aws

The generative AI advantage:

A leader's guide to using
your data as a differentiator

CONTENT

A leader's guide to using your data as a differentiator

The world has been captivated by generative artificial intelligence (AI) and its potential to transform how we communicate, create, and operate. In large part, the fascination with generative AI has been spurred by consumer-facing applications, such as ChatGPT and YouChat. These groundbreaking chatbots have revealed an awe-inspiring ability to mimic human creativity and conversation.

But the true strength of generative AI goes beyond general-knowledge chatbots. Organizations are just beginning to peel back the layers and uncover ways generative AI can help them innovate. They are eager to embrace these possibilities—and for good reason. According to Goldman Sachs, generative AI could raise the global gross domestic product (GDP) by almost $7 trillion and increase productivity growth by 1.5 percentage points over a 10-year period.

Given these findings, it is no surprise that business and data leaders want to move quickly with their own generative AI applications. They want to know not only how to take the next best step, but also how to gain a competitive advantage in this emerging space. The key to unlocking the potential of generative AI is your organization's own data.

Your data is your differentiator.

A brief primer

At a high level, generative AI can be defined as a type of AI used to produce new content and ideas. Generative AI applications can, for example, write stories, generate code, and design digital images. These applications also make it possible to automate cumbersome tasks, such as taking lengthy documents and condensing them into brief summaries.

Artificial intelligence

Machine learning

Neural networks

Generative models
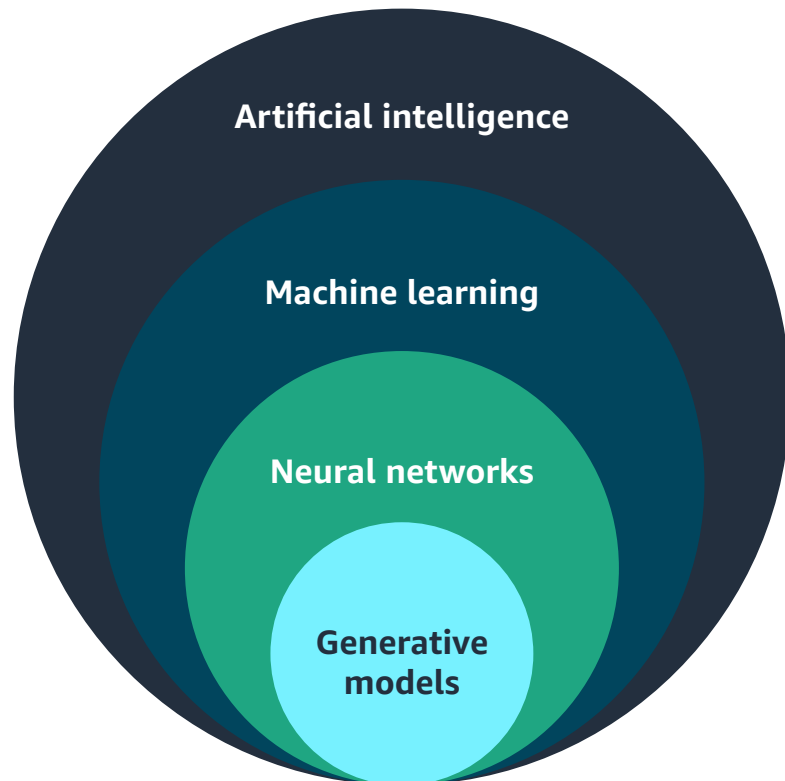
While this list is far from exhaustive, we commonly see our customers use generative AI applications to:

- Improve the customer experience through capabilities such as chatbots, virtual assistants, intelligent contact centers, personalization, or content moderation

- Boost employee productivity with conversational search, content creation, text summarization, or code creation

- Turbocharge production for all types of creative content like art, music, or animation

- Streamline business operations with intelligent document processing, predictive maintenance, quality control and visual inspection, or data augmentation

Like all AI, generative AI is powered by machine learning (ML) models—very large models that are pretrained on massive amounts of data. These models are commonly referred to as foundation models (FMs).

It's important to note that at its core, a FM leverages the latest advances in ML. A class of FMs, such as the generative pretrained transformer (GPT) models, are commonly referred to as large language models (LLMs) and are specifically focused on language-based tasks, such as summarization, text generation, and open-ended Q and A. LLMs are special because they contain a large number of parameters that make them capable of learning advanced concepts.

Put data at the center of your generative AI approach

We've long known that data is a strategic asset to businesses. Yet, according to an Accenture study, only one in three, or 32 percent, of companies are able to realize tangible and measurable value from data—even after investing in data infrastructure.

Generative AI can help move these numbers in the right direction. Generative AI presents an opportunity for you to tap into your data in a new way and get more value from it. The technology enables you to innovate on top of your data more quickly, leverage this data in new types of applications, and unlock the value of data that has traditionally been hard to work with, like unstructured data.

We're already seeing some of our customers combine data with generative AI to improve business outcomes and customer experiences. For instance, Intuit built Intuit Assist, a new generative AI–powered assistant that uses relevant contextual datasets spanning small business, consumer finance, and tax to deliver personalized financial insights to customers. And United Airlines has developed generative AI applications to help employees improve efficiency by providing summarized updates on flight status and issues that could lead to delays by pulling together data across the operations.

These applications are exciting and represent perhaps just a fraction of what generative AI can deliver for companies. The ways in which generative AI will change our world are still coming into focus. You may be grappling with how to capture this enormous potential when it feels as though we're entering uncharted territory. In reality, the path to realizing business value with generative AI is not much different than it is with any other technology. It comes down to the strength of your data strategy and how you leverage your data as a differentiator within that strategy.

You may already have a data strategy in place or are just starting to build one. In either case, there's never been a better time than now to bring this strategy together in a way that allows you to turbocharge your business value with generative AI. You have an unprecedented opportunity to gain a sustainable competitive advantage by differentiating with your data.

In this whitepaper, we provide business and data leaders with insights and next steps for using data to create generative AI applications that are unique to their organizations. To innovate and compete in this arena, you need to develop a broad data strategy that includes technology, as well as your business priorities and use cases, your employees, and your governance guardrails. Taken together, this strategy represents a modern view of data that ensures you can realize business value from your generative AI applications.

We'll focus on three areas to help you create this modern data strategy:

1 Turning your data into a differentiator for your generative AI applications

2 Establishing the right data foundation to unlock the value of your current data through generative AI

3 Thinking beyond technology to create a competitive advantage with generative AI

Turn data into a differentiator for generative AI applications

Your data is the difference between having a generic application and one that knows your company and customers deeply. Because of this, you'll have to determine how to best use your data to capture and showcase your company's uniqueness. For most companies, the starting point for deploying generative AI applications is an out-of-the-box foundation model (FM). A small number of companies will choose to build their own FMs to power their generative AI applications, but that requires extensive computing resources and highly specialized staff.

While FMs are powerful out of the box, they're generalized by design. They are—as the name implies—foundational. This means they are not tuned to your business needs because they can't access your most recent company data or perform domain-specific tasks to fulfill user requests. Your own data is key to aligning generative AI applications to your customer experience, internal knowledge, brand voice, and ethical parameters.

For instance, if you are an online travel agency that wants to provide better travel recommendations to your customers through a generative AI application, you would likely want to leverage data that is specific to your individual customers, such as past trips, web histories, and travel preferences. You would also want to access aggregate data on similar traveler patterns and trip inventories to create a better recommendation. By using your own data, you create a personalized and unique customer experience.

Also, out-of-the-box FMs are widely available, so customizing them with your own data allows you to differentiate your generative AI applications. Let's say you are also using an out-of-the-box FM to draft marketing copy for your online travel agency. Your competitors may be using the same model to the same effect. These models largely pull from the same general knowledge pool. So, without customization, you could end up creating content that is nearly identical to theirs and the other way around.

Customization creates a sustainable competitive advantage. You can customize your FM using a few methods, including fine-tuning and in-context learning.
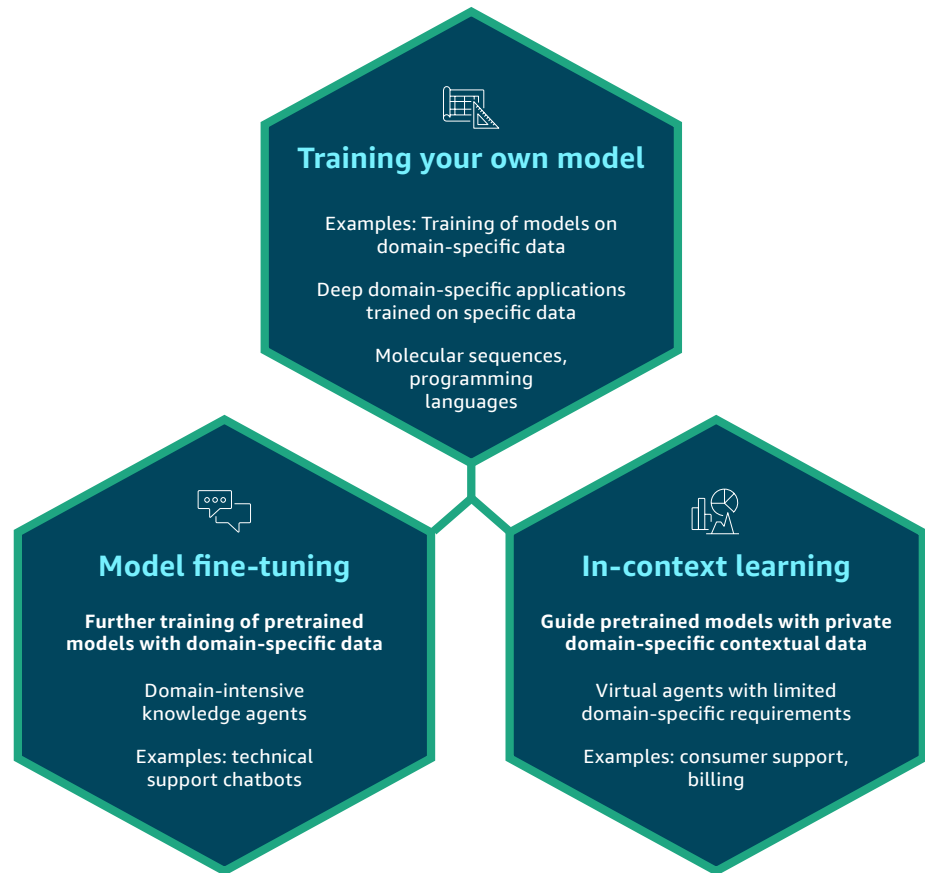
Customer spotlight

INRIX, a global provider of transportation data and analytics, is building a new solution—centered on Amazon Bedrock—that will deliver up-to-date, real-time information to help traffic and safety engineers understand where, when, and why something is happening on our streets—and what to do about it.

Its new Amazon Bedrock solution uses Retrieval Augmented Generation, or RAG, to augment prompts to the underlying FM with historic data like history of speeding incidents and crashes, and recent data like congestion status and current weather conditions.

By using its own data to augment its FM in Amazon Bedrock, INRIX can provide its customers with fast answers to complex questions like how roads should be changed to alleviate congestion and minimize accidents, how to determine the ideal location of a new retail store, or even how to mitigate traffic and parking issues for the next concert.

Emerging generative AI data patterns

Training your own model

Examples: Training of models on domain-specific data

Deep domain-specific applications trained on specific data

Molecular sequences, programming languages

Model fine-tuning

Further training of pretrained models with domain-specific data

Domain-intensive knowledge agents

Examples: technical support chatbots

In-context learning

Guide pretrained models with private domain-specific contextual data

Virtual agents with limited domain-specific requirements

Examples: consumer support, billing

Fine-tuning

With an out-of-the-box FM, you must leverage your data to customize the model for your unique business needs. Fine-tuning is a good option for domain-intensive applications, such as technical support agents or content creation unique to your business. With Amazon Bedrock, you can securely customize an FM with your data and use other built-in tools to build applications that know your business, your data, and your customers.

Imagine a content marketing manager who works at a leading fashion retailer and needs to develop fresh, targeted ad and campaign copy for an upcoming new line of handbags. To do this, they provide Amazon Bedrock a few labeled examples housed in a data lake on Amazon Simple Storage Service (Amazon S3) of their best-performing taglines from past campaigns, along with the associated product descriptions. Amazon Bedrock makes a separate copy of the base model that is accessible only to the customer for model training. After training, Amazon Bedrock automatically generates effective social media, display ads, and web copy for the new handbags.

In-context learning

Foundation models are trained at a moment in time and it is not practical to fine-tune them every time a dataset changes. Once they are trained, they're no longer ingesting new knowledge or data. They are also unable to locate and access real-time information if they need additional context to solve a problem.

To make responses more relevant and contextual, you can ground your FM with data through in-context learning, a technique in which the FM is guided to domain-specific, contextual data either through prompt engineering or Retrieval Augmented Generation (RAG). Many companies will use RAG as their main method to perform in-context learning. RAG enables your FM to access your company's most recent data to provide a more accurate, more relevant response. Often, RAG uses vector embeddings, or numerical representations of words, phrases, or images. Embeddings encode the semantic meaning of the source text or images so that FMs can more easily find relationships between similar vectors and improve responses to prompts.

While you can use each of these techniques separately, the combination of fine-tuning and RAG can help you turn your data into a differentiator for your generative AI applications.
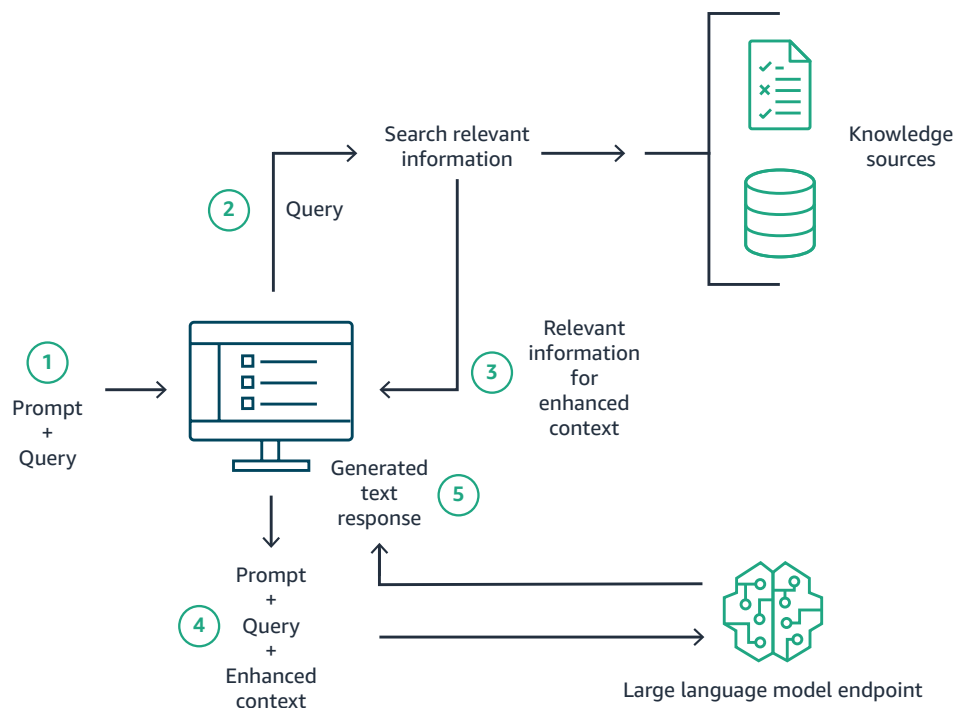
Establish a data foundation for generative AI

Your data is key for creating value with generative AI applications. It then becomes crucial to customize your models with high-quality, relevant, readily accessible, and available-to-use data. You meet these benchmarks by first having a strong data foundation. This foundation includes a comprehensive, integrated set of data services for all workloads, use cases, and types of data, as well as tools to govern that data. The following is a high-level overview of what this data foundation looks like:

Comprehensive

For generative AI, you need to store various types of data, including unstructured, structured, streaming, and vector data, that can be used for building and customizing models and adding context to prompts with or without RAG. A comprehensive set of data services makes it possible to store all of this data as well as query and analyze it at scale.

Typically, a comprehensive set of data services for generative AI includes a highly durable and scalable data lake. This data lake stores the domain-specific data you need to build and customize your FMs. AWS has been helping customers build a strong foundation for data lakes to store structured and unstructured data with services like Amazon S3, AWS Glue, and AWS Lake Formation for years. Customers use Amazon S3 to create hundreds of thousands of data lakes.

A data foundation for generative AI also includes high-performance knowledge stores for RAG. AWS provides several options depending on your use case. For instance, NoSQL databases store conversation state and history, so a chatbot can remember prior responses. Transactional databases store context and customer information to create more personalized responses. You can also use a knowledge store like Amazon Kendra that connects to multiple structured and unstructured content repositories, providing a document-based knowledge source for your FMs. Or databases that have vector search capabilities, specifically designed to be efficient at storing and retrieving embeddings. Being able to use vector search within the databases you already use has several advantages. For instance, it eliminates the steep learning curve for new programming tools, APIs, and SDKs. You can also feel confident knowing that your existing databases are proven in production and meet requirements for scalability, availability, storage, and compute. And when your vectors and business data are stored in the same place, your applications can run faster—and there's no data sync or data movement to worry about. AWS offers vector search capabilities for many of our popular data stores to give customers even more flexibility as they build their generative AI applications.

A comprehensive data foundation accounts for both data analytics and data storage. You can use data warehouses to customize your FMs and use cases that require up-to-date operational data, such as building an FM or other LLM to provide insights on business data through natural language queries. Amazon Redshift is a fast, petabyte-scale data warehouse delivering up to six times better price performance than other cloud data warehouses.[1] Amazon Redshift integrates with many of your data sources, including Amazon S3 and Amazon Aurora, so you can get a more complete look at your data.

> **"Companies that have not yet found ways to effectively harmonize and provide ready access to their data will be unable to fine-tune generative AI to unlock more of its potentially transformative uses."**
>
> What every CEO should know about generative AI, McKinsey Digital, 2023

Integrated

Data integration gives you a complete view of your business and ensures your data is readily accessible for your generative AI applications. With direct integrations between AWS services, we're reducing and eliminating the extract, transform, and load process (ETL) for common use cases so teams can move faster. You can also use AWS Glue, our serverless and scalable ETL and data integration service that makes it easier to discover, prepare, move, and integrate data from multiple sources for analytics and machine learning. AWS connects to hundreds of data sources, including software as a service, on premises, and other clouds, as well as third-party data from more than 300 data providers.

Secure and governed

Your data needs to be secure and governed throughout the lifecycle of building a generative AI application. AWS offers several tools to help ensure data quality, privacy, and access controls, so the data you use for generative AI applications is high-quality and compliant.

With Amazon Bedrock, all your data is encrypted at rest using your own key management service (KMS) keys, which provide full control and visibility into how you store and access your data and custom models. With AWS PrivateLink, you can pass your data on AWS to Amazon Bedrock exclusively through the AWS network and never via public internet. You can privately customize an FM on your own virtual private cloud (VPC), so no data is ever leaked and no data is ever used to train or customize a model that would be available to other companies. Amazon Titan FMs are built to detect and remove harmful content in the data you want to use for customization, reject inappropriate content in the user input, and filter the model's outputs that contain inappropriate content, such as profanity or hate speech.

Earlier in this whitepaper, we also discussed the importance of customizing your FM to reflect your specific brand and customer experience, as well as to root out inaccurate or irrelevant content. While customization is important to address these challenges, so is data governance in the form of human oversight and feedback. You will need people to ensure your output reflects how you want the world to perceive you. For example, reinforcement learning from the human feedback method enables you to train an FM to make decisions and act while receiving guidance from human experts. These experts look for potential complications, such as bias in the data, data quality issues, and data gaps, and help align your FM to your brand voice, corporate guidelines, ethics, and policies.

Think beyond technology to create a competitive advantage

Technology is a fantastic enabler for generative AI, but it's only one component of a strategy to embed data as a strategic asset in your company.

At AWS, we take a modern view of data strategy that is more expansive in nature. An end-to-end data strategy is one that encompasses technology as well as mindset, people, and process. This combination paves the way for your company to become data-driven by weaving data into every aspect of your business and operations.

At AWS, we define a data-driven organization as:

"An organization that harnesses data as an asset, to drive sustained innovation and create actionable insights to supercharge the experience for their customers so they demand more."

Accelerating business value with a modern data strategy, AWS, 2023

Mindset

Mindset refers to the way a company thinks about and treats data. A company's mindset is reflected in the beliefs, values, and behaviors that create their data-driven culture with aligned use cases.

Traditionally, leaders have built their data strategies with the mindset that data is a platform and a means to build solutions. This viewpoint consistently results in a mismatch between IT investment and improved business outcomes.
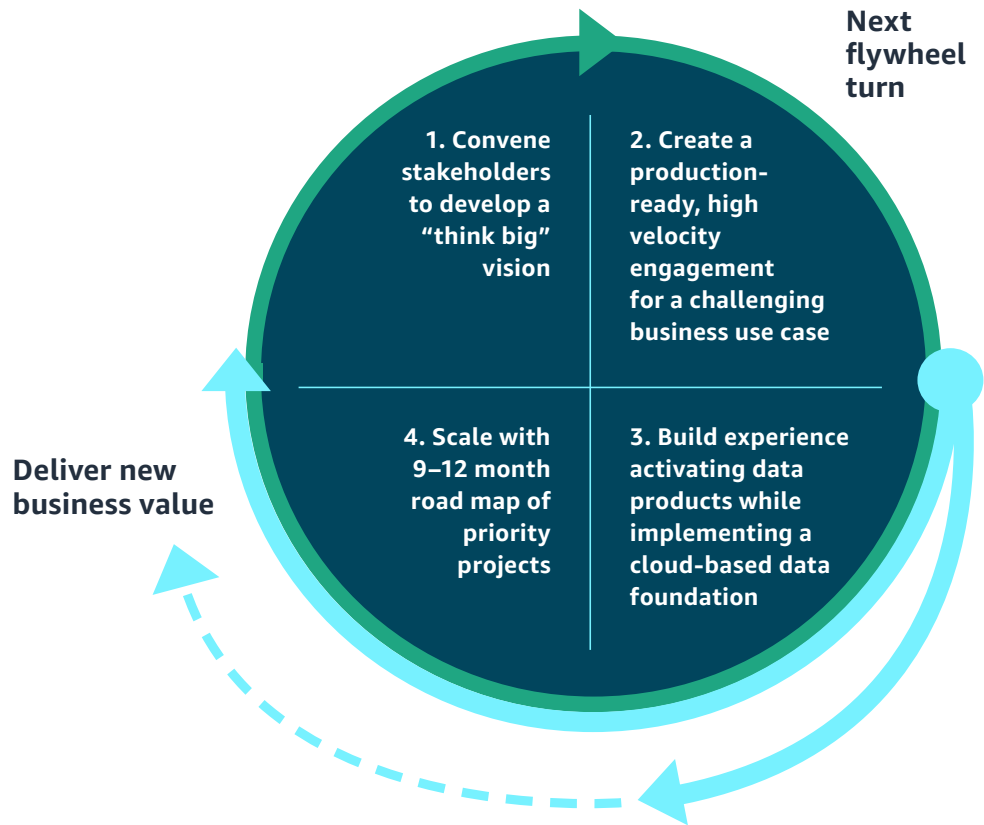
As interest in generative AI grows, we're seeing more of this type of thinking. This trend is understandable as leaders want to do something with generative AI and ensure they can remain competitive. However, you must first explore how these applications can help you solve a problem or differentiate your business. You must approach generative AI as an evolving data product that adds value to your business. In doing so, you put the focus on your customers, not solutions, and close the critical gap between your data initiatives and business results.

The flywheel approach

We help our customers bring data-driven products to life for their use cases through a flywheel approach. This approach accelerates the pace by which you identify a high-value business or customer opportunity and create a data product or experience to address it. Once the flywheel starts spinning, you continue to identify new use cases to deliver more products or experiences.

The flywheel speeds transformation and creates incremental value. We summarize it as "think big, start small, and scale fast" and have illustrated it briefly in the following example:

Next flywheel turn

1. Convene stakeholders to develop a "think big" vision

2. Create a production-ready, high velocity engagement for a challenging business use case

3. Build experience activating data products while implementing a cloud-based data foundation

4. Scale with 9–12 month road map of priority projects

Deliver new business value

NatWest Group, one of the largest banks in the United Kingdom, wanted to deploy personalized solutions at enterprise scale. During a series of workshops with AWS, the bank identified areas of improvement within their ML landscape and created a comprehensive plan for development. The resulting solution was an enterprise-wide, centralized ML workflow powered by Amazon SageMaker.

The workflow equipped the NatWest data science teams with easier access to data and ML tools, giving them the space to build numerous ML use cases in just four months. The use cases included a solution that tailors marketing campaigns to specific customer segments and an application that automates simple fraud detection tasks, so investigators can focus on difficult, high-value use cases. Using a flywheel approach, NatWest was able to reduce its time to value and increase its number of use cases.

Empower your people

Generative AI—and AI in general—can help your teams move faster and more purposefully with data. This shift can ultimately lead your company to realize greater efficiency and productivity, as well as better employee and customer experiences.

For example, Omnicom Group, a leading global marketing and corporate communications company, uses generative AI and ML services from AWS to help accelerate the transformation of its advertising campaign development. Omnicom will advance its AI-powered platform, Omni, by creating new FMs that help automate activities such as developing creative briefs, media plans, ad creative, audience segmentation, and performance measurement.

Philips uses Amazon Bedrock to accelerate the development of cloud-based generative AI applications that provide clinical decision support, help enable more accurate diagnoses, and automate administrative tasks. These new ways of working help simplify radiologists' workflow and reduce cognitive burden and clinician burnout.

Emerging roles in AI and ML

Generative AI Engineer: Specializes in developing and implementing generative AI models and systems, including designing, training, fine-tuning, and optimizing models for specific applications

Data Curator: Responsible for sourcing, collecting, and organizing high-quality datasets that are diverse, representative, and properly labeled

Generative AI Consultant: Provides expert guidance and advice on using generative AI technologies across various industries

Generative AI Artist: Explores the creative possibilities of generative AI technologies to produce unique and innovative art pieces, music compositions, or visual designs

AI Policy and Regulation Specialist: Shapes policies, guidelines, and regulations that govern the responsible development, deployment, and use of generative AI systems

Accenture currently uses Amazon CodeWhisperer, an AI coding companion, to accelerate coding as part of its software engineering best practices initiative in its Velocity platform. The Velocity team was able to reduce their development efforts by 30 percent and is now focusing more on improving security, quality, and performance.

Generative AI can help make it possible for more employees to innovate with data. To prevent roadblocks and bottlenecks, you must make it easier for them to discover, consume, share, and manage data—with appropriate policies. Traditionally, the IT department has held almost complete control over the dissemination of data in an organization. An end-to-end data strategy pushes this responsibility to the edges, giving it to the teams that produce and consume data. We often refer to this model as a modern data community and see customers using it to empower more employees to make data-driven decisions that affect their specific business function.

New skills and roles

Generative AI will open the door for entirely new roles and create greater demand for existing roles. According to the World Economic Forum's *Future of Jobs Report 2023*, the majority of the fastest-growing roles are technology-based and include AI and ML specialists, as well as data analysts, data scientists, and information security analysts.

Currently, however, the talent pool for AI, ML, and cloud expertise is limited. Companies often lack the skilled and diverse workforce to fully adopt their data strategies. To ensure you keep moving ahead, you need to invest in upskilling your current workforce, including those outside of the IT department in teams such as finance or marketing. Training these teams to interact with data removes bottlenecks by giving them access to the right data products at the right time to make business decisions.

Upskilling employees is an important part of your data strategy. These exercises are an investment but aren't the heavy lift some leaders may imagine. Again, thanks to AI/ML, you can more easily equip employees who have all levels of technical skill with the tools to analyze data, uncover insights, and construct narratives. For example, Amazon Q, our new generative AI assistant, helps you in Amazon QuickSight to author dashboards and create compelling visual stories from your dashboard data using natural language. We also announced that Amazon Q can help you create data integration pipelines using natural language. For example, you can ask Q to "read JSON files from S3, join on 'accountid', and load into DynamoDB," and Q will return an end-to-end data integration job to perform this action. With Amazon Q, data analysts, scientists, and engineers can also be more productive using generative AI text-to-SQL functionality in Amazon Redshift to query data in your data warehouse.

Create new processes or optimize existing ones to keep up with innovation

Generative AI has brought the phrase "responsible AI" to the forefront and added a new dimension to data governance. You must still consider issues not only of data sharing, security, and privacy, but also those related to bias, ethics, and false information.

In the previous section, we discussed some of the tools that help mitigate the common risks associated with generative AI. But, as data leaders know, data governance is more about strategy than it is tool selection. Data governance needs to be architected to support wider data and AI strategy, but implemented incrementally based on use cases and business priorities.

Data leaders often tell us they find it challenging to show business value from their data governance initiatives. When they look more broadly at their data strategy and align it with business priorities, however, they move closer to showing this value.

A broad data strategy is important because it ensures data governance policies are in step with organizational structures. To innovate and move quickly with generative AI, your teams need ready access to data. They also need to operate within the confines of security, sharing, and privacy policies. You need to strike a balance between the non-negotiables and autonomy and speed. You can achieve this balance when you view data governance as an enabler of your broader data strategy.

Tap into your data in a new way to deliver more value

We're just beginning to uncover the possibilities for generative AI, but can already sense how it will transform nearly every application, industry, and company. Generative AI offers fresh and exciting ways to tap into your existing data and gain more value from it. It can enable you to unlock new data products and experiences that delight your customers and set your brand apart.

We know that nearly every leader wants to explore the potential of generative AI, and we also know that your next steps will be critical. Rather than use a generic application, you should focus on what makes your company unique and use it to guide your decision making. That uniqueness exists within your data. And to fully use your data as a strategic asset, you need to create a data strategy that encompasses not just your technology but also your mindset, people, and process. Each layer of this strategy firmly cements your data as a differentiator for how you build and customize your generative AI applications and how you empower your entire organization to innovate.

Learn more about building a data foundation to gain a competitive advantage with generative AI

GLOSSARY

Generative AI is a type of AI that can create new content like text, code, images, and video using patterns it has learned by training on extensive public data with ML techniques.

Foundation models (FMs) are deep learning models trained on vast quantities of unstructured, unlabeled data that can be used for a wide range of tasks out of the box or adapted to specific tasks through customization.

Large language models (LLMs) make up a class of foundation models that can process massive amounts of unstructured text and learn the relationships between words or portions of words. This enables LLMs to generate natural language text, performing tasks such as summarization or knowledge extraction.

Retrieval Augmented Generation (RAG) is the process of increasing the contextual knowledge of LLMs and reducing factual errors, or hallucinations, by enabling prompts to be augmented with context and domain-specific data, such as through vector databases.

Customization is the process of adapting a pretrained foundation model to perform better in a specific task via a short period of training on a smaller labeled dataset. This additional training allows the model to learn and adapt to the nuances, terminology, and specific patterns found in the smaller dataset.

Prompt engineering refers to the process of designing, refining, and optimizing input prompts to guide a foundation model toward producing desired (that is, accurate) outputs.