Cost Containment

A Critical Piece of Data Team ROI

Lindsay Murphy

Head of Data

Secoda

Instructor, Advanced dbt

Uplimit

New episodes every Wednesday!

Listen on Spotify

Listen on Apple Podcasts

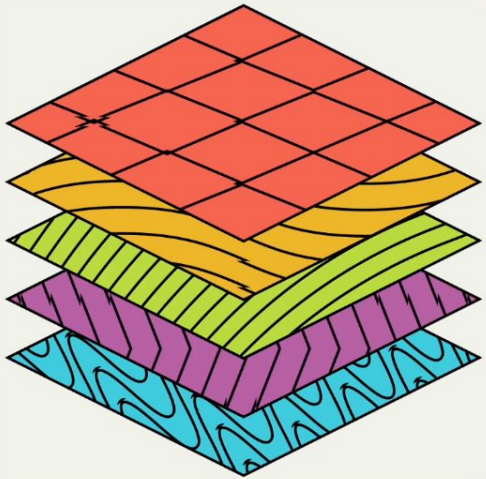
A VIRTUAL DATA CONFERENCE

PRESENTED BY Secoda

MDS FEST 2.0

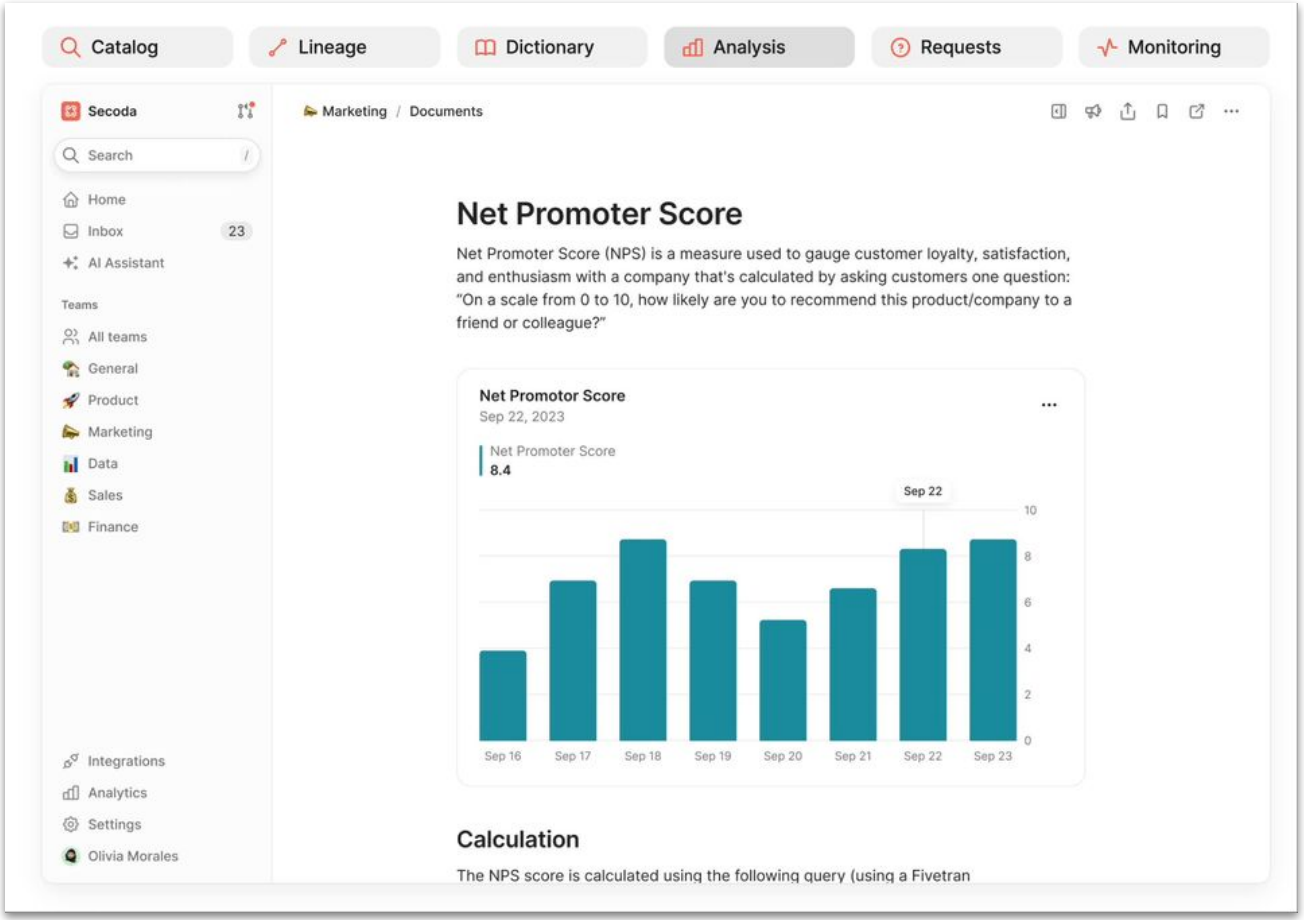
2024 LINEUP

OLEG AGAPOV • DYLAN ANDERSON • JAI BALANI
CHRISTOPHER BERRY • JULIE BEYNON
VIPUL BHARAT MARLECHA • CHRISTOPHE BLEFARI
MATTHEW BRANDT • TAYLOR BROWNLOW • KELLY BURDINE
STÉPHANE BURWASH • ERIC CALLAHAN
CYNTHIA CARIDAD • JES CARNEY • TIM CASTILLO
CHRISTOPHER CHIN • BRYCE CODELL • TIMO DECHAU
SHRAVAN DEOLALIKAR • RYAN DOLLEY • CECILIA DONES
TIANKAI FENG • KATIE HINDSON • KAREN HSIEH
MONICA KAY ROYAL • JERRIE KUMALAH
FAITH LIERHEIMER • MEGAN LIEU • ANDREW MADSON
PHOENIX MILLACY JAY • ETAI MIZRAHI
JORDAN MORROW • LINDSAY MURPHY
NATALIE NAKAMINE • PEDRAM NAVID • PARDIS NOORZAD
ZACK OBER • STEFANÍA ÓLAFSDÓTTIR • ABISOLA ONI
MEHDI OUAZZA • JAKE PETERSON • EMILY RIEDERER
BEN ROGOJAN • AUGUSTO ROSA • CHAD SANDERSON
ARCHIE SARRE WOOD • MADISON SCHOTT
EVA SCHREYER • SUMI SINGH • ABHI SIVASAILAM
PÁDRAIC SLATTERY • JEFF SLOAN • MARISA SMITH
BIJAN SOLTANI • GABI STEELE • JACK SWEENEY
MATT WEINGARTEN

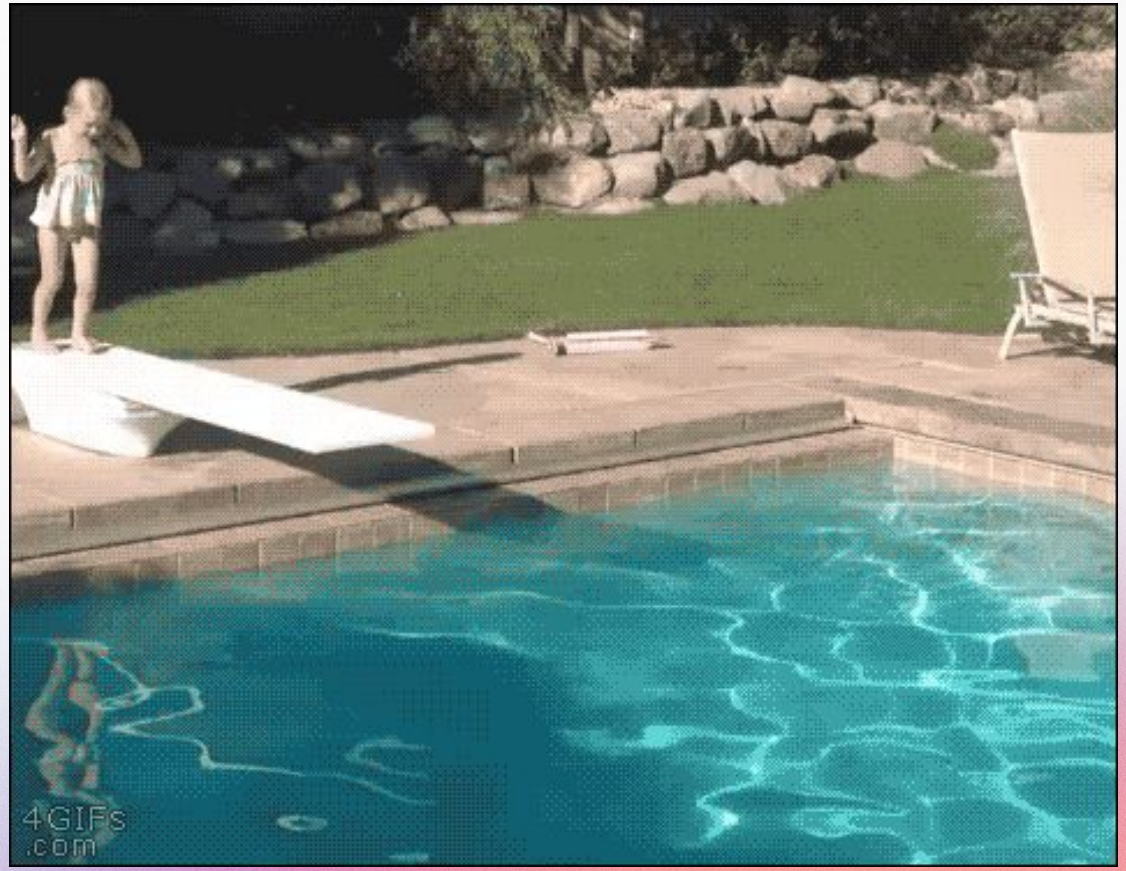


Secoda

- Catalog
- Lineage
- Governance
- Monitoring
- Documentation

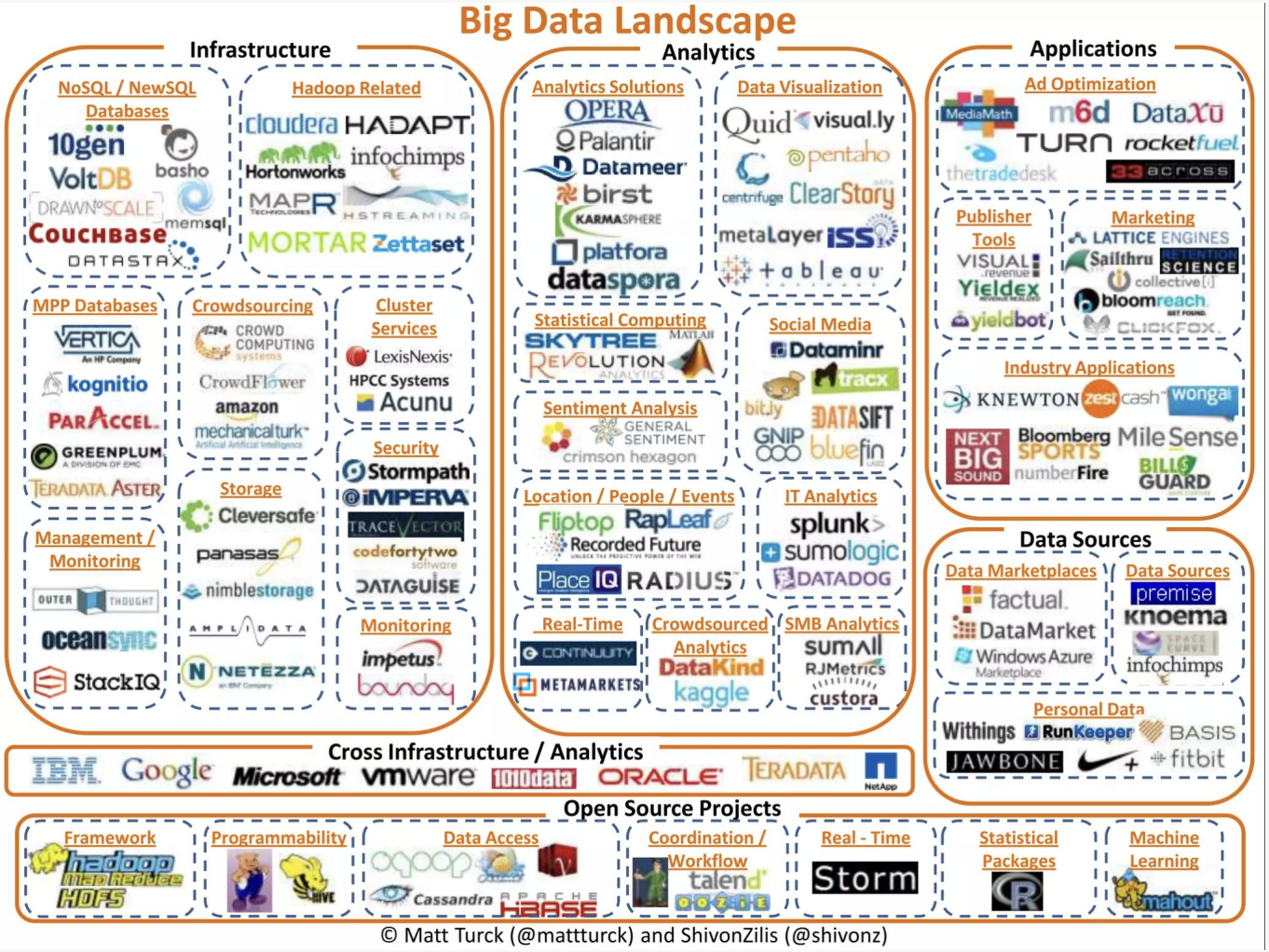


Let’s dive in



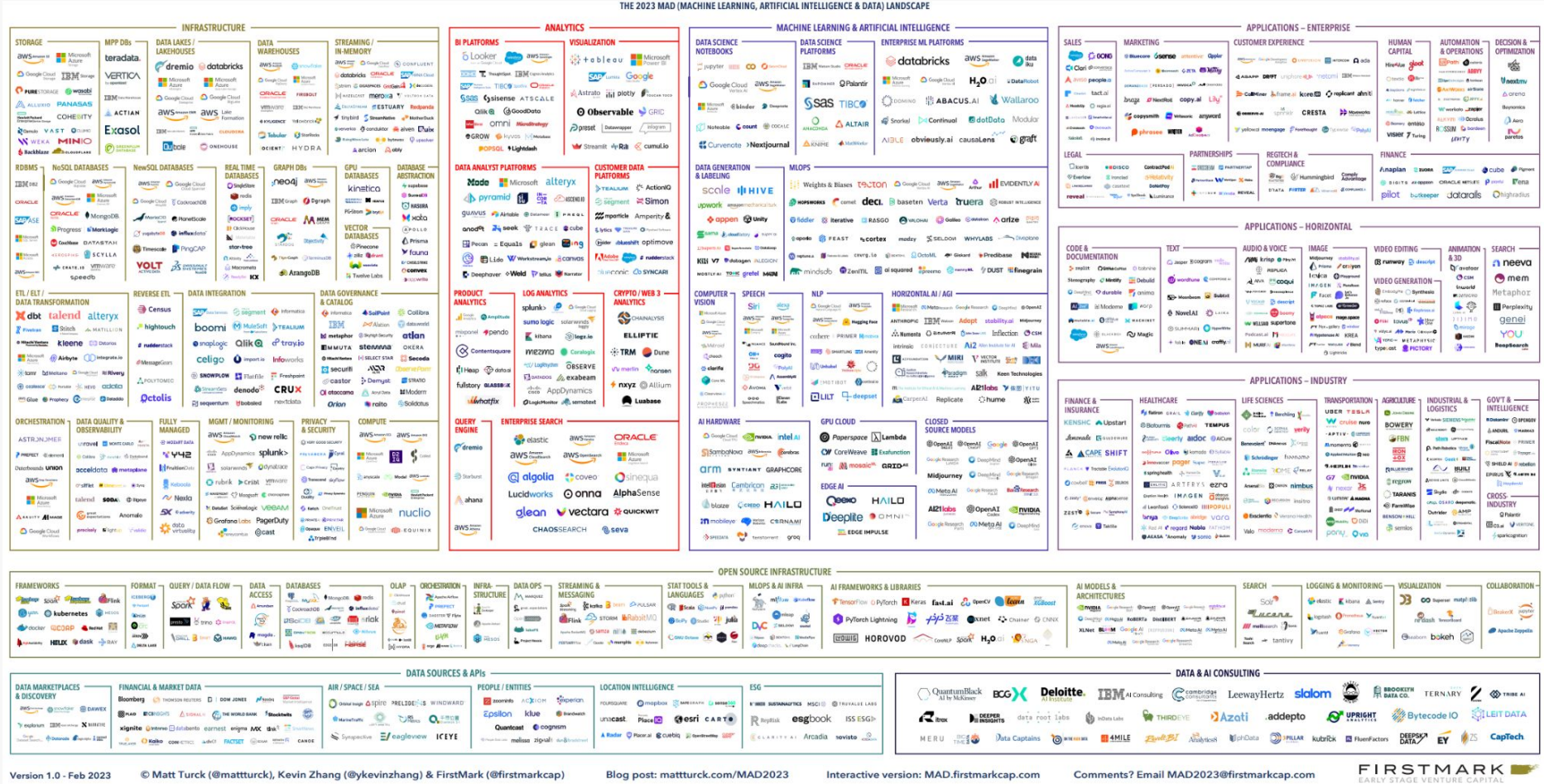
A lot has happened in our industry over the last 10+ years

We got some new tools



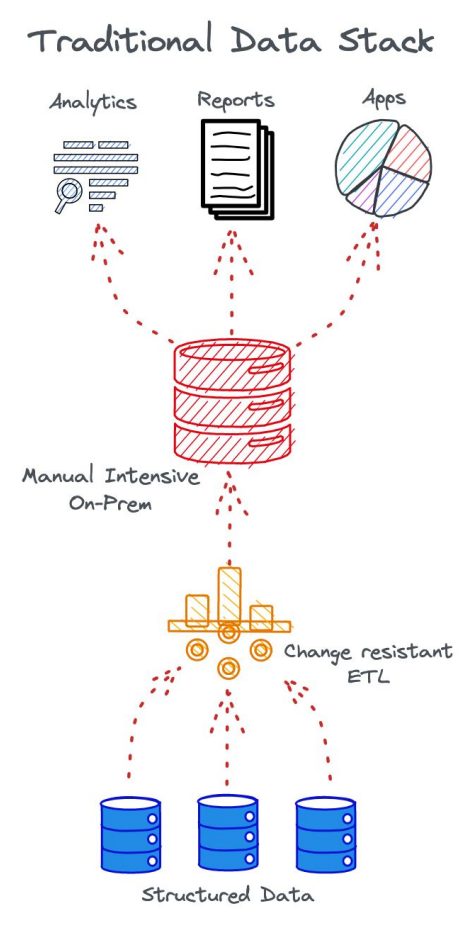
© Matt Turck (@mattturck) and ShivonZilis (@shivonz)

Source: https://mattturck.com/a-chart-of-the-big-data-ecosystem/

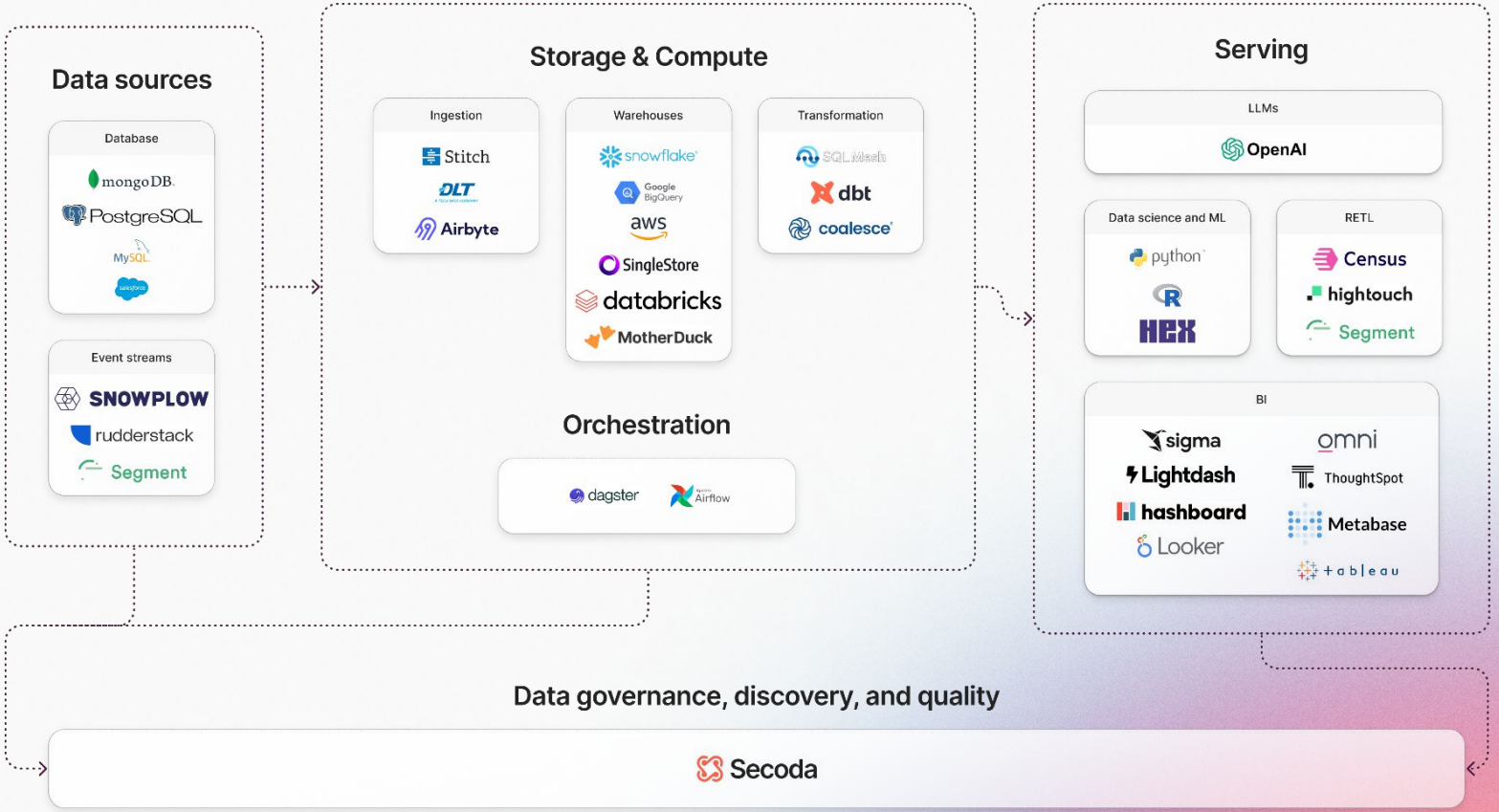


Source: https://mattturck.com/mad2023/

It’s gotten a lot easier to buy your entire data stack



Source: https://moderndata101.substack.com/p/evolution-of-the-data-stack-the-story

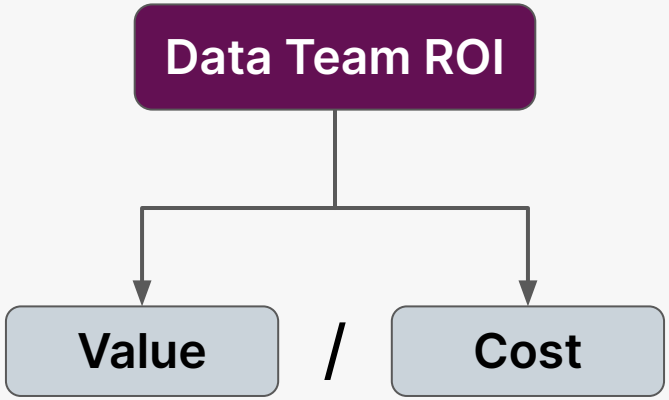

Data sources
Database
mongoDB.
PostgreSQL
MySQL
Event streams
SNOWPLOW
rudderstack
Segment
Storage & Compute
Ingestion
Stitch
DLT
Airbyte
Warehouses
snowflake
Google BigQuery
aws
SingleStore
databricks
MotherDuck
Transformation
SQLMesh
dbt
coalesce
Orchestration
dagster
Airflow
Serving
LLMs
OpenAI
Data science and ML
python
R
HEX
RETL
Census
hightouch
Segment
BI
sigma
Lightdash
hashboard
Looker
omni
ThoughtSpot
Metabase
tableau
Data governance, discovery, and quality
Secoda

Companies have invested a lot in becoming “data-driven”


Data Scientist:
The Sexiest Job of the 21st Century
Meet the people who can coax treasure out of messy, unstructured data.
by Thomas H. Davenport and D.J. Patil
When Jonathan Goldman arrived for work in June 2006 at LinkedIn, the business networking site, the place still felt like a start-up. The company had just under 8 million accounts, and the number was growing quickly as existing members invited their friends and col-
Data Engineer: The Sexiest Job of the 21st Century
Why We Need Them Before Data Scientists
Move Over Data Scientists, Analytics Engineers Have the Sexiest Job

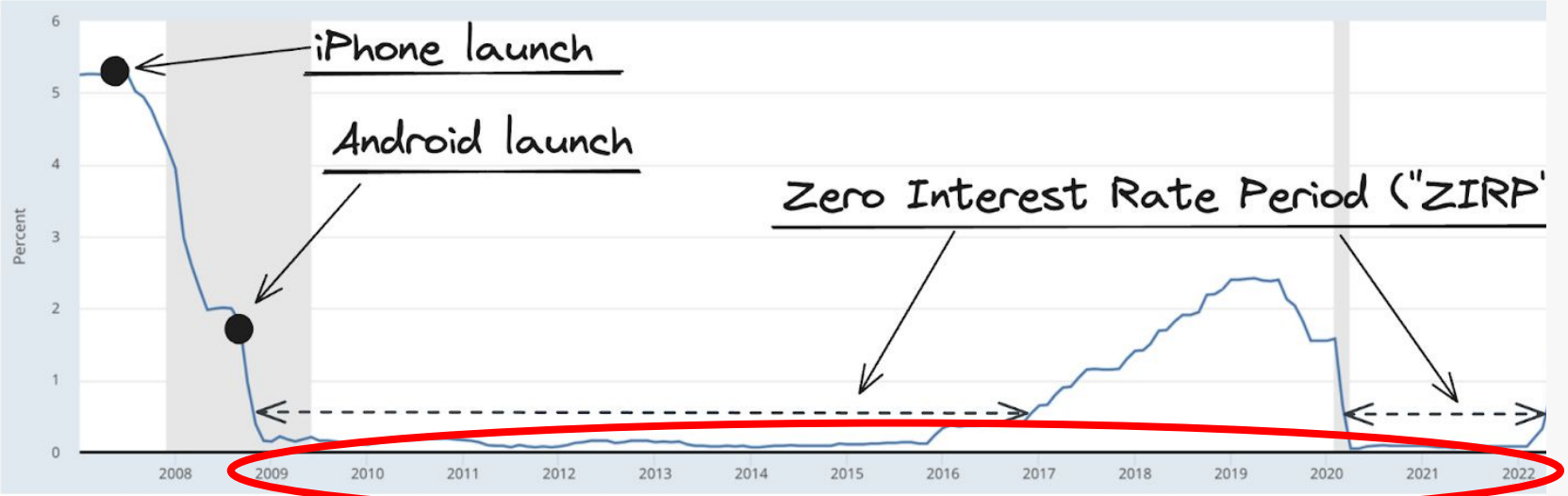
But some things haven’t changed…

Data teams still aren’t very good at measuring the value we deliver



Complicated to measure

Pretty easy to measure

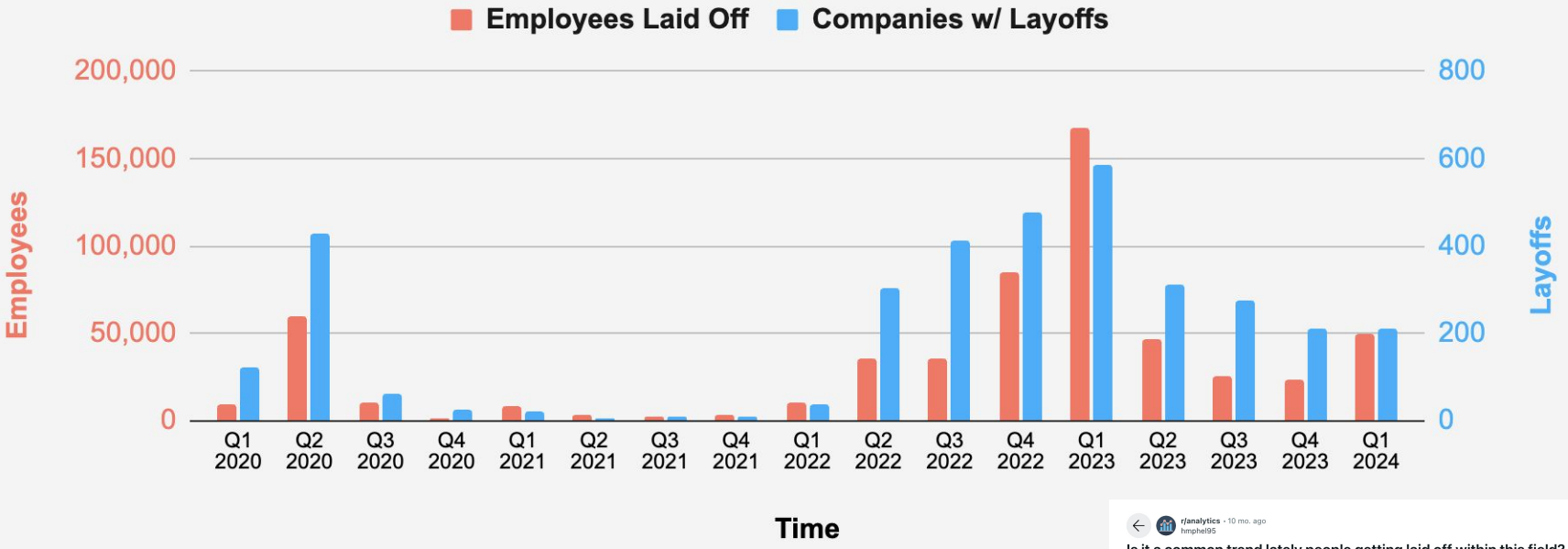

iPhone launch
Android launch
Zero Interest Rate Period ("ZIRP"
Percent
6
5
4
3
2
1
0
2008
2009
2010
2011
2012
2013
2014
2015
2016
2017
2018
2019
2020
2021
2022
pragma


iPhone launch
Android launch
Zero Interest Rate Period ("ZIRP")
Percent
6
5
4
3
2
1
0
2008
2009
2010
2011
2012
2013
2014
2015
2016
2017
2018
2019
2020
2021
2022
2023
pragmaticengineer.com

Data teams get labelled as cost centres, instead of value drivers

Tech layoffs since COVID-19

Source: https://layoffs.fyi



r/analytics • 10 mo. ago
hmphel95

Is it a common trend lately people getting laid off within this field? I've been seeing it a lot in fb groups. I plan to start diving into analytics and skill building relatively soon so just curious

Career Advice

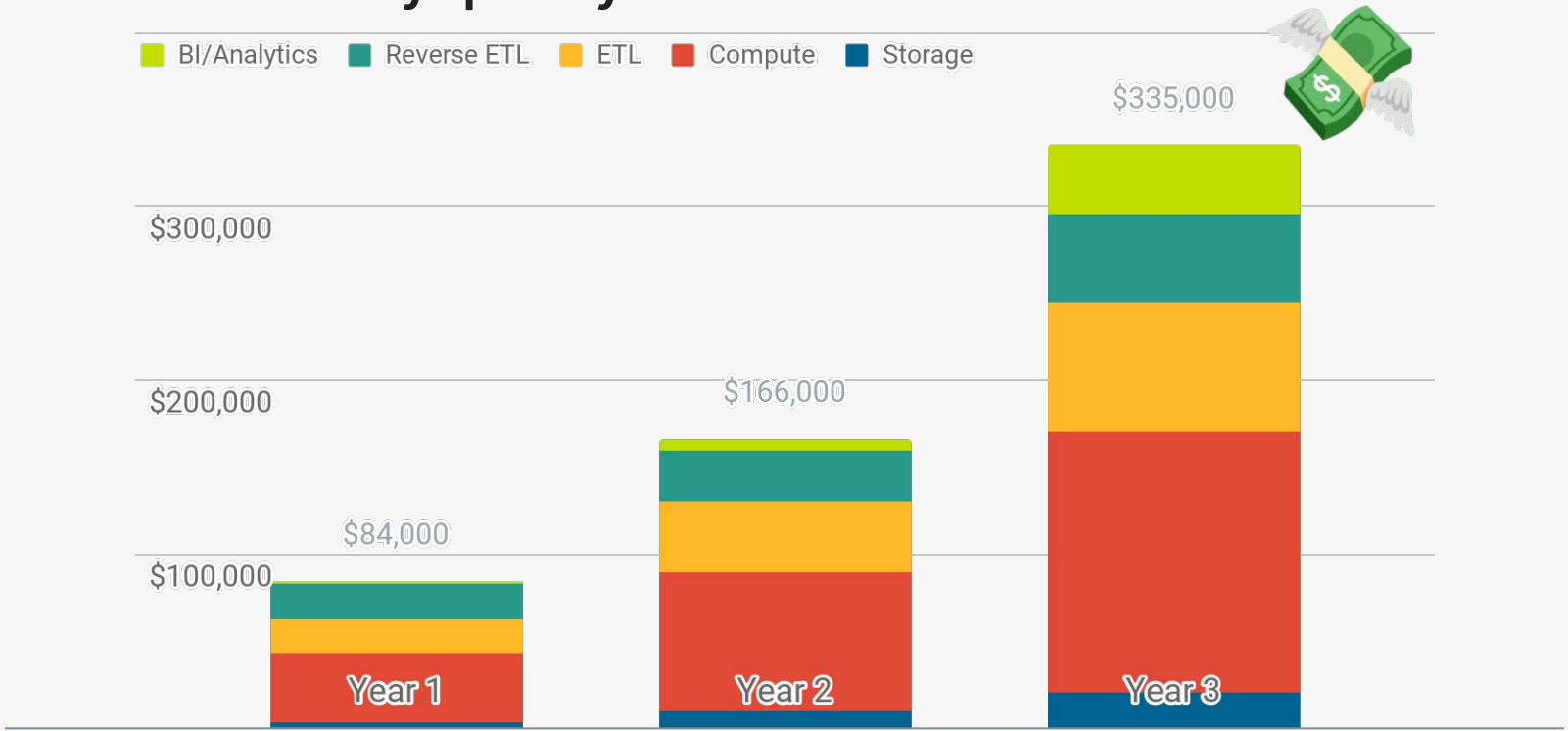

Data teams
Post ZIRP economy
2024
Now you're all in big, big trouble.

And we’ve got other problems too…

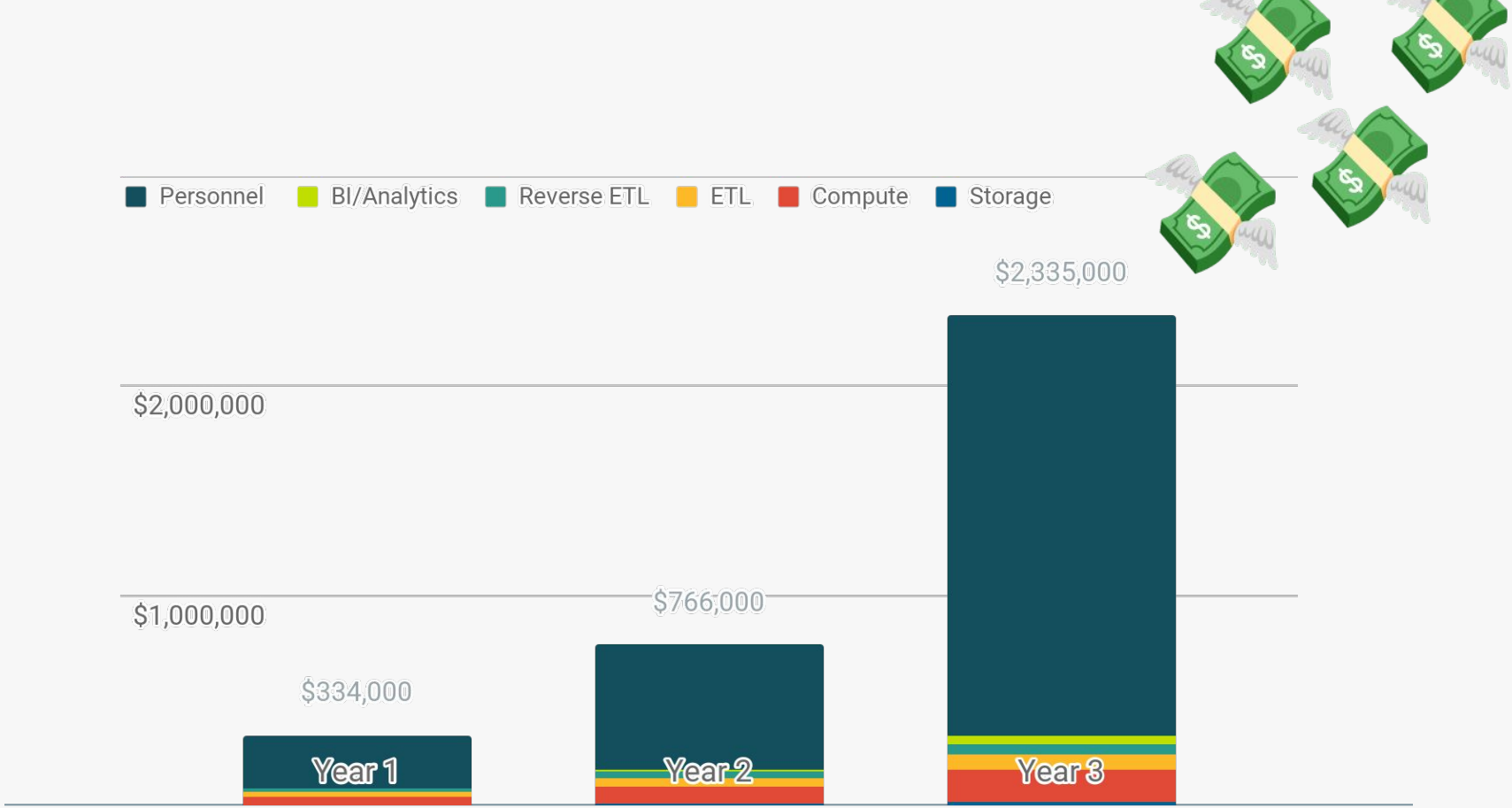
The “modern data stack” is designed to help you spend $$$

Initial costs are typically low...

But can scale very quickly...



Source: https://syncari.com/blog/modern-data-stack-cost/



Source: https://syncari.com/blog/modern-data-stack-cost/

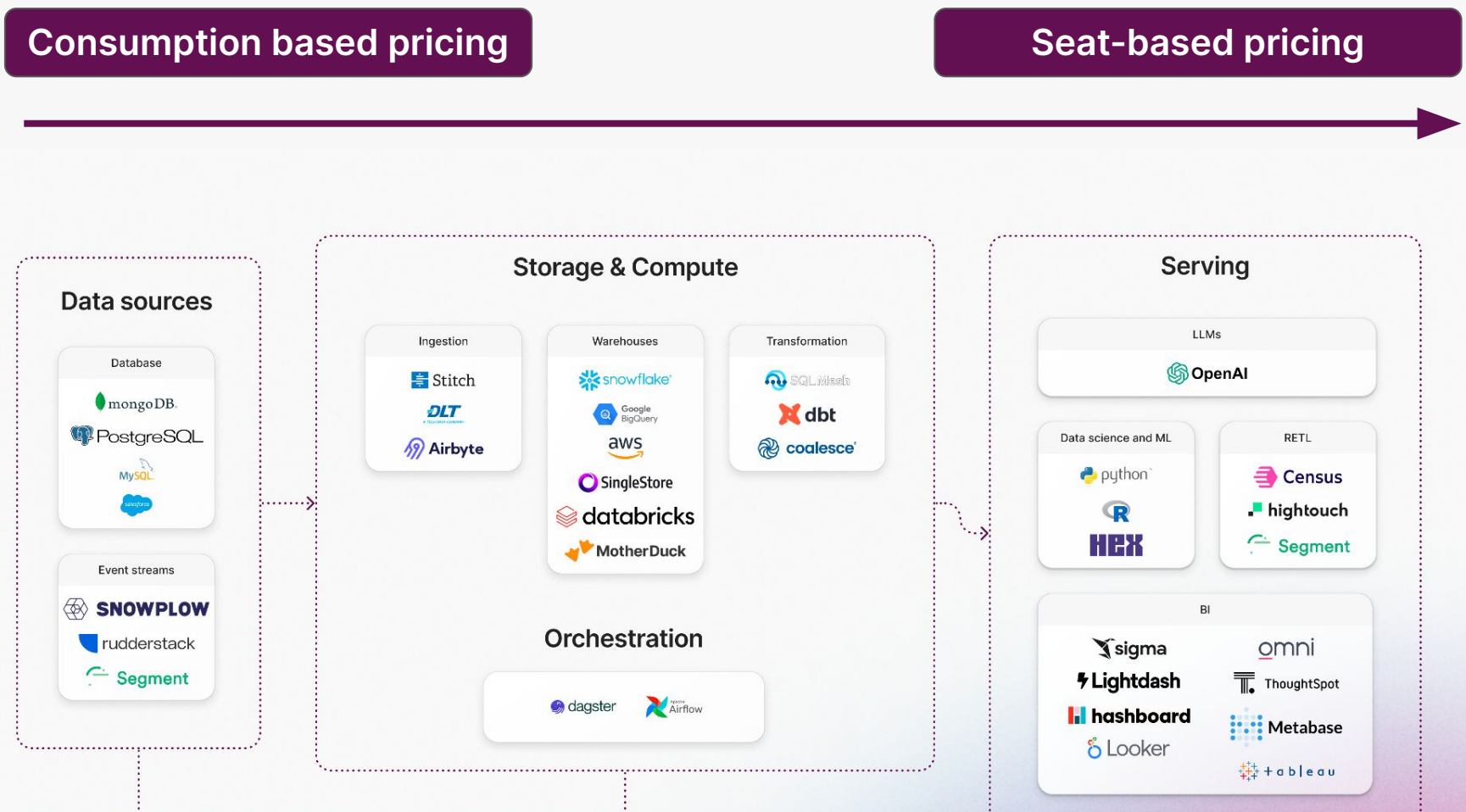
Tools abstract away complexity in favour of convenience

Addition is easy,
but deprecation can be...

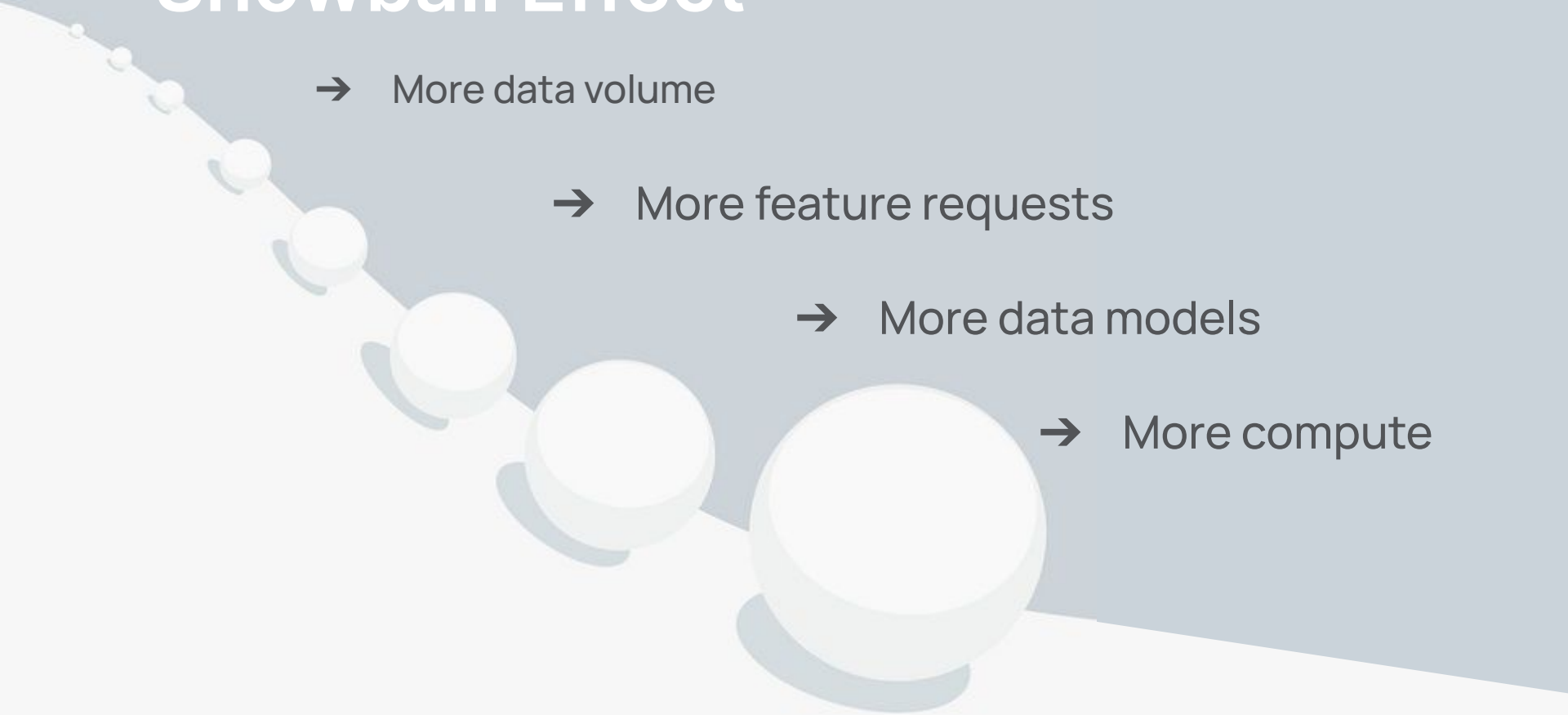
...a little scary

Source: dbt Labs product spotlight & keynote, Coalesce 2023

Interplay of pricing models


Consumption based pricing
Seat-based pricing
Data sources
Database
mongoDB.
PostgreSQL
MySQL
Event streams
SNOWPLOW
rudderstack
Segment
Storage & Compute
Ingestion
Stitch
DLT
Airbyte
Warehouses
snowflake
Google BigQuery
aws
SingleStore
databricks
MotherDuck
Transformation
SQLMesh
dbt
coalesce
Orchestration
dagster
Airflow
Serving
LLMs
OpenAI
Data science and ML
python
HEX
RETL
Census
hightouch
Segment
BI
sigma
Lightdash
hashboard
Looker
omni
ThoughtSpot
Metabase
tableau

Snowball Effect



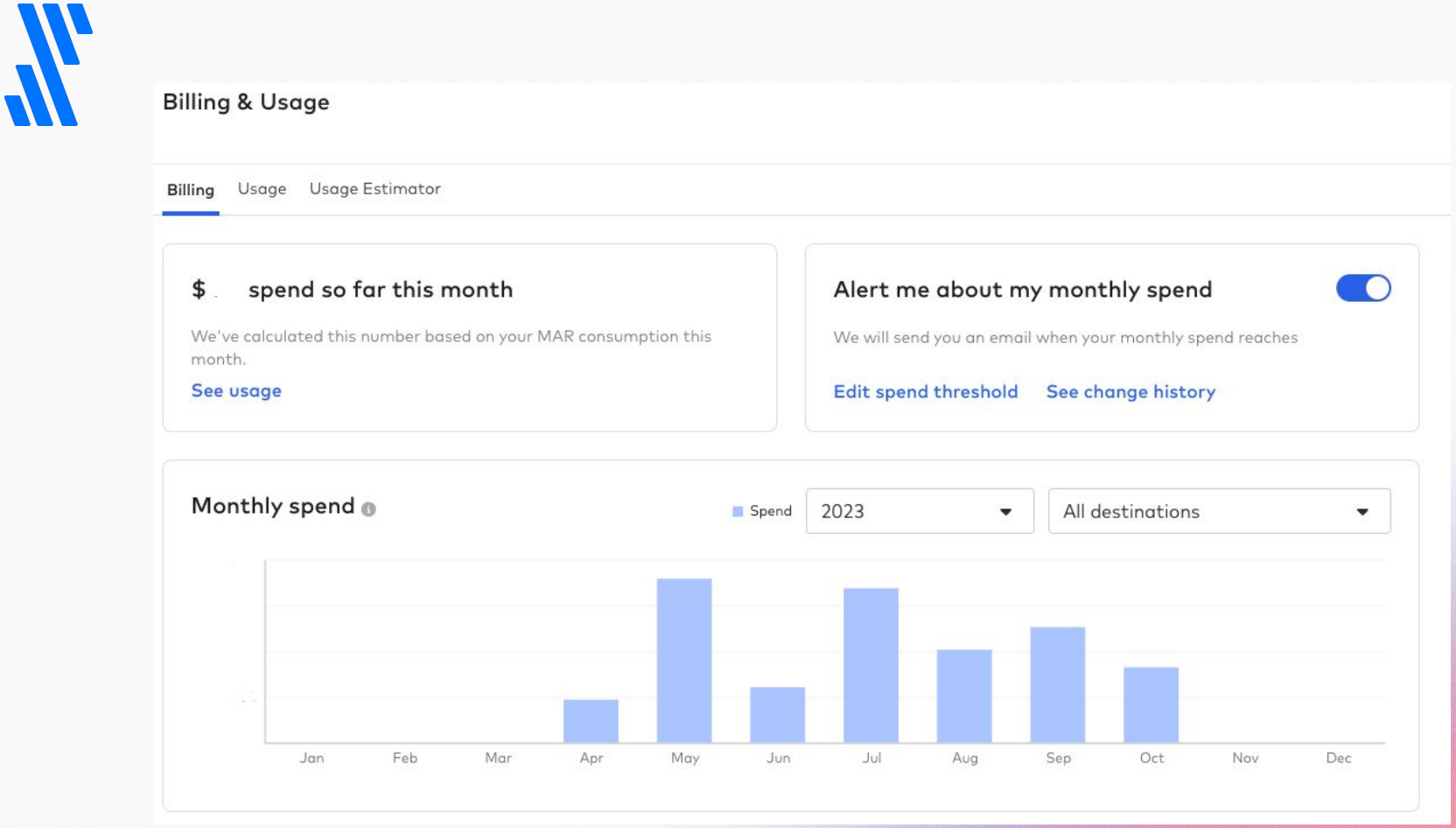
➔ More data volume

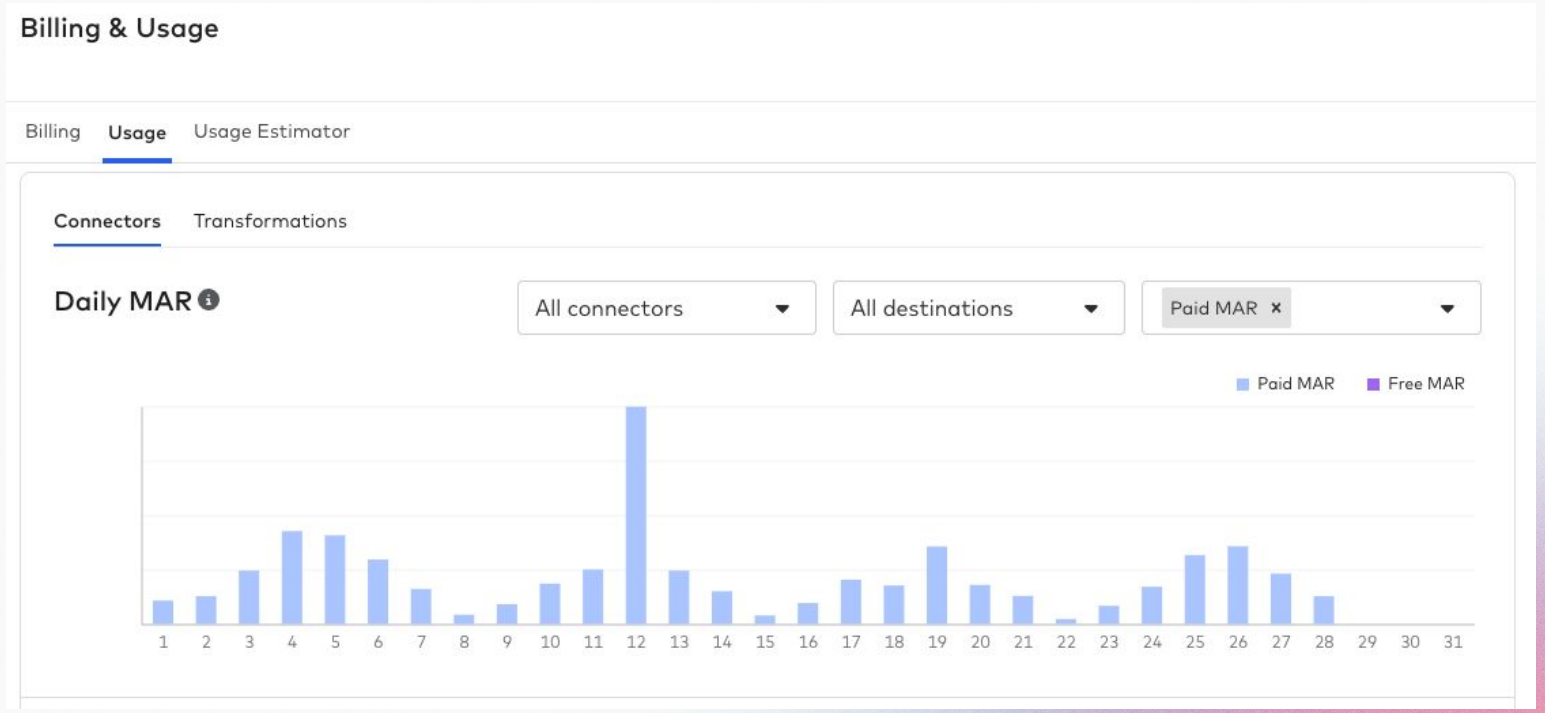
➔ More feature requests

➔ More data models

➔ More compute

Cost and usage data is not easily exposed or aggregated


Billing & Usage
Billing
Usage
Usage Estimator
$ spend so far this month
We've calculated this number based on your MAR consumption this month.
See usage
Alert me about my monthly spend
We will send you an email when your monthly spend reaches
Edit spend threshold
See change history
Monthly spend
Spend
2023
All destinations
Jan
Feb
Mar
Apr
May
Jun
Jul
Aug
Sep
Oct
Nov
Dec


Billing & Usage
Billing
Usage
Usage Estimator
Connectors
Transformations
Daily MAR
All connectors
All destinations
Paid MAR
Paid MAR
Free MAR
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 30 31

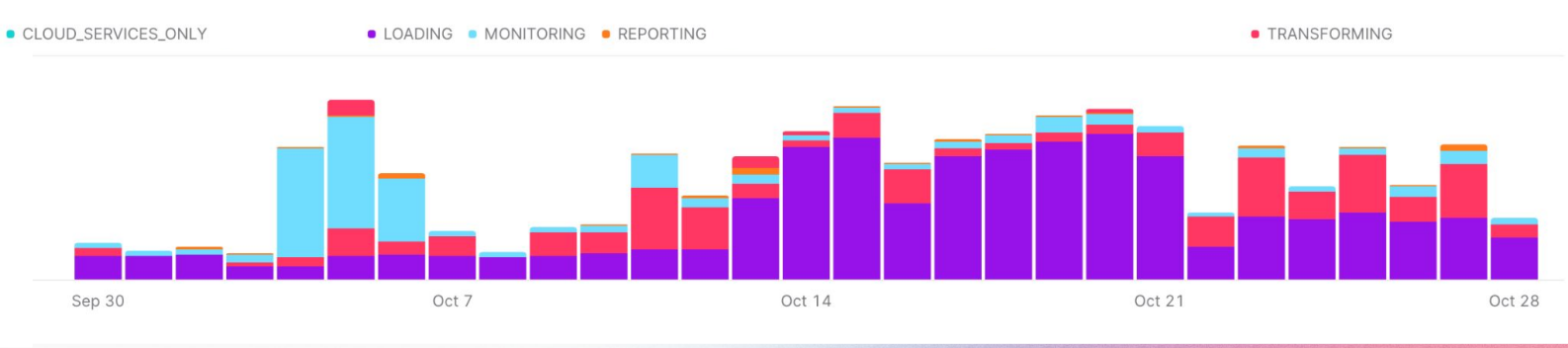
Time period: Last 30 Days | Source: All Sources | Destination: All Destinations

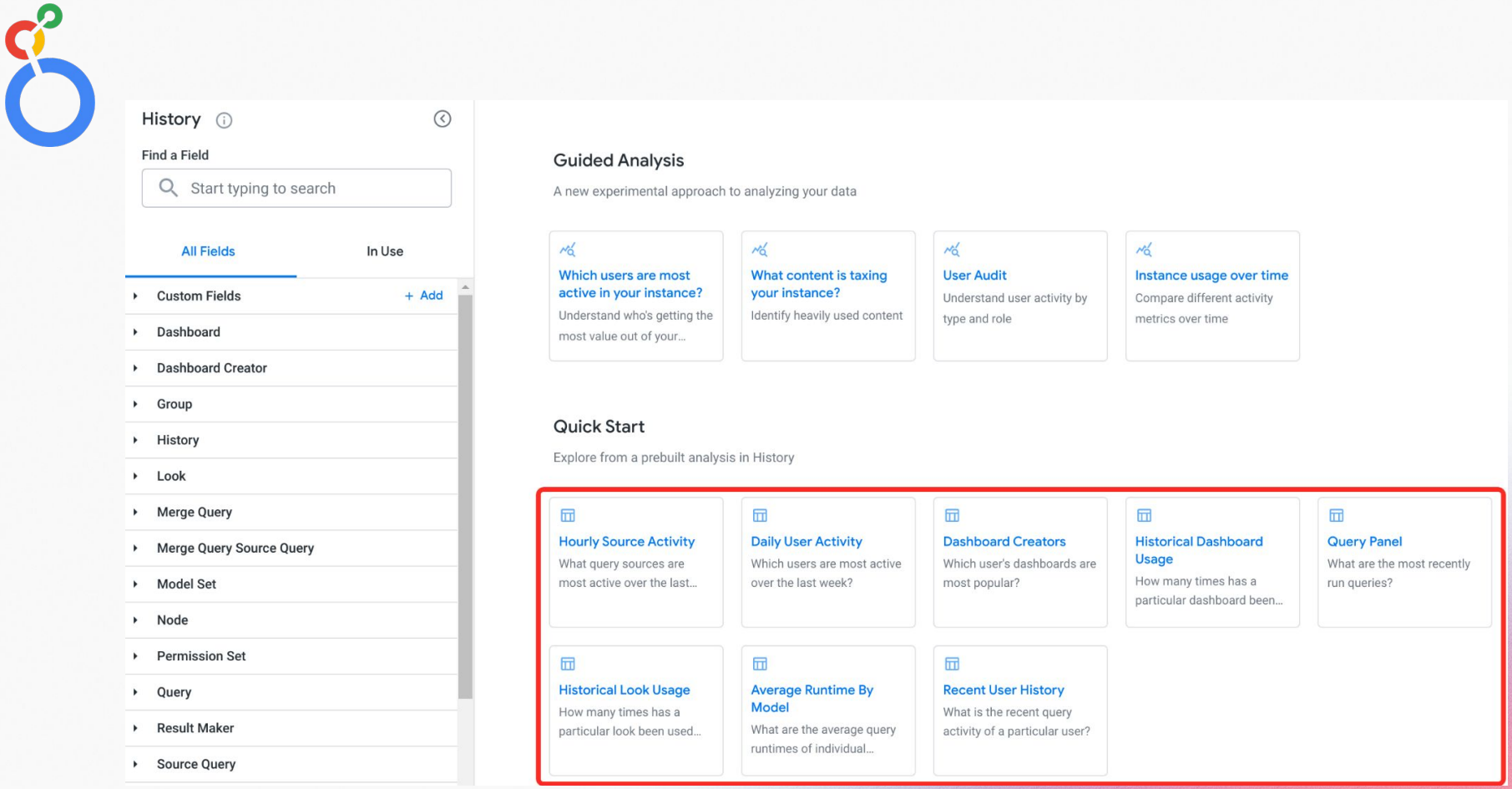
Total credits usage

Usage per connection

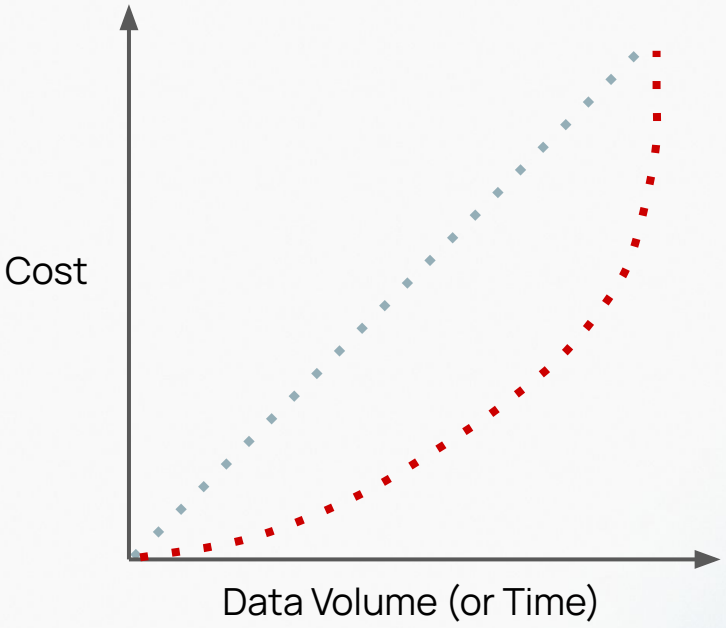
CONNECTION	SOURCE		DESTINATION	SCHEDULE	USAGE
US Read Replica → Snowflake	US Read Replica CERTIFIED	→	Snowflake CERTIFIED	6 hours	
APAC Read Replica → Snowflake	APAC Read Replica CERTIFIED	→	Snowflake CERTIFIED	6 hours	
EU Read Replica → Snowflake	EU Read Replica CERTIFIED	→	Snowflake CERTIFIED	6 hours	

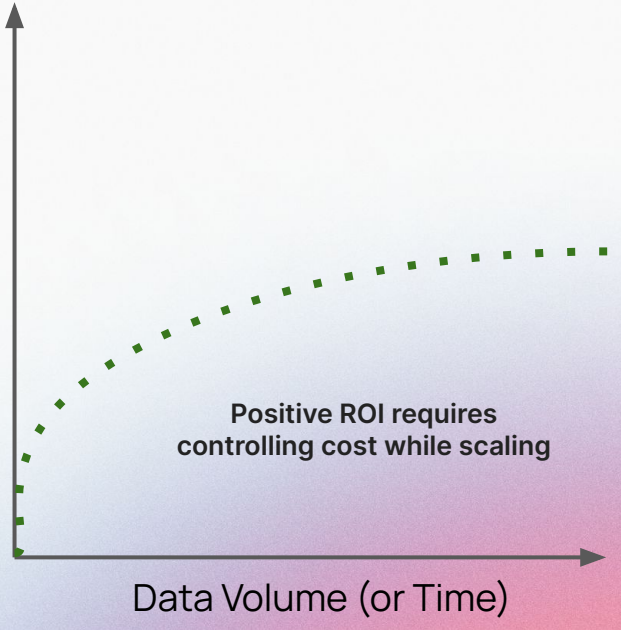

COMPUTE
CLOUD
Sep 30
Oct 7
Oct 14
Oct 21
Oct 28


CLOUD_SERVICES_ONLY
LOADING
MONITORING
REPORTING
TRANSFORMING
Sep 30
Oct 7
Oct 14
Oct 21
Oct 28


History
Find a Field
Start typing to search
All Fields
In Use
Custom Fields
+ Add
Dashboard
Dashboard Creator
Group
History
Look
Merge Query
Merge Query Source Query
Model Set
Node
Permission Set
Query
Result Maker
Source Query
Guided Analysis
A new experimental approach to analyzing your data
Which users are most active in your instance?
Understand who's getting the most value out of your...
What content is taxing your instance?
Identify heavily used content
User Audit
Understand user activity by type and role
Instance usage over time
Compare different activity metrics over time
Quick Start
Explore from a prebuilt analysis in History
Hourly Source Activity
What query sources are most active over the last...
Daily User Activity
Which users are most active over the last week?
Dashboard Creators
Which user's dashboards are most popular?
Historical Dashboard Usage
How many times has a particular dashboard been...
Query Panel
What are the most recently run queries?
Historical Look Usage
How many times has a particular look been used...
Average Runtime By Model
What are the average query runtimes of individual...
Recent User History
What is the recent query activity of a particular user?

Relationship between cost, volume or time

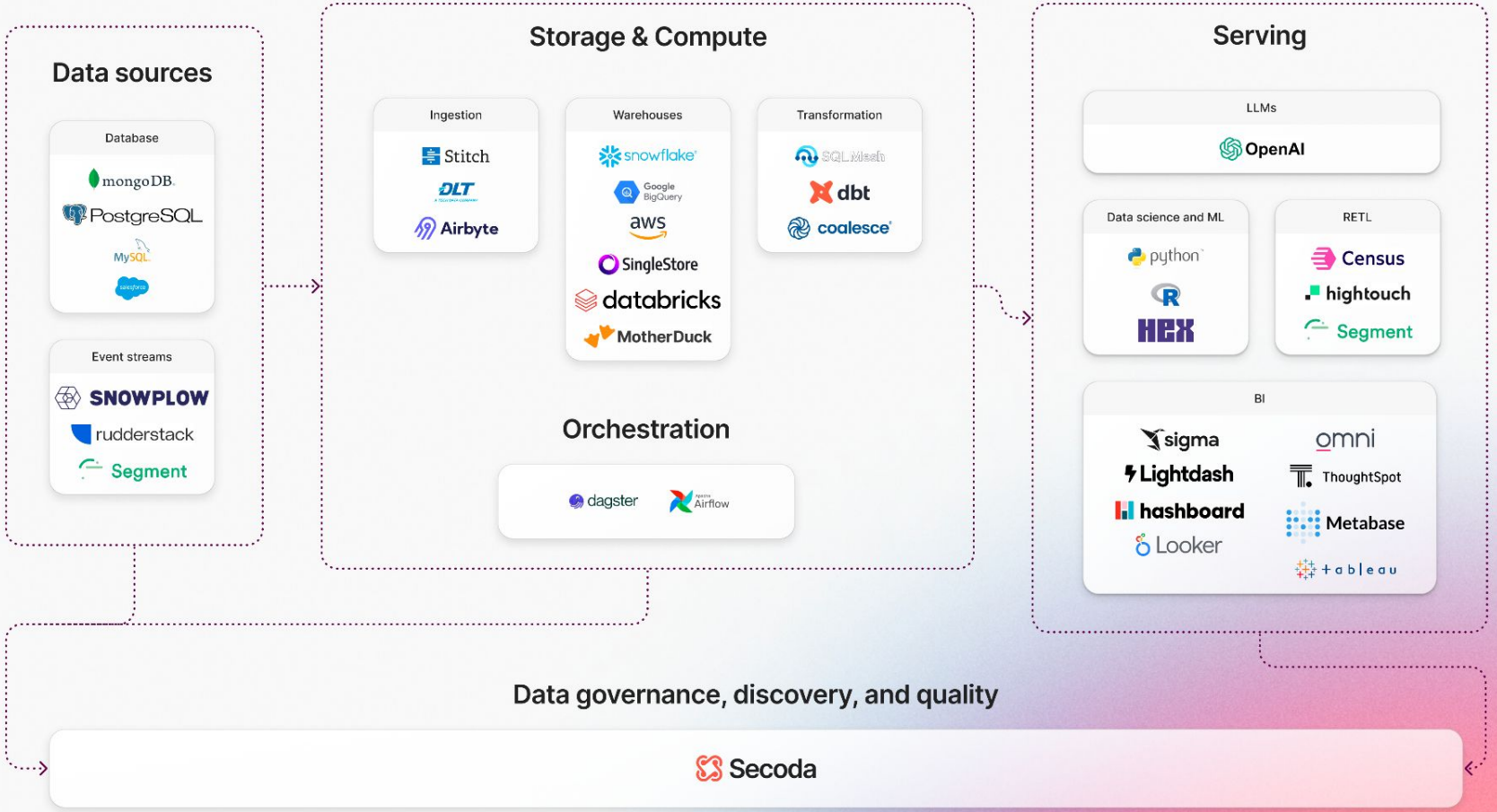




Recommendations

Measure

1. Map out your tech stack


Data sources
Database
mongoDB.
PostgreSQL
MySQL
Event streams
SNOWPLOW
rudderstack
Segment
Storage & Compute
Ingestion
Stitch
DLT
Airbyte
Warehouses
snowflake
Google BigQuery
aws
SingleStore
databricks
MotherDuck
Transformation
SQLMesh
dbt
coalesce
Orchestration
dagster
Airflow
Serving
LLMs
OpenAI
Data science and ML
python
R
HEX
RETL
Census
hightouch
Segment
BI
sigma
Lightdash
hashboard
Looker
omni
ThoughtSpot
Metabase
tableau
Data governance, discovery, and quality
Secoda

2. Categorize pricing models and identify sources of cost related data

Tool	Type	Pricing Model	Pricing Model Category	Billing Cadence	Data Source
Fivetran	ETL	usage based	variable	monthly	Fivetran logs
Airbyte	ETL	open source	variable	monthly	AWS billing API
Snowflake	Warehouse	usage based	variable	monthly	Snowflake information schema
GitHub Actions	Orchestration	plan based	static	monthly	N/A
Lightdash	BI	open source	variable	monthly	AWS billing API
Retool	Data Applications	plan based	static (currently credit-based)	monthly	Float card
Segment	CDP	plan based	static (up to usage tier)	monthly	Float card
Mixpanel	Product analytics (quant)	annual license based	static	annually	Contract
Zapier	Automation	plan based	static (up to usage tier)	monthly	Float card

Static

Annual Contracts

Monthly Plans

Salaries

Variable

ELT

Warehouse

Transformation

Orchestration

Headcount

4. Source and model cost data

Sourcing variable cost data

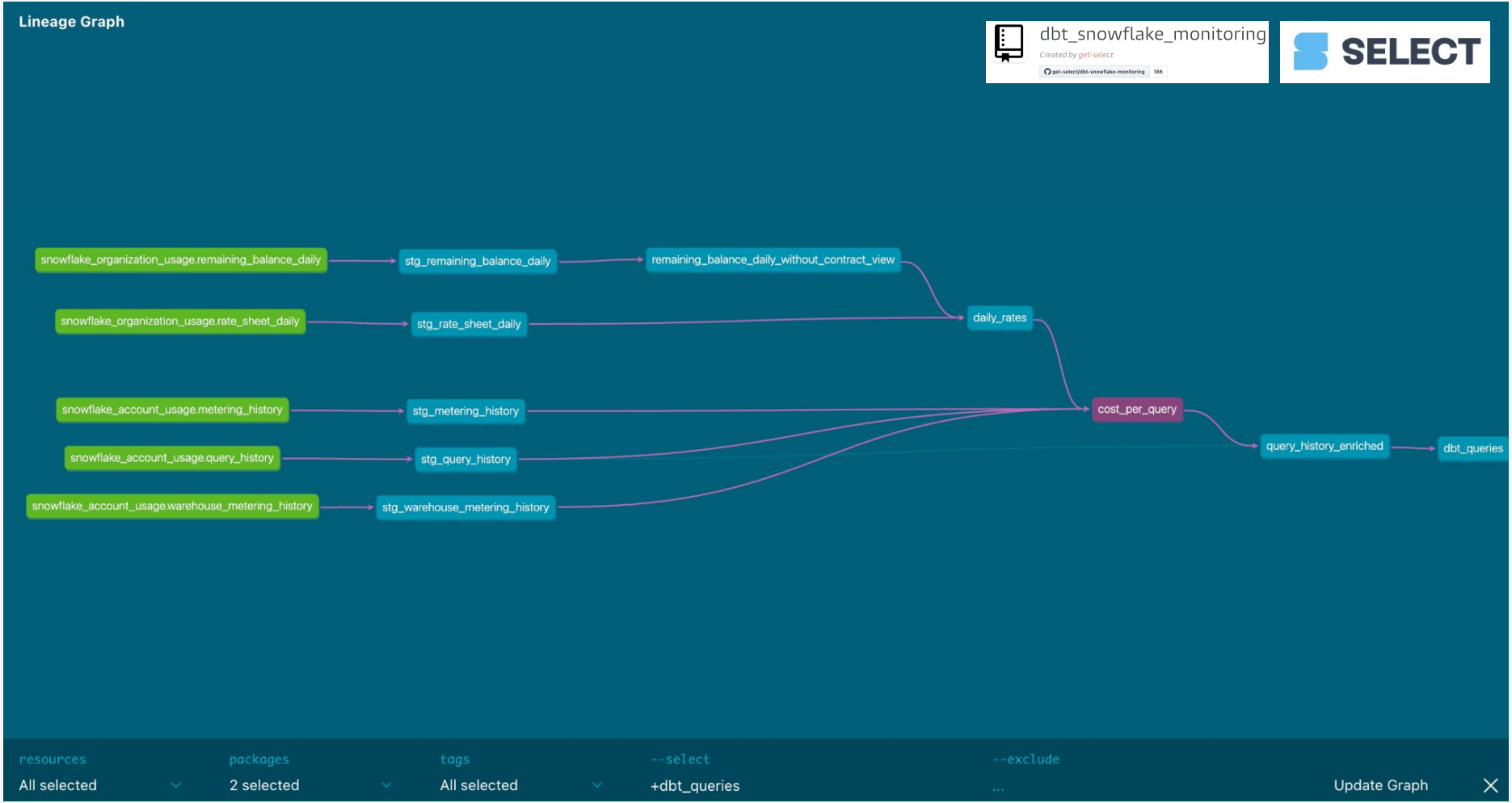
Warehouses:

- https://github.com/get-select/dbt-snowflake-monitoring
- https://github.com/kayrnt/dbt-bigquery-monitoring
- https://github.com/dbt-labs/redshift/

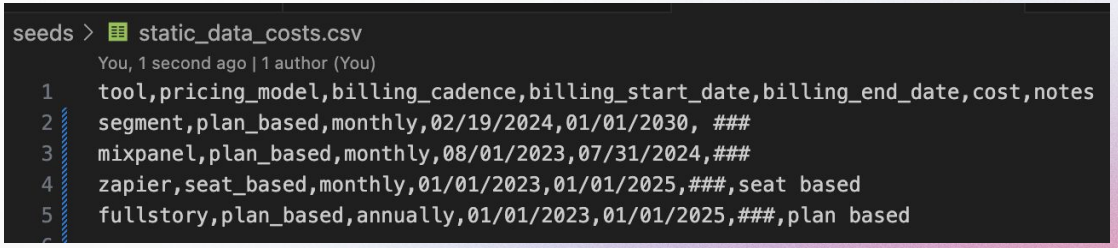
ETL:

- https://github.com/fivetran/dbt_fivetran_log


Lineage Graph
dbt_snowflake_monitoring
Created by get-select
get-select/dbt-snowflake-monitoring 186
SELECT
warehouse_credits_map
snowflake_account_usage.warehouse_events_history
stg_warehouse_events_history
warehouse_cluster_status
snowflake_organization_usage.usage_in_currency_daily
stg_usage_in_currency_daily
snowflake_account_usage.stage_storage_usage_history
stg_stage_storage_usage_history
snowflake_account_usage.serverless_task_history
stg_serverless_task_history
snowflake_account_usage.metering_daily_history
stg_metering_daily_history
snowflake_account_usage.database_storage_usage_history
stg_database_storage_usage_history
snowflake_account_usage.query_history
stg_query_history
query_base_object_access
query_base_table_access
snowflake_organization_usage.rate_sheet_daily
snowflake_account_usage.metering_history
stg_metering_history
snowflake_account_usage.access_history
stg_access_history
snowflake_account_usage.warehouse_metering_history
stg_warehouse_metering_history
stg_rate_sheet_daily
daily_rates
snowflake_organization_usage.remaining_balance_daily
stg_remaining_balance_daily
remaining_balance_daily_without_contract_view
query_direct_object_access
query_direct_table_access
warehouses_type2_dimension
hourly_spend
daily_spend
cost_per_query
query_history_enriched
dbt_queries
resources
All selected
packages
dbt_snowflake_monito...
tags
All selected
--select
...
--exclude
...
Update Graph


Lineage Graph
dbt_snowflake_monitoring
Created by get-select
get-select/dbt-snowflake-monitoring 186
SELECT
snowflake_organization_usage.remaining_balance_daily
stg_remaining_balance_daily
remaining_balance_daily_without_contract_view
snowflake_organization_usage.rate_sheet_daily
stg_rate_sheet_daily
daily_rates
snowflake_account_usage.metering_history
stg_metering_history
cost_per_query
snowflake_account_usage.query_history
stg_query_history
query_history_enriched
dbt_queries
snowflake_account_usage.warehouse_metering_history
stg_warehouse_metering_history
resources
All selected
packages
2 selected
tags
All selected
--select
+dbt_queries
--exclude
...
Update Graph

Building static costs from contracts



Building static costs from contracts

models > intermediate > int_static_data_tooling_cost_per_month.sql

You, 20 hours ago | 1 author (You)

```sql
WITH

date_spine AS (
    SELECT
        calendar_date AS month_start_date
    FROM
        {{ ref ('dim_dates') }}
    WHERE
        day_of_month = 1
        AND month_start_date <= CURRENT_DATE
        AND month_start_date >= '2023-01-01'
),

static_cost_tools AS (
    SELECT
        tool,
        pricing_model,
        billing_cadence,
        DATE_TRUNC('month', billing_start_date) AS billing_month_start_date,
        DATE_TRUNC('month', billing_end_date) AS billing_month_end_date,
        cost
    FROM {{ ref ('static_data_costs') }}
),

final AS (      You, 20 hours ago • add cost model and metrics
    SELECT
        date_spine.month_start_date,
        static_cost_tools.tool AS spend_source,

        CASE
            WHEN static_cost_tools.billing_cadence = 'monthly' THEN static_cost_tools.cost
            WHEN static_cost_tools.billing_cadence = 'annually' THEN static_cost_tools.cost / 12.0
            ELSE NULL
        END AS monthly_spend
    FROM
        date_spine
    LEFT JOIN
        static_cost_tools ON date_spine.month_start_date BETWEEN static_cost_tools.billing_month_start_date AND static_cost_tools.billi
)

```

Create a cost model

4. Build Operating Cost Assets

New Browse Search Secoda ⌘+K Secoda LM

Data Team Operating Cost Dashboard

Add filter | **End time** in the last **180 completed days** | **Warehouse name** is any value | Date Zoom

Dashboard Details

This dashboard contains spend information from Snowflake and specifically dbt queries.

Total Warehouse Spend

$

Snowflake Annualized Total Spend

$

+3% ↗ vs. Previous Day

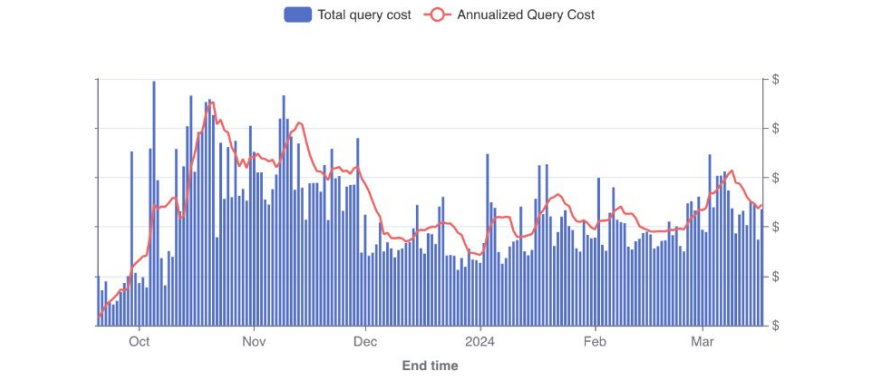
Total Spend on dbt

$

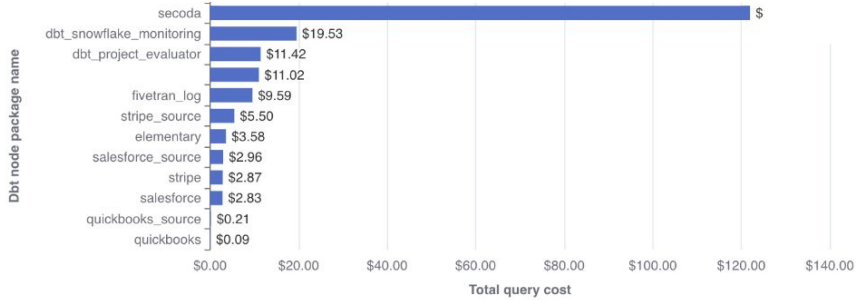
Count of dbt Queries

1,698,871

Snowflake: Daily and Annualized Query Cost



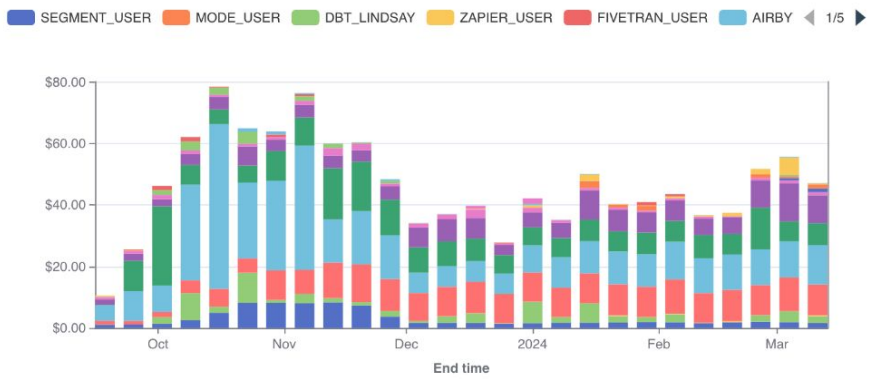
Total Query Cost by dbt Package



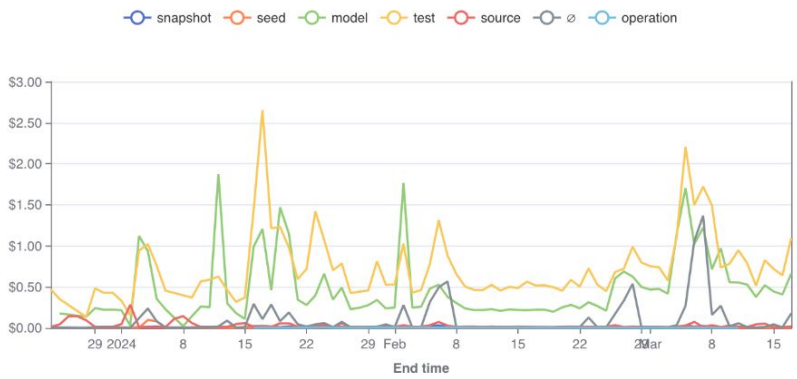
New Browse Search Secoda ⌘+K Secoda LM

Data Team Operating Cost Dashboard

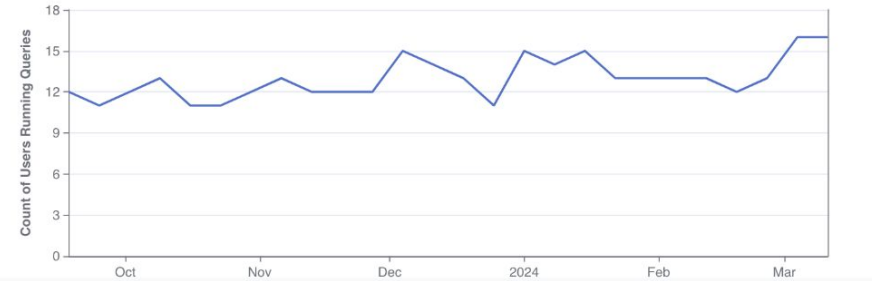
Daily Query Cost by User Name



Daily dbt Cost by Resource Type



Unique Users Running Queries

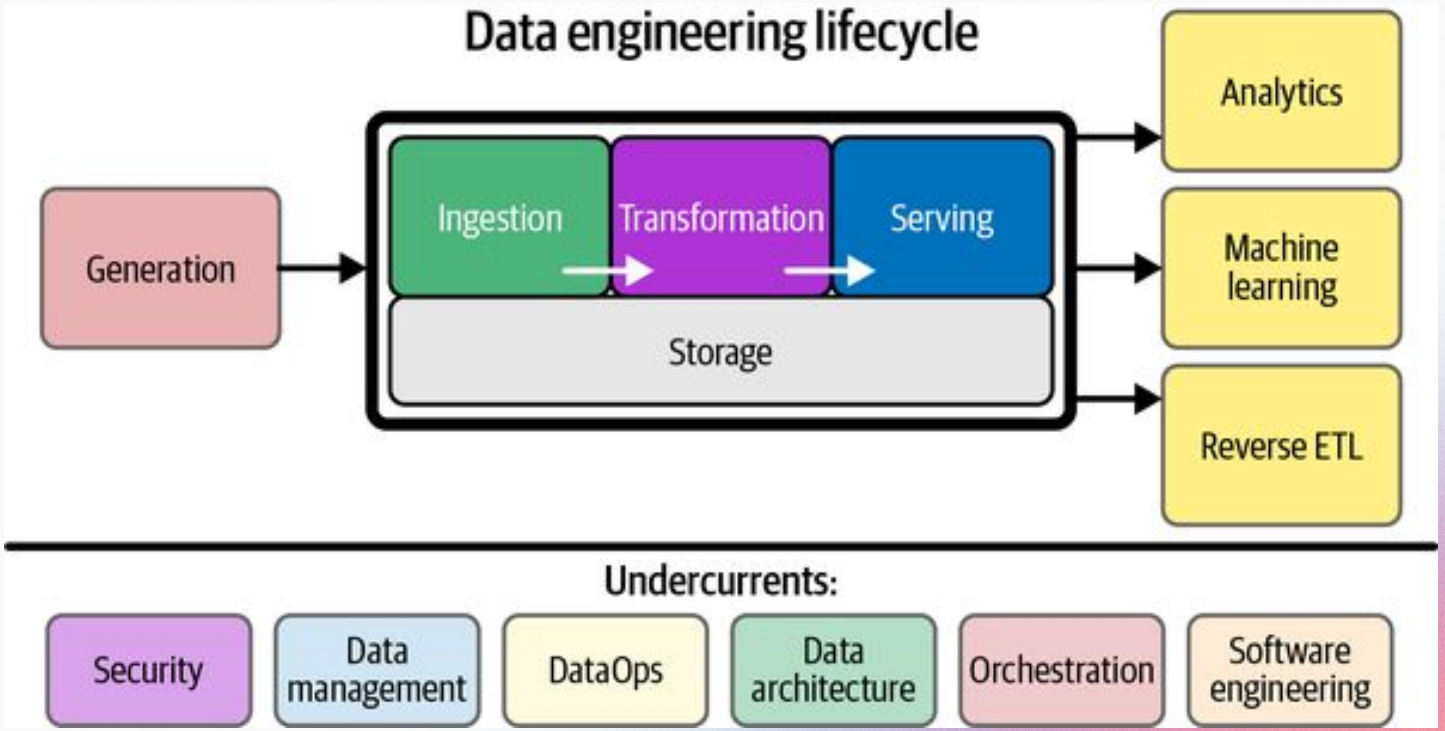


Most Expensive dbt Resources Last 30 Days

#	dbt Query Monitoring **Dbt node name**	dbt Query Monitoring **Dbt node resource type**	dbt Query Monitoring **Total query cost**	**Annualized Query Cost**
1	ø	ø		
2	stg_mixpanel_app__events	model		
3	hourly_spend	model		
4	query_history_enriched	model		
5	fct_marketing_conversions	model		
6	stg_secoda__admin_analyticsmetric	model		

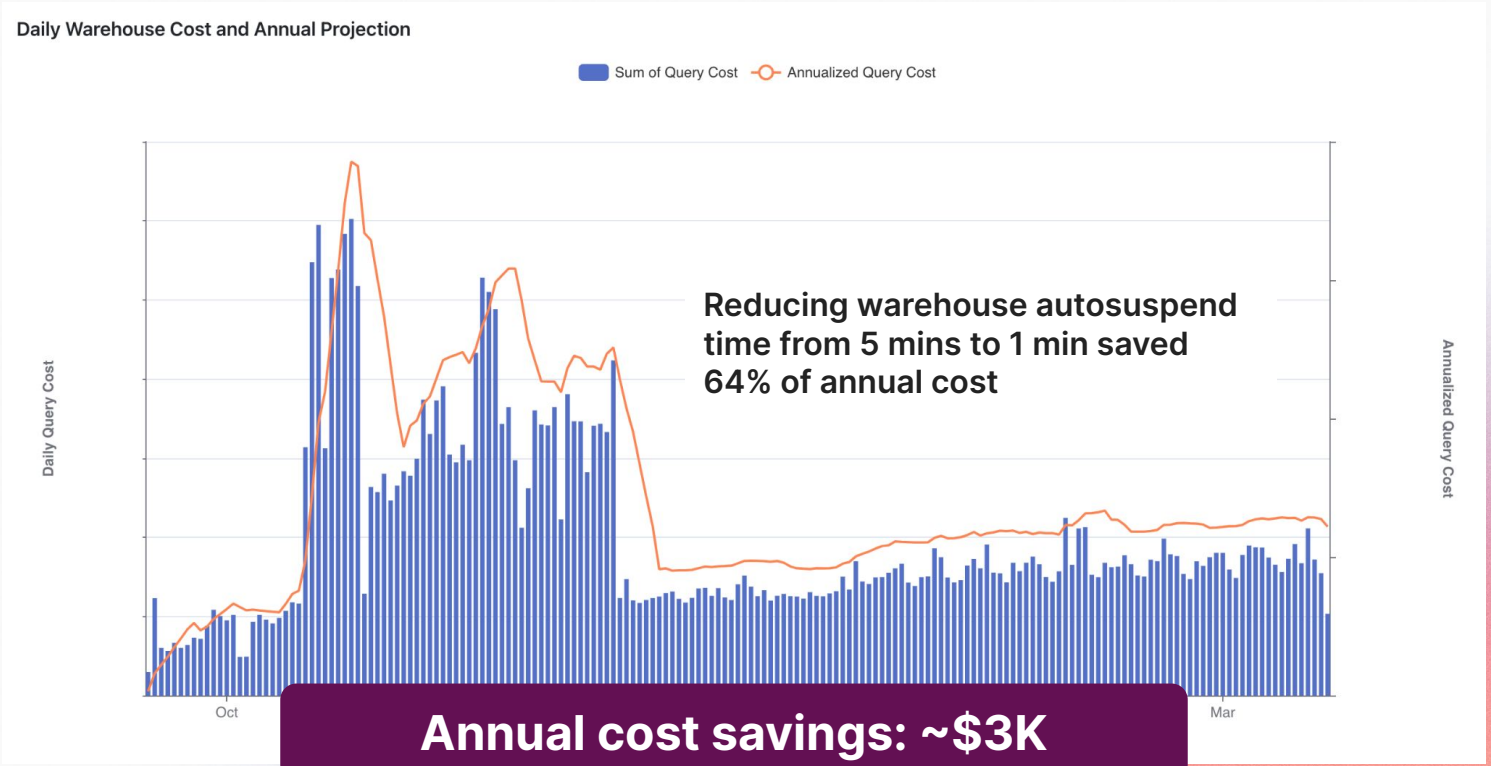
Optimize

Controlling Costs by Lifecycle Stage

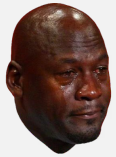


Source: Fundamentals of Data Engineering (Joe Reis and Matt Housley, 2022)

Snowflake Autosuspend Configuration



Open Source and Free Tiers: ETL



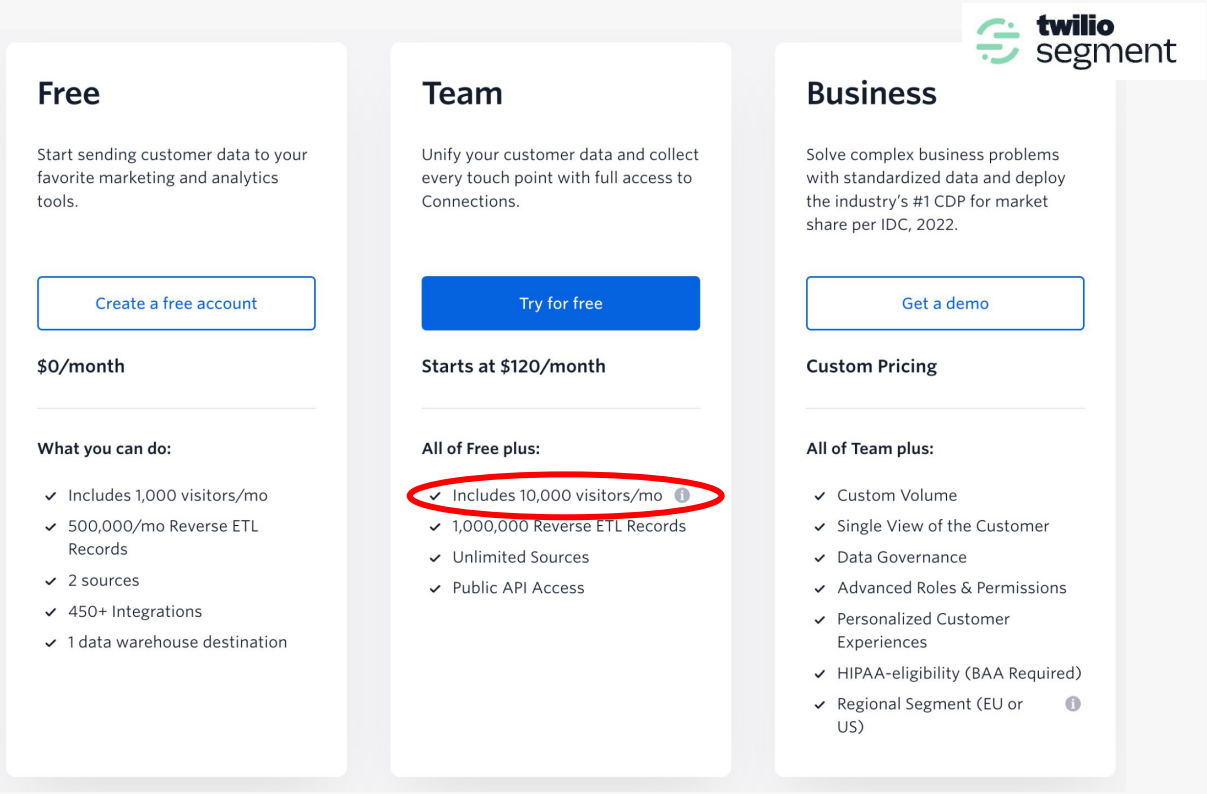
We got burned by Fivetran one too many times, switch to Airbyte OS

Moved to Fivetran Free Tier for SaaS applications

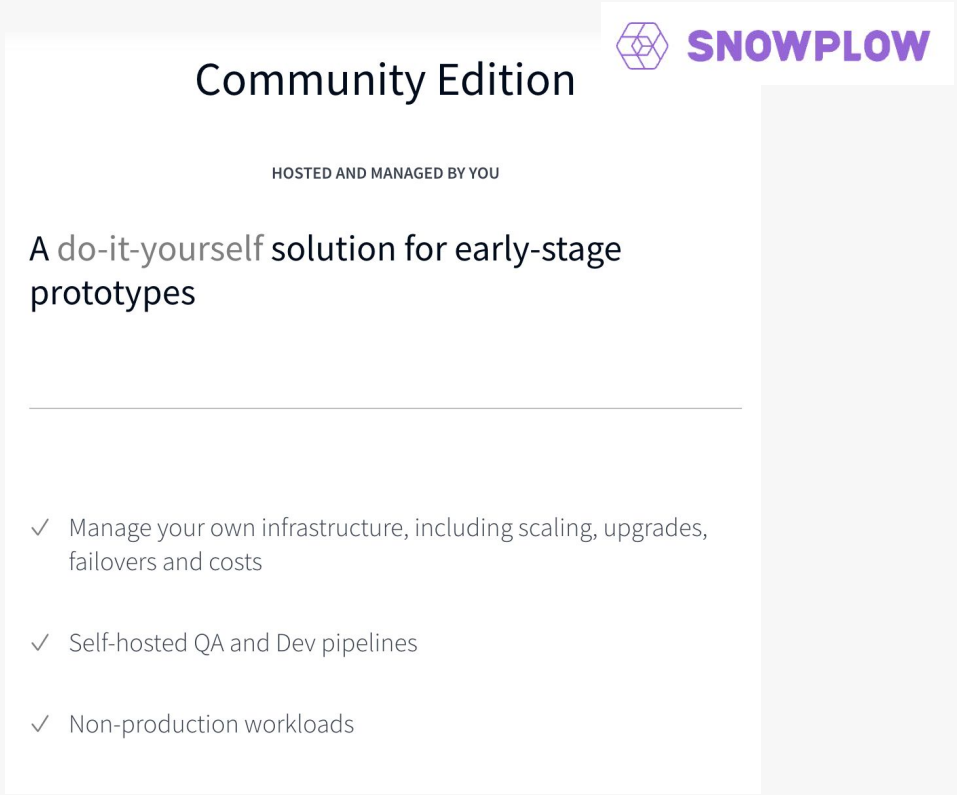
Switching power and familiarity with multiple ETL tools

Annual cost savings: ~$7.5K

Consider Total Cost of Ownership



Consider Total Cost of Ownership



Human Behaviour: Processes, Workflows, and Accountability

Processes

- “Use only what you need”
- Business need driven SLAs
- Data exhaust vs. data creation
- Asset deprecation process

Workflows

- Query optimization training
- Training for data producers and consumers
- Build cost optimization into your workflows
- Set cost safeguards where possible (e.g. resource monitors)
- Build cost monitoring dashboard and set up alerting
- Code reviews and feedback loops

Filters

All of the following dimension conditions match:

Abc Dbt node name	is	stg_mixpanel_app__events ×
Abc Dbt target schema	includes	dbt_ci_run_ ×

+ Add filter

Chart

Close configure

Accountability

- Cost of feature delivery
- Share accountability
- Normalize cost discussions
- Educate on cost reduction

User name	Total query cost
AIRBYTE_USER	$
FIVETRAN_USER	$
GITHUB_ACTIONS_DBT	$
SEGMENT_USER	
DBT_LINDSAY	
RETOOL_USER	
LINDSAY	
DBT_CAMERON	
DBT_LIGHTDASH_USER	
ANDREW	

Scale

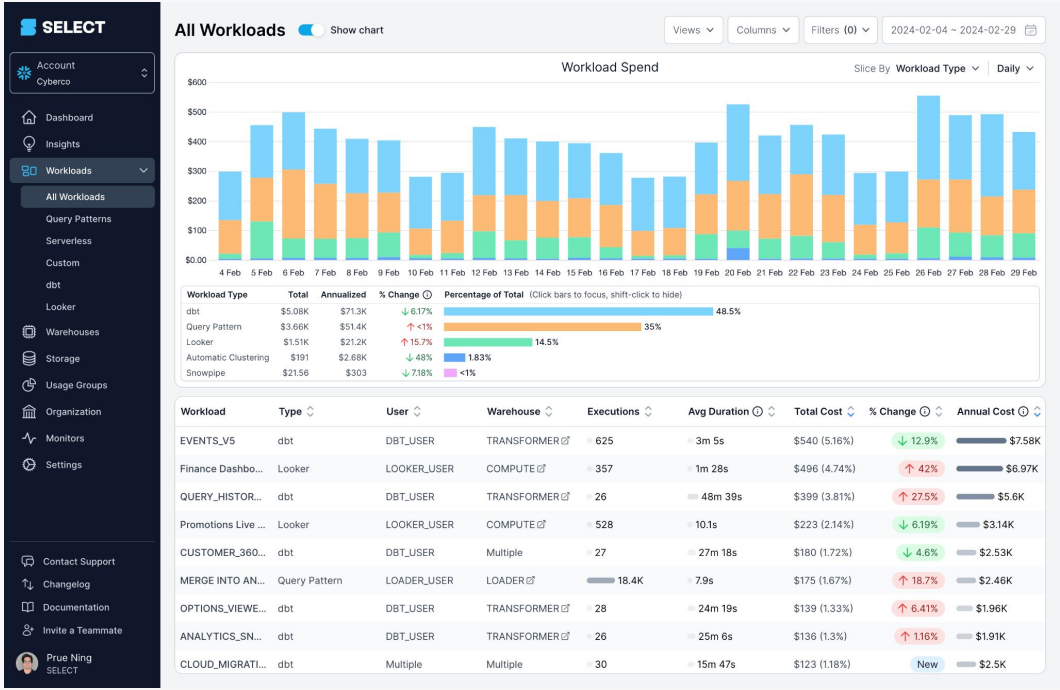
Forecasting

- Create targets and projections
- Use annualized metrics
- Track monthly and quarterly trends
- Alerting

Watch Outs

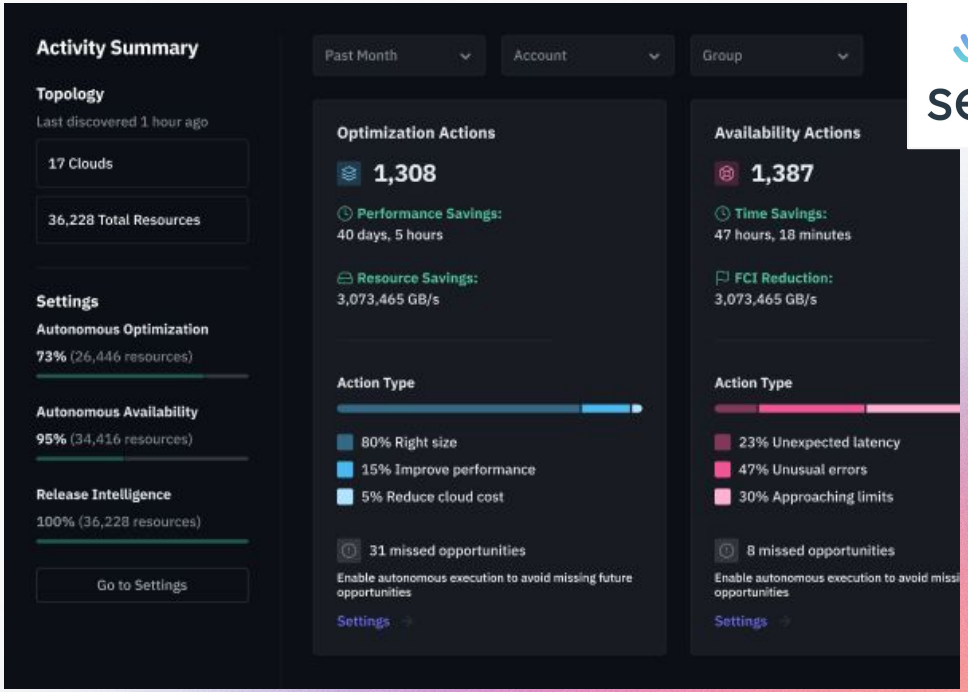
- Be skeptical and avoid linear scaling
- Ensure tool owners understand TCO and scaling costs
- Negotiate contract terms to reduce cost
- Avoid tool lock-in and weigh migration efforts

Cost Reduction Tools



*Disclaimer: I'm not sponsored by SELECT

Autonomous AI



*Disclaimer: I'm not sponsored by Sedai

Takeaways

Takeaways

1. Invest in cost measurement and monitoring
2. Build cost containment feedback loops
3. Drive accountability for cost containment

If you’d like to chat, you can find me on LinkedIn 🥳