Enterprise

Security Guide

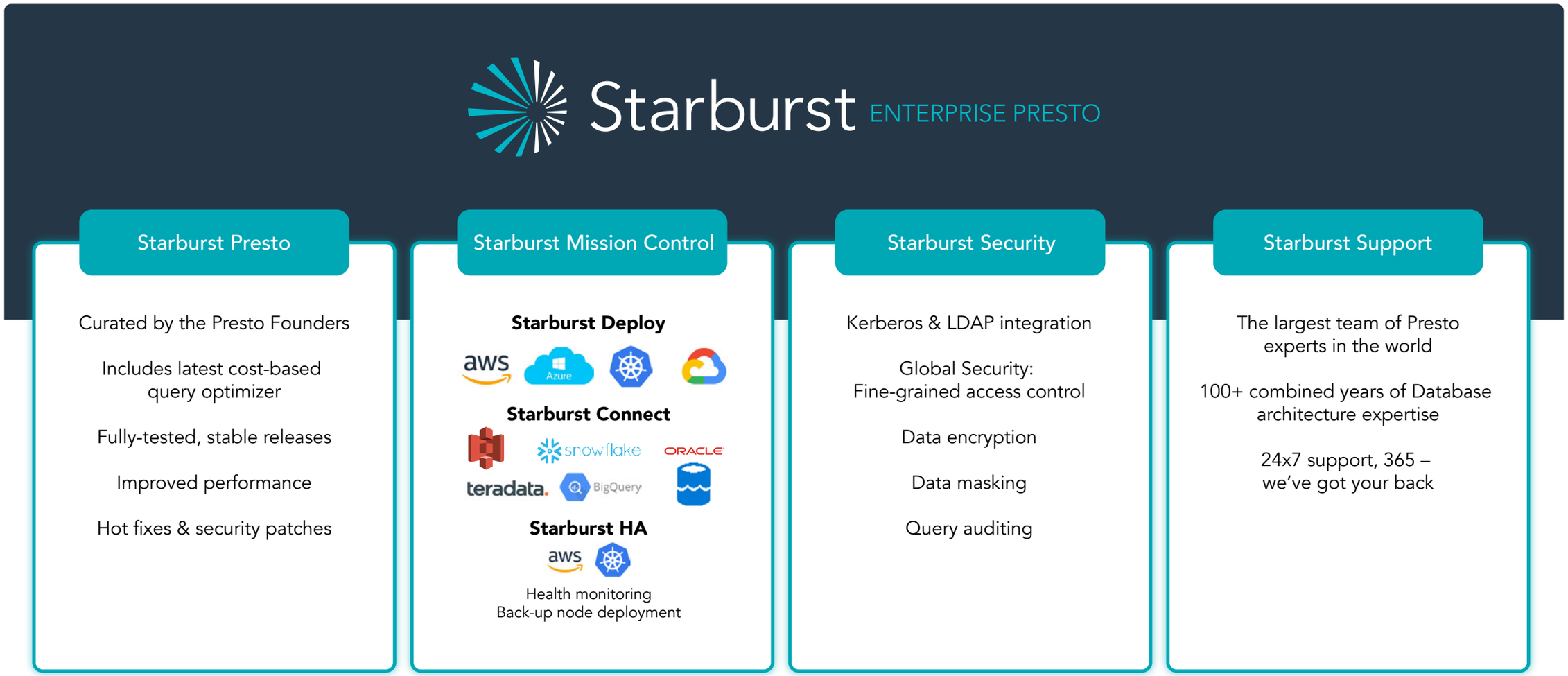
Starburst
ANALYTICS ANYWHERE

TABLE OF CONTENTS

INTRODUCTION

Starburst Enterprise Presto is a distributed SQL query engine that can be deployed in any infrastructure. Built on the open source Presto project developed at Facebook, Starburst Enterprise Presto is used by some of the largest, well known companies in the world such as Slack, Comcast, Zalando and FINRA.

Starburst Enterprise Presto is a fully supported, production-tested and enterprise-grade distribution of the open source Presto SQL query engine. It includes additional connectors for commercial database systems, query optimization, as well as management tools. One of the core reasons organizations select Starburst is the added security features that we've built into the Presto SQL engine. These include fine-grained access control, data masking and encryption, column and row-level security, query auditing as a few examples. This guide provides details on all of the enterprise-grade security features in Starburst Enterprise Presto.



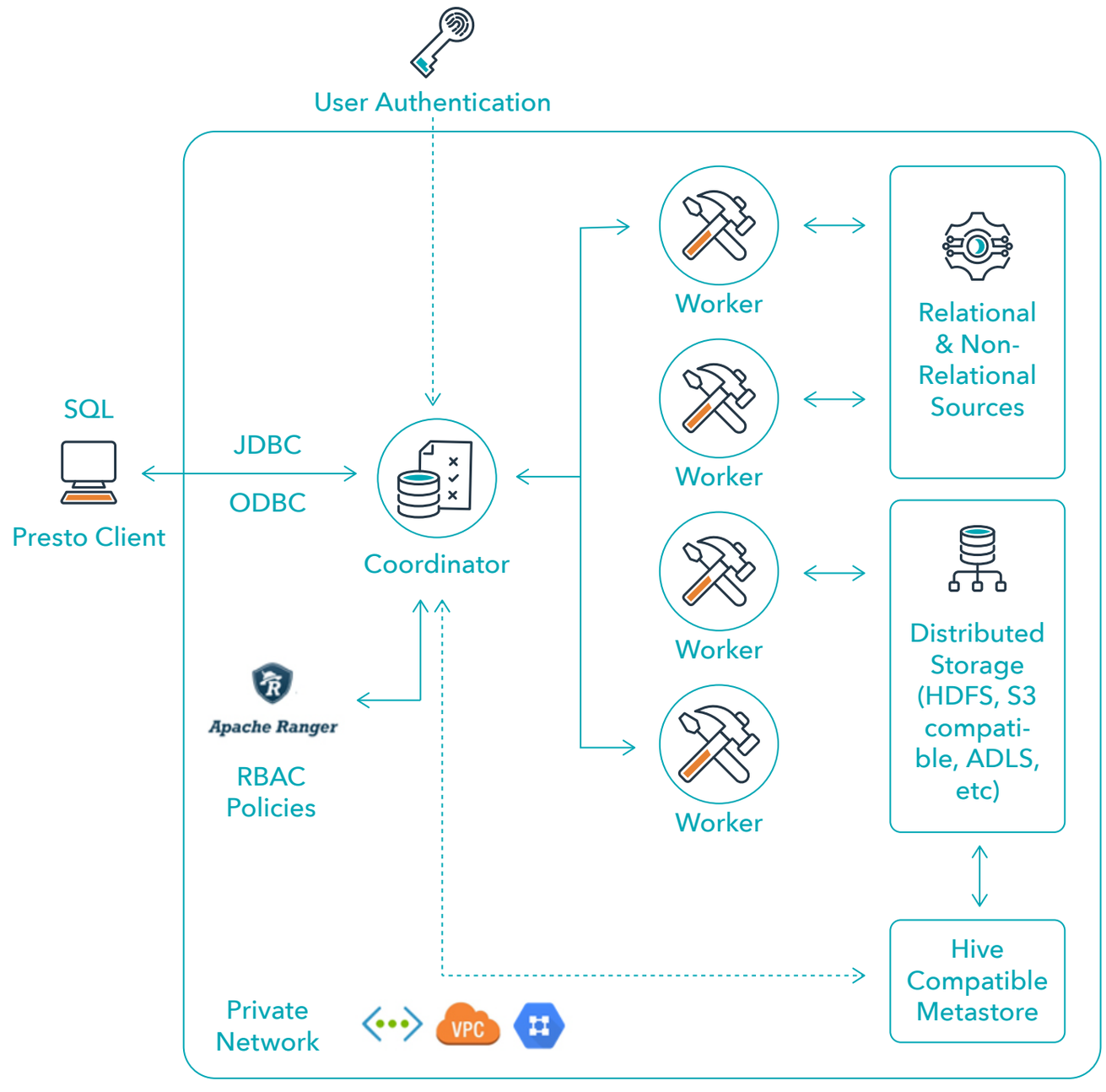
This guide provides details on the enterprise security features provided by Starburst Enterprise Presto along with extensive event logging and robust fine-grained access control.

Starburst Enterprise Presto Architecture

The lightweight, standalone architecture of Starburst Enterprise Presto makes it simple to install, secure, maintain and scale. Since there is no storage of data and it can be installed in any location including cloud or on-premises, security is simple to maintain and enforce.

Starburst Enterprise Presto's architecture consists of a coordinator and worker processes, which are configured with connectors. Each component can be secured using industry standard techniques as is best practice when deployed in a production environment. The diagram below illustrates the different components of a Starburst Enterprise Presto cluster.

1. The coordinator is the brain behind a Presto cluster. It's responsible for:
   - Accepting client connections to execute queries.
   - Parsing, analyzing, planning, and optimizing query plans. This uses Presto connectors to retrieve metadata about tables such as columns types, and data statistics.
   - Scheduling query data retrieval tasks on workers nodes.
   - Returning the query results to the client.
2. The workers are responsible for the heavy lifting. Their job is to retrieve data using the Presto connectors, filtering, joining, aggregating, and exchanging the intermediate data before streaming back the final result to the client via the coordinator.
3. Connectors translate the data source objects into something Presto can operate on when executing standard SQL. The categories for the connectors fall into the following:
   - Distributed Storage (HDFS, Amazon S3, Azure Blob and Data Lake Storage, Google Cloud Storage, and S3-compatible Object Storage) using Hive compatible metadata (Hive Metastore, Amazon Glue Data Catalog)
   - Relational Database Management Systems (Oracle, PostgreSQL, MySQL, etc.)
   - Key-Value Stores (Cassandra, Accumulo, Redis, etc.)
   - Document Stores (MongoDB, Elasticsearch)
   - Streaming Systems (Kafka, Kinesis)
   - Built-in Utility (System, Memory, TPCH/DS, etc.)
4. Lastly, fine-grained access control policies are enforced during query time using Starburst Global Security. This includes column and row level authorization and data masking

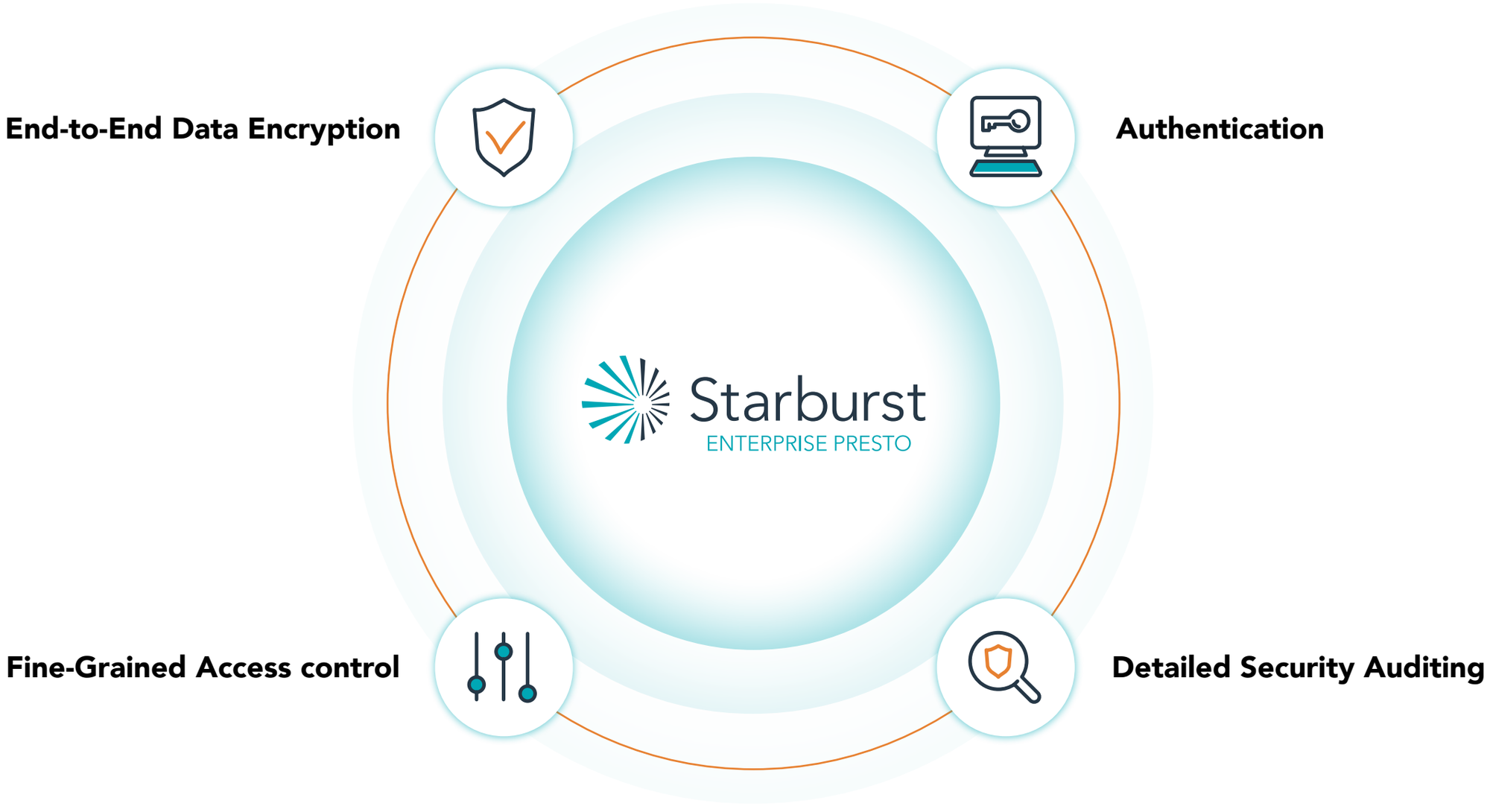


Achieving Enterprise-Grade Security with Starburst

We know very well that data security is critical and that sensitive information can be highly destructive and costly to an organization when it falls into the wrong hands. Starburst Enterprise Presto contains many security features that enterprise companies expect, but are not available in open source Presto.

Ensuring data is encrypted from the sources to the end user is now a standard in enterprise environments. Controlling access down to the column and row level usually is the function of third party software but is included in Starburst Enterprise Presto and is constantly being improved. Providing a full data access audit trail is also essential to ensure companies comply with state and federal compliance regulations.

The following sections detail out how Starburst Enterprise Presto implements the following enterprise features in order to secure Presto:



End-to-End Encryption and Authentication

The leading reason Starburst clients choose Starburst Enterprise Presto over open source Presto, is it comes with comprehensive security features and configurations. This document highlights the key elements of end-to-end security.

Encrypting User Access

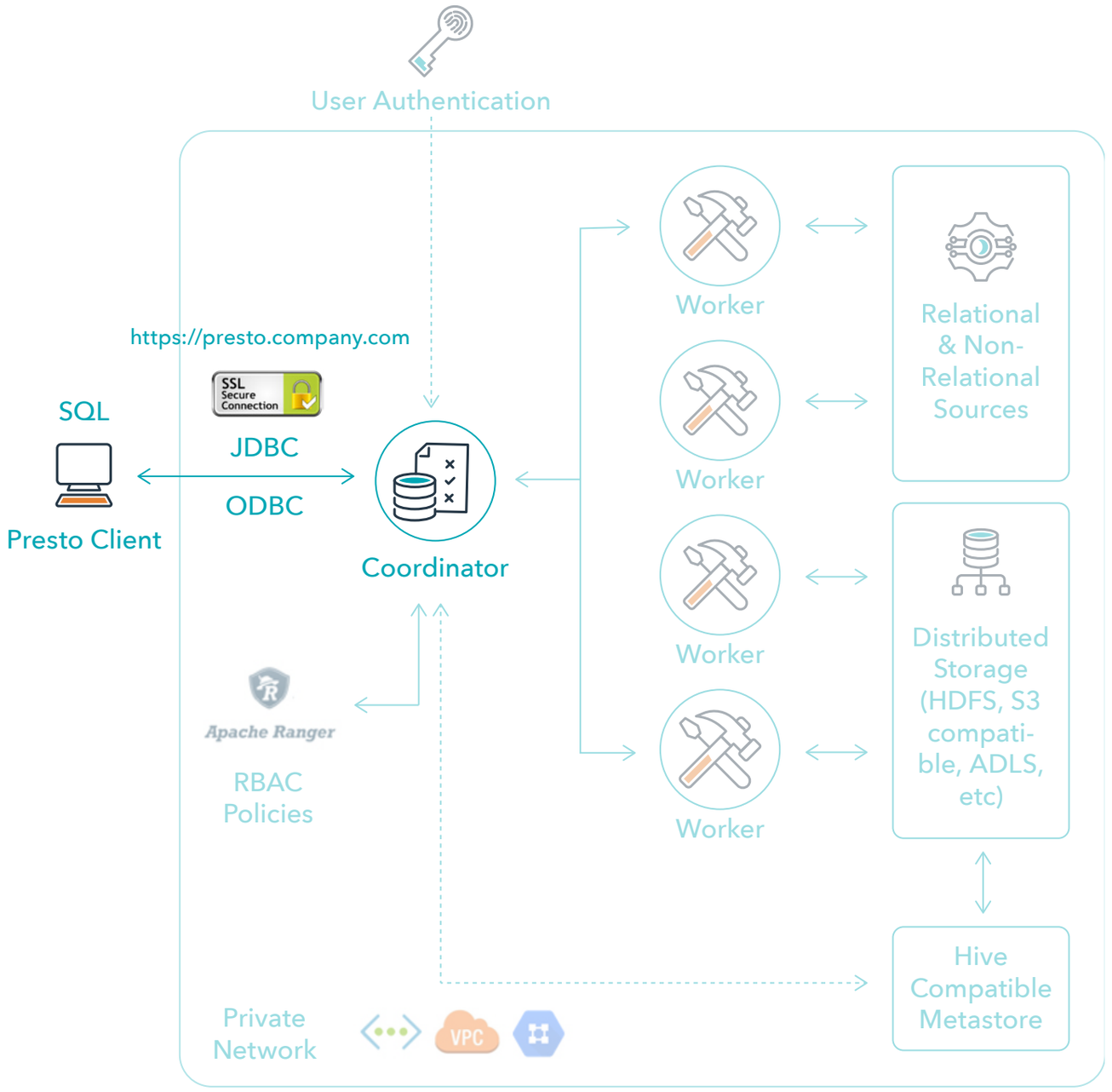
When users connect to Presto to issue queries, they are connecting to the coordinator node.

Starburst Enterprise Presto supports both trusted CA and self-signed certificates. When using a self-signed certificate, clients must have a password-protected truststore file containing the coordinator's certificate. Most ODBC/JDBC clients that will connect to Presto will not require the Truststore file when using a trusted certificate.

This will encrypt traffic from the end users to the Coordinator node. Traffic from the coordinator to the worker nodes will still remain unencrypted.

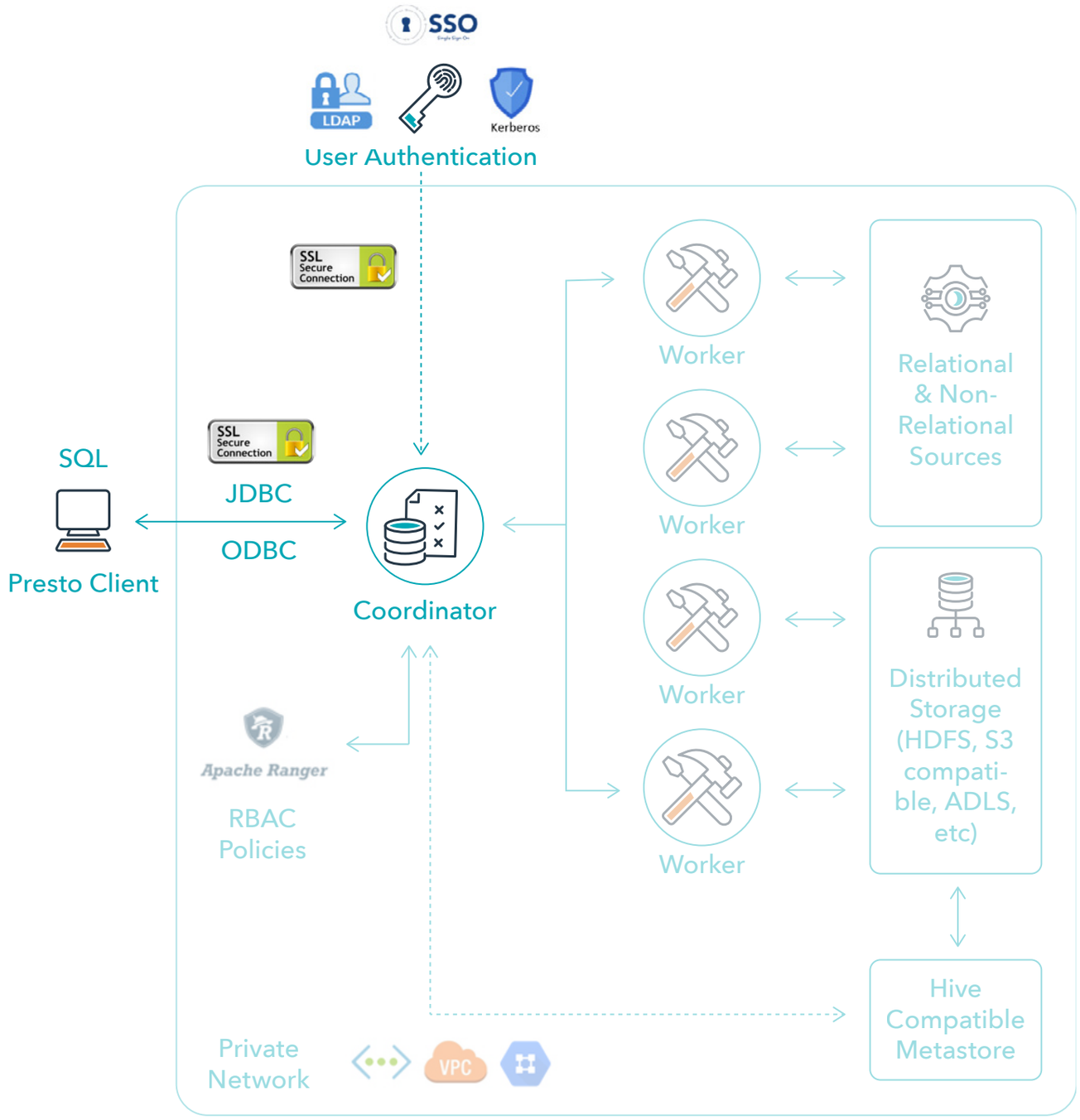
Configuring HTTPS between the Coordinator node and the workers is similar to securing the Coordinator. Workers would be configured to communicate with coordinator using standard HTTPS methods.

This is covered in section titled: Secure Internal Communication further down in this document.



Authenticating Users

Authenticating users to access Presto can be handled using a few different methods. Currently, Presto supports authenticating against external sources through LDAP, Kerberos or SSO (single sign-on) using HTTPS as illustrated in the diagram below:



LDAP

Authentication through the industry standard LDAP protocol supports many different providers such as Active Directory and OpenLDAP. Presto supports authenticating users using Secure LDAP (LDAPS) which requires the external LDAP server to be configured with TLS. (an industry standard practice)

NOTE: Enabling LDAP authentication requires HTTPS to be configured on the Coordinator node as described earlier in this document. Once the LDAP server's TLS certificate is imported into Presto's Truststore, users will be authenticated to login to the coordinator.

Kerberos

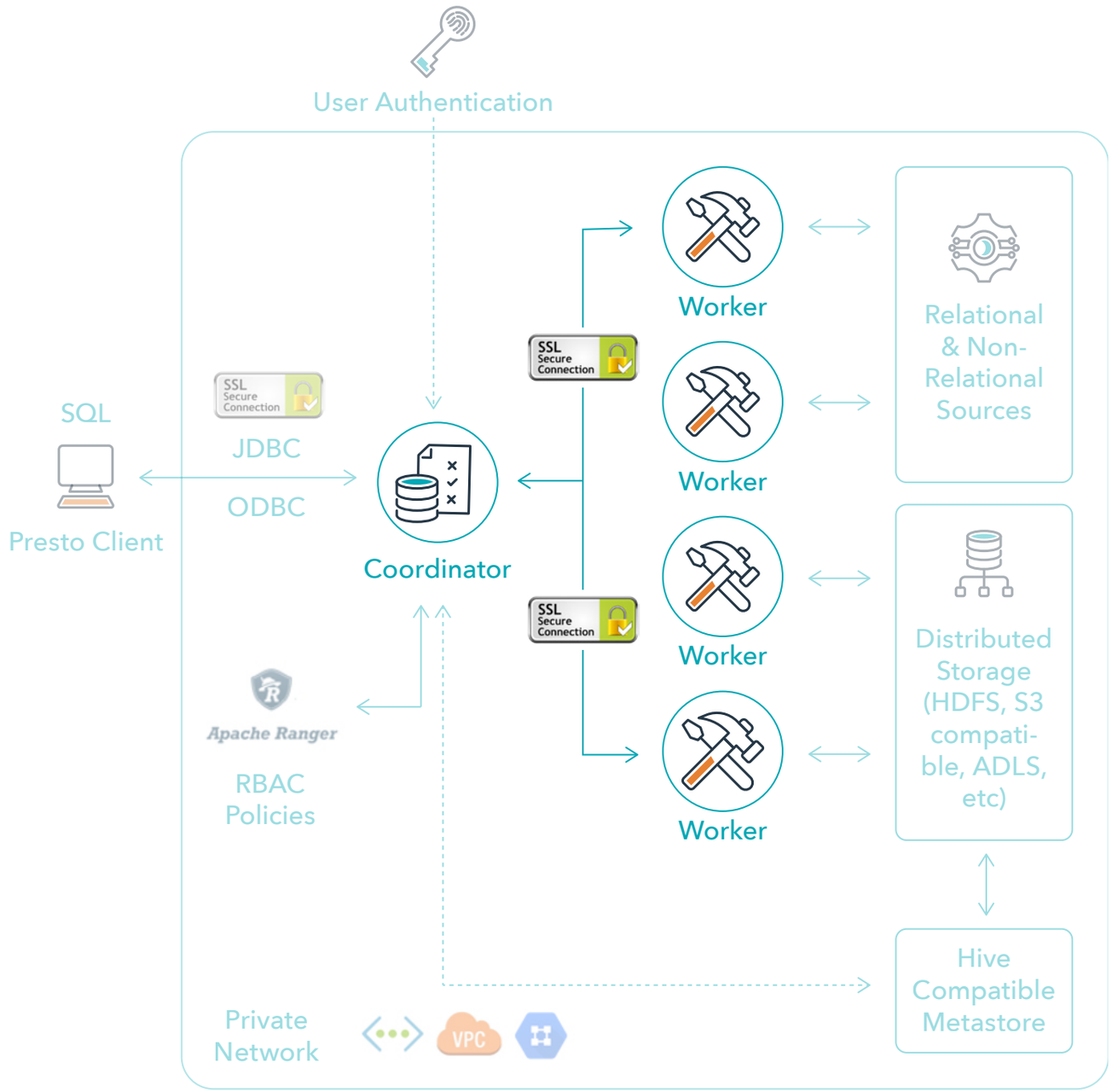
Alternatively, user authentication can be achieved via Kerberos. A Kerberos KDC service needs to be accessible by the Presto coordinator server. HTTPS must be enabled on the coordinator in order to secure user authentication requests. More detailed information can be found in the Kerberos documentation.

Identity Provider

Starburst Enterprise Presto supports Identity Provider (IdP) through Okta as of this writing. More providers will be added in coming releases. Identity Provider enables end-users to authenticate using a provider such as Okta or Ping.

Securing Internal Communication

Full end-to-end encryption of internal traffic between Presto nodes is possible for highly secure environments. Securing internal communication ensures that intermediate data exchanged between Workers and Coordinator during query processing is encrypted in transit.



Securing Connectors

Secure access to your data source depends on the particular connector and the source system capabilities. Often, several different methods are supported to fit various customer setups.

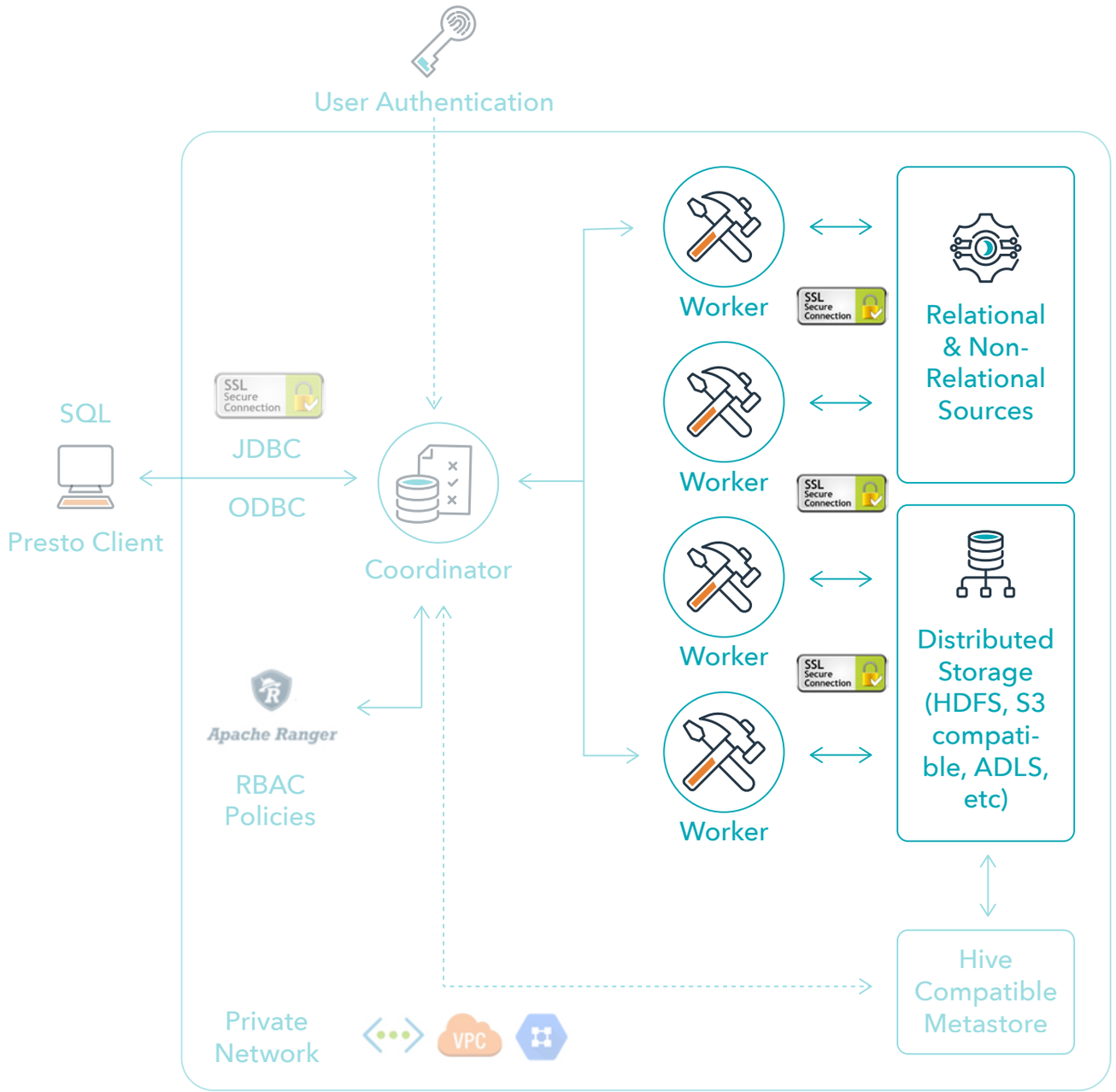
In most cases, the connector configuration file contains some form of security credentials allowing Presto to access the data source. Often the access to data is driven via so-called service accounts that channel all Presto end-users data access requests. We recommend the service account read-only access to objects in the source system.

End-user Impersonation

Starburst Enterprise Presto includes a feature in some of the connectors called end-user impersonation. Typically, a service account is used to access the data source. For auditing purposes this may not be desirable as auditing shows a single user accessing the data. However, end-user impersonation will run the queries as the user who executed the query. This means the privileges for that user are applied and also audited appropriately.

Securing Sensitive Data

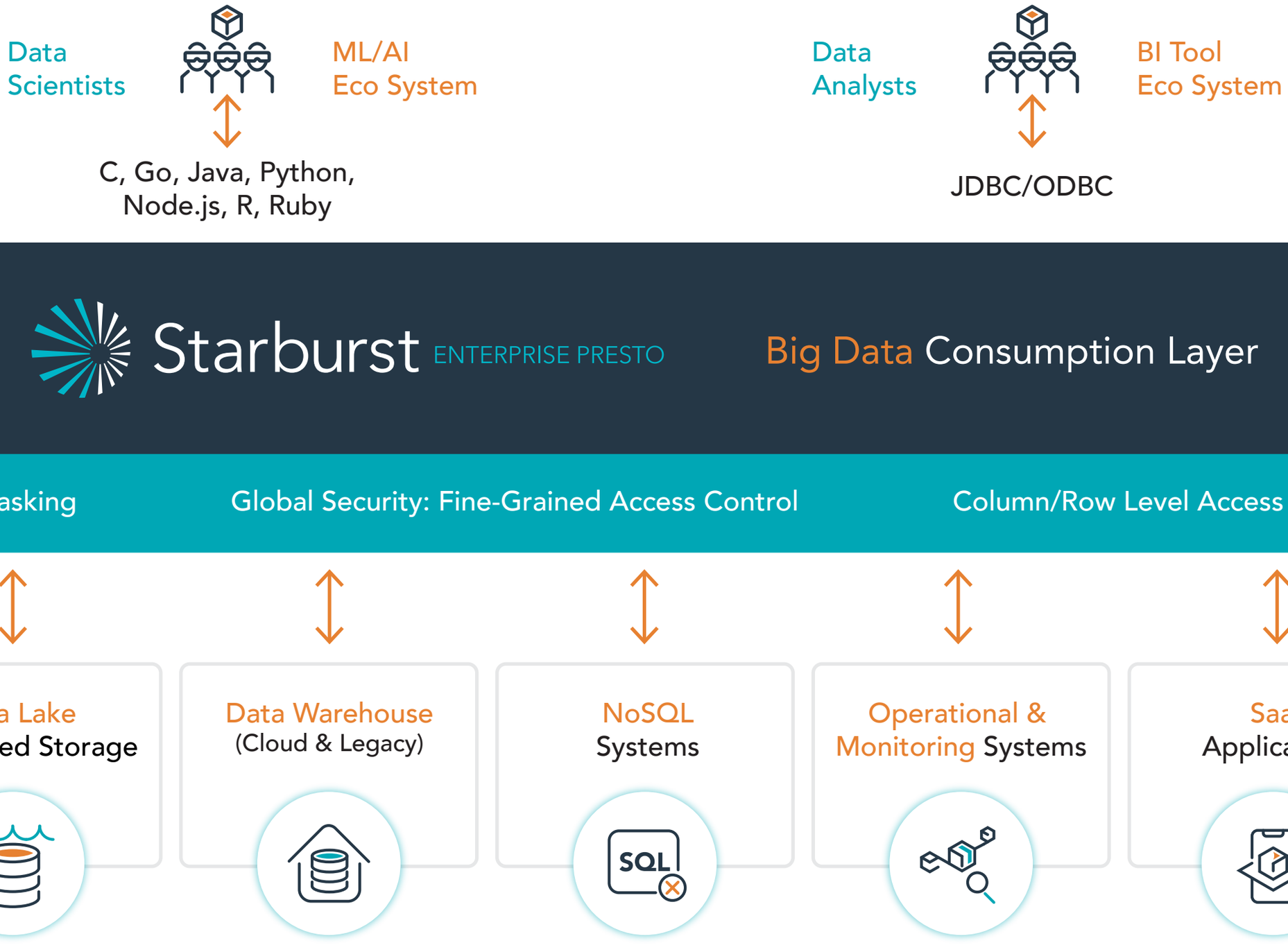
Presto's configuration files may include sensitive information, such as plaintext passwords and usernames needed by the service account to authenticate to the data sources.. In many enterprise organizations, this violates security policies. Starburst Enterprise Presto provides the ability to securely authenticate to these data sources without having to store the plaintext passwords and usernames in the configuration files.



Global Security: Fine-Grained Access Control

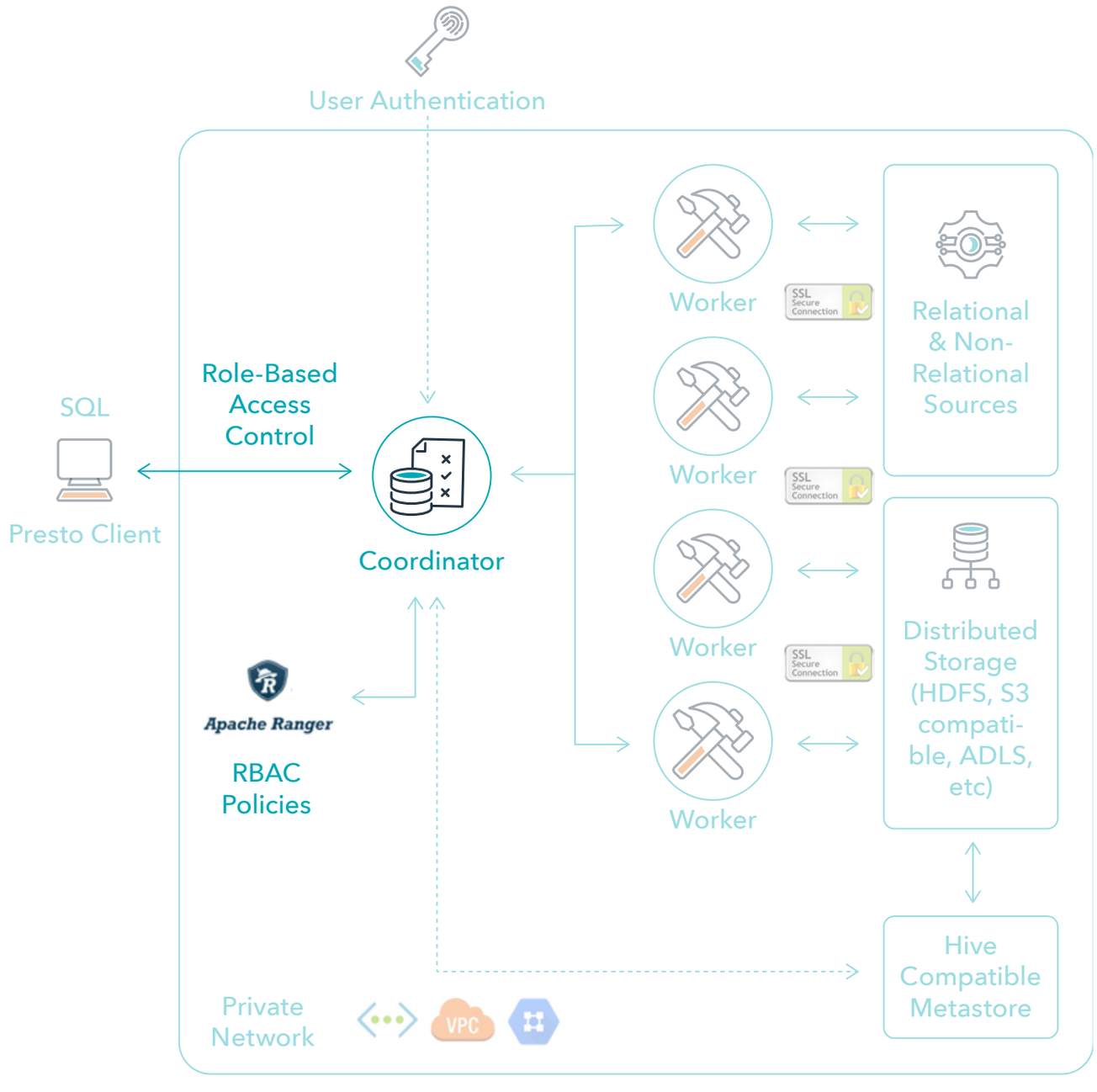
Fine-Grained access control allows column and row level control against data sources. With the amount of data in a data lake and other source systems, limiting certain users and groups to sensitive data is a requirement for many organizations.

Starburst Enterprise Presto provides Global Security over all data source connectors. This allows fine-grained access control including column and row level policy enforcement as well as column level data masking.



Global Security

Global Security integrates with open source Apache Ranger to provide a greater level of control access control as well as offering column level data masking. Groups and users can be synchronized to a central LDAP repository and policies can be set to only allow approved users and groups access to data within the source.



Groups and users can be synchronized to a central LDAP repository and policies can be set to only allow approved users and groups access to data within the source.

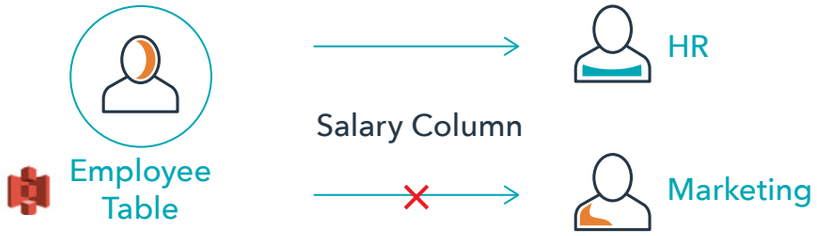
Catalog, Schema and Table Level Control

When providing access to many different data sources, it's crucial to an organization to be able to control access to this data. This becomes even more of a challenge when it comes to data residing in a data lake. These are often file/object based and existing policy enforcement is limited to the folder or bucket level.

Starburst Enterprise Presto provides the ability to limit groups and users by the catalog, schema and table. The diagram below illustrates a common scenario where one group of users do not have access to certain tables.

Column Level Control

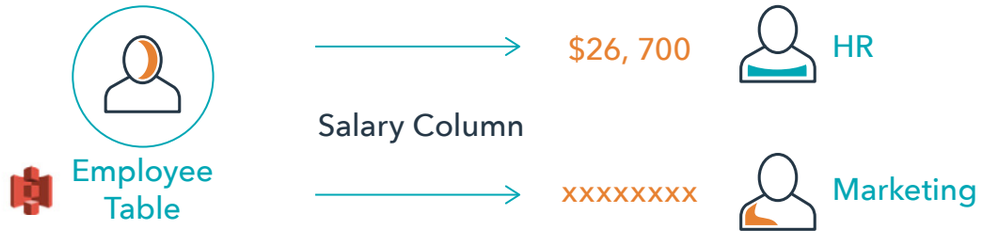
Tables in data lakes are often denormalized for performance and include many columns. In the case of many data lakes, this could be hundreds. Providing access to a large group of users can be problematic if there is no functionality to limit access at the column level. Starburst Enterprise Presto provides column level policy enforcement at the group or user level. If the user submitting the query doesn't have access to a column or set of columns in the table, they will receive a message that based on an existing policy, they do not have access.



Data Masking

Column data can additionally be masked using Starburst Enterprise Presto. Based on a user or group, columns of any tables can be masked using popular masking techniques such as redaction, partial masking and blanking out sensitive data.

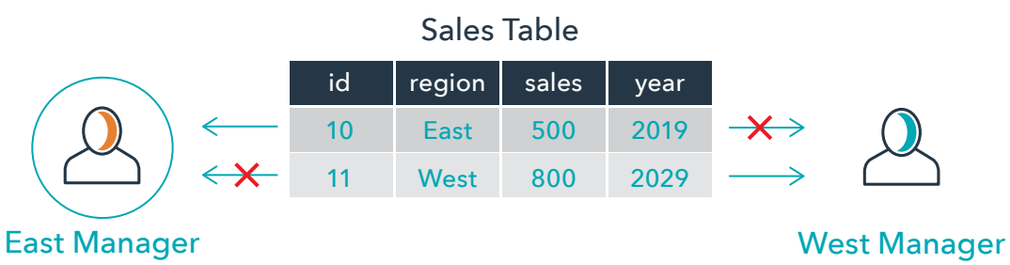
In the diagram below, when the HR and Marketing user query the Employee table, their results will vary based on their privileges. The HR user will be able to access the salary column but the Marketing user's results will be masked.



Row-Level Security

In a data lake or relational data source, there may be sensitive data that encomposes the entire row. Limiting access to this data for certain groups or users is a challenge in most systems. Starburst Enterprise Presto provides the ability to limit rows of data within a table by enforcing row-level based policies.

In the diagram below, the East Manager would not be able to view rows in the sales table with a region = "West". Likewise, the West Manager wouldn't be able view rows that belong to the East region.



Detailed Security Auditing

Logging information about queries is a requirement for many organizations. Starburst Enterprise Presto includes two different audit logging capabilities.

Event Logging

Starburst Enterprise Presto offers the ability to log each query in great detail to a remote database. From here, this data can be queried with Starburst Enterprise Presto or pulled into an existing event logging system.

Example fields that are available:

Column Name	Description
`query_id`	Randomly generated id of the query
`execution_time`	How long the query took to complete
`user`	The user that executed the query
`query`	The text of the query
`total_rows`	How many rows the query produced
`written_rows`	How many rows the query wrote to the target
`cpu_time`	Total cpu consumed on the cluster
`client_info`	Detailed information about the client. (JDBC,etc..)
`query_plan`	Plan the cost based optimized produced

Compliance

The ability to provide real-time query logging has become the standard in the database industry for years.

As data volume constantly increases, companies are under more and more pressure to monitor data access within their organization. With Starburst Enterprise Presto's event logging functionality, a full, GDPR level audit trail is available in real-time. This allows tracking access to all data sources that are a result of queries submitted to Starburst Enterprise Presto.

Chargeback

Starburst Enterprise Presto is used by many different departments and user groups. It can be difficult for a centralized IT organization to determine the resource usage of these different users. With event logging, each query is logged into a database. In addition to the user that executed the query, the text of the query and the elapsed time, other metrics such as RAM and CPU are available which can provide a more granular level of detail of actual usage.

Performance Tuning

The data collected for each query includes resource utilization. This data enables resource usage per query and can be used to determine queries, users and data sources where performance tuning might be considered. Reporting can easily be created to monitor Starburst Enterprise Presto usage based on users, connectors and tables.

Audit Logging

File based audit logging can be enabled within Starburst Enterprise Presto. This adds query execution information into a file on the coordinator which defaults to /var/log/presto/security.log. The fields collected into this file are: timestamp, initiating query user, query id and the sql statement.

This feature has been part of Starburst Enterprise Presto and is an alternative to a database logging approach detailed above.

SUMMARY

Presto is adopted by some of the world's most innovative companies such as Airbnb, Uber, Twitter, LinkedIn and many more. As the awareness and adoption of Presto has grown, Starburst has worked to deliver enterprise-grade features and support to ensure organizations are successful with Presto.

A critical part of ensuring Presto is successful at scale is providing a centralized security framework that meets enterprise standards, and Starburst Enterprise Presto is designed for that. Controlling access to data wherever it lies within a data-intensive organization conforms to strict new data governance and security regulations.

STARBURSTDATA.COM

Copyright © 2020 Starburst