Informatica

Informatica® PowerExchange for Amazon Redshift

10.1.1

User Guide for PowerCenter

Informatica PowerExchange for Amazon Redshift User Guide for PowerCenter
10.1.1
December 2016

Publication Date: 2018-09-27

Table of Contents

Preface

The *Informatica PowerExchange® for Amazon Redshift User Guide for PowerCenter®* describes how to read data from and write data to an Amazon Redshift target. The guide is written for database administrators and developers who are responsible for moving data from a source to an Amazon Redshift target, and from an Amazon Redshift source to a target. This guide assumes that you have knowledge of database engines, Amazon Redshift, and PowerCenter.

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.

You can also read data from or write data to the Amazon Redshift cluster that reside in a Virtual Private Cloud (VPC).

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 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 configure HTTPS proxy to connect to Amazon Redshift. You can also configure an SSL connection to connect to Amazon Redshift. The PowerCenter Integration Service uses the Amazon driver to communicate with Amazon Redshift.

Note: PowerExchange for Amazon Redshift does not support real-time processing.

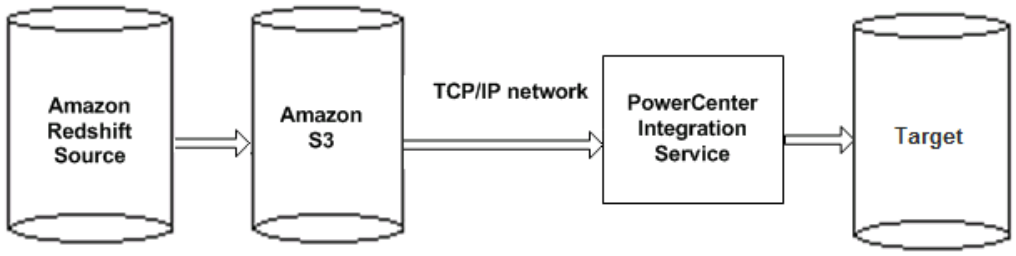
Example

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 an Amazon Redshift target for data analysis.

PowerCenter Integration Service and Amazon Redshift Integration

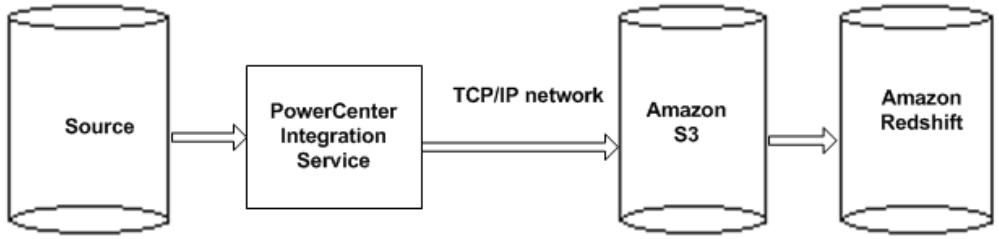
The PowerCenter Integration Service uses the Amazon Redshift connection to connect to Amazon Redshift.

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



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

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



When you run the Amazon Redshift session, the PowerCenter Integration Service writes data to Amazon Redshift based on the workflow and Amazon Redshift connection configuration. The PowerCenter Integration Service stores data in a staging directory on the PowerCenter machine. The PowerCenter 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 PowerCenter 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

You can use PowerExchange for Amazon Redshift on Windows or Linux. You must configure PowerExchange for Amazon Redshift before you can extract data from or load data to Amazon Redshift.

Prerequisites

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

1. Install or upgrade to PowerCenter 10.1.1.
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. Verify that you have read, write, and execute permissions on the following directories:
   `<PowerCenter Installation Directory>/server/bin`
4. Verify that you have read and write permissions on the following directories:
   `<PowerCenter Client Installation Directory>/Clients/PowerCenterClient/client/bin`

The organization administrator must also perform the following tasks:

- Mandatory. Get the Amazon Redshift JDBC URL.

- Mandatory. Manage Authentication. Use either of the following two methods:
  - Create an Access Key ID and Secret Access Key. Applicable when the PowerCenter client is not installed on Amazon Elastic Compute Cloud (EC2) system. Provide the values for Access Key ID and Secret Access Key when you configure the Amazon Redshift connection.
  - Configure AWS Identity and Access Management (IAM) Authentication. For enhanced security. Applicable when you install the PowerCenter client on Amazon EC2 system and you want to run tasks sessions on the EC2 system. If you use IAM authentication, do not provide Access Key ID and Secret Access Key explicitly in the Amazon Redshift connection. Instead, you must create an Redshift Role Amazon Resource Name (ARN), add the minimal Amazon S3 bucket policy to the Redshift Role ARN, and add the Redshift Role ARN to the Redshift cluster. For more information, see **IAM Authentication** section.
    Provide the Redshift Role ARN in the AWS_IAM_ROLE option in the UNLOAD and COPY commands when you create a session.

    If you specify both, Access Key ID and Secret Access Key in the connection properties and AWS_IAM_ROLE in the UNLOAD and COPY commands, AWS_IAM_ROLE takes the precedence.
- Optional. Create an Amazon Redshift master symmetric key to enable client-side encryption.

IAM Authentication (Optional)

You can configure IAM authentication when Secure Agent is installed the PowerCenter Integration Service runs on an Amazon Elastic Compute Cloud (EC2) system. Use IAM authentication for secure and controlled access to Amazon Redshift resources when you run Data Synchronization and Mapping Configuration tasks a session.

Use IAM authentication when you want to run the Data Synchronization and Mapping Configuration tasks on Secure agent installed a session on an EC2 system. Perform the following steps to configure IAM authentication:

- Step 1: Create Minimal Amazon S3 Bucket Policy. For more information, see "Create Minimal Amazon S3 Bucket Policy" on page 13.
- Step 2: Create the Amazon EC2 role. For more information, see "Create the Amazon EC2 Role" on page 13
- Step 3: Create an EC2 instance. Assign the Amazon EC2 role that you created in step #2 to the EC2 instance.
- Step 4: Create the Amazon Redshift Role ARN. For more information, see "Create the Amazon Redshift Role" on page 14
- Step 5: Add the Amazon Redshift Role ARN to the Amazon Redshift cluster. For more information, see "Add Amazon Redshift Role to the Redshift Cluster" on page 14
- Step 6: Install Secure Agent the PowerCenter Integration Service on the EC2 system.

Create Minimal Amazon S3 Bucket Policy

The minimal Amazon S3 bucket policy ensures PowerExchange for Amazon Redshift Connector 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
- GetBucketPolicy

Sample Policy:

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

Note: The **Test Connection** does not validate the IAM policy assigned to users. The Amazon S3 bucket name is available in the advanced properties for source and target.

You must make sure that the Amazon S3 bucket and Amazon Redshift cluster reside in the same region to run the Data Synchronization and Mapping configuration tasks a session successfully.

Create the Amazon EC2 Role

The Amazon EC2 role is used when you create an EC2 system in the Redshift cluster.

Perform the following steps to create the Amazon EC2 role:

1. Log in to the AWS Console.
2. Click Identity & Access Management.
3. Select Role under the Details menu and click **Create New Role**.
4. Specify the name of the role in the Set Role Name page.
5. Click **Next Step**.
6. Select the Amazon EC2 role type in the Select Role Type page.
7. Select the required Amazon S3 policy in the Attach Policy page.
8. Click **Next Step**.
9. Review the Role Name, Role ARN, Trusted Entities, and Policies values in the Review page.
10. Click **Create Role**.

After creating the Amazon EC2 role, create an EC2 instance. Assign the Amazon EC2 role to the EC2 instance.

Create the Amazon Redshift Role

Use Amazon Redshift Role for secure access to Amazon Redshift resources.

Perform the following steps to create the Amazon Redshift Role:

1. Log in to the AWS Console.
2. Click Identity & Access Management.
3. Select Role under the Details menu and click **Create New Role**.
4. Specify the name of the role in the Set Role Name page.
5. Click **Next Step**.
6. Select the Amazon Redshift role type in the Select Role Type page.
7. Select the required Amazon S3 policy in the Attach Policy page.
8. Click **Next Step**.
9. Review the Role Name, Role ARN, Trusted Entities, and Policies values in the Review page.
10. Click **Create Role**.

After you create the Amazon Redshift Role, verify that you assign this role to the Amazon Redshift cluster to successfully perform the read and write operations.

You can use the Role ARN in the UNLOAD and COPY commands.

Add Amazon Redshift Role to the Redshift Cluster

Perform the following steps to add the Amazon Redshift Role to the Redshift cluster:

1. Log in to the AWS Console.
2. Click Redshift under the Database option.
3. Click Clusters under Dashboard and select your cluster.
4. Click **Manage IAM Roles**. The Manage IAM roles dialog box displays.
5. Select the required Amazon Redshift role.
6. Click **Apply changes**.

After you add the Amazon Redshift Role to the Redshift cluster, install Secure Agent the client on the EC2 instance.

Registering the Plug-in

After you install or upgrade to PowerCenter 10.1.1, you must register the plug-in with the PowerCenter repository.

A plug-in is an XML file that defines the functionality of PowerExchange for Amazon Redshift. To register the plug-in, the repository must be running in exclusive mode. Use the Administrator tool or the pmrep RegisterPlugin command to register the plug-in.

The plug-in file for PowerExchange for Amazon Redshift is `AmazonRSCloudAdapterPlugin.xml`. When you install PowerExchange for Amazon Redshift, the installer copies the `AmazonRSCloudAdapterPlugin.xml` file to the following directory:

```
<Informatica Installation Directory>\server\bin\Plugin
```

Note: If you do not have the correct privileges to register the plug-in, contact the user who manages the PowerCenter Repository Service.

CHAPTER 3

Amazon Redshift Sources and Targets

This chapter includes the following topics:

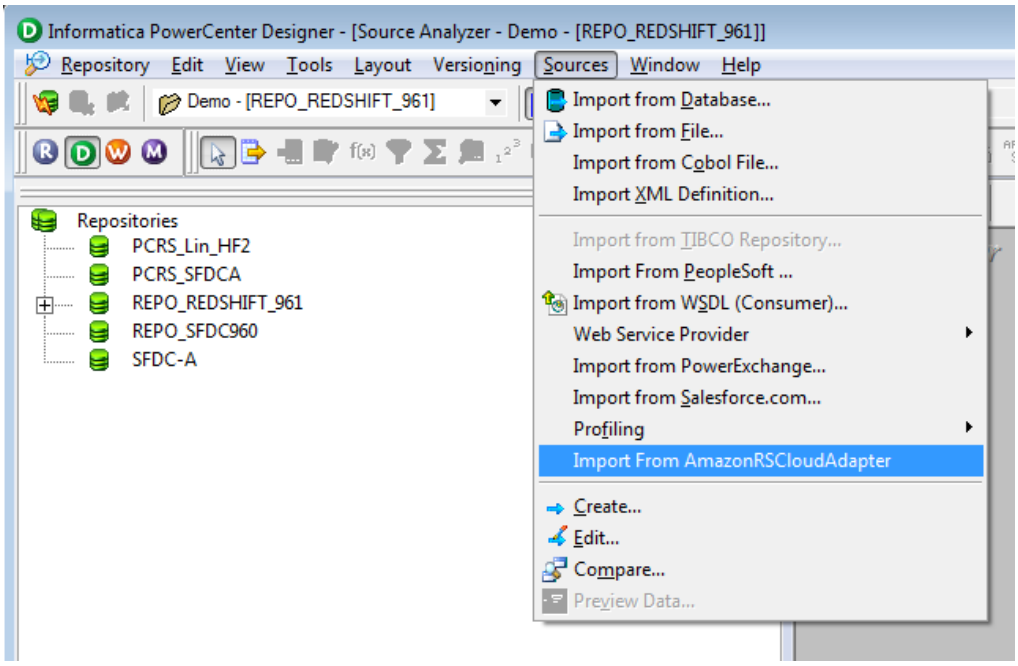
Amazon Redshift Sources and Targets

Create a mapping with an Amazon Redshift source to read data from Amazon Redshift. Create a mapping with any source and an Amazon Redshift target to write data to Amazon Redshift.

Import Amazon Redshift Objects

You can import Amazon Redshift source and target objects before you create a mapping.

1. Start PowerCenter Designer, and connect to a PowerCenter repository configured with an Amazon Redshift instance.
2. Open a source or target folder.
3. Select **Source Analyzer** or **Target Designer**.
4. Click **Sources** or **Targets**, and then click **Import from AmazonRSCloud Adapter**.



The **Establish Connection** dialog box appears.

5. Specify the following information and click **Connect**.

Connection Property	Description
Username	User name of the Amazon Redshift account.
Password	Password for the Amazon Redshift account.
Schema	Amazon Redshift schema name. 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. Amazon S3 encryption key. Provide a 256-bit AES encryption key in the Base64 format.
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.

Connection Property	Description
Number of Nodes in the 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. If you configure the Amazon Redshift cluster for SSL, you can specify the secure URL.

6. Click **Next**.
7. Select the table that you want to import, and then click **Finish**. If you want to see the table metadata, select the table, and click the table name.

Amazon Redshift Lookup Transformation

You can use the imported Amazon Redshift source in a lookup transformation.

CHAPTER 4

Amazon Redshift Mappings

This chapter includes the following topics:

Amazon Redshift Mappings Overview

After you import an Amazon Redshift source or target definition into the PowerCenter repository, you can create a mapping to read data from an Amazon Redshift source or write data to an Amazon Redshift target.

You can read data from a single Amazon Redshift source and write data to a multiple Amazon Redshift targets.

Configuring the Source Qualifier

When you import a source to create a mapping for Amazon Redshift source , you must configure the source qualifier to create the mapping.

1. In the mapping, click **Source Qualifier**
2. Select the **Configure** tab
3. Specify the Amazon Redshift connection details.
4. Save the mapping.

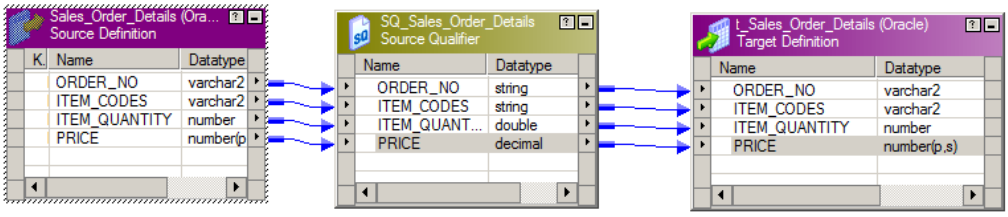
Amazon Redshift Mapping Example

Your e-commerce organization stores sales order details in an Oracle database. Your organization needs to move the data from the Oracle database to an Amazon Redshift target.

The following procedure shows how to move data from the Oracle database to Amazon Redshift:

1. Import the Oracle source.

2. Import an Amazon Redshift target.
3. Create a mapping with a source and an Amazon Redshift target.
   The following image shows the example mapping:



4. Create a session and configure it to write the data to the Amazon Redshift target.

The mapping contains the following objects:

Source Definition

The mapping source definition is a relational Oracle database. In the Source Analyzer, import the Oracle source. The PowerCenter Integration Service reads the sales order details from the Oracle source.

The following table describes the structure of the source definition called Sales_Order_Details:

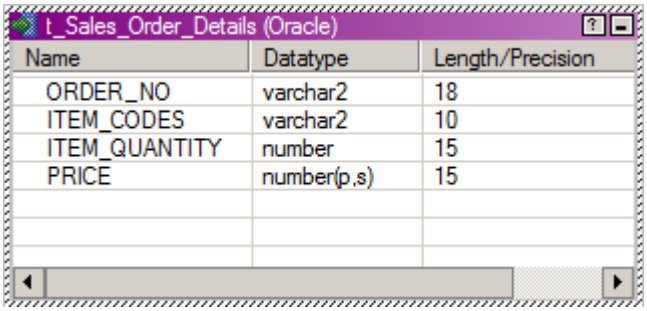
Field	Data Type
Order_No	Varchar
Item_Codes	Varchar
Item_Quantity	Number
Price	Number (p,s)

Mapping Target

The mapping contains an Amazon Redshift target definition.

In the Target Designer, import an Amazon Redshift target definition.

The following image shows the Amazon Redshift target definition t_Sales_Order_Details:



CHAPTER 5

Amazon Redshift Sessions

This chapter includes the following topics:

Amazon Redshift Sessions Overview

You must configure an Amazon Redshift connection in the Workflow Manager to read data from or write data to an Amazon Redshift table.

The PowerCenter Integration Service writes the data to a staging directory and then to an Amazon S3 bucket before it writes the data to Amazon Redshift. You must specify the location of the staging directory in the session properties. You must also specify an Amazon S3 bucket name in the session properties. You must have write access to the Amazon S3 bucket.

Amazon Redshift Connections

Use an Amazon Redshift connection to connect to the Amazon Redshift database. The PowerCenter Integration Service uses the connection when you run an Amazon Redshift session.

Amazon Redshift Connection Properties

When you configure an Amazon Redshift connection, you define the connection attributes that the PowerCenter Integration Service uses to connect to the Amazon Redshift database.

The following table describes the application connection properties:

Property	Description
Name	Name of the Amazon Redshift connection.
Type	The AmazonRSCloudAdapter connection type.
User Name	User name to access the Amazon Redshift database.
Password	Password for the Amazon Redshift database user name.

The following table describes the Amazon Redshift connection attributes:

Property	Description
Schema	Amazon Redshift schema name. 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. Amazon S3 encryption key. Provide a 256-bit AES encryption key in the Base64 format.
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 clusters, see the Amazon Redshift documentation.
Number of Nodes in the 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.
Number of bytes needed to support multibytes for varchar	Not applicable.

Configuring an Amazon Redshift Connection

Configure an Amazon Redshift connection in the Workflow Manager to define the connection attributes that the PowerCenter Integration Services uses to connect to the Amazon Redshift database.

1. In the Workflow Manager, click **Connections** > **Application**.

   The **Application Connection Browser** dialog box appears.
2. Click **New**.

   The **Select Subtype** dialog box appears.
3. Select **AmazonRSCloudAdapter** and click **OK**.

   The **Connection Object Definition** dialog box appears.
4. Enter a name for the Amazon Redshift connection.
5. Enter the application properties for the connection.
6. Enter the Amazon Redshift connection attributes.
7. Click **OK**.

Configuring the Source Qualifier

When you import a source to create a mapping for Amazon Redshift source , you must configure the source qualifier to create the mapping.

1. In the mapping, click **Source Qualifier**
2. Select the **Configure** tab
3. Specify the Amazon Redshift connection details.
4. Save the mapping.

Amazon Redshift Source Sessions

Create a mapping with an Amazon Redshift source and a target to read data from Amazon Redshift.

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

Client-side Encryption

Client-side encryption is a technique to encrypt data while writing the data to Amazon S3.

Create a Master Symmetric Key, which is a 256-bit AES encryption key in Base64 format. To enable client-side encryption, provide a Master Symmetric Key you created in the connection properties. The PowerCenter Integration Service encrypts the data by using the Master Symmetric Key. 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 support encryption with maximum security, you must update the security policy .jar files `local_policy.jar` and `US_export_policy.jar`. The .jar files are located at `<Informatica Installation Directory>\ava\jre\lib\security`. You can download the .jar files supported by the JAVA environment from the Oracle website.

Identity Columns

An identity column contains unique values that are automatically generated.

Rules and Guidelines for Identity Columns

- The data type for an identity column must be either int or bigint.
- When you create a mapping for an insert operation, you must link either all the source and target identity columns or none.
- When you create a mapping for an update, upsert or delete operation, you cannot map the identity columns that are not part of the primary key.
- If an identity column is part of the primary key, you must map the column for update, upsert, and delete operations, or the task session fails. However, you cannot set a source value for these columns.
- The ExplicitID and MaxError count options are removed for the upsert, update, and delete operations.

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 **UnloadOptions Property File** field. Enter the options in uppercase and delimit the options by using a new line space. The Unload command has the following options and default values:

```
DELIMITER=\036 DELIMITER=\036 ESCAPE=OFF PARALLEL=ON AWS_IAM_ROLE=arn:aws:iam::<account ID>:role/<role-name>

ESCAPE=OFF

PARALLEL=ON

AWS_IAM_ROLE=arn:aws:iam::<account ID>:role/<role-name>
```

When you run a task in the Secure Agent runtime environment, you can create a property file. You can also create a property file. The property file contains the Unload command options. Include the property file path in the **UnloadOptions Property File** field. For example:

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

In the property file, delimit the options by using a new line. For example:

```
DELIMITER=\036

ESCAPE=OFF

PARALLEL=ON

AWS_IAM_ROLE=arn:aws:iam::<account ID>:role/<role-name>
```

It is recommended to use octal representation of non-printable characters as DELIMITER and QUOTE.

If you run the Unload command as a pre-SQL or post-SQL command, specify the `ALLOWOVERWRITE` option to overwrite the existing objects.

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 **UnloadOptions Property File** 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 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. Default is \036, the octal representation of the non-printable character, record separator.

ESCAPE

You can add an escape character for CHAR and VARCHAR columns in delimited unload files before occurrences of the following characters:

- Linefeed \n
- Carriage return \r
- Delimiter character specified for the unloaded data
- Escape character \
- Single- or double-quote character

Default is OFF.

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.

AWS_IAM_ROLE

Specify the Amazon Redshift Role Resource Name (ARN) to run the task session on agent PowerCenter Integration Service installed on an Amazon EC2 system in the following format:

```
AWS_IAM_ROLE=arn:aws:iam::<account ID>:role/<role-name>
```

For example: `arn:aws:iam::123123456789:role/redshift_read`

ADDQUOTES

ADDQUOTES is implemented with the UNLOAD command by default. Do not specify the ADDQUOTES option in the advanced source properties. The Unload command adds quotation marks to each data field. With added quotation marks, the UNLOAD command can read data values that contain the delimiter. If double quote (") is a part of data, use ESCAPE to read the double quote as a regular character.

Partitioning

If you need to extract a large amount of source data, you can partition the sources to improve session performance. Partitioning sources allows the PowerCenter Integration Service to create multiple connections to sources and process partitions of source data concurrently. You can partition sources if the PowerCenter Integration Service can maintain data consistency when it processes the partitioned data.

By default, the Workflow Manager sets the partition type to pass-through for Amazon Redshift tables. In pass-through partitioning, the PowerCenter Integration Service passes all rows at one partition point to the next partition point without redistributing them.

If you create multiple partitions for an Amazon Redshift source session, the PowerCenter Integration Service evaluates the session properties in the following order to run the session:

1. SQL Query
2. INFA ADVANCED FILTER
3. Slices on Amazon Redshift Nodes

Amazon Redshift Source Session Configuration

You can configure a session to read data from Amazon Redshift. Define the properties for each source in the session.

The following table describes the session properties:

Advanced Property	Description
S3 Bucket Name	Amazon S3 bucket name for the Amazon Redshift source data. 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. Session performance improves when the PowerCenter Integration Service compresses the staged files. Default is selected.
Staging Directory Location	Amazon Redshift staging directory. Specify a directory on the machine that hosts the PowerCenter Integration Service.
UnloadOptions Property File	Path to the property file. Enables you to 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 - ESCAPE - AWS_IAM_ROLE Either specify the path of the property file that contains the unload options or specify the unload options directly in the **UnloadOptions Property File** field. Specify a directory on the machine that hosts the PowerCenter Integration Service.
Turn on S3 Client Side Encryption	Indicates that the PowerCenter Integration Service encrypts data before writing the data to Amazon S3 by using a private encryption key.
Enable Downloading 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.
Infa Advanced Filter	SQL filter command to divide the source database into multiple segments.
Pre-SQL	The UNLOAD or COPY commands to read from or write to Amazon Redshift. The command you specify here is processed as a plain text.
Post-SQL	The UNLOAD or COPY commands to read from or write to Amazon Redshift. The command you specify here is processed as a plain text.

Advanced Property	Description
SQL Query	Overrides the default query. Enclose column names in double quotes. The SQL query is case sensitive. Specify an SQL statement supported by the Amazon Redshift database.
Number of Sorted Ports	Number of columns used when sorting rows queried from the source. The PowerCenter Integration Service adds an ORDER BY clause to the default query when it reads source rows. The ORDER BY clause includes the number of ports specified, starting from the top of the transformation. When you specify the number of sorted ports, the database sort order must match the session sort order. Default is 0.
Select Distinct	Selects unique values. The PowerCenter Integration Service includes a SELECT DISTINCT statement if you choose this option. Amazon Redshift ignores trailing spaces. Therefore, the PowerCenter Integration Service might extract fewer rows than expected.
Source Table Name	You can override the default source table name.

Amazon Redshift Target Sessions

Create a session and associate it with the mapping that you created to move data to an Amazon Redshift table. Change the connection to an Amazon Redshift connection, and define the session 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

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

The PowerCenter Integration Service deletes the staged files from the staging directory after it writes the data to Amazon S3. Specify a staging directory in the session 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 PowerCenter Integration Service.

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

```
<staging directory>/infaRedShiftStaging<MMDDHHmmSS>
```

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.

You can configure vacuum table recovery options in the session properties. You can choose to recover disk space for the entire database or for individual tables in a database. Vacuum databases or tables often to

maintain consistent query performance. 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 a session that writes to Amazon Redshift. 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.

Retain Staging Files

You can retain staging files on Amazon S3 after the agent PowerCenter 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 create a target connection, you can configure a file prefix or directory prefix to save the staging files. After you provide the prefixes, the agent PowerCenter 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 agent PowerCenter Integration Servicecreates 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 agentPowerCenter 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 agentPowerCenter Integration Service does not save the staging files.

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 source to Amazon Redshift.

You can specify the Copy command options directly in the **CopyOptions Property File** field. Enter the options in uppercase and delimit the options by using a new line space. The Copy command has the following options and default values:

```
DELIMITER=\036 DELIMITER=\036 ACCEPTINVCHARS=? QUOTE=\037 COMPUPDATE=OFF
AWS_IAM_ROLE=arn:aws:iam::<account ID>:role/<role-name>

ACCEPTINVCHARS=?
```

```
QUOTE=\037

COMPUPDATE=OFF

AWS_IAM_ROLE=arn:aws:iam::<account ID>:role/<role-name>
```

When you run a task in the Secure Agent runtime environment, you can create a property file. You can also create a property file. The property file contains the Copy command options. Include the property file path in the **CopyOptions Property File** field. For example:

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

In the property file, delimit the options by using a new line. For example:

```
DELIMITER=\036

ACCEPTINVCHARS=?

QUOTE=\037

COMPUPDATE=OFF

AWS_IAM_ROLE=arn:aws:iam::<account ID>:role/<role-name>
```

It is recommended to use octal representation of non-printable characters as DELIMITER and QUOTE.

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 **CopyOptions Property File** 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. Default is \036, the octal representation of the non-printable character, record separator.

ACCEPTINVCHARS

Loads data into VARCHAR columns even if the data contains UTF-8 characters that are not valid. When you specify ACCEPTINCHARS, the agent PowerCenter 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 agent PowerCenter 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. Default is question mark (?).

QUOTE

Specifies the quote character to use with comma separated values. Default is \037, the octal representation of the non-printable character, unit separator.

COMPUPDATE

Overrides current compression encoding and applies compression to an empty table. Use the COMPUPDATE option in an insert operation 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.

Default is OFF.

AWS_IAM_ROLE

Specify the Amazon Redshift Role Resource Name (ARN) to run the task session on agent PowerCenter Integration Service installed on an Amazon EC2 system in the following format:

```
AWS_IAM_ROLE=arn:aws:iam::<account ID>:role/<role-name>
```

For example: `arn:aws:iam::123123456789:role/redshift_write`

Amazon Redshift Target Session Configuration

You can configure a session to write data to Amazon Redshift. Define the properties for each target instance in the session.

The following table describes the session properties:

Property	Description
S3 Bucket Name	Amazon S3 bucket name for the Amazon Redshift target data. 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. Session performance improves when the PowerCenter Integration Service compresses the staged files. Default is selected.
Staging Directory Location	Amazon Redshift staging directory. Specify a directory on the machine that hosts the PowerCenter Integration Service.
Batch Size	Minimum number of rows in a batch. Enter a number greater than 0. Default is 1000.
Max Redshift 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 PowerCenter Integration Service writes the entire batch to the error file. Default is 1.
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	String value used to represent null values in CHAR and VARCHAR fields in Amazon Redshift targets, such as NULL or a space character. Default is an empty string.
Wait time in seconds for file consistency on S3	Number of seconds to wait for the PowerCenter Integration Service to make the staged files consistent with the list of files available on Amazon S3. Default is 5.

Property	Description
CopyOptions Property File	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 - AWS_IAM_ROLE Either specify the path of the property file that contains the copy options or specify the copy options directly in the **CopyOptions Property File** field. Specify a directory on the machine that hosts the PowerCenter 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 PowerCenter Integration Service encrypts data by using a private encryption key. If you enable both server side and client side encryption, the PowerCenter Integration Service ignores the server side encryption.
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.
Prefix for Retaining 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`.
Pre-SQL	The UNLOAD or COPY commands to read from or write to Amazon Redshift. The command you specify here is processed as a plain text.
Post-SQL	The UNLOAD or COPY commands to read from or write to Amazon Redshift. The command you specify here is processed as a plain text.
Target Table Name	You can override the default target table name.

Property	Description
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.
INSERT	If enabled, the PowerCenter Integration Service inserts all rows flagged for insert. If disabled, the PowerCenter Integration Service rejects the rows flagged for insert. By default, the insert operation is enabled.
DELETE	If enabled, the PowerCenter Integration Service deletes all rows flagged for delete. If disabled, the PowerCenter Integration Service rejects all rows flagged for delete.
UPDATE	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 session target attributes: **Update as Update** The PowerCenter Integration Service updates all rows as updates. **Update else Insert** The PowerCenter Integration Service updates existing rows and inserts other rows as if marked for insert.
Success File Directory	Directory for the Amazon Redshift success file. Specify a directory on the machine that hosts the PowerCenter Integration Service.
Error File Directory	Directory for the Amazon Redshift error file. Specify a directory on the machine that hosts the PowerCenter Integration Service.

Configuring HTTP Proxy Options

If your organization uses a proxy server to access the internet, you can configure the HTTP proxy server authentication settings at design time or at run time to read data from or write data to Amazon Redshift.

Configuring HTTP Proxy Options at Design-Time

If your organization uses a proxy server to access the internet, you can configure the proxy server authentication settings in the jvmoptions.ini file to read data from or write data to Amazon Redshift.

- Ensure that you enable the proxy server settings from your web browser.
- Access the jvmoptions.ini file from the following location: `<Informatica Installation Location>\clients\PowerCenterClient\client\bin`
- Add the following properties to the jvmoptions.ini file:

Property	Description
-Dhttps.proxyHost=	Name of the HTTP proxy server.
-Dhttps.proxyPort=	Port number of the HTTP proxy server.

Property	Description
-Dhttps.proxyUser=	Authenticated user name for the HTTP proxy server. This is required if the proxy server requires authentication.
-Dhttps.proxyPassword=	Password for the authenticated user. This is required if the proxy server requires authentication. **Note:** The password is in plain text and not encrypted.

Configuring HTTP Proxy Options at Run-Time

If your organization uses a proxy server to access the internet, you can configure the proxy server authentication settings in the Administrator Console to read data from or write data to Amazon Redshift.

- Ensure that you enable the proxy server settings from your web browser.
- In the Administrator Console, navigate to the Data Integration Service for which you want to set the proxy server settings.
- Add the following custom properties as JVMOptions:

Property	Description
-Dhttps.proxyHost=	Name of the HTTP proxy server.
-Dhttps.proxyPort=	Port number of the HTTP proxy server.
-Dhttps.proxyUser=	Authenticated user name for the HTTP proxy server. This is required if the proxy server requires authentication.
-Dhttps.proxyPassword=	Password for the authenticated user. This is required if the proxy server requires authentication. **Note:** The password is in plain text and not encrypted.

For example:

```
JVMOption1=-Dhttps.proxyHost=<proxy_server_hostname>
JVMOption2=-Dhttps.proxyPort=8081
JVMOption3=-Dhttps.proxyUser=adminuser
JVMOption4=-Dhttps.proxyPassword=password
```

Octal Values as DELIMITER and QUOTE

In addition to printable ASCII characters, you can use octal values for printable and non-printable ASCII characters as DELIMITER and QUOTE.

To use a printable character as DELIMITER or QUOTE, you can either specify the ASCII character or the respective octal value. However, to use a non-printable character as DELIMITER or QUOTE, you must specify the respective octal value.

Example for a printable character:

`DELIMITER=#` or `DELIMITER=\043`

Example for a non-printable character, file separator:

```
QUOTE=\034
```

Octal values 000-037 and 177 represent non-printable characters and 040-176 represent printable characters. The following table lists the recommended octal values, for QUOTE and DELIMITER in the Copy command and as DELIMITER in the Unload command, supported by Amazon Redshift:

Command Option	Recommended Octal Values
COPY QUOTE	001-010, 016-037, 041-054, 057, 073-100,133, 135-140, 173-177
COPY DELIMITER	001-011, 013, 014, 016, 017, 020-046, 050-054, 057, 073-133, 135-177
UNLOAD DELIMITER	001-011, 013, 014, 016, 017, 020-041, 043-045, 050-054, 056-133, 135-177

Success and Error Files

The PowerCenter Integration Service generates success and error files after you run a session. Success and error files are .csv files that contain row-level details.

The PowerCenter Integration Service does not overwrite success or error files. You can manually delete the files that you no longer need.

Success Files

The PowerCenter Integration Service generates a success file after you run a session. The success file contains an entry for each record that successfully writes into Amazon Redshift. Each entry contains the values that are written for all the fields of the record. Use this file to understand the data that the PowerCenter Integration Service writes to the Amazon S3 bucket and then to the Amazon Redshift target.

If you want the PowerCenter Integration Service to generate a success file, specify a directory for the success file in the session properties. Specify a directory on the machine that hosts the PowerCenter Integration Service.

Rules and Guidelines for Success Files

Consider the following guidelines when you configure the session properties for success files:

- If you do not specify a directory for the success file as a session property, the PowerCenter Integration Service writes an empty file to the following directory: `$PMTargetFileDir`.
- The PowerCenter Integration Service generates the success file with the following naming convention: `<sessionName>_<timestamp>_success.csv`.
- If you do not map a primary key column in a mapping, the insert, update, upsert, and delete operations fail.

Error Files

The error file contains an entry for each data error. Each entry in the file contains the values for all fields of the record and the error message. Use the error file to understand why the PowerCenter Integration Service does not write data to the Amazon Redshift target.

If you want the PowerCenter Integration Service to generate an error file, specify a directory for the error file in the session properties. You must not specify a network path or shared path.

Consider the following guidelines when you configure the session properties for error files:

- If you do not specify a directory for the error file as a session property, the PowerCenter Integration Service writes a blank file to the following directory: `$PMBadFileDir`.
- The PowerCenter Integration Service generates errors file with the following naming convention: `<sessionName>_<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 PowerCenter 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""	""001E000000SI3jIIAT""
"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""

APPENDIX A

Amazon Redshift Data Type Reference

This appendix includes the following topics:

Data Type Reference Overview

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

Amazon Redshift and Transformation Data Types

The Amazon Redshift data types are the names and the aliases represent how the PowerCenter 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