solution of open-source software and selected services, allows you to choose your own technology for machine learning (ML) frameworks, MLOps platforms, data science platforms, and the rest of your AI stack.

Enact Stringent Data Security, Privacy, and Sovereignty Standards

All organizations, whether regional or global in scope, must embrace data regulations to avoid stiff penalties. AI broadens the concern for enterprises because it might involve leveraging sensitive data that could potentially leak into the wild when training and fine-tuning models.

Regulatory compliance helps to ensure information security by requiring organizations to adhere to rules designed to protect their assets from threat actors. Those responsible for data security and compliance should ensure they are familiar with and adhere to the regulatory requirements in every country in which they operate.

Protect your AI models and data by stringently adhering to one or more security frameworks, including National Institute of Standards and Technology Cybersecurity 2 (NIST CSF 2), Open Worldwide Application Security Project (OWASP), and AI TRiSM as defined in the Gartner® article "Tackling Trust, Risk and Security in AI Models".[1] Many organizations leverage a combination of these and other frameworks in their security practices. Specific works such as OWASP's top 10 list of AI security threats, which focuses on LLM applications, are also valuable resources for your AI security practice.

Also, remember the criticality of securing all of your sensitive data with encryption. In a 2022 survey, 55% of respondents reported that their organization transfers sensitive or confidential data to the cloud, whether it is encrypted or made unreadable in some other fashion. Do not make your data easy for malicious actors to view and handle.

Apply AI to Promote Organizational Efficiency

AI has become a revolutionizing force across numerous industries, having the potential to massively reshape the future of work. While companies are eager to realize the economic benefits of AI, they should also be aware of AI's unique ability to augment employee work, and not replace it, and look for ways to enact this in the enterprise. AI can be used to automate routine tasks, allowing employees to focus on more creative and productive endeavors.

AI Assistants

One intriguing example of this is an AI assistant, a conversational interface that aids workers with tasks and decision-making processes. AI assistants can help draft emails, answer specific questions, provide context-aware support, and guide workers through complex tasks and processes. They can even perform highly specific tasks such as data analysis, unifying disparate systems to connect tools and applications from different platforms, and generating code for development practices.

AIOps

AI operations is a process that leverages AI techniques to maintain and improve IT infrastructure, including automating critical tasks like performance monitoring, data backups, and workload scheduling. By collecting and analyzing data from many sources, AIOps can bring proactive real-time insights to your AI practice. This helps to reduce operational costs, reduce issue mitigation time, streamline AI operations, and enable predictive service management throughout the AI platform.

Demand Forecasting

Demand forecasting, the process of estimating the future demand for a product or service, can be greatly enhanced by AI. It can help to improve accuracy and reduce errors; incorporate external factors such as weather, events, holidays, and promotions; enhance agility and responsiveness by forecasting changes in supply and demand; enable more effective data-driven decisions; and increase innovation by testing and measuring the effectiveness of initiatives.

[1]Gartner Article. Tackling Trust, Risk and Security in AI Models, 5 September 2023

Enhancing Your AI Practice

Leveraging the benefits of AI in your organization and overcoming the inevitable challenges that arise require continual emphasis on applying new ideas and innovation to your AI practice. Here are a few key areas that can help you evolve your AI operations.

Develop Your Workforce Skills

The lack of available workers with Science, Technology, Engineering, and Mathematics (STEM) skills is well known, and finding skilled AI workers can be equally difficult. Therefore, as previously discussed, it is paramount to upskill your existing workforce to the greatest extent possible in AI knowledge and techniques, to help offset what 68% of executives see as a moderate-to-extreme AI skills gap in their workforces.

This practice will go a long way in establishing and cementing your AI efforts enterprise-wide. You can begin by simply training your workers to enter the most effective prompts into AI assistants and other AI tools to generate the best results. Workers can then become more familiar with AI and begin to treat it as a benefit to their positions.

Elevating this practice, prompt engineering as a specialized job role is also growing in popularity, requiring domain knowledge, creativity, and the understanding of how AI models work. This represents one way that AI is elevating workforce openings in the industry.

Additional specialized AI roles emerging in the workplace include AI-focused trainers, AI engineers, AI researchers, AI ethicists, AI consultants, and AI analysts. Potential new roles could include AI sentiment analyzers, AI input and output managers with expertise in bias and regulatory measures, AI compliance managers, and many more.

Automation

Automation is key as you build out your AI practice. For example, it is highly important to make AI as self-service as possible for data scientists and other AI practitioners in your organization. Projects are slowed when workers are forced to open tickets and wait for IT to deploy a container or virtual machine (VM) loaded with the libraries and tools they need. Automating this process so data scientists can deploy their own containers and VMs is vastly preferable. In addition, automation reduces human error.

Ideal Data Placement

The location and ease of access to the data used in training your models can significantly impact the performance, efficiency, and cost of your AI systems. The closer your data is to your AI models, the speedier they run, and the more control you have over it. But more importantly, take caution not to store or utilize your AI data in a public location, as this can expose it to potential threats and leakage.

It is worthwhile to re-emphasize the importance of reducing or eliminating data silos throughout your organization, and cleansing your data regularly, to impart more efficiency to your AI operations.

Single Control Plane

You can impart even more efficiency to your AI operations by using a single control plane to centralize the management of your AI platform. Leveraging a common interface helps to reduce complexity by unifying role-based access and policy-driven controls, for example, allowing you to operate with a standardized set of policies and controls to govern your entire AI estate. This control plane unites AI applications, workloads, and data mobility across every environment your AI platform operates in.

39%

of businesses hired software engineers and 35% hired data engineers for AI-related positions in 2022.

The Last Word

AI is a fascinating and powerful technology that has the potential to transform organizations and the world in endless ways. From enhancing creativity and productivity, to solving critical business challenges and innovating across industries, AI can be a positive force of change if leveraged wisely and ethically in the enterprise. As we close, here are a couple of final thoughts to ponder as you begin to build your AI practice.

Create Purpose in Your AI Practice

You can start with small AI projects and use them as early wins to maximize the value of data and insights quickly. Make every effort to adopt a creative mindset for use cases, data, and analytics in your AI endeavors. And build a holistic AI practice that touches every part of your enterprise, empowering your workers to be more productive than ever before.

Your Data is the Window to Your Success in AI

Your organizational data is unique and has limitless value. Start to tap into it with AI and you will derive ever more value from it—data with previously small value will suddenly become very interesting. Because your data is the lifeline of your AI operations, protect it with the strongest security possible and surround it with strict governance and compliance measures. Treating your on-premises data in this fashion allows you more control over it and helps to reduce risk in your AI operations.

Thank You

To learn more about how Nutanix can help you jumpstart your AI journey with security, privacy, and control, visit nutanix.com/ai.

NUTANIX

info@nutanix.com | www.nutanix.com | @nutanix