databricks | TECH mahindra

WHITEPAPER

Grounding Agentic AI with Semara:

Building the Smart Semantic Data Layer for Agentic Enterprise AI

Author Details

Saurabh Jha, Tech Mahindra

Kemalcan Jimmerson, Data Bricks

Executive Summary

There's a version of the agentic AI story that's relatively straightforward:

you buy a platform, bring your data together, deploy agents, and watch costs drop. But CIOs and CDOs who've undertaken this journey recognize the difference between that pitch and operational reality. Agents act erratically. Outputs require constant human review, and the ROI that may seem clear in a vendor demo becomes impossible to achieve when the system runs against real enterprise data. The root cause, and almost always the bottom line, is the same. The agents in practice do not really understand the business. They enter data without understanding what it means, move through schemas without comprehending how those concepts relate to one another, and answer questions with unverified logic that isn't validated against anything.

Semara by Tech Mahindra helps solve that issue.

It helps create a semantic data layer atop existing enterprise data infrastructure, providing AI agents with what they've been lacking. A governed, constantly updated representation of business meaning. It constructs and sustains the ontology and knowledge graph, converting a competent yet unreliable agent into one that truly understands how the business operates. What sets Semara apart is the speed and cost at which it arrives at that point.

Legacy enterprise ontologies development methods are slow and expensive. Semara provides a foundation in weeks rather than years, at a fraction of the cost, without forcing organizations to rebuild their data estate or hire teams of knowledge engineers to make it work. This foundation can then continue to grow as the number of use cases increases. The paper presents an overview of why the semantic layer has become a deciding factor in ensuring the success of agentic AI, what Semara brings to the table, how it compares against the available choices, and what CIOs and CDOs need to do to get started.

Table of Contents

Why Agentic AI Keeps Disappointing

By the end of 2026, the vast majority of large enterprises will have at least one agentic AI deployment running in production. And the majority of those deployments will remain far below expected performance levels. The gap between what was promised and what's being delivered has become one of the more uncomfortable conversations in enterprise technology.

An agent who traverses a data warehouse sees rows and columns. It overlooks that 'net_rev_adj' and 'net_revenue_adjusted' in two separate schemas refer to the same thing. It doesn't understand that a 'stall event' within the manufacturing system is something that your maintenance team calls 'a downtime incident'. It cannot be reasoned that a surge in sensor reading S-14 on line 3 is historically associated with a particular failure mode, which maintenance fixed three times during the preceding quarter. That is knowledge that has been laid out in an environment, but is wrapped tightly around systems, in industry knowledge, buried in BI software that no one else can decipher.

The Grounding Gap: From Approximation to Accuracy

WITHOUT SEMANTIC GROUNDING	WITH SEMARA
Raw data tables & column names	Named business entities with formal relationships
Inferred join paths — often wrong	Governed join paths extracted from existing data estate
Metric definitions approximated per query	Validated metric logic surfaced from BI investments
No memory of past events or context	Living Knowledge Graph updated continuously
Hallucinations — confident but wrong	Grounded answers traceable to ontology & graph paths

The difference between these two columns is the semantic layer. Semara builds it — automatically, from the data estate organizations already have.

According to existing benchmarks, less than 20% of unstructured enterprise data and 32% of structured data are ready for AI workloads. The issue is not with storage or compute. It means data exists, but context doesn't.

An agent without semantic context is not incompetent. It approximates confidently. At the transaction volumes at which agentic systems operate, confident approximations can propagate thousands of errors before the human recognizes them. This is the 'hallucination tax', the cost of oversight and correction, and the loss of trust. The fix is not a better prompt or a different model. It's grounding, providing the agent with a more formal, well-governed representation of how the enterprise's data actually works and how all its data ties into everything else inside.

Ontology and Knowledge Graph: The Foundation That's Been Lacking

These concepts are not new and are deep in academic and enterprise history, stretching back decades, where they have been too expensive to scale or too slow to build. That's changing, and for a good reason.

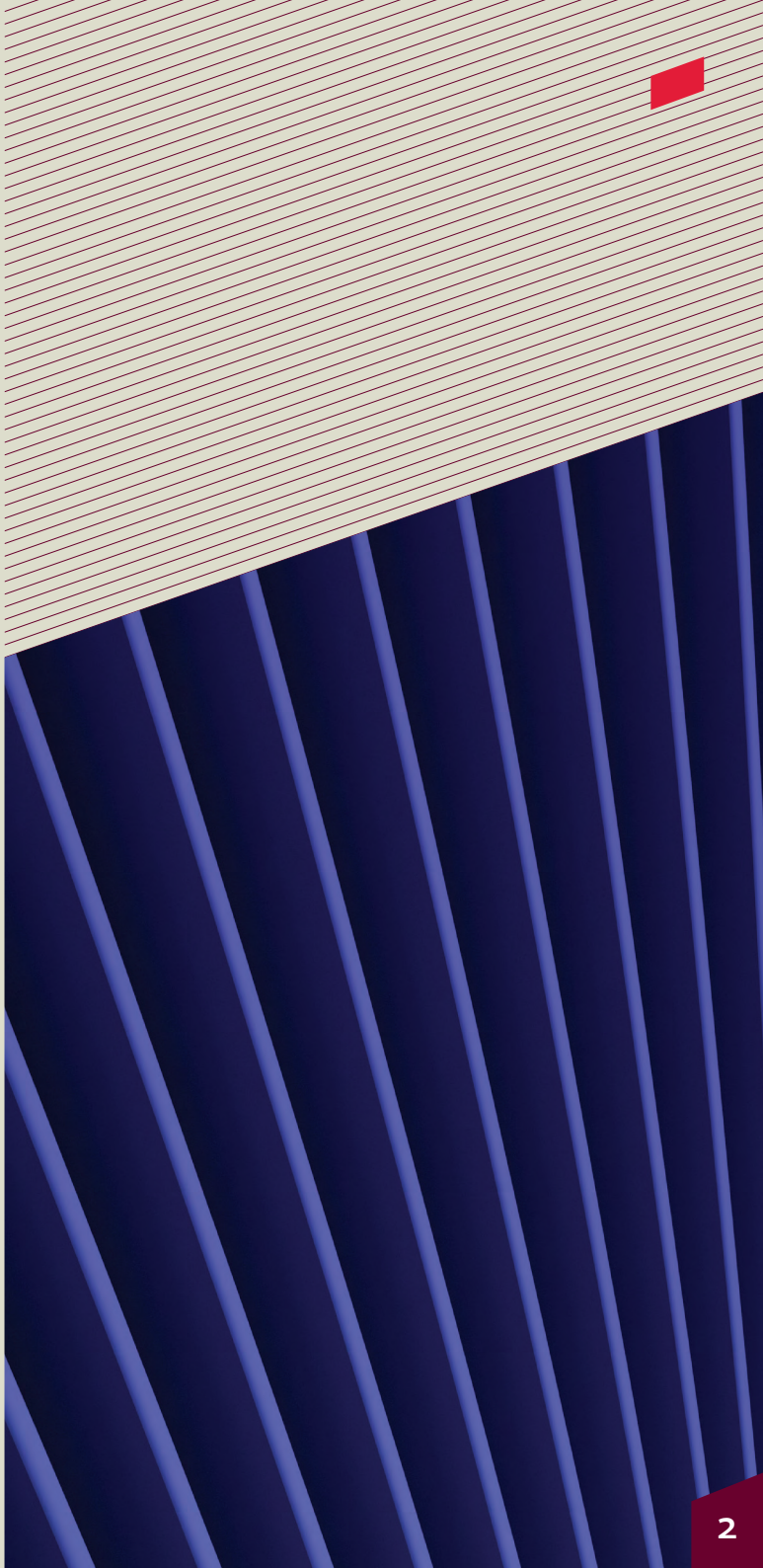
Ontology: A Shared Understanding of the Business

The concept of ontology in information systems was formalized by Tom Gruber, who in 1993 defined it as "an explicit specification of a conceptualization." Gruber's insight was that for knowledge to be shared across systems and agents, the vocabulary, relationships, and constraints governing a domain must be made explicit rather than left implicit in code or schema design.

An ontology is a formal model of the business domain. It describes every entity, every attribute, every relationship, every rule.

- In logistics, it indicates that a shipment belongs to an Order, an Order is linked to a Customer, and delivery delays are associated with combinations of carriers on individual routes.
- In finance, that means a transaction is associated with a product. Products are clustered within portfolios, and specific transaction patterns will attract regulatory scrutiny.

The ontology in itself is not a data dictionary. In addition to clearly enumerating each entity, it also provides descriptions of how closely they relate, how rules dictate their behavior, and what can be logically deduced from pairs of facts. It is the common language that enables a human analyst and an AI agent to ask the same question and obtain the same response, regardless of which system the underlying data is in. Without an ontology, any AI model that interacts with an enterprise's data is, in some way, making up its meaning. It's taking meaning from things like column names and schema structure, which shines in demos but is terrible at scale in production.



Knowledge Graph: The Living Memory Behind the Enterprise

A knowledge graph is where ontology comes to life. It stores every entity instance, including every machine, customer, product SKU, transaction, and the specific relationships between them, with timestamps, provenance, and context attached.

What that allows for is multi-hop reasoning; not just ‘which machine broke down?’ but 'what was the first sensor signal that came before the fault, which shift was running, has this very same failure sequence happened on this line before, and what was done about it?' Multi-hop reasoning refers to answering a question through a sequence of dependent inference steps, where each step builds on the result of the previous one. In a knowledge graph, each “hop” corresponds to traversing an edge between entities.

A flat, relational query can’t solve for that. A vector search on the embeddings can approximate it, but it's not reliable. A knowledge graph does it cleanly, traceably, and with the right auditability. The other thing a knowledge graph does that static models can’t do is that it compounds. All newly created events enrich it. Every agent query that adds an inferred relationship makes the next query smarter. When companies construct a knowledge graph at an early stage, they build a structural advantage that competitors without such information cannot replicate without delay.

Component	What It Does
Ontology	Formal model of business concepts, relationships, and rules.The shared vocabulary for humans and agents.
Knowledge Graph	Every entity instance and relationship is stored and queryable. Enables multi-hop reasoning and continuous learning.
Together	Give AI agents the contextual memory and logical grounding to reason accurately — and explain their reasoning.

Why Haven't Enterprises Built It Yet?

The real answer is cost and time. Developing a business-grade ontology and knowledge graph has traditionally been a major program, requiring dedicated budget lines, specialized teams, and multi-year roadmaps. Most organizations have determined that, while it is strategically desirable, it is unattainable at the pace required by their AI ambitions.

The Economics Are Punishing

An enterprise spanning four or five domains is considering a seven-figure investment before a single AI agent has been deployed to deliver the outcome. That is a hard business case to make when the underlying agentic platform is already being funded, and the impact is indirect rather than immediate and quantifiable. The platforms that have dominated this space are strong and powerful, but they were designed for organizations with the budget, time, and specialist capacity to match. Most enterprises, including big ones, are seldom able to meet all these criteria at once. The result is a capability that's now broadly acknowledged as key and yet greatly postponed.

Business Logic Exists, but Locked

There is a further complication, which is seldom openly discussed. Most of an enterprise's most validated business logic, the agreed definition of revenue, margin, customer health, and inventory efficiency, does not live in a database. It resides in business intelligence platforms. Years of fine-tuning, wrangling, and sign-offs from finance and business leadership are enshrined in formula logic that nothing outside the BI platform can natively access or integrate. Getting that logic out of there, standardizing it, and getting it in front of the AI agent has always come with specialized data engineering skills, something that sits at the crossroads of BI skills, data engineering, and AI architecture. This combination of skills is rarely readily available. As a result, either it gets done poorly or it gets deprioritized more often.

This is precisely the problem Semara was designed to solve.

What is Semara?

Semara is Tech Mahindra's proprietary IP that can help build the semantic data layer. It works in tandem with the enterprise data estate as it stands today; no migrations, rebuilds, or re-architecting of what's already been built and paid for. It connects to existing relational systems, data warehouses, lakehouses, and business intelligence environments to build and continuously update the ontology and knowledge graph. This gives AI agents the grounding that they need.

An Ontology Built in Weeks, Not Years

While traditional approaches require weeks or years of knowledge engineering to manually define business concepts and relationships, Semara creates a functional ontology from existing enterprise systems in a matter of weeks. The first output isn't a final product, no first-pass ontology of a complex business domain ever will be, but it's a substantial, controlled starting point that domain experts can check and tweak rather than construct from nothing.

This fundamentally changes the economics of the exercise. The first building block of ontology has been a slow, costly step for quite some time. Semara makes that quick and automated. What is left is the human work of validation and extension, which, if done adequately, is an authentic, meaningful effort by domain experts.

Enterprise Stack

Semara integrates with existing infrastructure — no data migrations or rebuilds required. Every agent query passes through the semantic layer, ensuring governed, consistent responses.

A Knowledge Graph that Grows With the Business

Semara is designed to periodically update the knowledge graph across the enterprise data estate, operational systems, sensor inputs, ERP history, CRM data, and maintenance data. The graph isn't a snapshot. It changes when the company is running. New nodes and edges emerge from newly created events. Agent queries that retrieve records of previously unrecorded relationships return results to the graph. It grows into a richer, more accurate representation of how the organization actually behaves. This is a representation that static data models, no matter how clever, absolutely cannot reproduce.

Built-in Governance and Security

With the Semara semantic layer applied to every AI agent query, access controls are applied consistently and automatically. An agent fetching supply chain information for a regional operations manager observes only what that person is authorized to see, no more. By default, audit trails on what data was accessed, what metrics were used, and what logic underpinned an agent's response are maintained. In a context of heightened regulatory scrutiny of AI systems, this is non-negotiable.

Compatibility with the Modern AI Stack

Semara's approach — connecting to existing data infrastructure without requiring migrations — fits naturally atop a lakehouse architecture, where structured and unstructured data coexist under a single governance layer built on open formats.

On Databricks, Unity Catalog already governs tables, models, and functions across the data estate. It also supports metric views — governed definitions of business metrics such as revenue, churn, or utilization — that represent the validated business logic the paper identifies as locked inside BI platforms. When Semara surfaces ontological relationships, and unity catalog governs the underlying metric definitions, agents get both structural context and definitional precision in a single stack.

Semara comes with an integrated MCP, the open standard that helps AI agents connect securely and consistently to enterprise systems. In this way, the semantic layer of the Semara code is made available as a queryable capability to all other agents or AI clients currently available on a compatible platform.

MCP is an open standard, originally introduced by Anthropic, that defines how AI agents discover, authenticate with, and interact with external tools and data sources through a unified interface. Rather than building custom integrations for every system an agent needs to access, MCP provides a standardized contract: the agent describes what it needs, and the MCP server exposes available capabilities — databases, APIs, knowledge graphs, semantic layers — in a consistent, secure format.

Semara's MCP integration means its semantic layer — the ontology, knowledge graph, and governed metric definitions — is exposed as a queryable capability that any MCP-compatible agent or orchestration platform can consume. An agent built on any compliant framework can discover Semara's semantic context, query relationships in the knowledge graph, and retrieve governed metric definitions without requiring Semara-specific client code. This composability is what makes Semara a platform-level capability rather than a standalone tool: it plugs into the broader agent ecosystem rather than requiring the ecosystem to adapt to it.

Governing the Agent-to-Semantic-Layer Connection with Unity AI Gateway

Exposing a semantic layer via MCP solves the connectivity problem. But in an enterprise with dozens of agents, multiple teams, and sensitive data domains, connectivity without governance creates a new risk surface. Which agents can access which parts of the knowledge graph? Who bears the cost when an agent makes thousands of inference calls to reason over the ontology? What happens when an agent's query should be blocked by policy?

Unity AI Gateway — now part of the Databricks Unity Catalog product umbrella — addresses this directly. It is the governance layer that sits between AI agents and the tools and models they use, including MCP-exposed services such as Semara. Key capabilities relevant to this architecture:

Guardrails powered by LLM judges:

Declarative policy enforcement can be applied to agent interactions with the semantic layer. Organizations can define rules — for example, that agents must not surface certain financial metrics outside of the finance domain, or that responses involving regulated data must include compliance disclaimers — and have those rules enforced automatically at the gateway level.

End-to-end request/response logging:

Every agent interaction with the semantic layer — MCP tool calls, LLM inference requests, knowledge graph traversals — is logged with full traceability. This extends the audit trail described in the paper from Semara's internal layer to the entire agent interaction chain, providing the end-to-end observability that compliance and security teams require.

Granular cost observability:

As agents scale from pilot to production, the cost of LLM inference and tool calls becomes a material concern. Unity AI Gateway provides attribution of spend across users, teams, and models, giving CIOs the financial visibility to manage agentic AI as a governed operational capability rather than an unmetered experiment.

The combination of Semara and Unity AI Gateway creates a complete governance stack for agentic AI: Semara provides the semantic grounding so agents reason correctly, and Unity AI Gateway ensures they do so within the boundaries the organization has defined — with the right permissions, the right policies, and full visibility into cost and behavior

Semara, part of TechM's D.A.H.L.I.A suite of IPs and accelerators serves as the semantic foundation beneath TechM Orion (TechM's enterprise AI orchestration platform) and provides a clean, governed context for the agent capabilities of TechM AmplifAI.

Semara Against the Alternatives

CIOs assessing this space are generally considering four alternatives. There is a real case for each; each has limitations worth clearly understanding.

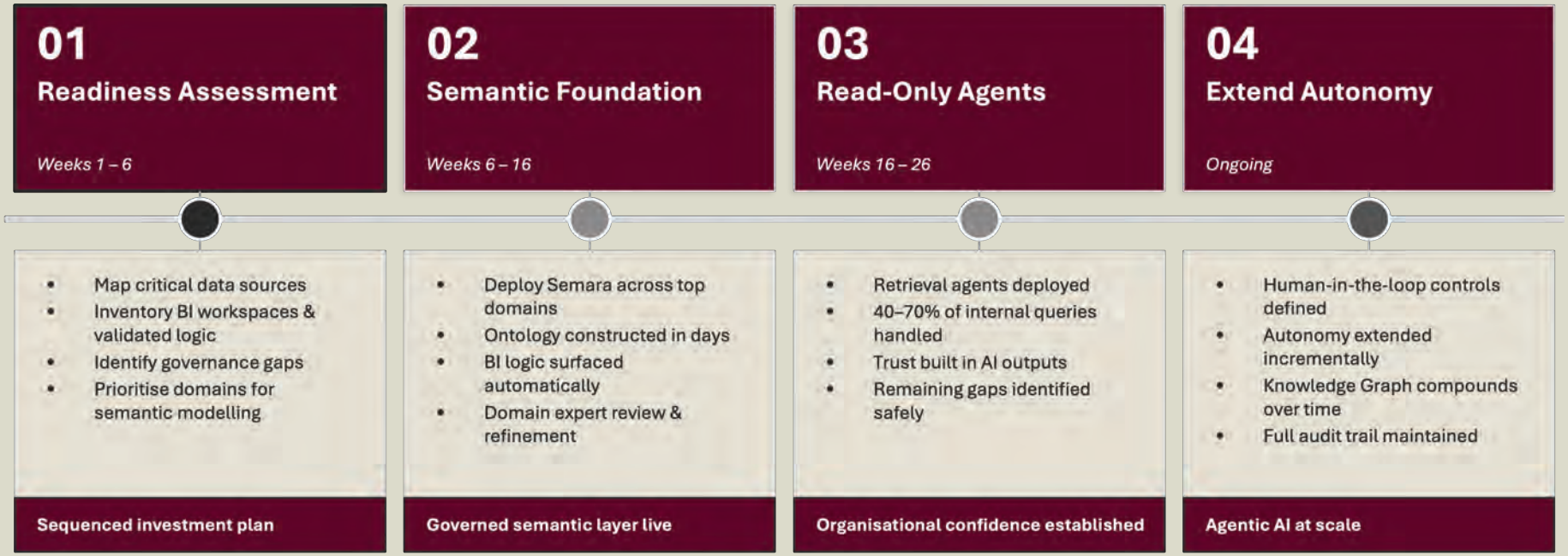
Option	Strengths	Limitations / Considerations
Build manually	Full control; fits exactly the organization's domain model	12–24 months; seven-figure cost; brittle without sustained specialist investment
Buy an established platform	Full control; fits exactly the organization's domain model	Designed for large budgets and specialist teams; slow to deploy; BI logic integration is typically custom
Rely on vector search / RAG	Fast to deploy; works well for document retrieval	Doesn't support structured reasoning; inconsistent on
		metric definitions; no graph traversal
Semara	Automated construction, BI logic surfaced natively, governed, and composable with the existing stack	Methodology is proprietary; best suited to organizations with an existing structured data estate

A Practical Path Forward

There is a temptation in enterprise technology to project a capability like this as something to achieve in a future state, a goal to plan around, a business case for, and start to deploy properly after the current wave of AI investment has settled. That instinct makes sense, and it's an expensive one in this instance. In that way, those organizations that benefit most from agentic AI in the near future are those that got started, even imperfectly, sooner than waiting for a full solution.

Here's a step-by-step process for CIOs and CDOs who want to move fast without creating problems they'll need the next two years to get used to.

Implementation Roadmap: From Readiness to Autonomous AI



These stages are concurrent design principles, not purely sequential phases. Governance, domain expert involvement, and data quality investment should run throughout.

Stage 1: Readiness Assessment (Week 1-6)

First and foremost, have a clear view of the data estate: which systems contain the highest level of decision-critical data, where the current validated business logic resides (e.g., in BI), and what governance mechanisms are already in place. This is not an extended discovery exercise; six weeks will suffice to identify the highest-value domains and significant gaps. The output must be a prioritized list of domains for semantic modeling, not a complete data list.

Stage 2: Build the Semantic Foundation (Week 6-16)

Use Semara up front to run over the top two or three priority domains. The automated construction functionality will generate a functioning ontology and initial knowledge graph population in days. Schedule two to four weeks of domain expert review and refinement, that's the part where machine output intersects the tribal knowledge residing within the organization. Together, this is faster and more accurate than working alone. At the end of week 16, you should have a governed semantic layer ready for agent deployment.

Stage 3: Start With Read-Only Agents (Week 16-26)

Don't let agents have "write" access to production systems to start with. Begin with retrieval agents: systems that bring information from the grounded semantic layer out, synthesize, and display it, but do not act autonomously. These retrieval agents handle a useful fraction of internal queries, ranging from supply chain status to customer account summaries to operational performance report generation. They will help build organizational trust in the outputs of AI and bring up any remaining holes in the semantic model before there is anything meaningful that relies on it.

Stage 4: Gradually Extend Autonomy (Ongoing)

With agent success verified by operational experience, gradually expand the boundaries of autonomy. For high-value or high-risk decisions, specify human-in-the-loop imperatives upfront before removing them. The semantic layer provides the audit trail that makes this progression auditable; every agent decision is traceable to the ontology logic and to the knowledge graph paths that generated it. This traceability is how we can cast trust in agents, rational rather than optimistic. Tools like MLflow Tracing support this by capturing the full execution path of each agent interaction — which data sources were consulted, which relationships were traversed, and what logic produced the output — giving teams a quantitative basis for deciding when and where to extend autonomy further.

Why Tech Mahindra?

Semara doesn't exist in isolation. It's part of an integrated data and AI capability that Tech Mahindra has been building for over two decades, and which now covers the full stack from raw data infrastructure to autonomous agent deployment.

TechM Orion is the enterprise AI orchestration platform, the environment in which agents are built, deployed, evaluated, and governed. Semara is the semantic grounding layer beneath it

TechM AmplifAI is the suite of AI and GenAI assets that operate above the semantic layer, including VerifAI for validation and guardrail configuration

D.A.H.L.I.A is a suite of IPs and accelerators that covers the entire data value chain, including ingestion, data products, quality and governance, and the data marketplace that will help feed clean, structured data into the knowledge graph. Semara is a part of this suite.

Tech Mahindra's partnerships with NVIDIA and commitments from all major hyperscalers will ensure that Semara's deployments are cloud-agnostic and not dependent on any single infrastructure provider.

This matters because a semantic data layer is only as useful as the agents and data pipelines connected to it. Deploying Semara alongside TechM Orion and D.A.H.L.I.A would mean that the semantic foundation, the data feeding it, and the agents consuming it are designed to work together, rather than being integrated after the fact by a systems integrator trying to make components from different vendors behave consistently.

Conclusion

Successful agentic AI programs share a similar property. They're grounded. The agents that work inside them know what the business's data means, understand the relationship between its concepts, and reason based on definitions that have been validated rather than inferred. That isn't an accident due to model choice or prompt engineering. It's all down to a strategic commitment to semantic infrastructure.

Most of the enterprise AI programs are not yet there. The data exists. The models are capable. The infrastructure is available. What's missing in most cases is the layer of meaning that actually makes all of that reliable, the ontology that provides the business with a shared vocabulary, and the knowledge graph of how that vocabulary applies to operational reality.

Semara makes that layer buildable. Not in the 18-24 months and at the price that conventional approaches require, but in weeks and at a price point that makes the business case simple. Investing in semantic grounding isn't another cost on top of the agentic AI program. This is what makes the agentic AI program worth doing.

The ROI of agentic AI depends entirely on data quality, governance, and context. Data and AI Infrastructure, AI models and tools, RAG, or any number of tools won't be able to supply it on their own. Grounding will. And Semara is how you get there.

Author Details

Saurabh Jha
SVP & Global Head - Data & AI, Tech Mahindra

With over 24 years of global experience, Saurabh has worked across India, Europe, the UK, and the US. He leads Tech Mahindra's Data and AI practice, which helps enterprises strategize, design, implement, and deliver data and analytics, cloud-based data, and AI-related transformation initiatives. He has a wide experience ranging from setting up new teams and practices, planning and executing go-to-market strategies, leading global alliances, and advising customers on effective alignment between their business goals and the latest digital technologies. Previously, he held strategic roles at Oracle, KPMG, and Mphasis, where he advised clients across industries and spearheaded regional expansions.

Kemalcan Jimmerson
Solutions Architect, Databricks

Kemalcan Jimmerson is a Solutions Architect at Databricks, specializing in enterprise AI, generative AI, and modern data platforms. He works with global consulting firms to design and deliver AI solutions for their most strategic clients, translating complex technology into measurable business outcomes. Previously, he led machine learning initiatives at Slalom Consulting, building AI platforms for Fortune 100 companies, and earlier developed analytics capabilities at Apple. He holds graduate degrees from University of California, Berkeley and New York Institute of Technology.

References

1. Gruber, T. R. (1993). "A Translation Approach to Portable Ontology Specifications." Knowledge Acquisition

About **Tech Mahindra**

Tech Mahindra (NSE: TECHM) offers technology consulting and digital solutions to global enterprises across industries, enabling transformative scale at unparalleled speed. With 147,000+ professionals across 90+ countries helping 1100+ clients, Tech Mahindra provides a full spectrum of services including consulting, information technology, enterprise applications, business process services, engineering services, network services, customer experience & design, AI & analytics, and cloud & infrastructure services. It is the first Indian company in the world to have been awarded the Sustainable Markets Initiative's Terra Carta Seal, which recognizes global companies that are actively leading the charge to create a climate and nature-positive future. Tech Mahindra is part of the Mahindra Group, founded in 1945, one of the largest and most admired multinational federation of companies. For more information on how TechM can partner with you to meet your Scale at Speed™ imperatives, please visit https://www.techmahindra.com/.

www.techmahindra.com
www.linkedin.com/company/tech-mahindra
www.x.com/tech_mahindra