Informatica® PowerExchange for Amazon Redshift

10.1.1 Update 2

User Guide

Informatica PowerExchange for Amazon Redshift User Guide
10.1.1 Update 2
April 2017

Publication Date: 2018-09-27

Table of Contents

Preface

The *Informatica PowerExchange® for Amazon Redshift User Guide* provides information about reading data from and writing data to Amazon Redshift. It is written for Amazon Redshift administrators and developers who create mappings to read from or write data from a Amazon Redshift resource.

This guide assumes that you have knowledge of Amazon Redshift and Informatica.

Informatica Resources

Informatica Network

Informatica Network hosts Informatica Global Customer Support, the Informatica Knowledge Base, and other product resources. To access Informatica Network, visit https://network.informatica.com.

As a member, you can:

- Access all of your Informatica resources in one place.
- Search the Knowledge Base for product resources, including documentation, FAQs, and best practices.
- View product availability information.
- Review your support cases.
- Find your local Informatica User Group Network and collaborate with your peers.

Informatica Knowledge Base

Use the Informatica Knowledge Base to search Informatica Network for product resources such as documentation, how-to articles, best practices, and PAMs.

To access the Knowledge Base, visit https://kb.informatica.com. If you have questions, comments, or ideas about the Knowledge Base, contact the Informatica Knowledge Base team at KB_Feedback@informatica.com.

Informatica Documentation

To get the latest documentation for your product, browse the Informatica Knowledge Base at https://kb.informatica.com/_layouts/ProductDocumentation/Page/ProductDocumentSearch.aspx.

If you have questions, comments, or ideas about this documentation, contact the Informatica Documentation team through email at infa_documentation@informatica.com.

Informatica Product Availability Matrixes

Product Availability Matrixes (PAMs) indicate the versions of operating systems, databases, and other types of data sources and targets that a product release supports. If you are an Informatica Network member, you can access PAMs at https://network.informatica.com/community/informatica-network/product-availability-matrices.

Informatica Velocity

Informatica Velocity is a collection of tips and best practices developed by Informatica Professional Services. Developed from the real-world experience of hundreds of data management projects, Informatica Velocity represents the collective knowledge of our consultants who have worked with organizations from around the world to plan, develop, deploy, and maintain successful data management solutions.

If you are an Informatica Network member, you can access Informatica Velocity resources at http://velocity.informatica.com.

If you have questions, comments, or ideas about Informatica Velocity, contact Informatica Professional Services at ips@informatica.com.

Informatica Marketplace

The Informatica Marketplace is a forum where you can find solutions that augment, extend, or enhance your Informatica implementations. By leveraging any of the hundreds of solutions from Informatica developers and partners, you can improve your productivity and speed up time to implementation on your projects. You can access Informatica Marketplace at https://marketplace.informatica.com.

Informatica Global Customer Support

You can contact a Global Support Center by telephone or through Online Support on Informatica Network.

To find your local Informatica Global Customer Support telephone number, visit the Informatica website at the following link:
http://www.informatica.com/us/services-and-training/support-services/global-support-centers.

If you are an Informatica Network member, you can use Online Support at http://network.informatica.com.

CHAPTER 1

Introduction to PowerExchange for Amazon Redshift

This chapter includes the following topics:

PowerExchange for Amazon Redshift Overview

You can use PowerExchange for Amazon Redshift to read data from or write data to Amazon Redshift. You can also use PowerExchange for Amazon Redshift to read data from Amazon Redshift views.

Amazon Redshift views contain information about the functioning of the Amazon Redshift system. You can run a query on views like you run a query on database tables. You can select multiple schemas for Amazon Redshift objects.

You can use Amazon Redshift objects as sources and targets in mappings. When you use Amazon Redshift objects in mappings, you must configure properties specific to Amazon Redshift. You can validate and run mappings in native or Hadoop environments.

You can configure HTTPS proxy to connect to Amazon Redshift. You can also configure an SSL connection to connect to Amazon Redshift.

The Data Integration Service uses the Amazon driver to communicate with Amazon Redshift.

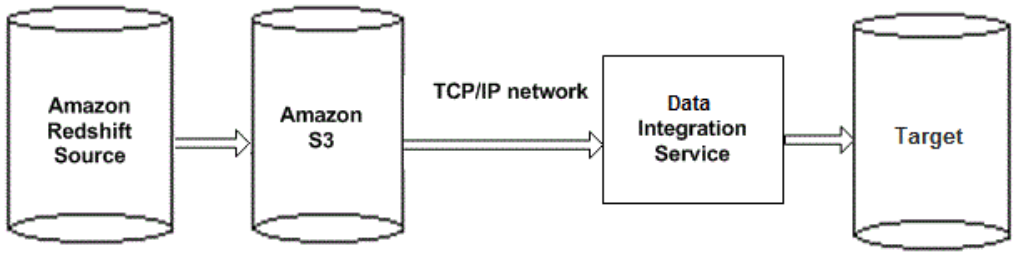
Example

You work for an organization that stores purchase order details, such as customer ID, item codes, and item quantity in an on-premise MySQL database. You need to analyze purchase order details and move data from the on-premise MySQL database to an affordable cloud-based environment. Create a mapping to read all the purchase records from the MySQL database and write them to Amazon Redshift for data analysis.

Data Integration Service and Amazon Redshift Integration

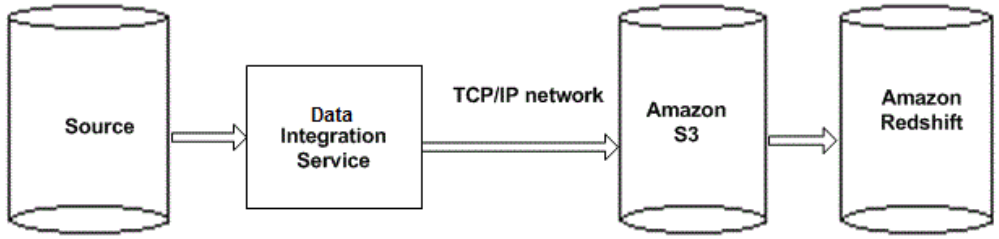
The Data Integration Service uses the Amazon Redshift connection to connect to Amazon Redshift.

The following image shows how Informatica connects to Amazon Redshift to read data:



When you run the Amazon Redshift mapping, the Data Integration Service reads data from Amazon Redshift based on the workflow and Amazon Redshift connection configuration. The Data Integration Service connects and reads data from Amazon Simple Storage Service (Amazon S3) through a TCP/IP network. The Data Integration Service then stores data in a staging directory on the Informatica machine. Amazon S3 is a storage service in which you can copy data from source and simultaneously move data to any target. The Data Integration Service issues a copy command that copies data from Amazon S3 to the target.

The following image shows how Informatica connects to Amazon Redshift to write data:



When you run the Amazon Redshift mapping, the Data Integration Service writes data to Amazon Redshift based on the workflow and Amazon Redshift connection configuration. The Data Integration Service stores data in a staging directory on the Informatica machine. The Data Integration Service then connects and writes data to Amazon Simple Storage Service (Amazon S3) through a TCP/IP network. Amazon S3 is a storage service in which you can copy data from source and simultaneously move data to Amazon Redshift clusters. The Data Integration Service issues a copy command that copies data from Amazon S3 to the Amazon Redshift target table.

Introduction to Amazon Redshift

Amazon Redshift is a cloud-based petabyte-scale data warehouse service that organizations can use to analyze and store data.

Amazon Redshift uses columnar data storage, parallel processing, and data compression to store data and to achieve fast query execution. Amazon Redshift uses a cluster-based architecture that consists of a leader

node and compute nodes. The leader node manages the compute nodes and communicates with the external client programs. The leader node interacts with the client applications and communicates with compute nodes. A compute node stores data and runs queries for the leader node. Any client that uses a PostgreSQL driver can communicate with Amazon Redshift.

CHAPTER 2

PowerExchange for Amazon Redshift Configuration

This chapter includes the following topics:

PowerExchange for Amazon Redshift Configuration Overview

PowerExchange for Amazon Redshift is installed with Informatica Data Services. You enable PowerExchange for Amazon Redshift with a license key.

Prerequisites

Before you can use PowerExchange for Amazon Redshift, perform the following tasks:

1. Install Informatica Big Data Management™.
2. Verify that you can connect to Amazon Redshift with an SQL client that uses the PostgreSQL driver. For example, you can use SQL Workbench/J to connect to Amazon Redshift.
3. Download the required Amazon Redshift JDBC jars from http://docs.aws.amazon.com/redshift/latest/mgmt/configure-jdbc-connection.html. Copy the jar files to the following location on the node that runs the Data Integration Service: `<Informatica installation directory>\connectors\thirdparty\informatica.amazonredshift\common`
4. Copy Amazon Redshift JDBC jars to the following location on the client machine: `<Informatica installation directory>\DeveloperClient\connectors\thirdparty\informatica.amazonredshift\common`
5. Copy Amazon Redshift JDBC jars to the following location on all Hadoop cluster nodes: `<Informatica installation directory>/connectors/thirdparty/informatica.amazonredshift/common>`

6. To run mappings on Hortonworks, Amazon EMR, and IBM BigInsights distributions that use non-kerberos authentication, configure user impersonation.

For more information about product requirements and supported platforms, see the Product Availability Matrix on Informatica Network:
https://network.informatica.com/community/informatica-network/product-availability-matrices/overview

Configure User Impersonation

Ensure that the Hadoop administrator creates a proxy user to impersonate other users if you run a mapping on Hortonworks, Amazon EMR, or IBM BigInsights distributions that use non-kerberos authentication.

The Hadoop administrator uses Apache Ambari to configure impersonation properties for Hortonworks HDP and IBM BigInsights.

Note: If Apache Ambari is not used, the Hadoop administrator configures configure the impersonation properties in core-site.xml on each node of the Hadoop cluster and restart Hadoop services along with the cluster.

The Hadoop administrator must configure the following user impersonation properties:

hadoop.proxyuser.yarn.groups

Comma-separated list of groups that you want to allow the YARN user to impersonate on a non-secure cluster.

Set to group names of impersonation users separated by commas. If less security is preferred, use the wildcard " * " to allow impersonation from any group.

hadoop.proxyuser.yarn.hosts

Comma-separated list of hosts that you want to allow the YARN user to impersonate on a non-secure cluster.

Set to a single host name or IP address, or set to a comma-separated list. If less security is preferred, use the wildcard " * " to allow impersonation from any host.

Using Apache Ambari to Configure User Impersonation

To enable user impersonation, use Apache Ambari to add or update the `hadoop.proxyuser.yarn.groups` and `hadoop.proxyuser.yarn.hosts` properties. If the properties are already added, you must change the value for the properties to the appropriate value. For example: *

1. Log on to Apache Ambari.
2. Click **HDFS service** > **Configs** > **Advanced**.
3. Navigate to the custom core-site section.
4. Click **Add Property**.
5. Add the following properties with the appropriate value:
   - `hadoop.proxyuser.yarn.groups`
   - `hadoop.proxyuser.yarn.hosts`
6. Save and restart the Hadoop services and the Hadoop cluster.

Configuring MapR Secure Clusters

Before you run an Amazon Redshift mapping, perform the following steps on all the nodes of the MapR secure clusters:

1. Take backup of ssl_truststore from the following location: `<mapR_Conf>/ssl_truststore`
2. Run the following command on all the nodes of the MapR cluster:

```
keytool -importkeystore -srckeystore <JDK_HOME>/jre/lib/security/cacerts -destkeystore
<mapR_Conf>/ssl_truststore -deststorepass mapr123 -v
```

Configuring Cloudera Clusters

You can configure additional functionality through the configuration files the in Hadoop environment.

When you use Cloudera Manager to update properties, updates are propagated to each node in the cluster. After you make updates, you can restart the Hadoop cluster and services through Cloudera Manager.

Update mapred-site.xml

Update the `mapred-site.xml` file on the Hadoop environment to configure functionality for Amazon Redshift connectivity.

To enable Amazon Redshift connectivity for Cloudera clusters, configure MapReduce properties in the **mapred-site.xml** file on the Hadoop cluster, and restart Hadoop services and the cluster.

1. Open the Yarn Configuration in Cloudera Manager.
2. Find the property named **NodeManager Advanced Configuration Snippet (Safety Valve) for mapred-site.xml**.
3. Click **+** and configure the following properties:

Property	Value

mapreduce.application.classpath	`$HADOOP_MAPRED_HOME/*,`
	`$HADOOP_MAPRED_HOME/lib/*,`
	`$MR2_CLASSPATH,`
	`$CDH_MR2_HOME`

mapreduce.jobhistory.intermediate-done-dir	`<Directory where the map-reduce jobs write history files>`
mapreduce.jobhistory.address	`<MapReduce JobHistory Server hostname>:<port>` **Note:** Configure this property if you want to run Amazon Redshift mappings on Cloudera Kerberos clusters that are not enabled with NameNode high availability.

4. Select the **Final** check box.
5. Redeploy the client configurations.
6. Restart Hadoop services and the cluster.

Update yarn-site.xml

Update the `yarn-site.xml` file on the domain to configure functionality such as Amazon Redshift for Kerberos authentication.

To run Amazon Redshift mappings on Cloudera Kerberos clusters that are not enabled with high availability, you must configure properties in the `yarn-site.xml` file on the Data Integration Service node and restart the Data Integration Service.

After you configure `mapred-site.xml` on the Hadoop environment, copy the following properties to the `yarn-site.xml` file on the Data Integration Service node:

mapreduce.jobhistory.address

Location of the MapReduce JobHistory Server. The default port is 10020.

```
<property>
<name>mapreduce.jobhistory.address</name>
<value><host name>:port</value>
<description>MapReduce JobHistory Server IPC host:port</description>
</property>
```

mapreduce.jobhistory.principal

SPN for the MapReduce JobHistory server.

```
<property>
<name>mapreduce.jobhistory.principal</name>
<value>mapred/_HOST@YOUR-REALM</value>
<description>SPN for the MapReduce JobHistory server</description>
</property>
```

mapreduce.jobhistory.webapp.address

Web address of the MapReduce JobHistory Server. The default value is 19888.

```
<property>
<name>mapreduce.jobhistory.webapp.address</name>
<value>hostname:port</value>
<description>MapReduce JobHistory Server Web UI host:port</description>
</property>
```

Create Minimal Amazon S3 Bucket Policy

The minimal Amazon S3 bucket policy ensures PowerExchange for Amazon Redshift performs read and write operations successfully.

You can restrict user operations and user access to particular Amazon S3 buckets by assigning an AWS IAM policy to users. Configure the AWS IAM policy through the AWS console. Following are the minimum required permissions for users to successfully read data from and write data to Amazon Redshift resources.

- PutObject
- GetObject
- GetObjectVersion
- DeleteObject
- DeleteObjectVersion
- ListBucket

Sample Policy:

```
{
"Version": "2012-10-17", "Statement": [
{ "Effect": "Allow", "Action": [ "s3:PutObject", "s3:GetObject", "s3:GetObjectVersion",
"s3:DeleteObject", "s3:DeleteObjectVersion", "s3:ListBucket" ], "Resource":
[ "arn:aws:s3:::<specify_bucket_name>", "arn:aws:s3:::<specify_bucket_name>/*" ] }
]
}
```

CHAPTER 3

Amazon Redshift Connections

This chapter includes the following topics:

Amazon Redshift Connection Overview

Amazon Redshift connection enables you to read data from or write data to Amazon Redshift.

You can use Amazon Redshift connections to create data objects and run mappings. The Developer tool uses the connection when you create a data object. The Data Integration Service uses the connection when you run mappings.

You can create an Amazon Redshift connection from the Developer tool or the Administrator tool. The Developer tool stores connections in the domain configuration repository. Create and manage connections in the connection preferences.

Amazon Redshift Connection Properties

When you set up an Amazon Redshift connection, you must configure the connection properties.

The following table describes the Amazon Redshift connection properties:

Property	Description
Name	The name of the connection. The name is not case sensitive and must be unique within the domain. You can change this property after you create the connection. The name cannot exceed 128 characters, contain spaces, or contain the following special characters:~ ` ! $ % ^ & * ( ) - + = { [}] | \ : ; " ' < , > . ? /
ID	String that the Data Integration Service uses to identify the connection. The ID is not case sensitive. It must be 255 characters or less and must be unique in the domain. You cannot change this property after you create the connection. Default value is the connection name.
Description	The description of the connection. The description cannot exceed 4,000 characters.

Property	Description
Location	The domain where you want to create the connection.
Type	The connection type. Select Amazon Redshift in the Database.

The **Details** tab contains the connection attributes of the Amazon Redshift connection. The following table describes the connection attributes:

Property	Description
Username	User name of the Amazon Redshift account.
Password	Password for the Amazon Redshift account.
Schema	Optional. Amazon Redshift schema name. Do not specify the schema name if you want to use multiple schema. The Data Object wizard displays all the user-defined schemas available for the Amazon Redshift objects. Default is public.
AWS Access Key ID	Amazon S3 bucket access key ID.
AWS Secret Access Key	Amazon S3 bucket secret access key ID.
Master Symmetric Key	Optional. Provide a 256-bit AES encryption key in the Base64 format when you enable client-side encryption. You can generate a key using a third-party tool. If you specify a value, ensure that you specify the encryption type as client side encryption in the advanced target properties.
Cluster Node Type	Node type of the Amazon Redshift cluster. You can select the following options: - ds1.xlarge - ds1.8xlarge - dc1.large - dc1.8xlarge - ds2.xlarge - ds2.8xlarge For more information about nodes in the cluster, see the Amazon Redshift documentation.
Number of Nodes in Cluster	Number of nodes in the Amazon Redshift cluster. For more information about nodes in the cluster, see the Amazon Redshift documentation.
JDBC URL	Amazon Redshift connection URL.

Note: If you upgrade the mappings created in earlier versions, you must select the relevant schema in the connection property. Else, the mappings fail when you run them on current version.

Creating an Amazon Redshift Connection

Create an Amazon Redshift connection before you create an Amazon Redshift data object.

1. In the Developer tool, click **Window** > **Preferences**.
2. Select **Informatica** > **Connections**.
3. Expand the domain in the **Available Connections**.
4. Select the connection type **Database** > **Amazon Redshift**, and click **Add**.
5. Enter a connection name and an optional description.
6. Select Amazon Redshift as the connection type.
7. Click **Next**.
8. Configure the connection properties.
9. Click **Test Connection** to verify the connection to Amazon Redshift.
10. Click **Finish**.

CHAPTER 4

PowerExchange for Amazon Redshift Data Objects

This chapter includes the following topics:

Amazon Redshift Data Object Overview

An Amazon Redshift data object is a physical data object that uses Amazon Redshift as a source or target. An Amazon Redshift data object represents the data in an Amazon Redshift data source.

You can configure the data object read and write operation properties that determine how data can be read from Amazon sources and loaded to Amazon Redshift targets. You first create a connection to create an Amazon Redshift data object. Create a data object operation for the Amazon Redshift data object. Then, you can add the data object read or write operation to a mapping.

Amazon Redshift Data Object Properties

Specify the data object properties when you create the data object.

The following table describes the properties that you configure for the Amazon Redshift data objects:

Property	Description
Name	Name of the Amazon Redshift data object.
Location	The project or folder in the Model Repository where you want to store the Amazon Redshift data object.
Connection	Name of the Amazon Redshift connection.

Amazon Redshift Data Object Read Operation

Create a mapping with an Amazon Redshift data object read operation to read data from Amazon Redshift.

You can encrypt data, specify the location of the staging directory, and securely unload the results to Amazon Redshift.

Amazon Redshift Staging Directory for Amazon Redshift Sources

The Data Integration Service creates a staging file in the directory that you specify in the source properties. The Data Integration Service reads the data from Amazon Redshift source and writes the data to the staging directory before it writes data to Amazon S3.

The Data Integration Service deletes the staged files from the staging directory after it writes the data to Amazon S3. Specify a staging directory in the mapping properties with an appropriate amount of disk space for the volume of data that you want to process. Specify a directory on the machine that hosts the Data Integration Service.

The Data Integration Service creates subdirectories in the staging directory. Subdirectories use the following naming convention:

```
<staging directory>/infaRedShiftStaging<MMddHHmmssSSS+xyz>
```

Client-side Encryption for Amazon Redshift Sources

Client-side encryption is a technique to encrypt data before transmitting the data to the Amazon Redshift server.

When you enable client-side encryption for Amazon Redshift sources, Amazon Redshift unloads the data in encrypted format, and then pushes the data to the Data Integration Service. The Data Integration Service writes the data to the target based on the mapping logic.

To enable client-side encryption, you must provide a master symmetric key in the connection properties. The Data Integration Service encrypts the data by using the master symmetric key. The master symmetric key is a 256-bit AES encryption key in the Base64 format. PowerExchange for Amazon Redshift uploads the data to the Amazon S3 server by using the master symmetric key and then loads the data by using the copy command with the Encrypted option and a private encryption key for additional security.

To enable client-side encryption, perform the following tasks:

1. Provide the master symmetric key when you create an Amazon Redshift connection. Ensure that you provide a 256-bit AES encryption key in Base64 format.

2. Update the `local_policy.jar` and the `US_export_policy.jar` files in the following directory:`<JAVA_HOME>\lib\security`. You can download the JAR files supported by your JAVA environment from the Oracle website.
3. Select **S3 Client Encryption** in the Data Object Operation Advanced properties.

Unload Command

You can use the Unload command to extract data from Amazon Redshift and create staging files on Amazon S3. The Unload command uses a secure connection to load data into one or more files on Amazon S3.

You can specify the Unload command options directly in the **Unload Options** field. Enter the options in uppercase in a separate line. For example:

```
DELIMITER = |
PARALLEL = ON
```

You can create a property file. The property file contains the Unload command options. Include the property file path in the **Unload Options** field. For example:

```
C:\Temp\Redshift\unloadoptions.txt
```

Unload Command Options

The Unload command options extract data from Amazon Redshift and load data to staging files on Amazon S3 in a particular format. You can delimit the data with a particular character or load data to multiple files in parallel.

To add options to the Unload command, use the **Unload Options** option. The ESCAPE option is set by default. You can set the following options:

DELIMITER

A single ASCII character to separate fields in the input file. You can use characters such as pipe (|), tilde (~), or a tab (\t). The delimiter you specify should not be a part of the data. If the delimiter is a part of data, use ESCAPE to read the delimiter character as a regular character.

PARALLEL

The Unload command writes data in parallel to multiple files, according to the number of slices in the cluster. Default is On. If you turn the Parallel option off, the Unload command writes data serially. The maximum size of a data file is 6.5 GB.

Do not use `PARALLEL OFF` if you run a mapping in the Blaze engine.

Amazon Redshift Data Object Read Operation Properties

The Data Integration Service reads data from Amazon Redshift based on the data object read operation.

The Developer tool displays advanced properties for the Amazon Redshift data object operation in the Advanced view.

The following table describes the Advanced properties for an Amazon Redshift data object read operation:

Property	Description
S3 Bucket Name	Amazon S3 bucket name for the Amazon Redshift source data. You can also specify the bucket name with the folder path. Use an S3 bucket in the same region as your Amazon Redshift cluster.
Enable Compression	Compresses staging files before writing the files to Amazon Redshift. Task performance improves when the Data Integration Service compresses the staging files. Not applicable if you run a mapping in the Blaze engine.
Staging Directory Location	Amazon Redshift staging directory. Specify a directory on the machine that hosts the Data Integration Service. Not applicable if you run a mapping in the Blaze engine.

Unload Options	Unload command options.
	Add options to the Unload command to write data from an Amazon Redshift object to an S3 bucket. You can add the following options:
	- DELIMITER
	- PARALLEL
	For example:
	`DELIMITER = |`
	`PARALLEL = ON`
	Specify a directory on the machine that hosts the Data Integration Service.
	Do not use `PARALLEL OFF` if you run a mapping in the Blaze engine.

S3 Client Encryption	Indicates that the Data Integration Service encrypts data by using a private encryption key. Not applicable if you run a mapping in the Blaze engine.
Download S3 Files in Multiple Parts	Downloads large Amazon S3 objects in multiple parts. When the file size of an Amazon S3 object is greater than 8 MB, you can choose to download the object in multiple parts in parallel. Not applicable if you run a mapping in the Blaze engine.

Amazon Redshift Data Object Write Operation

Create a mapping to move data to an Amazon Redshift table. Change the connection to an Amazon Redshift connection, and define the write operation properties to write data to Amazon Redshift.

You can perform insert, update, delete, and upsert operations on an Amazon Redshift target.

Amazon Redshift Staging Directory for Amazon Redshift Targets

The Data Integration Service creates a staging file in the directory that you specify in the target properties. The Data Integration Service writes the data to the staging directory before it writes data to Amazon Redshift.

The Data Integration Service deletes the staged files from the staging directory after it writes the data to Amazon S3. Specify a staging directory in the mapping properties with an appropriate amount of disk space for the volume of data that you want to process. Specify a directory on the machine that hosts the Data Integration Service.

The Data Integration Service creates subdirectories in the staging directory. Subdirectories use the following naming convention:

```
<staging directory>/infaRedShiftStaging<MMddHHmmssSSS+xyz>
```

Analyze Target Table

To optimize query performance, you can configure a mapping to analyze the target table. Target table analysis updates statistical metadata of the database tables.

You can use the Analyze Target Table option to extract sample rows from the table, analyze the samples, and save the column statistics. Amazon Redshift then updates the query planner with the statistical metadata. The query planner uses the statistical metadata to build and choose optimal plans to improve the efficiency of queries.

You can run the Analyze Target Table option after you load data to an existing table by using the Copy command. If you load data to a new table, the Copy command performs an analysis by default.

Data Encryption in Amazon Redshift Targets

To protect data, you can enable server-side encryption or client-side encryption to encrypt the data that you insert in Amazon Redshift.

If you enable both server-side and client-side encryption for an Amazon Redshift target, then the client-side encryption is used for data load.

Server-side Encryption for Amazon Redshift Targets

If you want Amazon Redshift to encrypt data while uploading the .csv files to Amazon Redshift, you must enable server-side encryption. To enable server-side encryption, select S3 Server Side Encryption in the Data Object Operation Advanced properties.

Client-side Encryption for Amazon Redshift Targets

Client-side encryption is a technique to encrypt data before transmitting the data to the Amazon Redshift server.

When you enable client-side encryption for Amazon Redshift targets, the Data Integration Service fetches the data from the source, writes the data to the staging directory, encrypts the data, and then writes the data to an Amazon S3 bucket. The Amazon S3 bucket then writes the data to Amazon Redshift.

If you enable both server-side and client-side encryption for an Amazon Redshift target, then the client-side encryption is used for data load.

To enable client-side encryption, you must provide a master symmetric key in the connection properties. The Data Integration Service encrypts the data by using the master symmetric key. The master symmetric key is a 256-bit AES encryption key in the Base64 format. PowerExchange for Amazon Redshift uploads the data to the Amazon S3 server by using the master symmetric key and then loads the data to Amazon Redshift by using the copy command with the Encrypted option and a private encryption key for additional security. To enable client-side encryption, perform the following tasks:

1. Provide the master symmetric key when you create an Amazon Redshift connection. Ensure that you provide a 256-bit AES encryption key in Base64 format.
2. Download the `local_policy.jar` and the `US_export_policy.jar` files for your JAVA environment from the Oracle website. Replace the existing `local_policy.jar` and the `US_export_policy.jar` files in the following directory: `<JAVA_HOME>\lib\security`.

3. Select **S3 Client Side Encryption** in the Data Object Operation Advanced properties.

Retain Staging Files

You can retain staging files on Amazon S3 after the Data Integration Service writes data to the target. You can retain files to create a data lake of your organizational data on Amazon S3. The files you retain can also serve as a backup of your data.

When you configure the write operation, you can provide a file prefix or directory prefix to save the staging files. After you provide the prefixes, the Data Integration Service creates files within the directories at Amazon S3 location specified in the target connection. Configure one of the following options for the **Prefix for Retaining Staging Files on S3** property:

- Provide a directory prefix and a file prefix. For example, backup_dir/backup_file. The Data Integration Service creates the following directories and files:

  - backup_dir_<year>_<month>_<date>_<timestamp_inLong>

  - backup_file.batch_<batch_number>.csv.<file_number>.<encryption_if_applicable>

- Provide a file prefix. For example, backup_file. The Data Integration Service creates the following directories and files:

  - <year>_<month>_<date>_<timestamp_inLong>

  - backup_file.batch_<batch_number>.csv.<file_number>.<encryption_if_applicable>

- Do not provide a prefix. The Data Integration Service does not save the staging files.

If you run a mapping in the Blaze mode, you must only provide a directory prefix. The file prefix is ignored. The file creation is dynamic in the Blaze engine.

Copy Command

You can use the Copy command to append data in a table. The Copy command uses a secure connection to load data from flat files in an Amazon S3 bucket to Amazon Redshift.

You can specify the Copy command options directly in the **Copy Options** field. Enter the options in uppercase in separate lines. For example:

```
DELIMITER = |
ACCEPTINVCHARS = #
QUOTE = '
COMPUPDATE = ON
```

You can create a property file. The property file contains the Copy command options. Include the property file path in the **Copy Options** field. For example:

```
C:\Temp\Redshift\copyoptions.txt
```

Copy Command Options

The Copy command options read data from Amazon S3 and write data to Amazon Redshift in a particular format. You can apply compression to data in the tables or delimit the data with a particular character.

To add options to the Copy command, use the **Copy Options** option. You can set the following options:

DELIMITER

A single ASCII character to separate fields in the input file. You can use characters such as pipe (|), tilde (~), or a tab (\t). The delimiter must not be a part of the data.

ACCEPTINVCHARS

Loads data into VARCHAR columns even if the data contains UTF-8 characters that are not valid. When you specify ACCEPTINCHARS, the Data Integration Service replaces UTF-8 character that is not valid with an equal length string consisting of the character specified in ACCEPTINVCHARS. If you have specified '|' in ACCEPTINVCHARS, the Data Integration Service replaces the three-byte UTF-8 character with '|||'.

If you do not specify ACCEPTINVCHARS, the COPY command returns an error when it encounters an UTF-8 character that is not valid. You can use the ACCEPTINVCHARS option on VARCHAR columns.

QUOTE

Specifies the quote character to use with comma separated values. Default is a double quote (").

COMPUPDATE

Overrides current compression encoding and applies compression to an empty table. Use the COMPUPDATE option in an insert task when the rows in a table are more than 100,000. The behavior of COMPUPDATE depends on how it is configured:

- If you do not specify COMPUPDATE, the COPY command applies compression if the target table is empty and all columns in the table have either RAW or no encoding.
- If you specify COMPUPDATE ON, the COPY command replaces the existing encodings if the target table is empty and the columns in the table have encodings other than RAW.
- If you specify COMPUPDATE OFF, the COPY command does not apply compression.

Vacuum Tables

You can use vacuum tables to recover disk space and sorts rows in a specified table or all tables in the database.

After you run bulk operations, such as delete or load, or after you run incremental updates, you must clean the database tables to recover disk space and to improve query performance on Amazon Redshift. Amazon Redshift does not reclaim and reuse free space when you delete and update rows.

Vacuum databases or tables often to maintain consistent query performance. You can recover disk space for the entire database or for individual tables in a database. You must run vacuum when you expect minimal activity on the database or during designated database administration schedules. Long durations of vacuum might impact database operations. Run vacuum often because large unsorted regions result in longer vacuum times.

You can enable the vacuum tables option when you configure the advanced target properties. You can select the following recovery options:

None

Does not sort rows or recover disk space.

Full

Sorts the specified table or all tables in the database and recovers disk space occupied by rows marked for deletion by previous update and delete operations.

Sort Only

Sorts the specified table or all tables in the database without recovering space freed by deleted rows.

Delete Only

Recovers disk space occupied by rows marked for deletion by previous update and delete operations, and compresses the table to free up used space.

Amazon Redshift Data Object Write Operation Properties

Amazon Redshift data object write operation properties include run-time properties that apply to the Amazon Redshift data object.

The Developer tool displays advanced properties for the Amazon Redshift data object operation in the Advanced view. The following table describes the Advanced properties for an Amazon Redshift data object write operation:

Property	Description
S3 Bucket Name	Amazon S3 bucket name for the Amazon Redshift target data. You can also specify the bucket name with the folder path. Use an S3 bucket in the same region as your Amazon Redshift cluster.
Enable Compression	Compresses staged files before writing the files to Amazon Redshift. Mapping performance improves when the Data Integration Service compresses the staged files. Not applicable if you run a mapping in the Blaze engine.
Staging Directory Location	Amazon Redshift staging directory. Specify a directory on the machine that hosts the Data Integration Service. Not applicable if you run a mapping in the Blaze engine.
Batch Size	Minimum number of rows in a batch. Enter a number greater than 0. Default is 2000000. Not applicable if you run a mapping in the Blaze engine.
Max Errors per Upload Batch for INSERT	Number of errors within a batch that causes a batch to fail. Enter a positive integer. If the number of errors is equal to or greater than the property value, the Data Integration Service writes the entire batch to the error file. Default is 0.
Truncate Target Table Before Data Load	Truncates an Amazon Redshift target before writing data to the target.
Null value for CHAR and VARCHAR data types	A string value that you want to replace as NULL when data is uploaded to Amazon Redshift. Default is an empty string.
Wait time in seconds for file consistency on S3	Number of seconds to wait for the Data Integration Service to make the staged files consistent with the list of files available on Amazon S3. Default is 0.

Property	Description

Copy Options	Name of the property file.
	Enables you to add additional options to the copy command for writing data from an Amazon S3 source to an Amazon Redshift target when the default delimiter comma (,) or double-quote (") is used in the data.
	You can add the following options:
	- DELIMITER
	- ACCEPTINVCHARS
	- QUOTE
	- COMPUPDATE
	For example:
	`DELIMITER = |`
	`ACCEPTINVCHARS = #`
	`QUOTE = '`
	`COMPUPDATE = ON`
	Specify a directory on the machine that hosts the Data Integration Service.

Turn on S3 Server Side Encryption	Indicates that Amazon S3 encrypts data during upload and decrypts data at the time of access.
Turn on S3 Client Side Encryption	Indicates that the Data Integration Service encrypts data by using a private encryption key. If you enable both server side and client side encryption, the Data Integration Service ignores the server side encryption. Not applicable if you run a mapping in the Blaze engine.
Analyze Target Table	Improve the efficiency of the write operations. The query planner on Amazon Redshift updates the statistical metadata to build and choose optimal plans to improve the efficiency of queries.
Vacuum Target Table	Recovers disk space and sorts rows in a specified table or all tables in the database. You can select the following recovery options: **None** Does not sort rows or recover disk space. **Full** Sorts the specified table or all tables in the database and recovers disk space occupied by rows marked for deletion by previous update and delete operations. **Sort Only** Sorts the specified table or all tables in the database without recovering space freed by deleted rows. **Delete Only** Recovers disk space occupied by rows marked for deletion by previous update and delete operations, and compresses the table to free up used space. Default is None.

Property	Description
Prefix to retain staging files on S3	Retains staging files on Amazon S3. Provide both a directory prefix and a file prefix separated by a slash (/) or only a file prefix to retain staging files on Amazon S3. For example, `backup_dir/backup_file` or `backup_file`. If you run a mapping in the Blaze engine, the file prefix is not retained because the file creation is dynamic in the Blaze engine. You must specify only a directory prefix. For example, `backup_dir`.
Success File Directory	Directory for the Amazon Redshift success file. Specify a directory on the machine that hosts the Data Integration Service. Not applicable if you run a mapping in the Blaze engine.
Error File Directory	Directory for the Amazon Redshift error file. Specify a directory on the machine that hosts the Data Integration Service. Not applicable if you run a mapping in the Blaze engine.
Treat Source Rows As	Select one of the following options: **INSERT** If enabled, the Data Integration Service inserts all rows flagged for insert. If disabled, the Data Integration Service rejects the rows flagged for insert. By default, the insert operation is enabled. **DELETE** If enabled, the Data Integration Service deletes all rows flagged for delete. If disabled, the Data Integration Service rejects all rows flagged for delete. **UPDATE and UPSERT** Performs update and upsert operations. To perform an update operation, you must map the primary key column and at least one column other than primary key column. You can select the following data object operation attributes: - Update as Update: The Data Integration Service updates all rows as updates. - Update else Insert: The Data Integration Service updates existing rows and inserts other rows as if marked for insert.

Creating an Amazon Redshift Data Object

Create an Amazon Redshift data object to add to a mapping.

1. Select a project or folder in the **Object Explorer** view.
2. Click **File** > **New** > **Data Object**.
3. Select **AmazonRedshift Data Object** and click **Next**.

   The **AmazonRedshift Data Object** dialog box appears.
4. Enter a name for the data object.
5. Click **Browse** next to the **Location** option and select the target project or folder.

6. Click **Browse** next to the **Connection** option and select the Amazon Redshift connection from which you want to import the Amazon Redshift object.
7. To add a resource, click **Add** next to the **Selected Resources** option.

   The **Add Resource** dialog box appears.
8. Select the checkbox next to the Amazon Redshift object you want to add and click **OK**.
9. Click **Finish**.
   The data object appears under Data Objects in the project or folder in the **Object Explorer** view.

Creating a Data Object Operation

You can create the data object read, write, or lookup operation for an Amazon Redshift data objects. You can the add the Amazon Redshift data object operation to a mapping.

1. Select the data object in the **Object Explorer** view.
2. Right-click and select **New** > **Data Object Operation**.

   The **Data Object Operation** dialog box appears.
3. Enter a name for the data object operation.
4. Select the type of data object operation. You can choose to create a read or write operation.
5. Click **Add**.

   The **Select Resources** dialog box appears.
6. Select the Amazon Redshift data object for which you want to create the data object operation and click **OK**.
7. Click **Finish**.

The Developer tool creates the data object operation for the selected data object.

Success and Error Files

The success file contain an entry for each record that successfully writes into Amazon Redshift. The error file contains an entry for each data error.

The Data Integration Service generates success and error files after you run a mapping. Success and error files are .csv files that contain row-level details. The Data Integration Service does not overwrite success or error files. Access the error rows files and success rows files directly from the directories where they are generated. You can manually delete the files that you no longer need.

Consider the following guidelines when you configure the data object operation properties for success files:

- Specify the Success File Directory in the data object operation properties. Specify a directory on the machine that hosts the Data Integration Service.
- The success rows file uses the following naming convention:
  `infa_rs_<operation>_<schema.table_name>.batch_<batch_number>_file_<file_number>_<timestamp>_success.csv`.

Consider the following guidelines when you configure the data object operation properties for error files:

- Specify the Error File Directory in the data object operation properties. Specify a directory on the machine that hosts the Data Integration Service.
- For insert tasks, the error rows file uses the following naming convention: `infa_rs_<operation>_<schema.table>.batch_<batch_number>_file_<file_number>_<timestamp>_error.csv`. For upsert tasks, the error rows file uses the following naming convention: `infa_rs_<operation>_<schema.table>_<timestamp_inLong>.batch_<batch_number>_file_<file_number>_<timestamp>_error.csv`.

Sample Error File

If a target table has the fields f_integer, f_char, and f_varchar, and if a row is rejected, the Data Integration Service generates an error file in the following format:

Errors Details	f_integer	f_char	f_varchar
"Query Start Time: 2014-03-24 11:41:30.629 Offending File: INSERT_bdt_with_composite_key.batch_0.csv.0.gz Line Number: 4 Column Name: f_char Column Type: char Offending Value: .....Furniture Values Intl LLC_upd_upd ERROR Reason: Multibyte character not supported for CHAR (Hint: try using VARCHAR). Invalid char: c3 a6"	"3"	""æˆ'Furniture Values Intl LLC_upd_upd""	""001E000000Sl3jIIAT""
"Query Start Time: 2014-03-24 11:42:00.763 Offending File: INSERT_bdt_with_composite_key.batch_8.csv.0.gz Line Number: 80 Column Name: f_char Column Type: char Offending Value: .....Heitkamp Inc_upd_upd ERROR Reason: Multibyte character not supported for CHAR (Hint: try using VARCHAR). Invalid char: c3 a6"	"9999"	""æˆ'Heitkamp Inc_upd_upd""	""001E000000SHd7ZIAT""

CHAPTER 5

Amazon Redshift Mappings

This chapter includes the following topics:

Amazon Redshift Mapping Overview

After you create the Amazon Redshift data object with a Amazon Redshift connection, you can develop a mapping. You can define the following types of objects in the mapping:

- A Read transformation of the Amazon Redshift data object to read data from Amazon Redshift in native or Hadoop run-time environment.
- A Write transformation of the Amazon Redshift data object to write data to Amazon Redshift in native or Hadoop run-time environment.

Validate and run the mapping. You can deploy the mapping and run it or add the mapping to a Mapping task in a workflow.

Mapping Validation and Run-time Environments

You can validate and run mappings in the native environment or a Hadoop environment.

You can validate a mapping in the native environment, Hadoop environment, or both. The Data Integration Service validates whether the mapping can run in the selected environment. You must validate the mapping for an environment before you run the mapping in that environment.

When you run a mapping in the native environment, the Data Integration Service runs the mapping from the Developer tool.

When you run a mapping in a Hadoop environment, the Data Integration Service pushes the mapping to a Hadoop cluster and processes the mapping on the Blaze or the Hive engine.

Amazon Redshift Mapping Example

Your organization has a large amount of customers data from for all regions in flat files. You organization needs to analyze customer data in the US region in a short span of time. Create a mapping that reads all the customer records and write the records to Amazon Redshift table.

You can use the following objects in a Amazon Redshift mapping:

Flat file input

The input file is a flat file that contains the customer names and other details about customers.

Create a flat file data object. Configure the flat file connection and specify the flat file that contains the customer data as a resource for the data object. Drag the data object into a mapping as a read data object.

Transformations

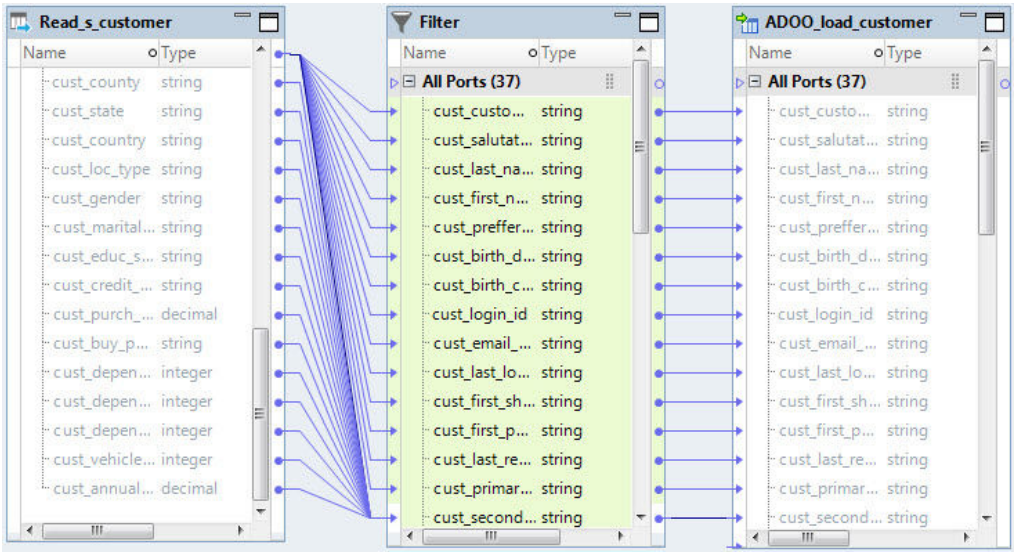
Add Filter transformation to get customer data in a particular region.

The Filter transformation filters the source data based on the value you specify for the region ID column. The Data Integration Service returns the rows that meet the filter condition.

Amazon Redshift output

Create an Amazon Redshift data object write operation. Configure the Amazon Redshift connection and specify the Amazon Redshift object as a target for the data object. Drag the data object into a mapping as a target data object.

The following image shows the Amazon Redshift mapping example:



When you run the mapping, the customer records are read from the flat file and written to the Amazon Redshift table.

APPENDIX A

Amazon Redshift Datatype Reference

This appendix includes the following topics:

Datatype Reference Overview

When you run the mapping to read data from or write data to Amazon Redshift, the Data Integration Service converts the transformation data types to comparable native Amazon Redshift data types.

Amazon Redshift and Transformation Datatypes

The Amazon Redshift data types are the names and the aliases represent how the Data Integration Service stores the data types.

For example, SMALLINT is the Amazon Redshift data type name. The data type is stored as a 2-byte integer. Here, SMALLINT is the Amazon Redshift data type name and INT2 is the Amazon Redshift data type alias.

The following table compares the Amazon Redshift data types and the transformation data types:

Amazon Redshift Data Type	Amazon Redshift Data Type Aliases	Description	Transformation Data Type
SMALLINT	INT2	Signed two-byte integer.	Small Integer
INTEGER	INT, INT4	Signed four-byte integer.	Integer
BIGINT	INT8	Signed eight-byte integer.	Bigint
DECIMAL	NUMERIC	Exact numeric of selectable precision.	Decimal

Amazon Redshift Data Type	Amazon Redshift Data Type Aliases	Description	Transformation Data Type
REAL	FLOAT4	Single precision floating-point number.	Double
DOUBLE PRECISION	FLOAT8, FLOAT	Double precision floating-point number.	Double
BOOLEAN	BOOL	Logical Boolean (true/false).	Small Integer
CHAR	CHARACTER, NCHAR, BPCHAR	Fixed-length character string.	String
VARCHAR	CHARACTER VARYING, NVARCHAR, TEXT	Variable-length character string with a user-defined limit.	String
DATE	NA	Calendar date (year, month, day).	Timestamp
TIMESTAMP	TIMESTAMP WITHOUT TIME ZONE	Date and time (without time zone).	Timestamp

INDEX