Requirements for S3

Contents

S3

Capture	Hub	Integrate
✓	✗	✓

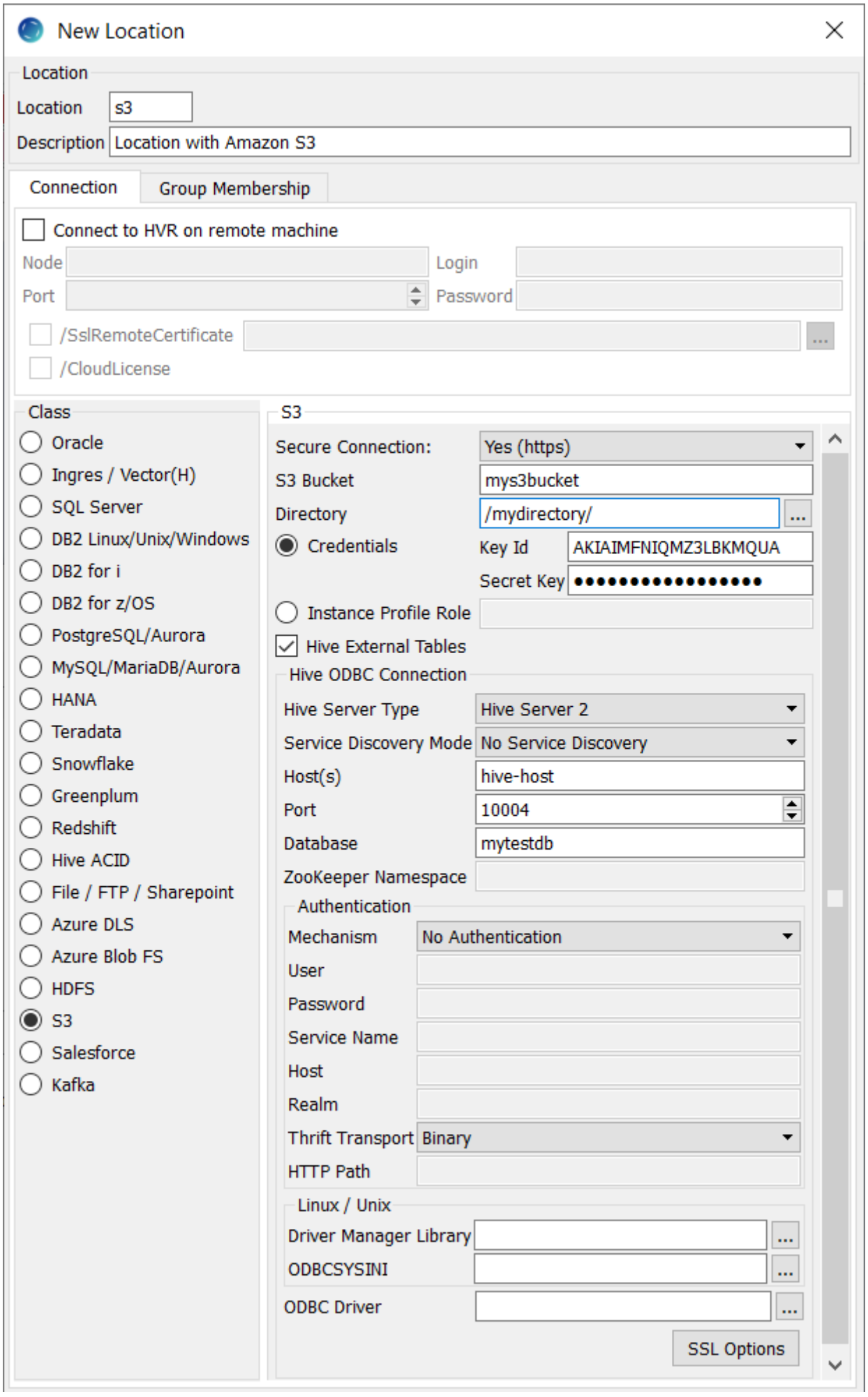
This section describes the requirements, access privileges, and other features of HVR when using Amazon S3(Simple Storage Service) for replication.

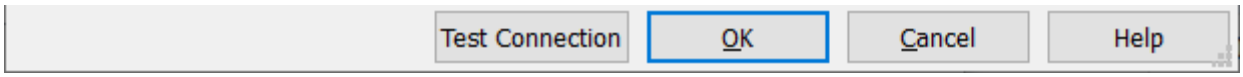
For information about compatibility and support for S3 with HVR platforms, see Platform Compatibility Matrix.

If there is an HVR agent running on Amazon EC2 node, which is in the AWS network together with the S3 bucket, then the communication between the HUB and AWS network is done via HVR protocol, which is more efficient than direct S3 transfer. Another approach to avoid the described bottleneck is to configure the HUB on an EC2 node.

Location Connection

This section lists and describes the connection details/parameters required for creating an S3 location in HVR. HVR uses the S3 REST interface (cURL library) to connect, read and write data to S3 during capture, integrate (continuous), refresh (bulk) and compare (direct file compare).


New Location
Location
Location
s3
Description
Location with Amazon S3
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
S3
Secure Connection:
Yes (https)
S3 Bucket
mys3bucket
Directory
/mydirectory/
Credentials
Key Id
AKIAIMFNIQMZ3LBKMQUA
Secret Key
Instance Profile Role
Hive External Tables
Hive ODBC Connection
Hive Server Type
Hive Server 2
Service Discovery Mode
No Service Discovery
Host(s)
hive-host
Port
10004
Database
mytestdb
ZooKeeper Namespace
Authentication
Mechanism
No Authentication
User
Password
Service Name
Host
Realm
Thrift Transport
Binary
HTTP Path
Linux / Unix
Driver Manager Library
ODBCSYSINI
ODBC Driver
SSL Options



Field	Description
S3	
Secure Connection	The type of security to be used for connecting to S3 Server. Available options: • **Yes (https)** (**default**): HVR will connect to S3 Server using HTTPS. • **No (http)**: HVR will connect to S3 Server using HTTP.
S3 Bucket	The IP address or hostname of the S3 bucket. **Example:** rs-bulk-load
Directory	The directory path in **S3 Bucket** which is to be used for replication. **Example:** /myserver/hvr/s3
Credentials	The authentication mode for connecting HVR to S3 by using IAM User Access Keys (**Key ID** and **Secret Key**). For more information about Access Keys, refer to Access Keys (Access Key ID and Secret Access Key) in section 'Understanding and Getting Your Security Credentials' of AWS documentation.
Key ID	The access key ID of IAM user to connect HVR to S3. This field is enabled only if **Credentials** is selected. **Example:** AKIAIMFNIQMZ2LBKMQUA
Secret Key	The secret access key of IAM user to connect HVR to S3. This field is enabled only if **Credentials** is selected.
Instance Profile Role	The AWS IAM role name. This authentication mode is used when connecting HVR to S3 by using AWS Identity and Access Management (IAM) Role. This option can be used only if the HVR remote agent or the HVR Hub is running inside the AWS network on an EC2 instance and the AWS IAM role specified here should be attached to this EC2 instance. When a role is used, HVR obtains temporary Access Keys Pair from the EC2 machine. For more information about IAM Role, refer to IAM Roles in AWS documentation. **Example:** Role1 or PRODROLE
Hive External Tables	Enable/Disable Hive ODBC connection configuration for creating Hive external tables above S3.

Hive ODBC Connection

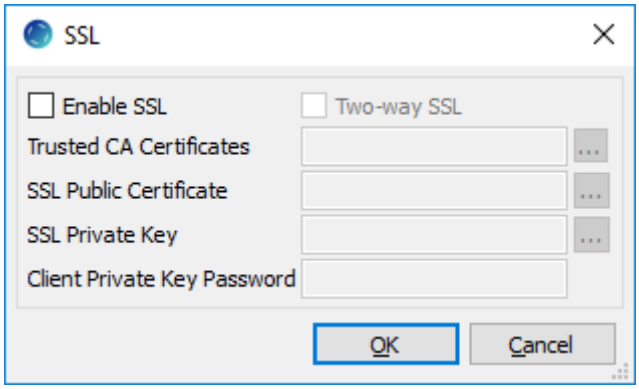
Following are the connection details/parameters required for connecting HVR to the Hive server.

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 **Mechanism** 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



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.

Permissions

To run a capture or integration with Amazon S3 location, it is recommended that the AWS User has the **AmazonS3FullAccess** permission policy.

AmazonS3ReadOnlyAccess policy is enough only for capture locations, which have a **LocationProperties /StateDirectory** defined.

The minimal permission set for integrate location are:

- **s3:GetBucketLocation**
- **s3:ListBucket**
- **s3:ListBucketMultipartUploads**
- **s3:AbortMultipartUpload**
- **s3:GetObject**
- **s3:PutObject**
- **s3:DeleteObject**

```
{
    "Statement": [
        {
            "Sid": <identifier>,
            "Effect": "Allow",
            "Principal": {
                    "AWS": "arn:aws:iam::<account_id>:<user>/<username>",
            },
            "Action": [
                    "s3:GetObject",
                    "s3:GetObjectVersion",
                    "s3:PutObject",
                    "s3:DeleteObject",
                    "s3:DeleteObjectVersion",
                    "s3:AbortMultipartUpload"
            ],
            "Resource": "arn:aws:s3:::<bucket_name>/*"
        },
        {
            "Sid": <identifier>,
            "Effect": "Allow",
            "Principal": {
                    "AWS": "arn:aws:iam::<account_id>:<user>/<username>"
            },
            "Action": [
                    "s3:ListBucket",
                    "s3:GetBucketLocation",
                    "s3:ListBucketMultipartUploads"
            ],
            "Resource": "arn:aws:s3:::<bucket_name>"
        }
    ]
}
```

For more information on the Amazon S3 permissions policy, refer to the AWS S3 documentation.

S3 Encryption

HVR supports client or server-side encryption for uploading files into S3 locations. To enable the client or server-side encryption for S3, see action **LocationProperties /S3Encryption**.

AWS China

For enabling HVR to interact with AWS China cloud, define the **Environment** variable **HVR_AWS_CLOUD** with value **CHINA** on the hub and remote machine.

S3 encryption with Key Management Service (KMS) is not supported in the AWS China cloud.

Hive External Tables

To **Compare** files that reside on the S3 location, HVR allows you to create Hive external tables above S3. The connection details/parameters for Hive ODBC can be enabled for S3 in the location creation screen by selecting the **Hive External Tables** field (see section Location Connection). For more information about configuring Hive external tables, refer to Hadoop Amazon Web Services Support and Apache Hadoop - Amazon EMR documentation.

ODBC Connection

HVR uses an ODBC connection to the Amazon EMR cluster for which it requires the ODBC driver (Amazon ODBC or HortonWorks ODBC) for Hive installed on the machine (or in the same network). The Amazon and HortonWorks ODBC drivers are similar and compatible to work with Hive 2.x release. However, it is recommended to use the Amazon ODBC driver for Amazon Hive and the Hortonworks ODBC driver for HortonWorks Hive. 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.

On Linux, HVR additionally requires unixODBC.

By default, HVR uses Amazon ODBC driver for connecting to Hadoop. To use the Hortonworks ODBC driver:

- For HVR versions since 5.3.1/25.1, use the **ODBC Driver** field available in the **New Location** screen to select the (user installed) Hortonworks ODBC driver.
- Prior to HVR 5.3.1/25.1, the following action definition is required:

Linux

Group	Table	Action
S3	*	**Environment /Name=HVR_ODBC_CONNECT_STRING_DRIVER /Value=Hortonworks Hive ODBC Driver 64-bit**

Windows

Group	Table	Action
S3	*	**Environment /Name=HVR_ODBC_CONNECT_STRING_DRIVER /Value=Hortonworks Hive ODBC Driver**

Amazon does not recommend changing the security policy of the EMR. This is the reason why it is required to create a tunnel between the machine where the ODBC driver is installed and the EMR cluster. On Linux, Unix and macOS you can create the tunnel with the following command:

```
ssh -i ~/mykeypair.pem -N -L 8157:ec2-###-##-##-###.compute-1.amazonaws.
com:8088 hadoop@ec2-###-##-##-###.compute-1.amazonaws.com
```

Channel Configuration

For the file formats (CSV, JSON, and AVRO) the following action definitions are required to handle certain limitations of the Hive deserialization implementation during Bulk or Row-wise **Compare**:

- For CSV

Group	Table	Action
S3	*	**FileFormat /NullRepresentation=\\N**
S3	*	**TableProperties /CharacterMapping="\x00>\\0;\n>\\n;\r>\\r;">\""**
S3	*	**TableProperties /MapBinary=BASE64**

- For JSON

Group	Table	Action
S3	*	**TableProperties** /MapBinary=BASE64
S3	*	**FileFormat /JsonMode=ROW_FRAGMENTS**

- For Avro

Group	Table	Action
S3	*	**FileFormat /AvroVersion=v1_8**

v1_8 is the default value for **FileFormat /AvroVersion**, so it is not mandatory to define this action.

Integrate

HVR allows you to perform **HVR Refresh** or **Integrate** changes into an S3 location. This section describes the configuration requirements for integrating changes (using **HVR Refresh** or **Integrate**) into the S3 location.

Customize Integrate

Defining action **Integrate** is sufficient for integrating changes into an S3 location. However, the default file format written into a target file location is HVR's own XML format and the changes captured from multiple tables are integrated as files into one directory. The integrated files are named using the integrate timestamp.

You may define other actions for customizing the default behavior of integration mentioned above. Following are few examples that can be used for customizing integration into the S3 location:

Group	Table	Action	Annotation
S3	*	**FileFormat**	This action may be defined to: • specify the format (**Xml**, **Csv**, **Avro**, **Json**, or **Parquet**) of the files integrated into the target location. • escape any delimiters (e.g. comma) present in a column using the parameter **/QuoteCharacter**. • escape the quote character (**/QuoteCharacter**) defined, using the parameter **/EscapeCharacter**.
S3	*	**Integrate /RenameExpression**	To segregate and name the files integrated into the target location. For example, if **/RenameExpression={hvr_tbl_name}/{hvr_integ_tstamp}.csv** is defined, then for each table in the source, a separate folder (with the same name as the table name) is created in the target location, and the files replicated for each table are saved into these folders. This also enforces unique name for the files by naming them with a timestamp of the moment when the file was integrated into the target location.

S3	*	**ColumnProperties**	This action defines properties for a column being replicated. This action may be defined to:
			• integrate the delete operation. By default, for file-based target locations, HVR does not replicate the **delete** operation performed at the source location. So to integrate the delete operation, an extra column for timekey (**/TimeKey**) needs to be added in the target location. For this, action **ColumnProperties** may be defined with the following parameters:
			  • **/Name**: This parameter defines the name for the extra column in the target location.
			  • **/Extra**: This parameter defines that this is an extra column in the target location (a column which is not present in the source location).
			  • **/IntegrateExpression**: This parameter defines the expression to be used for generating the **TimeKey** value. For example, **{hvr_integ_seq}** can be used here. This is a 36 byte string value (hex characters) which is unique and continuously increasing for a specific source location.
			  • **/TimeKey**: This parameter defines that this is a **TimeKey** column.
			  • **/Datatype=varchar**: This parameter defines the data type for the extra column.
			  • **/Length=36**: This parameter defines the data type length for the extra column.
			• add the source operation type (using **hvr_op**) information in the target location. This action definition is required for performing **HVR Compare** if **ColumnProperties /TimeKey** column is defined on a target file location. For this, action **ColumnProperties** may be defined with the following parameters:
			  • **/Name**: This parameter defines the name for the extra column in the target location.
			  • **/Extra**: This parameter defines that this is an extra column in the target location (a column which is not present in the source location).
			  • **/IntegrateExpression={hvr_op}**: This parameter defines the expression to be used for generating the information about source operation type.
			  • **/Datatype=integer**: This parameter defines the data type for this extra column.

Integrate Limitations

By default, for file-based target locations, HVR does not replicate the **delete** operation performed at the source location.