A Data Strategy for the AI World

John W. Moran, Ph.D.

Senior AI and Quality Advisor, Public Health Foundation

December 2025

Introduction

Organizations are sitting on heaps of data in emails, spreadsheets, reports, meeting notes, and more. To fully leverage this data in the Artificial Intelligence (AI) world, organizations need a clear data strategy. A data strategy is a comprehensive plan that outlines how an organization will collect, manage, analyze, and use its data to achieve specific strategic and business goals. It serves as a blueprint that connects people, processes, and technology to support better decision-making and improve operational processes.

A successful data strategy aligns data initiatives with overall objectives and includes key aspects such as data quality, governance, security, and the right technology infrastructure. With the advent of AI and the need to collect and prepare data for Large Language Models (LLMs), having an effective data strategy is now a top priority for organizations. That, in turn, requires a renewed emphasis on the importance of data engineering across the organization.

Why Do We Need a Data Strategy?

The goal of a data strategy is to make relevant information available to stakeholders across the organization so they can use it to drive decisions. Achieving this requires actively breaking down information silos between business units and adopting uniform policies for data types, storage architectures, and workloads.

One of the biggest challenges organizations face is reducing data silos, as shown in **Figure 1**. Over time, silos develop across different business units, work processes, architectures, languages, and tools. These disconnected data silos make it difficult to realize the full potential of AI. Integrating siloed data can be complex and may require costly resources to support the transformation.

Public health departments collect some type of personally identifiable, proprietary, or regulated information, making strong, unified governance policies essential to protect that sensitive data.

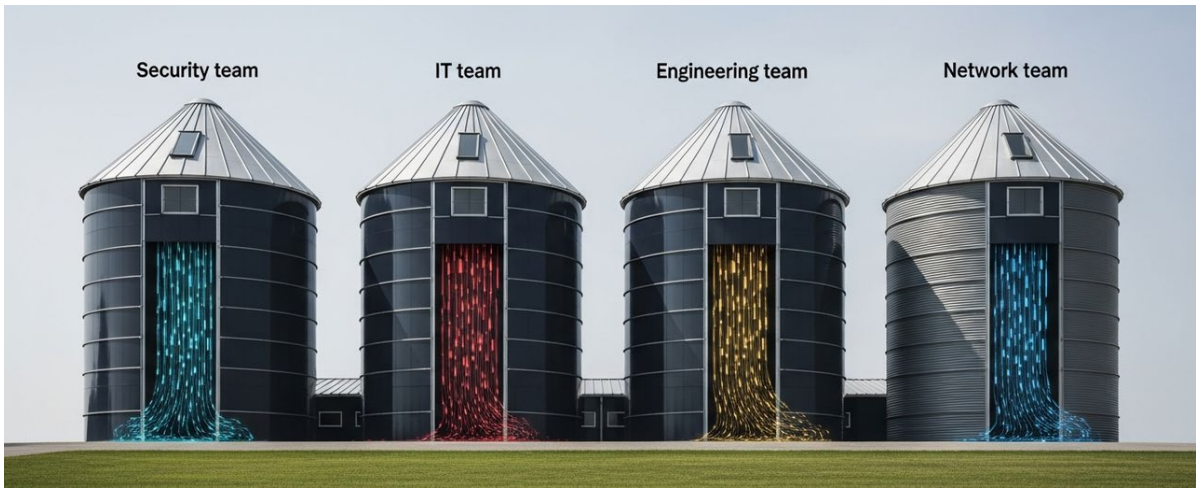

Figure 1

What Should a Data Strategy Accomplish?

A data strategy should ensure that business users have the information they need, when they need it, to make decisions that are best for the organization.

It should allow for seamless data sharing between internal and external partners, with proper safeguards in place to prevent data leaks or unauthorized access. It should also enable better performance when accessing data and provide greater resilience during business disruptions.

A good data strategy should enable an organization to know:

1. Where their data originated – This helps with lineage tracing and error debugging.
2. Who has access to which data and for what purpose – Clear visibility into permissions and usage is essential.
3. How the data have changed over time – Tracking changes helps with accuracy and accountability.
4. What the data quality is – Low-quality data can result in faulty business decisions, potentially leading to increased costs, missed market opportunities, lost revenue, etc.
5. How data are cataloged – Data storage is only as good as its organizing principles, which help users find the data they need and evaluate its relevance for their intended use.
6. Who has access privileges – Data governance policies should clearly define who can access, manipulate, or change any element of data within a repository.

7. How to maintain a comprehensive audit trail – This should identify what changes were made to the data, who made them, when they occurred, and which applications those changes may impact. This can reduce the potential for unauthorized alterations and errors.

There are data cleaning programs for transforming messy data into reliable assets. Tools such as Informatica PowerCenter, OpenRefine, Trifacta, and Talend can help transform messy data into reliable assets.

Summary

The purpose of a comprehensive data strategy is to develop a reliable framework that the organization can rely upon to make good decisions.

Trusted data should be:

- Accurate
- Consistent
- Relevant
- Rich in context

What's at stake if you lack a trusted data foundation?

- Potential for biased practices
- Inaccurate guidance
- Flawed decisions guided by AI
- Inability to safeguard sensitive information and ensure compliance with regulations and internal policies
- Once untrusted data is in the AI model, it is hard to get it out!

Remember, AI is only as powerful as the data behind it. *Power your applications with Trusted Data.*