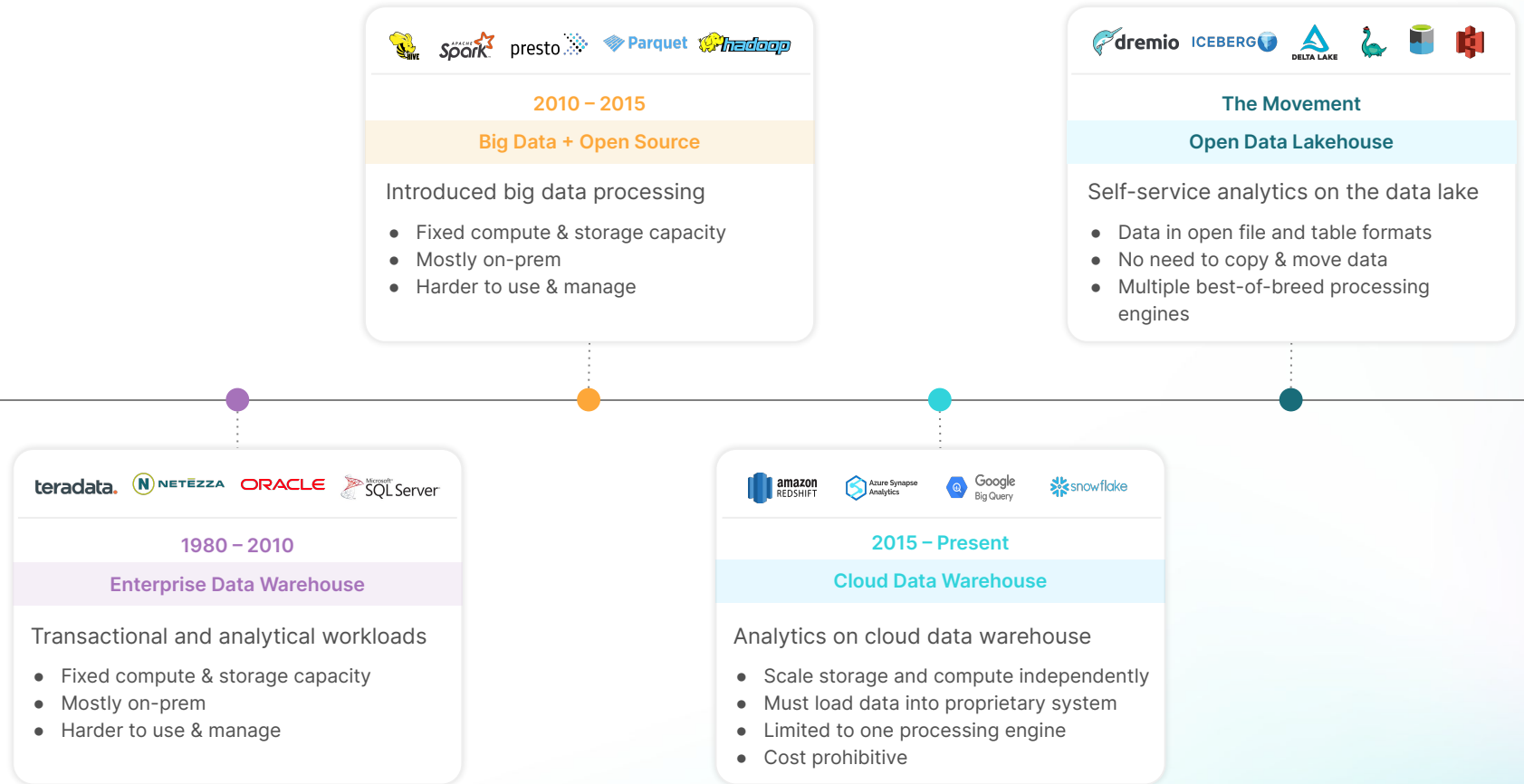




# Dremio

The Easy and Open Data Lakehouse

# Data Analytics - A History



# Intensity of Competing Data Priorities is Increasing

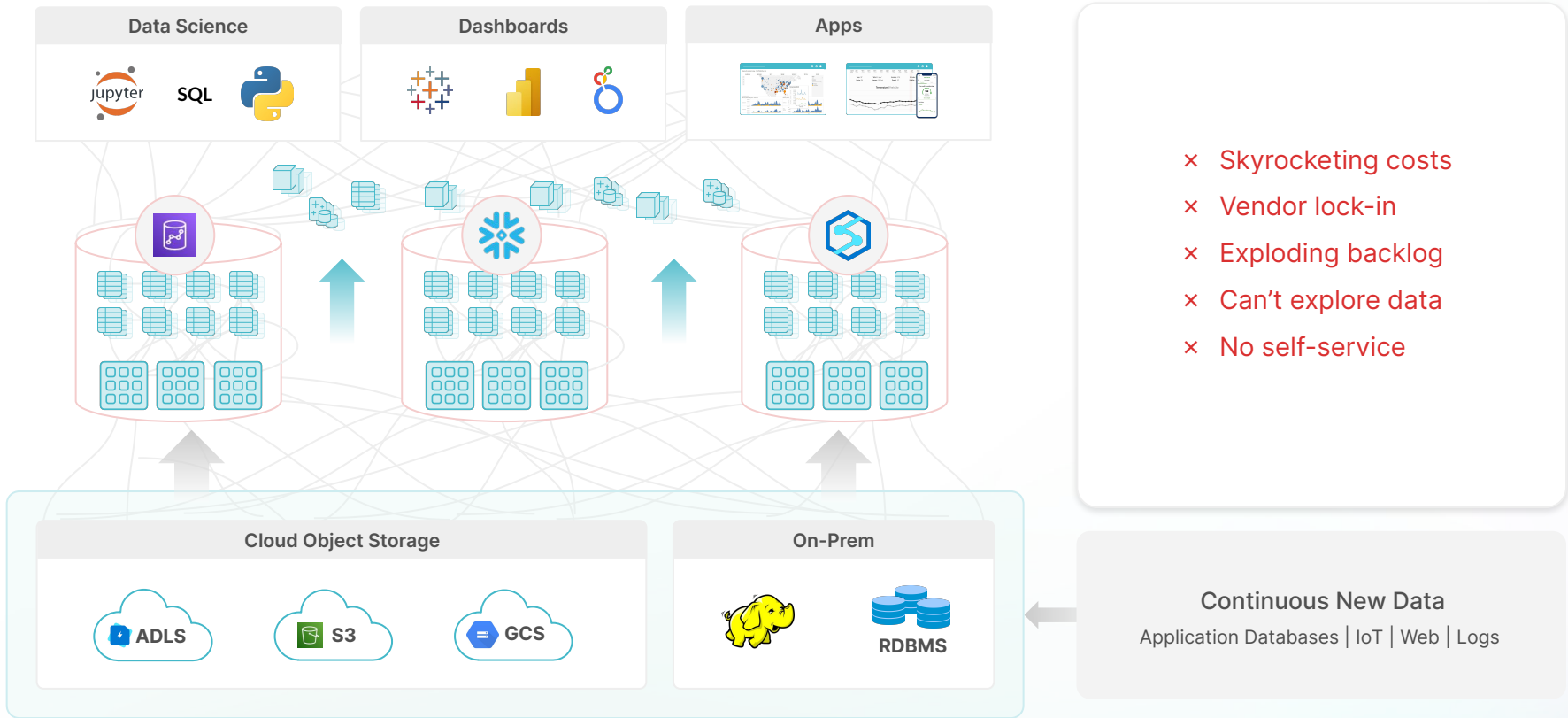
## Line of Business

- Access
- Speed + Agility

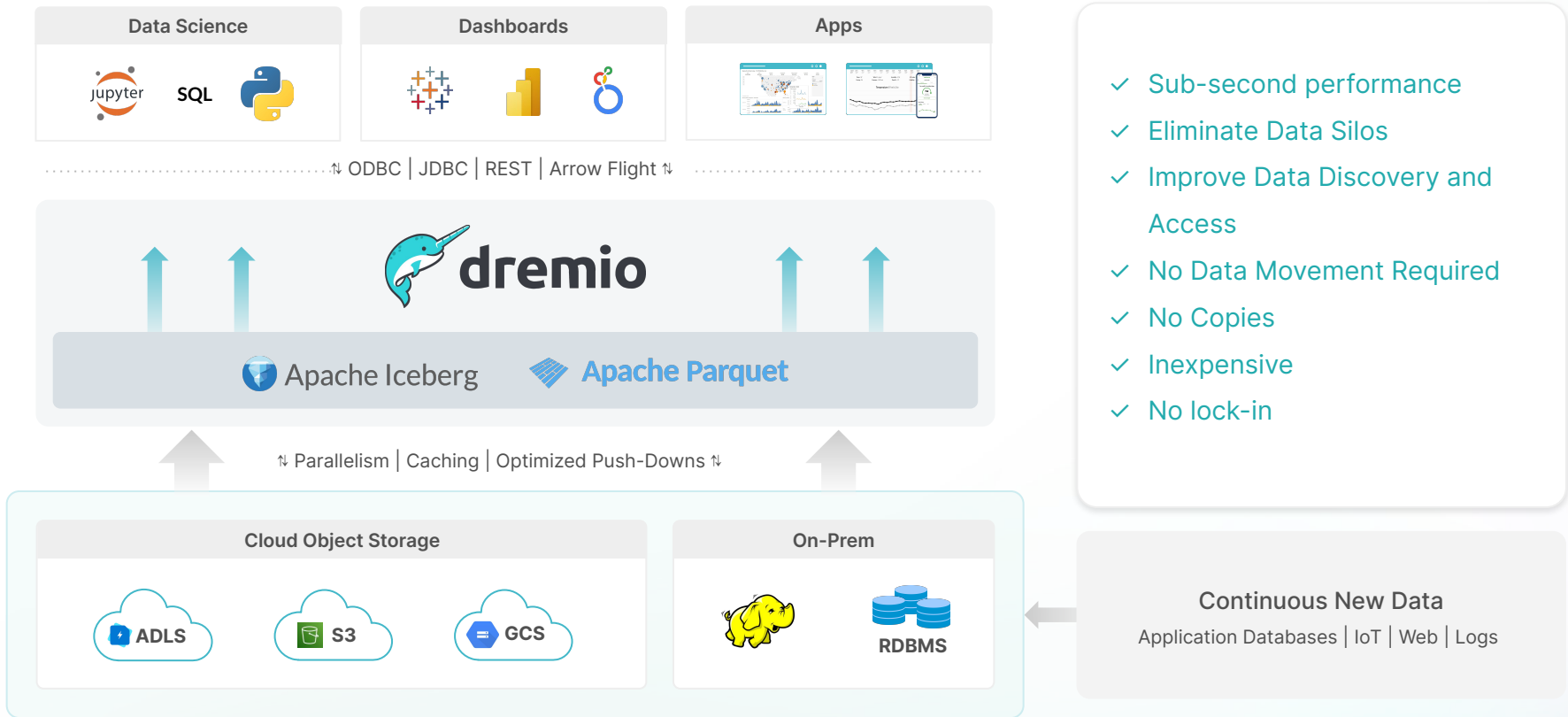
## Centralized Teams

- Governance
- Security

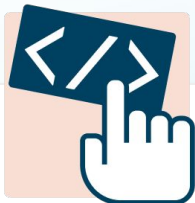
# Data Warehouses: Expensive, Proprietary, Complex



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



# The Dremio Advantage

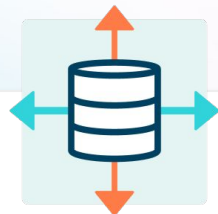


## Self-Service Analytics

Modern and Intuitive User Interface

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

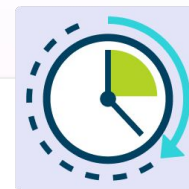
Federated Queries



## Open Data, No Lock-In

Based on community-driven standards, including:

- Apache Parquet
- Apache Iceberg
- Apache Arrow



## The Fastest BI Performance at 1/10th the Cost

Lightning-fast queries

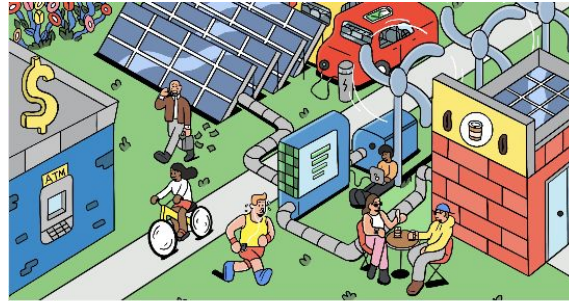
High concurrency

No expensive data copies to manage

# Recent Awards



# Forbes



America's Best Startup Employers 2023  
forbes.com



G2 Spring Report: Big Data Analytics



# 500<sup>TM</sup>

## Technology Fast 500

2022 NORTH AMERICA

**Deloitte.**

# Dremio's Data Lakehouse





# Dremio Sonar

# Dremio Sonar - Lakehouse Engine

BI tools, data science notebooks, SQL editors

↕ ODBC | JDBC | REST | Arrow Flight ↕



## Intuitive U/I

No-code UI for self-service data curation/sharing, SQL Runner

## Semantic Layer

Unified view of data with view/table hierarchy, built-in data catalog and lineage

## Query Engine

- Apache Arrow-based, vectorized execution
- Query acceleration - Data Reflections, Federated queries
- Auto-scaling/elastic engines, Multi-engine architecture, workload management, query routing

## Governance & Security

RBAC, fine-grained access control, authentication, auditing and query history

↕ Iceberg, Delta Lake | Parquet, ORC, JSON, CSV Readers | ARP Connectors ↕

## Cloud Object Storage



ADLS



S3



GCS

## On-Prem

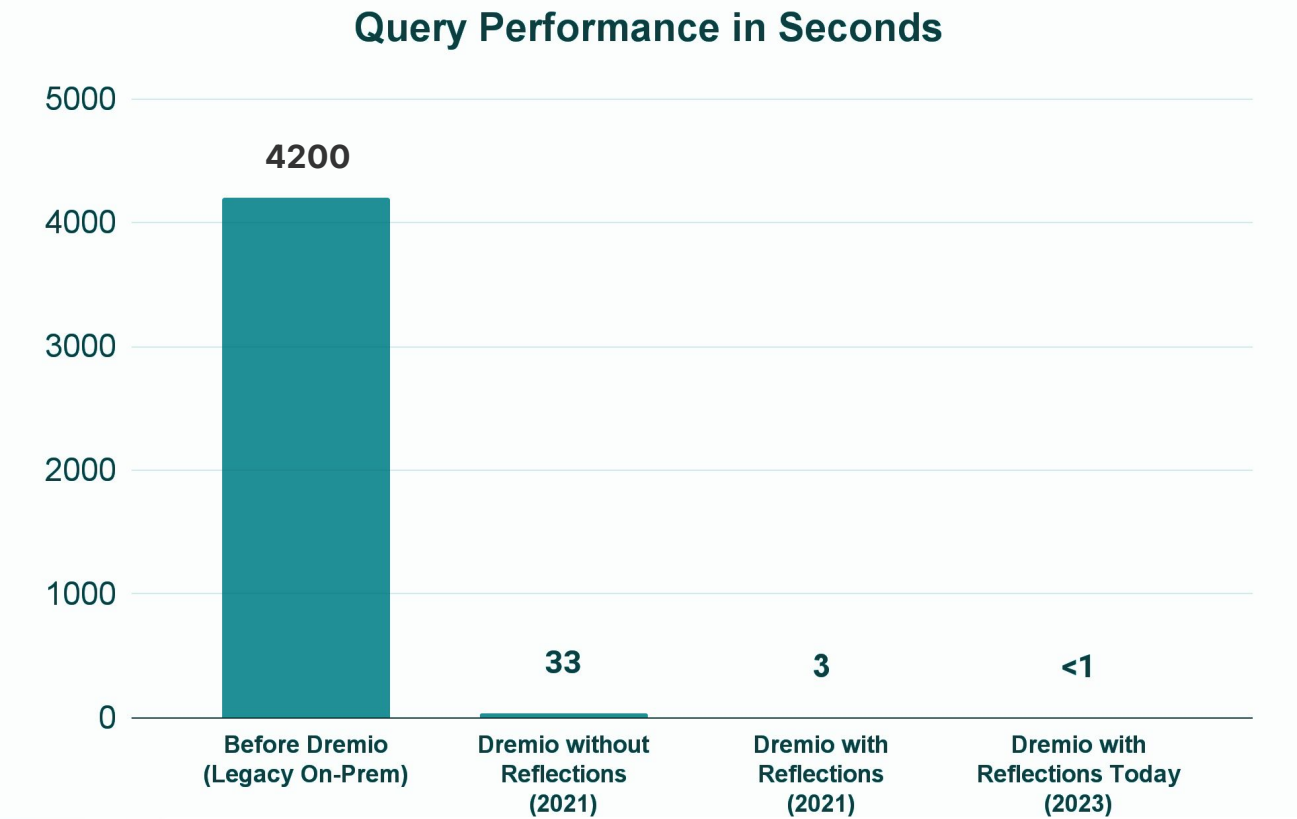


RDBMS

# RenaissanceRe Query Performance with Dremio + Amazon S3



Leading global provider of reinsurance and insurance



# 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"*

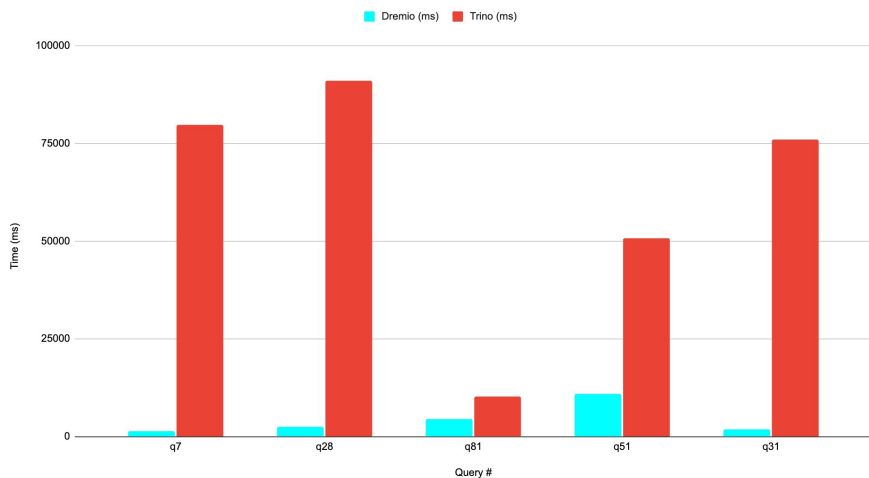
Ivan Alvarez  
IT vice president, big data and analytics  
NCR Corporation

Business Problem	Why Dremio?	Results
<ul style="list-style-type: none"><li>▪ Support the business's ability to cross-sell, up-sell, and service their customer base</li><li>▪ Moving data pipelines took 2-3 months for critical and large datasets</li><li>▪ Slow analytics development due to functional silos created among experts in different data repositories</li><li>▪ Long turnaround time for data requests</li></ul>	<ul style="list-style-type: none"><li>▪ Self-service data analytics</li><li>▪ Modernize data infrastructure on data lake</li><li>▪ Cost-effective solution that replaces expensive on-prem DW</li><li>▪ Immediate performance gains on Hadoop</li></ul>	<p><b>Cost reduction</b></p> <ul style="list-style-type: none"><li>▪ Reduced cost &amp; dependency on external data engineering consultants</li><li>▪ Retire EDW in 2 years</li></ul> <p><b>Faster time-to-insight</b></p> <ul style="list-style-type: none"><li>▪ Minimize "revenue leakage" by not having to wait to run analyses</li></ul>

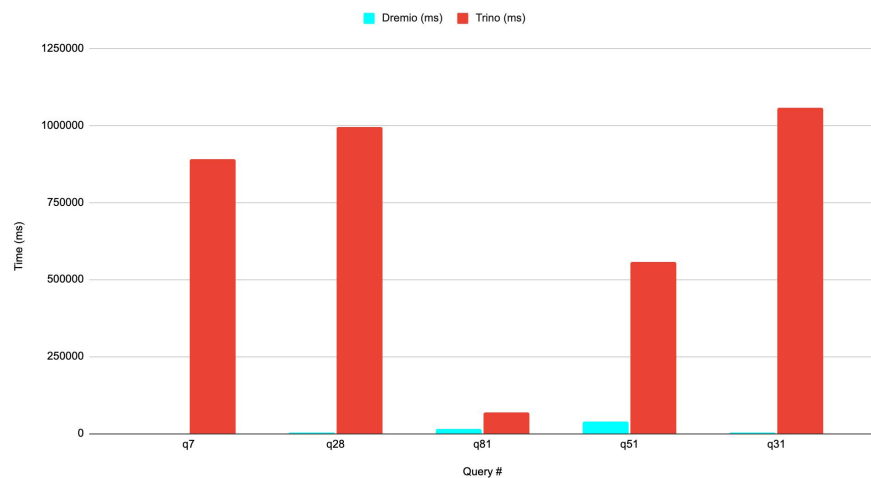
# >10x Faster Than Trino for BI Workloads (1TB)

## >100x Faster (10TB)

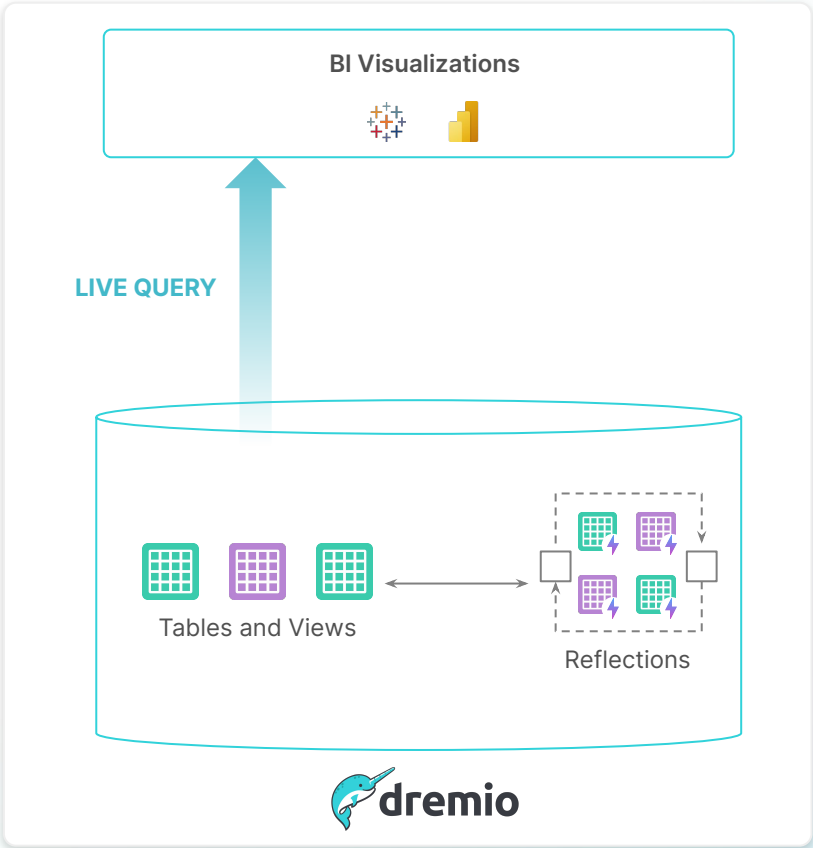
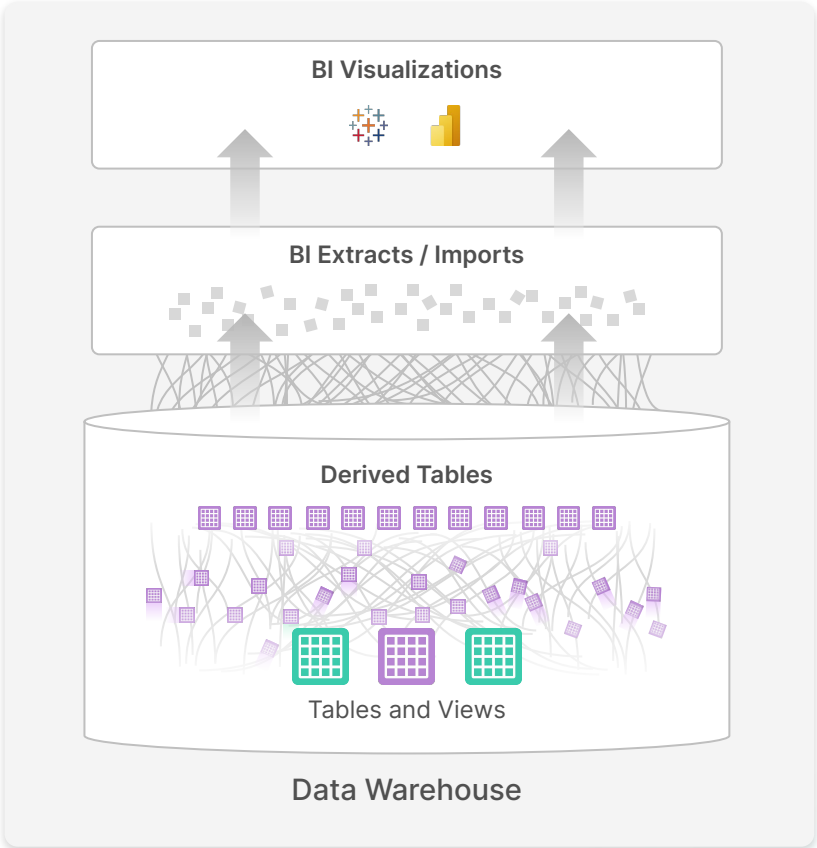
Dremio vs. Trino - SF1000 (1TB), 4 Nodes



Dremio vs. Trino - SF10000 (10TB), 4 Nodes



# Reflections Eliminate the Need for BI Extracts/Imports





# Dremio Sonar Demo



# Customer Use Cases

# Dremio Use Cases

## Data Analytics Modernization

Make your current system work better

Data Mesh

Migrate Enterprise Data  
Warehouses and DBs to  
Lakehouses

Migrate Hadoop

Modern Data  
Virtualization

Optimize Existing Cloud  
Data Lakehouse

## New Projects

Build new analytics capabilities

Departmental  
Lakehouses

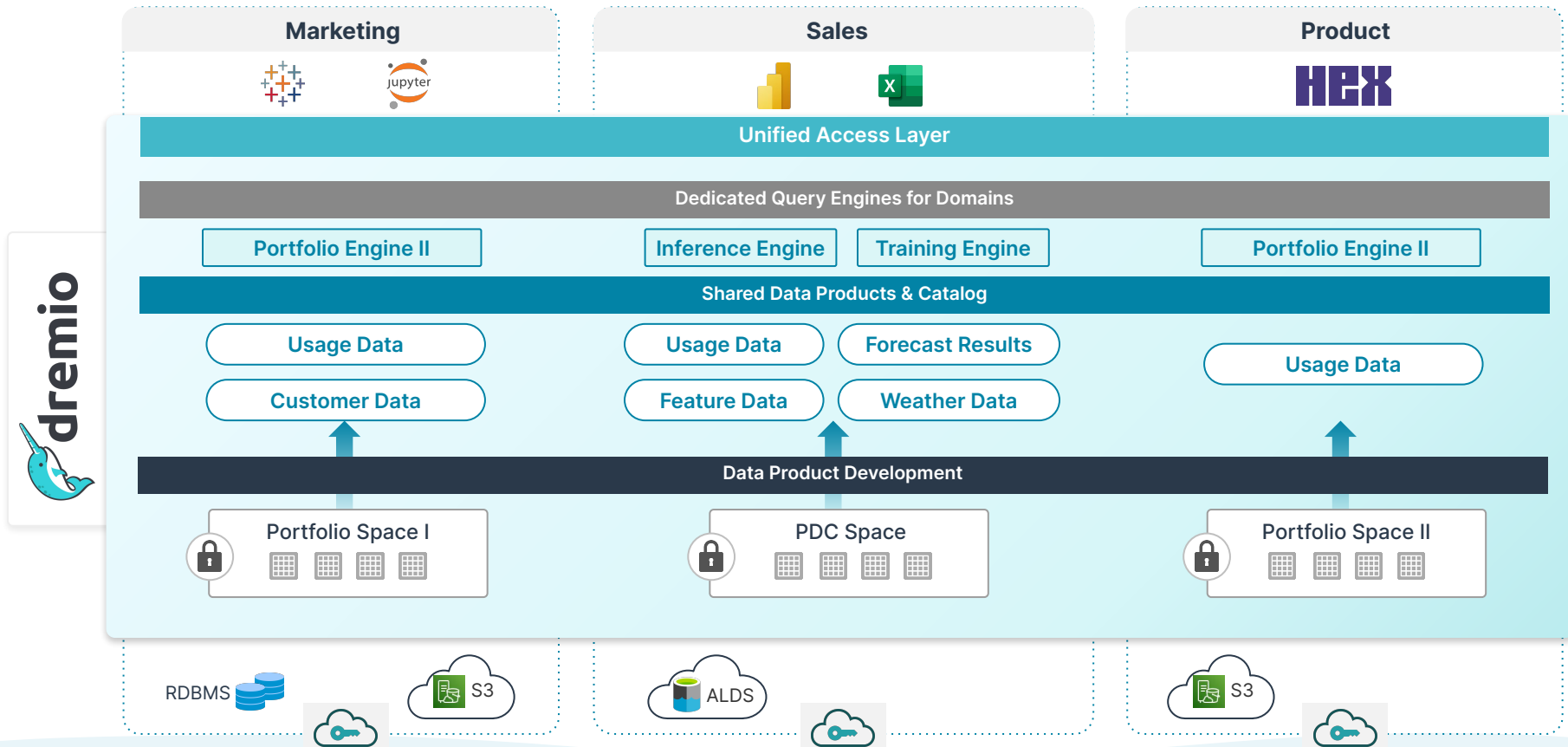
- Customer 360 / CDP
- Supply Chain
- Trading Data for Quants
- Product Analytics

Data Science

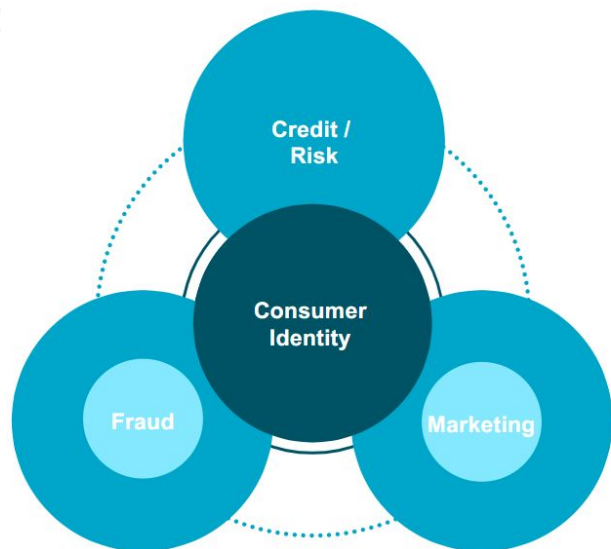
Internal Analytics  
Apps / Custom  
Dashboards

Customer-Facing  
Analytics Apps

# Dremio as the Technical Platform for Data Mesh



TransUnion's **Tru™ picture** is an actionable and robust view of each consumer built through our unique combination of data, accuracy and relevance



**Great Experiences**



**Economic Opportunity**



**Personal Empowerment**

# Dremio enables a seamless, governed self-service data and analytics interface across our entire data ecosystem



Data Analysts & Data Scientists



alteryx



Self Service, Simplified Access:



Global Data Mesh:



Proprietary Structured



Proprietary UnStructured



Internal Cloud Storage



Furnished



Public Records



API Access

Integrated Hybrid Cloud Control Plane:



Private

## Outcomes

Self-serve & domain-agnostic

Access to data where it resides

Efficient collaboration

Data governance

# We are the forefront of credit inclusion by enabling our customers to score more customers

## Trended Credit Data

### 30 months of account history

- Payment History
- Mortgages
- Car loans
- Credit cards



## Alternate Credit Data

### 3B records on >260M US consumers

- Deposit Accounts
- Address stability
- Public records
- Microloan activity

# 60M+

unscorable consumers  
gain access to credit

# TransUnion CIBIL and SatSure launched the Credit and Farm Report to promote financial inclusion and economic opportunity in India

## Bureau Information

- 1 Identity Verification
- 2 Demographic Details
- 3 Consumer Loan Summary
- 4 Consumer Score
- 5 Derogatory Instances (Write off/Settled)
- 6 Delinquency Trend (36 Months)
- 7 Enquiry Trend

## Agri Information

- 1 Crops Growth
- 2 Crop Yield Performance
- 3 Cropping Intensity; Irrigated/Rain Fed
- 4 Farm Score
- 5 Regional Matrix
- 6 Weather, Water Condition, Rainfall Trend
- 7 Farm Ownership & Land Verification



**55%**

of India's total workforce is its rural economy

**63%**

of farmers are currently unable to borrow

**~89M**

farmers will gain access to formal credit

# 7-Eleven



## Modernized Customer 360 Cloud Data Lakehouse

### Customer Overview

- 7-Eleven, Inc. is an American chain of convenience stores, headquartered in Dallas, TX
- Today 7-Eleven operates, franchises, and licenses 71,100 stores in 17 countries as of July 2020.

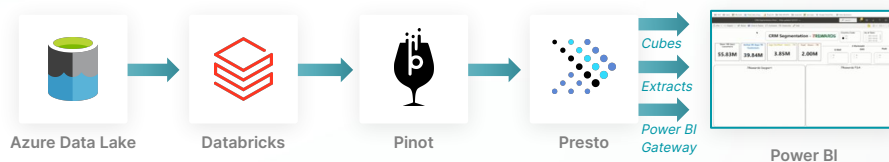
### Pain Points

- Rising resource costs & architecture complexity
- Suboptimal performance for dashboards & reporting tools
- Poor user experience & lack of self-service analysis against raw data

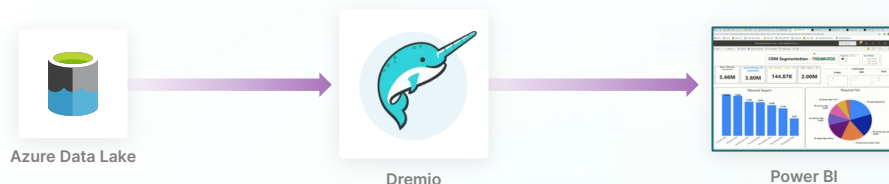
### Ideal Solution

- Ability to query data directly in the data lake, reducing the need to copy/move data
- Semantic layer providing a consistent view of the data and self service experience for the users
- Ability to query from heterogeneous data sources (multi-cloud & relational databases i.e.: SQL Server)

### ARCHITECTURE BEFORE DREMIO



### ARCHITECTURE AFTER DREMIO



### Results (Dremio Value)

Lower cost & complexity of modern data architecture

Empower data consumers with more self-sufficiency

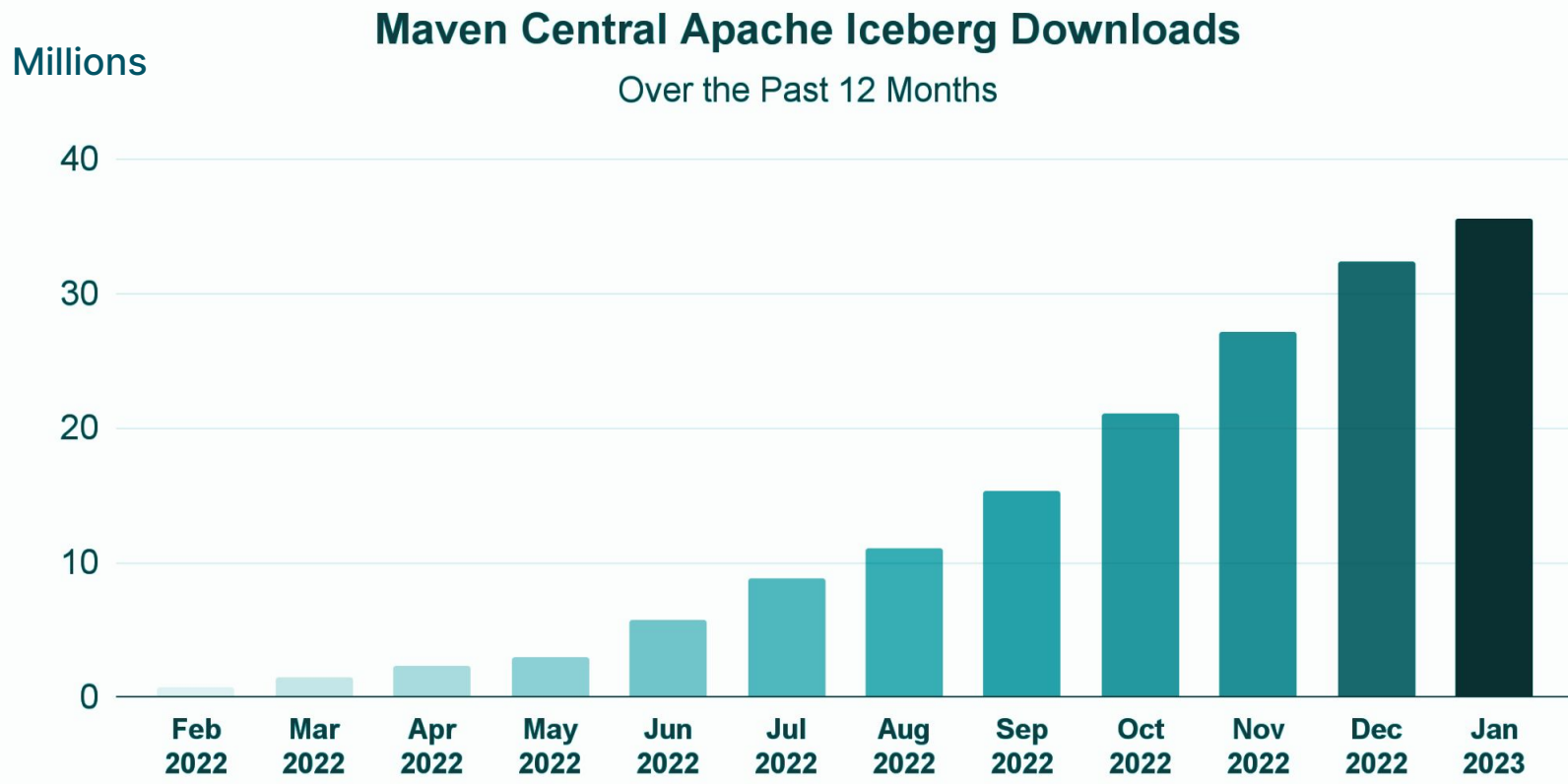
Provide faster time to analytics to facilitate better data-driven decisions



# Apache Iceberg

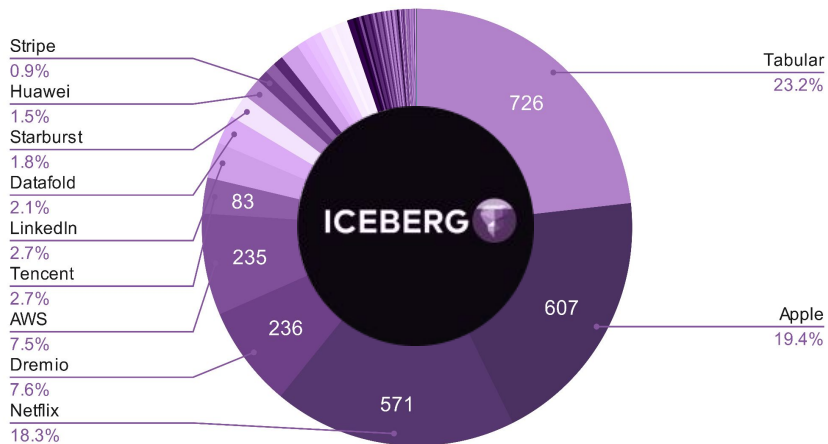
Enabling Full Data Warehouse Capabilities on the Data Lake

# Apache Iceberg Has Taken Off in the Last Year

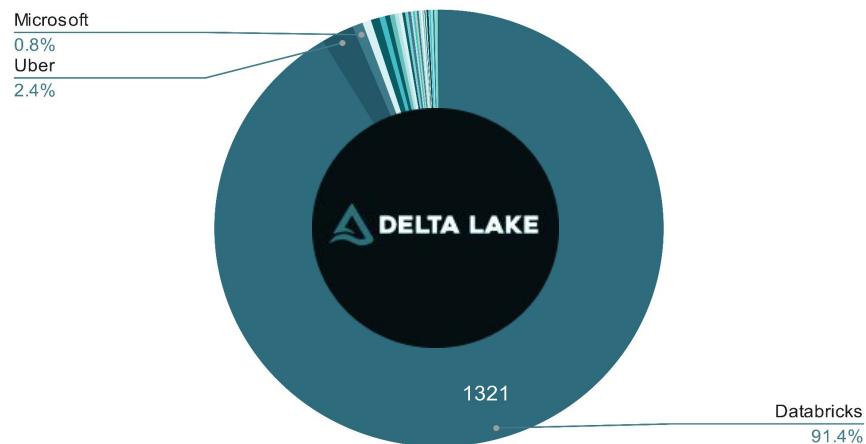


# Iceberg is a Community-Built, Vendor-Agnostic Table Format

% of Attributable Contributions to Apache Iceberg by Company



% Attributable Contributions to Delta Lake by Company



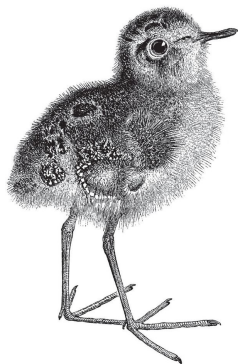
Source: Github repository contribution statistics, data from January 13, 2023

# Apache Iceberg: The Definitive Guide

O'REILLY®

## Apache Iceberg The Definitive Guide

Data Lakehouse Functionality, Performance,  
and Scalability on the Data Lake



Tomer Shiran,  
Jason Hughes,  
Alex Merced &  
Dipankar Mazumdar



# Dremio Makes Iceberg Easy

# Ingest Data into Iceberg Tables

## COPY INTO

Ingest existing data into an Iceberg table

```
{"id": 1, "name": "Bob", "age": 46}  
{"id": 2, "name": "Josie", "age": 65}  
{"id": 3, "name": "Gene", "age": 30}
```

```
COPY INTO mydomain.mytable  
FROM @SOURCE/bucket/path/folder]  
[ FILES ('foo.json');
```



id	name	age
1	Bob	46
2	Josie	65
3	Gene	30

# Manipulate Iceberg Tables

id	name	email
1	Alex Merced	alex.merced@dremio.com
2	Bob Jones	

+

id	name	email
2	Bob Jones	Bob@SomeDomain.xyz
3	Gina Somebody	GSomebody@Domain.xyz

## DML

Insert, update, delete and merge records in an Iceberg table

```
MERGE INTO names n
USING (SELECT * FROM names_staging) s
ON n.id = s.id
WHEN MATCHED THEN UPDATE SET name = s.name, age, s.email
WHEN NOT MATCHED THEN INSERT (id, name, email) VALUES (s.id, s.name,
s.email)
```

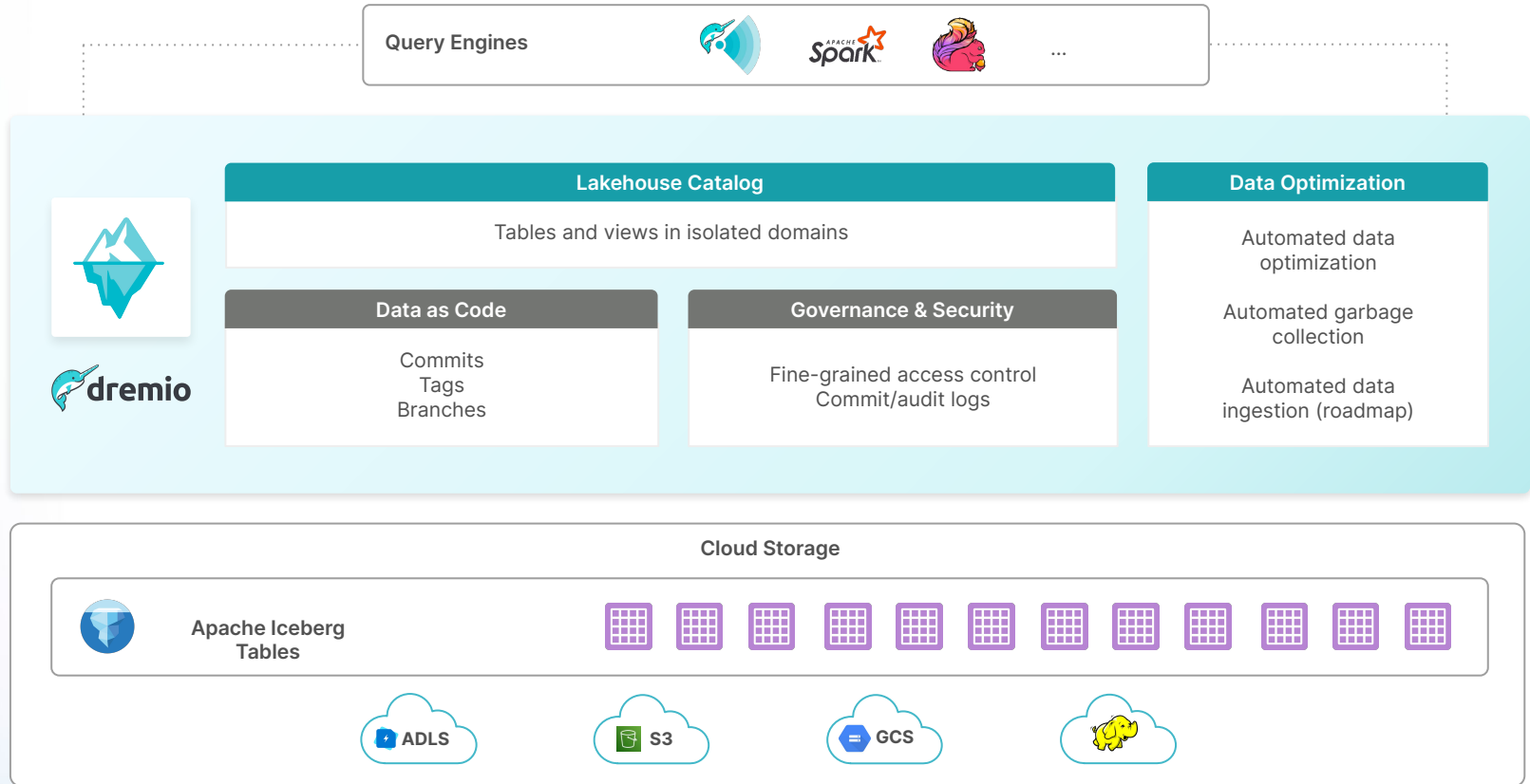
id	name	email
1	Alex Merced	alex.merced@dremio.com
2	Bob Jones	Bob@SomeDomain.xyz
3	Gina Somebody	GSomebody@Domain.xyz

# Dremio Arctic

A Lakehouse Management Service



# Dremio Arctic is a Data Lakehouse Management Service



# 5 Use Cases for Data as Code

# 1: Ensure data quality with ETL branches

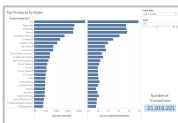
Create an ETL branch and ingest the data with COPY INTO, CTAS or Spark:

```
CREATE BRANCH events_etl_9_28_22
USE BRANCH events_etl_9_28_22
COPY INTO web.events ...
```

Run queries to test data quality:

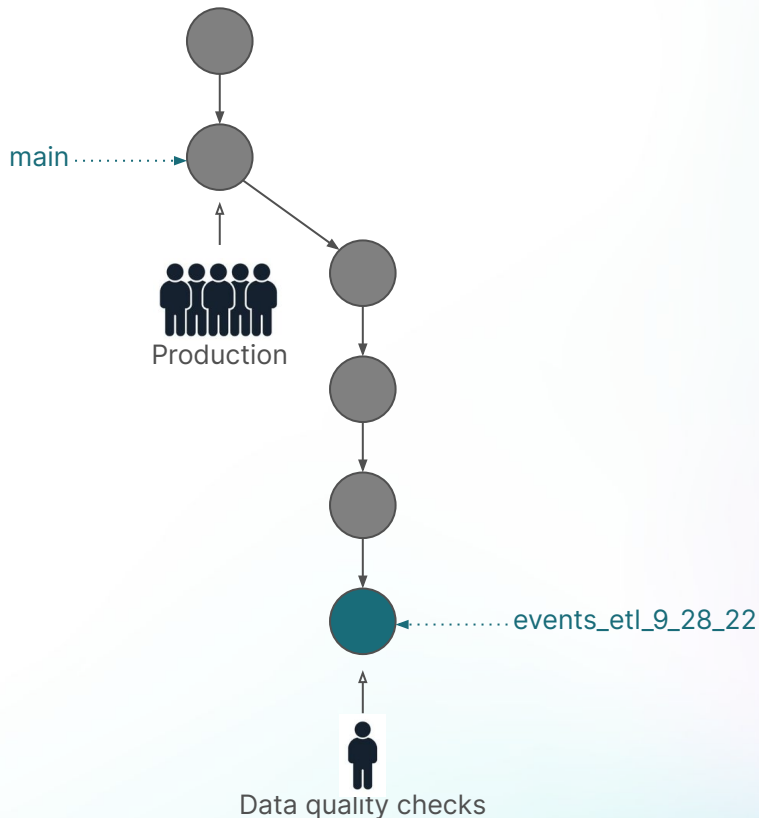
```
SELECT COUNT(*) FROM web.events WHERE
length(ip_address) >= 7
```

Test the dashboard to see that it looks okay:



Fix the problems and merge into main:

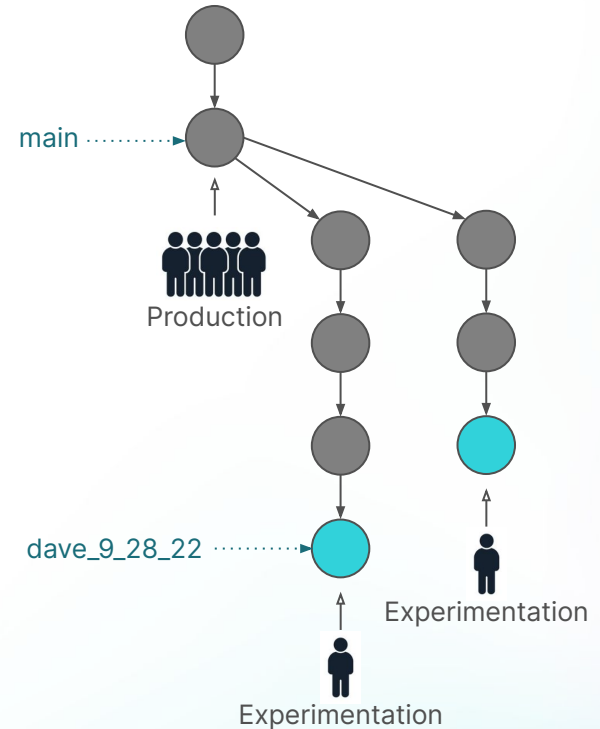
```
DELETE FROM web.events WHERE length(ip_address) >= 7
USE BRANCH main
MERGE BRANCH events_etl_9_28_22
```



## 2: Experiment with data in transient branches

Create a transient branch and perform data explorations and transformations in it:

```
CREATE BRANCH dave_9_28_22
USE BRANCH dave_9_28_22
CREATE TABLE t AS SELECT ...
UPDATE t ... SET ...
```



### 3: Reproduce models or analysis

Change context to a named tag:

```
spark.sql("USE REFERENCE modelA IN arctic")
```

Create ML model based on historic data:

```
val trainingData = spark.read.table("arctic.t")
val lr = new LogisticRegression()
// configure logistic regression...
val paramMap = ParamMap(...)
val model = lr.fit(trainingData, paramMap)
```

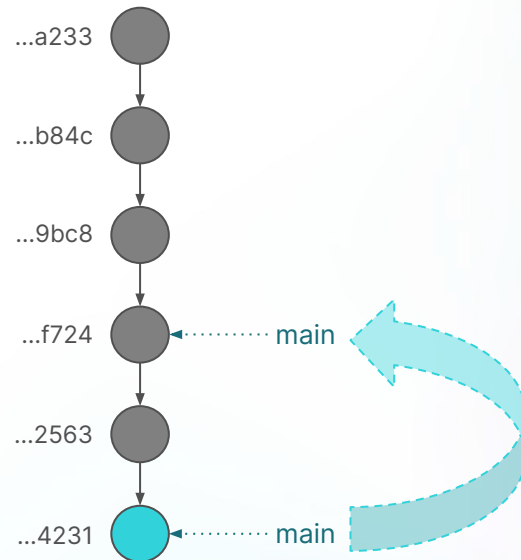
## 4: Recover from mistakes

If you accidentally mess up the data or schemas in your lakehouse:

```
INSERT INTO sales
  (SELECT * FROM
   sales_last_quarter_unaudited)
DROP TABLE customers
```

Move the branch head to a historical commit:

```
ALTER BRANCH main ASSIGN COMMIT ...f724
```



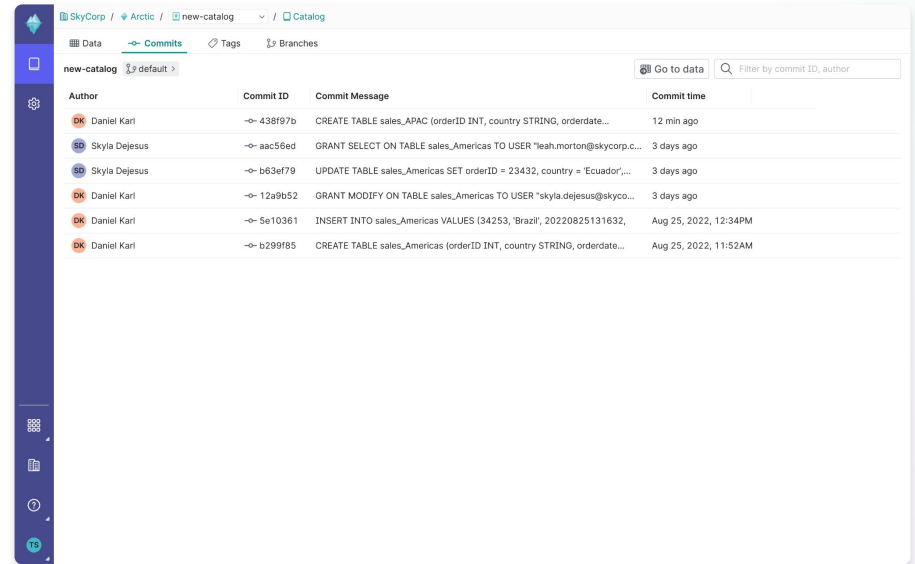
# 5: Troubleshooting (see who changed the data)

Get the commit history for a branch:

```
SHOW LOGS AT REFERENCE etl;
```

Get the commit history for a specific table:

```
curl -X GET -H 'Authorization: Bearer <PAT>' <Catalog API Endpoint>/trees/tree/<reference name>/log\?filter="operations.exists(op,op.key=='<table name>')"
```



Author	Commit ID	Commit Message	Commit time
Daniel Karl	438f97b	CREATE TABLE sales_APAC (orderID INT, country STRING, orderdate...	12 min ago
Skyla Dejesus	aac56ed	GRANT SELECT ON TABLE sales_Americas TO USER "heah.morton@skycorp.c...	3 days ago
Skyla Dejesus	b63ef79	UPDATE TABLE sales_Americas SET orderID = 23432, country = 'Ecuador',...	3 days ago
Daniel Karl	12a9b52	GRANT MODIFY ON TABLE sales_Americas TO USER "skyla.dejesus@skycor...	3 days ago
Daniel Karl	5e10361	INSERT INTO sales_Americas VALUES (34253, 'Brazil', 20220825131632,	Aug 25, 2022, 12:34PM
Daniel Karl	b299f85	CREATE TABLE sales_Americas (orderID INT, country STRING, orderdate...	Aug 25, 2022, 11:52AM

# Dremio Arctic Demo



# MERLIN

**Merlin Networks is the largest and most trusted international digital music licensing partner for independents.**

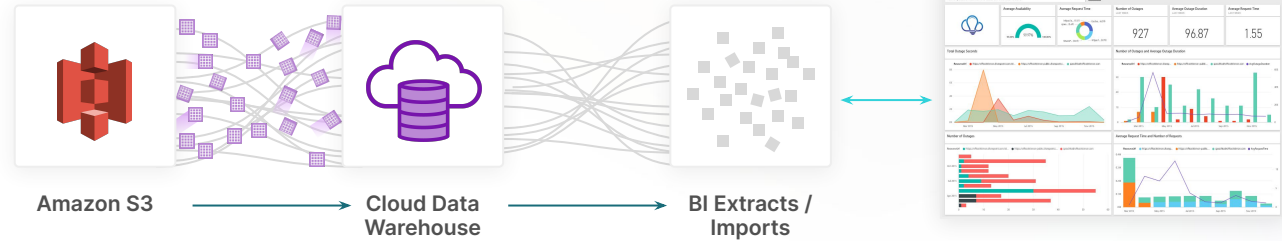
Merlin Networks helps independent music labels and artists:

- Sign distribution deals with digital service providers (e.g. Spotify, Apple Music)
- Track “listens” across platforms
- Pay independents for their music

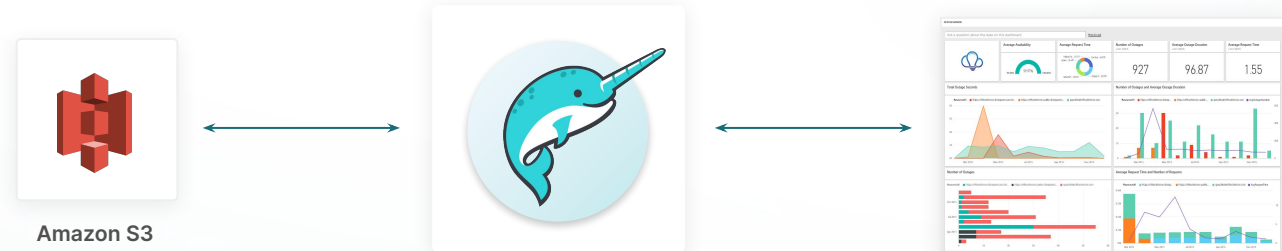
# Merlin Started By Moving to Dremio



## Architecture Before Dremio



## Architecture After Dremio



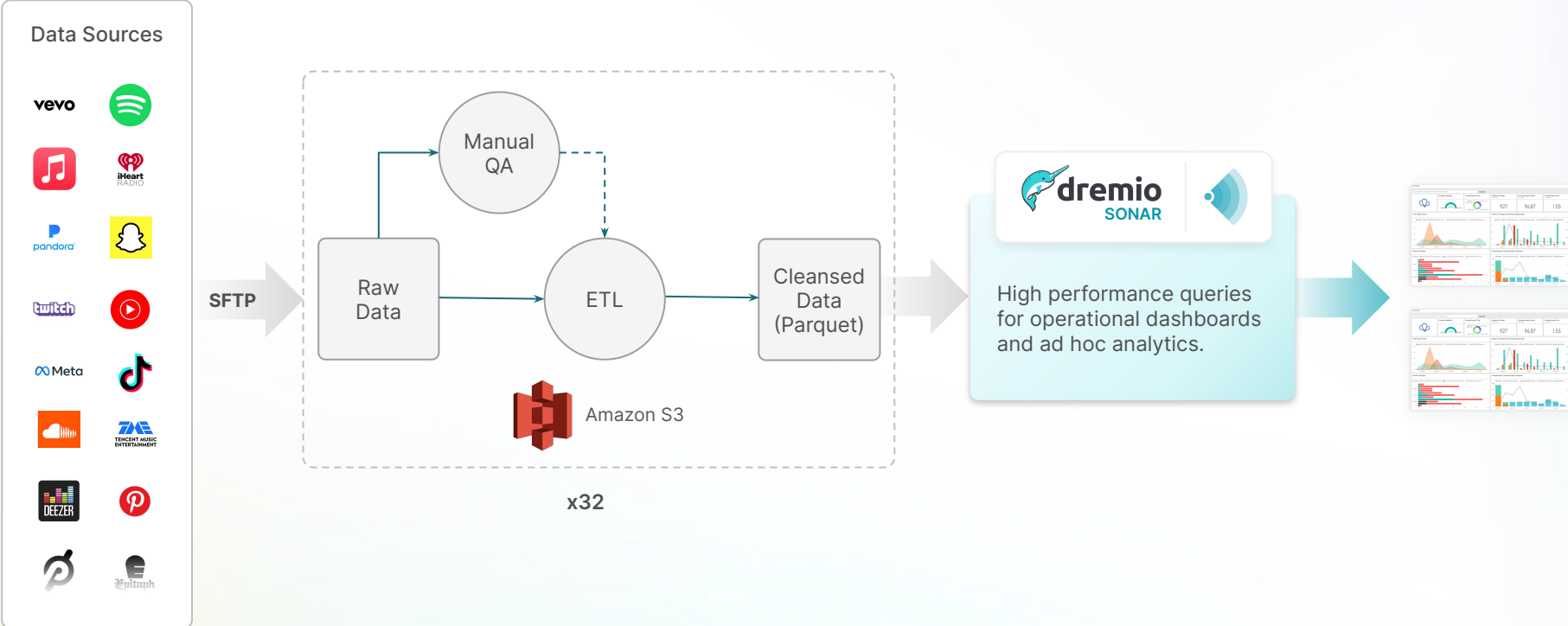
## Results

Empowered data consumers with self-service layer for BI team to access data

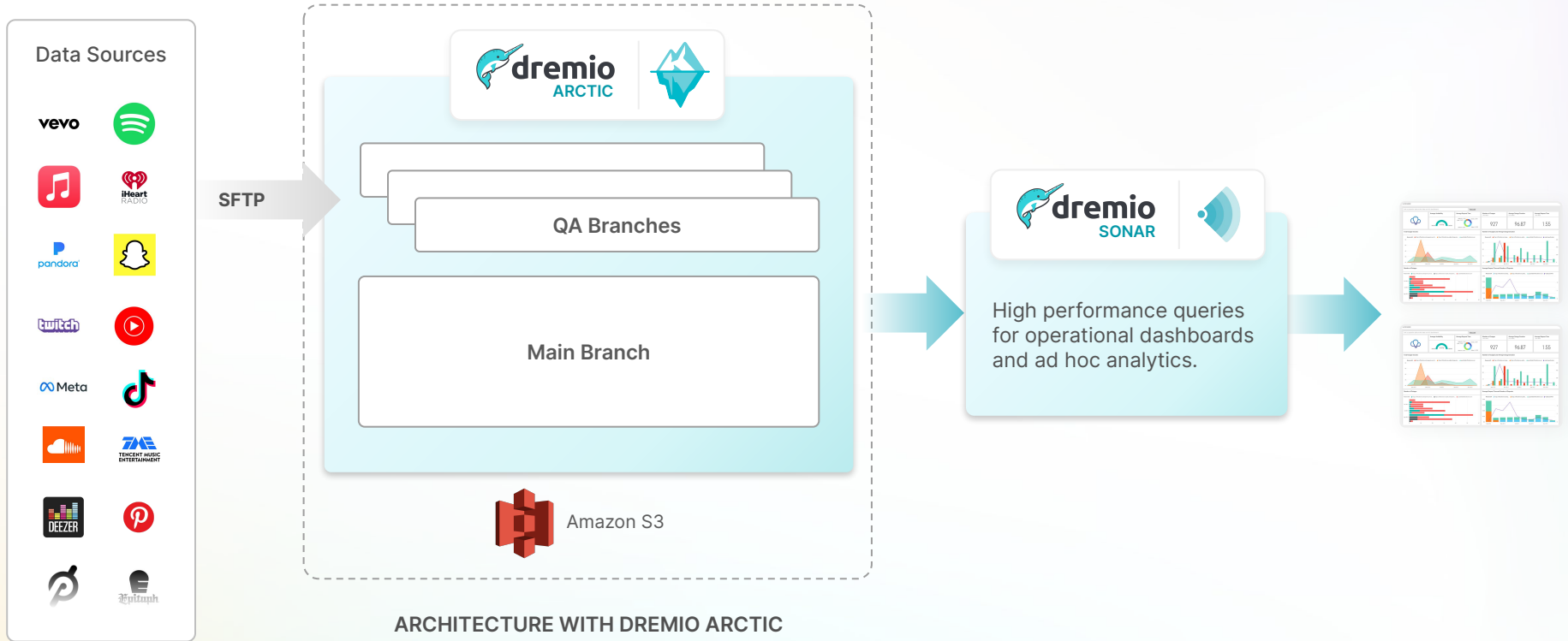
Elimination of data copies resulting in improved security and governance

Increased speed of BI reporting on royalty data

# Merlin's Manual Data Quality Process

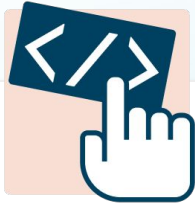


# Merlin's Data Quality Process with Dremio Arctic



ARCHITECTURE WITH DREMIO ARCTIC  
Data Quality Assurance with Data as Code

# The Dremio Advantage

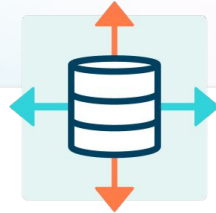


## Self-Service Analytics

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

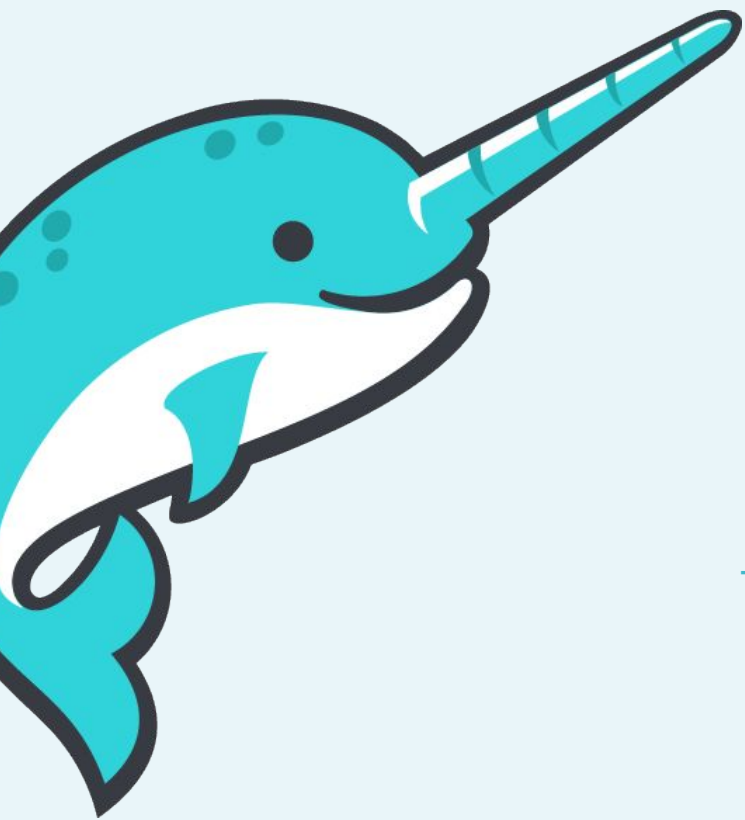


## The Fastest BI Performance at 1/10th the Cost

Lightning-fast queries

High concurrency

No expensive data copies to manage



# Dremio

The Easy and Open Data Lakehouse