

Data Lake / Lakehouse Health Pack

Discovery questionnaire — score each item honestly; low scores are the point: they show where the program will hurt.

Pack	Data Lake / Lakehouse Health Pack
Document	Discovery questionnaire (24 questions)
Response scales	Maturity 1-5 · Yes / Partially / No · Written
Date	2026-09

AnovaCloud — Engineering the AI-Native Enterprise. Frisco, Texas. This pack is a working instrument, not a brochure: complete it with your team, score honestly, and bring it to your discovery call. © 2026 AnovaCloud.

HOW TO USE

Run this like a workshop, not a survey

Complete it with the stakeholders named in each section's intro. One scribe, one room (or call), 90 minutes. Score from evidence — monitoring data, documents, demos — not opinions. Any item scoring 1–2 is a program risk: it needs an owner and a close date before you commit budget. Bring the scored questionnaire to your discovery call; it compresses weeks of fact-finding into one session.

1 · SOURCE INVENTORY & INGESTION**Source inventory & ingestion**

Every pipeline starts with a source nobody documented.

Q1. Are all sources inventoried with owner, format, volume, and freshness SLA?

1	2	3	4	5
---	---	---	---	---

Maturity 1-5 (1=Ad hoc, 2=Repeatable, 3=Defined, 4=Managed, 5=Optimized)

Evidence: source register.

Q2. Is ingestion incremental and idempotent (CDC, watermarks) — or full reloads?

1	2	3	4	5
---	---	---	---	---

Maturity 1-5 (1=Ad hoc, 2=Repeatable, 3=Defined, 4=Managed, 5=Optimized)

Evidence: per-pipeline ingestion pattern.

Q3. Are schema changes in sources detected and handled, or do they break pipelines silently?

1	2	3	4	5
---	---	---	---	---

Maturity 1-5 (1=Ad hoc, 2=Repeatable, 3=Defined, 4=Managed, 5=Optimized)

Evidence: schema-drift handling + recent incidents.

Q4. Is there exactly one sanctioned ingestion path per source type (no shadow pipelines)?

☐	Yes	☐	Partially	☐	No
--------------------------	-----	--------------------------	-----------	--------------------------	----

Evidence: pipeline inventory vs. source register.

2 · STORAGE & TABLE FORMAT

Storage & table format

The format decision compounds for years.

Q5. What table format(s) are in use, and is the choice deliberate and documented?

Evidence: format decision record.

Q6. Are small-file problems, partition skew, or compaction debt present?

1	2	3	4	5
---	---	---	---	---

Maturity 1-5 (1=Ad hoc, 2=Repeatable, 3=Defined, 4=Managed, 5=Optimized)

Evidence: file-size distribution + compaction jobs.

Q7. Is time travel / versioning used operationally (audits, rollbacks) or merely available?

☐	Yes	☐	Partially	☐	No
--------------------------	-----	--------------------------	-----------	--------------------------	----

Evidence: examples of operational use.

Q8. Is storage lifecycle managed (tiering, retention, deletion) with cost visibility?

1	2	3	4	5
---	---	---	---	---

Maturity 1-5 (1=Ad hoc, 2=Repeatable, 3=Defined, 4=Managed, 5=Optimized)

Evidence: lifecycle policy + storage cost trend.

3 · TRANSFORMATION DISCIPLINE**Transformation discipline**

Medallion is a discipline, not a folder structure.

Q9. Do bronze/silver/gold (or equivalent) layers exist with enforced contracts between them?

1	2	3	4	5
---	---	---	---	---

Maturity 1-5 (1=Ad hoc, 2=Repeatable, 3=Defined, 4=Managed, 5=Optimized)

Evidence: layer contracts + violation examples.

Q10. Are transformations tested (unit/integration) and code-reviewed like software?

1	2	3	4	5
---	---	---	---	---

Maturity 1-5 (1=Ad hoc, 2=Repeatable, 3=Defined, 4=Managed, 5=Optimized)

Evidence: test coverage + review practice.

Q11. Is orchestration centralized with SLAs, retries, and alerting — or cron and hope?

1	2	3	4	5
---	---	---	---	---

Maturity 1-5 (1=Ad hoc, 2=Repeatable, 3=Defined, 4=Managed, 5=Optimized)

Evidence: orchestrator + SLA attainment.

Q12. Can any analyst trace a gold-table metric back to its bronze source?

☐	Yes	☐	Partially	☐	No
--------------------------	-----	--------------------------	-----------	--------------------------	----

Evidence: lineage demo on one metric.

4 · GOVERNANCE & QUALITY**Governance & quality**

Trust is the product.

Q13. Is there a data catalog with ownership, definitions, and freshness per dataset?

1	2	3	4	5
---	---	---	---	---

Maturity 1-5 (1=Ad hoc, 2=Repeatable, 3=Defined, 4=Managed, 5=Optimized)

Evidence: catalog coverage % of gold tables.

Q14. Are data-quality checks (freshness, volume, distribution, referential) automated on critical pipelines?

1	2	3	4	5
---	---	---	---	---

Maturity 1-5 (1=Ad hoc, 2=Repeatable, 3=Defined, 4=Managed, 5=Optimized)

Evidence: DQ check inventory + alert history.

Q15. Is PII classified and access-controlled at the table/column level?

☐	Yes	☐	Partially	☐	No
--------------------------	-----	--------------------------	-----------	--------------------------	----

Evidence: classification + access policy.

Q16. Are breaking changes to gold tables versioned and communicated to consumers?

1	2	3	4	5
---	---	---	---	---

Maturity 1-5 (1=Ad hoc, 2=Repeatable, 3=Defined, 4=Managed, 5=Optimized)

Evidence: change-log + consumer comms examples.

5 · SERVING & CONSUMERS**Serving & consumers**

A lakehouse nobody queries is an archive with ambition.

Q17. Who are the consumers (BI, ML, apps, partners) and are their SLAs written down?

Evidence: consumer/SLA list.

Q18. Are serving patterns fit for purpose (caching, marts, APIs — not raw bronze queries)?

☐	☐	☐	☐	☐
1	2	3	4	5

Maturity 1-5 (1=Ad hoc, 2=Repeatable, 3=Defined, 4=Managed, 5=Optimized)

Evidence: serving architecture.

Q19. Is query performance acceptable at p95 for the top consumer workloads?

Evidence: p95 timings for top 10 queries.

Q20. Is there a deprecation path for unused/duplicate datasets?

☐	Yes	☐	Partially	☐	No
--------------------------	-----	--------------------------	-----------	--------------------------	----

Evidence: deprecation log.

6 · COST & OPERATIONS**Cost & operations**

Untracked cost becomes untracked architecture.

Q21. Is storage + compute cost attributed per pipeline/team?

1	2	3	4	5
---	---	---	---	---

Maturity 1-5 (1=Ad hoc, 2=Repeatable, 3=Defined, 4=Managed, 5=Optimized)

Evidence: cost-attribution report.

Q22. Are the top-10 cost drivers known, with optimization owners?

	Yes		Partially		No
--	-----	--	-----------	--	----

Evidence: cost-driver list.

Q23. Is there an on-call rotation with runbooks for pipeline failures?

1	2	3	4	5
---	---	---	---	---

Maturity 1-5 (1=Ad hoc, 2=Repeatable, 3=Defined, 4=Managed, 5=Optimized)

Evidence: rotation + runbook links.

Q24. Are capacity and growth planned (not discovered at 3 a.m.)?

	Yes		Partially		No
--	-----	--	-----------	--	----

Evidence: capacity plan.

SCORING

Read the pattern, not just the average

Average each section. Sections averaging below 3 are your critical path — sequence the project plan so those risks close first. A single 1 on a load-bearing question (downtime budget, rollback, access control) outweighs a 4 average elsewhere. Record owners and close dates below.

Lowest-scoring section:

Single biggest risk:

Owner:

Close date:
