Starburst

E-book

Why a modern data lake architecture is essential for data-driven organizations

Overview

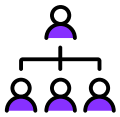
A data lake is a centralized repository that holds large amounts of structured, semi-structured, and unstructured data. This raw data is stored in its native format, regardless of data type or structure, and can be imported from multiple sources, including apps, clickstream data, social media, and IoT devices. All data in all its forms goes into one spot.

Centralizing most if not all of your data in one spot opens up exciting possibilities for data analytics, but establishing a lake is only the first step toward building a truly modern data lake architecture. A number of critical choices need to be made along the way to avoid letting your lake devolve into a data swamp that is more complex and expensive to manage than traditional systems. When properly designed, however, a modern data lake architecture eliminates such problems and enables a low-cost, high-performance, feature-rich, and future-proof data analytics stack.

Avoiding a data swamp

Originally, data lakes appealed to large organizations because unlike relational databases, they did not require data to be organized or transformed as it was ingested. Data could simply be stored in the lake in its native format, then transformed or prepared later. Organizations could leverage the lower costs of object storage and analysts could later extract, load, and transform (ELT) the data into desired formats for rapid analysis. As more companies adopted these early data lake architectures, however, they began to run into several problems.

Pitfalls of legacy data lakes



Poor management: With insufficient planning and foresight, data lakes can devolve data swamps beset by management and governance problems that limit the organization's ability to find and utilize data.

Governance & security: Most data lakes are designed to democratize data access, allowing for self-service access to data without the need for central IT. This has its benefits, as it enables more people to work with data, but it also creates security risks as more users may be able to access potentially sensitive datasets.

Performance: While data lakes are highly performant under many conditions, this is not a given – the scale of the lake and surrounding technologies deployed to organize and analyze the data inside can impact performance significantly.

Compliance: The use of a single repository for all data does not free organizations from adhering to compliance specifications and regulatory requirements; an ideal modern architecture strikes a balance between compliance and versatility.

Cost: The modern data lake offers a low-cost storage alternative to traditional warehouses, but this does not guarantee a lower TCO; management, engineering, compute and other costs can be higher than expected without the proper architectural considerations.

The data lake itself is a relatively new innovation, so what is a modern data lake architecture? Ultimately, a modern data lake architecture should provide the framework for a scalable, performant, flexible, cost-effective, and feature-rich big data storage and analytics stack.

What makes a data lake modern?

Before we discuss how to set up a modern data lake, let's look at what you should expect this architecture to deliver.

Cost-effective scalability

Data lakes allow you to store petabytes of data in a more cost-effective storage medium; their ability to accept data of any structure further lowers costs.

Separation of storage and compute

Previously, the same machines housed compute and storage resources, forcing them to scale alongside one another. Data lakes allow storage and compute to scale independently as needed, which helps control compute spend.

Availability to analytics engines

Data lakes are designed and optimized for querying and can be analyzed using a variety of analytics engines, including Hadoop, Trino, and Spark, plus machine learning, data visualization, and other tools.

Centralization of diverse datasets

Data lakes combine multiple datasets, formats, and data structures in one place, then connect with intelligent search and retrieval systems to make it easier to find the data you need.

Reduced need for ETL

Data moving into a data warehouse needs to be schematized through an Extract, Transform, and Load (ETL) process; data lakes accept data in its raw format. ETL operations may still be used on an ad hoc basis, but they are not required to stock the lake.

The next step is understanding how to leverage these benefits and design a data lake architecture optimized for analytics.

Modern data lake architecture

There are five core components to a modern data lake:

1. **Commodity storage and compute**
2. **Open file and table formats**
3. **High performance, scalable query engine**
4. **Fast, easy access to date outside the lake**
5. **Seamless connectivity to BI and data science tools**

Commodity storage and compute

Today cloud object storage is often the first choice for data lakes as it offers both cost-effective scale and separation of storage and compute. The three largest providers of cloud data lake storage services today are Amazon S3, Microsoft Azure Blob Storage or Azure Data Lake, and Google Cloud Storage. A data lake that is not built on commodity storage and compute cannot take full advantage of the benefits detailed previously.

Elastic compute is essential for scaling modern data lake analytics, allowing dynamic adjustments of computational resources based on demand. This flexibility ensures efficient processing of vast amounts of structured and semi-structured data while minimizing costs by only utilizing resources as needed.

Open file and table formats

When you stream data into a traditional or cloud data warehouse, your data is converted into a proprietary format, and is therefore not really your data any longer, as it is subject to vendor lock-in. A modern data lake leverages open file formats such as Parquet, ORC, and Avro.

Once you select an open file format, you need to select a table format to organize that data. The leading formats today are Apache Iceberg, Delta Lake, and Apache Hudi. Iceberg has emerged as an ideal high-performance format for very large datasets that performs efficiently at scale.

High performance, scalable query engine

Legacy data lakes and data warehouses such as Hadoop and Teradata attempted to address this problem by requiring all data to be landed in their systems. But it was hard for data analysts to run interactive queries on such large datasets without them failing. Trino was created in 2012 at Facebook to address this problem. It enabled Facebook to run analytics on their Hive/ Hadoop data lake at petabyte scale without the need for unnecessary, costly data movement.

Today, Trino is a widely used, massively parallel processing (MPP) query engine for rapid, SQL-based analytics of big, distributed data. Companies such as LinkedIn, Lyft, Netflix, GrubHub, and many others have embraced Trino as a critical data architecture component to accelerate their access to data.

The modern data lake is a great way to lower your storage costs and reduce ETL and maintenance. Ultimately, though, you want to be able to analyze this data in a highly performant way so you can extract business value out of all that data. For this you need to layer on a high-performance, scalable query engine that is optimized for the modern data lake.

Trino at Scale

25PB on S3

1 Exabyte of Data
100PB weekly data
1200 nodes
2.5M queries/week

600PB on S3
1000 nodes

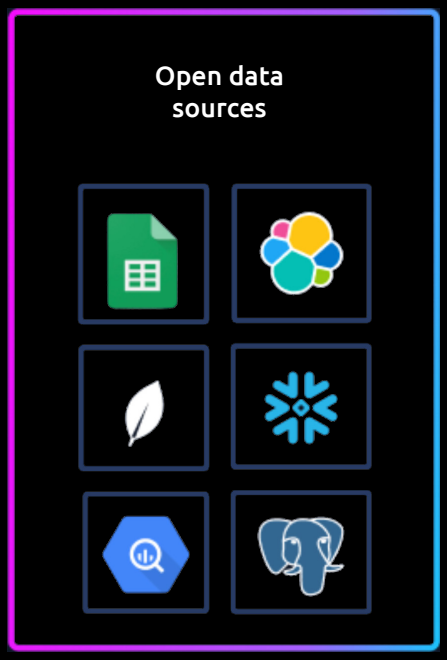
10PB daily read data
250k queries per day

300PB data lake

Fast, easy access to data outside the lake

The modern data lake has the capacity to store and maintain all enterprise data from all sources, but in reality, the vaunted single source of truth remains as elusive as ever. Large organizations will always have some (often critical) data residing in other systems outside the lake, whether that is a traditional data warehouse, a cloud data warehouse, the Salesforce cloud, or some other platform.

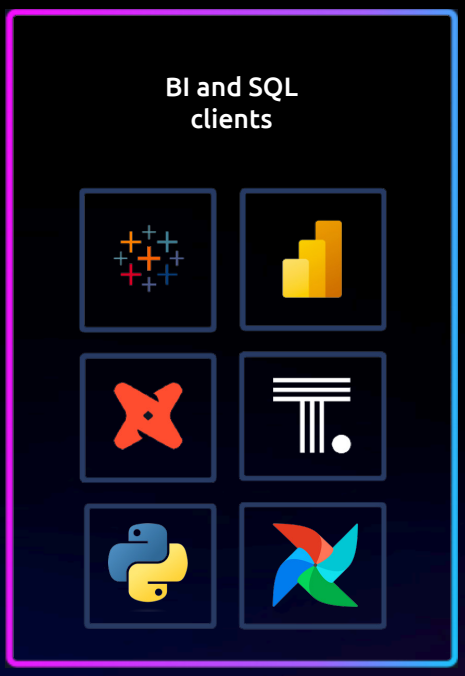
As such, implementing a data lake architecture that does not account for the presence of additional data sources will likely prove to be a shortsighted and costly mistake. This is where your choice of an analytics engine becomes very important, as it is possible to deploy a query engine that delivers all the capabilities of a standard engine while also providing federated access to distributed datasets. This allows you to simultaneously query data inside and outside of the data lake as if these volumes are in the same place.

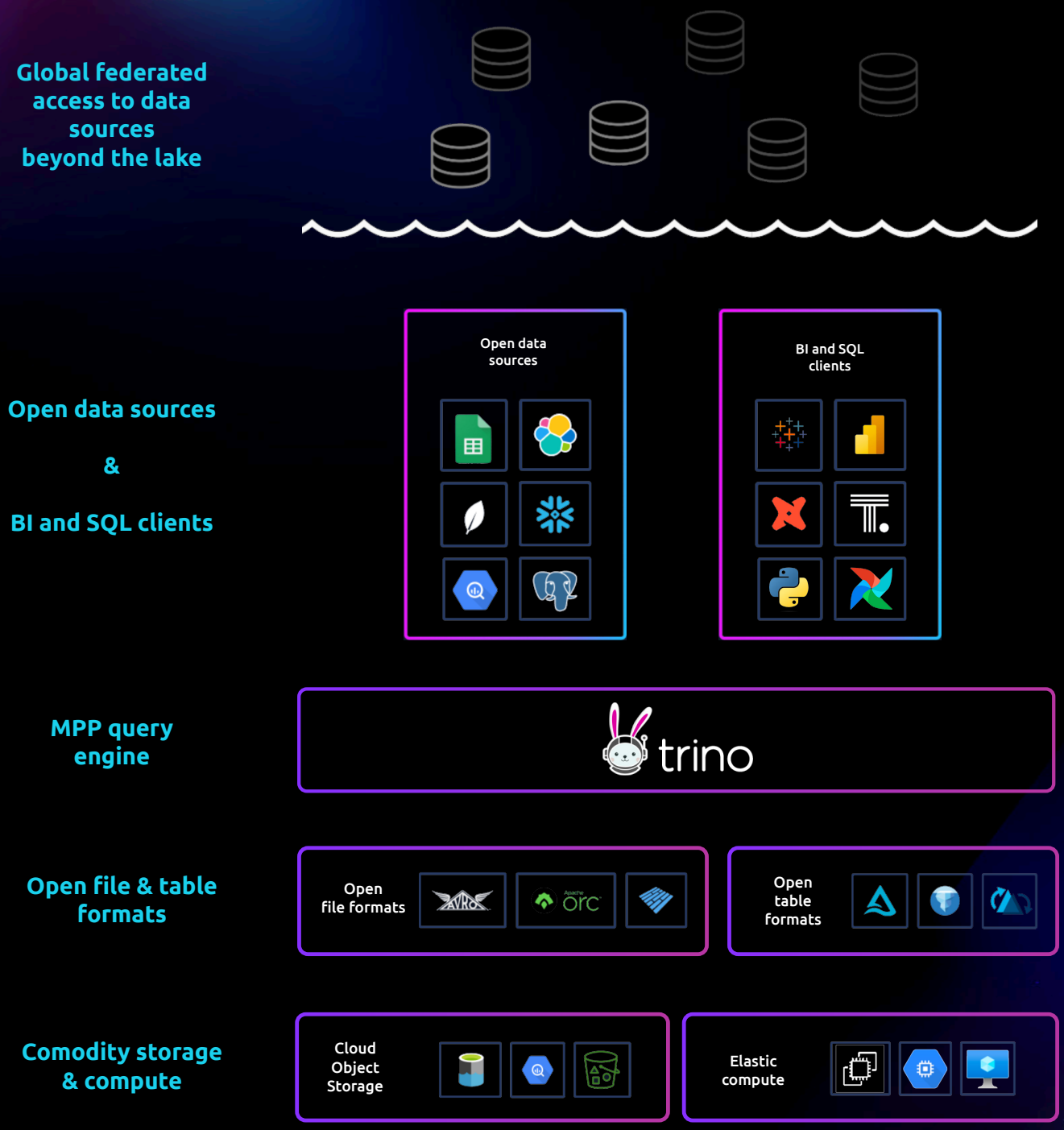


Seamless connectivity to BI & data science tools

Finally, a modern data lake architecture must integrate seamlessly with the tools your end users are accustomed to interacting with already. An ideal modern data lake architecture masks the underlying complexity for end users, allowing Business Intelligence (BI) analysts, data scientists, and others to work in familiar BI and data science and visualization tools with little or no knowledge of what is happening on the backend.

Trino's connectivity to BI and data science tools provides immense value by bridging the gap between data lakes and analytical applications. With seamless integration, users can directly query and analyze data stored in the data lake using familiar BI and data science tools, eliminating the need for complex data movement or transformation processes. This enhanced connectivity enhances collaboration, accelerates decision-making, and empowers data professionals to derive actionable insights from the data lake quickly and efficiently.

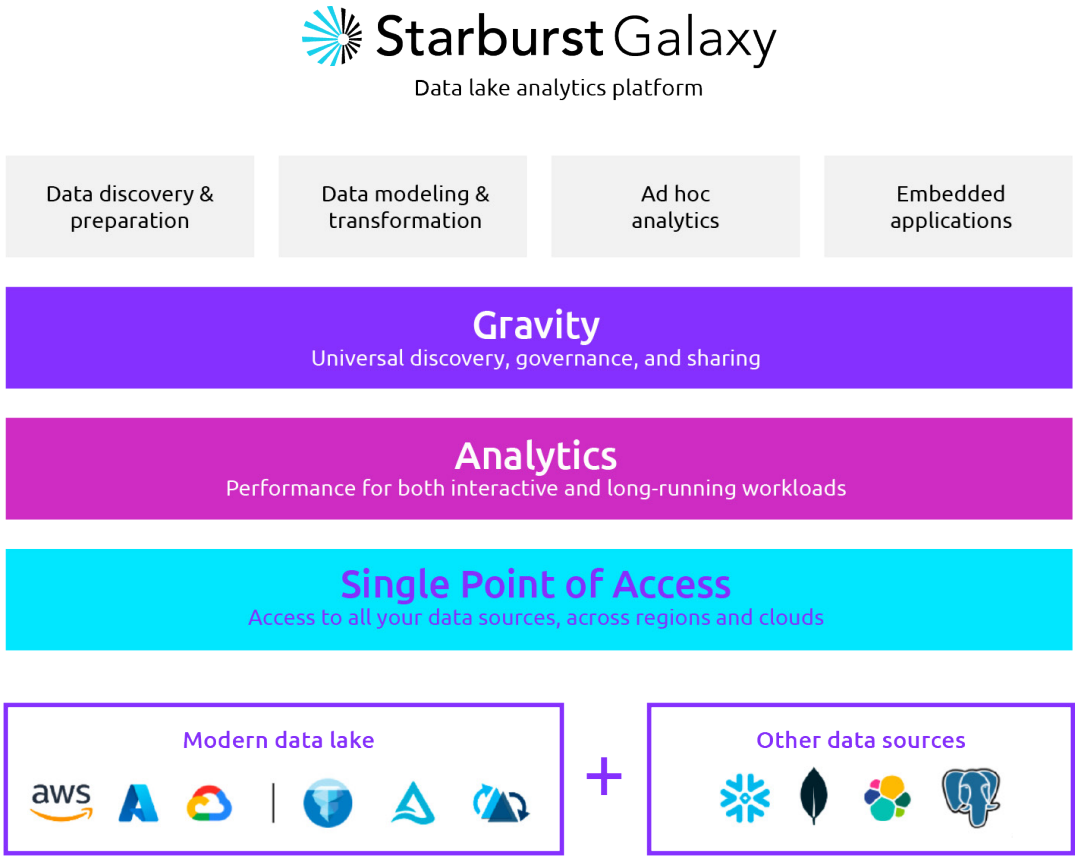




A modern data lake architecture designed around one of the leading cloud storage providers, open file formats, Apache Iceberg, and the Starburst analytics platform provides exactly what a company needs to extract as much value and insight as possible out of its data.

Query Engine for the Modern Data Lake

Starburst offers a full-featured data lake analytics platform for data-driven companies. Built on open source Trino, the Starburst platform makes it easy to discover, organize, and consume data without costly migrations. With Starburst, teams can access more complete data, lower the cost of infrastructure, use the tools best suited to their specific needs, and avoid vendor lock-in.



The Starburst-enhanced modern data lake provides:

- A single point of access and governance to all enterprise data
- Advanced warehouse-like capabilities directly on data in the lake
- Architecture that is vendor agnostic at every level
- A scalable and cost-effective analytics stack

As a core component of the modern data lake analytics architecture, Starburst functions as a single point of access to both the data lake and additional enterprise sources, and integrates seamlessly with leading tools, granting all authorized users within the organization fast access to the data they need to do their best work and uncover hidden value for their company.

The easiest way to build and manage your modern data lake

90%
Faster time to insight

53%
Lower TCO

100%
Future-proof architecture

Benefits of our approach

Simplicity

Take the fastest path from data to decision

Access

Future-proof your architecture with a single point of access and governance

Scalability

Operate efficiently and reliability at internet scale

Optionality

Never getb locked into a proprietary data ecosystem

Optum accelerates queries by 10X

Overview

Information technology service provider Optum is dedicated to shaping a healthcare system that gives patients a complete view of their health, providing them with personalized insights that lead to improved outcomes.

"Our data lake backbone was on a traditional Hadoop infrastructure. While that approach had its day, it's not flexible. We needed to scale out and separate our compute from our storage without moving the data."

--- Mike Prior, Principal IO Engineer, Optum

Challenge & Solution

Optum's data warehouse solution, Hive, could not support the growing demand for Optum's analytics systems. This led to a poor end-user experience and an inability to bring on new workloads. To solve this challenge, Optum replaced Hive with Starburst deployed on Hadoop. "Providing users with one endpoint is so much easier. They can use the same familiar tools, but everything is happening faster," shares Prior.

What Starburst delivered

10X
faster queries

30%
reduction in infrastructure costs

$8M
in projected savings

Unlocking new revenue streams with a 360-degree customer view

Overview

This telecommunications company is the leading pay-TV and cable TV company in the United States. With more than 15 million subscribers, this customer retains and ingests tremendous volumes of data across various platforms.

"When end users are going into on-prem or cloud environments, they will be presented with all the data sets they have access to, irrespective of where the data is located. This offers a huge value to our end users."

--- Anonymous, Director of Software and Engineering, Fortune 100 Telecom Giant

Challenge & Solution

The CMO wanted to run campaigns to upsell the existing subscriber base. With the billing data in Teradata and user data in a large Hadoop Cluster, being able to query the data to run this sales campaign would be impossible.

This telecommunications giant selected Starburst to migrate its data off Teradata to Amazon S3 and establish universal data access. With Starburst, end users are able to easily run queries against different data sources.

What Starburst delivered

>$200
million in new s ubscription revenue

>250 TB
of data ingested daily

Reduced time to insight from
18 months to 5 weeks

The New Center of Gravity

The modern data lake is the future of enterprise data storage and analytics and will become the center of gravity for any data-driven organization.

- Data lakes are scalable, cost effective, and open
- Data lakes now deliver warehouse-like capabilities
- Data lakes integrate with leading analytics engines and tools

	Legacy Data Lake	Modern Data Lake
Access	Limited to the data lake	Universal access to data in and around the lake
Table Formats	Limited to a single format (e.g. file formats in Hadoop)	Support for all the modern formats Iceberg, Delta Lake, Hudi
Scalability	Medium	High
Performance	Low	High
Cost	$ (can be expensice with proprietary vendors)	$
Use Cases	Raw data storage, ML	BI, SQL, ML, Real-Time Apps
Reliability	Low quality, data swamp	High-quality, reliable data with ACID transactions
Governance	Poor governance because security needs to be applied to files	Fine-grained security and governance for row/columnar level for tables

At Starburst, we believe your analytics architecture should be designed with the data lake as the new center of gravity. Yet our extensive experience working with companies at every scale - from agile startups to Fortune 100 giants - has demonstrated that there will always be additional datasets residing outside the lake, and that a truly modern analytics platform must provide access to all enterprise data, wherever it resides.

About Starburst

For data-drive companies, Starburst offers a full-featured analytics platform built on open-source Trino.

Our platform includes the capabilities needed to discover, organize, and consume data without the need for time-consuming and costly migrations. We believe the lake should be the center of gravity and be the starting point for querying disparate data.

With Starburst, teams can access more complete data, lower the cost of infrastructure, use the tools best suited to their specific needs, and avoid vendor lock-in.

Trusted by companies like Novant Health, Assurance, Optum, Pfizer, Sophia Genetics, EMIS Health, Gilead and Genius, Starburst helps companies make better decisions faster on all their data.

To learn more, visit **www.starburst.io** and follow Starburst on Twitter and Linkedin.

A single point of access to all your data

Learn more at www.starburst.io