Starburst
ANALYTICS ANYWHERE

Starburst
Technical Solution Brief

In this Document

Trino Overview **3**

Challenges: Solving for Access in a Big Data Environment **4**

Solution: Starburst Enterprise **6**

Key Benefits 7

How it Works 8

Starburst Enterprise Use Cases

Interactive Data Investigation 9

Business Intelligence 10

Data Science 11

Starburst Enterprise Advantages & Features **12**

Trino | Starburst Is Enterprise-Ready

Open-source, fast and scalable distributed SQL engine.

Trino is a high performance distributed SQL engine for running fast analytic queries against various data sources ranging in size from gigabytes to petabytes. Designed for the separation of storage and compute, Trino scales on demand, and eliminates the time and cost of integrating disparate data into a single data warehouse.

Analyze Anything

Allow your data scientists to query any data source, from traditional warehouses to cloud data lakes, through their favorite BI tools (Tableau, Power BI, Looker, Qlik, ThoughtSpot and many more). Trino's flexible architecture allows you to perform analytics federated across multiple data sources at the same time.

Deploy Anywhere

The separation of storage and compute makes Trino uniquely flexible. Deploy on public clouds such as AWS, Azure and GCP in addition to private clouds such as OpenStack, on premises on bare metal commodity hardware, or in a virtualized environment with a Kubernetes containerized deployment.

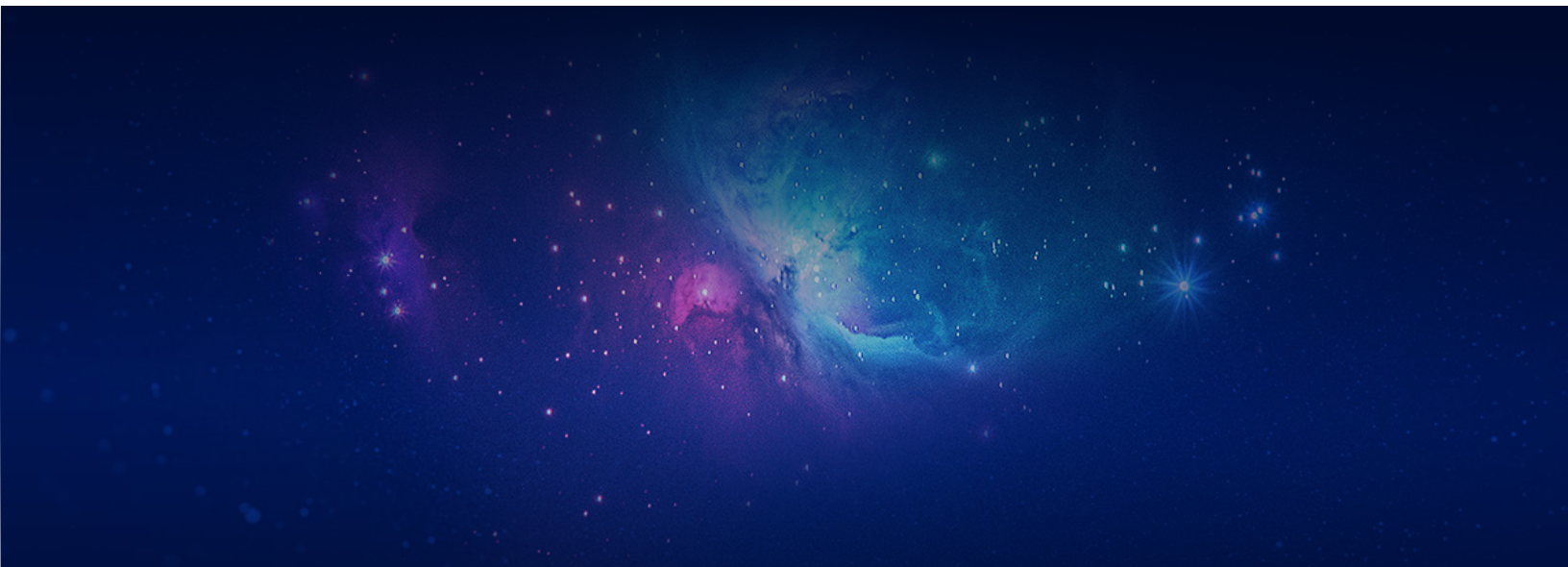
Separation of Compute & Storage

Trino can query data wherever it lies, allowing companies to separate storage and compute. There is no need to move your data, and you can provision compute to your exact needs, scaling up or down based on analytics demand, which results in significant cost savings.

Starburst offers the only enterprise-ready distribution of Trino, complete with stronger security, high-performance connectors, 24x7 support from the Trino experts, and more.

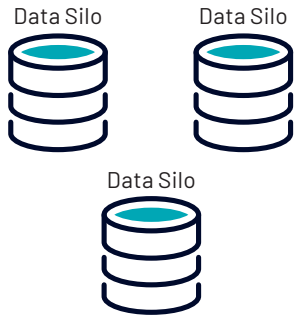
Starburst Enterprise

The Starburst Enterprise distribution of Trino was created to help large companies securely extract more value from their Trino deployments. With Starburst, enterprises have more Trino tools at their disposal, global security with fine-grained access controls, stable & reliable releases, and guidance and support from the most qualified group of Trino experts anywhere. With Starburst, your data architects, business intelligence analysts, and machine learning and AI teams have fast, reliable, stable access to the data they need to do their best work —no matter where that data resides.



Challenges | Solving for Access in a Big Data Environment

The Walls Between Analysts and Enterprise Data



The Consumers

Data analysts, data scientists and business intelligence analysts require access to the most complete, relevant data sets to do their jobs, regardless of where that data resides.

The Procurers

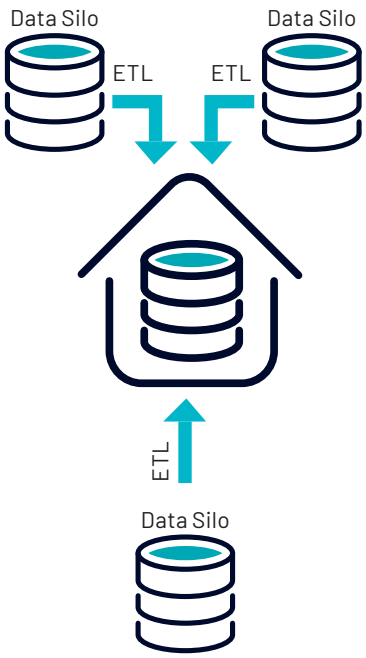
Data engineers and technical staff are responsible for making data available, but they need to maintain security and compliance, ensuring that only authorized individuals have access to specific data sets —or specific columns and rows within tables.

The Problem

In large enterprises, data typically exists in a variety of locations and formats. Analysts are often left without access to critical data—if they even know it exists—and technical staff are saddled with platforms that lack the necessary security and access controls.

The Traditional Data Warehouse Approach

Traditionally, companies try to solve this problem by building a data warehouse.



Advantages:

- Analysts have one place to go to find data for analytics.
- Security can be applied in one place, instead of bolted onto numerous systems.

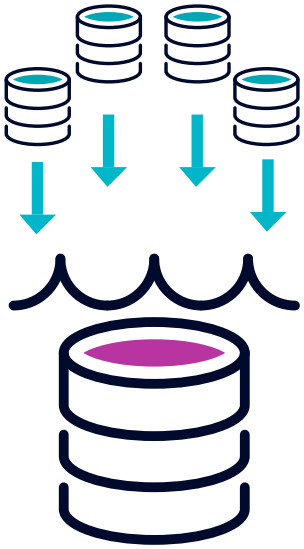
Disadvantages:

- Data warehouses are very costly to build and maintain.
- Expensive ETL pipelines are required to migrate data into the warehouse.
- Time delays: Source data must be accumulated before it is written to the warehouse.
- Storage and compute are one and the same, which means you must overbuy hardware relative to your anticipated requirements, and pay for expensive compute infrastructure that will often lie dormant.
- Vendor Lock-in: Once it's stored in a traditional data warehouse, your data isn't really your data anymore because it is locked in a proprietary format. This renders you less nimble, making it harder to extract your data and bring it into your larger ecosystem. Anything you do with that data has to be done through the vendor.

The Modern Data Lake Approach

A more modern solution is to build a data lake. This takes one of two forms:

- **On-premise** with Hadoop clusters or other on-prem, private cloud infrastructure.
- **In the cloud** with unstructured object storage via a cloud provider such as Amazon AWS, Microsoft Azure or Google Cloud.



Advantages:

- Cheaper storage and lower overall costs relative to the traditional data warehouse.

Disadvantages:

- Existing solutions for accessing data in the lakes are slow and unreliable.
- Making the right data available in the right form is time consuming.
- Data often remains federated, as it is generally not practical for all data to live in one location. Enterprises will still choose the optimal storage solution for a given dataset, and that means data may be stored in more than one place. A company might choose to store historical customer data in a low-cost data lake, whereas recent credit card transactions would more likely reside in a high-performance database, then get archived off after a certain number of days or weeks.

The modern data lake solves some of the issues associated with the traditional data warehouse, but it fails to offer the performance, flexibility, access, and fast time-to-insight modern enterprises deserve.

Solution | Integration in a Big Data Environment

Starburst Enterprise

The modern data lake began as an attempt to reach a state of big data nirvana, in which all of your analytics tools are accessing data in one place. But in practice this has not been realized. Even in the cloud data lake model, the connectors and/or retrieval methods typically used to access and query data do not meet enterprise standards, and the drive to use the optimal storage solution for each dataset results in federated data.

Starburst Enterprise simulates this everything-in-one-place nirvana state by operating as a consumption layer between your analysts and their disparate data sources. Yet Starburst does this with fast SQL performance, strong security controls, and other features one would normally associate with a traditional database. Instead of bringing users to their data, Trino erases the gap between them. This results in a simple but transformative paradigm shift in data analytics. Instead of a single source of truth, Starburst Enterprise gives companies a single point of access.

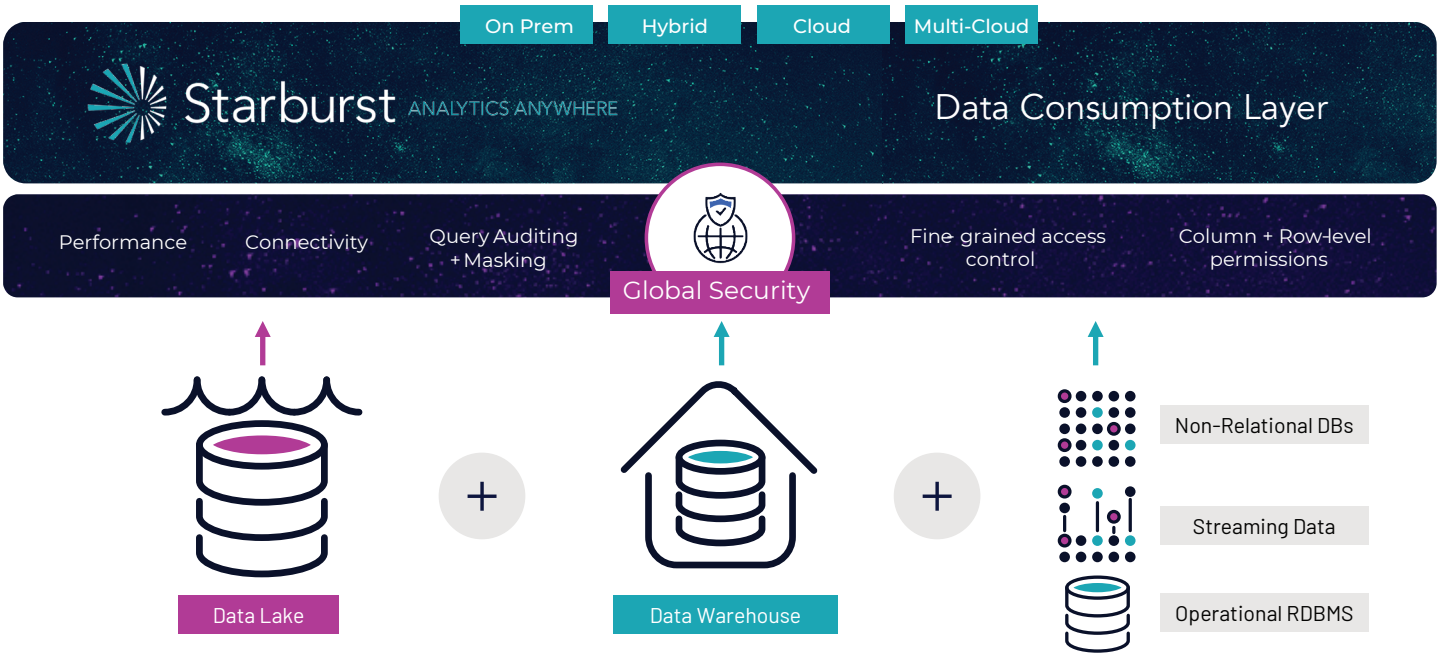
Starburst Enterprise gives companies a single point of access.

Trino was built to be a high-performance distributed query engine that solves the federated data access problem. Starburst Enterprise optimizes open-source Trino for the enterprise. Today, most BI Analysts, Data Scientists, and ML/AI experts are disconnected from the data they need, and Starburst delivers it to them quickly and securely, without the need for ETL or expensive warehouses. Starburst ensures that analysts have fast, secure access to the data they need.

A Single Point of Access with Starburst Enterprise

High-performance connectors serve data from various sources up to the client through a SQL interface. The source data itself is never actually stored within Starburst Enterprise, but once a connector has been registered, the data in the source immediately becomes available through an ODBC/JDBC connection or command line interface.

Starburst Enterprise is ANSI SQL compliant to support all queries and reporting requirements, and our developers, who have committed roughly 85% of the open source Trino code, are constantly developing new functionality and adding optimizations to the Starburst enterprise deployment package.



Key Benefits for Data Consumers

Familiar Interface: Access to data appears to be in a relational database irrespective of where the data actually resides, and data scientists and analysts can operate in their favorite BI tools and SQL clients, as well as R Studio, Jupyter, and more.

Federated Access: Data from disparate sources or stored in different formats can be joined and operated on as if it were in the same database.

Rapid Access: Trino is the high-performance distributed SQL query engine. Companies can query data in cloud data lakes as quickly as they can query other storage platforms, allowing them to save on storage without sacrificing insights.

Key Benefits for Enterprises

Cost Reduction: By eliminating the need to maintain a traditional data warehouse and separating storage from compute, Starburst allows companies to leverage low-cost storage without sacrificing insights. Enterprises can also dynamically scale compute to match the needs of their end users, ensuring optimal use of resources.

Immediate Productivity: Starburst Enterprise is fast, which means your analysts and ML/AI experts won't waste time waiting for results. Plus, the technical complexity is abstracted away, so technical teams merely need to set up and provide access. No one has to be retrained or educated extensively.

Eliminate Wasteful Efforts: Starburst eliminates data wrangling and other time-consuming tasks, allowing your technical teams and highly-trained data scientists to focus on what they do best.

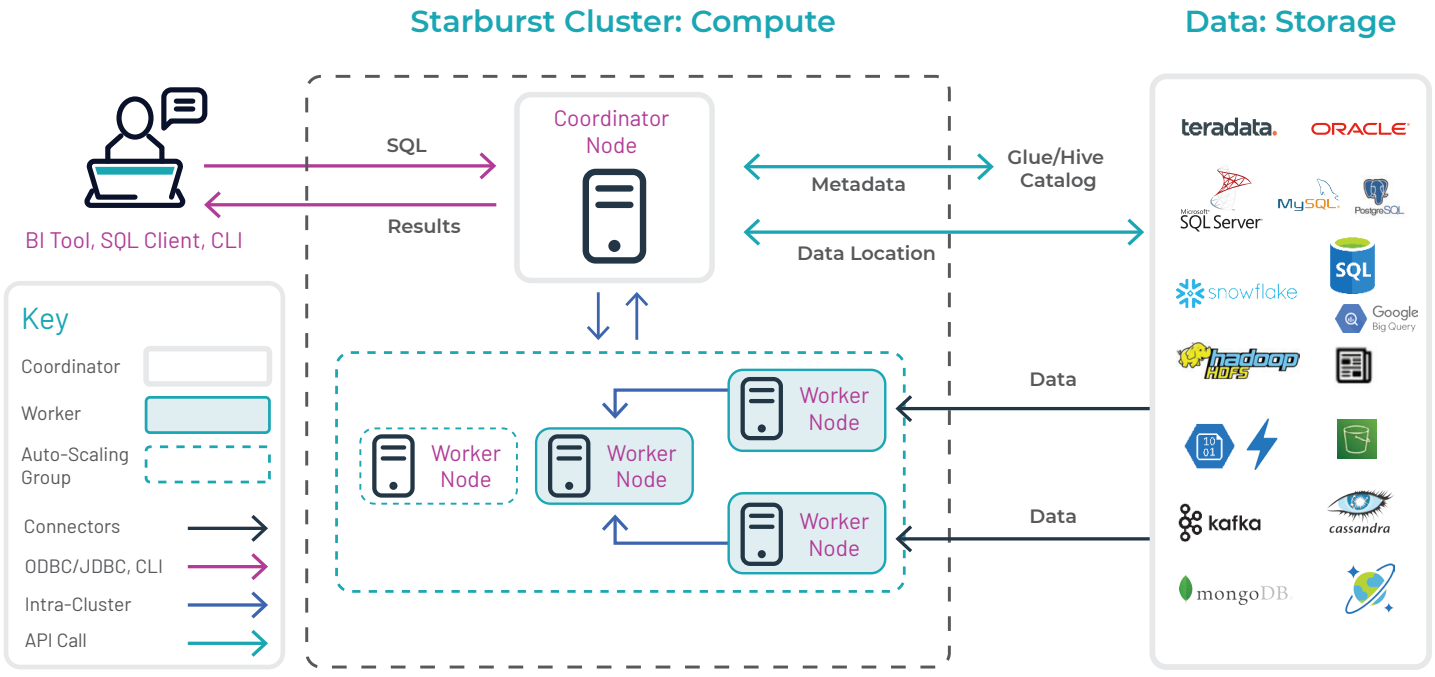
Reduced Time to Insight: Since companies no longer have to build ETL pipelines and wait for data to be migrated from one place to another, Starburst accelerates project timelines, reducing the time to generate actionable insights. As an example, one Starburst client, a large telecommunications provider, wanted to query two federated datasets to identify upsell opportunities within its subscriber base. The project timeline would have been 18 months if they'd opted for ETL. Starburst reduced it to a few weeks.

Data-driven Business Initiatives: The telecommunications example above isn't merely about reducing project timelines. By using Starburst to access, join and query previously siloed data sets, analysts were able to identify opportunities that drove $200M in new subscription revenue.

Architecture

How Starburst Works

1. The analyst submits a query or SQL statement through his or her chosen business intelligence tool, SQL client, or CLI. This statement then moves to the Trino cluster, which is composed of one Coordinator Node and one or more Worker Nodes. (These worker nodes can be scaled up or down to meet demand.)
2. The Coordinator Node parses the statement to ensure it's correct and that the analyst has permission to access the data requested. If these conditions are not met, the analyst will receive a message or alert. If everything is in order, the Coordinator Node identifies the most efficient execution path, optimizing for performance and cost.
3. Next, the Coordinator schedules or assigns the work to the appropriate Worker Node(s), and data is streamed out of the requested source(s) via connectors.
4. Data is processed in a series of stages according to the optimization plan. Processing is accomplished by a highly efficient, in-memory, pipelined distributed system. The results are streamed back to the analyst when processing is complete.



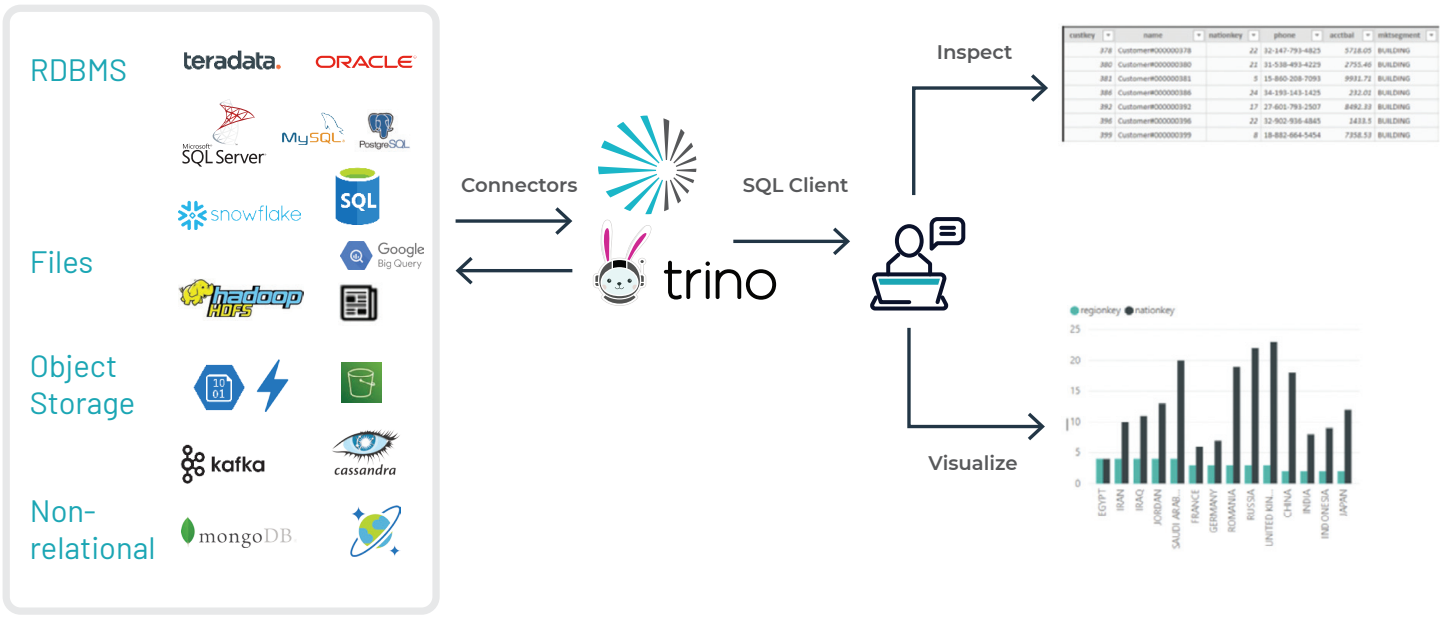
Use Cases | Integration in a Big Data Environment

Use Case #1: Interactive Data Investigation

Rapid ad-hoc interactive queries and analysis

BI analysts and data scientists require fast interactive communication to profile and understand their data before writing reports, performing analysis, and generating predictive models. They need this data back as quickly as possible, and Starburst Enterprise carries out this role better than any other solution.

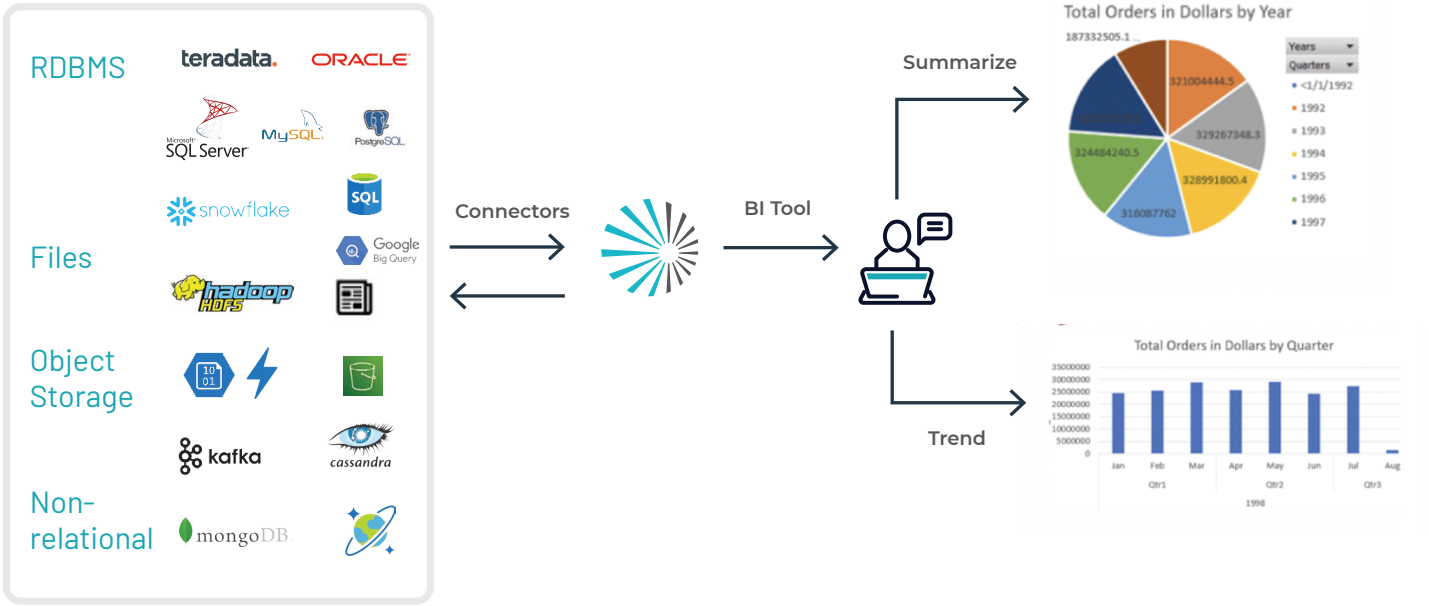
With Starburst, consumers can query underlying sources from their SQL or BI tool of choice. A typical example would be using your favorite SQL client such as SQuirreL or Toad or PowerBI to return and inspect raw data. Data can be queried rapidly from a single source or combined through federated joins, and results are returned within five to ten seconds or less.



Use Case #2: Business Intelligence Dashboards and Reports

Business Intelligence dashboarding and reporting requires rapid access to data in either a single source or federated sources.

Business intelligence analysts, risk and regulatory teams, and other users in this group need fast access to data for a variety of reporting functions. These analysts aren't necessarily writing SQL queries. They're building pie charts and dashboards in Tableau, Qlik, Looker, Power BI, ThoughtSpot, and other BI tools. Previously, it wasn't possible to query different data sources via the same connection in their chosen BI tool. Analysts resorted to creating mini-data marts. Starburst allows them to work in their preferred tool while vastly expanding their data access.



Use Case #3: Data Science

Data scientists need access to data for model development and machine learning purposes to support a variety of lines of business such as marketing and customer segmentation as well as attrition risk, financial risk, fraud detection and data quality.

Data scientists work with a variety of tools and languages such as Python, R and Spark to name a few. They need rapid ad-hoc access to data sources while constructing and investigating new data models. Once a model is in production they also require access in much larger data volumes. Historical data from archival sources often needs to be analyzed along with current data sources to build more accurate models, but this data is typically stored in a separate archival or cold storage medium. As a result, highly qualified data scientists may spend up to 50% of their time wrangling data.

This is a tremendous waste of time and resources. Starburst provides rapid access to both single source data like relational or unstructured blob storage and federated sources. Furthermore, Starburst is built to scale. **Data scientists can quickly access large volumes of source data via their tool or language of choice, and focus more of their time on actually doing data science.**

Starburst Enterprise Features

Starburst is committed to the open source Trino project. Our engineers frequently contribute back to Trino's core functionality, core performance, stability and bug fixes. Despite its strengths, the open source distribution of Trino is not optimized for today's time and resource constrained enterprises. Starburst Enterprise contains features not found in the open source version of Trino, including:

- Global Security & Fine Grained Access Controls
- High Performance & Secure Connectors to More Data Sources
- Secure Data Caching
- Additional Certified Connectors
- Enterprise ODBC and JDBC Drivers for Business Intelligence Tools
- Kubernetes Integration for Simplicity of Installation & Portability
- 24x7 Support from the Trino Experts
- Advanced Audit Logging for Security & Operational Metrics

Global Security & Access Controls

Any solution that provides access to enterprise data must maintain strong security and access controls. This is the primary driver for many of our customers. These enterprises store sensitive data such as credit card numbers or Personal Health Information (PHI) and need to ensure that this data is secure. Large enterprises also need to be certain that their analysts and data scientists can only access data they are authorized to investigate and query. Allowing everyone to view compensation and salary details, for example, could be damaging to an enterprise.

Security and access control are built into Starburst Enterprise. The first layer involves authentication, or determining whether an individual has the proper credentials to access data. This is part of the open source distribution as well. The second layer, authorization, is exclusive to Starburst, and essential to the enterprise, as it determines which data sources, and which tables, rows, and columns within those sources, each user is allowed to access.

Permissions Control

Apache Ranger provides role-based access control, and benefits include schema/table/column access control in addition to column-level masking and row-level filtering. Other Starburst-only features include Okta authentication.

Global Security

Starburst extends these security controls to all enterprise data sources. The ability to centralize policies for permissions and access results in stronger, easier-to-manage security.

AuthN & AuthZ

Credentials are maintained in your source system of choice, and those policies can only be altered through the Ranger or Sentry UI.

Query Audit Logs

Starburst keeps an audit log that details the time a query was submitted, the user who submitted the query and more.

Starburst Secrets

Starburst does not maintain credentials such as username and password – these are securely plugged in at runtime.

Encryption

Starburst supports encryption for intra-cluster data streaming and data at rest.

How It Works

When a query is submitted, Starburst inspects the identity of the user who submitted the query and determines if they have access privileges to the cluster. After successful authentication, fine-grained role-based access control privileges are applied to ensure that the user is allowed to query the requested sources, and tables, columns, and rows within those sources. A permissions error message is returned if insufficient privileges are detected at any stage.

Additional Certified Connectors & Drivers

Starburst Enterprise provides additional, certified proprietary connectors not available to the open source community. The list includes **Parallel Teradata, Parallel Snowflake, Oracle, BigQuery, MapR, Greenplum, SAP Hana, and DB2**. Additionally, **certified ODBC and JDBC drivers** are included as is **AWS Glue metastore integration** which is used to maintain data sources and statistics for cost based query optimization.

The Starburst team is always adding to our toolkit of certified, tested, high-performance connectors, ensuring enterprises can access all their data, no matter where it resides.

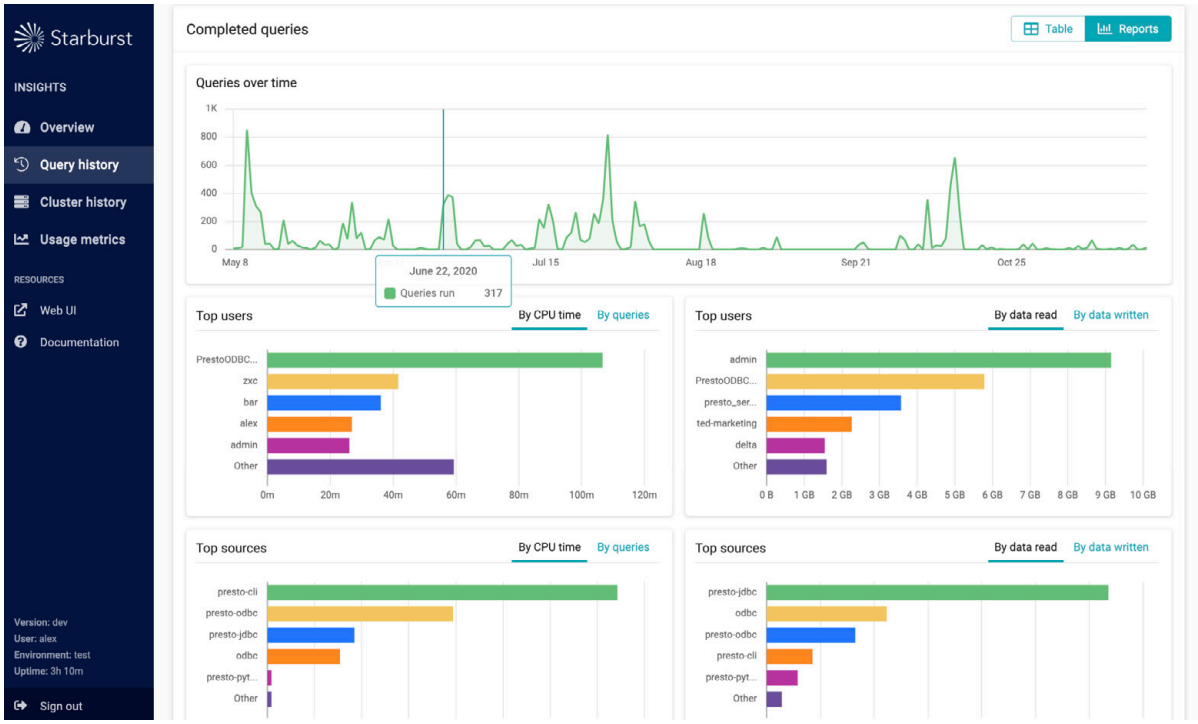
Kubernetes Integration

While open source Trino includes limited Kubernetes integration, Starburst Enterprise offers full integration with Trino on Kubernetes, Amazon Container Service for Kubernetes, Azure Kubernetes Service, and Google Kubernetes Engine. This has proven to be an extremely popular deployment method among our customers, and applies to all cloud, on-prem, and hybrid deployments.

1. **Cost Savings:** Running compute in the cloud is expensive, and Kubernetes augments Starburst's ability to scale compute up or down to meet the needs of your end users.
2. **Automatic Failure Detection:** If something goes wrong, Kubernetes immediately replaces the malfunctioning Worker Node or element.
3. **Simple Scalability:** With Kubernetes, replicating nodes and scaling up is as simple as clicking a button; Kubernetes automatically wires everything together, functioning as a highly efficient orchestration layer that manages the maintenance, running, expansion, and collapse of applications running in a Kubernetes cluster.
4. **Easy Deployment:** The initial spin-up and deployment of Starburst Enterprise is simple with Kubernetes, ensuring that your analysts and data scientists can start enjoying the benefits of Starburst even sooner.

Starburst Insights

Starburst Insights provides a visual overview of important metrics about your Starburst Enterprise cluster for all types of users, from platform administrators to data consumers. Within the interface, you can access detailed query history, including single-query statistics and query plans as well as cluster performance information across a selected date range. This allows operational teams to optimize performance, functionality, and projected capacity needs.



The Latest Cost-Based Query Optimization Features

As part of our commitment to open source Trino, Starburst chose to make the initial cost-based optimizer available to the OS community, but our engineers are continuously developing optimizations and add-ons that are available only to Starburst Enterprise customers.

24x7 Support for Starburst Enterprise

In addition to added features and enhancements, Starburst customers have access to the most experienced group of Trino experts in the world. Our engineers have contributed 85% of the open source Trino code. Our service includes 24x7 support, rapid severity 0 response times, and contact channels over email, WebEx and telecon. Additional benefits of Starburst support include:

- Faster Deployments
- Configuration Guidance
- Performance Tuning
- Cost Savings
- Rapid ROI

Training, ramping, and maintaining your own Trino support team would be massively expensive, and your internal teams still might not achieve the level of expertise of our group. With Starburst, your analysts and data scientists can focus on what they do best, and leave the configuration and optimization of your Trino deployment to us.

Starburst

ANALYTICS ANYWHERE

For more information on Starburst Enterprise, or to try the platform yourself, send us a note at

https://www.starburst.io/contact/