AWS Whitepaper

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

Copyright © 2024 Amazon Web Services, Inc. and/or its affiliates. All rights reserved.

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

Copyright © 2024 Amazon Web Services, Inc. and/or its affiliates. All rights reserved.

Amazon's trademarks and trade dress may not be used in connection with any product or service that is not Amazon's, in any manner that is likely to cause confusion among customers, or in any manner that disparages or discredits Amazon. All other trademarks not owned by Amazon are the property of their respective owners, who may or may not be affiliated with, connected to, or sponsored by Amazon.

Table of Contents

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

Accelerating Your Cloud-Powered AI Transformation

Publication date: **May 22, 2023** (*Document history*)

In this document, we outline the AWS Cloud Adoption Framework for Artificial Intelligence, Machine Learning, and Generative AI, a framework that describes a mental model for organizations that strive to generate business value from AI and ML. In this framework, we describe an AI/ML 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 CAF-AI

The AWS Cloud Adoption Framework for Artificial Intelligence, Machine Learning, and Generative AI (CAF-AI) is a starting and orientation point throughout your ML and AI journey. (We use AI and ML interchangeably in this document.) It is intended to be a document you draw inspiration from when you shape your midterm AI and ML agenda and try to understand the important topics and perspectives that influence it. It will help you in your conversation about your AI strategy with your team, coworkers, and AWS Partners.

Depending on where you are at 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), but for AI and ML, so that 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, AI and ML are 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 and ML 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.

AI/ML cloud transformation value chain

AI and ML have evolved from niche technologies to a powerful and broadly available business capability. ML is by now fueling a *new wave of innovation*, as described in the re:Invent 2022 Keynote with Swami Sivasubramanian, 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 and ML. 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/ML 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/ML does make an economical difference. As AI/ML enables solutions and solution paths to problems that have remained uneconomical to solve for decades or simply technically were impossible to tackle without AI/ML, the resulting business outcomes can be profound.

As an example, the emergent capabilities of large AI/ML models to perform domain-specific functions with little additional data are taking organizations by storm and help business to differentiate. One type of such large AI/ML models is *generative AI*, which has captured widespread attention and imagination. However, developing, applying, and tuning such models can be complex.

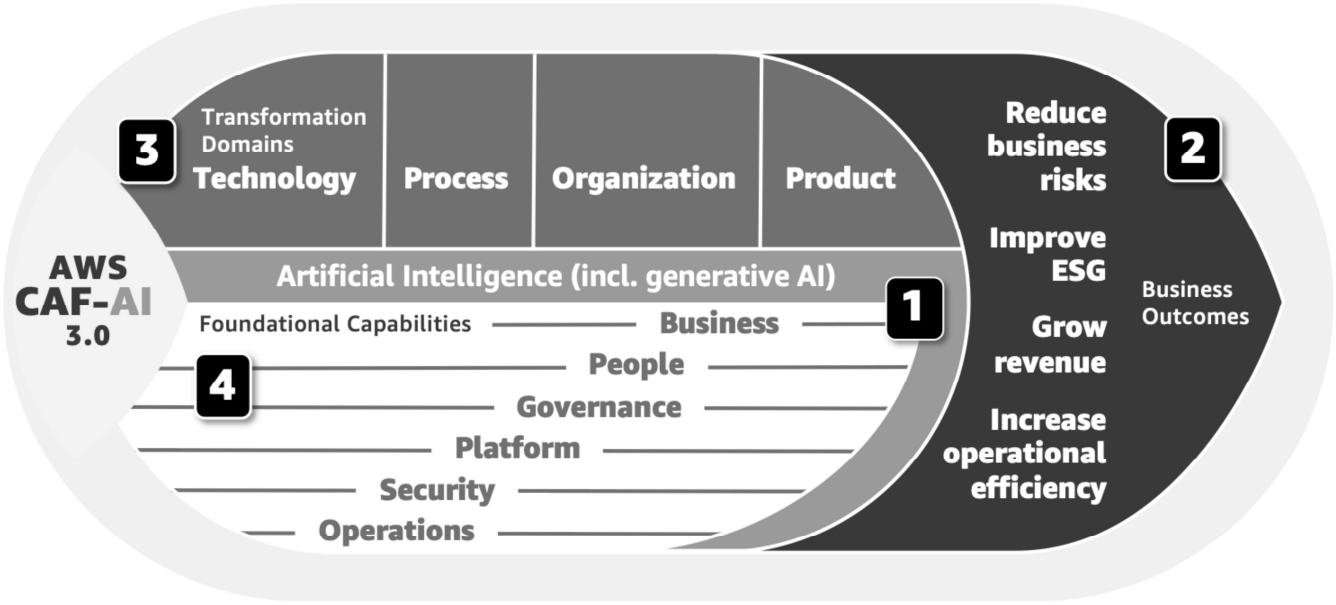

Figure 1: The AWS CAF-AI transformation value chain (all pink and magenta parts of it are dimensions taken from the original CAF that we build upon here).

The preceding figure provides an orientation on how to think about AI/ML adoption in the face of a changing market landscape and the rapidly accelerating field of applied AI/ML. AI/ML provides new capabilities to your organization (1). With these new capabilities, you and your organization strive to create tangible business outcomes (2). These 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 riles on your capability to adopt AI/ML. To adopt AI/ML, your organization needs to transform along at least four domains (3): technology, process, organization, and product:

- **Technology:** A domain that focuses on establishing the technological capability and then enabling the usage and adoption of AI/ML.
- **Process:** A domain that focuses on digitizing, automating, optimizing, and innovating on your business operations through the power of AI/ML.
- **Organization:** How your business and technology teams orchestrate their efforts to create customer value and meet your strategic intent, driven by AI/ML.

- **Product:** Reimagining your business model by creating new value propositions (products, services) and revenue models that capitalize on the capabilities of AI/ML.

Transforming these domains and enabling them to use AI/ML is depending on your foundational capabilities (4) in business, people, governance, platform, security, and operations.

To adopt AI successfully, plan out your journey:

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

Your AI/ML transformation journey

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

When embarking or advancing on your AI/ML transformation journey, consider *four critical elements*, which are also illustrated in Figure 2:

1. The destination of your journey, namely the **business outcome** that you seek to achieve and from which you work backwards.
2. The **AI/ML flywheel** as the motor of your journey. The AI/ML 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/ML 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 the fuel that keeps the AI/ML flywheel spinning.
4. Your **foundational capabilities** that, above all else, drive success or failure when adopting AI/ML.

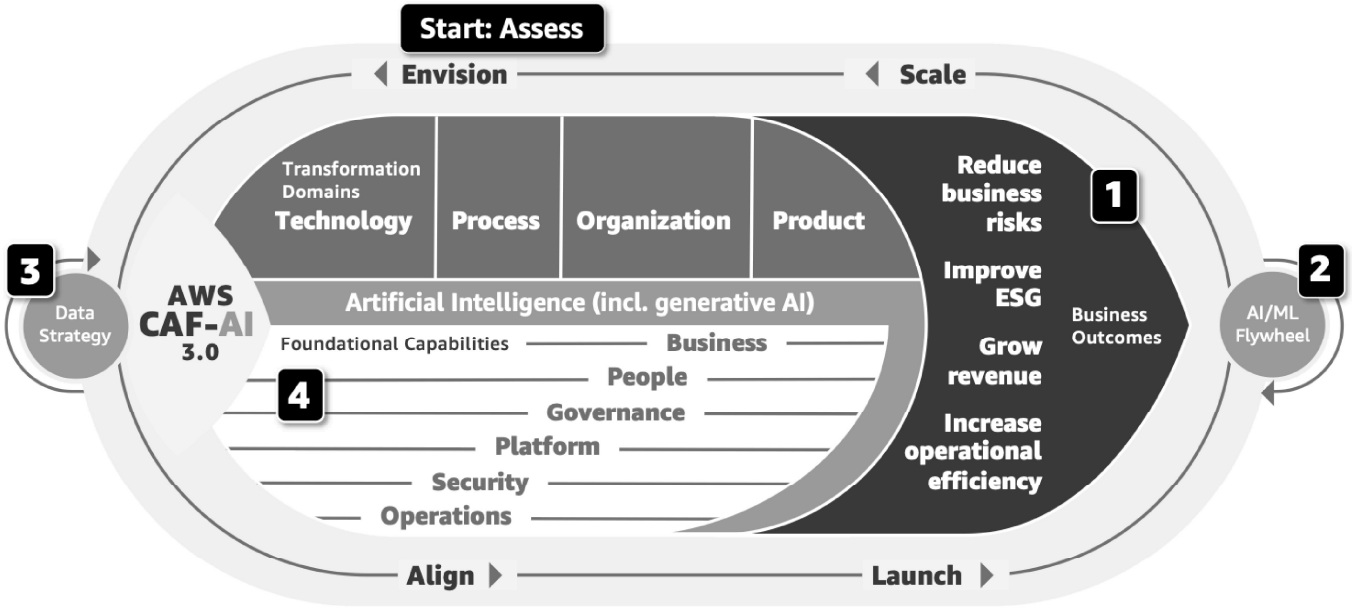

Figure 2: The AWS CAF-AI cloud transformation journey

When approaching this journey, base it on iterative and incremental improvement. The preceding image depicts the start of this journey with an **Assess** activity, where you examine your AI/ML readiness to understand where to start. This document will help you to begin assessing your state of maturity along the foundational capabilities. We also suggest you reach out to your AWS Support team to get assistance from AWS ML strategists, enterprise strategists, and ML advisors. After this initial assessment, the adoption cycle begins, based on four stages:

- **Envision:** This first phase focuses on envisioning how AI/ML 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/ML adoption is a cross-functional effort, much more so than this is the case for other technologies. Hence, aligning internally on the goals set in the envision phase is critical. Doing so helps you create strategies for improving your cloud and AI/ML 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/ML 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/ML readiness as you progress through your AI/ML transformation journey.

Foundational AI/ML capabilities

Iterating through your AI/ML transformation journey relies on your foundational capabilities to adopt AI/ML 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 for cloud and AI/ML adoption.

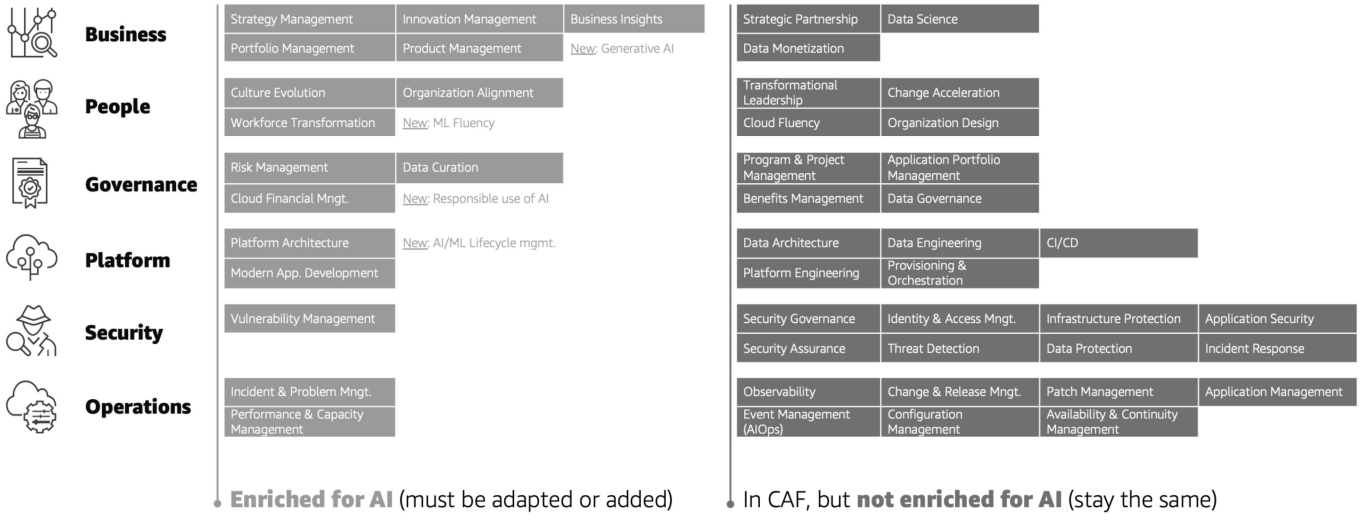

Figure 3: The AWS CAF-AI foundational capabilities

Take, for example, the capability product management (1) in the perspective business. While product management is already a needed capability to successfully develop cloud-based products, its implementation is significantly different when looking at AI/ML services in the cloud. The remainder of this document follows this logic: We call out the deviations and specific needs for AI/ML 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 interested 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/ML 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/ML 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/ML 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/ML initiatives while maximizing organizational benefits and minimizing transformation related risks. We pay special attention to the changing nature of the risk and hence 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/ML. 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/ML 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/ML 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/ML workloads, are delivered at a level that meets the needs of your business. We provide guidance on how to manage operational AI/ML workloads, how to keep them operational, and ensuring 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, which orders your areas of action for your AI/ML transformation journey in

time. The following image depicts this exemplary order and the assessment mentioned previously is best used to establish which of these capabilities already exist in the organization and to which degree from an AI/ML perspective.

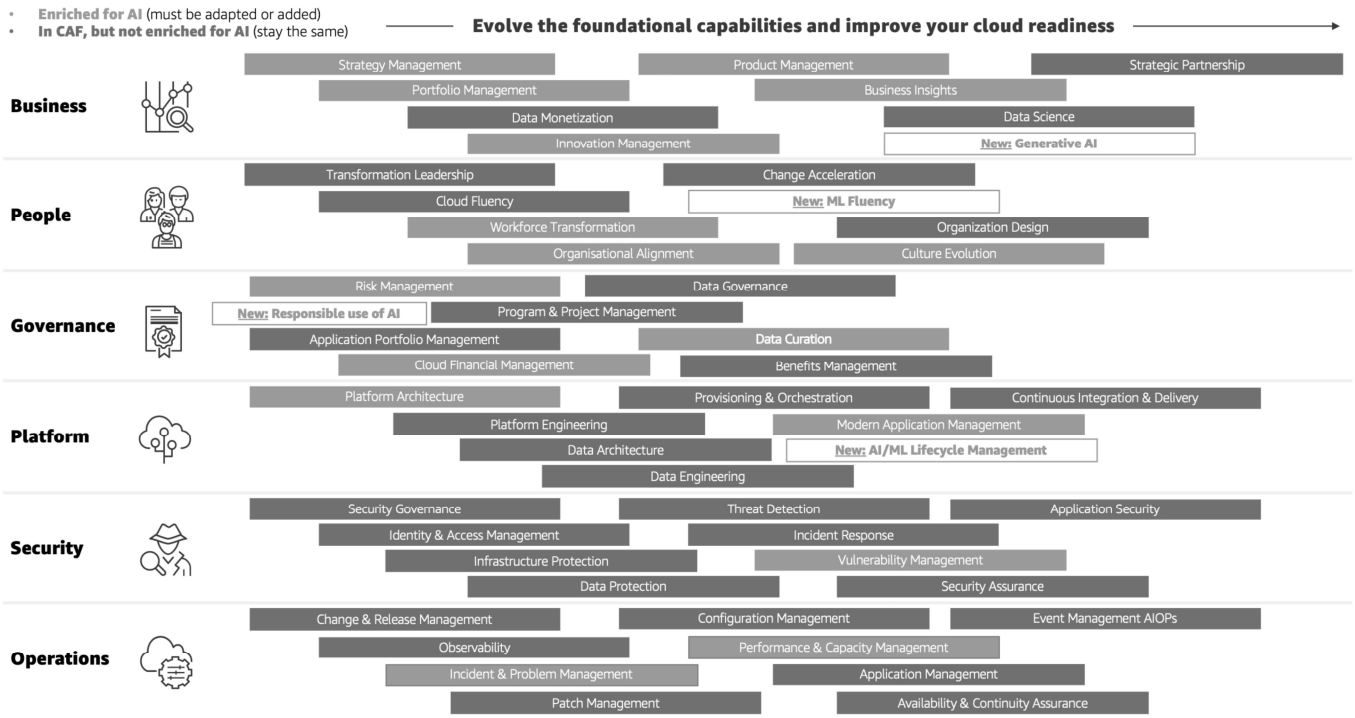

Figure 4: The AWS CAF-AI foundational capabilities ordered by maturity and evolution

Business perspective: The AI strategy in the age of AI/ML

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/ML infused or enabled products.
Business Insight	The power of AI/ML to answer ambiguous questions or predict from past data.
Portfolio Management	Identify and prioritize high-value AI/ML 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/ML 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/ML adoption and underpin it with an actionable strategy that moves step by step to adopting AI/ML 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/ML.

Work backwards from existing business and customer problems and the effect that AI/ML can have on them. When moving closer to prioritizing AI/ML 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/ML 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/ML capabilities are a part of that expectation.

For each opportunity, ask if you need to build, tune, or adopt an existing AI/ML 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/ML infused or enabled products.

Building and managing AI/ML-based products can be a significant challenge as the development and lifecycle of AI/ML 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/ML-based product include potentially costly uncertainties that require specific mitigation strategies.

When building or embedding AI/ML into products, work backwards from your customers' and users' expected value gain, and map measurable business proxies to individual decision points that an AI/ML 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/ML-based product through a proper lifecycle management, factor in how users interact with probability based output from AI/ML-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/ML 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/ML. 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/ML 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 Model Cards).

Business insights

The power of AI/ML 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/ML 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/ML-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/ML 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/ML suddenly becomes the driver for predictive decision making.

When preparing the transition of your BI practice to an AI/ML-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/ML journey.

In the early stages of transformation, any effective method can be to create a center of excellence for analytics (not necessarily AI/ML) 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/ML 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/ML 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/ML 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/ML in other portions of your business. At the same time, consider what larger customer and business problems you are solving through multiple AI/ML projects and products and combine those into a hierarchical portfolio where the lower layers of that portfolio enable the upper layers. Certain AI/ML 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.

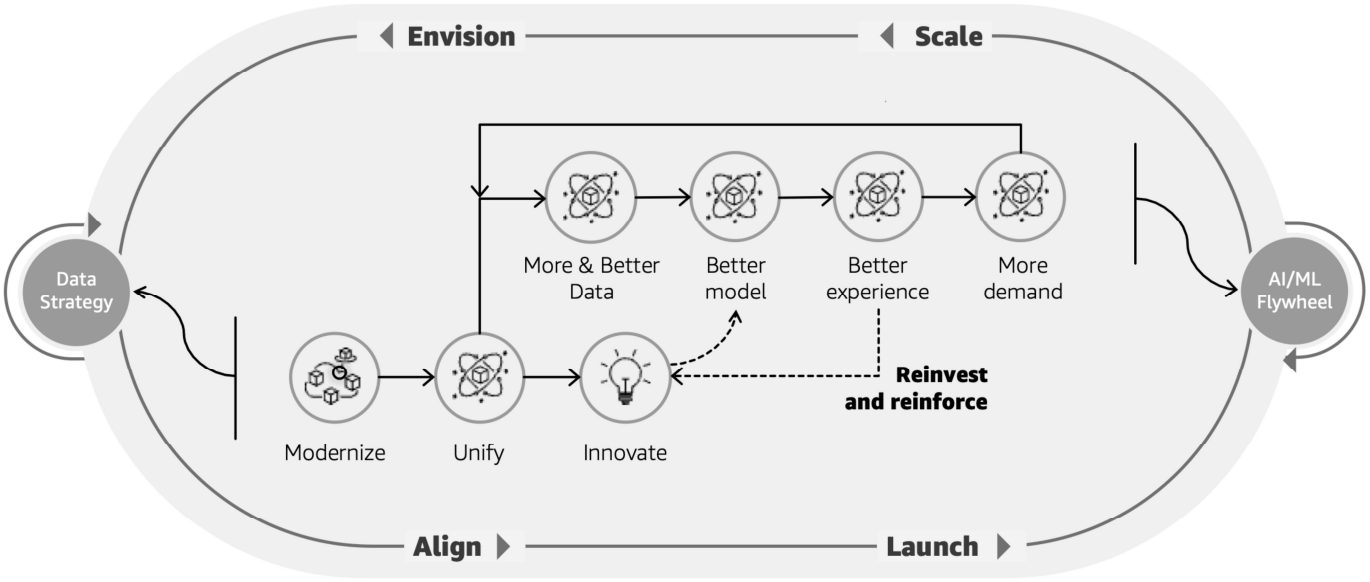

Figure 5: Create a self-reinforcing flywheel through your data strategy

Next, embed in this portfolio the design of an AI/ML 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 5). 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/ML workforce's efficiency by choosing the right AI/ML products and cloud environment. Realize how complex it is to even just technically govern your portfolio. To make sure you keep your scarce AI/ML 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/ML, 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/ML-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/ML 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/ML 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/ML 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/ML 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/ML ideas and projects.

New: Generative AI

Use the general-purpose capabilities of large AI/ML models.

The overall goal of AI/ML 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/ML 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/ML, 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/ML 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/ML-first

Adopting AI/ML and creating value reliably and repeatably is not a purely technological challenge. Any AI/ML initiative is crucially dependent on the people that guardrail and drive it. While AI/ML 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/ML systems come to live: Through collaboration between stakeholders, business units, and practices.

There often is talk about the potential of AI/ML 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/ML is largely about helping with tasks that are perceived as specifically complex for humans. We see that organizations that are AI/ML-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/ML. 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.*
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/ML champions.

Once that first layer of interpretation is spread across the organization, tackle the second, more technical one: AI/ML 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/ML outputs both in the organization and to customers. Consider that customers will have different mental models and terminology, so such as, letting an AI/ML 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/ML-talent from user to builder.

Being able to attract, retain, and retrain talent that can push your AI/ML strategy forward is one of the most crucial aspects of AI/ML success. There are many roles that are necessary for AI/ML 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/ML 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/ML 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/ML 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/ML roles is beneficial for organization-wide adoption.
- Hiring ML engineers and deep learning architects is most reasonable when you plan to base your AI/ML capabilities on established solutions, foundation models, or AI/ML 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/ML agenda fizzling out. When talent is not present, broadcast your AI/ML 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/ML is difficult, as supply has historically been outstripped by demand. Another factor is that real-world AI/ML differs significantly from the academic work that often drives talent into AI/ML. Counter this factor by having opportunities for your AI/ML 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/ML workforce through continuous re-training opportunities to learn new skills needed to perform well in the AI/ML 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/ML 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/ML 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/ML center of excellence (COE) is a unit that can fill this role, where AI/ML-focused teams are hired and evolved. Make sure that reporting lines in this organization align with those stakeholders that have ownership over the AI/ML 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/ML 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/ML 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/ML 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/ML.**

Developing an AI/ML-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/ML 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/ML 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/ML 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/ML-first culture:

- Experimental mindset paired with agile engineering practices

- Cross team and business unit collaboration and reliance
- Bottom-up and top-down AI/ML opportunity discovery
- Broad and inclusive AI/ML adoption solution design driven by customer value

Start expanding your AI/ML-first culture with the following:

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

  While you allow experiments, adjust your agile mindset toward the uncertainties of AI/ML. Recognize that you can't reliably define a time-effort estimate for complex projects, since many complex AI/ML 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/ML are not just heard but influence the decision-making process.

Governance perspective: Managing an AI/ML-driven organization

Just as people are one cornerstone of AI/ML adoption, another cornerstone is managing, optimizing, and scaling the organizational AI/ML initiative. Just as much as the cloud offered new ways of looking at operational expense (OPEX) versus capital expense (CAPEX), AI/ML requires adopters to look at cost structures in a different way while managing the organizational risks and opportunities that arise from that.

You need to consider that it's fiendishly easy to build a first proof of concept (POC) in AI/ML, but that scaling and growing both AI/ML solutions as well as the initiatives that enable them are a long-term investment. The cloud offers some solutions to these challenges, while others can only

be addressed by AI/ML leaders in your organization. In this perspective, we describe some solutions to these challenges and introduce a new capability: The Responsible use of AI, a decisive element for future competitive advantage in the AI/ML space.

Foundational Capability	Explanation
Cloud Financial Management (CFM)	Plan, measure and optimize the cost of AI/ML 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.
New: Responsible use of AI	Foster continual AI innovation through responsible use.
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/ML in the cloud.

The cost and investment profile of AI/ML can be somewhat surprising to adopters as there often is a zig-zag pattern (high/low/high/low) when developing such a system. This is true both on the individual use cases level, as well as on the bigger picture and all-entailing AI/ML initiative level. While cost is unsurprisingly associated with the concrete use case and your industry, it's also dependent on the state of the AI/ML system: Learning or inferencing. Start by analyzing the cost profile of AI/ML use cases over time and factor in this zig-zag cost-pattern that is inherent to many

cases: Typically, you will see a high initial cost to establish or increase the quality of the data that is needed to solve your problem, if not already readily available (then this initial cost is very low). This is followed by a surprisingly volatile proof-of-concept phase.

While most ML POCs will be relatively low cost (compute-wise), there are a few technical approaches that can become costly quickly. In such cases, refer to the AWS purpose-built AI/ML hardware (Trainium / `Trn1` and Inferentia / `Inf2`) to help keep costs down. If you have access to the right talent, AI/ML services, and AWS partners, let them estimate the resources needed for different phases of your use cases and overall AI/ML strategy. If feasible, calculate what an incremental improvement of an ML metric is worth to decide if it's worth the investment.

After the first system is built, the cost of the following minimum viable product (MVP) phase can, dependent on the use case then again have a relatively high cost, for example to generalize the systems capability or acquire edge case and long-tail data that is crucial for user adoption. This is where using or fine-tuning foundation models can make a large difference, as the initial training cost has been taken by you supplier or vendor (for example, AWS for the Titan foundation model).

After the model is deployed, inference itself is largely dependent on the volume of requests, and in many cases the inference cost itself is again relatively low. If not, refer to the purpose-built AWS Inferentia architecture. Monitoring model metrics and flagging drift will alert you to changes and potential need to retrain your algorithms. The then technical scaling on the cloud is cheap and resilience and reliability is backed in. Throughout this whole lifecycle keep cost visible and tag all resources and ML workloads.

Once you have cost-visibility measures in place, reason about data volume and training cost over time: There are a large quantity of problem types (text, forecasting, document processing) which in their infancy do not cost much and grow linearly with size. Other AI/ML problems that rely on audio and voice data have a much higher start-up cost and need well-defined goals even in the POC phase to not cause drastic charges. Aligning with your ML vision and where you need to be working should dictate how aggressively you scope the work.

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 ML (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, connect your ML 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 ML metrics. On top of this, factor in the often-underestimated cost of not recognizing the need for the

responsible use of AI/ML. Due to its importance, we have added The Responsible use of AI as 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/ML roadmap and you should ask yourself and other leaders: "Are we enabling AI/ML 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 re-enforcing 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 not 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 to avoid the tragedy of the commons, which in turn will help you build a robust data ecosystem. Start small and then continually add to your data mesh, as this will keep 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 and dictionaries will empower all skill levels to start using your data to create value. This will increase the speed to decide upon the additional investment cost needed for other cases considerably. There are many ways to go about increasing your data's potential, such as Buying external data sources, augmenting data via ML algorithms, crowdsourcing a team to label your internal data, or even changing your business practice to automate data generation and capture. 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 attached (and AI/ML is no different), don't let that fool you: Managing the risks involved both in the design and development process of AI/ML systems as well as in the deployment and long-term operations and application of AI/ML is challenging and not overly well understood in the industry yet. Start by factoring in the risk of sunken cost into the development process as the outcome of an AI/ML development initiative is hard to guarantee upfront (the nature of optimizing a system for output versus specifically building it to do so).

Establish solid practices, such as model cards and adversarial inputs) and mechanisms (such as proofs of concept (POCs), minimum loveable products, and minimum viable products (MVPs), to mitigate and control risks. This includes risks as classified by your local legislature, for example, the European Union, and those that are inherent to AI/ML, such as a hidden feedback loop or misinterpretation of uncalibrated outputs. 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.

Develop and adopt safeguards and architectures that constrain the system when necessary, not just in safety-critical environments. Make sure that subsystem failures don't propagate and compound downstream AI/ML systems. Consider which themes are relevant, such as explainability, transparency, and interpretability. Manage these risks, not just for a single AI/ML 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). Lastly, don't sugar coat the complexity of reaching human-level parity in certain domains.

New: 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 strive to reap the business benefits. However, as AI/ML systems learn from vast amounts of data, the knowledge this data embeds and hence what the system learns is not always what you might have intended. Responsible AI practices are therefore needed and key for fostering continuous AI innovation. 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/ML journey and throughout its lifecycle: Scale how it impacts your design, development and operations over time. Consider how your system will affect individuals, your users, customers, as well as society.

Given the speed at which AI/ML 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/ML vision, including well thought out principles and tenets around the responsible use of AI/ML and how it will affect your initiatives. In particular, include algorithmic fairness, diverse and inclusive representation, and bias detection.

Embed explainability by design in your ML 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/ML and build or use systems to enable your teams to inspect these factors. While this cost accumulates before the algorithms reach production state, it will pay 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 will trickle down directly into your business case.

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

Platform perspective: Infrastructure for and applications of AI/ML

With all the advances in algorithms and the problems they can solve, at times the systems and processes that produce and run those algorithms are left to decay. Like any good manufacturing process, you need systems and processes that deliver a uniform, consistent product, in our case, an algorithmic result leading to business value. Instilling processes for each facet of your platform, revisiting your infrastructure, and knowing how each portion contributes to value will keep you producing results at the front of the pack. We have seen that those who do not spend time on these steps will inevitably have to rip out systems and spend valuable time rebuilding to support their end business value.

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.
New: AI/ML Lifecycle Management and MLOps	Manage the lifecycle of machine learning workloads.
Data Architecture	*This capability is not enriched for AI, refer to the AWS CAF.*
Platform Engineering	*This capability is not enriched for AI, refer to the AWS CAF.*
Data Engineering	*This capability is not enriched for AI, refer to theAWS CAF.*
Provisioning and Orchestration	*This capability is not enriched for AI, refer to the AWS CAF.*
Continuous Integration and Continuous Delivery	*This capability is not enriched for AI, refer to the AWS CAF.*

Platform architecture

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

As machine learning is maturing from a research-driven technology to an engineering practice, the capability to reliably and repeatably create value from its application becomes more important. There are many mental models in the industry of how this capability should be implemented and supported. Since the ML lifecycle is long and complex, there are a variety of architectures that have been proposed, and they differ based on the intended outcome. To inform your future decisions, start by diving deep into broad and proven architectures that are independent of your specific use case and industry. At the same time, while building your first prototypes and applications, evolve your platform and adjust it to your industry- and workflow-specific needs.

Every platform strategy should be guided by a set of design principles that keep the overall system aligned in its purpose and intent. Over time, cover all aspects of the ML lifecycle and make sure that each component adheres to best practices to reduce the risk of failed investments. Make sure that these are, at the very least, a way to manage and access distributed and governed data in the organization, prepare and offer this data in a way to support the needs of individual consumers, support the development of novel AI/ML systems through and end-to-end development experience as well as a way to use existing AI/ML capabilities and foundation models, deploy and monitor the resulting models as well as sharing them with downstream consumers and iterate across the organization on them.

In many organizations, it's also valuable to enable those builders that don't have a deep background in data science through tools that still can cater to their individual and custom need, such as AutoML or No/Low-Code. While you do so, make sure that the components that cover each step can still be changed and evolved so that your ability to innovate is not hindered. This often results in component-based horizontal platforms that transfer ownership over different parts of the AI/ML value-chain to empower teams in the organization. Lastly, for a mature and scaled platform, find ways to offer human-in/on-the-loop functionality to improve on results and labeling capabilities to improve the inputs. The cloud offers these and more capabilities with the additional benefits of reduced adoption risks and cost-effectiveness.

Modern application development

Machine learning itself is a specific technical capability that enriches an application which aims to address a user, customer, or business need. The overall application consists of many more moving parts than only the ML model or ML capability. Still, in many cases, the capabilities that ML brings to the table are at the heart of the value proposition. This means that managing this relatively small (that is, when you consider the size of the code footprint) component of a much larger system is a make-or-break for its use, adoption, and ultimately value. Take a step back and consider the ML component in its own right and establish how you, your teams, and the organization iterate on the ML-driven application. Base your architectural decisions on best practices that integrate your cloud environment, your application development, and your user experience design. Reduce the gap between cloud, application development, and AI/ML to increase the speed at which your business operates and your teams iterate.

Consider from the beginning that the data you acquire and store in the cloud has gravity for other downstream uses and consumers in your organization. Aim to enable these consumers to deliver value by creating a seamless builder experience. When developing applications, consider how data moves through your systems, how it changes your AI/ML 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. This is an AI/ML feedback loop that is akin to patching your system through new data, for example, Example in healthcare.

Experiment not only with how AI/ML can deliver results without intervention, but also how humans in- / on- / off-the-loop can provide or increase its value. Finally, establish in your development teams a clear understanding that AI/ML systems are indeed perceived differently by customers and users, and that many users are lacking the mental models and metaphors that help them interact with these systems. This means that all AI/ML-based applications that customers and users interact with directly benefit from a fresh look at their user experience (UX) and customer experience (CX). Lastly, use generative AI to propel application development itself through tools that increase your velocity to write code.

> **Note**
>
> The AWS Well-Architected-Framework provides a Machine Learning Lens whitepaper to help you design, architect, and deploy your AI/ML workloads.

New: AI/ML lifecycle management and MLOps

Manage the lifecycle of machine learning workloads

The machine learning lifecycle is challenging to manage. It consists of three major components:

- Identifying, managing, and delivering the business results and customer value.
- Building and evolving the technological components of the AI/ML solution.
- Operating the AI/ML system over time (MLOps).

Each of which contain more fine-grained phases that we have mentioned in this framework regularly: Business goal identification, ML problem framing, data processing, model development, model deployment, model monitoring, and more. Consider AI/ML lifecycle management as its own practice and establish ML product managers and ML senior engineers that understand this bigger picture. While you have to start somewhere, make it your goal to advance your maturity in managing this lifecycle and consider the three components (business value, development, and operations) individually, each with their own maturity model. Different AI/ML strategies will require different perspectives on these three components. For example, if your overall goal is to enable

new products through custom models, you will see lifecycle management in a different light than if you strive to increase internal operational efficiency through publicly available services.

Align these three components with an independently stated lifecycle goal:

- Moving AI/ML to production and fine tune that statement by referring frequently to your vision and your principles included as well as your responsible AI perspective.
- Build bridges between these three components by checking if your AI/ML system is feasible to maintain in production whenever you move between them. Non-dogmatic estimates on how much it will cost to do model maintenance should be done early on as part of specifying the business case but inspected over time.
- Strategize over the effects of letting a model decay in operations as there are likely financial and business impacts or, more importantly, potential societal impacts. This bridge stands on the shoulders of a team culture where collaboration and empathic product development is normal.
- Consider single threaded teams that include business and subject matter experts throughout the entire ML lifecycle. While all phases are decisive, what customers struggle with the most is managing the complexity of the operations, or MLOps.

  To better understand where you are in relation to industry best-practices, assess your MLOps maturity with a AWS Partner or AWS, and base your decisions on a rigid MLOps and lifecycle framework. MLOps is much more than technology and needs a firm process to create value. It's common to see data science teams wrongfully focus only on hard ML metrics instead of how these metrics affect a business metric, which is a failure of lifecycle management. Whatever your path, make the process and standards you establish for MLOps repeatable. These processes and standards are the best defense against your system relying on tribal knowledge alone and reduces your AI/ML technical debt. Such MLOps best practices will also help you ensure that your science team does not get modeling fatigue and focuses on results instead of getting distracted with mass parallelization of experiments.

  > **Note**
  >
  > The AWS ML operations specialists have put together a maturity framework combined with architectures, roles, and guidance.

Security perspective: Compliance and assurance of AI/ML systems

Over the past decades AWS has invested big in the security of its cloud and services, and we put it above everything else we do. Earning customer trust is key for any technology, and this is even more true for AI/ML. This is why almost all of the security capabilities that are outlined in the AWS CAF do not need to change for AI/ML.

However, one single component that sits outside of the domain that the cloud can completely own for you has been enriched: Vulnerability management on the application level. With AI/ML having a strong reliance on unstructured data and learning relationships between input and outputs through a complex system, such as neural networks whose inner functions remain somewhat foggy (a property of these types of models) the attack vector for AI/ML models' customers build have increased. In this section, we share some perspectives on this challenge.

Foundational Capability	Explanation
Vulnerability Management	Principles, patterns, and best practices for repeatable AI value.
Security Governance	*This capability is not enriched for AI, refer to the AWS CAF.*
Security Assurance	*This capability is not enriched for AI, refer to the AWS CAF.*
Identity and Access Management	*This capability is not enriched for AI, refer to the AWS CAF.*
Threat Detection	*This capability is not enriched for AI, refer to the AWS CAF.*
Infrastructure Protection	*This capability is not enriched for AI, refer to the AWS CAF.*
Data Protection	*This capability is not enriched for AI, refer to the AWS CAF.*

Foundational Capability	Explanation
Application Security	*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/ML attacks.

AI/ML systems have additional attack vectors that bad actors can try to exploit. Implementing the existing best practices for keeping environments and data safe, such as the principle of least privilege, are the first step and a prerequisite. Start by balancing security and the speed of innovation. While security can be seen as an additional burden for builders, it will pay off in the mid-term and earn trust with your users and customers. At the same time, be sure to not place iron walls around tooling or data with no gate for allowing access as it can drastically hamper the effectiveness of AI/ML.

Once the basic mechanisms are in place, start to inspect the three critical components of any ML system, its inputs, model, and outputs:

- The **input** attack vector relates to all of the data that has an entry point to your algorithm. This input is often the goal of targeted attacks, such as targeted model and distribution drift where bad actors try to influence from what data the system learns over time or purposefully introducing a hidden bias or sensitivity to certain data. Harden these inputs through data quality automation and constant monitoring.
- The **model** attack vector relates to exploiting misrepresentations of the real world or the seen data in the model, most famously adversarial attacks. In these cases, the model itself often has shown no indication of unusual behavior. Harden your model against such attacks through a dedicated expert.
- The **output** attack vector relates to interacting with the system over a long period of time, which can allow bad actors to infer critical information about the inputs and properties of your model, often called *data leakage*. The damage this can do to your customers trust can hardly be overstated.

These are just a few of the attack vectors that you need to factor in. While not every AI/ML system exposes these attack vectors, and many are not directly accessible by bad actors, once you move models in front of customers, it's worth investing in a dedicated practice.

Operations perspective: Health and availability of the AI/ML landscape

Operating ML applications is new for many customers. In the new CAF-AI capability AI/ML 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/ML behavior.
Performance and Capacity	Monitor and handle AI/ML workload performance.
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.*

Foundational Capability	Explanation
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/ML behavior.

AI/ML 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/ML 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/ML systems through proxies and simplified statistics. When adopting AI/ML observing and monitoring, these simplified views into the AI/ML 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/ML 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/ML 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/ML system. For example, foundation models are significantly harder to correct and monitor than simple decision trees.

Performance and capacity

Monitor and handle AI/ML workload performance.

AI/ML 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 New: AI/ML lifecycle management and MLOps 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 provided an overview of the AWS Cloud Adoption Framework for Artificial Intelligence, Machine Learning, and Generative AI. CAF-AI provides a map of how a customer can organize and structure their AI journey, what capabilities are needed to successfully implement it, and a mental model for iteration over them. The foundational capabilities in this document are meant as an index for further investigation, learning, and conversations with your AI/ML experts. All of these capabilities tie into the AWS CAF and help organizations to think both about their cloud journey as well as their AI/ML journey.

Contributors

Contributors to this document include:

- Alexander Wöhlke, Sr. ML Strategist, Generative AI Incubator, AWS CAF-AI Lead
- Caleb Wilkinson, Sr. ML Strategist, Generative AI Incubator, AWS CAF-AI Lead
- Dr. Saša Baškarada, Worldwide Leader, AWS Cloud Adoption Framework
- Phil Le-Brun, Director, Enterprise Strategy
- Neil Mackin, Principal ML Strategist, Machine Learning Solutions Lab

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
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.

© 2023 Amazon Web Services, Inc. or its affiliates. All rights reserved.

AWS Glossary

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