



The Modern Data Stack: Past, Present, and Future

TRISTAN HANDY

CEO, @ Fishtown Analytics (maintainers of dbt)

This talk:





We are on the edge of another massive wave of innovation.

Let's make some predictions about where things are heading.

Why am I here pontificating about the past, present, and possible future of data?

Data practitioner

20 years as a data analyst

Founder of a data company

Fishtown Analytics
(we make dbt)

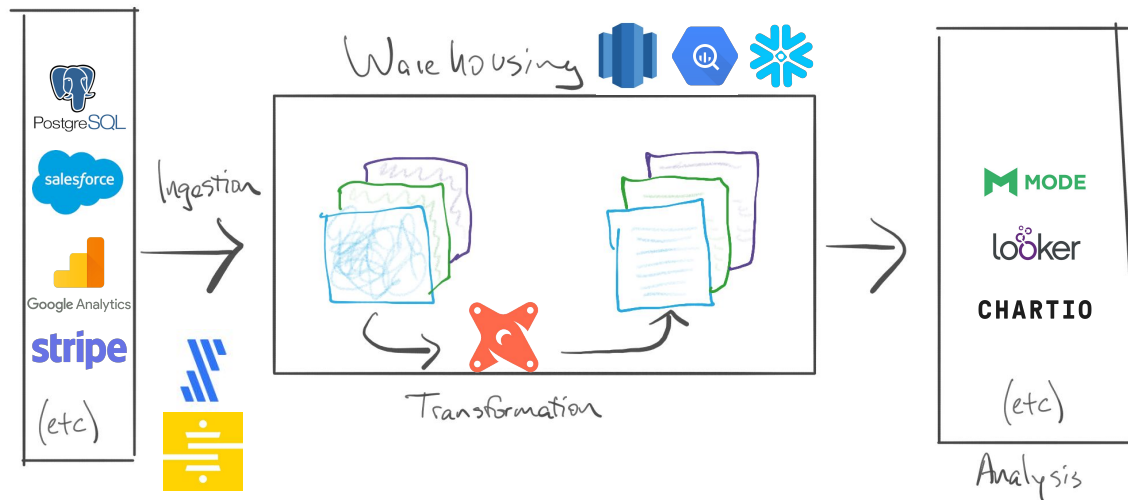
Avid follower of data trends

Data science roundup

Three Eras



The "Modern Data Stack"





Caveats

- I love all of these products! Literally, they've changed my career.
...I don't speak for them, though.
- No list of products can be all-inclusive! Sorry :(

Cambrian explosion I, from 2012–2016



**In the beginning,
there was Redshift.**



Redshift was a world apart from existing solutions.

Redshift vs:

- **Horizontal tools** (GA / Mixpanel / Salesforce):
Analyze *all* data, derive far more insights. No “silos”.
- **OLTP databases:**
Get answers 100–1000x faster.
- **Enterprise OLAP** (Vertica, Netezza):
Starting price of \$160 / month, down from \$100k / year.
- **Hadoop ecosystem:**
95%+ reduction in cost of ownership
- **Excel:**
...don't even start.

How to build a BI tool: pre-Redshift

Ingestion + Storage + Processing + Transformation + Analysis

How to build a BI tool: post-Redshift

Ingestion

Storage + Processing

Transformation

Analysis



I personally experienced this shift.





BY BOB MOORE
December 9, 2019



MOST READ

1

My \$2.6 Billion Ecosystem Fail: an RJMetrics Post Mortem

BY BOB MOORE

2

Partnerships 101: ISVs, VARs, SIs, MSPs, and the Glue that Holds them Together

BY BOB MOORE



Partnerships 101: Account

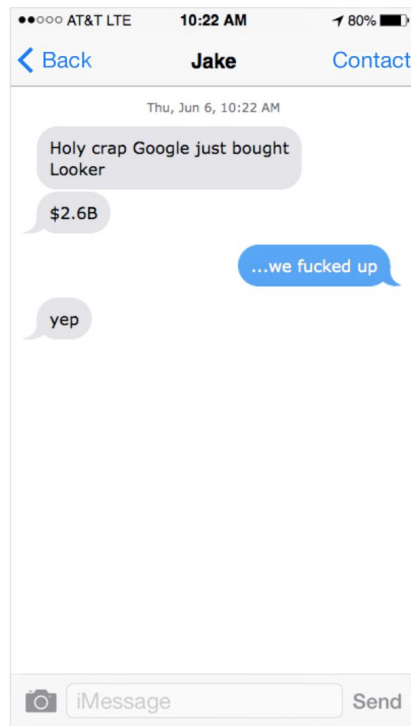
My \$2.6 Billion Ecosystem Fail: an RJMetrics Post Mortem

This June, a headline flashed across my phone and I knew one of my most serious business missteps had finally been finalized: Google had purchased Looker for \$2.6 Billion.

Before founding Crossbeam, I was the co-founder of RJMetrics and Looker was one of our main competitors in the business intelligence space. They had a great product and their team was first-class. Even so, I was disappointed at the outcome: We had all of those things *plus* a four-year head start.

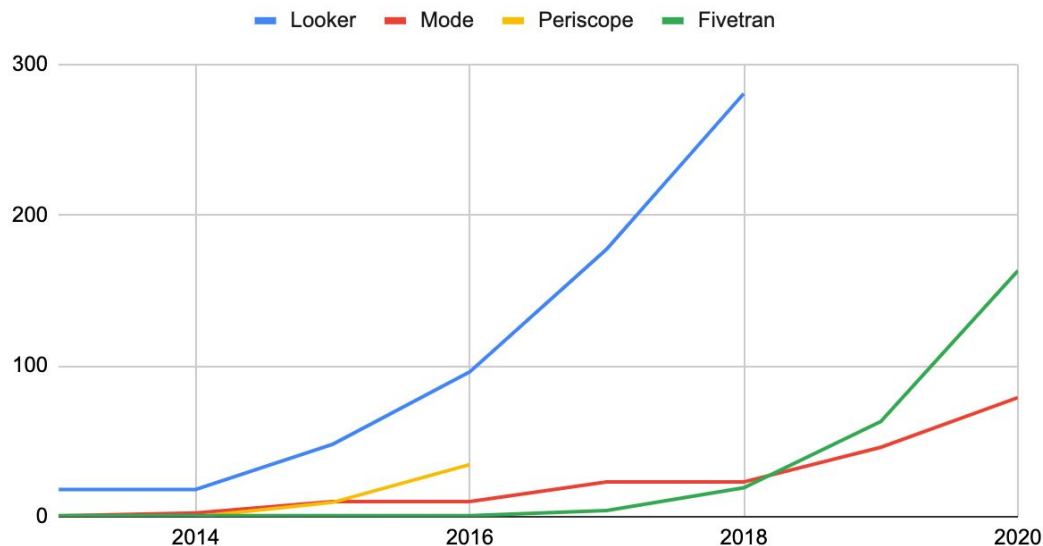
And yet, we ultimately sold RJMetrics to Magento in a modest transaction that was orders of magnitude away from the \$2.6 Billion windfall earned by Looker. What was the difference between our companies?

You could point to a hundred things, but if you dig deeply enough one core product decision is at the root of most of them: Looker placed itself at the center of a massive ecosystem, while RJMetrics operated as a silo. They made other products more valuable, and we were where your data went to die. What felt like a strategic advantage — we were a one-stop shop, the only thing you would need—ended up being our downfall.



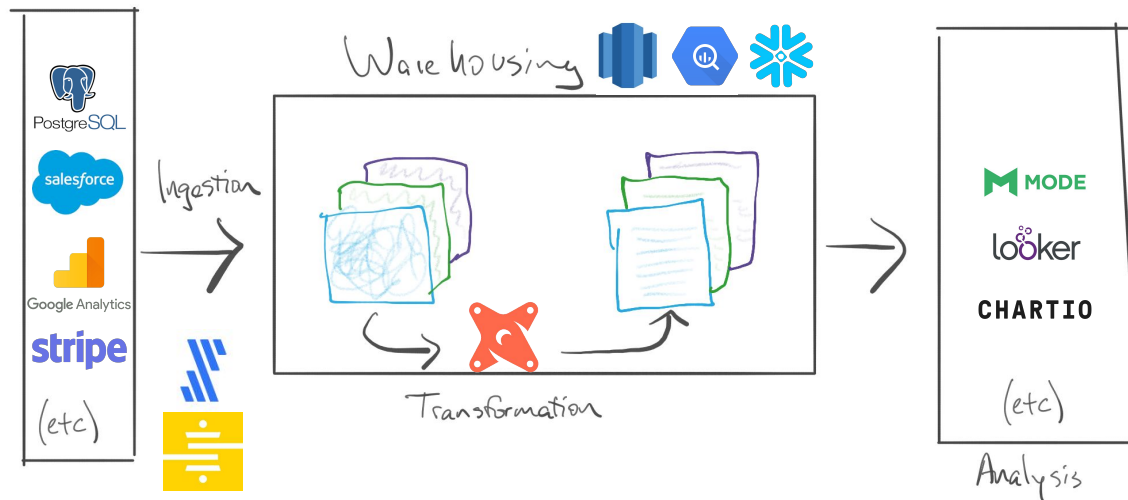
MDS Members Accelerate Post-Redshift

Funding Raised: Selected MDS Members

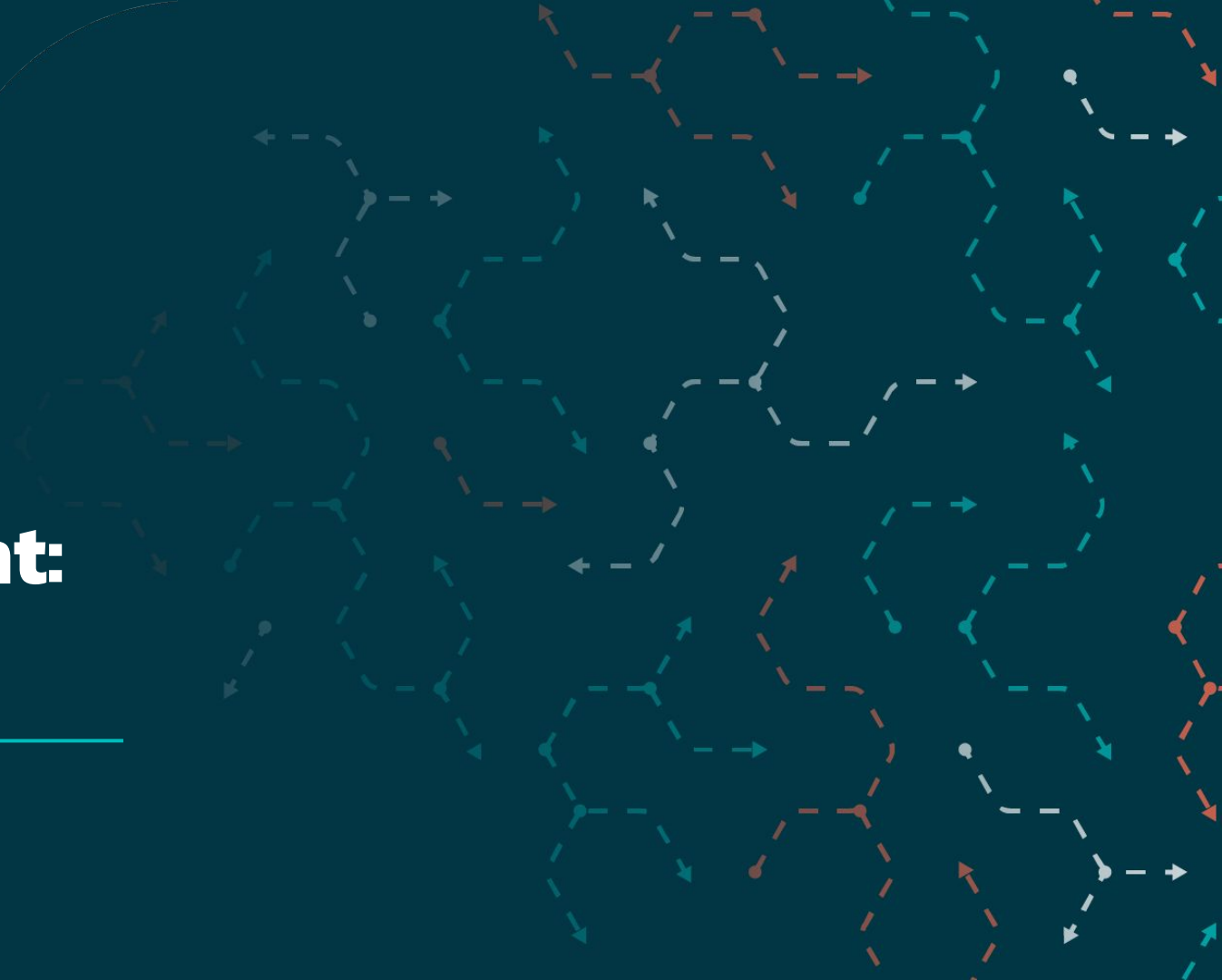


Redshift
launched!

The "Modern Data Stack"



Deployment: 2016–2020



During the next four years at Fishtown Analytics, we implemented the same familiar mix of tools.

Ingestion

- Fivetran
- Stitch

Warehouse

- Redshift
- Snowflake
- Bigquery

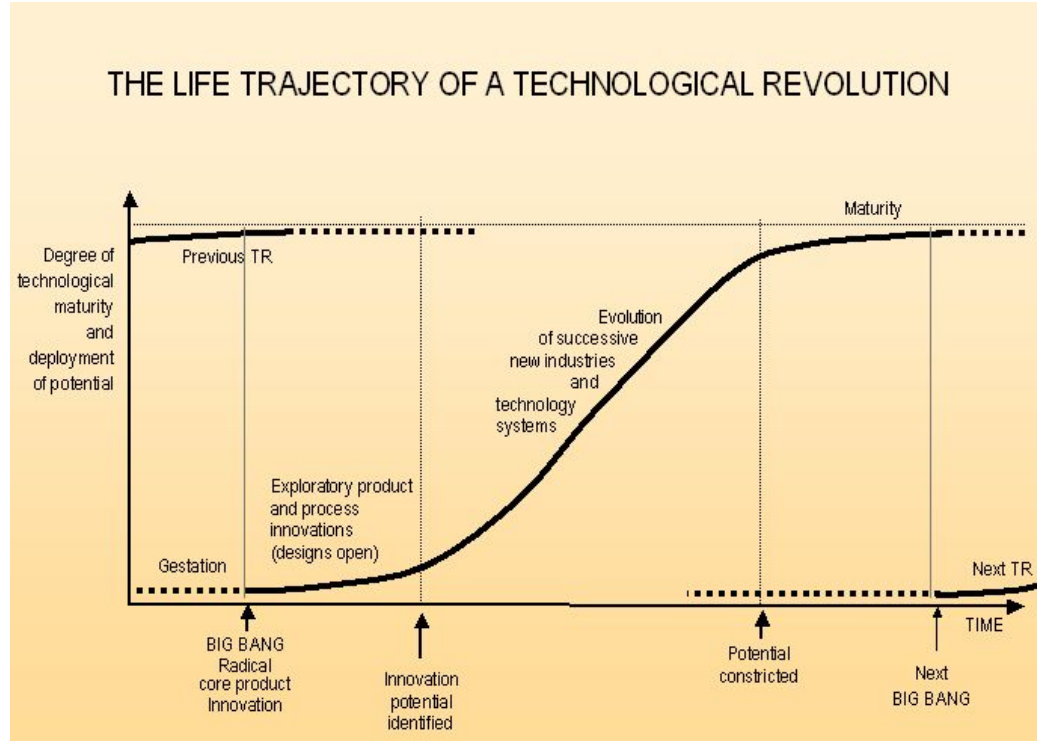
Modeling

- dbt

BI

- Looker
- Mode
- A sprinkle of Redash + Metabase

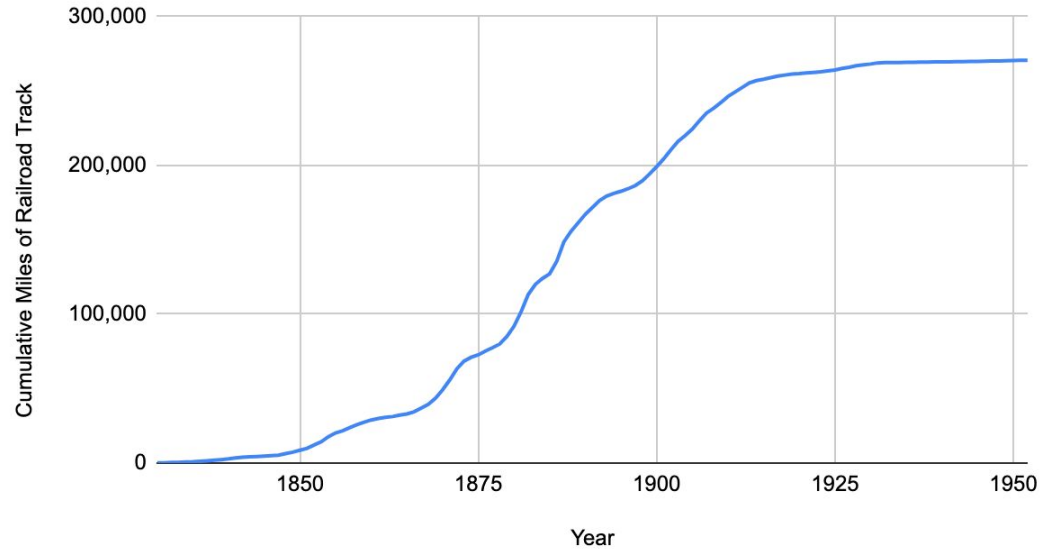
Innovation theorists call this pattern "The S-Curve"



All credit (and so much respect!) to Carlota Perez.

Textbook Example: The Railroad

Cumulative Miles of Railroad Track vs. Year

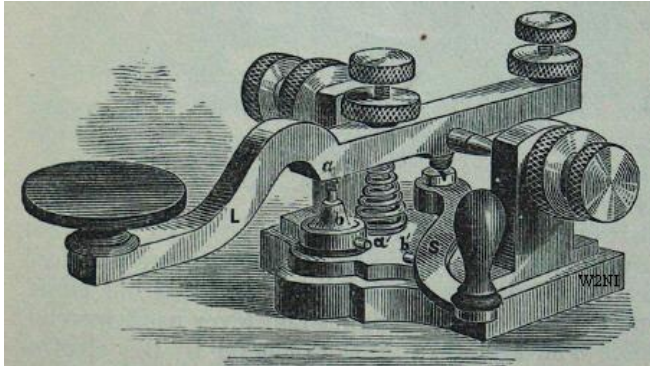


Source: <https://fred.stlouisfed.org/series/A02F2AUSA374NNBR>

Early adopters are forgiving, but technology needs to improve to reach mass adoption.

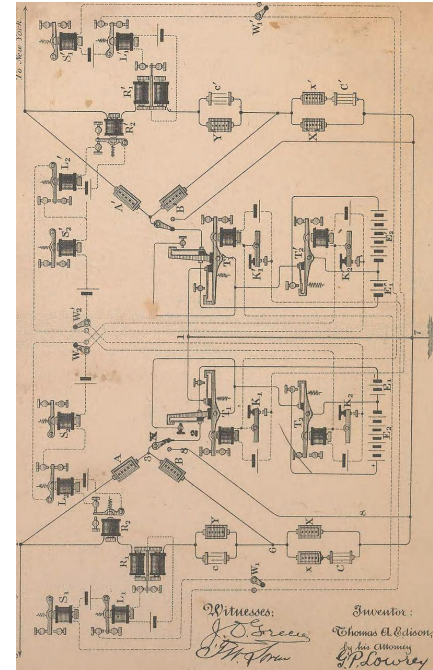
New Machine

The telegraph (1844)



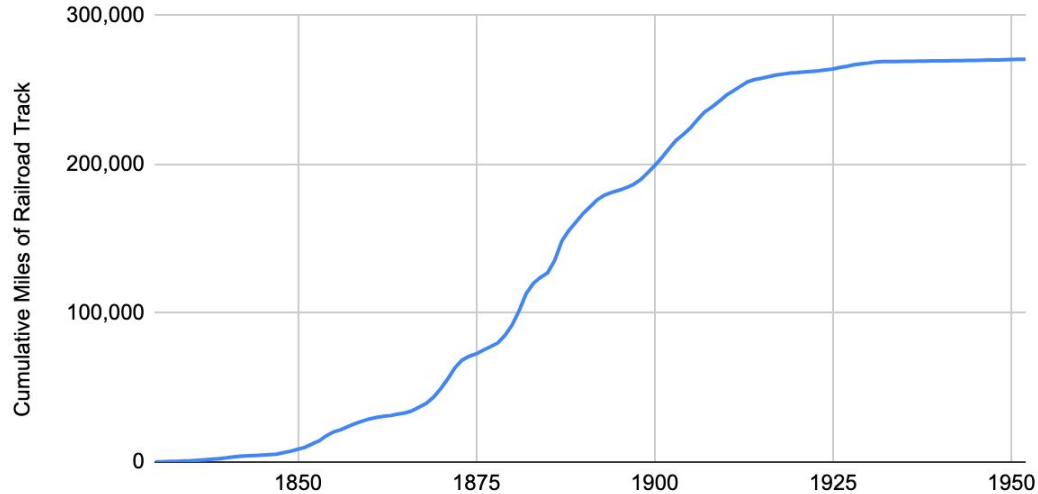
Better Machine

Quadruplex telegraph (1874)



Railroad and Telegraph: Inextricably Intertwined

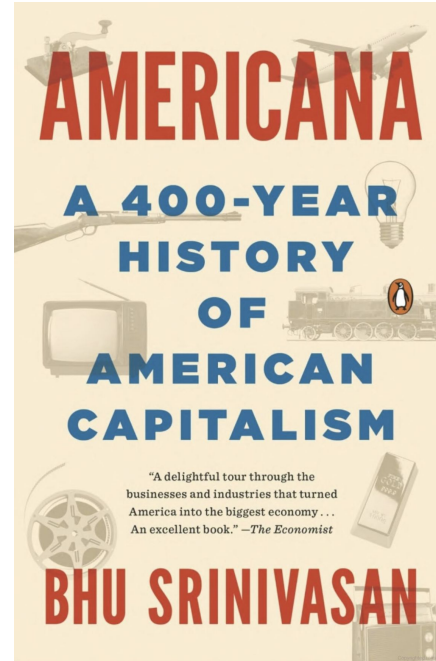
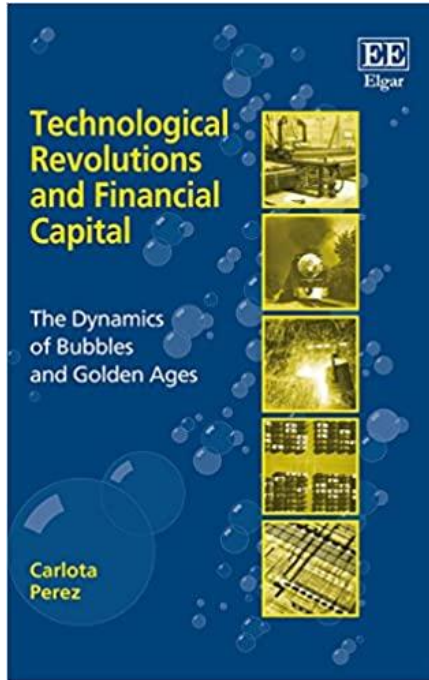
Cumulative Miles of Railroad Track vs. Year



telegraph

quadruplex
telegraph

**Turns out, this pattern is
very normal.**





What's actually going on
right now?

Products are maturing.

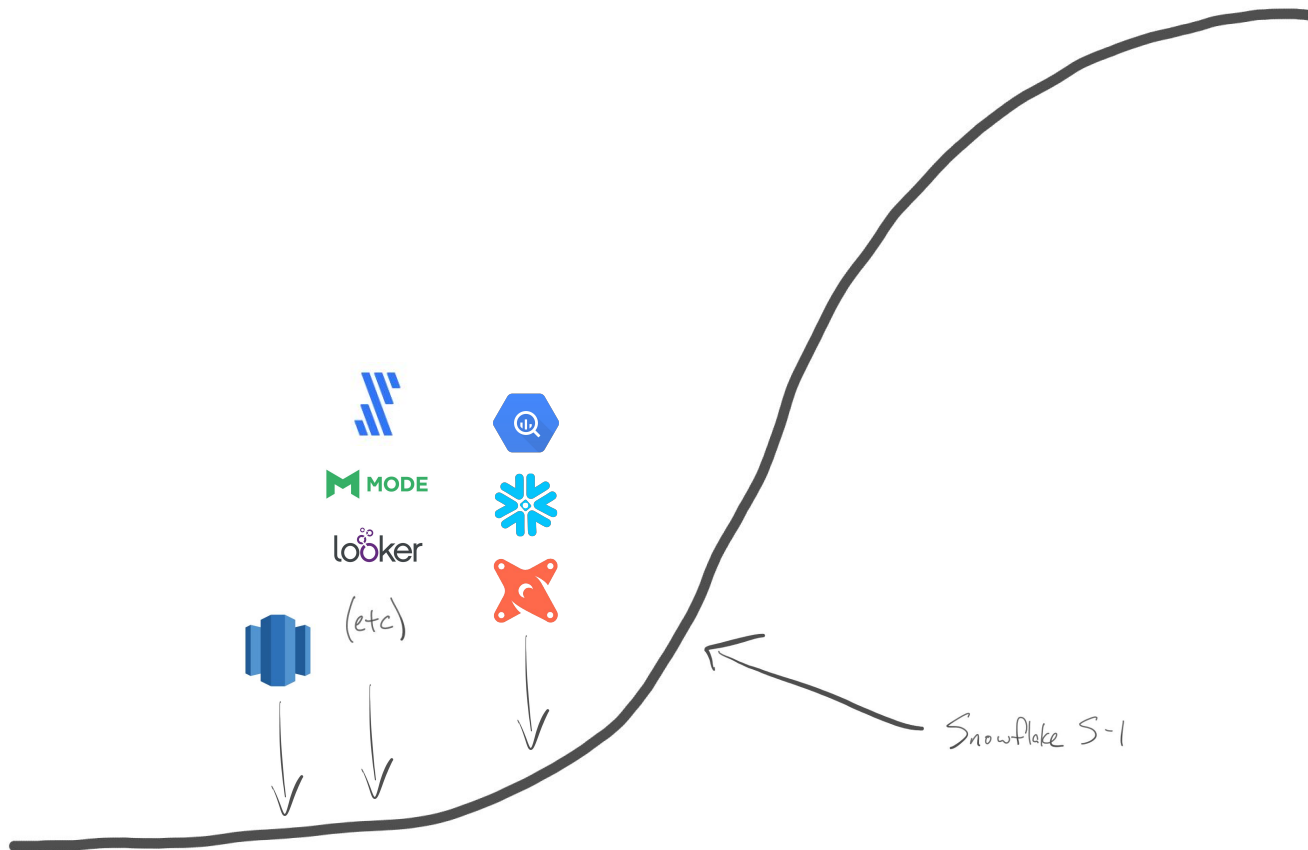
The next phase is being unlocked.

A case study!



dbt

So, where are we?



So, where are we?

Horizontal Products

One warehouse. One set of tools to analyze all data.

Speed

Every part of the data iteration cycle has sped up.

Unlimited Scale

Cost is the only limiting constraint on data processing.

Low Overhead

The modern data stack is incredibly easy to set up and manage.

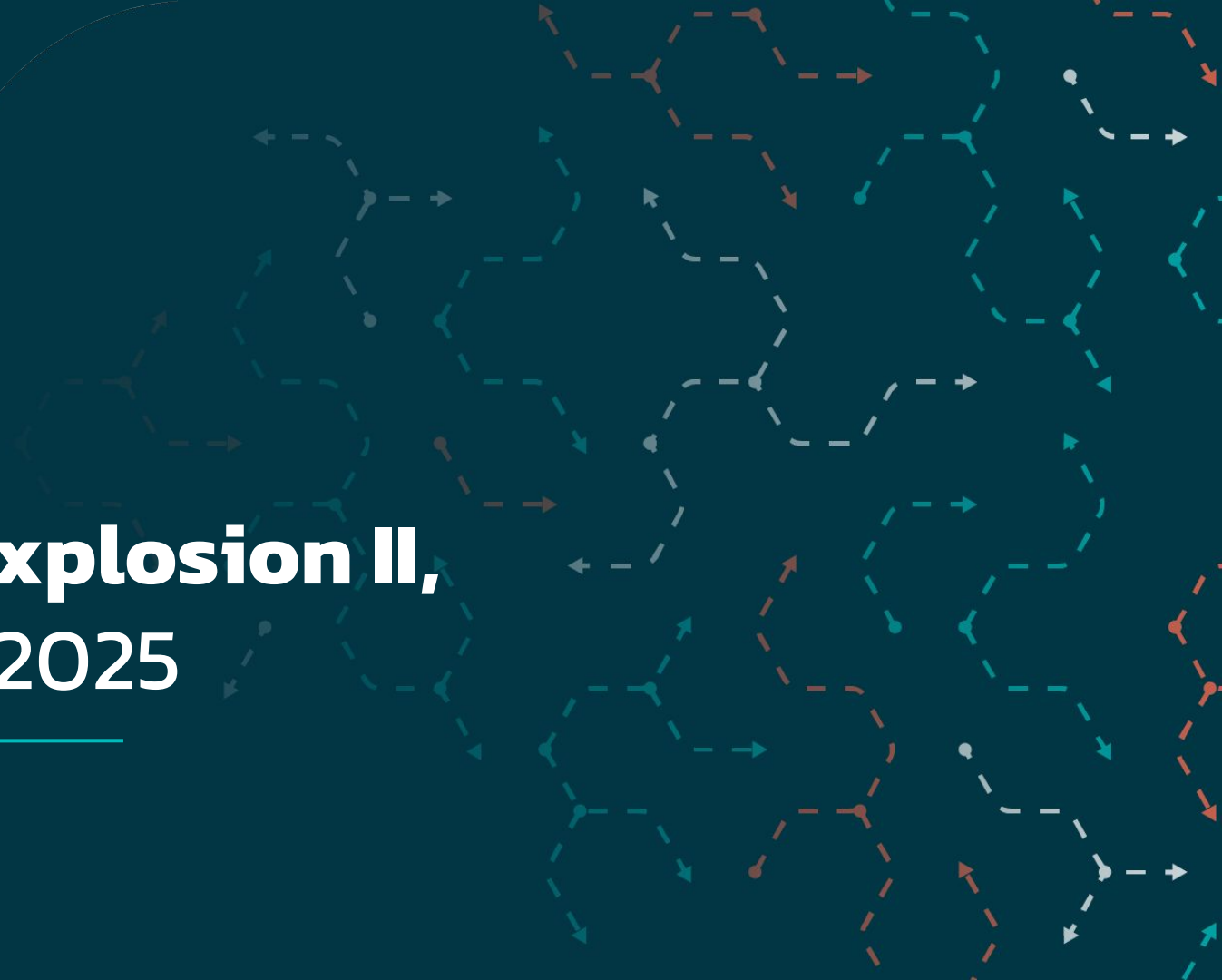
United by SQL

SQL has become the standard data language.

Widespread Integrations

Customers have come to expect that most datasets can be piped off-the-shelf.

Cambrian explosion II, from 2021 – 2025





We are on the edge of another massive wave of innovation.

Let's make some predictions about where things are heading.

What are the next big categories?

Let's look at the painful problems.

Type I

- **Governance is immature**
- **Self-service vision not truly realized**
- **Vertical analytical experiences**

Type II

- **Batch-based processing limits operational use**
- **Data doesn't feed back into operational tools**

Opportunity #1:
Governance

Governance

- **Who made this?**
- **Can I trust it?**
- **How should it be used?**
- **When was it last updated?**
- **Is this frequently used or can I delete it?**
- **Is the data current?**
- **Are there any data issues?**
- **Where can I find data about [topic]?**
- **...**

Governance

Every data-driven company is trying to solve it

DataHub

By LinkedIn

Amundsen

By Lyft

Marquez

By WeWork

Dataportal

By Airbnb

Lexicon

By Spotify

Metacat

By Netflix

Databook

By Uber

Arti.FACT

By Shopify

explorers




Sep 13, 2004 – Mar 21, 2019

Data for famous world explorers

Columns

explorer_id <small>int</small>	▼
primary key for explorers	
first_name <small>string</small>	▼
Explorer's given name	
last_name <small>string</small>	▼
Explorer's family name	
birthday <small>date</small>	▼
Explorer's date of birth	
place_of_origin <small>string</small>	▼
Country of birth	
regions_explored <small>int</small>	▼
Count of regions explored. Regions defined in explored_regions	
distance_travelled <small>int</small>	▼
Total kilometers travelled based only on entries in trips table	
nationality <small>string</small>	▼
Country of citizenship, may be different from place_of_origin	

OWNED BY

-  email@lyft.com
-  user@lyft.com
-  Add

FREQUENT USERS



GENERATED BY

 explorers.all_explorers



SOURCE CODE

 explorers.all_explorers

TABLE LINEAGE (BETA)

 explorers.all_explorers

TABLE PROFILE (BETA)

-  Preview Data
-  Explore with SQL

TAGS

 explorers



This is happening **now**,
unlocked by talent and VC \$\$.

Early commercial products
coming to you soon.

Opportunity #2:
True Self-Service

Democratized data exploration

🌶️ The modern data stack has disempowered day-to-day decision makers

Democratized data exploration

🌶️ The modern data stack has disempowered day-to-day decision makers

🌶️🌶️ For many decision-makers, Excel-land was actually better than the modern data stack

Democratized data exploration

🌶️ The modern data stack has disempowered day-to-day decision makers

🌶️🌶️ For many decision-makers, Excel-land was actually better than the modern data stack

🌶️🌶️🌶️ What if a spreadsheet-like interface is still the best option?



What's the unlock?

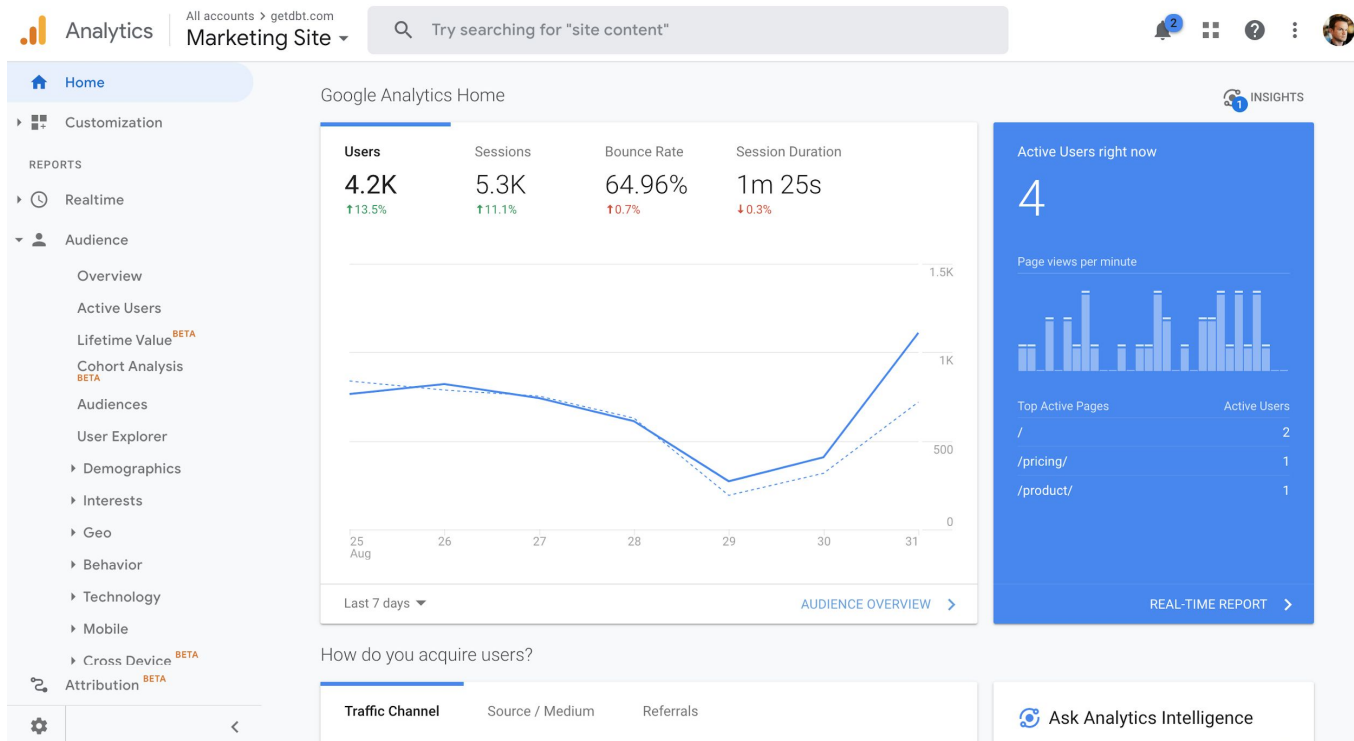
The right user experience

Who I'm watching

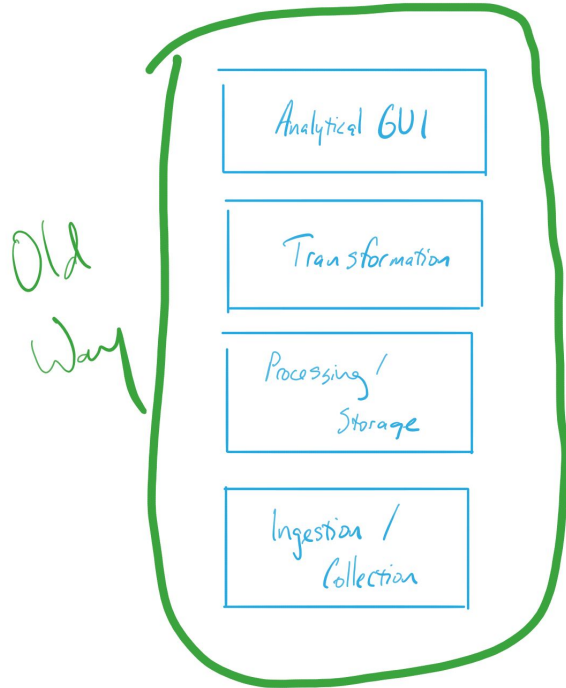
SeekWell, Sigma Computing,
new players

Opportunity #3:
Vertical Analytical Experiences

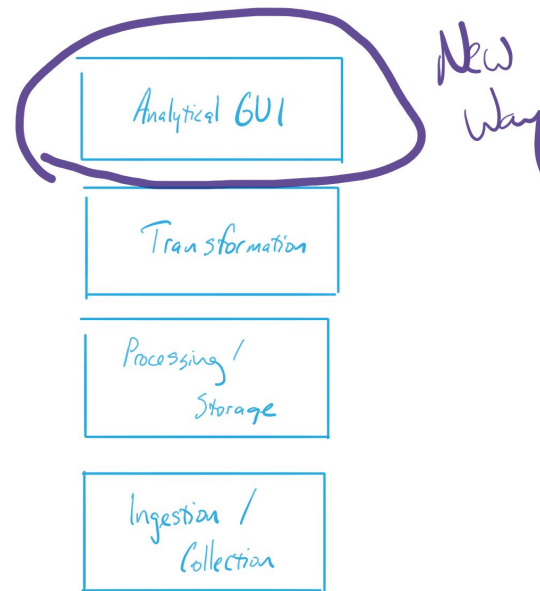
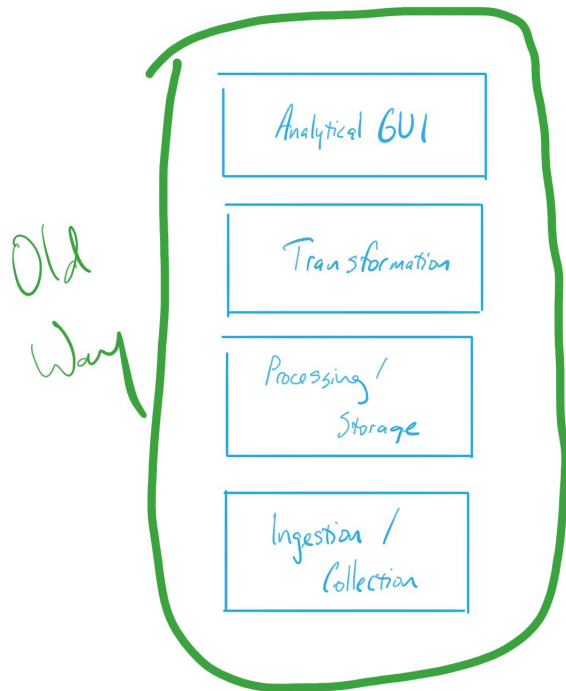
What is a “vertical” analytical experience?



Vertical analytical experiences



Vertical analytical experiences





What's the unlock?

Modern data stack market size

Who I'm watching

Still waiting...

Opportunity #4:
Real Time & Operational

What are “operational” use cases?

In-product analytics

Power dashboards inside your product

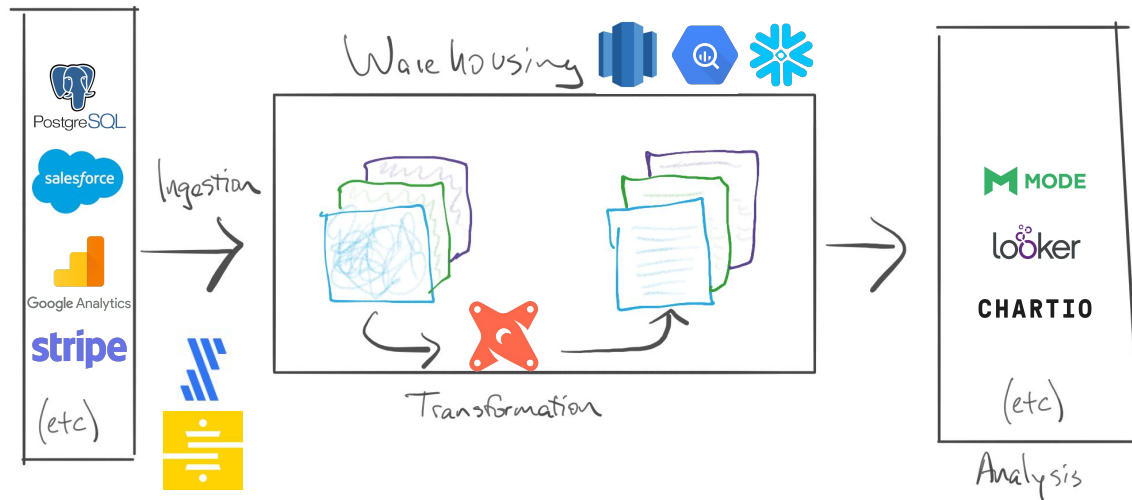
Operational intelligence

Deliver inventory and logistics information to frontline employees.

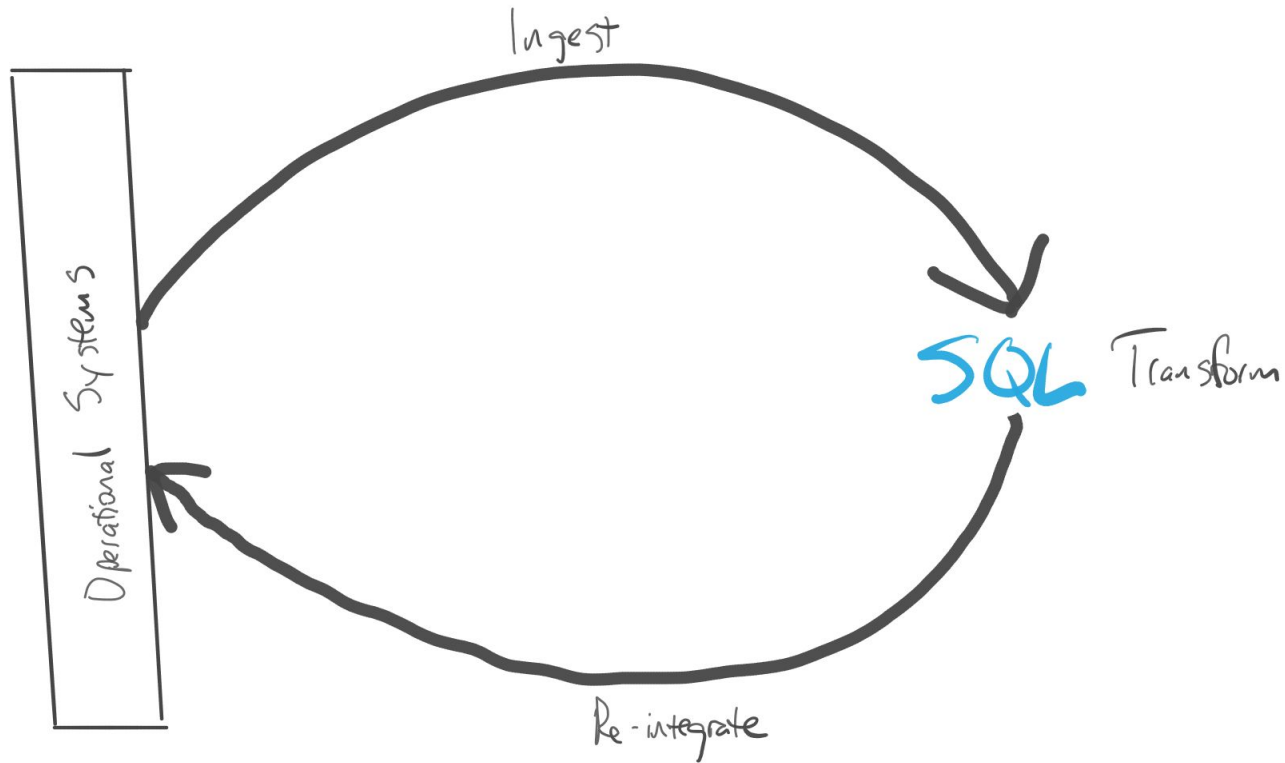
Process automation

Push data back into CRM / messaging / other operational apps to trigger workflows.

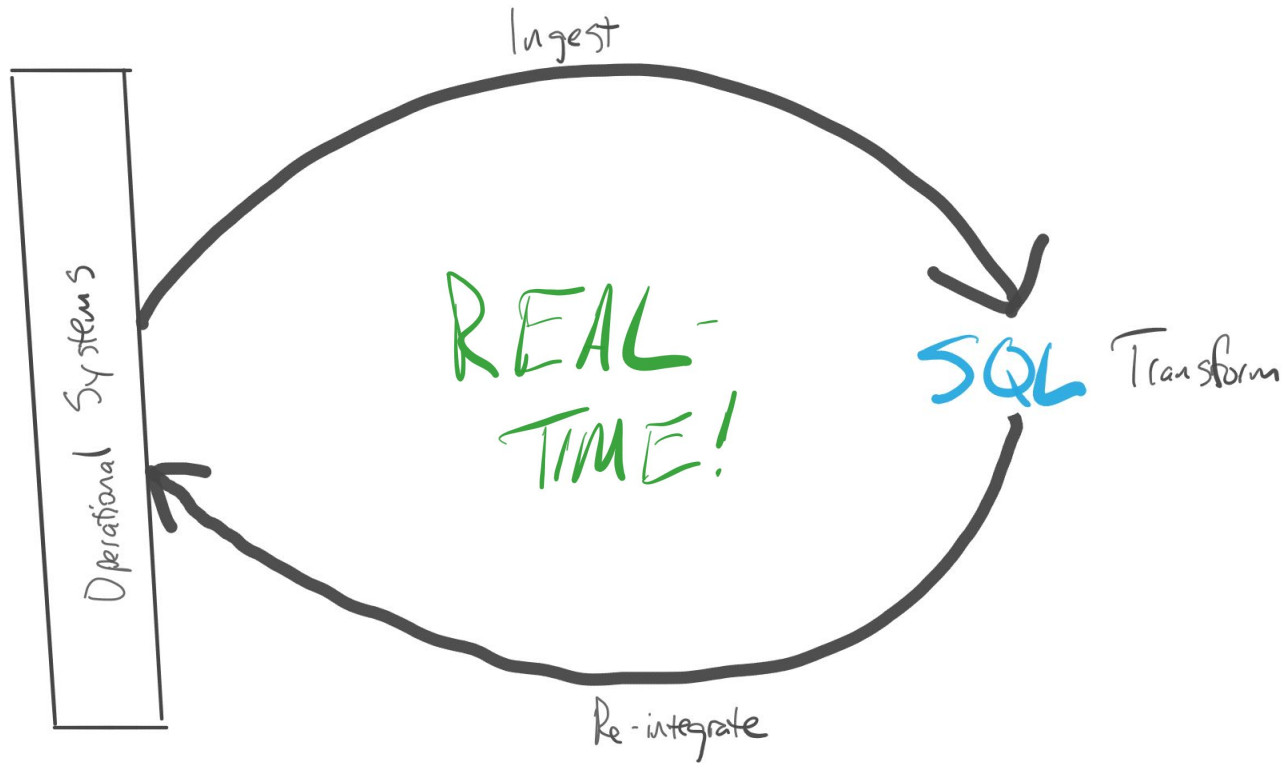
All the arrows go left-to-right!



Rewire the Modern Data Stack



Rewire the Modern Data Stack



What needs to be built?

- Streaming Ingestion
 - Debezium
 - Meroxa
- SQL-based real-time processing
 - “Big 3” warehouse providers
 - KSQL
 - Materialize
- Data re-integration
 - Census
 - Tray

Moving further up the curve



Drawing a new curve





What's the unlock?

Tech maturity.

We need 1m end-to-end pipeline
latency.

Who I'm watching

Materialize, Meroxa, warehouses,
KSQL, Census, Tray...

In the future...

- ...you'll have tools to manage the chaos.
- ...self-service will be realized.
- ...lightweight, vertical-specific tools will become commonplace.
- ...your pipelines will flow in near-real-time
and will feed back into your operational systems,
allowing data professionals to "program your business".



THANK YOU

@jthandy on Twitter
getdbt.com