dremio

EBOOK

Data as Its Own Tier

Data is its own tier — consisting of multiple layers of open source technologies.

Data is the key asset for all organizations across industries, and over the years leading companies have realized the need to democratize it, in order to provide increased data access to more and more people, tools and applications. These leading companies have **four main principles** that have shaped the trends and evolution of data analytics and data infrastructure over the past decade:

Cloud	Flexibility and agility
Scalability	Availability

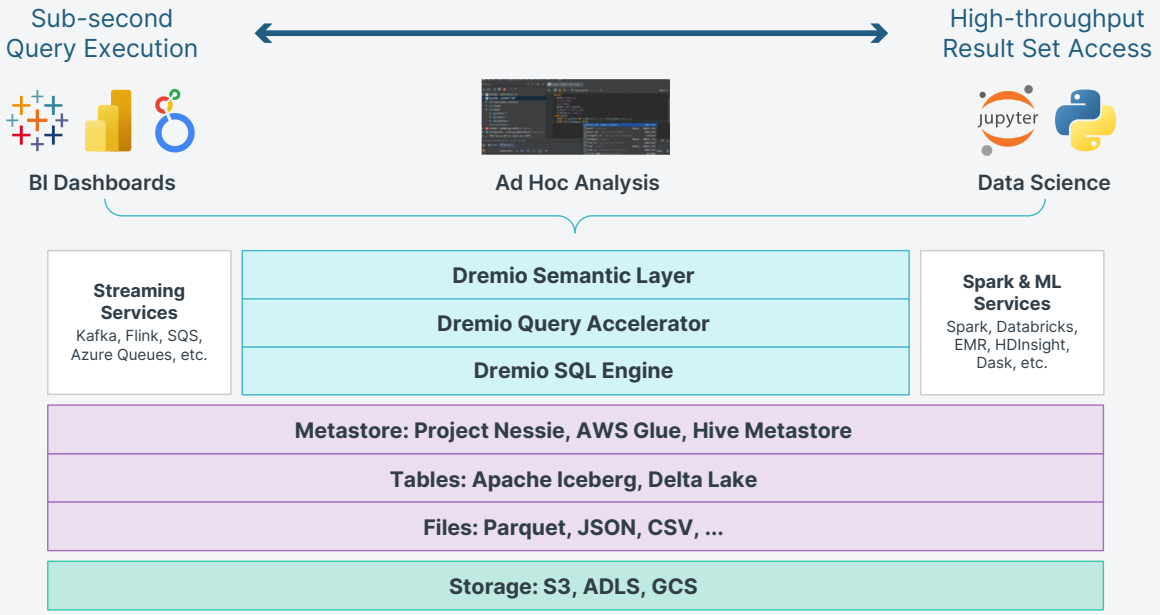
Data architectures have **evolved** from:

- The co-location of storage and compute in purely monolithic architectures (prior to 2015)
- To the separation of compute and storage in cloud data warehouses (from 2015-2020)
- To the separation of data and compute in today's data architectures (post-2020)

So, data now resides as its own separate tier.

This eBook explains how multiple open source technologies make up this new data tier and the advantages of this revolutionary architecture to your organization.

Dremio Is the Data Lake Service for SQL-Based Workloads

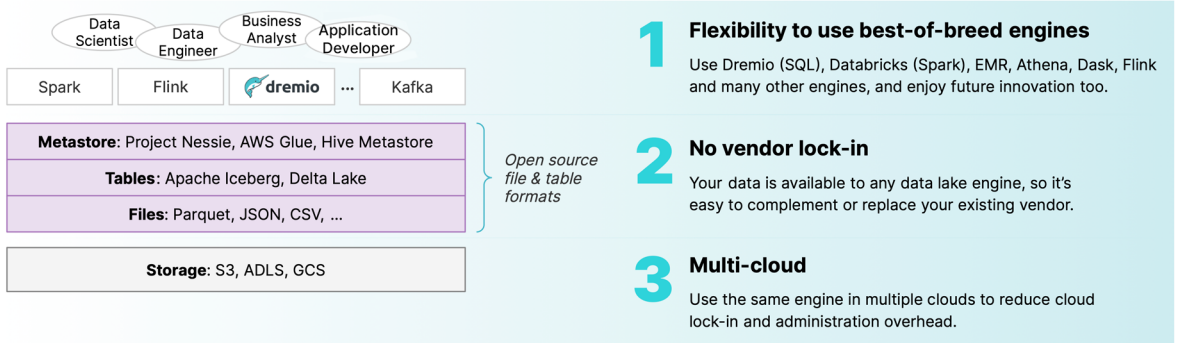


The modern (post-2020) data lake architecture consists of **three layers of open formats and standards** that not only bring the capabilities of a traditional data warehouse to the data lake as well as net-new capabilities that data warehouses can't provide, but also allow users to pick and choose the technology tools and engines best suited for each need or that they are most comfortable with.

One of the key benefits of the new data architecture is that it is open. The storage system could be S3, ADLS, GCS, etc., and all the layers above the storage layer — the data layers that determine how the data is represented — are all **open** source.

Open source file formats such as Parquet, open source table formats such as Apache Iceberg or Delta Lake and even the metastores are all open source and industry standards.

The Open Data Architecture



This open architecture has the following **advantages**:

- **Flexibility to use best-of-breed engines:** Users have the flexibility to use any surface tool or engine of their choice (such as Dremio, Spark, EMR, Athena, Dask, Flink, etc.), and as noted by Werner Vogels, users also have the assurance to enjoy future innovations.
- **No vendor lock-in:** An open architecture helps companies break away from the "vendor lock-in" situation. When companies put their data into a vendor's system, there is a dependency on the vendor to access that data, which locks the company with the vendor. This leads to disadvantages such as the lack of flexibility to move to better alternative systems and innovations and unavoidable financial burden over time. When companies moved to the cloud, those that decided to go with cloud data warehouses found themselves running into the same vendor lock-in and cost issues they had with their on-premises data warehouses. On the contrary, the new open data architecture gives companies the flexibility to switch to new tools or engines, and keep pace with the latest innovation without any forced dependencies on a single vendor.
- **Multi-cloud versatility:** The open data architecture is multi-cloud, which means that the majority of engines, such as Dremio, Spark, etc., work across different clouds. This makes it very easy for a company to use the same set of technologies, internal training and education, administration, and approaches, regardless of which cloud the company uses. This simplifies operations, as companies nowadays use several different clouds for different use cases and different groups.

> With a data lake, data is stored in an ***open format***, which makes it easier to work with different analytic services. Open format also makes it more likely for the data to be compatible with tools ***that don't even exist yet.***
>
> Various roles in your organization, like data scientists, data engineers, application developers, and business analysts, can access data with their choice of analytic tools and frameworks.
>
> **Werner Vogels,** CTO, Amazon

The **New Data Tier** is composed of ***three layers*** of **open source technologies:**

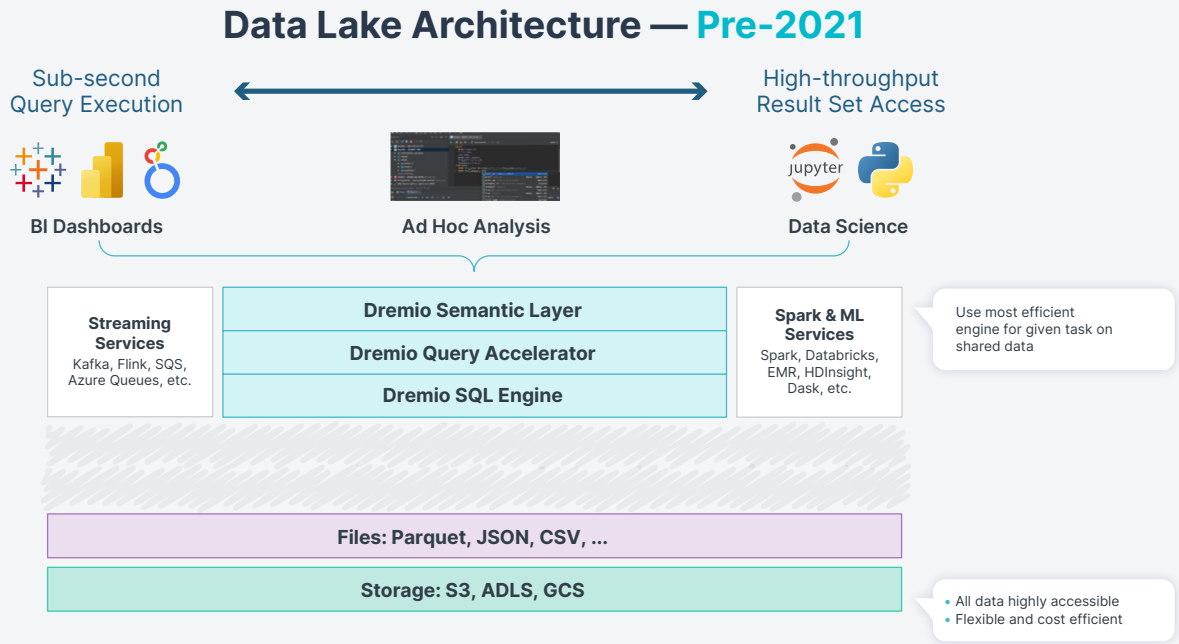
1. Open source file formats

Early on in the adoption of data lakes, standardized file formats emerged as the standard way to store data in them (e.g., Apache Parquet, JSON, CSV, etc.), and users had the flexibility to read these different file formats using various tools and engines, as per their convenience and for different, specific purposes.

In spite of the advantages open source file formats offer, there are some **challenges**, including:

- Synchronization of data and operations across the multiple tools and engines that are used
- Limitations on the types of functionality that you can get on top of the data lake tables

To address these constraints, additional open source technologies have been introduced, bringing the capabilities of the data lake beyond those of the data warehouse.



Open source file formats are ideal because:

- Data is highly accessible
- Data exists in an open format, offering flexibility as different tools can read the data
- Data storage is affordable and cost-efficient
- Users can take advantage of different engines based on the specific need and operation to be performed; for instance, Dremio may be the ideal engine for one operation such as SQL, and Spark may work better for another

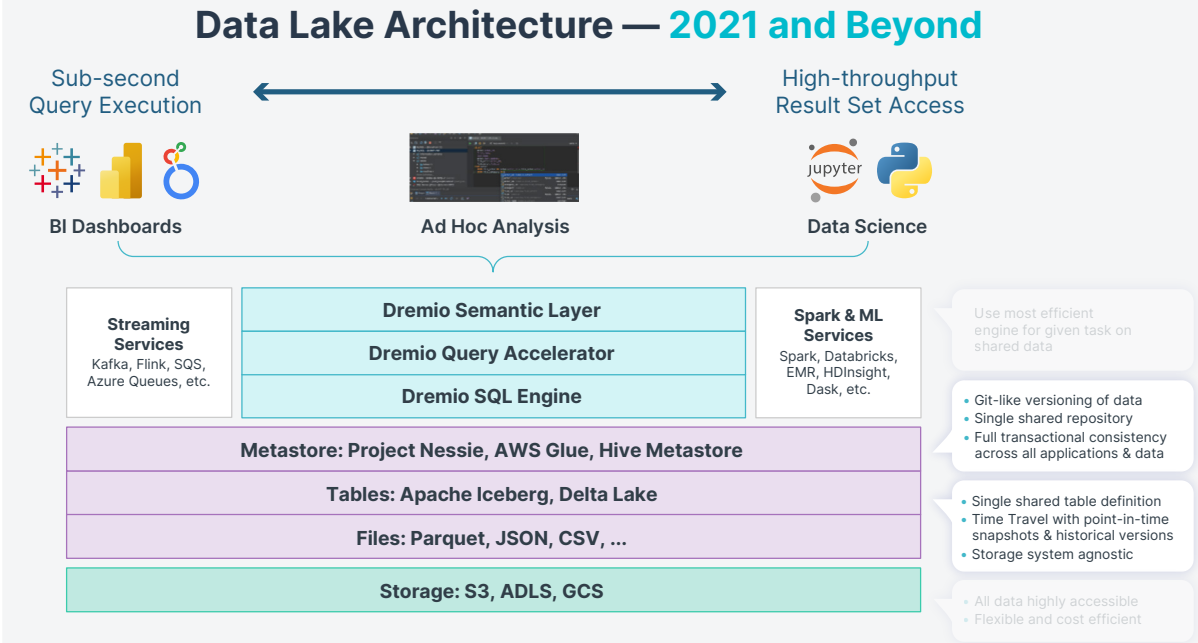
2. Open source table formats

One of the revolutionary changes happening in the world of data lakes in 2021 is the adoption of **open data table formats**. Open source projects such as Apache Iceberg and Delta Lake by Databricks are the new revolutionary industry standard projects this year that essentially allow you to think of your data lake datasets as tables, not just as a collection of files — it's a high-level abstraction.

Netflix created Iceberg, and this open source project has generated a lot of traction recently and has been embraced by several tech companies such as Amazon, Alibaba, Airbnb, Dremio, Expedia, Adobe, LinkedIn, Apple, and more.

Dremio will be integrating the concept of open source table formats, such as Apache Iceberg, which will have the following advantages:

- Open table formats create a single shared definition of a table across multiple different tools and engines. For instance, with open table formats such as Iceberg, users can leverage capabilities such as transactional consistency, record-level updates and record-level deletions, across different engines.
- Open table formats enable transactions and data mutation with SQL DML (insert, update and delete records).
- Open table formats also offer time travel, the ability to query data as it was in the past, with point-in-time snapshots and historical versions.
- Dremio facilitates easy promotion of data lake files and folders to Iceberg datasets. With a single click, the user can simply treat a directory of files as an Iceberg table, making Iceberg the single source of truth for metadata.
- Dremio also leverages similar technologies internally to solve some of the scalability problems that are inherent to data lakes. For example, there is no longer a data limit on metadata size in Dremio. Your data can have millions of splits and millions of files, without having to wait for the metadata to be refreshed.



- Dremio, in conjunction with Iceberg, will also be storage-system agnostic. Hence the SQL DML transactions and the time travel capabilities will be possible just like standard SQL, but in the data lake, this will be available across multiple engines.

Essentially, open table formats provide a coordination layer between multiple different tools and engines. For example, suppose Spark or Presto makes updates to a table, such as the addition of a partition, changing a column, addition of a row, etc., Dremio will immediately be in sync with those updates regardless of the velocity and nature of the table updates. This helps to create transactional consistency in the organization, and across the multiple tools, no matter which engine is used for the transaction.

Leveraging the capabilities of open table formats such as Iceberg enables multiple applications to work together on the same data in a transactionally consistent manner, which solves several operational complexities for an organization and ultimately leads to enormous time and cost savings.

Apache Iceberg: Backed and Used by Major Tech Companies

Learn more about Apache Iceberg in our article: What Is Apache Iceberg?

3. Open metastore: Nessie

Dremio takes the user experience beyond what you can do in a data warehouse, with the creation of an **open source project** called **Nessie**, which has revolutionized what is possible in the world of data infrastructure.

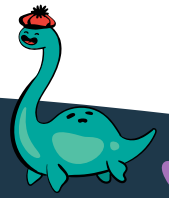
Nessie is a modern and much more sophisticated **metastore**, as compared to its predecessors. As the open source metastore layer, Nessie builds on top of and integrates with the following open table formats: Apache Iceberg, Delta Lake and Hive.

Designed to meet today's infrastructure needs, such as cloud and open table formats (e.g., Iceberg), project Nessie offers a **Git-like experience** for the data lake. Similar to Git, the concepts of branches, tag, release, etc. are part of the Nessie experience.

Nessie provides two categories of advantages:

1. Nessie enables the ability to query branches, tags and times in the past.

- **Reproducibility:** Provides the ability to go back in time and look at data in the past by tagging a specific point in time and querying that tag. For example, a user can go back and see how they got a specific number or display a dashboard based on yesterday's or last year's data. In the world of machine learning, the user can recreate a model from the past.



- **Dataset comparison:** This enables users to analyze how a particular dataset changed from yesterday to today, in order to find out if something was broken or if someone tampered with the data. The ability to see exactly what changed, when and who changed it, is invaluable information.
- **Compliance:** In the world of compliance and regulated industries, the ability to understand the whole history of what happened to data is very powerful.

2. Nessie also enables both multi-table and multi-engine transactions.

- **Multi-table consistency across multiple engines:** This is very powerful because it not only provides the ability to query multiple tables, but multiple tables from multiple different engines such as Dremio, Spark, etc.
- **Experimentation:** If the company wants to try something new, they can create a separate branch and try out their idea. Depending on the success of the idea, they can either merge that experimental branch into the main branch, or delete or save it.
- **Data promotion workflows:** It is common for companies to have separate environments for dev → stage → prod. With Nessie, instead of having separate environments with three separate copies of data, leveraging branches is much more efficient than creating multiple copies of the same data.

Nessie: A Git-like Experience for the Data Lake

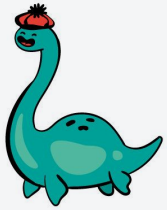
Queries on branches/tags/times

- Reproducibility
- Compare datasets
- Compliance

Multi-table & multi-engine transactions

- Multi-table consistency
- Experimentation
- Data promotion workflows (dev→stage→prod)

```
USE BRANCH 'main'
SELECT * FROM t1 // main implicit
SELECT * FROM t1@etl
SELECT * FROM t1 AT '2020-10-26'
SELECT * FROM t1@etl '2020-10-26'
```



```
CREATE BRANCH etl
[etl] Spark Job 1
[etl] Spark Job 2
[etl] Reflection Refresh 1
[etl] Reflection Refresh 2
USE BRANCH main
MERGE BRANCH etl
```

This is a Git concept copied from the past decade of software development into the world of data lakes and data architecture, which has helped bring data into the next era.

Learn more about Project Nessie in our Nessie: Git for Data Lakes article.

Reimagine Your Data Architecture with the New Data Tier

Data has emerged as one of the key assets of every company and, as a result, exists as its own separate tier within an organization's architecture.

In response, multiple open source technologies have emerged to enable a completely new data architecture — one that doesn't just separate the compute from the storage, but actually separates the compute *from the data*.

By separating the compute, data is placed squarely in the center of the data lake architecture — the New Data Tier.

The New Data Tier offers today's organizations more substantial and immediate benefits:

- Eliminates data silos and redundancies
- Facilitates Dremio's "no-copy" architecture approach
- Offers users the flexibility of universal data access from unlimited engines and applications via open data standards and formats
- Gives organizations the freedom to choose and use the best-of-breed solutions they prefer
- Removes vendor lock-in so companies are not bound to one vendor or technology and can adopt the latest and best innovations

With all these advances, data essentially becomes its own tier, enabling us to think about data architectures in a completely revolutionary way.