Deloitte.

Together makes progress

From AI pilots to production: getting the tech right

Artificial intelligence (AI) experiments abound. But most enterprises have yet to realise major business benefits. That's partly because choosing the right technical architecture to scale AI quickly and safely isn't easy. This article explains how to make that choice.

Key Recommendations

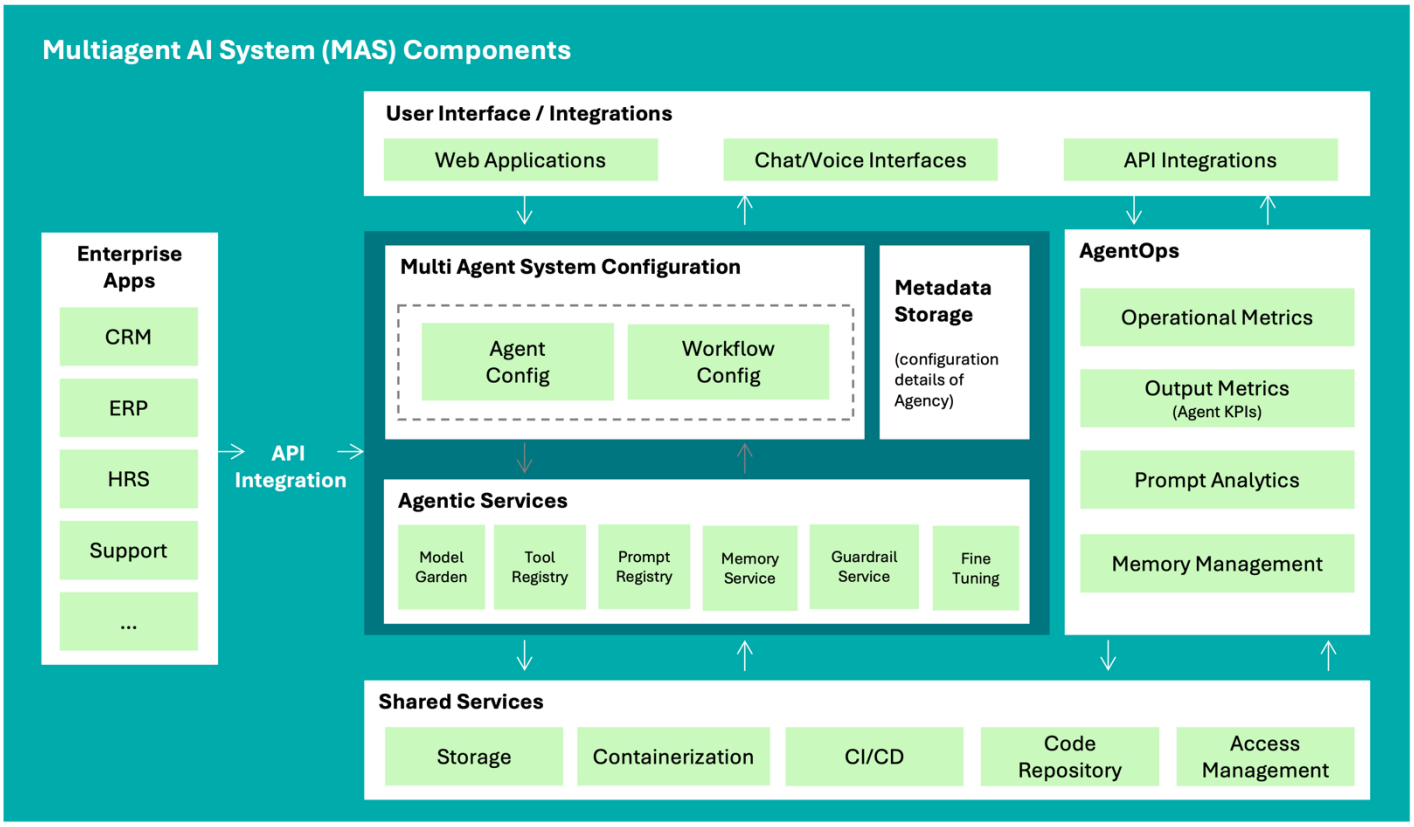
- Define the use cases where AI can add value and choose the technologies accordingly
- AI production systems need to be underpinned by a consistent enterprise-wide foundation,
- A cloud-native multi-AI agent system architecture can help deployments to scale
- Layered governance, encompassing policies, testing, verification and humans-in-the-loop, will reduce AI errors and bias
- Invest in data and knowledge management, so AI agents can retrieve, recall and reason consistently.

Choose the right architecture

When first deploying AI, start by identifying where automation can add significant value and then match the technologies to the job. There are typically two major elements in an enterprise-grade AI deployment. Foundation models (such as large language models) provide the reasoning core. These models are increasingly supplemented by "agentic" capabilities—systems systems that understand context, plan workflows, integrate with external tools and data sources, execute actions to achieve defined objectives, reflect on outcomes, and improve over time. For content creation and research, a well-aligned model with retrieval can suffice. For complex, long-running, multi-step tasks, agents with planning and tool integration are more effective.

When piloting a new AI system, it is important to define the success criteria, stakeholders, KPIs and a clear scope. While such pilots will inform technology choices, don't try to scale an isolated proof of concept (PoC) into production. All production systems should be underpinned by a consistent enterprise-wide foundation, consisting of tool registries, model routing, memory services, guardrails and observability. Ready-to-deploy platform components across Azure, AWS, or Google Cloud can help to scale quickly.

Ideally, enterprise-grade AI deployments will be based on a cloud-native multi agent system architecture, as illustrated in the diagram below. This architecture includes an interaction layer for users and applications; a workflow layer to control how tasks traverse agents; an agent layer (agentic services) where role-based agents run with defined tools and memory; and an operations layer (AgentOps) for monitoring, metrics and improvement. Enterprise architects will need to investigate what additional capabilities are needed for agentic AI and which to build (de)centrally.



To avoid overlap, keep agents role-based and group similar activities, while limiting each agent to only the tools and data it needs, and adopting the creator-validator pattern[1] for higher-risk content. A disciplined delivery model, coupled with architectural building blocks, paves the way for the data governance necessary to sustain quality.

[1] One agent (the creator) produces an output or proposes an action. A separate agent (the validator) checks that output against defined criteria before it reaches a user, triggers a tool, or updates a system.

Build for quality, reliability and safety

Indeed, robust governance is crucial. People will only use tools they trust. To build that trust, give humans responsibility for oversight (training, deployment, and monitoring), and ensure they remain in control for high-risk scenarios, while delegating routine steps to agents. For each agent, define autonomy levels, permissible actions, escalation paths and accountability.

An operational control centre, known as AgentOps, can be used to monitor how agents perform and flag anomalies. Agents' higher-risk outputs can be reviewed before being acted upon via feedback loops, human-in-the-loop checkpoints, and creator-validator patterns. For example, logistics company Flexport has implemented[2] an agentic AI system to automate repetitive freight tasks—such as data entry and terminal calls—while routing complex decisions and high-stakes exceptions to human operators for final judgment.

Memory systems, encompassing short-term working memory and longer-term semantic or episodic memory, help agents stay consistent across tasks and time, while prompt and tool registries standardise how capabilities are used. These registries (prompts, tools, models) can reduce drift and variation.

Guardrails, such as topic restrictions, prompt injection protection, and least-privilege tool access[3], are important to constrain agent behaviour. Agent telemetry[4], output metrics, prompt analytics and thought metrics[5] can be used to support observability and auditability. With the benefit of continuous insights, you can pinpoint potential improvements. Note, it is important to build these guardrails and observability into your AI stack from the start —do not bolt them on later.

In addition to keeping humans in the loop, employ strong prompt engineering to pre-empt mistakes, test agents for bias and misuse, and evaluate how they behave under real-world edge cases.

Don't overlook multi-agent risks, such as miscoordination, conflict, collusion, information asymmetries, deadlocks, cascading failures and destabilising dynamics. Such risks can be mitigated by clear communication protocols (orchestration or choreography), shared state where needed, and fallback patterns (including to another agent or a human).

Decide what to build, what to buy—and how to deliver

Should you buy or build your AI stack? The answer is very organisation-, situation- and use-case specific. For standard, well-trodden workflows, it can make sense to buy an AI system, but for complex, domain-specific processes, you may need to build one. If your use case has a native fit with your core packaged enterprise software, you would consider buying the native solution. If it needs bespoke integration patterns, you would build. For example, one leading media and technology firm opted to build its own causal AI platform because commercial products were unable to manage the immense scale, deep integration with proprietary systems, and the specialised analytical capabilities required for its core strategic business decisions.

In practice, many organisations will likely choose to combine general-purpose agent platforms with industry-specific solutions, open-source libraries, and hyper-scaler services. Sovereignty plays an increasingly important role in the decision.

---

2 Source: Air Cargo Week

3 Least privilege tool access is the principle that each agent can only access the tools, data and memory it needs to perform its defined role—nothing more.

4 Agent telemetry shows how the system runs, where it spends time, and where it fails, so you can keep agents reliable and efficient.

5 "Thought" metrics (reasoning trace signals) make reasoning transparent without exposing private content—they show how an agent planned, reflected and chose actions.

Make data your durable advantage

The ultimate performance of AI agents depends on high-quality, well-governed data. Therefore, organisations need to invest in data and knowledge management—structures and ontologies, knowledge bases, vector stores[6] and memory design—so agents can retrieve, recall and reason consistently. A robust data platform and data management capability will help organise both structured and unstructured data sources for discoverability and control. When data quality is designed in, performance follows—and the conversation can move from "if" to "how fast?"

Conclusions

Firstly, do not invest in a proof of concept (PoC) unless you have a clear vision of what you want to prove, and how you will scale and operationalise the solution if the PoC succeeds.

A successful shift from AI pilots to AI production depends, in part, on your technology and governance choices. To perform, protect and scale, agents need to be underpinned by a robust enterprise-wide foundation—architecture, guardrails, observability and data governance. With that in place, you can move faster with fewer surprises.

Finally, there is no need to reinvent the AI wheel. With so many enterprises now looking to deploy AI at scale, there is a fast-growing body of expertise about what works and what doesn't. Be sure to tap that expertise.

Contacts

For further information, or to discuss your AI challenges, please contact us.

Jorg Schalekamp

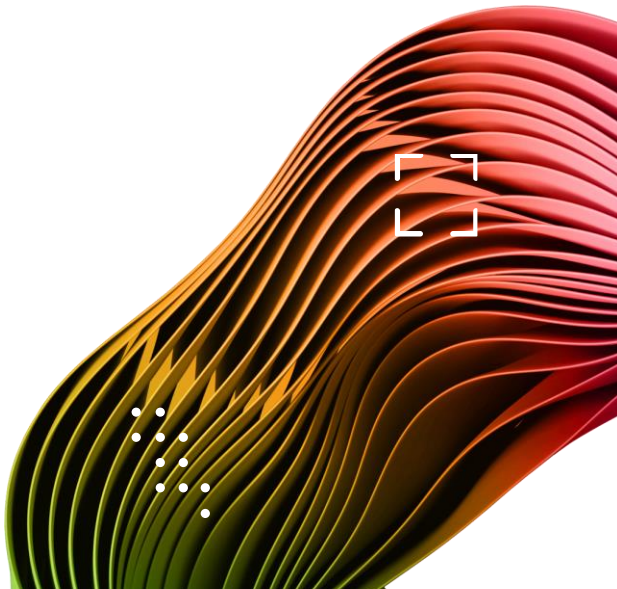
Partner Engineering, AI and Data & AI Leader

JSchalekamp@deloitte.nl

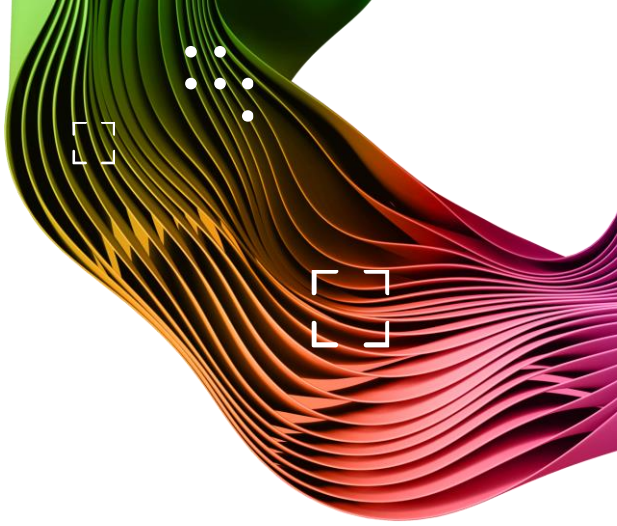
Stefan van Duin

Partner AI and Data

SvanDuin@deloitte.nl



[6] Vector stores are specialised databases that index numerical embeddings of approved content to enable fast semantic similarity search, so AI agents can retrieve the most relevant information to ground their answers and actions.



Deloitte.

CoRe Creative Services. RITM2402012