

DATA STRATEGY TRACKS

Architecture and principles

Govern every data initiative with a small set of stable principles, then shape the architecture around how well the data and the questions are already understood.

THE SO WHAT OF THIS TRACK

Agree eight to ten principles and a one-page target state before choosing tools, and test every new initiative against both at the stage gate.

Part of stage 4: Design

Roadmaps change every quarter; principles should not. A short, agreed set of principles lets many teams build in different places and still end up with one coherent estate. Architecture that is chosen use case by use case, without them, ends in duplicated pipelines, competing copies of the truth and a platform bill nobody can explain. The principles and the target state belong in the strategy because they are how the strategy is enforced between reviews.



QUESTIONS THIS TRACK ANSWERS

- Which principles will every data initiative be held to, whoever runs it?
- How well understood are the data and the questions behind each priority use case?
- Which integration, storage and analytics styles fit each kind of work?
- What must the target-state architecture contain, and what is deliberately left out?
- How does our value proposition change where the architecture puts its effort?

1

The method

Principles, four zones and a target-state checklist

Nine principles that govern the data estate

Our starting set. Keep the ones that fit, reverse the ones that do not, and write down why.

Decisions first

Every dataset, pipeline and model exists to improve a named decision; if none can be named, it is not built.

Enterprise asset, domain owned

Data belongs to the organisation, and a named business domain is accountable for each data product.

Open by default, protected by design

Data is discoverable and shareable unless a stated reason says otherwise; classification and access rules travel with it.

One meaning per term

Shared business terms, entities and metrics are defined once and reused everywhere they appear.

Fit for its purpose

Quality is set by the decision the data serves, published as a score, and checked where the data is produced.

Nine principles that govern the data estate (continued)

Reuse before rebuild

New use cases start from certified data products; a new copy needs a reason the architecture board accepts.

Connect before you move

Query data where it lives when that is good enough; copy or materialise it only for performance, history or isolation.

Traceable end to end

Every published number can be traced to its sources, its transformations and its owner.

Responsible use

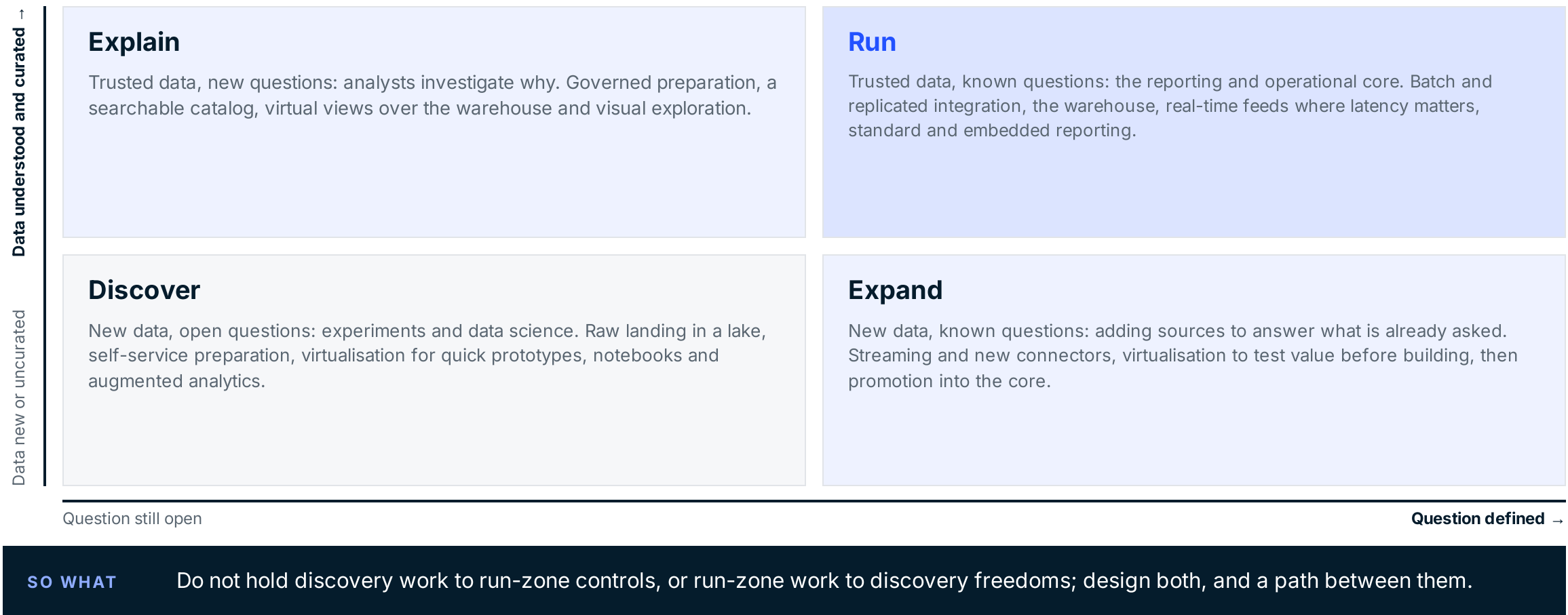
Lawful is the floor, not the bar: sensitive and automated uses pass an ethics and privacy check before release.

SO WHAT

Principles only govern if the stage gate checks them; add a principles review to every funding decision.

Four zones: how well do we know the data and the question?

Place each use case by asking two things: is the data understood and curated, and is the question already defined?



Which styles fit each zone

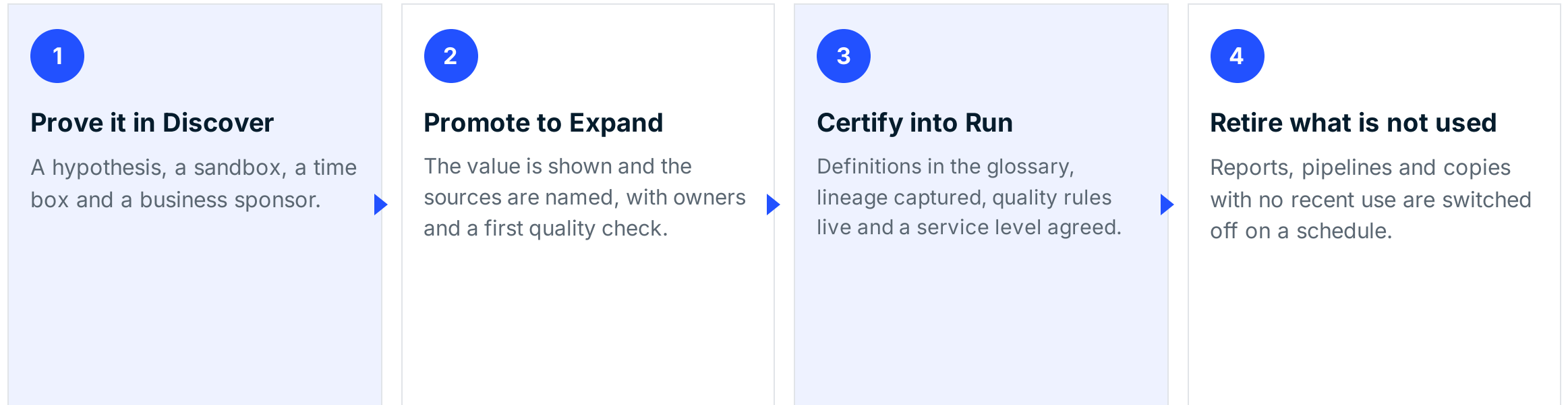
Zone	Integration	Storage and management	Access and analytics
Run	Scheduled batch, replication, change capture	Warehouse, real-time store, master data	Standard reports, dashboards, embedded analytics
Explain	Virtual views over curated data, governed preparation	Warehouse plus semantic layer	Visual exploration, natural-language query
Expand	New connectors, streaming, virtualisation to test value	Staging area, then promotion to the core	Prototype reports, then the standard set
Discover	Self-service preparation, raw ingestion	Lake or lakehouse sandbox	Notebooks, augmented analytics, model training

SO WHAT

Most estates need all four zones; the strategy decides how much of each, and who may move work between them.

The path from discovery to the core

Work moves zone to zone as it proves its worth; each move has an entry test.



Target-state architecture checklist

A one-page target state must answer each of these.

Area	What to capture	Test of done
Sources	Internal systems by domain, and external data with its licence terms	Every critical source has an owner and a criticality rating
Integration styles	Batch, replication, streaming, virtualisation and APIs, and when each is used	One written rule per style, not one per project
Management styles	Warehouse, lake or lakehouse, semantic layer and data products	Each data product has a home and a service level
Access styles	Portal, analyst workbench, data science lab, embedded analytics, APIs to other systems	Each persona has a named way in
Master and reference data	Golden records for core entities and shared code lists	One agreed customer, product and supplier record
Quality and observability	Rules at source, published scores, pipeline monitoring	Scores visible to the people who use the data
Security and privacy	Classification, access control, masking, consent and retention	Policies enforced by the platform, not by habit
Metadata	Catalog, glossary, lineage and usage, kept current automatically	Any published number traceable to source

Your value proposition changes the emphasis

Value proposition	Architecture puts its effort into	Principle that dominates
Foundation: data the business runs on	Availability, cost, speed of onboarding new sources, service levels	Reuse before rebuild
Enabler: data that improves existing processes	Trusted definitions, quality at source, embedded analytics in workflows	Fit for its purpose
Growth engine: data that creates new offers	Sandboxes, fast experimentation, external data, APIs and data sharing	Connect before you move

SO WHAT If the architecture effort does not match the value proposition the executives chose, one of the two is wrong.

2

Tools and evidence

Score it, measure it, see it in six industries

Zone finder: where does this use case belong?

Answer five questions about one use case; the finder suggests its zone and the patterns to start from.

Question	Options
How well defined is the question?	We ask it every week or month; We know the question but not the answer; We are looking for what to ask
Is the data already curated and owned?	Yes, certified and owned; Partly, some sources are new; No, it is new or raw
How often will the answer be used?	Continuously, in an operational process; Regularly, by a known group; Once or a few times, to decide something
What happens if the answer is wrong?	A regulatory, financial or safety impact; A poorer decision we would catch later; Little; it is an experiment
How fast does the data need to be fresh?	Within seconds or minutes; Daily is fine; A one-off extract is enough

What each answer points to

Run

A known question over trusted data. Build it for reliability and reuse in the reporting and operational core. Integrate by scheduled batch or change capture into the warehouse. Certify definitions and quality rules before go-live.

Explain

New questions over trusted data. Give analysts governed freedom on the curated core. Expose curated data through a semantic layer and virtual views. Make the catalog the starting point for every analysis.

Expand

A known question that needs new data. Test the value cheaply, then promote it into the core. Connect the new source virtually first to test its value. Name an owner and run a first quality check.

Discover

An open question over new data. Time-box the experiment and keep it out of production paths. Land raw data in a sandbox with a time limit and a sponsor. Use self-service preparation, notebooks and augmented analytics.

SELF-CHECK

Is this in place today?

Score each statement 1 (not at all) to 4 (fully, evidenced).

#	Statement	Score
1	We have a written set of data principles that leaders have agreed and can repeat.	
2	Every new initiative is checked against the principles before it is funded.	
3	Our target-state architecture fits on one page and is kept current.	
4	We distinguish trusted core work from exploratory work and govern them differently.	
5	There is a defined path for promoting proven experiments into the core.	

ILLUSTRATIVE

How it looks in six industries

Industry	Situation	The move	Measure
Manufacturer	Plant historians and quality data sit in each plant; field data on the installed base is unseen.	Run zone for delivery and quality reporting across plants; a Discover sandbox for installed-base and sensor analytics.	Share of plants reporting from one certified quality data product
Retail bank	Risk reports are rebuilt by hand from several ledgers, and lineage cannot be shown to the supervisor.	Principles of traceability and one meaning per term enforced first; regulatory metrics moved into the certified Run zone.	Share of regulatory metrics with end-to-end lineage
Retailer	Stores and online plan stock separately; pricing experiments run on ad hoc extracts.	One product and stock data product in the Run zone; pricing and range tests run in Discover with a promotion path.	Time from proven pricing test to production use
Hospital group	Patient-flow data is spread across hospital systems, and a privacy incident has raised the bar.	Protected-by-design principle applied first; bed and discharge data virtualised into an Expand view before building the core.	Share of sensitive data assets classified with access policies enforced
Telecom operator	Mobile, broadband and business run on separate billing and CRM stacks; inventory drifts from the field.	Connect before you move: a virtual customer view across the three stacks, then a mastered customer record in the Run zone.	Share of customers with one resolved identity across products
Insurer	Actuarial and regulatory reporting rely on manual reconciliations between policy and claims systems.	Certified policy and claims data products in the Run zone; fraud and leakage models built in Discover and promoted.	Manual reconciliation steps removed from the reporting cycle

Measures for the data scorecard

Measure	Definition	Lead or lag
Principles compliance at gate	Share of funded initiatives that passed the principles review	Lead
Source onboarding time	Median days from request to a governed, usable new source	Lead
Data product reuse	Average number of use cases served by each certified data product	Lead
Duplicate pipelines retired	Redundant pipelines and copies switched off in the period	Lead
Promotion cycle time	Median days from proven experiment to certified production use	Lead
Lineage coverage	Share of critical reports traceable to source	Lag
Platform cost per active user	Run cost of the data platform divided by monthly active users	Lag

3

Making it happen

Traps, the first 90 days and where SCIKIQ fits

Traps to avoid

Principles on a poster

Agreed in a workshop, never checked again.
Counter-move: Make the principles a stage-gate criterion and log every exception with an owner and expiry.

Tool-led target state

The target architecture is a list of products.
Counter-move: Describe styles and capabilities first; products are chosen to fit them, not the reverse.

One zone for everything

Experiments are forced through production controls, or production runs on sandbox code.
Counter-move: Design the four zones explicitly, with an entry test for each move between them.

Copy by default

Every project builds its own extract, and nobody knows which copy is right. Counter-move: Connect in place first and require a reason for each new copy.

No retirement plan

The estate only grows; cost and confusion grow with it. Counter-move: Schedule retirement reviews and switch off unused reports and pipelines.

The first 90 days

Days 1-30: Draft and agree the principles

- Inventory current architecture decisions and exceptions
- Draft eight to ten principles with the architecture board
- Test them against three live initiatives
- Agree them at the executive council

Days 31-60: Map the zones and the target state

- Place the priority use cases in the four zones
- Draft the one-page target state against the checklist
- Name owners for critical sources and data products

Days 61-90: Make them enforceable

- Add the principles review to the stage gates
- Publish the promotion path from Discover to Run
- Start the first retirement review
- Report principles compliance on the data scorecard

Where SCIKIQ fits

Virtual Data Fabric

Connects, federates and queries data where it lives, with caching and on-demand materialisation, so connect-before-you-move is the default.

Semantic Data Layer

Business objects, semantic models and natural language to SQL give the Explain zone one meaning over curated data.

Enterprise Discovery Engine

Discovers technical and business metadata, classification and lineage across sources, keeping the target state honest.

Governance Gate

Classification, access control, masking and policy checks enforce the protected-by-design principle in the platform.

Take it further

- The SCIKIQ Data Strategy Framework: the cascade this track plugs into
- Strategy Cascade and Scorecard Workbook: add this track's objectives and measures
- Run the interactive tool and self-check at scikiq.org/data-strategy#track-architecture

scikiq.org/data-strategy