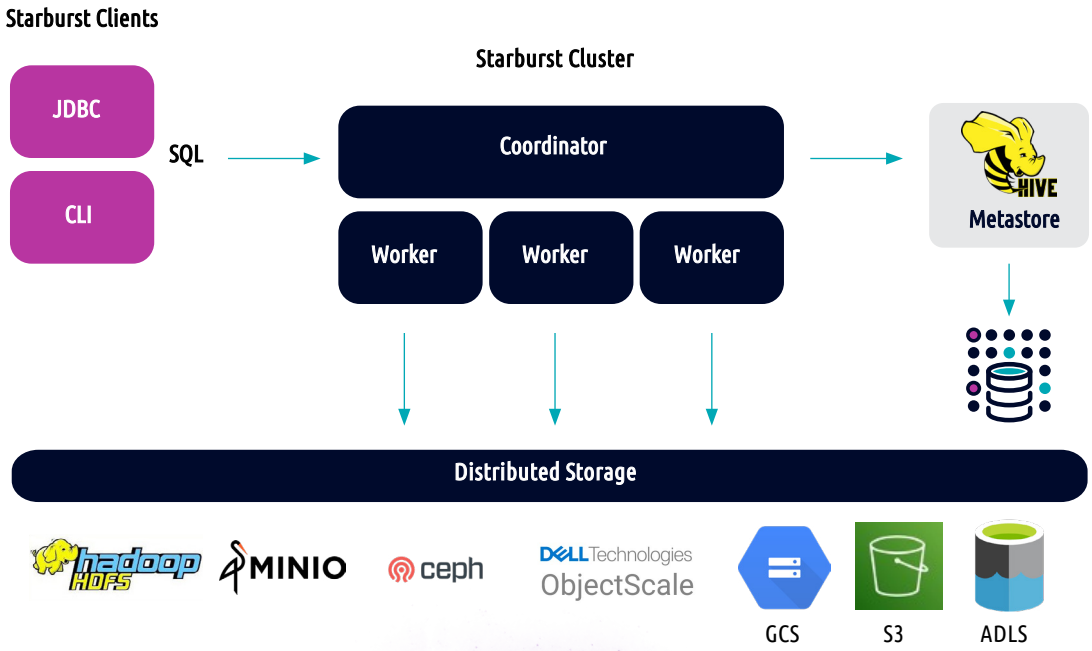
Starburst

Starburst Enterprise Improves Performance Over Hive and Impala

Customers look to Starburst Enterprise when they are experiencing an intensely slow query turnaround from their existing Hadoop, Spark, or Hive infrastructures. Hive was created in 2008 at Facebook for long-running ETL workloads, and was used for querying and analytics. In fact, the genesis of Starburst Enterprise's core open source engine Trino came about due to these slow Hive query conditions at Facebook back in 2012.

Starburst's runtime replaces Hive runtime

In the early days of big data systems, query turnaround was expected to take a long time due to the high volume of unstructured data in ETL workloads. Now, more businesses want to run fast interactive queries over their big data instead of running jobs that take hours and produce possibly undesirable results. The Trino engine utilizes the existing metastore metadata and files residing in storage, and the Trino runtime effectively replaces the Hive runtime responsible for analyzing the data. It's the fast access businesses require for the insights to make better decisions.



Architecture differences:

The largest difference architecturally between Starburst Enterprise and Hive is the dependency of Hadoop. Without the administrative overhead and constant upgrades, data teams can focus their efforts elsewhere using Starburst. While Hive was built for batch-processing, Starburst was built for fast, federated queries with high concurrency and low latency. The approach Trino takes dealing with data in distributed storage is reusing all of the components of Hive except for the runtime. This also simplifies the migration from using Hive to using Trino. The Starburst Enterprise Hive connector, and open source Trino, have a lot of similarities, however Starburst has added several very important features such as Ranger security integration for role-based access control.

Feature differences:

Features	Hive	Impala	Starburst Enterprise
Framework	Hadoop	Hadoop	**Distributed**
Connectivity	HDFS, ADLS, GCS, & S3	HDFS, ADLS, GCS & S3	**HDFS, ADLS, GCS, S3 +50 other enterprise sources**
Object storage	Yes	Yes	**Yes**
Operational RDBMS	No	No	**Yes**
Non-Relational DBs	No	No	**Yes**
Streaming sources	No	No	**Yes**
Data Federation	No	No	**Yes**
Concurrency	Low	Medium	**High**
Dependent on Hadoop	Yes	Yes	**No**
Requires YARN	Yes	Yes	**No**
Ranger/Sentry Integration	Yes	Yes	**Yes**
Latency	High	Medium	**Low**
Analytic workloads	Batch processing	Interactive analytics	**Interactive analytics & batch**

While Impala improves upon some of Hive's shortfalls, it's best used as a departmental solution with smaller data sets and which don't require high concurrency. Unlike Impala, Starburst Enterprise is not bound by Hadoop, and can federate across numerous other sources. Starburst is a stand alone technology not tied to any legacy technology. We enable businesses to leverage a real RDBMS, in addition to distributed storage, and offload the majority of workload they have on Hive/Impala onto something that delivers the responsiveness they need. On average, Starburst has helped customers gain 10x the performance of Hive and Impala at scale, at 1/3rd the infrastructure cost.

Graceful Scaledown and High Availability

When AWS auto scaling resizes a cluster, it starts decommissioning worker nodes. Starburst Enterprise has features to make sure this process doesn't disrupt the usage of a cluster. Most importantly, that no queries fail in the autoscaling. In addition to graceful scaledown, Starburst offers the high availability of the coordinator node. In the event the coordinator becomes unavailable, Starburst ensures the cluster automatically switches to a new coordinator and continues to accept new queries.

Advanced Connectivity and Performance

Starburst Enterprise includes a number of enhanced and enterprise-only connectors such as Delta Lake, Snowflake, Synapse, BigQuery, Redshift, Oracle, Teradata, and more. Starburst connectors additionally include enhancements such as parallelism, table statistics for cost-based optimizer, Kerberos support, and user impersonation and credential passthrough for more robust security.

Common Enterprise Use Cases

Data Lake Query Engine

Trino has become the de facto standard data lake query engine. A Starburst deployment enhances its functionality for the enterprise with role-based access control, autoscaling, high concurrency, ANSI SQL compatibility, and other benefits.

Data modernization

Starburst Enterprise lets you modernize data at your own pace, even as you work with the environment you have. Organizations can update, migrate, and move data as it makes sense for the business—without forced data migrations.

ETL workloads

Starburst Enterprise is ANSI SQL compliant and supports create table, and insert statements. It can act as the SQL engine for ETL jobs, providing a single platform for both query and migration needs. For example, archive data from an Apache Hadoop cluster could be moved to a data lake on Dell Object Storage, allowing federated Trino queries against that data as well as data from other sources that are not ready for migration.

- Get 10x the performance with 1/3 of the resources
- Less infrastructure cost
- Easier to manage
- Higher concurrency
- More connectivity
- Reduce ETL
- Hadoop not required
- Data federation
- Many file formats
- Fast and scalable

Interactive data investigation and long-running workloads

Starburst Enterprise enables rapid ad-hoc interactive queries, and long running ETL data processing from a range of data sources—including traditional, real-time, object stores, and so on. DBAs can query underlying sources from their SQL or business intelligence tools of choice. Data can be queried rapidly from a single source, or combined through federated joins.

Business intelligence dashboarding and reporting

Data consumers can work with their favorite BI tool of choice, such as Tableau, ThoughSpot, or Microsoft PowerBI for dashboarding and reporting. Because Starburst Enterprise separates compute and storage resources, it provides the interactive responsiveness that these tools require.

Data science

Data scientists need access to data for model development and machine learning purposes to support a variety of lines of business. Starburst Enterprise fulfills these requirements enabling data scientists to rapidly access large volumes of source data into their tool or language of choice through a standard ODBC/JDBC package interface.

Learn more about the differences between Hive, Impala, and Starburst Enterprise by contacting us today at starburst.io

Starburst