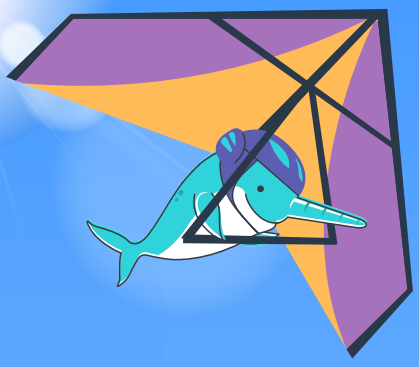
dremio

dremio.com

Get Started



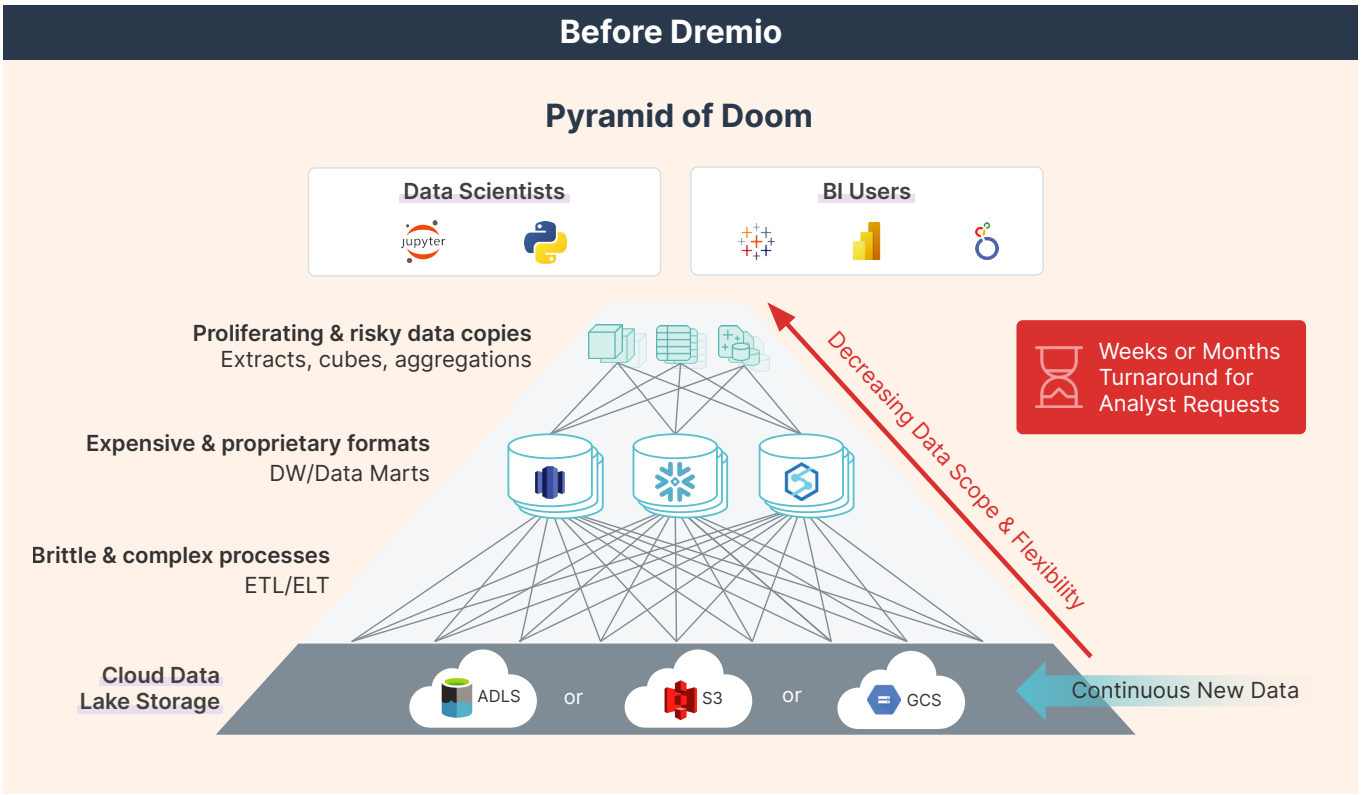
# Dremio Cloud Service

## Executive Summary

Data architectures today are required to rapidly adapt to the enormous zettabytes of data created by modern-day users and applications. Not only is the data growing exponentially at an unprecedented rate, but the user requests for data are also growing very fast. Everybody wants access to the data and wants to query it themselves. Executives and IT leaders want to empower their business users to have timely and self-service access to organizational data, so that they can support new initiatives and evolving business needs.

Moreover, nowadays since it has become incredibly cheap and easy to store near-infinite amounts of data in cloud data lake storage like AWS S3, Azure ADLS and Google Cloud Storage (GCS), companies end up dumping a lot of data in their storage. And everyone is well aware of the painful and time-consuming process required to set up *and* maintain complex ETL/ELT processes to copy multiple subsets of that data into a data warehouse for different use cases and reporting purposes. On top of these copies of data, data analysts and BI users end up creating additional copies of data for personalized needs, aggregated analysis, reporting and dashboards, etc. This ends up in a very complicated system and multiple data copies.

Data consumers such as BI and analytics users and data scientists cannot access this data themselves and end up relying on data teams, who in turn get overwhelmed with weeks or months of data request backlogs. This leads to a situation we call the "Pyramid of Doom":




Organizations of all sizes know that data is one of the key enablers to increase and sustain innovation, and drive value for their customers as well as their own business. As companies look to make business decisions driven by data, they know that they can be agile and productive by modernizing traditional data platforms with cloud-native technologies that are highly scalable, feature-rich and cost-effective.

As more and more organizations migrate to the cloud, or aspire to optimize their existing investments in the cloud, decision-makers are trying to break free from the intrinsic dependencies on brittle, proprietary data warehouse vendors and liberate their organizations and budgets from the lock-in and economic extortion of proprietary data vendors.

Leaders are looking for solutions that provide them the flexibility, cost-effectiveness and scalability of data lakes as well as the performance and data management capabilities that traditional data warehouses have provided.

## Data Teams Across Companies Face the Same Challenges.

Data teams in most companies face a common problem: they are being asked to do the impossible — to liberate data and put it in the hands of as many people as possible and to quickly respond to continuous and ever-changing requests from the business within the confines of stringent security and governance requirements. Furthermore, the amount of data and expectations for self-service access to that data is growing exponentially. Unfortunately, most IT budgets today remain flat or are even shrinking, so data teams are finding it difficult to meet these requirements.

> We used to have to wait 3-6 weeks for even the smallest changes to our BI dashboards because our engineers were so backlogged with data requests. We simply couldn't afford to wait that long to make critical business decisions.
>
> **- Multinational Technology Company**

## Data Trends Over the Years

With the cost of data storage decreasing and the amount of data generated by modern users and applications increasing exponentially, massive amounts of data have been dumped into data lakes without the accompanying value of delivering real insight to the business. On top of this, the limitations of traditional data warehouses have forced companies to create complicated webs of ETL pipelines and multiple copies of data at different levels, to meet the requirements for different users and use cases — leading to data drift, high infrastructure costs, irreparable inefficiencies and dependencies on proprietary data warehouse vendors.

In response, companies are moving away from the constraints and limitations of traditional, proprietary data warehouses and data marts, and are heading toward data lakes and lakehouses and open data architectures that enable them to be future ready for new innovations and technologies.

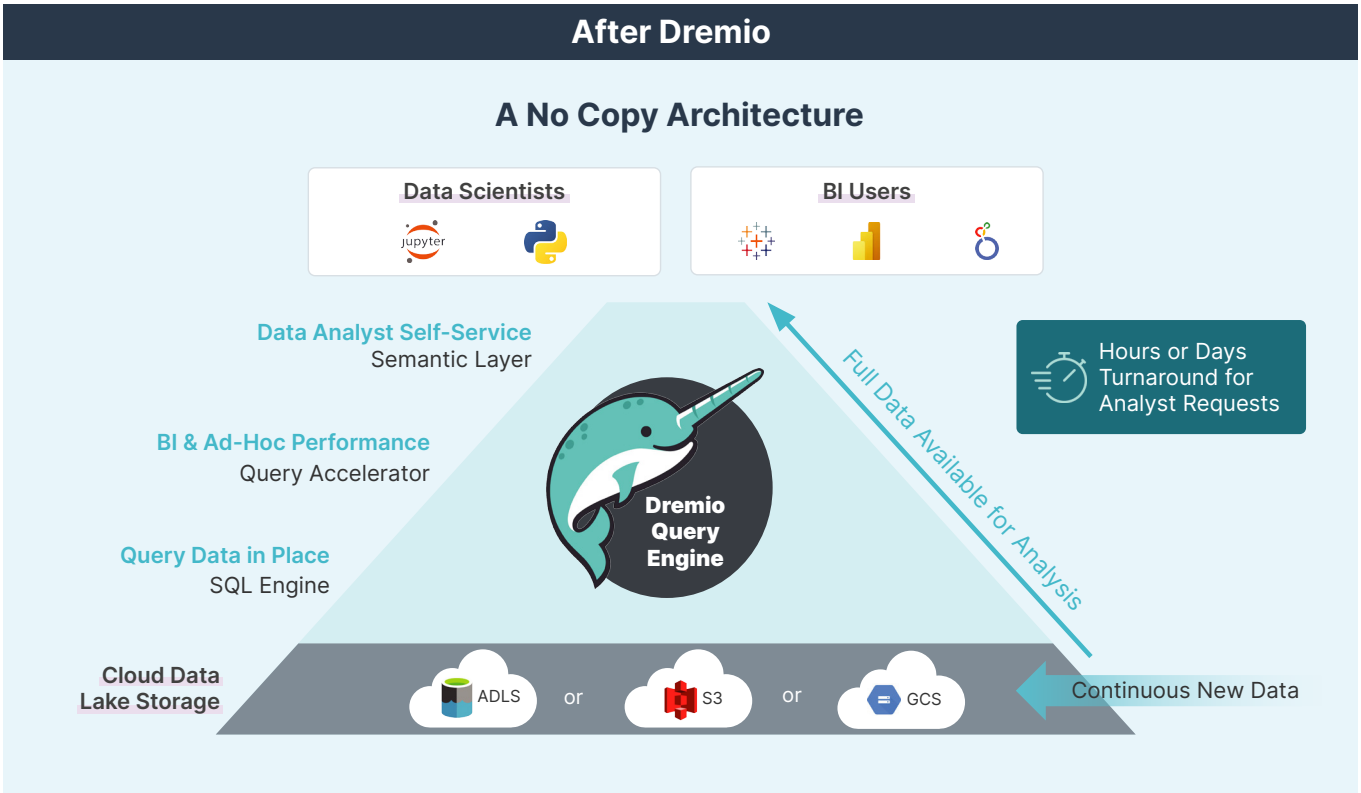
Worldwide organizations across all industry verticals leverage Dremio — a SQL lakehouse platform that is built to automatically handle any scale with consistent query performance — to achieve data democratization, lightning-fast analytics and dashboards and self-service data access, while only paying for what they truly need.

**According to a March 2021 IDC study, in 2020, 64.2 ZB of data was created or replicated and the amount of digital data created over the next five years will be greater than twice the amount of data created since the advent of digital storage.**

# The Dremio SQL Lakehouse Platform

Dremio is a high-performance **SQL lakehouse platform** built on an **open data architecture**.

Dremio's SQL lakehouse platform simplifies data engineering and eliminates the need to copy and move data to proprietary data warehouses or create cubes, aggregation tables and BI extracts, providing flexibility and control for data architects and data engineers, and self-service for data consumers. It enables high-performing BI dashboards and interactive analytics directly on data lake storage through seamless integrations with Tableau, Power BI and other BI tools.




Unlike traditional data warehouse vendors, which require companies to load data into their vendor-owned storage as a prerequisite to using their SaaS offering, Dremio Cloud allows companies to keep the data in their own cloud account. This not only improves security and data governance, but also eliminates vendor lock-in, and allows companies to be present and future-ready and use other best-of-breed compute engines on that data.

Dremio provides the next-generation cloud data lakehouse architecture that brings together the best of the data warehouse and the data lake, eliminates multiple data copies, reduces cost, promotes simplicity and boosts efficiency with a "no-copy" strategy and enables companies to enjoy high-performing dashboards and interactive analytics directly on the data lakehouse, with enterprise-grade security and data governance.

Dremio enables **high-performance BI dashboards and lightning-fast interactive analytics** directly on data lake storage through **seamless integrations** with Tableau, Power BI and other BI tools, through the following industry-defining **Dremio innovations**:

**I.** Dremio's **shared semantic layer** provides **consistent semantics** across all users, applications and tools, and empowers **self-service data access** for data analysts and data scientists while **centralizing security and governance**. Dremio's semantic layer empowers data analysts and data scientists to discover, curate, analyze and share datasets in a self-service manner.

The inflexibility of traditional data warehouses have historically forced users to define isolated definitions of data, calculated field and virtual datasets in individual BI tools, which cannot be accessed by other users, BI tools and apps.

**Why define isolated definitions of data, calculated fields, and virtual datasets in individual BI tools where they can't be used by your teammates?**

**Dremio's semantic layer to create dataset definitions, calculated fields, and security rules that can be leveraged by any downstream application.**

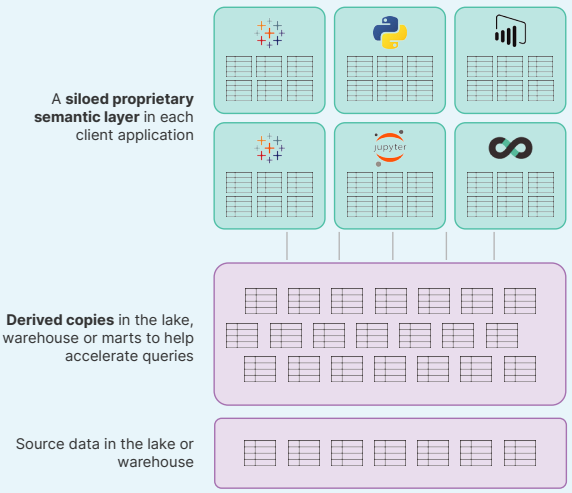
**A single, consistent, secure view of data:** In contrast to traditional data warehouses, you can use Dremio's semantic layer to create dataset definitions, calculated fields and security rules that can be leveraged by any downstream applications. Centralized, consistent data semantics eliminates data sprawl and inconsistent, siloed definitions of data across your company.

**Transparent acceleration:** We do not expect you to know and remember which view of data you should connect your dashboards to in order to get the performance you need! With Dremio's semantic layer, you don't have to worry about connecting to various materialized views in order to make your dashboards run fast. Simply create dashboards and reports on the tables you want, and the Dremio query optimizer will work behind the scenes to accelerate your queries.

**Data lineage:** The relationships between your data sources, virtual datasets and all your queries are maintained in Dremio's data graph, telling you exactly where each dataset came from.

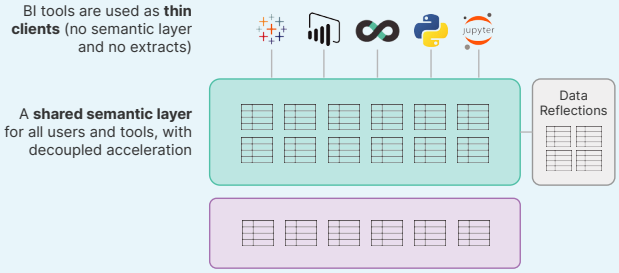
## Shared Semantic Layer for All Users & Tools

### Without Dremio




### With Dremio

- ✓ Consistent semantics across all client applications
- ✓ Centralized security & governance
- ✓ Fast time to value



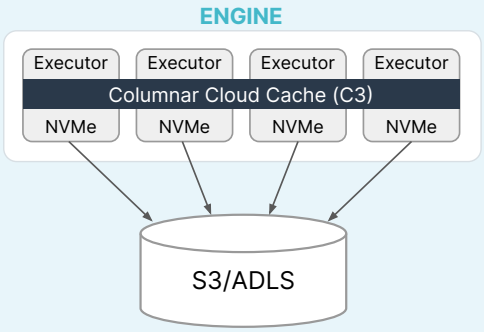
II. Dremio is the only SQL lakehouse platform that has been purposefully designed and built its **SQL engine** from the **ground up** to enable customers with high-performing BI and interactive dashboards and interactive analytics directly on data lake storage by thoughtfully innovating at all stages of query execution. Everything from query planning and caching technologies to generating hardware-optimized machine code has been thoughtfully designed and built from the ground up to optimize queries and analytics for customers.

The following **innovations within Dremio's query engine** make Dremio the industry's fastest SQL lakehouse platform on cloud data storage:

**Columnar Cloud Cache (C3)** enables customers to accelerate reads from cloud data lake storage such as S3, ADLS and GCS via distributed NVMe cache. Dremio's Predictive Pipelining technology fetches data just before the execution engine needs it, dramatically reducing the time the engine spends waiting for data. And our real-time C3 automatically caches data on local NVMe as it's being accessed, enabling NVMe-level performance on your data lake storage.

## Columnar Cloud Cache (C3)

- NVMe-level I/O performance on S3/ADLS/GCS
- Eliminate S3/ADLS I/O costs (10-15% of cost per query)
- Use existing NVMe/SSD on EC2 instances & Azure VMs
- Transparent to analysts and engineers




C3 enables Dremio to achieve NVMe-level I/O performance on S3/ADLS/GCS by leveraging the NVMe/SSD built into cloud compute instances, like Amazon EC2 and Azure Virtual Machines.

C3 only caches data required to satisfy your workloads and can even cache individual microblocks within

> Made us rethink our whole architecture. Intuitive, effective and transformative.
>
> **- Systems Architect, Large Enterprise**

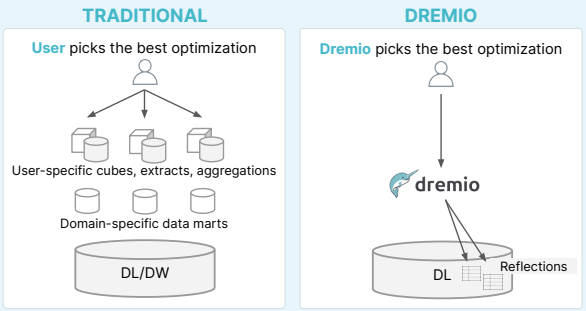
datasets. If your table has 1,000 columns and you only query a subset of those columns and filter for data within a certain timeframe, then C3 will just cache that portion of your table.

You are enabled to not just run lightning-fast queries, but you save 10-15% of your infrastructure costs for running each query! By selectively caching data, C3 eliminates over 90% of S3/ADLS/GCS I/O costs, which can make up 10-15% of the costs for each query you run.

**Data Reflections** are data structures that intelligently precompute aggregations and other operations on data, so you don't have to do complex aggregations and drilldowns on the fly.

## Data Reflections

- Enable low-latency (including sub-second) BI queries
- Eliminate cubes and BI extracts
- Reduce infrastructure costs by up to 100x
- Persisted on S3/ADLS as Parquet/Iceberg tables
- Transparent to analysts (advanced query plan rewrites)




**Reflections provide 100x speedup of BI workloads by materializing aggregations on S3 and ADLS**

Reflections organize and optimize the data close to Dremio's query execution engine, using techniques such

as columnarization, compression, sorting, partitioning and aggregating data.

Reflections are completely transparent to end users. Users do not have to worry about connecting to a

Dremio automatically applies these filters on joins without any user involvement and provides up to 100x improved performance when working with traditional star or snowflake schemas.

> FactSet enables tens of thousands of investment professionals around the world with the data and analytics they need to make crucial decisions. Dremio helped us modernize our existing data infrastructure, achieving dramatic SQL query acceleration, with its reflections capability on top of AWS data lake. Furthermore, Dremio's single pane of glass for analytical insights and graph data lineage eliminated the complexity of joining data silos, and shed light on the origins of data sources.
>
> **- Wilson Tsai, VP, Director – Data Platform Infrastructure & Services, FactSet**

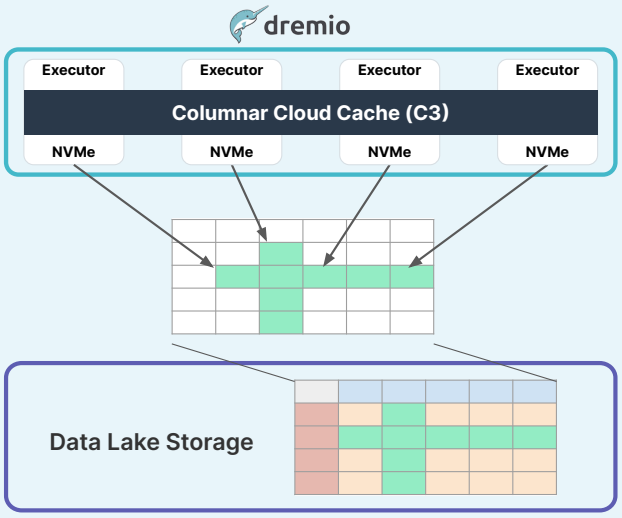
specific materialization. Data analysts query tables and views, and the Dremio optimizer picks the best Reflections to satisfy and accelerate the query.

Aside from simplicity for data analysts, Reflections are also incredibly easy to create and maintain! You can use a UI or REST API to administer Reflections, instead of having to write complicated SQL statements to define materialized views and refresh rules.

A **Cost-Based Optimizer** picks the fastest path to complete your query by understanding deep statistics about the data you want to query, including location, cardinality and distribution. It uses that data to accurately predict how much data will flow through the query's operators so that it can choose the best plan. It also takes into account the Reflections in the system, and rewrites the query plan to use them.

Data Reflections, which are similar to indexes in relational databases, provide Dremio's cost-based query optimizer useful information that can result in dramatically more cost-effective query plans than performing query push-downs to the data source.

**Granular Pruning**: Runtime filtering enables Dremio to dynamically apply filters from a smaller joined table to a larger table to enhance filtering on larger tables.

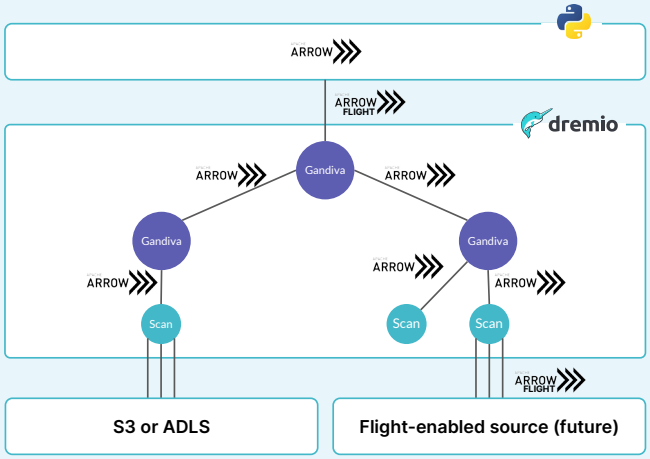



**Apache Arrow Gandiva** provides customers with just-in-time (JIT) compilation and runtime code generation. Dremio is a columnar engine powered by Apache Arrow, the open source standard for columnar, in-memory computing (which we co-created!).

Dremio leverages Gandiva, an LLVM-based library for runtime code generation, to create machine code that efficiently evaluates arbitrary expressions on batches of columnar Arrow data, rather than row-based execution. This not only improves resource utilization, but also provides faster, lower-cost operations of analytical workloads.

Generally speaking, Gandiva provides performance advantages across the board with very few compromises. First, it reduces the time to compile most queries to less than 10 milliseconds. In Dremio, Gandiva also improves the performance of creating and maintaining Data Reflections, which can be used by Dremio's query planner to create a more intelligent query plan that involves less work and hence accelerates queries by orders of magnitude.

Gandiva maximizes CPU utilization and leverages optimizations like vectorized processing and SIMD execution to make your queries fly!




**Apache Arrow and Arrow Flight enable 2-5x faster processing and 20-100x faster result transfer**

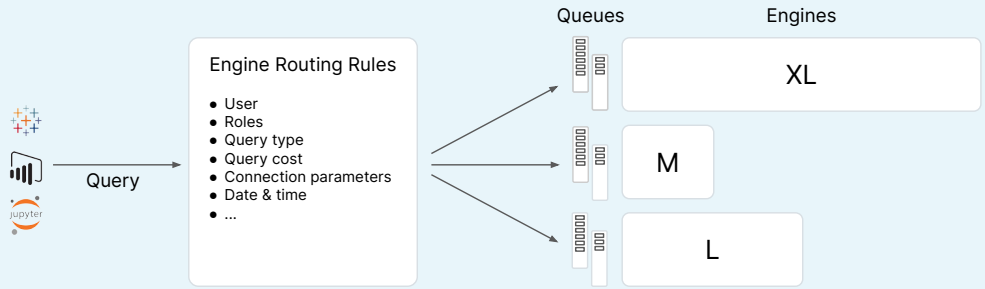
**Apache Arrow** was created by Dremio and, with over 20 million downloads a month, is used by almost every data scientist in the world! Apache Arrow is Dremio's internal memory format, and it's also the standard for Python and R developers. **Arrow Flight** is a modern, open source RPC framework that was co-created by Dremio to enable ultra-fast data transfer between Arrow-enabled systems.

Flight eliminates serialization and deserialization, enables parallelism, and avoids the need for proprietary client-side drivers. The result: 20-100x faster access to query results compared to traditional JDBC and ODBC interfaces.

**Multi-engine architecture and workload management**: Dremio features a multi-engine architecture, so you can create multiple right-sized, physically isolated engines for various workloads in your organization. This avoids the "noisy neighbor problem." You can easily set up workload management rules to route queries to the engines you define, so you'll never have to worry again about complex data science workloads preventing an executive's dashboard from loading.

Aside from eliminating resource contention, engines can quickly resize to tackle workloads of any concurrency and throughput, and auto-stop when you're not running queries.

## Multi-Engine Architecture




**60%**
**LOWER COMPUTE COSTS**

Auto-stop/start and right-sized engines eliminate the need to over-provision infrastructure.

**0**
**NOISY NEIGHBOR CONCERNS**

Workloads are physically separated so one workload can't impact the performance of another workload.

**100%**
**CONTROL OF RESOURCES**

Control resource allocation with policies such as query priority, max query cost, max queue time, max runtime, etc.

The elastic, multi-engine architecture scales infinitely and helps manage workloads with the highest concurrency and lowest latency.

You can eliminate under- and over-provisioning of compute resources, eliminate workload contention and further slash cloud costs by 60% or more with elastic engines. Configure any number of query engines, each one not only sized and tailored to the workload it supports, but equipped with elastic, on-demand scale.

> With its strong query performance and semantic layer capabilities, Dremio is the perfect backbone for our Henkel data lake.
>
> **- Thomas Zeutschler, Director of Data and Application Foundation at Henkel**

**III.** Dremio is **designed from the ground up** to deliver industry-leading, enterprise-grade security and is purpose-built to provide the industry's highest levels of security with multiple layers of protection.

Dremio provides end-to-end data security to customers. Your data and data processing stays in your cloud account, where data is encrypted at rest and in transit. Easily safeguard your data further with dynamic data masking and support for customer-managed encryption keys.

In Dremio's security model, customers can implement fine-grained, database-style permissions across all Dremio objects with role-based access control (RBAC). Use ANSI standard SQL to manage users, grant/revoke privileges, define custom roles and more.

Dremio's user security is easily extensible, so you can sync existing roles and users in enterprise identity providers like Azure Active Directory, Okta and Ping via SCIM. In addition, Dremio integrates with tools like Okera and Privacera, so you can apply any existing access policies in those tools to Dremio.

Dremio provides seamless authentication by supporting OAuth 2.0 on all interfaces, including programmatic interfaces (ODBC, JDBC, REST API and Arrow Flight), and native single sign-on (SSO) integrations with Tableau and Power BI.

Dremio also supports generating access tokens via custom processes, so you can adhere to your corporate security guidelines.

To top it all, Dremio's consistent, centralized shared semantic layer facilitates centralized security and governance over your data lake, instead of having to define individual policies across multiple tools.

## Dremio Cloud

Dremio is a **fully managed SQL lakehouse platform** built on an **open architecture**. It simplifies data engineering by eliminating the need to copy data into proprietary data warehouses or create BI extracts, cubes and other copies. It enables **high-performing BI dashboards** and **interactive analytics** directly on data lake storage through **seamless integration** with Tableau, Power BI and other BI tools. Dremio Cloud is an infinitely scalable, frictionless service across multi-cloud.

Dremio Cloud's frictionless user experience starts on **day one**: you can start in seconds just with your email address. We also offer free in-product and online education to help you maximize your experience with Dremio.

Dremio launched Dremio Cloud in July 2021, since a large portion of our customers already have their data in cloud data lake storage. However, more companies than ever are looking for fully managed services that enable them to focus on deriving value from data instead of worrying about system setup and administration. Also, for the fifth year in a row, optimizing the existing use of cloud (cost savings) is the top initiative for the year for over 60% of organizations of all sizes.

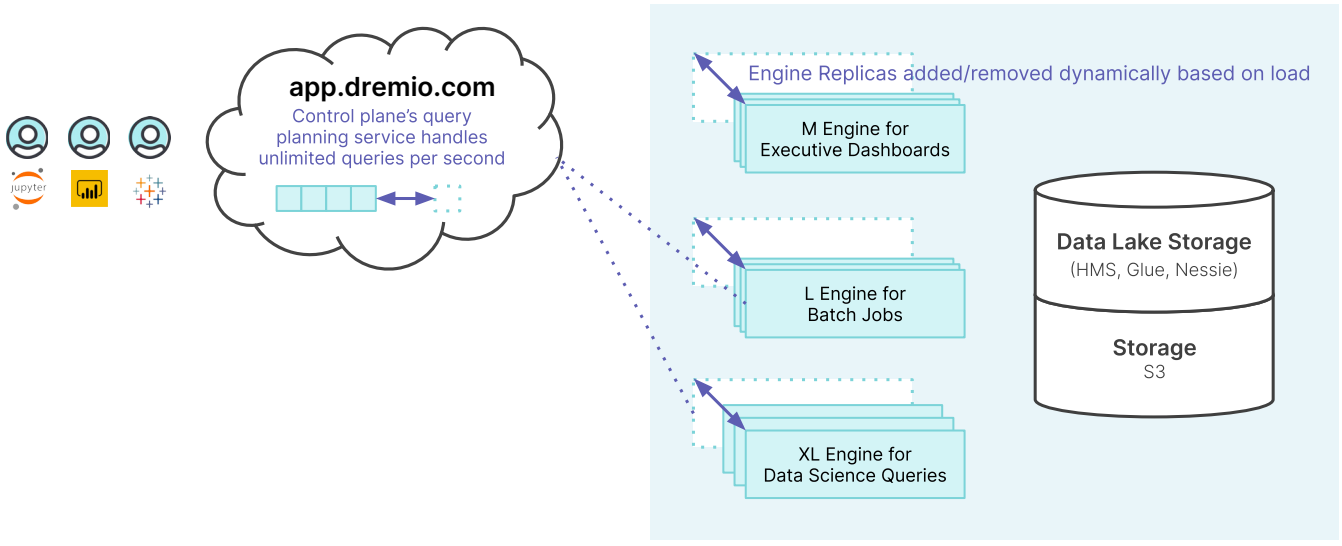
Dremio Cloud is a **fully managed** Dremio service! You do not have to install any software or manage and monitor any infrastructure, config files or log files or worry about complex sizing exercises. You bring the data, your analysts bring the queries, we take care of the rest!

Dremio Cloud reimagines the data lake by combining the best of traditional data warehouses and data lakes into a SQL lakehouse, while removing the limitations of traditional data warehouses resulting from closed data architectures.

Dremio Cloud is an **infinitely scalable service** that enables **high-performance SQL workloads directly on cloud data storage**, eliminating the cost and complexity of copying and moving data. Dremio Cloud hence **eliminates multiple data copies**, reduces cost, promotes simplicity and boosts efficiency with a "no-copy" strategy and enables companies to enjoy high-performing dashboards and interactive analytics directly on the data lakehouse, with enterprise-grade security and data governance.

At its core, **Dremio Cloud** consists of two main components:

1. A multi-region **control plane** that receives queries from clients and is responsible for query planning and engine management:
    - The control plane is an always-on, multi-tenant service which resides in a Dremio-managed VPC, and is designed to handle an unlimited number of queries per second
    - The control plane is central to the Dremio Cloud customer experience and includes all client-facing interfaces, including the UI and REST API, as well as the JDBC, ODBC and Arrow Flight query endpoints
    - When a user wants to use any client tool with Dremio Cloud, such as Tableau, Power BI, Looker, SageMaker or a Jupyter Notebook, they simply connect it to the control plane, and the control plane routes queries accordingly
    - The control plane securely delegates query execution to the compute engines




2. A **data plane** comprising compute engines that are responsible for query execution:
    - The query execution happens in the compute engines, which are automatically provisioned in the customer's AWS VPCs, so all data processing happens within the customer's account
    - This is in stark contrast to cloud data warehouses, where everything happens in the vendor's account instead of the customer's own account, thereby requiring data to be copied and moved into the vendor's account — creating the problem of multiple data copies

Only the control plane resides in a Dremio-managed cloud account, while compute engines reside in the customer's account alongside your data. Assigning ownership this way ensures all data storage and processing occurs within the **customer's account**, while maintaining a seamless SaaS experience that removes the burden of deploying, configuring, managing, monitoring or scaling clusters.

Dremio Cloud ensures that customers have **full control of their data**, resulting in stronger security and governance, and avoiding the vendor lock-in that many companies have experienced in the past with databases and data warehouses.

Besides the enhanced data security, Dremio Cloud's architecture also provides a **drastically simplified analyst experience**. With **bidirectional integration** with Tableau and Power BI, analysts can easily connect their BI tools to Dremio Cloud with **passwordless** SSO and zero effort. The control plane automatically routes client queries to their corresponding engines according to a set of user-defined routing rules, ensuring maximum performance while **avoiding "noisy neighbor" issues**.

Dremio Cloud helps customers optimize and secure their data and workloads with a **highly scalable multi-engine architecture**:

- Multi-engine, multi-cluster data plane architecture consists of multiple right-sized engines to support different workloads in the company. For example, a company can have a medium engine for executive dashboards, a large engine for batch jobs and an extra-large engine for data science queries.
- Dremio Cloud introduces the capability for engines to have multiple replicas, where engine replicas are added and removed dynamically based on the workload. This helps companies tackle any level of concurrency while maintaining consistent performance.

**The elastic, multi-engine architecture scales infinitely and helps manage workloads with the highest concurrency and lowest latency.**

> We have to constantly size capacity for peak loads and run our engines 24x7. Auto-scaling from Dremio will help manage this problem.
>
> **- Dremio customer**

Dremio Cloud is an **elastic service** that can scale up and down. This elasticity can help you tackle workloads of any concurrency or throughput based on these features:

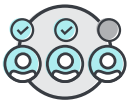
- SaaS control plane can scale-out to handle any concurrency
- Engines can be created and sized to meet throughput requirements for different workloads
- Engines replicate to meet any concurrency with consistent performance
- Engines can be easily resized to meet throughput requirements
- Zero cost for the control plane and idle engines
- Pay based on usage and only when you run queries

dremio

The **security** in Dremio Cloud is built for the **enterprise** from the **ground up**, and supports **enterprise security requirements across the stack** — from the underlying infrastructure, to data encryption, to access control, to authentication with widely used identity providers, etc.

### Infrastructure

- Regional control planes
- TLS encryption for all communication
- No incoming connections to customer VPC/data plane
- Multiple engines with privileges to control user access
- Private IP for engines



### Authorization

- OAuth 2.0 on all interfaces including SSO from BI tools
- Fine-grained privileges on all objects (data and administration)
- Custom roles
- Projects for compute & data isolation by team

### Authentication

- Enterprise & social Identity Providers (AAD, Okta, Ping, Google, etc.)
- Authenticate with OpenID Connect to your IdPs
- Existing users and groups with SCIM

### Data

- Data stays in your cloud account
- End to end encryption of data
- Customer keys for encryption
- Dynamic data masking

Dremio has **two offerings:**

### Dremio Software

With **Dremio Software**, free yourself from expensive, proprietary data warehouses and enjoy timely decision-making with sub-second BI queries and instantaneous dashboards directly on data lake storage at the speed of business, a much more efficient and cost-effective no-copy data architecture, consistent business logic and enterprise-grade governance and security.

### Dremio Cloud

With **Dremio Cloud**, enjoy an infinitely scalable SaaS with a frictionless experience across multi-cloud, enterprise-grade security built from the ground up and high-performance BI dashboards and interactive analytics directly on data lake storage.

## The Dremio Advantage

With Dremio, shorten time to insight from days or weeks to minutes with a simplified, no-copy data architecture, high-performance BI dashboards and interactive analytics directly on the SQL lakehouse:

- The **no-copy architecture** eliminates the long backlog of requests and empowers data scientists, data analysts and data consumers through self-service access to business-ready data in real time.
- The industry's only vertically integrated **semantic layer**, with shared business logic, data access, KPIs and security, provides a single, consistent view of the data for all teams and tools and applies security and data governance across the company.
- This shared logic and security eliminates business logic duplication across multiple BI tools and **prevents multiple data copies** at different levels — no more cubes, aggregation tables or extracts — and recurring data drift.
- **Seamless** SSO **integrations with BI tools** provide lightning-fast queries directly on the cloud data lake with sub-second responses from any BI or dashboard.
- Dremio's **high-performance execution engine**, powered by Apache Arrow, speeds queries by 100x and eliminates >90% of compute costs.
- With Dremio Cloud, your data and data processing **stays in your cloud account**. Within your account, data is encrypted at rest and in transit. Secure your data further with support for customer-managed encryption keys and dynamic data masking.
- You can deploy Dremio **anywhere**! Run **Dremio Software** in your own environment, or choose **Dremio Cloud**, a fully managed SQL lakehouse platform built on an open architecture.

## Get Started today with Dremio Software or Dremio Cloud.

### ABOUT DREMIO

Dremio is a SQL Lakehouse Platform company enabling organizations to leverage open data architectures. Dremio's SQL Lakehouse Platform simplifies data engineering and eliminates the need to copy and move data to proprietary data warehouses or create cubes, aggregation tables and BI extracts, providing flexibility and control for data architects and data engineers, and self-service for data consumers. Dremio Cloud, a frictionless, infinitely scalable service, enables high performance SQL workloads directly on cloud storage, eliminating the cost and complexity of copying and moving data. Dremio Cloud reimagines the traditional data lake by combining the best of traditional data warehouses and data lakes into a SQL lakehouse, while removing the limitations of traditional data warehouses resulting from closed data architectures.