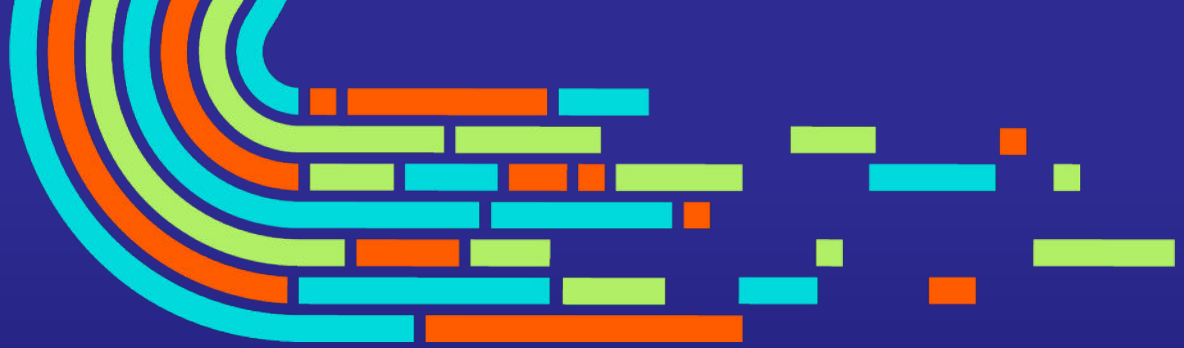


GNARLY Data_Waves

PRESENTED BY dremio

EPISODE 07

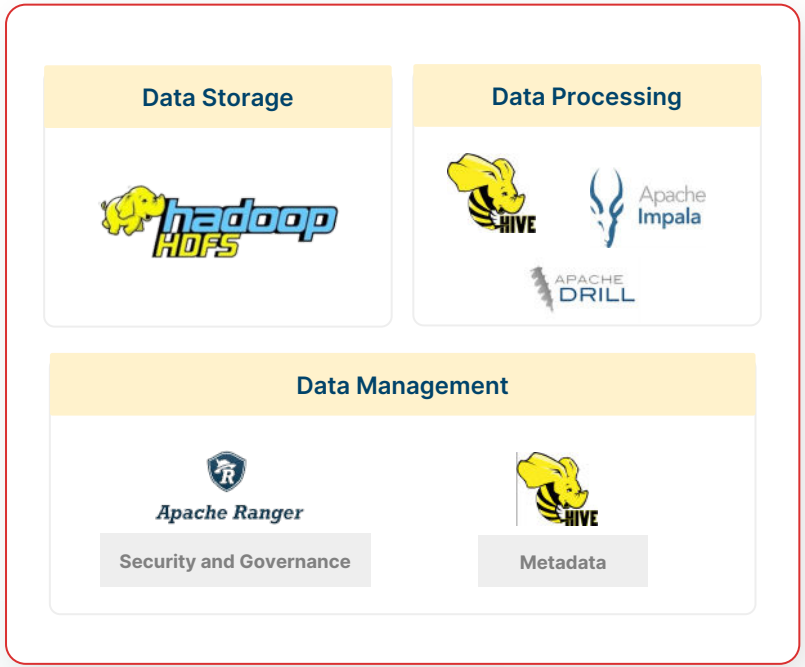
# Getting Started with Hadoop Migration and Modernization

# Today’s Agenda

## Getting Started with Hadoop Migration and Modernization

- Challenges With Hadoop
- Options For Migrating
- The Path to Hadoop Migration and Modernization
- Demo!

# Why organizations get off Hadoop




## Challenges

- Requires deep expertise in the Hadoop ecosystem to maintain
- High cost of scalability as your data grows
- Query performance management
- Difficult to enable governed self-service analytics

# About Dremio

## The Easy and Open Data Lakehouse

Self-service analytics with data warehouse functionality and data lake flexibility across all of your data.

**The Only Data Lakehouse with Self-Service SQL Analytics**

**Your Data Forever, No Lock In**

**Sub-Second Performance, 1/10th the Cost of Data Warehouses**

## Open Source & Community

Apache Arrow (**60M+ downloads/m**), Apache Iceberg, Nessie

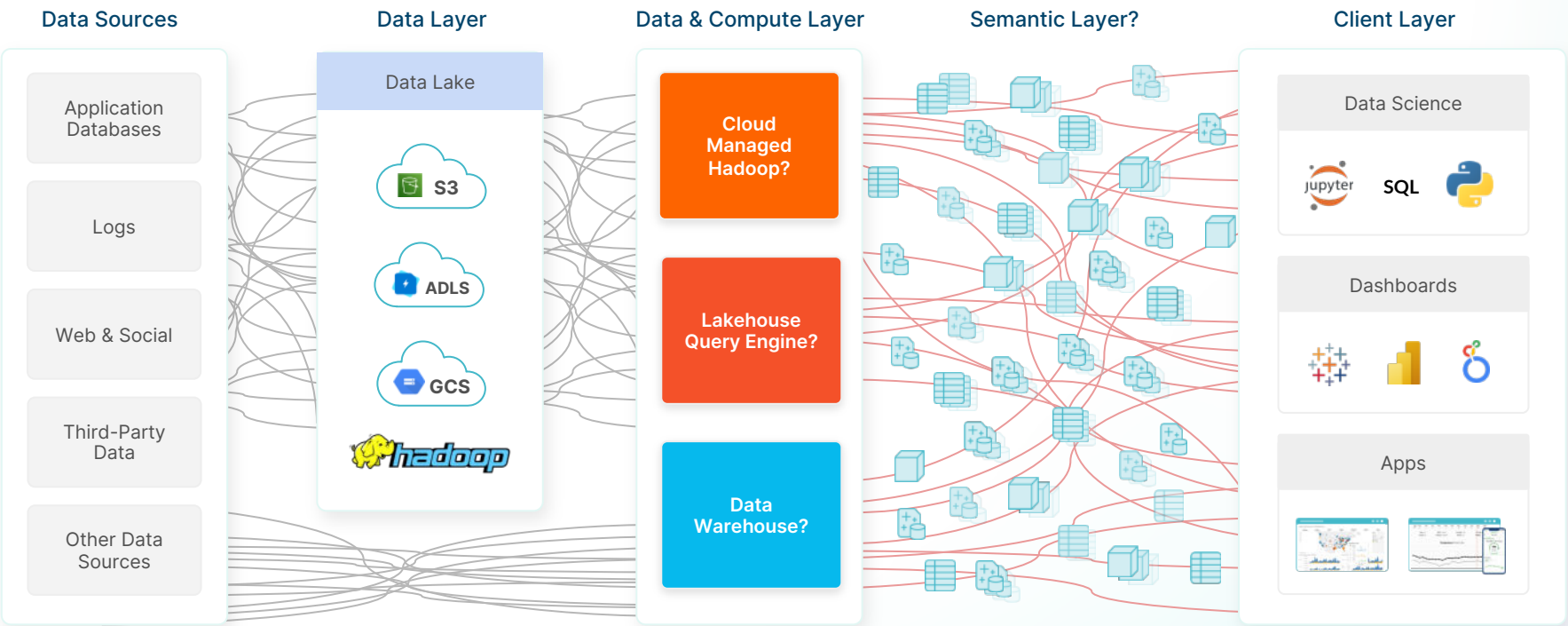
Creator and host of **Subsurface LIVE** conference

## Enterprise Adoption

**1000s**
of companies across all industries

**5**
of the Fortune 10

# What are your options?



# The Path to Hadoop Migration and Modernization

Stage 1

Modernize Hadoop Query Engine & Provide Self-Service Analytics

# The Path to Hadoop Migration and Modernization

Stage 1

Modernize Hadoop Query Engine & Provide Self-Service Analytics

Stage 2

Migrate off HDFS

# The Path to Hadoop Migration and Modernization

| Stage 1 | Stage 2 | Stage 3 |
| --- | --- | --- |
| Modernize Hadoop Query Engine & Provide Self-Service Analytics | Migrate off HDFS | Create Your Open Data Lakehouse |

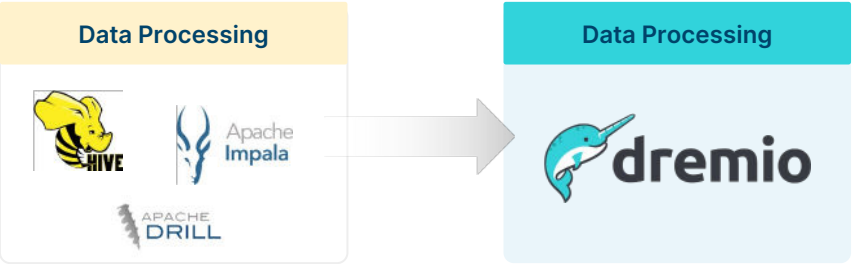
# **Step 1a:** Modernize Hadoop Query Engine

## What happens?

- Connect Dremio to existing Hadoop clusters and simplify the transition to modern cloud object storage.
- **Minimize impact** to production system.

## Results:

- Immediately improve query performance over Hive, Drill, and Impala



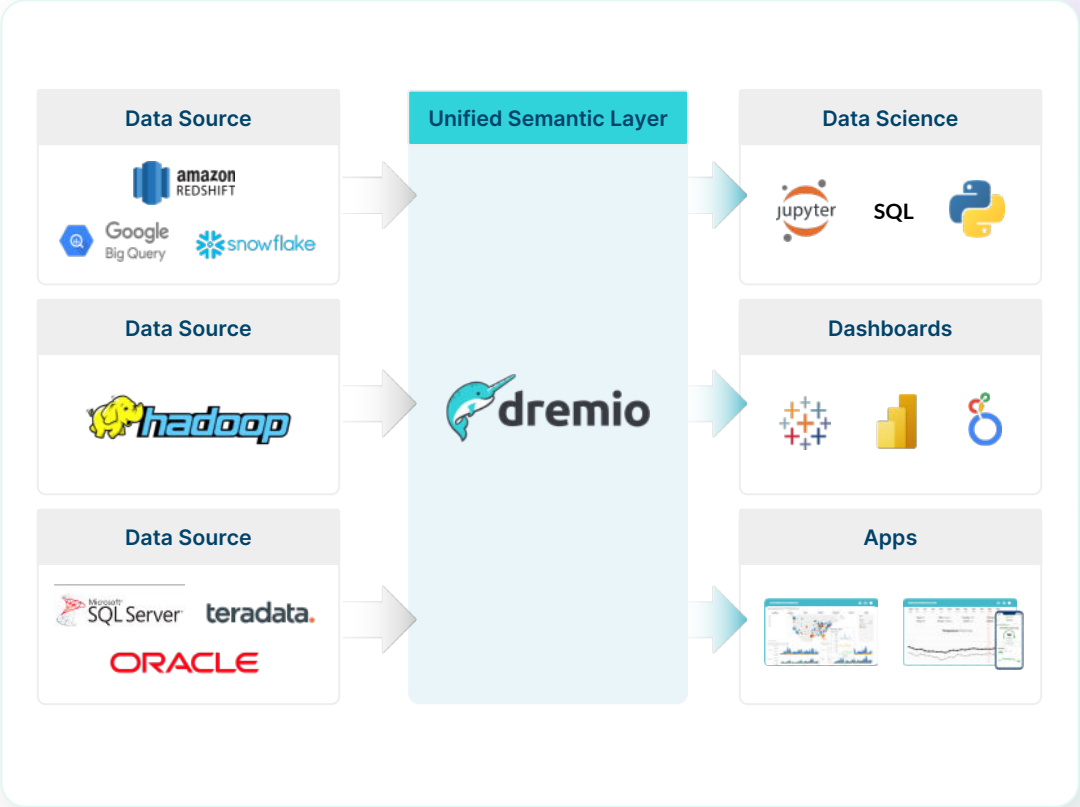
# **Step 1b:** Provide Self-Service Analytics

## What happens?

- Unify all your data for self-service analytics using Dremio's semantic layer
- Connect and federate queries across other data sources
- Minimize impact to production system and **reduce complex ETL footprint**

## Results:

- With Dremio's semantic layer, these sources are given business-friendly names, helping to deliver reliable data products across all your downstream applications.



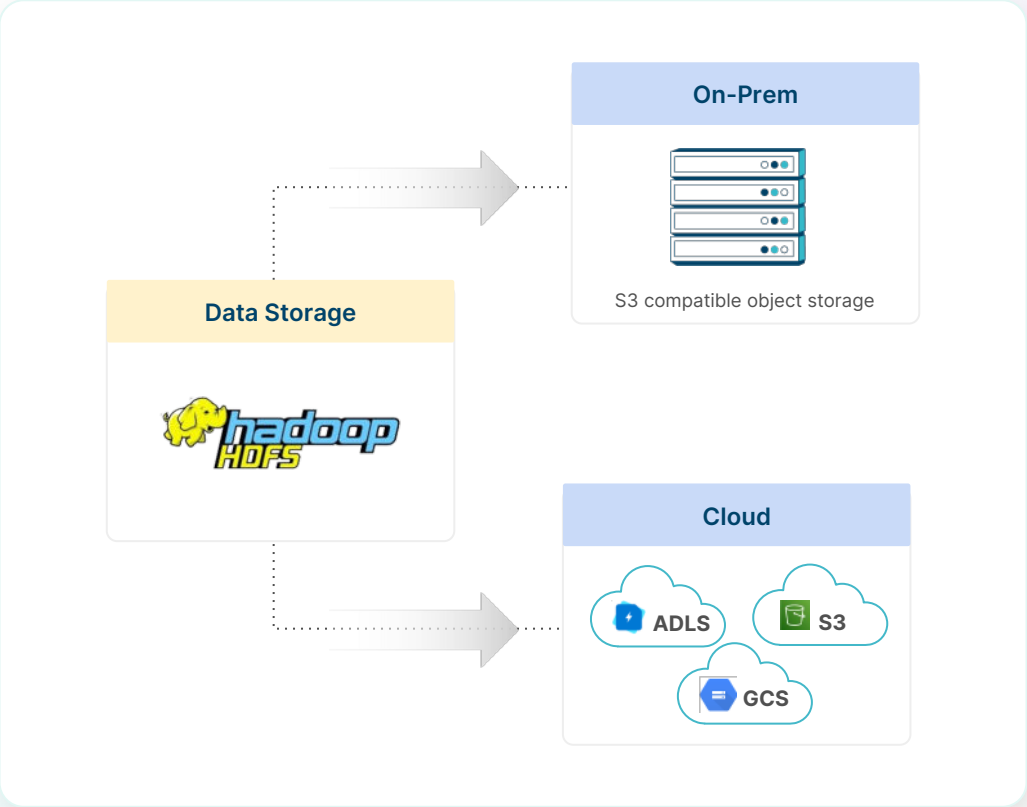
# **Step 2:** Migrate off HDFS to Object Storage

## What happens?

- Start migrating off HDFS to object storage
  - Data that needs to be on-prem can migrate to S3 compatible object storage
  - Everything else can go to cloud object storage

## Results:

- Eliminate Hadoop costs (Cloudera license and underutilized servers)
- Minimize impact to business continuity on HDFS with this phased approach
- Scalability



# **Step 3:** Create your open data lakehouse

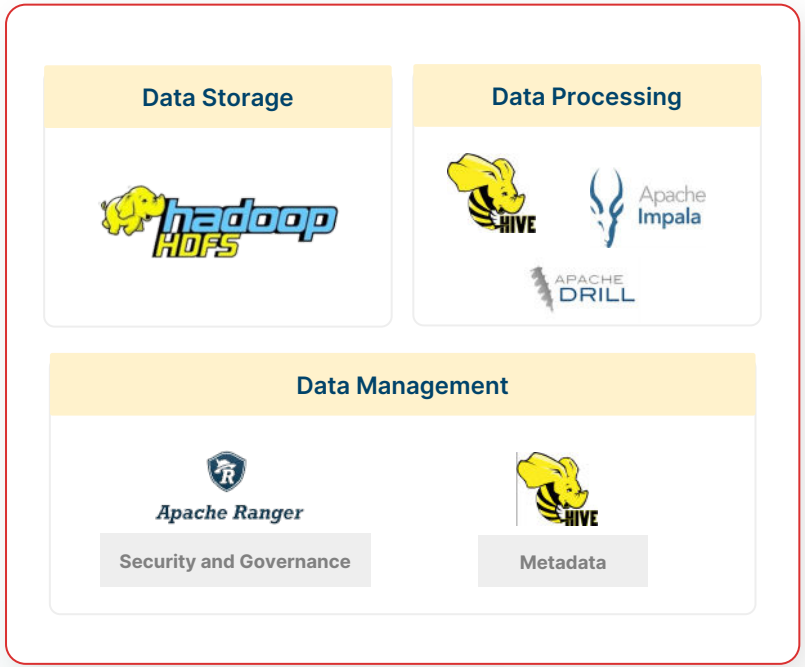
## What happens?

- Data is in cloud object storage
- Migrate Hive tables to open table format like Apache Iceberg

## Results:

- Future proof your data architecture
- Achieve higher performance, DML, schema evolution, time-travel, and other data warehouse functionality
- Avoid vendor lock-in from proprietary table formats
- Make data accessible to your query engine(s)

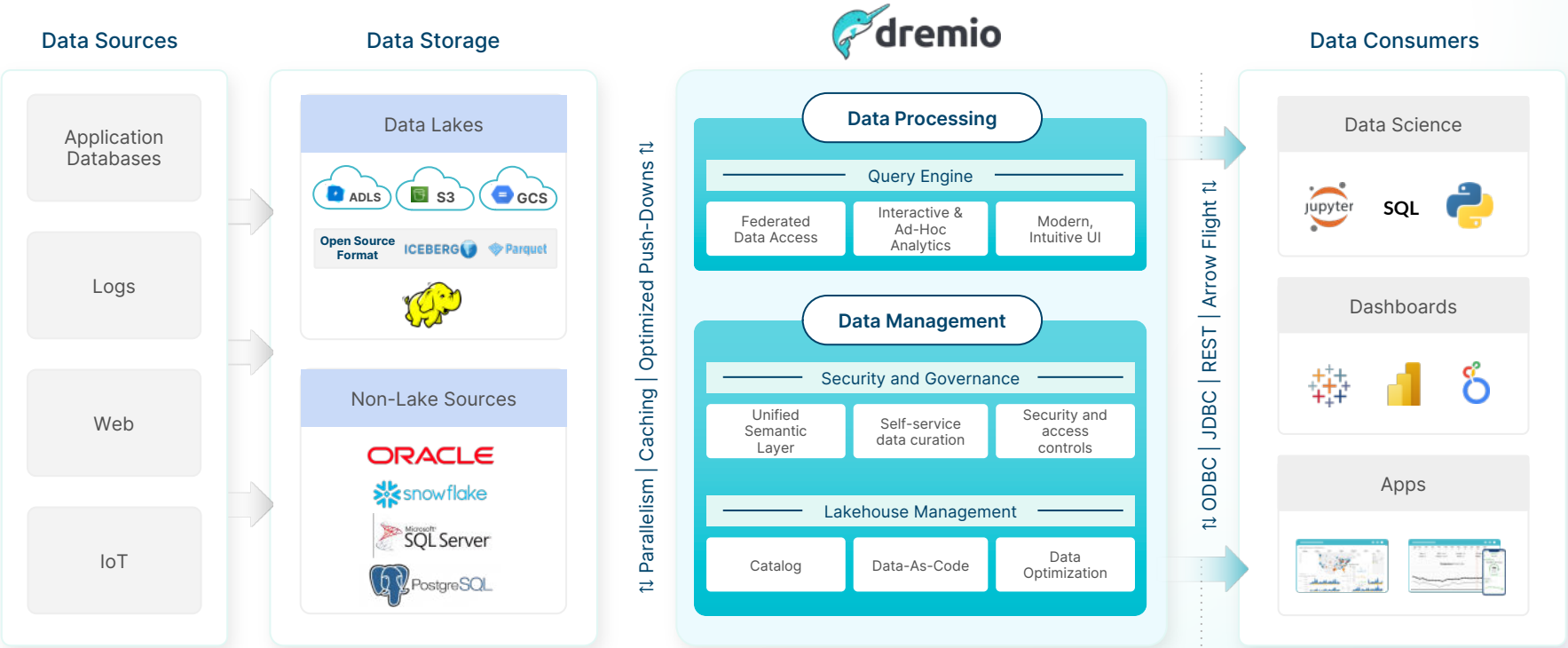
# Remember this?



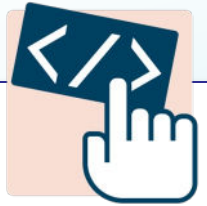

## Challenges

- Requires deep expertise in the Hadoop ecosystem to maintain
- High cost of scalability as your data grows
- Query performance management
- Difficult to enable governed self-service analytics

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



# The Dremio Advantage

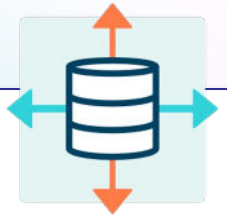


## Self–Service Analytics

Modern and Intuitive User Interface

Unified View of Data
*(on-prem, hybrid and Cloud)*

Federated Queries



## Open Data, No Lock-In

Based on community-driven standards, including:

- Apache Parquet
- Apache Iceberg
- Apache Arrow

## Sub-Second Performance at 1/10th the Cost

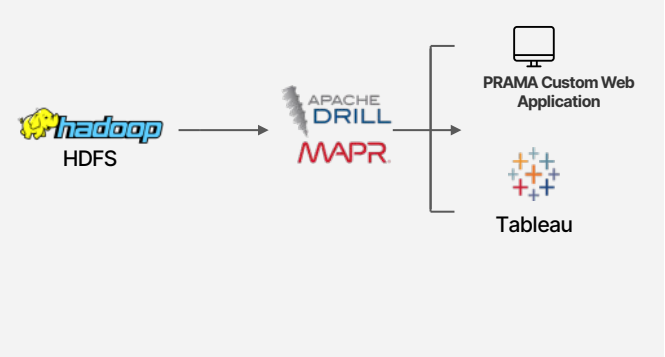
Lightning-fast queries

High concurrency

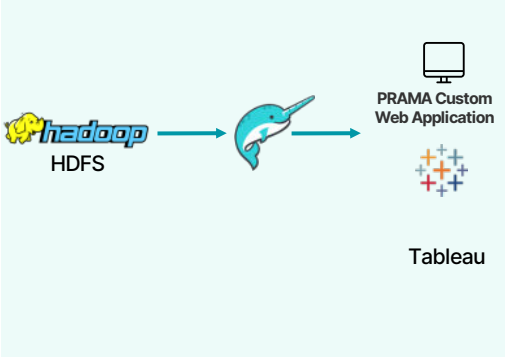
No expensive data copies to manage

# Empower Analysts and Customers to Do Self-Service Analytics

**Before Dremio**




**After Dremio**




## Business Problem

- Need to manage **30 petabytes of data** from 90,000 data sources and more than three billion updates per month from their data providers.
- Customers expect a fast response for queries, and **slow performance of SQL on Hadoop** resulted in a poor experience for analysts and end customers.
- Employees felt that IT was slowing things down.
- Forced to have **entire agile teams** devoted to maintenance.

## Why Dremio?

- Data reflections and virtual datasets provide acceleration and a handoff between IT and the end users.
- Can easily join customer datasets into the solution and it scales easily with their Hadoop cluster.

## Results

**Self-Service Access**

- Empowers analysts with **self-service ability** to explore data without having to wait for data engineers.
- Gives analysts and customers **individualized interactive dashboards.**

**Reduced Data Engineering Workload**

- Provides **5-10x immediate performance gain**, before implementing reflections.
- Reduces overhead and increases product development agility, able to redeploy 14 data engineers from maintenance to building new products.

# Dashboards Running Up To 30x Faster

The company began in Ohio as "National Manufacturing Company" in 1879 to manufacture and sell the first mechanical cash register. Today NCR has annual revenues >$6B and is at the cutting edge of hardware and software business solutions for banking, restaurants, grocery stores, airlines and modern stadiums and arenas.

*"Dremio bridges the data warehouse and the data lake, enabling NCR to derive more value between the two data sources. Most importantly, to deliver faster data insights to our internal and external customers"*

Ivan Alvarez
IT vice president, big data and analytics
NCR Corporation

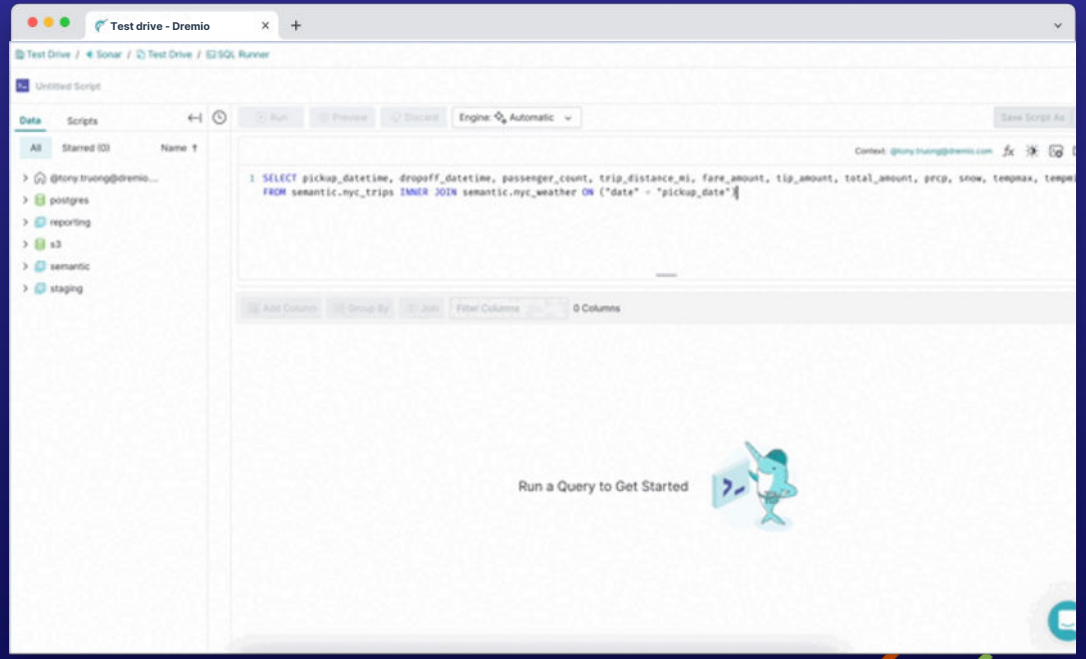
| Business Problem | Why Dremio? | Results |
|---|---|---|
| ▪ Support the business's ability to cross-sell, up-sell, and service their customer base<br>▪ Moving data pipelines took 2-3 months for critical and large datasets<br>▪ Slow analytics development due to functional silos created among experts in different data repositories<br>▪ Long turnaround time for data requests | ▪ Self-service data analytics<br>▪ Modernize data infrastructure on data lake<br>▪ Cost-effective solution that replaces expensive on-prem DW<br>▪ Immediate performance gains on Hadoop | **Cost reduction**<br>▪ Reduced cost & dependency on external data engineering consultants<br>▪ Retire EDW in 2 years<br><br>**Faster time-to-insight**<br>▪ Minimize "revenue leakage" by not having to wait to run analyses |

# Demo

# Experience the data lakehouse with Dremio Test Drive

- ✓ Sub-second query on 1 million rows of data joining Amazon S3 with a Postgres database
- ✓ Connect to Tableau or Power BI and build a dashboard with this dataset
- ✓ Everything hosted by Dremio - 100% free for you

**Start Test Drive**




GNARLY Data_Waves
PRESENTED BY dremio