dremio | NetApp

With Dremio's Unified Lakehouse Platform, NetApp achieves 95% faster time-to-insight

At a Glance

The Customer

NetApp

Challenge

NetApp's ActiveIQ team needed to modernize its Hadoop infrastructure to reduce costs and improve performance in supporting expanding data requirements across its portfolio of product offerings. Crucially, it required that the implementation of any upgrades would involve minimal or no changes to existing data pipelines.

Solution

Dremio provided NetApp with a phased approach to Hadoop modernization and migration, providing a single access layer for self-service over NetApp's on-prem data lakehouse.
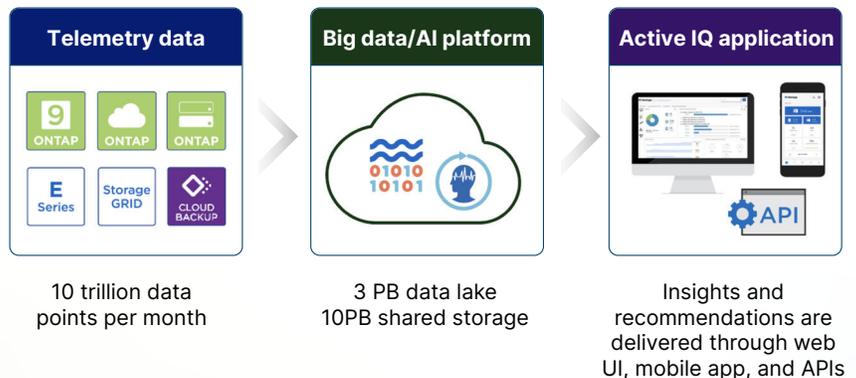
Results

- Reduced storage cost using NetApp's StorageGrid E-Series.
- Modernized Hive query engine with Dremio, resulting in 20x faster queries for BI analysts and data science users.
- Provide improved self-service end-user experience, including the ability for data engineers to easily curate, publish, and make virtual data marts easily discoverable and explorable for end users from Dremio.

The Business:

NetApp is a leading provider of data solutions with a portfolio of offerings that span data management, application, and storage, addressing enterprise requirements across a range of environments, from on-premises to hybrid and multi-cloud. NetApp develops leading all-flash data storage hardware and the only enterprise-grade storage OS, available on the world's leading public clouds.

NetApp's Customer Experience business unit oversees the ActiveIQ Data Lake. Its Active IQ platform is a digital advisory platform that simplifies and proactively manages the customer experience across NetApp's suite of services. The team analyzes over 10 trillion data points per month coming from customer environments, data and AI operations, as well as the Active IQ application for receiving insights and recommendations delivered to customers via a web UI, mobile app, and APIs.

Active IQ: Digital Advisory for Predictive Maintenance and Optimization

Telemetry data	Big data/AI platform	Active IQ application
10 trillion data points per month	3 PB data lake 10PB shared storage	Insights and recommendations are delivered through web UI, mobile app, and APIs

The Challenge:

NetApp's Active IQ solution started as a tool for integrating and analyzing telemetry data for its support use cases, eventually evolving into a broader offering for both NetApp-internal users as well as customers. The underlying backend—a Hadoop/MapReduce-based data infrastructure developed over a decade ago—posed significant challenges with the growth of data and need for data access.

For example, its storage needs were expanding far more rapidly than its compute needs; however, because compute was directly attached to storage, adding more of the latter meant scaling horizontally and adding more unneeded compute, and with that hardware and Cloudera licensing-related costs.

Before Dremio, NetApp's Active IQ data infrastructure consisted of 33 mini-clusters, over 4,000 cores, and more than 7 petabytes of data. Creating and maintaining the Hadoop cluster requires a Hadoop expert, and maintenance was time consuming as the cluster grew.

Along with the cost of compute, data performance and management were also increasingly problematic. Queries on average took 45 minutes, and Hive's course-grained configurations meant that misconfigurations and sub-optimal settings could result in starving out other tasks like Hive queries from required resources. NetApp therefore evaluated solutions based on these cost reduction and related storage/compute decoupling requirements, as well as performance improvements (i.e., reducing the 45-minute average query time), features for simplifying data and resource management, the availability of more fine-grained controls, and disaster recovery capabilities.

The Solution:

Dremio provided NetApp a roadmap for its journey to unified analytics using a phased approach for modernizing its Hadoop-based data infrastructure. Dremio required minimal changes to existing pipelines. During its evaluation of the solution space, other vendors required substantial changes to how it processed data, resulting in significant time and expenses added to the migration project.

ActiveIQ's old environment was running on top of bare metal, which made patching and overall management a difficult affair; by moving to Dremio and a fully containerized environment, they could drastically reduce their management overhead while improving security and resilience. Also, Dremio adoption of open ecosystem around Apache Iceberg and Arrow, meant the solution was future-proof, transparent, and extensible, and as a replacement for their Hadoop/Hive infrastructure, could provide functionality for various secondary use cases via the semantic layer.

The existing Spark-based ETL and data ingestion mechanisms would remain in place, but Dremio would provide a unified access layer that makes data easier to discover and explore for end users without data duplication. This allowed for a drastic data replication factor reduction as well as the de-coupling of storage and compute.
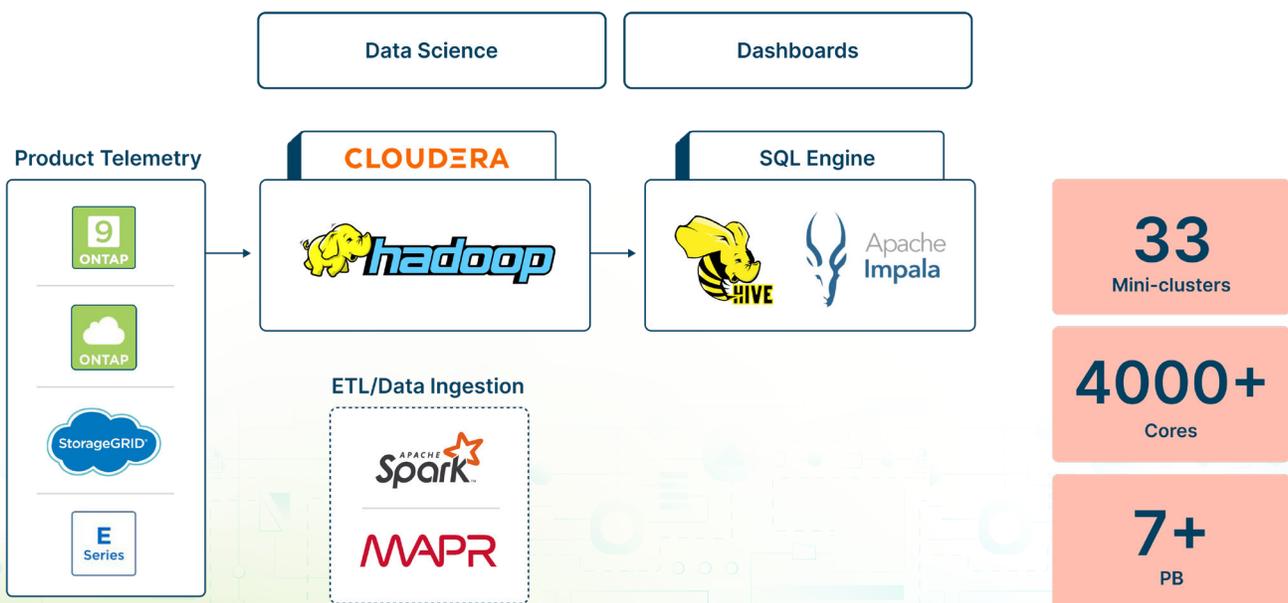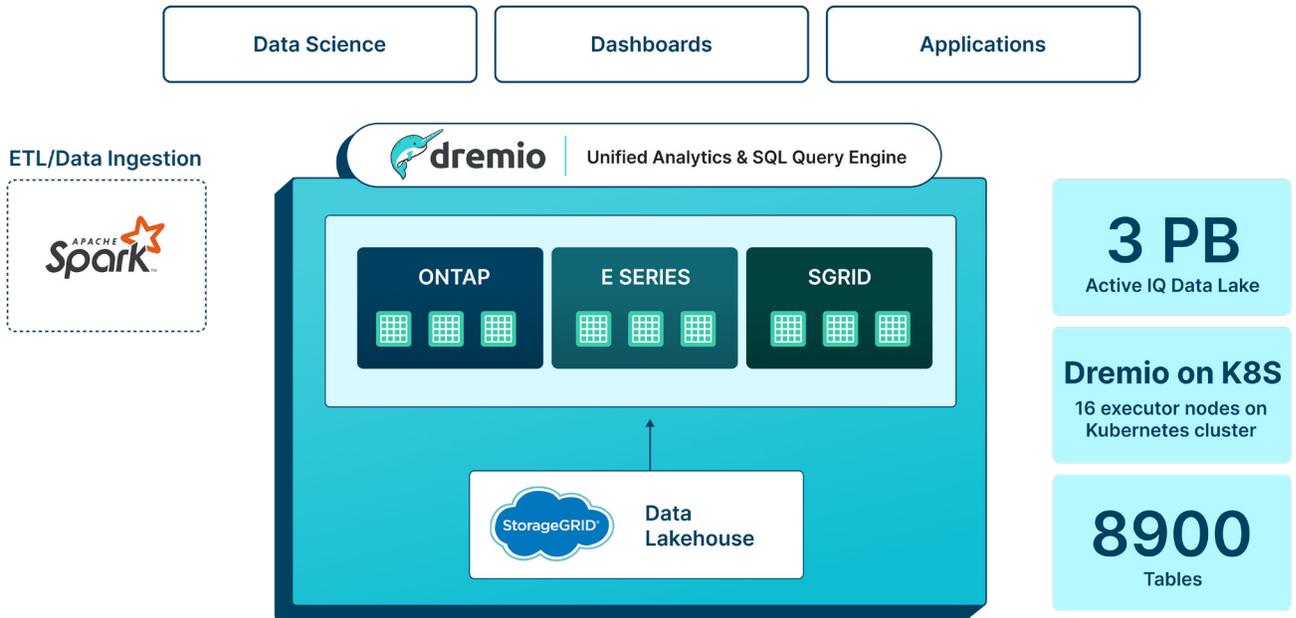
Results:

With Dremio in place, NetApp was able to significantly cut its costs by drastically reducing both compute consumption as well as the amount of disk space needed in their data environments. The resulting data infrastructure consisted of 8,900 tables holding 3 petabytes of data, in contrast to the previously over 7 petabytes of data; the new Active IQ Data Lake was supported by 16 executor nodes on Kubernetes clusters versus the previous data infrastructure of 33 mini-clusters and over 4,000 cores. Along with compute-related cost savings, NetApp also saw drastic performance increases — even with the decrease in compute resources.

By accessing data directly over their data lakehouse with Dremio, query runtime was reduced from 45 minutes to 2 minutes, a 95% faster time to insight for predictive maintenance and optimization across NetApp's product telemetry data. The migration resulted in an over 60% reduction in compute costs compared to its previous data infrastructure, over 20 times faster queries, and over 30% in TCO savings.

Conclusion:

With Dremio's Unified Lakehouse Platform, NetApp achieves 95% faster time to insight while simplifying proactive customer care. There is now a better way to empower data consumers to self-service data, proactively manage the customer experience, and optimize and identify problems before they happen. Dremio's data lakehouse enabled the ActiveIQ team to leverage telemetry data, reduce risks of customer churn, and provide higher product availability across the customer journey.