



BEYOND THE DATA LAKE

Big Data's Unfunded Twin — Why Data Governance Keeps Losing the Argument

An essay on why the distance between big data's promise and data governance's reality is designed in — not something organisations simply mature past.

Contents

1. Executive Summary	3
2. The two ledgers	3
3. The seduction we should take seriously	4
4. What the discipline actually holds together	5
5. Why the gap is structural	5
6. The strongest case for building first	7
7. What the pattern tells us about transformation	8

Big Data's Unfunded Twin — Why Data Governance Keeps Losing the Argument

Transformation · No. 211

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

“An immaturity is something you grow out of. A structure is something that reproduces itself until someone changes the structure.”

1. Executive Summary

Every organisation now has a big-data ambition and almost none has a governance ambition to match it. The imbalance is not an accident of immaturity that time will correct; it is a structural feature of how these programmes are funded, staffed, and rewarded. The promise — that abundant data, cheaply stored and rapidly processed, will yield insight and advantage — is genuine, and worth taking seriously on its own terms. The reality is that the abundance arrives faster than the discipline needed to make it trustworthy, and the discipline is the part no one wants to pay for. This essay argues that the gap between the two is designed in: the same incentives that make big data attractive to sponsors are the incentives that starve governance. It examines why the argument for governance keeps losing even when everyone agrees, in principle, that it should win; it takes seriously the strongest case for building capability first and governing later; and it treats the whole pattern as a small, sharp instance of a much larger truth about why transformation intent and transformation reality so rarely meet.

2. The two ledgers

A programme is six months in. The cluster is running; commodity storage has swallowed years of transaction logs, clickstreams, and machine output that the old warehouse could never have held. The dashboards multiply weekly. A newly hired analyst — one of the scarce, expensive kind the trade press keeps calling the most sought-after hire of the decade — is asked a plain question by the board: how many customers do we have?

The answer takes eleven days to produce and arrives in three versions. The finance warehouse says 4.1 million. The billing system says 3.6 million. The new platform, which ingests both and more, says 5.2 million, because it has been faithfully counting every identifier it was ever handed and has no way of knowing that a person, a household, and an account are not the same thing. Nobody in the room is surprised. Everybody in the room is quietly appalled.

This is the moment the two ledgers of a data programme become visible at once. On the first ledger is everything that was built: the storage, the pipelines, the processing, the visible and celebrated machinery of having the data. On the second ledger is everything that would make the first one mean something: the agreement on what a customer is, the lineage that lets you trust a number, the ownership that lets you fix it when it breaks. The first ledger is full and growing. The second is thin, and — this is the part that matters — it was never seriously funded to be anything else.

We should resist the temptation to read this as a failure of competence. The people who built the platform did what they were asked to do, and did it well. The gap is not between good engineers and bad ones. It is between the thing that was bought and the thing that was needed, and those two were never the same purchase.

3. The seduction we should take seriously

It is easy, from the governance chair, to be sniffy about the enthusiasm for volume. That is a mistake, because the enthusiasm is well founded. The case for big data is not a fashion; it is a real shift in what is economically possible, and any honest essay has to grant it its full strength before arguing with it.

For most of the history of enterprise information, storage was expensive enough that the first question about any dataset was what to throw away. Schemas were designed in advance, at the point of capture, precisely because you could not afford to keep everything and decide later. That constraint has largely dissolved. When storage is cheap and processing can be thrown at data in parallel across ordinary machines, the old discipline of decide the structure first inverts into keep everything, impose meaning on the way out. The fashionable name for the destination — the lake into which every stream is allowed to run, untreated — captures the appeal exactly: hold the raw material at full fidelity, and let a thousand later questions find their own answers in it.

The prize is not imaginary. Patterns that were invisible in sampled, pre-aggregated, schema-bound data become visible when the grain is fine enough and the history long enough. The influential studies of the preceding years put real numbers against the prize — productivity and margin advantages for the organisations that learn to compete on analytics rather than on instinct — and every executive has read at least the summary. The fear underneath the enthusiasm is just as real: the sense that a competitor who masters this will see the market a quarter before you do, and that the cost of being late is not a missed project but a structural disadvantage.

So the sponsor who fights for the platform is not a fool chasing a trend. The seduction is rational. That is exactly what makes it dangerous, because a rational, well-defended enthusiasm crowds out the unglamorous work that would make it pay.

The promise of big data is not overstated. The problem is that everything which makes the promise true rather than merely possible sits on the second ledger — and the second ledger is the one no sponsor is fighting for.

4. What the discipline actually holds together

Ask what governance is for and you get a weary catalogue: policies, councils, standards, stewardship. The catalogue is why the discipline loses arguments — it sounds like administration, and administration is the first thing a hungry programme cuts. It is worth restating what the discipline actually holds together, because stated plainly it is not administration at all; it is the difference between data and trustworthy data.

- **Definition.** Somebody has decided, and written down, what a customer is — and a product, an account, an active user. Without this the same word means three things in three systems and every cross-system number is a negotiation.
- **Lineage.** For any figure on any dashboard, you can trace where it came from, what was done to it, and when. Without this you cannot answer the only question that matters about a number under challenge: can I trust it?
- **Ownership.** Some named person is accountable for the quality of a given domain of data — not for the servers it sits on, but for whether it is right. Without this, errors have no home and no one to fix them.
- **Quality.** There is a known, measured standard for completeness, accuracy, and timeliness, and a way of knowing when it slips. Without this, decay is invisible until a number embarrasses someone in public.

None of this is exotic. The body of practice for it was mature well before the fashion for volume arrived; the reference frameworks had been written, the roles named, the disciplines codified. The knowledge is not missing. What is missing is the will to fund the knowledge, and that is a different problem entirely — a problem of structure, not of understanding.

Consider the counting failure from the opening again. Every one of those three numbers was correct against its own system. The failure was not in the arithmetic; it was in the absence of an agreed definition sitting above all three, owned by someone, with the authority to say which one the board should believe. That agreement is cheap to describe and expensive to establish, because establishing it means asking three parts of the business to give up their private version of the truth. The platform did not create that problem. It industrialised it — took a latent disagreement and reproduced it at machine speed across every downstream report.

5. Why the gap is structural

Here is the heart of it. The reason governance is under-resourced is not that leaders fail to understand its value. Most of them understand it perfectly well and will say so, sincerely, in any meeting where it is raised. The reason is that the incentives around a data programme are systematically asymmetric, and the asymmetry runs against the discipline every time.

Dimension	The platform (first ledger)	The discipline (second ledger)
Visibility	A cluster, a dashboard, a demo you can show a board	Definitions and lineage no one can see working
Time to credit	Weeks — the demo lands this quarter	Years — trust compounds slowly and silently
Failure mode	Loud and rare — the platform is down	Quiet and constant — the number is subtly wrong
Ownership	Obvious — the technology function builds it	Contested — it belongs to the business, which did not ask for it
Budget line	Capital: an asset, capitalised and celebrated	Operating cost: a tax on someone else's project

Read that table as a single sentence and the structure becomes plain. The platform is visible, fast, ownable, and capitalisable. The discipline is invisible, slow, contested, and expensed. In any competition for finite attention and money, the first profile wins — not because anyone chose to neglect governance, but because every gravitational force in the organisation pulls the other way.

Three of those forces deserve naming, because they recur with almost boring reliability.

1. The demo beats the definition. A sponsor can stand in front of a steering committee and show a working platform. No one has ever received applause for a data dictionary. The reward system runs on the demonstrable, and governance is, almost by nature, the work whose success looks like nothing going wrong — the least demonstrable outcome there is.
2. The cost lands on the unwilling. The platform is built by the technology function, which wants to build it. Governance has to be done by the business — the people who own the meaning of the data — and to them it arrives as an unfunded obligation attached to someone else's initiative. The party that must do the work is not the party that wanted the programme.
3. The accounting favours the asset. A platform can be treated as a capital investment: a thing, on a balance sheet, depreciating respectably. The ongoing labour of stewardship is an operating expense with no asset to show for it. One of these is easy to approve. The other is easy to cut.

We are fluent, as a profession, in the language of the first ledger and nearly mute in the language of the second. We can specify a platform to the decimal place and cannot write a compelling business case for a definition. This is not a personal failing of any leader; it is a property of the system they operate inside. And because it is a property of the system, it does not dissolve as the organisation matures. A more mature organisation simply builds a bigger platform on the same thin foundation. The gap does not close with time. Left alone, it widens, because the first ledger scales cheaply and the second does not.

That is what it means to call the gap structural rather than immature. An immaturity is something you grow out of. A structure is something that reproduces itself until

someone changes the structure.

6. The strongest case for building first

The honest objection to everything above is not weak, and it deserves to be met at its strongest rather than dismissed.

It runs like this. Governance imposed too early is worse than governance imposed too late. You cannot write sensible definitions for data whose value you have not yet discovered; you cannot design stewardship for uses that have not yet emerged. Insist on full governance before the platform exists and you will govern a great deal of data that turns out to be worthless, while smothering exactly the exploratory, schema-on-the-way-out freedom that made the whole approach valuable in the first place. Premature governance is how you get a two-year data-modelling exercise that delivers nothing and a business that has quietly routed around you. Build first. Let the valuable data reveal itself through use. Then govern the part that proved it mattered.

This is a serious argument and, within limits, it is correct. Governing everything up front is a real failure mode, and the heavy, ceremonial data programmes of the previous era earned their bad reputation honestly. Discovery does need slack. A lake that could hold nothing until every stream was certified would hold nothing worth having.

But notice what the argument actually justifies, and what it does not. It justifies sequencing governance — letting discipline follow discovery, applying it to the data that earns it rather than to all data on principle. It does not justify never funding governance, and in practice the two are constantly confused. "We'll govern it later" is a defensible plan when later is a funded, scheduled, owned commitment with a trigger — this dataset, once it is feeding a decision that matters, gets a definition and an owner. It is a fiction when later is a word used to win this quarter's budget with no intention of returning. The structural forces described above ensure that, absent a deliberate counterweight, later almost always means never: the next demo is always more fundable than the definition you deferred.

So the counter-argument, taken seriously, does not overturn the thesis. It sharpens it. The task is not to govern first or to build first. It is to build in such a way that governance is triggered — to decide, at the outset, that the moment a dataset crosses from exploration into a decision the business relies on, it acquires an owner, a definition, and a lineage as a condition of that reliance. Sequencing is legitimate. Starvation dressed as sequencing is the failure. The difference between them is whether later has a date and a name against it.

*"Sequencing governance is a defensible strategy.
Starving it and calling that sequencing is how the
strategy becomes an alibi."*

7. What the pattern tells us about transformation

Step back from data and the shape of this story is familiar, because it is the shape of almost every transformation that disappoints.

There is a visible thing and an invisible thing. The visible thing is the platform, the system, the reorganisation, the new operating model — the part you can point to, demonstrate, capitalise, and celebrate. The invisible thing is the discipline that makes the visible thing actually work: the definitions, the behaviours, the ownership, the slow accumulation of trust. Transformation intent attaches to the visible thing, because that is what can be specified, funded, and shown. Transformation reality depends on the invisible thing, because that is what determines whether any of it means anything a year later. And the gap between intent and reality is, over and over, exactly the second ledger — the underfunded, contested, unglamorous work that no one fought for.

The data case is a clean specimen of the disease precisely because the failure is so measurable. When the same customer is counted three ways, the gap between what was promised and what was built is a number you can put on a slide. In most transformations the gap is real but harder to see, which is why it survives longer. What data makes visible, other programmes merely feel.

The corrective is not more enthusiasm for the platform, and it is certainly not less. It is a refusal to treat the two ledgers as separable purchases. A data programme that funds storage and processing but not definition, lineage, and ownership has not bought a cheap version of the capability; it has bought the appearance of the capability and deferred the substance onto a ledger it does not intend to pay. The mature move — and here maturity is exactly the right word — is to insist that the second ledger is part of the price of the first, written into the same business case, owned by the same sponsor, funded on the same day.

The question to ask of any data ambition is not how much can we store and process but who will own what this data means, and when does that obligation start. An organisation that cannot answer the second question has not yet begun the programme it thinks it has funded.

We know all this. That is the uncomfortable conclusion. The knowledge is not the missing ingredient; the frameworks were written, the roles were named, the lesson was available to anyone who wanted it. What is missing is a structure that funds what it knows to be true against the constant pull of what is easier to show. Until that structure changes, the promise and the reality will keep drifting apart at the speed the platform can scale — which, now that storage is cheap and processing is abundant, is very fast indeed.