Hylaine

AI Success Starts with Your Data

AI won't wait, and neither will your executive team. This paper provides **4 practical and actionable approaches** to solve the AI data readiness problem.

Introduction

In the past 3 years—since generative AI tools like ChatGPT showed business leaders how they might easily use AI to automate processes and functions—IT departments have been scrambling to bring AI applications to life. While some organizations are reporting substantial ROI from their AI investments, many more are hitting roadblocks. An MIT study[1] found that about 95% of enterprise AI pilot projects fail to deliver measurable financial returns.

Often, the biggest problem is lack of data readiness. Companies find out too late that the data needed for training machine learning models (much less large language models or agentic AI) is stuck in silos, riddled with errors and redundancies, or simply not available. In one survey, 42% of enterprise tech leaders reported that more than half of their AI projects have been delayed, failed, or underperformed due to data readiness issues[2].

On the other hand, when companies invest in a more robust, AI-ready data architecture, they have the foundation for reliable, repeatable processes to develop AI systems that retain feedback, remember context, and improve over time—and produce positive ROI.

American Express[3], for example, has developed an AI system that looks at transactions from more than 130 million cardholders and 160 million merchants to continuously spot telltale patterns that indicate potential fraud. The AI has not only improved fraud detection, but it has also reduced losses, and enhanced customer loyalty. Astra Zeneca[4] has invested in enhancing its data foundation for AI and is using generative AI across operations, from identifying novel targets to informing clinical trial design and improving efficiency of regulatory submissions. The pressure to build the data systems to enable AI success will only increase. Despite disappointing results, nearly two-thirds of top executives continue to push aggressively for AI adoption, because they fear falling behind competitors, according to a recent IBM report[5].

In this white paper, we look at the common challenges in preparing data for AI and how they can be addressed in cost-effective ways. We also show how companies can develop approaches to efficiently manage future demands for AI-ready data. Based on our work with clients in the banking, insurance, healthcare, and life sciences industries, we find that successful data prep for AI starts with a robust and up-to-date data infrastructure, the engineering capability and tools to stage and certify a tsunami of data for AI, and rigorous governance. Finally, return on investment (ROI) from AI investments depends on adoption. IT must communicate and collaborate with the business to build trust in the data so employees and customers will use the AI. Explainability is essential.

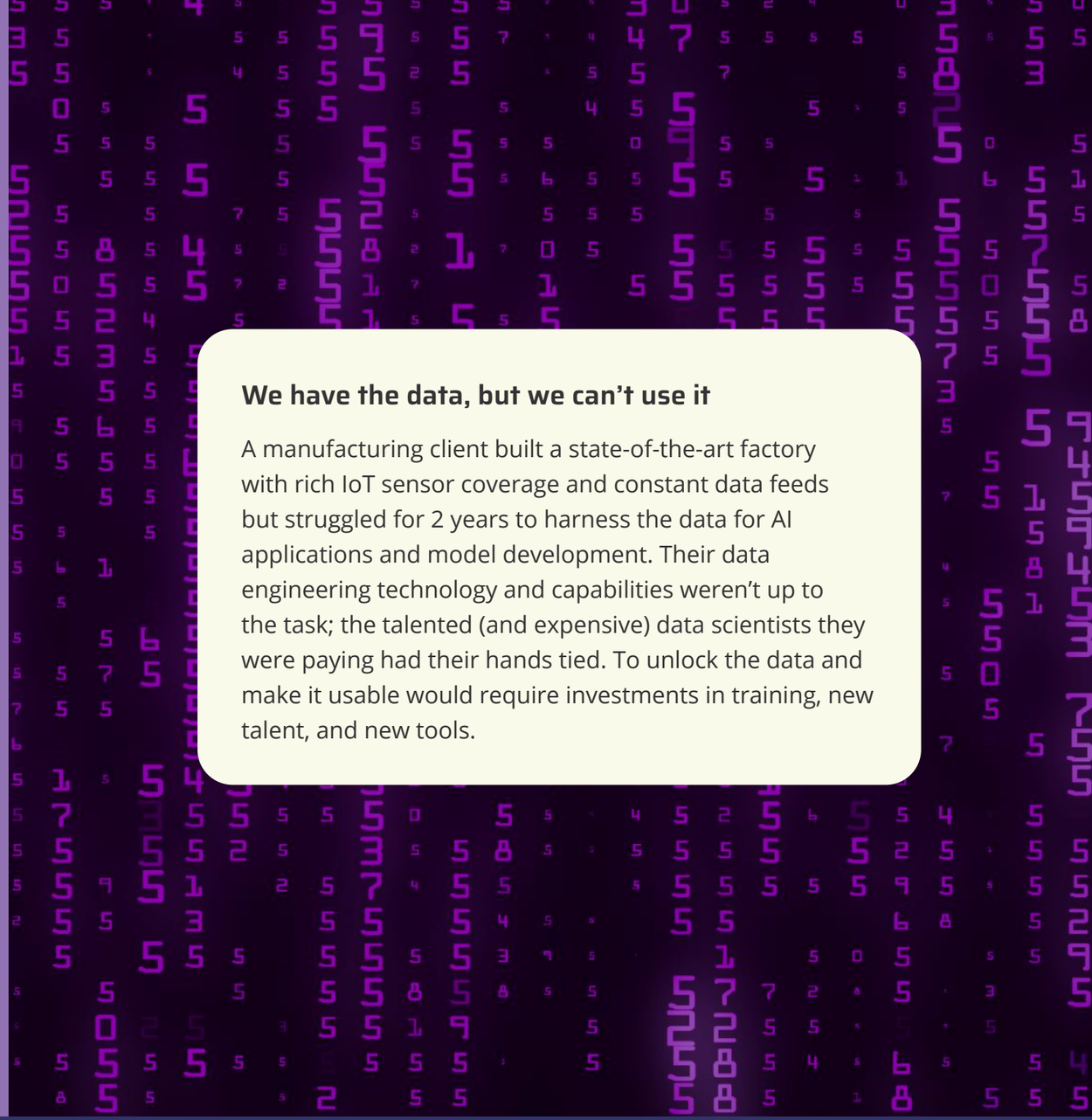
Common Hurdles for Successful Data Preparation

Across industries and types of AI implementations we see common challenges for successful data preparation. We have organized these challenges into 5 broad groups: Data Access, Siloed Data, Data Quality, Governance, and The Human Factor.

Data access: The most obvious hurdle is lack of data access—you know there is data that would enable your AI application but can't get the data you need. It's critical to determine if this is a hard block, where the data simply does not exist, or a soft block, such as security or legal objections that may be resolved through negotiation. Then there are datasets that you would love to get, but they just are not available. For example, an insurer underwriting coverage for medical providers would love to have every malpractice claim for an individual provider, group, or hospital. The problem is that this sort of data does not exist in a sharable way across the industry. Sometimes, valuable data exists in third-party systems, but access will require contract negotiation and ingestion.

Another distinct possibility is that there's particularly juicy data—say records about customers who were cross-sold or up-sold during a helpline call—but the number of records is simply too small for an AI model to reliably answer queries to drive more cross-selling. In such cases, synthetic data can be used to augment the actual data and train the model.

Data access is complicated by format and compatibility issues. AI developers have made huge strides in teaching machines to understand unstructured data—to "read" text, recognize images, listen to speech. Sometimes the data is there, but the organization has not figured out how to organize it and use it. For an AI model to work properly, it needs lots and lots of data, from many sources, in a consistent format.



We have the data, but we can't use it

A manufacturing client built a state-of-the-art factory with rich IoT sensor coverage and constant data feeds but struggled for 2 years to harness the data for AI applications and model development. Their data engineering technology and capabilities weren't up to the task; the talented (and expensive) data scientists they were paying had their hands tied. To unlock the data and make it usable would require investments in training, new talent, and new tools.

Siloed data. This topic has long been a bane of IT departments. Not only is corporate data spread across different systems in different functions and business units, but today data is likely to be housed on multiple cloud platforms, too. Even getting the corporate data for the AI to ingest—say data to build a 360-degree view of the customer profile across multiple lines of business—can be difficult because data remains siloed, or data is not available in a usable format. AI applications may need to pull in data from a number of on-prem and off-prem systems as well as from third-party sources. The complications—and costs—quickly multiply. Say you're building the AI application on the cloud, but your data is on-prem. In many cases, you'll need to move that data up to the cloud for cost-effective analysis, then pay to move it back. Alternatively, if your company has invested in its own GPU hardware for AI work, you will have to pay an "exit fee" to move the data back in-house.

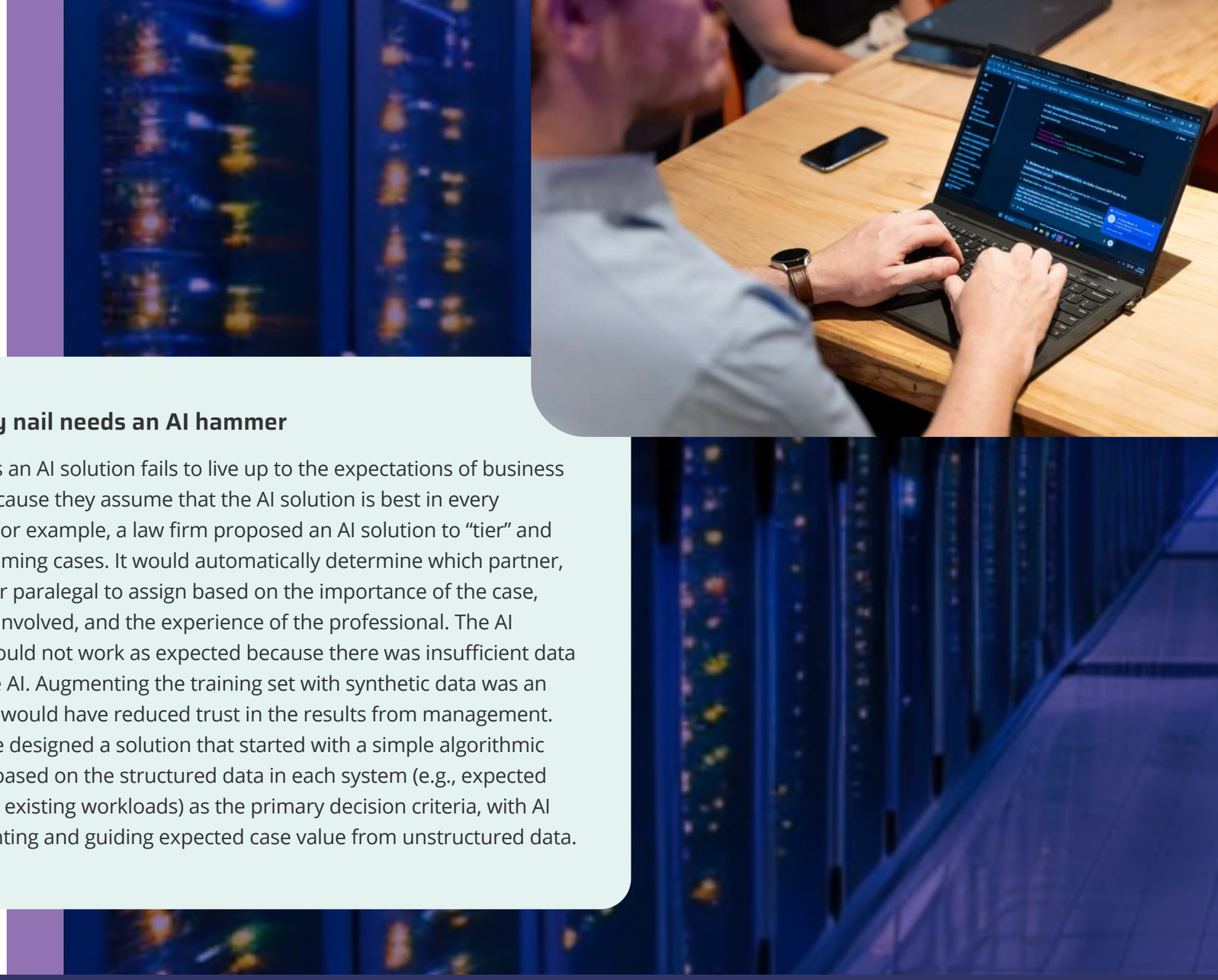
Another common challenge: getting data from critical legacy systems to a location where it can be accessed quickly for training AI. The data needs to be pulled, mapped, and reformatted into modern interchange standards. And it also may need to be augmented—databases on old mainframe systems often use limited data fields. There is no quick and easy way to do this. Developers of AI applications have to be prepared for hard limitations when it comes to ingesting data from these systems—think daily feeds, not real-time.

Data quality. The risks of garbage-in/garbage-out are greater than ever with AI. Even when the data going in is good, an AI system is non-deterministic, meaning there can be more than one answer. If you ask the same question of an AI system twice, you will likely get two slightly different results. This means that shakiness on the left side of this equation (garbage in) is a multiplicative problem.

Inaccurate, stale, incomplete, and redundant data can wreak havoc with model accuracy, causing hallucinations, bias, and bad answers. It's not enough for IT to build the pipes to bring in data from various sources. IT must be able to certify that the data is reliable—including third-party data. Imagine building an AI tool to help business development fine-tune the M&A strategy, only to discover that one of the datasets had inaccurate information about deal multiples in your sector.

Governance. Once clean data is available, there's another stumbling block: Is there risk in using it? Frequently, the data for AI applications can't have PII (Personally Identifiable Information), like social security numbers, confidential medical information, or anything that would allow a model to pick a person out of a crowd or provide details that would enable bias. For example, an insurer developing an AI underwriting tool typically has to remove not just direct PII, but also secondary data indicating that an individual is a member of a protected class. If this precaution is not taken, a well-developed underwriting model might come up with the "right" but illegal answer.

The human factor. Things go sideways in AI projects when the business is not clear about what it needs for its application, where the data resides, or if the data even exists. This problem may arise because IT departments do not communicate effectively what is and is not possible. Projects often die in the pilot phase when it becomes clear that the data can't deliver the answers the business wants—even if the data is complete, squeaky clean, and up to date. In the frenzy not to be left behind, business leaders may be imagining AI capabilities that are impossible or impossibly expensive. IT departments have to make sure that they set expectations correctly with the business at the start, rather than having to scrap AI pilots after time and money have been invested.



Not every nail needs an AI hammer

Sometimes an AI solution fails to live up to the expectations of business leaders because they assume that the AI solution is best in every situation. For example, a law firm proposed an AI solution to "tier" and assign incoming cases. It would automatically determine which partner, attorney, or paralegal to assign based on the importance of the case, the issues involved, and the experience of the professional. The AI solution would not work as expected because there was insufficient data to train the AI. Augmenting the training set with synthetic data was an option but would have reduced trust in the results from management. Instead, we designed a solution that started with a simple algorithmic approach based on the structured data in each system (e.g., expected case value, existing workloads) as the primary decision criteria, with AI supplementing and guiding expected case value from unstructured data.

Take Action on 4 Fronts to Solve the AI Data Readiness Problem

The challenges for preparing AI data are structural, technical, and organizational. So are the solutions. Together the solutions we describe—building a mature data infrastructure, adopting the necessary tools, developing proper governance, and addressing talent and organizational requirements—can help companies build a reliable and repeatable AI development process.

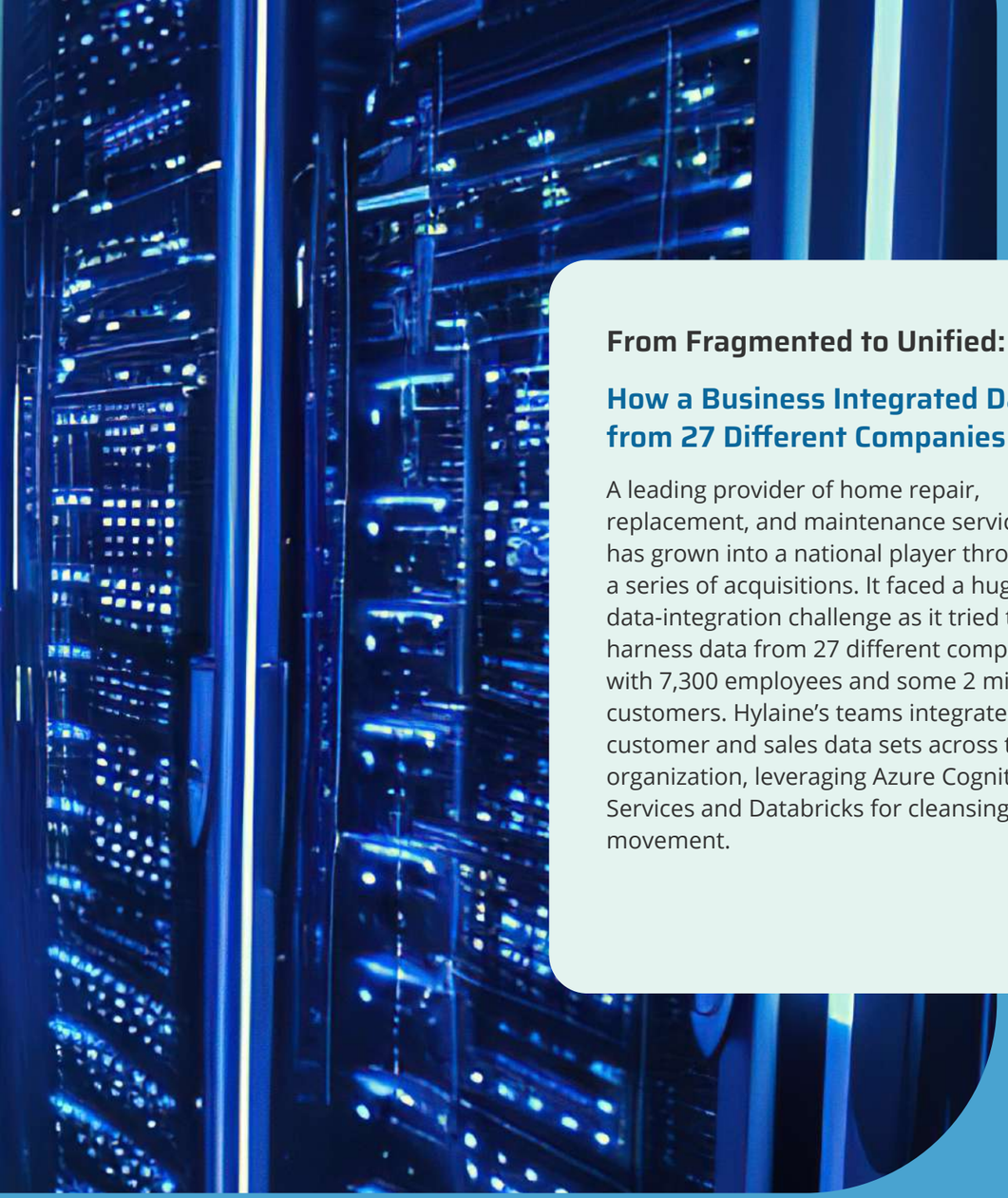
1. Build a mature data infrastructure. Many large organizations have well developed data infrastructures and have invested in data engineering tools and talent—the foundations for a data-driven business. Companies that have both data warehouses (with curated, reliable, and structured data sets) and data lakes (built to accommodate diverse data types) have a head start. But developing and running AI applications requires additional ways of collecting, processing, and storing data at greater scale and velocity. Often, traditional data architecture is not optimized for this.

For example, an AI application for a life sciences company might need to continuously ingest clinical trial results from multiple research sites, patient monitoring data from wearable devices, updates from regulatory agencies, and even information about emerging public health trends—all of which enables the AI to automatically adjust trial protocols, flag safety concerns, and provide researchers with real-time insights. To support such data streaming and real-time analytics, data infrastructure must support event-driven pipelines, stream ingestion, and dynamic transformation. This includes support for RAG (retrieval-augmented generation), a technique for bringing in new data for the AI, to keep the analysis current and avoid model drift and other issues. Organizations with a traditional medallion architecture may need to develop and adopt a parallel architecture to support AI.

Finally, data reliability engineering (DRE) should be a core capability in the data organization, providing strategies and processes for ensuring data quality, availability, observability, testing and root cause analysis of errors.

"Often what this ends up looking like is data engineering. And it's not always sexy. Sometimes it is just building a new pipeline that's moving data from one location to another and cleaning it at the same time. This is work that's been going on for a long time for different reasons, but it's just more urgent now, and having that expertise is important."

—Ryan McElroy, Vice President Technology, Hylaine



From Fragmented to Unified:

How a Business Integrated Data from 27 Different Companies

A leading provider of home repair, replacement, and maintenance services has grown into a national player through a series of acquisitions. It faced a huge data-integration challenge as it tried to harness data from 27 different companies with 7,300 employees and some 2 million customers. Hylaine's teams integrated key customer and sales data sets across the organization, leveraging Azure Cognitive Services and Databricks for cleansing and movement.

2. Adopt necessary technical tools and processes. There is a rapidly expanding toolkit available for companies to acquire, verify, clean, and format data for building AI models and running AI applications. There are off-the-shelf tools for data integration, such as FiveTran or Airbyte, that can streamline many ETL (Extract, Transform, Load) development projects. These fully managed ETL tools are more expensive than traditional technology over the long run, but their cost is justified by excellent standardization and lower investment in upfront development. When they're enabling expensive, high-priority AI systems, their ROI is even easier to justify.

We've found such ETL tools are particularly adept at ingesting data from tediously complex but common systems like ERPs. On the other hand, we've found that some fully managed ETLs may promise more than they can deliver. When working with more complex AI datasets that change often or require rich cleansing and mapping, it's best to use traditional ETL tools. If you're still relying on legacy platforms, it's worth taking a look at the latest versions of what's available such as Azure Data Factory, Snowflake, or Databricks. These newer tools have vastly improved DevOps capabilities, performance, and cloud-first architecture that can reduce costs.

While ETL (Extract, Transform, Load) applies transformations before loading into the destination system, ELT (Extract, Load, Transform) first loads raw data into the data lake or warehouse and then performs transformations within that environment.

Regardless of what ETL or ELT your organization uses, keep in mind there are other approaches for pre-processing and cleaning data, too. These tools often use AI or ML to increase quality and reduce time to deliver. It's often a worthwhile trade for expensive developer time.

Selecting the best tools for your organization—and sticking to a budget that will enable positive ROI—can be a fraught process and may require outside assistance. To make the right choice of vendor, we urge clients to arrange for "bake offs" between competing products using real data and scenarios. Whatever tools you select, it is important to make sure your data infrastructure is flexible, adaptable, and able to accommodate new technologies as they are needed.

3. Develop proper governance. We believe that developing proper governance approaches and mechanisms may have the greatest impact on improving data readiness and ensuring ongoing success of

AI implementations. Risk management is, of course, fundamental—the business needs to know that there are rules in place to protect personal data and prevent unauthorized use of proprietary content or data. We recently helped a health insurer devise a DRE solution that automatically traps PII data and notifies teams before it can be ingested. A simple way to make sure Social Security data numbers are not ingested is to substitute "XXX" for the last three digits in any nine-digit record. Off-the-shelf solutions like Perforce's Delphix offer "continuous compliance" through tokenizing real data. For example, Social Security numbers could be masked by converting them into a different series of digits, so the records could be used without risk of privacy violations.

Companies should think about data governance for AI in broader terms. The governance model should establish the rules of the road for the company's data practices, and keep the data strategy on track through monitoring, auditing, tracking KPIs (including metrics for ROI), and reporting. We often recommend creating a governance council that includes representatives from different parts of the business, as well as IT experts. Once you've created a data governance framework, it's important to note that AI governance is still a separate effort, even if it's tightly connected with overall data governance.

4. Address organizational issues: *talent and adoption.* Data readiness ensures that you have the necessary ingredients to deliver an AI implementation. But AI success requires adoption—if employees don't use the agentic tool to perform a task or customers don't find an AI chatbot useful, the investment will be wasted. As the MIT study confirms, it's repeatable and scalable adoption—not just isolated wins—that drives sustained ROI from AI.

Building trust in the data—leaning into explainability and transparency—can go a long way to ensure adoption. Leadership is also important. AI initiatives seem to succeed when three personas are leading: an executive champion, the relevant business process owner, and the technical lead. Of course, getting the right talent and skills to manage data for AI is also a critical organizational need. Companies may have to train or hire new talent (already in short supply) and call in outside help. For data teams, we've found that the most common skillset gaps exist in DevOps and cloud infrastructure. However, once teams are knowledgeable about the state of the art, they're reluctant to go back to the old ways of working with data.

One effective way to train your employees is to create hybrid teams, where your team works as co-equals with outside experts. This is the best of both worlds: Your team can leverage their deep business knowledge, and the specialists can help you move to AI data preparedness more quickly. Long after the consultants are gone, your company team is able to keep these systems running and delivering valuable business insights.

"The companies that have generated positive ROI from AI investments have been the ones that found ways to make AI initiatives repeatable and scalable across the enterprise. In other words, they have achieved adoption at scale—not just one-off use cases. Companies that treated AI as a strategic program (with executive champions, cross-functional buy-in, and ongoing iteration) have reaped returns, over and over."

—Justin Goff, Director of Technical Delivery, Hylaine

Ready to Get Started?

Many companies that have failed to build the data foundations for successful AI implementations have come up short because they tried to do it all by themselves. In-house development succeeds only about one-third of the time, according to Aditya Challapally, head of the MIT NANDA (Networked AI Agents in Decentralized Architecture), an agentic AI project. A faster and more reliable route to AI success is through partnerships, with tools vendors and outside data infrastructure experts.

The first step is getting an objective assessment of current capabilities and benchmarking against best practice. And, once a roadmap to the right data strategy for AI preparedness is laid out, use a phased approach to increase chances of success and limit risks.

Sources

1. Sheryl Estrada. MIT report: 95% of generative AI pilots at companies are failing. Fortune CFO Daily. August 18, 2025.
2. Fivetran. Fivetran Report Finds Nearly Half of Enterprise AI Projects Fail Due to Poor Data Readiness. Fivetran press release. May 13, 2025
3. American Express. American Express is Accepted at 160 Million Merchants Around the World; Since 2017, Amex-Accepting Locations Have Increased by Nearly 5x. American Express press release. September 9, 2025
4. Renee Iacona, Francis Kendall, Sajan Khosla. Transforming patient outcomes with generative AI. Astra Zeneca blog. July 11, 2024.
5. IBM. IBM Study: CEOs Double Down on AI While Navigating Enterprise Hurdles. IBM press release. May 6, 2025.

Meet the Authors

Ryan McElroy
Vice President of Technology
Hylaine

Justin Goff
Director of Technical Delivery
Hylaine

About Hylaine

Hylaine is a values-first technology consulting firm that stands for partnerships over transactions, doing what's right over what's easy, honesty without exception—no bait and switch—ever, and transparency in everything. We help Fortune 1000 and high-growth enterprises solve problems like outdated tech systems, slow software delivery and time-to-market, and data that's unreliable or scattered. We modernize systems, accelerate software delivery, and drive data accuracy to use AI effectively—and realize extraordinary results.

We're trusted in regulated, data-intensive industries, especially in banking, insurance, healthcare, and life sciences—where compliance, performance, and scalability are non-negotiable. Founded in 2017, Hylaine is headquartered in Charlotte, NC, operates across 5 U.S. regional hubs, and has earned a reputation as a true partner aligned to its clients' success.

Our Expertise, Tech & Tools

We bring deep expertise across modern platforms, tools, and technologies—backed by senior technologists with an average of 15 years' experience. Our approach is platform-agnostic and integration-focused, so we recommend what fits, not what's familiar to us. Wondering if we work with a specific platform or stack? Just ask. Chances are, we already know it well.

Let's Build Something Better Together

At Hylaine, we're setting a new gold standard for the technology consulting industry by redefining what values-led leadership looks like. Since launching in 2017, we believed that true partnership, putting client needs ahead of our own, prioritizing doing what's right, and telling the truth early—even when it's hard—should be the norm, not the exception.

Connect with us and see what better looks like.

hylaine.com

info@hylaine.com

Schedule a Complimentary Discovery Session

HQ: Charlotte | Atlanta | Dallas-Fort Worth | Indianapolis | Raleigh