



databricks

The Data Unification Diagnostic

What This Diagnostic Is

Most public sector data teams are not choosing between a “warehouse” or a “lake.” They’re paying to run both.

Plain Language Definitions

Data Warehouse

A system optimized for trusted, repeatable reporting on structured data (budgets, programs, operations).

Data Lake

A low-cost store for large volumes of raw and semi-structured data, often where teams run advanced analytics and AI.

Why teams end up with both: Reporting and governance needs push them toward a warehouse, while AI and unstructured workloads push them toward a lake. The result is often duplicate storage, duplicate pipelines, and inconsistent definitions.

This diagnostic helps you pinpoint where fragmentation’s creating the biggest operational and cost drag across your data-to-decision workflow, and it gives you a clear set of next steps based on what you find.

Use it when you are:

- consolidating platforms,
- planning a modernization roadmap,
- trying to reduce time-to-insight,
- or preparing a business case for unification.

Who It's For

- CIOs, CTOs, CDOs, and enterprise architects accountable for platform direction
- Data engineering and analytics leaders responsible for delivery reliability
- Security and governance leaders responsible for access control, lineage, and auditability

How To Use The Diagnostic

1. Work through the four sections below.
2. Check the boxes that reflect your current reality.
3. Add your score at the end.
4. Use the interpretation and next steps to decide what to do first.

What “Done” Looks Like

You can clearly answer, “Where does fragmentation hurt us most, and what is the smallest pilot that will prove unification value?”

Section 1

Map The Data Movement Journey

Start with a simple, honest map of how data becomes a dashboard, report, model, or operational decision.

In one sentence:
"Our typical journey looks like:"



Quick prompts (write short answers):

Where does data usually land first?

How many systems does it pass through before someone uses it?

Where do you create additional copies "just to make it work?"

Where do handoffs happen between teams?

Scoreboard

0	1	2	3
Mostly direct path, minimal copies	Some duplication, but controlled	Multiple hops/copies are normal	You regularly recreate datasets because it is easier than tracing what exists

Section 2

Inventory Friction Points (Checklist)

Check every statement that is true today.

Architecture & Duplication

- ☐ The same dataset exists in multiple places because different tools require different formats or locations.
- ☐ Different teams maintain their own “trusted” copies of the same dataset.
- ☐ The same KPI can show different numbers depending on which system or report you use.
- ☐ Data quality issues are discovered late because it is hard to trace where a metric came from.

Delivery Reliability & Time-To-Insight

- ☐ Pipeline failures happen often enough that “break/fix” work is part of the normal week.
- ☐ Schema changes regularly break downstream dashboards, reports, or models.
- ☐ New reporting requests get delayed because the data must be copied, re-modeled, or re-approved in multiple places.
- ☐ Most engineering effort goes to maintenance and data movement, not net-new capabilities. Example: more than half the team’s sprint capacity is spent on break/fix, re-runs, and data movement work.

Governance, Security, and Auditability

- ☐ Access control and policy enforcement differ across systems (for example, between SQL reporting and engineering or ML workflows).
- ☐ If asked today, you couldn't quickly answer "who accessed this sensitive field, when, and how."
- ☐ Lineage is incomplete, inconsistent, or manual (you cannot trace from dashboard to raw source without detective work).
- ☐ Governance work slows delivery because controls are bolted on after the fact instead of enforced by the platform. Example: access reviews or security approvals happen late, requiring rework before something can go live.

AI and Unstructured Readiness

- ☐ Important unstructured data (documents, logs, images, audio/video) is not usable in governed analytics workflows.
- ☐ Data science work typically happens on copied datasets because production access is slow or hard to govern.
- ☐ Introducing AI workflows creates new security and compliance risk because governance is inconsistent across tools.
- ☐ Moving an AI use case from experimentation to production is hard because the operational path is unclear.

Score: Count checked boxes. (0–16)



Tip

If a statement feels "sometimes true," check it if it's true for your most mission-critical workflow.

Section 3

The Integration Tax

You don't need perfect numbers; you need an honest range.

1. About how many full-time equivalents (FTEs) touch data movement and reliability work?

- Engineering FTEs: _____
- Analytics/BI support FTEs: _____
- Security/governance support FTEs: _____

2. If you had to estimate time allocation (relative to total data team delivery capacity):

- % spent maintaining / troubleshooting pipelines: _____ %
- % spent adding net-new capabilities: _____ %

3. What is your typical “time-to-insight” for a new request?

- Simple dashboard / report: _____ days
- Cross-domain or multi-source request: _____ days
- Production ML / AI use case: _____ weeks

0	1	2	3
Most requests ship quickly, reliability work is contained	Some bottlenecks, but manageable	Bottlenecks are the norm, time-to-insight is inconsistent	Delivery speed and reliability are a major constraint on mission outcomes

Section 4

Readiness Score and Interpretation

You don't need perfect numbers; you need an honest range.

Add your scores:

- Section 1 (0-3): _____
- Section 2 (0-16): _____
- Section 3 (0-3): _____

Total score (0-22): _____

What Your Score Means

Score	Urgency	What It Means
0-6	Low Urgency	• Fragmentation exists, but it is not a top constraint. • Focus on cleaning up lineage, standardizing definitions, and reducing duplicative reporting.
7-13	Moderate Urgency	• Fragmentation is slowing delivery and increasing risk. • Prioritize a targeted unification pilot that reduces copies and standardizes governance.
14-22	High Urgency	• Fragmentation is creating a measurable operational tax. • A unification effort is likely to unlock major time-to-insight gains, governance consistency, and reliability improvements.

What you are solving for (and how the right platform helps)

The point of unification isn't "one tool." It's reducing the operational tax created when data, governance, and delivery live in different places.

Put simply: unification is less about standardizing on a single product, and more about ensuring teams can use the same governed data for reporting, engineering, and AI without creating extra copies, handoffs, and reconciliation work.

At a practical level, most teams are solving for three outcomes:

1) Faster time-to-insight (without brittle pipelines)

What you are solving for: fewer hops, fewer copies, and less break/fix work to get from raw data to a trusted output.

How the right platform can help: a unified data foundation and consistent execution environment can reduce the number of handoffs and transformations required to deliver analytics and AI.

2) Consistent governance and auditability

What you are solving for: one set of access controls, visibility, and lineage expectations across teams and workloads.

How the right platform can help: centralized policy enforcement and end-to-end visibility can make it easier to answer, "Who accessed what, when, and under what policy?" across SQL, engineering, and AI workflows.

3) AI readiness that does not create new risk

What you are solving for: bringing unstructured and structured data into analytics and AI workflows without creating a separate, less governed "AI stack."

How the right platform can help: shared data + shared governance can make it easier to move from experiments to production while maintaining security controls.

Recommended Next Steps

Choose the smallest proving move

Score	What It Means
0-6	• Pick 1 cross-domain reporting workflow and document end-to-end lineage. • Standardize “system of record” definitions for a short list of mission-critical datasets. • Identify where duplication is intentional and where it is accidental.
7-13	• Select one high-value dataset that is currently copied across systems. • Define a pilot that proves you can: • reduce copies, • improve freshness, • and keep governance consistent. • Establish success metrics (examples): • reduce pipeline break/fix time, • reduce time-to-insight, • increase auditability and lineage coverage.
14-22	• Create a 30-day pilot plan with: • a single mission-relevant workflow, • clear decision makers, • and a defined “exit criteria” for success. • Inventory governance requirements up front (access control, audit logging, retention). • Identify the two or three integration choke points that consume the most engineering time, and prioritize eliminating them first.

Questions to take into vendor evaluation conversations

1. Where will data physically live, and what is the practical path to reduce copies?

2. How do you enforce governance consistently across SQL, notebooks, and AI workflows?

3. What lineage and audit visibility do you provide out of the box?

4. What does a 30-day pilot look like, and what are the success criteria?

5. What changes when you introduce unstructured data into governed analytics and AI workflows?

Explore how
Databricks is
supporting the
public sector.



databricks

2026