Starburst

The Ultimate Blueprint for Cloud Data Migrations

Ensure success before, during, and after your migrations

Forward

Data will make or break any modern business, especially as cognitive analytics and generative AI gain traction. Well-used data enables insights, innovation, and competitive advantage. But as data volumes grow and business needs evolve, many organizations face the challenge of migrating to the cloud. Cloud migration is moving workloads from an on-premises environment into a cloud mix (single, multi-cloud, or hybrid cloud), where they can benefit from scalability, flexibility, reliability, and operational and cost-efficiency. However, cloud migration is not a simple, straightforward task. It involves detailed planning, strategy, execution, and optimization to mitigate significant risks to the business, all while adhering to strict compliance and governance requirements. It can also take multiple years to complete for a set of workloads, and the business demands of data continue to change and evolve.

At Starburst, we also recognize that not all data will be migrated to the cloud. Each business has different needs, and we disagree with forced ‘standards’ of moving everything to the cloud. We believe that every business should make sound assessments to determine which data should and shouldn’t be migrated and ensure that their data and analytics architecture supports their federated data landscape. Of course, there are also compliance reasons why data cannot always be moved to a cloud or region and needs to be held in a single location.

Given there are so many existing resources available on cloud migration, we won’t aim to recreate the wheel but instead focus on two core areas: best practices for data migration and optimizing your data workloads before, during, and after migration.

Contents

Contents Continued

01

Why Migrate?

Let’s get the fundamentals out of the way. The data landscape for most organizations is complex. It involves collecting high volumes of data spread across multiple applications and use cases supported by various infrastructure deployment models. It doesn’t take long before disparate databases support a multitude of applications, data warehouses are supporting analytics initiatives, data lakes are supporting data science use cases. In such scenarios, data silos emerge naturally (some by design but many unintentionally), preventing an effective utilization of the diversity of data across the company. This data infrastructure becomes extremely complicated and data pipelines are stitched together to drive downstream insight and analysis.

Ultimately, a decision is made to migrate a subset or all of the company's data to take advantage of all the benefits of a modern data lake architecture. **More than 70% of companies have now migrated at least some workloads into the public cloud, according to Gartner research.**[1]

Some main drivers are:

- **Business drivers:** These include mergers and acquisitions and new business models from cloud-built products and services. By migrating to the cloud, businesses can accelerate digital transformation and adapt to changing market conditions and customer expectations. They can also foster innovation and experimentation by enabling faster development and deployment of new products and services. However, 70 percent of CXOs lack confidence in their organizations' ability to pivot and adapt to disruptive events.[2]

- **Digital transformation**: Businesses can leverage the cloud to modernize their core functionalities, such as CRM, ERP, analytics, and more. The cloud enables businesses to access the latest technologies and tools, such as artificial intelligence (AI), machine learning (ML), and big data analytics. In a 2021 Deloitte and Fortune survey of CEOs about their leadership through the COVID-19 pandemic, 85 percent indicated their organizations had significantly accelerated digital transformation during the crisis.[3]

- **Growth and agility:** Businesses can scale their data and applications on-demand with the cloud without worrying about the physical limitations or costs of on-premises infrastructure. When combined with open file and table formats, the cloud also allows businesses to respond faster to changing market conditions and customer needs while maintaining optionality and ownership of their data. According to a 2018 report by McKinsey, which still holds true today, cloud adoption can increase IT productivity by 30-40 percent and business agility by 20-30 percent.[4]

1 Cloud Migration Costs and Avoiding Overspend | Gartner

2 Five characteristics of resilient organizations | Deloitte Insights

3 2023 CEO Priorities Survey | Deloitte US

4 Cloud adoption to accelerate IT modernization | McKinsey

- **Security and reliability:** The cloud provides a secure and reliable environment for data storage and processing, with built-in features such as encryption, backup, disaster recovery, and compliance. The cloud also reduces the risk of data breaches, hardware failures, and human errors. According to a report by IBM, the average cost of a data breach in 2022 was USD 4.35 million (up 12.7 percent from USD 3.86 million in 2020). The average time to identify and contain a breach was 304 days (up 8.6 percent from 280 days in 2020).[5]

- **Cost reduction:** The cloud offers a pay-as-you-go model that eliminates the upfront capital expenditure (Capex) and reduces the operational expenditure (Opex) of managing on-premises infrastructure. When implemented correctly, the cloud also optimizes data and applications' resource utilization and performance. On the flip side, for every dollar an organization spends on capital expenses to upgrade its IT infrastructure, it can also expect to pay about $2 to manage, maintain and secure that infrastructure.[6]

- **Data consolidation:** Businesses can migrate their data from multiple sources and locations to a single cloud platform, creating a central source of truth for their data. This can improve data quality, accessibility, and governance. However, organizations should tread cautiously if this is the only or primary driver. The single source of truth for many remains elusive because of changing technological landscapes and migrations that many organizations never realize this state of being. The truth is, centralizing and consolidating all data is a road to failure.

---

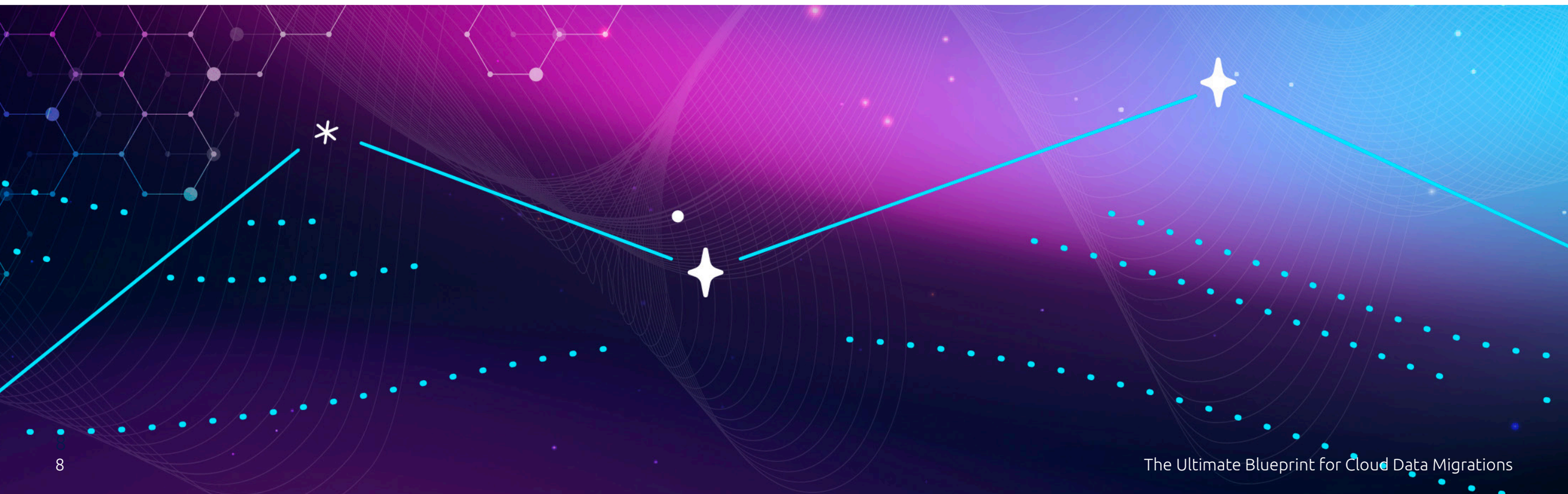
5 Cloud Migration Costs and Avoiding Overspend | Gartner

6 Five characteristics of resilient organizations | Deloitte Insights

The unspoken reality - not all data gets migrated, and that is fine

One of the most common challenges in data migration projects is deciding what data to migrate and what to leave behind. Data migration is not simply copying and pasting data from one system to another. If a vendor or an employee ever says it is or is just a simple lift and shift, there should be red flags everywhere. A full migration project involves planning, moving, testing, troubleshooting, productionalizing, and monitoring data workloads across numerous systems, user roles, and security policies to ensure quality, compatibility, and data security.

However, two groups of data may not be worth the migration headaches. It is crucial to have a clear data migration strategy that defines the scope, criteria, and methods for selecting and migrating data. This strategy should consider the trade-offs between the costs and benefits of migrating different data workloads.



The first group includes data that may still be needed to support ongoing business needs or applications managed on-premise. For example, a product team may continue to run a web application on-premise while its smartphone app sibling is deployed to the cloud. Tables supporting both applications, like “customers” or “order history,” may need to persist on-premise for the web application until the web application is also ready to be migrated to the cloud.

The second group includes data that may be subjected to regulatory requirements or could be sensitive, outdated, inaccurate, incomplete, or irrelevant to the business’s future operations but must be maintained. Migrating this data can cause errors, performance issues, and compliance risks and have significant cost implications.

Based on the situation, it may be advantageous to persist data on-premises and utilize vendors or solutions that bridge the cloud and on-premise data sets but at a lower cost or risk than a complete migration. This option can increase the capacity of a migration team, speed the migration, prevent data or workload outages and help avoid data copying, performance, or compliance issues. This decision should not be taken lightly. Leaving data temporarily or permanently on-prem could similarly impact workloads and teams dependent on that data. Therefore, monitoring and reviewing the data migration outcomes and planning for future data needs is essential.

02

The migration paths

There is no one-size-fits-all approach to data migration. Depending on the current state, goals, and requirements, businesses can choose from different migration paths, such as:

- On-premise data lake or data warehouse to a cloud data lake: This involves retiring the legacy data lake or warehouse infrastructure and moving data from on-premises to a cloud-based storage service, such as Amazon S3 or Azure Blob Storage. This option offers the benefits of low costs, high scalability, better performance, and easier access to cloud-native analytics tools. However, it may require significant changes in the data architecture, governance, security policies, and ETL pipeline code. This may put data quality at risk, creating data swamps, based on format conversions or schema evolution.

- On-premise to a modern data lake architecture: This approach puts the data lake as the center of gravity but can be either in a single or multi-cloud or a hybrid model (ex., Dell ECS and Amazon S3). It doesn't discriminate whose servers contain the data but focuses on what you can do with the data. A modern data lake has four primary requirements:

    1. It has to be built on commodity storage and compute, which means you can scale up and down cost-effectively.

    2. It has to be built on open file and table formats, which means your data is portable and stays yours.

    3. A high-performance and scalable query engine to query all the data.

    4. It must access all the data that won't live in the lake.

We explore the pros and cons of this particular approach below

On-premise data warehouse to a cloud data warehouse (CDW): Moving data from on-premises to a cloud-based database service, such as Snowflake, Amazon Redshift, Azure Synapse Analytics, or Google BigQuery. Compared to on-premise options, this offers the benefits of lower costs (initially), higher scalability, better performance, and easier access to cloud-native analytics tools for your structured and semi-structured data. However, it also requires similar changes as those listed above of a cloud data lake. In the case of Snowflake, their own recommended best practices require you to move your data into object storage like Amazon S3 first as a staging environment and then move into their cloud data warehouse. Additionally, a CDW may limit the flexibility and diversity of the data sources and formats that can be ingested and processed in the cloud - this can have severe implications on the ability to use data for AI/ML-based applications. If pursuing this path, it's no longer an all-in requirement to benefit from the promises of a CDW. Instead, seriously consider what critical data needs to be in the CDW and what data can remain in the on-prem or a cloud (modern) data lake and still run SQL analytics on the data there. We explore this in a bit more detail below.

On-premise data lake or data warehouse to a hybrid architecture: Moving some of the data lake or warehouse infrastructure and data from on-premises to a cloud-based storage or database service while keeping some on-premises for specific reasons such as on-prem applications still in production, compliance, or latency. This option offers the benefits of leveraging both worlds: maintaining control over sensitive or mission-critical data while taking advantage of cloud scalability.

Why modern data lake?

Modern data lakes, or lakehouses, combine data warehousing-like high-concurrency SQL analytics with the flexibility and cost-effectiveness of a data lake. It becomes the center of gravity for data but not a dependency on being the only storage source system. With a modern data lake, organizations can leverage open table and file formats from technologies like Apache Iceberg, Delta Lake, or Apache Hudi, not only as a cost-effective object store and landing zone for interactive analytics but also for prescriptive, predictive, and cognitive analytics, and the descriptive and diagnostic analytics you would expect from a data warehouse.

A modern data lake offers several benefits:

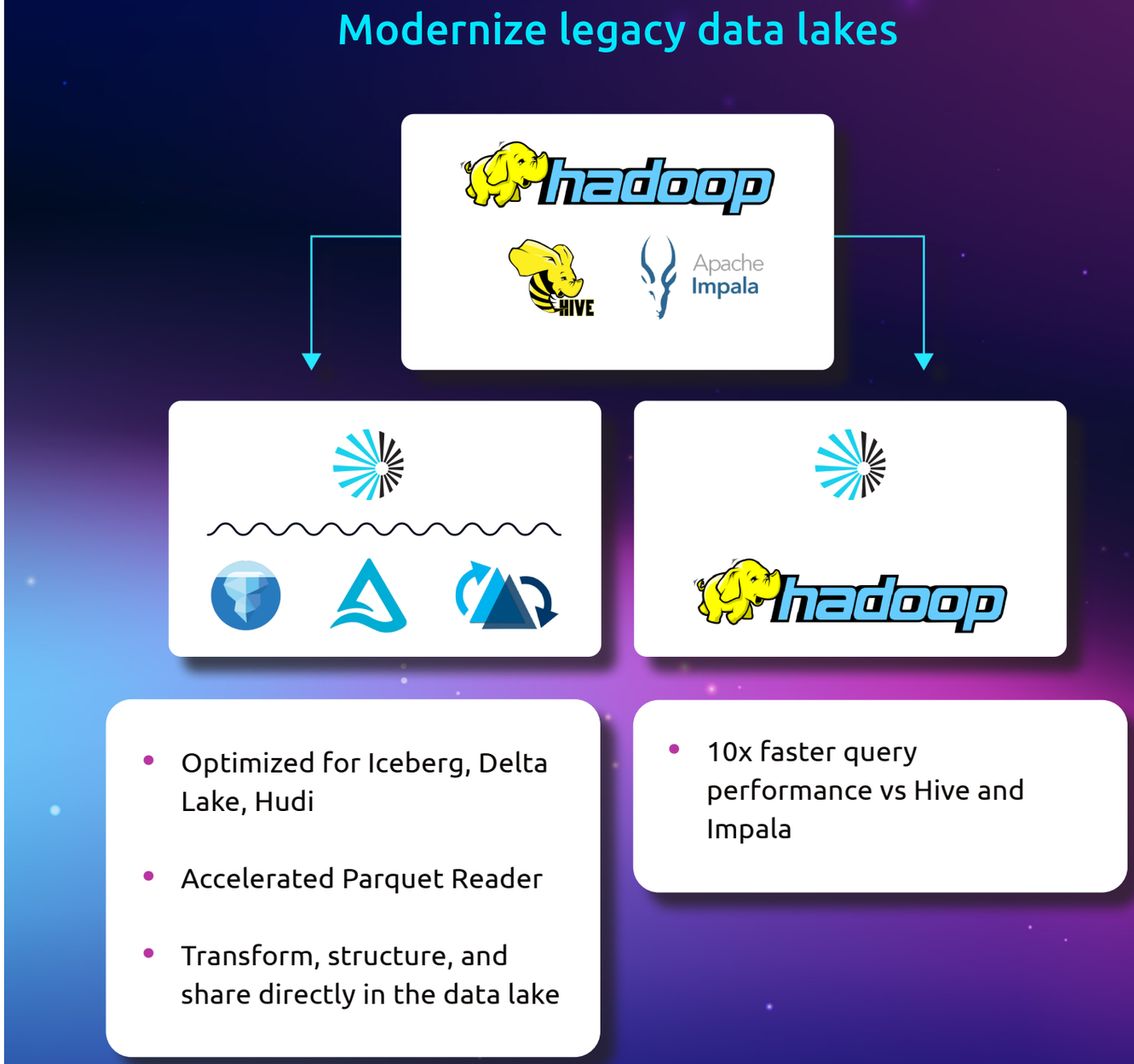
- **Flexibility:** A modern data lake can store data from any source without requiring schema definition or transformation upfront - schema on read. This allows for greater flexibility and diversity of the data sources and formats to be ingested and processed.
- **Scalability:** Compute and storage can scale elastically to accommodate data without compromising performance or availability, allowing for handling large volumes of diverse and complex data efficiently and cost-effectively.
- **Optionality:** It can provide more options for data consumers and producers. They can use different tools and frameworks to access, process, and analyze the data in and around the data lake, leveraging multiple distributed compute engines, and open file and table formats. They can also leverage cloud providers and services to optimize cost, performance, availability, and functionality.
- **Future-proofing:** A modern data lake can help future-proof the data strategy and architecture. It can accommodate new data sources, formats, and use cases without significant changes or migrations. It can also enable innovation and experimentation by allowing data consumers and producers to explore their data before moving it into the modern data lake.
- **Data warehouse capabilities on the lake:** Gain the performance, schema evolution, time travel, CRUD and ACID operations traditionally only available on an optimized data warehouse directly on your data in the data lake.

Let’s not forget about Hadoop

A lot of credit is owed to Hadoop, which has either influenced or been the foundation for many of the technological advancements businesses and society benefit from today. However, one of the common migration patterns tends to be migrating parts or all of the Hadoop environment to the cloud.

Hadoop is a popular open-source framework for storing and processing large-scale data on-premises. Many organizations that have invested in Hadoop-based data lakes can face challenges in maintaining and scaling their infrastructure and keeping up with the evolving analytics demands of the business.

Hadoop has limitations and challenges, such as high maintenance costs, complex administration, scalability issues, and a lack of cloud-native features. Many organizations are looking to migrate their Hadoop workloads to the cloud to overcome these challenges and benefit from the advantages of the cloud. Migrating from Hadoop to a cloud data lake can offer many benefits, such as improved performance, security, reliability, and cost-efficiency.



But, migrating from Hadoop to the cloud is not a trivial task. It involves moving the data and the applications, pipelines, and tools that depend on it. Moreover, it consists in choosing the right cloud destination and architecture that can support the existing and future needs of the business.

A modern data lake strategy can help simplify and streamline the migration from Hadoop to the cloud. It can enable a seamless data transition from HDFS to a cloud-based storage service without requiring any format or schema changes. It can also allow a seamless transition of the applications, pipelines, and tools from MapReduce, Hive, Spark, and other frameworks to Trino, which can query any data source without moving or copying the data. Trino can also integrate with other open-source tools and frameworks that can enhance the functionality and performance of the modern data lake, such as Iceberg, Delta Lake, and Hudi.

By adopting a modern data lake strategy with Trino, organizations can migrate from Hadoop to the cloud faster, easier, and cheaper. They can also enjoy the benefits of a modern data lake's flexibility, scalability, optionality, and future-proofing.

Here are some considerations for a successful data migration from Hadoop:

- **Evaluate current environment:** Begin with assessing your existing Hadoop setup, introducing Trino as the compute engine. Analyze data specifics, workflows, dependencies, and desired outcomes to define project scope and objectives.

- **Select cloud platform:** Compare cloud options based on features, compatibility, and costs. Match these with migration goals to identify the optimal cloud solution, potentially spanning multiple platforms.

- **Design cloud architecture:** Map out storage, compute, and analytics layers. Choose scalable storage (e.g., Azure Data Lake, Amazon S3), compute service (Trino and Hive), analytics tools, and account for security, governance, and observability.

- **Plan data migration:** Prioritize batch migration over simultaneous transfer for efficiency. Minimize disruption, monitor the process, and ensure business continuity by maintaining data federations between legacy and new systems.

- **Agile migration execution:** Prepare data by cleansing, transforming, and validating it. Choose migration tools like Azure Copy, AWS Transfer, or BigQuery Data Transfer, and ensure incremental data movement for accuracy. Consider managed options or manual scripts.

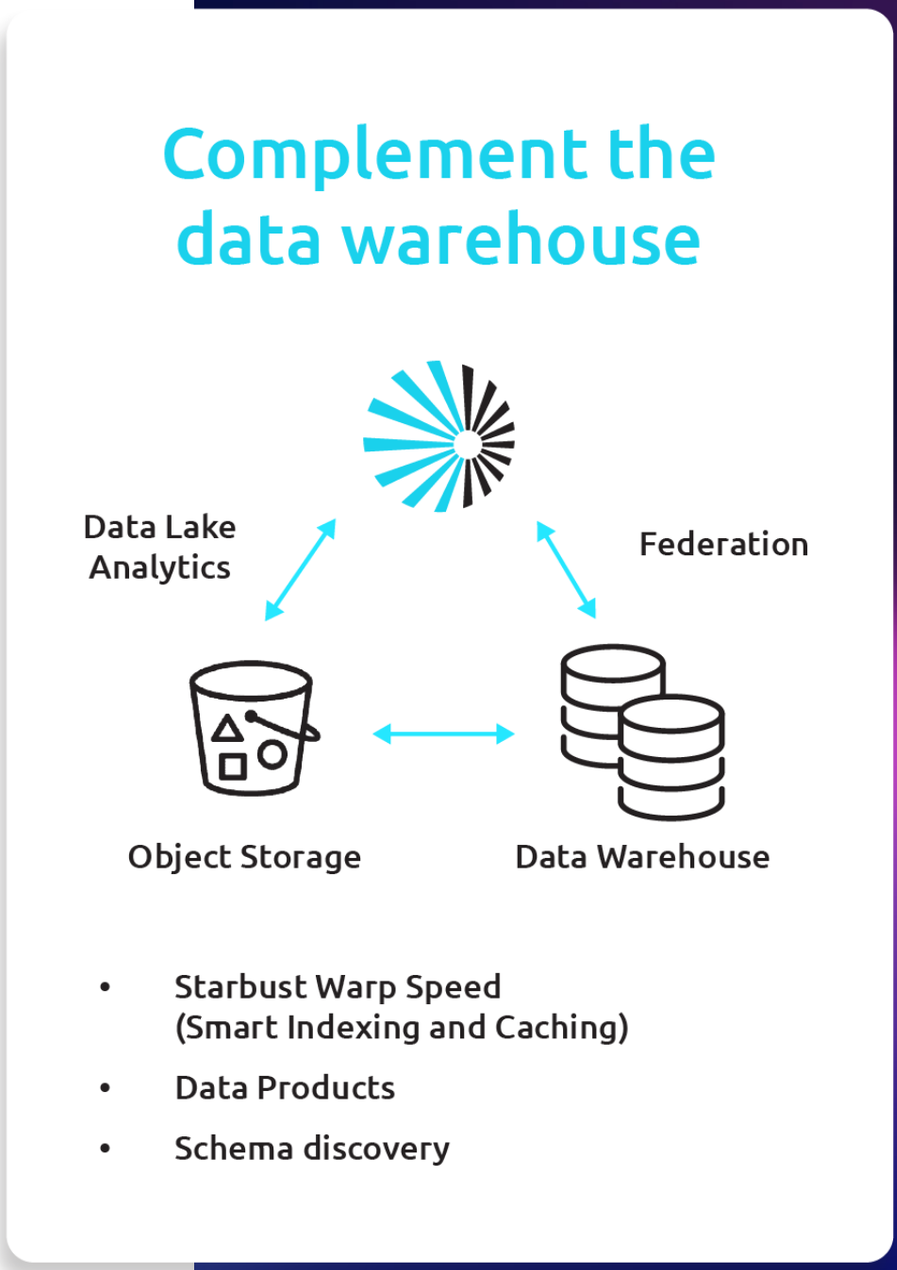
- **Optimize, test, and validate:** Enhance data storage efficiency by adopting best practices (partitioning, compression, indexing). Regularly test and validate data quality, integrity, and accessibility through queries and analytics.

Effective Hadoop migration entails careful assessment, strategic cloud selection, meticulous architecture design, phased data movement, agile execution, and ongoing optimization.

Already decided on a CDW? Consider how to get ahead of the inevitable ballooning costs

For those too dependent on the data warehouse, we help by unlocking the value in the object storage you likely already have. For example, Snowflake's recommended ingestion pattern on AWS is first to move all data into an S3 staging environment. In other words, move your data to a cloud data lake. Then, they recommend loading the full data set into Snowflake, where they charge you big bucks to do all the cleanup from raw data to consumable tables. Here's where Starburst helps.

With Starburst, we can run high-performance data lake analytics directly on S3 and help you choose which data should move into the data warehouse. Because Starburst provides warehouse-like capabilities on the data lake, you don't need to put all your data into an expensive and rigid data warehouse. Maybe you still want to put some first-class BI and reporting data there, and that's fine, but for everything else, it's usually more scalable and more efficient to leave it in the lake.



03

Risks and challenges of data migrations

Even with meticulous planning, data migrations entail inherent challenges and risks. Organizations grapple with evolving priorities, technological shifts, security vulnerabilities, unexpected expenses, and more.

Common hurdles include:

Integration strategy: A robust blueprint for harmonizing data sources, formats, schemas, pipelines, and tools within the cloud ecosystem is vital. Without integration, migrations can foster data silos, inconsistencies, errors, and inefficiencies.

Data security: Migrating data to the cloud demands stringent security measures. Protecting against unauthorized access, alteration, or loss is paramount. Adhering to data privacy regulations mandates careful cloud provider selection, encryption, authentication, and auditing.

Risk factors: Inherent uncertainty surrounds data migrations. Potential data loss, corruption, or leakage loom. Poorly planned migrations risk business disruption and unmet expectations if cloud environments misalign with objectives.

Migration timeline: Migrations vary in duration based on data complexity and size. This downtime might impact user access, affecting business operations and analytics performance.

Cost challenges: Migration expenses, notably hidden costs, pose significant challenges. Adequate planning for data transfer fees, operational outlays, and skilled labor diversion from essential tasks is imperative. Additional factors include:

- Bandwidth and resource provisioning
- Post-migration processes
- Training for the new system
- Cloud service provider fees
- Data storage expenses
- Security and compliance investments

Strategizing to mitigate these issues is essential for successful data migrations.

04

Best practices for a data migration

The following process flow provides a logical framework to help define and manage your data migration. The high-level categories offer a starting point for determining detailed activities and specifications to be managed throughout the migration. Visualizing the interrelationships between the steps helps to ensure dependencies are accurately defined. Within each step, objectives should be set with the corresponding assignment of responsibilities and workloads across your data migration team.

Regardless of which path you choose for data migration, there are some best practices to follow to ensure a successful outcome.

General best practices for people and processes

Assemble your data migration team of internal and external resources, including stakeholders and vendors. For users, you'll need to inventory their roles to ensure equivalent permission levels are set up.

In addition to the people, ensure you have a bulletproof process, including migration specification development, setting objectives, assembling a migration team, data mapping, backup, testing and validation, and roll-out plan. Configuring Starburst early on the cloud(s) of choice will help minimize operational disruptions.

Have an aligned precise cut-off date to shut down the legacy platforms. The project plan and timeline must include the steps, sequencing, and communication plan for shutting down the old systems once the migration is complete. This will help prevent the emergence of new siloes resulting from the remanence of legacy systems.

Data Best Practices for Migration:

Understand your data:
Prior to migration, comprehensively assess source and target systems, data formats, types, volumes, dependencies, and relationships. Evaluate data quality and integrity.

Define a strategic plan:
Formulate a clear strategy aligning with business goals, outlining scope, objectives, timeline, budget, and roles. Migration is an ongoing cycle demanding careful planning.

Implement disaster recovery:
Safeguard against data loss or corruption by creating backups and setting up source data replication. Preparedness is crucial to recover from migration errors or failures.

Migrate in batches:
Opt for a phased approach to minimize disruption. Batch migration reduces downtime, accelerates the process, and aids in efficient monitoring and troubleshooting.

Use appropriate tools:
Leverage data migration tools for streamlined and automated processes. Cloud-specific transfer services, migration software, and professional services simplify migration without excessive costs.

Thoroughly test migration:
Validate migration process in pre-production phase. Test data accuracy, system functionality, performance, security, and compatibility.

Rigorous migrated data testing: Before deploying to production, rigorously test migrated data for accuracy, completeness, consistency, and integrity. Ensure smooth business outcomes.

Ensure uninterrupted data flow: Maintain continuous data access for downstream applications and teams during migration. Avoid disrupting business operations and duplicating architecture

Efficient data migration hinges on understanding, strategy, preparation, testing, and seamless continuity.

Technology best practices

As with your people, process, and data, you must be equally deliberate about your technology stack decisions when building your modern data lake architecture—you can argue this element is possibly the most critical as the implications of hasty or wrong choices can have lasting negative consequences for years. Here are seven essential questions to consider.

1. Which public cloud platform?

Choosing the right cloud platform provider can be daunting, but you can't go wrong with the big three, AWS, Azure, and Google Cloud Platform. Each with its own massively scalable object storage solution, data lake orchestration solution, and managed Spark, Trino, and Hadoop services. Depending on your business requirements, you may combine the three to pursue a multi-cloud architecture.

Let’s look at each of the big three:

AWS Lake Formation & S3.

AWS Lake Formation provides a wizard-like interface over various pieces of the AWS ecosystem, allowing organizations to build a data lake quickly. The primary backend storage of an AWS data lake is its S3 storage. S3 storage is highly scalable and available and can be redundant across several availability zones. Depending on availability, S3 has three tiers (Standard, IA, and Glacier), with lower storage costs and higher read/write costs. S3 also has automatic object versioning, where each version is addressable so it can be retrieved at any time. AWS S3 storage offers rich functionality, it’s been around the longest, and many applications have been developed to run on it.

Azure Data Lake & Blob Storage.

Azure Data Lake is centered around its storage capacity, with Azure blob storage equivalent to Amazon S3 storage. It offers three classes of storage (Hot, Cool, and Archive) that differ mainly in price, with lower storage costs but additional read and write costs for infrequently or rarely accessed data. Azure Data Lakes rely heavily on the Hadoop architecture. Azure Blob Storage can also be integrated with Azure Search, allowing one to search the contents of stored documents, including PDF, Word, PowerPoint, and Excel. Although Azure provides some level of versioning enabling users to snapshot blobs, unlike AWS, it is not automatic.

Google Cloud Storage.

Google Cloud Storage is the backend storage mechanism driving data lakes built on the Google Cloud Platform. As with other cloud vendors, Google Cloud Storage is divided into tiers (Standard, Durable Reduced Availability, and Nearline) by availability and access time (with less accessible storage being much cheaper). Like AWS, Google supports automatic object versioning.

2. Which open table format is the right one for your business?

Table formats are an integral part of a modern data lake architecture. They enable the scalability benefits of low-cost data lakes and the underlying object store while getting the data quality and governance once only possible with data warehouses on large data sets. These formats also provide table schema compatible with existing data processing tools and support advanced features such as versioning, indexing, and ACID transactions. Several key factors must be considered when deciding on open table formats between Apache Iceberg, Apache Hudu, and Delta Lake.

First, look at the overall feature comparison and how each format addresses some of the most pressing issues with data lakes, such as atomic transactions, consistent updates, and data and metadata scalability. Second, consider platform compatibility and how well each format integrates with other tools, such as massive parallel processing (MPP) query engines. This also ties in with the ecosystem supporting the open-table format and the community momentum. Third, concurrency guarantees matter and refers to the ability of a format to handle multiple concurrent operations without conflicts or corruption. Most important, it's critical to evaluate the features and capabilities of each format to your specific use cases and data architecture.

3. Which open-source query engine and data virtualization technology?

Similar to selecting the open table format, the choice of an appropriate query engine and data virtualization technology is crucial.
A plethora of open-source and commercial options exist. Notably:

Trino (formerly PrestoSQL):** Originating from Facebook, Trino is a distributed query engine built on ANSI SQL. It seamlessly integrates with various BI tools and efficiently handles querying vast data volumes. Trino, an extension of Presto, has broadened its scope to accommodate diverse analytics and customer needs. Its strengths encompass user-friendliness, robust performance, high interoperability, and strong community support. Trino enables combining data from multiple sources, supporting various data stores, formats, and connectors (Hive, Iceberg, Delta Lake, Elasticsearch, Postgres, MySQL, Kafka, etc.).

Apache Drill: Drill, an open-source distributed query engine, facilitates interactive analysis of extensive datasets. It mirrors Google's Dremel and utilizes Apache Arrow for in-memory computations and Calcite for query optimization. While sharing certain features with Trino, Drill faces limited adoption due to performance and concurrency constraints, despite its compatibility with various data stores and evolving structures.

Spark: Apache Spark, a unified analytics engine, handles large-scale data processing through diverse APIs (Java, Scala, Python, R) and supports general execution graphs. Spark SQL facilitates SQL querying, executed by a distributed in-memory computation engine. While versatile, Spark's broader application scope encompasses data transformation, machine learning, batch queries, streaming, and more. However, its adoption for interactive queries lags behind Trino or Drill.

The selection process involves understanding these options and aligning them with your specific requirements and preferences.

4. Do you want to become your own software vendor?

There is a certain allure of building something from the ground up, the challenge, the pride, the glory. However, building and managing a modern data lake or any technology stack comes with increased responsibility and unknown and uncontrollable variables, which can be time and resource intensive. Teams need to plan and resources for managing the development and testing, along with increased project timelines and costs. They must also know implementation, documentation, training and enablement, ongoing maintenance, and updates, especially around security vulnerabilities. But more importantly, many out-of-the-box capabilities that simplify things become your responsibility:

Management: everything is manual from a technology perspective unless you have the resources similar to the largest internet companies in the world to build it yourself. There is no autoscaling.

Security: no built-in security integrations; this requires hiring and retaining top infosec talent to build and manage the capabilities.

Access control: Requires third parties for role-based access control (RBAC).

Support and maintenance: there is no emergency hotline or a professional services team to call. Ultimately you become reliant on community responsiveness.

Unless there are truly unique capabilities and competitive advantages to be gained by building something completely customized with open-source technologies, put some serious thought into this decision and the time-to-market factor. By taking on the approach to building in-house, plan to be your software vendor for off-the-shelf solutions that, in many cases, are loaded with significant enhancements compared to the open-source alternatives.

5. Which use cases are served best by the modern data lake analytics stack?

By answering this question, you'll also help guide your data prioritization strategy of which data to move when. The modern data lake stack can be used for a wide range of analytics use cases:

Business use cases:

- **Customer experience** - next best offer/ recommendation engines, customer 360, A/B testing, supply chain analytics, demand planning
- **Risk mitigation** - fraud detection, anti-money laundering, underwriting, and log analytics
- **Product and applications** - data application, clickstream, AdTech, IoT, product 36

Technical use cases:

- **Federated querying**
- **Ad hoc queriesw**

The modern data lake stack dramatically improves the speed of ad hoc queries, dashboards, and reports. It enables you to operationalize all your data and run existing BI tools on lower-cost data lakes without compromising performance or data quality while avoiding costly delays when adding new data sources and reports.

6. How to serve various use cases on the modern data lake and avoid data silos?

To serve as many use cases as possible and shift your workloads to the modern data lake, you need to avoid data silos. You'll also need to ensure your stack is analytics-ready with workload observability and acceleration capabilities to easily integrate with niche analytics technologies such as text analytics for folder and log analysis.

Consider how each key component of the modern data lake is an enabler to activate use cases and yield optimal output.

Data access
should seamlessly connect to all your data sources in and around your lake. This also includes supporting connectivity to sources across regions, clouds, and between cloud and on-prem environments.

Security and governance
should provide out-of-the-box capabilities to manage access, privacy, encryption, monitoring, and logging.

The query engine
it should be elastic as your business needs, efficiently scaling up to internet scale as needed. It should also allow you to get query response times like an optimized data warehouse while running long-running queries and complex transformation jobs without fail due to out-of-memory limitations.

Modeling and semantic
includes all the features to help you build, organize, and share data models with standard SQL.

The ecosystem
3rd party integrations for preference for a specific vendor to bring your use cases to life.

7. How to achieve optimal query performance and price balance?

The agility and flexibility benefits of the modern data lake are clear. But delivering performance and cost are the critical driving forces behind the massive adoption of data lakes. As analytics use cases grow across every business unit, data teams will continue to struggle while balancing performance and costs.

Manual query prioritization and performance optimization are time-consuming, not scalable, and often result in heavy DataOps. To expand the open data lake concept across the entire organization, data teams should seek an intelligent and dynamic solution that will autonomously accelerate queries using advanced techniques such as micro-partitioning or dynamic indexing.

The migration checklist

This checklist provides eight critical steps to help data lake architects successfully migrate to the finalized cloud data lake.

It discusses how Starburst can improve analytics operations throughout the data migration process.

Build the business case for the migration

- ☐ Prioritize and assess systems and data impact on the business
- ☐ Establish the processes for measuring performance and ROI gains
- ☐ Develop KPIs for the migration
- ☐ Establish baselines pre-migration to determine post-migration performance improvements
- ☐ Set the objectives and anticipated gains

Explore, inventory, and back up the data

- ☐ Assess and prioritize systems and data to minimize business impact
- ☐ Align migration timing and critical business processes
- ☐ Map the architecture of your data and data stores, and establish your requirements
- ☐ Plan to migrate with lower-value data first to reduce risk to business operations
- ☐ Establish a governance plan to ensure compliance with PII, PHI, and PCI regulations
- ☐ Back up the data and set a restoration plan

Initiate a Starburst trial

- ☐ Determine which flavor of Starburst meets your requirements, Starburst Galaxy, a fully managed SaaS offering, or Starburst Enterprise, self-hosted and self-managed.
- ☐ Sustains data analytics operations during a migration
- ☐ Delivers access to data where it lives with no copying required
- ☐ Facilitates transfer and test protocols to ensure data integrity is not impacted by the migration
- ☐ Reduces migration costs and the total cost of ownership (TCO) of data analytics
- ☐ Seamlessly integrated with your cloud

Capture stakeholder buy-in

- ☐ Identify and engage all stakeholders across the enterprise
- ☐ Coordinate migration planning based on the objectives and impact on each department/business unit
- ☐ Prioritize planning around high-value, data-rich systems, and departments
- ☐ Assess and minimize the impact on users

Training and documentation

- ☐ Determine the process changes resulting from the move to the selected cloud platform(s) - AWS, Azure, or GCP
- ☐ Update and communicate documentation changes
- ☐ Set up training sessions to ensure all impacted users experience minimal productivity loss

Mitigate service disruptions

- ☐ Schedule targeted data set migrations during times of low utilization
- ☐ Maintain data on source systems
- ☐ Anticipate disruptions and proactively reach out to impacted users

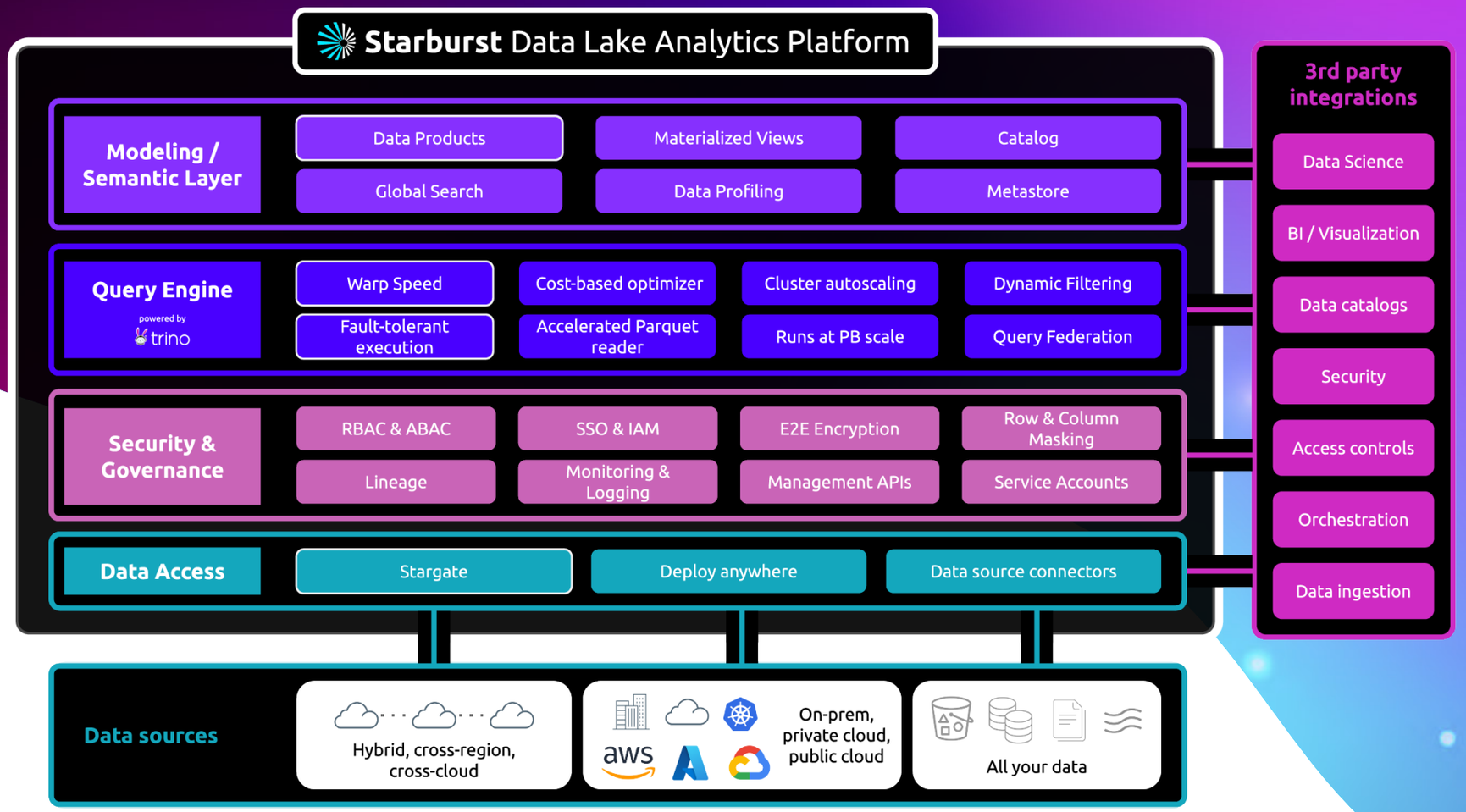
Maximize aata value during and post migration

- ☐ Establish a thorough quality assurance process for testing migrated data
- ☐ Ensure data integrity and system performance before roll-out
- ☐ If using a Data Mesh strategy, set up a Data Mesh layer to access data where it resides currently and in the future state
- ☐ Optimize the cloud data lake resources for ROI-driven performance gains

Solicit stakeholder feedback

- ☐ With each functional dataset migration, ensure stakeholders report on their experiences and adjust as needed
- ☐ Determine the next steps for performance and business intelligence gains

05

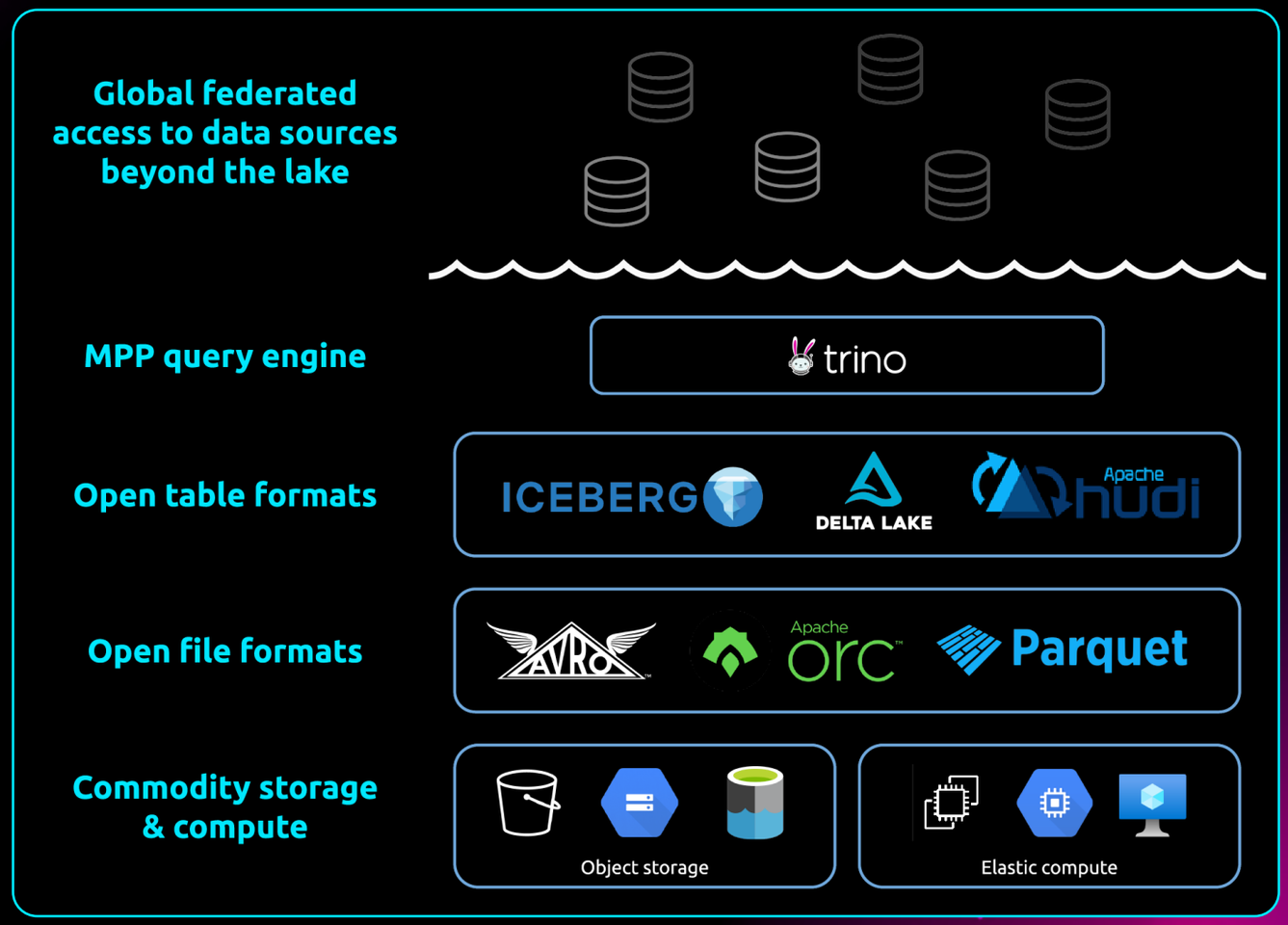


Starburst for data migration

Enterprises wanting to modernize to cloud data storage face two complex tasks: staging data migrations and, perhaps more complicated, moving and rebuilding the analytical business logic—the data pipeline—that connects existing BI and data science tools. Starburst leverages Trino, an open-source, distributed SQL query engine, to make better decisions with lightning-fast access to all data, no matter where it lives. Starburst helps solve these problems with a data consumption layer that enables seamless operation of analytics workloads during migration and eliminates the need to reconstruct data pipelines on cloud data lake storage. With Starburst, a semantic layer abstracts the datasets queried from the underlying datasets and storage. Thus, even as the location of the data changes during migrations, queries continue against the same virtual dataset without disruption.

The modern data lake architecture

The average data architecture is fraught with data silos everywhere. Organizations typically have their data center of gravity in a data lake but depend on many other data sources outside the lake, which may evolve as their needs change. For instance, modern applications often require a combination of data technologies, including historical data stored on a lake, streaming event data, log data, application data, etc. And there will always be new data types, and data sources created by different business needs to develop new technologies and applications. At Starburst, we believe a modern data lake (or data lakehouse) architecture solves many of the problems legacy data architectures are plagued with.



Four core components make up a modern data lake:

1

Commodity cloud storage

It has to be built on commodity storage and compute, which means you can scale up and down cost-effectively.

2

Open file and table formats

It has to be built on open file and table formats, which means your data is portable and stays yours.

3

Query engine

A high-performance and scalable query engine to query all that data.

4

Federation

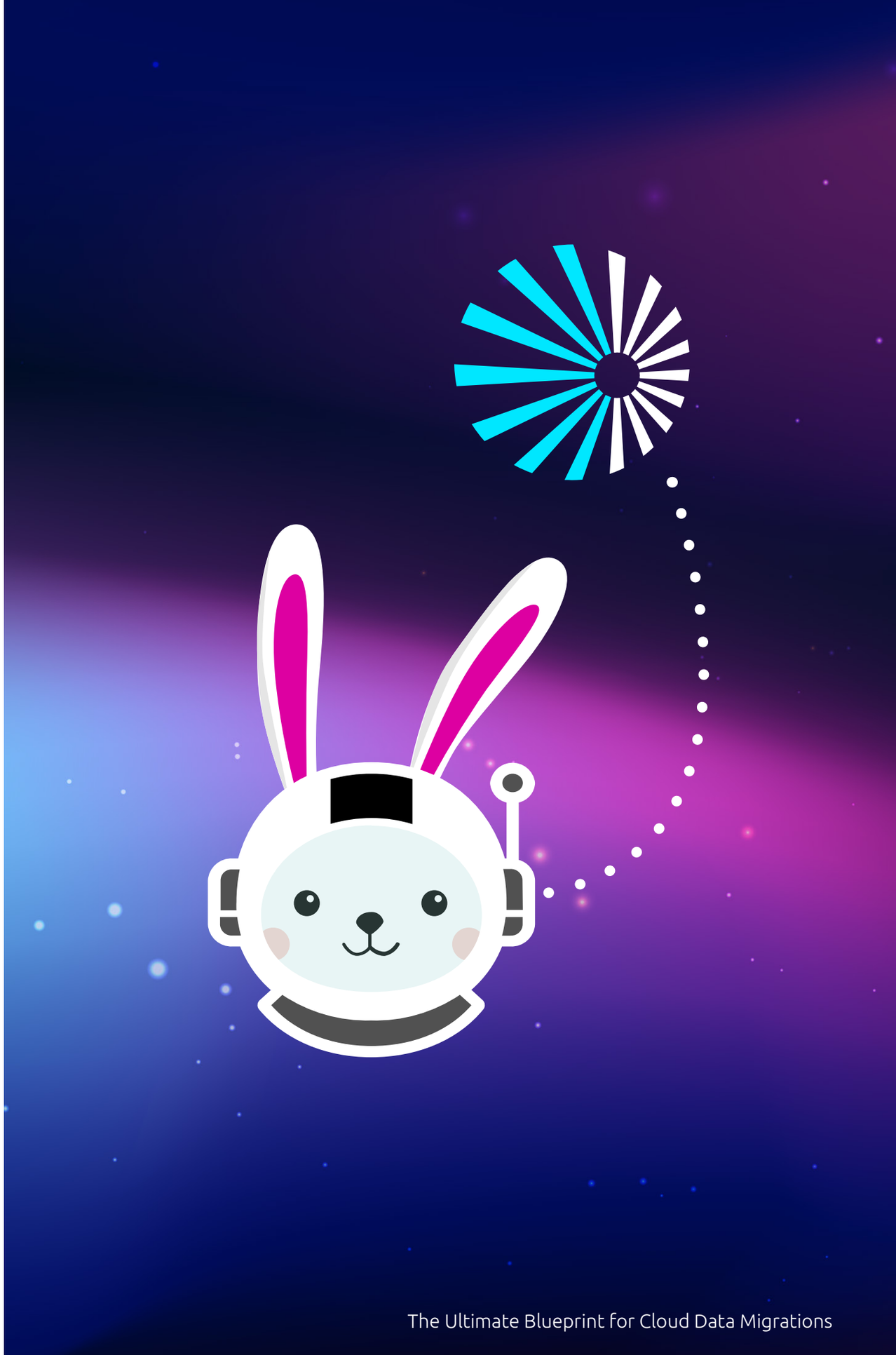
The ability to access all the data that won't live in the lake.

Ultimately, this unlocks organizations' ability to unlock capabilities to execute SQL analytics on the data lake, query federation, enhanced ad-hoc analytics, and ability to embed Starburst with modern apps to improve performance and customer experience.

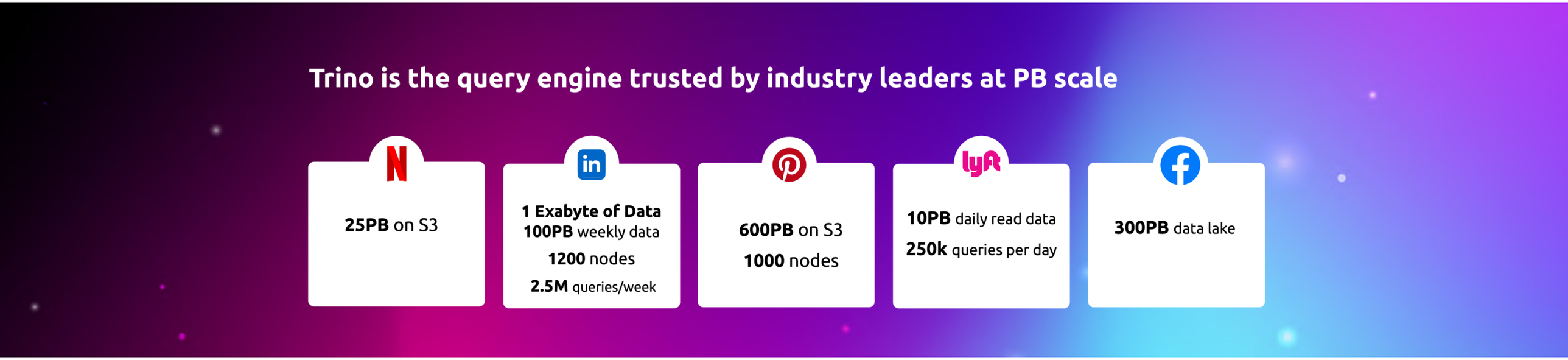
The Starburst and Trino difference

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 needing unnecessary, costly data movement.

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



With universal access to your data, your company will have the visibility it needs to make improved business decisions in a fraction of the time and migrate data when it's best for you. Trino eliminates the need to maintain a traditional data warehouse and separates storage from compute, allowing organizations to leverage low-cost storage without sacrificing insights. In addition to tremendous cost savings, Trino improves productivity and time to insight from actionable data, enabling decision-makers to impact the business with faster, more intelligent, data-driven outcomes. Although Starburst is built on open-source Trino, the two have several significant differences.



Managing your Trino deployments is a complex task. Trino is an incredibly powerful engine that requires extensive resources to manage and scale. Inadequately resourced teams are challenged with deploying and scaling infrastructure and managing day-tday operations. That's why so many Data teams turn to Starburst to ease that burden and help reduce their total cost of ownership (TCO). Managing your Trino deployments is a complex task. Trino is an incredibly powerful engine that requires extensive resources to manage and scale. Inadequately resourced teams are challenged with deploying and scaling infrastructure and managing day-tday operations. That's why so many Data teams turn to Starburst to ease that burden and help reduce their total cost of ownership (TCO).

Starburst offers a full-featured data lake analytics platform built on open-source Trino

With Starburst, teams can access their data, lower infrastructure costs, use the tools best suited to their specific needs, and avoid vendor lock-in—data teams from the most prominent enterprises to recently minted digital native businesses trust Starburst. With Starburst, you can evolve your Trino deployments to include enterprise-ready enhancements for improved performance, better connectivity, more security controls, and expert support.

To illustrate the total business value of Starburst's data lake analytics platform, we look at the following benefits:

Simplicity — Operate more efficiently with an intuitive interface and out-of-the-box query editing, data discovery, and data product capabilities.

Access — Access enhanced and additional connectors for data sources like Snowflake, Salesforce, Teradata, etc.

Scalability and performance — Easily manage and scale your architecture to support ever-growing demand while gaining autonomous workload acceleration for your data lake queries with advanced indexing and caching technology.

Optionality — Break away from the cycle of data siloes and a single source of truth approaches by taking control and ownership of your data and never get locked into proprietary data ecosystems.

Security and governance — Leverage built-in access controls with role and attribute-based configurations and guarantee enterprise-grade security from the client to the underlying data source.

Support — An enterprise-grade professional services team and the world's largest group of Trino experts back you.

Empower access to all distributed data through every stage of the migration process

Starburst unlocks access to data across disparate clouds, regions, and ground-to-cloud. It is cluster-to-cluster connectivity. On the slide below each of these clusters sits numerous data stores. With Starburst, customers link catalogs or sources supported in one cluster, with those data sources supported in remote clusters. Starburst becomes a gateway for unlocking data access across geographies while ensuring access controls are in place and data residency requirements are honored—no more cloud data lock-in, empowering organizations to take control of constantly increasing egress fees.

Before your migration — connect your data silos

Organizations house diverse data across sources such as databases, warehouses, and open-format storage, spread across locations including on-premises and the public cloud. By implementing data federation before migration, data teams can unify data silos through a SQL-based semantic layer, establishing a single access point model without undue complexity. This layer enables seamless transitions in data locations and formats without disrupting users, freeing IT to focus on migration strategy and architecture.

Transitioning users, tools, and services to this unified layer simplifies their experience. Instead of interacting with direct data sources, they interact with the data access layer. When sources or formats change, repointing within Starburst becomes effortless, ensuring simultaneous, smooth migration for all users.

Leveraging Starburst product capabilities at this juncture amplifies advantages:

Stargate:
Facilitates linking remote Starburst clusters, allowing data access across geographies and clouds while maintaining access controls and data sovereignty compliance. Multi-cloud and hybrid analytics is enabled.

Data products:
Merging curated data with APIs creates publishable, manageable data assets for downstream users. This simplifies data access and management during migration.

Warp speed:
Optimize query performance and costs via Starburst's caching and autonomous updates, ensuring efficient analytics while managing expenses.

During migration preparation, seamlessly transfer Starburst's semantic layer to new cloud data lake storage without disruptions or pipeline rebuilding. This accelerates modernization post-migration, enabling enterprise-wide access to data. Starburst empowers analytics across datasets, both on-premises and in the cloud, facilitating migration without impeding workloads. Migrate your data and workloads to the cloud progressively, aligning with your pace and needs.

During your migration — ensure business continuity

Flexibility for optionality. As data is migrated from one system to another, views are updated to point to the new location with zero downtime or interruption for the end users. In other words, it keeps the business moving forward.

Since Starburst can query low-cost data lake storage at terabyte scale, this allows more flexibility in determining where your data lives. With the ability to query across different platforms, choosing the lowest cost storage with open file and table formats will future-proof your architecture, avoiding vendor “lock-in.” This also allows the combination of historical, operational, and real-time data sources for the most complete and accurate view throughout the data life cycle.

While continuing to leverage the SQL-based semantic layer created before the migration, IT teams can concentrate on moving the data, and analytics users can get fast, reliable, accurate, and timely data without interruption. The data products created prior to the migration can be constructed such that the analytics users do not even know there is an underlying migration happening. That is because data products can process data in place and do not require any movement of the data not prioritized to be migrated. It reduces the need for techniques such as Extract-Transform-Load (ETL) or Extract-Load-Transform (ELT) which create numerous data paths, vulnerabilities, and the complexity of the overall data landscape. Time to insight shrinks as data movement before executing a query is no longer necessary. Users can explore data sets and join them with little effort, avoiding the natural silos that can reemerge.

Finally, through the migration, data federation with Starburst is accompanied by data virtualization techniques that enhance performance. The ability to create materialized views and cache both raw and intermediate data enable efficient execution of production queries and exploration queries leveraging popular data sets.

Enhancing business with modern data lake analytics post-migration

Completion of a data migration marks just the beginning of leveraging Starburst's value within your data architecture. A host of new capabilities and benefits emerge:

Unlocked capabilities and realized value:

Data lake query engine:
Rapidly access data within the data lake with high efficiency.

Federated queries:
Seamlessly join data across various sources and types without physical data movement.

Cross-cloud queries:
Federate queries spanning on-premises and public clouds, reducing data transfers between locations.

Starburst Gravity:
Empowers data asset management via universal discovery, governance, and sharing features like universal search, cataloging, and data products.

Materialized views:
Establish persistent integrated data copies for low-latency, repetitive use scenarios like reporting.

Data products:
Facilitate organized, certified, and managed data assets publication for enhanced reuse.

Data localization:
Ensure compliance with regional data processing regulations.

User benefits:

Data analysts, scientists, and analytics engineers

- High-performance data access and exploration
- In-place data exploration without central storage dependency
- Reduced reliance on IT for provisioning centralized data

Data engineers:

- Develop data products for frequently used data assets
- Explore, sample, and model data assets without data movement

As you plan future data migrations, Starburst's presence streamlines the process, thanks to its pre-migration and during-migration capabilities.

Starburst empowers your analytics journey to create a more efficient and powerful data environment.

Taking the next step in your transformation to build and enhance AI/ML applications

Progressing to build and enhance AI/ML applications is your next transformational step. This aligns with tapping into new capabilities offered by modern data lake architecture. With AI's advancement and the widespread presence of data science teams, some facets have become standardized. However, the crux of AI and machine learning's success is the quality of underlying data. Merely applying technology won't suffice. Despite hiring skilled data scientists, investing in cutting-edge software, and even organizational restructuring for AI integration, many efforts still fall short.

Unlike the past, accessing algorithms or tools is now less arduous. Open-source languages like R and Python, alongside vibrant communities and expanding vendor technologies, aim to optimize data science and AI resource utilization. Yet, mastery in this realm remains a distant horizon.

Amid today's AI applications, the question arises: how effectively are AI-generated insights employed? The current AI challenges revolve around scaling applications and ensuring usable outputs, which starts with comprehensive training data.

Data solutions for your AI teams

Supporting a range of user tools presents challenges for data engineers due to AI outcomes' historical confinement. Enter "data products." This trend involves sharing reusable data products across AI engines and teams, encapsulating both outputs and training data. These curated datasets transcend systems, harmonizing legacy and modern systems. Tailor them to cater to user needs and interlink old and new seamlessly.

Data products should reside within a consumption layer, offering flexibility, reusability, and security. Extracting aggregation rules from data sources and client tools to the intermediate consumption layer ensures uniformity across AI solutions and data sources. It also establishes a centralized security checkpoint for data access. This approach bridges legacy and modern infrastructures, simplifying data comprehension for end users.

Expanding AI organization-wide

Scaling data and analytics faces common hurdles in legacy environments, especially within large enterprises and government agencies. Effectively powering an AI engine becomes challenging when data is spread across modern and legacy systems. While the optimal solution varies, centralizing all data isn't practical. Consider ChatGPT's power, derived not just from code but vast data access. Waiting for data centralization is unrealistic for ChatGPT, organizations, and departments alike.

The core challenge lies in scaling applications amidst complex data ecosystems.

The present race centers on data accessibility. AI's value hinges on timely, high-quality data. Modern environments consist of intricate technology ecosystems, intensifying the challenge of delivering accurate data, fittingly formatted, to the right individuals as needed.

Conclusion

Data migrations are an inevitable reality for every growing and transforming business. At Starburst, we fundamentally believe that the future success, and this isn't decades out, but instead starting within the next year, will be dependent on businesses, departments, and their teams' ability to quickly and securely access all the relevant data to extract the most meaningful insights that move the needle on core business KPS. To achieve this, a modern data lake (or lakehouse) architecture will propel businesses faster because they have gained the following:

A system that is scalable and cost-effective because it is built on commodity storage and compute

An architecture that is vendor agnostic because it is built on open file and table formats bringing the control and ownership back into your hands

A single point of access and governance for data in and around the lake provides speed and peace of mind

To learn more, visit www.starburst.io or test out Starburst for yourself with a free trial to experience the possibilities of a modern data lake analytics platform and what it can do for your business before, during, and after your upcoming data migration.

Advanced warehouse-like capabilities directly on the data in the lake using SQL.

Starburst

Data Lake Analytics Platform

$3.3B
valuation

$414M
raised

Leading investors

andreessen.horowitz
Index Ventures
COATUE
salesforce ventures
ALTIMETER
ALKEON CAPITAL MANAGEMENT
B Capital Group

Original creators of

trino

Deployed at exabyte scale at 4 out of 5 FAANG companies; adoption across thousands of companies globally.

presto

Originally created at Facebook to query **300PB Hadoop cluster**; thousands of active users today.

200+
customers

100%
YoY growth

85
NPS

Industry-recognized Leader

Gartner Market Guide for Analytics Query Accelerators
GigaOm Radar Report for Data Lakes and Lakehouses
G2 Enterprise Grid for Big Data Analytics Software

About Starburst

For data-driven companies, Starburst offers a full-featured data lake analytics platform built on open-source Trino. Our platform includes the capabilities to discover, organize, and consume data without time-consuming and costly migrations. The lake should be the center of gravity but support accessing data outside the lake when needed. With Starburst, teams can access more complete data, lower the infrastructure cost, use the tools best suited to their specific needs, and avoid vendor lock-in. Trusted by companies like Comcast, Grubhub, and Priceline, Starburst helps companies make faster decisions on all their data.

Learn more about Starburst Data Products at starburst.io

Starburst

Copyright © 2023 Starburst

STARBURST.IO