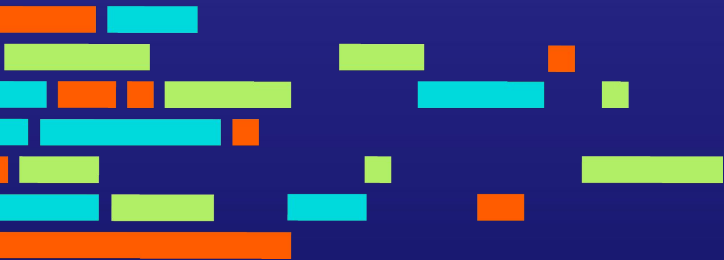


**GNARLY**  
Data\_Waves

PRESENTED BY  **dremio**

EPISODE 24

# Simplifying Data Mesh with Dremio's Open Data Lakehouse

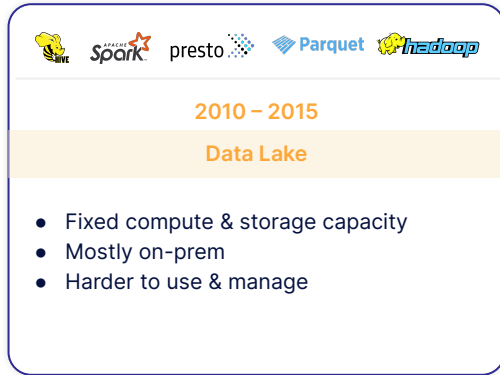


SPEAKER

## Nik Acheson

Sr. Director,  
Product Management and GTM Strategy  
Dremio

# Leading the Present and Future of Data Analytics



2010 – 2015

**Data Lake**

- Fixed compute & storage capacity
- Mostly on-prem
- Harder to use & manage

Logos: Hive, Spark, presto, Parquet, Hadoop

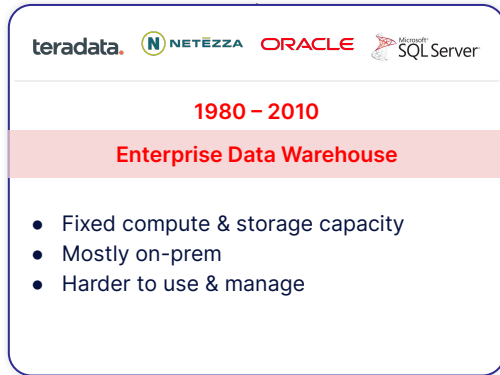


2022-...

**Data Lakehouse**

- Data in open file and table formats
- No need to copy & move data
- Multiple best-of-breed processing engines

Logos: dremio, ICEBERG, DELTA LAKE, Redshift, Snowflake

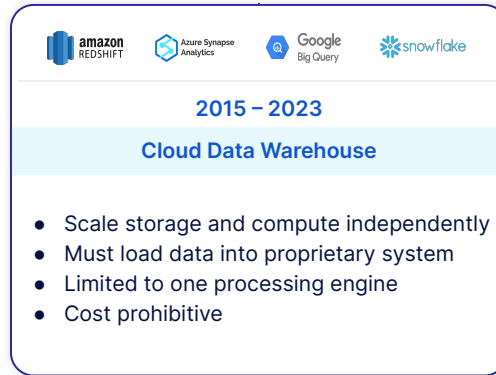


1980 – 2010

**Enterprise Data Warehouse**

- Fixed compute & storage capacity
- Mostly on-prem
- Harder to use & manage

Logos: teradata, NETEZZA, ORACLE, Microsoft SQL Server

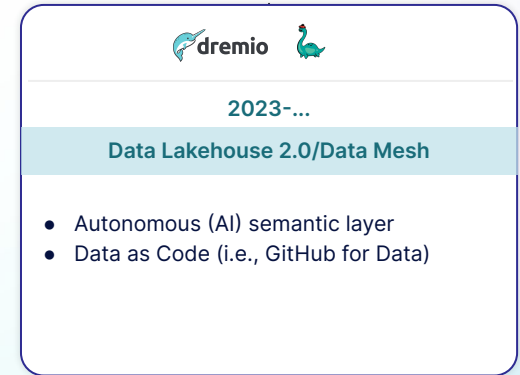


2015 – 2023

**Cloud Data Warehouse**

- Scale storage and compute independently
- Must load data into proprietary system
- Limited to one processing engine
- Cost prohibitive

Logos: amazon REDSHIFT, Azure Synapse Analytics, Google Big Query, snowflake



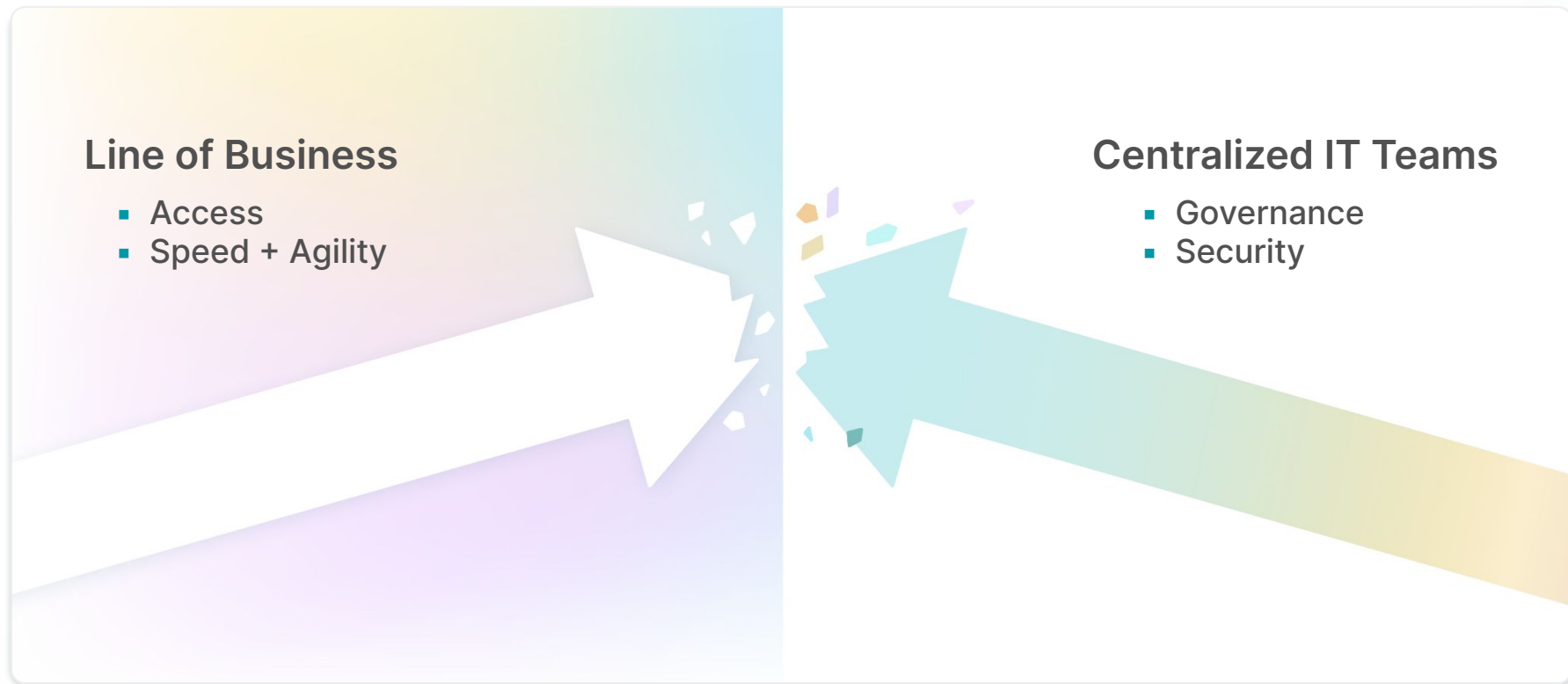
2023-...

**Data Lakehouse 2.0/Data Mesh**

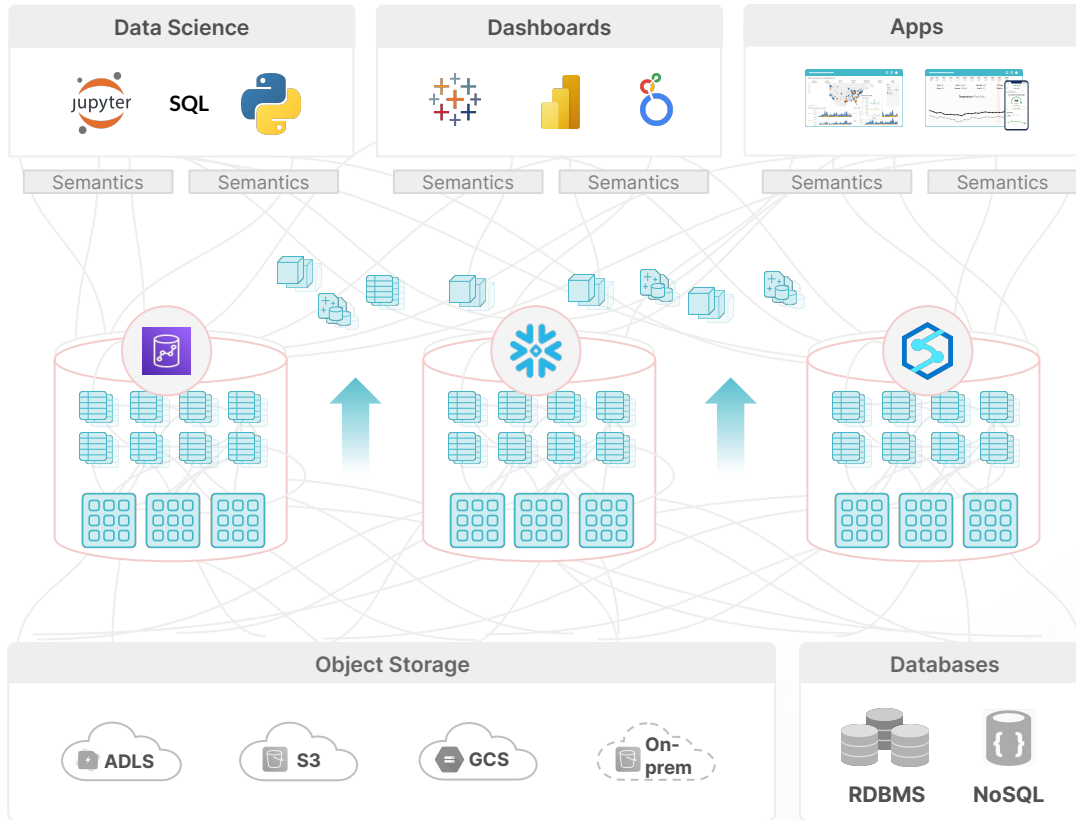
- Autonomous (AI) semantic layer
- Data as Code (i.e., GitHub for Data)

Logos: dremio, Dremio logo

# Intensity of Competing Data Priorities is Increasing

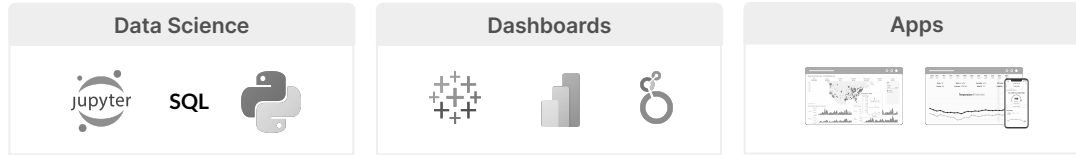


# Data Warehouses: Complex, Proprietary, Expensive

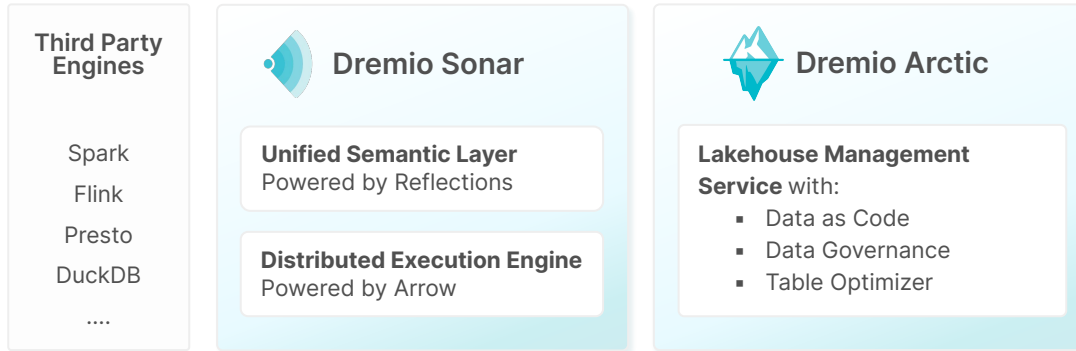


- × Complex
- × Expensive
- × Lock-in
- × Impossible to secure
- × No self-service
- × Limited data exploration
- × Inconsistent data

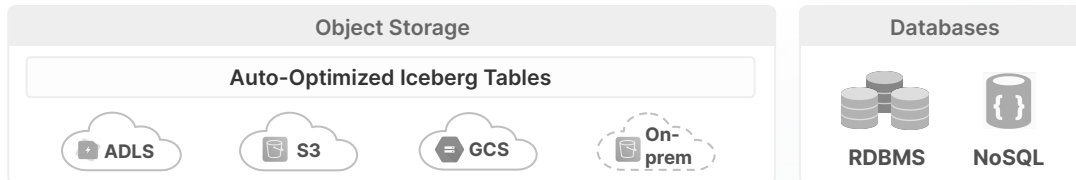
# Dremio Data Lakehouse: Easy, Open, 1/10th the Cost

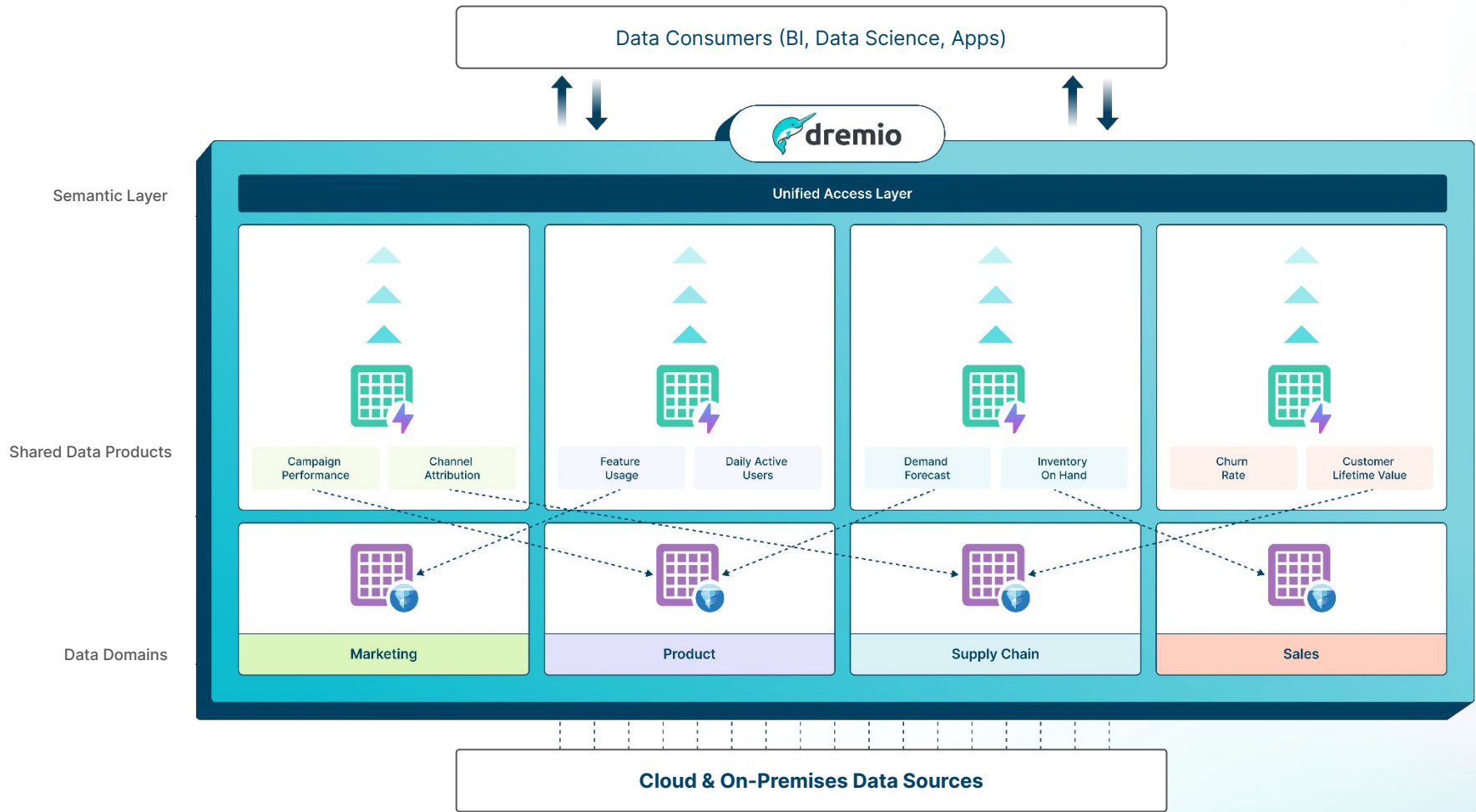


↕ ODBC | JDBC | REST | Arrow Flight ↕



- ✓ No data copies
- ✓ No complex ETL pipelines
- ✓ Inexpensive
- ✓ Self-service analytics
- ✓ Data engineering productivity
- ✓ Security
- ✓ Governance
- ✓ Consistent data





# Organizations Building their Data Mesh with Dremio

*“Dremio is literally going to help MSK save the lives of cancer patients.”*

Arfath Pasha, Lead Software Engineer at Memorial Sloan Kettering Cancer Center



Reshaping the clean energy system  
on Dremio's Open Data Lakehouse

[▶ Hear from Shell](#)



Memorial Sloan Kettering  
Cancer Center

Creating a purposeful scientific data  
management system that accelerates  
cancer research for scientists

[▶ Hear from MSK](#)



Building products for financial  
inclusion and to promote economic  
opportunity

[▶ Hear from Transunion](#)



**Data as a Product**

**Self-Serve Data Platform**

**Data Mesh**

**Domain Ownership**

**Federated Computational  
Governance**

**Data as a Product**

**Self-Serve Data Platform**

**Data Mesh**

**Domain Ownership**

**Federated Computational  
Governance**

# How Dremio Delivers Data Products

**Dataset Graph** Reset focus to "NYC Taxi Trips"

**Parents(2)**

- zips.json**  
Samples: samples.dremio.com.myfol...  
Ref: ↗ mybranch2  
Jobs (last 30 days): 21  
Descendants: 7  
Columns: 3
- zips2.json**  
Samples: samples.dremio.com.myfol...  
Ref: ↗ mybranch  
Jobs (last 30 days): 21  
Descendants: 7  
Columns: 3

**NYC Taxi Trips**  
GCS:"Level 1":"Level 2":"NYC Taxi Trips"  
Jobs (last 30 days): 21  
Descendants: 3  
Columns: 8  
**Columns (8)**  
# NYCTaxiTrips  
date  
pickup\_datetime  
dropoff\_datetime  
# passenger\_count  
# fare\_amount  
# total\_amount

**Children(2)**

- test1**  
Samples: sa...  
Ref: ↗ myf...  
Jobs (last 30 days): 3  
Descendants: 3  
Columns: 3
- test2**  
Samples: sa...  
Ref: ↗ myf...  
Jobs (last 30 days): 3  
Descendants: 3  
Columns: 3

**Edit Wiki**

B I S U

**Volume**  
1 Billion records stored in

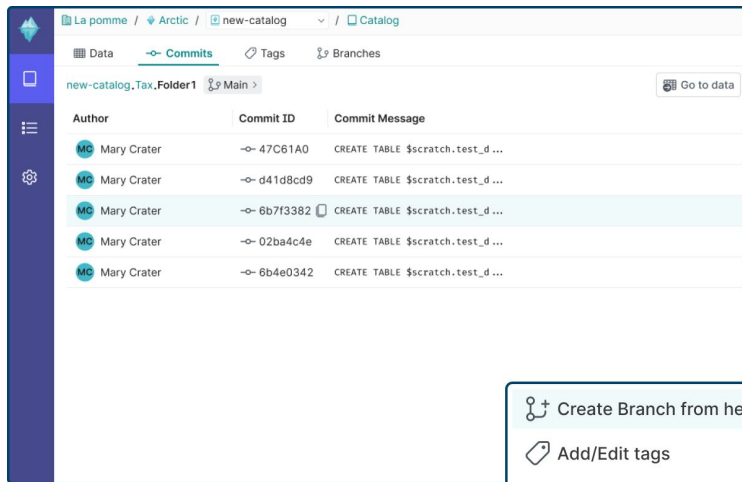
**Data Quality**  
This dataset has been vali

**Description**  
The yellow and green taxi trip records include fields capturing pick-up and drop-off dates/time reported passenger counts. The data used in the attached datasets were collected and provid Taxicab & Livery Passenger Enhancement Programs (TPEP/LPEP). The trip data was not create

## Search, Create, Share, Measure & Manage

- Easily search and access data products without needing help from engineering using the searchable catalog experience.
- Create and publish data across domains and users using the semantic layer. Register data for consumption using the wiki and tags.

# How Dremio Treats Data as a Product



## Data as Code

- Treat data products as code and give data consumers a **consistent & accurate view** of their data at all times.
  - Isolate experiments with branches without impacting other users
  - Easily reproduce models and BI dashboards from previous states with version control
- Use the same data (lake) across your environments without risking the integrity of production data.

**Data as a Product**

**Self-Serve Data Platform**

**Data Mesh**

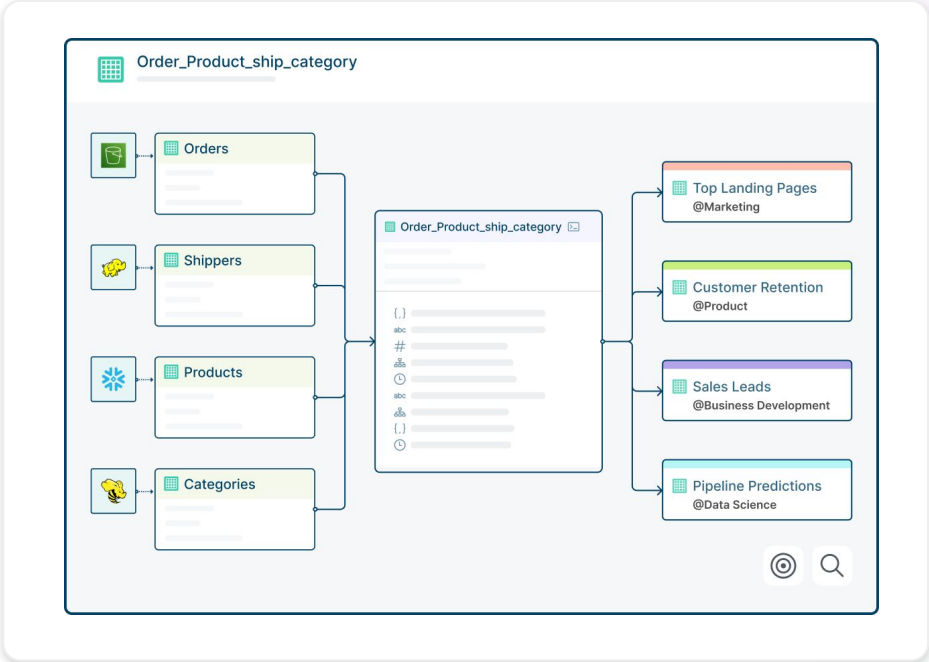
**Domain Ownership**

**Federated Computational  
Governance**

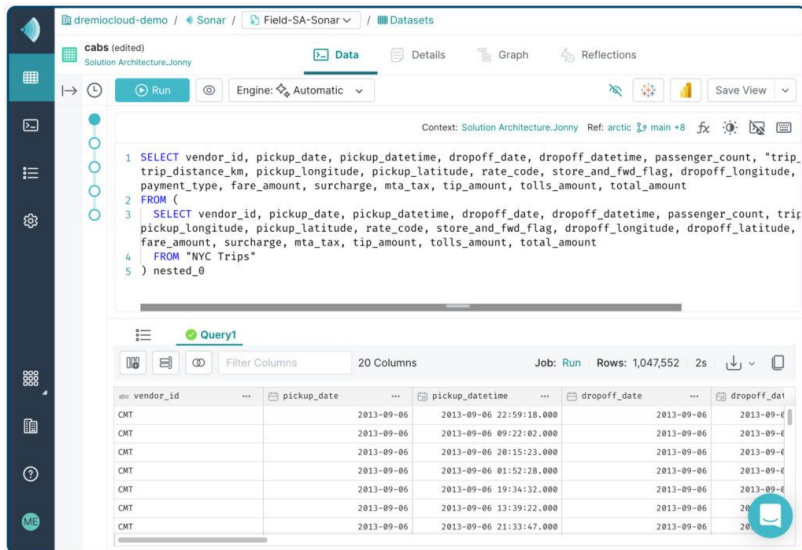
# Self-Service Data Platform with a Unified Semantic Layer

## Unified Access Layer

- Minimize siloed ETL pipelines and federate distributed data across on-premises, hybrid, and cloud environments.
- Dremio's semantic layer improves data discovery, ensures consistent reporting and enables governed self-service data access.



# Self-Service Data Platform



The screenshot displays the Dremio user interface. At the top, the breadcrumb navigation shows 'dremiocloud-demo / Sonar / Field-SA-Sonar / Datasets'. The main workspace is titled 'cabs (edited)' and contains a SQL query editor with the following code:

```
1 SELECT vendor_id, pickup_date, pickup_datetime, dropoff_date, dropoff_datetime, passenger_count, *trip,
trip_distance_km, pickup_longitude, pickup_latitude, rate_code, store_and_fwd_flag, dropoff_longitude,
payment_type, fare_amount, surcharge, mta_tax, tip_amount, tolls_amount, total_amount
2 FROM (
3 SELECT vendor_id, pickup_date, pickup_datetime, dropoff_date, dropoff_datetime, passenger_count, trip,
pickup_longitude, pickup_latitude, rate_code, store_and_fwd_flag, dropoff_longitude, dropoff_latitude,
fare_amount, surcharge, mta_tax, tip_amount, tolls_amount, total_amount
4 FROM "NYC Trips"
5 ) nested_0
```

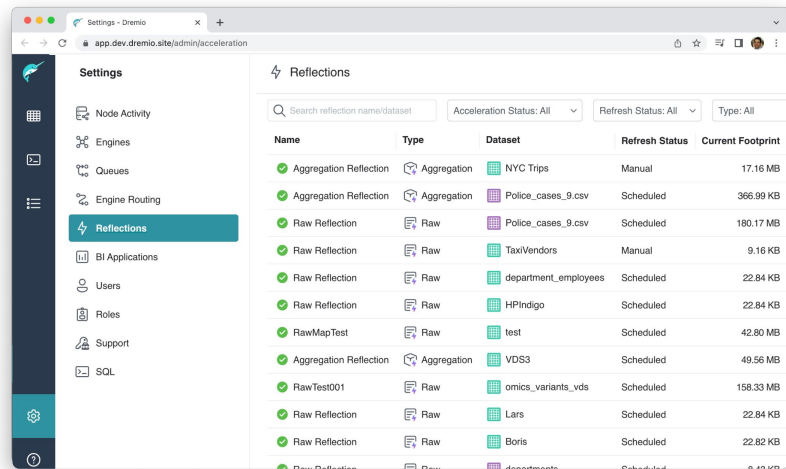
Below the query editor, a 'Query1' results pane is visible. It shows a table with 20 columns and 1,047,552 rows. The table headers are: vendor\_id, pickup\_date, pickup\_datetime, dropoff\_date, and dropoff\_datetime. The first few rows of data are:

vendor_id	pickup_date	pickup_datetime	dropoff_date	dropoff_datetime
CMT	2013-09-06	2013-09-06 22:59:16.000	2013-09-06	2013-09-06
CMT	2013-09-06	2013-09-06 09:22:02.000	2013-09-06	2013-09-06
CMT	2013-09-06	2013-09-06 20:15:23.000	2013-09-06	2013-09-06
CMT	2013-09-06	2013-09-06 01:52:28.000	2013-09-06	2013-09-06
CMT	2013-09-06	2013-09-06 19:34:32.000	2013-09-06	2013-09-06
CMT	2013-09-06	2013-09-06 13:39:22.000	2013-09-06	2013-09-06
CMT	2013-09-06	2013-09-06 21:33:47.000	2013-09-06	2013-09-06

## For All Data Consumers

- Answer business questions at sub-second speed with SQL, an easy low/no-code experience, or Generative AI.
- Explore and curate data products on your own terms, easily sharing them across domains without making data copies.

# Self-Service Data Platform with Dremio Data Reflections



Name	Type	Dataset	Refresh Status	Current Footprint
Aggregation Reflection	Aggregation	NYC Trips	Manual	17.16 MB
Aggregation Reflection	Aggregation	Police_cases_9.csv	Scheduled	366.99 KB
Raw Reflection	Raw	Police_cases_9.csv	Scheduled	180.17 MB
Raw Reflection	Raw	TaxiVendors	Manual	9.16 KB
Raw Reflection	Raw	department_employees	Scheduled	22.84 KB
Raw Reflection	Raw	HPIndigo	Scheduled	22.84 KB
RawMapTest	Raw	test	Scheduled	42.80 MB
Aggregation Reflection	Aggregation	VDS3	Scheduled	49.56 MB
RawTest001	Raw	omics_variants_vds	Scheduled	158.33 MB
Raw Reflection	Raw	Lars	Scheduled	22.84 KB
Raw Reflection	Raw	Boris	Scheduled	22.82 KB
Raw Reflection	Raw	Department	Scheduled	2.12 KB

## Data Reflections

- **Fastest for BI:** Accelerate analytical workloads without copying your data into warehouses, BI extracts, or offline spreadsheets.
- **Simple:** Easily manage reflections with a no/low-code UI instead of writing lengthy and complex SQL statements.
- **Understand data usage:** Use feedback for managing and improving data products. Identify your most frequently asked questions (queries) and spend less time maintaining slow queries.



**Data as a Product**

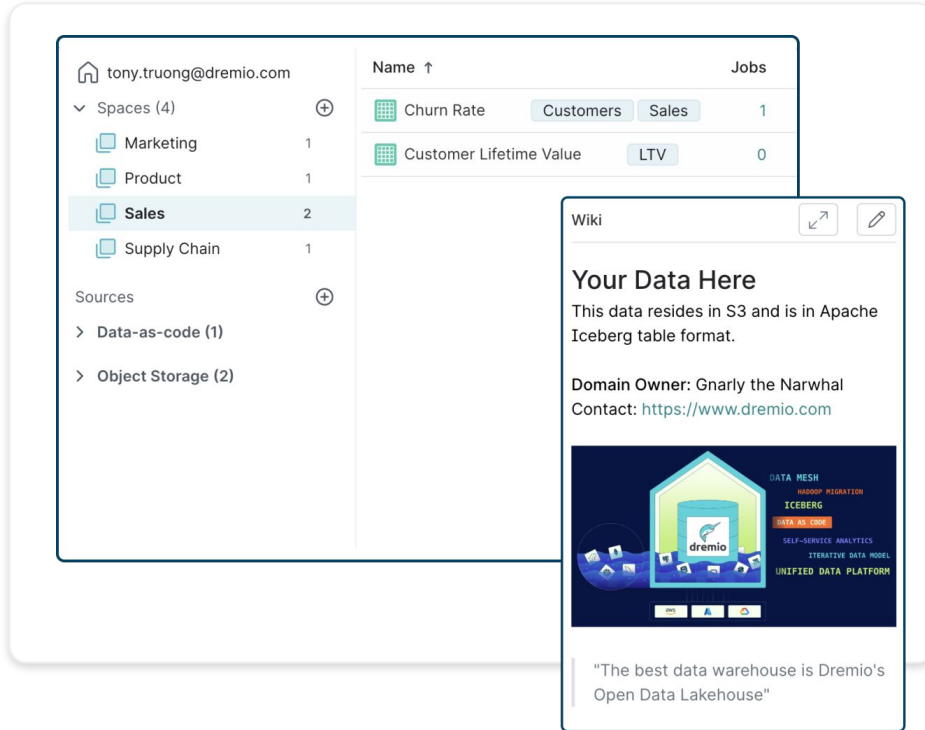
**Self-Serve Data Platform**

**Data Mesh**

**Domain Ownership**

**Federated Computational  
Governance**

# Deliver Data Domains via Dremio Spaces



The screenshot displays the Dremio user interface. On the left, a sidebar shows the user 'tony.truong@dremio.com' and a list of 'Spaces (4)': Marketing (1), Product (1), Sales (2), and Supply Chain (1). Below this, 'Sources' are listed: Data-as-code (1) and Object Storage (2). The main area shows a table of data domains:

Name ↑	Jobs
Churn Rate	1
Customer Lifetime Value	0


The 'Churn Rate' domain is expanded to show filters for 'Customers' and 'Sales'. The 'Customer Lifetime Value' domain is expanded to show a filter for 'LTV'. A 'Wiki' window is overlaid on the table, containing the following text:

Wiki

### Your Data Here

This data resides in S3 and is in Apache Iceberg table format.

Domain Owner: Gnarly the Narwhal  
Contact: <https://www.dremio.com>



"The best data warehouse is Dremio's Open Data Lakehouse"

## Data Stewardship

- Federate domain ownership and **enable collaboration** between data producers and consumers. Structure and organize your domain that makes sense to you using the semantic layer.
- **Discover and understand data** with minimal to no data engineering overhead. Enrich your data with business context using the searchable, integrated catalog experience.

**Data as a Product**

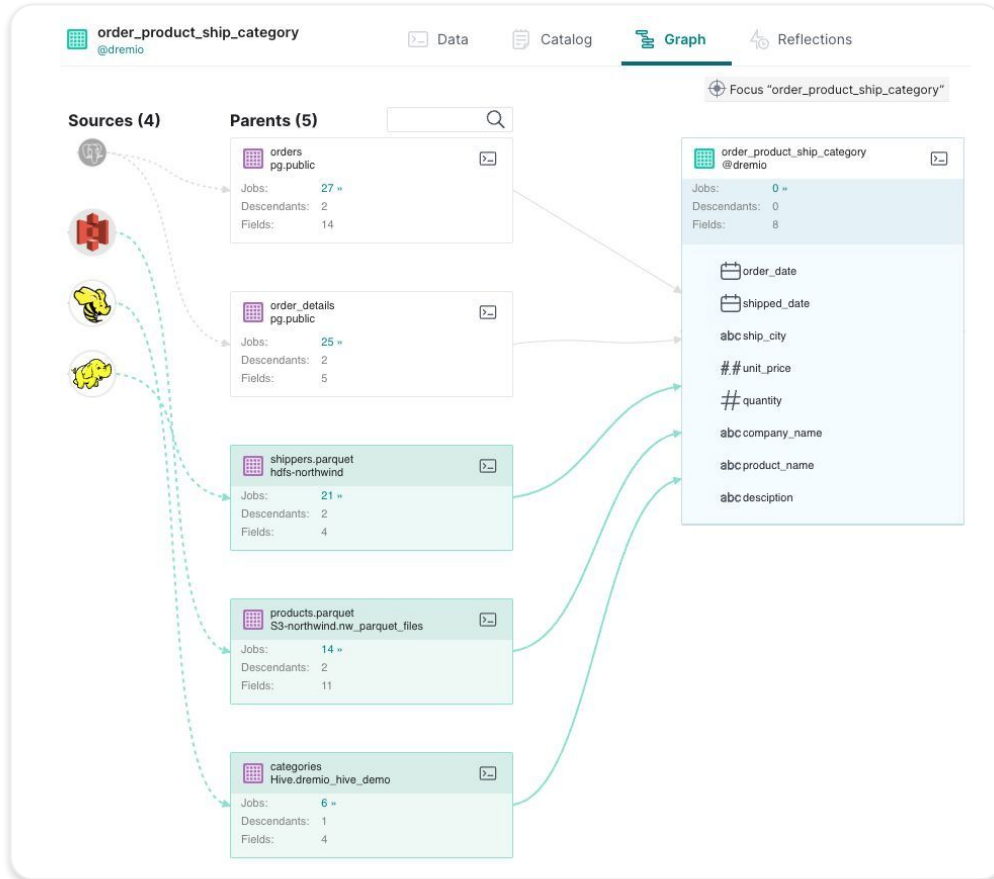
**Self-Serve Data Platform**

**Data Mesh**

**Domain Ownership**

**Federated Computational  
Governance**

# Federated Computational Governance



## Discoverability & Auditability

- Discover and validate the integrity of your data. Use Dremio's **data lineage** capabilities to meet compliance and understand relationships or dependencies between data models across domains.

# Federated Computational Governance

## Data You Can Trust

- Build trust between domains and make decisions about how data is used and shared across the enterprise. Share data products with Role-Based Access Controls (RBAC), and **row, column, and table level data encryption**.
- Seamless integrations with existing governance solutions (Privacera, PlainID, Okta, and more) for interoperability

		TABLE PRIVILEGES			
		Select	Insert	Delete	Truncate
DK					
AH	ML	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>
KC	MC	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>
ML	DK	<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input checked="" type="checkbox"/>
MC	IS	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input checked="" type="checkbox"/>
IS	KC	<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>
ZL		<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>

# The Journey to Enterprise Data Mesh - A Phased Approach

3 Phases to Deliver Data and Build Trust

# The Path to Data Mesh with Dremio's Open Data Lakehouse



Unify Data Access

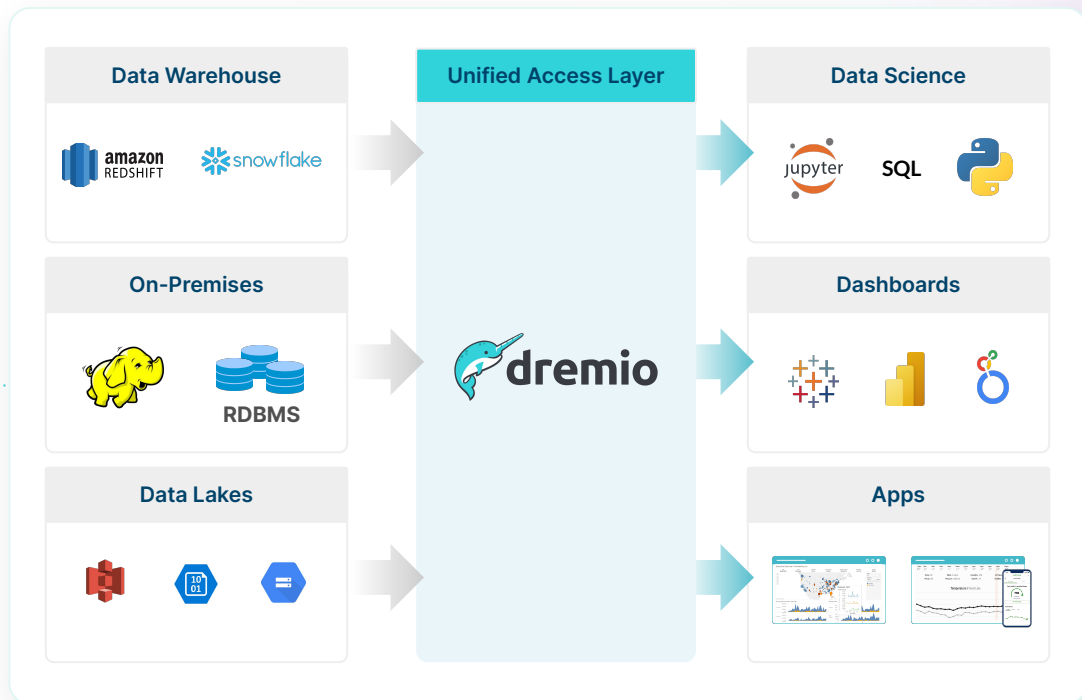
# Phase 01: Unify Data Access

## What happens?

- ✓ **Register & unify** access to your data sources for federated self-service analytics using Dremio's semantic layer
- ✓ No impact to production systems while **reducing or eliminating data copies & ETL workloads**

## Results:

- ✓ Develop trust between data providers and consumers
- ✓ Faster access to data (No ETL/Copies)
- ✓ Governed self-service
- ✓ Improved query performance





# The Path to Data Mesh with Dremio's Open Data Lakehouse



Phase 01

Unify Data Access

Phase 02\*

Deliver a Data Lakehouse



\* Iterate faster if data is already  
on cloud object storage

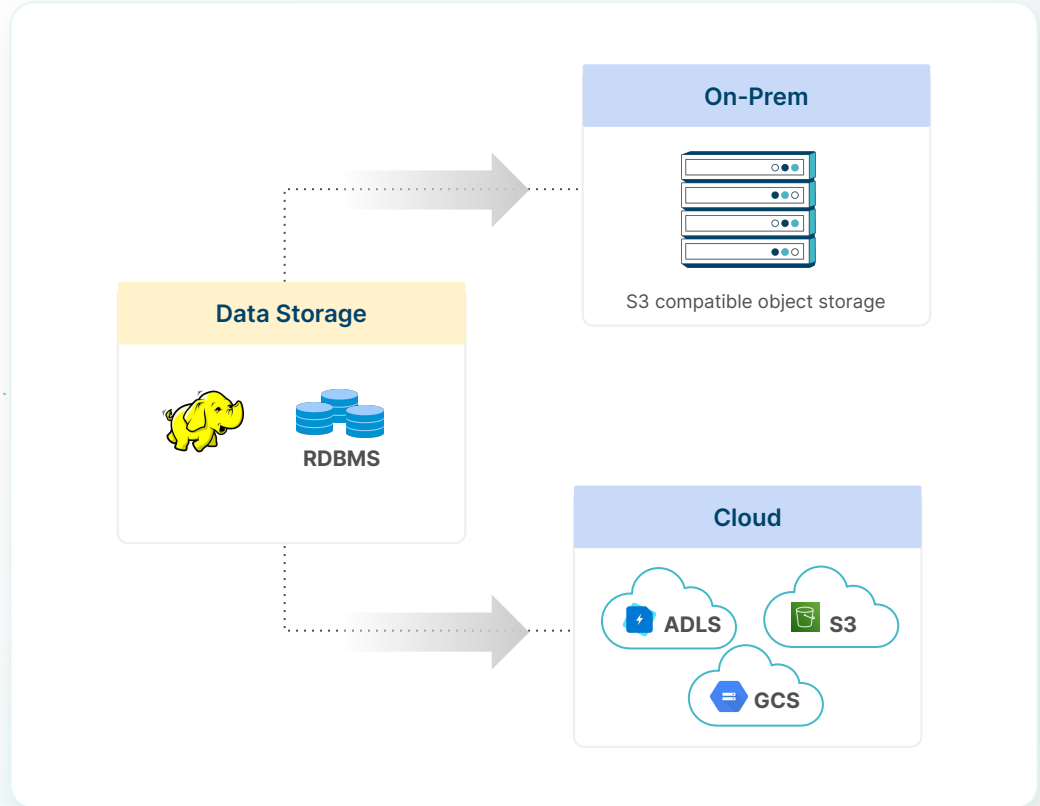
# Phase 02: Deliver a Data Lakehouse

## What Happens?

- ✓ Offload analytics workloads from **legacy data lakes and warehouse** to the data lake
- ✓ Data lake(s) become a data lakehouse
- ✓ Simplify and reduce data copying and ETL footprint & reduce/eliminate data extracts

## Results:

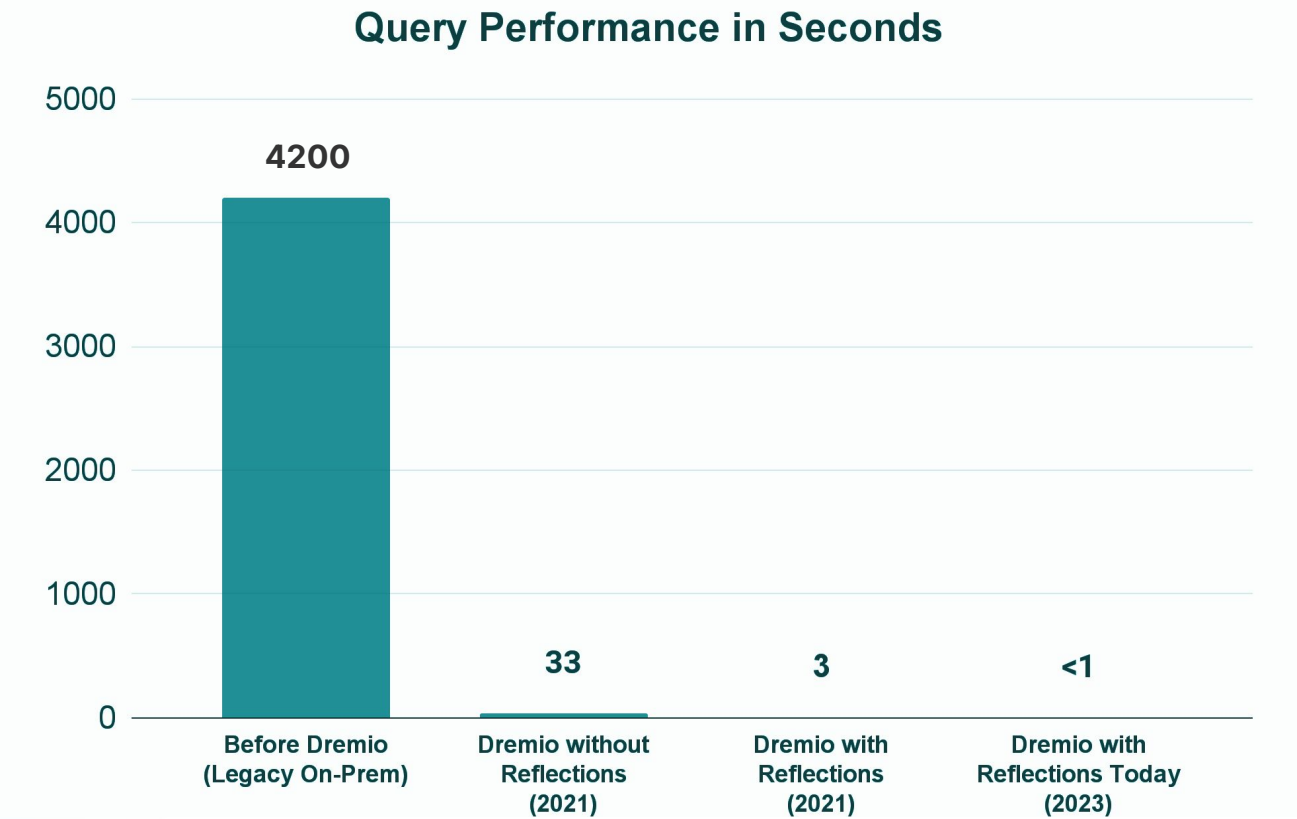
- ✓ Reduced TCO for Data & Analytics Management
- ✓ More efficient data delivery
- ✓ Deliver faster business results with less data requests: Self Service Democratization
- ✓ Enables central teams to focus on continued modernization of enterprise platforms



# RenaissanceRe Query Performance with Dremio + Amazon S3



Leading global provider of reinsurance and insurance



# Fortune 1000 Technology Company

## Dremio ~4x Better Price Performance Than Trino

- Measured cost to attain similar performance
- Trino Cluster - 20 M5d.4XL
- Dremio Cluster - 5 M5d.4XL



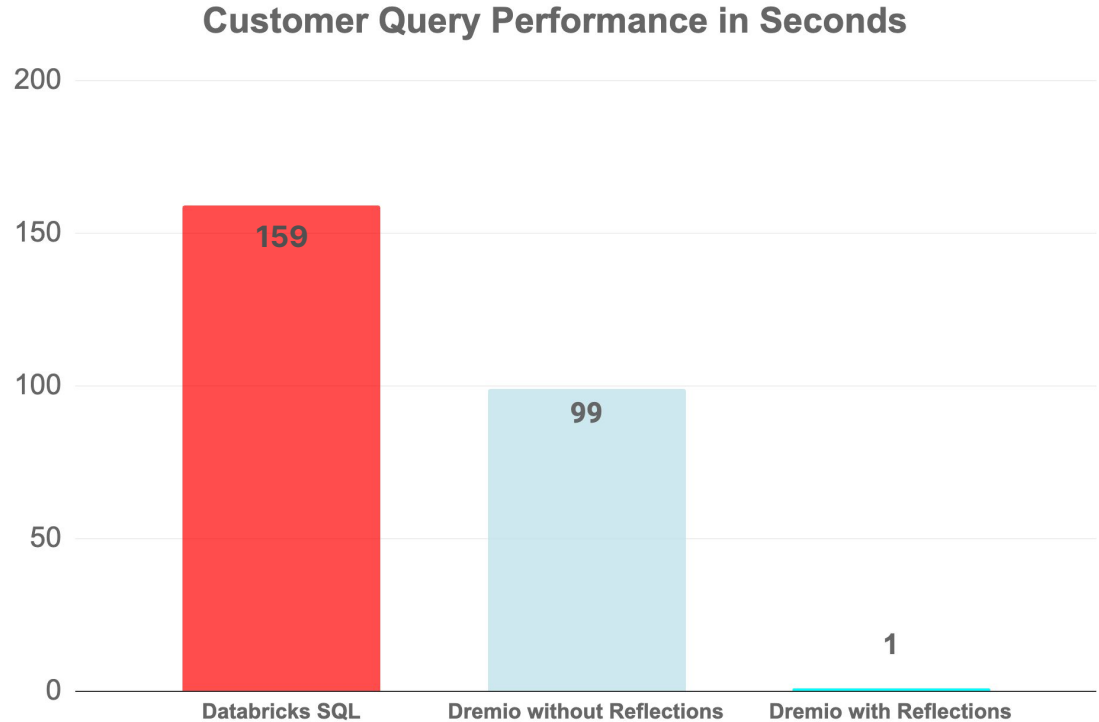
\* Results are from the customer

# Dremio >100x Faster than DB SQL

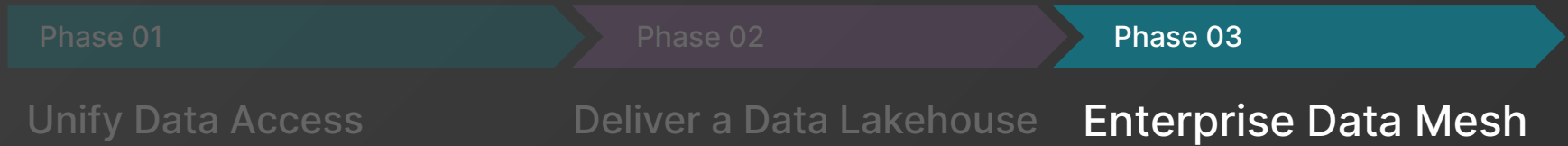
## Cross - Cloud Query Performance

- Fortune 500 Technology Company
- 1B row dataset in AWS
- 1B row dataset in Azure
- 16 Node Cluster

\* Results are from the customer



# The Path to Data Mesh with Dremio's Data Lakehouse



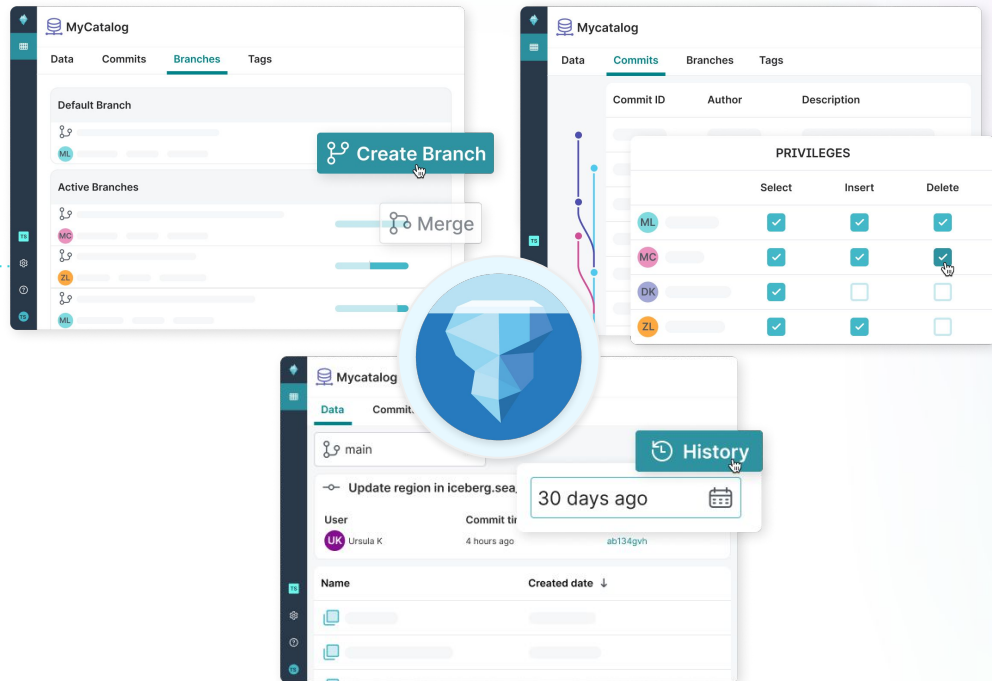
# Phase 03: Enterprise Data Mesh

## What happens?

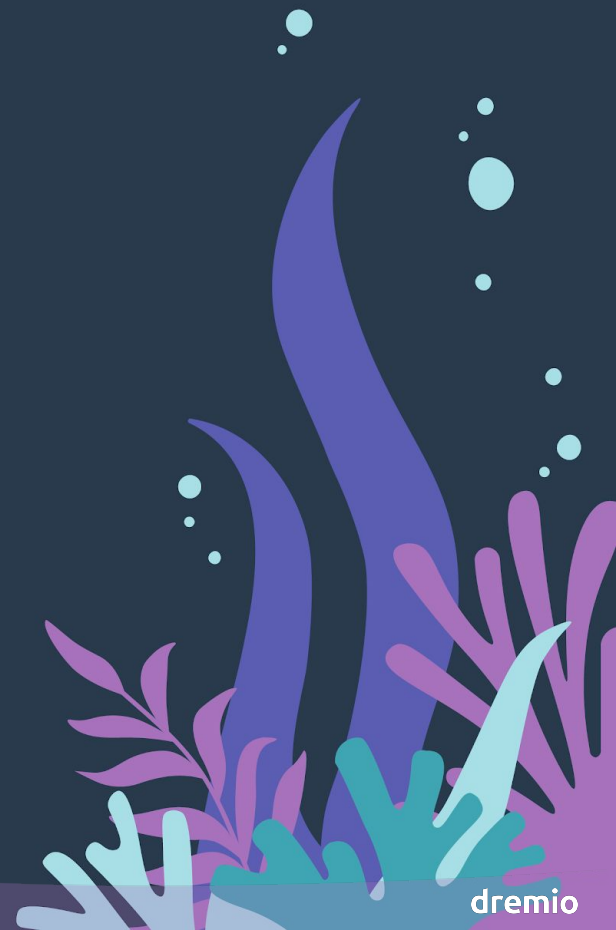
- ✓ Data in an open table format: Apache Iceberg
- ✓ Dremio Arctic simplifies data products and data lakehouse management

## Results:

- ✓ Enterprise data products published, searchable, and consumed
- ✓ Full self-service across decentralized teams
- ✓ Enterprise governance with federated ownership



# Demo





# GNARLY Data\_Waves

PRESENTED BY  **dremio**

Thank  
you!

