RESEARCH NOTE

Analyst+Consultant Advisory Network | MAY 2024

UNIFIED ANALYTICS WITH OPEN DATA

Dremio: An Open Approach to a Next Generation Lakehouse

RICHARD WINTER

More than 30 years after the invention of the data warehouse, and several years into the era of cloud data warehousing, many customer organizations still struggle with the fundamental challenges of managing data for analytic use.

People say it takes too long to bring new data online, that they find it too difficult to get answers to straightforward business questions, or that they have moved to the cloud and are deeply distressed by the size of the monthly cloud bill.

Dremio's answer to these issues is a new approach to data: a second-generation lakehouse based on open data formats, a massively parallel query engine, and patented technology to make data much more easily used and managed. Dremio's solution is available in the cloud, on prem and in hybrid configurations. It provides unified access to data in Iceberg format in object storage and to data in a wide range of other repositories and architectures.

Dremio says its approach provides customers with advantages in ease of use, time to value, speed of access and cost of operation. Dremio's solution is in use by blue-chip customers, including some operating at large scale. Its claims of cost and performance advantage are backed by benchmark results and customer studies.

My opinion: Customers open to an alternative or a supplement to their current warehouse or lakehouse platform ought to take a look at Dremio.

Dremio: The Next Generation of Data Lakehouse

The data lake emerged about ten years ago primarily as a way to manage "big data" and the ever-growing requirements to more rapidly analyze proliferating corporate business data. But, as implemented by most customers, the data lake concept fell short when it came to data governance, ease of use, time to value, and a range of other modern analytic needs. Further, many customers have data both on prem and in the cloud, presenting a challenge to cloud-only architectures.

The data lakehouse seeks to address those issues by combining the best features of the data warehouse and the data lake.

Now Dremio is on the market with a next-generation data lakehouse approach that uses open data and advanced technology to overcome limitations of earlier efforts. Dremio says its data lakehouse is much easier to use, providing higher performance and creating a more agile entity than similar systems customers have been using.

Dremio aims to deliver on more of the advantages of the data warehouse, while still storing data in open table formats and providing the flexibility and low cost of earlier data lake products.

In addition, Dremio provides novel capabilities for experimenting with and managing data to shorten cycle times and make data more readily consumable by end users. In short, Dremio has rethought enterprise data and analytics in a modern framework that applies equally well in the cloud, on prem and in hybrid environments.

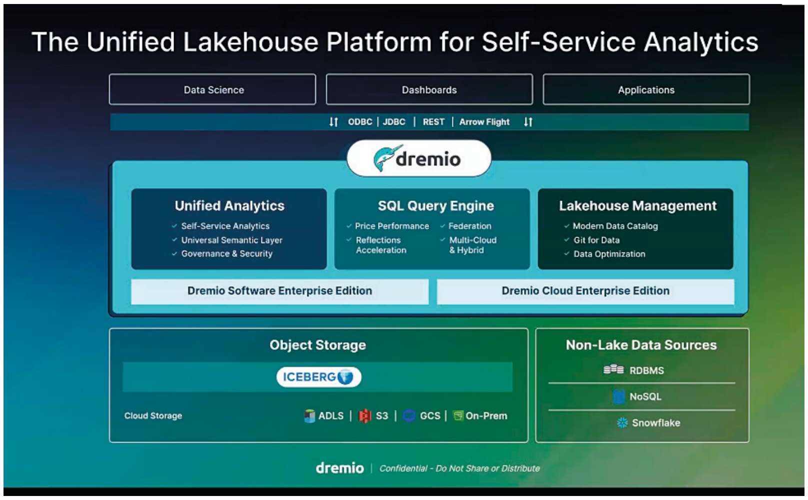
This report provides an introduction to Dremio and comments on its significance following a discussion of the product with twelve senior independent experts in a recent ACAN private forum with Dremio product and marketing leadership.

The Dremio Product Concept

Dremio provides a next generation approach to the data lakehouse that includes:

Optimized Support for Iceberg, the Widely Popular New Open Data Format. Data is stored in the Iceberg open source format where it can be accessed and updated by Dremio as well as by other query engines, such as Spark, Flink and an increasing number of data warehouse query engines. Dremio's SQL query engine, built on Apache Arrow, is optimized to take full advantage of Iceberg data while maintaining ACID properties, meaning that transactional integrity is preserved as data is changed.

There are four key implications of storing data in Iceberg format in object storage: (a) no vendor lockin; (b) lower cost than storing data ingested into a data warehouse engine; (c) concurrent access by other query engines to enable specialized techniques when applicable; and, (d) simplified data pipelines, because data does not have to be moved or ingested once delivered to the object store.



The Unified Lakehouse Platform for Self-Service Analytics

Data Science | Dashboards | Applications

ODBC | JDBC | REST | Arrow Flight

dremio

Unified Analytics
- ✓ Self-Service Analytics
- ✓ Universal Semantic Layer
- ✓ Governance & Security

SQL Query Engine
- ✓ Price Performance
- ✓ Reflections Acceleration
- ✓ Federation
- ✓ Multi-Cloud & Hybrid

Lakehouse Management
- ✓ Modern Data Catalog
- ✓ Git for Data
- ✓ Data Optimization

Dremio Software Enterprise Edition | Dremio Cloud Enterprise Edition

Object Storage

ICEBERG

Cloud Storage: ADLS | S3 | GCS | On-Prem

Non-Lake Data Sources

RDBMS

NoSQL

Snowflake

dremio | *Confidential - Do Not Share or Distribute*

Figure 1: The Dremio Unified Lakehouse Platform (Source: Dremio)

High Performance, Massively Parallel Query Engine. Dremio says its query engine, Arrow, delivers high performance with massively parallel architecture that supports scaling up, scaling out and federated query.

Dremio asserts major performance advantages, backed by benchmark data, when the Dremio SQL engine is compared to leading cloud query engines. The Dremio SQLengine features a high performance NVMe-based data cache and supports *Reflections*, a patented query acceleration technology invented by Dremio, which acts as a transparent materialized view over data that may be distributed over a large and diverse data ecosystem. Reflections can rewrite queries at run time,

ACAN Research Note Methodology

PURPOSE AND METHODOLOGY FOR THIS REPORT

This **ACAN Research Note** describes **Dremio** and its significance for data management and analytics. In developing this report, the author drew on his own independent research and experience, moderated the ACAN expert forum on Dremio, interviewed Dremio leadership and analyzed Dremio documentation and literature. Dremio was provided an opportunity to comment on the paper with respect to facts in its capacity as the sponsor of this research. The author has final editorial control over the content of this publication and is solely responsible for any opinions expressed.

eliminating costly operations, including joins, aggregations and data movement among distributed data stores.

Semantic Layer for Self-Service. A semantic layer is a business-friendly presentation of data designed for consumption by end users. Dremio has built in facilities to create and maintain a semantic layer of data that is designed to make data easier to find, easier to use and faster to deliver. End users interact with data through the Dremio UI, which supports SQL, low code drag and drop, and natural language text to SQL. Dremio provides a data catalog, based on the Nessie open source standard, for documenting the business meaning and lineage of data. GenAI enablement of the data tagging and wiki descriptions are supported.

Data Versioning. Another Dremio invention, data versioning, allows users to create their own virtual versions of data objects (e.g., without physical data copy or movement) for experimentation, testing, time travel, model validation and other uses. This is uniquely helpful in experimenting with classic machine learning models, explaining and documenting decisions, traceability, and other common problems in data and analytics.

Diverse Deployment Models. Dremio is available on prem, in the major public clouds and in hybrid and multi-cloud configurations. In the cloud, Dremio is available in a customer-managed arrangement and as a fully managed cloud service on AWS and Azure. Free versions are available for experimentation, evaluation and production use. Dremio reports that thousands of customers use the free version to support business activities.

Customers

Dremio cites blue-chip customers as well as small- and medium-sized businesses among its substantial base of users. In addition to the customer examples on its website, Dremio can provide data concerning major customers who have realized large savings and other benefits by migrating from popular, widely used data warehouses and data lakehouses.

Recommendation

Dremio provides a next-generation implementation of the data lakehouse, featuring many of the advantages of a data warehouse. Dremio operates on data in open formats and aims to accelerate time to value while delivering data to end users that is business friendly and readily consumed.

As with any analytic data platform, a decision to adopt Dremio should be based on a careful evaluation driven by the customer's specific requirements. Before relying on cost savings or performance advantages, I recommend that customers conduct realistic benchmarks at scale to validate that they will actually realize the benefits they seek.

With that approach in mind, customers who want to store and manage some or all of their enterprise data in open formats such as Iceberg — or who want to explore lower cost alternatives to popular data warehouse or data lakehouse platforms — should to take a look at Dremio. +++

acadvisorynetwork.com

About ACAN and the Author

ACAN is a network of independent analysts and consultants in data, analytics + AI.

The network holds private forums for vendors with significant innovations in these areas. In the forums, designed for an expert audience, ACAN members analyze major developments and provide feedback and guidance to the sponsoring vendor. In addition, ACAN produces podcasts and other publications resulting from the research and hands-on consulting work of its members.

Richard Winter, the author of this report, is an independent consultant in analytic data management at scale. The Founder and CEO of WinterCorp since 1992, he has led data warehouse evaluations and benchmarks for more than 50 leading enterprises, architecting solutions to some of the largest scale and most demanding data requirements in business and government.