Infomerica™
The Right Mind Set

Getting Started with Generative AI on AWS

EBOOK

aws PARTNER Select Tier Services

Msp partner
SaaS partner
Training partner
Marketplace seller
Network competency

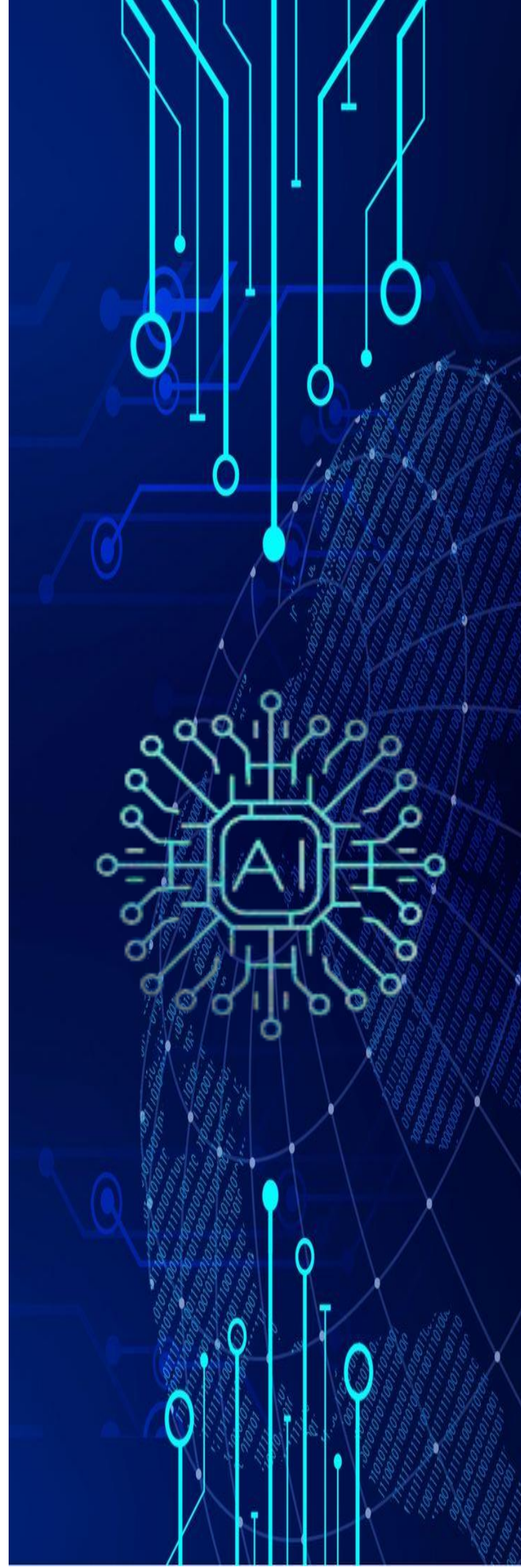
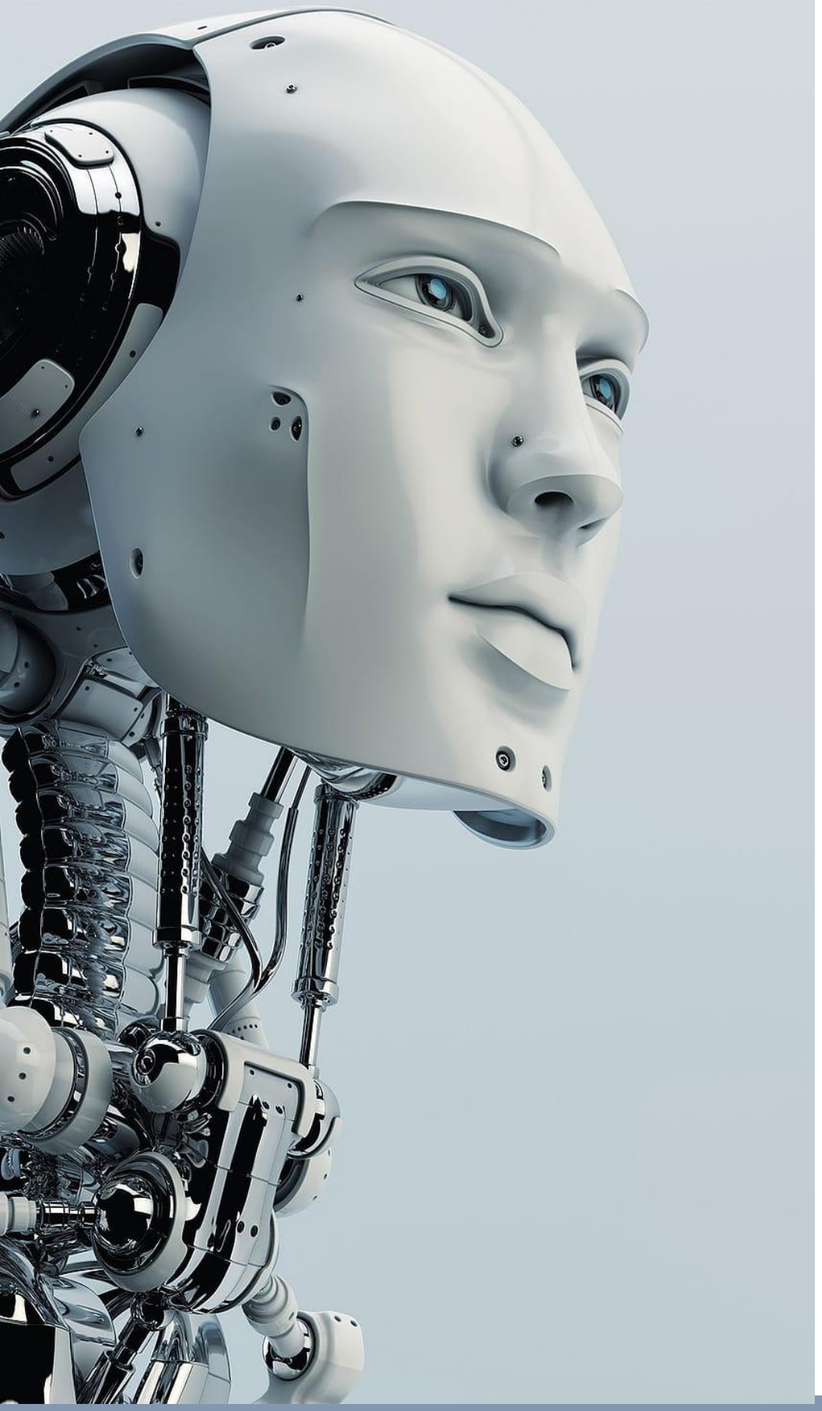


Table of Contents



Introduction

The introduction further emphasizes the significance of Infomerica's Generative AI in harnessing the power of foundation models (FMs). With a focus on customization without starting from scratch, Infomerica's Generative AI offers a unique approach to leveraging the capabilities of LLMs and multi-modal models.

The integration of this technology with AWS provides users with a seamless and secure environment for building distinctive solutions. Infomerica's Generative AI plays a pivotal role in the pre-training phase, ensuring that the resulting FMs excel in capturing nuanced context across diverse datasets, contributing to their impressive performance across various tasks and domains.

The introduction invites readers to explore a comprehensive e-book that not only provides insights into the FM landscape, opportunities, and risks associated with LLMs but also offers practical guidance on utilizing these advanced models using AWS technologies, with a specific focus on Infomerica's Generative AI. This collaborative synergy between Infomerica and AWS underscores the potential for innovation and the creation of tailored solutions in the ever-evolving field of generative artificial intelligence.

Selecting and Customizing Foundational Models

Zero-Shot Learning:
Zero-shot learning refers to the capability of FMs to be easily interacted with by both machine learning (ML) practitioners and non-ML experts. Infomerica, offers accessible interfaces like web playgrounds or chat interfaces such as ChatGPT. Examples include tasks like listing action items from a meeting transcript or translating a document into another language.

In-Context Learning:
In in-context learning approach, users, whether they are developers or ML practitioners, can improve the relevance of FM outputs by incorporating specific examples in their input prompts. This allows the model to adapt and generate outputs that align more closely with the provided context.

Fine-Tuning:
FMs can be further customized for specific tasks through a process known as fine-tuning. ML practitioners can train the FM using a small number of labeled examples specific to their use case. As an example, a professional recruiting firm could customize an FM to automatically process and summarize incoming resumes at scale by fine-tuning the model with a few examples of candidate resumes.

"AI foundation models are the backbone of various AI applications, but their true potential is unlocked through customization"

Issues to Consider with Foundational Models

RESPONSIBLE AI

Infomerica knows that using big AI models can bring up problems like accuracy, fairness, and privacy issues because these models are huge and trained with lots of data. We are aware of concerns, like the models creating offensive content or breaking intellectual property rules. Infomerica is on it, working on solutions such as educating users, filtering content, and using technical tools like watermarking and differential privacy .our aim is to make sure AI is not just powerful but also responsible and safe for everyone..

Hallucinations

Infomerica recognizes that large language models can make mistakes known as "hallucinations," where they generate inaccurate responses inconsistent with their training data. These errors occur because the models have difficulty distinguishing between different numbers or names and may invent information to meet the expected output. For example, they might create citations or author names when providing evidence. Infomerica is aware of these issues and is likely working on improving the accuracy of these models.

Costs

Infomerica's costs involve training models periodically, but during live applications, the model makes constant predictions (inferences) potentially reaching millions per hour. Real-time predictions require fast and efficient networking. Customizing models for specific uses can create smaller, more accurate models that scale efficiently, possibly reducing operational costs.

The Opportunities Ahead with Generative AI

Infomerica's Generative AI holds transformative potential for the global economy. According to Goldman Sachs, it could contribute to a substantial 7% increase in global GDP, amounting to nearly $7 trillion, and boost productivity growth by 1.5 percentage points over a decade. The growth is primarily attributed to investments in generative AI cloud services, projected by Bloomberg to surpass $109 billion by 2030, indicating a remarkable Compound Annual Growth Rate (CAGR) of 34.6% from 2022 to 2030.

Infomerica, in collaboration with AWS, has been pivotal in democratizing machine learning (ML), ensuring accessibility for a diverse clientele of over 100,000 customers across various industries and scales. Notable entities such as Intuit, Thomson Reuters, AstraZeneca, Ferrari, Bundesliga, 3M, BMW, along with numerous startups, and government agencies globally, are leveraging AWS capabilities facilitated by Infomerica. This collaboration has empowered organizations to undergo transformative journeys in their industries and missions, utilizing ML through AWS infrastructure, managed services, and access to a range of Foundation Models (FMs).

Generative AI will play a transformational role in industries like

- Healthcare and Life Sciences
- Financial Services
- Media and Entertainment
- Education
- Automotive and Manufacturing

Build with Generative AI on AWS

“Today, many AWS customers are seeing an impact from generative AI. Here are the top reasons why customers choose AWS to build generative AI applications"

Innovate with generative AI

With enterprise grade security and privacy, a choice of leading foundation models, a data first approach, and the most performant, low-cost infrastructure, organizations trust AWS to deliver generative AI fueled innovation at every layer of the technology stack.

Securely build and scale generative AI applications

Customers trust AWS to build generative AI services and capabilities responsibly and securely. AWS also offers the most comprehensive set of services, tooling, and expertise to help you protect your data, so it remains secure and private when you customize and fine tune foundation models.

The most performant, low-cost infrastructure

Train your own models and run interference at scale.
With AWS, you get the most performant and low-cost infrastructure for generative AI and the broadest choice of accelerators in the cloud.

Data as your differentiator

With AWS, it’s easy to use your organization’s data as a strategic asset to customize foundation models and build more differentiated experiences. Securely customize a foundation model on AWS with your data and built generative AI applications that truly know your business and customers.

About Infomerica

Unveil the Future with Our AWS Services: A Symphony of Innovation, Scalability, and Excellence

Welcome to Infomerica.Inc, where your visions meet execution. As an AWS Select Consulting Partner, we are adept at molding the sophisticated architecture of AWS services to offer transformative solutions, designed meticulously to suit your distinctive business paradigms.

Navigating the Digital Elysium with AWS

Delve deeper into the transformative power of AWS with us! Our spectrum of AWS services is broad and versatile, covering a myriad of domains such as Cloud Migration, Advanced Analytics, DevOps, and Security Compliance. Our focal point is to infuse innovation, fortify security, and instigate scalable solutions that bolster your operational fortitude and business growth.

Innovate, Transform, and Excel with AWS

Engage with us to elevate your journey in the cloud. We stand by you from conceptualization to realization, delivering a seamless experience and ensuring your enterprise’s strategic alignment with the evolving digital epoch. Embrace the future with unparalleled innovation, agility, and resilience with Infomerica.Inc as your beacon in the AWS landscape...

Benefits of building generative AI applications on AWS with Infomerica

Scalability: AWS offers scalable infrastructure, allowing your generative AI applications to handle varying workloads and scale resources as needed. Being part of the AWS ecosystem provides access to a supportive community of developers and resources, ensuring that your team has the necessary support and information

Diverse AI Services: Access a wide range of AI services on AWS, including machine learning frameworks, pre-trained models, and tools that enhance the capabilities of generative AI applications. AWS has a global network of data centers, ensuring low-latency access and global deployment options for your generative AI applications

Security and Compliance: AWS prioritizes security, providing features like encryption, access controls, and compliance certifications, ensuring a secure environment for your AI applications and sensitive data. AWS's pay-as-you-go model enables cost optimization, allowing you to pay only for the resources you use and scale costs based on demand.

Managed Services: Leverage managed services on AWS to simplify infrastructure management, allowing your team to focus on developing and improving generative AI models. AWS integrates seamlessly with various services and tools, facilitating the integration of generative AI applications into existing workflows and systems

Continuous Innovation: AWS regularly introduces new services and updates existing ones, allowing your generative AI applications to leverage the latest advancements and innovations in the field. Collaborating with an AWS Partner brings additional expertise in building and deploying generative AI applications, ensuring that your project benefits from best practices and industry knowledge.

CASE STUDY

Solution

Developed a self-service platform fortified with advanced technologies including Large Language Models (LLM) and Generative AI (GenAI). This comprehensive system serves as a dynamic data injection pipeline. It acquires diverse datasets, conducts pre-processing tasks such as standardizing columns and taxonomies, and leverages LLT for enhanced language understanding. Additionally, it merges relevant datasets using association rules and fills in missing data utilizing supplemental datasets like census demographic data. Through the integration of LLM and GenAI, this tool not only aggregates and standardizes multiple datasets but also generates actionable insights and recommendations. Ultimately, it culminates in the creation of an interactive dashboard, providing users with intuitive access to comprehensive and insightful data analytics.

Benefits

- Data ingestion
- Interactive dashboard.
- AI-based rules engine
- Report Automation & NLP (ML) Search-Based Analytics

Cutting Edge Technologies

Kubernetes, Docker, Data Analytics , elk stack, Python, Golang, Gen AI

Business problem

"Examining a dataset of emergency calls and extracting pertinent information to uncover concealed trends and patterns can significantly enhance the preparedness of the emergency response team, enabling them to more effectively address emergencies”

CASE STUDY

Solution

We moved its systems and data to AWS cloud gradually. And we also used AWS services like Amazon Sage Maker for building and running machine learning models, Amazon EC2 virtual servers for AI workloads, Amazon Redshift for big data analysis, Amazon Aurora for scalable databases, and Amazon VPC for secure resource isolation. We leverage AWS auto-scaling to automatically adjust resources based on demand, and optimize its AI algorithms to run faster on AWS's powerful computing instances. For security, we use AWS access management and encryption services, following relevant data protection regulations. To save costs, we monitor and optimize spending using AWS cost tools, adopt pay-as-you-go pricing and reserved instances, and explore serverless computing with AWS Lambda. Overall, moving to AWS would enable Infomerica to support its AI applications efficiently, process data at scale, ensure scalability and security, and optimize costs flexibly.

Benefits

- Advanced AI Capabilities
- **Scalability and Performance.**
- **Security and Compliance**
- **Cost Efficiency & Innovation and Growth.**

Cutting Edge Technologies

Amazon [Sage Maker, EC2, Redshift, Aurora, VPC], AWS Identity and Access Management (IAM), AWS Key Management Service (KMS)

Business problem

"To harness the power of Amazon Web Services (AWS) to support Gen AI's AI-driven applications, enhance data processing capabilities, ensure scalability, and maintain high levels of security and compliance"

Infomerica™

The Right Mind Set

Learn more

https://www.infomericainc.com/AWS-Services

aws PARTNER Select Tier Services

- Msp partner
- SaaS partner
- Training partner
- Marketplace seller
- Network competency