Starburst Enterprise on Red Hat OpenShift

Federated queries across multiple software-defined data sources

Highlights

Stop moving data and start unlocking its value by accessing and analyzing data from any data source at scale.

Gain insights quickly without worrying about where data is stored, what form it is stored in, or whether it is in flight or at rest.

Modernize your data infrastructure at your own pace while answering your most critical business questions.

Deploy compute and storage resources efficiently and cost-effectively to perform rapid reporting, business intelligence analysis, and SQL queries.

Introduction

Making better and more timely decisions is essential to improving performance and remaining competitive. At the same time, many large organizations struggle with how best to query and analyze isolated and disparate data sets in combination. Traditional data warehouse products approach the problem with an extract, transform, and load (ETL) process that copies data from one or more sources into a destination system. Unfortunately, this outdated and monolithic technique is inefficient and keeps business analysts from running fast analytics on their data, potentially delaying critical insights.

In contrast, Starburst Enterprise gives analysts the freedom to interrogate diverse data sets wherever they are located, without building a data warehouse. Organizations can run multiple clusters, scaling up or down dynamically and optimizing for query speed and cost as desired. Critically, Starburst Enterprise lets organizations access multiple software-defined data storage platforms deployed with Red Hat® OpenShift® Container Platform. For instance, data can be queried instantly and simultaneously from a data lake built on Red Hat Ceph® Storage and a SQL or NoSQL database running on Red Hat OpenShift Data Foundation in addition to data sources existing in myriad other environments.

Starburst Enterprise on Red Hat OpenShift

Distributed cloud and hybrid cloud applications are increasingly popular, but the transition to cloud deployments cannot take place instantly. As a result, many organizations rely on a combination of traditional applications as well as more modern applications to run their business and make critical decisions. Likewise, most employ a combination of both traditional and modern data sources with data scattered across datacenters, cloud, and vendor environments.

For example, an analyst might need to combine data from a PostgreSQL app on a Kubernetes persistent volume (PV), general ledger running in Microsoft SQL Server, and archived client data in an object store. Enterprises need a reliable and consistent user and operational experience that lets them develop applications and analyze data from diverse sources rapidly while managing infrastructure effectively.

Starburst Enterprise

Starburst Enterprise provides a modern solution built on the open source Trino (formerly known as PrestoSQL) distributed SQL query engine. Trino was designed and written for interactive analytic queries against data sources of all sizes, ranging from gigabytes to petabytes. It approaches the speed of commercial data warehouses while scaling to the size of large organizations. Starburst Enterprise adds the tools and 24x7 support that organizations need for big data access at scale.

The Starburst Enterprise platform provides distributed query support for varied data sources, including:

- NoSQL systems (MongoDB, Cassandra, Redis).

- SQL databases (Microsoft SQL Server, MySQL, PostgreSQL).

- Data warehouses (IBM Db2 Warehouse, Teradata, Oracle Exadata, Snowflake).

- Hive (HDFS, Cloudera, MapR).

- Data services (Kafka, Elasticsearch, OpenShift Data Foundation).

- Cloud object storage (AWS S3, ADLS, Azure Blob, Ceph, IBM COS).

Starburst Enterprise lets organizations run analytics anywhere to make better business decisions.

Starburst on Red Hat OpenShift Container Platform

Operating Trino at scale can present challenges, especially for those attempting to size and configure their environments manually. Organizations need ways to achieve petabyte scale with autoscaling and to decommission resources gracefully when they are no longer needed. Combining Starburst Enterprise with Red Hat OpenShift Container Platform addresses these needs by offering automation, high availability, elasticity, and monitoring for Trino clusters.

Kubernetes operators delivered with Red Hat OpenShift Container Platform automate installation, upgrades, and life-cycle management throughout the container stack. A Kubernetes operator for Starburst Enterprise lets Red Hat OpenShift greatly simplify the administration of a Starburst Enterprise cluster. Together, these operators offer benefits that include:

- **Automation.** Red Hat OpenShift and Starburst Enterprise operators provide automatic configuration, tuning, and management of Starburst Enterprise clusters. Red Hat OpenShift operators determine what to deploy, including identifying the hardware and provisioning new instances. Starburst Enterprise operators manage updates to the environment.

- **High availability.** Continuous operation of the Trino coordinator is essential. Using liveness probes, the Red Hat OpenShift load balancer can keep services like the Trino coordinator in an always-on state.

- **Elastic scalability.** Red Hat OpenShift can automatically scale the Trino worker cluster based on query load. Using the Kubernetes Horizontal Pod Autoscaler (HPA), organizations can specify thresholds for Trino worker pods. As the number of queries increases, the HPA will automatically spin up additional worker pods based on specified system constraints.

- **Graceful scale-down and decommissioning.** With Red Hat OpenShift, reduced load does not imply system downtime or killed queries. The Kubernetes HPA will gracefully decommission unused Trino worker pods and free system resources for other tasks without service interruptions.

- **Monitoring for all hardware and software layers.** Prometheus, the cluster monitoring service for Red Hat OpenShift, delivers metrics and alerts that inform Kubernetes orchestration and populate the Red Hat OpenShift dashboard. Prometheus informs Kubernetes if a pod is offline and provides metrics to the HPA to let it know whether to commission or decommission additional Trino pods.

Red Hat OpenShift Container Platform automates the provisioning, management, and scaling of applications so that you can focus on writing the code for your next big idea.

▸ **Support for Red Hat Data Services.** Starburst Enterprise lets organizations use data storage platforms associated with Red Hat OpenShift. Application developers can make use of SQL and NoSQL databases backed by OpenShift Data Foundation. Businesses can pull data from data lakes running on Red Hat Ceph Storage archives. Parquet files on Red Hat Ceph Storage are also useful for data analytics.

Figure 1 illustrates how Starburst Enterprise works with Red Hat OpenShift Container Platform and related Red Hat Storage data storage platforms.
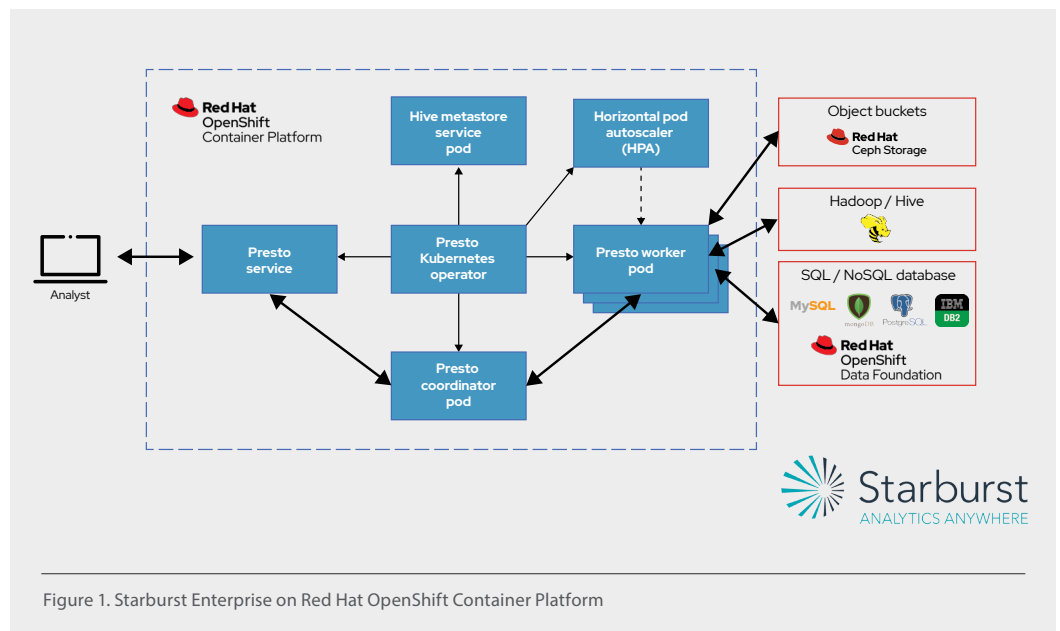

Figure 1. Starburst Enterprise on Red Hat OpenShift Container Platform

Use cases

Running on Red Hat OpenShift, Starburst Enterprise accommodates a wide range of use cases, including:

▸ **Data modernization.** Starburst Enterprise lets you modernize data at your own pace while 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 for support of 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, archived data from an Apache Hadoop cluster could be moved to a data lake on Red Hat Ceph Storage, allowing federated Trino queries against that data as well as data from other sources that are not ready to migrate.

- **Interactive data investigation.** Starburst Enterprise enables rapid ad hoc interactive queries from a range of data sources — including traditional, real-time, and object stores. Database administrators 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 (BI) dashboarding and reporting.** Data consumers can work with their favorite BI tool such as Tableau, Microstrategy, or Qlik 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 to support a variety of lines of business. Starburst Enterprise fulfills these requirements, allowing data scientists to rapidly ingest large volumes of source data into their tool or language of choice through a standard ODBC/JDBC package interface.

Flexible container-native storage services for Trino

The combination of Red Hat OpenShift Container Platform and Red Hat Data Services provides extensive flexibility for Starburst Enterprise, allowing it to access data stored on software-defined storage in a wide range of formats.

Red Hat OpenShift Data Foundation

Red Hat OpenShift Data Foundation is software-defined storage integrated with and optimized for Red Hat OpenShift Container Platform. It is built on Ceph , the Rook Kubernetes operator, and the NooBaa multicloud object gateway to provide container-native storage that supports a wide range of access methods, including:

- **Block storage** for stateful cloud-native applications including databases, document stores, and messaging systems.

- **File storage** for continuous integration/continuous delivery (CI/CD) build environments, web application storage, and for ingest and aggregation of datasets for machine learning.

- **Multicloud object storage** for CI/CD build artifacts, origin storage, data archives, and pre-trained machine learning models that are ready for serving.

Red Hat Ceph Storage

Red Hat Ceph Storage is an open, massively scalable, software-defined storage platform for petabyte-scale deployments. It provides performance at scale for file, block, and object data protocols along with a Ceph management platform, deployment utilities, and support services. The platform is engineered to be flexible and is intended for modern workloads including data analytics, artificial intelligence and machine learning (AI/ML), cloud infrastructure, media repositories, and backup and restore systems.

Conclusion

Starburst Enterprise and Red Hat OpenShift promote better and more timely insights by letting organizations analyze data across multiple disparate and distributed data platforms rapidly. The combination provides critical automation, high availability, elasticity, and monitoring that meets the demands of enterprise organizations. Backed by Red Hat OpenShift Data Foundation and Red Hat Ceph Storage, the solution supports a wide range of data sources with storage solutions that are designed and tested to work with Red Hat OpenShift Container Platform.

Find out more

Read the blog post
Watch the demo
See the webinar
Follow us on the web

About Red Hat

Red Hat is the world's leading provider of enterprise open source software solutions, using a community-powered approach to deliver reliable and high-performing Linux, hybrid cloud, container, and Kubernetes technologies. Red Hat helps customers integrate new and existing IT applications, develop cloud-native applications, standardize on our industry-leading operating system, and automate, secure, and manage complex environments. Award-winning support, training, and consulting services make Red Hat a trusted adviser to the Fortune 500. As a strategic partner to cloud providers, system integrators, application vendors, customers, and open source communities, Red Hat can help organizations prepare for the digital future.

facebook.com/redhatinc
@RedHat
linkedin.com/company/red-hat

North America
1 888 REDHAT1
www.redhat.com

Europe, Middle East, and Africa
00800 7334 2835
europe@redhat.com

Asia Pacific
+65 6490 4200
apac@redhat.com

Latin America
+54 11 4329 7300
info-latam@redhat.com

redhat.com
#F27299_0221