

AI GUARDRAILS #6

GETTING YOUR DATA HOUSE IN ORDER

THE PRINCIPAL AGENTIC PREREQUISITES

Andy Bochman | Resilience Strategic Lead | West Yost

One thing we've heard from organizations that have attempted to deploy agentic AI in their organizations, especially those that have experienced initial setbacks, is that the preparation of their data assets is the single most important determinant of success. This is painstaking, unglamorous work, but it's essential and pays benefits that go far beyond supporting successful internal AI launches. In the conversation below with West Yost Data Scientist and Data Strategic Lead, Micah Krider, he unpacks what to do, how to do it at a high level, and the multifold benefits that accrue to utilities ready to bring order to the chaos. As utilities explore the opportunities presented by AI, many are discovering that one of the greatest long-term investments they can make is strengthening the way they manage their data. AI promises many benefits, but perhaps one of its most valuable contributions is accelerating attention to data management practices that improve operational efficiency, decision making, and organizational resilience, regardless of whether AI is ultimately deployed.

Q1: To get started, what does the phrase "get your data house in order" mean to you?

MK: To me, "getting your data house in order" starts with having Good Data. Without good data, organizations have less visibility into their operations and fewer reliable insights to inform decision making and continuous improvement. Technologies such as machine learning, artificial intelligence, and agentic systems depend on the same foundation, and when data quality is poor, they are far more likely to produce unreliable or misleading insights. Incorrect insights can be devastating to operations, and working with bad data wastes time, which ultimately wastes money.

My philosophy is simple: **Good Data + Good People + Good Tools = Success**. All three elements matter, but good data is the foundation that enables the other two.

I define good data by five key characteristics: it must be Clean, Normalized, Aggregated, Abundant, and Available.

Clean data is data that has been reviewed for errors, inconsistencies, duplicates, corruption, and other quality issues. Cleaning data is often the most time-consuming

and expensive phase of a data initiative, yet it is also the most critical. Whether data originates from human entry, sensors, or multiple source systems, organizations must identify and correct problems rather than simply discard data. Just as importantly, they must put processes in place to keep data clean by guiding users toward accurate and consistent data entry.

Normalized data means data is collected and defined consistently across the organization. Organizations must first determine what information is truly important to capture, then establish common standards for collecting and defining that information across departments, projects, and systems. Standardized formats, definitions, and collection methods ensure that data from different sources has the same meaning and can be compared reliably.

Aggregated data brings information together from multiple systems into a centralized location where it can be efficiently analyzed. Most organizations store information across numerous applications, spreadsheets, databases, and operational systems. Aggregating that data into a consolidated environment, such as a data warehouse, allows people and tools to work from a single source of truth rather than spending time searching for and reconciling information across multiple locations.

Abundant data means having sufficient information to identify meaningful patterns and trends. Small datasets can be useful, but larger collections of clean, normalized, and aggregated data create greater opportunities for deeper analysis and more reliable insights. Whether through traditional reporting, statistical analysis, machine learning, or artificial intelligence, abundant data allows organizations to uncover inefficiencies, identify anomalies, and continuously improve their operations.

Available data is data that can be accessed and used when it is needed. Data loses much of its value if it is locked away in isolated systems, hidden in spreadsheets, or difficult for staff to obtain. Information must be accessible to the right people at the right time, while remaining appropriately governed and protected, so that it can support decision making and drive action.

When data is Clean, Normalized, Aggregated, Abundant, and Available, organizations establish the foundation needed to improve operations, strengthen decision making, and successfully adopt advanced technologies such as AI.

Q2: From your experience at Jacobs and now at West Yost, when you first begin working with an organization, how do you assess the maturity of its data management practices, and what are some of the common patterns you find?

MK: *In my experience, the best way to assess an organization's data maturity is through structured interviews with people across the organization, combined with a survey that captures how employees perceive the organization's data capabilities.*

I view these as two complementary exercises. The interviews reveal how data is actually collected, managed, shared, and used. The survey captures how people believe the organization is performing. Comparing those two perspectives often reveals opportunities that might otherwise be overlooked.

Whether I was working at Jacobs, West Yost, or with utility clients, I found that the same fundamental questions tend to apply regardless of organization size or industry. I typically begin by speaking with leaders and subject matter experts to understand what the organization does, what information it collects, where that information originates, how it is stored, how quality is maintained, what tools are used, how data is shared, and what challenges people encounter in their daily work.

A key part of the interview process is asking about work products, not just data. Many people do not naturally think in terms of data flows, databases, or system architecture. They think in terms of the things they are responsible for producing: models, maps, assessments, reports, dashboards, regulatory documents, planning documents, or decisions. Starting with work products helps make the conversation concrete. Once you understand what someone produces, you can work backward to identify what data supports it, where that data comes from, how it is collected, how it is stored, what tools are used, how quality is checked, and where the process breaks down.

I also spend significant time understanding processes. Organizations often focus on the technology they own, but technology is rarely the primary issue. More commonly, I find inconsistent processes, duplicate data collection efforts, unclear ownership, disconnected systems, and heavy reliance on spreadsheets and manual workarounds. These challenges often develop gradually over many years as individual teams solve problems independently without a broader data strategy.

One of the most common observations is that different groups frequently have different versions of the same information, each believing their version is correct. When that occurs, discussions become centered on reconciling data rather than using data to make decisions. Another common finding is that staff spend substantial amounts of time manually collecting, cleaning, and reformatting information before they can perform any meaningful analysis.

The interview process helps uncover these realities while also identifying opportunities for improvement. By understanding the people, processes, data sources, tools, reporting needs, and pain points across the organization, it becomes possible to develop a roadmap that improves data management while supporting the organization's operational goals.

Q3: You've been following the evolution of AI from machine learning to generative AI and now agentic AI. From your perspective, what's the relationship between the quality of an organization's data and the quality of the results AI can produce?

MK: *There is a direct relationship between the quality of an organization's data and the quality of the results produced by artificial intelligence. AI can only learn from and reason about the information it is given. If the underlying data is incomplete, inconsistent, poorly documented, or incorrect, the resulting insights will reflect those same shortcomings.*

This is one of the reasons I place such a strong emphasis on getting your data house in order before pursuing AI initiatives. The same characteristics that define good data—clean, normalized, aggregated, abundant, and available—also create the foundation for successful AI adoption.

Today, AI is very good at finding patterns, summarizing information, identifying anomalies, and helping people make sense of large volumes of data. However, it still struggles when data lacks context. If information is scattered across multiple systems, uses inconsistent terminology, contains quality issues, or lacks metadata that explains what the data means, AI often has difficulty producing reliable results.

AI is becoming increasingly capable of helping organizations clean, classify, and organize their data. While these tools continue to improve, they still require human oversight and organizational context. AI cannot reliably infer business meaning, relationships, or operational intent from disconnected datasets on its own.

For that reason, I view data readiness as a prerequisite for AI readiness. Organizations that invest in understanding, standardizing, documenting, and making their data accessible will be far better positioned to realize the benefits of AI. In my experience, AI works best when it is applied to good data, not when it is asked to create good data from chaos.

Q4: Is there anything you've seen with water utility data policies and processes that strikes you as distinctly different from other domains you've worked with closely, including military?

MK: *One of the biggest differences I have observed between the water sector and other environments I have worked in, particularly the Department of Defense, is the level of organizational maturity around data management and cybersecurity.*

In many water utilities, I find that organizations generally have confidence in both their data quality and data security. However, a closer look often reveals opportunities to improve data governance, strengthen data quality, and refine security practices. Organizations often believe they have good data because they have been collecting it for years, but closer examination may uncover inconsistencies, duplicates, missing information, or undocumented business rules that make analysis difficult. Similarly, organizations frequently believe their information is secure because it resides on internal systems or SharePoint sites, only to discover that permissions are broader than intended or that sensitive information is more accessible than they realized.

This is not a reflection of a lack of commitment or professionalism. Rather, it is often the result of limited resources. Most utilities are focused on delivering safe and reliable water service, not maintaining large data governance or cybersecurity teams. As a result, data management practices and security controls are frequently developed over time through practical necessity rather than through a comprehensive enterprise strategy.

In the Department of Defense, by comparison, there are extensive policies, dedicated personnel, significant funding, and strict controls governing how information is collected, classified, stored, transmitted, and protected. Data governance and cybersecurity are treated as mission-critical functions and are resourced and supported accordingly.

Water utilities face many of the same risks but often with far fewer resources. Many operational technology systems were originally designed with reliability and ease of operation as the primary objectives, omitting today's heightened cybersecurity requirements. As utilities modernize these systems, balancing operational reliability with evolving security requirements becomes increasingly important.

As water utilities continue to modernize and adopt AI, advanced analytics, and greater system connectivity, I believe data governance and cybersecurity will become increasingly important. Before organizations can fully trust the outputs of advanced technologies, they must first be confident that the underlying data is accurate, well-managed, and appropriately protected.

Q5: Many utilities are interested in AI but want to move thoughtfully. If they focus first on improving their data and governance, what value can they expect to realize even before deploying AI?

MK: *One of the biggest misconceptions is that getting your data house in order only matters if you're planning to deploy AI. In reality, many of the benefits are immediate, regardless of when or even whether AI is adopted.*

First, the utility will be much better positioned when it decides to adopt AI. By investing in clean, normalized, aggregated, abundant, and available data now, the organization increases the likelihood that future AI initiatives will be successful, without having to address years of accumulated data issues during implementation.

More importantly, many of the benefits of good data can be realized without AI at all. Organizations with well-managed data typically experience improved reporting accuracy, faster access to information, and greater confidence in decision making. Staff spend less time searching for information, reconciling conflicting reports, and manually manipulating spreadsheets, allowing them to focus more on analysis and decision making. When everyone is working from the same trusted information, confusion is reduced and discussions can focus on solving problems rather than debating whose numbers are correct.

Getting the data house in order also improves operational efficiency. Standardized collection processes, clear ownership, documented definitions, and improved data quality controls help ensure that information is captured correctly the first time. This reduces rework, minimizes errors, and allows staff to spend more time using information rather than correcting it.

There are also governance and security benefits. Organizations gain a better understanding of what data they have, where it resides, who has access to it, and how sensitive information should be protected. This improves compliance, reduces risk, and strengthens organizational trust in the data itself.

In many ways, the benefits of getting the data house in order are immediate, while the benefits of AI are often future-oriented. Utilities that invest in data quality, governance, accessibility, and standardization today will see improvements in efficiency, reporting, and decision making regardless of whether and when they choose to adopt AI. If they do decide to move forward with AI, they will be doing so from a position of strength rather than trying to build the foundation while simultaneously deploying the technology. Ultimately, organizations that invest in their data today are making an investment in better decisions, more efficient operations, and greater resilience, regardless of where they are on their AI journey.

Comments and questions: resilience@westyost.com