
Starburst

Starburst Proof of Value Pre-Requisites

Know Before You Go

Checklist

✓	Requirement	Description
	Available Compute Resources	Available compute resources for the SEP cluster and tools. **See the Compute Requirements section below.**
	Your Data Sources, Queries, and Client Tools	Your list of **data sources**, with login info, access keys for the backend. For the frontend, the planned **queries** and **client tools**.
	Confirmed Connectivity	**Tested connectivity** between the test lab/cloud, the data sources, and your laptops. This might include testing VPNs and cloud networking firewall rules. See the **Network section** for required ports and connectivity.

Table of Contents

Documentation

Educate yourself on Starburst Enterprise Platform (SEP) using these handy links and guides before you get started with running it in your environment.

Starburst Documentation

- Starburst Enterprise Platform (SEP) Documentation
- Trino: The definitive Guide (The Starburst Trino O'Reilly book downloadable for free)
- Starburst Enterprise Security Guide

Installation

- Starburst Installation

Security and Operations

- SEP Security
- SEP Administration
- SEP Query Optimizer
- SEP Connectors

Deployment

We can start smaller, with the minimum configuration than grow based on the workload

Resource	Notes
Coordinator	1 Coordinator 16/32 x vCPUs, 64/128/256 GB RAM, ~ GBs of filesystem headroom
Workers	N number of Workers 16/32 x vCPUs, 64/128/256 GB RAM, ~ GBs of filesystem headroom
Ranger	If required. Native Built-in or File-based Access Control can also be used. Postgres can be used as an external backend.
Postgres, MySQL, or Oracle	Required for Insights (Query History, Cluster Resources, etc.)
Hive Metastore	Required for Object Store, Views, Materialized Views, etc. Glue or Dataproc can also be used. MySQL, Postgres, and SQL Server can be used as external databases.

Compute Requirements

Starburst Enterprise can be deployed on numerous systems with different methods. That includes bare metal, virtual machines, running from docker containers (for exploration only, and Kubernetes environments.

Stand Alone requirements:

Starburst will provide access to download either the Tarball or RPM packages.

The RPM installer is slightly easier to work with – but does require sudo access to root.

For a Tarball or RPM install – we recommend using the ansible based playbooks to install, configure, and manage the cluster:

https://docs.starburst.io/starburst-enterprise/starburst-admin/index.html

Note: For ansible to work, the ansible server (e.g., coordinator node) will need to be able to access the worker nodes.

Latest requirements

You can always find the latest requirements in the link below:

https://docs.starburst.io/latest/installation.html

Operating System

RHEL 7 or 8, CentOS 7 or 8 64-bit can use the RPM installer. A current version of python is required, i.e., 3.x.

For more information, please refer to:
https://docs.starburst.io/starburst-enterprise/starburst-admin/

JDK

Current versions of Starburst (390 and beyond) require java 17, so this will need to be installed on each node.

https://cdn.azul.com/zulu/bin/zulu17.36.13-ca-jre17.0.4-linux.x86_64.rpm

Hardware

Although a specific hardware vendor is not required, Starburst Enterprise supports the following architectures:

- x86_64 and AWS Graviton

Scaling requirements

Scaling is readily automated in a k8s environment. In a bare-metal or VM scenario – the scaling would need to be scripted.

Sizing

SEP Worker and Coordinator pods can require a significant amount of memory and CPU to work efficiently.

Ideally, a SEP Worker pod requires a minimum of 128GB of memory and 16 vCPUs which provide good overall performance.

Performance

As Starburst is a high-performance computing platform, it is highly recommended to place the Coordinator and Worker pods on dedicated hardware.

Consolidation of workloads can be explored during the Testing or if very larger servers have been allocated (> 256GB memory).

Internet Access

The RPMs will be required to be downloaded from the Internet, or via an intermediate layer.

Storage Requirements

Other than install files, log files, and OS – there is no hard requirement to store data on the Starburst Nodes. It would be advisable to have some storage headroom on each node, to avoid any filesystem full errors.

Reference

https://starburstdata.github.io/latest/installation.html

No	Title Description
IN01	**Starburst License** The license file for Starburst. Required for cluster to start with starburst features. A license key is required for all deployments – except for Marketplace PAYGO deployments – as the license component is included in the consumption. Starburst can provide license keys under agreement.
IN02	**Harbor Credentials** This is required to extract the helm charts and images from Starburst. Starburst harbor is provided: https://harbor.starburstdata.net/ Access is required by the default helm charts. Starburst can provide access under agreement.
IN03	**Bare Metal Images** To install information is provided below: https://docs.starburst.io/starburst-enterprise/try/rpm.html https://www.starburst.io/platform/starburst-enterprise/download/ The images can be requested online above, or the direct link provided by Starburst.

No	Title Description
K01	**K8S Cluster / VM** Recommend node size of - 16 vCPUs minimum / 32vCPU recommended - 64-128GB RAM / 128-256 RAM - Network 10 Gb minimum - Disks minimum possible to it operating system - Disks extra required to provide hive caching. If no pre-sizing has been done, the following are good starting points - 2 x worker nodes for functional testing - 4 or 8 worker nodes for capacity testing

	The setup will include secrets, for simplicity please ensure this is not disabled.
K02	**Load Balancer / Ingress Point** It is recommended to have a load-balancer, or an ingress point in Production. This is dependent on your environment. For a PoV this can be optional. In the first instance there is limited security. The setup of security for simplicity we recommend terminating TLS here.
C01	**Database (Query History)** Starburst will log data to a database. The supported databases are listed below: https://docs.starburst.io/latest/admin/query-logger.html Postgres or MySQL are the most popular, we suggest you use whichever you are more familiar with. The instance is dependent on query concurrency it should be - Between 2 to 8 CPU - Between 16 to 64GB The database should be setup and a username and password to have full access for the database, should be available.
C02	**Expose ports** Starburst by default is exposed on port - 8080 without TLS. - 8443 with TLS It is recommended that at least one port is exposed to access Starburst.

C03	**Data Sources**
	Network access and credentials are required for access to data sources. Please consider the following data source types:
	
	Object Storage
	• Have available one of the following
	  ○ ACCESS KEYS and SECRET KEYS
	  ○ Roles to provide to Starburst configuration
	  ○ Permit machines / pods access to object storage
	  ○ Kerberos keytab available (HDFS)
	• Ensure cluster has access to any encryption
	• Ensure in the same region to minimize egress costs

	RDBMS, for initial setup use a service principal user and password. If authentication is different, please advise before the installation. • Have available URL • Ensure Cluster has network access to URL • Have available username and password
C04	**Data Clients** We recommend starting by using the Starburst GUI. Any additional clients must be able to see the load balancer or ingress port on the exposed port.

Metastore

No	Title Description
M01	**Choice** There are Various options for a Metastore. Please advise if Starburst Hive Metastore is not going to be employed.
M02	**Starburst Hive Metastore** If Starburst Hive Metastore is required, the helm chart/docker build will require to deploy a pod. The evaluation node recommended is - 2 CPU - 8GB
M03	**Starburst Hive Metastore External DB** If deploying Starburst Hive Metastore to an External DB. Create a database in the query logger database instance. Have available Username, password, database name

Access Control

No	Title Description
AC01	**Choice** Various options for Access Control: - File Based - Ranger - Starburst Built-In Access Control

	- 3rd Party (Immuta, Privacera) For evaluation it is suggested to use a Starburst Access Control. There are edge cases where File Based Access Control can be used. Only use Ranger if this has been decided as the method of deployment
AC02	**Starburst Ranger** If Starburst Ranger is required, the helm chart/docker file will require to deploy a pod. The evaluation node size recommended is - 2 CPU - 8GB
A03	**Starburst Ranger DB** If deploying Starburst Ranger. Create a database in the query logger database instance. Have available Username, password, and database name.

Security Features

Note: Is Production Level Security required for the PoV?

No	Title Description
ID01	**IdP integration to Starburst –** For the purposes of a PoV – using File-based Authentication is an option. This avoids integration issues and is suited if no sensitive data is in use. 2 weeks should be allowed for credentials and firewalls to be opened. Confirmation of IdP, please confirm if TLS and certificates are required.
ID02	**IdP Integration for Access Control** 2 weeks should be allowed for credentials and firewalls to be opened
ID03	**Data Sources** The default is to use a security principal for testing underlying data sources. If a different method is required. This should be stipulated to ensure that the setup is reflected in the configuration.

Features

No	Title Description
F01	**Hive Caching** Only advised for bare metal install
F02	**Materialized Views, Table Re-Direction, Cache Service** Advised for testing enhanced performance techniques
F03	**Smart Indexing and Caching** Only available for preview in AWS
F04	**Data Products** Requires users to test
F05	**Stargate** Requires multiple install
F06	**Workload Isolation** Basic install for performance tests
F07	**Dynamic filter** Enabled by default. Do not disable
F08	**Parquet Accelerated Reader** Enabled by default. Do not Disable
F09	**Pushdown** Depends on Connector. Sometimes default, sometimes more forceful (setting to EAGER).
F10	**Fault Tolerant Execution** Only required for batch workloads. Do not use for a user query deployment.
F11	**Auto Scaling** Can be configured with k8s.

Networking Requirements

Network speeds

For testing purposes, 1GbE would be sufficient, but 10GbE for hosts is recommended to reduce network latency. Check with your network administrator to identify potential bottlenecks or limitations before proceeding with a deployment.

Network Proxies

Are Network Proxies used between the Starburst Cluster and the Data Sources?

Open Ports and Connectivity

Open and test the connections between the following in your lab. Note that some of these components are optional.

Component	Source	Destination	Default Port
Tools connection	Your Tools	Coordinator	JDBC (8080, 443)
SEP UI	Your browser	Coordinator	8080, 8443, 443 (TLS ports)
Cluster Communications	Cluster Nodes	Cluster Nodes	8080
Data Sources Connections	Cluster Nodes	Data Sources	Various
MySQL	MySQL Service	Coordinator	3306
PostgreSQL	Postgres Database	Coordinator	5432

Data Source Access

Storage Access

For example – if S3 is the Object Store – IAM credentials, or S3 secret keys can be used to provide access. IAM credentials are preferred.

Data Source Requirements

AWS S3

Standard Hive Connector. Important information on getting the correct IAM Roles. Keys and Secrets can also be used:
https://docs.starburst.io/latest/connector/hive-s3.html

Security Requirements

Identity and Authentication:

Acme use Okta for SSO and Authentication.

For the purposes of the Test Drive – it would deliver faster time to value by implementing a basic local user database with File based access control, with Built-in Access Control (native capabilities in Starburst).

This involves a couple steps to implement.

Okta is supported and could be implemented with Starburst and could be deployed in a subsequent phase.

Starburst and Okta Support:

https://docs.starburst.io/latest/security/okta-authentication.html

TLS

TLS expected from Client/Tools to the Cluster.
TLS Terminated at the Coordinator.
Enterprise Certificates need to be made available for TLS.

LDAP/AD

Integrating LDAP/AD with Starburst is a common task.
It does not add a lot of value in terms of Data Discovery and becoming familiar with the Technology.

Testing

Resources & Tools

To execute testing against Starburst, access to one or all the following will be required:

- Starburst UI (provided by default)
- BI Tools
- SQL workbench tool: DBeaver, etc.

For any benchmarking – JMeter is the most used tool.
This will require a host with Java and JMeter installed as well as connectivity to the cluster.

Your Teams

To make sure your lab gets set up properly, you might need to schedule your teammates to help install and configure SEP and connect it to your own infrastructure and resources and let them know they may need to be available to answer questions in case their expertise is needed.

Role	Resource
Starburst Account Team	To assist with understanding Starburst – from the Commercial and Technology viewpoints.
Infrastructure Team	Your resource to help you get your hardware or VMs to install in AWS.
Cloud Team	Your cloud teammates to get the required user rights and cloud resources so you can install SEP in the cloud.
DBAs	Your DBAs give you access to the data sources so you can connect them to SEP. You'll need logins/passwords and will possibly need to grant these users appropriate rights.
Active Directory/LDAP Admins	If you are connecting your lab to an LDAP backend, you may need your AD/LDAP administrator to help with the connection info.
Security Team	If you are using your own CA or are doing security testing, you may need to schedule time with your security team.
BI team/Data Analysts	Your BI and Data Analysts who want to test SEP and run through the test plan.
Network Team	Your local or cloud network team might be needed to **open up ports,** so the cluster has the proper communications. (See the network section in this document.)

Operations Team	Your operations team might be needed if they are required to do operational tests such as monitoring, HA, backup, etc.

Sample Starburst Deployment Architecture

A sample deployment is based on Kubernetes (k8s) is provided below:

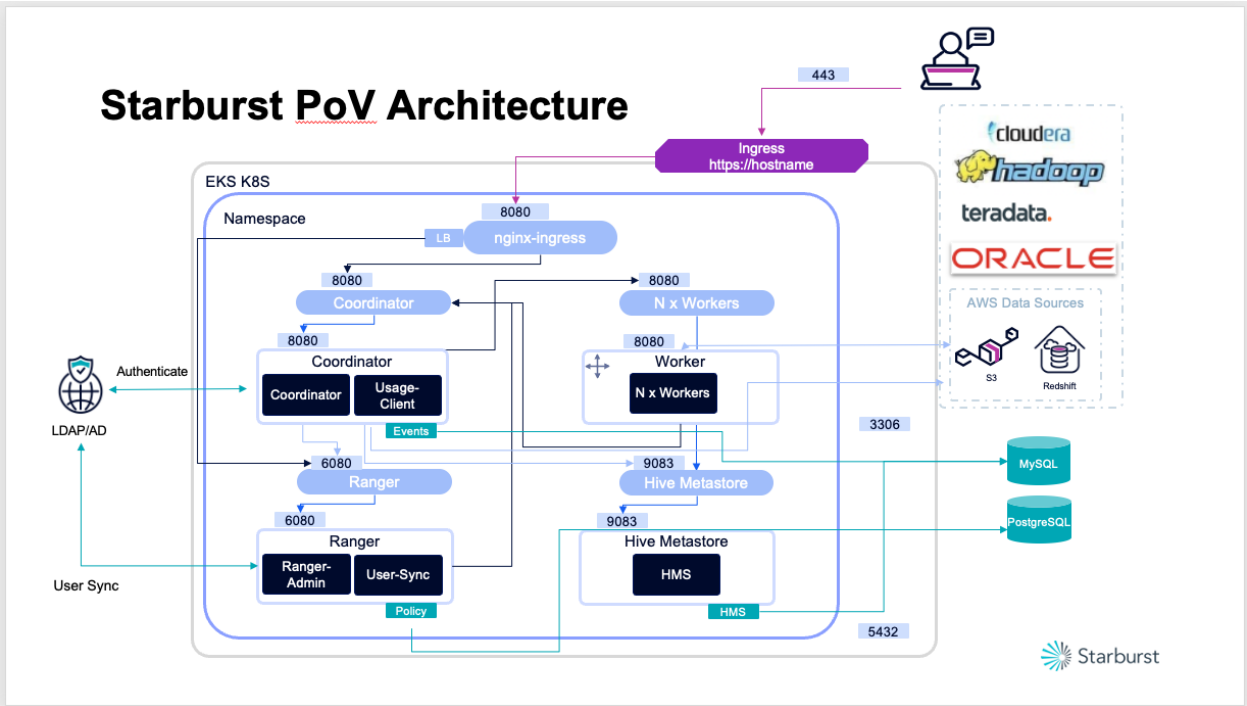

Figure 1 – Starburst Deployment

Starburst would be deployed into an k8s cluster with a 1:1 correlation of k8s node to pod to container. Starburst's Coordinator and Workers would each be a single container deployed into a single pod. With auto-scaling the number of k8s nodes can scale up and down, based upon cluster usage, given a defined range. Both the Coordinator and Worker NodeGroups can be deployed into a Starburst-specific namespace in an existing cluster, if required.

The Starburst Coordinator would connect to each Worker and each Worker would connect to each data source.

In the figure above, the Hive Metastore is stored in an external MySQL Database, and the RDS Postgres is acting as the cluster event logging database (logs all the Queries run in the cluster).
Note: The Hive Metastore can also be persisted in an Internal PostgreSQL database, located on a Container Image in the Cluster. This can be used for testing, but it is recommended to use a more resilient service.

Additional Details

See below for deeper details on the above pattern.

Security

LDAP and Active Directory are being used as the Identity Provider. Apache Ranger is being used to Control Access Policies. Ranger has an external Database to store the Policies.
Note: Ranger can also use an Internal PostgreSQL Database, running on a Container Image in the Cluster. This can be used for testing, but it is recommended to use a more resilient service.

External Databases

In the example - MySQL – Hive and Events, and Postgres – Ranger, are used. One could be used instead.
For most POV deployments a small instance size of at least 2 CPUs / 16 GB memory should suffice.
For a PoV, the Hive, Events, and Ranger databases are under a lot less stress than a large-scale production deployment.