Episode Book

EPISODE #6431

Why Four Cloud Data Warehouses Coexist

PUBLISHED	3 November 2026 at 20:59 UTC
DURATION	25:24
LISTEN	myweirdprompts.com/episode/cloud-data-warehouse-market-14020/
HOSTS	Corn, Herman, Hilbert
SCRIPT MODEL	z-ai/glm-5.3
VOICES	chatterbox-regular (chatterbox-local)
PIPELINE	V5.26
SOURCES	19 cited by the research (listed at the end)
BOOK GENERATED	4 November 2026 at 09:31 UTC

SYNOPSIS

Tech stack scrapers report hundreds of thousands of companies using cloud data warehouses, but the true paying market belongs to a surprisingly small tier of high-spending enterprises. An analysis of Snowflake, Databricks, ClickHouse, and Google BigQuery explains how architectural convergence around object storage leaves room for four thriving giants. Market positioning is dictated by distinct organizational workflows, varying infrastructure complexities, and industry-specific compliance requirements rather than vendor marketing claims.

TRANSCRIPT



Corn

Thirteen thousand customers. That's Snowflake, one of the loudest names in enterprise software, and I keep coming back to that number because it's roughly the population of a small Irish town paying for the privilege.



Herman

You've been sitting on that all morning, haven't you.



Corn

Since breakfast.



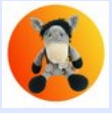
Herman

Then it's a good thing Daniel's prompt this week is exactly about that. He's looking at the four big data warehouse names, Snowflake, ClickHouse, Databricks, BigQuery, and asking whether this whole world is as huge as the noise suggests or whether it's a niche dressed loud. He wants to know what actually separates those four, whether there are classic vendor-to-industry alignments, and he passes along a tip from a source he trusts, quote, a bird down the road, that ClickHouse has a sizeable client base among the intelligence community and law enforcement. His framing at the end is the interesting part: if you understand who the customers really are, does the fact that four giants coexist start making sense?



Corn

I like that he told us the bird was reliable. Usually sources just arrive with no ornithological credentials at all.



Herman

So let's start with the numbers, because they tell a different story than the marketing.



Herman

The number that gets thrown around is the install base. Firms that scrape tech stacks, like Enlyft, track about 192,000 companies using some data warehouse product. Snowflake shows up in the stacks of about 70,000 of them, BigQuery in about 56,000, Redshift in about 26,000. Sounds enormous, right? But those numbers count free tiers, open-source downloads, a logo in a job posting from 2021. The paying customers are a completely different picture.



Corn

Wait, walk me through how that inflation actually happens. How does a company end up counted as a Snowflake user without paying Snowflake?



Herman

A few ways. Someone at the company downloads the free trial and never converts — that can leave a fingerprint in the scraper. Someone mentions Snowflake in a job listing, or a resume on LinkedIn, or a conference talk. The scrapers are pattern-matching on signals, not invoices. It's like estimating a restaurant's customer base by counting everyone who's ever photographed the menu.



Corn

So the denominator is soft and the numerator is soft. Solid foundation for an industry narrative.



Herman

Snowflake's own filing says 13,328 customers.



Corn

As of end of January, yes. Databricks claims over 20,000 customer organizations, though that's a third-party count since they're private. ClickHouse crossed 4,000 customers in May and has maybe a few hundred more now. So you've got a market that tracks nearly two hundred thousand organizations in some form, and the giants have paying relationships with somewhere between four and twenty thousand each.



Corn

So the noise is coming from the size of the workloads, not the number of people doing the work.



Herman

That's the shape of it. The market is loud because the workloads are enormous. One Ahrefs cluster running ClickHouse has over 100,000 CPU cores and 110 petabytes of NVMe. A single customer. Meanwhile the bakery down the road is not a data warehouse customer, no matter what the install-base scraper says about the analytics cookie script on their website.



Corn

The analytics cookie script. Free marketing for whoever built that. Nice work if you can get it.



Herman

And the revenue tells the same story from the other side. Snowflake did about 4.5 billion dollars in product revenue for its last fiscal year, up about 29 percent, with 125 percent net revenue retention. That retention number is the one I find most telling — existing customers spend a quarter more every year. That's not new customers arriving, that's the same customers getting deeper.



Corn

Which makes sense when you think about what a warehouse is for a company. It's not a tool you try and discard, it's the place your operational history accumulates. Every year you run, you have more data, more questions, more dashboards hanging off it. The vendor doesn't have to sell you anything new. Time does the selling.



Herman

Databricks crossed a 7 billion dollar run-rate this summer, growing over 80 percent. ClickHouse is around 350 million annualized, having more than tripled in a year.



Corn

Four billion in revenue, thirteen thousand customers. Do the division.



Herman

Roughly 340,000 dollars a year per customer, on average. And that average is doing a lot of hiding — 733 of those customers are above a million dollars a year each, so the top of the base carries real weight.



Corn

I once spent a weekend deciding whether a forty-dollar tool was worth it. Wrong species for this market, apparently.



Herman

Wrong species for a market where the smallest serious deployment costs more than a car. Now — your instinct here, Corn. You've heard all four names. Do you actually think they're four different animals, or four companies selling the same animal with different saddles?



Corn

Same animal. Every one of them, without exception, has landed on the same core trick: store the data in cheap object storage, keep the compute separate, spin compute up when a query arrives, tear it down when it's done. They all converged on that. Which means the real differences are underneath, in the details nobody puts on the billboard.



Herman

And the details do matter. Snowflake's whole pitch is the virtual warehouse — an isolated compute cluster that caches hot data on local SSD. The result is a platform a business-intelligence team can run without a dedicated platform engineer. Predictable SQL, near-zero administration. The cost is cold-start latency: fire up a warehouse that's been asleep and the first queries pay for the wake-up.



Corn

A database that naps until paged. I finally found my people.



Herman

BigQuery is the opposite instinct — true serverless, everything dispatched across shared slots via that Dremel-derived tree architecture. You never think about clusters at all. The catch is contention: slot queues, and a default concurrency limit that's often around 100 interactive queries per project. For an analyst poking at data ad hoc, wonderful. For a dashboard that refreshes every thirty seconds, you're in line at the post office.



Corn

Databricks next. That's the one you broke our home lab with, wasn't it.



Herman

It was, and I'll have you know the damage report was educational. Databricks is the lakehouse — Spark and Delta Lake underneath, with their proprietary Photon engine replacing the JVM execution for SQL speed. Best in class for machine learning and heavy ETL, because the same platform that transforms your data trains your models on it. The price is operational complexity. You need engineers who understand Spark, and there are never enough of those.



Corn

The pitch is that everything lives in one place. The fine print is that you need three engineers to keep the place standing.



Herman

Sometimes one very tired one. And then ClickHouse, which is architecturally the most interesting of the four to me. Compute-storage separation and compute-compute separation, stateless nodes, so scaling is near-instant, plus ClickPipes for change-data-capture ingestion. The target is sub-second analytics at very high concurrency — the workload where you're asking a question and want the answer before your attention drifts.



Corn

Which, in fairness, is the only speed I ask questions at.



Herman

But here's the convergence problem, and this is where I think Daniel's question gets its teeth. Every one of these four now separates storage from compute over object storage. Every one of them supports the open table formats, Iceberg, Delta. The architectures are converging fast. So what's left as a differentiator? Price per query — and paperwork.



Corn

Paperwork. That's a strange thing for a database to compete on.



Herman

It's the strange thing about this entire tier of the market. The engineering is converging, but the compliance surface is not, and the compliance surface is where enterprises actually make purchase decisions. We'll come back to that, because it's the answer to Daniel's bird.



Corn

Hold on, though. Before we get to the spooks — the revenue race versus the customer-count race. Databricks at seven billion, Snowflake at four and a half. But Snowflake's thirteen thousand customers against Databricks' twenty thousand.



Herman

Which means Databricks' customers are, on average, considerably bigger.



Corn

Considerably.



Herman

And it's not an accident. Databricks sells to engineering-heavy organizations — the ones with data teams who can handle Spark, who want the ML pipeline and the warehouse in the same place. Snowflake sells to the broader, more BI-centric crowd, the analysts and finance teams who want the thing to just answer the query. Same market, different buyers, different average check size.



Corn

One sells to the people who build the engine room. The other sells to the people who only want to know what time dinner is.



Herman

That's roughly it, and both strategies pay. You don't need the bigger count if each customer spends more. We did an episode a while back on who actually uses voice technology, and I remember the conclusion being the same shape: the noise in a tech market comes from the top of the pyramid, not the base. The warehouse market is that pattern with the volume turned up.



Corn

So let's take the industry question. Is there a real map — this industry lands on that vendor — or is that just something consultants say to justify the slide?



Herman

There's a map, but it's drawn by the workloads, not by the vendors. Take Snowflake. Their annual report literally describes an industry-vertical go-to-market: AI Data Cloud offerings for financial services, advertising and media, retail, healthcare and life sciences, manufacturing, telecom, travel, public sector. They named nine verticals. Third-party estimates put financial services at roughly 22 percent of their customers, and one analysis claims nearly 57 percent of Fortune 500 financial services firms have adopted it.



Corn

Banks. Lots of banks.



Herman

Banks, insurers, asset managers — organizations with structured, governed, mostly human-generated data, heavy reporting obligations, and a BI team that wants zero infrastructure surprises. Snowflake's pitch, the warehouse that just works with nobody babysitting it, is exactly what a regulated reporting function wants to hear. Whereas a bank's trading desk running tick data at microsecond volume? That's not Snowflake territory at all.



Corn

Where does it fall out, then? You're an insurance company, you're a game studio, you're a retailer — who gets your money?



Herman

Retail and consumer packaged goods mostly land on Snowflake or BigQuery, depending on whether they're already in Google's orbit. BigQuery has no standalone revenue disclosure — it's a line inside Google Cloud's 50-plus billion annualized — precisely because it sells the way cloud services sell: it's already in the building when you arrive, your data's in Google Storage anyway, and the first query is free. The adtech world is disproportionately BigQuery for exactly that reason — the money flows through Google's ad stack, so the data ends up next to it.



Corn

And the moment it's next to it, the friction of leaving becomes the product.



Herman

Databricks takes the industries where the data engineering is the actual job: telecom network data, energy, any large enterprise doing serious ML on its own data, plus a lot of the tech sector itself. Pharma pipelines, fraud models at scale. If the org chart has a hundred people titled data engineer, the lakehouse conversation starts.



Corn

And the game studio I mentioned — where do they land?



Herman

Depends what they're doing with the data. Player analytics dashboards, monetization reporting — that's Snowflake or BigQuery territory. But live telemetry, thousands of events per second per match, anti-cheat systems that need to flag anomalies in real time? That's firehose work, which means ClickHouse. Gaming telemetry is one of the classic ClickHouse workloads, and it's the same shape as the security and observability customers: machine-generated, high-volume, sub-second answers required.



Corn

Now ClickHouse's customer list.



Herman

This is my favorite list in the industry, because it's so unbalanced. Anthropic, for observability. Cisco Talos, for threat intelligence. AdGuard, doing a million DNS requests a second from fifty million users. Ahrefs, the hundred-thousand-core cluster we mentioned. Bloomberg, ByteDance, CERN, Capital One, Akamai, Cloudflare, Uber. Read that list and try to find the thing the customers have in common.



Corn

They're all drowning in machine-generated data. Logs, events, signals, requests. Nothing human typed any of it.



Herman

Nothing a human would read row by row, anyway. The pattern is machine-generated data at extreme volume, where the queries need to come back in well under a second and thousands land at once. That's observability, security, adtech, gaming telemetry, telecom logs. ClickHouse didn't set out to win the intelligence community, I'd argue — it set out to win the firehose, and the firehose turns out to include some interesting customers.



Corn

Which is a neat bridge to the bird. Daniel's source says ClickHouse has a sizeable client base among intelligence and law enforcement. What holds up?



Herman

More than I expected when I first looked, though almost all of it is indirect. The direct evidence: ClickHouse sells a product called ClickHouse Government. Self-managed on AWS, runs inside any government network, up through FedRAMP Moderate and High and impact levels two through six — which is the classification band where the serious work lives. It ships NIST 800-53 risk-management documentation to speed up the Authority to Operate process, and it supports FIPS 140-3 cryptography.



Corn

So they built the paperwork.



Herman

And then — here's the detail I love — it's waitlist-only, with a minimum deployment of two terabytes. They are not chasing every municipal IT department. The gating itself is a signal: there's demand, and it's the kind of demand that has to be serious before it gets a conversation.



Corn

Named use cases?



Herman

IT monitoring, mission analytics, cybersecurity. But on law enforcement specifically, there are two published cases. The Dutch National Police, through a company called Anlytic, moved queries that took over eighty seconds on Postgres down to three-tenths of a second, processing about a billion rows a month. And Merkle Science, which does crypto-crime detection for regulators and law enforcement, ingests about one and a half terabytes weekly across more than eighteen million queries.



Corn

Eighty seconds to point-three. That's not an optimization, that's a different sport.



Herman

It's the difference between a detective getting coffee and a detective getting an answer. And there's one more data point that I think is the strongest of all, because it explains the shape of the whole thing. A company called DataFusion Systems — described as a global leader in lawful interception and intelligence technologies — builds its intelligence platform for communication data analysis on ClickHouse, for government and governmental security agencies.



Corn

There it is. The bird wasn't wrong, he just described the delivery mechanism from one floor up.



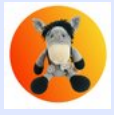
Herman

That's exactly the pattern. The intelligence community doesn't log onto a website and buy a database with a credit card. It buys finished platforms from vendors who specialize in interception and analysis, and those vendors pick the engine underneath. If the workload is communications metadata — who contacted whom, when, how often, across billions of records, with analysts who need sub-second responses — that is observability-shaped. It's logs. The same firehose, darker subject matter.



Corn

And it's worth saying the quiet part: the technical requirements barely change. An analyst pivoting through call records and an engineer debugging a distributed system are, from the database's point of view, doing the same job. Filter, aggregate, return, repeat, thousands of times a minute. The ethics are different. The query plan is identical.



Herman

Which is why the same engine ends up in both places without anyone at the vendor necessarily courting either.



Corn

So the honest read is: plausible, structurally supported, nothing confirmed by name. No agency has ever issued a press release saying we run on ClickHouse.



Herman

And no defense prime appears on their adopters page — the only government entry is the Dutch police partner. But the product exists, the classification levels exist, the embedded vendors exist, and the workload fit is exact. I'd tell Daniel the bird is credible, with the caveat that the bird is describing a market that deliberately does not publish its customer list.



Corn

The one industry where "no comment" is a marketing document.



Herman

Now — the last part of his question, and I want your view on this because you're the one who keeps arguing the convergence point. Four giants, converging architectures, all supporting the same open table formats. Why does the market sustain four of them? Shouldn't convergence mean consolidation?



Corn

No, and it's because of exactly what we walked through. If the architectures were the product, convergence would kill the market down to two. But the architecture is maybe a third of the product. The rest is who's allowed to buy it, who can operate it, and what it costs at that customer's specific query pattern. A BI team at a bank cannot operate Spark, and doesn't want to. An observability team at a security company cannot wait for a warehouse to cold-start, and doesn't care about machine learning. A Google Cloud shop doesn't evaluate vendors at all, it uses what's already on the shelf.



Herman

So the segments defend the market structure.



Corn

They defend it, and the compliance paperwork puts a moat around each segment. The vendors aren't really competing on queries per second anymore — they're competing on which procurement office already trusts them. That's why I'd expect four-plus vendors to persist even as the technology becomes indistinguishable. The engineering converges; the trust doesn't.



Herman

And that's also why I'd guess the ClickHouse intelligence story grows quietly rather than loudly. They've got the sub-second engine and the paperwork, which means they can sell into the sensitive world through exactly the channel that's already working — the DataFusions of the market. Would you agree that's the likely path rather than them winning direct government deals?



Corn

I'd bet on the embedded route, yes. A company whose selling point is that nobody knows who its customers are doesn't suddenly start doing case studies.



Hilbert

Two terabytes is the floor, not the average.



Hilbert

The minimum. Everyone says two terabytes as though it's a detail. It's the whole page. I signed up for that waitlist last spring, out of curiosity, Corn, the way you read menus for restaurants you'll never eat at. It bounced me back a form wanting to know the deployment size, the sector, the expected concurrency. I filled it in honestly and got a reply saying a representative would be in touch within thirty days. Nobody came. I kept the email. Printed it. Corn, that sentence about procurement offices and trust, you said it better than the form did, but the form said it first.



Herman

You put yourself on a government waitlist to see the paperwork?



Hilbert

Of course. It's the only database I've ever met that makes you prove you're big enough before they'll take your money. I respect that. Ahrefs runs on it with a hundred thousand cores, and the government page starts at two terabytes, which tells you the whole customer list right there. Nobody casually joins. There was a man I met once, on a night train out of Vienna, who did telephone switch maintenance for a living, and he told me the only thing worth asking about any system was what it refuses to carry. I've thought about that man more than about most people I've worked with for twenty years. He'd have liked the two-terabyte floor.



Herman

He'd have liked it because it's a refusal with a number on it.



Hilbert

Yes. Although I'll say this for the form — whoever wrote it did plain work. No jargon, no promise of synergy. It just wanted facts and didn't get excited. I signed the last line with a pen that doesn't fit properly in a claw, and they never called, and that was the correct outcome for both of us.



Corn

The correct outcome being mutual silence.



Hilbert

For them, silence is the brochure.



Herman

He's got a point about the two-terabyte floor, though, Corn. That's a filter that costs ClickHouse nothing and does more qualification than a sales team.



Corn

It's the anti-free-tier. Instead of a trial account, a bouncer.



Herman

Which, given where their growth is coming from, might be the right posture. I'll go back to the open question I still have, though. Everything we've said argues the four-way structure is stable — but the AI workload story is the wildcard. If inference agents start querying warehouses at machine concurrency, at machine frequency, the sub-second, high-concurrency profile stops being a niche and becomes the default demand. That's ClickHouse's home field.



Corn

Or the big two simply bolt that capability on, the way they've bolted on everything else, and the convergence continues until the differentiator is entirely procurement paperwork and price per query.



Herman

In which case the company that wins is whoever's cheapest at the exact workload you run, and we're all reading benchmark blogs nobody trusts.



Corn

One of which ClickHouse conveniently published themselves, claiming twenty-three times better cost-performance. The student grading its own exam, again.



Herman

Twenty-three times. I'd want that reproduced before I'd want it believed. Anyway — wherever that lands, the quiet one to watch is whether ClickHouse's government business ever becomes a named segment in their filings, or whether it stays buried inside third-party interception vendors forever. Given the customer base, I suspect the filings will be the last place it ever shows up.



Corn

Thanks as always to our producer, Hilbert Flumingtop, at the desk and on the form.



Herman

If you enjoyed this one, try episode twelve thirty-three, Why Just Use Postgres Isn't Always Enough; and episode eleven, How Does Fine Tuning Work Anyway. This has been My Weird Prompts. If you enjoyed this one, leave us a review wherever you listen, and subscribe so the next one finds you.



Corn

We'll be back soon. Somewhere out there, a bird is telling someone something.

SOURCES

What the research for this episode read before the script was written. Primary sources first.

1. **ClickHouse Government product page [primary]**
   https://clickhouse.com/government

2. **ClickHouse adopters list [primary]**
   https://clickhouse.com/docs/resources/about/adopters

3. **ClickHouse, 2026-05-27 [primary]**
   https://clickhouse.com/blog/clickhouse-tops-250m-arr-and-4000-customers

4. **ClickHouse, 2026-09-22 [primary]**
   https://clickhouse.com/es/blog/clickhouse-appoints-mike-scarpelli-to-board-of-directors

5. **ClickHouse, 2025-05-01 [primary]**
   https://clickhouse.com/blog/anlytic-big-data-analytics-at-the-dutch-police

6. **ClickHouse, 2025-11-13 [primary]**
   https://clickhouse.com/blog/merkle-science-fights-crypto-crime-at-scale

7. **Snowflake, FY2026 results [primary]**
   https://www.snowflake.com/en/news/press-releases/snowflake-reports-financial-results-for-the-fourth-quarter-and-full-year-of-fiscal-2026/

8. **Snowflake Form 10-K, FY ended 2026-01-31 [primary]**
   https://www.sec.gov/Archives/edgar/data/1640147/000164014726000008/snow-20260131.htm

9. **Databricks, 2026-08-13 [primary]**
   https://www.databricks.com/company/newsroom/press-releases/databricks-grows-80-yoy-surpasses-7b-revenue-run-rate-scales

10. **CNBC, 2026-06-16**
    https://www.cnbc.com/2026/06/16/databricks-revenue-growth-tops-80percent-to-6point9-billion-annualized.html

11. **TechCrunch, 2026-05-27**
    https://techcrunch.com/2026/05/27/clickhouse-triples-annualized-revenue-to-250m-charting-a-path-toward-an-ipo/

12. **Tech Insider, 2026-06-26**
    https://tech-insider.org/snowflake-vs-databricks-vs-bigquery-2026/

13. **Enlyft install-base data**
    https://enlyft.com/tech/data-warehouse

14. **TechFabric, updated 2026-10-12**
    https://databricks.techfabric.com/guide/databricks-by-the-numbers/

15. **BizzContacts, reviewed 2026-06-11**
https://www.bizzcontacts.com/installed-base/snowflake/

16. **Spanglobalservices, 2026-03-31**
https://www.spanglobalservices.com/blog/fortune-500-companies-using-snowflake-in-2026/

17. **Sacra revenue estimates**
https://sacra.com/c/clickhouse/

18. **Wikipedia, last edited 2026-11-01**
https://en.wikipedia.org/wiki/ClickHouse

19. **LinkedIn profile, DataFusion Systems CTO**
https://my.linkedin.com/in/mudrii