Episode Book

EPISODE #6430

Why Data Pipeline Tools Cost Billions With So Few Engineers

PUBLISHED	3 November 2026 at 20:57 UTC
DURATION	24:20
LISTEN	myweirdprompts.com/episode/data-pipeline-etl-headcount-61587/
HOSTS	Corn, Herman, Vex, Hilbert
SCRIPT MODEL	z-ai/glm-5.3
VOICES	chatterbox-regular (chatterbox-local)
PIPELINE	V5.26
SOURCES	25 cited by the research (listed at the end)
BOOK GENERATED	4 November 2026 at 09:31 UTC

SYNOPSIS

ETL and data pipeline platforms command billions in venture valuations and market forecasts, yet data engineers represent a tiny fraction of global software developers. The gap reveals an industry built to address corporate fragmentation rather than pure technical scale. This look into the ETL landscape breaks down the five distinct product categories—from ingestion engines to orchestrators—and examines how consolidation, usage-based pricing, and automation shape modern enterprise data architecture.

TRANSCRIPT



Corn

Twelve billion dollars, minimum. That's the floor on what the analysts say this category is worth, and I'd never heard a normal person mention it at dinner.



Herman

Which category?



Corn

ETL. Data pipeline tools. Fivetran, Informatica, Airflow, dbt — the entire shelf. That's actually why I had it on my mind, because Daniel's written in on exactly this. His angle is that it's a whole industry of big names, and yet from the outside it feels niche. So he wants to know why the category exists at all, in what contexts, how many people are really moving data at the scale that justifies any of it — and then, from that vantage point, the top tools and who actually needs them.



Herman

He's put his finger on the thing that makes this field strange to cover, which is that the gap between the noise and the headcount is enormous. You've got a market the analysts size at twelve to fifteen billion dollars, a dozen well-funded vendors, mergers in the billions — and then you look at who does the work and it's a rounding error among developers.



Corn

One point nine percent, if Stack Overflow's survey is to be believed.



Herman

Of forty-nine thousand respondents, yes. So before any tool names, the question is why this category is necessary at all. And the answer is boring, which I find convincing — boring answers usually are.



Corn

Go on.



Herman

Enterprise data doesn't live anywhere. That's the whole thing. Your payroll runs on one vendor's system, sales runs on another, purchasing on a third, each on separate hardware, each owned by a different department with a different person who says no. The textbook example is a cost accounting system: to answer one question — what did it actually cost us to run this business — you need payroll, sales and purchasing data that were never designed to sit in the same room. ETL is the machinery that reconciles that fragmentation into a single target. A warehouse, a data mart, a lake.



Corn

So it's not a technology problem, it's an organizational one wearing a technology costume.



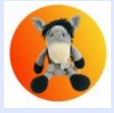
Herman

Largely, yes. If your company runs one database, none of this exists. The moment you have a second system owned by a second team, you have an integration problem, and it only grows from there. Forrester did a survey years back — I want to say twenty fifteen — that found half of companies were blending fifty or more data sources, and twelve percent over a thousand.



Corn

A thousand sources.



Herman

At that point you're not hand-writing anything. And the pain is measurable in the vendors' own numbers, which is the only kind of vendor number I half trust, because it's marketing that admits the product is needed. Matillion surveyed around three hundred data decision-makers last year: seventy percent called pipeline management somewhat or extremely complex, sixty-four percent said their teams spend over half their time on repetitive tasks. Fivetran's older survey, from twenty twenty-one, had ninety-eight percent reporting pipeline-building challenges and about half saying pipelines break monthly or more.



Corn

Ninety-eight percent of people who bought the tool that solves this say it's hard. That's either honesty or the best upsell ever written.



Herman

It's both, which is my favorite kind of statistic.



Corn

There's a definitional trap here too, before we get deeper. "ETL tool" is doing duty for at least five different products. Ingestion tools that copy data into a warehouse. Transformation tools that run inside it. Full enterprise suites. Orchestrators that don't move data at all. And the cloud-native services each hyperscaler bundles. Comparing across those kinds is where most bad buying decisions come from, and it's why any article that ranks "the top ten ETL tools" as if they're one species is basically worthless.



Herman

Right — the acronym stopped meaning the architecture it describes too. Classic ETL transforms before loading. The cloud-data-lake era flipped it to ELT: you land the raw data first and run the transformations inside the warehouse engine, on demand. That fork is the whole reason dbt exists as a company.



Corn

dbt being the transformation layer that assumes somebody else did the loading.



Herman

Yes. It runs SQL inside the warehouse after the data's already there. It transforms only, it needs an ingestion tool, and together those two facts built an entire ecosystem of companies pointing at each other.



Corn

So the category exists because of fragmentation, the tooling is a mess of five species wearing one label, and the market-size numbers — tell me about those, because you said twelve to fifteen billion with a face that suggested you didn't believe it either.



Herman

I don't believe any single number, and the disagreement itself is the finding. Grand View Research says the data pipeline tools market was about fourteen point seven billion in twenty twenty-five and grows at thirty percent a year to over a hundred billion by 2033. Fact.MR says twelve billion growing at twelve percent. The Business Research Company, using a narrower definition of just ETL software, says under six billion. Grand View's trajectory is roughly two and a half times Fact.MR's for the same physical reality.



Corn

Two analysts, same market, one of them predicts a world two and a half times richer. That's not a forecast, that's a temperament.



Herman

It tells you the category is being sized by narrative. When the story is "data is the new oil," you get thirty percent compounding. When the story is "this is plumbing," you get twelve.



Corn

Which takes us to Daniel's actual suspicion — how many people are really doing this? Because the noise suggests a city, and the headcount suggests a village.



Herman

A small village. Stack Overflow's twenty twenty-five survey, forty-nine thousand developers: one point nine percent call themselves data engineers. StartUs Insights puts the global data engineering workforce at just over a hundred and fifty thousand people, growing at about twenty-three percent. Boot.dev counted around fifteen hundred job postings in a November snapshot, average US salary a hundred and fifty thousand dollars.



Corn

A hundred and fifty thousand people worldwide. That's a mid-sized software company's worth of practitioners supporting a twelve-billion-dollar shelf.



Herman

And there's a counter-signal inside the same survey, which is what makes it interesting rather than just damning. Data engineers had the lowest salary growth of any role in Stack Overflow's twenty twenty-five results — five percent, against twenty-nine percent for product managers. If demand were exploding the way the market reports imply, pay would move first.



Corn

Unless the role is being automated by the very tools we're discussing.



Herman

Which is a real possibility! Every one of these products is, at bottom, a machine for needing fewer hands per pipeline. The vendor pitch is literally "what used to take months now happens in weeks" — that's a Zendesk data engineering director quoted in the Fivetran merger materials, and it's meant as praise.



Corn

The most honest sentence in the whole press release and nobody noticed.



Herman

So the numbers can be reconciled: the category is real, the spend is real, and the workforce is thin because the entire value proposition is thinning it. My question to you, though, Corn — you've run your own pipelines for the show, the n8n setup, the episode indexer. Where would you put the threshold? When does somebody actually need this stuff?



Corn

Boringly low, honestly. Multiple source systems owned by different teams, plus a warehouse, plus someone who needs the numbers every day. All three or none. One database and one SaaS app — and that describes an enormous number of businesses — a cron job and some SQL is the honest answer, and anything more is decoration.



Herman

That's the conclusion I'd defend too. Below that threshold, none of this. Above it, the real-world pattern isn't one tool anyway — it's one ingestion tool, plus dbt, plus an orchestrator. Nobody buys one product and is done.



Corn

So the category clearly has a lot of noise around it, but I keep coming back to Daniel's question of who is really doing this work. Because on the headcount evidence, it's either a hundred and fifty thousand engineers or a great many refrigerators.



Herman

...I'm sorry, we have a caller, and I recognize him, because it's him again. Corn, how does he keep getting through? We changed the screening number after the condiment incident.



Corn

He memorized Hannah's extension last spring. I've stopped asking how.

v

Vex

Listen to me, because I have been saying this for years and nobody will print it. You know that hum your fridge makes, that low steady drone at night, that is not a compressor, that is a data pipeline running in your kitchen, and I have the receipts, I have the maintenance logs, I have the little sticker on the back that nobody reads!

Vex

And that is why the whole ETL industry exists, because every appliance in every home is quietly reconciling your grocery receipts into a warehouse nobody asked for, and the one point nine percent of data engineers are just the ones who noticed, so I need you two to look into this on air, right now, loudly!



Herman

The hum is a compressor, Vex. It's on the sticker. The fridge is the one appliance not sending your groceries anywhere — that honor goes to the smart meter next to it.



Corn

Moving on before he asks about the washing machine.



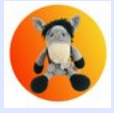
Herman

So — the tools. Five species, one label. Let me walk them. Ingestion, the EL tools: Fivetran, Hevo, Airbyte, Meltano, dlt. Their whole job is copying data out of SaaS apps and databases into a warehouse. Transformation: dbt, as covered. Enterprise integration suites: Informatica, Qlik Talend, Matillion — the ones that bundle ingestion, quality, governance, the works, and bill accordingly. Orchestrators: Airflow, Dagster, Prefect — and these don't move any data themselves, they schedule, retry and monitor the steps that do. And cloud-native: AWS Glue, Azure Data Factory, Google's Dataflow and Datastream.



Corn

The pricing tells you what each species is afraid of. Fivetran bills by monthly active rows per connection, free tier half a million rows. Which means your invoice scales with usage you can't forecast.



Herman

That's the documented trade-off, yes. Airbyte gives away the self-managed core under an elastic license and sells cloud credits. dbt core is free Apache two-point-zero, and the platform starts around a hundred dollars a seat plus what you build. Dagster's hosted tier is a hundred a month plus credits per materialization. Airflow is entirely free — you pay only if you want somebody else to run it, which is the cleanest licensing story in the category.



Corn

And Informatica bills in "Informatica Processing Units."



Herman

They do. A currency you cannot convert to dollars without a sales call.



Corn

The tell of every enterprise contract: the vendor invented its own money.



Herman

Then there's the graveyard side of the shelf. Talend Open Studio, the free edition, was retired at the end of January twenty twenty-four. Stitch got absorbed into Qlik and isn't recommended for new projects. If you built on the free tier of either a few years ago, you've been quietly migrated into a pricing conversation.



Corn

And that gets us to the actual story of the last eighteen months, which is consolidation. Fivetran and dbt Labs completed their all-stock merger on the first of June this year. Fivetran was valued at five point six billion back in twenty twenty-one, dbt at four point two in early twenty-two — and the combined business is projected at "approaching" six hundred million in annual recurring revenue.



Herman

Read those numbers twice, because the spread is the story. Nearly ten billion of peak paper valuation against six hundred million of actual revenue, and the merger was all-stock, meaning both sides looked at that gap and decided their best move was to stand closer together.



Corn

It's the corporate version of two people in a sinking boat agreeing the boat is fine.



Herman

Salesforce closed its Informatica purchase last November for a reported eight billion, folding catalog, quality, governance, MDM into Data 360 and Tableau. Prefect bought Dagster in July. Airbyte stands alone — a hundred and eighty-one million raised, a one and a half billion valuation from the peak, and about a hundred and twenty-six employees, which at that valuation is roughly twelve million dollars of paper value per person.



Corn

I'd want a very good chair at that company.



Herman

And here's what the consolidation means, to me. The "modern data stack" — the whole pitch of the last decade — was best-of-breed point tools wired together by you. That pitch is now collapsing into suites. Fivetran absorbed its own complement. Salesforce bought the suite that already was one. The category that sold itself on "nobody bundles this well" is bundling, because the buyers got tired of assembling furniture.



Corn

There's a version of that argument we've made before, on the redaction pipeline episode — the choice of what gets transformed where always has consequences downstream that the tool doesn't advertise. The suite hides that choice entirely.



Herman

The orchestration fight is a nice little case study in contested metrics, by the way. GitHub stars: Airflow about forty-six thousand, Prefect twenty-three and a half thousand, Dagster just under sixteen thousand — so Prefect plus Dagster combined still trails Airflow alone, which prompted one writer's line that Airflow is a nine-year-old project neither company managed to out-adopt alone, and now can't out-adopt together either.



Corn

But?



Herman

But weekly PyPI downloads as of October have Dagster at one point nine million, Prefect one point seven, Airflow one point six. Prefect's CEO points to that as the new center of gravity — with the caveat that the PyPI counting methodology changed, and some trackers still show Airflow well ahead on monthly downloads. There is no neutral referee. The two public metrics point in opposite directions, and each side quotes the one it likes.



Corn

Two teams, one scoreboard each, both leading.



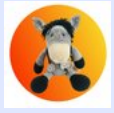
Herman

What I find more consequential is the messaging pivot, because it's happened everywhere at once. Fivetran's merger release doesn't talk about analysts anymore. It says AI agents are quickly becoming the primary consumers of enterprise data, and that they behave differently from the human analysts the stack was built for — agents operate continuously, in parallel, at machine speed. And dbt State is pitched as a caching layer that only builds what's changed and skips what hasn't, cutting infrastructure costs by thirty percent or more.



Corn

My question about all of that is whether an agent reading a warehouse actually needs a different pipeline, or whether the vendors need a different story after the best-of-breed one expired.



Herman

I don't know, and I've gone back and forth. There's a real technical claim buried in there — a human analyst runs a query at nine in the morning and goes home; a thousand agents polling continuously does change the load pattern, and caching what hasn't changed is a legitimate answer to that.



Corn

It's also a legitimate answer to a bill that grew out of nowhere. "Only build what's changed" is what a warehouse does when it's trying not to be blamed.



Herman

That's the cynical reading and I hold it about forty percent. The other sixty percent says agents will consume data differently and the first vendor to serve that well wins the next decade — which is what every one of these mergers is betting on. Ten billion of combined paper valuation is a very large bet on a customer who hasn't shown up yet.



Corn

So who actually needs these tools, now that we've got the taxonomy in hand?



Herman

Same threshold as before, and I'd keep it strict. Multiple sources with different owners, a warehouse, daily consumers. Below that, none of it. Above it, the pattern is still one ingestion tool, dbt, and an orchestrator — except the mergers are now pushing against exactly that pattern, which is the tension to watch. Fivetran plus dbt claims over a hundred thousand data teams globally, and that's the loosest possible unit of measurement — a "team" can be three people at a logistics firm or one analyst who knows SQL — but it's the best upper bound anyone's offered on the addressable population.



Corn

A hundred thousand teams at a hundred and fifty thousand engineers. The arithmetic there is either very flattering to the engineers or very loose about what a team is.



Herman

Both, I'd say. And it's the number I'd least trust, since it's the one that prices the IPO story.



Corn

The other thing I'd hold onto from the pricing discussion is that the honest cost of this category is never on the pricing page. Rows per month, seats, credits per materialization — every one of those scales with success, which is a strange way to sell infrastructure. The more your business grows, the more moving your own data costs you.



Herman

Which is the answer to why the market reports disagree so wildly, actually. If billing is usage-based and usage compounds with adoption, the optimists just extrapolate the invoice. Thirty percent a year isn't a prediction about the world, it's a prediction about the meter.



Corn

The meter with no ceiling on it. Lovely industry you've chosen to defend.



Herman

I don't defend it, I describe it. There's a difference, and the meter can tell.



Hilbert

The meter never sleeps, that's the part your download counts missed. Herman, you said agents poll continuously — well, the rows don't know it's an agent. We priced one at the club. The darts team, back in the spring, because Reg at the team wanted the scores in a dashboard instead of the chalkboard, and I got a quote through for the ingestion side. Monthly active rows, per connection. It came to more than the shirts, and I said no, and we kept the chalk. I still know the number. Four hundred and ten a month, and that was before the scores went in — that was just the roster and the fixture list moving once a week. We call it the sugar wagon at the club, Reg and me, because it comes around whether you ordered anything or not.



Corn

The sugar wagon.



Hilbert

That's what we call the bill. Reg's got it written inside the scoreboard lid. And the bit nobody's priced — Corn, you were nearly there with the meter — if the agents do start reading the warehouse at machine speed, they'll be chalking up active rows at three in the morning. Continuous polling means continuous billing. Every agent left running overnight is a small customer you've enrolled without asking.



Herman

That is a ugly corner of the pricing model I hadn't followed to the end. Active rows don't care whether the reader has a pulse.



Hilbert

No. And I didn't want the wagon in the first place, which is why I never bought one. The quote's still in my coat. You two want the funny part — the chalkboard's never broken once.



Corn

Monthly breakage rate of zero. Beats the surveyed fifty-one percent.



Hilbert

Well. It has broken once, now you mention it. Somebody's elbow, final of the spring league. Cost a tin of chalk. I'll grant you the wagon wouldn't have done that.



Corn

The misconception, then, before we close it out.



Herman

Go on, which one?



Corn

That this is a mass-market category. It's the assumption underneath all the noise — that everybody's moving data at scale, so everybody should be shopping here. One point nine percent of developers, a hundred and fifty thousand people worldwide, the lowest salary growth of any role in the field.



Herman

And the second one, tangled into it — that "ETL tool" names one kind of product, when it names five. Comparing an orchestrator to an ingestion tool is how the bad decisions get made. If you remember nothing else: it's a niche trade with enterprise-grade marketing, and the threshold — multiple sources, different owners, a warehouse, daily consumers — is narrower than anything a sales call will tell you.



Corn

The open question I'd leave with Daniel is the agent one. If the machines really do become the primary readers of all this data, the hundred-and-fifty-thousand-strong workforce and the twelve-billion-dollar market might both be transitional numbers — heading in opposite directions, and there's no telling which arrives first.



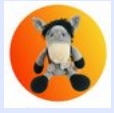
Herman

And whether the suites hold. The whole category was built on the promise that you'd assemble it yourself. Fivetran plus dbt is a bet that you'd rather not.



Corn

Our producer is Hilbert Flumingtop — the meter, for the record, has never billed us for an episode.



Herman

It's a fixed fee. He shows us the quote.



Corn

If you want more of this, try episode twelve thirty-four, Why Hashing Fails; and episode twelve, The AI Breakthrough. This has been My Weird Prompts. If the show earns its chalk, leave us a review on Spotify — it helps.



Herman

We'll be back soon. See you tomorrow.

SOURCES

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

1. **Matillion, 2025-03-24 [primary]**
   https://www.matillion.com/news/data-ai-readiness-survey

2. **Fivetran, 2021 [primary]**
   https://get.fivetran.com/rs/353-UTB-444/images/Data-Engineers-Survey-Report.pdf

3. **Fivetran, 2026-06-01 [primary]**
   https://www.fivetran.com/press/fivetran-dbt-labs-complete-merger-to-create-the-data-infrastructure-for-trusted-ai-agents

4. **dbt Labs, last edited 2026-06-08 [primary]**
   https://www.getdbt.com/blog/fivetran-and-dbt-are-one-company-now-here-s-what-that-means

5. **Salesforce, 2025-11-18 [primary]**
   https://www.salesforce.com/news/press-releases/2025/11/18/salesforce-completes-acquisition-of-informatica/

6. **Astronomer, 2026-01-22 [primary]**
   https://www.astronomer.io/press-releases/astronomer-releases-state-of-apache-airflow-2026-report/

7. **Airbyte, 2021-12-20, updated 2026-06-02 [primary]**
   https://airbyte.com/blog/a-150m-series-b-to-power-the-movement-of-data

8. **Grand View Research, Data Pipeline Tools Market 2026**
   2033
   https://www.grandviewresearch.com/industry-analysis/data-pipeline-tools-market-report

9. **Fact.MR, published 2026-07-29**
   https://www.factmr.com/report/data-pipeline-tools-market

10. **The Business Research Company, Feb 2026**
    https://www.researchandmarkets.com/reports/5971042/etl-software-market-report

11. **Market Glass, published 2026-10-01**
    https://www.marketresearch.com/Market-Glass-Inc-v1039/Data-Pipeline-Tools-46585673/

12. **TechSci Research**
https://www.techsciresearch.com/report/data-pipeline-tools-market/19512.html

13. **Wikipedia, last edited 2026-07-16**
https://en.wikipedia.org/wiki/Extract,_transform,_load

14. **Wikipedia, last edited 2026-05-13**
https://en.wikipedia.org/wiki/Extract,_load,_transform

15. **Wikipedia, last edited 2026-06-10**
https://en.wikipedia.org/wiki/Data_blending

16. **StartUs Insights, 2025**
https://www.startus-insights.com/innovators-guide/data-engineering-market-report/

17. **IT Pro, 2025-10-23**
https://www.itpro.com/business/data-and-insights/data-engineers-have-never-been-more-important-and-businesses-are-starting-to-find-out

18. **Boot.dev, last published 2026-02-21**
https://www.boot.dev/blog/jobs/supply-demand-broken-programming-education

19. **IT Brief UK, 2025-10-30**
https://itbrief.co.uk/story/uk-developer-pay-hits-record-levels-as-ai-skills-drive-demand

20. **Stack Overflow, 2025-08-01**
https://stackoverflow.blog/2025/08/01/diving-into-the-results-of-the-2025-developer-survey/

21. **codewithsql.com, verified 2026-10-07**
https://codewithsql.com/sql-tools/best-etl-tools

22. **TechTarget, 2026-06-01**
https://www.techtarget.com/news/Fivetran-DBT-Labs-complete-merger-to-form-data-layer-for-AI/842165/

23. **DataDriven**
https://datadriven.io/blog/fivetran-and-dbt-labs-merge-what-data-engineers-should-know

24. **TechTarget**
https://www.techtarget.com/news/Salesforce-adds-Informatica-to-Data-360-MuleSoft-fold/848740/

25. **TechTarget, 2026-05-20**
https://www.techtarget.com/news/Informatica-update-aims-to-provide-trust-foundation-for-AI/842176/