SAMPLE REPORT — NOT A CLIENT ENGAGEMENT

The Data-Rich, Insight-Poor Company

A diagnostic framework and 12-month roadmap for organisations that collect far more data than they use

Empirica Technologies Pty Ltd · empiricaai.org

June 2026 · Condensed Focused-tier sample (commissioned reports run 10–15 pages)

Every statistic in this document is cited to a checkable published source.

Contact: empirica@empiricaai.org ·

ABOUT THIS DOCUMENT

What this sample is — and what it is not

This document demonstrates, in condensed form, the format and sourcing discipline of an Empirica research report at the Focused tier (commissioned Focused reports run 10–15 pages). The topic — what to change when an organisation collects far more data than it uses — was chosen because it is the kind of question our report service is built for.

It is a sample, not a client engagement. It contains no client data and describes no client outcomes; the diagnosis and recommendations are built from the published research cited throughout and from our own operating experience running an autonomous research fleet in production. Where a commissioned report would analyse *your* systems, metrics and constraints, this sample necessarily speaks to the general pattern.

Three habits of every Empirica report are on display here and worth checking as you read: every statistic is cited to a checkable published source (the Sources section at the back — each entry notes why it mattered); claims we could not verify were cut rather than hedged; and the machine-readable companion (a structured JSON of the findings, roadmap and sources) ships alongside this PDF at empiricaai.org/samples/empirica-sample-report.json.

A commissioned report adds: analysis of your actual data estate, models scoped to your stack, a bespoke machine-readable companion, and (on Standard and Deep tiers) a walkthrough call with the founder — the firm's lead researcher. Reports run $2,500–$12,000 depending on scope; you get a fixed quote within 2 business days, before any work begins.

SECTION 1

Executive summary

The question this report answers: *“We collect a lot of data, but nobody uses it properly or to our advantage. What should we change — modelling, structure, people, anything?”*

The short answer: in the published evidence, this is an organisational and decision-process failure before it is a technology failure. An estimated 68% of data available to enterprises goes unleveraged[S1], while the share of surveyed firms describing themselves as data-driven *fell* in successive annual surveys — 37.1% to 32.4% to 31.0%[S3] — over a period when tooling only got better. The constraint is rarely collection or compute; it is the absence of ownership, trusted definitions, and a named path from data to a recurring decision. The association suggests the prize is material: firms that adopted data-driven decision making showed output and productivity 5–6% higher than expected given their other investments and IT usage[S2].

Five recommendations, in order of leverage:

1. **Map decisions before data.** Inventory the recurring decisions the business makes, their cadence and value-at-stake — then work backwards to the data each needs. Data-forward programmes produce dashboards; decision-back programmes produce changed behaviour (§3).
2. **Create one source of truth with named owners.** A single analytical store, metric contracts for the top KPIs, and domain owners accountable for data quality — product thinking applied to data, scaled to mid-market reality (§4).
3. **Model baseline-first against two or three mapped decisions.** Start with simple, explainable models wired to real decisions; add sophistication only where the baseline demonstrably leaves value on the table (§5).
4. **Close the translation gap before hiring modellers.** The scarce role is the translator between analytics and operations; we recommend an analytics engineer plus a translator function before a first data-science hire — and AI agents can absorb much of the repeatable layer (§6).

5. Sequence over 12 months with a 90-day proof loop. One use case end-to-end in the first quarter beats a platform programme that shows value in year two (§7).

SECTION 2

The problem, in numbers

Four published findings frame the situation of a data-rich, insight-poor organisation. Each is cited to its primary source; together they locate the bottleneck.

Finding	Figure	Source
Share of enterprise-available data that goes unleveraged	**68%** (32% is put to work)	Seagate / IDC, *Rethink Data*, 2020 [S1]
Output / productivity premium of firms adopting data-driven decision making, beyond other investments	**5–6%**	Brynjolfsson, Hitt & Kim, ICIS/SSRN, 2011 [S2]
Surveyed firms identifying as data-driven — falling across successive annual surveys	**37.1%** → **32.4%** → **31.0%** (2017–19)	Bean & Davenport, HBR / NewVantage survey, 2019 [S3]
Working time data scientists report spending on data preparation (loading + cleansing)	**~45%** (n = 2,360)	Anaconda, *State of Data Science*, 2020 [S4]

Read together: most data available to enterprises goes unused[S1]; using it well is associated with materially better firm performance[S2]; the failure mode is organisational — firms went *backwards* on self-assessed data-driven status while technology improved[S3]; and data scientists across industries report ~45% of their time going to preparation rather than analysis[S4] — a cross-industry average, not the unstructured-estate worst case. Any credible fix therefore has to touch structure, modelling *and* the operating model — in that order of dependency, which is how this report is sequenced.

A note on these figures: they are cross-industry survey and panel estimates, useful for sizing the pattern rather than your firm specifically. A commissioned report replaces them with measurements from your own estate — utilisation of your tables, latency of your decisions, time-cost of your reporting cycle.

SECTION 3

Diagnostic: where data value actually stalls

This report assesses a data estate as five layers. Value only flows when all five connect, and the published failure modes — unleveraged data[S1], stalled data-driven adoption[S3], analyst time lost to preparation[S4] — cluster in the middle and top of the stack (structure, access, decisions), rather than in capture.

Layer	The question to ask	Common failure	Symptom you'll recognise
1 · Capture	Are the events that matter recorded at all?	Over-collection of easy data, gaps in valuable data	Terabytes of logs; no record of quote-to-loss reasons
2 · Storage & structure	Is there one analytical home with usable schemas?	Data marooned across SaaS tools and spreadsheets	Five versions of “revenue”, none reconciled
3 · Access & trust	Can a motivated manager self-serve a trusted number?	Gatekeeping, undefined metrics, no catalogue	Every question becomes a ticket to one overloaded analyst
4 · Modelling	Do models exist where they would change a decision?	Models built tech-first, never wired to a decision	A churn model nobody's retention workflow consumes
5 · Decisions	Is any recurring decision contractually fed by data?	Reporting culture instead of decision culture	Dashboards reviewed monthly; decisions still made on instinct

The practical diagnostic is a **decision audit**, run in week one of the roadmap: list the ten most consequential recurring decisions (pricing reviews, stock re-orders, campaign allocation, hiring triggers, credit terms…), and for

each record who makes it, how often, on what information today, and what its value-at-stake is. Two facts fall out immediately: which decisions are flying blind despite relevant data existing (the use-case shortlist for §5), and which collected data feeds no decision at all — a candidate for deletion rather than storage spend.

SECTION 4

Structure: one source of truth, named owners

4.1 · Consolidate into a single analytical store

For a mid-market organisation the answer is rarely exotic: one cloud warehouse or lakehouse as the analytical home, fed by managed extract-load connectors from the operational systems, with transformations versioned in code rather than performed by hand in spreadsheets. The components are commodities now; the decision that matters is *that there is exactly one home*, so that every downstream number has a single auditable lineage. The industry's ~45% preparation average[S4] shows the scale of that tax: skilled people reassembling the estate by hand, per analysis, forever.

4.2 · Metric contracts — definitions are the real asset

The cheapest high-leverage artifact in this entire report is a **metric contract** for each of the organisation's top five to ten KPIs: a one-page, version-controlled definition — owner, formula, source tables, inclusion rules, refresh cadence. When “revenue” means one thing, the weekly argument about whose number is right disappears, and with it the deepest source of executive distrust in the data.

4.3 · Ownership: data as a product, scaled to your size

The data-mesh literature[S6] argues for domain-oriented ownership, data treated as a product, and self-serve infrastructure. A 100–500-person company should not build the full architecture — but the two principles transfer directly: **each domain's data has a named owner** accountable for its quality and documentation (sales owns CRM hygiene; operations owns inventory accuracy), and **datasets are published for consumers** — discoverable, documented, trustworthy — rather than dumped. Ownership is an accountability change, not a hiring programme; it costs a paragraph in four job descriptions.

4.4 · Governance-lite

Default-open access to the analytical store inside the company, with a short audited exception list (payroll, personal data, M&A). A lightweight catalogue — even a well-kept index page — beats an unused enterprise governance suite. Gatekeeping is layer-3 failure dressed up as prudence; the audit trail, not the lock, is what scales.

SECTION 5

Modelling: baseline-first, decision-backed

5.1 · Choose use cases from the decision audit

Score each candidate decision on value-at-stake, decision frequency, and data readiness, and take the top two or three. Frequency matters more than glamour: a model that improves a weekly decision by a few percent compounds faster than one that informs an annual strategy offsite. For most mid-market firms the shortlist lands on some subset of: **demand forecasting** (feeds purchasing and rostering), **churn / retention risk** (feeds a named save-workflow), **pricing response** (feeds quarterly price reviews), and **anomaly detection** on cost or quality streams (feeds an operational alert someone owns).

5.2 · Baselines before sophistication

Start every use case with the simplest defensible model — seasonal-naïve or moving-average forecasts, single-table regression or scorecard-style churn flags — and measure against the decision as it is made today. The evidence favours this ordering: with enough of the right data, simple models tend to outperform elaborate models built on less[S7], so the early money is better spent widening trusted data (§4) than deepening model machinery. The baseline also gives you an honest yardstick — if gradient boosting can't beat seasonal-naïve on your demand data, that is a finding about your data, learned cheaply.

5.3 · Respect the hidden debt

Production ML carries most of its cost outside the model: configuration, data dependencies, glue code, monitoring, retraining[S5]. Three rules keep it honest: **no model ships without a named decision consuming it**; **no model ships without monitoring** (input drift and outcome accuracy, reviewed on a calendar); and **prefer one model in production to three in notebooks**. A model that is not wired into a workflow is not an asset — it is deferred maintenance.

SECTION 6

People: close the translation gap first

“Nobody to use the data” is usually diagnosed as a data-scientist shortage and treated with a data-science hire. That hire, parachuted into an estate with no structure (§4) and no mapped decisions (§3), inherits the industry's ~45% preparation average[S4] as a starting point — and the rest of the archetype is familiar: models nobody consumes, then a departure. McKinsey's diagnosis points elsewhere: the role to fill first is the **translator** — the person who carries a business decision to the technical work and carries the result back into the workflow[S8].

Option	What it buys	When it's the right first move
Analytics engineer (builds §4: pipelines, warehouse, metric contracts)	Multiplies everyone downstream by shrinking the prep tax that averages ~45% across the industry[S4]	Usually the right first technical hire for a data-rich, insight-poor estate
Translator function — the translator role[S8], filled part-time by a trained-up domain manager (our mid-market adaptation)	Decisions actually consume the data; use cases chosen by value, not novelty	Immediately, from existing staff — it is a role, not necessarily a hire
Data scientist	Modelling depth beyond baselines	After §4 exists and a baseline model has hit its ceiling on a valuable decision
External / fractional specialists	Burst capacity for the warehouse build or first models, without permanent cost	When the 12-month plan needs skills the team won't permanently need
AI agent automation	The repeatable layer: monitoring, anomaly flagging, scheduled reporting, literature and citation work, first-draft analysis	Once structure exists — agents consume clean, documented data; they do not fix ownership or definitions

On the last row we speak from direct experience rather than literature: Empirica itself is one human and a fleet of autonomous research agents, with every published output scored 0–100 by a public-rubric validator before release. That experience cuts both ways and we report it honestly — agents are strong at the high-volume repeatable layer (monitoring, synthesis, drafting, citation verification) and they do not substitute for the human work in this report: deciding ownership, negotiating metric definitions, or changing how a decision is made. Automation amplifies a working operating model; it cannot create one.

Operating shape for mid-market: a **small hub** (analytics engineer + translator function, plus agent automation for the repeatable layer) serving **spokes in each domain** (the data owners from §4.3). No analytics department required.

SECTION 7

The 12-month sequence

Sequenced so every quarter ships something a decision-maker uses — the antidote to the platform programme that promises value in year two. The 90-day phase is deliberately narrow: one use case, end to end, visible.

Phase	Workstream	Concrete deliverables	Exit test
Days 0–90 Prove the loop	Decisions + structure + first baseline	Decision audit (§3) · metric contracts for top-5 KPIs · warehouse live with 2–3 core sources · baseline model on use-case #1 wired into its decision · one legacy report retired per new one shipped	A named manager makes a recurring decision differently, and says so
Months 4–6 Widen	Use cases #2–3 + operating model	Translator function named and trained[S8] · domain data owners in role · default-open access policy · baseline models #2–3 in their workflows	A second team self-serves a trusted number without a ticket
Months 7–12 Industrialise	Model lifecycle + automation	Monitoring and retraining calendar for every production model[S5] · agent automation of the repeatable layer · catalogue covering all published datasets · ROI review against the decision audit's value-at-stake column	Estate survives the loss of any single person; review quantifies value per use case

Budget shape, not budget numbers: the heavy spend in this plan is people-time, not licences. The warehouse-and-connectors bill for a mid-market estate is generally small next to the fully-loaded cost of a single mis-hired analytics role — which is why the sequencing (structure → translation → modelling depth) is also the cheapest path.

SECTION 8

What a commissioned version of this report contains

This sample speaks to the general pattern because it has no client data. A commissioned Empirica report on the same question is built on your estate:

- **Your decision audit, done for you** — interviews structured into the §3 framework, with value-at-stake estimates per decision.
- **Estate measurement** — actual utilisation of your stores and reports, metric-definition conflicts found in your BI layer, and the prep-tax measured on your team, replacing this sample's cross-industry figures.
- **A scored use-case shortlist and baseline-model scoping** for your top two or three decisions, including what each needs from §4 before it can ship.
- **A tailored 12-month sequence** with the hiring/role plan of §6 adapted to the team you already have.
- **The machine-readable companion** — the full findings, roadmap and source list as structured JSON, ready for your own tooling (this sample ships one alongside, so you can see the format).
- **Every source cited**, with a note on why it mattered — the same discipline as the Sources page here, applied to your domain's literature.

Reports run $2,500–$12,000 depending on scope; you get a fixed quote within 2 business days, before any work begins. Scope conversations start at **empirica@empiricaai.org** or empiricaai.org/reports.

SOURCES

Every source, and why it mattered

All sources verified against the cited publication on 2026-06-10. Claims we could not verify were removed from this report rather than hedged — mirroring the citation discipline our public-rubric validator enforces on our research.

[S1] Seagate Technology / IDC (2020). *Rethink Data: Put More of Your Business Data to Work — From Edge to Cloud*. Seagate-commissioned IDC report.
https://www.seagate.com/our-story/rethink-data/
Used for: Only 32% of data available to enterprises is put to work; the remaining 68% goes unleveraged. **Why it mattered:** The clearest published estimate we found of how much enterprise-available data goes unused — the report's core problem statement.

[S2] Brynjolfsson, E., Hitt, L. M. & Kim, H. H. (2011). *Strength in Numbers: How Does Data-Driven Decisionmaking Affect Firm Performance?*. ICIS 2011 Proceedings / SSRN working paper. · DOI 10.2139/ssrn.1819486
https://aisel.aisnet.org/icis2011/proceedings/economicvalueIS/13/
Used for: Firms that adopted data-driven decision making showed output and productivity 5–6% higher than expected given their other investments and IT usage (an observed association, not a measured return on remediation). **Why it mattered:** A widely cited firm-level estimate of what data-driven decision making is associated with — it sizes the prize.

[S3] Bean, R. & Davenport, T. H. (2019). *Companies Are Failing in Their Efforts to Become Data-Driven*. Harvard Business Review (reporting the NewVantage Partners 2019 executive survey).
https://hbr.org/2019/02/companies-are-failing-in-their-efforts-to-become-data-driven
Used for: The percentage of surveyed firms identifying as data-driven fell across successive annual surveys: 37.1% (2017), 32.4% (2018), 31.0% (2019). Verified against HBR's published summary of the survey. **Why it mattered:** Evidence that the binding constraint is organisational, not technical — self-assessed adoption went backwards while tooling improved.

[S4] Anaconda, Inc. (2020). *The State of Data Science 2020*. Practitioner survey, n = 2,360 (Feb–Apr 2020).
https://www.anaconda.com/resources/whitepaper/state-of-data-science-2020
Used for: Data scientists report spending roughly 45% of their time on data preparation (loading and cleansing) — a cross-industry average. **Why it mattered:** Quantifies why analytics hires underdeliver when the data layer is unstructured — nearly half their time goes to plumbing.

[S5] Sculley, D. et al. (Google) (2015). *Hidden Technical Debt in Machine Learning Systems*. Advances in Neural Information Processing Systems 28 (NeurIPS).
https://papers.nips.cc/paper/5656-hidden-technical-debt-in-machine-learning-systems
Used for: In production ML systems the learning code is a small fraction of the whole; configuration, data dependencies and glue code dominate long-run cost. **Why it mattered:** The canonical warning against buying models before plumbing — it shapes the build order in sections 5 and 7.

[S6] Dehghani, Z. (2019). *How to Move Beyond a Monolithic Data Lake to a Distributed Data Mesh*. martinfowler.com.
https://martinfowler.com/articles/data-monolith-to-mesh.html
Used for: Argues for domain-oriented data ownership, data treated as a product, and self-serve data infrastructure (published 20 May 2019). **Why it mattered:** Source of the ownership principles in section 4 — applied here at mid-market scale without the full mesh architecture.

[S7] Halevy, A., Norvig, P. & Pereira, F. (2009). *The Unreasonable Effectiveness of Data*. IEEE Intelligent Systems 24(2). · DOI 10.1109/MIS.2009.36
https://doi.org/10.1109/MIS.2009.36
Used for: Simple models trained on large datasets tend to outperform more elaborate models trained on less data. **Why it mattered:** Grounds the baseline-first modelling discipline in section 5 — spend on data reach before model sophistication.

[S8] Henke, N., Levine, J. & McInerney, P. (2018). *Analytics Translator: The New Must-Have Role*. McKinsey & Company (first published in Harvard Business Review).
https://www.mckinsey.com/capabilities/quantumblack/our-insights/analytics-translator
Used for: Defines the translator role: bridging technical analytics expertise and the operational managers whose decisions analytics is meant to serve. **Why it mattered:** Names the role most data-rich companies are actually missing — the centrepiece of section 6.

ABOUT EMPIRICA · DISCLOSURES