dremio

The Unexpected Cost of Data Copies

REALIZING A MORE EFFICIENT DATA LAKE ARCHITECTURE WITH DREMIO

April 2021

Table of Contents

Introduction

While modern data lakes provide many advantages, data copies are a pervasive problem. Data is copied for several reasons, including ingesting data into data warehouses, creating performance-optimized copies and building OLAP cubes and BI extracts for analysis.

Excessive replication and data transformation can result in longer time to value, reduced operational efficiency and skyrocketing storage, platform and labor costs. Data copies also pose a security threat that can result in increased risk and compliance costs. Fortunately, new data lake technologies can dramatically reduce the need for replicated data, providing the opportunity to simplify and streamline data management environments.

This paper will discuss why organizations frequently end up with multiple data copies and how a secure "no-copy" data strategy enabled by the Dremio data lake service can help reduce complexity, boost efficiency and dramatically reduce costs.

Modern Data Management Challenges

As businesses become increasingly dependent on timely, high-quality data to enable business decisions, data teams are under enormous pressure. They need to democratize data access, enable secure self-service access to more users and do so quickly to minimize time to value. As if this were not hard enough, they also need to comply with industry-specific regulations such as HIPAA and Basel and fast-evolving privacy laws similar to EU's GDPR being enacted in various regions around the world.[1] Data teams also need to deal with increasingly diverse, fast-growing datasets and limited IT budgets.

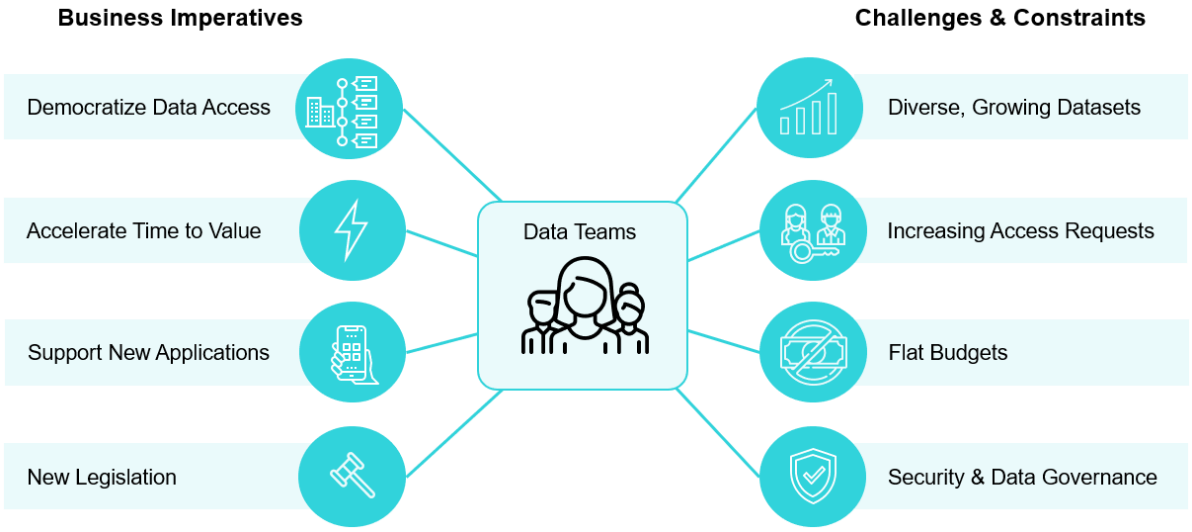

Figure 1 – Data teams face multiple business challenges

[1] Laws similar to GDPR include Australia's Privacy Act (APP), Canada's Personal Information Protection and Electronic Documents Act (PIPEDA) and the California Consumer Privacy Act (CCPA). A partial list of privacy legislation is presented in the article, 12 Countries with GDPR-like Data Privacy Laws.

To address the challenges illustrated in Figure 1, data teams have had to make compromises using different data management technologies for different purposes. For example, organizations often deploy a cloud-based data lake to store and analyze large volumes of raw data in various formats economically. They also typically operate a data warehouse for applications that require better performance and security and data governance controls. Before discussing the challenge of data copies, it is helpful to review how modern data management environments have evolved and how we came to have this challenge in the first place.

The Evolving Data Warehouse

For decades, enterprises have relied on data warehouses to collect information into a single location where it could be analyzed to help make better business decisions. While data warehouses have existed in various forms since the 1980s, technologies have evolved in distinct phases. Figure 2 provides a simplified view of how data warehouses and data lakes have matured and evolved.

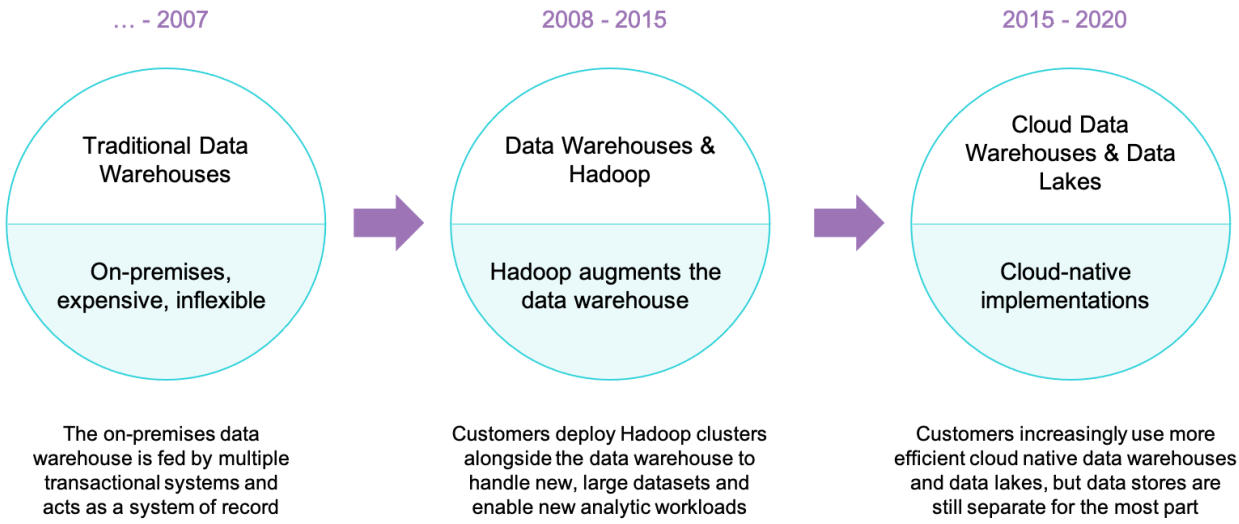

Figure 2 – The evolution of data warehouse and data lake technologies

Traditional Data Warehouses

In the early days, most data warehouses resided in on-premises data centers. Data warehouses were large relational databases that consolidated data from multiple operational systems to support decision-making efforts. At a time when storage was relatively expensive, organizations were careful about what information they warehoused. Data teams frequently handcrafted table schemas storing only the data that they needed. They also made sure tables and indexes were structured to support anticipated reporting requirements.

While early data warehouses generally provided good query performance and security features, they had drawbacks as well. They were expensive, ill-suited to storing unstructured or semistructured data, and SQL schemas needed to be tailored to support specific BI and reporting requirements.

Maintaining a data warehouse typically involved managing automated batch processing to extract, transform and load data from external data sources, and periodically purge or re-aggregate data in tables. Every time a

data source or table schema changed, ETL workflows needed to be modified and tested. Maintaining ETL workflows was manual, time-consuming and expensive — a challenge that persists to this day with many modern data warehouse solutions.

Data Warehouses and Hadoop

With the growth of the internet and cloud providers, customer interactions increasingly shifted online. As storage costs plummeted and the volume of easily collectible electronic information soared, organizations increasingly saw value in collecting and retaining this data for analysis. These datasets were diverse, including everything from document scans to log files to call center recordings to video files — all unstructured data types not easily stored in a traditional data warehouse.

Hadoop, initially developed in 2005, was one of the first technologies that made it practical to store and analyze vast amounts of unstructured and semistructured data at scale. Storing data in Hadoop was less costly than the data warehouse, so many organizations deployed Hadoop-based "data lakes" alongside the data warehouses continually fed from various sources. Data in Hadoop could be analyzed using native tools without impacting business operations. If the data proved valuable to the business, it could be loaded into the data warehouse.

While the idea was sound, and tools in the Hadoop ecosystem improved dramatically between 2008 and 2015, Hadoop clusters and tools proved to be too difficult to use and manage for most enterprises. The rise of cloud computing occurring at the same time impacted both data warehouses and data lakes. As cloud-based object stores matured and cloud data lakes became easier to deploy and manage, data lakes and data warehouses increasingly shifted to the cloud.

Cloud Data Warehouses and Data Lakes

Today, cloud providers typically offer their own cloud-native data warehouses. Other data warehouse solutions typically run across multiple clouds. Data lakes are usually deployed on scalable cloud object stores such as Amazon S3 or Azure Data Lake Services (ADLS). These cloud object stores have the advantage that storage scales independent of compute capacity, making them more economical than their Hadoop-based predecessors.

While modern cloud warehouses and data lakes have come a long way, there are still challenges. Organizations typically deploy both data warehouses and data lakes because they have complementary strengths. Data lakes are far more economical per TB stored and can more easily accommodate diverse, quickly evolving datasets. On the other hand, data warehouses typically provide better support for critical reporting and business intelligence applications and more robust security and data governance controls.

Despite the added cost and complexity of maintaining multiple data stores and storing data in proprietary data warehouse formats, both technologies have generally been needed to support the full range of business requirements and use cases faced by data teams.

Data Copies Are a Key Obstacle

An obvious problem with operating both a data lake and a data warehouse, besides the added cost, is that organizations end up replicating data across multiple systems. However, data copies occur for a whole variety of additional reasons, as illustrated in Figure 3.

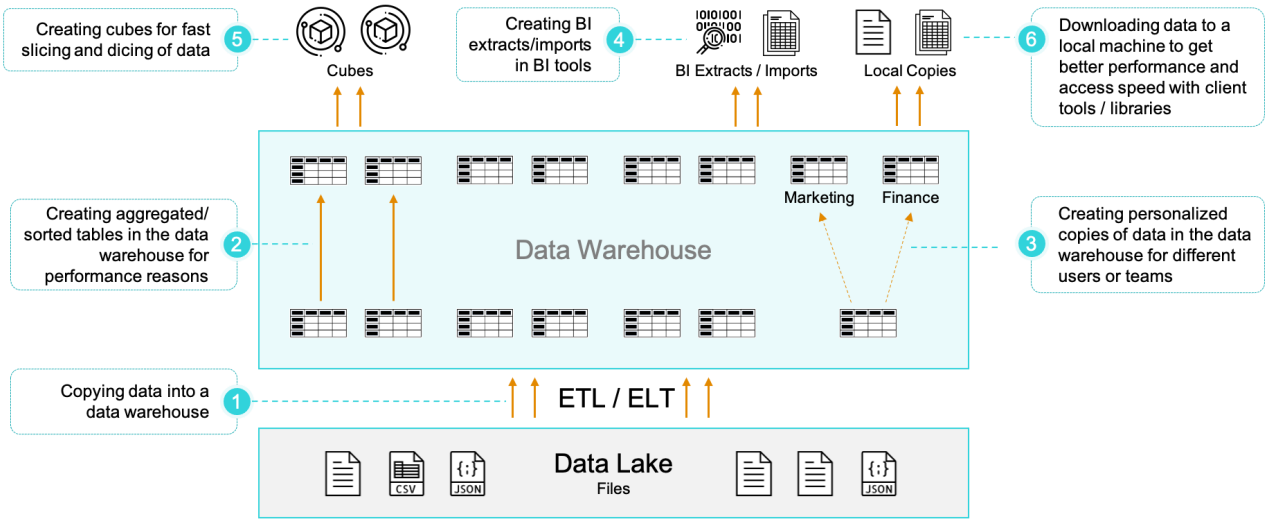

Figure 3 – Many copies of data exist in modern data warehouses and data lakes

Each of the following are potential sources of copies:

1. **Copying data into a data warehouse** – Raw data residing in a data lake is typically loaded into data warehouse tables via ETL/ELT batch processing resulting in additional copies of data.
2. **Performance-optimized copies** – Data is frequently replicated in a data warehouse for performance reasons. For example, data teams may load tables with aggregated or pre-sorted data so that particular queries and BI dashboards run more quickly.
3. **Personalized data copies** – Another source of replication is when copies of tables are provided for individuals or teams. For example, suppose a department wants access to a table with sensitive financial information. In that case, data teams may provide a copy of a cleansed data subset rather than risk allowing direct access to source data.
4. **BI extracts / imports** – BI tools such as Tableau and Power BI frequently operate on data extracts rather than maintaining a live connection to the data warehouse. Extracts are often used because analysts realize that they can work faster with local copies of data. In other cases, data warehouse operators may be concerned about live BI queries interfering with warehouse operations.
5. **OLAP cubes** – Batch ETL flows or BI tools may be used to create multidimensional OLAP cubes. Pre-computed cubes enable analysts to quickly slice and dice data and drill down or roll up data across multiple dimensions. OLAP cubes have become less popular with the advent of fast columnar databases. However, they still represent yet another copy of disconnected data.
6. **Local extracts and copies** – A final category of data copies occurs when users download extracts from the data warehouse for local processing using external libraries or tools.

Multiple copies of the same data are frequently stored in different formats across multiple systems for all the reasons above. In many cases, these copies need to be created, managed and maintained by data teams. Even worse, since datasets often change with time, data copies are frequently out of date. Not only do data copies add cost, but they are a nightmare for organizations concerned about data veracity, security, compliance and governance.

The High Cost of Data Copies

Data copies are created because data lake storage alone cannot deliver the query performance required by reporting, BI and data science applications. Unfortunately, these data copies come at a high cost. IDC estimates that as much as 60% of all storage is dedicated to managing copies of data at an estimated cost of USD 55B annually.[2] To make matters worse, organizations are collecting data at an increasing rate. According to Statista, data is projected to grow at a CAGR of 26% between 2020 and 2024, compounding this problem even further.[3]

Data copies add direct costs and impact the bottom line in multiple ways:

- **Direct storage costs** – Cloud storage costs vary in direct proportion to the amount of data stored. Users are typically charged per GB hour both in the data lake and for object or block storage consumed by the data warehouse at rates that depend on the storage's quality of service.[4] Data copies result in significantly higher bills from the cloud provider.
- **Increased data warehouse costs** – Replicating data in the data warehouse is expensive. While managed storage costs in the data warehouse may seem comparable to the data lake at first glance, the data warehouse is typically much more expensive. Users will pay incrementally based on the number and type of compute instances comprising the data warehouse.[5] Also, fees may apply every time the data warehouse scans the object-store. Additional usage-based fees may apply for data catalogs and key management services.[6]
- **ETL/ELT processing** – The ETL/ELT processing used to transform and load data also adds to the cost. In ETL processing, data is typically extracted into intermediate storage such as S3 buckets before being processed and copied into the data warehouse. Also, cloud providers levy fees for ETL pipeline execution in addition to the fees above related to storage, compute and data scanned per operation.
- **Lost productivity and slower time to value** – As more stakeholders require access to enterprise data, data engineering teams can find themselves backlogged with requests. Providing users with access to data in suitable formats often involves altering data schemas and building or modifying ETL workflows. For example, particular clients may need data tables with columns indexed or aggregated on particular fields, or table schemas may be intentionally de-normalized to accelerate particular

---

[2] IDC data copy estimates

[3] CAGR calculation based on project data growth of 59 ZB in 2020 to 149 ZB in 2024 (source: Statista)

[4] Block storage costs typically vary based on factors such as IOPS and throughput in addition to the amount of storage. Object storage pricing varies depending on the storage tier and the frequency with with data is accessed.

[5] Data warehouse node costs can be high. For example, an RA3 node type typically used with AWS Redshift (ra3.4xlarge) will cost $3.26 per hour at on-demand published rates.

[6] Policies may vary depending on the data warehouse and cloud provider. See AWS Redshift Spectrum pricing. Redshift Spectrum charges are approximately $5.00 per TB scanned depending on the cloud region and incremental charges for catalog access and key management services.

types of queries.[7] Creating and validating new workflows can take weeks or even months, placing a burden on data engineering teams. It also results in delays and slower time to value for business analysts and data scientists.

- **Security and compliance risks** – Security and auditing are critical requirements in many regulated industries such as financial services and healthcare. Having multiple copies of the same data makes data security and governance a nightmare. When multiple inconsistent copies of data proliferate, it becomes difficult to control access and ensure accuracy. Also, extracting data into ungoverned platforms (BI extracts, cubes and local data copies as examples) makes it impossible to secure and audit data access. Data copies can lead directly to compliance-related risks.

Analysts and data scientists need a better solution rather than creating layers of copies. Ideally, they would analyze their data directly in the data lake at interactive speed, without needing to copy data into expensive data warehouses or create performance-optimized copies in other data stores. To achieve this, they need a high-speed query engine that works with existing BI tools and provides security and data governance features equivalent to those in a data warehouse.

Fortunately, modern data architecture makes it possible to query data lakes directly at interactive speed, avoiding the need for a data warehouse and redundant data copies.

The Dremio Data Lake Service

Dremio's data lake service supports lightning-fast queries directly on the data lake, self-service data access and advanced data governance and security features. With these capabilities, Dremio avoids the need for complex ETL/ELT processing, data warehouses and performance-optimized data copies.

Rather than loading data into an intermediate data warehouse to achieve adequate query performance and data governance controls, Dremio queries the data lake directly. Dremio provides rich security and data governance features and supports multiple workloads, including reporting, BI, ad hoc queries and analysis and data science applications. Dremio also works seamlessly alongside streaming services used to ingest data or machine learning services such as Spark, Databricks, HDInsight and others.

Key components that comprise the Dremio data lake service are shown in Figure 4.

[7] Source: When and How You Should Denormalize a Relational Database

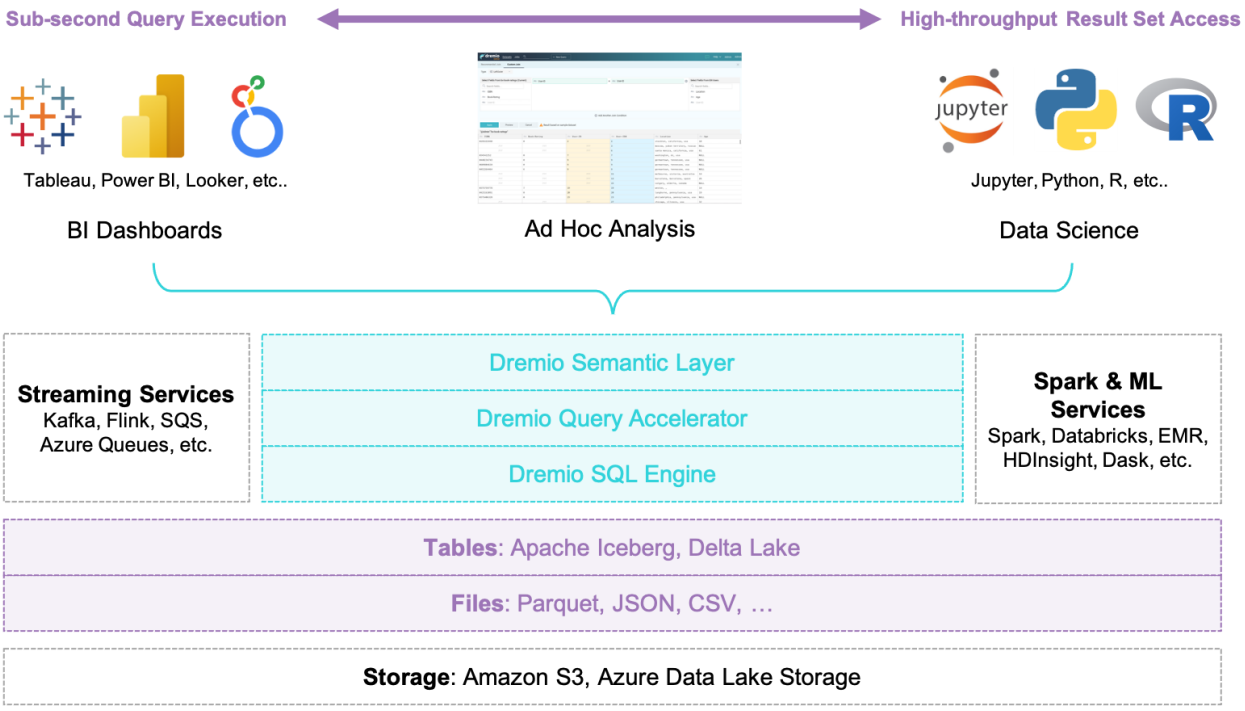

Figure 4 – A simplified view of the Dremio data lake service architecture

Lightning-Fast SQL Queries

Traditional SQL Engines such as Presto and Athena are too slow to query data lakes directly. Dremio leverages various open-source technologies, including Apache Arrow and Gandiva, to accelerate SQL queries dramatically. These technologies combine with Dremio features such as data reflections to deliver between 4-100x the performance of traditional data lake SQL engines.

Dremio works seamlessly with standard open file formats such as Parquet, ORC, Avro and JSON. It also supports multiple metastores, including the Hive Metastore (HMS) and AWS Glue Catalog. Dremio can also query a wide variety of non-data lake sources, including relational databases, data warehouses, file systems and NoSQL data sources such as MongoDB, Cassandra and Elasticsearch.

A Self-Service Semantic Layer

In addition to delivering superior performance, Dremio includes a self-service semantic layer providing multiple virtual views into the data in the data lake. Virtual datasets enable data analysts and engineers to manage, curate and share data while adhering to centralized data governance and security policies. This is achieved without the overhead and complexity of copying and moving data. Virtual datasets are fully indexed and searchable.

The semantic layer also supports granular role-based access controls integrating with LDAP, Active Directory and OpenID-based services for a seamless user experience. Business applications can access these virtual datasets just as they would access a traditional SQL data warehouse via industry-standard ODBC/JDBC connectors, REST API calls or optimized Arrow Flight-based interfaces.

Open Table Formats, Advanced Transactional Capabilities

Open table formats, including open source Apache Iceberg and Delta Lake (from Databricks) mean multiple engines (Spark, Dremio, etc.) can operate on the same datasets. These table formats bring capabilities to the data lake previously found only in proprietary data warehouses, including:

- Concurrent transactions
- Record-level mutations (updates, deletes, etc.) across large datasets
- Time travel to query historical databases
- Data versioning

Key Technologies

Dremio has helped pioneer many of the new technologies that make a modern data lake engine possible:

- *Apache Arrow* – an open source project enabling columnar in-memory data processing and interchange with distributed, vectorized execution (SIMD or GPU)
- *Arrow Flight* – A parallel zero-copy RPC protocol that exchanges Arrow format memory buffers between a client and a data lake engine
- *Gandiva* –An open source LLVM-based compiler in Apache Arrow that translates queries into vectorized execution kernels for efficient computations
- *Columnar Cloud Cache (C3)* – A real-time, distributed, NVMe-based caching
- *Data Reflections* – An optional query acceleration technique in Dremio that transparently re-writes queries to use internal pre-aggregated and sorted materialized views maintained by Dremio, providing a highly optimized physical representation of source data with granular reuse
- *Massively parallel high-performance readers* – enabling fast parallel queries using Arrow Flight

Dremio continues to deliver new technologies and innovations to accelerate data lake queries.

Enabling a "No-Copy" Data Strategy

Among the many advantages that Dremio offers for data lake and data warehouse environments, it enables a no-copy data strategy. It does this by sidestepping the various technical and performance limitations that have historically required data engineers and users to rely on data copies.

Dremio enables a no-copy data strategy by:

- **Avoiding replication in a data warehouse** – Rather than loading copies of data into curated tables in the data warehouse, BI and reporting applications can query data directly in S3 or ADLS via Dremio, delivering comparable performance at a fraction of the cost.
- **Eliminating the need for performance-optimized copies** – Dremio makes optimized, aggregated or sorted tables in the data warehouse a thing of the past. Rather than create these additional data copies, users can use data reflections, fully managed by Dremio and transparent to the user. Data reflections dramatically accelerate queries while leaving data in place.
- **No personalized copies** – The semantic layer in Dremio makes it easy to quickly provision different logical views without physically copying the underlying data.
- **Avoid BI extracts/imports** –Analysts can achieve performance goals by querying the data lake with a live connection from the BI tool avoiding the need for disconnected data extracts.

- **No data science exports** –Key Dremio technologies such as Apache Arrow Flight enable 10-50x the performance with pyodbc, enabling live data access and avoiding the need to work on local copies.[8]

Dremio Provides Clear Business Advantages

By deploying Dremio, organizations can dramatically reduce the need for data copies while enjoying additional benefits such as:

- Reduce or eliminate the need for costly data warehouse infrastructure by running queries directly against the data lake
- Reduce the cost, overhead and inefficiencies of complex and brittle ETL/ELT workflows to transform data by querying data in place via high-performance query optimizations
- Easily accommodate new data access requests via Dremio's semantic layer, without the need for extensive data engineering work leading to faster time to value and increased productivity
- Provide the security and data governance features of a modern data warehouse without the cost and complexity of replicating data into a cloud data warehouse

Additional Sources of Savings

With a next-gen data lake architecture, organizations can reduce costs in multiple areas. Not only will organizations avoid storage costs due to replicated data, but they will avoid the costs associated with a cloud data warehouse, including infrastructure costs as well as additional fees for table scans, catalogs and other usage-based fees that can quickly add up.

Dremio provides a multi-engine execution model deployable on premises, in the cloud or on Kubernetes cloud services to achieve exceptional performance and scalability. With a query speed that is 4-100x faster than traditional SQL engines, compute instance sizes and costs can be reduced by over 75% while delivering the same performance. Dremio eliminates both the under- and over-provisioning of compute resources by taking advantage of AWS's underlying elasticity. Elastic engines scale automatically, maximize concurrency and performance while reducing cloud costs by an additional 60% or more. When coupled with Dremio's dramatic performance advantages, AWS users can save up to 90% of total infrastructure cloud costs.[9]

Your Choice of Clouds, Tools and Data Sources

Many cloud data warehouses are specific to a single cloud provider, and store data in their own internal formats locking organizations into a particular cloud platform and storage service. Other data warehouse solutions may support multiple clouds. However, they still require that data be ingested into the data warehouse where it resides in a proprietary format, effectively locking the customer into the data warehouse platform. Neither of these solutions addresses the problem of data copies, and both can result in vendor lock-in.

---

[8] Performance advantages of Apache Arrow with pyodbc are discussed in this webinar and transcript.

[9] Costs savings are explained in the article Introducing Dremio AWS Edition, Delivering Data Lake Insights On Demand.

Dremio is unique in that it decouples compute from data, providing flexibility for current and future needs. Customers can store their data in open formats in their preferred cloud. Dremio users can run their chosen analytic and data science tools, avoiding lock-in to a particular cloud platform or data warehouse provider.

With Dremio, organizations can:

- Run workloads with their preferred cloud provider, including AWS and Azure, or on premises. Customers can also deploy Dremio on cloud VMs, using Kubernetes services such as AKS or EKS, or deploy on YARN (Hadoop) or Docker-based environments.
- Use their choice of BI and analytic tools, including Tableau, Power BI, Looker, Jupyter and others, to directly query data residing in the data lake with full-featured data security and governance controls.
- Query external data sources directly from Dremio, such as cloud databases, data warehouses and NoSQL data stores, simplifying operations and centralizing data management.

Conclusion

Disconnected and unmanaged data copies are a challenge for all organizations. Data copies are frequently required to meet performance objectives, provide personalized data copies, and support business intelligence, reporting and data science requirements. The ETL/ELT workflows frequently used to create data copies result in slower access to data, significant data engineering work, and business delays. All of these factors lead to increased costs and reduced business efficiency.

Fortunately, new developments in data lake engine technologies can help organizations implement a "zero-copy" architecture. Users can enjoy lightning-fast queries and sophisticated data governance and security features while querying open table formats directly on the data lake. Data teams are freed from the burden of creating query-optimized data tables and custom workflows for users. Instead, they can quickly create secure, high-performance data views easily accessible from popular BI, reporting and data science tools.

Dremio data lake service enables enterprises to realize a zero-copy data architecture, keeping a single source of data in the data lake. Analysts, business users and data teams enjoy excellent query performance, a simpler lower-cost architecture, sophisticated data governance features. Dremio also enables faster time to value, increased operational efficiency, and the flexibility to use their choice of clouds, tools and data sources.

To learn more about Dremio and its no-copy data architecture, visit http://dremio.com. To quickly launch a free on-demand instance to test drive Dremio, visit http://dremio.com/testdrive/.

About Dremio Corporation

Dremio reimagines the cloud data lake to deliver faster time to analytics by eliminating the need to copy and move data to proprietary data warehouses, or create cubes, aggregation tables and BI extracts. A self-service semantic layer provides flexibility and control for data architects, and self-service for data consumers.

Founded in 2015, Dremio is headquartered in Santa Clara, CA. Investors include Cisco Investments, Insight Partners, Lightspeed Venture Partners, Norwest Venture Partners, Redpoint Ventures and Sapphire Ventures. For more information, visit www.dremio.com.

Connect with Dremio on GitHub, LinkedIn, Twitter and Facebook.