aws FOR FINANCIAL SERVICES

Embracing AI-Driven Innovation:

AWS Generative AI and AI/ML for Capital Markets

Notices

Customers are responsible for making their own independent assessment of the information in this document. This document: (a) is for informational purposes only, (b) represents current AWS product offerings and practices, which are subject to change without notice, and (c) does not create any commitments or assurances from AWS and its affiliates, suppliers or licensors. AWS products or services are provided “as is” without warranties, representations, or conditions of any kind, whether express or implied. The responsibilities and liabilities of AWS to its customers are controlled by AWS agreements, and this document is not part of, nor does it modify, any agreement between AWS and its customers.

Aug 2024

Contents

Forward

> “The capital markets industry is in a transformative era. The convergence of technological innovation and evolving client expectations is reshaping how firms operate. The integration of generative artificial intelligence (AI) and AWS technologies has become a catalyst for unlocking new areas of operational efficiency, personalization of content, new analytics, and streamlining knowledge management workflows. This re-invention of legacy processes and capabilities with generative AI on AWS is becoming a competitive advantage and helping firms attract, engage, and retain customers. Today, it is crucial for our customers to chart a path towards the future where data-driven insights, streamlined infrastructure, and innovative offerings redefine the capital markets experience for their customers."

Brian Cassin–Global Head of Market Development, Capital Markets, AWS

The capital markets landscape is undergoing a transformative shift, driven by evolving client demands, heightened regulatory scrutiny, and the imperative to drive operational efficiencies. Capital markets have been leading the charge for AI/ML adoption, and generative AI applications such as ChatGPT have prompted them to explore this transformative technology with even faster pace and innovate rapidly. The innovative force of generative AI has permeated the capital markets, prompting organizations to explore ways to leverage this transformative technology.

Generative AI is a type of AI that can create new content and ideas, including conversations, stories, images, videos, and music. Like AI, generative AI is powered by machine learning (ML) models—known as large language models (LLMs) that are pre-trained on vast amounts of data and also commonly referred to as foundation models (FMs). Generative AI holds the potential to boost workforce productivity, catalyze innovation, understand market and customer sentiments, and build new products across the trade lifecycle. It can create additional revenue streams through the development of innovative products and services underpinned by data-driven insights. Cloud technology is poised to be a pivotal enabler in the successful adoption of generative AI within the capital markets realm. At Amazon Web Services (AWS), we are actively collaborating with our customers to navigate this rapidly evolving landscape.

This whitepaper is meant to help technical and business decision makers, and will explore the diverse participants and personas that shape the capital markets ecosystem, from the buy-side and sell-side, to industry utilities and financial information providers. The section on capital markets participants and personas provides an overview of the capital markets for those less familiar. For those already knowledgeable about capital markets, the value chain section begins the generative AI use cases along with traditional AI/ML. These use cases hold the potential to revolutionize capital markets across the entire trade lifecycle. The paper also outlines comprehensive AWS services, strategy and offerings tailored to addressing the unique needs of the capital markets customers in their generative AI journey. It highlights the pivotal role of cloud technology as an enabler for the successful adoption of generative AI.

Capital Markets Participants and Personas

The capital markets industry is the business of issuing, pricing, buying, and selling financial products such as securities and derivatives. It provides services and platforms across the entire trade lifecycle from the inception of securities trading idea, to its execution, to post trade analytics.

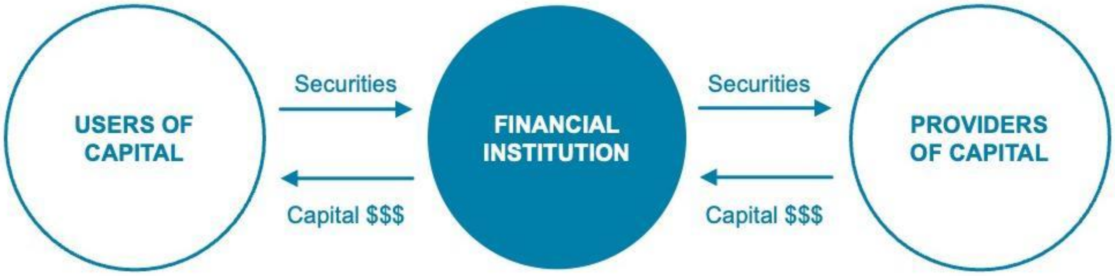


Figure 1: Capital markets participants

The trade between counterparties includes different market participants and a series of functions that together form the capital market value chain. Based on their functions, there are four categories of market participants who are part of capital markets value chain:

1. **Buy-side** refers to organizations like asset managers, wealth managers, and hedge funds that offer investment services to customers ranging from mass affluent to high-net-worth individuals and institutional investors**.** Example organizations include asset managers (such as Vanguard, BlackRock, PIMCO), wealth managers (such as LPL, Edward Jones, Stifel, Saint James Place, Hargreaves Lansdown), and hedge funds (such as Bridgewater, Millennium Partners, Point72).

2. **Sell-side** refers to organizations like investment banks and broker-dealers that provide investment advisory and engage in the business of trading securities of their accounts or on behalf of their customers. Example organizations include Goldman Sachs, JP Morgan, and Citi.

3. **Industry Utilities** refers to organizations like trading exchanges, clearinghouses, and custodians. *Trading exchanges* are the marketplaces to trade financial products (specifically, equities, derivatives, cryptocurrency, and so on). Examples of exchanges include Intercontinental Exchange (ICE), the Chicago Mercantile Exchange (CME), the London Stock Exchange (LSE), and

the NASDAQ. *Clearinghouses* serve as an intermediary between buyers and sellers of financial products and provide services for central margining, reporting, and reconciliation of orders. Examples of clearinghouses include DTCC and Euroclear. *Custodians* safeguard and administer the assets traded in the market. They hold and protect securities, ensure accurate record-keeping, and facilitate the settlement of trades. Examples of custodians include Bank of New York, Fidelity.

4. **Financial Information Providers** refer to data providers of market pricing, tick data, earning reports, and regulatory filings. They offer data infrastructure, analytics services, and technologies for buy-side and sell-side firms. Examples include Thompson Reuters, FactSet and Bloomberg.

The Capital Markets industry is highly regulated, competitive, and operates on tight margins. The capital markets are affected by movement of macroeconomic indicators such as gross domestic product (GDP), growth rate, unemployment rate, inflation rate, consumer confidence index, and the Federal Reserve interest rate. These indicators provide insight into the overall health of the economy, influencing investor sentiment and market behavior. For example, high GDP growth and low unemployment typically signal a strong economy, leading to increased investor confidence and higher stock prices. Conversely, high inflation along with rising interest rates can dampen investor sentiment, leading to market corrections or declines as borrowing becomes more expensive and consumer spending decreases. Events like geopolitics, wars or a pandemic affect the capital market in different degrees.

The global markets have shifted over the last two years, driven by rising inflation, higher interest rates, geopolitical conflicts in Ukraine and the Middle East, and uncertainty regarding a recession. These shifts have created more volatility in the markets and uncertainty in the economy. For different segments, the capital markets identify certain macro themes such as AI, sustainability, environmental, social, and governance (ESG), Internet of Things (IoT), and more. These macro themes play a key role in large allocation of capital and investments from the capital markets investors to participating firms. Investors look to generate returns on their investments over time, as these macro themes are executed and translated into products and services.

Market volatility, while viewed as a challenge, can actually present opportunities within the capital markets. Periods of high volatility, characterized by large price fluctuations in assets and wider performance spreads between and within asset classes, create more trading opportunities (shown in Figure 2), potentially leading to higher returns. However, it is important to note that high or sudden volatility also comes with its own set of risks, such as large drawdowns (significant declines in investment value), margin calls, and counterparty risks (the risk of default on contractual obligations). As a result, effective risk management is crucial to capitalize on the opportunities that market volatility offers. By striking the right balance and implementing appropriate risk management strategies, portfolio and asset managers navigate the dynamic capital markets landscape and potentially benefit from the increased trading opportunities that volatility can provide. Figure 3 illustrates different market scenarios based on various factors.

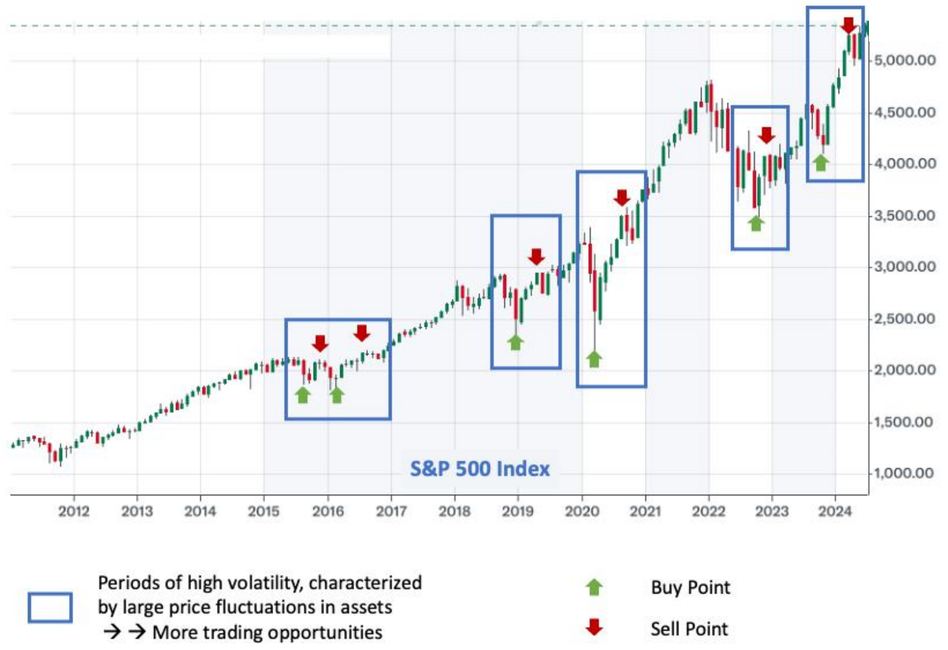
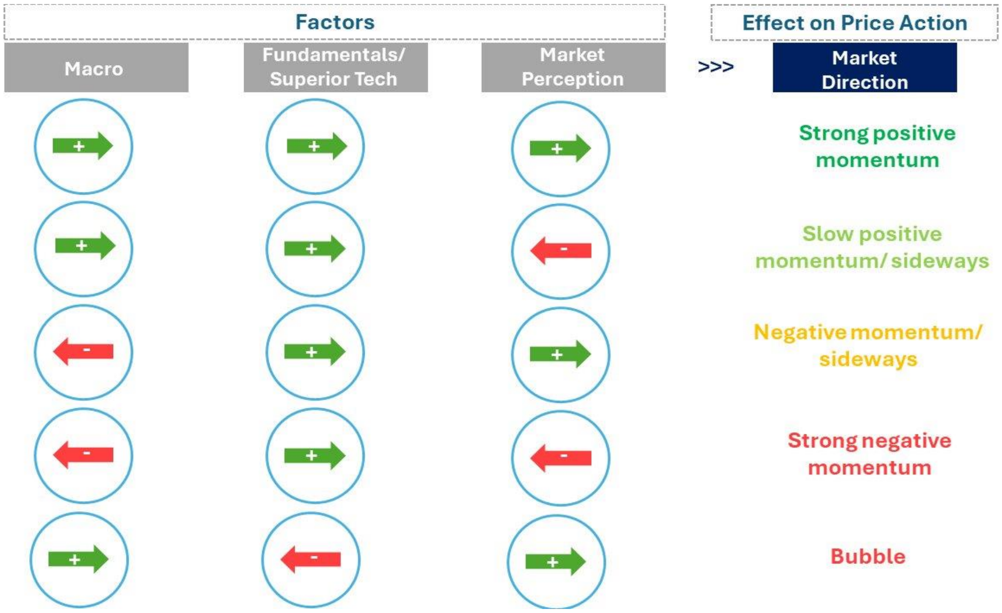


Figure 2: S&P Index - periods of high volatility (illustrative)



Factors				Effect on Price Action
Macro	Fundamentals/ Superior Tech	Market Perception	>>>	Market Direction
+ →	+ →	+ →		Strong positive momentum
+ →	+ →	← -		Slow positive momentum/ sideways
← -	+ →	+ →		Negative momentum/ sideways
← -	+ →	← -		Strong negative momentum
+ →	← -	+ →		Bubble

Figure 3: Different market scenarios based on various factors (illustrative)

Capital markets organizations make money in various ways. One way is by making markets: simultaneously buying and selling financial products (equities, fixed income, commodities, crypto, credit, currencies and their derivatives) and profiting on the spread between the two. Another way is through investing or speculating: buying or selling products for extended periods of time and profiting on movements in price. Moreover, they generate revenue through fees charged for financial transactions or for the management of funds. They also make money by selling financial data used for pricing or investment decisions.

The key personas involved in the buying and selling of financial products in the capital markets are portfolio managers, asset managers, research analysts, quantitative analysts (quants), and traders. Investment bankers and financial analysts play important roles as well. There are also additional personas who support the activities, such as risk managers, compliance officers, and investor relations officers (IROs). Investments in asset management firms and hedge funds are often overseen by chief investment officers (CIOs).

Asset management and hedge-funds have technology, platform and cloud engineering, and innovations teams, which are often led by chief technology officers (CTOs). These teams include personas like enterprise architects, data engineers, data scientists, and software developers, who assess and build technologies and innovative solutions to support different core line of businesses (LoBs) and functionalities. Capital market firms allocate information technology (IT) budgets to support growth and initiatives, such as innovation and product enhancement. These firms maintain IT systems that span trading and investment management on the frontend, as well as risk management and post-trade workflow processes in the middle. On the backend, these systems handle clearing and settlement operations.

AI/ML and Generative AI Use Cases in Capital Markets Value Chain

As mentioned earlier, market participants in the capital markets value chain are divided into four categories: buy-side, sell-side, industry utilities, and financial information providers. There are functional commonalities and dissimilarities in the value chain of these participants. Generative AI and AI/ML bring beneficial impacts with many use cases across the value chain. In this section, we highlight the potential high-impact business use cases across the entire value chain common within capital markets.

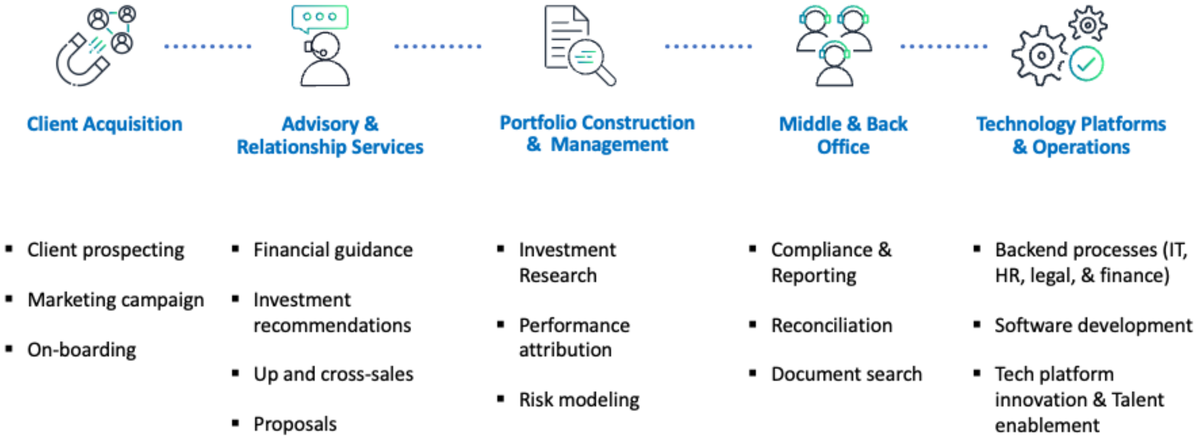


Figure 4: Generative AI use cases across the Capital Markets industry

The transformative potential of generative AI and AI/ML is reverberating across capital markets. It has begun a new era of intelligent automation that is driving efficiency, personalization, and innovation throughout the entire value chain.

Client acquisition stands to be improved as generative AI enables the creation of high-propensity lead generation solutions. These cutting-edge solutions can identify ideal client profiles by ingesting and analyzing vast troves of data, from internal sources like transaction histories and client interactions, to external feeds encompassing market events, news, and alternative data sets. Relationship managers, once burdened by manually collating this disparate information, can now leverage generative AI to surface hyper-targeted prospects through simple natural language queries. "Show me individuals over 50, with a net worth exceeding $5 million, based in California, and interested in golf," for instance. Generative AI can then dynamically generate personalized pitches and meeting materials tailored to each prospect's unique profile and circumstances.

Within **advisory and relationship services**, digital advisor assistants are emerging as indispensable tools. By summarizing the latest research tailored to specific client needs, contextualizing market commentary, and offering on-the-fly insights during client meetings, these generative AI-powered solutions allow human advisors to deliver more customized, high-value guidance. Upsell and cross-sell opportunities are identified proactively, while customer support is elevated through context-aware, omni-channel content serving up expert insights and next-best-action recommendations.

For **portfolio construction and management**, generative AI can revolutionize the ideation process for fund managers. Sophisticated large language models can dynamically generate data visualizations, textual summaries, and thematic analysis by querying structured and unstructured data sources in response to natural language prompts. This frees analysts from tedious data wrangling tasks, allowing them to focus on higher-order strategy and insight generation. Organizations can also explore new product opportunities by expanding their aperture to alternative data, with generative AI surfacing investible themes from vast textual repositories such as news, filings, earnings transcripts, and more.

The **middle and back offices** are also being transformed as knowledge management systems, turbo-charged by generative AI, streamline document processing and regulatory compliance. AI assistants can intelligently search, summarize, and extract key information from emails, chat logs, voice recordings, and other unstructured data sources. This unlocks significant operational efficiencies for processes like trade reconciliation, while enabling firms to get ahead of evolving regulatory landscapes through proactive reporting and analysis.

Finally, on the **technology platforms and operations** front, services like Amazon Q Developer are accelerating the software development lifecycle by automating large swaths of undifferentiated coding tasks. This frees engineering teams to concentrate on higher-value solution design, while AI-powered security scanning mitigates risks. Platform modernization initiatives are streamlined through automated legacy code conversion, allowing organizations to embrace agile cloud-native architectures.

Across these use cases, three common themes emerge: **summarization, conversational interfaces**, and **hyper-personalization**. The ability to interact with AI assistants through natural language queries is revolutionizing user experiences and democratizing access to powerful capabilities. By dynamically tailoring content, insights, and recommendations to specific contexts and profiles, generative AI is enabling a new frontier of personalized client engagement.

This is merely the beginning. As language models grow more sophisticated and firms iterate on real-world deployments, we can expect an accelerating wave of innovation rolling across the capital markets landscape. Firms embracing generative AI will be positioned to ride this wave, outpacing competitors through intelligent automation that boosts efficiency, sharpens personalization, and sparks continuous reinvention.

Considerations for Designing Generative AI Strategy

Now that we've discussed various use cases throughout the capital markets value chain, let's explore the factors to consider for a generative AI strategy. Developing a generative AI strategy is essential to successful generative AI adoption. A well-rounded strategy can help you identify your ideal use cases, achieve quick wins to build momentum, maintain stakeholder trust, and continuously measure and iterate to drive impactful results.

The specifics of your generative AI strategy will vary according to your organization's unique needs. Several best practices are virtually universal. First, an end-to-end data lifecycle should live at the core of your Generative AI strategy. This practice ensures that you can unify data from disparate sources and transform it into high-quality, structured datasets to train the foundational models that power generative AI applications. Second, it is important to understand that the adoption of generative AI technology is driven by diverse personas. Each of them has distinct goals and expectations from the technology provider. Recognizing and accommodating the unique needs of these personas is crucial for seamless and effective generative AI adoption. The third decision point is defining your generative AI investment strategy. Your strategy should include how you will construct your generative AI solution, either by building, buying, or constructing a hybrid of generative AI systems and services. Figure 5 shows the different offerings AWS has which cater across the buyer to builder spectrum.

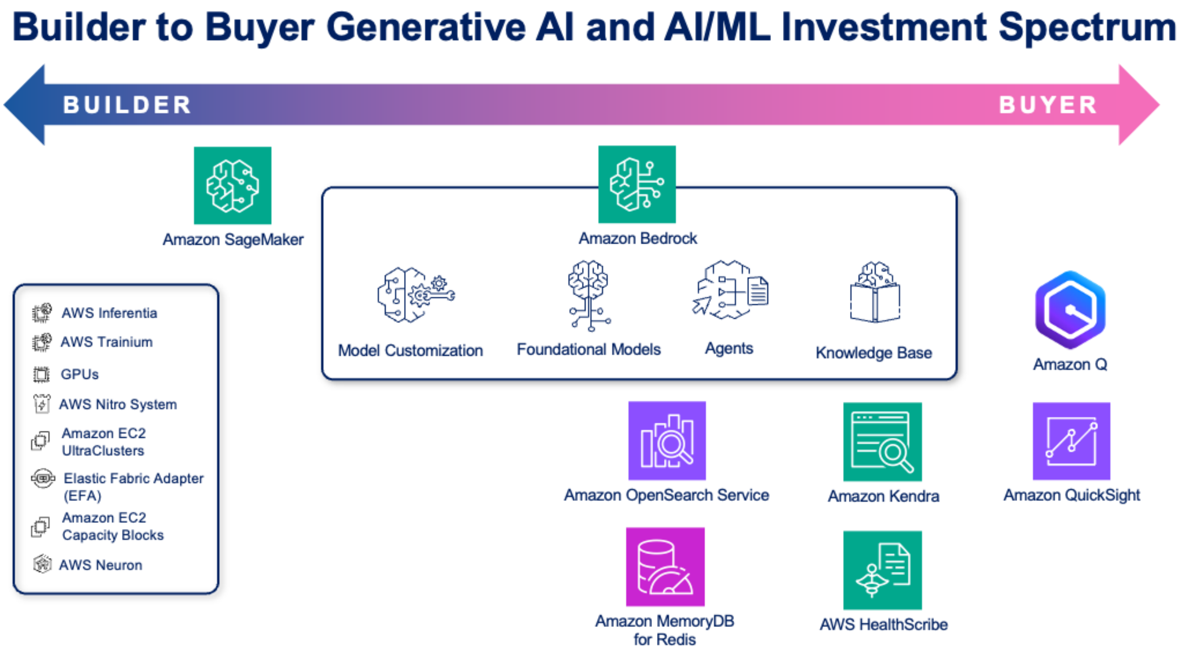


Figure 5: Builder to Buyer Generative AI and AI/ML Investment Spectrum

Broadly, we see organizations with personas who are **model builders**, **model consumers** and **application consumers**. Let us understand their needs.

Model builders are classified into two categories: foundation model builders and custom model builders. *Foundation model builders*, such as Anthropic, AI21 Labs, Cohere, Mistral, OpenAI, Stability.AI, and TII, focus on developing large-scale foundation models. These organizations typically require access to high-performance computing (HPC) clusters and employ advanced applied scientists to optimize the performance of these clusters as they pre-train foundation models on vast amounts of unlabeled data. AWS provides a comprehensive suite of offerings, including a choice of chips (such as AWS Inferentia and AWS Trainium) and cluster management capabilities (such as Amazon SageMaker HyperPod[1]), to meet the demanding needs of this customer segment. On the other hand, *custom model builders* generally work with smaller amounts of labeled or unlabeled data and relatively less compute resources to create models tailored for their specific domain or use case. These organizations often rely on their data science teams rather than employing dedicated advanced applied scientists. Consequently, custom model builders typically prefer solutions that abstract away the complexities of cluster management and other routine tasks. In this regard, Amazon SageMaker supports the needs of custom model builders, enabling them to focus on their core model development efforts while leveraging a fully managed service.

Model Consumers can be further subdivided into two cohorts: generative AI platform builders and generative AI application builders. The *generative AI platform builders* typically reside within the chief information officers' (CIO) organization or a dedicated AI center of excellence. Their primary aim is to encapsulate best operational and security practices around generative AI technology, making them available company-wide. These teams build their platform based on a variety of choices, depending on whether they have a preferred technology provider or wish to adopt a multi-provider approach. We have seen customers leverage both Amazon SageMaker and Amazon Bedrock for these purposes, with the decision driven by factors such as regional availability and flexibility required. The *generative AI application builders*' cohort usually operates under the chief technology officer (CTO) or specific lines of business (LoBs). Their focus is on producing generative AI-powered software applications that offer differentiating capabilities and deliver tangible business value. Where a company intends to build its own generative AI platform, these application development teams will rely on the internally developed platform. However, in scenarios where software engineers are given the flexibility to choose their own stack many prefer Amazon Bedrock, as it provides non-data scientists with the tools and capabilities to effectively use generative AI features, such as model customization and guardrails.

There will be instances where customers prefer to acquire generative AI capabilities through software-as-a-service (SaaS) offerings, which we call **Application Consumers**. As AWS identifies common use cases, we package some applications into services, enabling customers to consume these capabilities out-of-the-box. Services like Amazon Q Business, Amazon Q in QuickSight and Amazon Q in Connect exemplify this approach. This path represents the most accessible entry point for organizations to begin

[1] Amazon SageMaker HyperPod

their journey with generative AI, allowing them to achieve early wins, build confidence, and develop the skills required for more advanced implementations. Organizations cannot afford to ignore generative AI offerings, as doing so would put them at a disadvantage compared to their competitors who may have already embraced these technologies. Over time, most firms will adopt a multi-faceted approach to generative AI adoption consuming SaaS offerings, directly or through specialized independent software vendor (ISV) products.

Leveraging a comprehensive data strategy as the foundation for your generative AI implementation will help you deliver the highly accurate market predictions and relevant financial insights your workforce requires and your customers' demand. Further, it can help to build trust for generative AI solutions across your organization and client base—driving smarter financial decisions and a competitive advantage for your services. In addition, your strategy should consider various "human-in-the-loop" scenarios that add manual review by subject matter experts (SMEs) to the generative AI workflow. This will help you address potential risks, improve security, and maintain compliance.

AWS Generative AI Offerings

Amazon Bedrock offers a comprehensive solution for the capital markets customers seeking a secure, flexible, and responsible AI platform. Amazon Bedrock is a fully managed service that offers a choice of high-performing foundation models (FMs) from leading AI companies (such as AI21 Labs, Anthropic, Cohere, Meta, Mistral AI, Stability AI, and Amazon) through a single API, along with a broad set of capabilities needed to build generative AI applications with security, privacy, and responsible AI. Amazon Bedrock uses single API access, regardless of the models you choose, giving you the flexibility to use different FMs and upgrade to the latest model versions with minimal code changes.

Now let us understand the versatility of the different areas Amazon Bedrock offers:

- **Model customization** enables you to deliver differentiated and personalized user experiences. To customize models for specific tasks, you can privately fine-tune FMs using your own labeled datasets with a few clicks. Amazon Bedrock supports fine-tuning for Anthropic Claude 3 Haiku, Cohere Command, Meta Llama 3, Amazon Titan Text Lite, Amazon Titan Text Express, Amazon Titan Multimodal Embeddings, and Amazon Titan Image Generator.
- **Agents for Amazon Bedrock** plan and execute multi-step tasks using company systems and data sources from answering customer questions about your product availability to taking their orders. With Amazon Bedrock, you can create an agent in just a few clicks by first selecting an FM and providing it access to your enterprise systems, knowledge bases, and AWS Lambda functions to execute securely your APIs. An agent analyzes the user's request and automatically calls the necessary APIs and data sources to fulfill the request. Agents for Amazon Bedrock enables you to do

all this securely and privately—no need for you to engineer prompts, manage session context, or manually orchestrate tasks. Agents can also interpret code to tackle complex data-driven use cases, such as data analysis, data visualization, text processing, solving equations, and optimization problems.

- **Amazon Bedrock Developer Experience** makes it straightforward for developers to work with a broad range of high-performing foundation models using an API call. You can experiment with different FMs using interactive playgrounds for various modalities, including text, chat, and image. Model evaluation on Amazon Bedrock allows you to use automatic and human evaluations to select FMs for a specific use case.
- **Guardrails for Amazon Bedrock** evaluates user inputs and FM responses based on use case specific policies and provides an additional layer of safeguards regardless of the underlying FM. Using a short natural language description, Guardrails for Amazon Bedrock allows you to define a set of topics to avoid within the context of your application. Guardrails detects and blocks user inputs and FM responses that fall into the restricted topics. Guardrails for Amazon Bedrock provides content filters with configurable thresholds to filter harmful content across hate, insults, sexual, and violence categories. Additionally, contextual grounding in Guardrails for Amazon Bedrock ensures the large language model (LLM) response is based on the right enterprise source data and evaluates the LLM response to confirm that it’s relevant to the user’s query or instruction. Customers can use Guardrails across any foundation model—even those not supported by Amazon Bedrock.
- **Knowledge Bases for Amazon Bedrock** is a fully managed capability that helps you implement the entire retrieval augmented generation (RAG) workflow from ingestion to retrieval and prompts augmentation without having to build custom integrations to data sources and managed data flows. Session context management is built in, so your app can readily support multi-turn conversations. You can use the Retrieve API to fetch relevant results for a user query from knowledge bases. You can also add knowledge bases to agents for Amazon Bedrock to provide contextual information to agents. All the information retrieved from Knowledge Bases for Amazon Bedrock is provided with citations to improve transparency and minimize hallucinations. Organizations can leverage more business data to customize models for their specific needs with connectors for Salesforce, Confluence, and Microsoft SharePoint.
- **Custom Model Import on Amazon Bedrock** enables you to bring your own custom models and use them seamlessly on Amazon Bedrock. Whether you've fine-tuned Meta Llama or Mistral AI models to suit your specific needs, or developed a proprietary model based on popular open architectures, you can now import those custom models and use them alongside the foundation models.
- **Amazon Bedrock Studio** is a new single sign on-enabled web interface that provides a way for developers across an organization to experiment with large language models and other foundation models, collaborate on projects, and iterate on generative AI applications. It offers a rapid prototyping environment and streamlines access to multiple FMs and developer tools in Amazon Bedrock. To enable Amazon Bedrock Studio, AWS administrators can configure one or more

workspaces for their organization in the AWS Management Console for Amazon Bedrock, and grant permissions to individuals or groups to use the workspace.

Architecture Patterns and Approaches to Generative AI Solutions

There are different approaches to solve the architectural patterns emerging as the Capital Markets industry experiments with generative AI. These approaches include: prompt engineering, retrieval-augmented generation (RAG), fine-tuning, and continued pre-training (commonly ranked by their level of complexity, from the easiest to the most complex).

Prompt engineering is the practice of designing inputs (prompts) for FMs and LLMs that will produce optimal outputs from these models. There are different techniques associated with prompt engineering such as zero-shot, few-shot, chain-of-thought (CoT) and different combinations of these techniques. With **retrieval augmented generation (RAG),** there is no need to retrain the model. The contextual data used to augment your prompts comes from multiple data sources (for example, document repositories, databases, APIs and more) and is passed on as context to the LLM. Next is **fine-tuning**, where you fine-tune an existing FM using a small sample from your domain-specific labelled data and create a new model with your proprietary data-set for specific tasks. Finally, to adapt FMs with knowledge more relevant to a domain, you can engage **continued pre-training,** which leverages vast sets of unlabeled data.

Let’s dive deep into building AI-powered assistants using multi-model data based on the most commonly used RAG architectural pattern with generative AI agents.

AI-powered assistant for investment research with multi-modal data

Capital Markets organizations generate, collect, and use multi-modal data (including market, economic, customer, news, social media, and risk data) to gain insights, make better decisions, and improve performance. However, they face challenges because of the complexity and lack of standardization in financial systems, data formats, and quality, as well as the fragmented and unstructured nature of the data. There is difficulty in combining data from multiple modalities (text, images, audio, video) before gaining useable insights. This operational overhead causes complex extraction and transformation logic, leading to increased effort and costs.

Investment research analysts in the capital markets distill business insights from various data sources, including public filings, earnings call recordings, market research publications, and economic reports.

They face challenges because of the increasing variety of tools and the amount of data—needing to synthesize massive amounts of data from multiple sources. Analysts need to learn new tools and programming languages, such as SQL (with different variations). To add to these challenges, they must think critically under a time pressure and perform their tasks quickly to keep up with the pace of the market.

AI-powered assistants can address the challenges and improve the efficiency of investment research with multi-modal data. AI-powered assistants are advanced AI systems, powered by generative AI and large language models, that can understand goals from natural language prompts, create plans and tasks, complete these tasks, and orchestrate the results to reach the goal. As AI technology advances, the abilities of these generative AI agents are expected to grow, providing more opportunities to gain a competitive advantage in the financial industry. AI-powered assistants can boost the productivity of financial analysts, research analysts, and quantitative traders by automating many tasks, allowing them to focus on high-value creative work.

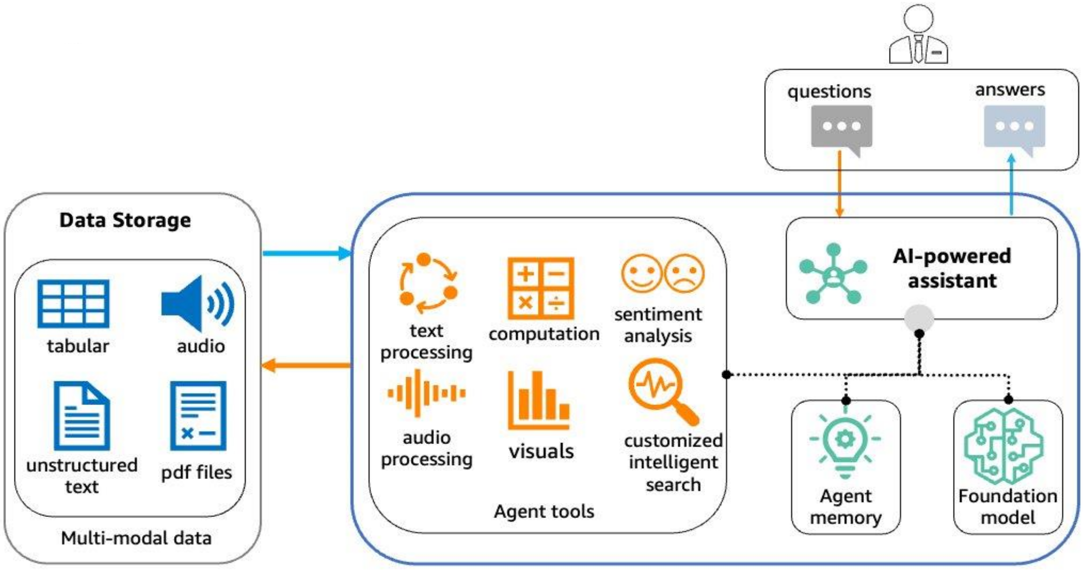


Figure 6: Architecture Pattern - AI-powered assistant for investment research

There are different ways of implementing AI-powered assistants using LLM agents—Agents for Bedrock[2], Langchain agents[3], LlamaIndex agents[4], and more. The technical architecture (Figure 7) shows implementation using Agents for Bedrock.

The key components of the technical architecture are as follows:

- **Data storage and analytics:** The quarterly financial earning recordings as audio files, financial annual reports as PDF files, and S&P stock data as CSV files are hosted on Amazon Simple Storage Service (Amazon S3). Data exploration of stock data is done using Amazon Athena.
- **Large language models:** The LLMs available to be used by Agents for Amazon Bedrock are Anthropic Claude Instant, v3, v2.1.
- **Agents**: Agents for Amazon Bedrock are used to build and configure autonomous agents. Agents orchestrate interactions between FMs, data sources, software applications, and user conversations. Depending on the user input, the agent decides the action or knowledge base to call to answer the question. We created the following purpose-built agent actions using AWS Lambda and Agents for Amazon Bedrock for our scenario:
    - **Stocks querying**: To query S&P stocks data using Athena and SQLAlchemy
    - **Portfolio optimization**: To build a portfolio based on the chosen stocks
    - **Sentiment analysis**: To identify and score sentiments on a topic using Amazon Comprehend
    - **Detect phrases**: To find key phrases in recent quarterly reports using Amazon Comprehend
- **Knowledge base**: To search for financial earnings information stored in multi-page PDF files, we use a knowledge base (using an Amazon OpenSearch Serverless vector store)

[2] Agents for Bedrock

[3] Langchain agents

[4] LlamaIndex agents

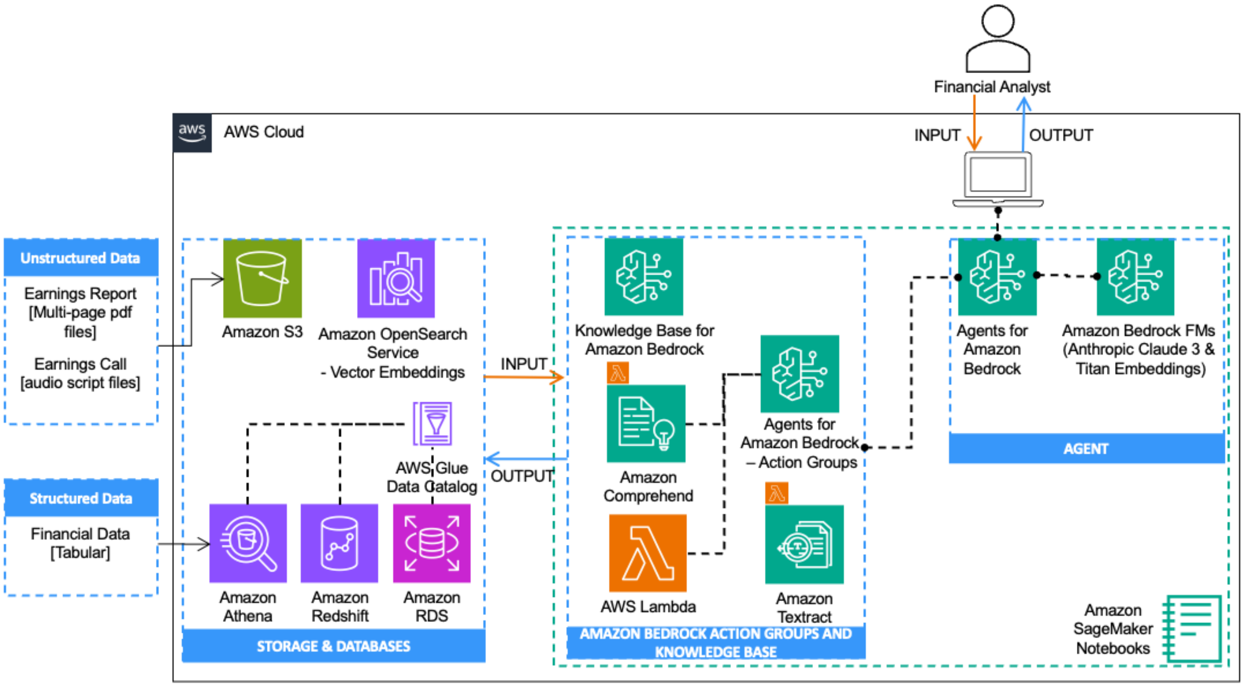

Financial Analyst
INPUT
OUTPUT
AWS Cloud
Unstructured Data
Earnings Report [Multi-page pdf files]
Earnings Call [audio script files]
Structured Data
Financial Data [Tabular]
Amazon S3
Amazon OpenSearch Service - Vector Embeddings
AWS Glue Data Catalog
Amazon Athena
Amazon Redshift
Amazon RDS
STORAGE & DATABASES
INPUT
OUTPUT
Knowledge Base for Amazon Bedrock
Amazon Comprehend
AWS Lambda
Agents for Amazon Bedrock – Action Groups
Amazon Textract
AMAZON BEDROCK ACTION GROUPS AND KNOWLEDGE BASE
Agents for Amazon Bedrock
Amazon Bedrock FMs (Anthropic Claude 3 & Titan Embeddings)
AGENT
Amazon SageMaker Notebooks

Figure 7: Technical architecture - AI-powered assistant for investment research

To find out more about applying generative AI for investment research, please refer to AI-powered assistants for investment research with multi-modal data: An application of Agents for Amazon Bedrock[5]

[5] AI-powered assistants for investment research with multi-modal data: An application of Agents for Amazon Bedrock

Customer Case Studies

The following case studies show examples of AWS customers taking generative AI use cases and putting them into production with success.

Jefferies[6]:

Jefferies is a global, full-service investment banking firm with a client-driven culture. The company takes pride in its entrepreneurial spirit, focusing on teamwork, innovation, and creative thinking.

Jefferies needed to put that spirit to work to meet a new Securities and Exchange Commission (SEC) regulation that shortens the settlement period for most routine securities transactions from two business days to one (T+1). While this change will reduce risk and better protect investors' money, it will also challenge post-trade operations teams to accelerate and automate many manual processes to ensure trades settle in time.

At Jefferies, processing derivative trade confirms was particularly cumbersome. Middle office managers had to manually review emails from various counterparties in many different formats with a lot of unstructured data. Further complicating the process, the confirmations not only varied by type of product and counterparty but also by different events in a trade life cycle, such as cancellation or amendment.

Once middle office managers extracted the necessary data into a standardized format, they started the reconciliation against internal systems. Then, they would send the confirmation back to the counterparty or request an amendment. This was a very complex and difficult process, ripe for automation.

Jefferies began working with AWS in 2022 to scale its infrastructure and migrate to the cloud. Today, Jefferies has applications built natively on AWS in every aspect of its business, from banking to trading to the back office.

Jefferies developed a custom intelligent document processing (IDP) framework based on AWS solutions like Amazon Textract to eliminate the need to manually monitor mailboxes, scan unstructured documents, parse information from them, and reconcile the data against internal systems. The document processing framework is adaptable to any file format and has greatly reduced the burden on Jefferies' middle office team.

The framework also includes a "human-in-the-loop" user interface to enable Jefferies' middle office team to review processing outputs and identify opportunities to train the tool on new tasks. This UI also ensures there is a full audit trail to meet compliance requirements.

[6] AWS Financial Services Symposium 2024, How Jefferies is reducing derivative settlements risk and AI/ML and IDP

The framework is fully API-based, enabling integration with internal platforms and automatic reconciliations. All data is stored in Amazon DynamoDB and Jefferies uses Amazon Cognito for authentication.

The solution has reduced Jefferies' confirmation processing time by 80-90%, while freeing its operations team to focus on other priority tasks. Jefferies has streamlined its processes with better audit trails and controls, hence improving accuracy.

The project's first phase focused on extracting data from unstructured documents. In the next phase of the implementation, Jefferies intends to address automated reconciliation against internal systems, which will double both time and cost savings. The company also plans to roll the solution out across equities, effects, and fixed income derivatives, as well as make it truly end-to-end by integrating with other areas of the trade cycle, such as the trade capture and settlement process.

NatWest[7]:

NatWest, a major British banking and financial services company, are using generative AI to create personalized product messaging and fraud detection solutions. Generative AI allows NatWest to derive better insights from customer data, market information, and their own internal data. The bank's goal is to drive customer interactions and engagement. However, as a bank, they are not primarily a content creation company, and generating all the personalized, helpful content for customers was a bottleneck.

To address this challenge, NatWest built strong AI and machine learning capabilities over the last few years, utilizing AWS offerings, such as Amazon SageMaker, to build and deploy models rapidly. What used to take 6-12 months can now be accomplished in just a couple of weeks, enabling greater agility in areas like fraud detection and personalized customer content that needs to evolve constantly.

Using generative AI, NatWest can now create highly personalized marketing content at scale. They can reduce customer segment size and craft unique, contextual messages tailored to each individual customer. The bank has built tools that leverage large language models to generate the content, while also checking it against brand guidelines, compliance rules, and allowing for human approval. Applying the RAG technique in AWS, NatWest can bring contextual information securely to large language models and generate personalized content.

This approach has allowed NatWest to review a larger set of messages, catch violations in human-generated content, and move towards production with increased controls. The results have been promising so far. For the bank's credit score product, the AI-generated personalized messages drove 4x higher engagement compared to human-written control messages, and they also saw a 900% growth in applications for high-interest rate accounts.

[7] AWS re:Invent 2023 - How to deliver business value in financial services with generative AI (FSI201)

NatWest recognizes that harnessing the power of responsible and ethical AI use will be essential to achieving the bank's commitment to helping customers manage their financial well-being. They are doing this through personalized engagement tools, such as the NatWest Digital Financial Health Check and the extended Know Your Credit Score service.

Linedata:

Linedata, a leading provider of asset management technology, AI analytics and services, collaborated with the AWS Generative AI Innovation Center to improve efficiency in investment compliance workflows for its clients.

Compliance teams at large asset management firms, wealth management firms, fund administrators, and alternative investment managers typically face a manual and error-prone process when scanning investment management agreements, identifying key terms, and creating monitoring rules.

This challenge is exacerbated as agreements and regulations frequently change, making it difficult to keep up.

To address this, Linedata leveraged AWS offerings, including Knowledge Bases for Amazon Bedrock and large language models (LLMs), to automate compliance processes. They first built an intelligent interface powered by LLMs to answer user questions about key terms in client investment management agreements. To provide authoritative responses, Linedata used Knowledge Bases for Amazon Bedrock to create a knowledge repository indexing underlying documents and guidelines. The LLMs enabled inferencing to generate accurate responses to new questions by leveraging the knowledge in the Knowledge Bases for Amazon Bedrock. Linedata also created a method to automatically generate compliance rules using API endpoints on Amazon Elastic Compute Cloud (Amazon EC2) instances. Linedata built an efficient and scalable question-answering intelligent interface for various compliance data modalities by seamlessly integrating LLMs and Knowledge Bases for Amazon Bedrock within the AWS cloud architecture.

Linedata’s AI-powered solution has significantly improved efficiency and accuracy in onboarding new clients and updating terms for existing clients. It is more scalable, allowing the system to dynamically update as new agreements come in, saving substantial time. The intelligent interface provides explainable responses to increase user confidence. With faster agreement review, portfolio managers can capitalize on time-sensitive market opportunities more quickly.

Through this innovative collaboration with the AWS Generative AI Innovation Center, Linedata is leveraging AI to automate a critical compliance workflow, significantly improving client efficiency, scalability, and accuracy.

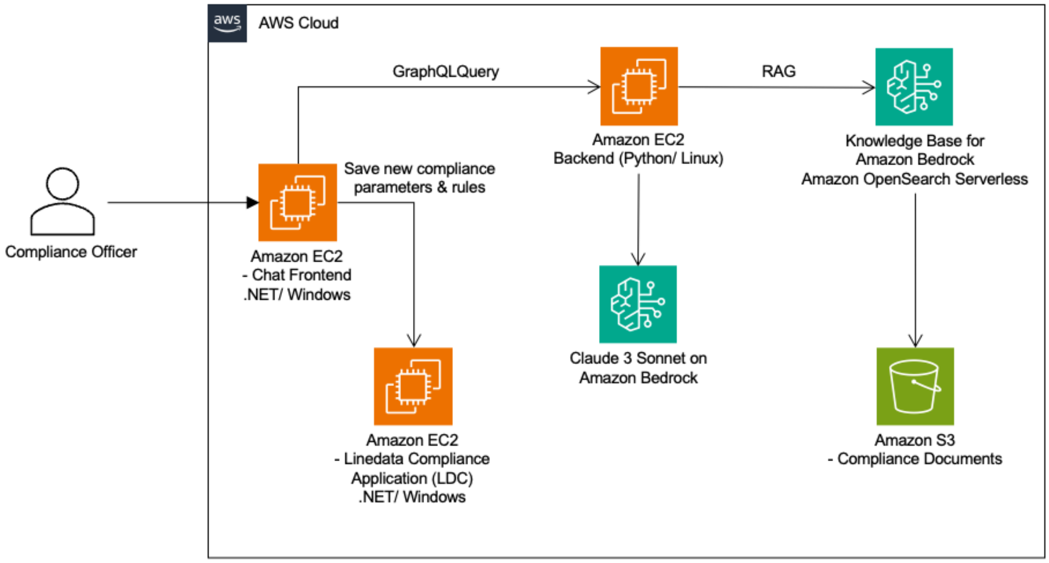


Figure 8 : Linedata's AI-powered solution

Lessons Learned and Path to Production

AWS works with thousands of customers globally concerning their generative AI initiatives. We observe the prototypes start with plenty of excitement. However, it quickly dissipates once teams get to the implementation phase as they realize that the required skill sets are not there in the teams, data is not easily accessible because of siloed databases, and business teams are difficult to onboard without a sound business case. We've seen these and organization cultural challenges as a common pattern globally across industry verticals. Establishing the right foundations in people, process, technology, and mindset/culture is crucial for driving technological innovation and business value through generative AI at scale.

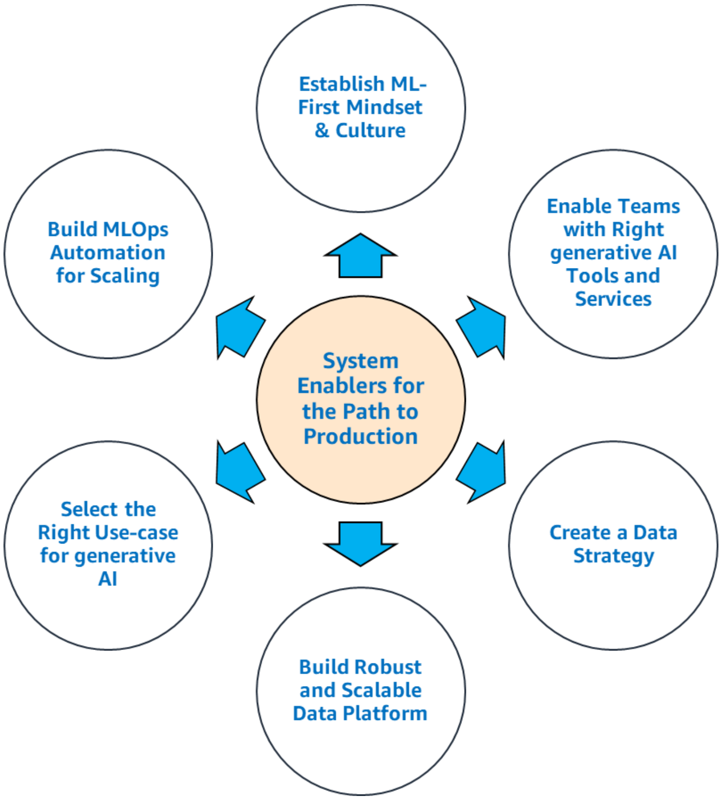

System Enablers for the Path to Production
Establish ML-First Mindset & Culture
Enable Teams with Right generative AI Tools and Services
Create a Data Strategy
Build Robust and Scalable Data Platform
Select the Right Use-case for generative AI
Build MLOps Automation for Scaling

Figure 9: Enablers for scaling AI initiatives

While there are many things requiring attention on the path to production, the following six enablers are the key to successfully scaling generative AI initiatives and increasing the chances of making it to production:

1. **Establish ML-first Mindset and Culture:** The ML-first mindset is an approach to embrace generative AI and AI/ML in the fabric or organization vision, processes, and operations. It is about applying these technologies to gain a competitive advantage and boost organization productivity. The organizations invest time, money, and resources to implement this consistently. Setting a clear vision and business drivers for generative AI is paramount to the long-term success. In a Gartner poll[8] of 1400 executives, the organizations innovating with generative AI showed growth initiatives (30%), cost optimization (26%), and customer experience/retention (24%) as the biggest business drivers. Our ecommerce business, Amazon, had been using AI/ML in every aspect of the business operations today. Every team at Amazon is asked to identify clear customer opportunities for leveraging AI/ML. It brings the business and technology teams together to articulate the biggest opportunities, compelling customer benefits and needs to champion the ML skills within their teams.

2. **Enable Teams with Right generative AI Tools and Services:** The democratization of generative AI services is vital for the broader adoption of the technology. “Right tool for the job” goes beyond merely exploring technology features and capabilities. It’s about what the team needs are in short term and long-term, fit-for-purpose skillset, and their view about toolset evolution over time. AWS provides a complete stack of managed and unmanaged generative AI and AI/ML services that are straightforward and powerful to use toward common use cases. Low-level services provide more granular controls, but require substantial expertise. Organizations with early success have correctly determined their selection mix of services from the start. Their “generative AI platform” provides a comprehensive solution with governance, guardrails, and service endpoints, granting access to diverse LLMs for builders, innovators, and data science teams.

3. **Create a Data Strategy:** Secure, enriched, accurate, and governed data is vital to the success of any generative AI application at scale**.** Organizations that have not yet effectively harmonized and provide ready access to their data cannot fine-tune generative AI to unlock more of its potentially transformative uses. Our suggestion is to create a flexible data strategy that considers dynamic internal and external factors while aligning with your organization’s business objectives for AI/ML and generative AI. Organizations often over index their data strategy towards the technology for their data platform. While technology is an important aspect of an organization’s data strategy, it’s equally important to focus on mindset, people, and processes that are tailored to the organization’s requirements in order to effectively scale generative AI and AI/ML adoption and derive maximum value from data.

[8] Gartner poll

4. **Build Robust and Scalable Data Platform:** The lifeblood of generative AI is fluid access to data honed for a specific business context or problem. So, contextually relevant, accurate, governed, and secured data is crucial for the success of a generative AI application. A scalable cloud-based fit-for-purpose data platform (for example, a data lake) is important to achieve this. By leveraging a cloud-based data platform, organizations can scale, capture, and store any amount of data, at low cost, and in open standard formats. The data platform has four components: 1) A cloud-based database optimized for the specific workload needs (specifically, key-value pair, columnar databases and so on); 2) A scalable data lake component to ingest and store high volumes of structured, unstructured, and semi-structured data at low cost and with high durability; 3) Native integration between services that span the data lake, databases, and machine learning services; 4) A catalog of business and technical metadata that registers data products. It uses a contract mechanism to support governance automation. For example, Amazon reimagined the traditional data warehouse and created a new data lake that scales with its diverse business needs. It paved the way to transform on-premises monolithic data warehouse that stifled innovation to a 2 exabyte micro-service-based, modern data platform, where teams are experimenting and innovating at the speed of their individual businesses.

5. **Select the Right Use Case for generative AI:** Choosing the first generative AI use case for an organization is like choosing the first dish at a new restaurant. Just as the first dish colors your impression of a restaurant's food, the first use case sets the tone and shapes perceptions of the organization's capabilities. Asking these two questions can make you successful in selecting the use cases with the biggest potential to make to production: 1) “Does it solve a real problem for the customer and your business?” The first use case cannot be a proof-of-concept (PoC) lacking substance and business value. 2) “How does generative AI unlock big new opportunities?” Often our customers struggle to define the problem that generative AI will be focusing on. For example, if traditional approaches to AI/ML solve 90% of the use case, then it is hard to argue for a compelling generative AI business case. Once you ask these two questions, it allows your organization to define a clear vision and focus on quality use case for prioritization. As you prioritize, we recommend asking secondary questions such as: “Is there good quality data available for the use case?”, “Can you deliver success in 3-6 months?”, and “is the use case important enough to get business attention and promote adoption?” These questions help make your prioritization geared towards delivering business value. Finally, it is important to realize that the business cannot start trusting generative AI from the first attempt. It is also important to be thoughtful of business and regulatory risks in the early efforts and focus on what can be delivered successfully. Then start expanding quickly with a better chance of success in the long term.

6. **Build MLOps Automation for Scaling:** The ML lifecycle is a multi-step process. Once you pass the PoC phase, the right tools for development and automation becomes critical. It requires a lot of effort to go into production once the model is validated through PoC. MLOps focuses on automating the ML lifecycle. It helps ensure that models are not just developed but also deployed, monitored, and retrained systematically and repeatedly. It brings DevOps principles to ML. FMOps and LLMOps are practices that extends MLOps to foundation models and large language models, and bring

additional setups for generative AI solutions. MLOps results in faster deployment of ML models, better accuracy over time, and stronger assurance that they provide real business value. For example, multiple teams at Amazon are enjoying the benefits of quicker model delivery at lower costs by migrating their MLOps to Amazon SageMaker. These teams can minimize time spent on repeated tasks (such as model retraining and deployment) by being able to do them with "mouse clicks" on these ML platforms. This also means they are minimizing human errors and the downstream cost associated with poor quality—ultimately resulting in an ability to identify and pursue new business opportunities that can be unlocked through AI/ML and generative AI.

Conclusion

Generative AI has brought a paradigm shift for the Capital Markets industry in the way it operates, innovates, and automates across the entire trade lifecycle. Cloud computing has emerged as a pivotal enabler for the successful adoption of generative AI at scale. AWS offers a comprehensive suite of AI/ML and generative AI services, tools, and solutions tailored to address the unique needs of the Capital Markets industry. From customizable AI platforms to pre-trained foundational models, to out-of-the-box applications, AWS provides capital markets customers the flexibility to chart their own generative AI journey aligned with their specific requirements.

As highlighted through the architectural patterns and case studies, generative AI can streamline and enhance capabilities across the entire trade lifecycle, from intelligent lead generation and personalized client advisory, to automated portfolio construction and management, regulatory compliance, and agile software development operations. The ability to interact through natural language, gain insights from multi-modal data, and dynamically tailor content drives a new frontier of hyper-personalized customer engagement.

While early successes show the tangible impact, widespread production deployment requires a robust generative AI strategy, responsible AI practices, data strategy, and seamless integration with existing workflows. AWS takes a comprehensive approach focused on cloud adoption, scalable data infrastructure, model management lifecycles, and decision augmentation—helping position capital markets firms to outpace competitors through intelligent automation.

Additional Readings

- Analyst Report - Celent | Generative AI Making Waves | Adoption waves in banking and capital markets
- AWS Whitepaper - AWS Cloud Adoption Framework for Artificial Intelligence, Machine Learning, and Generative AI
- Implementation Guide - Generative AI Application Builder on AWS (AWS Well-Architected design considerations)
- AWS Well-Architected Framework - Machine Learning Lens
- Generative AI for Financial Services
- Amazon Bedrock
- Amazon SageMaker
- AI-powered assistants for investment research with multi-modal data: An application of Agents for Amazon Bedrock
- Generative AI and multi-modal agents in AWS: The key to unlocking new value in financial markets
- Empowering analysts to perform financial statement analysis, hypothesis testing, and cause-effect analysis with Amazon Bedrock and prompt engineering
- Establishing an AI/ML center of excellence
- Empowering everyone with GenAI to rapidly build, customize, and deploy apps securely: Highlights from the AWS New York Summit
- Automate derivative confirms processing using AWS AI services for the capital markets industry
- Anthropic’s Claude 3.5 Sonnet ranks number 1 for business and finance in S&P AI Benchmarks by Kensho
- Reinventing the data experience: Use generative AI and modern data architecture to unlock insights

Contributors

Sovik Kumar Nath - Generative AI and AI/ML Specialist Senior Solution Architect | Capital Markets Lead | Amazon Web Services

At AWS, Sovik leads generative AI and AI/ML solution architecture for capital markets customers. He has extensive experience designing end-to-end machine learning and business analytics solutions in finance, operations, marketing, supply chain management, and internet of things. Sovik has published articles, AWS blogs and holds a patent in ML model monitoring. He has double master's degrees from the University of South Florida. U.S.A., University of Fribourg, Switzerland, and a bachelor's degree from the Indian Institute of Technology, Kharagpur.

Nimit Jain - US, Generative AI and AI/ML Specialist Senior SA Leader | Financial Services | Amazon Web Services

Nimit is a seasoned technology and business leader with 20+ years of experience in the field of artificial intelligence (AI). His strategic focus and innovative approach have enabled large-scale AI and data initiatives, solving real-world problems and unlocking high-impact improvement opportunities across various industries, including banking, pharmaceuticals, fast-moving consumer goods (FMCG), and retail.

Vipul Parekh - Senior Customer Solutions Manager | Capital Markets | Amazon Web Services

Vipul Parekh is a senior customer solutions manager at AWS guiding our capital markets customers in accelerating their business transformation journey on Cloud. He is a generative AI ambassador and a member of AWS AI/ML technical field community. Prior to AWS, Vipul played various roles at top investment banks, leading transformations spanning from front office to back office, and regulatory compliance areas.

Chris McDonald | Capital Markets Specialist | Amazon Web Services

Chris is a Capital Markets Specialist at AWS focusing on Industry trends, wealth management and cyber event recovery. Chris joined AWS from Bloomberg where he held various senior global positions in business and product development focusing on enterprise data, regulatory reporting, regulation and compliance solutions. Previously Chris was a senior VP with Goldman Sachs Investment Bank and also JPMorgan Investment Bank where he managed and was responsible for various back and middle office functions.

Renee Lau - Financial Services Industry Specialist | Amazon Web Services

Renee Lau is a Principal Financial Services Specialist covering capital markets and banking customers at AWS. Prior to joining AWS, Renee led Risk Management teams at MUFG and Goldman Sachs. Her experience also includes investment banking advisory, residential MBS origination, and middle office roles.

Kimberly Hatton – Principal Product Marketing Manager, Financial Services | Amazon Web Services

Kimberly is a Principal Product Marketing professional for the Financial Services Industry at AWS. Prior to joining AWS, Kimberly led marketing at Bloomberg for Financial Products and Compliance, and product marketing at J.P. Morgan Chase, Guardian Life Insurance, and UBS. She has played a pivotal role in guiding firms across various sectors to establish impactful brands, design distinctive customer experiences, and implement digital strategies that drive rapid growth.