

Secure Business Intelligence in the Age of Hadoop

How AtScale Delivers Secure Self-Service BI

'HOW IT'S DONE' BY ATSCALE: The purpose of our 'How it's Done' series is to share a deeper look into AtScale's unique and comprehensive approach to BI on Hadoop. Should you have further questions, please let us know. We look forward to helping you make your BI work on Hadoop, securely, with all the scope and depth required by today's modern and big data world.

SECURITY

As more and more enterprises adopt Hadoop and the Hadoop Distributed File System (HDFS) in pursuit of a Data Lake or Enterprise Data Hub strategy; data security has become increasingly important. In this article we share how AtScale enables secure, self-service access to Hadoop data for Business Intelligence users.

Key Components of Data Security

For the enterprises the world over, in every industry, including Healthcare, Financial Services, Telecommunications, and Retail, data security is of paramount importance. IT departments are faced with a difficult challenge. They need to ensure that business users have access to the data they need to make decisions and do their job, while at the same time enforcing the appropriate controls to prevent unauthorized users from viewing or modifying sensitive data or business models. At a high level, the security requirements that enterprises require include but may not be limited to the following:

- **Enterprise Directory Integration:** Users (both design-time and query-time users) must be able to be sourced from and authenticated against an enterprise directory service, such as LDAP or Active Directory.
- **Role Based Authorization:** Users must be able to be assigned to groups. Additionally users and/or groups must be able to be granted specific roles (such as a Cube Designer, a Query-Only User, or an Organization Administrator).
- **Object Level Security:** In a multi-domain and multi-tenant environment it is critical to be able to control which users and groups have access to specific 'objects' – in the case of AtScale, 'objects' are represented by Virtual Cubes.
- **Secure Cluster Connections:** As data access controls are increasingly managed at the Hadoop cluster level using HDFS permissions, Apache Sentry, or Apache Ranger enterprise architects are paying close attention to the mechanisms and approaches that are used by applications that consume data from the cluster. Supporting the appropriate security for these connections is imperative.
- **Delegation and Impersonation:** It is important that users who query AtScale cubes should only be able to query data elements (both raw and aggregated) that they are authorized to see. Additionally, for audit purposes it is required that end-user queries are logged when these users access Hadoop data.

- **Row Level Access Controls:** Row-Level Security must enable customers to control access to rows in at AtScale Cube based on the characteristics of the user executing a query (e.g., group membership or execution context).
- **Secure Data Transport Protocols:** Our customers want to ensure that AtScale provides privacy and data integrity between all components of AtScale as well as the data sources and clients with which we communicate.

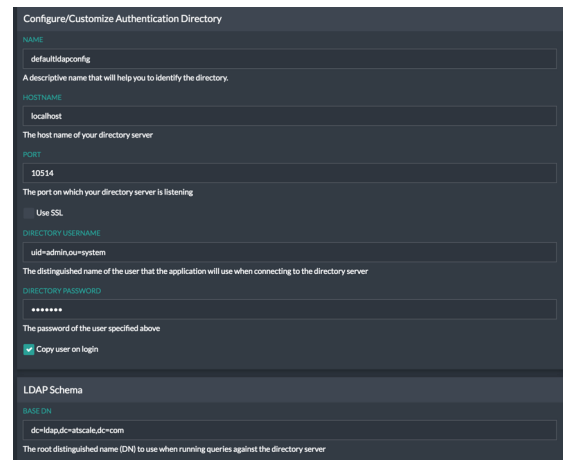
The AtScale approach to security takes these requirements into consideration, and as such delivers the security controls that are described in more detail in the following sections.

Enterprise Directory Integration

While AtScale ships with its own embedded [Apache Directory Server](#), customers have the option to choose to use their existing LDAP or Active Directory service for user and group management and authentication. In addition, AtScale provides a rich set of additional features and functionality, including:

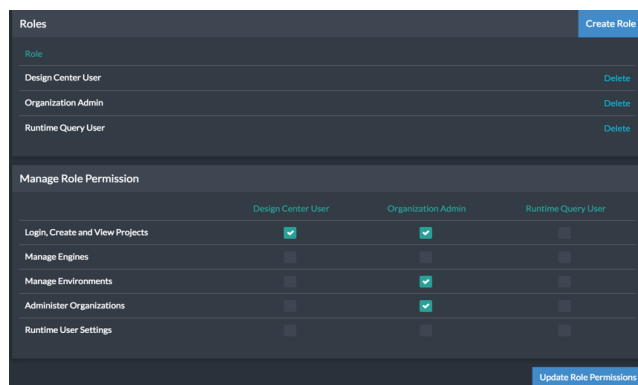
- Bulk or 'per sign-on' user synchronization
- Directory Group to AtScale Role Mapping
- Multiple User and Group DN Search Paths
- Directory Synchronization Filters

This support for enterprise directories allows IT organizations to maintain users, roles, and permissions in a single location, significantly reducing the cost of user onboarding and maintenance.



Role Based Authorization

In a Hadoop-based Business Intelligence environment there are a number of different user roles that need to be provisioned and enforced.



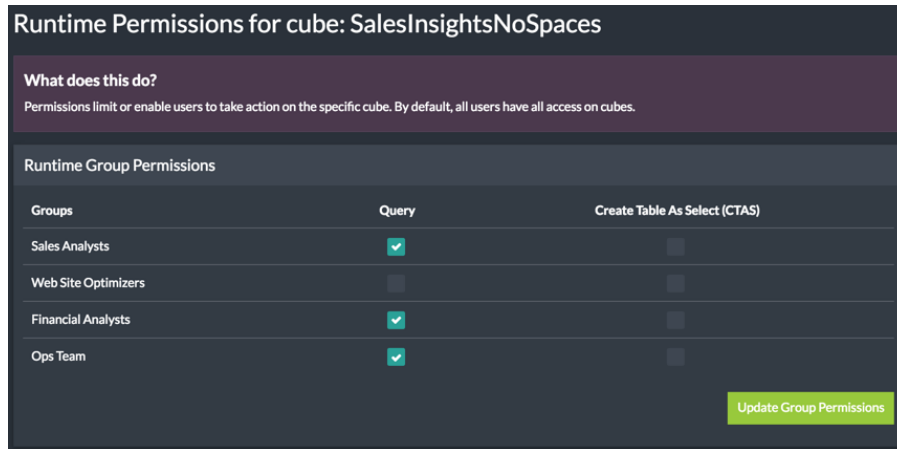
	Design Center User	Organization Admin	Runtime Query User
Login, Create and View Projects	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input type="checkbox"/>
Manage Engines	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
Manage Environments	<input type="checkbox"/>	<input checked="" type="checkbox"/>	<input type="checkbox"/>
Administer Organizations	<input type="checkbox"/>	<input checked="" type="checkbox"/>	<input type="checkbox"/>
Runtime User Settings	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>

For example, Data Architect within the Claims Analysis team might be authorized to create, edit, and publish AtScale Virtual Cubes within the "Claims Project" in AtScale. However, the Claims BI Team may only be able to query the elements that are exposed in these Virtual Cubes (but not edit them). Similarly, the AtScale Administrator may be allowed to create new users, enable/disable features, and configure Environments while the Data Architect may not.

AtScale includes a robust set of Role-level security that can be used to configure the right level of application authorization for different user types.

Object Level Security

In addition to controlling roles, users, and group authorization, AtScale's security model supports Object Level permissions. Essentially this determines which users and groups are able to access specific AtScale Projects and Virtual cubes.



The AtScale Object Level Security model supports two levels of permissions:

- **Design Time Permissions** control which Users and Groups are able to Create, Modify, and Publish cubes within AtScale's Design Center application.
- **Runtime Permissions** control which Users and Groups are able to query AtScale virtual cubes.

Within AtScale, a user's net permissions are the intersection of their Role Based Authorization (e.g. Cube Designer) and their Object Level Permissions (e.g. Design Time access to the Sales Cube). This means that the right users in the organization are able to access only the data assets that they are authorized to query and manage.

Secure Cluster Connection

When you connect AtScale to a Hadoop cluster, there are two approaches that can be used to create a secure, authorized, and authenticated connection:

1. A username/password can be provided for the JDBC connection.
2. The connection can be secured using Kerberos. This is the most common approach. More detail on this is provided below.

If Hadoop has been configured to run in Kerberos-secured mode, each Hadoop client connection must be authenticated by Kerberos in order to access the core Hadoop services. The Hadoop services leverage Kerberos to perform user authentication on all remote procedure calls (RPCs). Group resolution is then performed on the Hadoop NameNode and ResourceManager nodes respectively. Query processing tasks are executed using the operating system account of the user who submits a job (or by an impersonated or delegated user, if configured).

Configuring AtScale with a Kerberos-secured connection involves the following:

- **On the Hadoop cluster**
 - Create AtScale User on Hadoop Nodes.
 - Generate a Kerberos principal name for AtScale in your KDC (key distribution center). Principal names are usually in the format of:
os_username/client_node_hostname@your_company.com.
 - Create a single keytab file that contains the credentials for the following Hadoop services: hdfs, http, yarn, hive, impala (if using).
- **On the AtScale Node**
 - Install and configure the Kerberos client packages for your environment (follow the instructions for your operating system).
 - Obtain a .keytab file from your Kerberos administrator, and save it to a location on your AtScale machine (for example, /etc/conf/hadoop). The keytab file should contain the credentials for the following Hadoop service principals.

With AtScale's approach to Secure Cluster Connection the security of Kerberos can be leveraged while still allowing self-service BI access to authorization data assets.

Delegation and Impersonation

In designing the mechanisms that AtScale uses to query secured Hadoop clusters, we paid special attention to the enterprise security requirements that are consistent across industries. These security requirements include:

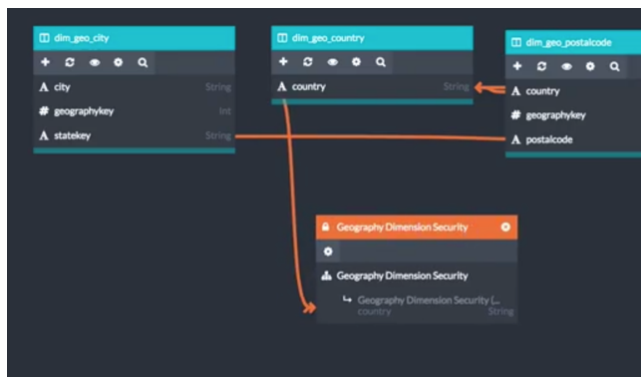
1. Users who query AtScale cubes should only be able to query data elements (both raw and aggregated) that they are authorized to see. For our customers, data authorization may mean one or several of the below:
 - a. Users are authorized to access only data in specific HDFS files or directories, for example /hdfs/landingzone/clinicaldata/chartlogs
 - b. Users are authorized to see only specific rows of data, for example rows of data where provider_id = "XGH-KPROC-978"
 - c. Users are only authorized to see specific columns of data, for example users in the "Non-Finance Analysts" group cannot see the "Claim Paid Amount" column
2. Our customers want to be able to manage data authorizations in a single location, and are often requiring that this "single source" of data authorizations be their Hadoop cluster. Currently we are aware of several approaches that customers are using to define their Hadoop-level data authorizations
 - a. HDFS file permissions using Unix users
 - b. HDFS file permissions using HDFS ACLs
 - c. Hadoop or Hive table permissions using Cloudera Sentry
 - d. Hadoop or Hive table permissions using Apache Ranger
3. Our customers want to be able to support query audit requirements that enable them to determine which users accessed which data elements, and when (where 'users' are BI/client users, not service account users).

To support the above requirements **AtScale supports both delegation** (a mechanism to pass Kerberos credentials from BI clients all the way through to the Hadoop query engine without requiring additional authentication) **and impersonation** (a mechanism to execute queries in “run as” mode, where the AtScale service account executes queries as the BI client user).

Because AtScale supports both Delegation and Impersonation, enterprises can maintain their user-level security policies in a single location (on the Hadoop cluster) while still enabling appropriate access for BI.

Row Level Data Access Security

In addition to role based, object based, and cluster-level security within the AtScale application, many businesses also want to be able to constrain the scope of the underlying data that end users are able to see. For example, in a multi-brand environment there may be a need to ensure that the manager for Brand A cannot see data for Brand B, and vice versa. Or, in a multi-tenant environment there may be a need to have a single data repository support multiple tenants or clients, each with their own restricted view of the underlying data.



To support this level of functionality AtScale supports the concept of ‘Per User Attribute Value Security’. Through the use of a special type of dimension, known as a Security Dimension, AtScale administrators can provide a lookup table that maps user ids to allowable values (such as brand_id or client_id) to ensure that all queries executed by these listed users are scoped to only include the allowable value(s).

Secure Data Transport Protocols

A final component of any secure business intelligence is the secure transport of data; query requests, usernames and passwords, result sets, and more. To support this you can configure AtScale so that all communication between every component of the AtScale system, as well as with the Hadoop cluster and BI clients, is secured using TLS (or [Transport Layer Security](#)).

Summary

As explained above, AtScale supports myriad security options across a variety of platform, users, roles, objects and systems that modern enterprises need and require in the world of big data today.

The intent of this document was to summarize AtScale’s unique and comprehensive approach to secure BI on Hadoop. At the same time we realize that each of the topics covered above is worthy of its own specific and detailed technical discussion. If you have further questions, let us know. We look forward to helping you make your BI work on Hadoop, securely, with all the scope and depth required by today’s modern and big data world.

Visit us at www.atscale.com to learn more.