Infosys®
Navigate your next

VIEWPOINT

AN ARCHITECTURE FOR MATURE ENTERPRISE AI

Enterprise artificial intelligence (AI) has generally not scaled well or worked across business functions. It also struggles to react quickly enough to fast-changing markets. A reference architecture can help businesses scale AI that is more agile, holistic, and future-proof.

Infosys® | Knowledge Institute

The current generation of enterprise AI has hit a wall — a result of both business practices and technological limitations. Operating mostly in isolated pockets, different business functions take a piecemeal approach to develop their respective AI models. And solution architects deliver only what is needed for individual projects instead of considering the bigger picture (see Figure 1). The siloed systems make it difficult for organizations to adopt AI best practices and limit the technology's effectiveness. These structural barriers also slow organizations' efforts to keep pace with technological change.

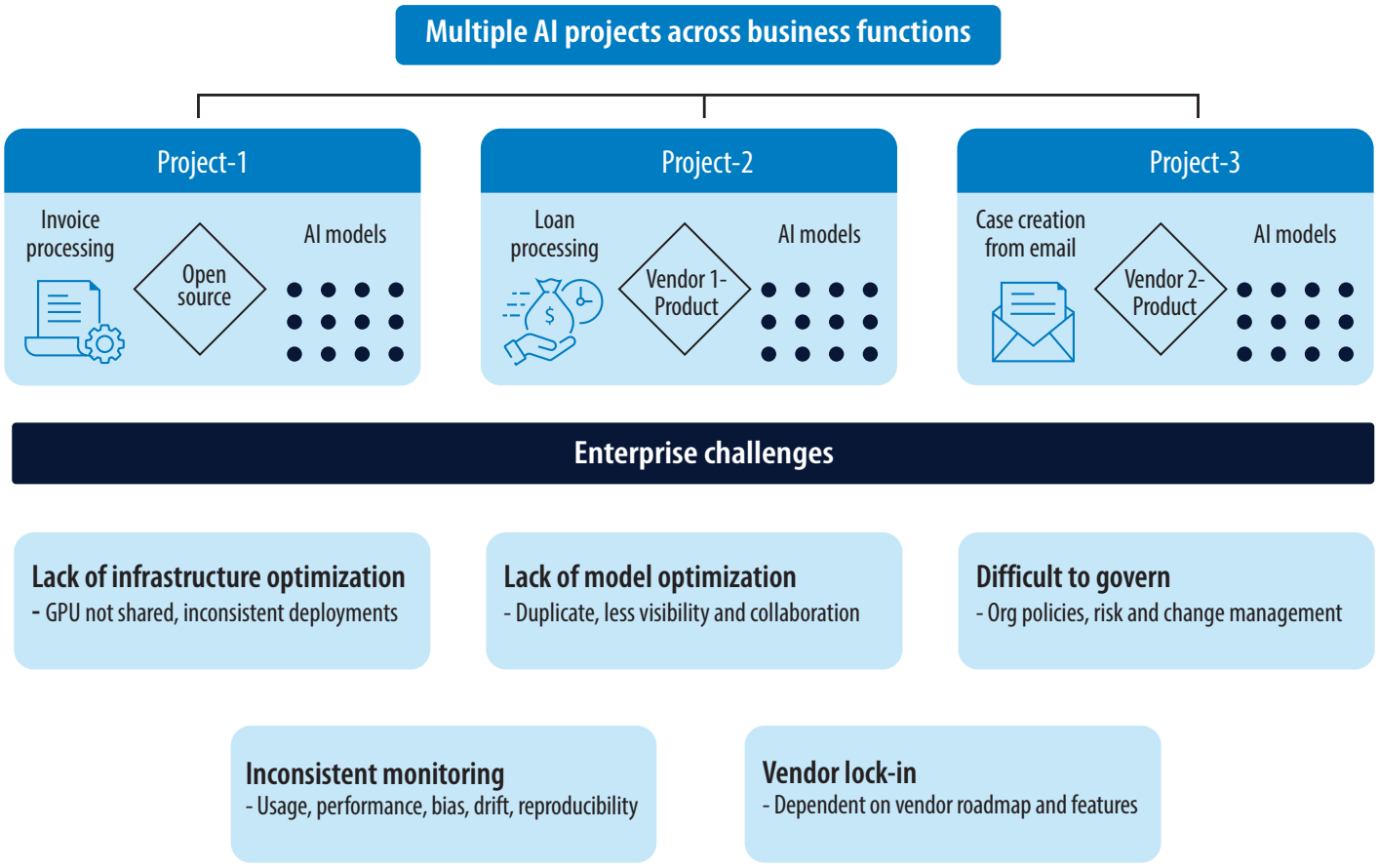
Despite its patchy adoption, AI is evolving rapidly. AI models developed just three years ago now require significant upgrades to utilize the underlying technology or are already irrelevant. Just 17% of companies reported fully mature AI and machine learning (ML) capabilities, according to a 2021 global survey by Rackspace Technology.[1]

AI has been criticized for years about its unfulfilled promises. However, companies are now starting to understand the shortcomings of how they've used the technology. And instead of retreating from AI, businesses are searching for ways to expand its use. To avoid previous failings, organizations need to build AI systems that are not dependent on the underlying technologies or tied to a single vendor.

Companies must also develop new governance and guiding principles that account for the expanded use of AI. In addition, AI developers and users need to be able to share information about the models, their underlying logic, and data usage. Collectively, this new approach will allow organizations to create systems that can grow and respond faster as new AI models reach the market.

For example, a large U.S.-based telecom company struggled with its lack of standardized AI development. Business functions — including sales, customer service, and finance — worked independently to develop their own AI models. These systems did not communicate internally, which limited their capabilities and benefits. To solve these problems, the company developed an enterprisewide self-service AI platform. This allowed data scientists, data engineers, business analysts, and other users to easily collaborate and develop cross-functional AI solutions.

Figure 1. Challenges of disconnected AI model development



Source: Infosys

Pillars for long-term AI prosperity

To ensure the next generation of AI is actually transformational, companies need to invest in models that can grow and learn from all of a business's data rather than in multiple models working in isolation. Through our work with clients, we have found that the most effective reference architecture requires three pillars:

- Future-proof the AI with a layered architecture.
- Democratize the technology through crowdsourcing.
- Scale AI enterprisewide using cloud native Agile systems.

> AI delivers most value when it is flexible to adapt to technological developments which comes from strong governance and a modular setup

Future-proof

A new AI platform should be able to keep pace with rapid developments in the technology (see Table 1). A strong set of governing principles can ensure that all future AI development is connected throughout the organization and is simultaneously transparent. Additionally, the platform would need a modular setup that allows developers to easily remove or modify models that do not fit current needs.

Democratize

To democratize AI development and usage, the platform should allow and encourage participation from all AI stakeholders, including users, developers, vendors, and business leaders (see Table 2). A robust AI system can take input from individuals with varying skills and roles, AI vendors, and the open-source community. This will promote healthy competition among providers and ensure the technology is easier to use enterprisewide.

Table 1. Principles to future-proof AI

	Principle	What does this mean?	Why is this important?
1	Layered architecture	Platform is divided into clear layers with predefined responsibilities and boundaries	Localizes the impact of future changes to a few components in a single layer and helps avoid vendor lock-in
2	Ability to switch in and out from the underlying technology and vendors	AI systems and services interfaces support multiple vendors and technologies	Allows AI components and vendors to be switched based on quality of service and price
3	Early adoption of responsible AI	Models support processes and technology for responsible AI metrics including performance, drift,* and bias	Keeps the governance model nimble and responsive to any changes in technology or regulations; encourages faster and easier regulatory compliance
4	Evolvability	Platform allows the addition of new capabilities and functionalities without depending on a specific technology or vendor	Enables faster adoption of emerging AI concepts and best practices, while learning from adjacent technologies

* Drift happens when an AI system starts delivering results that may not have been contemplated by its human developers/handlers.

Source: Infosys

Scale

The platform should be flexible so it can scale up or down as the organization's use of AI matures and needs change (see Table 3). Also, cloud and other adjacent technologies should be able to grow and support model development across uses, products, geographies, languages, and stakeholders.

> Active participation from all AI stakeholders is important for quick identification of new capabilities

Table 2. Principles to democratize AI

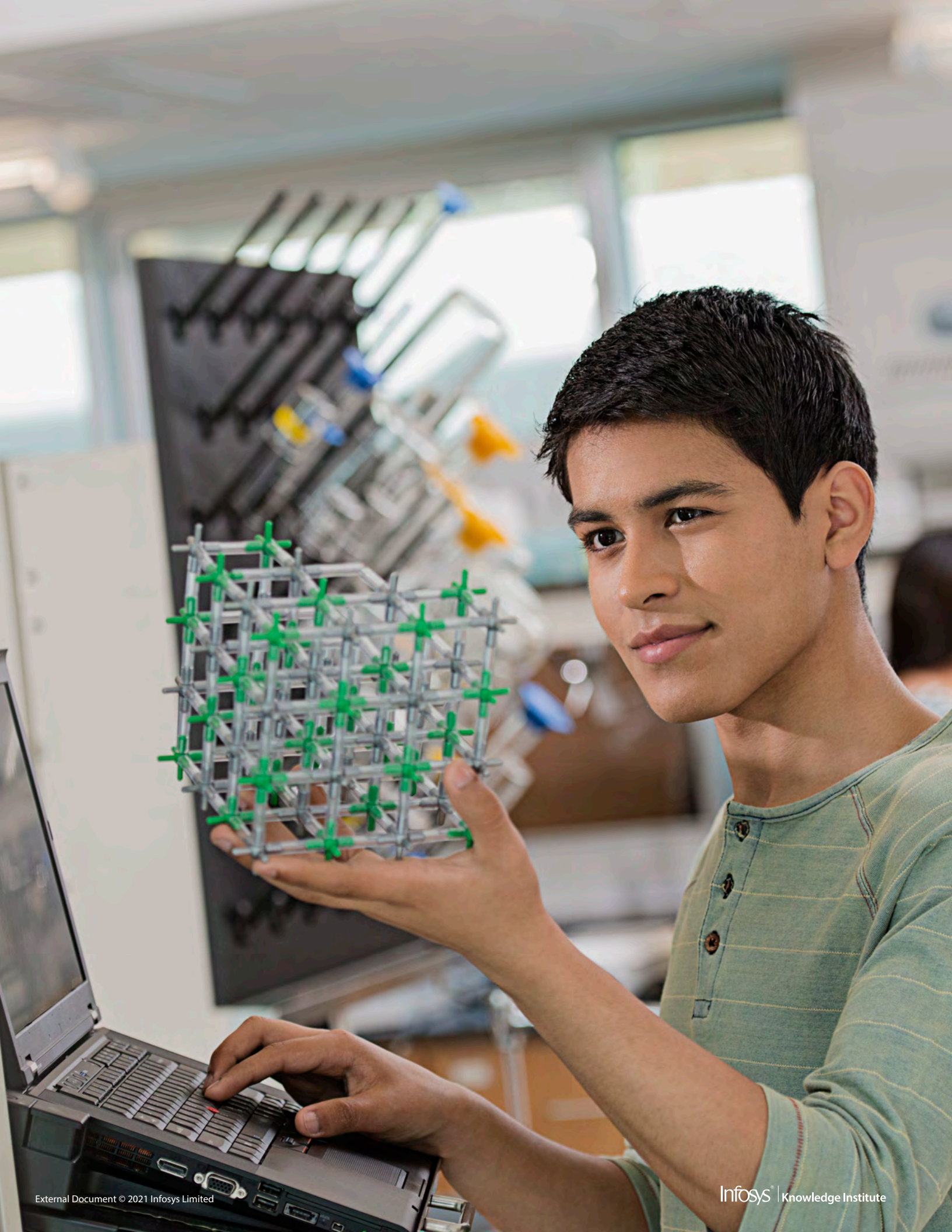
	Principle	What does this mean?	Why is this important?
1	Unified visibility	Publish AI metrics and use cases internally, including usage, performance, and interfaces between AI consumers and providers	Provides visibility to the business community about enterprise AI adoption and maturity; also creates opportunities for both small and large technology developers to easily contribute
2	Self-service	Break the silos between data engineers, data scientists, and ML engineers; developers and users with different skill sets can operate independently	Enables both AI users and developers to easily consume, contribute, and collaborate
3	Crowdsourcing	Use hackathons and competitions to develop datasets and models; also provide tools and incentives for different roles	Sources diverse AI assets through community engagement while allowing quick identification and incorporation of new capabilities
4	Choice for users	Support diverse AI tool sets, software development kits, and user channels	Facilitates easy use by all participants, enabling contribution from different skill and technology groups in AI development

Source: Infosys

Table 3. Principles to scale AI

	Principle	What does this mean?	Why is this important?
1	Cloud native	Microservices-based architecture deployed and scaled independent of other services; should include containerized applications with smaller footprints	Ensures easy scalability of technology components, with flexibility to add new capabilities; also enables higher efficiency and faster performance
2	Trustworthy	Publication of explanation and trust scores of the knowledge and logic used at the time of AI decision-making, along with data history and accuracy metrics	Establishes trust with users and helps accelerate AI adoption; also ensures easier regulatory compliance
3	Self-governed	Policy-based development of models and infrastructure to enable early data and model debiasing	Provides easier enterprisewide optimization and governance of AI components and infrastructure
4	Agile and iterative	Agile methodologies used for experimentation and proofs of concept for AI development and deployment	Enables rapid changes during development and easy onboarding of new AI technology

Source: Infosys



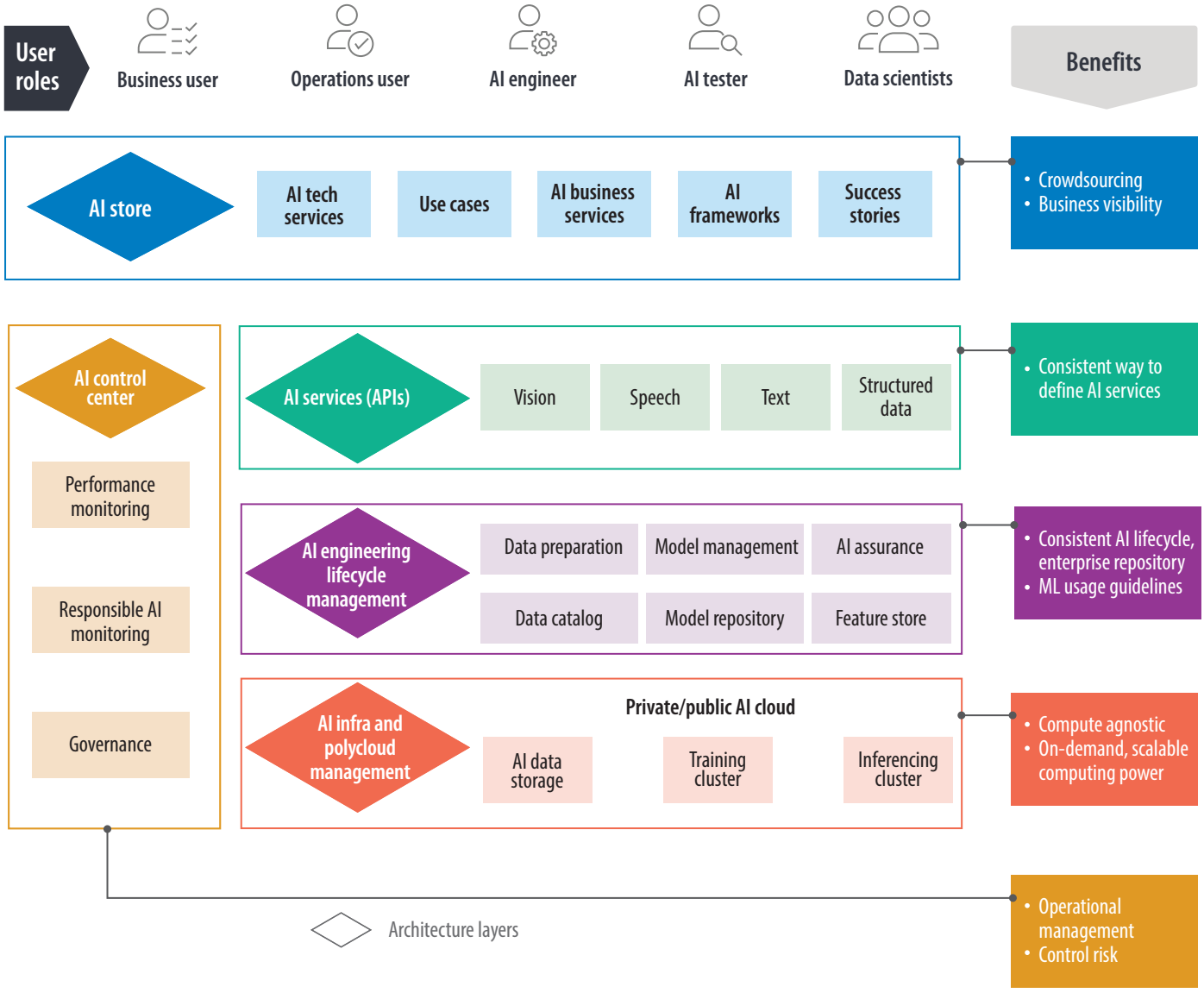
Reference architecture for the AI of the future

These three pillars, their associated principles, and a strong AI platform can ensure that enterprise AI is relevant to a company's business goals. With cloud native and Agile methodologies, the platform will keep AI development nimble and ready to accommodate future technology upgrades . In addition, the layered format (see Figure 2) enables many of the benefits found in these pillars and the other principles. This high-level logical view is composed of blocks that can be built using a phased strategy, much like Infosys used in creating its own Live Enterprise model.[2]

> Early adoption of Agile methodologies and cloud ensures AI stays nimble and ready for future developments in technology

This architecture consists of five layers, each of which can be developed independently with its own user personas, interface, technology, services, and deployment. Implementation of each layer connects with the organization's current technology stack, irrespective of vendors and preference for build versus buy. This allows the organization to adopt best-in-class solutions without depending on a single technology or vendor. Below are those five layers.

Figure 2. Reference architecture for an enterprise AI platform



Source: Infosys

AI infrastructure management

This layer optimizes infrastructure across multiple providers while ensuring there is sufficient processing power for model training. Cross-functional teams develop these models so they will be relevant to units throughout the enterprise. The layer manages data storage, hosts applications (on-premises and in the cloud), trains AI models, and conducts inference. Further, it helps reduce complexity across public and private cloud systems. The users are often teams that configure and operate infrastructure from multiple AI vendors, and ML operations teams that deploy AI artifacts and workloads.

AI engineering lifecycle management (ML operations)

This next layer supports multiple development tools and frameworks that can standardize the AI lifecycle. The layer collects AI artifacts (versions and metadata) across models for reuse. To enable faster development and deployment, it takes advantage of open-source and third-party frameworks supported by microservices-based cloud native architecture. The layer enables data scientists to use ML development tools to train and deploy models at enterprise scale. Further, it helps testers validate AI artifacts and helps IT operations teams configure and deploy AI policies.

AI services (APIs)

This layer requires a standardized API and catalog to enable users to access the AI services. The microservices-based design allows each API to provide small but well-defined functions. This layer also permits one AI service to be called from multiple applications. This function is based on the organizational policies and credentials of the calling application and provides flexibility to switch models without changing the underlying system. It also acts as a catalog of enterprise AI requirements for the internal data science community as well as for third-party AI vendors and hyperscalers. Designers and developers use this layer to understand specifications and integrate AI capabilities with applications. At the same time, data scientists can use it to develop new services, and AI monitoring teams can analyze the inventory of AI services and conduct gap analyses.

AI control center

This layer ensures that AI systems are consistent and optimized across business functions. It collects and reports AI metrics and measures them against business key performance indicators. The layer allows business teams to understand the model usage and intervene if the AI model is developing bias or not delivering the intended results. IT support teams use it to understand quality metrics, manage service disruptions, and optimize requirements. Data scientists use this layer to monitor and optimize the AI model's performance. And the layer can help organizations comply with existing regulations and adapt to future changes in regulatory requirements.

AI store

This layer offers a unified view of all AI artifacts, including success stories, capabilities, and innovations. It promotes quality content by seeking social feedback and monitoring quality metrics and ratings. Here, telemetry helps track user behavior. The integrated development and ML operations collect automatic updates from various AI tools and services, which can then be published. The layer is primarily designed to promote AI adoption and is typically used by all stakeholders, including end users, testers, data scientists, operations teams, infrastructure teams, and IT managers.

Working toward better enterprise AI

Developing a mature version of AI is now within reach, but it requires a significant investment and commitment. Firms must act decisively to move away from the siloed approach and disconnected use cases that have held them back. However, there is always risk when adopting enterprise AI. To mitigate that risk, organizations should seek a reference architecture platform that is free from vendor limitations and develop a culture where IT and business are aligned. In such an environment, AI will evolve and inspire the whole organization to become more productive, innovative, and customer centric.

Although enterprise AI is highly technical, business executives must be on board and willing to take the lead. An architecture without a high-level strategy or vision will fail. AI customers need benefits and value across all applications, such as sales predictions, cost optimization, and supply chain management. For this to happen quickly, the AI operating model must be built for cross-functional Agile teams.

In the next wave of advances, AI will be the linchpin for other technologies. But it can't stand alone. The success of enterprise AI depends on how well an organization ensures that its functions are connected and collaborative as it braces for the technology's next evolution.

References

1. Assessing AI and machine learning adoption hurdles across industries, May 12, 2021, Rackspace Technology.
2. Infosys Live Enterprise – A continuously evolving and learning organization, Mohammed Rafee Tarafdar, Jeff Kavanaugh, and Harry Keir Hughes, 2019, Infosys Knowledge Institute.

Authors

Amit Gaonkar

AVP - Senior Principal Technology Architect

Kaushal Desai

Principal Technology Architect

Harry Kier Hughes

Infosys Knowledge Institute

Abhinav Shrivastava

Infosys Knowledge Institute

About Infosys Knowledge Institute

The Infosys Knowledge Institute helps industry leaders develop a deeper understanding of business and technology trends through compelling thought leadership. Our researchers and subject matter experts provide a fact base that aids decision making on critical business and technology issues.
To view our research, visit Infosys Knowledge Institute at infosys.com/IKI

For more information, contact askus@infosys.com

Infosys®
Navigate your next