



BEYOND THE DATA LAKE

The Big Data Promise and the Governance Deficit — Why Organisations Are Building Lakes They Cannot Navigate

A Practitioner's Perspective on Data Ambition Outpacing Data Discipline in Large Enterprises

Giovanni Leonardi

January 2013

Contents

1. Executive Summary	3
2. The Seduction of Volume	3
3. The Governance Gap Is Not New — But Big Data Has Made It Visible	4
4. Why the Pattern Persists	5
4.1 The Incentive Asymmetry	5
4.2 The Organisational Ownership Problem	5
4.3 The Skills Mismatch	5
5. The Data Lake as Organisational Mirror	6
6. What Good Looks Like — and Why It Is Rare	6
7. The Deeper Pattern	7
8. Where This Leaves Us	7

The Big Data Promise and the Governance Deficit — Why Organisations Are Building Lakes They Cannot Navigate

Transformation · No. 212

Originally written January 2013 | Re-edited and re-rendered for the WhereBeyond Edition, 2025

“The organisations investing most heavily in data are, paradoxically, the ones least able to tell you what data they actually hold.”

1. Executive Summary

Across sectors, large organisations are committing significant capital to big data programmes — investing in Hadoop clusters, hiring data scientists, and building what the industry has started calling data lakes. The promise is compelling: unprecedented volumes of operational, transactional, and behavioural data will unlock competitive insight, drive efficiency, and transform decision-making. Yet the pattern that is emerging, repeatedly, is one of ambitious data infrastructure deployed into organisations that lack the governance, quality frameworks, and operating disciplines to use it. The result is not transformation but frustration — expensive platforms that ingest everything and deliver little, surrounded by data scientists who spend most of their time cleaning data rather than analysing it. This essay examines why the gap between data ambition and data governance has become structural, what forces sustain it, and what it reveals about a deeper pattern in how organisations approach transformation.

2. The Seduction of Volume

The big data narrative, as it has gained momentum over the past two years, carries an implicit assumption that has gone largely unchallenged: that the primary constraint on data-driven decision-making is the volume and variety of data available. If only we could capture more data, store it more cheaply, and process it more quickly, the argument runs, then insight would follow.

This assumption has driven a wave of infrastructure investment that is remarkable in its speed and scale. Organisations that five years ago struggled to justify a data warehouse upgrade are now commissioning Hadoop environments, experimenting with MapReduce, and standing up data science teams. The technology vendors have been effective evangelists, and the early adopter case studies — largely from born-digital

companies with fundamentally different data cultures — have created a powerful sense of urgency.

But the assumption is wrong, or at least dangerously incomplete. The binding constraint on most organisations' ability to use data is not volume but discipline. It is not the absence of a data lake but the absence of a shared understanding of what the data means, where it comes from, who is responsible for it, and whether it can be trusted.

In my experience, the organisations investing most heavily in data are, paradoxically, the ones least able to tell you what data they actually hold. They can point to petabytes of storage. They cannot point to a reliable, agreed definition of what constitutes a customer, a transaction, or a product across their own business units.

3. The Governance Gap Is Not New — But Big Data Has Made It Visible

Data governance has been a recognised discipline for at least a decade. Master data management, data quality, and metadata management are not new concepts. What has changed is the consequence of ignoring them.

In the era of traditional data warehousing, poor governance manifested as reconciliation problems between reports, protracted ETL development cycles, and periodic data quality crises that were contained within the business intelligence team. These were painful but bounded. The data warehouse, by its nature, imposed a degree of structure — data had to be modelled, transformed, and loaded through controlled processes. The structure was expensive and slow, which was the complaint that drove the big data movement, but it also forced a minimum standard of curation.

The data lake inverts this discipline. Its foundational principle is that data should be ingested in its raw form, without prior transformation, and that structure should be applied at the point of use rather than the point of capture. This is technically elegant and economically attractive. It is also, in an organisation without strong governance, an invitation to create a data swamp.

The pattern I have observed across multiple large enterprises is consistent:

- Data is ingested from dozens of source systems with minimal documentation of lineage, meaning, or quality
- No authoritative business glossary exists, or if one does, it is maintained by a small team and ignored by the data engineering function
- Data scientists are hired at considerable expense and then spend sixty to eighty per cent of their time on data preparation — finding, understanding, cleaning, and reconciling data before any analysis can begin
- Business users, promised self-service analytics, find that the data available to them is incomprehensible without deep technical knowledge of the source systems
- The data lake grows in size but not in utility, and the organisation's actual decision-making continues to rely on the same spreadsheets and tribal knowledge it always has

4. Why the Pattern Persists

If the governance deficit is so predictable, why does it persist? The answer lies in a set of structural forces that are easier to describe than to overcome.

4.1 The Incentive Asymmetry

Data governance is a classic example of a discipline where the costs are immediate and visible but the benefits are diffuse and delayed. Building a business glossary, establishing data stewardship roles, implementing data quality rules — these are slow, unglamorous activities that require sustained organisational commitment. They do not produce a demonstration that can be shown to the board. They do not generate the kind of excitement that a proof-of-concept predictive model does.

Big data infrastructure, by contrast, is tangible. A Hadoop cluster can be commissioned. A data science team can be recruited. A proof of concept can be delivered in weeks. The technology creates the appearance of progress even when the underlying data problems remain unresolved.

The result is a persistent bias towards infrastructure investment over governance investment. Organisations will spend millions on platforms and pennies on the disciplines needed to make those platforms useful.

4.2 The Organisational Ownership Problem

Data governance requires clear ownership, and clear ownership of data is something most large organisations have never achieved. Data is generated by operational processes owned by one function, stored in systems managed by another, and consumed by a third. The question of who is responsible for the quality, meaning, and lifecycle of a given data entity — a customer record, a product code, a financial transaction — typically has no clear answer.

The appointment of Chief Data Officers is beginning to address this in some organisations, but the role is new and its authority is often ambiguous. In practice, data governance initiatives frequently stall because no single function has the authority to impose standards across organisational boundaries, and the executive sponsor either lacks the patience for the work or is distracted by the more visible data infrastructure programme.

4.3 The Skills Mismatch

The market for data scientists has become intensely competitive. Organisations are hiring quantitative analysts, machine learning specialists, and statisticians at premium salaries. What they are not hiring, in anything like the same numbers, are data architects, data stewards, and information managers — the people who build and maintain the governance infrastructure that data science depends on.

This is not a failure of recruitment strategy alone. It reflects a deeper misunderstanding of what data-driven transformation actually requires. The popular narrative emphasises algorithms and analytics. The less exciting reality is that algorithms are only as good as the data they consume, and that data quality is an organisational discipline, not a technical one.

5. The Data Lake as Organisational Mirror

There is a revealing pattern in how organisations describe their data lake strategies. The language is almost always technological: platforms, clusters, processing frameworks, ingestion pipelines. Rarely does the strategy articulate the governance model, the data ownership structure, or the quality standards that will apply.

This is not a technology problem. It is an organisational problem expressed through technology. The data lake becomes a mirror of the organisation's existing data culture — and in most large enterprises, that culture is one of fragmentation, duplication, and ambiguity.

Consider the typical large bank, insurer, or utility. These organisations have grown through merger and acquisition, through decades of system proliferation, and through organisational restructuring that has repeatedly redrawn the boundaries between business units without rationalising the data that flows across them. They do not have one customer database but dozens. They do not have one product hierarchy but several. The definitions of apparently simple concepts — what counts as revenue, what constitutes an active account, when a customer relationship begins and ends — vary across functions and sometimes across teams within the same function.

A data lake built on this foundation does not resolve the fragmentation. It replicates it at scale, in a less structured environment, with less visibility into the contradictions.

6. What Good Looks Like — and Why It Is Rare

The organisations that are making genuine progress with data — and they do exist, though they are fewer than the conference presentations would suggest — share certain characteristics that are worth noting.

- They treat data governance as a business discipline, not a technology initiative, and they fund it accordingly
- They have established clear data ownership at the business level, with named data stewards who have both the authority and the accountability to maintain quality standards
- They invest in a business glossary and metadata management before they invest in a data lake — or at least in parallel with it
- They measure data quality systematically and report it with the same rigour they apply to financial metrics
- They resist the pressure to demonstrate quick wins at the expense of sustainable foundations

These organisations are rare because the approach they take requires patience, sustained executive commitment, and a willingness to invest in work that is invisible to most stakeholders. It requires leaders who understand that the value of data is not in its volume but in its trustworthiness.

7. The Deeper Pattern

The big data governance gap is, in the end, a specific instance of a much broader pattern in how organisations approach transformation. The pattern is this: organisations consistently invest in the visible, tangible, technology-enabled elements of change while underinvesting in the organisational disciplines, cultural shifts, and governance structures that determine whether the technology delivers value.

We see it in enterprise resource planning programmes that deploy the platform but never achieve the process standardisation it was designed to enable. We see it in customer relationship management implementations that automate existing dysfunction rather than transforming customer engagement. And we see it now in big data, where the infrastructure is being built at pace but the organisational capability to use it is lagging far behind.

The lesson is not that big data is overhyped — the analytical potential of large-scale data processing is real and significant. The lesson is that technology capability without organisational capability is not transformation. It is expensive infrastructure.

The binding constraint on most organisations' ability to use data is not volume but discipline — not the absence of a data lake but the absence of a shared understanding of what the data means, where it comes from, and whether it can be trusted.

8. Where This Leaves Us

As we enter 2013 with big data investment accelerating across sectors, the organisations that will extract genuine value are those that resist the temptation to lead with technology. They will ask the unglamorous questions first: do we know what data we have? Do we agree on what it means? Do we know who is responsible for it? Can we trust it?

These are not exciting questions. They do not feature in vendor presentations or analyst reports. But they are the questions that determine whether a data lake becomes a strategic asset or a very expensive swamp.

The practitioners who have been through previous waves of technology-led transformation — ERP, CRM, business intelligence — will recognise the pattern. The technology is new. The organisational challenge is not. And until we learn to address the organisational challenge with the same energy and investment we bring to the technology, we will continue to build platforms that promise transformation and

deliver disappointment.

*“The technology is new. The organisational challenge
is not.”*
