



Pharma lab AI scales only when data architecture comes first

By Anshul Agarwal, Sanjiv Chinnappan, Varun Chadha and Nirmal Venkatachalam



KEY TAKEAWAYS

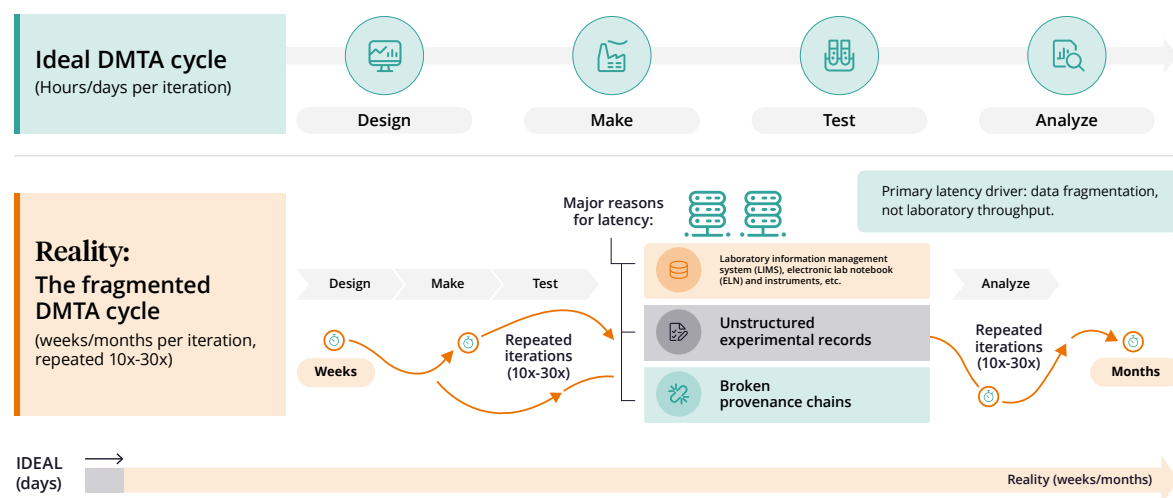
- **A lab of the future starts with a connected lab data architecture.** Trusted, machine-readable scientific information gives AI and laboratory automation a reliable foundation across the Design-Make-Test-Analyze (DMTA) cycle.
- **Successful pharma lab transformation follows architectural dependencies.** Connect and govern scientific information, integrate systems and orchestrate workflows before adding advanced AI.
- **AI-ready lab data turns faster scientific cycles into economic value.** Closed-loop experimentation reduces manual handoffs, accelerates decisions and can lower cost per insight across pharma R&D.

The pharmaceutical industry faces an unavoidable strategic pivot. By 2030, patent expirations on roughly 190 drugs, representing an estimated \$236 billion in annual revenue, will expose a significant pipeline gap. Organizations are filling this gap through robust pipelines, but bringing those products to market faster is equally critical. Compressing the timeline from scientific hypothesis to launched product takes more than headcount or capital. Digitizing the development life cycle can help compress drug discovery cycle time.

That pressure collides with a research environment where the two most valuable assets a lab has, time and data, are trapped in legacy processes. The industry's real constraint isn't scientific talent or analytics tooling. It's the data infrastructure for both. The DMTA cycle, the basic unit of drug discovery, is throttled by fragmented data, disconnected systems and broken provenance chains that keep AI and automation from operating reliably at scale (Figure 1).

FIGURE 1:

Fragmented systems, manual handoffs and broken provenance can stretch the DMTA cycle from hours or days to weeks or months



Fixing that bottleneck is fundamentally a sequencing problem. Build the data foundation first, then scale orchestration, AI and automation and measure the return in cycle time and unit economics.

Technical debt blocks AI and automation in brownfield labs

Walk into a typical pharma lab beginning a lab of the future program today, and the picture is familiar: instruments from eight to 15 different vendors, each outputting proprietary formats. Multiple electronic lab notebook (ELN) systems across functions, each with its own metadata schema and no shared ontology. A laboratory execution system (LES) and an ELN connected by brittle, point-to-point code written by someone who left the organization years ago. This is technical debt in its purest form: every workaround that solved yesterday's problem becomes a liability someone needs to service today.

This is the "brownfield reality." Unlike a greenfield build, where capabilities like robotic workcells or real-time digital twins can be architected from scratch, brownfield transformation means modernizing deeply entrenched infrastructure without disrupting ongoing science. In this environment, the barrier to AI and automation isn't a technology deficit. It's an architecture deficit.

Data format heterogeneity and system fragmentation are the two challenges organizations cite most often as barriers to scaling digital initiatives. Ironically, the heaviest past investors in point solutions are often the most constrained. Each tool solved a local problem while adding a new layer of debt on top of the last. Figure 2 summarizes seven recurring failure patterns and the architectural consequences they create.

FIGURE 2:

Seven recurring brownfield failure patterns turn local workarounds into architectural debt that blocks scalable AI, automation and cross-domain analytics

The challenge	The brownfield reality	The architectural consequences
1. The integration spaghetti web	Systems are connected via point-to-point APIs built for specific, one-off use cases.	Adding a new instrument or application requires geometric effort. Maintenance consumes engineering capacity.
2. The identifier crisis	No stable, enterprisewide experiment or run ID standard is propagated at execution.	AI models can't definitively link results back to design intent; Closed-loop automation is impossible, as the robot cannot find the data to determine the next action.
3. The "unstructured diary" ELN	Scientists record deviations, rationales, and protocol changes as free-text prose.	Orchestrators cannot parse the data, preventing automated workflow scaling and making the creation of useful process digital twins computationally prohibitive.
4. Ontology fragmentation	Each lab function or vendor system utilizes its own controlled vocabulary and naming conventions.	Cross-domain analytics fail. Data scientists spend 80% of their time manually mapping and normalizing terms.
5. Blanket GxP validation	The organization applies full regulatory validation overhead to all workflows by default.	Exploratory science is burdened with an "innovation tax," slowing technology adoption and iteration.
6. Disconnected CRO data	External partner data arrives via nonstandard PDFs, CSVs or emails with inconsistent metadata envelopes.	Internal AI models cannot use external data without massive manual reconciliation and reformatting.
7. Siloed model sprawl	Isolated data science teams build machine learning (ML) models locally with no centralized registry or ML operations (MLOps) framework.	Models become scientifically functional but regulatorily indefensible and organizationally un reusable.

These are symptoms of one root problem: lab systems were built for local workflows, never designed as nodes in a connected scientific supply chain. Pharma R&D doesn't need another point solution or another AI pilot. It needs architectural thinking at scale.



A 7-layer lab data architecture creates the foundation for AI

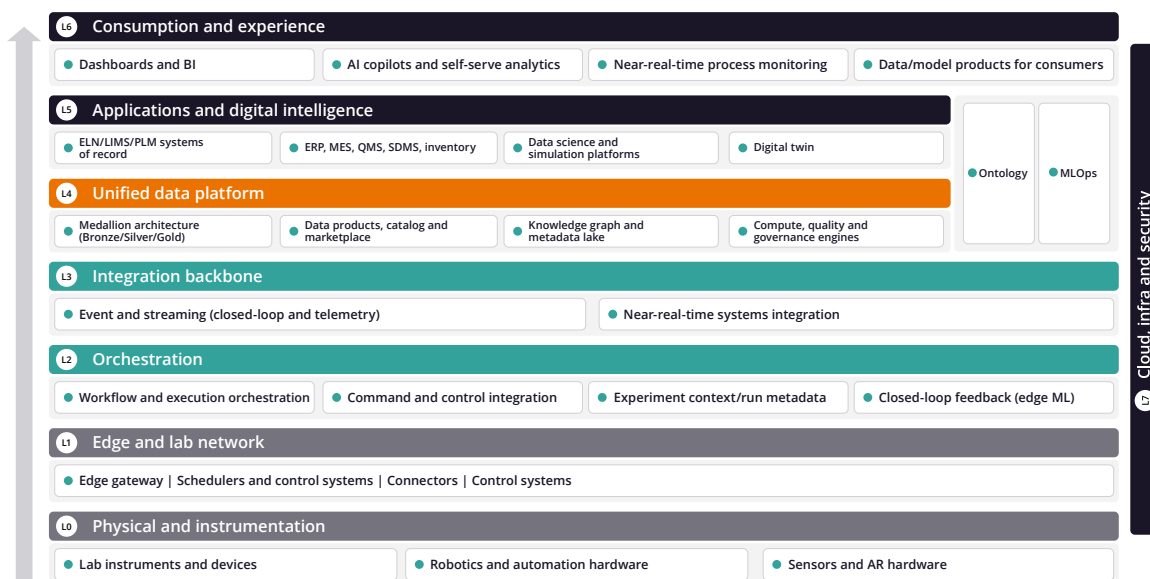
The lab of the future isn't a collection of standalone technologies bolted together. Think of it the way modern logistics solved the same problem for physical goods: a barcode, a container standard and a conveyor system that moves goods without a manual handoff at every stage. Lab data needs the equivalent—a single, cohesive architecture where every layer exists to keep information moving without someone reentering it by hand.

At a conceptual level, that architecture is organized into seven layers, running from the physical instruments on the bench up to the AI and analytics that scientists interact with (Figure 3). Data flows up. Orchestration and intelligence flow back down. The structural point worth understanding before anything else: AI sits at the top of this stack, but it depends entirely on the layers beneath it working—trusted data, machine-readable systems of record and a functioning integration layer. It can't substitute for them.

Why does the order matter so much? Consider what it takes to trust a single AI-recommended reaction parameter. That recommendation is only as good as the model that produced it. The model is only as good as the training data that shaped it. That data is only trustworthy if the pipelines that processed it did their job. And those pipelines can only standardize data that was structured at the point of capture, which requires the instrument to output something machine readable in the first place. Skip a layer, and everything built on top of it degrades immediately, often invisibly. This is why so many AI pilots produce confident predictions on dirty data, and no one can explain why the model's answers don't match lab reality.

FIGURE 3:

AI and advanced analytics depend on seven connected architecture layers, from physical instruments through orchestration, integration and governed data

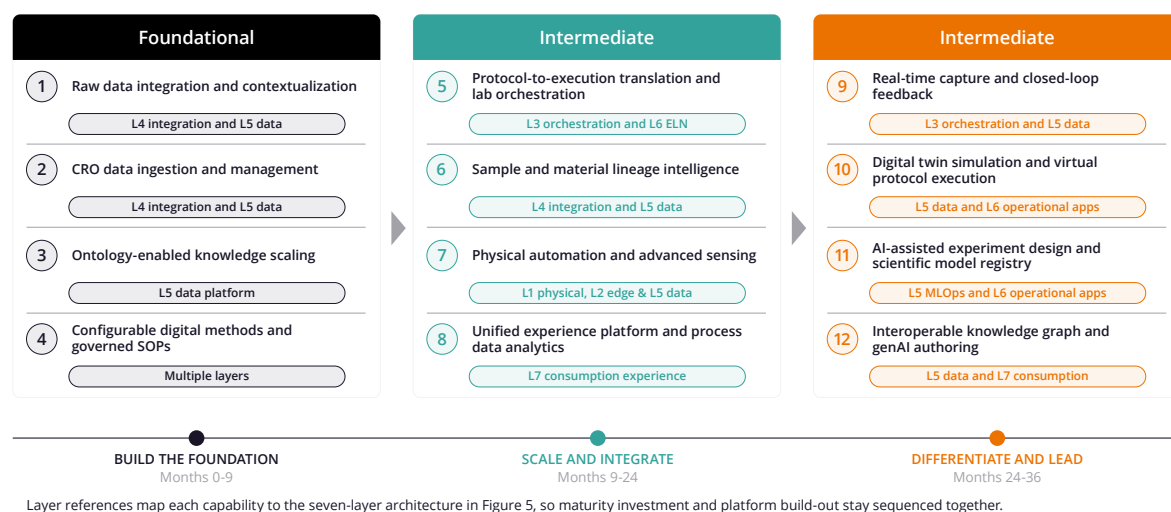


Building the seven-layer architecture is not the destination. It is the prerequisite condition. The advanced scientific capabilities that organizations want to buy (AI, digital twins, closed-loop autonomous execution) are differentiators that are strictly unlocked by specific foundational layers.

Organizations that pursue these differentiating outputs before building their prerequisites often discover that a highly capable tool has nothing trustworthy to work with. To move from pilot purgatory to scalable impact, organizations must view digital capability deployment as a maturity curve. The critical capabilities required for a modern lab map directly to this progression—starting with foundational data logistics, advancing to orchestrated execution, and culminating in AI-driven scientific acceleration (Figure 4). It’s worth grounding that progression in something concrete: what it looks like when a single experiment runs through this architecture on the bench.

FIGURE 4:

Lab transformation scales in sequence—connect and govern data first, integrate and orchestrate workflows next, then activate AI-driven closed-loop experimentation



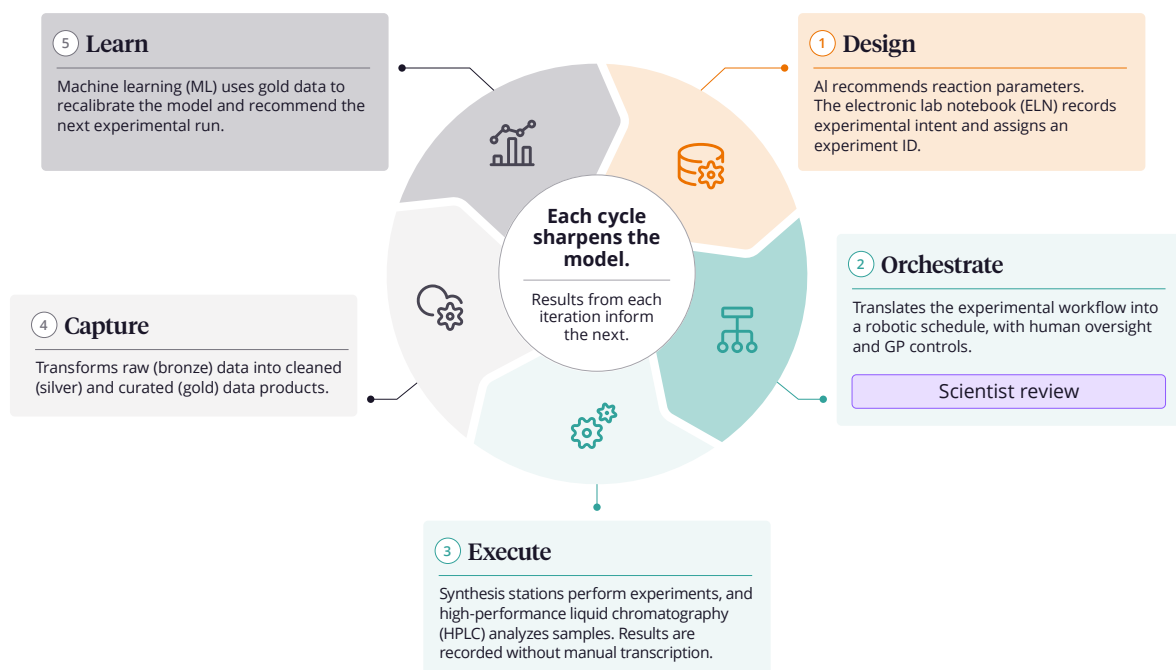
Connected lab data turns experimentation into a closed loop

Consider an Active Pharmaceutical Ingredient (API) development lab trying to find the optimal conditions for a complex synthesis route. In a connected architecture, this becomes a continuous loop rather than a manual cycle. An AI model reviews past runs and proposes a new set of reaction parameters. That recommendation flows automatically into a scheduled lab run, with a mandatory human checkpoint for review before execution proceeds. The physical run happens with minimal manual transcription. Results land in a governed data layer rather than a scientist’s personal spreadsheet. The model ingests the new results, evaluates them against the original hypothesis, and proposes the next condition, restarting the loop shown in Figure 5.



FIGURE 5:

A connected architecture turns experimentation into a closed loop in which AI designs, orchestrated systems execute, governed data captures results and models learn



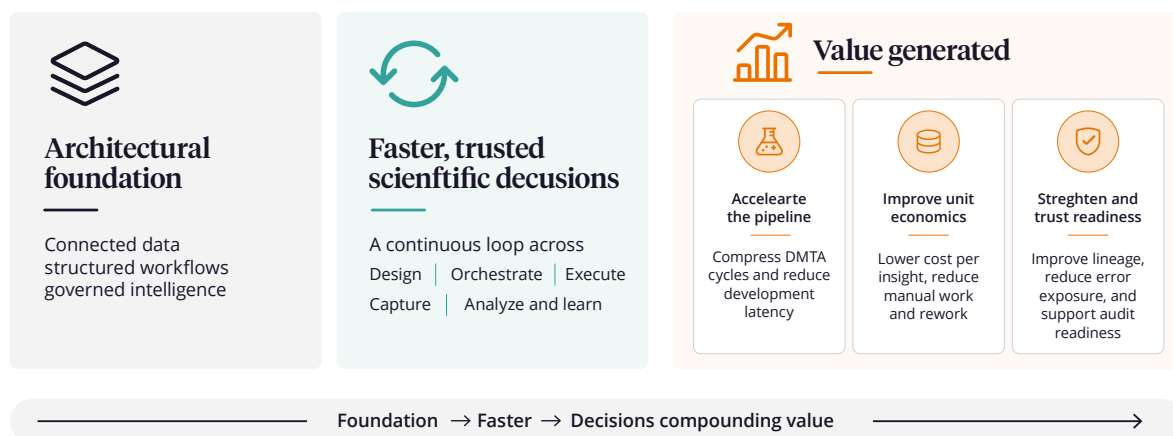
Every dollar this loop saves, whether in scientist hours, reagent cost or review cycles, is the same mechanism repeating at scale across the organization.

Lab architecture creates value by compressing scientific cycle time

Architectural investment in the lab isn't an IT modernization cost. It's a mechanism for compressing the scientific and regulatory life cycle at a unit-economics level.

FIGURE 6:

Architectural foundations turn faster, trusted scientific decisions into compounding value



1. DMTA cycle compression reduces development latency

The DMTA cycle is iterative and may repeat many times per development program, and its primary latency driver isn't instrument throughput. It's data reconciliation and manual transcription across disconnected systems. In one documented program, integrating ELN-to-orchestrator data flow accelerated molecule optimization by multiple months.

2. Automated experimentation lowers cost per insight

Closed-loop execution changes the cost-per-insight model of drug discovery. When harmonized data and machine-readable protocols connect design, execution and analysis, the marginal cost of an additional experimental condition approaches reagent and instrument cost—not scientist time. In one documented implementation, this kind of architecture enabled 36,000 reactions across six iterative rounds, achieving a 40% reduction in protein production cost. The enabler was data infrastructure, not automation hardware.

3. Structured workflows reduce compliance cost and human error

Human error drives more than 80% of process deviations in pharmaceutical manufacturing, with direct cost implications across rework, batch failures and GxP remediation. Routing review steps through structured workflows and replacing free-text deviation logs with controlled fields removes the conditions that generate errors in the first place.

4. Integrated data accelerates time to market

Integrated, orchestrated data infrastructure can accelerate time to market. That advantage isn't driven by which AI model an organization picks. It's driven by how fast reliable data reaches decision-makers and reenters the design cycle.

These four levers don't activate independently, and they don't activate once. Organizations that build the foundation compound all four simultaneously, every quarter.

Architecture-first lab transformation creates a widening, compounding advantage over time

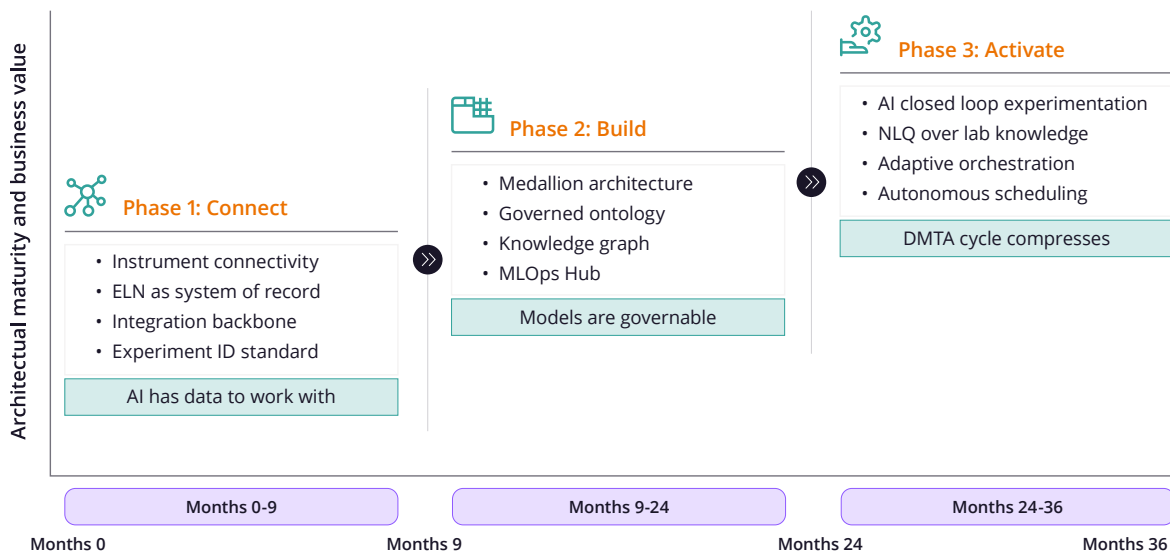
The industry is effectively running two divergent experiments, and the results are already visible in how the two cohorts' performance separates over time. The maturity curve in Figure 7 lays out the sequence: each layer of architecture unlocks the next tier of capability, and organizations that skip ahead don't move faster, they just hit the ceiling of whatever layer they skipped.

One cohort is building genuine digital transformation—stabilizing the integration backbone first, turning the ELN into a machine-readable decision system, building harmonized data models before layering on AI. They're not ahead on technology procurement. They're ahead on architectural compounding. Every governed data product becomes permanent infrastructure. Every closed-loop workflow teaches the organization something a competitor's siloed pilot can't. The lead widens as those capabilities compound.

The second cohort is digitally fragmented and accelerating in the wrong direction—buying AI platforms before resolving basic identifier problems, deploying robotic schedulers that can't read the ELN because no one enforced structured templates before, running pilots that work in isolation and can't scale. The tragedy isn't a lack of ambition. It's that every capability layered onto a fragmented foundation makes that fragmentation more expensive to unwind.

FIGURE 7:

A phased architecture roadmap moves from connected instruments and systems of record to governable models and, ultimately, closed-loop experimentation and autonomous scheduling



What separates these cohorts isn't which AI model or platform they select. It's whether their data infrastructure can actually feed it. As access to generative chemistry models, predictive assay systems and autonomous scheduling tools broadens, operating rigor and data infrastructure will increasingly determine how effectively organizations can use them. The firms that built the foundation will use those tools at full power. The firms that didn't will run the same tools on fragmented, untrustworthy data—and get fragmented, untrustworthy results. The technology will be equal but the operating rigor beneath it will not.

The question isn't whether to build the foundation. It's whether you're still positioned to lead, or already managing the consequences of not having started.

A technical blueprint carries the architecture into execution

We go considerably further on the full technical blueprint—including the detailed layer-by-layer architecture, implementation sequencing, deep dives and validation by capability—in our work directly with clients. If this framing matches what you're navigating in your organization, we'd welcome the conversation. [Reach out to compare notes](#) on where your architecture stands today.

About the authors



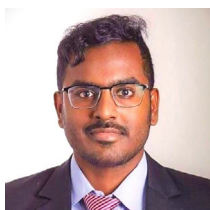
Anshul Agarwal Anshul leads ZS's global supply chain and manufacturing practice, driving large-scale manufacturing transformation for life sciences organizations through AI, advanced analytics and intelligent digital ecosystems. He modernizes operations by building contextualized manufacturing data hubs grounded in ISA-88, ISA-95, and Unified Namespace, enabling seamless connectivity from the factory floor to enterprise leadership. Anshul transforms both brownfield and greenfield sites by deploying agentic AI solutions and natural-language-enabled co-operators to advance intelligent batch release, streamline nonconformance and corrective action and preventive action management, accelerate deviation resolution, and improve overall equipment effectiveness and yield. He partners with clients to move beyond automation toward adaptive, insight-driven manufacturing enterprises that enable faster, smarter decisions from plant floor to HQ.



Sanjiv Chinnappan is the executive director of architecture and engineering at ZS, bringing over 23 years of experience in enterprise and solution architecture, with a focus on digital transformation in the biopharma sector. He specializes in scaled agile processes and innovation, leveraging frameworks like The Open Group Architecture Framework and platforms such as Amazon Web Services and Azure.



Varun Chadha is a director at ZS with over 12 years of experience in designing and delivering large-scale enterprise data and cloud solutions. He specializes in information management, enterprise integration, and digital transformation initiatives, with deep expertise in big data platforms, cloud architectures, and enterprise search solutions. He focuses on building scalable architectures that align business objectives with long-term technology strategy.



Nirmal Venkatachalam is a data and AI architect with deep expertise in designing and deploying AI, machine learning and optimization solutions across healthcare, industrial products and consumer retail. He has led engagements across diverse business functions, including supply chain, manufacturing marketing, sales, finance, aftermarket and customer service. He has worked across the technology value chain from digital strategy, architecture and design to large-scale implementations across different cloud technologies. He is Claude certified.



About ZS

ZS is a management consulting and technology firm that partners with companies to improve life and how we live it. We transform ideas into impact by bringing together data, science, technology and human ingenuity to deliver better outcomes for all. Founded in 1983, ZS has more than 15,000 employees in over 40 offices worldwide.

Learn more about ZS's supply chain and manufacturing capabilities:
visit www.zs.com or follow us on [LinkedIn](#).

