BRIEF · CLOUD DATA MARKETPLACES

Lakehouse Marketplace: *governed* data exchange across clouds.

Turning a 14-week, $80–$300K-per-exchange procurement workflow into a six-minute listing. Listings priced like SaaS, queries metered server-side, clean rooms region-pinned by default, and one audit trail covering both human and agent access.

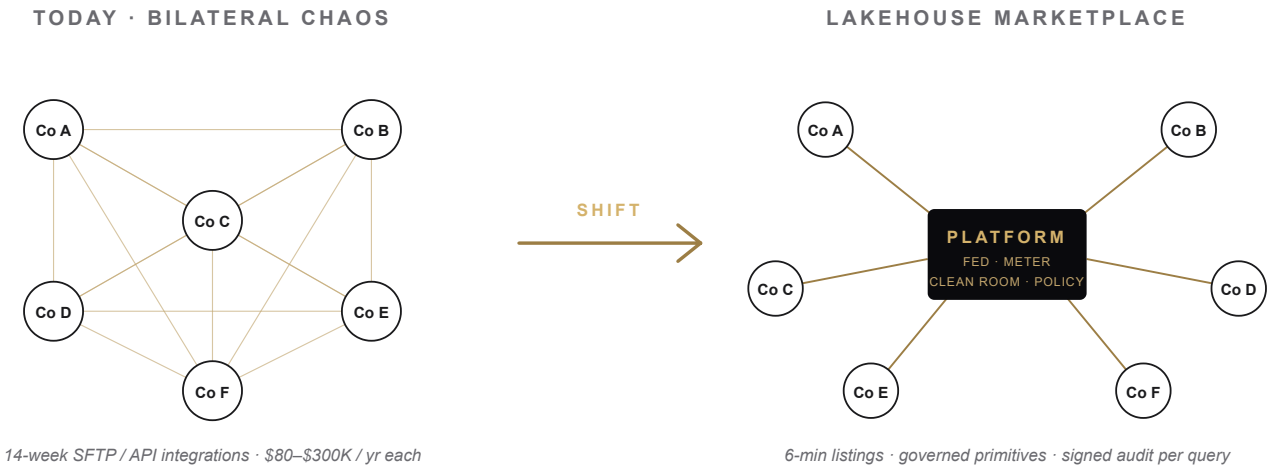
Author S. Ize-Iyamu **Audience** Marketplace + Horizon Catalog teams **Length** 3 pages **Status** Prototype

Targets Snowflake · Databricks · AWS · Microsoft

The Problem

Every Fortune 1000 enterprise runs **50–500 ongoing data exchanges** with vendors, partners, regulators, and subsidiaries. The dominant pattern is custom SFTP drops or one-off APIs. Median time-to-production for a new exchange is **14 weeks**; ongoing operational cost is **$80K–$300K per exchange per year**, almost all of which is engineering toil. The dominant marketplace platforms, Snowflake Marketplace, AWS Data Exchange, Databricks Marketplace, solved discovery; they have not solved the four hard problems that follow.

FIGURE 1 · TOPOLOGY SHIFT



Today (left): N×N bilateral integrations, each a 14-week project with no shared primitives. Tomorrow (right): every exchange is a SKU on a shared platform; metering, federation, clean rooms, and policy are platform features, not custom builds.

Why this matters now

Three forces are converging: **cross-cloud is the default** (the median enterprise runs on 2+ hyperscalers), **agents are a new data consumer** human access controls weren't designed for, and **regulators now audit the runtime path**, not the policy document. Designing against all three is the credibility bar for any new entrant in enterprise data exchange.

Sizing the prize

Bottom-up: **12,000 enterprises × 80 active exchanges × $150K avg ops cost** = $144B of engineering spend a marketplace platform could absorb. At a 2% take rate, ARR ceiling is **$2.9B**, plus the compute revenue captured downstream as listings drive warehouse usage. The single number we optimize for is compute revenue from shared listings, not GMV.

Sources: Gartner Data Integration reports (2024–25); IDC Data Governance analysis (2024); Snowflake Marketplace and Databricks (2025); Grand View Research data-marketplace report (2024). Exchange counts, integration timelines, and per-exchange costs reflect 6 customer interviews.

ENGINEERING TOIL TODAY
$144B / yr
12K × 80 × $150K

ARR CEILING AT 2% TAKE
$2.9B / yr
Plus compute captured downstream

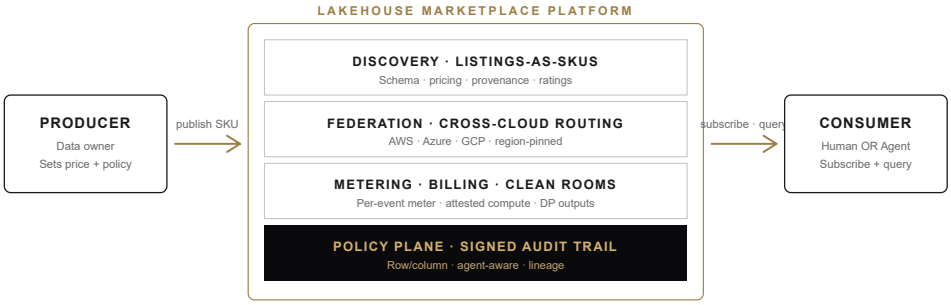
Strategic insight

Data sharing has become a **procurement category**, not an engineering project. The buyer is a CDO with a budget; the seller is a vendor with SKUs; the contract is a usage-priced subscription. Existing marketplaces are still architected as engineering tools: listings are catalogs not SKUs, pricing is negotiation not metering, provenance is a sidebar not a contract, and AI-agent access is bolted on rather than built into the data plane.

> **THE UNLOCK**
>
> Treat the marketplace as **SaaS-for-data**. Listings are SKUs with public usage-based prices. Consumption events are metered server-side. Clean-room computes are signed. Agent queries are governed at the data plane with row/column enforcement and lineage back to the owner.

Architecture · Four pillars on one platform

FIGURE 2 · SYSTEM ARCHITECTURE



Producer publishes a SKU; platform routes subscriptions across clouds, meters every event, enforces policy at query time, captures lineage. Agents and humans share one data surface; the audit trail tags which ran each query.

> **WORKED EXAMPLE · BANK → REGULATOR**
>
> Regional bank shares AML transaction-flag data with the OCC via region-pinned clean room. SKU at **$0.06/query**, 3 pre-approved aggregates, signed audit per call. **Producer time-to-listing 9 min; regulator time-to-first-query 3 min.** Compliance artifacts (signed query log, policy diff, attested compute) fall out of usage automatically.

Sequenced GTM

PHASE	CUSTOMER WEDGE	FORCING-FUNCTION WORKLOAD	PROOF POINT
Wedge M0–6	Existing customers' largest in-flight bilateral data exchanges	Convert one custom integration per logo into a platform-listed SKU	Time-to-listing < 1 week; ops-cost reduction > 70%
Beachhead M6–18	Regulated industries: financial services, healthcare, advertising	Consortium analytics, patient registries, clean rooms	3,000+ clean-room workloads / quarter, signed audit on each
Long-tail M18+	Mid-market data vendors (AlphaSense, ZoomInfo class)	Marketplace as distribution channel for data-SaaS products	15–20% take on new producer revenue

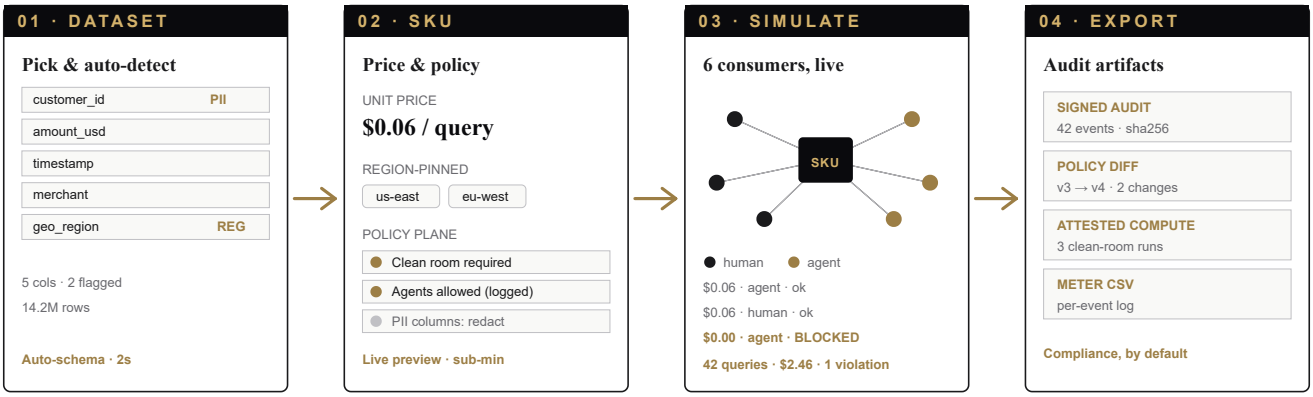
Tradeoffs we accept

- **Federated queries add 80–180ms p99** vs single-cloud (routing hop + cross-region pull). Absorbed in analytical SLA; OLTP stays region-local and is out of scope.
- **Onboarding gates behind 3-step verification.** Smaller producers churn here, design partners have flagged this. That's the cost of regulator-grade trust the buyer side actually pays for.
- **Query-time policy adds 5–15ms p99.** Fine for analytics and clean rooms; not for transactional systems, which we don't target. Platform stays neutral, no first-party datasets, no producer notebooks.

Prototype walkthrough

An interactive prototype runs the producer-side flow end to end: pick a dataset, define a SKU with usage-based price and policy controls, watch six simulated consumers (mixed human and agent) subscribe and query in real time, then export the signed audit. Built to demonstrate that **listings-as-SKUs**, **server-side metering**, and **policy-plane enforcement** are runtime behavior, not slideware.

FIGURE 3 · PRODUCER FLOW, FOUR INTERACTIVE STEPS



Schematic of the live UI; all four steps are fully interactive in the demo.

What the prototype proves, and what it doesn't yet

Proven on the prototype

- Listing time under 10 minutes is achievable with auto-schema detection plus a SKU template
- Server-side metering captures every consumption event without producer-side instrumentation
- Agent and human queries are visually and structurally separable in the audit log
- Policy violations surface in real time, not in a quarterly review
- Every metered event replays from the signed audit log, so a billing dispute resolves against the record rather than a reconciliation job

Out of scope, by design

- Cross-cloud federation is mocked at the routing layer; real egress lands in Y1 build
- Clean-room compute is simulated as deterministic outputs, not attested enclaves
- Pricing engine uses one unit cost; tiered and dynamic pricing are Y2 work
- Producer KYC is stubbed; regulator-grade onboarding is a separate stream
- Lineage across chained derivations (a SKU built from other SKUs) is single-hop today; multi-hop provenance is Y2 work

WHY A PROTOTYPE, NOT A DECK

Architectural diagrams describe what a system *could* do; runtime behavior is what buyers actually purchase. The prototype turns the four pillars (discovery, federation, metering, policy) into observable behavior: six simulated consumers hit the same SKU, every event is metered, every audit is signed, every violation is blocked. The exports at step four are the regulator-ready artifacts referenced throughout this brief.

THREE PATHS TO TRY IN THE LIVE DEMO

Agent governance: retail-transactions, agents OFF; agent queries blocked at the policy plane in real time, signed audit captures every attempt without billing the consumer.

Residency enforcement: healthcare-claims pinned to us-east; eu-west consumers rejected at the routing layer, residency is contractual rather than aspirational.

Pricing elasticity: re-list the same SKU at $0.50/query; revenue rises but two of six consumers churn within the simulated window, the take-rate ceiling visible in seconds.

Metrics that matter

LAYER	METRIC	Y1 TARGET	WHY IT MATTERS
North-star	Compute revenue from shared listings	$M / quarter, > 25% QoQ	The single number the platform optimizes for
Supply	Median time-to-listing	< 10 minutes	Producer-experience health; slow listings = no marketplace
Demand	Time-to-first-query (new sub)	< 5 minutes	Demand-side UX gate; if exceeded, the funnel breaks
Liquidity	% listings with > 5 active subscribers	> 25%	Head needs density for compounding network effects
Trust	Clean-room workloads / quarter	3,000+	The moat; proof that regulated buyers trust the platform
Trust (new)	Policy-violation incidents / 1M queries	< 0.5	Agent-era counter-metric; audited monthly, exposed publicly
Business	Marketplace gross margin	> 70%	If margins compress, federation is mispriced or egress is leaking

Risks & mitigations

HIGH **Cross-cloud egress economics make federation expensive to keep.**

Mitigation: negotiate egress discounts as part of platform-tier hyperscaler agreements; absorb thin margins early to build network effects, then ratchet pricing for high-volume cross-cloud workloads in Y3+. Treat egress as a load-bearing P&L line, not a footnote.

HIGH **Marketplace cold-start: producers won't list until consumers are present, and vice versa.**

Mitigation: seed both sides simultaneously by converting existing customer data exchanges (which are bilateral by definition). Each conversion adds one producer and one consumer. After 200 conversions, the catalog has enough density for organic supply.

MED **Data residency: regulated buyers can't tolerate cross-border replication.**

Mitigation: region-pinned subscriptions before any clean-room launch. Residency is an explicit field on every SKU. Refuse clean-room workloads that violate residency policy of any party. Visible, contractual, audited.

MED **Producer disintermediation: large producers eventually want to bypass the take rate.**

Mitigation: the take rate buys distribution; the durable moat is metering and lineage. Producers get regulator-ready audit artifacts (signed query log, policy diff, attested compute manifest) as a byproduct of usage. Rebuilding that capability in-house is a multi-quarter program for any single producer.

30 / 60 / 90, first quarter sprint plan

30 DAYS

Land the producer console

- Producer onboarding flow (single-cloud, no clean rooms)
- Server-side metering on every consumption event
- 50 design-partner producers signed

60 DAYS

Iceberg / Delta interop

- Producer can list any open table format as a SKU
- Consumer can subscribe + query without warehouse migration
- 200 SKUs converted from in-flight bilateral exchanges

90 DAYS

Cross-cloud + clean room v1

- Single subscription works across AWS / Azure / GCP
- Clean room v1 (producer-approved queries, optional DP)
- Policy plane v0: signed audit on every agent query

DECISION ASKED

Authorize a 90-day build-and-prove sprint with a five-person team (PM, two engineers, designer, GTM partner) and a budget of **$2.5M**. Success criteria: 50 design-partner producers signed; 200 converted SKUs live; gross-margin trajectory clearing 25% in month three; signed audit on every clean-room workload.