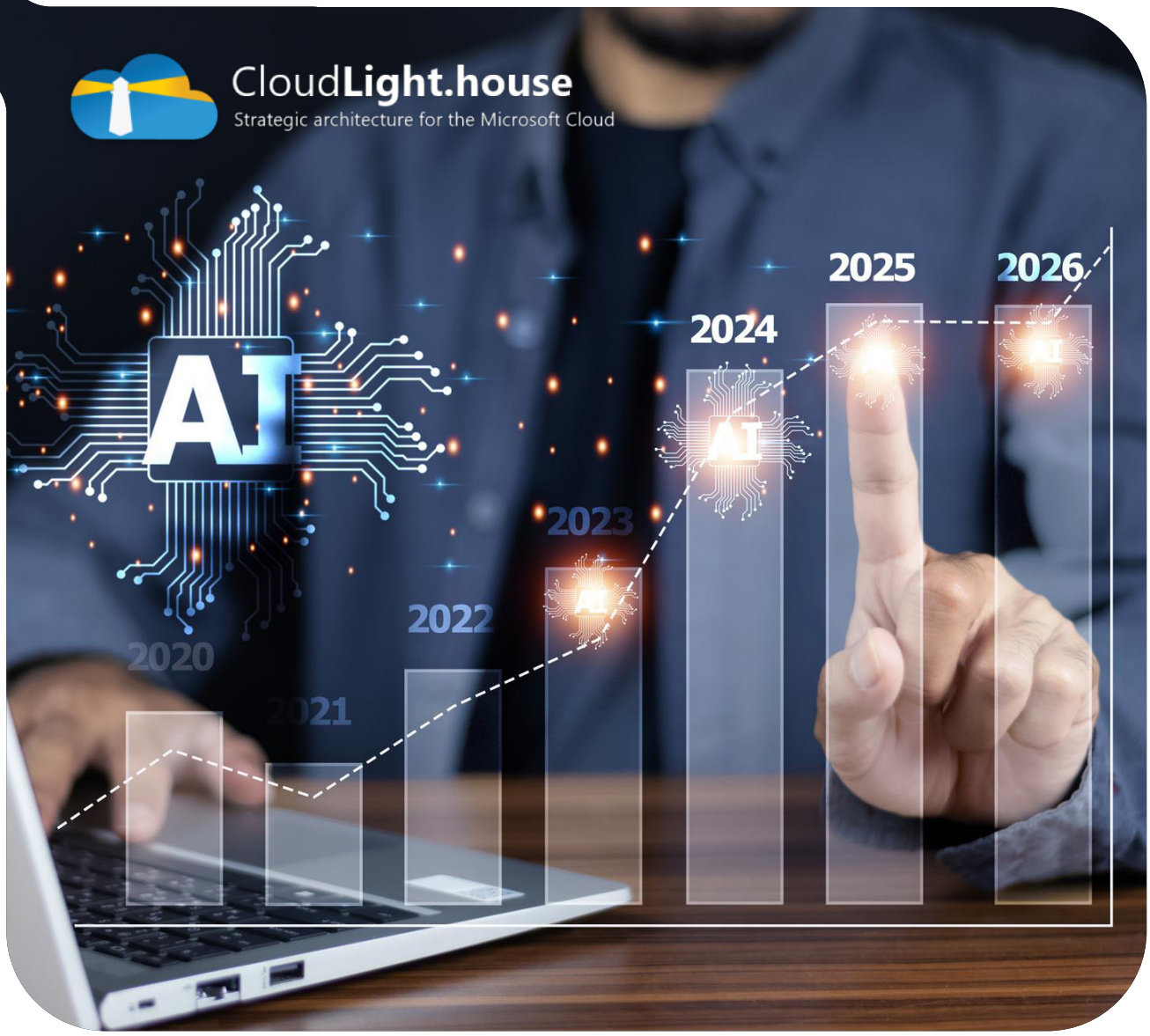
CloudLight.house
Strategic architecture for the Microsoft Cloud



Crafting your Future-Ready Enterprise AI Strategy

Second Edition

Microsoft

CloudLight.house
Strategic architecture for the Microsoft Cloud

Wave periods between major innovation in the cloud are growing shorter. We no longer have the luxury of waiting it out, of adopting later. Cloud ecosystems built on strategic foundations create the conditions to absorb successive waves of change.

CloudLight.house/strategy

Follow on LinkedIn



Contents

Introduction 5
Chapter 1: Background and Foundational Considerations 6
Background 7
Foundational Considerations 9
You are not ready; almost nobody is 10
Future-ready, not future-proof 10
Guiding Principles 10
How generative AI acts on enterprise data 11
Chapter 2: AI Strategy Framework 13
Pillar One: Strategy and Vision 16
Executive Vision 17
Actionable Roadmap 19
Ecosystem Map 21
Programmatic Rigor 23
Center for Enablement 24
Pillar Two: Ecosystem Architecture 27
Core Platform Services 31
Data Distribution 32
Integration 38
AI Development Tools 43
Pillar Three: Workloads 47
Workload Prioritization 48
Incremental AI 52
Extensible AI 54
Differential AI 56
Power Users 60



© Cloud Lighthouse LTD, 2024

Pillar Four: Responsible AI 62
Reliability and Safety 63
Privacy and Security 65
Fairness and Inclusivity 66
Transparency 67
Accountability 67
Pillar Five: Scaling AI 68
AI Operations (AIOps) 69
Data Governance 70
Technical Debt 71
Monitoring and Metrics 74
Digital Literacy 76
Chapter 3: AI Maturity Model 80
Chapter 4: Onwards 85
Organizational Obstacles 86
About Cloud Lighthouse 90
About the Authors 91
References 93

© Cloud Lighthouse LTD, 2024

Introduction

I have spent the last two years crisscrossing the world helping C-level leaders craft their AI strategies, building the actionable roadmaps and technical architectures necessary to bring those strategies to fruition. The first edition of Crafting your Future-Ready Enterprise AI Strategy, released in January 2024, made the pillars of this strategic thinking publicly available, compiling written guidance to get organizations started on their journeys.

Eight months later, I find myself sitting on an airplane in the skies somewhere between

Bangkok and Paris writing the introduction to the paper's second edition. Why so soon?

It has become obvious to my co-authors and me how difficult so many organizations are finding it to actually craft and execute their AI strategy, in part because of the (often) decades they've spent kicking their proverbial data can down the road, in part because it turns out that enterprise-grade AI really does require the adoption of ecosystem-oriented architecture to truly scale, but largely because many organizations have no idea where to start. Many lack the wherewithal to really assess where they stand on day one, and to identify areas where they must mature to get to day 100 (and beyond).

We've learned a great deal about maturity and readiness for - and responsibility to the ethics of - AI over the past year, as well. It's now time to broaden the thesis, so in this second edition we offer a model through which organizations may realistically assess their current maturity to adopt and scale artificial intelligence, and then identify specific areas to invest time, talent, and funding along their journey.

The first edition suggested five pillars for a proper AI strategy: Data Consolidation, Data Readiness, Incremental AI, Differential AI, and Scaling AI. These tenets are important as ever, but in expanding upon them we have re-ordered them into a broader AI Strategy Framework detailed in the pages that follow.

Think of this second edition as an expansion, not a revision, of the first. We've preserved much of the content and insights from its predecessor just as we've focused on the Strategy Framework and the accompanying AI Maturity Model as the blueprint for organizations building holistic, highly scalable, and, of course, future-ready enterprise AI strategies.

Cheers,

Andrew Welch
Author | CTO
Cloud Lighthouse
Follow on LinkedIn

© Cloud Lighthouse LTD, 2024

CloudLight.house
Strategic architecture for the Microsoft Cloud

Chapter 1

Background & Foundational Considerations



Background

We're often asked, "What should our AI strategy look like?".

In co-author Andrew Welch's early-2023 essay, Strategic thinking for the Microsoft Cloud, he mused that, "I really wonder how many will miss out on the AI wave because they lack the wisdom or the willpower to make the most of it."

Then, in June 2023, under the headline Your employer is (probably) unprepared for artificial intelligence, The Economist suggested that divergence in technology investment - and the success of that investment across firms - has resulted in a "two-tier economy" wherein "firms that embrace tech are pulling away from the competition." The piece cited several examples:

- A nearly 11% rise in average worker productivity in Britain's most productive firms from 2010 to 2019, a period where the least productive firms saw no rise;
- A tripling of productivity growth in Canada's most productive vs. least productive firms from 2000 to 2015;
- Tim Koller's (McKinsey) findings that when ranking firms based on return on invested capital, "the 75th percentile had a return 20 percentage points higher than the median in 2017," which was itself double the gap in 2000.

Discussing AI in recent years, we have often thought about the fable of the boiled frog, whereby a frog placed in boiling water jumps out, but a frog placed in warm water that is gradually heated lacks awareness of his impending demise until it is too late. Indeed, The Economist piece goes on to explain:



"Robert Gordon of Northwestern University has argued that the 'great inventions' of the 19th and 20th centuries had a far bigger impact on productivity than more recent ones. The problem is that as technological progress becomes more incremental, diffusion also slows, since companies have less incentive and face less competitive pressure to upgrade."

—The Economist, 16th June 2023

© Cloud Lighthouse LTD, 2024

It is yet unknown if artificial intelligence is more akin to the “great inventions” of the 19th and 20th centuries, or if it will ultimately represent another more incremental evolution of existing capabilities. The former - as seems more likely given the immense investments being made today - will present significant challenges to nearly every organization that, having become accustomed to incremental change, is suddenly faced with a “great inventions” caliber paradigm shift that AI seems to portend.

Or, as we continue to remind the CIOs with whom we work closely, **the grace period for organizations to get their act together and position themselves for the next wave is growing much shorter, and the margin for error much narrower.**

© Cloud Lighthouse LTD, 2024

Foundational Considerations

So it is that we've been spending significant time and mental energy thinking about a proper "AI strategy" for organizations that wish to escape the sad fate of the frog, or that of the world's organizations being left behind by the pace of technological change.

Let's start by defining what we're talking about when we discuss "artificial intelligence." There's an argument that we could trace the lineage of AI back to the Turing Machine of the 1940s, and then pull that string through the history of computing to include such logic-based "if this, then that" applications all the way through robotic process automation (RPA) technologies of recent years. These technologies lacked real intelligence, given that they were rather extraordinarily intelligent machines created by extraordinarily intelligent people. These machines were designed to make more efficient (or make possible) the intelligence of their creators at scale.

Generative AI is what we have in mind when we think about artificial intelligence today. Here we define 'AI' as the ability of the machine to think independently of its creators or of the parameters its creators have set forth. Generative AI describes the ability of AI to generate unique and original responses - be they textual, imagery, or in some other medium - based on its index of accumulated knowledge. In other words, for the machine to assimilate data in unpredictable ways that conjure new responses, rather than to simply navigate a logical process for which it has been pre-programmed.

It's worth noting that we've also seen what we call "multi-modal" AI capabilities markedly mature in 2024. Multi-modal extends previously existing abilities to learn from and generate new text, imagery, video, and other mediums as separate scenarios such that single models can understand and generate across multiple modes. This is a significant development in the ability of an AI workload to wrap its inorganic head around a vastly expanded variety of mediums, which in turn expands both its comprehension and its generative abilities.

All of this is new. It is groundbreaking. Historically speaking, we are in a very different era in late 2024 than we were two years ago, a fact that causes me to encounter two foundational realities time and again.

© Cloud Lighthouse LTD, 2024

You are (probably) not ready; almost nobody is

First, very few - if any - organizations are truly prepared to make the most of the AI wave crashing on their shore. Very few have done the hard work to build the kind of proper, modern data platform required to make AI work at scale across their organization. For this reason, nearly everyone's "AI strategy" will at this point look very much like most everyone else's AI strategy, as organizations across the economy and around the world scurry to get their house in order.

Future-ready, not future-proof

Second, nobody truly knows exactly what a mature AI capability will look like or exactly how this will play out in practice. Today we're at a point with AI reminiscent of where we were with the "world wide web" in the late 1990s when organizations rushed to digitize their physical identity. Back then we - predictably - experienced a lot of websites that looked like someone had relocated their back-of-the-phonebook advertising to a screen. We went through a similar chain of events with the advent of the smartphone when developers first tried to cram desktop apps into a smaller form factor, and then again with smart watches when developers tried to cram smartphone apps onto our wrists. Comparably, many today are busy making yesterday's business processes incrementally more efficient by grafting AI onto them. Each of these cases illustrates our tendency to transpose legacy paradigms to new technologies. That is, at least until we gain a sufficient understanding of the new capability to grasp how transformative it really is… and how best to exploit it.

Guiding principles

Any future-ready AI strategy must be flexible, meaning it is able to absorb tomorrow what we don't fully grasp today. Your strategy should also **offer immediate value to the organization beyond specific AI-driven workloads because the nature and value of these workloads will remain unclear for some time.** For example, "data readiness" is indispensable to your AI strategy because it is likely to yield better AI workloads in the future, and because it offers real value in terms of security, accuracy, discoverability, and analytical integrity separate and apart from AI today.

This explains my fondness for the phrase future-ready, and why I cringe when I hear people say "future-proof." The former describes a cloud ecosystem built with modern technologies using best practices that are most likely to absorb whatever future innovations come our way. The latter is unachievable in all circumstances.



© Cloud Lighthouse LTD, 2024

How generative AI acts on enterprise data

Let's establish a basic understanding of how AI uses and acts on enterprise data. We will define 'enterprise data' as data that is proprietary to a specific organization, kept and (I certainly hope!) secured inside the boundary of the organization's data estate.

You may be familiar with the term "RAG", an acronym for "retrieval augmented generation". While this is not the sole means through which AI acts on organizational data - and new and evolving patterns now emerge regularly - RAG represents a good baseline for the general concept through which nearly all AI workloads essentially augment an existing model with an organization's proprietary data.

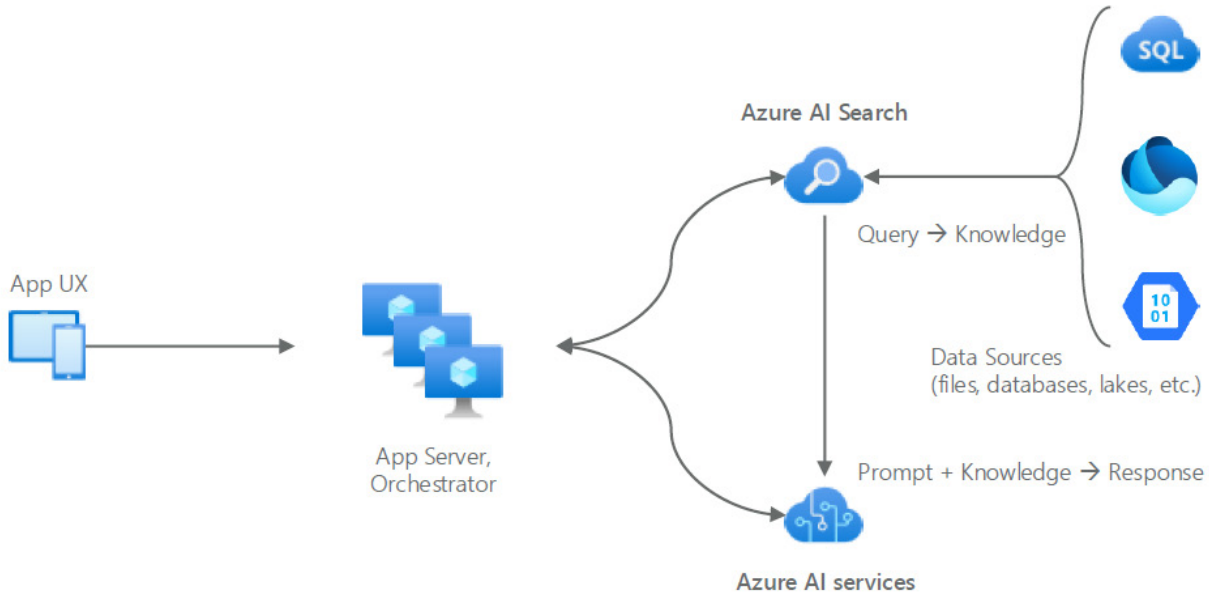


Figure 1: A basic RAG pattern through which enterprise data (raw knowledge) is stored such that it can be (a) indexed and (b) accessed by AI. Capabilities such as Azure AI services act on that knowledge to produce a response before passing that response back to the user (via the App UX.)

In the top-right of the diagram we're looking at various data sources sitting in a modern data platform (Azure SQL, OneLake, and Blob Storage are shown top to bottom for representative purposes). Blob Storage is a highly efficient way to store unstructured data, that is, files, images, videos, documents, etc. In this simple scenario we'll say that unstructured data is drawn from Blob.

These data sources are indexed by Azure AI Search (formerly called "Cognitive Search"), which also provides an enterprise-wide single search capability. Moving to the far left we see an application user experience (UX) e.g., a mobile, tablet, or web app that provides an end user the ability to interact with our workload.

The application sitting beneath the UX queries the knowledge contained in Cognitive Search's index (as derived from the data sources on the right). It then passes that prompt and knowledge to Azure AI services to generate an appropriate response to be fed back to the user.

CIOs and enterprise architects need not be experts in the technical mechanics of AI to formulate and execute an effective AI strategy. That said, it is critical that leaders driving this strategy must understand this basic concept of how institutional AI - that is to say, AI workloads specific to your organization - both requires and acts on enterprise data.

Without that data, it's just AI, unspecific to the organization it is serving.

© Cloud Lighthouse LTD, 2024

CloudLight.house
Strategic architecture for the Microsoft Cloud

Chapter 2
AI Strategy Framework



AI Strategy Framework

An organization's AI strategy ought to be constructed atop five pillars, each with five dimensions to be considered, matured, and regularly evaluated. This model has the benefit of shaping (a) how you evaluate your organization's maturity, risks, and opportunities in AI at any point in time - including when just getting started - and (b) how you organize your strategy to mitigate those risks, seize those opportunities, and mature the organization's use of AI over time.

Incidentally, because AI depends on a sound technical foundation in terms of data estate, application portfolio, governance, security, etc., those who embrace this model will find that they significantly mature the strategic architecture of their IT ecosystem overall. Reference our guiding principle that your investments in AI ought to, "offer immediate value to the organization beyond specific AI-driven workloads". In other words, invest in AI such that the investment pays off in other ways, as well.



Figure 2: Use these five pillars to evaluate an organization's maturity, risks, and opportunities in AI at any point in time. Then build your strategy to mitigate those risks, seize those opportunities, and mature the organization's use of AI over time.

These pillars address five broad questions that an organization ought to continually ask itself:

- Are our investments in AI strategically driven by a coherent vision for how we wish to use it rather than driven by the arrival of the latest trend or "shiny object"?
- Do we build AI capabilities atop a solid ecosystem-oriented architecture across our IT estate rather than grafting AI capabilities onto a fragmented IT estate that will be difficult to maintain in the future?
- Have we effectively balanced AI's risk and reward across incremental, extensible, and differential workloads?
- Do we embrace the principles of "responsible AI" (RAI), and – importantly - are we doing the never-ending hard work of making those principles actionable in our organization?
- Are we positioned to scale AI across the organization, including our ability to manage and govern AI and the data upon which it relies?

© Cloud Lighthouse LTD, 2024

These broad questions offer helpful guidance, but on their own lack the specificity that a truly actionable strategy requires. Each pillar, therefor, is supported by five component dimensions.

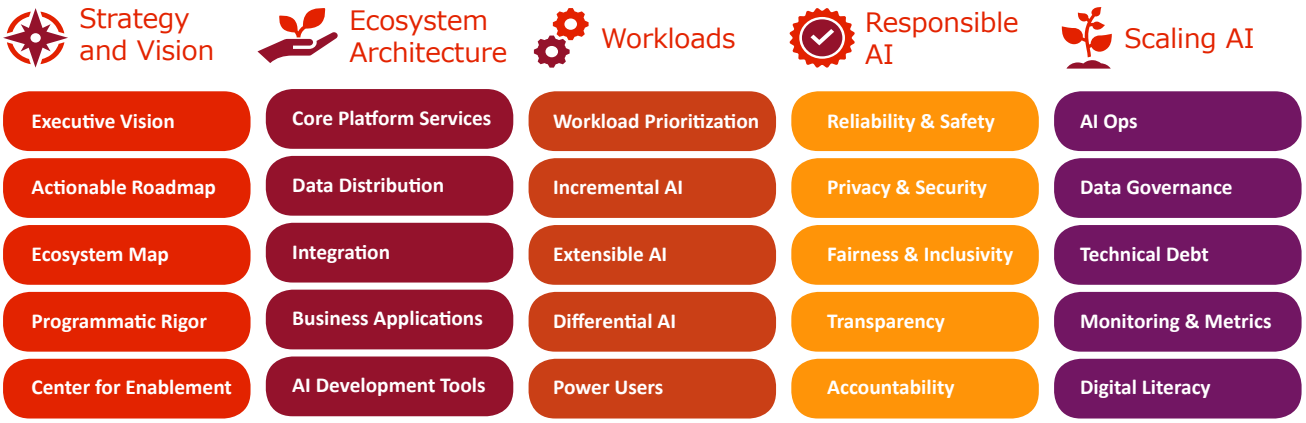


Figure 3: Each pillar is supported by five component dimensions that offer greater specificity through which AI maturity can be measured. Low maturity dimensions generally indicate risks to be mitigated, whilst higher maturity dimensions generally indic ate strengths or opportunities to be leveraged.

Look no further to understand how significantly we've expanded our model of AI strategy from this white paper's first edition to this second edition. The first edition discussed the pillars Data Consolidation, Data Readiness, incremental AI, Differential AI, and Scaling AI. This second edition preserves and expands "Scaling AI" as its own pillar, preserves "Incremental AI" and "Differential AI" as dimensions in the new and much expanded "Workloads" pillar, and preserves "Data Distribution" whilst changing "Data Readiness" to "Core Platform Services" as part of the new and expanded "Ecosystem Architecture" pillar. Moving parts to be sure, but we thought it helpful to map the evolution of AI strategy from early to late 2024.

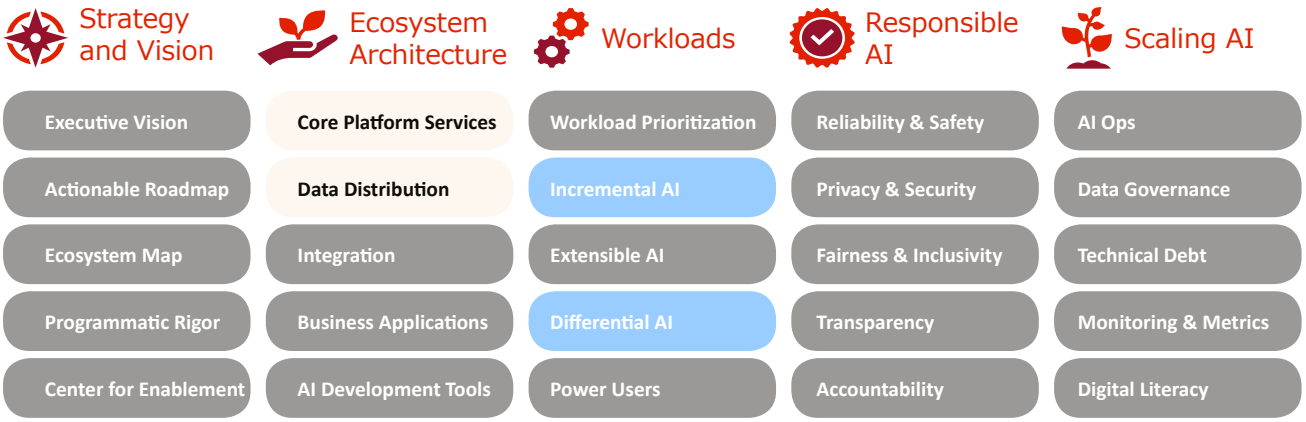


Figure 4: Core Platform Services (formerly "Data Readiness"), Data Distribution, Incremental AI, Differential AI, and Scaling AI have been revised and reworked from this white paper's first to second edition.

We'll explore each pillar and its dimensions in the subsections that follow.



CloudLight.house
Strategic architecture for the Microsoft Cloud

Pillar One
Strategy & Vision

Strategy and Vision



There's an incredibly important transition in the broad information technology space that is often lost in the furor and excitement over generative AI.

You see, since IT time immemorial most chief information officers and those in similar roles have been called on by their organizations to essentially function as superintendents of utility companies. Their charge has been to keep the phones ringing, the emails sending and receiving, and to prevent data from leaking.

AI is upending this paradigm, even though many still don't yet realize it.

As AI and its related technologies become more pivotal to the success of an organization - see our earlier statistics regarding productivity and Investment - technology leaders are finding that they must transition from being superintendents of utility companies to being strategic leaders of the organizations they serve.

But crafting, executing, and making smart investments in scalable cloud and AI strategy is hard. Leading strategically - and empowering your people to implement the vision - can seem overwhelming.

Simply "wanting AI" doesn't cut it. So, our Strategy and Vision pillar sets forth five dimensions which begins with vision, extends to creating the actionable roadmap and architecture necessary to actualize that vision, and finally establishes the programmatic elements necessary to drive that vision to fruition. These dimensions help organizations formulate and take action on their big ideas.

Strategy and Vision is thus the first indispensable pillar of any future-ready enterprise AI strategy.

Looking for the place to start with your AI ambitions? This is it.

Executive Vision

We've tried in vain over the years to accommodate shortcuts demanded by various organizations with whom we've worked. Alas, we've reached the same conclusion each time: Technology adoption fails when not driven by executive vision. Adopting AI is simply too challenging for most organizations to do when absent of long-term vision supported from top-down. You simply must define the organizational direction of travel for AI at the CXO level.

© Cloud Lighthouse LTD, 2024

This is the stuff of many, many business leadership books written over the years, so we don't want to be too prescriptive here. Executive vision can take many forms, but the bottom line is that your executive vision for AI (or any technology) must frame everything that follows so that it is crystal clear why the organization is embracing this technology and what the organization collectively aspires to achieve from its adoption.

We've helped many organizations craft their vision for AI. The anonymized aspirations below provide a great example of a top-level executive vision at a real-world enterprise firm.

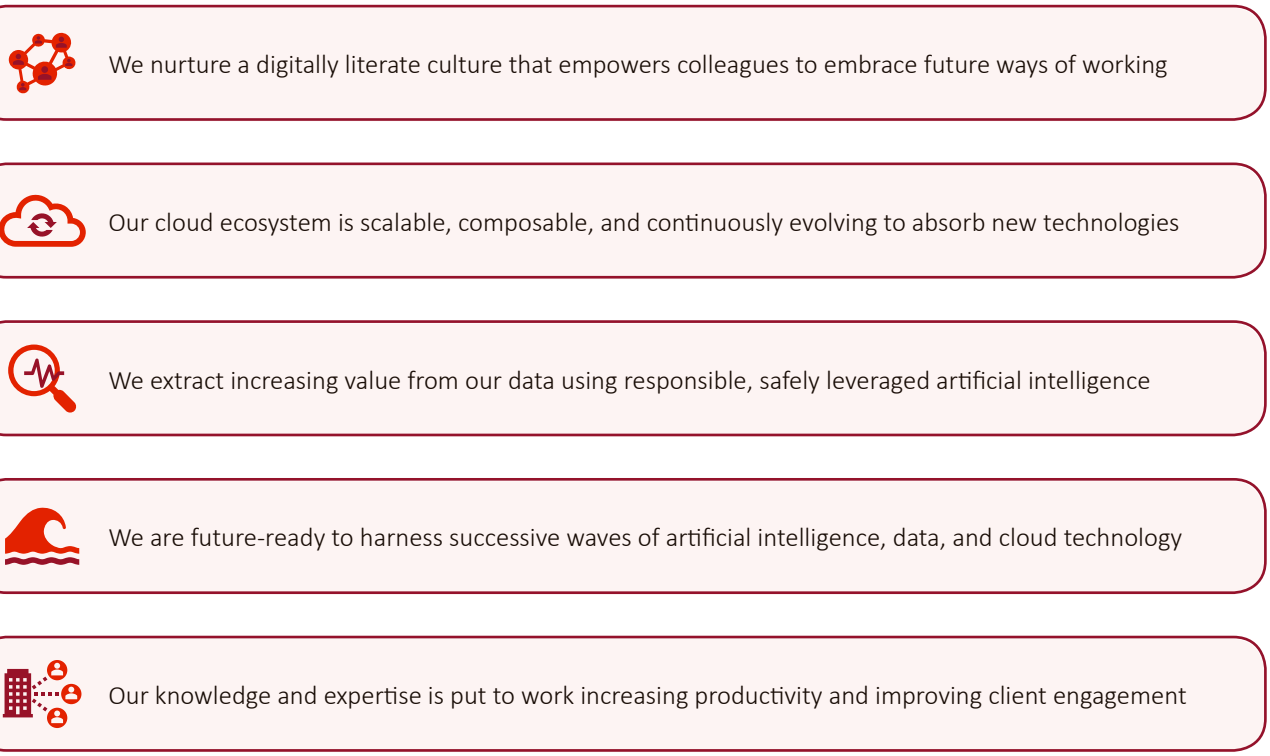


Figure 5: A top-level executive vision for AI framing the aspirations for an organization embracing AI. This example includes building a digitally literate culture, creating a scalable and composable cloud ecosystem, extracting value from data safely, adopting a future-ready mindset and increasing AI knowledge and expertise.

Notice that our vision is aspirational, succinctly describing not just what we hope to achieve with artificial intelligence, but what we hope to be as an organization that has embraced artificial intelligence. Further, only two of our five aspirations explicitly mention AI at all. This is important: We often hear folks talk of AI as if it were a product, but it's not a product at all. AI is quickly being woven through nearly every aspect of our work lives (and our lives in general), and it equally depends on the proper functioning of other domains including data, applications, technical governance, business process, digital culture, and the mission of the organization itself ("improving client engagement", in the case of the example above).

© Cloud Lighthouse LTD, 2024

Finally, a well-crafted executive vision ought to go beyond headline aspirations to describe what we call “targeted outcomes”, which is to say, to define the outcomes the organization hopes to achieve in actualizing its aspirations. Think of targeted outcomes as adding specificity to your aspirations, not necessarily hard, quantifiable specificity, but a clear articulation of what it means to (for example) “Extract increasing value from our data using responsible, safely leveraged artificial intelligence”:

- The data platform offers a mastered single source of truth for the most mission critical data domains;
- Data is addressable by AI and aggregated from different sources as part of our data platform;
- AI is deployed consistently and with governance guardrails in place;
- “Low-hanging fruit” (incremental) AI capabilities quickly deliver lower-risk capabilities to our colleagues;
- We pursue a risk-sensitive portfolio of “differential AI” customized for the firm.

Whatever your executive vision, it is important to lead with it, to prioritize the AI investments that best align to it, and to evangelize it such that colleagues both in IT and the wider business understand the all-important “why”.

Actionable Roadmap

Strategy without action is like the rule of law on a deserted island. Irrelevant, even to the birds.

The trick to making strategy relevant is to pair it with an actionable roadmap, really the actions, activities, even full-blown projects that will be undertaken to actualize our aspirations and achieve our targeted outcomes.

There’s an old adage attributed to American General and later President Dwight D. Eisenhower that “plans are useless, but planning is indispensable”. Take it to heart. Firm roadmaps quickly grow obsolete even under stable conditions, and the only thing stable about the evolution of AI is its acceleration. An actionable roadmap for your AI strategy that runs more than 12 to 24 months into the future is far too long. We’re only able to achieve that level of durability by taking to heart our first principles:

- **AI strategy should offer immediate value to the organization beyond specific AI- driven workloads because the nature and value of these workloads will remain unclear for some time.** In other words, make investments in modern data platform technology that will pay dividends not just in AI but in analytics, business intelligence, search, etc.;



© Cloud Lighthouse LTD, 2024

- **AI strategy must be flexible: able to absorb tomorrow what we don't fully grasp today.** It's wise to plan 24 months in advance, but it is equally unwise to assume that you'll not be regularly revising those plans as things evolve.

Start by formulating up to five big priorities, inspired of course by your executive vision. If, for example, you have established five aspirations as part of your vision, try first to devise one major priority aligned with each aspiration. For example, referring to the executive vision shared earlier, we might establish the following topline priorities:

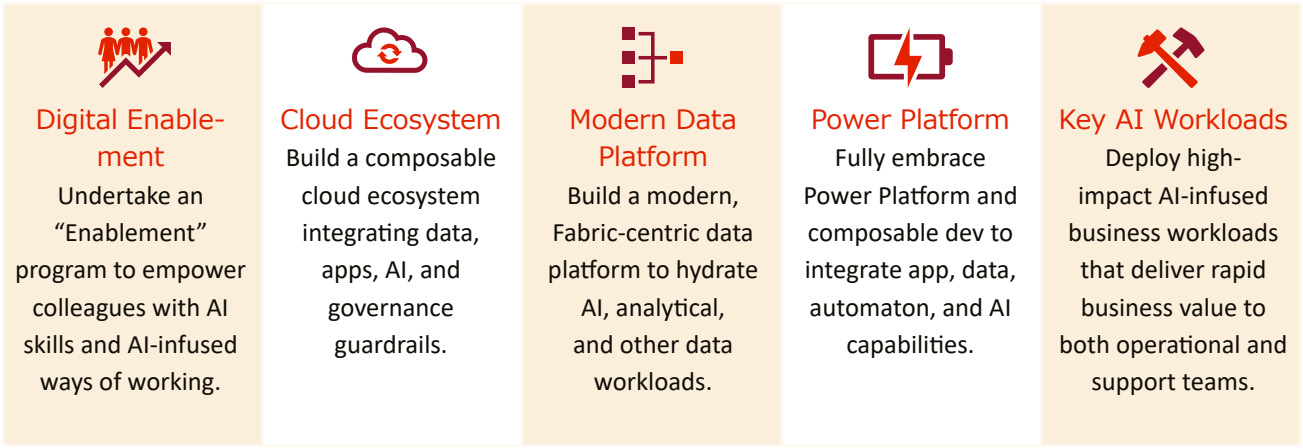


Figure 6: These five top-line priorities are representative samples similar to those that we see many organizations priorities as part of their early AI strategy.

Then, add specificity to these priority buckets with 3-5 milestones that the organization will achieve in the next 18 (give or take) months. It's helpful to break these down into three horizons of three to six months each, and be prepared to drastically rework the milestones in the third horizon given that they're likely at least 12 months out.

Finally, keep in mind that you are likely to uncover specific actions or milestones you need to undertake simply by evaluating where the organization is in each of the twenty-five AI maturity dimensions outlined earlier. For example, if you assess early on that the organization is particularly immature in the dimensions of "AI Development Tools" and "Digital Literacy", it's wise to prioritize milestones that are likely to close those maturity gaps as part of your actionable roadmap. Finally, invest in your stakeholder relationships to ensure that your roadmap is mapped back to those stakeholders, clear feedback loops are in place, and updates are shared so that you bring colleagues on the proverbial journey.

© Cloud Lighthouse LTD, 2024

Ecosystem Map

An ecosystem map is a high-level architectural diagram of an organization's cloud ecosystem, and something that every organization ought to create at the start and continuously evolve as they progress on their AI journey.

The "map" metaphor is instructive here. It is used to distinguish an ecosystem map from the various forms of architectural diagrams, nearly all of which tend to include more technical minutiae than a typical ecosystem map. Whereas an architectural diagram provides specific parameters for specific technical solutions, an ecosystem map presents a higher-level, more visionary view of an organization's cloud ecosystem.

To make an analogy to architecture in the physical world:

- **Solution architecture** provides schematics - floor plan, dimensions, electrical wiring, ventilation, plumbing - from which a building is constructed;
- **Enterprise architecture** provides plans for specific neighborhoods or systems such as a subway or electrical grid;
- **Ecosystem architecture** and, by extension, an ecosystem map shows us the entire city.

Thinking of an organization's cloud ecosystem as a city, we then conceptualize the next-level-down component parts of the ecosystem as "neighborhoods" (we might have also called them "boroughs"). Cities the world over are pieced together this way: Downtown, Seaport, Southie, etc. in Boston; Greenwich, Soho, Canary Wharf, etc. in London (though you're forgiven if you thought I was talking about New York until you got to "Canar y Wharf"); Palermo, Recoleta, Puerto Madero, etc. in Buenos Aires; Norrmalm, Gamla Stan, Kungsholmen, etc. in Stockholm. The list goes on.

Each of these neighborhoods share the quality of dividing their city into smaller pieces, each often with their own distinct culture, aesthetic, or purpose.

Like cities, ecosystem maps are constantly evolving and changing. To prevent your ecosystem from becoming overcrowded, stagnant, or unable to meet the needs of its expanding 'population,' it's essential to revisit, revise, and adapt your Ecosystem Map on a regular basis.

© Cloud Lighthouse LTD, 2024

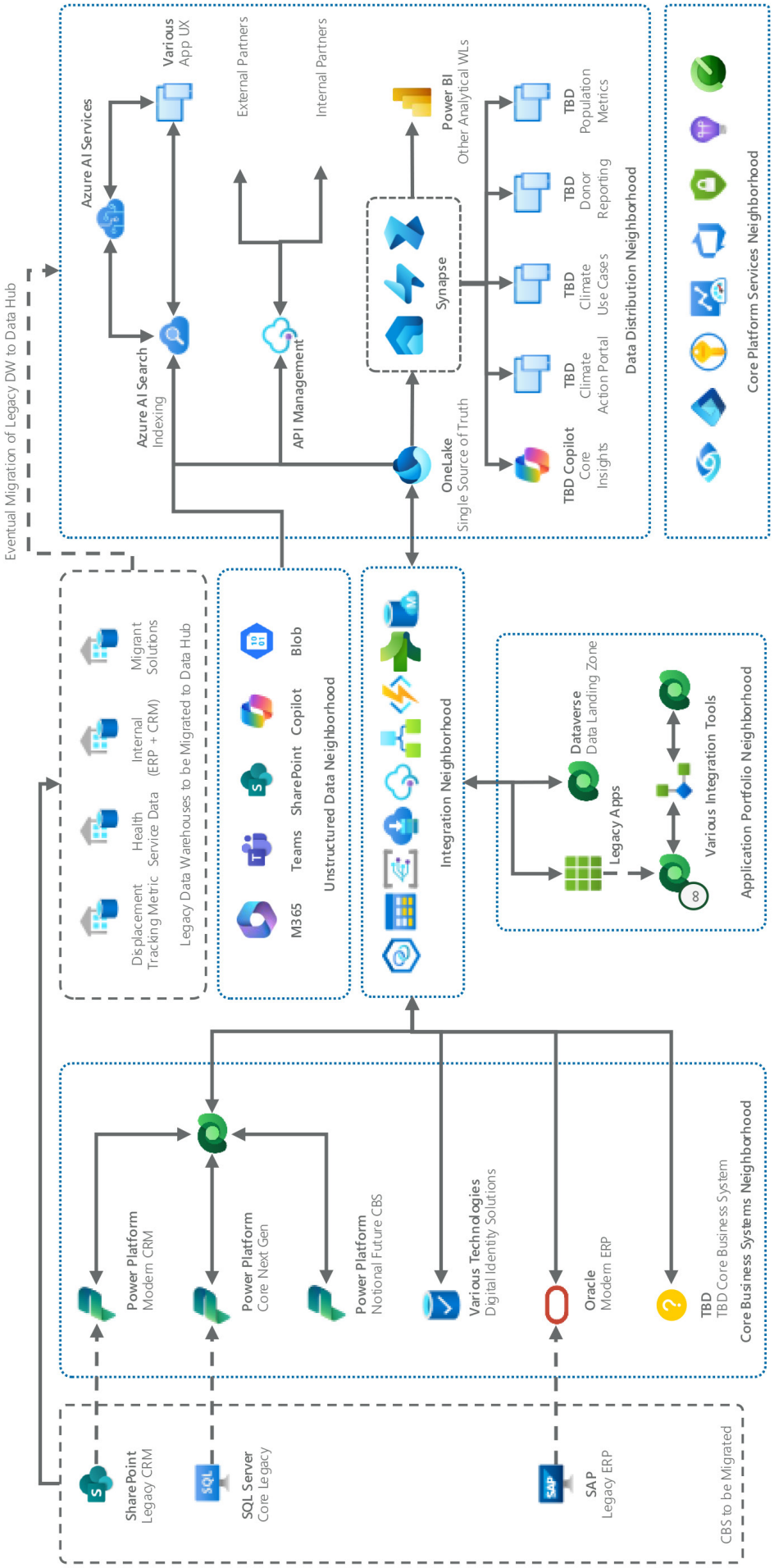


Figure 7: This is a good, representative ecosystem map that a global enterprise organization recently produced as part of its AI strategy.

The example above is one of our recent favorites produced in collaboration with a global enterprise organization as part of their AI and broader cloud strategy. Note that:

- Specific AI and other data products are identified in the Data Distribution Neighborhood (bottom of the box on the far right);
- The ecosystem map shows how data will flow from the organizations' core business systems (including some yet to be identified, which is just fine) and application portfolio such that it hydrates various data distribution points, including but not limited to AI workloads;
- Migration of legacy applications and legacy data warehouses is identified as a priority - a part of data consolidation that we will discuss later.

Mapping the cloud ecosystem is a key element of our AI strategy because the efficacy of any AI workload is directly related to the quality of the data upon which the workload's model is trained or augmented. Think back to our earlier foundational discussion of how generative AI acts on enterprise data. Mapping, evolving, and maintaining the organization's cloud ecosystem map provides the essential high-level technical architecture underpinning our AI strategy.

An ecosystem map is part of a broader approach to what we call ecosystem-oriented architecture (EOA) that we will discuss more deeply as part of the Ecosystem Architecture pillar in the next section.

Programmatic Rigor

Roadmaps don't get driven on their own, and architecture doesn't build itself. With an executive vision, actionable roadmap, and ecosystem map in hand, it's crucial that the organization institute the programmatic rigor required to navigate that roadmap and bring its vision to fruition.

We could have called this "programmatic discipline". In this dimension, leaders ought to ask themselves if their organization broadly, and their IT teams specifically, are sufficiently focused, possess the rigor and discipline, and operate at a cadence necessary to achieve the milestones they've set forth in the time planned. This is program management 101 stuff, so we'll not rehash it here. Suffice it to say that the organization must act rigorously and consistently to marshal the resources, direct action, monitor progress, and feed lessons learned back into its AI strategy. Organizations that lack this discipline will entirely fail to achieve their vision for AI (or any technology).

As the organization's AI strategy and program mature it is important to ensure clear lines of communication and feedback are established across organizational lines and stakeholders, ensuring that there is visibility of the program advancement and that success metrics and milestones are being achieved. Regular self-checks are essential to maintain a coherent DevOps strategy. Regular peer reviews, relevant testing, and comprehensive documentation must be standard practices. On the delivery side, IT teams need the discipline to adhere to delivery methodologies, avoiding the temptation to overproduce or create unnecessary deliverables out of an overzealous desire to future-proof. Being future-ready does not mean deploying every conceivable capability just in case it might be needed. Instead, it's about maintaining high standards in delivery and resisting the use of patchwork or temporary solutions that could undermine the very foundation of your platform. Effective programmatic rigor requires not just planning but an ongoing commitment to quality and strategic alignment throughout the development lifecycle.

Center for Enablement

The final dimension in our Strategy and Vision pillar is organizational, putting in place the team or organizational unit required to drive our AI strategy forward. The Center for Enablement (CFE or C4E) concept is rather a departure from IT organizational concepts of old, though, representing a shift from controlling processes to enabling people.

Contrast it to the well-worn "center of excellence" that has historically focused on maintaining standards and enforcing compliance within various technology domains. A Center for Enablement, however, is dynamic: continuously evolving to adapt to new technologies and business needs. It moves beyond reactive governance to more proactively drive the organization's executive vision for AI.

A well-rounded C4E will broadly focus on the following activities:

- **Strategic refresh** ensuring that the organization's AI strategy is continually re-evaluated and refreshed to reflect changes in the technology, business environment, and the performance of the organization's portfolio of AI initiatives;
- **Programmatic rigor**, with the C4E taking responsibility for managing the portfolio of AI initiatives across the organization. This must include primary accountability for the achievement of milestones on the actionable roadmap, the development of key workloads, and the organization's ongoing maturation across the entire AI Strategy Framework;

© Cloud Lighthouse LTD, 2024

- **Facilitate human connection**, creating opportunities for colleagues to collaborate, innovate, and build communities of practice across the organization. This promotes collaboration, networking, brainstorming and building new skills in accordance with an ever-changing environment. The C4E must enable colleagues to succeed in their use of AI;
- **Drive a culture of continuous improvement and innovation**, encouraging a mindset of experimentation and learning, where insights are not just consumed but acted upon, iterated, and improved;
- **Monitoring and metrics** of the organization's AI initiatives through advanced analytics and AI monitoring itself, identifying patterns and trends in a continuous loop to inform strategy. This must include continual assessment of the organization's aggregate AI maturity using the AI Maturity Model (discussed in a later section). Learn more about this in the Monitoring and Metrics dimension in the Scaling AI pillar.

The **Ecosystem Design Authority** (EDA) offers a sound model through which the C4E can facilitate the success of the organization's AI strategy across technical domains that it may not directly control. Think of the EDA as a collaborative, standing working group; "air traffic control" for the cloud, landing the workloads and technical services of different technology domains in the cloud ecosystem. The Ecosystem Design Authority (EDA) ensures that (a) architecture and technical decisions are aligned with the cloud and AI strategy, and (b) that workloads and technical services are assembled for the benefit of the whole ecosystem, not in service to a specific technical domain.

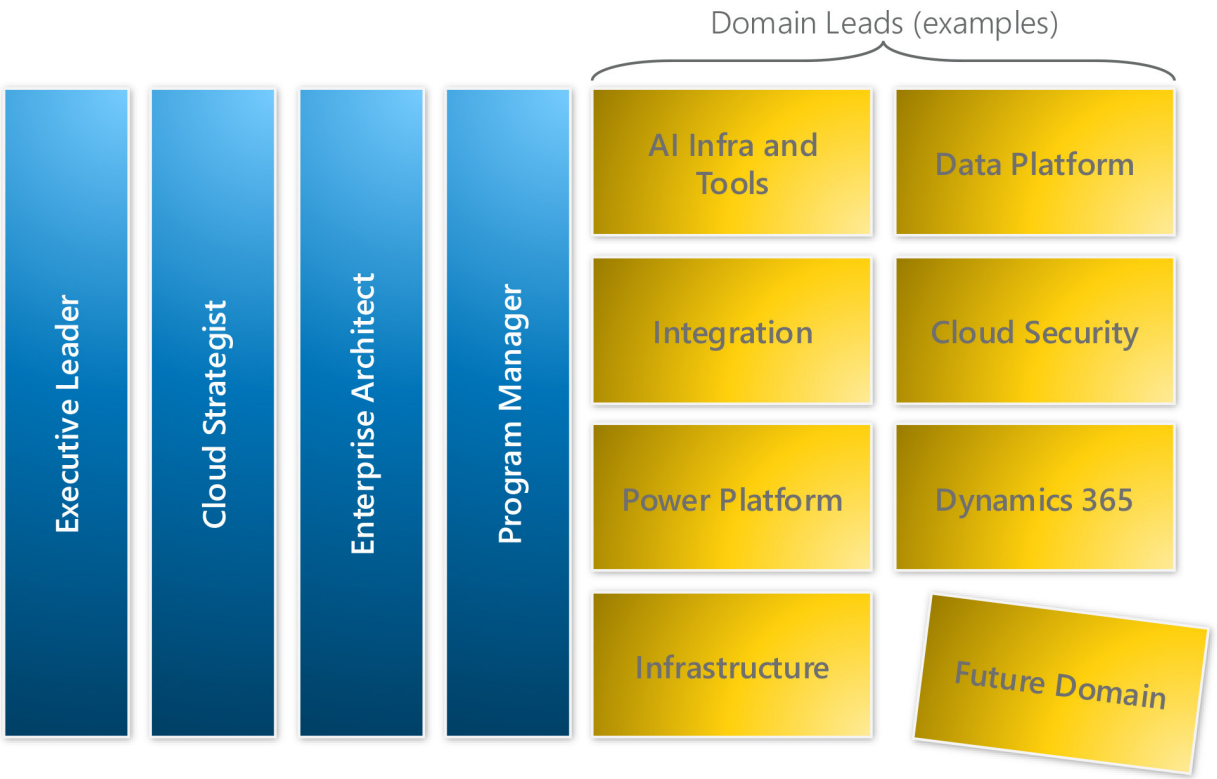


Figure 8: A notional Ecosystem Design Authority model, though note that the specific choice of technical domains should be tailored for the organization specifically.

© Cloud Lighthouse LTD, 2024

A few notes on the EDA model depicted above:

- ”Domains” segment technical disciplines within the ecosystem, and are fluid over time;
- Each domain is represented by one cloud solution architect or technical leader, regardless of the number of projects or work streams in the domain;
- Domains work together across ecosystem neighborhoods and the ecosystem at large;
- The executive leader is a CIO-level or direct report able to make decisions on behalf of the organization;
- The following are key roles within the Center for Enablement:
    - Cloud Strategist is “air traffic control” for the ecosystem, ensuring that technical services fit together and are aligned to strategic priorities;
    - Enterprise Architect oversees architecture and technical work on a day-to-day basis;
    - Program Manager is responsible for the non-technical programmatic rigor and execution of the AI strategy.

Here’s where a Center for Enablement becomes truly essential not just to the AI strategy but to the organization’s ecosystem more broadly, where the focus is building adaptable scalable cloud ecosystems that grow with the organization itself.

© Cloud Lighthouse LTD, 2024

CloudLight.house
Strategic architecture for the Microsoft Cloud

Pillar Two
Ecosystem Architecture



Ecosystem Architecture



Think about the way most IT has evolved over the course of several decades. For much of this history, organizations acquired software by ginning up a specific need or "use case," which was often followed by a basket of requirements pertaining to that use case either given to a team for development or turning into a procurement. Infrastructure, whether on-premise or cloud, was then deployed to accommodate the specific, forthcoming, solution.

Ecosystem-oriented architecture (EOA) inverts this approach. Ecosystem architects seek first to build a cloud ecosystem, that is, a collection of interconnected technical services that are flexible or "composable," re-usable, and highly scalable. The ecosystem then expands, contracts, and is adapted over time to accommodate the workloads deployed within it.

EOA is ideal for scaling AI because it promotes data consolidation as a first principle, avoiding the de-consolidation that point solutions tend to promote via the use of data services tied specifically to the application itself and point-to-point data integrations with other point solutions.

Earlier we shared a sample ecosystem map (see the Ecosystem Map dimension in the Strategy and Vision pillar). Let's now dig into the concept of EOA a bit further, as it is essential to understanding the Ecosystem Architecture pillar.

We've created what we call the "Reference Ecosystem", essentially a composite of dozens of enterprise organizations' cloud ecosystems that we've studied across many industry sectors and geographies.

© Cloud Lighthouse LTD, 2024

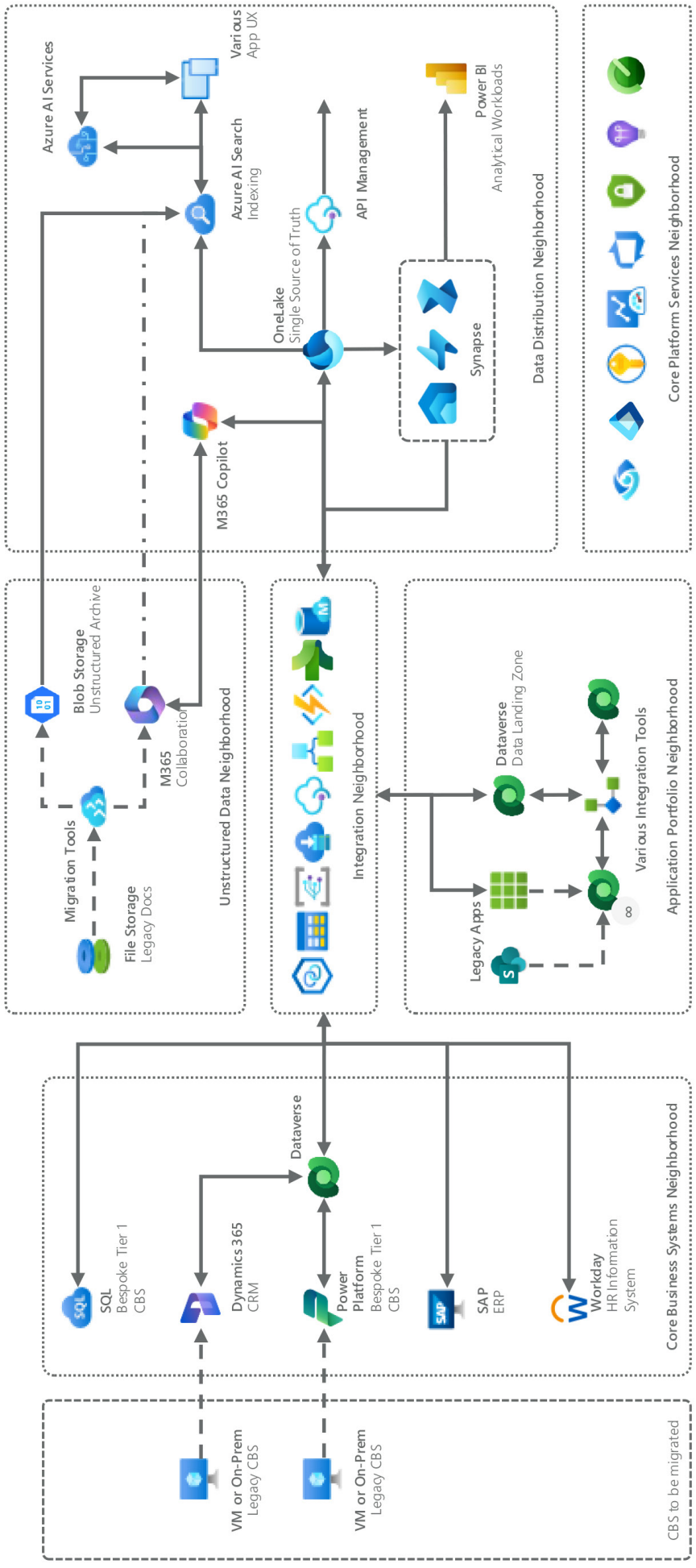


Figure 9: The "Reference Ecosystem" is a composite of ecosystem-oriented architecture (EOA) across many different organizations in many different industries. Consider it the "North Star" for an effective EOA across an entire cloud estate and use it as a reference to build your own.

To orient you, the map makes a clear analogy between the cloud ecosystem and a city divided into “neighborhoods.” These neighborhoods are conceptual, in other words, one should not necessarily construe them as hard logical boundaries such as environments, subscriptions, or tenants. Rather, the ”neighborhood” concept helps us understand the categories, relationships, and boundaries of technical services and workloads, in what can be a vast cloud ecosystem, in a relatable, clear way.

Adopting an ecosystem-oriented architecture across an enterprise organization supports your AI strategy in many ways. Fully adopted EOA is the pinnacle of many of the strategies we’ve already discussed, the enterprise architecture “North Star,” if you will, in the era of AI:

- Use this Reference Ecosystem to orient you to the first four dimensions of our Ecosystem Architecture pillar, specifically Core Platform Services, Data Distribution, Integration, and Business Applications (which combines the Core Business Systems and Application Portfolio neighborhoods shown in the reference);
- EOA speeds the deployment of AI workloads, creating the conditions for those “quick wins” that many think they can achieve with AI only to find out that they’ve not done the work necessary around data consolidation, readiness, and scaling to make this work. Time to value is much shorter when your data is already consolidated, indexed, governed, secured, and you have supporting services such as application lifecycle management ALM (and MLOps) in place;
- EOA organizes and integrates “traditional” workloads found primarily in the Core Business Systems, Application Portfolio, and Unstructured Data neighborhoods in a way that supports AI workloads downstream. This integration mitigates the struggles many will have preventing new silos of unconsolidated data emerging as they scale;
- Undertaking a transition to ecosystem-oriented architecture aligns well with the two core principles that underpin the AI strategy:
    - Your AI strategy must be flexible, able to absorb tomorrow what we don’t fully grasp today . Cloud ecosystems are designed to be metaphorically living, breathing entities that evolve to meet today’s needs and achieve the promise of future innovation;
    - Your strategy should offer value to the organization beyond specific AI-driven workloads because the nature and value of these workloads will remain unclear for some time. Your transition to EOA is a great investment in AI, but also brings value to the organization in terms of time to value, resolution of technical debt, retirement of legacy licensing and capacity costs, reduction of organizational risk around data governance and security, and in the form of workloads such as search, outside integration, analytics, and reporting.



© Cloud Lighthouse LTD, 2024

Core Platform Services

Core Platform Services include many of the infrastructure, security, governance, management, and monitoring services used across a cloud ecosystem. Largely synonymous with a “cloud landing zone”, the Reference Ecosystem shows (left to right) Purview, Entra ID (formerly Azure Active Directory), Key Vaults, Azure Monitor, Azure DevOps (ADO), Security Center, Sentinel, Application Insights, and Power Platform Managed Environments as examples. There will surely be others, but we have chosen these as representative “core services” that nearly every modern cloud ecosystem should contain to technically support an organization’s AI strategy.



Figure 10: A magnified view of the Core Platform Services Neighborhood from the Reference Ecosystem.

Core Platform Services, and Microsoft Purview in particular, also play a vital role in data governance, which refer to measures taken to secure, govern, cleanse, establish lineage and compliance, manage metadata, etc. In other words, have you established the conditions for your data to produce quality responses when consumed by AI?

Though establishing baseline data governance through Microsoft Purview is very much a part of your core platform services, it’s too important a topic to hide away. So, we’ve broken it into its own dimension which we’ll discuss separately as part of the Scaling AI pillar later in the paper.

In any case, our core platform services approach to data readiness involves good, old- fashioned best practices around building and maturing your cloud landing zone and the ongoing maturation of your cloud estate. Given the convergence of data platform infrastructure and business applications, this must necessarily include deployment of typical Azure as well as Power Platform security and infrastructure measures. Here we’re talking about nuts-and-bolts capabilities such as identity management via Microsoft Entra ID (formerly “Azure Active Directory”), as well as services that are less well-known outside of the circles of technologists who specialize in them. Examples of these are Key Vaults, Azure Monitor, Security Center, Managed Environments, etc. The essential task here is to mature your cloud infrastructure in accordance with industry best practices, with an emphasis on data security and governance.

© Cloud Lighthouse LTD, 2024

Technical documentation on this subject abounds, so we'll not overextend this discussion. The real question that every organization must answer in evaluating its current state and direction for core platform services is whether it has built a broadly based cloud landing zone that provides for the technical availability, management, and governance of its cloud infrastructure (Azure, in the case of Microsoft), data platform (Fabric and related services), and business applications (Power Platform, Dynamics 365, and Azure application services), as well as the application lifecycle management (ALM) necessary to shepherd both traditional and AI-based workloads to production.[1]

Data Distribution

This is amongst the lengthiest discussions of any single dimension in this white paper, owing to the centrality of data to any AI strategy.

Data Distribution provides for data consolidation in (for example) OneLake, and for all manner of "downstream" data distribution such as search, APIs, data warehousing, analytical workloads, data science and data engineering workloads, and use of data in AI-driven scenarios. We also place Microsoft 365 Copilot in the Data Distribution neighborhood as it - like many other scenario-specific Copilots that also reside there - relies on consuming, interpreting, and otherwise distributing data in response to user prompts. Note that M365 Copilot is hydrated with data from Microsoft 365 via the Microsoft Graph and may be configured to consume data from elsewhere in the data estate, as well.

Pre-requisite to your use of AI is consolidation of data in storage services where it can be aggregated and accessed by AI services. Data is the essential fuel without which AI models cannot be trained nor have the capacity to act on the information that makes them valuable. For all the advancements in cloud technology of the last decade, most organizations are home to vast unconsolidated stores of data. Your data lives in OneDrive, spreadsheets, desktops and one-off databases often sitting beneath point solutions, and - if you're lucky - some of it lives in lakes, warehouses, lake houses, and properly managed databases.

[1] The technologies listed here are Microsoft-specific, so note that organizations with a more technical landscape may be using alternatives to the services listed.



© Cloud Lighthouse LTD, 2024



Figure 11: Consider this model when mapping out your data consolidation approach. Remember that not all data needs to be consolidated, for example, in the case of "seemingly trivial datasets" that need to be secured, but if not being used in AI or other distributive workloads, may not need to be consolidated.

This disarray results from the convergence of a lack of capability and a lack of will. Microsoft Fabric, announced in May 2023, represents a significant investment and resulting leap forward in the capability side of the equation. Microsoft's Corporate VP of Azure Data, Arun Ulagaratchagan, heralded Fabric as "empowering data and business professionals alike to unlock the potential of their data and lay the foundation for the era of AI." Though Fabric offers capabilities beyond data consolidation, a significant share of its value comes from its capabilities to move and store consolidated data. OneLake, built atop Azure Data Lake Storage (ADLS) Gen2, provides one single lake for the entire organization (with a 1:1 relationship between OneLake and the tenant) and one copy of data for use in multiple data distribution scenarios including AI and beyond, such as in enterprise search, distribution via (say) API, and analytical workloads in Power BI.

Data consolidation refers specifically to the consolidation of data from across your cloud estate into storage technologies that can be accessed and used by AI. This is likely to be achieved through a variety of techniques including: copying, one-time migration with the intent to retire the legacy data source, data integration (which is, implicitly, ongoing), standardization on one or a small number of future-ready transactional data services for app dev, and employing "shortcuts" in Fabric through which data is shortcut from its source into OneLake (analogous to how a file in OneDrive may be shortcut from its source to another location).

© Cloud Lighthouse LTD, 2024

To be clear: We are not suggesting that all data needs to be copied to a single location, nor are we advocating for the use of a single data storage technology within any one organization (which is, frankly, a preposterous idea). We'll spare you a deep dive into each of the consolidation techniques mentioned above and say simply that ecosystem, enterprise, and solution architects possess a variety of context-appropriate data consolidation techniques that may be employed to get from your current state of disarray into a consolidated, future-ready state.

Think of data consolidation not as a single, fixed end state but rather the existence of your enterprise data estate along a spectrum. At one end we have a wholly unconsolidated stat - — the technical term for this ought to be "epic mess" - and at the other end we have a truly single source of truth for data.

Though that "truly single source of truth for data" sounds romantic, and is a fantastic aspiration, it's also not likely to be achieved by most organizations in anything resembling the short term.

Each organization's happy medium is likely to fall somewhere on this spectrum.

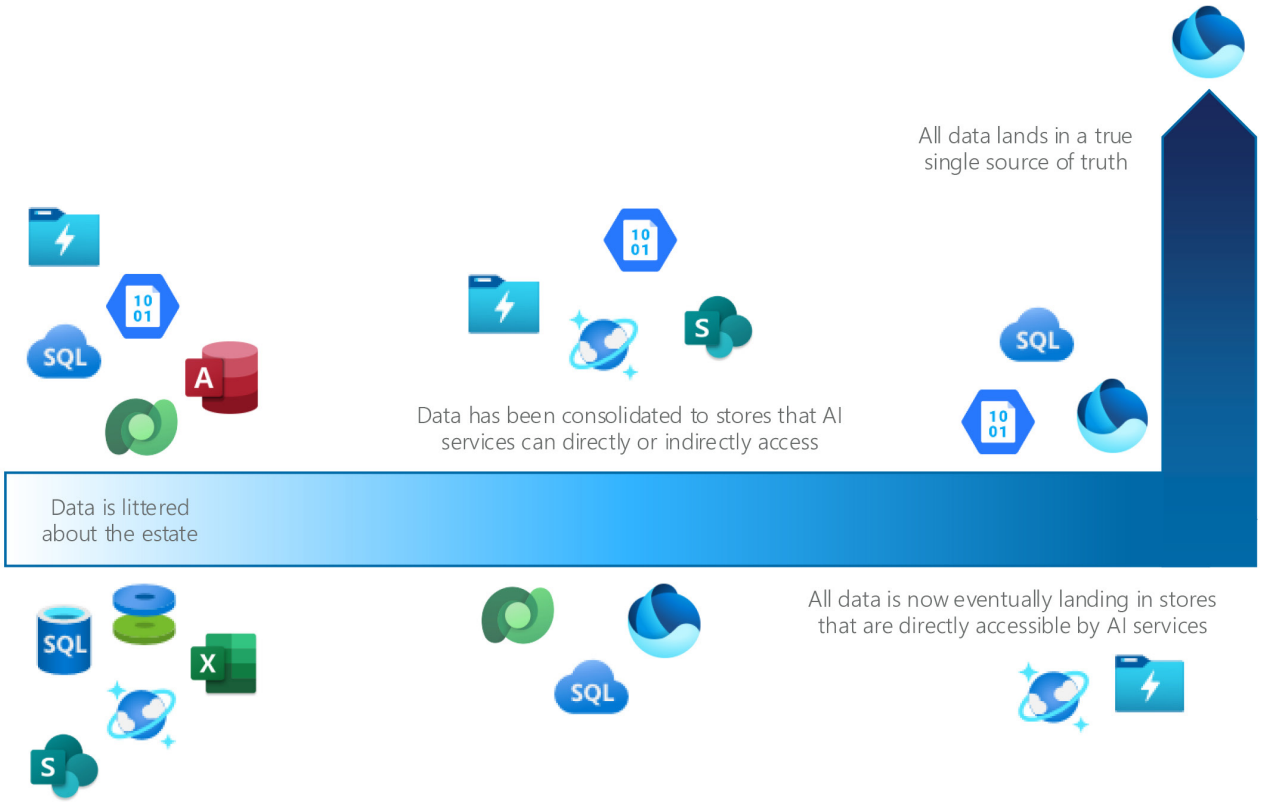


Figure 12: Data consolidation occurs on a spectrum. Do not imagine that a truly single data store is the goal.

We also need to distinguish between application data vs. data stored and staged for downstream distribution (including for use in AI workloads). For example, Dataverse (the green vortex icon in the previous diagram) is an excellent data service atop which to build applications, though it is more efficient to consolidate data from Dataverse into the lake when staging it for downstream distribution. This is the expected pattern in Microsoft business applications—be it custom solutions built with Power Platform or Microsoft's pre-built Dynamics solutions—where data is pushed via copy or shortcut from Dataverse (the applications' transactional data service) into the lake for downstream distribution.

Most importantly, regardless of the upstream data services you're employing, the data that you wish to be addressable by downstream workloads, be they AI, enterprise search, analytical (Power BI), etc., must ultimately land in storage services that can feed a multitude of downstream distribution scenarios (including, but not limited to, AI).

So, returning to our earlier model, most organizations are likely to land on a data consolidation architecture that looks something like the diagram shown below. Here we see data migrated (dotted line) out of, for example, Access databases and Excel files into Dataverse, on-premise SQL into Azure SQL, or network storage into Azure Blob. You'll then find yourself with a fairly sizable transactional data estate underpinning most of your applications whose data flows downstream to services such as OneLake. For example, data from Azure SQL is pushed or shortcut into the lake.

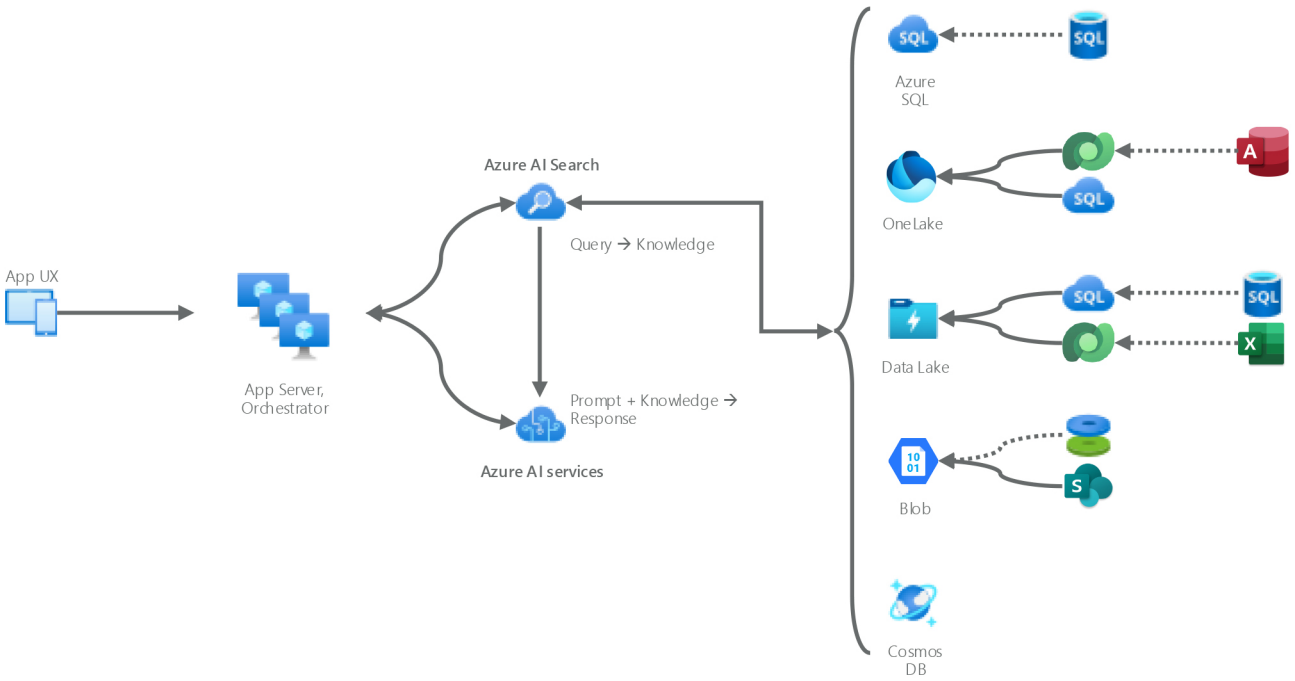


Figure 13: A notional architecture for data consolidation in practice working with the AI model we discussed earlier.

Think of these migration and integration paths as the rivers and tributaries by which data flows to the lake and other related, consolidated services. It's rather basic data platform 101 stuff. Given how incomplete this architecture is in most organizations - and how essential it is given that it lays the foundation for your AI strategy - it really does bear including here.

Though we possess increasingly sophisticated technical capabilities to make data consolidation a reality, successfully executing on this pillar of your AI strategy requires the organizational will to do so. Grassroots action within a typical organization is insufficient and likely to result in further, randomized diffusion of data across the estate. So too is point-solution oriented architecture, which tends to neglect data consolidation in favor of data storage for application- specific data services.

Data consolidation is, for this reason, one of the most important executive-led priorities in technology today as IT organizations the world over race to build future-ready ecosystems.

Let's now turn our attention towards what should happen with the data once it has been consolidated.

The architecture below shows a fairly typical "data distribution neighborhood", largely as shown in the reference ecosystem. We've simplified it by showing some undefined number of unstructured and structured data sources being consolidated into the neighborhood.

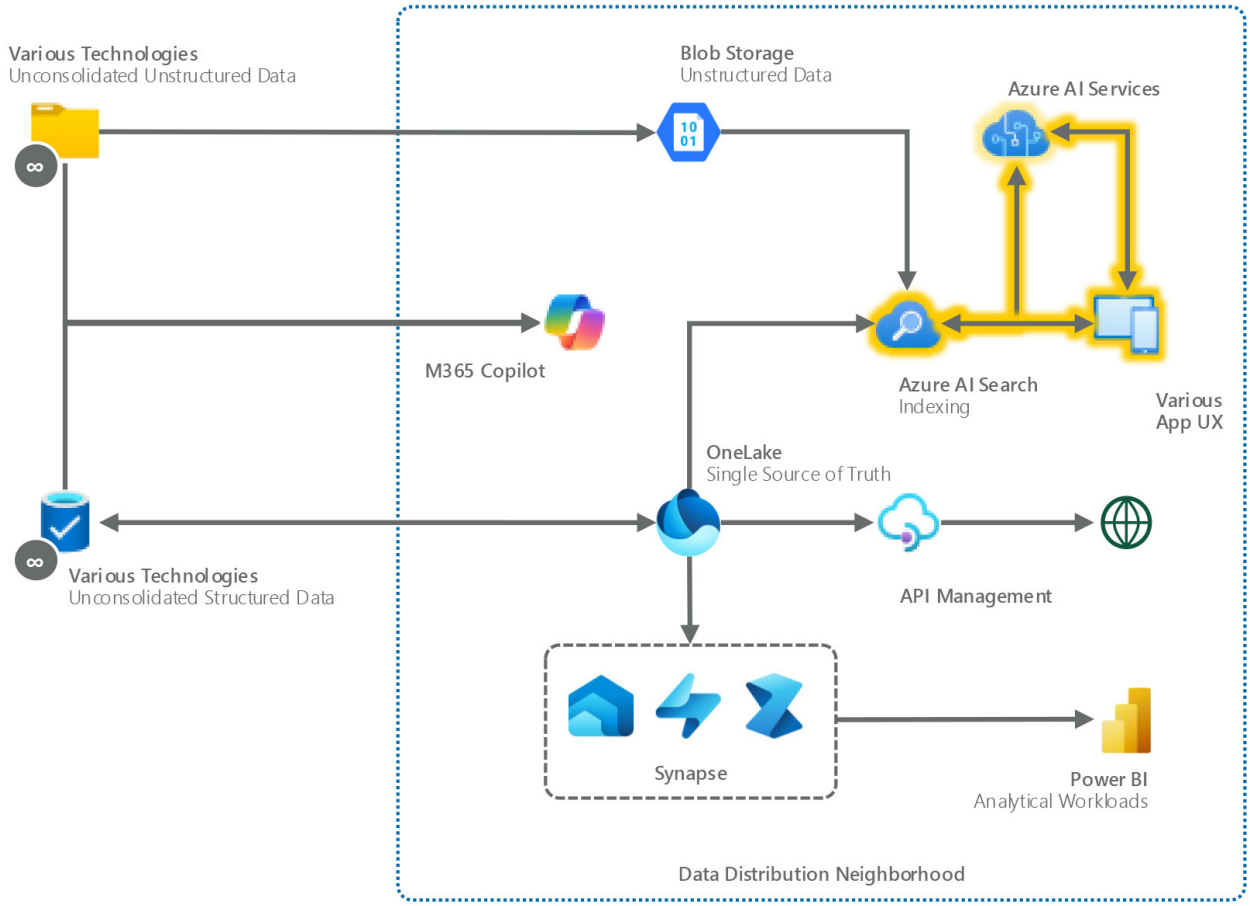


Figure 14: Notional data distribution "neighborhood" with Azure AI Search (glowing yellow) called out as the linchpin for indexing data for use in AI workloads. Other examples of data distribution are shown illustratively

There are several key components to note below:

- We're representatively showing most structured data consolidated to OneLake and most unstructured data consolidated to Blob Storage. We could have made other technology choices here (though OneLake is pretty undisputable, in our opinion), but have chosen representative examples;
- Azure AI Search (formerly "Azure Cognitive Search") is important, and we've come a long way since the days when most people knew Cognitive Search as an indexer and provider of traditional end user search across enterprise data. Yes, it still does that, but AI Search is increasingly becoming one of the principal "front doors" through which AI walks in order to use enterprise data. It turns out that a vast index of enterprise data is extraordinarily useful to AI workloads, hence the RAG pattern discussed much earlier. In any case, the implementation of AI Search ought to rank alongside Purview and OneLake in terms of specific data platform services required by an organization seeking to become future- ready in the age of AI. Get it. Implement and configure it. Index your consolidated data now and continually into the future.
- The RAG pattern itself is shown in a simplified form, glowing in yellow as a callout.
- We're also extracting value from our consolidated data in other distribute workloads as well, including integration with outside third parties via API, analytical workloads in Power BI, and some combination of data warehouse, data engineering, and data science workloads using the Synapse technologies integrated with Microsoft Fabric.

High quality data distribution supports our objective for an AI strategy that, "offers immediate value to the organization beyond specific AI-driven workloads because the nature and value of these workloads will remain unclear for some time."

The diagram above illustrates how sound data readiness - and data distribution, specifically, shown here - supports non-AI workloads as well, including enterprise search, data distribution to third parties, and analytical workloads. The particulars of your data platform architecture may vary, but data distribution is a sound investment within and outside the scope of AI.

Integration

Integration includes (left to right in the diagram) technology-specific integration services such as OneLake shortcuts in Microsoft Fabric, virtual tables in Microsoft Dataverse, event-driven integration services such as Event Grid and Service Bus, use of APIs via API Management, logic-driven integration such as Logic Apps and Azure Functions, batch integration relying on Azure Data Factory, and enterprise master data management (MDM).

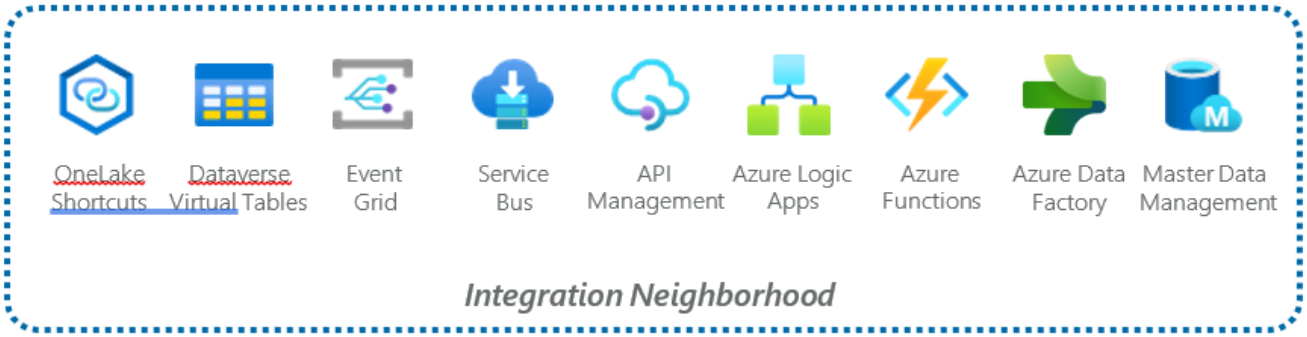


Figure 15: A magnified view of the Integration Neighborhood from the Reference Ecosystem.

A lineup of icons does not an architecture make, of course. An Integration Neighborhood - also known as “integration platform” or an “integration layer” - can be incredibly architecturally complex. Organizations that have not already established one using some combination of these and related technologies, alongside best practices and patterns reusable across their ecosystem, will certainly want to do so in order to facilitate data consolidation and scale AI.

The organization’s AI strategy, however, need not be concerned with architectural specifics at a detailed level, rather with establishing and maturing their Integration Neighborhood, and normalizing its use across the ecosystem to avoid the types of “spaghetti web” integrations that are costly to maintain and the antithesis of what AI needs to scale.

Business Applications

Our Business Applications dimension broadly consists of two neighborhoods from the Reference Ecosystem:

- **Core Business Systems** which includes the “Tier 1” business applications common in many organizations such as ERP, CRM, HRMS, etc.;
- **Application Portfolio** which may include “Tier 1” applications, but often includes solutions aimed at smaller audiences or more niche business processes of the Tier 2 (“business important) and Tier 3 (“productivity”) variety.

© Cloud Lighthouse LTD, 2024

Modern business applications are important to the enterprise AI strategy because they are both (a) the user experience or touchpoint that most business users have with the technology ecosystem, and (b) often the principal generator of data, or collection point for the organizational data upon which AI will rely.

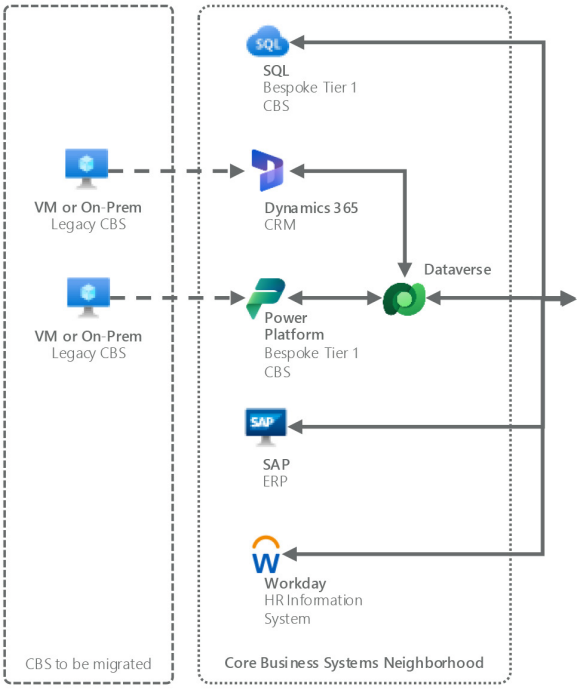


Figure 16: A magnified view of the Core Business Systems Neighborhood from the Reference Ecosystem including legacy workloads to be migrated to modern, future-ready technologies Power Platform and Dynamics 365 (apps that sit atop Power Platform).

The Core Business Systems Neighborhood shown here from the Reference Ecosystem includes custom applications built atop Azure SQL and Power Platform, CRM in Dynamics 365, ERP in SAP, and HR in Workday. These are illustrative examples of the many solutions organizations rely upon in their core business systems neighborhood. We've also included references to migrating legacy workloads to technologies whose underlying data sources (Dataverse, in the case of both Dynamics 365 and bespoke Power Platform solutions) are easily consolidated into our Data Distribution Neighborhood.

Yours may look very different, and that's (probably) just fine.

We'll spare you a lengthy discussion of core business systems, here, and say only that (a) use of modern data services that are readily accessible by AI such as Azure SQL and Dataverse are ideal, and (b) over-fragmentation of core business systems, which is to say, the use of many different technologies and vendors in the Core Business Systems Neighborhood, can introduce unwelcome barriers to AI adoption including added integration needs between systems, denormalized data structures across the ecosystem, and disparate security models.

The Application Portfolio Neighborhood receives separate treatment from its core business systems counterpart.

This is to distinguish between the business important and team productivity solutions often residing in the former, and the business-critical workloads that reside in the latter.

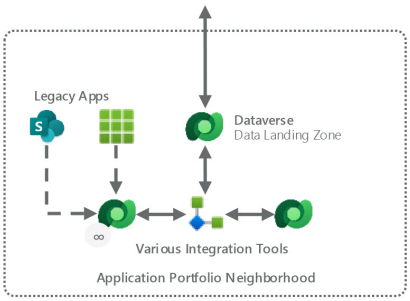


Figure 17: A magnified view of the Application Portfolio Neighborhood from the Reference Ecosystem, including the need for legacy "SharePoint" apps to be migrated to modern, secure technologies such as Dataverse.

© Cloud Lighthouse LTD, 2024

We see the application portfolio as relevant to an organization's AI strategy for two reasons - one of opportunity and the other of risk mitigation:

- First, most organizations are littered with dozens, hundreds, even thousands of apps in this Tier 2-3 space. Many contain more trivial datasets that are either not relevant to AI (at least for the moment) or too unwieldy to consolidate. Many, though, transact data that could be quite valuable in AI scenarios, but has previously been kept in data sources that are inaccessible to AI;
- Second, many organizations have overengineered applications (Power Apps atop SharePoint and older style "SharePoint apps" themselves) that sit atop SharePoint lists as their data source. SharePoint may be appropriate as a data store for small apps that don't transact sensitive data but is wholly inappropriate for apps that require any degree of security. Data in these lists also hydrates the Microsoft Graph with data consumed by Microsoft 365 Copilot, which creates a significant security risk when sensitive data is stored there.

In both cases, your AI strategy ought to reflect moving large chunks of application portfolio data into Power Platform solutions built using Dataverse as a data orchestration layer either to facilitate the secure consolidation of data to the Data Distribution Neighborhood, or to secure data from inadvertent leakage altogether.

Then there is the discipline required to maintain a healthy data platform over time, for example by avoiding the temptation to revert to the legacy pattern of siloed, inconsistent data services or point-to-point integration. Instead, architecting individual applications such that they don't run off and spawn a new generation of one-off data siloes, just as you've exorcised the data demons of your past, maintaining the index across your estate, etc. Today's generation of low- code services, such as Microsoft's Power Platform, can help immensely.

Forrester's 2022 study, The Total Economic Impact of Microsoft Power Platform Premium Capabilities, found 63% of IT decision makers reported that Power Platform helped them eliminate or rein in shadow IT, which is good news for an AI-powered future where data consolidation, governance, security, etc. is paramount.

In fact, it is difficult to scale AI across an ecosystem-oriented architecture absent of Power Platform for three reasons:

- Power Platform enables software and data engineers to build workloads faster, whilst simultaneously allowing non-technical "citizen developers" the ability to self-service many of their own productivity needs in a safe, secure way (thereby freeing professional engineering time to focus on more complex workloads);



© Cloud Lighthouse LTD, 2024

- Every organization we have ever encountered has vast swathes of its data estate buried in the types of productivity solutions (e.g., spreadsheets, Access databases, small third-party "shadow IT" solutions) that are not addressable by AI because they are not consolidated. As such, it is cost and time prohibitive to bring these data out of the shadows and to address the tier one and tier two workloads with professional engineering talent alone;
- Microsoft is building AI capabilities into Power Platform itself, essentially making Power Platform as indispensable an ingredient to your AI strategy as Azure AI Search and Azure AI services themselves. For example:
    - Copilot Studio, Microsoft's service for organizations building their own agents, is built atop what was previously known as "Power Virtual Agents," which is architecturally ingrained with Power Platform. In other words, we now build agents in Power Platform itself;
    - Power Platform includes AI capabilities that developers can embed into their solutions, allowing users to, for example, chat with the app and have AI-driven insights returned to them using data in the application itself;
    - Investing in and standardizing Microsoft Dataverse (the principal data service underlying Power Platform solutions) as your first port of call for transactional application data. This offers a repeatable pattern where apps are built atop a data service that, in one hop, consolidates data into the lake. This pattern is readily applicable to tier one or critical core business systems as well as to tier two and three (or below) - more productivity centric - applications.

Let's explore an applied example of this.



© Cloud Lighthouse LTD, 2024

The diagram below shows a series of workloads (blue icons) each representing an “app,” or a series of apps, placed in the hands of various users in a typical globally distributed organization. Note that most of these workloads would be classified as “core business systems”, though different organizations may prioritize each of its functions in different ways. These workloads use Microsoft Dataverse as their “single source of truth” for transactional application data.

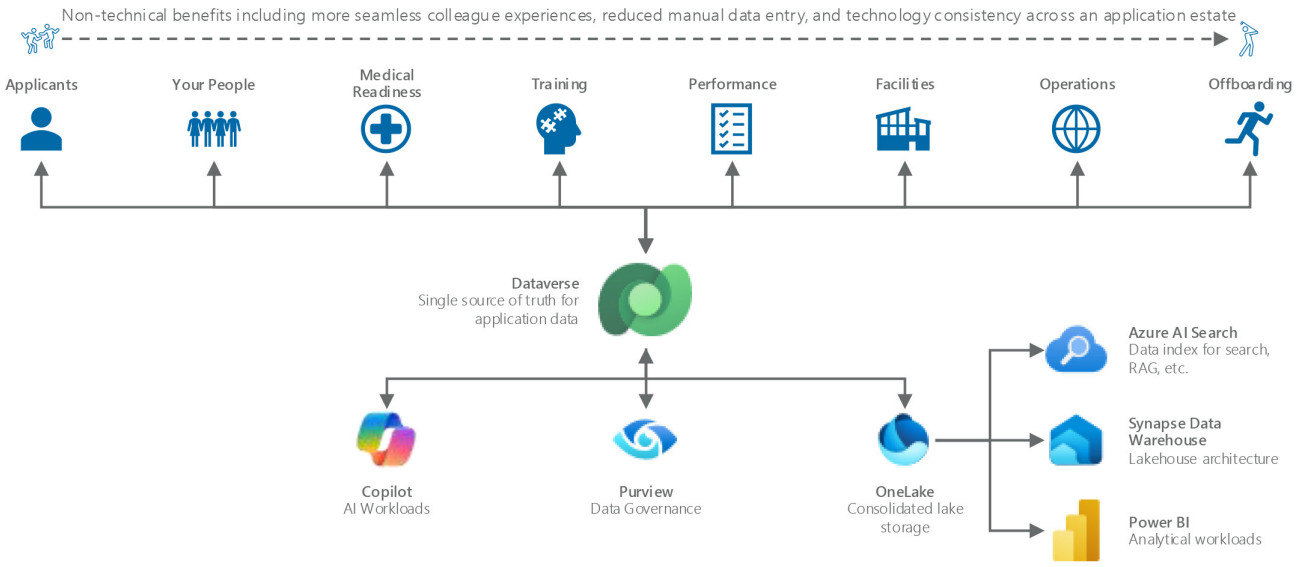


Figure 18: Different workloads (blue icons) that might typically be found in an enterprise organization. The map is illustrative, though it’s highly likely that most organizations will see something of their own needs here.

Try to take this in without getting hung up on the specifics of each workload. This diagram depicts a generic organization that could be found in nearly any industry. It’s really a composite of workloads that I have personally been involved with implementing in sectors as diverse as financial services, professional services such as big law firms, manufacturing, retail and distribution, emergency services (e.g., law enforcement, medical, rescue, coast guard, etc.), the military, and other public sector and non-governmental organizations that deploy or maintain personnel globally.

For example, the “Medical” workload may not be relevant to an insurance firm, but is certainly relevant to a law enforcement agency, the military, or even an industrial firm where safe operation of heavy equipment is a concern. We could have easily substituted that “Medical” workload in favor of an “Underwriter Workbench” workload for an insurance firm.

Now let’s reshuffle those workloads and combine them with the RAG pattern discussed previously.

© Cloud Lighthouse LTD, 2024

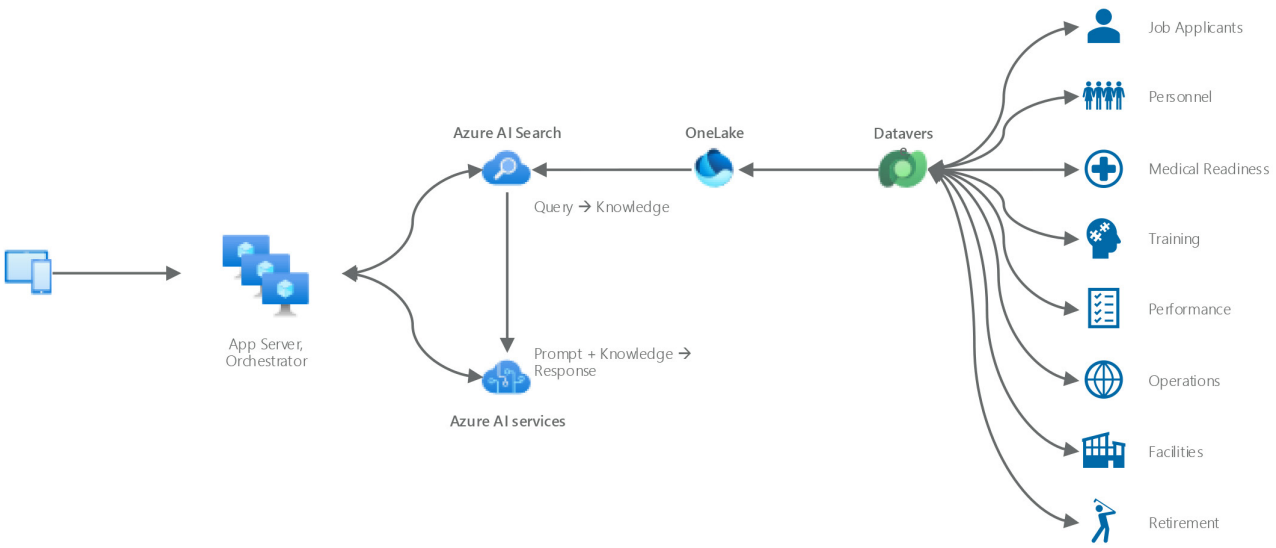


Figure 19: Functional workloads (blue icons) from our previous diagram have been arranged with Dataverse and combined with our AI architecture to demonstrate how this data may flow to AI workloads.

Each of these workloads can be built with Power Platform. So, as you can see, Power Platform operating alongside Azure AI services and Azure data services (Microsoft Fabric) makes for a compelling combination. Modernizing both tier 1/2 and productivity "shadow" apps to Power Platform provides outsized value in scaling AI across the organization, plus many other benefits of this app modernization approach (which are new stories for other times).

AI Development Tools

AI Development Tools account for the specific tooling used by developers to build AI capabilities, and whether the organization has deployed, configured, and made them generally available to their developer population. We can think of these tools as belonging to one of three classes:

- Large language models (LLMs) and other AI models themselves;
- AI tools like GitHub Copilot or Power Apps Copilot that increase developer productivity across many technologies, AI or not, by reducing time spent writing code;
- Developer tools like Azure AI Studio or Copilot Studio that enable or accelerate development of AI workloads specifically.

Let's take each type in turn.

© Cloud Lighthouse LTD, 2024

As of late 2024 most organizations were not training their own large language models (LLMs), but rather meta-prompting existing LLMs through various AI studios, leveraging pre-existing retrieval-augmented generation (RAG) and other patterns, fine-tuning models, and extending the models to which they already had access. This is the same as the operating system paradigm. Every company, or even every sector, does not have a custom operating system that they build. Instead, they rely heavily on custom apps for their industry and their specific business.

Given that existing AI models are already quite diverse, and the diversity of capability is likely to increase as startups, the open-source community and other innovators push the bounds of what is possible. It seems unlikely that this status quo will change significantly in the near term. Technology firms and open-source communities will continue to build and mature their LLMs and other models, but investments from Microsoft and its competitors portend a world in which typical firms across "non- tech" sectors are able to leverage pre-existing models in increasingly sophisticated, and easily accessible, ways. In other words, in almost all cases it is both unnecessary and unlikely that a firm in a non-technology sector will achieve results - at least not with any reasonable level of investment—that surpass those of the specialist firms in the near future.

We've worked with some organizations eager to employ multiple LLMs or to train their own but were struck by how immature all of these organizations were across the 25 AI Strategy Framework dimensions. We'll sum it up by saying, "Better to learn how to play the piano before you attempt to compose a symphony".

Just as most organizations are not training their own LLMs, so should nearly every organization assume that they'll leverage the LLMs provided by their preferred technology vendor - be it Microsoft, Google, AWS, etc. - and avoid oversubscribing to too many until they've reached a sophisticated point of maturity across their AI portfolio (a consistent score of "4" or "Proactive" in the AI Maturity Model that we'll discuss later).

We'll not spend much time on the second class of AI development tools, those like GitHub Copilot or Power Apps Copilot - those that increase developer productivity across many technologies, AI or not, by reducing time spent developing software. These tools can be as valuable to a software developer as tools like ChatGPT or Copilot for Microsoft 365 are to colleagues in other disciplines, but a certain sense of stating the obvious abounds: Organizations with colleagues developing technologies using tools that offer an AI-enabled assistant should go ahead and turn the assistant on. There are few more straightforward ways of wringing value from artificial intelligence than to put purpose-built AI into the hands of technically sophisticated users.

Which brings us to our third class of AI development tools, those like Azure AI Studio or Copilot Studio that enable or accelerate development of AI workloads specifically. Given the current rapid advancement in this area, it can be challenging to discern which tools are worthy of investment both in budget and in the time required to deploy them. We'll consider two candidates below.



© Cloud Lighthouse LTD, 2024

Copilot Studio is an AI development tool used to create agents (formerly "custom Copilots"), as well as agents that extend existing Copilots.



"Agents" are an emerging concept in AI best thought of as workloads falling along a spectrum, ranging from AI performing simple reasoning over data and user prompts to produce responses, all the way to agents capable of interacting with one another, using their "memory" of contextualized learnings and even requesting for help when it's needed.

"Low-code" to a large extent, Copilot Studio, offers tools that allow developers to efficiently design, train, and deploy agents to reason over data housed in a range of data storage technologies.

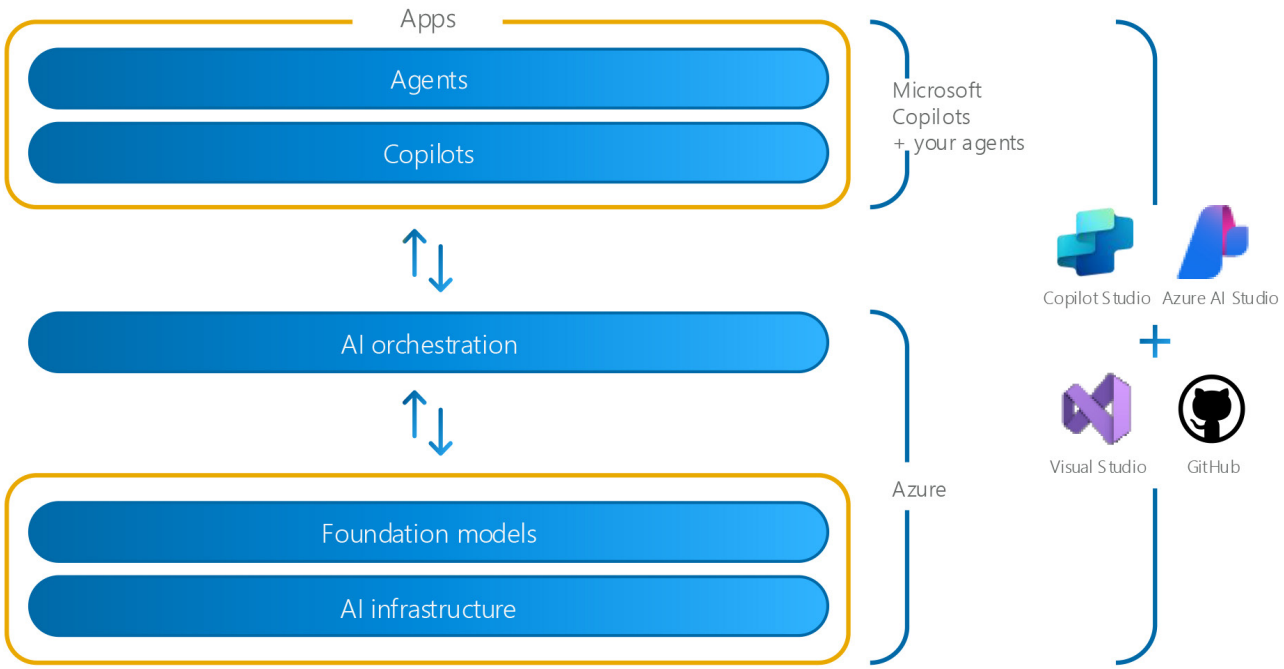


Figure 20: Simple architecture depicting how Azure provides AI infrastructure, foundation models, and AI orchestration both to Microsoft Copilots and agents created by developers using a combination of Copilot Studio, Azure AI Studio, and development tools you're likely already familiar with (Visual Studio and GitHub).

© Cloud Lighthouse LTD, 2024

Copilot Studio is amongst the most accessible and easiest to implement AI development tools so is a good place to begin. Though it is absolutely accessible to less technical “citizen developers” in personal or small team productivity scenarios, experienced engineers will be able to develop more enterprise grade workloads thanks to familiarity with programming languages such as Python or R, ability to preprocess and analyze data, and knowledge of machine learning.

Azure AI Studio, by contrast, is firmly the domain of experienced engineers. If Copilot Studio is the first port-of-call for the development of extensible AI workloads (see the later dimension Extensible AI), then Azure AI Studio is best employed for more advanced extensible AI as well as Differential AI workloads that exceed the former’s capabilities. It provides a robust environment for developing, training, and deploying AI models, making it a popular choice among developers. Azure AI Studio offers additional benefits including scalability for larger or more complex AI development projects and integration with technical services across all of Azure.

Developers here will benefit from their proficiency in other cloud technologies where an understanding of cloud infrastructure and the data platform is essential for effectively leveraging Azure AI Studio - proficiency in languages such at Python, experience with machine learning libraries, and data science skills such as preprocessing, model training and evaluation. What’s mandatory is proficiency in building end-to-end apps since AI apps are still...apps.

The “better to learn how to play the piano before you attempt to compose a symphony” wisdom rings true with these AI development tools, as well. Though there are others available and, in some cases, required to get the job done, organizations ought not invest significantly in pre-deploying them. For those choosing to rely on Microsoft technology, the duo of Copilot Studio and Azure AI Studio is a solid combination, at least in the early days.

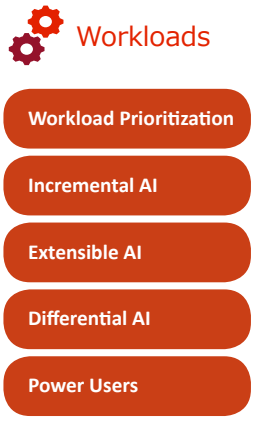
© Cloud Lighthouse LTD, 2024

CloudLight.house
Strategic architecture for the Microsoft Cloud

Pillar Three
Workloads



Workloads



Strategy and Vision created the aspirational, architectural, and programmatic framework for our AI strategy; Ecosystem Architecture built the infrastructure, data, and application foundations. No doubt you have noticed that we've not yet discussed any specific AI-driven workloads. Take this as a lesson concerning the foundational nature of data and your data platform to any big dreams you have about artificial intelligence, for the success of any AI workload is absolutely and critically dependent on your success with strategy, vision, and – particularly - ecosystem architecture discussed above.

Our Workloads pillar gets to what's on the mind of most folks when they think about artificial intelligence: How will we use AI to solve real-world challenges?

"Workload" is not a throwaway word. It is rather a specific term that we use precisely because I value its imprecision. App-centric people speak in terms of apps, integration-centric people speak in terms of integrations, etc. Workloads cover it all. They are, simply put, a collection of one or more apps, chatbots, visualizations, integrations, data models, etc. working towards the same end. "Workload" is essentially the combination of the front-end and back-end required to produce an AI-driven response or action.

This pillar broadly addresses three topics:

- Identifying and road mapping the best candidate workloads for development via **Workload Prioritization**;
- Understanding the spectrum of different workloads through which AI can be used, and why balancing your portfolio across **Incremental AI**, **Extensible AI**, and **Differential AI** is so important;
- Enabling the organization's **Power Users** - also called "Citizen Developers" or "Communities of Practice" - to extend and even develop AI capabilities for themselves.

Workload Prioritization

We (the authors) often say that "We're not concerned with one app. We're interested in one thousand workloads." Maximizing the use of AI throughout an organization, truly weaving it into the culture and ways of doing business. That's how we achieve real value - how we maximize return on investment - in artificial intelligence.

© Cloud Lighthouse LTD, 2024

Our goal in Workload Prioritization (also known as “workload road mapping” or “app rationalization”, depending on which circle you’re running in) is to create a prioritized roadmap of specific workloads to be modernized with an infusion of AI or built anew to solve an emerging problem or a challenge whose solution may have been out of reach without AI. This prioritization is an indispensable part of an organization's ongoing AI journey. Prioritization results in a workload roadmap, a backlog of workloads that are candidates for development with AI capabilities. It allows the organization to project AI’s business value over time, and is a core driver of return on investment (ROI).

There are many techniques and patterns through which you might prioritize and re-prioritize workloads, including:

- **Alignment** of the workload to the desired guiding principles or outcomes defined in your executive vision;
- **Legacy location** in that where the workload lives today, e.g. evacuating the data center on the third floor of your headquarters building may present an excellent opportunity to rebuild previously on-premise workloads to be “AI native”;
- **Legacy technologies**, similar to “legacy location” above, but related to technologies you wish to sunset rather than locations you wish to evacuate;
- **Telemetry** such as monthly or daily active users, last active use, data volume (remember to consider both structured and unstructured);
- **Security or compliance risks** that exist in whatever systems or processes you are currently employing relative to a given workload (i.e. is our current model too risky for the organization?);
- **Target technologies** in that you may highly prioritize workloads that are targeted for the same or similar technologies, i.e. if you’ve just made a significant capacity investment in a particular AI development tool on which you’d like to maximize the return;
- **Qualitative Assessment** wherein we compare each of a series of workloads’ relative impact or value to the organization with the complexity (technical, organizational, political, budgetary) of implementation to determine which workloads are big wins, quick wins, nice to have, or wastes of time.

There are likely more considerations, but these are our favorites. “Qualitative Assessment”, in particular, offers a relatively quick and effective way to elicit real business challenges from colleagues and to then prioritize them so that we can make investment decisions as to which AI workloads we will invest in next.

© Cloud Lighthouse LTD, 2024

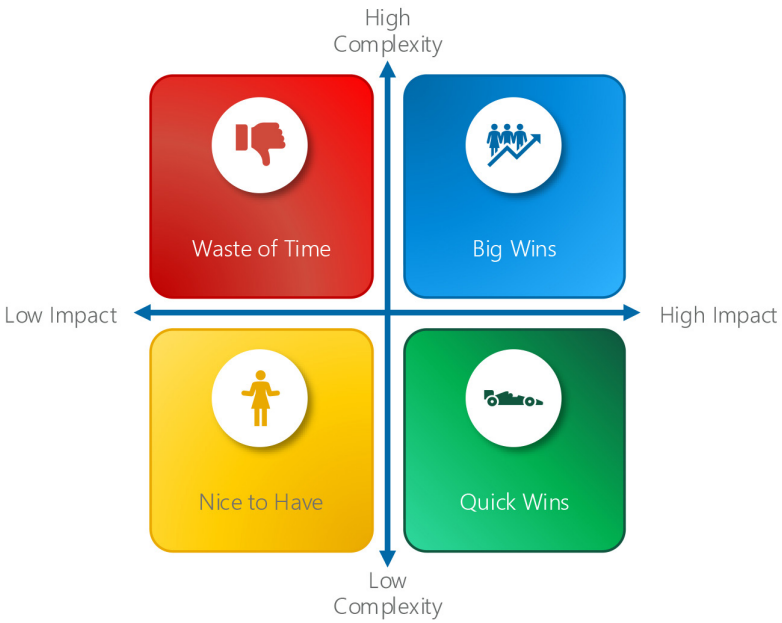


Figure 21: The qualitative method for workload prioritization asks stakeholders to make qualitative judgements as to the relative impact or business value and complexity of implementation for each workload.

The roadmap is pivotal to ensuring that there is a flow of workloads inbound to the AI and data technologies in which we've invested, and that the organization is focusing its development resources on the workloads that will provide the highest value once deployed. In short, a healthy roadmap allows the organization to:

- Continuously decrease marginal costs per workload whilst increasing overall return on investment;
- Provide a trajectory of anticipated usage so that capacity can be tapered up over time;
- Achieve quick wins and big wins to justify early and long-term investment;
- Focus development resources on the most valuable workloads;
- Avoid costly development quagmires by de-prioritizing the least valuable or riskiest workloads.

We must therefore work directly with business stakeholders, users, and IT to identify, prioritize, and categorize candidate workloads and to create a roadmap for building them. These candidate workloads should be prioritized and re-prioritized on an ongoing basis so that we remain focused on the most important next efforts.

Finally, it is useful to categorize each AI workload on your roadmap as incremental, extensible, or differential.

© Cloud Lighthouse LTD, 2024

There is some risk here that specific AI workloads won't work as their creators intended. No, we don't mean that the AI will turn evil and obliterate its makers, rather that because we are (still) in the early days and because the technology's true capabilities are not entirely known, organizations may find that a particular AI-driven workload just doesn't produce the hoped-for results either because it lacks the data it needs or because the technology just isn't ready for what it's being asked to do. Wise organizations can therefore hedge this risk in several ways:

- Balance your roadmap of AI workloads in development between incremental, extensible, and differential. Think of this as a spectrum where incremental incurs the least risk to success and differential incurs higher risk;
- Relatedly, balance your roadmap of AI workloads in development between those that are medium-to-high impact but low complexity, as well as those that are high impact and high complexity. In other words, take a few moonshots, but balance the risk of your moonshots not panning out by also investing in less complex yet still valuable workloads. Distribute risk across your workload roadmap;
- Make measured, incremental investment, and evaluate progress regularly. This is not the space for lengthy IT projects with big bang go-lives at the end. Instead, develop these workloads such that each sprint makes observable progress, preferably solving specific development challenges such that you either (a) build confidence in the solution as you move towards a minimum viable product, or (b) can quickly recognize when it's time to pull the plug and invest elsewhere;
- Augment your AI development with expertise, particularly around specific types of workloads where others have made good progress. Partner with others outside of your own firm, learn from the successes and failures of others, and don't be afraid to learn and share;
- Educate stakeholders and investors; set their expectations. AI can seem magical, but it is not magic. "Quick wins" are not particularly attainable if you've not done the hard work upfront around your data platform, and it's still difficult for almost everyone to prognosticate about outcomes in such an unknown space.

Finally, many organizations will want to conduct a business value assessment (BVA) to make a business case based on specific goals and expected outcomes of their Power Platform adoption generally, and their development efforts for important and critical workloads specifically. Such assessments allow mature organizations to determine and make sound financial choices around considerations such as return on investment (ROI), net present value (NPV), and other industry-specific factors.



© Cloud Lighthouse LTD, 2024

Incremental AI

Incremental AI is a broad, conceptual category of **workloads in which AI is applied to bring speed, efficiency, scale, accuracy, quality, etc. to activities that a human would have otherwise performed.**

Microsoft's Copilots generally fall squarely into this bucket as they help their end user to reason over information more granularly, identify highest-potential sales targets more accurately, create content more quickly, book appointments more efficiently, write code more effectively, recap meeting action items easily, etc. These use cases tend to share two things in common:

- They apply artificial intelligence to scenarios that were already being performed by human, and would have gone on being performed by a human with or without AI;
- They are often* performed in support of the organization's overall purpose, not as the organization's primary function.

* Big asterisk here, so let us explain what we mean by comparing the "book meetings more efficiently" and "write code more effectively" examples above.

In the case of the former, organizations do not typically exist for the purpose of having meetings (though there are days that this would feel like revelatory news to many of us), so booking an appointment is a necessary task which creates the medium through which a service is provided; but is not the service itself. This is a typical scenario for so-called incremental AI, that AI acts as a "Copilot" helping the human cut through various tasks to get to the real point of their work. At the risk of being called out here, we will acknowledge that there are some cases - such as AI writing code for a developer at a software company - where AI is actually creating the products or rendering the services that are the ends, rather than a means to the ends. It's a bit of a semantic discussion, but worth keeping mind as you organize your thoughts.

It's important to remember that "Copilot" is less a standalone, specific product and a technology that is woven throughout many Microsoft products. Actually, there were 108 or more Copilots available from Microsoft at the time of this writing,

the most well-known perhaps being Microsoft 365 Copilot integrated with Microsoft's 365 suite. Other examples include Microsoft Copilot, GitHub Copilot, Power Apps Copilot, Copilot for Power BI, and (do the math) 104 others.

More info on these many Copilots can be found on Copilot Learning Hub.

© Cloud Lighthouse LTD, 2024

In any case, the central question to ask yourself, your colleagues, and those formulating your AI strategy is, “Which activities currently being performed by humans could we make better in ABC ways by turning some or all of the task over to AI?”

These answers will often come easily, but it’s worthwhile to cast a wide net to ensure you’re maximizing the potential of AI in your incremental workloads. Consider…

Turn on AI you already own (or can easily acquire)
Implementing Microsoft’s extensive (and growing) range of “Copilot” products that are likely to be most impactful to you.

Ask your people
Directly ask end-users and line of business owners to identify pain points in their work and their wish list for AI assistance.

Rationalize your workloads
Consider AI capabilities as part of any app rationalization or workload prioritization exercise you conduct.

Mine your processes
Use tools like ”Process Advisor” in Power Automate to ID pain points in business processes that AI could mitigate.

Engage with your peers (and competitors)
Keep your ear to the ground with competitors, “friendlies,” and industry groups to understand how they’re using AI.

This is ultimately a workload prioritization and road mapping activity (reference the earlier Workload Prioritization dimension), so we recommend having a look at Andrew Welch’s “ One Thousand Workloads” piece from several years ago (which addressed road mapping for Power Platform but is also quite applicable here), as well as more recent thoughts on “Workload Prioritization” which you’ll find in his “Strategic thinking for the Microsoft Cloud” piece from early 2023. The workloads that make it onto your roadmap should absolutely be included for implementation as part of your AI strategy.

Extensible AI

Though it barely registered as a concept in early 2024, by late 2024 it has become apparent that Extensible AI may be where the bulk of an organization's AI workloads are categorized in the coming years.

Admittedly, there is a lot of grey area here, and in any case, the spectrum of incremental, extensible, and differential workloads is really meant more as a conceptual framework than any hard technical boundary. That said, extensible AI occupies the broad middle range where more incremental workloads are extended to suit an organization's specific scenarios.

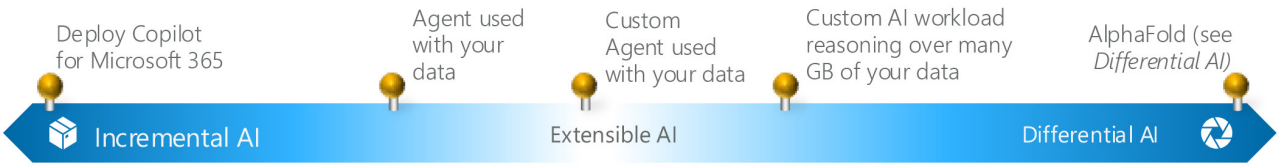


Figure 22: Incremental, Extensible, and Differential AI workloads provide a conceptual framework—a spectrum—for evaluating an organization's portfolio of investment in AI workloads. Health portfolios are well balanced, though different organizations and industries will have varying risk tolerances.

For example, consider the scenario above where an organization has consolidated a store of its proprietary data and desires that AI reason over this data to produce generative responses via a chat-based user interface, including directly through Copilot for Microsoft 365. This would have required engineering a bespoke RAG-based workload even as recently as early 2024, complexity that would have landed this scenario squarely in the realm of Differential AI.

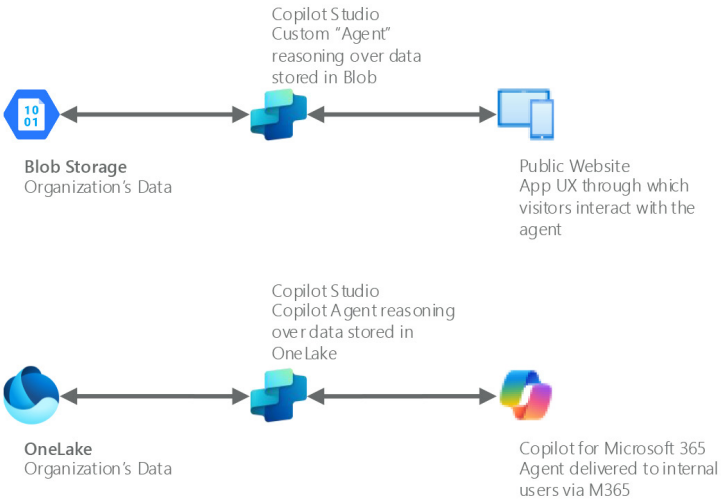


Figure 23: Simplified models of typical Extensible AI scenarios.

Developers can extend many Copilots using their organization's data to further enhance the functionality of these Copilots, making them even more relevant and effective for specific business needs. Copilot Studio is increasingly purpose-built for just these types of scenarios, so if you've not read more of the AI Developer Tools dimension, this is a good segue to that earlier topic. Let's now discuss several other key extensibility tools:

- **Graph connectors** allow data from external systems to be indexed and made searchable within the Microsoft ecosystem. By integrating these connectors, developers can extend the capabilities of Microsoft Copilots to access and utilize data from a variety of sources. Graph Connectors use APIs to crawl data from external sources such as file systems, databases, and SaaS applications. This data is then indexed and made available within Microsoft Search experiences. Consider a company that uses both SharePoint and an external CRM system. By creating a Graph Connector for the CRM, employees can search for customer data directly from their Microsoft 365 environment, enabling seamless access to critical information;
- **Teams message** extensions provide a way to extend the functionality of Microsoft Teams by allowing users to interact with your services and data directly within Teams messages. They can be categorized into Search Message Extensions and Action Message Extensions. These extensions allow users to search for information from external systems and insert the results into a Teams conversation:
    - A Search Message Extension could be developed to allow employees to search for knowledge base articles and insert relevant excerpts into a Teams chat, facilitating quick access to helpful information during discussions;
    - Action Message Extensions enable users to initiate workflows or perform actions based on message content, for example, allowing users to create a new task in a project management tool directly from a Teams message, streamlining task management and reducing context switching;
- **API plugins** (in preview as of fall 2024) enable developers to integrate third-party APIs with Microsoft Copilots, allowing these intelligent assistants to interact with external services and retrieve data in real-time. API plugins involve writing custom code that calls external APIs and processes the returned data. This data can then be used by the Copilot to provide informed responses or perform specific actions. A Copilot could be extended with an API plugin that retrieves weather data from an external service. Users could ask the Copilot for the current weather conditions or a forecast, and the Copilot would provide accurate, up-to-date information;

© Cloud Lighthouse LTD, 2024

- **Copilot Studio Agents** allow for the definition of Copilot behaviors through configuration rather than code. This approach can simplify the process of extending Copilots and make it accessible to non-developers. Copilot Studio Agents use configuration files or low-code tools to define how they should interact with users and data sources. These configurations can specify triggers, responses, and integrations with other services. A Copilot Studio Agent could be configured to monitor a SharePoint list for new entries and send a notification to a Teams channel whenever a new item is added. This setup requires minimal coding and can be quickly adapted to changing business needs;
- The **Teams AI Library** provides tools and frameworks for building intelligent bots and Copilots within Microsoft Teams. By leveraging these resources, developers can create sophisticated, AI-powered assistants that enhance collaboration and productivity. The Teams AI library includes pre-built models for natural language understanding, tools for building conversational interfaces, and integration capabilities with Microsoft Graph and other services. Using the Teams AI library, for example, a developer could create a Copilot that helps employees schedule meetings. The Copilot could understand natural language requests, check participants' availability, and suggest suitable meeting times, all within a Teams conversation.

The bottom line of extensible AI, though, is that architects need to quickly transition from making black and white assessments of whether a pre-built AI tool such as Copilot for Microsoft 365 can do the job, or if an entirely bespoke AI development is needed to increasingly embrace extensibility scenarios that fall in the broad middle of the spectrum. In other words, seek first to extend, then consider entirely custom development.

Differential AI

We're finally on to what popular culture might consider the "fun" bit, or at least the bit that dominates the imagination when it comes to artificial intelligence's supposedly boundless possibilities. Whereas Incremental AI includes the scenarios that improve upon solo human performance of activities that would have been performed anyway, Differential AI broadly encapsulates **workloads that would not have likely been performed by humans alone, scenarios that are valuable to the organization because they allow you to jump out ahead of your competition,** to offer your customers something that you'd not have otherwise been able to provide. We once toyed with the idea of calling this "Secret Sauce AI" or "Moonshot AI" to underscore the point.

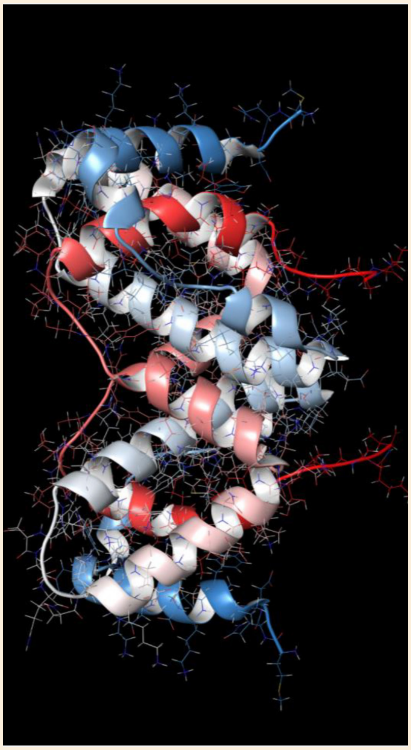
Hallmarks of these differentiating, accelerative workloads are that they require a degree of creative thinking to dream up, can be challenging to implement, often involve deriving insights by mixing data that you already own but never had the ability to co-mingle, and operate along a time dimension, that is to say, involve some sort of computation or connection that must be completed within a window of time that makes human intervention more challenging.



© Cloud Lighthouse LTD, 2024

On top of this, these workloads will require a degree of flexibility on your part, at least in the early days as you figure out exactly how to harness the power of this newfangled thing you've built.

AlphaFold, developed by Alphabet's DeepMind lab to make predictions about protein structure, is one of the world's more successful examples of this type of differential AI. We will not - nor are we qualified to - dig too deeply into the biological particulars of what the thing does, but we will draw on some incremental AI to summarize it for us below.



"AlphaFold AI is a groundbreaking deep learning system designed by DeepMind that brings tremendous advancements to the field of protein structure prediction. By harnessing the power of neural networks and sophisticated algorithms, AlphaFold AI analyzes amino acid sequences and accurately predicts the three-dimensional structure of proteins, revolutionizing our understanding of their functions and enabling significant breakthroughs in drug discovery, disease research, and biotechnology. Its exceptional precision and speed promise to reshape the way scientists study protein structure, opening up new possibilities for tackling complex biological challenges and accelerating scientific progress."

— A slick bit of incremental AI (ChatGPT)

In theory, many humans might have spent many years working this out in the absence of artificial intelligence. But this would have been impractical, if even achievable, and in this way AlphaFold serves as a (ridiculously brilliant) example of differential AI offering a benefit to its creator far beyond what would have otherwise been possible.

That said, differential AI need not offer groundbreaking promise for the future of humanity to

be useful to one organization, team, or even to an individual's world of work.



Figure 24: The practical examples discussed here each, in their own way, assimilate and reason over a combination of proprietary customer information, internal documentation, unstructured data, publicly available data, consolidated data stores, and observable characteristics of the physical world to produce a response or outcome.

Let's consider some practical examples below.

Professional Services

For example, a law firm, accountancy, or consultancy might use AI to produce regular guidance advising its clients of upcoming regulatory or legal changes that may impact their business in the various jurisdictions in which they operate. A typical firm would know who its clients are, in which countries and sub-national regions (e.g. states, provinces, territories, counties, autonomies, etc.) its clients operate, and possess data concerning its clients' products or services. This proprietary information could be used in conjunction with (a) publicly available information and (b) internal documentation - indexed by Azure AI Search, of course - concerning regulatory or legal changes taking effect over (say) the next 6-12 months to produce tailored guidance for individual clients.

Public Sector

A public sector law enforcement or foreign affairs organization might possess information in one quasi-consolidated data store that Firm A is owned by Firm B, which is itself a shell company owned by (say) a sanctioned Russian official. Another quasi-consolidated data store, or some selection of unstructured data may also indicate that Firm B owns a yacht that the global automatic identification system (AIS) knows is presently docked in an allied port. This serves as a great example of, among other things, workloads with a time dimension. A human could connect these dots, but it is less likely that a human could connect all these dots in a potentially narrow window of time during which that particular yacht was docked in that particular port.

Agriculture

An agricultural firm depends on phenotyping (observing and analyzing an organism's observable traits or characteristics to gain insights into its genetic, physiological, or environmental attributes) to breed and produce new vegetable varieties. This extremely laborious process can be significantly enhanced by image analysis assessing thousands of images per minute.

…and on it goes.

There are clearly overlaps between incremental, extensible, and differential AI workloads. Think of these as conceptual categories meant to help you place them in context of your AI strategy (rather than hard technical boundaries). The three examples above share, however, that they move their organization's use of AI from that which betters or enhances something that would have been done anyway to the realm of achieving results that could not have (realistically) been otherwise produced.

As AI technology evolves, though, it seems certain that the line between these different flavors of AI workloads will blur further. We will move from "bolt-on" architectures and user experiences wherein AI is grafted onto pre-existing workloads or used in legacy form factors such as "the app" to "born-in-AI" solutions that have been architected from the ground up with AI in mind.

There is no perfect historical comparison here, but let's illustrate the point by thinking back to the early days of smart watches. When this new form factor first became available, many of its early applications simply shrank that which was available in the mobile phone form factor onto a tiny, wearable screen. These apps were predictably terrible until developers began thinking of the wrist as an entirely new form factor with different constraints and opportunities.

AI represents a much more fundamental paradigm shift than a screen that's similar to other screens but smaller; running apps that are similar to other apps but on your wrist. Suffice it to say, though, expect the discipline of born-in-AI workload design to mature in the years ahead until the industry at large accumulates enough experience to really understand what's possible, what works, and – importantly - what doesn't.

One can get a sense of where this may be going in the form of Project Sophia, which Microsoft describes as aiming "to help our customers solve complex, cross-domain business problems with AI, by enabling them to interact with data in new ways and answer strategic questions that drive better outcomes." There's no way to know if this is a form factor that will stick, but Sophia's very existence points the way towards an era of creativity where even the future of the venerable "app" is up for grabs.



© Cloud Lighthouse LTD, 2024

In any eventuality, it seems likely that those of us whose idea of what computing can be was shaped by our years growing up watching the crew of Star Trek verbally converse with “the computer” and be handed back responses that we’d now call “generative” will not have to wait long to see these dreams become reality.

What differential AI looks like in any specific organization is, well, specific to each organization. This of course risks spiraling into its own form of the use case death spiral wherein fixation on an endless stack of workloads prevents action on workloads one, two, and three. Differential AI also takes us into the realm of trying to solve today’s problems with tomorrow’s promises, reference one of our original, foundational principles:

“Any future-ready AI strategy must be flexible, meaning it is able to absorb tomorrow what we don’t fully grasp today.’’

Your differential AI workloads ought to be baked into your AI strategy’s roadmap from the start, shepherded by technology and business leaders with a healthy tolerance for flexibility.

Power Users

Power Users are pivotal in the successful adoption and scaling of AI technologies within organizations. They are typically advanced users who leverage AI tools beyond their standard functionality, often discovering new use cases and pushing the boundaries of what is possible. By acting as early adopters, Power Users not only validate the utility of AI tools but also drive their evolution by providing critical feedback. They represent a bridge between the general user base and technical teams, translating complex AI functionalities into practical, everyday applications.

We’ve included this dimension as a part of the Workloads pillar because these power users should be enabled to (a) evangelize and promote the use of AI workloads amongst their colleagues and teams, and (b) develop AI workloads - particularly Extensible AI workloads - themselves using tools like Copilot Studio.

To further bridge that gap, it is important we introduce yet another concept: Communities of Practice (CoPs).

Communities of Practice (CoPs) are vital in embedding AI into the organizational culture. These communities bring together individuals across various roles and departments to share knowledge, best practices, and insights about AI tools like Microsoft Copilot. CoPs facilitate continuous learning, enabling members to stay current with AI advancements and to collaboratively solve challenges related to AI adoption. By fostering a culture of open dialogue and experimentation, CoPs help organizations navigate the complexities of AI, making its deployment more strategic and effective.

© Cloud Lighthouse LTD, 2024

Organizations must provide access to the right tools, training and resources that enable Power Users and the CoPs they lead. In other words, investing in them and setting them up for success by establishing a robust support framework that covers everything from understanding data security within AI contexts, getting skilled up with a selection of copilots and, or course learning the art of prompt engineering.

It sounds like a lot, and that's because it is, so to maximize the impact, and avoid potential chaos, organizations should:

- Establish clear roles: Define roles and responsibilities for specific Power Users and CoPs to clearly define their focus within AI initiatives. Examples include testing new features, providing feedback or mentoring other users.
- Facilitate engagement: It's important to encourage active participation in CoPs through regular meetups, workshops and collaborative platforms that support knowledge sharing and problem solving.
- Support continuous learning: Offer ongoing training, access to AI tools and resources and give them the opportunity to apply their skills to projects or proofs of concept, keeping Power Users at the forefront of AI innovation.
- Be strategic about the impact of Power Users and CoPs: Harness their capabilities and use them to ease user adoption and reduce resistance to change, ensuring AI adoption is aligned with business objectives (record the benefits of using AI in said objectives and track their progress).

© Cloud Lighthouse LTD, 2024

CloudLight.house
Strategic architecture for the Microsoft Cloud

Pillar Four
Responsible AI



Responsible AI



Science fiction abounds with tales of the computer surpassing and, eventually, dethroning the human. In the real world, though, the need to regulate and moderate artificial intelligence is enacted through the discipline of Responsible AI (RAI).

Microsoft has established a series of RAI principles that guide the ethical development and deployment of AI. These principles - Reliability and Safety, Privacy and Security, Fairness and Inclusivity, Transparency, and Accountability - are essential to ensure that AI is used safely within an organization. These principles join the AI Strategy Framework as dimensions in our Responsible AI pillar.

We need to be unequivocal, here, lest organizations foolishly treat RAI as the first thing to be cut when budgets tighten:

Responsible AI is not optional. Omitting it from your AI strategy is, in fact, irresponsible, and exposes the organization to intolerable levels of risk. You must either take RAI seriously or walk away from AI altogether.

This is not to say that RAI is more important than the other four pillars, rather to say that organizations failing to take - for example - workload prioritization seriously are likely to waste time and money. Organizations that don't take RAI seriously face the possibility of being sued and regulated out of existence.

Microsoft is researching and analyzing various RAI scenarios with the goal of defining risks and then measuring them using annotation techniques in simulated real-world interactions. It's product and policy teams, and the company's AI Red Team area group of technologists and other experts who poke and prod AI systems to see where things might go wrong.

Reliability and Safety

AI workloads and their underlying infrastructure, models, and use of data must be reliable and safe in any scenario into which they are deployed. This principle emphasizes the importance of building AI systems that are dependable and secure, capable of functioning correctly under diverse and unforeseen circumstances. It also involves rigorous testing and continuous monitoring to prevent failures and mitigate risks.

For example, AI is increasingly used in healthcare for diagnostic purposes. To ensure reliability and safety, a hospital might implement an AI-based diagnostic tool and conduct extensive testing in controlled environments before full deployment. Continuous monitoring and updates ensure that the tool performs accurately and safely in real-world medical scenarios.

© Cloud Lighthouse LTD, 2024

Software was nearly entirely deterministic prior to the advent of generative AI, which is to say that its programming provided for a specific number of defined outcomes from any input it was provided. When new lead is created, check if contact exists. If contact does not exist, create contact.

Deterministic programs can be tested for every possible outcome, because the outcomes can be quantified and defined.

Modern AI is largely non-deterministic, meaning that the program chooses its own path, its own adventure if you will, each time that it is run. Responses, even to an identical prompt, vary each time the prompt is given and the response is generated.

Let's illustrate this non-deterministic phenomenon with an innocent example.

In writing this chapter, we provided Microsoft Copilot with the following simple prompt:

Please paint me a picture of a lighthouse.

Copilot returned a response several seconds later, generated by Microsoft Designer using DALL-E 3.

We then repeated the same prompt, only to have Copilot return a different generative image.



Figure 25: The first painting of a lighthouse that Microsoft Copilot returned.



Figure 26: Copilot then returned a different image when prompted again just a minute later.

© Cloud Lighthouse LTD, 2024

All of which is to say that reliability and safety - and really all five RAI dimensions - are not things that can be deterministically tested for in advance and then left to run on their own. RAI requires ongoing monitoring, correction and tuning, and testing again to produce responses that are ever more aligned to RAI principles. It also requires risk tolerance to the reality that AI will make it mistakes, it will produce irresponsible responses. All the more reason for the organization's collective digital literacy to be attended to, so that humans are able to recognize these errors and take part in continually refining the AI workloads with which they interact.

Privacy and Security

AI systems must be designed to protect individual privacy and ensure data security. This dimension focuses on safeguarding personal data against unauthorized access and misuse. It involves implementing robust security measures and ensuring transparency about data collection, usage, and storage practices.

Consider a smart home device that uses AI to learn and adapt to the user's preferences. To adhere to privacy and security principles, the manufacturer must encrypt all data transmissions, provide clear information on data usage, and give users control over their data. Regular security updates and vulnerability assessments also help protect user privacy.

Organizations that have robustly implemented Microsoft Azure across their cloud ecosystem have an in-built advantage here.

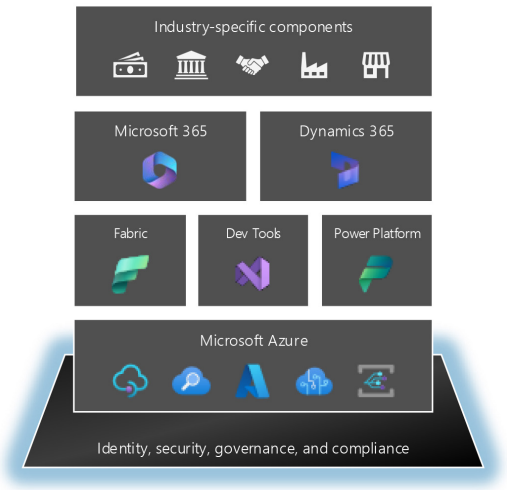


Figure 27: Microsoft's major platforms or product families build upon one another. Fabric, Dev Tools, and Power Platform sit atop Azure services, and in turn Microsoft 365, Dynamics 365, and the industry-specific components are end-user applications dependent on the platform services sitting beneath them.

You've likely seen a version of this building blocks diagram if you've been paying attention to Microsoft marketing over the last several years, but allow us to direct your attention to the foundational "identity, security, governance, and compliance" layer. Data security is emerging as one of the strongest cases to be made in favor of adopting a Microsoft-centric ecosystem rather than piecing together "best of breed" ecosystems from amongst disparate software vendors. Just as Apple has been able to create highly integrative user experiences by controlling its consumer technology end-to-end, so too do we believe Microsoft will increasingly create highly integrative data security experiences by (more or less) controlling its data platform technology end-to-end.

© Cloud Lighthouse LTD, 2024

Privacy and security has been a significant area where existing infrastructure has raced to keep up with the evolution of generative AI technology in recent years. No technology company has a silver bullet, but in integrating both the security infrastructure and the data platform sides of the coin, Microsoft has made significant strides through its investments in technologies including Entra ID (formerly Azure Active Directory), Purview for data governance, the security models built into technologies like Dataverse and OneLake, and more. We're far from the promised land of a fully integrated security model across the entire data estate, but we're getting there. This is an incredibly exciting (albeit likely thought of as quite niche) area to watch in the coming years.

Fairness and Inclusivity

Fairness and Inclusivity are two highly related yet subtly different concepts.

The more subjective of the two is fairness, which we'll define here as the importance of treating everyone the same, including bias against disability. Fairness in AI means that systems should treat all people equally and equitably, without bias or discrimination. Notoriously difficult to quantify, ensuring fairness involves identifying and eliminating biases in AI systems that might lead to unfair treatment of individuals based on attributes such as race, gender, age, or other protected characteristics. It is essential for maintaining trust and social harmony.

A typical employee hiring process offers a real-world example of this principle. Imagine a company using an AI-based recruitment tool. To ensure fairness, the company must regularly audit the AI system for biases and make necessary adjustments. For example, if the system favors candidates from certain demographics over others without a valid reason, the algorithms need to be revised to eliminate such biases.

Age, interestingly, has become quite a common manifestation of the fairness problem in AI wherein models are unduly influenced by the extraordinary amount of internet content produced by members of younger generations, and inadvertently favor members of those generations in processes such as hiring.

Inclusivity is relatively more straight forward to quantity and measure, ensuring that AI systems are accessible and beneficial to a diverse range of people, including those with disabilities. This principle underscores the importance of designing AI technologies that are usable by people from all backgrounds and abilities. It promotes equal access to AI's benefits and encourages diverse perspectives in AI development.

For example, consider the development of AI-powered language translation tools. By supporting multiple languages and dialects, these tools enable people from different linguistic backgrounds to communicate more effectively. Adding features such as voice recognition for people with speech impairments further enhances inclusiveness.

Transparency

Transparency involves making AI systems understandable and providing clear information about how they operate. This principle highlights the need for openness about AI decision- making processes, including that users ought to be informed about how AI systems work, the data they utilize, and the algorithms they employ. Transparency fosters trust and accountability.

In the financial sector, AI algorithms are often used for credit scoring. Adhering to transparency principles might cause a bank to provide customers with detailed explanations of how their credit scores are calculated, including the factors considered and their respective weights. This helps customers understand and trust the AI workload and its underlying infrastructure, models, and grounding data.

Accountability

Accountability ensures that organizations and individuals are responsible for the outcomes and responses produced by their AI workloads. This principle emphasizes the need for clear lines of responsibility and mechanisms for addressing issues that arise from AI deployment.

Organizations must be prepared to acknowledge and take corrective action when AI systems cause harm or operate incorrectly.

Further, corrective actions must be timely. In other words, organizations must be resourced such that they are able to resolve harmful, incorrect, or other issues that run counter to the RAI principles which ought to be treated with the urgency of a critical error in a Tier 1 core business system. The programmatic and technical proficiency to diagnose, triage, and act on these resolutions is an absolutely core competency for any organization deploying AI tools.

The need for accountability can be seen, for example, in self-driving cars. If an autonomous vehicle is involved in an accident, there must be a framework to determine responsibility. The car manufacturer would need to have processes in place to investigate the incident, address any faults in the system, and provide appropriate remedies to those affected.



© Cloud Lighthouse LTD, 2024



CloudLight.house
Strategic architecture for the Microsoft Cloud

Pillar Five
Scaling AI

Scaling AI



The Economist writer we cited in this white paper's Background section concluded by asserting that "the most significant benefits from new forms of AI will come when firms entirely reorganize themselves around the new technology," while cautioning that "gathering data is tiresome and running the best models expensive," and pointing out that 40% of American small businesses report being uninterested in AI tools.

In time, though, most organizations will turn their attention from future readiness and establishing themselves with AI to focusing instead on scaling (and sustaining) their investment in AI and the data platform upon which it depends. Put another way, one-time consolidation and readiness of data combined with a few AI-driven workloads does not a future-ready organization make.

First, organizations must tune their technical capabilities to support the scaling of AI. Some of this will be directly relevant to AI itself, for example, employing AI Operations (AIOps - including Machine Learning Operations (MLOps) - to build, deploy, monitor, and maintain production models, incorporating rapid advances in the tech to your enterprise machine learning platform where you'll land your workloads and operate them in your daily business.

Second, there are human Digital Literacy considerations that should be thought about as you progress AI across the organization. There is a significant element of people-centric scaling and change management required here. In other words, the baking of AI into the way people work.

We'll explore these and other dimensions in the pages that follow.

AI Operations (AIOps)

When we speak of AI Operations (AIOps), we're talking about the patterns, best practices, and enabling tools used to develop, tune, test, incrementally improve, and productionize these types of AI workloads and the models themselves in a scalable way, automating where possible to achieve efficiencies and reduce human error. In this way, AIOps is itself a sub-discipline within DevOps, and mirrors many of the patterns (e.g., CI/CD), best practices (e.g., automate where possible), and enabling tools (e.g., Azure DevOps) found in "traditional" DevOps.

Just as AIOps is a sub-discipline of DevOps, so too is Machine Learning Operations (MLOps) a component of AIOps. It's confusing, so let's set the record straight.

It's easy enough to use the terms "Artificial Intelligence" (AI) and "Machine Learning" (ML) interchangeably, so let's clearly define the difference.

© Cloud Lighthouse LTD, 2024

AI is a broad and evolving field wherein the technology mimics, and in some cases surpasses, the cognitive abilities of humans. This broad category encompasses everything from “if this, then that” scenarios where the seeming “intelligence” is the product of pre-determined decision trees and patterns that humans have themselves created, all the way to “generative AI” where the technology is able to generate bespoke answers to questions, images, insights, and other responses based on its index of accumulated knowledge.

ML is a sub-discipline within the broad category of AI, wherein the machine “learns” and refines its own capabilities based on the information and feedback it encounters, and based on the tuning that ML engineers apply over time.

Implementing and maturing AIOps will, for many organizations, involve a simultaneous maturing of their DevOps capabilities to be relevant in the age of AI, as well as maturing their MLOps to support their machine learning models.

Data Governance

We briefly discussed Data Governance as part of the Ecosystem Architecture pillar’s Core Platform Services dimension, particularly insofar as establishing baseline or minimum viable product data governance as an essential part of building and maturing an organization’s cloud landing zone.

Data Governance is so important to a future-ready enterprise AI strategy for several reasons:

- AI strategy elevates the centrality of the modern data platform in an organization’s technology ecosystem, bringing data out of the shadows such that we finally - at long last - replace the “security by obscurity” approach that has loomed in IT for many years, with a deliberate and rigorous approach to data governance;
- Earlier we said that “data is the essential fuel without which AI models cannot be trained nor have the capacity to act,” so, simply, data governance is essential to the care and safeguarding of AI’s most important asset, and to mitigating the risks of AI hallucination (incorrect or misleading responses), and the RAI topics of reliability and safety, privacy and security, inclusivity, transparency, and accountability gone awry;
- Finally, as we have said, the data distribution capabilities that we’ve instituted as part of our AI strategy will also be used in analytical workloads, search, integration with third parties, and more; strong data governance improves the outputs of these classes of workloads, as well.



© Cloud Lighthouse LTD, 2024

Microsoft continues to invest heavily in its Purview capability to provide for data governance, security, quality, lineage, compliance, etc. across the data estate. In accordance with the time-honored principle of "following the money," we recommend that implementation of Purview be an early-stage milestone in nearly every organization's AI strategy, and that your Purview implementation be matured and kept current with the evolution of the organization's data estate (and the product's latest capabilities) over time.

We're also still early days when it comes to anything like unified data security and role-based access controls (RBAC) across a large organization, so we both expect and are hopeful that the next couple of years will see increased convergence around a data security model that is established at the source and hydrated throughout our ecosystem. This is important so that information security teams can be confident that a nugget of data to which a user would never have access in the context of its source application, does not somehow pop up for that user in an AI or analytical scenario downstream. We'll cover these risks from another perspective in the Technical Debt dimension.

Microsoft promises these sorts of robust conditional access capabilities in Fabric, but we suggest a combination of caution and - of course - rigorous information security best practices as we see how this plays out.

Robust data governance should, at a minimum, be instituted for all data residing within the Core Business Systems and Data Distribution Neighborhoods in your cloud ecosystem. We also recommend serious thought be given to establishing data governance of data residing in the Tier 2 or "business important" applications discussed in the Ecosystem Architecture pillar.

Technical Debt

Gartner forecasted a 6.8% rise in global IT spending for 2024. Around the world and across industries, technology leaders struggle to bend their cost curve even as new investments in artificial intelligence and the data platform technologies required to power it become more pressing.

A Forrester Total Economic Impact Study in July 2024, also mentions technical debt as a risk associated with shadow IT and unauthorized software use. Before adopting Power Platform, organizations often face challenges with employees building their own solutions in tools like SharePoint and Excel, which can lead to technical debt. This debt arises because these makeshift solutions can create dependencies on unmaintained tools and processes, increasing risks related to security, compliance, and overall system reliability. The use of unauthorized software also complicates the governance and management of IT resources, further exacerbating technical debt within the organization.

Indeed, technical debt not only incinerates IT budgets and distracts from the hard work that organizations must undertake to modernize for the age of AI, but, more insidiously, AI itself exposes organizations to immense risk due to the technical debt found in their existing application estates.

Think back to the example from the Business Applications dimension…

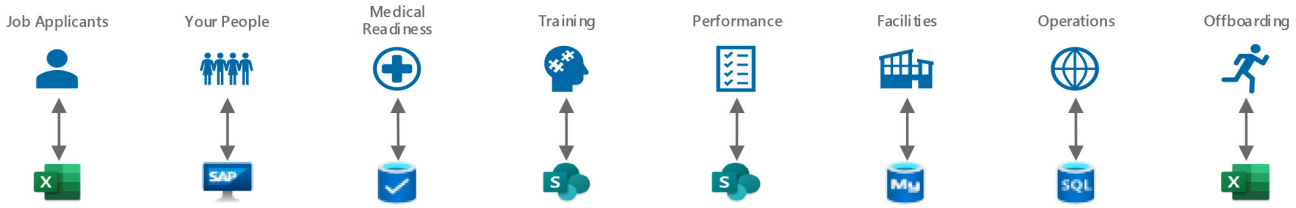


Figure 28: The above diagram shows (icons noted in blue) an assortment of workloads common across many organizations.

These workloads have grown over time as point solutions implemented largely in isolation of one another. Their specific data storage technologies differ between organizations, but we've used a combination of Excel, SAP, proprietary databases, SharePoint lists, MySQL, and SQL Server to provide a representative sample.

Now, in nearly every organization we've worked with, individual applications in their fragmented collection of point solutions require significant amounts of common data. Personnel data offers a great example, because each workload shown above requires some degree of data or knowledge about the people working there. So, IT organizations build "spaghetti web" point-to-point integrations between data stores using a variety of tools including Power Automate, scattershot use of actual integration tools (event, logic, or batch integration), Excel, and even what we used to jokingly call "sneaker net", in other words, manually moving data from one system to the next via physical media.

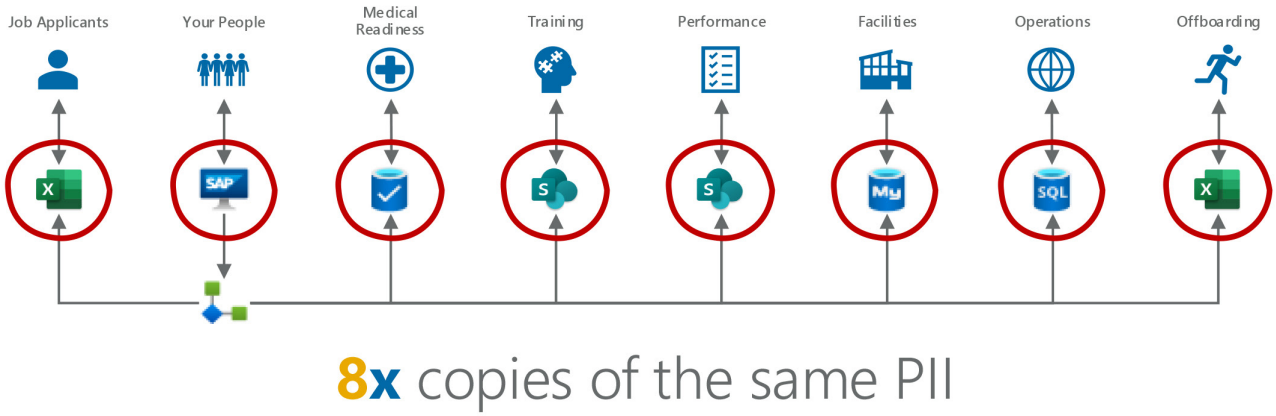


Figure 29: Working around technical debt limitations can often result in copying data multiple times from one location to another, multiplying data security risk and creating significant risk.

This copying of data - scratch that, this making copies of copies of data - often results in a catastrophic proliferation of (among other things) personally identifiable information (PII). Indeed, our scenario above has resulted in 8x copies of the same PII.

The phenomenon is even more insidious for organizations with large portfolios of “SharePoint apps” or Power Apps built atop SharePoint lists as their data source, overengineered workarounds to avoid the cost of properly licensing Power Platform. The diagram below replaces all of the data sources in the previous architecture with SharePoint, which may or may not be as ubiquitous in your organization but, let us tell you, it is in many.

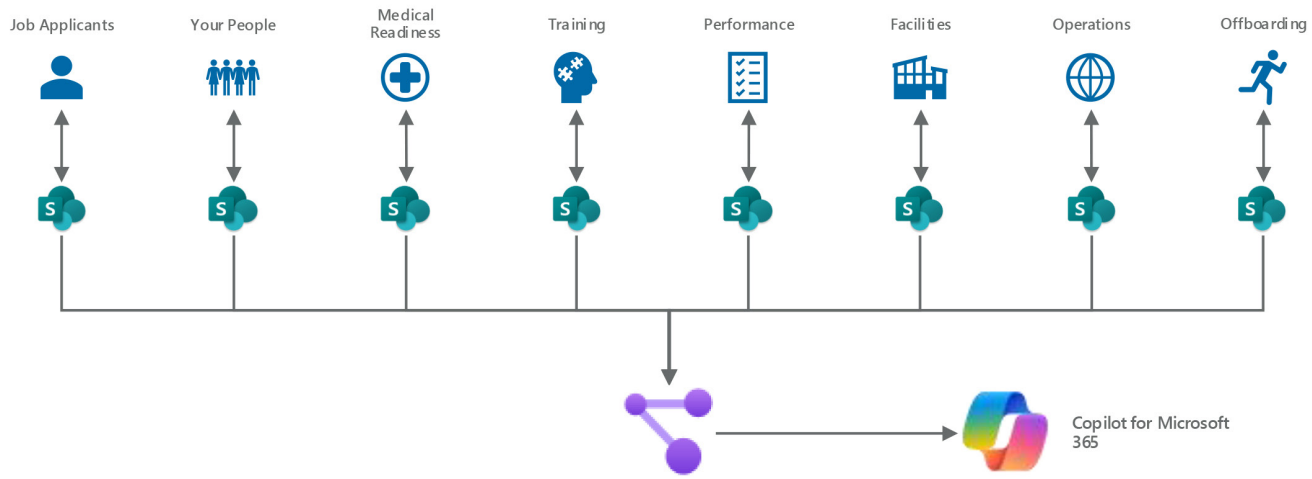


Figure 30: The diagram above replaces Excel, SAP, proprietary data storage technologies, MySQL, and SQL Server with SharePoint lists as the application data stores.

You see, the data kept in SharePoint is part of the Microsoft Graph, which itself hydrates Copilot for Microsoft 365 with your organization’s data. Those 8x copies of unsecured PII data have now been handed over to AI to craft generative responses for your users. Oh my.

To be clear, this is not a shortcoming of the technology. This is a shortcoming of poor architectural and security practices -fundamental misuses of the technology - that have created these security risks.

These are but examples of how technical debt hinders and creates significant risk when combined with artificial intelligence. Scaling AI requires these legacy technical debts to be retired to create budgetary space for new investments and to mitigate the risk that data from legacy, unsecured applications will improperly leak into your AI workloads.

© Cloud Lighthouse LTD, 2024

Monitoring and Metrics

Scaling AI is far from a purely technical endeavor. It's intuitive, in fact, to assume that long-term execution of strategy requires Monitoring and Metrics of the strategy's efficacy in producing business results. Well-rounded monitoring regimes should account for five key considerations:

- Maturity and risk;
- Adoption;
- Content moderation;
- Technical performance;
- Return on investment.

We're giving a bit away here in discussing maturity and risk at this stage of the paper. Skip ahead to the section presenting the AI Maturity Model to understand how to evaluate the organization's readiness or maturity for AI, as well as how to identify dimensions that present risk to be mitigated.

Adoption concerns the rate at which users take up a particular AI capability, how consistently they engage with the capability, and whether they continue engaging with the capability over the long term. We favor "weekly active users" (WAU), i.e., the number of users that actively use a given workload (let's say Microsoft 365 Copilot, for example) each week. Monthly active users (MAU) has been a favorite metric in previous waves of technology adoption, but we find MAU to be misleading in the case of AI because AI requires such a cultural shift; a user who only engages once or twice a month is not likely to adopt "born-in-AI" ways of working and is thus unlikely to make significant productivity gains thanks to AI.

We also recommend employing a "rings of release" method when releasing new capabilities, then monitoring the uptake amongst colleagues. This is standard fare in software deployments, but for the uninitiated, this approach groups users for whom the capability will be available into concentric rings, say a closed circle of technical users, then business early adopters, and an ever-widening pool of users until the capability has been released to all target users. Monitor for WAU (or daily active users, if it makes sense) in each ring, identify and remedy obstacles to adoption in the smaller rings, and avoid releasing more widely until you're satisfied with adoption in the predecessor ring.

Content moderation is crucial when implementing an AI product to ensure a safe, respectful, and legally compliant environment for users. Effective moderation addresses potentially harmful content, such as hate speech, explicit material, and misinformation, thereby preserving user trust and upholding community standards. For instance, input moderation may involve filtering user-uploaded images to detect obscene content, while output moderation could include analyzing text generated by AI to prevent the dissemination of inappropriate language. Content moderation is situational depending on your business, customers and goals. For example, requiring high amounts of blocking of violent imagery will be vital in business SaaS applications, but more nuanced in gaming.

Azure AI offers robust content moderation features, such as image moderation, text moderation, and video moderation, capable of detecting offensive content across multiple formats. Benefits include real-time monitoring, scalability to handle large volumes of content, and compliance with various international standards. These capabilities enable organizations to protect their brand reputation, enhance user experience, and foster a safe community.

Monitoring the **technical performance** of AI products is a multifaceted task that encompasses various metrics and benchmarks to ensure the systems are functioning optimally. Key performance indicators such as model accuracy, precision, recall, and F1 score are critical in evaluating the effectiveness of machine learning models.

Additionally, assessing workloads involves examining throughput and resource utilization to ensure the system can handle the expected volume of data and tasks. Responsiveness and latency are also vital metrics; low latency and high responsiveness indicate a well-optimized system capable of real-time or near-real-time processing.

Tools like performance dashboards, log analysis, and automated monitoring systems provide continuous insights into these parameters. Regular performance testing and anomaly detection are essential practices to preemptively identify and address potential issues, thereby maintaining the robustness and efficiency of AI products.

Azure AI Studio allows you to evaluate single-turn or complex, multi-turn conversations where you ground the generative AI model in your specific data (RAG). You can also evaluate general single-turn question answering scenarios, where no context is used to ground your generative AI model (non-RAG).

We learned through organizations' experience adopting Power Platform (an earlier Microsoft platform technology that entered the mainstream in the 2018-2019 timeframe) that many organizations crave **Return on Investment (ROI)** data for every minor workload that's deployed. This made sense in previous eras of big, monolithic software applications like ERP or CRM, but requires a more balanced, nuanced approach for AI (and for Power Platform, though this is a story for another time). Organizations that truly transform themselves for the age of AI will infuse AI throughout many, many aspects of its work.



© Cloud Lighthouse LTD, 2024

We therefor recommend that ROI be measured explicitly for major “anchor” workloads, and in the aggregate for more micro-workloads, in other words, an aggregate assessment of worker hours saved, or costs reduced across the workforce or department.

Digital Literacy

We’ll begin this final Digital Literacy dimension with a personal story from Andrew (one of this paper ’s co-authors, for those readers who didn’t skip straight to the biographies at the end).

In late 2023, my wife, Ana (another of the paper’s co-authors), and I found ourselves sitting at a café in Melbourne, Australia, with most of the day left before we needed to catch a flight. My first instinct was to spend thirty minutes scouring the internet for things to do or see in Melbourne, an idea that I did not relish because I dislike being glued to my phone when in good company.

Ana suggested that they seek advice from Microsoft Copilot (which was, at the time, branded simply as “Bing”). I had not thought of this, but curiously explained the situation to Bing and clarified where we were, how much time we had before we needed to go to the airport, and what kinds of sights we like to see when visiting a city. Much to my delight, in about fifteen seconds Bing suggested an entire itinerary for the day including sights to see and places to eat and drink. The itinerary was even organized according to a logical walking path from the place in the city where we were then sitting.

This capability had been in my pocket for months, but so ingrained was the impulse towards self-directed Googling that it had never occurred to me that AI-infused Bing could do the work for me so much faster. Off we went to explore Melbourne!

It is not enough that we put AI into our colleagues’ hands (or pockets), lest it stay there until some outside force compels them to give it a tr y.

To understand this predicament, let’s consider a 2023 study published by Boston Consulting Group (BCG) finding “that people mistrust generative AI in areas where it can contribute tremendous value and trust it too much where the technology isn’t competent.”

BCG goes on to enumerate the study’s key takeaways:

- “Around 90% of participants improved their performance when using GenAI for creative ideation. People did best when they did not attempt to edit GPT-4’s output;
- “When working on business problem solving, a task outside the tool’s current competence, many participants took GPT-4's misleading output at face value. Their performance was 23% worse than those who didn’t use the tool at all;

© Cloud Lighthouse LTD, 2024

- "Adopting generative AI is a massive change management effort. The job of the leader is to help people use the new technology in the right way, for the right tasks, and to continually adjust and adapt in the face of GenAI's ever-expanding frontier."

The change needed in most of your non-technical colleagues can be thusly understood as:

- **Knowledge:** Colleagues must know that a particular AI capability exists, what it does, where to find it (hopefully embedded in workstreams with which they are already comfortable), and - in some cases - be persuaded as to its merits over doing things "the old-fashioned way;"
- **Understanding:** Comfort breeds acceptance, so it is important that you help colleagues understand how one interacts with AI generally and a given workload specifically, which should include a healthy awareness of how to talk with AI, how to write an effective prompt, and an understanding that AI workloads thrive on better information (so, explain where you are in Melbourne, and what kinds of things you'd like to see);
- **Skepticism:** In particular, colleagues should have some grounding in how to be an ethical user of AI, recognition of possible hallucination, incorrect responses, or bias in training data, and an appreciation for the notion that AI workloads sometimes get things wrong, too.

Remember, also, that the leadership teams in most organizations have themselves not been immersed in concepts such as the latest technical skills, basic knowledge of data concepts, data literacy, responsible AI, how to use and apply AI in general, etc. Your digital literacy efforts ought to therefor include elements designed explicitly for senior and executive leaders such that they can develop the knowledge required to be the best possible organizational leaders in the age of AI.

Developers, engineers, IT colleagues, and others involved with creating AI capability ought to go several steps further.

Workloads and their user experiences should be designed to be increasingly born in AI rather than a traditional app with AI bolted on. As discussed earlier, this is a transition that will both take a bit of time to play out yet is likely to happen in earnest.

Microsoft's Project Sophia is already pushing boundaries that are likely to burst wide open as more and more architects and developers experiment, refine, and commercialize their born-in- AI solutions.

Whilst this first transition to born-in-AI workloads takes shape, the developer-equivalent to my Bing-powered exploration of Melbourne is already underway. For just as developers will learn how to build workloads that harness AI for end users, they, too, will continue to learn how to fully harness AI to help them build the workloads itself. It is difficult to land on a believable statistic here, but immense volumes of new code are now being written by AI in the form of tools such as GitHub Copilot.



© Cloud Lighthouse LTD, 2024

Meanwhile, Copilot for Power Apps and other Power Platform services are now creating significant pieces of the workloads themselves based on the Copilot's chatting with developers and citizen developers alike.

The human-centric change required of IT professionals will thus be two-fold:

- Learn to build workloads and user experiences that are born in AI; and
- Use AI to build the workloads and user experiences themselves.

CloudLight.house
Strategic architecture for the Microsoft Cloud

Chapter 3
AI Maturity Model



AI Maturity Model

The AI Strategy Framework offers a comprehensive blueprint through which organizations craft their future-ready enterprise AI strategy. Equally important is our ability to assess an organization's maturity or readiness for artificial intelligence, both in beginning to craft its strategy and regularly as it travels its roadmap.

We've developed the AI Maturity Model shown below to accompany the AI Strategy Framework previously discussed at length.

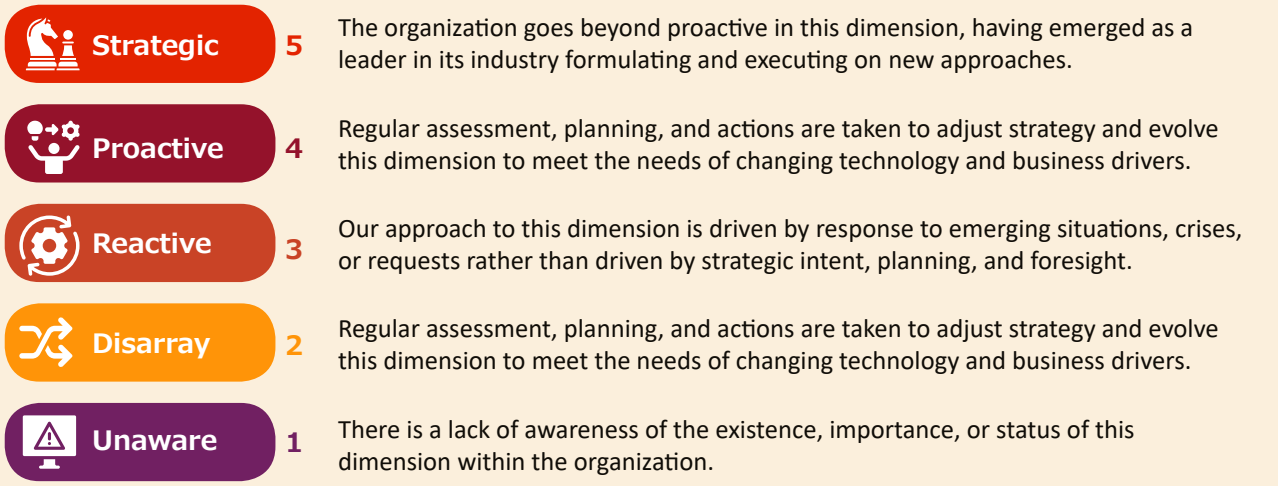


Figure 31: The AI Maturity Model allows organizations to assess their maturity or readiness for AI across each of the five pillars and twenty-five dimensions.

In the model, each dimension is reviewed with cognizant stakeholders - and your AI Center for Enablement team, we hope - to reach consensus on which maturity level and description fits best at the time of review. These ratings align to the 5-point scale shown, with "Strategic" (5) being the most mature and "Unaware" (1) being the least.

Apply the model to each dimension to determine each dimension's maturity relative to the others.

More mature dimensions are assets to be leveraged across the organization. They are also indicators of success that justify investment, in other words, where an investment has sufficiently matured a dimension and effectively lowered corporate risk. Less mature dimensions represent organizational risk and opportunity to unlock new capabilities, and should generally be a focus of investment.

Undertaking this assessment as you begin formulating your AI strategy promotes informed decisions as to which dimensions ought to receive early attention and be included in your actionable roadmap.

© Cloud Lighthouse LTD, 2024

Let's work through a practical example.

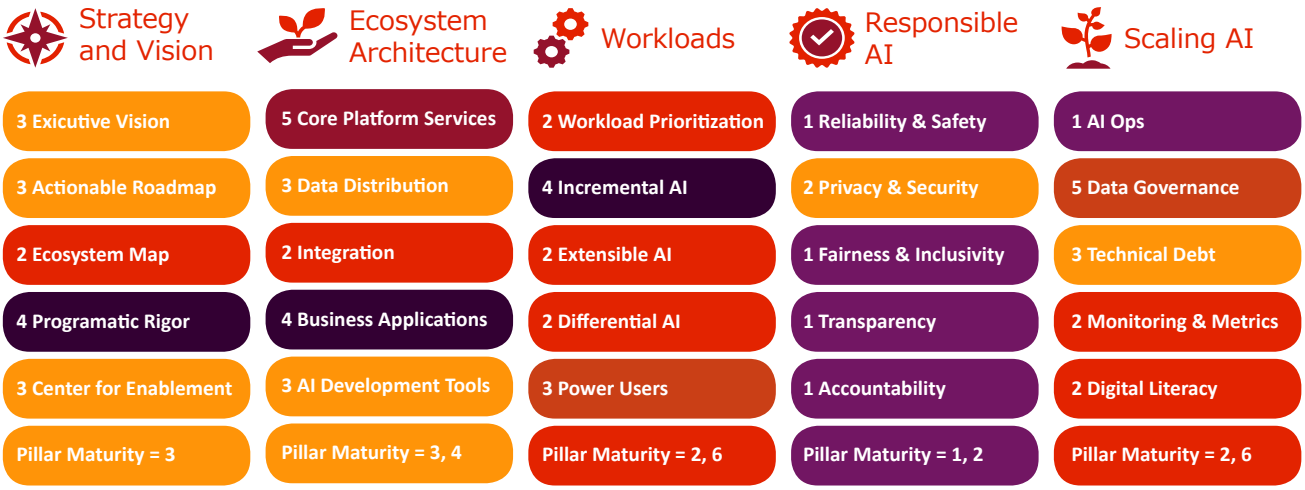


Figure 32: An example of the AI Strategy Framework with each dimension and pillar scored using the AI Maturity Model.

It's early days and we're just beginning to craft our AI strategy. We've worked through the dimensions one by one, giving a score to each. The diagram above reflects this, using averages to produce:

- Pillar Maturity scores of:
    - Strategy and Vision = 3
    - Ecosystem Architecture = 3.4
    - Workloads = 2.6
    - Responsible AI = 1.2
    - Scaling AI = 2.6
- Aggregate Maturity (the average of all dimensions) = 2.56, so, Disarray

Incidentally, we believe that any organization that achieves a score of 2.56 in 2024 should count itself lucky. Most are even less future-ready for AI. It's also worth noting that, based on our recent work with organizations around the world, a Pillar Maturity of 1.2 for Responsible AI is not hyperbole; most organizations are woefully unprepared for RAI.

© Cloud Lighthouse LTD, 2024

Apply this guidance when choosing which dimensions focus on in your actionable roadmap:

- Scores of less than "3" are high risk / high opportunity, so address these immediately;
- Scores of "3" are both a risk and opportunity for the organization, so address these when possible;
- Scores greater than "3" are lower risk and areas of strength, so protect them.

The model provides a common standard for assessing AI maturity and readiness, but it cannot be used on its own absent the insight and judgement that comes from professional expertise. The model is best used as a tool in the hands of experienced practitioners, not as a formulaic shortcut. In fact, Microsoft partners that take AI seriously should develop questions and methods that they can use to facilitate such assessments. Customers ought to challenge any Microsoft partner claiming expertise here to demonstrate it accordingly.

We recommend some ground rules when using this model:

- Round down when undecided between two maturity levels. It is better to overestimate risk than to ignore it;
- There is no shame in “Disarray”. It is better to admit where you are and fix it than to hope things magically improve;
- “Proactive” is a high bar to achieve. It means that you’ve planned and committed resources to evolving as AI technology and your business drivers change;
- “Strategic” is an even higher bar. Don’t award yourself lightly.

The pace at which an organization re-assesses itself is important. Too infrequent assessments can result in bad data that could skew risk management and resource allocation, whilst assessing too frequently can waste a lot of time in pursuit of only marginally more current results. In general, there are three reasons to update a dimension’s maturity assessment (up or down):

- When just starting out on your AI journey. It’s hard to know what to do next when you don’t have a firm understanding of where you are. We recommend assessing all dimensions in a single round;
- At a regular cadence that makes sense for your organization. This could be quarterly or half-yearly. It may also make sense to re-assess a different pillar each month to produce rolling maturity updates;
- When a compelling event occurs, which might include big Microsoft product updates, your organizations M&A events or major internal re-orgs, following an incident, major platform expansions in the organization, etc.



© Cloud Lighthouse LTD, 2024

If opting for a six-month regular re-assessment cadence, consider coinciding these with major Microsoft product announcements like Microsoft Ignite.

Rigorously applying this model and re-assessing yourself on a regular basis will not only equip you to keep the strategy fresh and relevant but will also demonstrate progress - and help to justify investment - from a less to a more mature state.



© Cloud Lighthouse LTD, 2024

CloudLight.house
Strategic architecture for the Microsoft Cloud

Chapter 4
Onwards



Onwards

We've used words like "journey" and "roadmap" to describe the path along which organizations execute their AI strategy. So, we'll conclude with a discussion of what that journey really looks like.

The Journey

The diagram below shows a simplified path that is reflective of that which most organizations are walking in their AI journey. Notice that we've tied it to the maturity levels in the AI Maturity Model, discussed earlier.

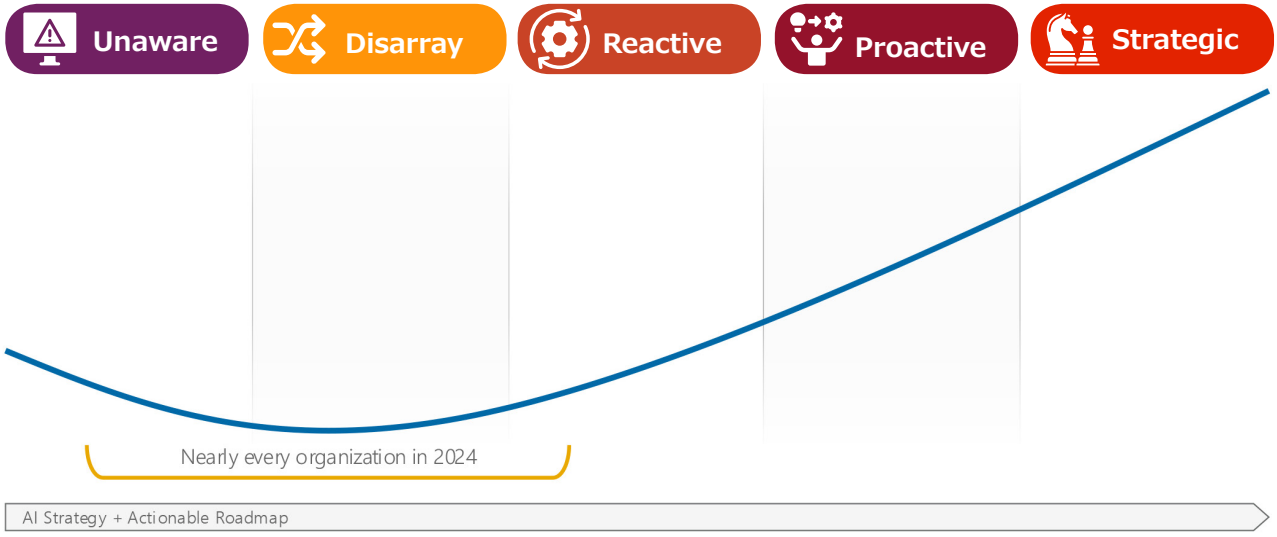


Figure 33: A simplified path that is reflective of that which most organizations are walking in their AI journey.

Consider, though:

- We see this as the general trend for organizations who embrace a strategic approach and rigorously commit to achieving their vision, the "real" path is likely to meander somewhat as you take two steps forward only to take a step back thanks to some unforeseen circumstance or technological development that takes you by surprise;
- The easiest maturity transition to make is from unaware to disarray, for it is easy to neglect a dimension to the point of letting it get out of control once you've realized that it exists at all;

© Cloud Lighthouse LTD, 2024

- Today, nearly every organization on the planet falls somewhere between unaware and reactive, and we have collectively yet to encounter one that has truly achieved a strategic maturity.

As organizations honestly assess their current maturity, they will be wise to also plot themselves on this diagram, drawing a line in the proverbial sand as to where they wish to be in twelve months' time. Hang it on your wall. Measure yourself. Be rigorous.

Organizational Obstacles

The best intentions and greatest attention to the people, platform, process, and AI portfolio at large are unlikely to suffice, though, in organizations that are not themselves structured for the era of AI (which is to say, "almost every firm on the planet"). Most IT organizations have carried forward significant relics from their legacy, pre-cloud models. Take care that you do not fall into these traps.

Use Case Death Spiral

First is the Use Case Death Spiral, wherein our classic IT approach focused on point solutions causing us to lose sight of the cloud ecosystem while we obsess about use cases. We always see knowing smiles and head nods when we discuss this topic at conferences, so you are likely to immediately recognize the phenomenon.

Most IT leaders "grew up" in our field during the long era of point solutions, so it is natural that they reflexively ask, "what's the use case?" when considering new technologies. The problem is that in platform technology - be it AI, ecosystem-oriented architecture, modern data platform technologies like Microsoft Fabric or low code technologies like Power Platform, and others - the use cases are essentially infinite, and they're driven by the organization adopting the technology, not by the vendor (Microsoft, in this case) providing it.

Organizations fall into the use case death spiral when they grow increasingly focused on enumerating and planning for every individual workload that they can conjure. They set to work identifying, designing, and prototyping the first workload. There is often, then, a demand to identify more workloads, so they workshop these until they have a big ol' stack o' use cases.

And just as they feel like they're close to a breakthrough, potentially with their dozens (or hundreds) of use cases identified, someone in the organization will pop up and ask:

But what are we going to do with AI after that?



© Cloud Lighthouse LTD, 2024

They go round and round on this so that months later they find that they've built nothing, achieved no value, and are little further than they were on day one. They will have produced fantastic shelfware in the form of analysis, lists of things, rumination of the art of the possible, etc. But they will have delivered no value to the organization.

You see, AI and other modern cloud technologies carry what Admiral James Loy, one of our co-author's long-ago mentors, called a "bias for action." Get as close as you can, analytically,

and then press forward. Incrementally, sure, so that you see your value grow over time. But you must get moving lest you fall into the use case death spiral where a fixation on workload number forty-two (and beyond) impedes development of workloads one, two, and three.

But there is a deeper problem at play here. Think about the story we've just shared and notice how it was largely a tale of pawing around for use cases, often going from one business stakeholder or group to the next asking them "what do you need?" or "how can we help you?" Whilst it's important to engage business stakeholders like this (see the Workload Prioritization pillar), the fallacy of the approach is that it positions AI and its ilk as a solution in search of a problem. You're asking, in other words, "hey, we have this thing that may be able to help you, but… umm, do you need help with anything?"

Overcoming the Use Case Death Spiral requires IT organizations to adopt a more rigorous approach to application rationalization, road mapping, and prioritization, and CI/CD around their AI and other cloud workloads, which leads us to our next organizational consideration…

IT Tower of Babel

Since (tech) time immemorial IT organizations have structured themselves in siloed, technology-specific teams. This organizational model tends to produce a phenomenon that we call the "IT Tower of Babel", wherein baskets of requirements are given to specific teams built around specific technologies. Give a basket to the RPA team, and they will build you a solution out of RPA. Give a basket to the data services team, and they will build you a solution out of data services. AI is a team sport requiring artificial intelligence and machine learning expertise alongside expertise in data science, data platform and integration, infrastructure, security, as well as application development.

AI is not a mountain made of its own tech somewhere off in the distance, visible through the clouds from the traditional IT mountain we're already standing on. Scaling AI is not about building out an AI team with AI baskets of requirements to produce AI workloads. Rather it is about weaving AI into the proverbial fabric of your cloud ecosystem.

Organizations that insist on treating AI as a distinct technology owned by a distinct team do so at their peril. We are working in an age where engineering teams must be as cross-pollinated as the technologies they represent. Better for leaders to spend their energy breaking down their IT Towers of Babel, not building new ones.



© Cloud Lighthouse LTD, 2024

Tyranny of the Deliverable

Re-building your IT organization to scale innovation by co-mingling different technical expertise throughout the org chart will be instrumental in creating a real culture of AI within any organization.

This is easier said than done thanks to the way that many of these teams are allocated funding and resourced from a budgeting perspective.

Many of the organizations we have worked with over the years build their annual budgets with line items tied to specific projects or deliverables, those “baskets of requirements” that we discussed. This approach is perhaps the single biggest way that organizational models from the pre-AI era prevent us from getting the most value from our AI investments. Consider an example of this model in action…

Your organization decides that it’s time to modernize its ERP solution. This is a worthy goal, so a bucket of money is created in next year’s (and likely a few years following) IT budget. This makes some sense in the context of big point solutions with multi-year implementation patterns. IT wants to reserve budget for an ongoing project, hold itself and its vendors accountable, monitor burn, correct for cost overruns, and in the end have some confidence that it will deliver a modern ERP solution to the business.

Unfortunately, this approach absolutely crushes innovation around AI and the development patterns through which you achieve it. Organizations that combine the IT Tower of Babel with the budget model from long-running point solution implementations applied to AI and other ecosystem-oriented technologies find themselves living something of the experience below…

Your organization has decided that it would like to embrace AI. This is a worthy goal, so a pile of workloads or business demands are prioritized on your roadmap. They each become a basket of requirements that get their own budget line item. Those requirements are then parceled out to

the AI team, whose leaders understand that they have X budget to “deliver” a solution that addresses Y basket of requirements. They (and their partners / consultancies / vendors / what-have-you) are now incentivized not to deliver truly valuable outcomes to the organization, but rather to check off as many pre-defined deliverables as possible. Management of these disconnected efforts causes overhead costs to skyrocket, as well.

And so it is that the organization succumbs to the Tyranny of the Deliverable, robbing itself the benefit of the shorter development cycles, opportunity to knit together multiple cloud technologies to solve problems, and agility that needs to be baked into your IT organization’s DNA if you’re to truly maximize the benefit of AI.



© Cloud Lighthouse LTD, 2024

This is, by far, the most difficult challenge to overcome of the several we've discussed. Difficult because this isn't just about adopting a technology or re-organizing a team, rather it's about fundamentally rethinking the way we fund the work of IT and measure its success. Consider several approaches to throw off the Tyranny of the Deliverable:

- Commitment to AI and your strategy around it absolutely must begin with executive vision that hangs a "north star" in the sky. This provides a clear answer to the question of why we are adopting this technology and what outcomes we seek as an IT organization and a broader business. What outcomes do we seek to achieve? And, importantly, are we prepared as an organization to measure our success in terms of outcomes achieved rather than deliverables crossed off a list?
- Start by taking some of those budget line items that you have allocated to specific baskets of requirements, and re-direct these funds to cross-technology solution teams and / or a trusted partner vendor whose mission is to execute on that vision and work towards the outcomes you've defined. Empower them to work flexibly, knock down problems quickly, modernize workloads rapidly, etc. And above all, to be outcome- focused rather than deliverable-constrained;
- This commitment must be sustained. Executive vision should be forward looking to not become obsolete next quarter. Your focus on outcomes needs to be sustained long enough to see those outcomes realized. In practice, if your commitment to the AI strategy - and to the executive vision you have articulated - can't be sustained for a year or more, then you have already failed.

Very early we explained that your organization is (probably) not ready for AI, because almost none are.

Very few - if any - organizations are truly prepared to make the most of the AI wave crashing on their shore. Very few have done the hard work to build the kind of proper, modern data platform required to make AI work at scale across their organization.

We've created the AI Strategy Framework and AI Maturity Model - and have written this extensive guidance - to prepare you, and the world's many other organizations like yours, to seize the moment and thrive in whatever future AI has to offer us. We will continue to evolve this guidance as the technology and our lessons learned about the technology evolve. Until then, remember that your AI strategy must be flexible, able to absorb tomorrow what we don't fully grasp today. Onwards.



© Cloud Lighthouse LTD, 2024

About Cloud Lighthouse

Wave periods between major innovation in the cloud are growing shorter. We no longer have the luxury of waiting it out, of adopting later.

Created by Andrew Welch - Microsoft MVP and experienced cloud technology leader - Cloud Lighthouse guides forward-looking organizations and their leaders crafting and executing their AI strategy based on the idea that cloud ecosystems built upon strategic foundations make organizations future-ready to absorb successive waves of change.

Cloud Lighthouse helps executive leaders become strategic leaders for the age of AI, to craft their AI strategy and then execute and scale it across their organization.

We serve:

- IT and business leaders, enabling them to become strategic leaders, crafting and executing their organization's journey in the Age of AI;
- Microsoft partners, enabling them to build new technology practices, nurture strategic client relationships, and grow their industry profile;
- Investors, boards, and portfolio companies, enabling investment decisions, steering portfolio companies, absorbing M&A, and investing smartly in AI and cloud technology;
- Technologists, enabling individuals through teaching, coaching, mentoring, and insights on strategic and ecosystem architecture.

Join us.

CloudLight.house/strategy

Follow on LinkedIn

© Cloud Lighthouse LTD, 2024

Author



Andrew Welch | Author

Founder + CTO @ Cloud Lighthouse

Cloud technology leader, founder and CTO of Cloud Lighthouse, Microsoft MVP (Most Valuable Professional), speaker, published author, and regular writer of essays and white papers on cloud technology including the multi-edition Crafting your Future-Ready Enterprise AI Strategy, Power Platform in a Modern Data Platform Architecture, Strategic Thinking for the Microsoft Cloud, The "Strategic Pyramid" of Cloud Ecosystem Architecture, and the multi-edition Power Platform Adoption Framework. Andrew works directly with CXOs, Microsoft partners, Microsoft product engineering groups, technologists, and investors to guide cloud + AI strategy and ecosystem architecture. His previous service includes that as CTO, global Vice President, Practice Director, and VP of Operations, and other tech and business focused roles at various Microsoft partners. He's led some of the world's largest Microsoft adoptions on all seven continents (yes, including Antarctica), serving startup to Fortune 100 organizations, public sector agencies, and global NGOs. Andrew is based in London and Boston with his wife, Ana, and daughter, Alexandra.

Follow on LinkedIn | Twitter

© Cloud Lighthouse LTD, 2024

Co-Authors



Ana Welch | Follow on LinkedIn



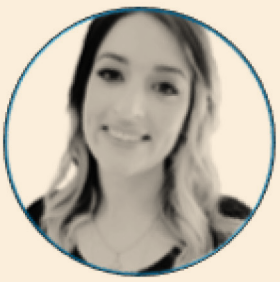
Dona Sarkar | Follow on LinkedIn



Chris Huntingford | Follow on LinkedIn



Samuel Smith | Follow on LinkedIn



Keegan Stanton | Follow on LinkedIn

© Cloud Lighthouse LTD, 2024

References

The below external articles are referenced in this white paper:

- BCG. (2023, 21 September). How People Create and Destroy Value with Generative AI.
- Castro, P. (2023, 9 March). Revolutionize your Enterprise Data with ChatGPT: Next-gen Apps w/ Azure OpenAI and Cognitive Search.
- The Economist. (2023, 16 July). Your employer is probably unprepared for artificial intelligence.
- Forrester Consulting. (2021, February). The Total Economic Impact™ Of Microsoft Power Platform.
- Microsoft (2023, 1 November). LinkedIn VP Aneesh Raman on Why Adaptability Is the Skill of the Moment
- Microsoft (2024, 18 September). Microsoft Purview.
- Microsoft (2024, 18 September). Copilot Learning Hub.
- Microsoft (2024, 18 September). Microsoft Project Sophia (in preview).
- Ray, S (2024, 9 September). Measurement is the key to helping keep AI on track.
- Ray, S (2024, 24 July). Red teams think like hackers to help keep AI safe.
- Ulagaratchagan, A (2023, 12 May). Introducing Microsoft Fabric: Data analytics for the era of AI.
- Wikipedia (2024, 15 September). Turing Machine.
- Wikipedia (2024, 31 July). AlphaFold.

Read on with the author's (Andrew Welch) previous essays referenced in this white paper:

- Welch, A (2021, 16 January) Crafting your Future Ready Enterprise AI Strategy
- Welch, A. (2021, 29 March). One Thousand Workloads: How your Roadmap maximizes Power Platform investment.
- Welch, A. (2023, 7 March). Strategic thinking for the Microsoft Cloud.
- Welch, A. (2023, 22 January). The "Tyranny of the Deliverable", and other short stories about why you're struggling to realize big value with Power Platform.
- Welch, A. (2023, 22 April). Understanding cloud ecosystem value and architecture via a "strategic pyramid".

© Cloud Lighthouse LTD, 2024