AWS Whitepaper

AWS Cloud Adoption Framework for Artificial Intelligence, Machine Learning, and Generative AI

AWS Cloud Adoption Framework for Artificial Intelligence, Machine Learning, and Generative AI: AWS Whitepaper

Table of Contents

This whitepaper is for historical reference only. Some content might be outdated and some links might not be available.

AWS Cloud Adoption Framework for Artificial Intelligence, Machine Learning, and Generative AI

Accelerating Your Cloud-Powered AI Transformation

Publication date: **February 13, 2024** (*Document history*)

In this document, we outline the AWS Cloud Adoption Framework for Artificial Intelligence (AI), Machine Learning (ML), and generative AI, a framework that describes a mental model for organizations that strive to generate business value from Artificial Intelligence. In this framework, we describe the Artificial Intelligence journey that customers go through as their organizational capabilities on AI and ML mature. We structure this journey by carving out foundational capabilities that assist an organization to grow its maturity in AI. Finally, we provide prescriptive guidance by providing an overview of the target state of these foundational capabilities and explaining how to evolve them step by step to generate business value along the way.

Introduction to Artificial Intelligence

Artificial Intelligence (AI) is a broad field that aims to create or at least imitate intelligent machines capable of performing tasks that traditionally require human intelligence. These tasks can include anything from understanding natural language and visual perception to decision-making and problem-solving. One commonality among many AI systems is the pursuit of probabilistic outcomes—essentially, generating predictions or decisions with a high degree of certainty, often mirroring the complexity of human judgment. Such systems can then be used to automate or augment knowledge work.

A significant portion of AI today is built upon machine learning (ML), a branch of AI that focuses on developing techniques that enable computers to learn from and make decisions based on data. Rather than relying on explicit programming, machine learning models generalize from examples, making them highly versatile for a myriad of applications. Among the various techniques within machine learning is deep learning, a specialized subset that uses neural networks with multiple layers to analyze complex factors in data. Deep learning is especially adept at handling unstructured data like images and text, and has led to breakthroughs in numerous complex tasks such as image and speech recognition.

An emerging capability within deep learning is generative AI, which makes AI generate or create new, potentially original content. This innovative sub-discipline is increasingly being recognized for

its ability to produce outputs that mimic aspects of human-like thought and reasoning capabilities. Advances in computing power, data availability, and algorithmic innovation have made generative AI possible, paving the way for a wide range of applications, from entertainment and art to scientific research.

Together, these sub-disciplines and techniques represent the layered and interrelated landscape of Artificial Intelligence, each contributing to the development of systems that can autonomously perform an increasingly broad array of tasks. The applications and capabilities of AI are likely to continue to expand quickly, making it an integral part of our daily lives and a crucial tool for solving complex problems.

"Generative AI has captured people's imagination in a way few innovations have. As it evolves beyond the realm of researchers and developers, it's proving to have broad applications that range from enhancing consumer experiences to solving complex enterprise problems. Whether it's producing human-like text, assisting coders with AI-driven code snippets, or automatin g customer interactions via intelligent chatbots, the possibilities seem endless. Beyond these applications, generative AI serves as a catalyst for reimagining how technology can augment human abilities and extend our reach, doing so with an unprecedented blend of scalability, customization, and intelligence. As we stand on the precipice of mass adoption, the technolog y's promise isn't just in accomplishing tasks more efficiently, but in fundamentally redefining what is possible across various sectors." **– Andy Jassy, CEO Amazon**

Note

Moving forward, the term *Artificial Intelligence (AI)* is used as an umbrella term encompassing all its various sub-disciplines. When referring to specialized areas within AI, they are specified by name, such as *generative AI* or *Machine Learning*, to distinguish them from the broader field of AI.

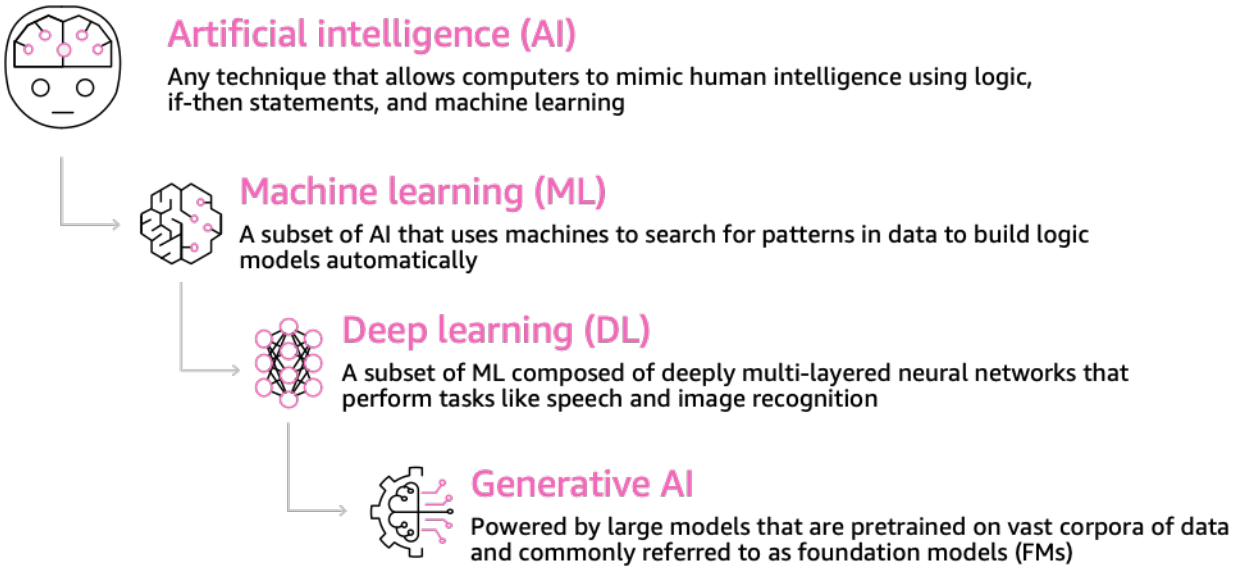

Figure 1: Taxonomy of artificial intelligence, machine learning, deep learning, and generative AI

Introduction to AWS CAF-AI

The AWS Cloud Adoption Framework for Artificial Intelligence, Machine Learning, and generative AI (CAF-AI) serves as both a starting point and a guide for your ongoing journey in AI, ML, and generative AI. This framework aims to inspire and inform your mid-term planning and strategy in these specialized fields. Use it as a resource for discussions about your AI strategy, not only within your team but also when collaborating with coworkers and AWS Partners.

Depending on where you are in your journey, you might focus on a specific section and hone your skills there, or you might use the whole document to judge maturity and help direct near-term improvement areas. CAF-AI is a constantly growing and updated summary and index of all the things that you need to consider when adopting AI at an enterprise level and help you to go beyond a single proof of concept (POC). Our goal is to give our customers the same prescriptive guidance they know and expect from the AWS Cloud Adoption Framework (AWS CAF) for AI, so they can successfully implement it. AWS CAF is underpinned by a set of foundational organizational capabilities and provides prescriptive guidance that thousands of organizations around the world have successfully used to accelerate their cloud transformation journeys.

In AWS CAF-AI, we remain reliant on these foundational capabilities but we enrich many of them so they include the changes that AI demands. In addition, we identify and add new foundational capabilities that organizations should consider as part of their AI journey.

AWS CAF: The Cloud Adoption Framework

At AWS over the last 10 years, we have built the AWS Cloud Adoption Framework (AWS CAF) as a cornerstone for our customers' cloud adoption strategy. While evolving this framework, we have kept it largely unbound to specific technologies beyond the cloud to make sure that its insights and mental model can be used by many of our diverse customers. However, Artificial Intelligence is an entirely new breed of technologies that has a large impact on all verticals and most of our customers. We have built CAF-AI to help our customers along their journey of AI adoption accelerated through cloud technology.

Are you Well-Architected?

The AWS Well-Architected Framework helps you understand the pros and cons of the decisions you make when building systems in the cloud. The six pillars of the Framework allow you to learn architectural best practices for designing and operating reliable, secure, efficient, cost-effective, and sustainable systems. Using the AWS Well-Architected Tool, available at no charge in the AWS Management Console, you can review your workloads against these best practices by answering a set of questions for each pillar.

In the Machine Learning Lens, we focus on how to design, deploy, and architect your machine learning workloads in the AWS Cloud. This lens adds to the best practices described in the Well-Architected Framework.

For more expert guidance and best practices for your cloud architecture—reference architecture deployments, diagrams, and whitepapers—refer to the AWS Architecture Center.

Artificial Intelligence cloud transformation value chain

Artificial Intelligence has evolved from niche technologies into a powerful and broadly available business capability. Machine Learning is by now fueling a new wave of innovation, where data is the genesis of invention, and where ML is a net-new capability for organizations not only to describe the past but also to predict the future and prescribe meaningful actions. Because of the impact this capability has on all markets and businesses, organizations across all industries are increasing their investment in AI. This investment can create a competitive advantage through improved customer insight, greater employee efficiency, and accelerated innovation. This is driven by the applicability of AI to a vast problem space that spans both vertical and horizontal use cases.

Notably, the business problem space to which AI can be applied is not a single function or domain, rather there is significant potential across all functions of businesses and all industry domains with the opportunity to reset the playing field in markets where AI does make an economical difference. As AI enables solutions and solution paths to problems that have remained uneconomical to solve for decades or simply technically were impossible to tackle without AI, the resulting business outcomes can be profound.

As an example, the emergent capabilities of large AI models to perform domain-specific functions with little additional data are taking organizations by storm and help business to differentiate. The discipline that these mainly fall into is generative AI, which has captured widespread attention and imagination. However, developing, applying, and tuning such models can be complex.

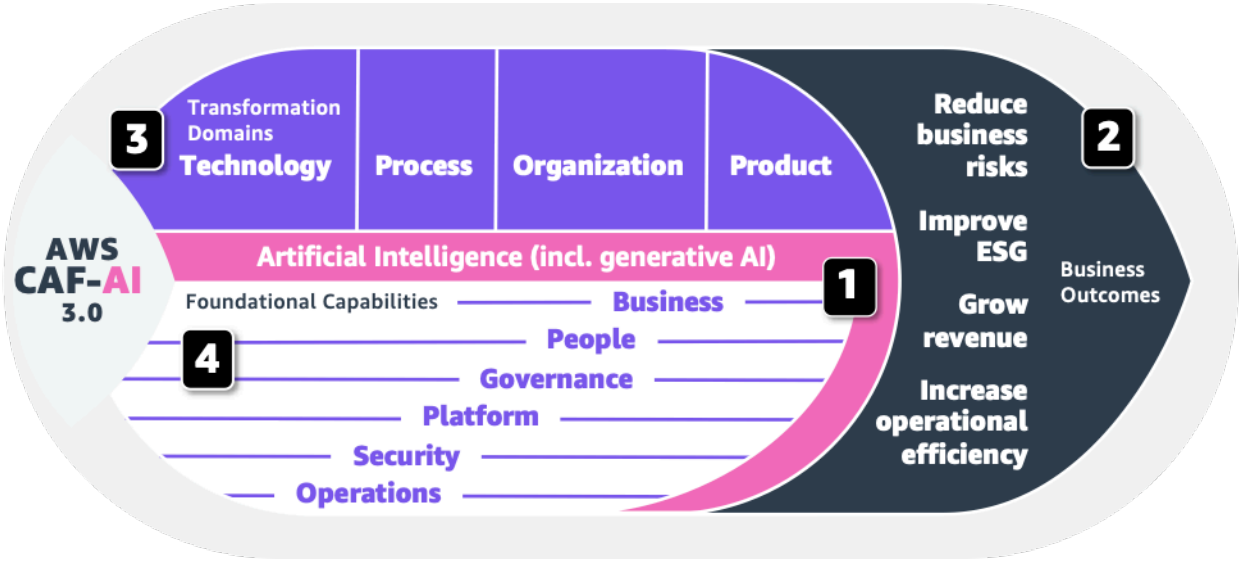

Figure 2: The AWS CAF-AI transformation value chain (all pink and magenta parts of it are dimensions taken from the original CAF that we build upon here).

The preceding figure provides an orientation on how to think about Artificial Intelligence adoption in the face of a changing market landscape and the rapidly accelerating field.

1. AI provides new capabilities to your organization.
2. With these new capabilities, you and your organization strive to create tangible business outcomes. These business outcomes can be many, such as reduced business risks (for example, by detecting broken or faulty parts in a production chain), by improving the environmental, social, and governance (ESG) performance (for example, by automatically summarizing and flagging environmental protection compliance reports), growing new and existing revenue (for example, by personalizing product and service recommendations to customers) or by increasing the operational efficiency (for example, by classifying and mapping travel receipts to internal booking codes). However, creating these business outcomes relies on your ability to adopt AI.
3. To adopt AI, your organization needs to transform along at least four domains:
    a. **Technology:** A domain that focuses on establishing the technological capability and then enabling the usage and adoption of AI.
    b. **Process:** A domain that focuses on digitizing, automating, optimizing, and innovating on your business operations through the power of AI.
    c. **Organization:** How your business and technology teams orchestrate their efforts to create customer value and meet your strategic intent, driven by AI.
    d. **Product:** Reimagining your business model by creating new value propositions (products, services) and revenue models that capitalize on the capabilities of AI.
4. Transforming these domains and enabling them to use AI depends on your foundational capabilities in business, people, governance, platform, security, and operations.

To adopt AI successfully, plan out your journey:

- Work backwards from your understanding of what AI enables you to do.
- Define what your expected business outcomes are over time.
- Carve out the transformation that your business has to go through.
- Develop the foundational capabilities that enable this journey.

Your AI transformation journey

Any large technological adoption agenda is a long journey, especially when adopting a technology that is rapidly evolving, such as AI. While transformation and adoption journeys are highly individual to the organization, we have observed patterns of successful AI adoption. Therefore, to de-risk this journey for customers, the AWS CAF-AI provides the following observations learned from thousands of customers as best practices. Still, each organization's AI journey remains a unique one.

When embarking or advancing on your AI transformation journey, consider *four critical elements*, which are also illustrated in Figure 3:

1. The destination of your journey, namely the **business outcomes** that you seek to achieve and from which you work backwards.
2. The **AI flywheel** as the motor of your journey. The AI flywheel is a virtuous cycle where initial high-quality data (which is timely, relevant, valuable, and valid) is used to train or tune an AI system that then delivers predictions. These predictions positively impact business outcomes that in turn lead to more or deeper customer relationships, sparking the creation of more or higher quality data (network and flywheel effect).
3. Your **data and data strategy** is what keeps the AI flywheel in motion.
4. Your **foundational capabilities** that, above all else, drive success or failure when adopting AI.

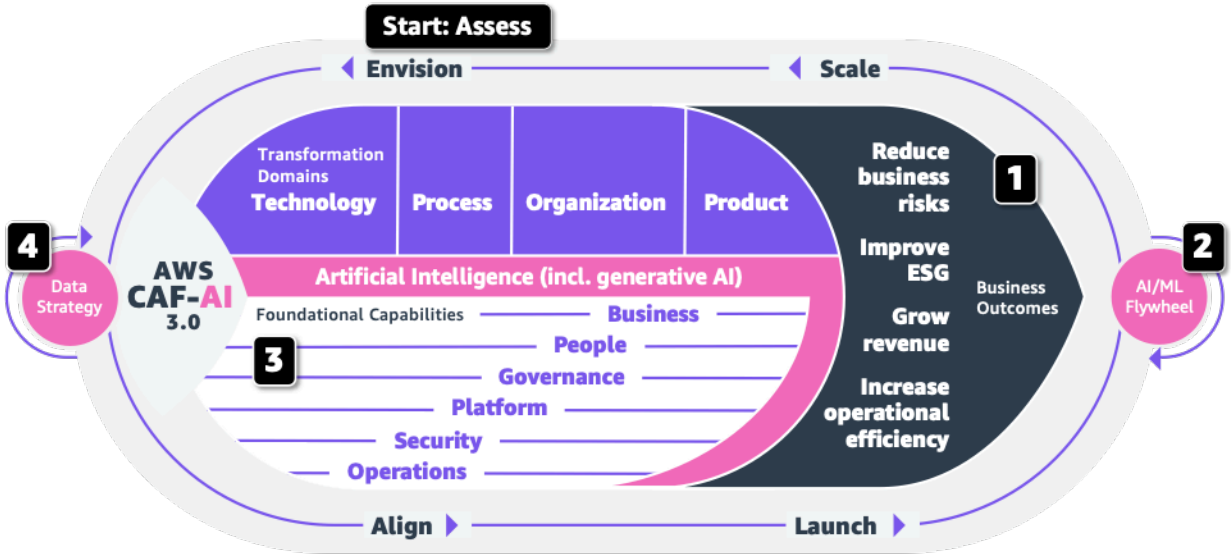

Figure 3: The AWS CAF-AI cloud transformation journey

When approaching this journey, base it on iterative and incremental improvements. We also suggest you reach out to your AWS contacts (for example, your account team) to get assistance from AWS ML strategists, enterprise strategists, and ML advisors. After an initial assessment, the adoption cycle begins, and it is based on four stages:

- **Envision:** This first phase focuses on envisioning how AI can help accelerate your business outcomes. This means identifying and prioritizing transformation opportunities in line with your business objectives. Associate your transformation initiatives with key stakeholders (that is, senior individuals capable of influencing and driving change) and measurable business outcomes. Be sure to also identify in this early phase what data assets and sources these initiatives and opportunities rely upon. Work backwards from your opportunities towards data requirements.
- **Align:** In this second phase, you focus on the foundational capabilities. You identify cross-organizational dependencies and surface stakeholder concerns and challenges. AI adoption is a cross-functional effort, much more so than this is the case for other technologies. Aligning internally on the goals set in the envision phase is critical. Doing so helps you create strategies for improving your cloud and AI readiness at large, ensure stakeholder alignment and future buy-in, and facilitate relevant organizational change management activities.
- **Launch:** In this phase, you focus on delivering pilot initiatives from early proofs of concept to production and demonstrate incremental business value. Pilots should be highly impactful on the organization and the business, as well as meaningfully benefit from AI being applied to it. Regardless of whether they are successful or not, they can help influence your future direction. Learning from them helps you adjust your approach before scaling to full production.
- **Scale:** This phase focuses on scaling pilots in production to achieve broad, sustained value. Scaling here can mean not only the technical capabilities of solutions or initiatives, but also the reach of them through the business and towards your customers. This activity translates your activities into customer value.

While you iterate through these cycles, recognize the limits of what you can achieve in a single cycle. It is important to be ambitious and aim high, but trying to do everything in the same cycle can lead to discouragement in the organization. This is why pairing a larger picture with many pragmatic and actionable steps and measurable KPIs on these smaller steps is crucial. Every step then brings the organization closer to its goal. Do not try to do everything at once. Rather, evolve the foundational capabilities and improve your AI readiness as you progress through your AI transformation journey.

Foundational AI capabilities

Iterating through your AI transformation journey relies on your foundational capabilities to adopt AI across business, people, governance, platform, security, and operations. A *capability* is an organizational ability to use processes to deploy resources (such as people, technology, and other tangible or intangible assets) to achieve an outcome. The following figure shows the list of foundational capabilities specifically relevant for AI adoption (pink). In gray are existing CAF capabilities that remain untouched for AI adoption.

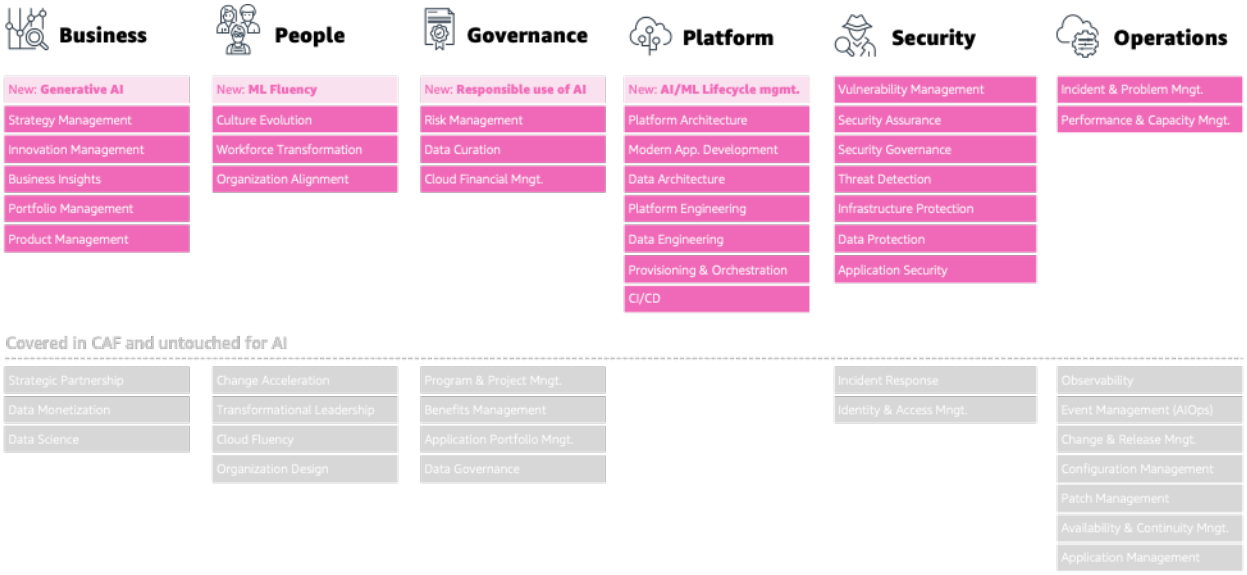

Figure 4: The AWS CAF-AI foundational capabilities

For example, the product management capability in the Business perspective section. Although product management is already a needed capability to successfully develop cloud-based products, its implementation is significantly different when looking at AI services in the cloud. The remainder of this document follows this logic: We call out the deviations and specific needs for AI adoption. The remaining capabilities are described in the original AWS Cloud Adoption Framework document. Which executive level stakeholder owns which of these capabilities depends on the organization. Often, multiple stakeholders have a shared interest in one or more capabilities. To support navigating this document, we provide a list of typical stakeholders that are concerned with a perspective:

- **Business perspective:** This perspective helps ensure that your AI investments accelerate your digital- and AI-transformation ambitions and business outcomes. In particular, we enrich many of the capabilities in this perspective to explain and share how to make AI center-stage, reduce risks, and increase outputs and outcomes for customers, effectively enabling the formulation

of an AI strategy. Common stakeholders include the chief executive officer (CEO), chief financial officer (CFO), chief operations officer (COO), chief information officer (CIO), and chief technology officer (CTO).

- **People perspective:** This perspective serves as a bridge between AI technology and business, and aims to evolve a culture of continual growth and learning, where change becomes business-as-normal. We are extending the AWS CAF by zooming in on capabilities that most impact future competitive advantage in the age of AI: the right talent, the language it speaks, and the culture that holds it together. Common stakeholders include the chief human resources officer (CHRO), CIO, COO, CTO, cloud director, and generally other cross-functional enterprise-wide leaders.
- **Governance perspective:** This perspective helps you orchestrate your AI initiatives while maximizing organizational benefits and minimizing transformation related risks. We pay special attention to the changing nature of the risk and therefore the cost that is associated both with the development as well as the scaling of AI. Additionally, we introduce a new CAF-AI capability to this perspective: The responsible use of AI. Common stakeholders include the chief transformation officer, CIO, CTO, CFO, chief data officer (CDO), and chief risk officer (CRO).
- **Platform perspective:** This perspective helps you build an enterprise-grade, scalable, cloud platform that enables you both to operate AI-enabled or infused services and products, but also provides you with the capability to develop new and custom AI solution. We enrich the capabilities to shine light on how AI development is different from typical development tasks and how practitioners can adapt to that change. Common stakeholders include CTO, technology leaders, ML operations engineers, and data scientists.
- **Security perspective:** This perspective helps you achieve the confidentiality, integrity, and availability of your data and cloud workloads. We largely rely on the best practices from the AWS CAF here but extend on how you can reason about the attack vectors that can affect AI systems and how to address them through the cloud. Common stakeholders include chief information security officer (CISO), chief compliance officer (CCO), internal audit leaders, and security architects and engineers.
- **Operations perspective:** This perspective helps ensure that your cloud services, and in particular your AI workloads, are delivered at a level that meets the needs of your business. We provide guidance on how to manage operational AI workloads, how to keep them operational, and how to ensure reliable value creation. Common stakeholders include infrastructure and operations leaders, ML operations engineers, site reliability engineers, and information technology service managers.

For each of these perspectives, there is a natural or logical order by which the capabilities are addressed or improved that orders your areas of action for your AI transformation journey in time. The following image depicts a sample order and an assessment together with experienced implementors of AI Strategies. It is best used to establish which of these capabilities already exist in your organization and how mature they are.

Figure 5: The AWS CAF-AI foundational capabilities ordered by maturity and evolution

Business perspective: The AI strategy in the age of AI

While the cloud enables organizations to innovate at an accelerated pace, new technological paradigms, such as AI and ML, enable net-new organizational capabilities, products, and services. For decades, those business problems where the decision-making process was complex, the data that informs it was unstructured, or where the environment of the decision was constantly changing, had proven elusive to be solved through the methods of computer science.

The recent advances in ML have changed this and suddenly, problems that require machines to see, or understand language, or learn from past data and predict outcomes can be addressed. The newly and readily available ML capabilities are questioning long standing market hypotheses of established organizations, such as, automotive companies that shy away from driver assistance and automation. This perspective therefore addresses those capabilities that directly enable the business to make the most of these use cases.

Foundational Capability	Explanation
Strategy Management	Unlock new business value through artificial intelligence and machine learning.
Product Management	Manage data-driven and AI infused or enabled products.
Business Insight	The power of AI to answer ambiguous questions or predict from past data.
Portfolio Management	Identify and prioritize high-value AI products and initiatives that are feasible.
Innovation Management	Question long-standing market hypotheses and innovate your current business.
New: Generative AI	Leverage the general-purpose capabilities of large AI models.
Data Monetization	*This capability is not enriched for AI, refer to the AWS CAF.*
Strategic Partnership	*This capability is not enriched for AI, refer to the AWS CAF.*
Data Science	*This capability is not enriched for AI, refer to the AWS CAF.*

Strategy management

Unlock new business value through artificial intelligence and machine learning.

Machine learning enables new value propositions that in turn lead to increased business outcomes, such as reduced business risk, growing revenue, operational efficiency, and improved ESG. Therefore, start by defining a business- and customer-centric north-star for your AI adoption and underpin it with an actionable strategy that moves step by step to adopting AI technology. Make sure that any adoption strategy is based on tangible (short term and measurable) or at least

aspirational (long term and harder to measure) business impact that capitalizes on these new capabilities. Factor in both short-term as well as long-term impact of adopting AI.

Work backwards from existing business and customer problems and the effect that AI can have on them. When moving closer to prioritizing AI opportunities, address how and what data will fuel the systems capability. Consider from the start the self-reinforcing properties of a data flywheel on any ML product or service, where new data leads to an improved system that grows your customer base, increasing the amount of data your business benefits from.

While building such a flywheel, consider if the data you acquire can provide a defensive moat around your value proposition (something that is rare and costly). Given the broad impact AI technology already has on the market landscape, consider that your customers are likely to raise their expectations towards your products and services capability in the near future and that AI capabilities are a part of that expectation.

For each opportunity, ask if you need to build, tune, or adopt an existing AI system. For example, if you expect to use the broad emergent capabilities of foundation models but lack the capabilities to create them from scratch, focus on customizing them for your specific needs. If your ambition is to create a domain-specific general system to propel your business, invest in the data foundations.

Product management

Manage data-driven and AI infused or enabled products.

Building and managing AI-based products can be a significant challenge as the development and lifecycle of AI systems differs from traditional software and cloud products. Both the development as well as the operation and continuous creation of results (such as direct predictions) of any AI-based product include potentially costly uncertainties that require specific mitigation strategies.

When building or embedding AI into products, work backwards from your customers' and users' expected value gain, and map measurable business proxies to individual decision points that an AI system can support, enrich, or automate. For each of those, define potential metrics in the ML solution domain (such as how the value gain of detecting fraudulent transactions in the financial sector translates to expected monetary gain and a correlating precision or recall of an ML-enabled transaction classifier) and the corresponding ML problem (such as a classification problem, an intent extraction problem, generative AI and many more). Together, these formulated ML problems and their individual solutions form the value-gain that ML brings to your product.

Crucially, these ML solutions impose certain data requirements on you and your product, hence you must investigate the 4 V's of Data for each of them. While you build this knowledge bottom

up, make sure to involve business, data, executive, and ML stakeholders in the assessment of your solution. Since ML products fuse data, domain, and technology into one predictive and sometimes prescriptive system, all of them are needed. Pave the path to evolve your AI-based product through a proper lifecycle management, factor in how users interact with probability based output from AI-systems (such as gracefully fail when the confidence of the system is low) and consider what the impact of your solution is when adopted to make sure you use AI responsibly.

Condense your understanding of which questions are critical to properly scope the ML capabilities of your product and improve your product management capability for AI. This means, for example, to take an experimental, often time-bound approach to de-risking the ML component and considering from the beginning how learnings from these experiments translate into a production-grade system. Equally it means designing feedback loops into the information flow of the system (or explicitly preventing them). Over time, enable the broader organization to build new AI products, based on the output of other ML systems through technologies such as data mesh (also see DataZone) and data lake architectures and by establishing proper knowledge transfer between teams and product groups (implemented such as through SageMaker AI Model Cards).

Business insights

The power of AI to answer ambiguous questions or predict from past data.

Business intelligence (BI), mostly including descriptive and diagnostic analytics, is frequently where companies begin their journey when preparing to use AI. However, beyond descriptive and diagnostic analytics, ML enables predictive and even prescriptive capabilities and together they form the AI journey. It is crucial to acknowledge that the scope of analytical and BI units has been a different one than what is expected organizationally from AI-driven ones.

Today, many companies require subject matter experts (SMEs) to sift through insights and pull out the cause for certain observations in the data (the *why*). However, using AI techniques, BI is starting to augment these SMEs and give them new insights to incorporate into their thought process by identifying the *why* and the *what if*. With this, data and AI suddenly becomes the driver for predictive decision making.

When preparing the transition of your BI practice to an AI-enabled one, and to higher level analytics in general, a great way to push the boundaries is using algorithms with diagnostic analytics to help you understand key variables or root causes influencing your problem statement. Make sure that organizational maturity in analytics is not siloed with each subsection of the organization and ponder how you can cross-pollinate your more mature organizations with the less mature to accelerate your AI journey.

In the early stages of transformation, any effective method can be to create a center of excellence for analytics (not necessarily AI) that is closely tied to your cloud initiatives. Such a center of excellence (COE) can provide immediate value through democratized access to data-driven predictions and analysis and propel your larger ambitions. Most importantly, create a rhythm of using AI to inform major business decisions as this will drive the recognition of its value to a true business outcome.

Portfolio management

Identify and prioritize high-value AI products and initiatives that are feasible.

The challenge of ML initiatives is that short-term results must be shown without sacrificing long-term value. In the worst case, short-term thinking can lead to technical AI proofs of concept (POCs) that never make it beyond that technical stage because they were focused on irrelevant business technicalities. Your first goal when identifying, prioritizing, and running ML projects and products must be to deliver on tangible business results.

Starting somewhere is crucial, and small wins can drive faith in your organization as it helps people connect to where they could use AI in other portions of your business. At the same time, consider what larger customer and business problems you are solving through multiple AI projects and products and combine those into a hierarchical portfolio where the lower layers of that portfolio enable the upper layers. Certain AI capabilities simply can't be built in one go. Rather, they build upon each other. For example, in the financial industry, before being able to recommend new products to customers, you must be able to categorize what is important today, so classifying transactions precedes next-best-offer actions. Each layer of your portfolio should add additional value to the organization at large.

Next, embed in this portfolio the design of an AI flywheel where the value that your portfolio provides propels business outcomes that, in turn, enable and create additional data from which your portfolio benefits (see Figure 6). This flywheel does not need to be on a single-product level but can reach through your portfolio. As your portfolio evolves and scales, prioritizing what to buy versus what to build becomes crucial. Push back on the *not invented here* syndrome.

Exploring which use cases and which solutions already exist in the market, and at what maturity level, should not be an afterthought. Also investigate which solutions require custom modeling, and raise your AI workforce's efficiency by choosing the right AI products and cloud environment. Realize how complex it is to even just technically govern your portfolio. To make sure you keep your scarce AI workforce efficient, be decisive and bold, and push back on analysis paralysis.

Finally, as your portfolio grows and more parts of the organization start to use AI, enable efficient collaboration between your business units, teams, and AWS partners that you rely on (see AWS DataZones, AWS Redshift and AWS CleanRoom) .

Innovation management

Question long-standing market hypothesis and innovate your current business.

As mentioned in the introduction to this perspective, ML offers new capabilities to businesses that can be and in many cases are disruptive to existing businesses and value chains. The power of this general-purpose technology is seen and felt across sectors and there is virtually no exception to that, as the long-term goal of AI-research is to replicate or at least imitate intelligence. The historically human capability to do knowledge work and process complex information, reason and derive insights, and take actions can now be tackled by advanced foundation models and generative AI. In your innovation roadmap and your innovation management practice, bridge to these mid- and long-term goals of AI research through short-term and real-world applicable value propositions.

To do this, start by exploring the evolving customer expectations and needs, both from an internal and external perspective. The business outcomes that the CAF-AI suggests can guide you in identifying these needs and expectations. Consider the value chain of ML-enabled or infused products, and differentiate between innovation for cost reductions (such as process improvements), revenue and profit gains (such as product improvements), or completely new income channels (such as new products and services).

Use and position ML as a unique differentiator to the respective internal and external stakeholders, and customers. Integrate ML to unlock new capabilities, augment existing ones, and reduce effort through automation. Capitalize and double down on domain-specific knowledge that is represented in the data you access. Design a healthy data value chain for your AI system to allow a long-lasting value generation. Don't get discouraged that some ML-based products only grow better over time, or that your innovation cycles might be longer than what some companies are used to. While you build up single lines of ML-enabled products, pave the way to innovation across the organization by raising data to a first-class citizen of the value-creation process and creating internal data products for consumption.

Additionally, to this top-down approach to innovation-management, get a grassroots movement going through internal AI champions. These champions can be business owners, product managers, technical experts, as well as the C-suite. Constantly keep a balance between audacious goals and the achievable ones. While typical software systems and environments grow their value with an

increasing number of users, the value of ML systems is driven largely by the data that makes it more effective. Therefore, managing AI innovation also means bringing your data strategy to life, not just archiving data that describes the past. With this growing high-quality and value data that is governed and accessible across organizational boundaries, you will create gravity for AI ideas and projects.

New: Generative AI

Use the general-purpose capabilities of large AI models.

The overall goal of AI is to create systems that are of general quality and can be applied to many complex problem spaces with little to no additional cost. One particularly powerful stream of this work is generative AI, a type of AI that can create new content and ideas, including conversations, stories, images, videos, and music. Generative AI is powered by very large models that are pre-trained on vast amounts of data and commonly referred to as foundation models (FMs). The potential of these FMs lies in their capability to generalize across domains and tasks. Such foundation models will, one way or another, influence your organization and business as they reduce the cost of knowledge work dramatically. When planning to adopt this powerful branch of AI, there are three considerations. Do you require to build such FMs:

1. From scratch and uniquely tailored to your business?
2. Fine-tune a pre-trained model and capitalize on the abilities it has already learned?
3. Use an existing FM from a supplier without further tuning?

Having the choice between these three is essential and the correct choice depends on your business case. Very often unlocking the true value of these large models means contextualizing them with your domain specific data (case 2) and applying them to a wide variety of tasks. This is the case because large and pre-trained models already possess emergent capabilities (for example, reasoning) that are costly to produce from scratch (case 1). Therefore, when using foundation models and generative AI, capitalize on a pre-trained model's ability to adapt and learn from little to no data.

For many businesses, this approach means selecting the right foundation model for their business problems and customizing (for example, through instruction tuning and few-shot learning) and fine-tuning them with domain- or customer-specific data. The effectiveness and differentiating capabilities of generative AI and foundation models, just as with other AI systems, will largely rely on your data strategy and data flywheel. Whichever path you choose, verify that you

are comfortable with the data you use, as the data influences how the model will behave in production, and establishing guardrails around generative AI systems is significantly hard.

People perspective: Culture and change towards AI-first

Adopting AI and creating value reliably and repeatably is not a purely technological challenge. Any AI initiative is crucially dependent on the people that guardrail and drive it. While AI as a general-purpose technology will impact sectors, those organizations where the workforce embraces its capabilities will be successful. This becomes all the truer when considering how good AI systems come to live: Through collaboration between stakeholders, business units, and practices.

There often is talk about the potential of AI to automate human labor, when in reality it enriches, supplements, or even augments human labor. While some domains are in reach for automation, today's AI is largely about helping with tasks that are perceived as specifically complex for humans. We see that organizations that are AI-first reduce operating costs, increase revenue, and give challenging, meaningful work to employees. Rallying the organization, building up the right talent, and speaking the same language when searching for valuable business problems is the focus of this perspective. *Culture is king*, even more so when adopting AI. This perspective comprises seven capabilities shown in the following table. Common stakeholders include the CIO, COO, CTO, cloud director, and cross-functional and enterprise-wide leaders.

Foundational Capability	Explanation
New: ML Fluency	Building a shared language and mental model.
Workforce Transformation	Attracting, enabling, and managing AI talent—from user to builder.
Organizational Alignment	Strengthening and relying on cross-org anizational collaboration.
Culture Evolution	*Culture is king*, even more so when adopting AI.
Transformational Leadership	*This capability is not enriched for AI, refer to the AWS CAF.*

Foundational Capability	Explanation
Cloud Fluency	*This capability is not enriched for AI, refer to the AWS CAF.*
Organization Design	*This capability is not enriched for AI, refer to the AWS CAF.*

New: ML fluency

Building a shared language and mental model.

The boundaries and semantic scope of artificial intelligence and machine learning is not well specified. Both terms are also overloaded with varying mental models and emotional interpretations, which is why it's key to align internally on what stakeholders mean by it. Spread a largely aligned perspective on what these words mean and identify those stakeholders that are intrigued by it as your future internal AI champions.

Once that first layer of interpretation is spread across the organization, tackle the second, more technical one: AI projects and requirements differ in terminology and what importance is assigned to them. From the product management practice to the engineering and data science practice, align on what joint understanding is needed to work effectively. An effective way is to define interface words between different practices, for example, how can success be measured in ML versus how can it be measured in the business domain.

Implement these alignments through ML fluency and ML culture trainings, as they will help you get buy-in throughout your organization. It's likely that this understanding will become crucial in helping business owners adapt to the unique aspects of ML use cases and setting expectations with customers.

Lastly, consider how to best communicate AI outputs both in the organization and to customers. Consider that customers will have different mental models and terminology, so such as, letting an AI system gracefully fail and keeping trust is challenging. With the right language and fluency, you will not only be more efficient but also reduce the risk of building systems that don’t align with interests of your customers.

Workforce transformation

Attracting, enabling and managing AI-talent from user to builder.

Being able to attract, retain, and retrain talent that can push your AI strategy forward is one of the most crucial aspects of AI success. There are many roles that are necessary for AI success, some of which you can outsource while others can only have their impact as the in-house workforce. As a first step, your AI strategy leaders need to be tightly connected to your business and drive value from within. This role can seldom be handled by an outsourced firm.

Enable these leaders by hiring or developing the many roles that are needed for successful AI adoption:

- Technical talents (such as data scientists, applied scientist, deep learning architects, and ML engineers.
- Non-technical product talents (such as ML product managers, ML strategists, and ML evangelists) that manage roadmaps and identify needs.

Tightly align your hiring strategy with your overall AI strategy and ambition:

- PhDs with years of experience might be appropriate for scientifically ambitious large-scale initiatives, though it's best to complement them with business-close counterparts, such as ML strategists.
- Transitioning some of your existing talent to AI roles is beneficial for organization-wide adoption.
- Hiring ML engineers and deep learning architects is most reasonable when you plan to base your AI capabilities on established solutions, foundation models, or AI work that is outside the reach of your organization.

In addition to this internal workforce, bet early on the right AWS Partners to not fall prey to your AI agenda fizzling out. When talent is not present, broadcast your AI vision externally and start to run initiatives that will yield both results and inspire new talent. Recognize from the beginning that retaining talent in AI is difficult, as supply has historically been outstripped by demand. Another factor is that real-world AI differs significantly from the academic work that often drives talent into AI. Counter this factor by having opportunities for your AI experts to collaborate, present at conferences, and write whitepapers.

Attrition, however, is unavoidable. Be flexible and establish processes to hire talent with proper timing and to keep resources on deck to fill in when attrition occurs. The processes we reference in other parts of CAF-AI are crucial to helping make your business robust against attrition. Fuel your AI workforce through continuous re-training opportunities to learn new skills needed to perform well in the AI space. This approach has an added advantage of being able to have a person that

has in-depth business knowledge as well as being able to run projects. Lastly, recognize that the headcount-to-value ratio in AI is lower than in other fields. A small team of strong practitioners typically outperforms larger teams as the work is less mechanical than intellectual.

Organizational alignment

Strengthening and relying on cross-organizational collaboration.

When AI becomes top-of-mind for organizations, providing an encapsulated and empowered separate unit that spreads and disseminates its value and knowledge across the organization is a typical first step. The AI center of excellence (COE) is a unit that can fill this role, where AI-focused teams are hired and evolved. Make sure that reporting lines in this organization align with those stakeholders that have ownership over the AI strategy in the organization and make sure that there are short paths to the C-suite. Do this to make sure decisions and changes can be made quickly when needed, and new teams can find their rhythm. At the same time, it's crucial to align the incentives of such a COE with your strategy, business, and most crucially your customers. A common mistake is to evolve AI units that do not deliver on business value.

Over time, your workforce transformation should enable your broader organization and other builders to effectively use the COE and existing AI services, as well as collaborate effectively. Be sure to prevent a *not invented here syndrome*, so the organization does not rebuild what is readily available in the cloud, provided it fulfills your business requirements. Make sure that your COE and talent develop an engineering mentality, recognize the cost of maintaining disparate systems, and establish an MLOps best practice that brings a DevOps mentality to the culture. While such units, other internal builders, and AI talent evolves, enable your data flywheel by establishing a data-driven product mentality. Permit businesses across the organization to not only share and govern data, but also establish a vivid ecosystem of data products. However, don’t build such data products for their own sake.

Culture evolution

***Culture is king*, even more so when adopting AI.**

Developing an AI-first culture is a long and challenging process as it often requires breaking up old mental models. In typical cloud and software development, the cultural focus is on empowering builders to codify complex rules and systems. AI relies much more on a culture of searching for the right inputs that generate the desired output. To circumvent a culture that is centered around technology, embrace a mentality where builders, the business, and other stakeholders work backwards from business opportunities and customer needs to all the AI challenges.

Working backwards means pre-formulating the expected result of a change in your business environment and then asking what needs to happen to achieve that change. In a way, this is how AI systems are built: Defining the expected outputs, and then searching for inputs that contain a signal to enable that output.

With such a value-driven mindset in place, zoom in on the cornerstones of an AI-first culture:

- Experimental mindset paired with agile engineering practices
- Cross team and business unit collaboration and reliance
- Bottom-up and top-down AI opportunity discovery
- Broad and inclusive AI adoption solution design driven by customer value

Start expanding your AI-first culture with the following:

- Empower your builders to experiment with AI systems, not for experimentation's sake, but because building an AI system involves exploring which solution pathways work and which are dead ends. It's helpful to consider the reduced risk in adopting existing AI services where the pathway is known.

  While you allow experiments, adjust your agile mindset toward the uncertainties of AI. Recognize that you can't reliably define a time-effort estimate for complex projects, since many complex AI problems with high business value have yet to be solved. When this is the case, double down on those where the expected customer value is the largest.
- Embrace a culture where data is the interface between teams and value is created in tandem with each other. Be careful not to build business-distant data science teams, but a culture where you create a flywheel of collaboration.
- Empower a culture where value is identified, recognized and enabled at all levels of the organization. This includes leadership incentivizing and elevating challenging the status quo.
- Build an environment where concerns about the impact and use of AI are not just heard but influence the decision-making process.

Governance perspective: Managing an AI-driven organization

Managing, optimizing, and scaling the organizational AI initiative is at the core of the governance perspective. Incorporating AI governance into an organization's AI strategy is instrumental in building trust, enabling the deployment of AI technologies at scale, and overcoming challenges to

drive business transformation and growth. By driving consistency, AI governance enables alignment with organizational goals, and ensures that AI technologies are ethically used and effectively managed. To that end, AI governance frameworks create consistent practices in the organization to address organizational risks, ethical deployment, data quality and usage, and even regulatory compliance, as well as managing the different cost patterns of AI workloads. This creation of scalable processes and standards for AI deployments allow organizations to expand initiatives across business units to create long term business value.

Building an AI governance practice requires close alignment with the organization's AI strategy. The first step is to identify all the key stakeholders and bring together a team with representation from multiple business units. This team will be responsible for:

- Defining governance goals, including compliance and ethical goals as well as identifying areas of potential risks.
- Developing policies and guidelines to include data, transparency, responsible AI and compliance.
- Defining mechanisms to monitor AI systems, performance, compliance, bias and determine actions based on predefined thresholds.
- Continuously revise results and existing policies to ensure alignment with business goals, and AI safety.

In this perspective, we describe some solutions to governance challenges and introduce a new capability: The Responsible use of AI, a decisive element for future competitive advantage in the AI space.

Foundational Capability	Explanation
Cloud Financial Management (CFM)	Plan, measure, and optimize the cost of AI in the cloud.
Data Curation	Create value from data catalogs and products.
Risk Management	Leverage the cloud to mitigate and manage the risks inherent to AI.
Responsible use of AI	Foster continual AI innovation through responsible use.

Foundational Capability	Explanation
Program and Project Management	*This capability is not enriched for AI, refer to the AWS CAF.*
Data Governance	*This capability is not enriched for AI, refer to the AWS CAF.*
Benefits Management	*This capability is not enriched for AI, refer to the AWS CAF.*
Application Portfolio Management	*This capability is not enriched for AI, refer to theAWS CAF.*

Cloud Financial Management (CFM)

Plan, measure, and optimize the cost of AI in the cloud.

Managing AI projects in the cloud involves planning for the cost structure of training and inference. This is important to consider in advance when budgeting for individual projects as well as for the overall funding of AI initiatives. An example of such a cost structure over the AI lifecycle, are *zig-zag* costs or phases of low/high/low/high costs:

- You might start off with a high initial cost to establish or increase the quality of the data that is needed to build your solution. However, if the data is ready, this initial cost may be very low. This is followed by a potentially volatile proof-of-concept phase.
- While most AI proof-of-concept (POC) initiatives may be relatively low-cost compute-wise, there are a few technical approaches that can quickly become costly, such as the training of larger models (in the context of generative AI) or constant retraining for domain-specific ML models. In such cases, you can leverage purpose-built AI hardware likeAmazon Elastic Compute Cloud (Amazon EC2) Trn1 instances powered by AWS Trainium or Amazon EC2 Inf2 instances powered by AWS Inferentia2 to help keep costs low. If you have access to the right talent, AI services, and AWS Partners, leverage their expertise to estimate the resources needed for different phases of your use cases and overall AI strategy. If feasible, work on calculating what an incremental improvement of an ML metric is worth to decide how to optimize your investment.
- After the first iteration of the system is built, the next phase of building a minimum viable product (MVP) may have a relatively high cost; for example, to generalize the system's capability

or acquire edge-case and long-tail data that is crucial for user adoption. If you are working on a use case that requires generative AI capabilities, you can evaluate using or fine-tuning foundation models, since that can have significant positive cost impact, as the initial training costs have been absorbed by your supplier or vendor (for example, Amazon BedrockTitan Foundation Model).

- After AI models are deployed, inference itself is largely dependent on the volume of requests, and in many cases the inference cost itself is relatively low. If not, you can leverage the purpose-built AWS Inferentia architecture. At this stage, monitoring model metrics and flagging drift alerts you to changes and the potential need to retrain your algorithms. You can leverage the low costs of scaling in the cloud. Throughout the AI lifecycle, it is important to track costs and tag all resources and ML workloads.

Once you have cost-visibility measures in place, it is critical to analyze the data , training , and inference costs over time . There is a large quantity of problem types (text, forecasting, document processing) , which in their infancy do not cost much , but their costs grow linearly with data size. There are other AI problems that rely on audio and voice data that have a much higher start-up cost and need well-defined goals even in the POC phase to not cause unexpected charges. Aligning your AI vision with the business goals should inform how you scope the work , and establishing mechanisms to calculate the tradeoffs between model costs and model performance is critical for maintaining positive ROI. Additionally, the cost of data acquisition is strongly influenced by the mechanisms that organizations establish around their data process. A standard process around acquiring new data, and master data, is key to keeping costs down, just as much as keeping data in formats where it can be used for AI (with reduced copy/read/copy or ETL needs). The cloud helps with all of these challenges through governed data-services and zero-ETL patterns .

Beyond this, always connect your AI initiative to an underlying business goal. If it relates to a new revenue stream, assume how much revenue will likely be associated to what success criteria and translate business value into your AI metrics. Factor in the often-underestimated cost of not recognizing the need for the responsible use of AI. Due to its importance, we have added the Responsible use of AI a s a new capability later in this perspective.

Data curation

Create value from data catalogs and products.

Your ability to acquire, label, clean, process, and interact with data will increase your speed, decrease time-to-value, and boost your model's performance (such as accuracy). When models stall

for accuracy, consider going back and enriching, growing, or improving the data you are feeding the algorithm. Doing so is often much easier than rearchitecting or squeezing out that next percent of performance with modeling alone.

Collecting data with ML in mind is crucial to achieving your AI roadmap and you should ask yourself and other leaders: "Are we enabling AI innovation through democratizing data?", "Does my organization think of my data as a product?" and, "Is my data discoverable across my organization?" While answers to these questions often sit on a spectrum between yes and no, the key thing to remember is that it's all about reinforcing a culture where data is recognized as the genesis of modern invention. Treating data as code and making it a first-class citizen in your business should not be an afterthought.

Data quality assessments and rules around the governance can either accelerate the use of your data or stop all progress. Balance these two and use proper tooling to allow your whole organization to innovate. Have direct owners of datasets or data stewards, which in turn helps you build a robust data ecosystem. Start small and then continually add to your data mesh, as this keeps the data flywheel spinning. Have your data accessible and discoverable by different means for different user types. This approach allows you to have greater visibility into work happening in your environment and avoid shadow DataOps.

Easy to use human readable data repositories, catalogs and dictionaries, can provide a centralized and organized repository of data and metadata about the organization's data assets, which empowers teams of all skill levels to discover, understand, collaborate on data, and start using your data to create business value. This increases the speed to decide upon the additional investment cost needed for other use cases considerably. There are many ways to go about increasing your data's potential, such as buying external data sources, augmenting or creating synthetic data through ML algorithms, crowdsourcing a team to label your internal data, or even changing your business practices to automate data generation and capture. It is strategic to develop practices to decide when to use each of these resources.

Risk management

Use the cloud to mitigate and manage the risks inherent to AI.

While every new technology comes with a new set of risks, managing the risks involved both in the design and development process of AI systems as well as in the deployment and long-term operations and application of AI is challenging due to the non-deterministic nature of AI models. Some risks are financial. Start by factoring in the risk of sunken cost into the development process as the outcome of an AI development initiative is hard to guarantee upfront (the nature

of optimizing a system for output compared with specifically building it to do so). Establish solid practices, such as model cards and adversarial inputs, and mechanisms such as POCs, minimum loveable products (MLPs), and MVPs, to mitigate and control risks.

Other risks are of legal and ethical nature. This includes risks as classified by your local legislature, for example, the European Union and those that are inherent to AI, such as a hidden feedback loop or misinterpretation of uncalibrated outputs, and unexpected outcomes that may impact different groups of people negatively. Also consider its professional, organizational, and even societal use and impact (such as echo chambers or long-term impact on customer behavior). For more information, see Responsible use of AI .

Developing and adopting safeguards and architectures that constrain the system when necessary, not just in safety-critical environments is a priority. Make sure that subsystem failures don't propagate and compound downstream AI systems. Consider which themes are relevant, such as explainability, transparency, and interpretability. Manage these risks, not just for a single AI-influenced decision or action, but across the process or larger system you act in. Capture the long-term challenges that drift of data and concepts in the world can have on your system and invest into hardening them against bad actors (see Security perspective: Compliance and assurance of AI/ML systems). Lastly, don't minimize the complexity of reaching human-level parity in certain domains.

Responsible use of AI

Foster continuous AI innovation through responsible AI practices.

Until recently, the responsible use of this powerful new technology was often an afterthought as organizations focused exclusively on the technical aspects of developing AI solutions, and the specific business goal desired. However, the recognition that AI systems learn from vast amounts of data, and what the system learns is not always what you might have intended, has made it critical to focus on Responsible AI practices. Responsible AI practices are key for fostering continuous AI innovation and to ensure that AI solutions are developed, deployed, and used ethically, transparently, and without bias. The broader the use and impact of your application, the more important it becomes. Therefore, consider and address the responsible use of AI (RAI) early on in your AI journey and throughout its lifecycle.

Establish an AI governance board with representation of multiple business units (like research, human resources, diversity and inclusion, legal, government and regulatory affairs, procurement, and communications) to work closely or as part of AI leadership teams to ensure AI solutions are safe and cause no harm to employees, customer, and society at large. This board should be

responsible for overseeing and guiding the ethical and responsible development , deployment, and use of AI technologies, and for driving alignment with industry regulations and compliance with AI-focused legislation. Scale how Responsible AI impacts your design, development and operations over time. Consider how your system affects individuals, subgroups of the population, your users, customers, as well as society. Given the speed at which AI can be scaled in the cloud, you need to consider how key responsible AI dimensions like explainability, fairness, governance, privacy, security, robustness, and transparency are being included, as well as how different cultures and demographics are impacted by the technology. Make it a key part of your AI vision, including well thought-out principles and tenets around the responsible use of AI and how it affects your initiatives. In particular, include algorithmic fairness, diverse and inclusive representation, and bias detection.

Embed explainability by design into your AI lifecycle where possible and establish practices to recognize and discover both intended and unintended biases. Consider using the right tools to help you monitor the status quo and inform risk. Use best practices that enable a culture of responsible use of AI and build or use systems to enable your teams to inspect these factors. While this cost accumulates before the algorithms reach production state, it pays off in the mid-term by mitigating damage. Especially when you are planning to build, tune, or use a foundation model inform yourself about new emerging concerns like hallucinations, copyright infringement, model data leakage, and model jailbreaks. Ask if, and how the original vendor or supplier has taken an RAI approach to the development as this trickles down directly into your business case.

> **Note**
>
> The AWS Responsible Use of AI team has written a whitepaper on this subject.

Platform perspective: Infrastructure for and applications of AI

With advances in AI and ML algorithms and their use cases, the systems and processes that are in place to run them may quickly become obsolete. Just as in any efficient manufacturing process, you need systems and platforms for AI development that ensure a uniform and consistent product. This product is an algorithmic result that drives business value. Developing a platform that is aligned to supporting your foundational capabilities helps carve out competitive advantages and accelerates innovation. A risk-reducing platform is reliable, extensible, and delivers on the promise of foundational capabilities that enable long-term business value in alignment with other perspectives in this paper.

AI-enabling platforms need to be guided by a set of design principles that keep components aligned in their purpose and intent, ensuring coverage of all aspects of the ML lifecycle over time. Central to this is the management and access of distributed and governed data, which is prepared and offered in ways that meet the specific needs of individual consumers. Furthermore, there's a requirement to support the development of novel AI systems through a comprehensive end-to-end development experience. It's also essential to harness existing AI capabilities and foundation models. After these models are trained, they can be orchestrated, monitored, and subsequently shared for integration into applications, systems, or processes with downstream consumers. These activities are overseen by platform enablement teams who continuously iterate on the feedback provided, aiming for consistent improvement.

Foundational Capability	Explanation
Platform Architecture	Principles, patterns, and best practices for repeatable AI value.
Modern Application Development	Build well-architected and AI-first applicati ons.
AI Lifecycle Management and MLOps	Manage the lifecycle of machine learning workloads.
Data Architecture	Design a fit-for-purpose AI data architecture.
Platform Engineering	Build an environment for AI with enhanced features.
Data Engineering	Automate data flows for AI development.
Provisioning and Orchestration	Create, manage, and distribute approved AI products.
Continuous Integration and Continuous Delivery	Accelerate the evolution of AI.

Platform architecture

Principles, patterns, and best practices for repeatable AI value.

As machine learning development matures from a research-driven technology to an engineering practice, the need to reliably and repeatably create value from its application becomes more important. The goal of Platform Architecture is to consider inputs from different CAF perspectives to design a foundation that aligns with your business objectives, ensuring adoption and enablement of the AI lifecycle. Start by understanding the maturity and capabilities of your platform stakeholders and what they require from the ML-stack: Are you trying to enable consumption of prebuilt off-the-shelf AI services, Low Code and Autopilot functionalities, making AI accessible to non-experts? Or are you aiming to support expert uses along their AI development lifecycle, including usage and customization of ML frameworks where there is direct access to infrastructure? Especially as you venture into the domain of generative AI, such questions make significant differences in platform architecture. Consider the specific AI-related requirements on three layers:

1. **The compute layer:** AI can have significant hardware demands for training and inference might require massive amounts of compute resources (for foundation models). Along with guardrails for consumption, price compared to performance is one of the key factors in setting standards for your organization. Consider using specialized hardware that can drive down costs with better price performance than traditional CPUs or GPUs.
2. **The ML- and AI-service layer:** Design how your platform supports the development, deployment and iteration on ML and AI services. ML services need to enable your expert stakeholders to, for example, train or tune custom models (such as foundation models), whereas AI services should be ready to consume models and capabilities (such as large and costly-to-train foundation models in the generative AI domain). While this separation is not always straightforward, requirements differ.
3. **The consumption layer:** The downstream consumers of your AI capabilities operate on this layer. This can be as simple as a dashboard or as complex as the augmentation of a foundation model through prompt engineering, or specific generative AI architecture such as Retrieval Augmented Generation (RAG) applications.

While building up the platform, analyze industry-specific legal requirements that have implications for your data, your model development process, and your deployment (for example, required segmentation of data), and steer through guardrails accordingly. Spend time on identifying standards and publishing those for downstream teams to consume, for example on data privacy and data governance. Next, streamline the provisioning of compliant environments and infrastructure, thereby speeding up the development and deployment of new AI use cases. Plan for the integration of feedback loops onto your platform by understanding how your teams might

use Human-in-the-loop (HITL) and human-on-the-loop (HOTL) functions as they serve as vital checkpoints in AI workflows. Lastly, identify any needs for ML-specific monitoring, such as bias detection, explainability, and human reviewers when a model behavior changes.

Adopting a modular design for the AI value chain is important as it allows for independent scaling and updates. This modular approach aids in quicker data labeling and facilitates ownership and accountability of different components. When deciding on which cloud-native solutions to standardize, factors like costs, reliability, resiliency, and performance must be considered. All these best practices, along with the design guidelines and standards, should be published to a central repository accessible by all practitioners in your organization. Implementing a feedback mechanism and metrics that measure platform adoption can provide continuous insights into your AI initiatives, helping you make informed decisions.

Modern application development

Build Well-Architected and AI-first applications.

> **Note**
>
> The AWS Well Architected Framework – Machine Learning Lens is the definitive source for workload and architecture design patterns and best practices.

As AI technology matures, it touches every aspect of application development:

1. AI-enhanced application development: Enhance your Software Development Lifecycle (SDLC) with AI. Use AI services and tools to propel applications with generative and autocomplete features or streamline the review process by identifying potential code issues, automating performance and testing by ensuring efficient and error-free development. Reimagine your SDLC with AI capabilities from ideation to maintenance of the software.
2. AI as a differentiation in your product: Integrating AI into your software enhances the user experience or can even be at the core of the value-proposition. AI can elevate the software's functionality, ensuring it aligns closely with user needs and expectations, ultimately delivering a product that deeply resonates with its audience. When developing such applications, consider how data moves through your systems, how it changes your AI systems, what outputs it produces, how these outputs are interpreted by consumers and customers, and how those outputs might lead to new data that you use. Base your architectural decisions on design principles that are already well established in AI.

3. AI model development: When integrating AI into your software, evaluate whether to adapt existing models, leverage open-source options, or craft bespoke solutions. Mastering AI becomes a routine necessity as modern application development evolves. You might need more customization with your use case, using specific data and fine tuning a model fit for your need.

For all three aspects, consider how to break your application and development processes into smaller and more manageable chunks. Align a micro-services or multi-model approach with agile practices to allow for more flexibility, faster delivery, and better response to changes. This approach is particularly beneficial in AI development, where the need for iterative testing, experimentation and refinement is high. Establish a clear understanding in your development teams that AI systems are indeed perceived differently by customers and users, and that many users are lacking mental models that help them interact with these systems. This means that all AI-based applications that customers and users interact with directly benefit from a fresh look at their user experience (UX).

AI lifecycle management

The AI lifecycle management decomposes into an architecture and engineering viewpoint that mature along with organizational capabilities.

The architectural viewpoint focuses on the design, planning, and conceptual aspects of AI lifecycle management. Managing the lifecycle of machine learning workloads is a complex task that requires a comprehensive approach. It consists of three major components:

1. Identifying, managing, and delivering the business results and customer value.
2. Building and evolving the technological components of the AI solution.
3. Operating the AI system over time, also known as Machine Learning Operations (MLOps) or for larger models Foundation Model Operations (FMOps).

As each of these components are complex, we have built detailed guidance in the Well Architected Framework: ML Lens. Different AI strategies require different perspectives on these three components. For example, if your overall goal is to enable new products through custom models, you see lifecycle management differently than if you strive to increase internal operational efficiency through publicly available services. Independently of your approach, use centralized repositories and version control to store your AI artifacts and track model-lineage and data-lineage.

The engineering viewpoint emphasizes the implementation and operations of the AI lifecycle management. To streamline this process, it's crucial to implement MLOps practices that automate the deployment and monitoring of AI models to reduce labor, improve reliability, reduce time-to-deployment and increasing observability. Make sure to align on a defined process for managing the AI lifecycle from ideation to deployment to monitoring. This process should include steps for collecting and storing data, training and deploying models, and monitoring and evaluating models (operations section of the CAF-AI), including performance monitoring. This helps you discover deficiencies early on and support the continuous evolution of your models. Lastly, establish an automation framework for retraining your AI models; for example, when performance degrades or new data arrives.

To better understand where you are in relation to industry best-practices, assess your MLOps maturity with an AWS Partner or AWS, and base your decisions on a MLOps and lifecycle framework. These processes and standards are the best defense against your system relying on institutional knowledge alone, helping you reduce AI technical debt. It's common to see data teams overly focusing on hard ML metrics instead of how these metrics affect a business metric, which is a failure of lifecycle management. Whatever your path, make sure that the process and standards you establish for MLOps are repeatable. Such MLOps best practices also helps you ensure that your science team does not get modeling fatigue and focuses on results instead of getting distracted with mass parallelization of experiments.

Data architecture

Design and evolve a fit-for-purpose AI data architecture.

Data is the key to AI, and with the explosion of data types and volume, traditional data architectures need to evolve. In particular, AI demands new approaches for storage, management, and analytics to meet AIs new complexities head-on, as it’s becoming central to business decision-making. Keep in mind that AI workloads demand not just massive volumes of data but also require diverse and high-quality data for model training and validation. As such data comes from multiple sources, often in varied formats and structures, traditional data architecture, with its constraints on data movement and types, often is not suitable to manage this kind of diversity and volume efficiently. Therefore, dive deeper into the evolving field of modern data architectures. These bring together data lakes, data warehouses, and other purpose-built data stores under one umbrella and reduce the complexity of governance while enabling data movement, a critical aspect for AI.

In today’s organizations, three architectures are dominant: The data warehouse (structured and throughput optimized stores), the data lake (aggregate data from diverse silos and serve as central

data repositories) and specialized stores for business applications (NoSQL databases, search services, and so on), each supporting different use-cases. However, moving data in and out of these stores can be challenging and costly. Therefore, as data movement is becoming more important for AI systems harden your architecture for data-movement requirements:

- **Inside-Out:** Data initially aggregates in the data lake from various sources—structured like databases and well-structured spreadsheets, or unstructured like media and text. A subset is then moved to specialized stores for dedicated analytics, such as search analytics or building knowledge graphs.
- **Outside-In:** Data starts in specialized stores suited to specific applications. For example, to support a game running in the cloud, the application might use a specific store to maintain game states and leaderboards. This data is later moved into a data lake where more comprehensive analytics can be conducted to enhance gaming experiences.
- **Around the Perimeter:** This involves moving data between specialized data stores, such as from a relational database to a NoSQL database, to serve specific needs like reporting dashboards.

To keep the velocity of AI teams high, such data movements need to be possible and happen seamlessly. As AI is evolving rapidly it's key to have this flexibility. Due to the importance of data in AI, where data can be considered like machine-code, the line between AI and data architecture is blurring. Modern data architectures then enable the organization to consider data itself as a product . A modern data architecture is not a static construct; it is designed to be fluid, adapting to new data types and technologies as they emerge. Therefore, investigate different emerging archetypes of data architectures such as the Modern Data Architecture, the distributed data meshes and the data mart , and envision a unified platform or ecosystem for all types of data. Finally, regularly reflect on your current architecture and consider the access patterns and needs upfront and pick an architecture that fits the purpose. Plan to ensure your data sets are discoverable , well-documented, and easy to understand. Establish metadata principles or data documentation to describe the data, including its meaning, relationships to other data, origin, usage, and format.

Platform engineering

Build a compliant environment with enhanced features for AI.

The cloud has revolutionized the way organizations access state-of-the-art AI infrastructure and services. By democratizing access to AI, organizations can simplify their AI workflows and harness the immense advantages of economies of scale. Well-reasoned AI platforms therefore enable your AI teams to do more at lower cost. Engineer your platform accordingly and provide simplification

and abstractions for your different stakeholders (such as developers, data teams, and operations), and reduce their cognitive load while increasing their capability to innovate on their ways of working:

- **AI services:** Enable your teams by simplifying the connection between your platform and off the shelf AI services with pre-built models and specific use cases in mind, tying directly into the modern data architecture.
- **ML services:** In the cloud, developers can use specialized environments designed specifically for AI application development and deployment. When considering the training and deployment of AI models,such managed machine learning services become indispensable. They efficiently handle the intricate and often prolonged processes inherent to ML system engineering. By employing these services, you can enable your AI teams to reallocate valuable time to more strategic initiatives.
- **ML infrastructure:** Enable your teams by taking over the heavy lifting in your platform by managing the highly specialized underlying AI infrastructure. Keep in mind that AI teams are often not empowered by needing to own infrastructure but rather often are held back by it, leaving business value on the table.

One of the principal merits of the cloud is its capacity to automate routine tasks. Wherever you can, automate your ML platform tasks as it accelerates processes, reduces human error, and ensures consistency. The more complex your AI solutions become, the more a dedicated MLOps practice becomes relevant. Embed AI-specific monitoring tools in your platform from the start. Such tools track the performance of AI workloads, providing valuable insights into their operation and helping to identify any problems early on. Feedback mechanisms influence model fine-tuning and hyperparameter configurations. Monitoring workloads in real-time, organizations are better positioned to ensure that their AI applications are performing optimally and can quickly address any issues that arise.

While the cloud offers extensive flexibility, it's essential to deploy guardrails. Employ such guardrails through guidelines or restrictions set in place to ensure that developers work within defined best practices and security parameters reducing risk and ensuring responsible use. Provide a safety net, ensuring that while innovation is encouraged, it never compromises the security, compliance, or performance standards of the organization.

Data engineering

Automate data flows for AI development.

As data is a first level citizen of any AI strategy and development process, data engineering becomes significantly more important and cannot be an afterthought, but a readily available capability within your organization and teams. As data is used to actively shape the behavior of an AI system, engineering it properly is decisive. Data preparation tools are an essential part of the development process. While the practice itself is not changing fundamentally, its importance and need for continuous evolution rises. Consider integrating your data pipelines and practice directly into your AI development process and model training through streamlined and seamless pre-processing. Consider transitioning from traditional Extraction, Transformation, and Load (ETL) processes to a zero-ETL approach . With such an approach to data engineering, you reduce the friction between your data practice and your AI practice. Enable and empower your AI team to combine data from multiple sources into a single, unified view as a self-service capability. Couple this with visualization tools and techniques that help your AI and data teams to explore and understand their data visually.

Focus on your data being accurate, complete, and reliable, when possible. Design data models or transformations as part of your workflows specifically for machine learning (normalized, consistent, and well-documented) to facilitate efficient data handling and processing. This significantly improves the performance of your AI applications and reduces friction in the development process.

Provisioning and orchestration

Create, manage, and distribute approved AI products.

As the infrastructure requirements of an AI system change significantly over the course of different development and deployment stages, provisioning and orchestration deserves a second view in your existing cloud strategy. Understand where you are in the AI transformation journey and how that relates to your MLOps maturity level. Consider your consumers, data engineers, data scientists, developers, and business analysts all have varying needs and requirements when they are executing in their roles. Identify ways to provide self-service provisioning of AI environments for your different users, particularly those with limited technical expertise. Do so by creating catalogs , portfolios , and products that have been approved by Platform Architecture. Catalogs can be distributed to end users and the products within them can be consumed. Products can be defined as infrastructure as code (IaC) and deployable either through a personalized portal or CI/CD pipeline in alignment with organizational policies administered by platform teams. A common use case is to provide a personalized portal that vends predefined notebook s and compute for data teams to experiment on a new business problem, and doing so quickly without waiting for platform teams to provision resources. For more advanced roles such as data scientists requiring a

suite of tools, catalogs can be configured that deploy an entire AI environment including granting access to accelerators for foundation models .

Consider that the training or tuning step of an AI model can require high-performance compute and automate its provisioning with preapproved services that align with budgetary and governance constraints. Use API- and Framework-level automation and orchestration where you can . Design mechanisms to manage deployment of your AI workloads and streamline the creation of underlying infrastructure.

Continuous integration and continuous delivery (CI/CD)

Accelerate the evolution of AI.

There are two fundamentally different perspectives on continuous integration and delivery in the context of AI: The first is to automate and harden the model development and deployment process as much as possible, for example, for the development of custom models. The second is to use AI itself as part of the DevOps experience and make CI/CD easier through it.

Focusing on the first, organizations may automate the deployment and testing of AI models and empower their teams to innovate at the speed of the cloud . In the context of a custom model, the goal is to automate the deployment and management of AI workloads along with managing complex workflows, for data processing, model training, model evaluation, post processing, model registration, and model deployments. As you automate your AI development process, use tooling specific for ML pipelines along with methods and tools common in traditional application development. With the right architecture and blueprints, data scientists are enabled to experiment with different models and ensure that models are thoroughly tested before going into production. Take time to consider if building this capability is right for your organization. Do so by understanding the speed with which ML models are produced, the need for them to be refreshed, and the criticality and impact of your use case. As model drift can and often does happen over time, consider in how far you can automate the validation process such as by defining thresholds for retraining. Automated validation checks the performance of the models against predefined criteria, and if a model's performance drifts beyond the acceptable threshold, it triggers automatic retraining or rollback to a previous version. Lastly, by incorporating human feedback and automating tasks like model validation, testing, and retraining, repeatability improves the reliability of AI workloads freeing up valuable time for data scientists and engineers to focus on more strategic tasks. Integrating these aspects, organizations can evolve AI models in a cost-effective way while ensuring that models and the encapsulating AI system stay relevant and effective even as data and requirements evolve.

Focusing on the second, use AI itself for your AI and non-AI related DevOps activities , enrich your development process and make use of generative AI where sensible. Some of the most significant business value stems from AI being applied to the development process itself. Therefore, explore how your stakeholders can embrace it in their technical workflows. This can mean using AI for analyzing workloads for anomalies , optimize code-level performance through AI or generating code based on developer prompts . At all times, make sure your use of AI for DevOps is enterprise ready with security in mind.

Security perspective: Compliance and assurance of AI systems

Security is top priority at AWS, and all customers, regardless of size, benefit from AWS's ongoing investment in our secure infrastructure and new offerings. For customers developing AI AWS workloads, security is an integral part of the overall AWS solution. Generative AI is a key enabler in scaling Foundation Models for realizing business outcomes and there are multiple ways to create a Generative AI workload. Integrating security and privacy in all aspects of AI is critical for the overall success of business outcomes. The underlying business case of using AI is to solve specific business problems that can range from simple automation of routine productivity tasks to complex healthcare or financial decisions containing sensitive data. Apply risk management techniques to implement security and privacy capabilities defined in this perspective to meet your business needs.

Foundational Capability	Explanation
Vulnerability Management	Continuously identify, classify, remediate, and mitigate AI vulnerabilities
Security Governance	Establish security policies, standards and guidelines along with roles and responsibilities related to AI workloads
Security Assurance	Apply, evaluate, and validate security and privacy measures against regulatory and compliance requirements for AI workloads
Threat Detection	Detect and mitigate potential AI-related security threats or unexpected behaviors in AI workloads

Foundational Capability	Explanation
Infrastructure Protection	Secure the systems and services used to operate AI workloads
Data Protection	Maintain visibility, secure access and control over data used for AI development and use
Application Security	Detect and mitigate vulnerabilities during the software development lifecycle process of AI workloads
Identity and Access Management (IAM)	*This capability is not enriched for AI, refer to the AWS CAF.*
Incident Response	*This capability is not enriched for AI, refer to the AWS CAF.*

Vulnerability management

Continuously identify, classify, remediate, and mitigate AI vulnerabilities.

AI systems may have technology -specific vulnerabilities that you should be aware of; for example, prompt injection, data poisoning, and model inversion vulnerabilities. The three critical components of any AI system are its inputs, model, and outputs. These components can be protected with the following best practices to mitigate potential vulnerabilities for your workloads.

- The **Input vulnerability** relates to all of the data that has an entry point to your model. This input can be a source of targeted model and distribution drift, where a threat actor attempts to influence decisions over time, or purposefully introduces a hidden bias or sensitivity to certain data. Harden these inputs through data quality automation and continuous monitoring. Model misuse is an example of a vulnerability that results from prompt injection in AI solutions since data and instructions cloud be interlaced with each other and it's worth paying special attention to the quickly evolving field of jailbreaking foundation models. Perform input validation to segregate data from instructions and use least privilege principles by limiting access of Large Language Models (LLMs) to specific authorizations. Avoid access to system commands, executable files, and log actions that have widespread operational impact.

- The **Model vulnerability** relates to exploiting misrepresentations of the real world or the seen data in the model. Harden your model by mitigating known documented threats using threat modeling. While using commercial generative AI models, review their sources of data, terms of use for model fine-tuning, and vulnerabilities that could impact you from the model itself or from the use of third-party libraries. Validate that model goals and their results are monitored and that they remain consistent over time to avoid model drift.
- The **Output vulnerability** relates to interacting with the system over a long period of time, which can allow critical information to be inferred about the inputs and properties of your model, often called *data leakage*. For generative AI, validate that its output is sanitized and not consumed directly to mitigate cross-site vulnerabilities and remote execution. These are just a few of the vulnerabilities that you need to consider for your workloads. While not every AI system exposes these vulnerabilities, be vigilant about risks that apply to your specific workload. Perform regular testing, Game Days, and table top exercises to validate remediation prescribed by playbooks.

Security governance

Establish security policies, standards, and guidelines along with roles and responsibilities related to AI workloads.

Validate that policies are clearly defined for use of commercial or open-source models hosted internally and externally. Similarly, for commercial generative AI model usage, consider the risks from your organization's sensitive data leakage to the commercial model's platform (see Data protection capability). Understand the assets, security risks, and compliance requirements associated with AI that apply to your industry or organization to help you prioritize your security efforts. Allocate sufficient security resources to identified roles and provide visibility.

Risks associated with AI can have far-reaching consequences, including privacy breaches, data manipulation, abuse, and compromised decision-making. Implementing robust encryption, multi-factor authentication, continuous monitoring and alignment to risk tolerance and frameworks (for example, NIST AI RMF) can be essential to safeguard the integrity and confidentiality of an AI environment.

Provide ongoing direction and advice for the three critical components for your workloads:

- **The Input** – Establish who can approve data sources and the use of AI. Consider in the approval process data aspects such as data classification or sensitivity, existence of regulated data within the data sets, data provenance, data obsolescence, or the lawful right to process the data. To

manage risks, evaluate the mechanisms used to source input data considering factors such as the reputation of the source, the manner in which it was received, and how it is being stored or secured. Validate that the data classification of the source data is in alignment with the solutions' classification, such as not allowing confidential data to be processed on a public AI solution.

- **The Model** – Establish roles and responsibilities for creating and training models. Establish roles associated with an author, approver, publisher approach to model release. To manage risks, evaluate the model training mechanisms, including the tools and individuals involved, to avoid an intentional or unintentional introduction of vulnerabilities. Evaluate the model's architecture for vulnerabilities that influence the output. Enable failure modes of any model to fail to a closed or secure state to avoid data exposure.
- **The Output** – Establish lifecycle management of created outputs. Establish classification criteria, paying close attention to the outcome of datasets that may have been potentially disparate datasets or of dissimilar data classifications. To manage risks, determine appropriate protection and retention controls, classify your data based on criticality and sensitivity, such as Personally Identifiable Information (PII), and define appropriate access controls. Define data protection controls and lifecycle management policies. Establish robust data sharing protocols with privacy regulations and other compliance alignment.

Security assurance

Apply, evaluate, and validate security and privacy measures against regulatory and compliance requirements for AI workloads.

Your organization, and the customers you serve, need to have trust and confidence in the controls that you have implemented. As your customers' and users' awareness and sensitivity for AI-related risks and potential misuse increases, so do their expectations that a high security bar is met. Design, develop, deploy and monitor solutions in a manner that prioritizes cybersecurity, meets regulatory requirements, and effectively and efficiently manages security risks that are specific to AI and are in line with your business objectives and risk tolerance. Meticulously monitoring, providing transparency and collaboration between legal experts, compliance professionals, data scientists and information technology professionals will help validate a holistic approach to assurance. Implementing testing procedures and remediation processes can enable a proactive approach to assurance. Continuously monitor and evaluate for the three critical components for your workloads:

- The **Input** – Because models often require vast amounts of data for training and analysis, you need to validate the type of data ingested is aligned to the model's goals and outcomes. Establish auditmechanisms to understand adherence to the established control framework.
- The **Model** – Certify that the users understand what is acceptable usage of AI in alignment with organizational policies. Implement policies and controls to validate that the organization understands where it is appropriate to use AI and where it is not. Establish audit mechanisms to identify how the model is using data and where AI capabilities are in use within the organization.
- The **Output** – Establish acceptable usage criteria for the output paying attention to where the data may be reused or reintroduced to additional AI models. Establish discoveryor audit mechanisms to review output data to validate that generated data cannot be used to infer or recreate sensitive or regulated data. Create mechanisms for validating the authenticity and origin of the output where trustworthiness is paramount, such as medical diagnoses.

Preserving individual privacy necessitates strict adherence to ethical and legal guidelines to prevent unauthorized access, misuse, or disclosure of the data. Balancing the potential of AI and respecting privacy rights fosters public trust and realizes the benefits of these capabilities. See MLSEC-05: Protect sensitive data privacy in the Well-Architected Framework for safeguard information. Establish transparency and mechanisms such as informed consent. Limit data retention to only what is necessary for functionality and implement data sharing agreements. Again, consider the privacy requirements associated with the three critical components of your workloads:

- The **Input** – Validate that you understand how data that is subject to privacy related regulations (for example - GDPR, CCPA, COPPA, PDPA) could be used and that legal basis for processing the data exists. Consider data residency and where data is stored or processed. Establish Privacy Impact Assessments (PIA) or similar processes for each use of regulated data.
- The **Model** – As the model is trained or tuned, consider whether or not the legal basis for processing data exists and that transparency for the data subject can be demonstrated. Establish Privacy Impact Assessments or similar processes related to potential leakage from the models.
- The **Output** – Consider whether regulated data is being used to train additional models and whether restricted secondary usages of personal data limitations apply. Establish mechanisms to accomplish *right to erasure* or *right to be forgotten* type requests. Establish discovery or audit mechanisms to review output data to validate that generated data cannot be used to infer or recreate previously de-identified data.

Threat detection

Detect and mitigate potential security threats or unexpected behaviors in AI workloads.

To improve the protection of the three critical components of any ML or generative AI system (its inputs, model, and outputs) use the following best practices to detect and mitigate threats to your workloads:

- The **Input** – Detection of threats for AI solutions is critical to mitigate vulnerabilities that could impact your business. Sanitize input data to detect threats at the onset of model usage. Continue to track input data for user sessions to detect and mitigate threats that could impact availability and misuse.
- The **Model** – Conduct threat modeling specific to the AI system and threat huntingexercises to detect and mitigate potential threats. Update threat models and monitoring to include AI threat concepts that include training models with unexpected user inputs, poisoning of data sets used for content or training, privacy breaches, and data tampering. Correlate input data and data used by the model to detect anomalous or malicious activity.
- The **Output** – Monitor for output anomalies that deviate from model goals, and enable checks for detecting sensitive data in model outputs. Build a threat catalog that includes identified applicable known threats that apply to your workloads. Create automated tests to validate detection capabilities and the integration of threat intelligence to increase efficacy and to reduce false positives. Consider usage of threat intelligence to increase efficacy and reduce false positives.

Infrastructure protection

Secure the systems and services used to operate AI workloads.

MLOps uses DevOps practices for AI workloads and security needs to be applied to the infrastructure that makes up the overall environment. Use secure endpoints for your AI model and Amazon API Gateway for rate-limiting model access. Use API security best practices for all internal and external APIs used and create an explicit allow-list of API calls from models outside of its own VPC.Begin with security capabilities as prescribed by the Security Reference Architecture and apply network, compute, and storage security controls based on your environment.

Models are distributed over multiple environments across networks and servers. Communication between these environments should be protected using encryption in transit. Use centralized configuration of development and production environments and apply preventive and detective

guardrails that are managed independently by security administrators. Isolate development environments for sensitive tasks such as model training. Validate that there is session isolation for end users to preserve experience integrity and prevent unintended data disclosure. Log output responses and related session data to Write Once Read Many (WORM) storage devices for compliance and troubleshooting purposes. Consider using a model bug-bounty program to uncover and mitigate edge use cases that could cause a security issue.

Data protection

Maintain visibility, secure access, and control over data used for AI development and use.

Data protection is critical throughout the AI development lifecycle, and where data protection policies defined by security governance are operationalized, like MLSEC-07: Keep only relevant data in the Machine Learning Lens of the Well-Architected Framework. If using commercial models for generative AI development, be aware that direct use of data as input to the model could disclose sensitive information. Likewise, letting your proprietary or self-hosted models access protected data can open the door for data-related privilege escalations. Evaluate model usage and service terms accordingly.Data collected for model development during the pre-training and fine-tuning phases of model development should be secured in transit, at rest, and in use. Consider using a data tokenization process to replace sensitive data with non-sensitive data tokens as part of a data preprocessing phase that includes cleaning, normalization, and transformation. Create verifiable mechanisms for all sources of data used by the models, especially inference data that is used to train the models. Monitor and create alerts for sensitive data or data that could result in sensitivity class escalation. Employ data activity monitoring techniques to detect access patterns by usage, frequency, and so on. Avoid using sensitive data to train the models, as that could cause unintended disclosure of data from the model output (for example, through data leakage during inference). Tag and label data that is used for training in all the different environments, and align data tags and labels to data classification policies and standards. Validate that data lineage and data access in non-production and development regions is controlled to prevent data manipulation that would introduce vulnerabilities in the model. Consider using CI/CD pipelines to promote data to testing and production environments to preserve integrity. Log and mask sensitive data while creating an audit trail for data access. Implement data loss prevention techniques on sensitive data stores and on data stores that are, by design, not supposed to store data of specified data classes (for example, Confidential), and monitor for unintended disclosure of sensitive data. Validate data quality for model outputs to enable trust and avoid hallucinations. Monitor sensitivity levels of model data output and trigger re-classification by redaction or quarantined response if the sensitivity levels rise. For example, if new input data sets are used by the model or used to train the model, validate that the output data conforms to the existing sensitivity level.

Application security

Detect and mitigate vulnerabilities during the software development lifecycle process of AI workloads.

Verify that model developers execute prompt testing and other security test cases locally in their environment and also in the CI/CD pipelines to validate model usage. Create and maintain test case libraries to validate coverage and to enable automation. Leverage data and model pipelines that are integrated with security scans across all development, test, and production environments and store all of your model artifacts in secure repositories . Maintain an inventory of AI models and assign model instances with specifically identified technical and business owners. Validate that known good trained models are backed up. Retain points-in-time recovery so that compromised models can return to a known good state. Protect access to model and data backups to validate that they are not compromised, and test model recovery periodically to enable full recovery to a known good state. Track data related to the model and data development including parameters, metadata, and so on for provenance in order to support the validity of the output results. Create and use operational runbooks and test roll-back mechanisms independently for data sets and models that can be executed in case of operational or security incidents to provide resilience.

Operations perspective: Health and availability of the AI landscape

Operating ML applications is new for many customers. In the new CAF-AI capability AI lifecycle management and MLOps, we have already introduced a few perspectives and guidance on tackling this. Beyond what has already been covered, what remains are considerations around incident management and performance. To dive deeper beyond this CAF-AI perspective, we recommend reviewing the MLOps Maturity Framework and the Machine Learning Lens of the AWS Well-Architected Framework, as they both provide extensive documentation and best practices on these challenges.

Foundational Capability	Explanation
Incident and Problem Management	Identify and manage unforeseen AI behavior.
Performance and Capacity	Monitor and handle AI workload performance.

Foundational Capability	Explanation
Observability	*This capability is not enriched for AI, refer to the AWS CAF.*
Event Management (AIOps)	*This capability is not enriched for AI, refer to theAWS CAF.*
Change and Release Management;	*This capability is not enriched for AI, refer to the AWS CAF.*
Configuration Management	*This capability is not enriched for AI, refer to the AWS CAF.*
Patch Management	*This capability is not enriched for AI, refer to the AWS CAF.*
Availability and Continuity Management	*This capability is not enriched for AI, refer to the AWS CAF.*

Incident and problem management

Identify and manage unforeseen AI behavior.

AI systems are often used in situations where the expertise of a single person is not enough to grasp or solve a problem. This nature of AI systems makes it hard to understand the general behavior of the system and the edge cases, making it difficult to foresee the potentially degrading performance over time. Therefore, practitioners look at AI systems through proxies and simplified statistics. When adopting AI observing and monitoring, these simplified views into the AI system become key. This is already true in the early phases of development, but is especially important when the system is used under real world conditions.

Make sure to establish practices that acknowledge that AI systems get validated but never verified, and that they need constant and ongoing control and observation. One example is training-serving skew, where the performance of the in-lab developed AI system significantly differs from what's being seen in production. When needed, allow your customers and users to flag results as unfavorable or wrong. Open up pathways for them to engage directly to report incidents. From the beginning, prepare or a change in data and hence performance through drift, training-serving

skew, black swan events, and unobserved data-points. Where the system allows for it, provide ways to gracefully fail and to report and react to such incidents and learn from them. Anticipate that customers and users for which the system does not work well will often not be represented in the data. Finally, expect such incidents to occur and be suspicious if none are reported. Expect this challenge to grow with the size and complexity of your AI system. For example, foundation models are significantly harder to correct and monitor than simple decision trees.

Performance and capacity

Monitor and handle AI workload performance.

AI follows different development cycles than traditional software and comes with different performance and workload profiles: In the early stages of development, data is explored and cost and performance require the capability to adapt to numerous and very different workloads, often dominated by experiments and training workloads that require strong machines, specialized hardware and memory-effective architectures. Use the cloud to enable this multitude of workloads as it delivers the capability to react dynamically to these workload profiles, each of which occur sparsely and only at certain points in the development lifecycle.

Over time, training and streamlined pre-processing takes over and dominates the workload profile, becoming more consistent and predictable. Your speed of innovation will be impacted by your ability to adapt to this new profile and move quickly and continuously between the two while keeping clear lines between development and production. Make sure that model artifacts and the data that has been fueling these streamlined workloads are available for potential fallbacks. Once a model moves into a deployed and operationalized stage, make sure that the inference gets optimized for non-functional requirements (such as, latency or throughput) cost and monitoring of performance and capacity are in place. In the AI lifecycle management capability, we introduced the MLOps maturity model, refer to it for deeper operations insights. Over time, multiple types of workload-profiles will mix and mingle and are rarely comparable to the ones data-scientists experience when they develop them in isolation before launching (often called *in the lab*). Take a deep-dive into the Well-Architected-Framework and its purpose-built ML Lens that addresses how to architect such systems in the cloud.

Conclusion

In this document we gave an overview of the CAF-AI, a map of how customers can organize and structure their AI journey, what capabilities are needed to succeed, and a mental model for iterating on them. The foundational capabilities in this document are meant as an index for further investigation, learning, and conversations with your AI experts. All of them tie into the AWS CAF and enable organizations to think both about their cloud journey as well as their AI journey.

Contributors

Contributors to this document include:

- Alexander Wöhlke, Sr. ML Strategist, Generative AI Innovation Center, AWS CAF-AI Lead
- Caleb Wilkinson, Sr. ML Strategist, Generative AI Innovation Center, AWS CAF-AI Lead
- Payal Vadhani, Director Security, Professional Services
- Mayank Jain, Sr. Manager, Principal, Professional Services
- Michael Sinnwell, Sr. Security CDA, Professional Services
- Mark Lieberg, Sr. Security Consultant, Professional Services
- Matias Undurraga, Transformation Architect, Modernization Innovation Transformation
- Tony Santiago, WW Partner Solution Architect, CAF Platform Perspective Lead
- Dr. Saša Baškarada, Worldwide Leader, AWS Cloud Adoption Framework
- Neil Mackin, Principal ML Strategist, Machine Learning Solutions Lab
- Shuja Sohrawardy, Sr. ML Strategist, Generative AI Innovation Center
- Emily Soward, Data Scientist, Professional Services
- Margaret Sharp, Technical Program Manager, Engagement Security, Professional Services
- Ana Echeverri, Sr. Specialist AI Services, World Wide Specialist Organization, CAF-AI Assessment Lead
- Phil Le-Brun, Director, Enterprise Strategy

Further reading

For additional information, refer to:

- AWS Cloud Adoption Framework (AWS CAF)
- Machine Learning Lens of the AWS Well-Architected Framework
- AWS Well-Architected
- AWS Architecture Center
- AWS Prescriptive Guidance
- AWS Whitepapers

Document history

To be notified about updates to this whitepaper, subscribe to the RSS feed.

Change	Description	Date
Update	Updating and expanding the Intro section, extending the security, platform, and governance perspective.	February 13, 2024
Initial publication	Whitepaper first published.	May 22, 2023

Note

To subscribe to RSS updates, you must have an RSS plug-in enabled for the browser that you are using.

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.

AWS Glossary

For the latest AWS terminology, see the AWS glossary in the *AWS Glossary Reference*.