

GNARLY Data_Waves

PRESENTED BY  **dremio**

Episode 53

Build the next-generation Iceberg lakehouse with Dremio and NetApp



July 9th, 2024 | 8AM PDT | 11AM EST | 4PM BDT



Vishnu Vardhan

Director of Product Management
StorageGRID
NetApp



Alex Merced

Senior Technical Evangelist
Dremio

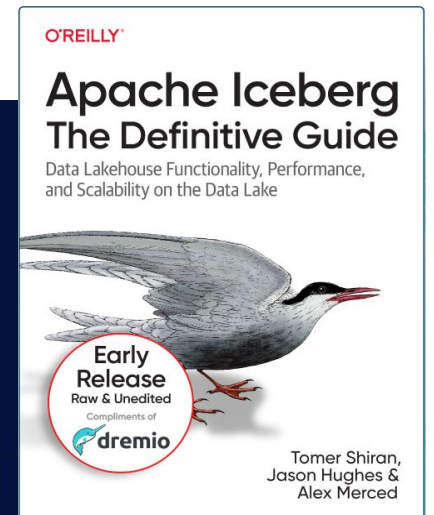
Dremio: Expert and Leader in the Apache Iceberg Lakehouse



- One of the first platforms to embrace Apache Iceberg
- Created some of the earliest online Apache Iceberg education materials
- Creators of O'reilly's "Apache Iceberg: The Definitive Guide"
- Creators of the first open-source Apache Iceberg Catalog (Nessie), and only catalog with catalog level Git-for-data features.



APACHE ICEBERG 101



An Apache Iceberg Crash Course

10 Sessions

July 11 - October 29



Alex Merced
Sr. Evangelist
Dremio



Agenda

- Market News
- Why Iceberg Lakehouses
- Dremio & NetApp Joint Offering
- NetApp StorageGRID & Dremio Deep Dive
- Joint Solution Benefits & Use Cases
- Case Studies

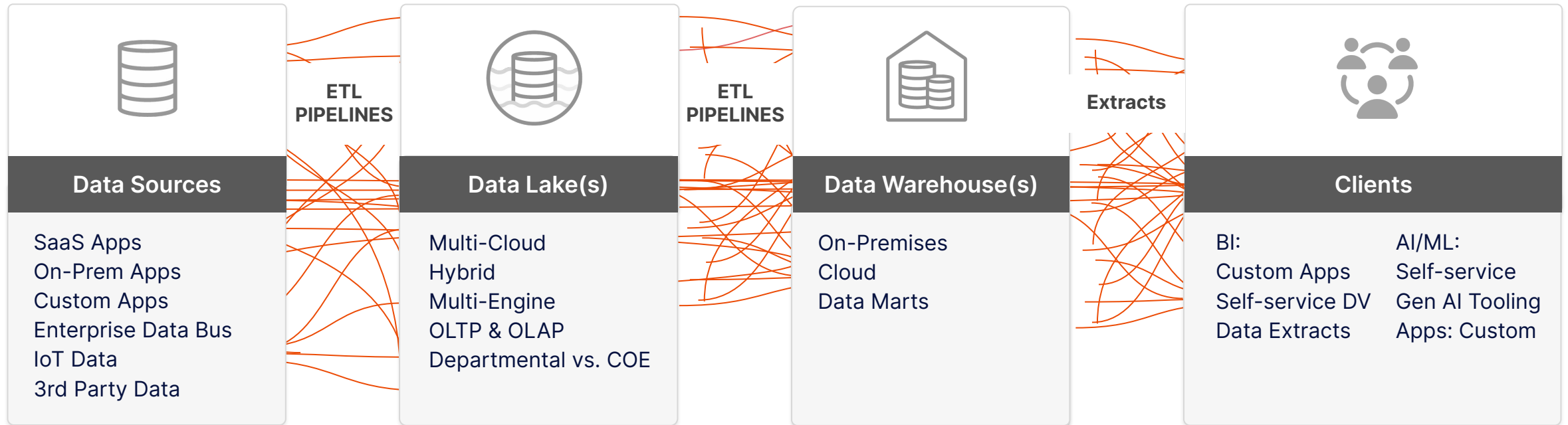


dremio



NetApp

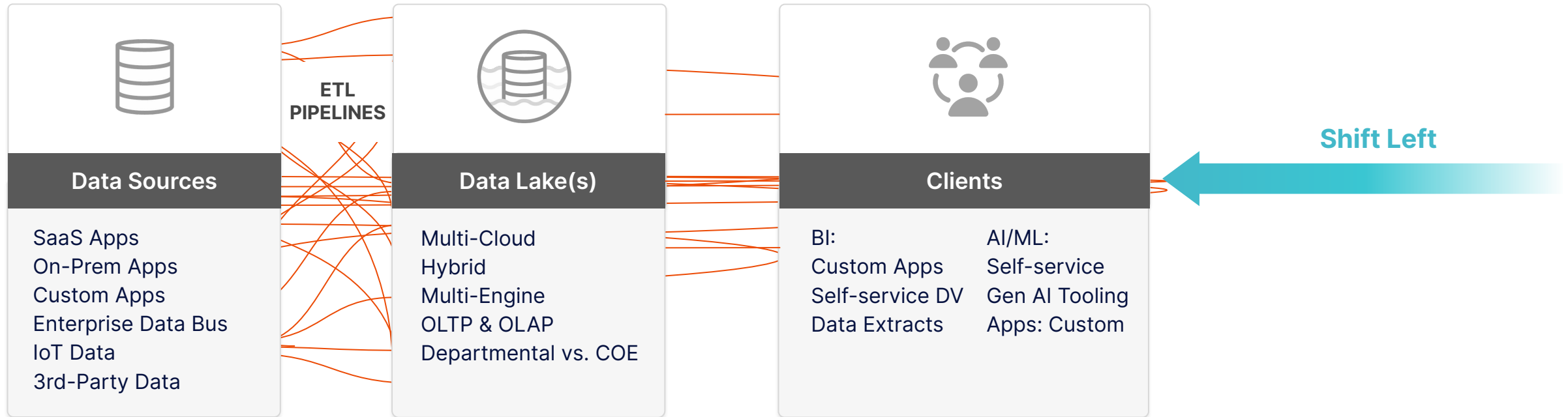
Data Lifecycle Remains Complex, Brittle, and Expensive



Data lifecycle and management remains complex, expensive, and time-consuming, especially for large organizations

Duplicate copies, 'expert' ETL, "dark data", governance complexity, not self-service

Dremio & NetApp Iceberg Data Lakehouses Help!



Reduced Cost, no ETL to a data warehouse, reduced extracts, no copies to manage
Simplifies data pipelines, improves “uptime” for data accelerates and time-to-insight

Databricks purchases Tabular

Deals

Databricks to buy data management firm Tabular for over \$1 bln

By Reuters

June 4, 2024 11:39 AM PDT · Updated 17 days ago

June 4 (Reuters) - Databricks said on Tuesday it would buy data-management startup Tabular for more than \$1 billion, as the privately held analytics platform looks to attract customers by helping them develop custom artificial intelligence (AI) ap

The announcement comes at a time clients to use open-source AI mode (SNOW.N) and Cloudera, among

Databricks acquires Tabular in a big move to close the compatibility gap with Apache Iceberg

BY PAUL GILLIN

Databricks Inc. today agreed to acquire Tabular Technologies Inc., developer of a universal storage platform based on the Apache Iceberg standard.

The move signals stepped-up efforts by Databricks to bridge the compatibility gap between its Delta Lake storage format and Iceberg. Terms weren't announced, but Databricks Chief Executive Ali Ghodsi (pictured) told CNBC the price tag was more than \$1 billion. Snowflake Inc. and Confluent were also reportedly in on the bidding.

Tabular was founded by three former Netflix Inc. employees who co-created Iceberg at that company.

Home > AI/ML > Databricks buys Tabular to win the Iceberg war

AI/ML Data Management

Databricks buys Tabular to win the Iceberg war

By Chris Mellor · June 5, 2024



APP DEV CLOUD GEN AI MACHINE LEARNING ANALYTICS NEWSLETTERS RESOURCES

NEWS

Databricks \$1B-plus Tabular acquisition adds Iceberg support

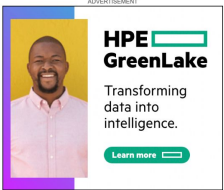
The lakehouse specialist's latest purchase adds support for Apache Iceberg to its existing support for Delta Lake and is also a direct confrontation of rival Snowflake.

By Eric Avidon, Senior News Writer

Published: 06 Jun 2024

Databricks continued its recent buying spree with the acquisition of Tabular, a move that adds support for Apache Iceberg storage to Databricks' existing support for Delta Lake storage.

Databricks did not specify what it paid for Tabular but confirmed that the cost was more than \$1 billion.



Databricks to acquire storage platform maker Tabular

While Snowflake is talking up its use of Iceberg to promote interoperability, Databricks is buying Tabular, the tool built on Iceberg's table format by Iceberg's creators.

Facebook Twitter LinkedIn Reddit Email Print

By Anirban Ghoshal

Senior Writer, InfoWorld | JUN 4, 2024 1:03 PM PDT

Snowflake Releases Polaris Iceberg Catalog

Snowflake Releases Polaris Catalog: Transforming Data Interoperability with Open Source Apache Iceberg Integration

By Asif Razzaq · June 4, 2024

Reddit

Y

F

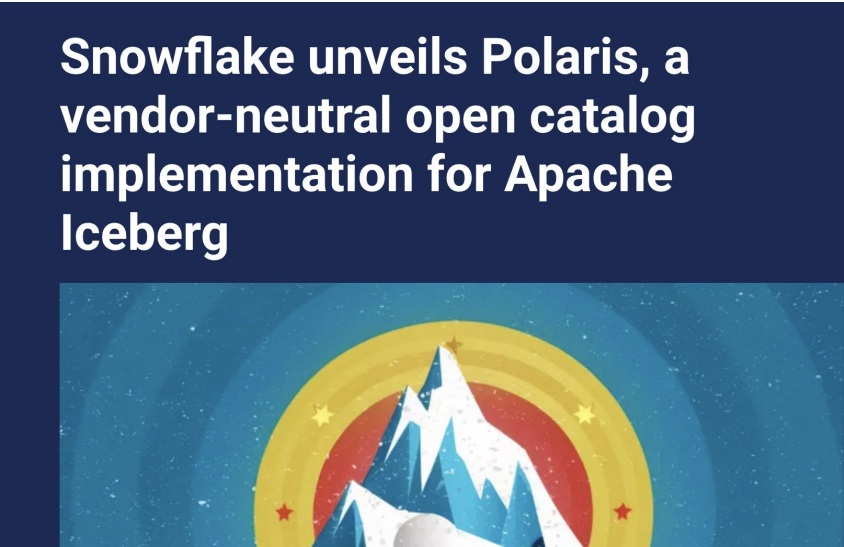
in

X

0 SHARE

Snowflake has unveiled the **Polaris Catalog**, an open-source catalog for Apache Iceberg that enhances data interoperability across various engines and cloud services. This launch signifies Snowflake's commitment to providing enterprises more control, flexibility, and security for their data management needs.

The data industry has increasingly embraced open-source file and table formats for their potential to improve interoperability. This capability allows multiple technologies to operate over a single copy of data, reducing complexity, costs, and risks associated with vendor



Snowflake, Databricks and the Fight for Apache Iceberg Tables

DATA

The market for data lakes and data lakehouses is clearly being disrupted by open source software, given recent news from Databricks and Snowflake.

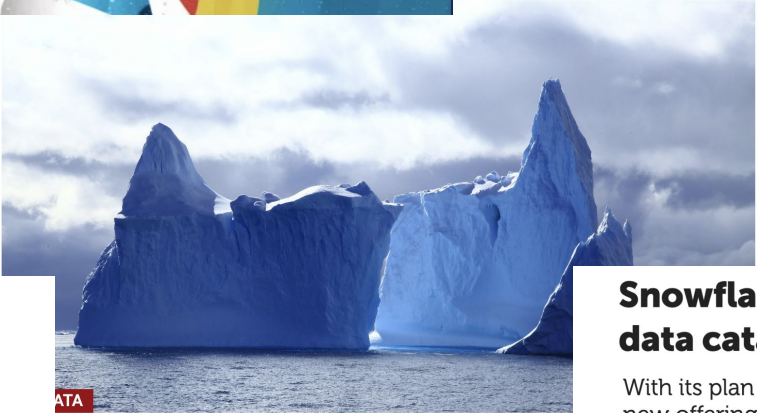
Jun 10th, 2024 9:46am by Joab Jackson

CLOUD SERVICES / DATA

Snowflake Polaris Aims for Multiquery Engine Interoperability

An open source catalog for Apache Iceberg, Snowflake Polaris provides a way for multiple query engines to write Apache Iceberg tables.

Jun 3rd, 2024 2:05pm by Joab Jackson



Snowflake catalog supports cross-engine access

By PAUL GILLIN



DATA SCIENCE • AI • ADVANCED ANALYTICS

AboutResourcesSubscribe

HOMECOVID-19FEATURES ▾SECTORS ▾APPLICATIONS ▾TECHNOLOGIES ▾

June 3, 2024

Snowflake Embraces Open Data with Polaris Catalog

Alex Woodie



Photo by Shutterstock

On the first day of its Data Cloud Summit today, Snowflake unveiled Polaris, a new data catalog that stores data in the Apache Iceberg format. In addition to contributing Polaris to the open source community, the catalog also enables Snowflake customers to use open compute engines with their Iceberg-based Snowflake data, including Apache Spark, Apache Flink, Presto, Trino, and Dremio.

The launch of Polaris represents a significant embrace of open source and open data on the part of [Snowflake](#), which grew its business predominantly through a closed data stack, including proprietary table format and a proprietary SQL processing engine.

Snowflake adopts open source strategy to grab data catalog mind share

With its plan to make its Polaris data catalog open source, Snowflake hopes the new offering will be seen as vendor-neutral, boosting its attractiveness when compared to Databricks' Unity Catalog.

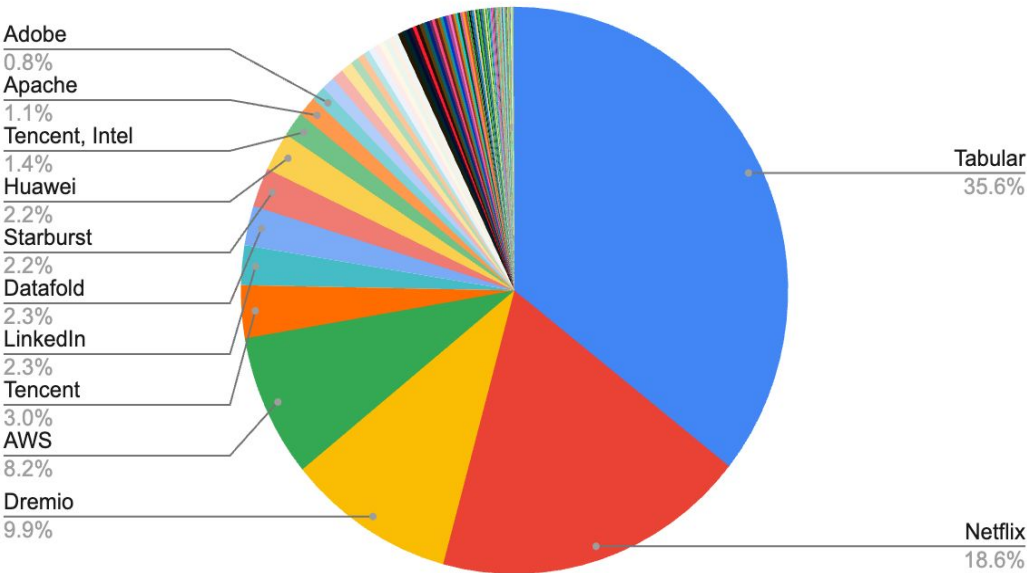


 By **Anirban Ghoshal**
Senior Writer, InfoWorld | JUN 3, 2024 6:00 AM PDT

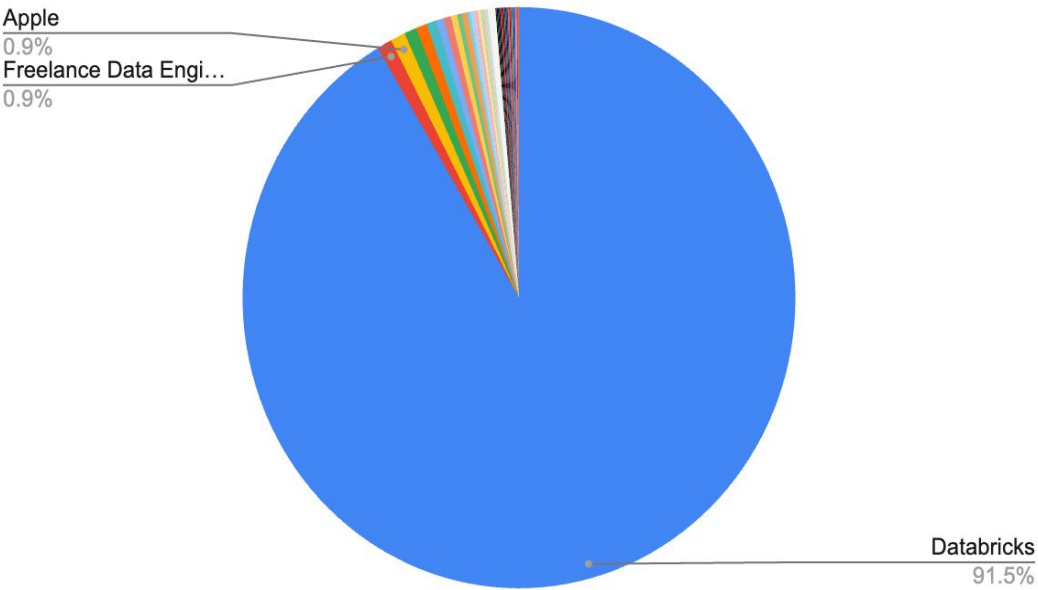
Why Iceberg? A Diverse Developer Community



% Attributable Contributions to Apache Iceberg by Company



% Contribution to Delta Lake by Company



Iceberg for Your Next Generation Lakehouse

Queries are faster (even without perfect users)

- Automatic up-to-date statistics
- Hidden partitioning reduce full table scans
- Optimal data layout (file sizes, prefixes, etc.)

Changes are easier

- No need to refresh metadata after updating a table
- Atomic updates in one command (INSERT/UPDATE/DELETE)
- Raw files (CSV/JSON) to queryable and transactional table in one command (COPY INTO)
- Schema/partition evolution with one command

Less work for the data team

- Automatic data optimization
- Automatic garbage collection
- No metadata refresh jobs to trigger and manage
- No manual work to insert/update data or evolve schema/partitions
- Recovering from mistakes and disasters is trivial
- **Fewer tickets!!**

Lower compute and storage costs

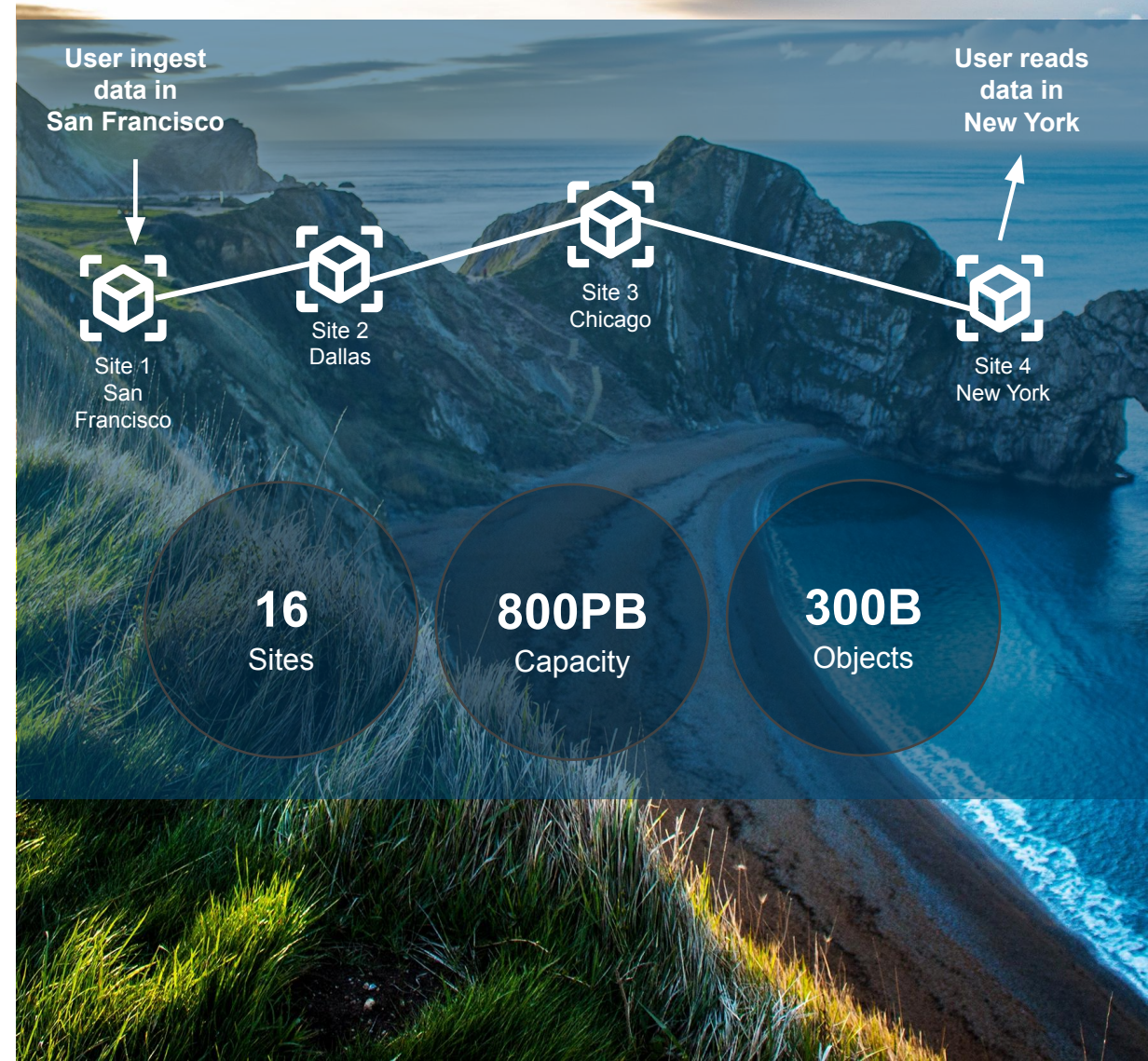
- Time travel reduces data copies
- Less compute per query (see “queries are faster”)
- No need to copy data into a data warehouse

- + No vendor lock-in
- + Broadest ecosystem of any table format

NetApp StorageGRID

Industry leading Object Storage solution

- **Single Global Active/Active Namespace**
- **Powerful Policy Engines**
 - Automatically place data to the selected media, data center, and/or protection scheme
- **Flexible & Simple to Deploy**
 - Mix & match software-defined and appliances
- **Unmatched Durability**
 - Achieve 15 9's of durability leveraging layered erasure coding
- **Scalable Architecture**
 - Supports 300 billion objects and 640+PBs in single namespace
- **Enabling the Hybrid Cloud**
 - Cloud integration with AWS workflows and services
 - Tier data to Microsoft Azure, AWS S3/Glacier & Google Cloud



Automate Lakehouse data management at scale with ILM policies

Policies determine

- Where data lives
 - Place objects in specific data centers or specific media (disk, tape, cloud)
- How long data is stored
 - Have different lifecycles for objects
- How secure the data is stored
 - Create more or fewer replicas of objects or erasure encode objects

Apply policies on existing objects

- StorageGRID[®] automatically brings objects into policy alignment with a fast background engine

The screenshot shows the 'German Data Policy' configuration page in the StorageGRID interface. At the top, there are buttons for '+ Create', 'Edit', 'Clone', and 'Remove'. Below these is a table with columns 'Name', 'In Active Policy', and 'In Proposed Policy'. The table lists two policies: 'Make 2 Copies' (which is active, indicated by a checkmark) and 'German Data Policy' (which is selected with a blue circle). Below the table, there are navigation arrows. The main section for the 'German Data Policy' (Version 1.0) includes a 'Description' (Data needs to stay in German Datacenters), 'Tenant Account' (94352640785552504147), 'For Object Type' (S3/Swift), 'Reference Time' (Ingest Time), and 'Filtering Criteria'. The filtering criteria are defined as 'Matches all of the following metadata: location equals germany'. At the bottom, a 'Retention Diagram' shows two data paths: one for 'Frankfurt DC' with a blue bar extending to 'Forever', and another for 'Berlin DC' with an orange bar extending to 'Forever'. A vertical dashed line marks 'Day 0'.

Name	In Active Policy	In Proposed Policy
Make 2 Copies	✓	
German Data Policy		

German Data Policy Version 1.0

Description: Data needs to stay in German Datacenters
Tenant Account: 94352640785552504147
For Object Type: S3/Swift
Reference Time: Ingest Time
Filtering Criteria: Matches all of the following metadata:
location equals germany

Retention Diagram:

Trigger Day 0

Frankfurt DC

Berlin DC

Duration Forever

Simple and feature rich software-defined object storage

ILM

- Granular, Rule based Hierarchical Storage Management



Load balancing

- Purpose-built Load Balancer with HA
- Advanced traffic QoS
- Endpoint port binding
- Multi-VLAN



Identity & Access management

- Federated: LDAP/AD
- S3 IAM bucket and group policies



Bucket management

- Versioning
- Object Lock
- Object Lifecycle
- Federation



Service Provider Ready

- Multi-tenant capable
- Audit logging & Metering support



Restfull APIs

- APIs for all actions by Grid Admins or Tenant Admin



Usage quota

- Grid admin-specified
- Total data stored in account



Security & Encryption

- Multi-level encryption
- KMS Integration



Cloud Integration

- Cloud Replication,
- Cloud Storage Pools
- Event Notifications

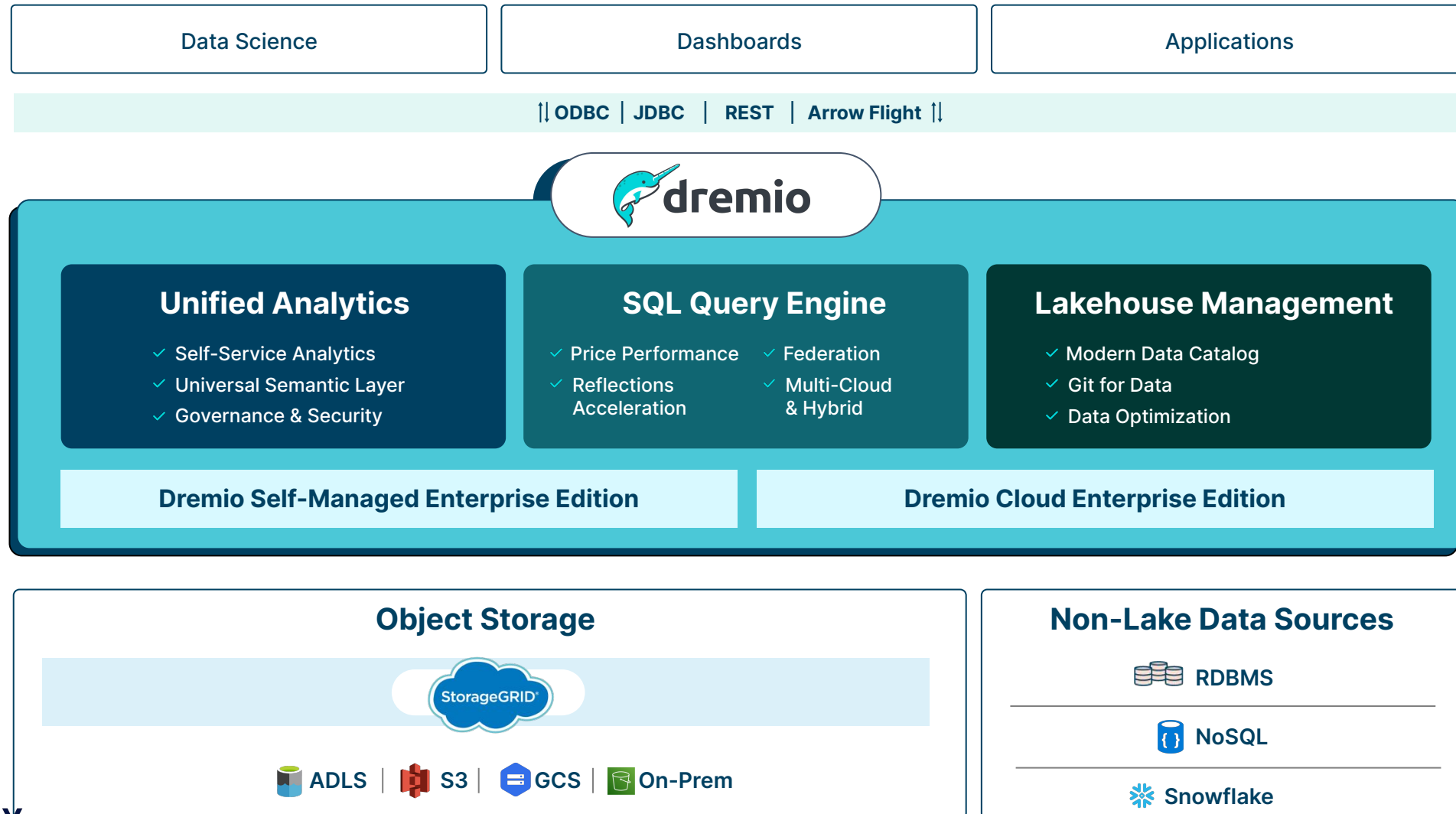


Deployment Options

- Appliances
- VMs
- Bare Metal



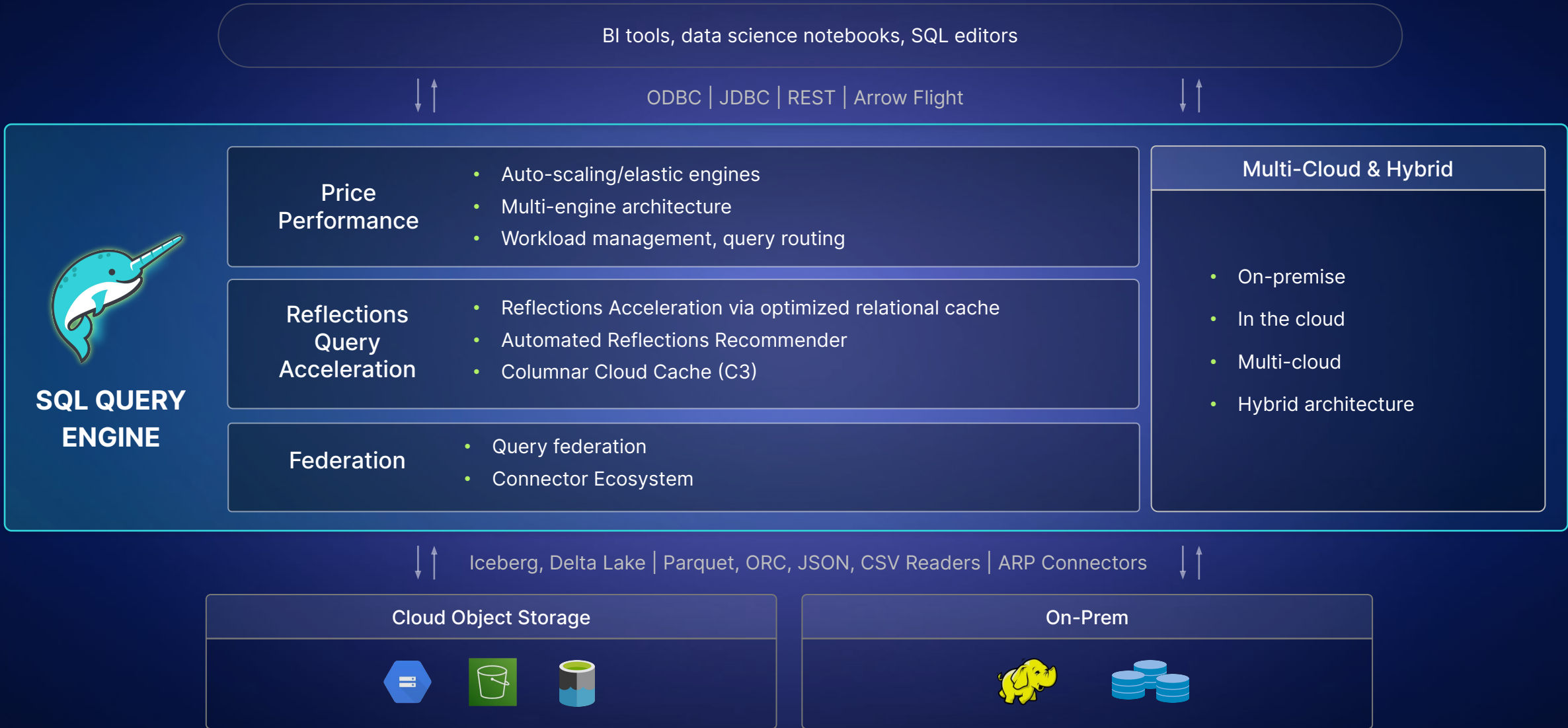
Dremio: The Unified Lakehouse Platform for Self-Service Analytics & AI



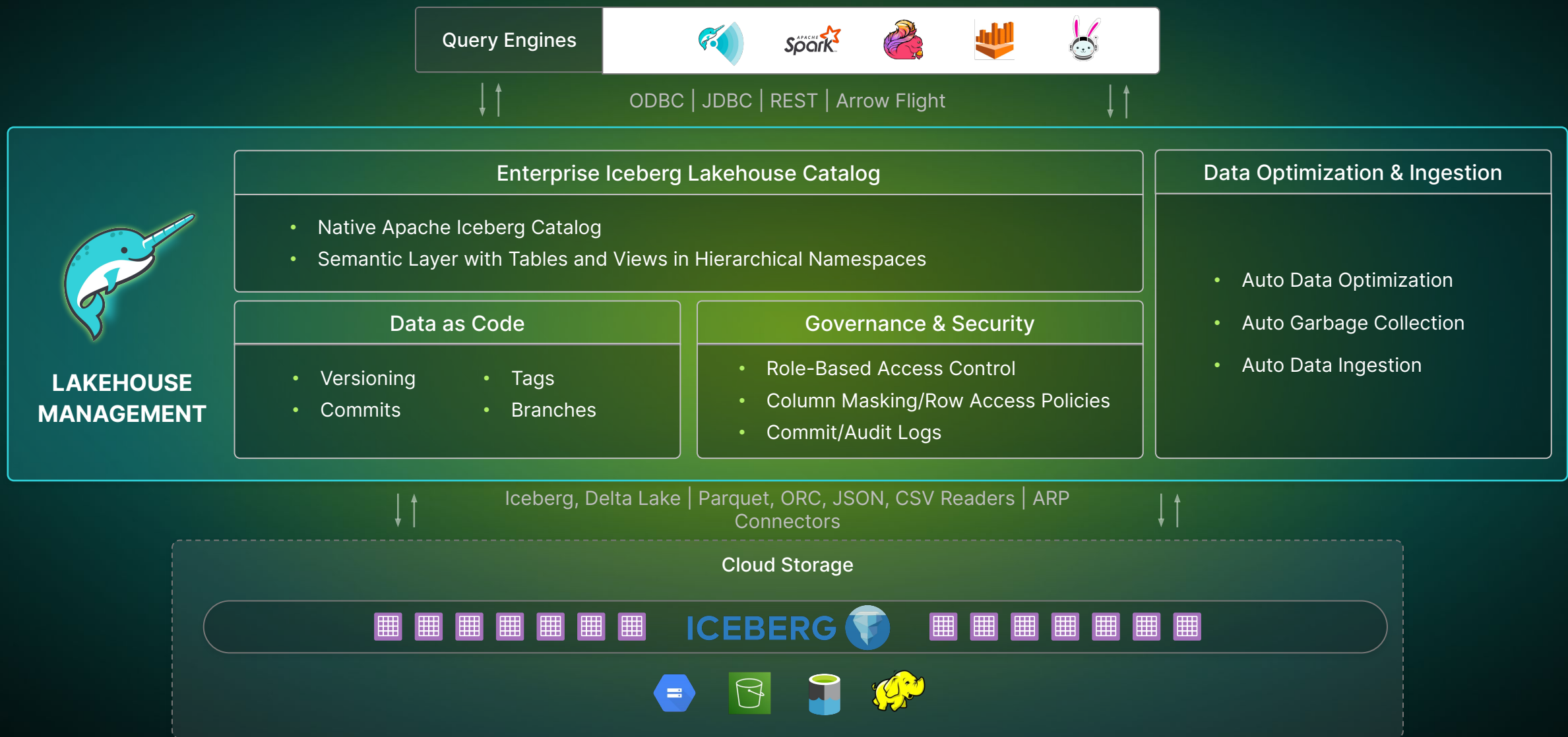
Unified Analytics



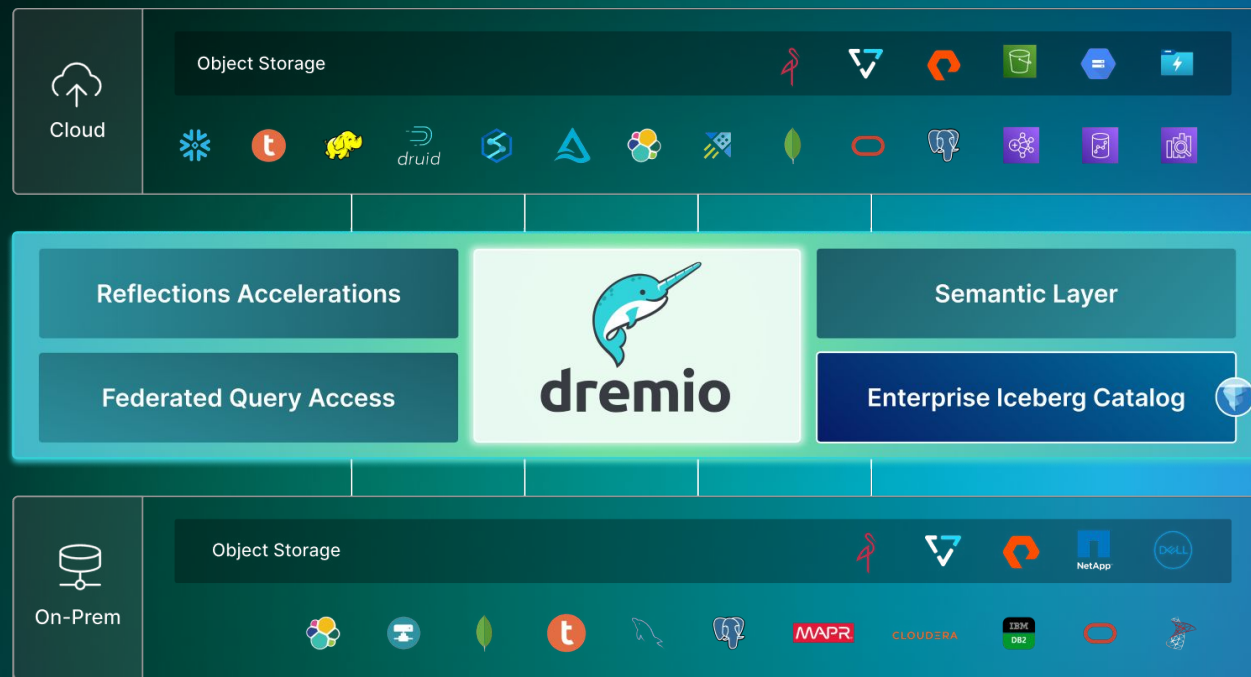
SQL Query Engine



Lakehouse Management



Dremio Hybrid Iceberg Lakehouse



- Enterprise Hybrid Cloud Iceberg Catalog
- Leverage on-prem and cloud based storage
- Enable self service across all data sources



Cloud

Object Storage



Reflections Accelerations

Federated Query Access



dremio

Semantic Layer

Enterprise Iceberg Catalog



On-Prem

Object Storage



MAPR

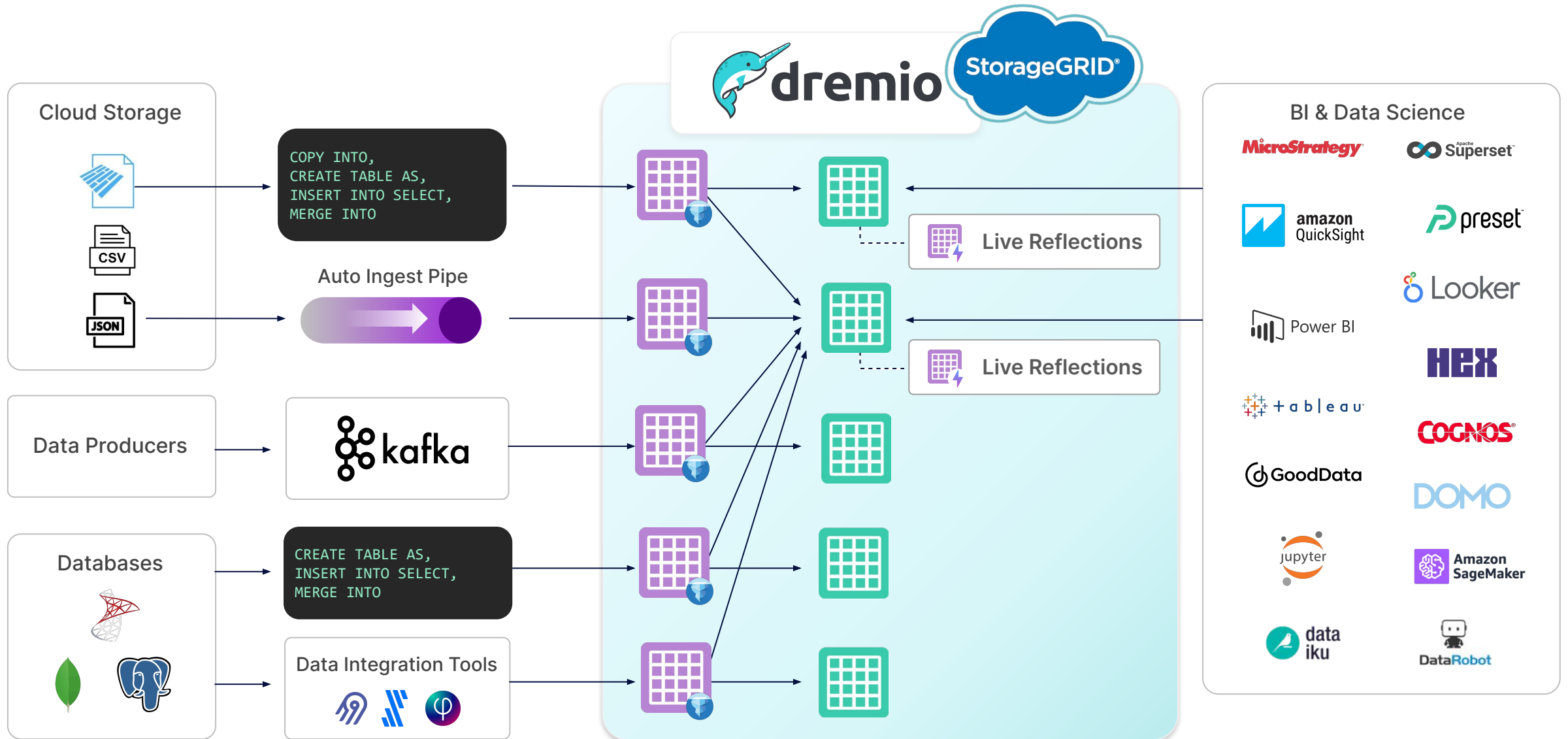
CLOUDERA



Why Dremio + NetApp



Iceberg Ingestion Options with Dremio



Dremio + NetApp StorageGRID Iceberg Lakehouse



Unified Analytics

Self-Service Analytics
Universal Semantic Layer
Governance & Security

SQL Query Engine

Price Performance Federation
Reflections Multi-Cloud &
Acceleration Hybrid

Lakehouse Management

Modern Data Catalog
Git for Data
Data Optimization



NetApp StorageGRID

ILM
Granular, Rule based
Hierarchical Storage
Management

Bucket Management
Versioning
Object Lock
Object Lifecycle
Federatio

Load Balancer with
HA
Advanced traffic QoS
Endpoint port
Multi-VLAN

Data Warehouse
performance at
Data Lake scale and
cost

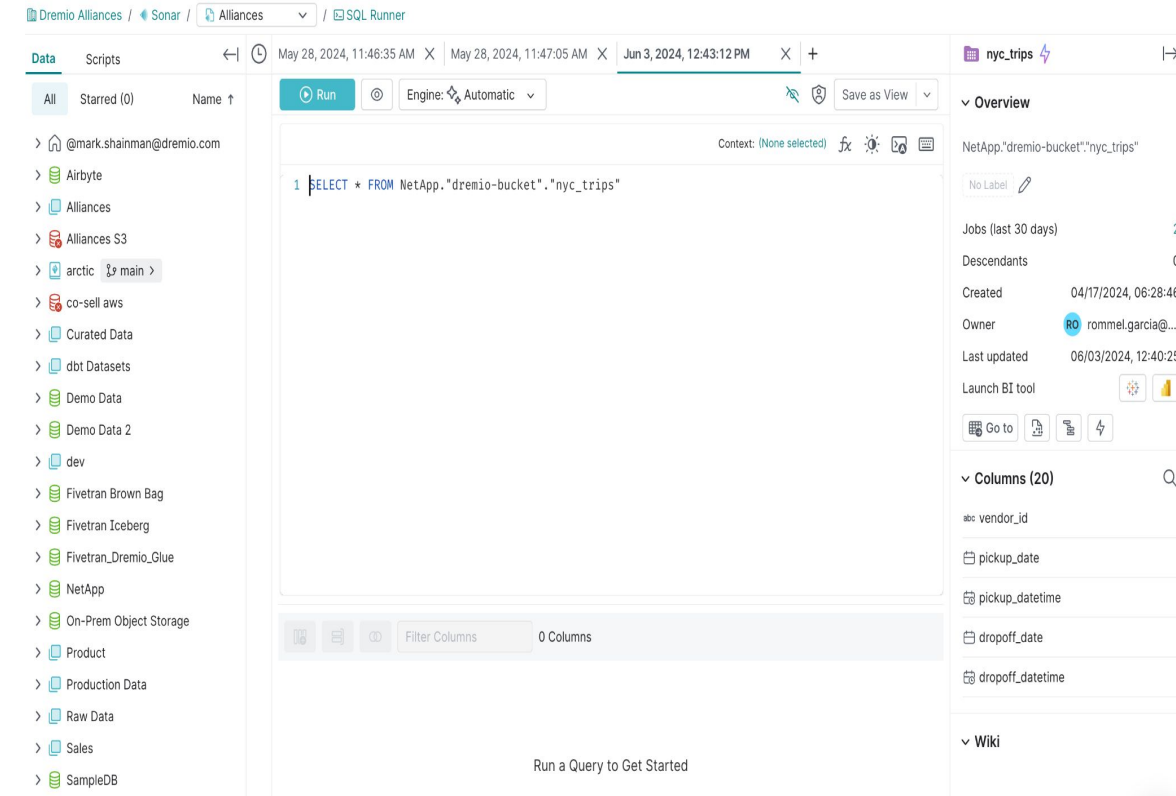
Multi-level encryption
KMS Integration

Dremio & StorageGRID Automate Data Management at Scale

Isolation, version control, governance & table optimization

Dremio Manages:

- User Access
 - Fine-grained privileges to control access to the data at the table, column and row level
- Version Control
 - Recover from any mistake by instantly undoing accidental data or metadata changes
- Table Optimization
 - Automatically compacts small files and group similar rows together
- Isolation
 - Enables experimentation with data without impacting other users



Benchmark Results: NetApp StorageGRID & Dremio win!

StorageGRID			
S3 Requests	Hive	Spark + Delta Lake	Dremio
GET (range read)	1,117,184	2,074,610	4,414,227
List Objects	312,053	24,158	240
HEAD (non-existent object)	156,027	12,103	192
HEAD (existent object)	982,126	922,732	1,845
Total Requests	2,567,390	3,033,603	4,416,504
99 Queries: Total Time (Minutes)	1084*	55	47

80% range read 2KB – 2MB from 32MB objects, 50 – 100 requests/sec

73% range read below 100KB from 32MB objects, 1000 – 1400 requests/sec

90% 1M byte range read from 256MB objects, 2000 – 2300 requests/sec

* Hive unable to complete query #72 in this test.



The NetApp & Dremio Difference

Best-in-Class TCO

>50%

Reduced analytics TCO in one year

Best Price-Performance

9x

Better Price-Performance than competitors

Fastest Query Performance

10x+

Faster query performance

Simplified data pipelines & Engineering

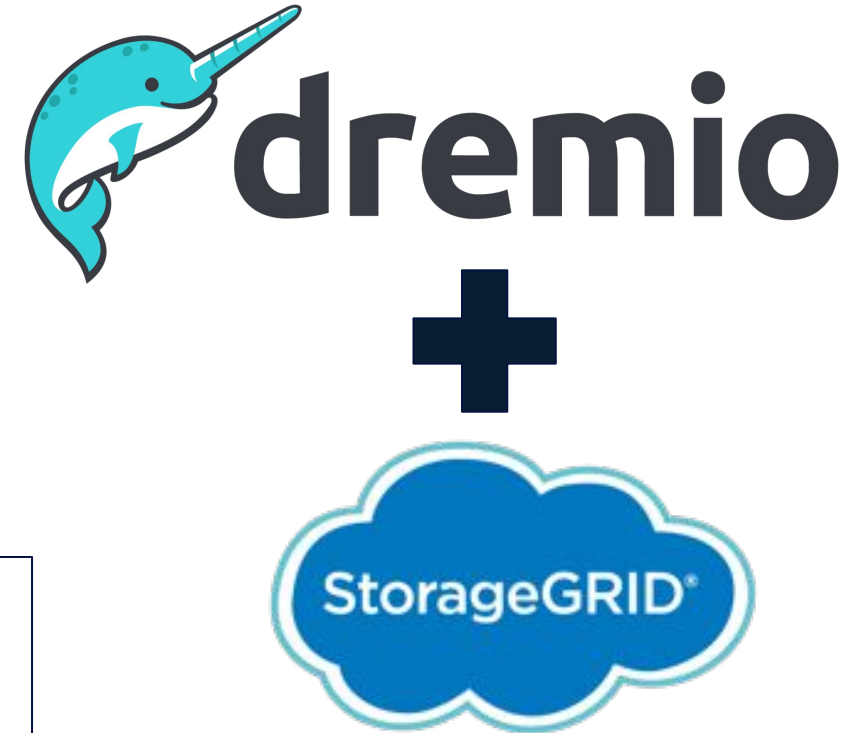
75%

ETL savings for a customer in one year

Dremio & NetApp: Hadoop Modernization

- Sub-second query performance & 10x better price/performance
- Governed self-service analytics
- Unified view & access across all of an organization's data
- Ability to scale compute and storage separately
- Improved overall data management
- Reduced complexity

Migrate from a legacy Hadoop data lake to a modern NetApp StorageGRID & Dremio lakehouse environment!



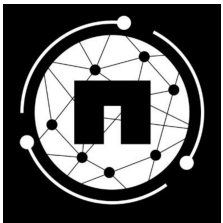
**Superior business insights &
faster time to value!**

Use Case: Hadoop Modernization for NetApp's Own Lakehouse!

Product 360: 95% faster time-to-insight while simplifying proactive customer care

Solution Overview

- Active IQ is NetApp's Digital Advisory solution for predictive maintenance and optimization



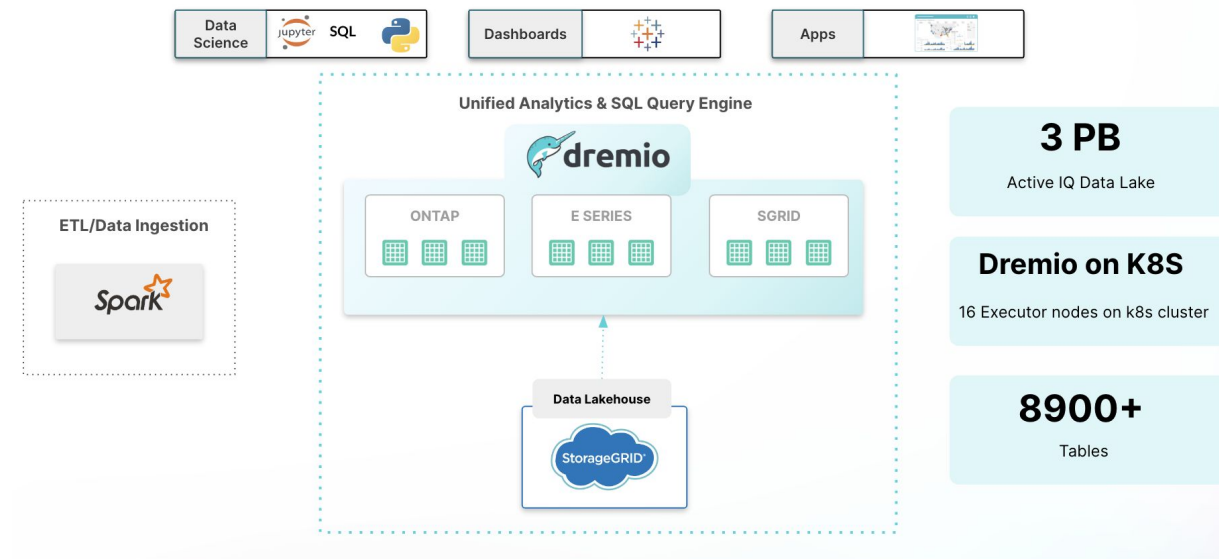
Pain Points

- Poor performance with legacy Cloudera/Hadoop technology
- Storage and compute tightly-coupled, leading to increased costs
- Scaling performance to meet workload demands was challenging: the average Hive query took more than 45 mins
- Dealing with costs of having to copy and move data between data sources

Why NetApp StorageGRID & Dremio

- Ability to query data directly in the datalake, reducing need to copy or move data
- Reuse existing investment of compute & storage, minimize changes to existing data pipelines
- Queries are 95% faster with Dremio & StorageGRID: 45 min down to 2 minutes!**
- 30+% analytics TCO savings in Year One!**

StorageGRID & Dremio Lakehouse Solution

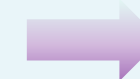


Results

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



Proactively manage customer care, optimize and identify problems before they happen.



Reduced risks of customer churn and higher product availability

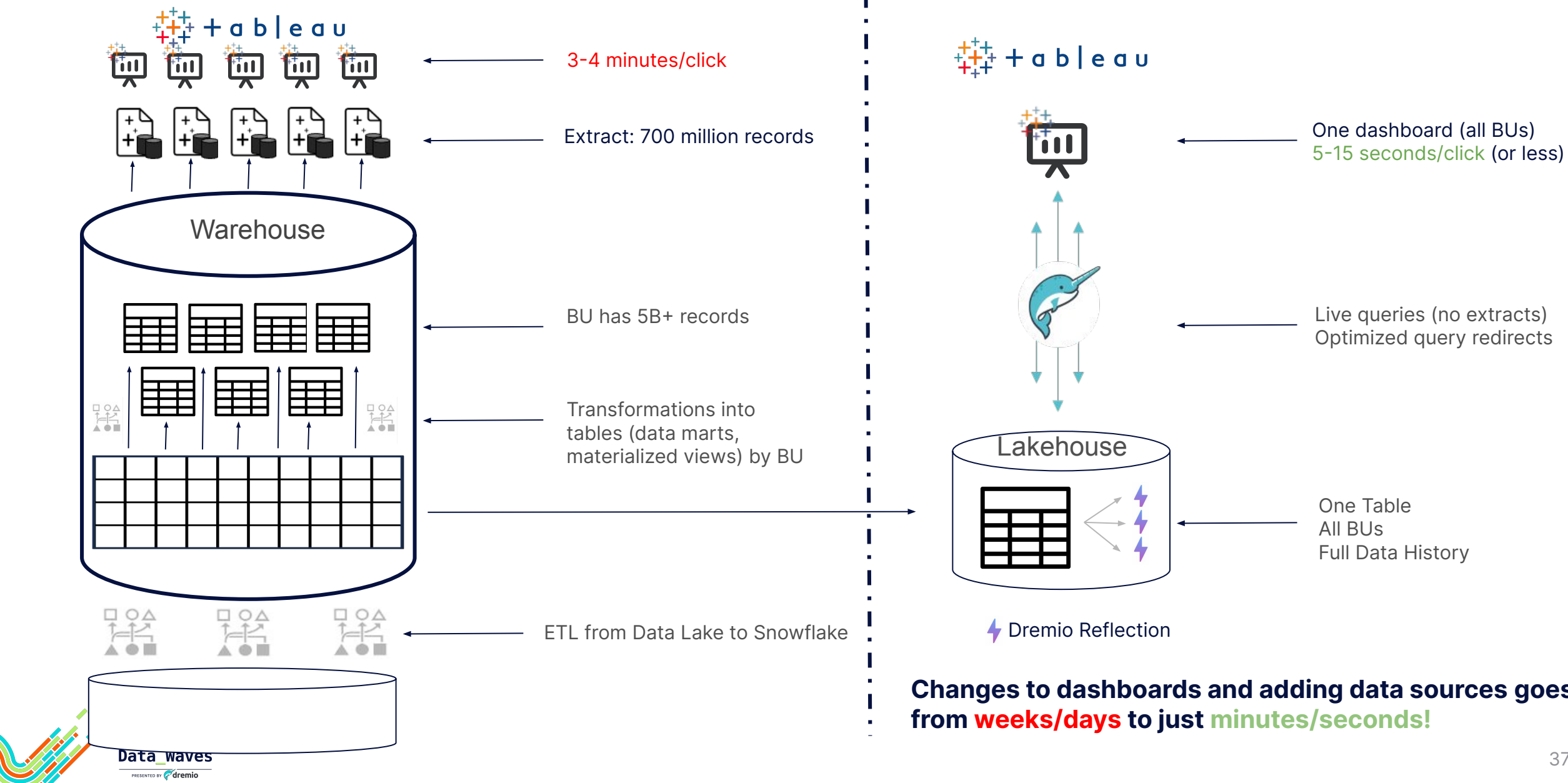
Use Case: Hadoop Modernization: Global 100 Financial Institution

- Existing Large, Legacy On-Prem Hadoop Data Lake
 - No plans to move it to the cloud
 - Concerns with regulations, security & PII
- Uses Dremio for their Financial Data Landing Zone
 - “Single source of truth” for all of their corporate treasury, liquidity, etc.
 - Able to join and query data from NetApp StorageGRID and existing Hadoop for high-performance BI & reporting
 - With Dremio semantic layer, they’re able to provide a self-service analytics layer and make cross-functional data easily-discoverable
 - Dremio makes the audit trail of data easier for mandatory reporting to the FTC
- Long-Term NetApp Plans
 - Reduce Cloudera Data Platform footprint and move off of DAS
 - As data moves off of DAS, it will go to NetApp StorageGRID
 - Majority of Treasury group’s new data moving to NetApp StorageGRID
- Long-Term Dremio Plans
 - Adding cloud data sources & relational database data sources with their on-prem data
 - Continue to expand existing on-prem footprint



Use Case: Warehouse to Lakehouse

Fortune 10 Customer: 75% TCO Savings, \$3M Savings In Just One Dept



The NetApp & Dremio Difference

**Best-in-Class
TCO**

**Best
Price-Performance**

**Fastest
Query Performance**

**Simplified Data
Engineering &
Management**

Get More Great Information



[NetApp Case Study](#)



[Dremio & Storage Grid Solution](#)



[Five Factors to Consider When Migrating from Hadoop to the Data Lakehouse](#)



Thank You