dremio

5 Best Practices to Completely Eliminate Costly Data Copies with a SQL Lakehouse Platform

Organizations constantly face the challenge of disconnected and unmanaged data copies. The inherent design of data warehouses and data warehousing processes leads to the creation of multiple, redundant data copies throughout the organization's architecture, exponentially increasing infrastructure setup and maintenance costs as well as time to value.

Fortunately, new developments in SQL lakehouse platform and data lake engine technologies help organizations implement a "zero-copy" architecture. You can enjoy lightning-fast queries and sophisticated, enterprise-grade data governance and security features by querying open table formats directly on the data lake, without the need to create multiple data copies and redundancies!

Top 5 Costs of Expensive Data Copies

1 Direct storage costs

Cloud storage costs vary in direct proportion to the amount of data stored and the multitude of data copies in the data warehouse result in significantly higher bills from the cloud provider.

2 Increased data warehouse costs

The inherent dependency on the proprietary data warehouse vendor results in several types of costs such as: incremental costs based on the number and type of compute instances in the data warehouse, additional fees to scan the object store, additional usage-based fees for data catalogs and key management services and more. Storing and retrieving every piece of data to and from the data warehouse for reporting or analytics dashboards ends up with a hefty price tag and impacts the company's bottom line.

3 ETL/ELT processing costs

In ETL processing, there are several costs that add up significantly, such as: cost of intermediate storage before data is processed and copied into the data warehouse, fees for ETL pipeline execution, fees for storage, compute, and data scanned per operation, etc.

4 Costs due to lost productivity and slower time to value

Productivity loss and the costs associated with backlogs and significantly slower time to insight are caused by: data access backlogs with increasing number of user requests, inability of data warehouses to provide self-service access to data, time and effort required to alter data schemas and build or modify ETL workflows, time required to create custom requests for indexes, cubes, and aggregations, etc. Delays of weeks or even months increase the burden on data engineering teams and also results in delays and slower time to value for business analysts and data scientists.

5 Security and compliance costs

The proliferation of multiple inconsistent copies of the same data makes security and governance a nightmare, especially in regulated industries such as financial services and healthcare. Extracting data into ungoverned platforms (BI extracts, cubes, and local data copies, etc.) makes it impossible to secure and audit data access and leads directly to compliance-related risks and regulatory fines.

5 Ways to Eliminate Data Copies with Dremio's "No-Copy" Data Strategy

1 Avoid duplication of data in a data warehouse

Ditch old, outdated technologies such as data warehouses that lead to multiple data copies! Rather than loading copies of data into curated tables in the data warehouse, Dremio queries data directly from cloud data storage such as S3, GCS, or ADLS, delivering comparable performance at a fraction of the cost.

2 Eliminate the need for performance-optimized copies

Dremio makes optimized, aggregated, or sorted tables in the data warehouse a thing of the past. Rather than creating additional data copies, you can take advantage of data reflections — transparent to the user and fully managed by Dremio. Data reflections dramatically accelerate queries while leaving data in place.

3 End the need for personalized copies

The self-service semantic layer in Dremio makes it easy to quickly provision different logical views without physically copying the underlying data.

4 Do away with BI extracts/imports

Analysts can achieve performance goals by querying the data lake directly with a live connection from the BI tool of their choice and and completely eliminate the need to create disconnected data extracts.

5 Remove the need for data science exports

Key Dremio technologies such as Apache Arrow Flight enable 10-50x faster result set transfer than JDBC and ODBC, enabling live data access and avoiding the need to work on local copies.

Read this whitepaper to learn how you can use the Dremio SQL Lakehouse Platform to enable your data engineers, data analysts, and business users to achieve their goals without the need to create multiple expensive data copies!

dremio

Dremio is the SQL Lakehouse company. Dremio simplifies data engineering and eliminates the need to copy and move data to data warehouses, providing flexibility and control for data architects and engineers, and self-service for data consumers. Organizations enjoy high-performing dashboards and interactive analytics directly on the data lakehouse, with enterprise-grade security and data governance.