Requirements for Hive ACID

Contents

Hive ACID

Capture	Hub	Integrate
✗	✗	✓

This section describes the requirements, access privileges, and other features of HVR when using Hive ACID (Atomicity, Consistency, Isolation, Durability) for replication. For information about compatibility and supported versions of Hive ACID with HVR platforms, see Platform Compatibility Matrix.

For the Capabilities supported by HVR on Hive ACID, see Capabilities for Hive ACID.

For information about the supported data types and mapping of data types in source DBMS to the corresponding data types in target DBMS or file format, see Data Type Mapping.

ODBC Connection

HVR uses ODBC connection to the Hive ACID server. The HortonWorks ODBC driver or Cloudera ODBC driver must be installed on the machine from which HVR connects to the Hive ACID server.

For information about the supported ODBC driver version, refer to the HVR release notes (**hvr.rel**) available in **hvr_home** directory or the download page.

HVR can deliver changes into Hive ACID tables as a target location for its refresh and integration. Delivery of changes into Hive ACID tables for Hive versions before 2.3 is only supported if the following action is defined:

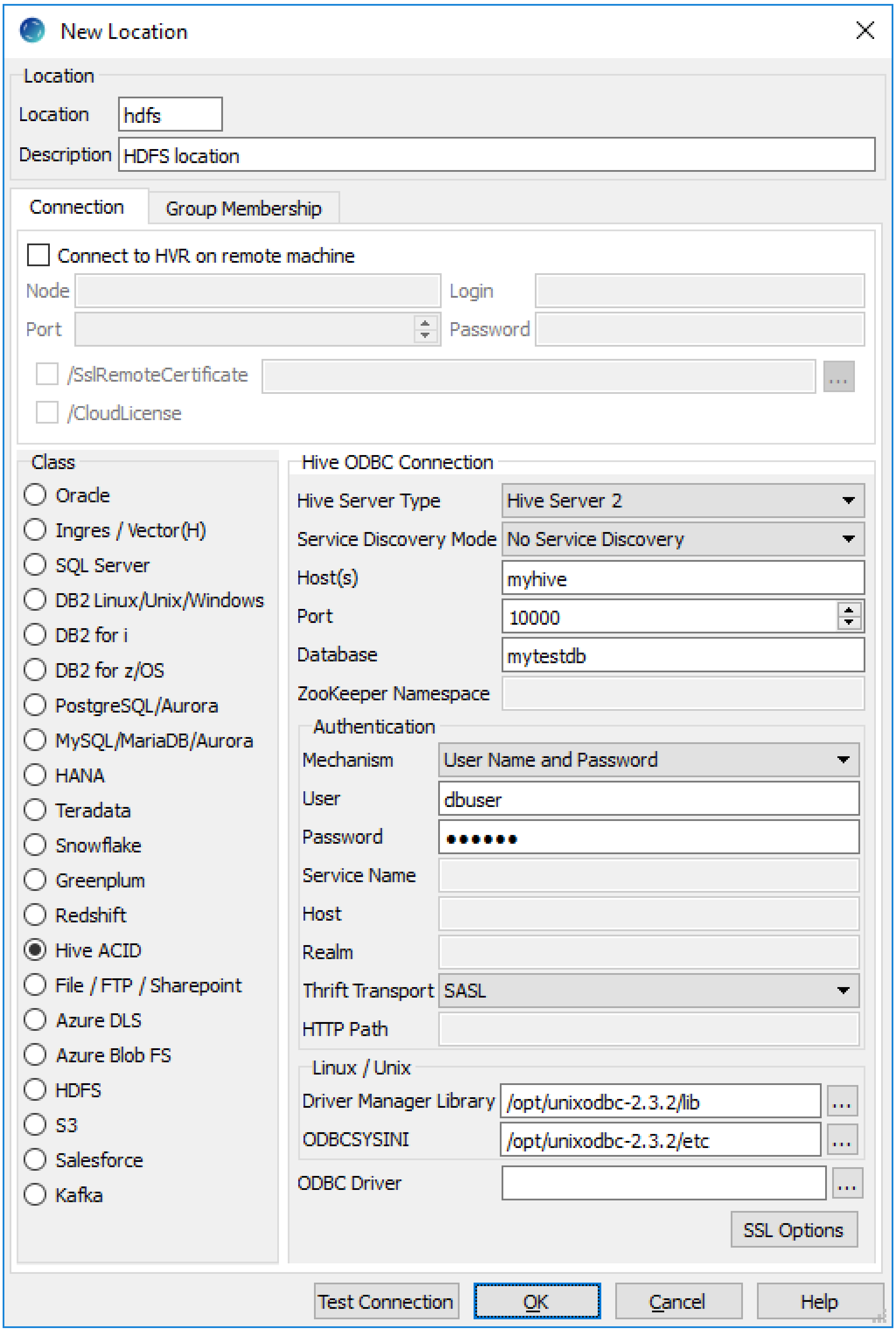
Group	Table	Action
Hive ACID	*	**ColumnProperties /TimeKey**

For file formats (JSON and Avro), the following action definition is required to handle certain limitations when execute any SQL statement against Hive external tables (due to compatibility issues with Hive 3 Metastore and Hive ODBC drivers):

Group	Table	Action
Hive ACID	*	**Environment /Name=HVR_ODBC_CONNECT_STRING_ADD /Value="UseNativeQuery=1"**

Location Connection

This section lists and describes the connection details required for creating Hive ACID location in HVR.



New Location

Location

Location: hdfs

Description: HDFS location

Connection | Group Membership

□ Connect to HVR on remote machine

Node: Login:

Port: Password:

□ /SslRemoteCertificate

□ /CloudLicense

Class

○ Oracle
○ Ingres / Vector(H)
○ SQL Server
○ DB2 Linux/Unix/Windows
○ DB2 for i
○ DB2 for z/OS
○ PostgreSQL/Aurora
○ MySQL/MariaDB/Aurora
○ HANA
○ Teradata
○ Snowflake
○ Greenplum
○ Redshift
● Hive ACID
○ File / FTP / Sharepoint
○ Azure DLS
○ Azure Blob FS
○ HDFS
○ S3
○ Salesforce
○ Kafka

Hive ODBC Connection

Hive Server Type: Hive Server 2

Service Discovery Mode: No Service Discovery

Host(s): myhive

Port: 10000

Database: mytestdb

ZooKeeper Namespace:

Authentication

Mechanism: User Name and Password

User: dbuser

Password: ••••••

Service Name:

Host:

Realm:

Thrift Transport: SASL

HTTP Path:

Linux / Unix

Driver Manager Library: /opt/unixodbc-2.3.2/lib

ODBCSYSINI: /opt/unixodbc-2.3.2/etc

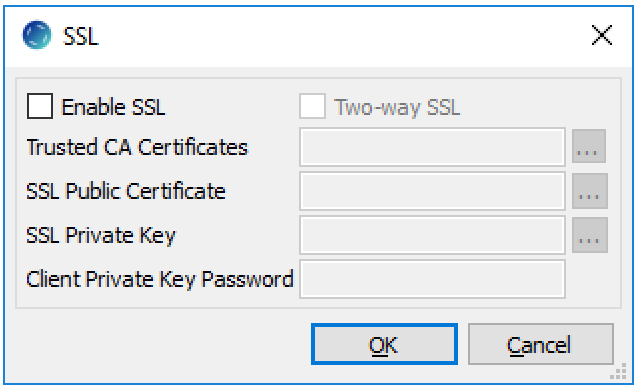
ODBC Driver:

Field	Description
Hive ODBC Connection	
Hive Server Type	The type of Hive server. Available options: - **Hive Server 1** (default): The driver connects to a Hive Server 1 instance. - **Hive Server 2**: The driver connects to a Hive Server 2 instance.

Service Discovery Mode	The mode for connecting to Hive. This field is enabled only if **Hive Server Type** is **Hive Server 2**. Available options: • **No Service Discovery** (default): The driver connects to Hive server without using the ZooKeeper service. • **ZooKeeper**: The driver discovers Hive Server 2 services using the ZooKeeper service.
Host(s)	The hostname or IP address of the Hive server. When Service Discovery Mode is ZooKeeper, specify the list of ZooKeeper servers in following format **[ZK_Host1]:[ZK_Port1],[ZK_Host2]:[ZK_Port2]**, where **[ZK_Host]** is the IP address or hostname of the ZooKeeper server and **[ZK _Port]** is the TCP port that the ZooKeeper server uses to listen for client connections. **Example:** hive-host
Port	The TCP port that the Hive server uses to listen for client connections. This field is enabled only if **Service Discovery Mode** is **No Service Discovery**. **Example:** 10000
Database	The name of the database schema to use when a schema is not explicitly specified in a query. **Example:** mytestdb
ZooKeeper Namespace	The namespace on ZooKeeper under which Hive Server 2 nodes are added. This field is enabled only if **Service Discovery Mode** is **ZooKeeper**.
Authentication	
Mechanism	The authentication mode for connecting HVR to **Hive Server 2**. This field is enabled only if **Hive Server Type** is **Hive Server 2**. Available options: • **No Authentication** (default) • **User Name** • **User Name and Password** • **Kerberos** • **Windows Azure HDInsight Service** **Since** v5.5.0/2
User	The username to connect HVR to Hive server. This field is enabled only if **Mechan ism** is **User Name** or **User Name and Password**. **Example:** dbuser
Password	The password of the **User** to connect HVR to Hive server. This field is enabled only if **Mechanism** is **User Name and Password**.
Service Name	The Kerberos service principal name of the Hive server. This field is enabled only if **Mechanism** is **Kerberos**.
Host	The Fully Qualified Domain Name (FQDN) of the Hive Server 2 host. The value of **Host** can be set as **_HOST** to use the Hive server hostname as the domain name for Kerberos authentication. If **Service Discovery Mode** is disabled, then the driver uses the value specified in the Host connection attribute. If **Service Discovery Mode** is enabled, then the driver uses the **Hive Server 2** host name returned by ZooKeeper. This field is enabled only if **Mechanism** is **Kerberos**.
Realm	The realm of the Hive Server 2 host. It is not required to specify any value in this field if the realm of the Hive Server 2 host is defined as the default realm in Kerberos configuration. This field is enabled only if **Mechanism** is **Kerberos**.

Thrift Transport **Since** v5.5.0/2	The transport protocol to use in the Thrift layer. This field is enabled only if **Hive Server Type** is **Hive Server 2**. Available options: - **Binary** (This option is available only if **Mechanism** is **No Authentication** or **User Name and Password**.) - **SASL** (This option is available only if **Mechanism** is **User Name** or **User Name and Password** or **Kerberos**.) - **HTTP** (This option is not available if **Mechanism** is **User Name**.) For information about determining which Thrift transport protocols your Hive server supports, refer to HiveServer2 Overview and Setting Up HiveServer2 sections in Hive documentation.
HTTP Path **Since** v5.5.0/2	The partial URL corresponding to the Hive server. This field is enabled only if **Thrift Transport** is **HTTP**.
Linux / Unix	
Driver Manager Library	The optional directory path where the ODBC Driver Manager Library is installed. This field is applicable only for Linux/Unix operating system. For a default installation, the ODBC Driver Manager Library is available at **/usr/lib64** and does not need to be specified. However, when UnixODBC is installed in for example **/opt/unixodbc** the value for this field would be **/opt/unixodbc/lib**.
ODBCSYSINI	The optional directory path where **odbc.ini** and **odbcinst.ini** files are located. This field is applicable only for Linux/Unix operating system. For a default installation, these files are available at **/etc** and do not need to be specified. However, when UnixODBC is installed in for example **/opt/unixodbc** the value for this field would be **/opt/unixodbc/etc**.
ODBC Driver	The user defined (installed) ODBC driver to connect HVR to the Hive server.
SSL Options	Show **SSL Options**.

SSL Options



SSL

□ Enable SSL □ Two-way SSL

Trusted CA Certificates

SSL Public Certificate

SSL Private Key

Client Private Key Password

Field	Description
Enable SSL	Enable/disable (one way) SSL. If enabled, HVR authenticates the Hive server by validating the SSL certificate shared by the Hive server.
Two-way SSL	Enable/disable two way SSL. If enabled, both HVR and Hive server authenticate each other by validating each others SSL certificate. This field is enabled only if **Enable SSL** is selected.

Trusted CA Certificates	The directory path where the **.pem** file containing the server's public SSL certificate signed by a trusted CA is located. This field is enabled only if **Enable SSL** is selected.
SSL Public Certificate	The directory path where the **.pem** file containing the client's SSL public certificate is located. This field is enabled only if **Two-way SSL** is selected.
SSL Private Key	The directory path where the **.pem** file containing the client's SSL private key is located. This field is enabled only if **Two-way SSL** is selected.
Client Private Key Password	The password of the private key file that is specified in **SSL Private Key**. This field is enabled only if **Two-way SSL** is selected.

Hive ACID on Amazon Elastic MapReduce (EMR)

To enable Hive ACID on Amazon EMR,

1. Add the following configuration details to the **hive-site.xml** file available in **/etc/hive/conf** on Amazon EMR:

```
<!-- Hive ACID support -->
<property>
   <name>hive.compactor.initiator.on</name>
   <value>true</value>
</property>
<property>
   <name>hive.compactor.worker.threads</name>
   <value>10</value>
</property>
<property>
   <name>hive.support.concurrency</name>
   <value>true</value>
</property>
<property>
   <name>hive.txn.manager</name>
   <value>org.apache.hadoop.hive.ql.lockmgr.DbTxnManager</value>
</property>
<property>
   <name>name>hive.enforce.bucketing</name>
   <value>true</value>
</property>
<property>
   <name>hive.exec.dynamic.partition.mode</name>
   <value>nostrict</value>
</property>
<!-- Hive ACID support end -->
```

2. Save the modified **hive-site.xml** file.
3. Restart Hive on Amazon EMR.
   For more information on restarting a service in Amazon EMR, refer to How do I restart a service in Amazon EMR? in AWS documentation.

Integrate and Refresh Target

HVR supports integrating changes into Hive ACID location. This section describes the configuration requirements for integrating changes (using **Integrate** and **refresh**) into Hive ACID location. For the list of supported Hive ACID versions, into which HVR can integrate changes, see Integrate changes into location in Capabilities.

Burst Integrate and Bulk Refresh

While **HVR Integrate** is running with parameter **/Burst** and Bulk **Refresh**, HVR can stream data into a target database straight over the network into a bulk loading interface specific for each DBMS (e.g. direct-path-load in Oracle), or else HVR puts data into a temporary directory (‘staging file') before loading data into a target database.

For best performance, HVR performs **Integrate** with **/Burst** and Bulk **Refresh** on Hive ACID location using staging files. HVR implements **Integrate** with **/Burst** and Bulk **Refresh** (with file staging) into Hive ACID as follows:

1. HVR first creates Hive external tables using Amazon/HortonWorks Hive ODBC driver
2. HVR then stages data into:
    - S3 using AWS S3 REST interface (cURL library) or
    - HDFS/Azure Blob FS/Azure Data Lake Storage using HDFS-compatible libhdfs API
3. HVR uses Hive SQL commands '**merge**' (**Integrate** with **/Burst**) or '**insert into**' (Bulk **Refresh**) against the Hive external tables linked to S3/HDFS/Azure Blob FS/Azure Data Lake Storage to ingest data into ACID Hive managed tables.

The following is required to perform **Integrate** with parameter **/Burst** and Bulk **Refresh** into Hive ACID:

1. HVR requires an AWS S3 or HDFS/Azure Blob FS/Azure Data Lake Storage location to store temporary data to be loaded into Hive ACID.

    If AWS S3 is used to store temporary data then HVR requires the AWS user with 'AmazonS3FullAccess' policy to access this location. For more information, refer to the following AWS documentation:
    - Amazon S3 and Tools for Windows PowerShell
    - Managing Access Keys for IAM Users
    - Creating a Role to Delegate Permissions to an AWS Service
2. Define action **LocationProperties** on Hive ACID location with the following parameters:
    - **/StagingDirectoryHvr**: the location where HVR will create the temporary staging files. The format for AWS S3 is **s3://***S3 Bucket***/***Directory* and for HDFS is **hdfs://***NameNode***:***Port* **/***Directory*
    - **/StagingDirectoryDb**: the location from where Hive ACID will access the temporary staging files.
      If **/StagingDirectoryHvr** is an AWS S3 location then the value for **/StagingDirectoryDb** should be same as **/StagingDirectoryHvr**.
    - **/StagingDirectoryCredentials**: the AWS security credentials. The supported formats are **'aws_access_key_id="***key***";aws_secret_access_key="***secret_key***"'** or **'role="***AWS_role* **"'**. How to get your AWS credential or Instance Profile Role can be found on the AWS documentation web page.
3. Since HVR uses CSV file format for staging, the following action definitions are required to handle certain limitations of the Hive deserialization implementation:

Group	Table	Action
Hive ACID	*	**TableProperties** **/CharacterMapping="\x00>\\0;\n>\\n;\r>\\r;">\""**
Hive ACID	*	**TableProperties** **/MapBinary=BASE64**