Adding Hadoop Data Warehouses

A Hadoop data warehouse is a Hadoop cluster that contains the tables and views that you want to access as cube facts and dimensions. It also contains aggregate-table instances that either you or the AtScale engine creates in a schema that you specify.

- For Hadoop installations, the ability to run User-Defined Functions (UDFs) is required. This feature is required because it computes important statistics about the contents of tables. AtScale uses these statistics to make important decisions about aggregates, hierarchy validation, and more. Consequently, the advanced Engine settings for Hyper Log Log ++ (HLL) must be set to true. The settings are: **HLL.CREATEJAR** and **HLL.ENABLED**.
- There are cases where you can setup Kerberos using Kerberos Credential Cache. For more information, see Using Kerberos Credential Cache.

Before You Begin

- Ensure that you have the URLs for connecting to the primary and optionally the secondary name-node of the Hadoop cluster that the data warehouse will connect to.
- If the Hadoop cluster is secured with Kerberos, ensure that you have the Kerberos principal name to use for connections to each secured namenode server. The name must be in this format:
  `hdfs/hdfs_hostname@your_company.com`
- For the primary node, ensure that you know which setting to use for Hadoop RPC protection.
  - **authentication**: Authentication only (default)
  - **integrity**: Integrity check in addition to authentication
  - **privacy**: Data encryption in addition to integrity
- Ensure that your user ID in the Design Center is assigned the Super User role or is assigned the Manage Data Warehouses role permission.
- Ensure that you are in the correct organization in the Design Center. Go to **Settings > Global Settings > Organizations** to verify the organization.
- **Aggregate Schema**: AtScale reads and writes aggregate data to this schema. The AtScale service account user must have additional permissions on this schema. See the Data Warehouse Preparation instructions for more details. BI tool user accounts should not have the select permission on this schema.
- **AtScale UDAF Schema**: This is the storage location for the AtScale UDAF functions. The AtScale service account user must have additional permissions on this schema. See the Data Warehouse Preparation instructions for more details. If using impersonation, BI tool users must have the select permission on this schema.

- **File System Install Path**: The data warehouse file system path (e.g. HDFS) where AtScale will write required files. The AtScale service account user must have the write permission for this location. If the directory does not exist, AtScale will attempt to create it.

Procedure

To add a new Hadoop data warehouse:

1. Go to **Settings > Data Warehouses** and click **Create Data Warehouse**.
2. In the Add a Data Warehouse dialog box, under **Type of Data Warehouse**, select **Hadoop**.
3. Enter a unique name for the data warehouse.
4. (Optional) Enter the **External Connection ID** (formerly Query Name). The External Connection ID defaults to the data warehouse name. Override by entering a different name and enabling **Override generated value**.
5. Enter the name of the schema for AtScale to use when creating aggregate tables.
6. Specify the Data Warehouse as a **Read-only source**.

    - Select this option if AtScale will be configured to access this database with credentials that do not have permission to create new tables, or if you need to prevent AtScale from placing any aggregate tables in this data warehouse.
    - If **Read-only source** is enabled upon editing an existing data warehouse with aggregates present, all existing aggregates stored in this data warehouse will be deactivated.

7. If desired, enable **Impersonation** to allow AtScale to communicate with the data platform using the BI tool end-user user credentials. This allows the data warehouse administrator to apply more detailed security policies when securing data.
8. If Impersonation is enabled, toggle the settings as needed:

    - **Always Use Canary Queries**: Whether canary queries are required in light of aggregate misses.
    - **Allow Partial Aggregate Usage**: Allows mixed aggregate and raw data queries.

    

    When running HDP SparkSQL, enabling **Always Use Canary Queries** is required because SparkSQL is not integrated with Ranger (See RANGER-2128) and the alternative method of configuring Spark to run in LLAP mode is not supported by AtScale. It is not necessary to enable "Always Use Canary Queries" if running AtScale 2021.2.0+ and connecting to Apache Livy.

    

    AtScale cannot import data-loader bundles while impersonation is enabled. It is suggested that you import your desired data-loader bundles, commonly used for training, before enabling impersonation or use a separate AtScale installation for training purposes.

9. (AtScale 7.4.0 and later) Select the **Custom Function Installation Mode**:

- **AtScale Managed** - If selected, AtScale installs and manages the custom functions. enter the name of the AtScale User-Defined Aggregate Functions (UDAF) schema and a file system installation path.
- **Customer Managed** - If selected, the Hadoop administrator installs the custom functions in a designated schema. See Customer-Managed UDAF Installation on a Hadoop Cluster. This is the required setting when connecting to MAPR. Enter the name of the AtScale UDAF schema to use.
- **None** - Available for certain management edge-cases. Do not use this settings when servicing end-user queries because system performance will be degraded.

10. If you selected **AtScale Managed**: Enter the file system install path.
    The AtScale engine uses this path to load data into the Hadoop Distributed File System (HDFS).
11. Configure the file system type. This step is optional if your **Custom Function Installation Mode** is set to Customer Managed or None.
12. Enter the URI for Hadoop NameNode 1.
13. (Optional) Enter the Kerberos principal (if AtScale is configured to use Kerberos for authentication). For details, see Configuring Kerberos.
14. Select the Hadoop RPC protection value.
15. (Optional) Enter the information for NameNode 2.
16. Click **Save**.

Add A Hadoop Connection

After setting up your data warehouse, you must add a SQL engine connection to the Hadoop data warehouse before you can run queries. You can establish multiple connections with different configurations to the same data warehouse. If multiple Hadoop connections are configured, you can assign which types of queries are permitted per connection. For more information on adding SQL-on-Hadoop engines, see Adding SQL-on-Hadoop Engines.

1. Enter a unique name for the connection.
2. Select the **Server Type** from the available selections.
3. Enter the **Hostname** of the hosts which have this SQL engine installed. Multiple hosts can be separated by commas. Ranges are supported such as 192.168.1.[1-10], or 192.168.1.[1, 3, 5].
4. Enter the **Port** number.
5. Specify any **Extra JDBC** flags.
6. Enter an **Engine Management Console URL**.
7. Select the **Authorization** type.

    - If Username and Password authentication is used, select **Password**.

      If you have external secret manager enabled and configured, enter the corresponding credentials.

    - Select **Kerberos** if Kerberos authentication is configured.

8. Test the connection.
9. Click **Save** to complete the setup.