Migrating from On-Prem to a Modern Cloud Data Lake: The Cost-Efficiency and Performance Benefits

dremio

Introduction

The need for new data analytics in the enterprise is growing now more than ever. What's not growing, however, are the funds and resources required to meet these needs — a problem only compounded by the recent pandemic and corresponding economic issues.

Organizations that already bought into the idea of a "data lake" — a centralized repository for collecting, storing and processing all structured, semi-structured and unstructured data at any scale — might believe their on-premises data lake is immune to any expected or unexpected shortfalls like the ones described above.

But it's not. On-prem data lakes have proven themselves to be inherently complex, costly and poor performing, despite heavy system management efforts.

Now many disillusioned and frustrated on-prem data lake veterans are eyeing cloud data lakes as the more flexible, affordable and future-proof option. And, while a cloud data lake certainly does offer up a myriad of advantages over its on-prem predecessor, only a *modern* cloud data lake can deliver on the data lake's original promise — and more.

That's because only a modern cloud data lake enables you to:

- Break down data silos, and maintain full and complete control of your data at all times
- Empower a broader group of people — from data scientists and business analysts to data engineers and architects — to easily make sense of data using the best-of-breed, advanced analytic tools and frameworks of their choice *against a single, cloud-based store of data*
- Future-proof your platform through an open cloud data lake architecture that makes it easy for users to efficiently work with any of the different analytic applications, tools and data formats available today, as well as those that become available in the future
- Grow storage and data processing needs cost-effectively over time because it's easier to configure, provision and scale for data replication and high availability
- Accelerate current and future machine learning and artificial intelligence (AI) needs

As a result, modern cloud data lakes deliver substantially more cost-efficiency and performance benefits than their on-prem counterparts.

Let's compare them side by side to find out why.

What Makes a Cloud Data Lake "Modern"?

Only "modern" cloud data lakes truly democratize the data stored there. That's because *only modern cloud data lakes use next-gen data lake query engines* to run live, interactive queries and business intelligence (BI) directly on cloud data lake storage. This modern approach makes complex data easily accessible and highly performant to a wide range of data users — without copying or moving it to a data warehouse or data mart.

The Performance Benefits of a Modern Cloud Data Lake

PERFORMANCE — Scalability and Throughput

Modern cloud data lakes can more easily and seamlessly adapt for data volume increases while simultaneously supporting multiple intensive workloads — without adversely impacting performance.

ON-PREMISES DATA LAKE

Legacy data lake solutions like Hadoop require compute and storage resources to be co-located, so you have to size clusters to accommodate peak processing loads, an exercise that creates significant processing overhead and constrains performance.

Engineers with deep technical skills and extensive knowledge of Hadoop's notoriously complex platform are required to build the custom tools and workarounds intended — but not guaranteed — to overcome Hadoop's limitations.

VS

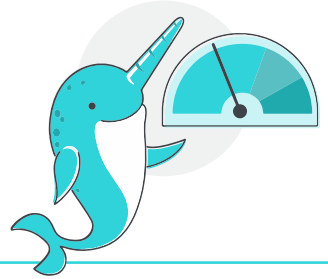
MODERN CLOUD DATA LAKE

A modern cloud data lake leverages flexible object storage services like Amazon S3 or Microsoft Azure Data Lake Storage (ADLS) for near-infinite, independent scalability of storage and compute resources.

A modern cloud data lake query engine decreases the amount of time required for data and analytics processing from hours to seconds.

A modern cloud data lake supports an unlimited number of users and workloads.

A modern cloud data lake provides a foundation for all aspects of data analysis — from data loading to reporting and data science.



Since many early on-prem data lakes were built on Apache Hadoop, we specify Hadoop in our on-prem data lake write-up.

PERFORMANCE — Query Response Times

Modern cloud data lakes make analysts significantly more productive while freeing them from the tedious operational tasks common to on-prem data lakes.

ON-PREMISES DATA LAKE

Slow analytics performance doesn't meet the low latency and volume requirements of modern BI queries and interactive workloads.

Long query response times result in delayed time to insights, preventing users from timely analysis and iteration.

VS

MODERN CLOUD DATA LAKE

Since you run BI directly on a modern cloud data lake, users can query data directly via a cloud data lake query engine without having to sacrifice the time and effort needed to move data into a data warehouse.

With a modern cloud data lake architecture, you can leverage a high-performance query engine to meet real-time or near-real-time SLAs.

Since a modern cloud data lake architecture includes a cloud data lake query engine, intelligent and transparent caching makes query response times faster when analyzing huge datasets.

With a modern cloud data lake, batch processing can be run during business hours without stalling or degrading the performance of other business systems.

A modern cloud data lake seamlessly manages data flow spikes without negatively impacting performance.

Apache Arrow Gives an Added Performance Boost

Modern cloud data lakes that use Apache Arrow in-memory columnar processing eliminate unnecessary I/O and serialization/deserialization of data. Plus, Arrow further improves a heterogeneous environment's performance by providing a common memory format enabling multiple processors to run on the same data.

Dremio: 3,000x Faster than Presto

Dremio's cloud data lake query engine was recently benchmarked against several Presto-based query engines, including PrestoDB, PrestoSQL, Starburst Presto and AWS Athena. The results?

- **Ad hoc queries** on Dremio are **3,000x** faster and use **1/3000th** of the EC2 compute infrastructure.*
- **BI and reporting queries** on Dremio are **1,700x** faster and only use **1/1700th** of the EC2 compute infrastructure.*
- **On average,** Dremio is **1,000x** faster and uses **1/1000** of the EC2 compute infrastructure.*

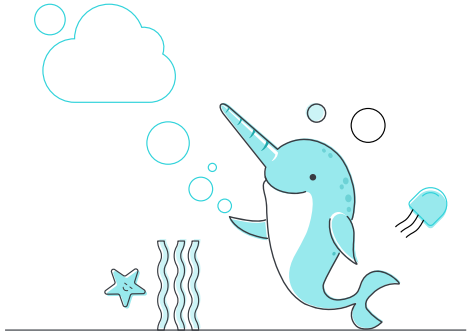
* At the same node count (since Dremio is so much faster, the only way to compare performance is by evaluating the same number of compute nodes).

The Cost-Efficiency Benefits of a Modern Cloud Data Lake

COST-EFFICIENCY — Initial Investment

A modern cloud data lake requires a much smaller and more predictable upfront investment than an on-prem data lake.

ON-PREMISES DATA LAKE	MODERN CLOUD DATA LAKE
An on-prem data lake requires you to make a large, upfront capital investment that's treated as an asset and depreciated over time.	A modern cloud data lake requires a much smaller upfront investment that's treated as an operating expense and deducted monthly against revenues.
You have to purchase (then maintain) hardware and software infrastructure.	You don't have to purchase any hardware or software — instead, you provision the cloud-based infrastructure, compute and storage resources you need.
It's difficult to correctly estimate the resource requirements needed to maintain and update an on-prem data lake, so you have to choose between overprovisioning — paying for resources that will never be used — or provisioning more resources after the fact, as they're needed, at a higher cost.	Multiple storage classes and pricing options let you pay only for the amount of storage you use, so you never have to pay for hardware that's not needed.



COST-EFFICIENCY — Ongoing

Compared to modern cloud data lakes, on-prem data lakes require substantially more ongoing investment because in-demand technical resources must dedicate a significant amount of their time to operating and maintaining on-prem data lake infrastructure, making those resources less productive overall.

ON-PREMISES DATA LAKE

Storage

You need to buy more compute and storage hardware as the datasets in your on-prem data lake grow in size or volume.

To increase capacity in order to meet growing demands, you need to add more physical storage to your on-prem data lake, and then manage and maintain the storage hardware.

Often it's harder to separate compute from data and accurately estimate how much storage will be needed for an on-prem data lake, so efforts to scale storage will be less efficient and more expensive — you end up paying for capacity you don't use because you buy infrastructure based upon the maximum capacity you think you will need.

VS

MODERN CLOUD DATA LAKE

Storage

Pay-as-you-go pricing makes it easy to expand a modern cloud data lake's capacity if/when necessary, ensuring you pay only for what you need.

It's easy to separate data from compute in a modern cloud data lake, so you can scale each one independently and much more efficiently, and also configure the data lake to automatically deactivate CPUs during idle time between jobs.

Less storage capacity is required for a modern cloud data lake because fewer data copies are generated, so less data redundancy is needed.

Instead of implementing a one-size-fits-all model, modern cloud data lakes that use elastic compute engines can right-size execution resources for each distinct workload by provisioning multiple, separate execution engines that can start/stop based upon predefined workload requirements at runtime.

Watch Out!

Data warehouses that separate storage and compute still lock your data into their architectures. Only *true separation of data* from compute — which can be provided by modern cloud data lake query engines — permits best-of-breed data lake engines to process your data.

> **Data lake early adopters have hit a ceiling, held back by Hadoop's numerous omissions and weaknesses in key areas such as cluster maintenance, admin cost, resource management, metadata management and support for SQL and other relational techniques.[1]**
>
> TDWI

ON-PREMISES DATA LAKE

VS

MODERN CLOUD DATA LAKE

Resources

Busy, highly paid IT and engineering team members have to:

- Manage, operate and maintain an on-prem data lake's infrastructure
- Deal with aging hardware, outages and downtime
- Integrate all of the tools used to ingest, organize, pre-process and query the data stored in the on-prem data lake

To support more users, more datasets or more workloads with an on-prem data lake, you must add and manage more infrastructure — which requires a significant manual effort, and adds a substantial amount of maintenance and operating costs to your initial investment.

With an on-prem approach, high availability of data and insights is severely restricted and delayed because only employees with highly technical skill sets can use the data lake.

Resources

Far fewer scarce, highly skilled engineers are needed to build and maintain a modern cloud data lake.

A modern cloud data lake's intuitive nature combined with open formats, open APIs and flexible capabilities that can integrate with many environments deliver operational simplicity.

It's easy to add new cloud services to a modern cloud data lake without having to change the architecture.

The cloud vendor handles all technical support, maintenance and upgrades for a modern cloud data lake per the SLA agreement.

What the Cloud Vendor Provides

Technical support

- Provisions resources, backs up data, optimizes performance and ensures data is secure (as defined by the shared responsibility model)
- Mitigates and contends with service interruptions
- Stores redundant copies of data across different availability zones
- Provides expert customer support

Maintenance

- Eliminates the traditional downtime required for hardware maintenance and software updates
- Ensures security patches are immediately applied upon release
- Makes upgrades easy and ensures you always work with the most efficient, up-to-date infrastructure and software

THE DATA TOOLKIT FOR DATA LAKES

Ingestion
Sqoop
Flume
Apache Kafka
Apache NiFi
StreamSets
Amazon Kinesis
DataTorrent
Apache Storm
Fluentd
Scribe

Processing
Spark
EMR
Qubole

Formats
CSV
Parquet
ORC
JSON

Queries
Dremio
Presto (PrestoSQL, PrestoDB, Starburst, AWS EMR, AWS Athena)
Hive
Spark

> **According to Google, Google Cloud TCO on average is 57% lower than an on-premises Hadoop deployment.[2]**

Henkel's Modern Cloud Data Lake Drives Accelerated Insights and Efficiency for Supply Chain Leaders

Generating nearly 6.6 million euros in sales and comprising one-third of the company's core business, Henkel's Laundry & Home Care division generates a massive number and scale of supply chain datasets — datasets that are key to effective demand forecasting, supply network planning, production scheduling, manufacturing and logistics for the company's 33 production plants, 70 contract manufacturers and 60 warehouses around the world.

But in 2016, Henkel lacked the real-time supply chain insights managers needed because data silos storing critical supply chain datasets weren't coherently connected. Forced to rely upon Microsoft Excel reports generated only weekly, monthly or quarterly — or turn to expensive, external consultants or vendors — to complete their analyses, leaders knew they needed to find a better way.

After learning a modern cloud data lake composed of ADLS, Dremio and Tableau would allow them to natively analyze data stored in the cloud, Henkel leaders decided it was the answer to their problem.

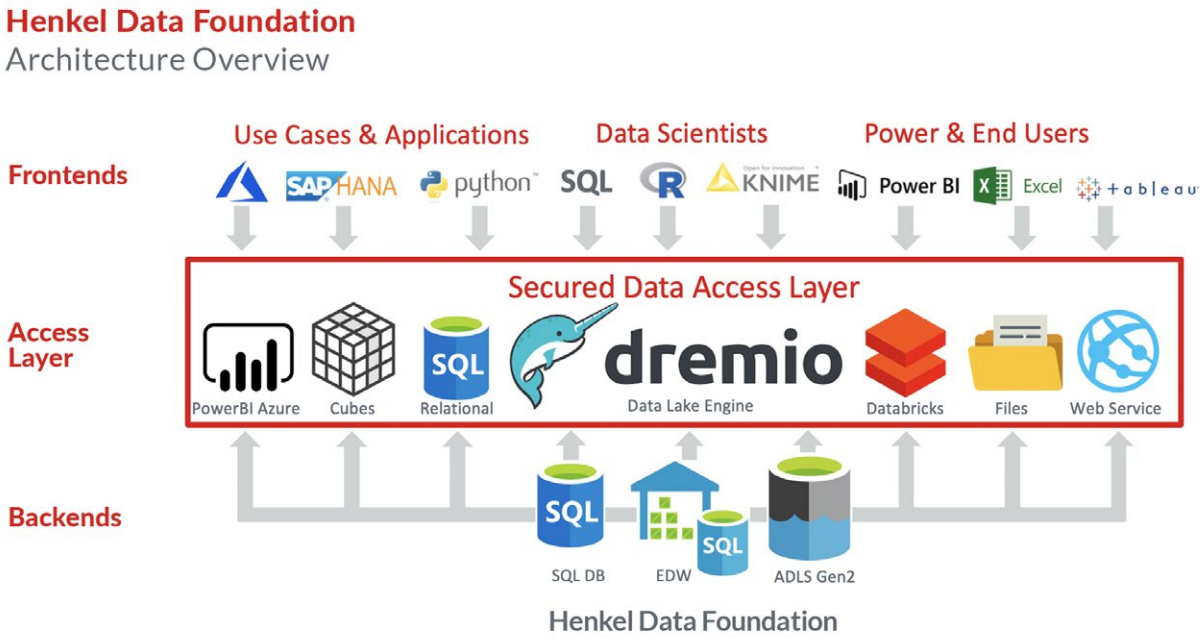
Now, Henkel leaders enjoy dramatically increased visibility and transparency into the company's supply chain data. They can self-serve the up-to-date insights they need, whenever they need them, while obtaining real-time answers to supply/demand planning, production or inventory questions from the more than 500 Tableau dashboards that are significantly accelerated by Dremio's data lake query engine.

Real Business Results, Real Fast

- **+10% increased Overall Equipment Effectiveness (OEE)**
- **-30x significantly reduced query time (from 3-4 minutes to only 8 seconds)**
- **Dramatically accelerated insights from >500 Tableau dashboards**
- **Reduced computing and maintenance costs**

The Data Foundation Architecture

Unique data join, filtering and transformation capabilities from Dremio — ones that enable the dynamic sharing of data across data silos — are key components of Henkel's modern cloud data lake.



The differentiation Dremio enables is evident in the live and dynamic parameter settings available in Henkel's new supply chain planning tool. According to Wolfgang Weber, Head of Digital Transformation for Henkel Laundry and Home Care, "This is huge. We have 33 production plants with around 400 lines, with 250 connected in real time. It's a huge productivity increase and advantage, and just one example of how data drives cost savings."

> **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 HENKEL

Use Dremio to Save Up to 90% in Modern Cloud Data Lake Infrastructure Costs

Traditional SQL query engines are suitable only for ad hoc queries run by highly technical users where low, inconsistent performance is tolerable. That's because those engines lack the query acceleration and semantic layer needed to make BI and interactive SQL work directly against cloud data lake storage.

Attempts have been made, of course, to increase SQL query performance — for example, by copying and moving significant volumes of data from the data lake into a proprietary, on-prem or cloud-based data warehouse, or by creating and maintaining external acceleration technologies like BI extracts, OLAP cubes and aggregation tables. But these attempts also fail to expediently and cost-efficiently support common, in-demand analytics use cases like BI and real-time reporting.

That's where Dremio's cloud data lake query engine comes in.

Only Dremio delivers secure, self-service data access and lightning-fast queries directly on AWS or Azure storage. And only Dremio delivers a vertically integrated semantic layer and Apache Arrow-based SQL engine — elements that reduce time to analytics insight while increasing data team productivity and lowering infrastructure cost and complexity.

This unique approach is how Dremio:

- Lets you keep all of your data in cloud data lake storage
- Enables BI users and data scientists to directly query data in the data lake
- Shrinks the amount of data (and the data's associated workloads) that's copied or moved into a data warehouse to just 20% or less

Dremio: The Key to Even Faster, More Affordable Insights from Cloud Data Lakes

Dremio makes evolving your on-prem data lake to a modern cloud data lake easy and affordable — and delivers fast time to value to boot. Here's why.

1. **It's open.** Dremio separates data — not just storage — from compute, so you can future-proof your analytics architecture to leverage best-of-breed applications and engines today and tomorrow.
2. **It's lightning-fast.** Dremio accelerates ad hoc queries by 3,000x and BI queries by 1,700x versus SQL engines, eliminating the need to ETL data into a data warehouse and eliminating the need for cubes, extracts or aggregation tables.
3. **It makes you more productive.** Dremio lets you provision new datasets with consistent KPIs and business logic in minutes — not days or weeks — and empower analysts to create their own derivative datasets, without making copies.
4. **It makes you more efficient.** Dremio lets you easily size the minimum compute needed for each workload and only consume compute when running queries, which reduces compute infrastructure and associated costs by up to 90%.

Conclusion

Cloud data lakes certainly offer many advantages over their on-prem counterparts. But simply residing in the cloud does not make a data lake "modern."

Rather, transitioning to a truly modern cloud data lake means you break down data silos to make a single source of hard-to-reach data easily accessible by the people who need it. You let them easily understand that data using whichever best-of-breed, advanced analytic tools, frameworks and services they prefer. And you keep complete control of the data while future-proofing your cloud data lake investment with an open, standards-based approach that ensures down-the-line compatibility.

Only then will you realize the significant cost and performance benefits a modern cloud data lake delivers over its on-prem counterpart.

Learn more about Dremio's unique approach to powering modern cloud data lakes — watch the on-demand webinar Build a Best-in-Class Data Lake.

CHECKLIST: Is the Cloud Data Lake You're Considering *Really* "Modern"?	Yes	No!
Is the cloud data lake based upon open standards — particularly regarding data formats and data access?	☐	☐
Does the cloud data lake offer independently scalable storage and compute resources?	☐	☐
Does the cloud data lake give you the option to deactivate compute instances during idle time between queries and jobs?	☐	☐
Does the cloud data lake use in-memory processing to facilitate real-time or near-real-time SLAs?	☐	☐
Does the cloud data lake eliminate the need for siloed semantic layers to be built into BI visualization tools?	☐	☐
Does the cloud data lake architecture eliminate the need for separate OLAP cubes, BI extracts, etc.?	☐	☐
Is it easy to optionally join external sources to data in the cloud data lake without needing ETL?	☐	☐
Does the cloud data lake automatically stay up to date with new capabilities?	☐	☐
Can data engineers provision new datasets to data analysts in only minutes, without copying or moving data?	☐	☐
Can BI analysts and/or data engineers:		
- Use a consolidated, self-service semantic layer to find, access and create new datasets without making copies of data?	☐	☐
- Run BI/reporting queries that return in only sub-seconds, and ad hoc queries that return in only seconds or minutes?	☐	☐
- Execute interactive queries on their own?	☐	☐
- Run queries directly upon the datasets stored in the cloud data lake without needing to ETL it into a data warehouse?	☐	☐

If you answered "No" to any of the left questions, the modernity of the cloud data lake you're considering is suspect. Dig deeper.

Want to learn more about how Dremio powers true modern cloud data lakes? Download the Dremio Architecture Guide to understand Dremio in depth.

[1] TDWI, "Data Lake Platform Modernization: 4 New Directions," March 15, 2019.

[2] Google Cloud, "Data Lake Modernization."

ABOUT DREMIO

Dremio delivers lightning-fast queries and a self-service semantic layer directly on your data lake storage. No moving data to proprietary data warehouses, no cubes, no aggregation tables or extracts. Just flexibility and control for data architects, and self-service for data consumers.

Deploy Dremio

Learn more at dremio.com

CONTACT SALES

contact@dremio.com