IBM Technology Atlas

Hybrid Cloud Roadmap

A distributed platform enables hybrid computing.

Realized
In progress

The journey	2026	2028	2029+
	Accelerate the development and delivery of AI solutions across hybrid cloud environments.	*Scale AI multi-agent systems to production using operating system for AI.*	*Empower agent web and quantum computing innovation with an operating system for hybrid computing.*
What are we doing and how we are doing it Implications for: Automation Data Security Systems	The AI innovation shift from models to AI systems that incorporate models with traditional components, and the applications built on top of these systems, represent a novel paradigm. Standards are emerging around these new concepts and interactions. We can view these trends as the beginning of a new application and protocol layer (layer 8) that leads to the evolution of middleware and platform. It creates a need to build an “operating system” for AI (AIOS) to manage and provide common services to this class of applications. AI innovation also intensifies the need for operating such non-deterministic systems at scale while integrating them into existing ecosystems. A wave of new security threats is also emerging stemming from cloud-specific agent indeterminacy in distributed environments. These create the need for novel observability and security mechanisms. The AIOS will support development, life-cycle management, and operations of AI workloads across the entire stack: models, tools, persistence, and applications. AIOS for enterprises will be implemented by extending and adapting an existing cloud platform rather than by building a new stack, as this allows the reuse of capabilities and best practices that provide reliability, security, and compliance. We will deliver a hardened, enterprise-grade inference stack for OpenShift AI with a production-ready control plane, SLO-driven autoscaler, pluggable scheduler, fast model actuation, and multi-tiered kv-cache offloading for heterogeneous hardware. We will develop a simplified and consistent experience for connecting models to data and will consistently and flexibly scale AI across hybrid cloud. Storage with AI capabilities (content-aware storage) will enable storage to process data effectively for agentic AI applications. We will have an MCP platform with registry and life-cycle management; an MCP gateway for agent-tool discovery, connectivity, and access control; and MCP servers for agents to manage hybrid cloud platform and infrastructure. Focused AI agents will assist with system operation by generating recommendations and remediations that build on top no/low-code techniques, supporting humans in the loop for change request approvals. AI agents will be able to define (code) development and operations guard-rails, observe their effectiveness in live systems autonomously, and fine-tune as needed. The ability to continuously and intelligently assess applications and operations against benchmark expectations will yield richer visibility into operations risk and deliver higher trust and confidence in AI agents in operations. We will deliver trusted agent identity support for agent-tool authorization flows based on existing and emerging standards. We will implement best-of-class security by integrating IBM enterprise security tools with AIOS to achieve identity management for agents; data protection; security posture management; and governance, risk, and compliance. Data-centric workloads will drive the need for batch inference using smaller-scale models to support various ingest and analytic flows in which millions of records or documents need to be processed. These scales also have significant cost implications, further driving the need for efficient bulk inference with bursty demand. Finally, the need for low-latency inference associated with data access will continue to drive the push of inference into data repositories. Our integrated AI agent platform will have agent life-cycle and layer-8 interconnect for security, safety, metering, quality enforcement, and observability, including seamless interactions of humans into agent life-cycle and operations. The llm-d "inference brain" will provide core scheduling, kv cache management, and autoscaling logic behind platform-agnostic APIs with key enterprise features for security and multi-tenancy. It will integrate with AI hardware like AIU Spyre. KV-cache-optimized storage will improve inference performance and cost.	Return on AI investment will depend on the ability to specialize AI capabilities to specific domains and enterprise workflows. For this, models and inferencing platforms will need to be customized to enterprise needs, AI platforms integrated with existing processes and systems, and AI agents will need to learn from feedback and adapt to context. Use of multiple models will become common as a way to optimize quality, performance, and cost for specific use cases. Agent memory will emerge as a new layer of the data stack. To accomplish complex tasks, agents will interact with one another without human coordination, leading to adoption of multi-agent systems (MAS). We expect further evolution of standards, focusing on reliability, security, observability, and trust. With more adoption of standards, ecosystems for tools, agents, evals, and others will expand . Production deployments will need to address concerns of infrastructure-level indeterminacy and security while adhering to compliance regimes. For instance, the scaling attack surface will expand across hybrid infrastructure, and manual operations will be overwhelmed by the distributed cloud complexity. DevOps practices and tools will change to incorporate the increasing use of coding agents without compromising software reliability, security, and performance. Autonomous, self-healing systems will emerge, and root cause detection will be more accurate and independent. Coding agents will provide autonomous remediation. Cloud platforms will be more frequently built by local vendors in small-medium size data centers driven by sovereignty concerns. AIOS will support multi-tenant deployments backed by multi-tenant infrastructure and platform packaged software. It will be operated by third-parties. In 2028, we will build a fully-formed distributed operating system for AI inference in the era of agentic workloads. We will showcase a unified control plane that federates and orchestrates entire graphs of multi-modal agents across a hybrid fabric, becoming the foundational substrate for next-generation AI applications. We will develop multi-party protocols for hybrid environments and automated policy management and compliance across clouds (policy as code). We will provide real-time monitoring, evidence collection, threat correlation, incident orchestration, distributed audit, and AI/ML analytics; and implement infrastructure-level consensus mechanisms. We will also lay a foundation for interoperability at scale for MAS, with a visibility and control layer across all interactions in MAS and tools to diagnose unexpected outcomes of AI workflows, assign accountability, and govern and audit MAS behavior. We will develop interfaces and workflows to integrate MAS with humans and systems into seamless multi-modal interactions. There will be persistent memory to capture and share an agent state and context. We will reinvent processes for systems and application management with infrastructure automation leveraging AI in a trustworthy and reusable manner. There will be tools to support agents and secure their life-cycle. Management and observability agents will handle reliability engineering, operational risk, security, compliance, and business impact assessments, and will provide recommendations based on business and technical KPIs. We will deliver Fusion storage with support for hybrid cloud multi-agents (kv cache and vectors sharing), support for agentic AI across structured and unstructured data, and integrated storage management into MAS for autonomous storage operation.	We expect the continued expansion of AI agents and the emergence of practically useful quantum computing to drive the technology landscape after 2028. The continued transformation of industries with AI will lead to a decentralized world-wide system of agents interacting in a variety of patterns. As was the case with the internet, this will require further evolution of protocols, standards, and control planes to support discovery and connectivity. Enterprises will require new mechanisms to control the information flow, prevent security exposures (including novel ones), and ensure alignment with organization objectives regardless of uncertainty. This agents web will drive a multi-fold increase of compute demand per knowledge worker across all aspects of AI: models, applications, state management, and operational management. This demand will likely lead to diversification of hardware specialized on specific tasks. If this holds, infrastructure will become even more hybrid, with emphasis on open architectures with open standard abstractions across the entire stack. The evolution of LLMs will significantly slow down and training costs will dramatically decrease. This will enable the creation of truly competitive open weights models. This may lead to the standardization of prompt formats, customization methods, and drive faster innovation in the application layer. As quantum computing matures to practical applicability, enterprise use cases will also emerge. Innovative organizations will want to integrate quantum computing capabilities into their applications and workflows, leveraging both AI and quantum computing for enterprise problems. We expect to see and influence the emergence of new applications that leverage high precision computing, AI, and quantum computing. Particularly in an enterprise context, leveraging these technologies will require improved alignment between the domains of enterprise application and high-performance computing. IBM infrastructure will support a quantum computing and agentic AI platform with a single storage/data management. Meanwhile, large-scale fault-tolerant quantum computers threaten to break hybrid cloud cryptography. This is addressed with post-quantum cryptography for hybrid infrastructure. Other security risks include potential cross-cloud data exposure from quantum computing advances, and infrastructure-level indeterminacy in a world of parallel execution and quantum workflows. This will be addressed with confidential computing, decentralized NHI permissions, and leveraging uncertainty for infrastructure verification.