SOLUTION BRIEF

Modern, powerful, and efficient data lake infrastructure with StorageGRID and Dremio

NetApp

dremio

Enable effective and impactful data consumption across the enterprise

Developing a modern data infrastructure isn't an easy task. Enterprises want to extract as much value as possible from their data, while minimizing complicated data pipelines. Traditional data warehouses can lead to data silos, vendor lock-in, inconsistent sources of truth, and tangled processes. Faced with head-spinning increases in data volume and variety year after year, enterprises often find that legacy data infrastructures like Hadoop just can't scale to provide the access they need. Data administrators are tasked with the complicated process of building out modern data lakes while maintaining legacy data pipelines that are critical for day-to-day success. NetApp has partnered with Dremio, the easy and open data lakehouse, to help enterprises face the challenge of building a future-proof, scalable, and efficient data infrastructure.

Unrivaled scalability and data management for the data lake

NetApp® StorageGRID® object storage is an enterprise-grade, on-premises solution that supports the native Amazon Simple Storage Service (S3) API. StorageGRID is software defined, which means that you can run it on different platforms—bare metal, VMware-based environments, or NetApp's purpose-built appliances—and mix platforms within a grid. StorageGRID offers massive S3 object storage and dynamic data management, enabling you to run next-generation workflows on premises alongside your public cloud. The solution's industry-leading data management policy engine helps you optimize performance and durability and adhere to data locality requirements. StorageGRID is extremely scalable, supporting low-touch, nondisruptive expansions, and can store billions of objects. In a single namespace, StorageGRID can scale up to 16 data centers worldwide.

Dremio helps data teams deliver faster access to their data. Deploying the Dremio open data lakehouse on top of StorageGRID object storage, you can maximize the value of your enterprise's data while exercising full control of data placement, lifecycle, and tiering. Dremio and StorageGRID put you in control of data management, driving impactful analysis and business intelligence (BI) while ensuring that data is properly placed to maximize value. StorageGRID information lifecycle management (ILM) policies allow complete granular control over how long data is retained, where the data sits, when it's tiered to lower-cost object storage, and more. As StorageGRID gives you unrivaled control over how your data lake is managed, Dremio facilitates simple access to the data across an entire enterprise.

Dremio: The easy and open data lakehouse

The Dremio open data lakehouse empowers you to make effective and impactful use of your enterprise's data. With Dremio's SQL query engine, data can be queried in place on StorageGRID, so data users across an enterprise have access to the data lake. Dremio can query data from a wide range of sources in addition to S3 object storage, including block and file storage, Hadoop Distributed File System (HDFS), and relational databases like Amazon Redshift and Postgres. With Dremio's semantic layer, it's easy to build and share data products over your current data infrastructure, which gives you direct access to all your data during phased migrations to a modern data lake.

Key benefits

- Build your data lake on StorageGRID for unmatched scalability and data management.
- Empower data users across the enterprise to derive full value from your data lake by querying data in place with Dremio.
- With the combination of StorageGRID and the Dremio open data lakehouse, enable enterprisewide infrastructure that galvanizes the grid's data for efficient consumption.

Dremio supports platform-agnostic data formats like Parquet and Apache Iceberg, making it easy to avoid vendor lock-in and to future-proof your data infrastructure.

Maximize the value of your data

Enterprises employ a wide variety of data users—from data scientists to business analysts to executives who need high-level BI dashboards—and traditional data warehouse infrastructures make it difficult to get the right data to the right users quickly. With Dremio, users can connect their BI tools and SQL clients directly to the data lake. This approach removes complicated pipelines, so the whole data lake is just a click away for any data user. Dremio natively connects with Tableau and Microsoft Power BI for BI teams, and it supports Apache Arrow Flight to easily connect the data lake to Python, R, and Jupyter Notebook. In addition, Dremio spaces provide a shared semantic layer for all users and tools. Spaces allow data analysts and scientists to create consistent dataset definitions, calculated fields, and security rules that downstream users and tools can use. By providing simple access to an organization's entire data infrastructure, Dremio maximizes the value of data stored on your StorageGRID system to enable high-level business intelligence and data analytics.

The Dremio SQL engine uses query-acceleration technology to achieve interactive-speed response times, opening the data lake to real-time data analysis and BI. Dremio supports Columnar Cloud Cache (C3), which uses NVMe SSD technology built into cloud compute instances to achieve NVMe-level I/O performance. Dremio also uses Data Reflections to help accelerate BI dashboards and help end users work freely in their semantic layer without ever needing to know about their physical data model.

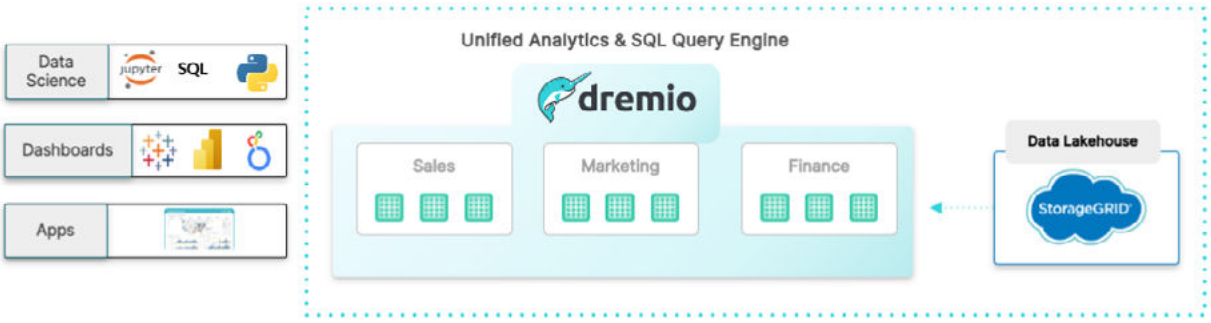

Figure 1: The StorageGRID and Dremio solution architecture enables data users across the enterprise to access the data lake.

This helps data engineers eliminate redundant data pipelines and physical data copies commonly found in maintaining materialized views and BI extracts.

Modernize your Hadoop cluster

Hadoop is a widely used legacy data analytics platform that is still employed by many enterprises across industries. Although Hadoop is a powerful tool for processing big data, it faces several challenges in the modern enterprise. To address these pain points, Hadoop workloads can be migrated to a StorageGRID and Dremio joint solution. This solution provides a modernized data analytics platform with improved performance, scalability, security, and simplicity, and offers several advantages over a legacy Hadoop cluster. These advantages include lightning-fast query performance, independent, and low-touch scaling of compute and storage, built-in future-proof security, and ease of administration. Migrating your Hadoop cluster to StorageGRID and Dremio can provide these benefits and more, making it a no-brainer decision.

Dremio and StorageGRID

Simple setup with incredible results

Dremio can be deployed as a software solution on your enterprise's hardware, or as a fully managed software-as-a-service (SaaS) cloud deployment. Dremio Cloud runs on AWS and handles software installation, configuration, and upgrade as well as compute engine management. Dremio software runs on premises and in multiple clouds, and brings data lakehouse functionality wherever your data resides, including your data center. StorageGRID can easily connect to the Dremio lakehouse as an S3 data source, and after this simple setup, the grid's data is available for democratized access across the enterprise. Deployed together, Dremio and StorageGRID powerfully enable an enterprisewide data infrastructure, galvanizing the grid's data for effective, impactful, and efficient consumption. Dremio is the perfect platform to make the most of your StorageGRID data lake.

About NetApp

In a world full of generalists, NetApp is a specialist. We're focused on one thing, helping your business get the most out of your data. NetApp brings the enterprise-grade data services you rely on into the cloud, and the simple flexibility of cloud into the data center. Our industry-leading solutions work across diverse customer environments and the world's biggest public clouds.

As a cloud-led, data-centric software company, only NetApp can help build your unique data fabric, simplify and connect your cloud, and securely deliver the right data, services, and applications to the right people—anytime, anywhere. www.netapp.com

About Dremio

Dremio is the easy and open data lakehouse. Data teams use Dremio to deliver self-service analytics while enjoying the flexibility to use Dremio's lightning-fast SQL query engine and any other processing engine on the same data. In addition, Dremio eliminates data silos by enabling queries across data lakes, databases, and data warehouses. Dremio helps organizations get started with analytics in minutes, and automatically optimizes data for every workload.

NetApp

+1 877 263 8277