# [AI Products] In the cloud

[PART 1]

## NAVIGATING THE COMPLEXITIES OF AI DEVELOPMENT AND CLOUD DEPLOYMENT

DEFIANCE DIGITAL

# Table of Contents



# Introduction

You're exploring artificial intelligence (AI) applications for many reasons, including the potential to enhance your offerings and competitive advantage. Businesses can indeed unlock untold product and customer engagement opportunities using mechanisms like generative AI and large language models. Aspiration and execution are distinct, however. Developing with AI involves intricate challenges across development and deployment, making it difficult for organizations to accomplish without thorough expertise and resources.

Cloud hosting further complicates AI products, as the rules and requirements of operating technology in a cloud environment versus one's systems change significantly. Beyond figuring out how to develop an innovative AI-based tool, businesses must also test and launch their products into the cloud without compromising data, code, or operating integrity. The most effective tools for accomplishing this are education, preparation, and knowing when to ask for help.

This preparation manual is delivered in two parts: Part 1 serves as a guide to the basics of AI in the cloud, and Part 2 explores developing AI products in the AWS cloud environment specifically (skip to Part 2 to learn more about AWS's AI offerings). This report includes best practices for ensuring secure and sustainably launched AI products. It also explores when and how to leverage outside expertise like a cloud-managed service provider (MSP) to streamline their AI-to-cloud journey.

# The Business Case [for AI and LLM ]

## An AI Primer

Artificial Intelligence is a broad field of computer science focused on creating systems that perform tasks that typically require human intelligence. AI encompasses a range of techniques and approaches, including machine learning, natural language processing, and problem-solving. Some key terms in AI and relevant to this paper are:

- Machine learning (ML) is a subset of AI that involves training algorithms on data to enable them to learn patterns and make predictions or decisions without explicit programming. ML techniques such as deep learning, or multi-layered neural networks create the foundation for models like Generative AI.

- Generative AI is a class of artificial intelligence algorithms designed to generate new, original content. Unlike traditional AI models that are typically task-specific and make decisions based on predefined rules, generative AI can create novel outputs such as images, text, or other data types.

- Large Language Models (LLMs) are a specific application of machine learning and Generative AI that leads to content creation within a text-based context, where models like OpenAI's GPT (Generative Pre-trained Transformer) are trained on vast amounts of text data, allowing them to interpret, generate, and manipulate human-like language for tasks such as text completion, translation, and question answering. LLMs showcase the power of machine learning in processing and generating human-like text, contributing to advancements in natural language interpretation and communication.

# The Business Case

Excellent product development is only as strong as the infrastructure on which the products are built, and one of the surest ways to gain and maintain a robust IT infrastructure is by leveraging CloudOps. The CloudOps process entails provisioning, monitoring, scaling, and optimizing cloud infrastructure to ensure the performance, availability, and security of applications and data hosted in the cloud. CloudOps teams are responsible for implementing best practices, automation, and tools to manage cloud environments efficiently and meet the needs of users and applications. Some of the main activities include:

# The Challenges of AI Tech

Excellent product development is only as strong as the infrastructure on which the products are built, and one of the surest ways to gain and maintain a robust IT infrastructure is by leveraging CloudOps. The CloudOps process entails provisioning, monitoring, scaling, and optimizing cloud infrastructure to ensure the performance, availability, and security of applications and data hosted in the cloud. CloudOps teams are responsible for implementing best practices, automation, and tools to manage cloud environments efficiently and meet the needs of users and applications. Some of the main activities include:

# The Building Blocks [of AI in the Cloud ]

Many decisions shape the trajectory of companies' AI development and cloud deployment process. Companies want to ensure responsible market entry, protect data and security, and effectively manage cloud infrastructure as AI technology evolves or they experience business growth. But there's a lot to learn and many nuanced choices, which can be overwhelming without direction. To help streamline this process, we've identified some of the most critical areas to focus on when starting your AI product strategy. They include:

- **Choosing your AI model**
- **Building your AI instances**
- **Compliance and security**
- **Developing a data strategy**
- **Choosing your AI partners**

**NEXT, WE DIVE DEEPER INTO THESE AREAS.**

# Choosing Your AI Model

Building your app on a custom or pre-trained AI model is one of the first significant decisions. Often, a company’s size, resources, industry, and product goals impact what type of model they choose.

The market offers diverse options for AI models, ranging from OpenAI's GPT models to more specialized or situationally trained models. Some organizations prefer a pre-trained model as a foundational solution, as they offer a convenient starting point, reduce needed training, and can be more cost-effective for many businesses. However, the size of these models corresponds to their capacity to address various use cases, and choosing a larger foundation model may lead to higher operating costs due to increased demand for accelerated compute instances during inference.

Building a custom AI model entails a significant investment in data, expertise, and computational resources, which is only feasible for some growing companies. However, if your use case requires a high level of customization or unique features, building a custom model may be the better option in the long run despite the initial investment.

When choosing an AI model, consider factors such as your organization's goals and the complexity of the problem you're trying to solve. You’ll need to assess the model's capabilities, computational requirements, operating costs, and how well it integrates with your existing infrastructure and data sources.

# Where to Build Your AI Instances

Another major decision is where and how you intend to build, train, and test your product– locally or in the cloud. You can take a hybrid approach based on the phase, but your strategy should be intentional.

One approach is to conduct the entire development cycle within a unified cloud environment. Cloud providers offer integration with various services required for product development, including data storage, model training, deployment, collaborative tools, and monitoring. This streamlined process eliminates the need for additional data transfers and accelerates the development cycle, providing immediate value by leveraging the cloud's inherent capabilities.

Conversely, developing and training a language model locally, without cloud infrastructure, simplifies stitching components together and involves fewer permissions. Another benefit is greater control over the development environment and data privacy. However, you may encounter challenges such as database issues and interaction complexities, and achieving deep integration with your cloud tools upon deployment may take time and effort. Another issue is the limited scalability of local infrastructure, which may hinder your ability to manage the workload as it grows.

A hybrid approach could be building and training products locally and testing them in the cloud. This takes advantage of a local environment's control and privacy while connecting to the cloud for scalability and access to a broader range of testing environments.

Your choice should align closely with your financial and technological resources, as where you host significantly shapes infrastructure requirements, including the need for specialized hardware, such as GPUs or TPUs, to accelerate training and inference processes. In addition, the AI model you choose will impact where you run your instances. Pre-trained models typically require less computational resources and expertise, making them well-suited for cloud environments. In contrast, custom models may benefit from the power of specialized hardware and on-premises infrastructure.

# Considering Compliance and Cyber Security

AI technology is changing fast, and navigating the complex landscape of data security, ethics, and cybersecurity is imperative. Build your solution carefully, and be mindful of shifting regulations.

However, also be aware that full compliance is a moving target. Governmental and regulatory bodies struggle to keep up with AI's evolution or instate meaningful protective measures. Ill-intended players are moving even faster, and AI is already used in questionable ways in public and business arenas. It's critical to be diligent and stay informed on the changing space and how experts recommend you handle security issues.

Compliance considerations include aligning with data protection regulations, ensuring ethical AI use through transparent guidelines, and prioritizing user privacy with robust consent mechanisms. Here are some helpful measures you can take to ensure your products are as secure as possible.

- Encrypting data both in transit and at rest, implementing strict access controls, and securing APIs with authentication mechanisms.
- Regular security audits, vulnerability assessments, and a well-defined incident response plan are critical for identifying and mitigating potential threats.
- Continuous employee training, monitoring, and logging mechanisms contribute to a security-conscious culture
- Collaborating with cloud service providers that stay informed about emerging risks around building and launching AI products in the cloud
- Regular reassessment and refinement of security measures to adapt to the dynamic cybersecurity landscape and regulatory change
- Regular reassessment and refinement of security measures to adapt to the dynamic cybersecurity landscape and regulatory change

# Forming Your Data Strategy

Because your AI models will contain an immense amount of data and require major computation power, you must establish a thoughtful data management strategy to use your product sustainably. This means cleaning it up, structuring it correctly, and using clever tools to reduce the strain on memory and processing capabilities.

The first step in a successful data strategy is ensuring that data is cleaned and prepared correctly. If the underlying system data fed into an AI model is inaccurate, outdated, or poorly structured, the outputs will be flawed, broken, or ineffective. Preparing your data correctly entails removing any duplicate or irrelevant data, ensuring consistency in formatting and labeling, and handling missing values appropriately.

Next, successfully using large datasets involves breaking them into smaller, more manageable chunks. It entails properly sizing, formatting, and segmenting datasets, which helps improve storage and retrieval processes and addresses model size limitations. Retrieval Augmented Generation (RAG) is one method for doing this, as it uses metadata with segmented chunks to provide context, and the metadata guides models to access relevant information efficiently, enhancing the overall performance of the AI system. In all, data formatting, sizing, and segmentation mitigate errors, minimizing the occurrence of "hallucinations" in model predictions.

# The [ Essential ] AI Partners

## Choosing a Cloud Provider

When choosing a cloud service provider, there are a lot of options and many aspects to evaluate. These tools are pivotal in ensuring the smooth operation, security, and reliability of AI applications. From managing data pipelines to monitoring and securing AI models, the suite of services offered by cloud providers contributes significantly to the overall success of an AI project.

Some baseline offerings you want to look for from a cloud provider include:

- **Scalability:** Ensure the cloud provider can scale resources (such as computing power and storage) based on demand to handle AI workloads efficiently.
- **Performance:** Look for providers that offer high-performance computing options, such as GPUs or TPUs, to accelerate AI model training and inference.
- **Data Management:** Ensure the provider offers robust data management capabilities, including data storage, processing, and integration services, to manage AI datasets effectively.
- **Security:** Look for providers that offer comprehensive security features, such as encryption, access controls, and threat detection, to protect AI applications and data.
- **Compliance:** Ensure the provider complies with relevant data protection regulations (e.g., GDPR, CCPA) and offers tools to help you maintain compliance in your AI projects.

- **Monitoring and Logging:** Look for providers that offer tools for monitoring AI applications and logging events, allowing you to track performance and troubleshoot issues.

- **Cost Management:** Choose a provider with transparent pricing and cost management tools to help you optimize AI project costs.

- **AI Services:** Look for providers that offer AI-specific services, such as pre-trained models, machine learning frameworks, and APIs, to accelerate AI development.

# Using an MSP

When launching an AI product in the cloud, a partnership with an MSP becomes a strategic advantage, enabling organizations to harness the full potential of AI in the cloud while navigating the intricacies of cloud infrastructure and services. Benefits include:

- **AI Model Selection:** MSPs contribute significantly to AI model selection by providing insights into better-performing models, presenting benchmarks, and guiding implementation on cloud platforms.

- **Data Management:** MSPs can help you create a robust data strategy, including assisting with setting up a retrieval augmented generation pipeline and optimizing vector databases. Their expertise ensures efficient handling of complexities associated with data management in the cloud.

- **Integration Assistance:** Leveraging established partnerships with cloud service providers, they assist in integrating existing resources with other offerings, simplifying the process for clients.

- **Ongoing Maintenance:** In the daily operations of AI products, MSPs troubleshoot issues with their specialized cloud provider knowledge, promptly identifying solutions or directing clients to specialized support when needed.

- **Cost Optimization:** MSPs optimize costs by analyzing cloud usage, recommending cost-effective services, and implementing savings plans, ensuring efficient utilization of resources while running specific instances.

# [Conclusion]

You aren't looking for a cloud provider that will build an AI product for you. Instead, you are working to optimize and manage your cloud environment so that your AI tool can work excellently and truly help you achieve your business objectives. Doing the proper prep work and making good decisions during development is essential, as is aligning with a suitable cloud provider. Partnering with an MSP is the final piece in securing your future with this exciting new technology.

# About [Defiance]

Founded in 2020 out of Defiance Ventures, Defiance Digital is an AWS managed services provider offering pay-as-you-grow cloud services and consulting for small and medium businesses. We focus on delivering personalized support and exceptional results through direct access to elite cloud engineers who embrace our 'customers as co-workers' ethos. Our mission is to maximize cloud benefits while minimizing complexity and costs, allowing our clients to focus on their core business.

Our team of cloud experts offers end-to-end support, from strategy to execution, providing our clients with reliable, secure, and scalable solutions tailored to their unique needs. We foster strong relationships with AWS, Datadog, Lacework, Clumio, and other strategic partners to provide the best-of-breed security, observability, automation, and public cloud solutions. We operate with transparency, thoughtfulness, proactivity, and agility and constantly evolve to remain valuable partners for our scaling customers.

## We'll Take It From Here


**WEBSITE**

DEFIANCEDIGITAL.COM

**EMAIL**

SALES@DEFIANCEDIGITAL.COM

**ADDRESS**

201 S COLLEGE ST - 1590,
CHARLOTTE, NC 28202