informatica Cloud

Informatica Cloud & Redshift Getting Started User Guide

© 2014 Informatica Corporation. No part of this document may be reproduced or transmitted in any form, by any means (electronic, photocopying, recording or otherwise) without prior consent of Informatica Corporation. All other company and product names may be trade names or trademarks of their respective owners and/or copyrighted materials of such owners.

INFORMATICA CLOUD & REDSHIFT GETTING STARTED USER GUIDE 1
OVERVIEW 4
REDSHIFT CONNECTOR OVERVIEW 4
INFORMATICA CLOUD ARCHITECTURE 4
REDSHIFT CONNECTOR PREREQUISITES 5
DOWNLOADING AND INSTALLING THE VIBE SECURE AGENT 6
REDSHIFT CONNECTOR CONFIGURATION 7
GET YOUR AWS ACCOUNT SECRET KEY 7
GET YOUR REDSHIFT JDBC URL 9
CONFIGURE THE CONNECTOR PROPERTIES IN INFORMATICA CLOUD 10
USING THE DATA SYNCHRONIZATION WIZARD WITH REDSHIFT 12
CREATE YOUR DSS TASK 12
READING DATA FROM REDSHIFT 18
ODBC CONFIGURATION 19
SECURITY CONSIDERATIONS 19
CONFIGURING THE REDSHIFT CLUSTER VPC'S INBOUND IP SECURITY 19
CONFIGURING FOR REDSHIFT SSL 21
REDSHIFT CONNECTOR BEST PRACTICES 23

Overview

Amazon Web Services Redshift is a fast, fully managed, petabyte-scale data warehouse optimized for business intelligence. The Informatica Cloud Redshift Connector is a native, high-volume data connector enabling users to quickly and easily design petabyte-scale data integrations from any cloud or on premise sources to any number of Redshift nodes.

Getting started with Amazon Redshift is now easier than ever thanks to the Informatica Cloud 60 day trial for Amazon Redshift. Easily and quickly move data from all of your on premise and Cloud data sources, without writing a single line of code and without being a data integration expert. You can use our 6 step wizard to quickly replicate your data or use our intuitive web based designer to tackle more advanced use cases, such as combining multiple data sources into one Redshift table.

In this document we will cover all aspects of using Informatica Cloud to quickly get started with loading data into Redshift.

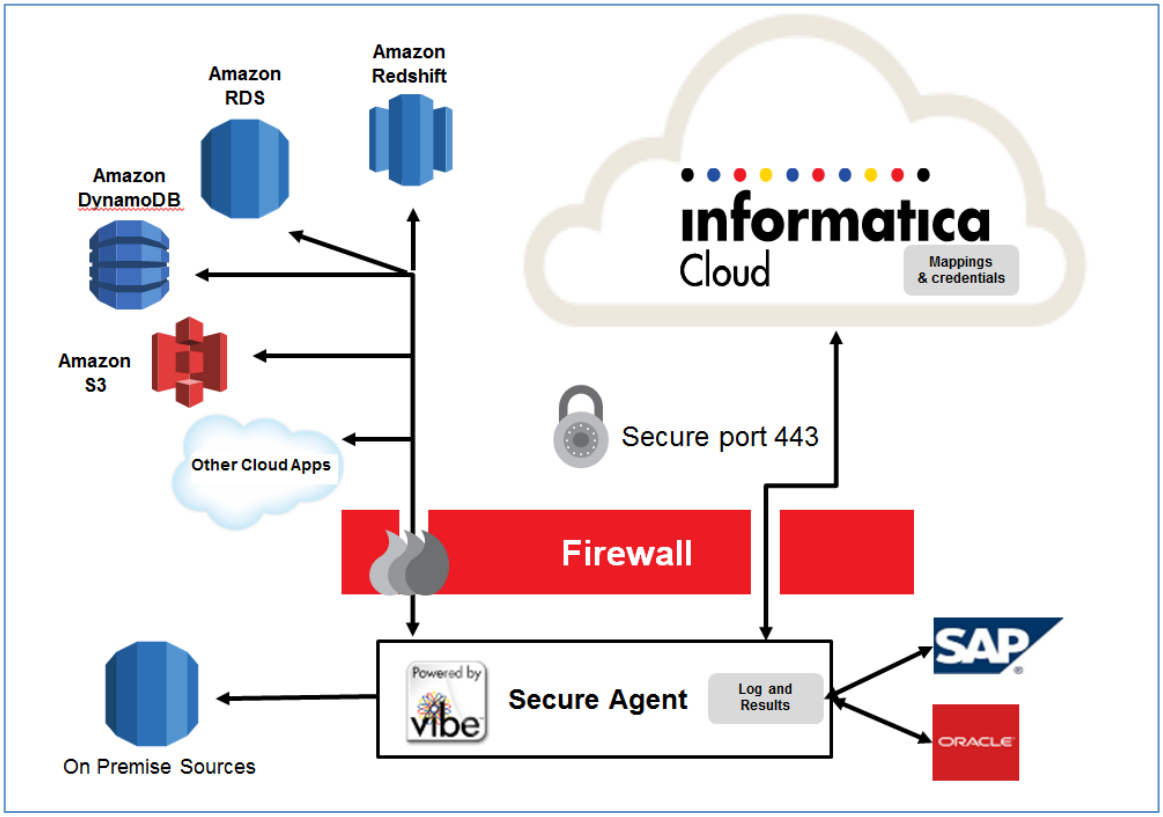
Redshift Connector Overview

The Redshift connector is a bulk-load type connector and allows you to perform inserts, deletes, and upserts (insert and/or update). Although Redshift does not natively support Upsert, the connector allows Upsert functionality by creating and loading a staging table first and then merging that with the existing table.

Access to Redshift data is available via ODBC or JDBC PostgreSQL drivers.

Informatica Cloud Architecture

The diagram below describes Informatica Cloud's high level architecture. It is important to note than none of your data flows through the cloud service; it all runs through the Vibe Secure Agent installed behind your firewall or on EC2.



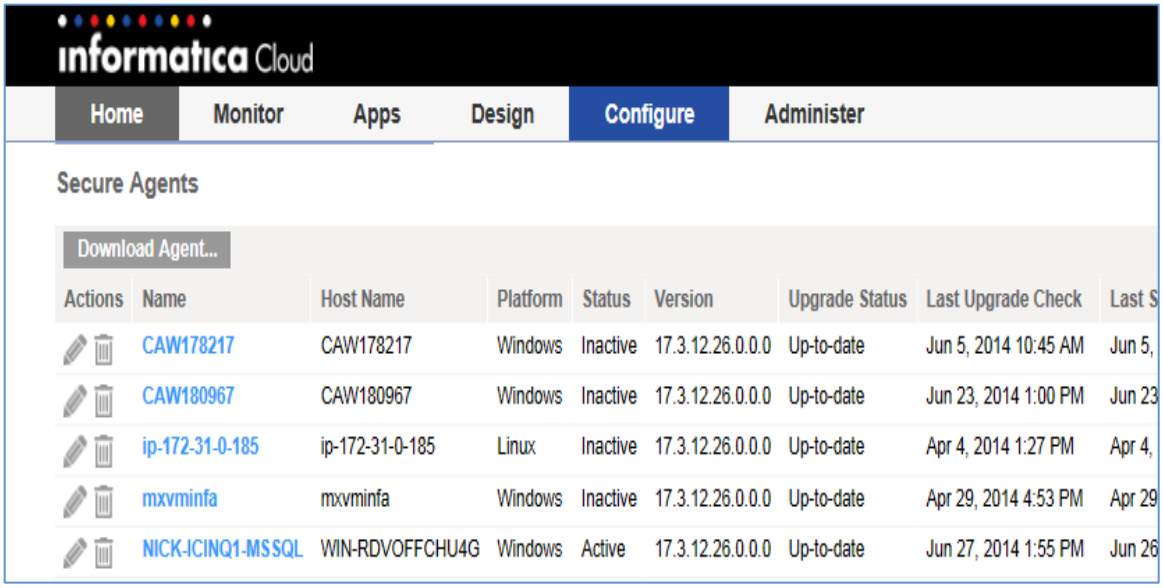
Redshift Connector Prerequisites

Before using the Redshift connector you will need the following prerequisites:

- An Informatica Cloud user account. You can sign up for a trial here: http://www.informaticacloud.com/
- An Amazon Web Services (AWS) Account .You can sign up here: http://aws.amazon.com/
- If you are not familiar with Redshift, it is recommended to go through the Amazon Get Started Guide here: http://docs.aws.amazon.com/redshift/latest/gsg/getting-started.html
- A Redshift Cluster with a schema that your user has CREATE and USAGE privileges to. By default all users have those privileges with the “public” schema.
- The user name and password for your Redshift cluster. These are not the same as your AWS account user credentials.
- An S3 bucket in the same region as your Redshift cluster
- An Informatica cloud agent that has access to the Redshift cluster. **IMPORTANT!** The IP of your Informatica Cloud Secure agent will need to be in the access inbound list of the VPC for your Redshift cluster.
- If you are running the agent on Windows make sure the 2010 Visual Studio C++ redistributables are installed. Please see this link: http://www.microsoft.com/en-us/download/details.aspx?id=5555
- It is recommended you set up a new IAM user for using with your Redshift cluster and Informatica Cloud. More information about IAM can be found here: http://aws.amazon.com/iam/

Downloading and Installing the Vibe Secure Agent

To download and install the agent, follow these steps:

1. Click on the Configuration tab.
2. Click on Agents.
3. Click the Download Agent button:
4. 

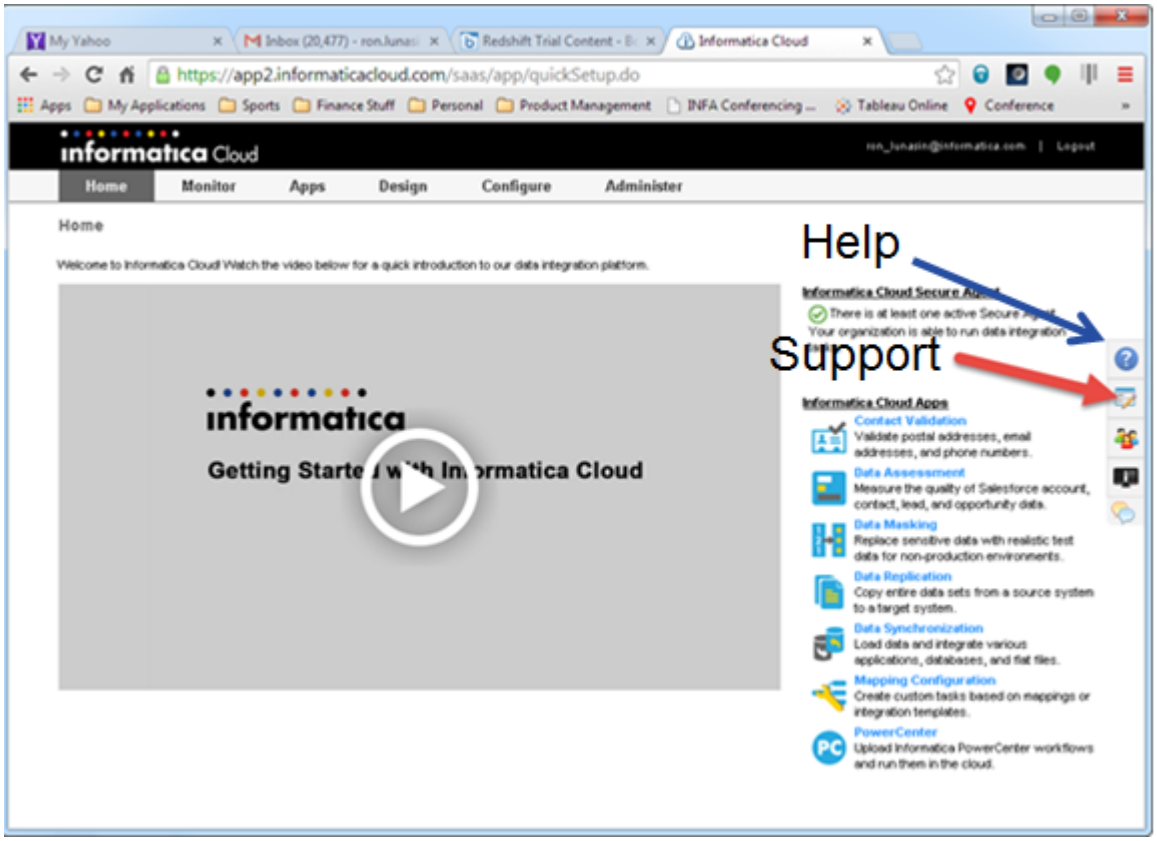
5. Select your operating system (in this guide we use Windows), and click the Download button:
6. Click the Save button to save the installer to your local machine:
7. You will be prompted to select a location for the file on your local machine.
8. Select a location and click the Save button:
9. When the installer has finished downloading, locate the file - named "agent_install.exe" - and double-click it to start the installation.
10. Click the Run button and follow the remaining steps in the installation wizard.
    a. We recommend that you accept the installation default values.
11. A registration page appears.
12. Enter your Informatica Cloud user name and password and click Register
13. The Secure Agent starts.
14. The Informatica Cloud Secure Agent window displays the status of the Secure Agent. You can restart, stop, and configure the Secure Agent proxy in this window. You can

close the window at any time. The Secure Agent continues to run as a service until stopped.

15. The Secure Agent Manager minimizes to the Windows taskbar notification area. Closing the Secure Agent Manager does not affect the Secure Agent status.

Help & Support

Informatica Cloud provides a number of getting started videos which are available in the Home tab of your Informatica Cloud Account. You can also click on the Help icon (see blue arrow below) from any page to access the online help documentation. If you need further assistance, click on the Support icon shown below (shown with the red arrow).

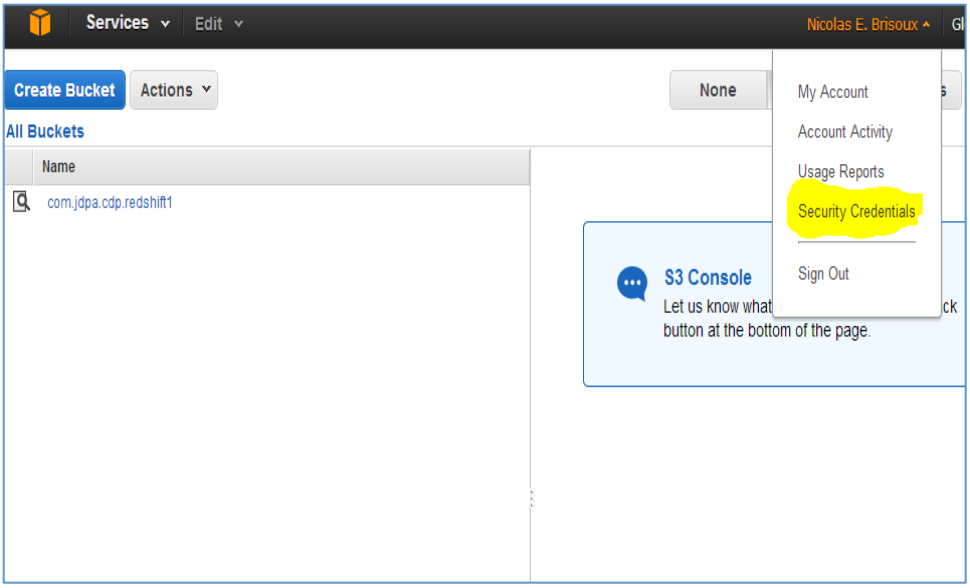


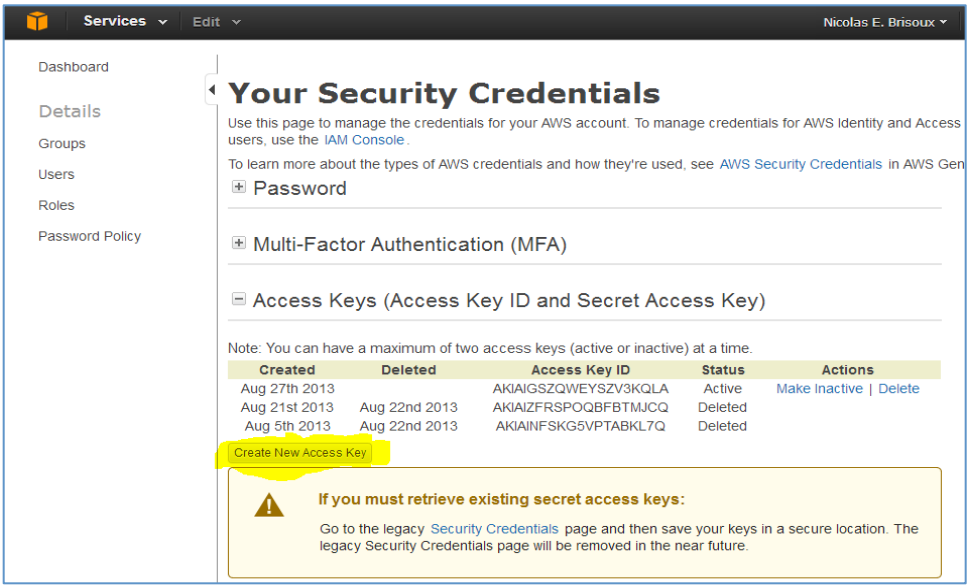
Redshift Connector Configuration

In order to configure the Redshift connector you will need to follow the steps below.

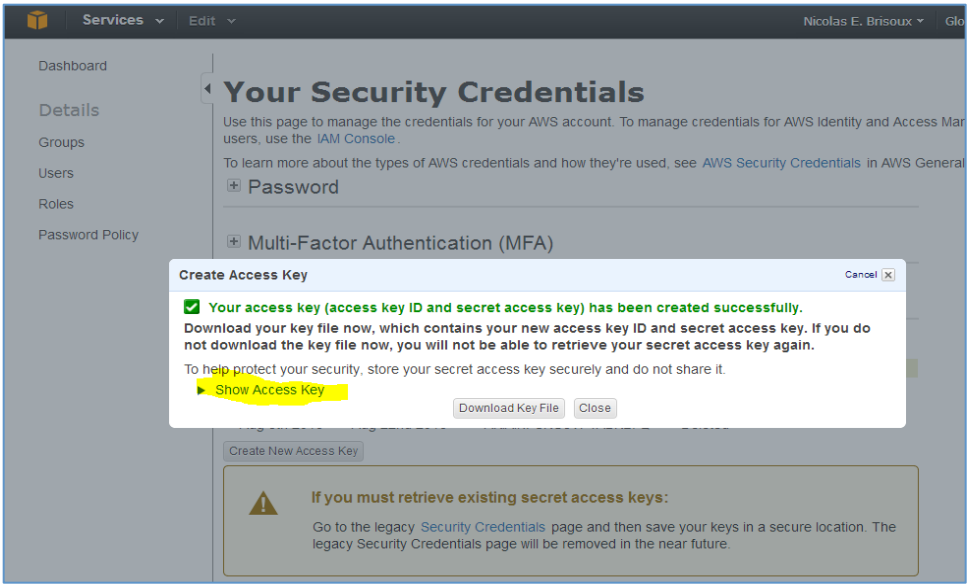
Get your AWS account secret key

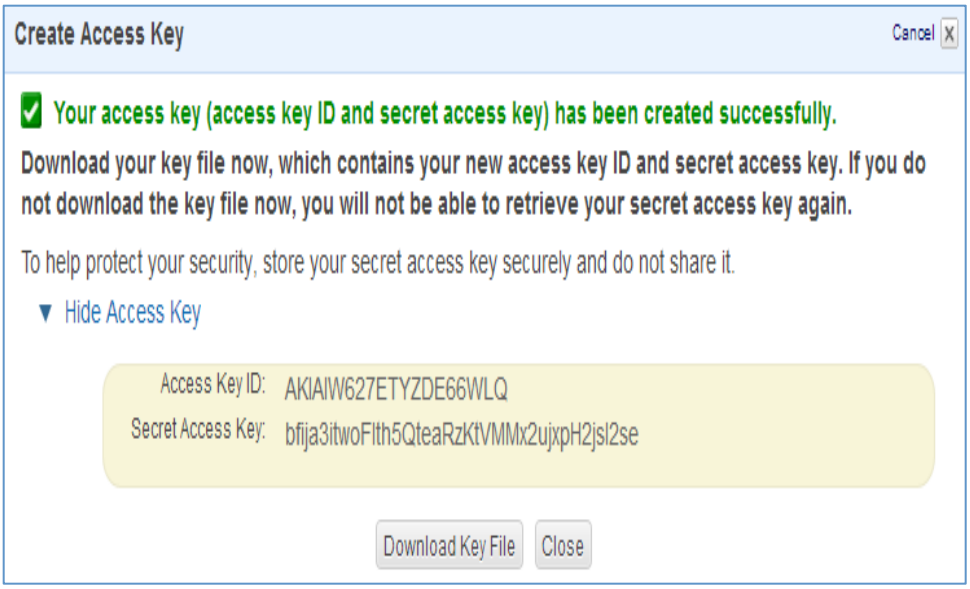
1. Go to your AWS account Security Credentials console as shown below:

2. 

3. Click on the Continue to Security Credentials button in the next dialog
4. Once in the console, expand the Access Keys section, and click on the Create New Access Key button:
5. 

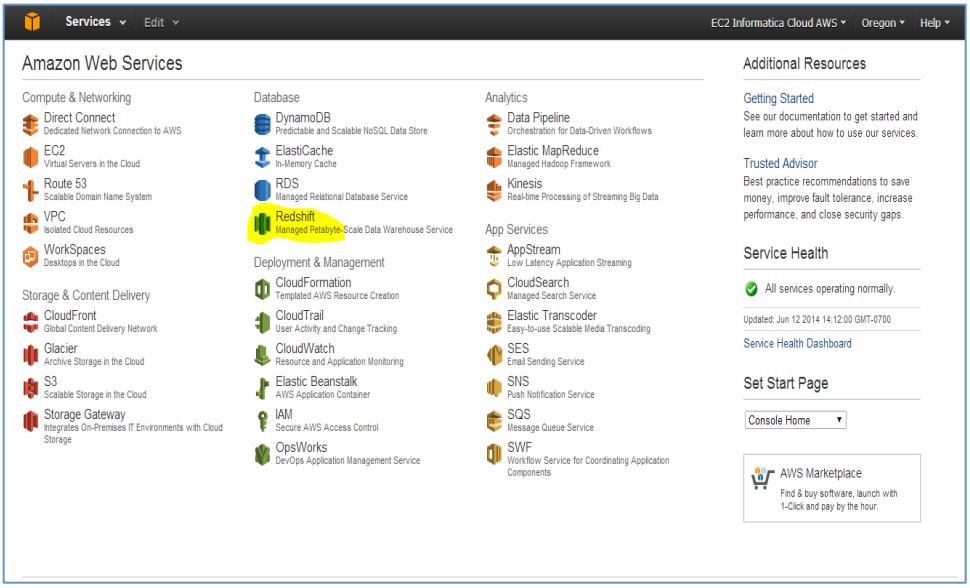
6. The following screen will appear. Click on the Show Access Key link

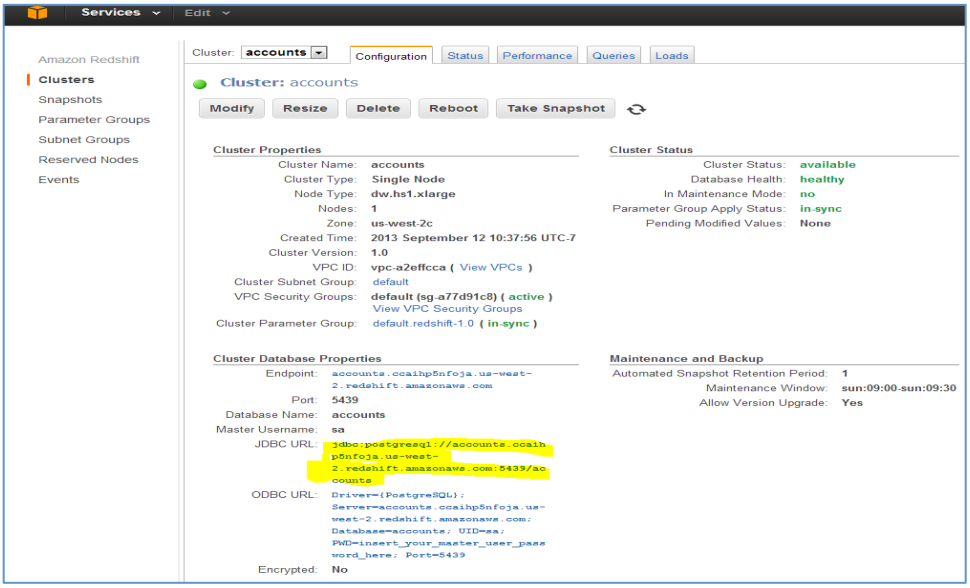
7. 

8. Note down your Access Key ID and the Secret Access Key:
9. 

Get your Redshift JDBC URL

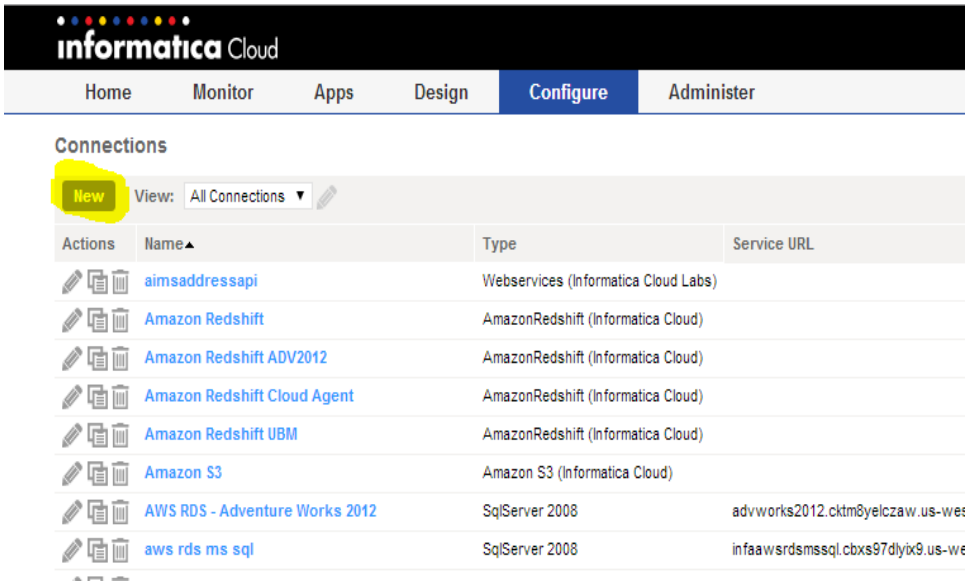
1. Go to the AWS management console: https://console.aws.amazon.com/console/home and from there go to the Redshift management page.

2. 

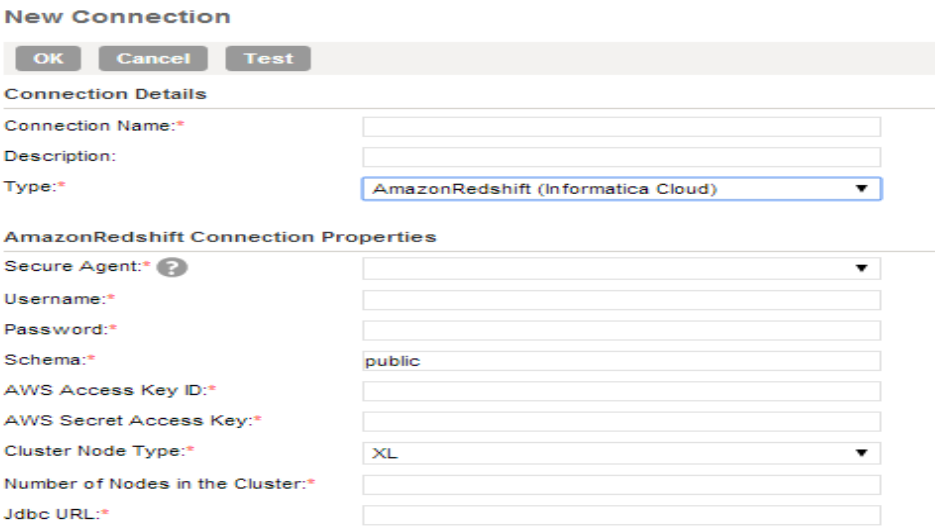
3. Bring up your cluster properties.
4. Note down the JDBC URL as shown below:
5. 

Configure the connector properties in Informatica Cloud

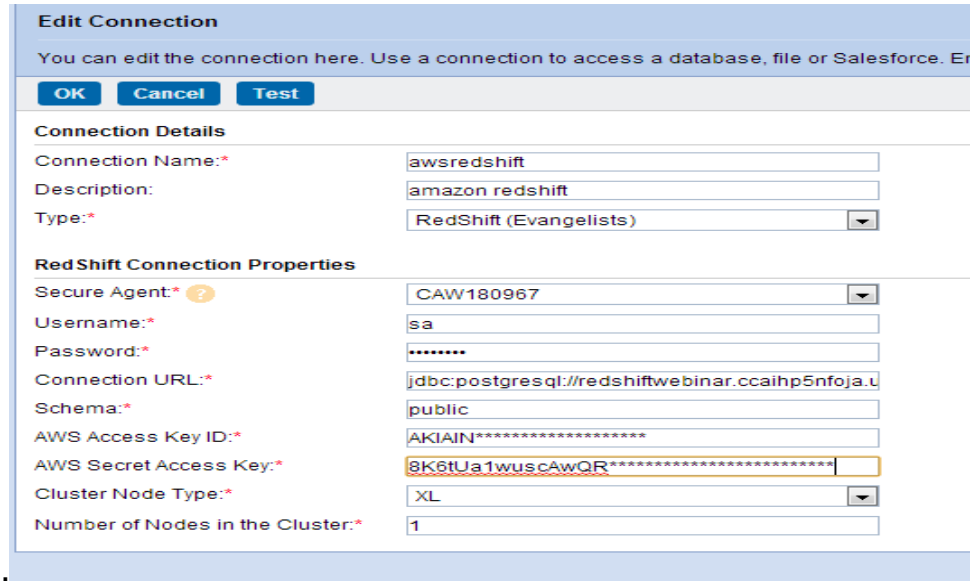
1. Log in to your Informatica Cloud account and go to your Connections page and click on New.



2.
3. Select Amazon Redshift as your connection type



4.
5. Enter the Redshift cluster username and password
6. Enter the schema name. If you did not create a specific schema for your cluster, you can use the “public” one.
7. Enter the cluster type, number of nodes, and the JDBC URL. See below for an example.



8. .
9. Click on the Test button to make sure you can connect to the Redshift Cluster.

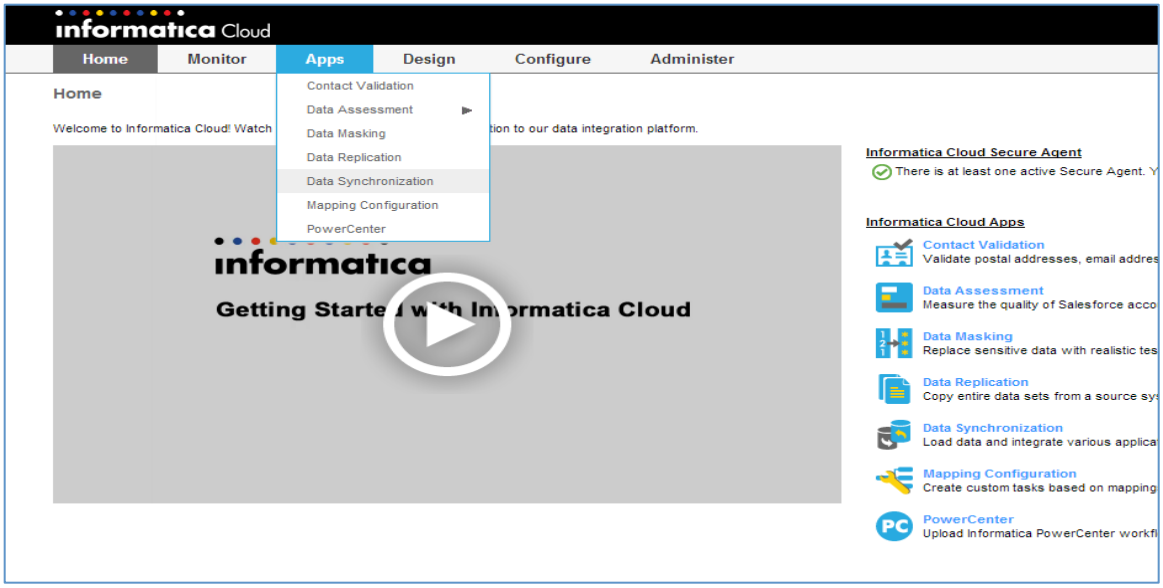
Using The Data Synchronization Wizard With Redshift

The Informatica Cloud data synchronization service (DSS) application delivers all of the key bi-directional synchronize data integration functions you need – and all through an intuitive web-based wizard. You can perform data transformation through a drag and drop web interface, perform lookups, as well as automate the running of your jobs on an hourly or to the minute schedule.

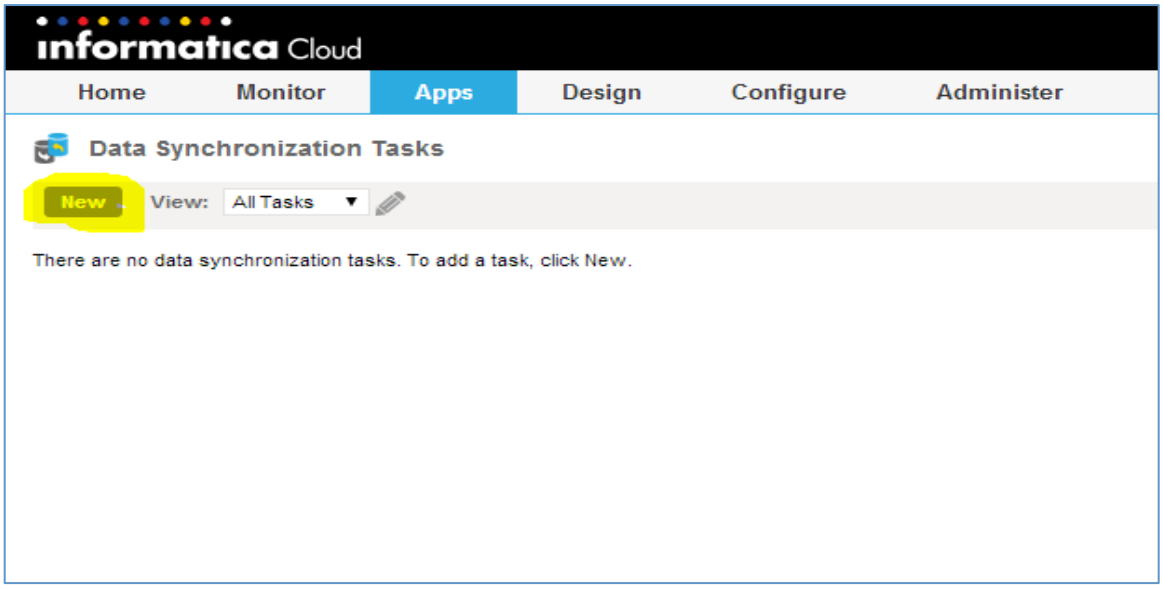
The guide below will show how to configure your first DSS task to load data into Redshift.

Create Your DSS Task

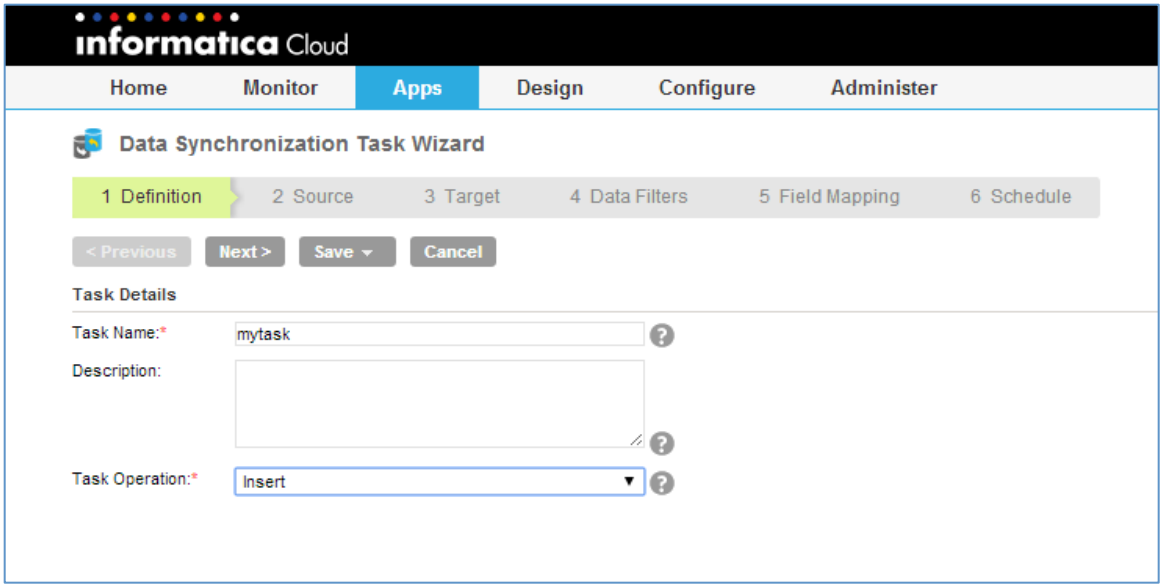
1. Go to the Apps menu and select the Data Synchronization application

2. 

3. Click the "New" button.

4. 

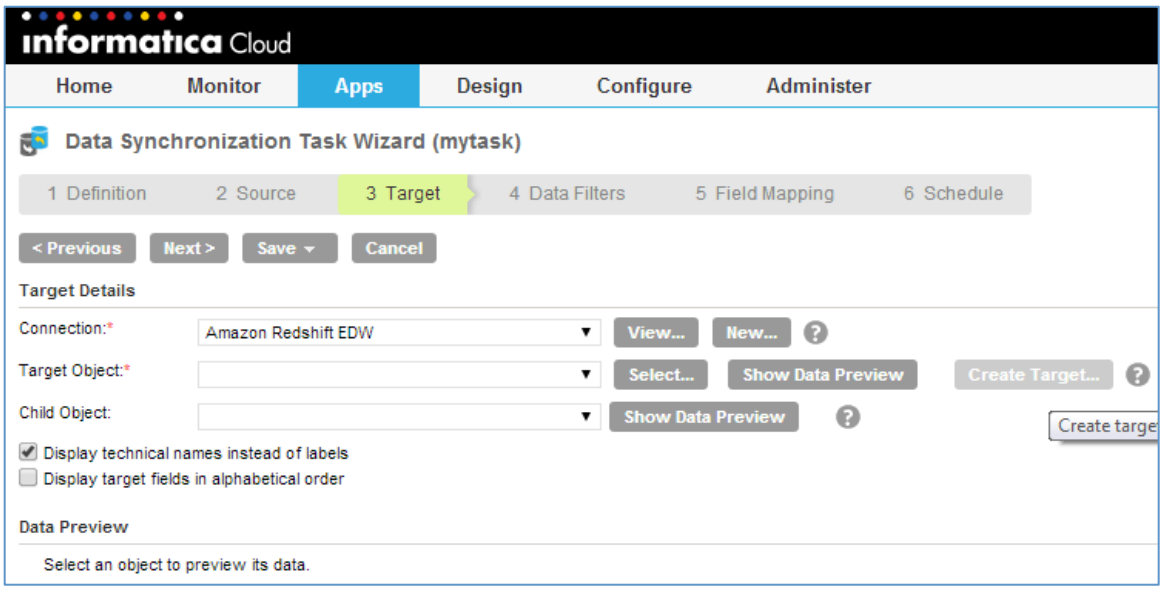
5. Choose a name for your task and from the Task Operation drop down selection box and choose "Insert"

6. 

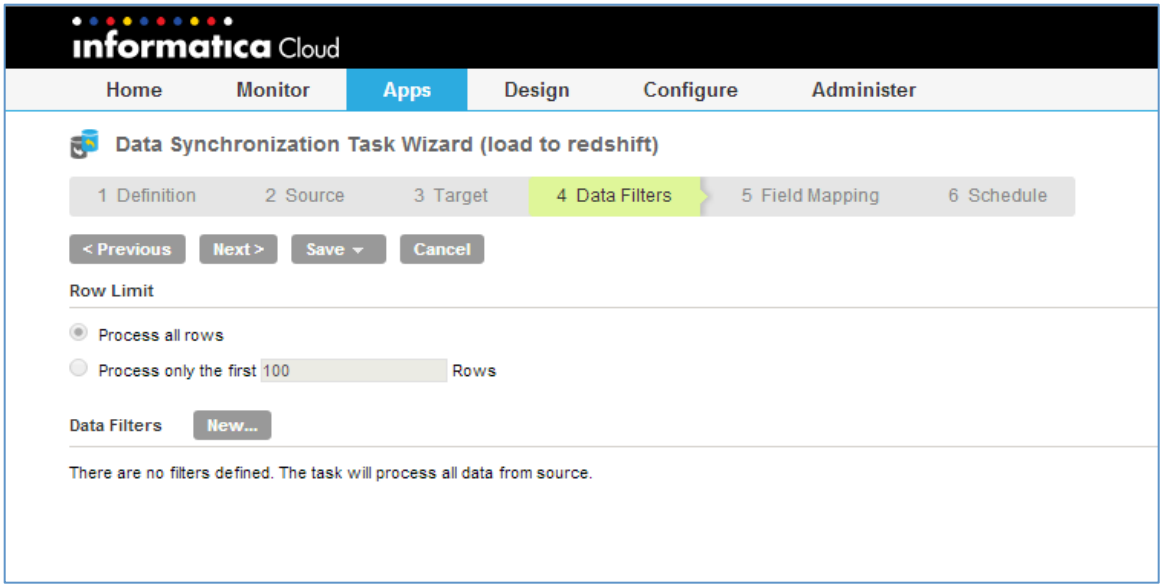
7. Click the "Next" button.

8. Choose your source connection for the data you will be loading into Redshift. Below is an example.

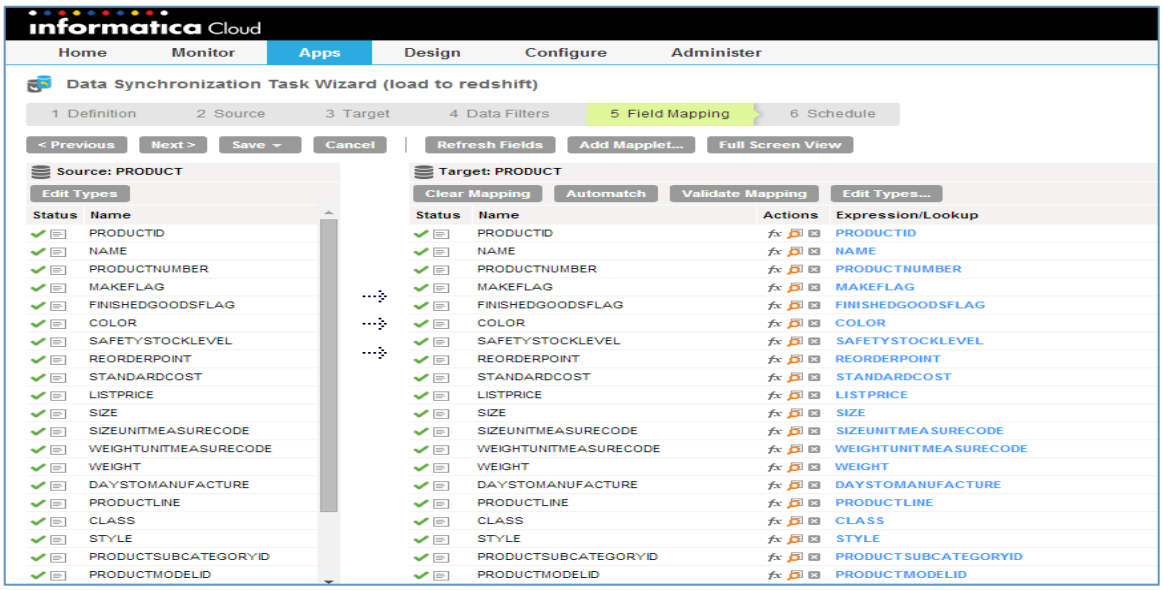
9. Pick your RS connection as the connetion type and click on the "Create Target" button.

10. 

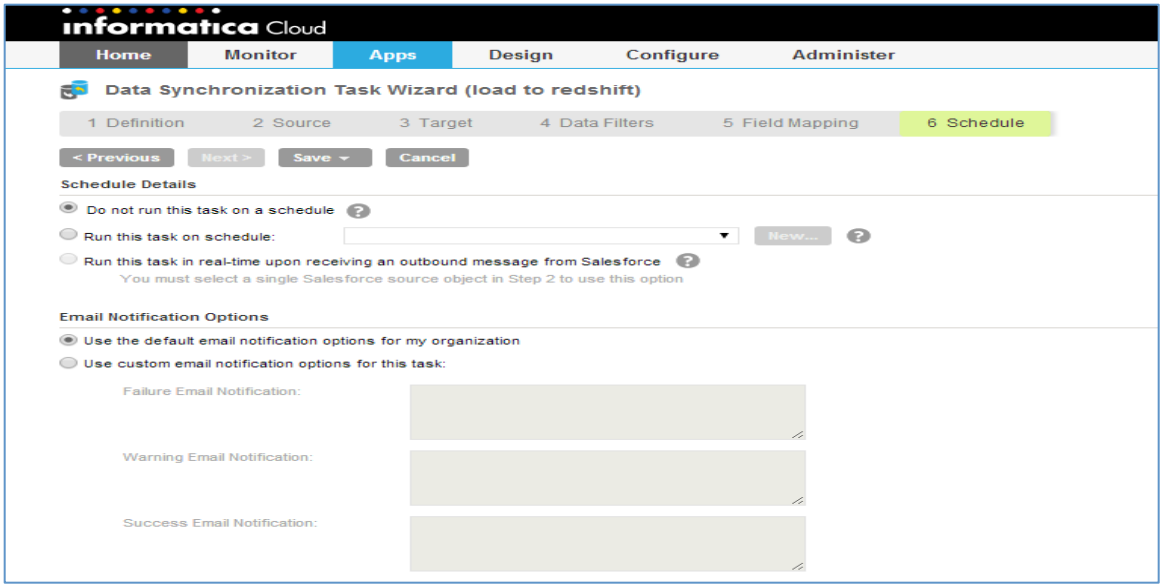
11. In Step 4 you can specify a source filter. This is optional. Click on the "Next" button.

12. 

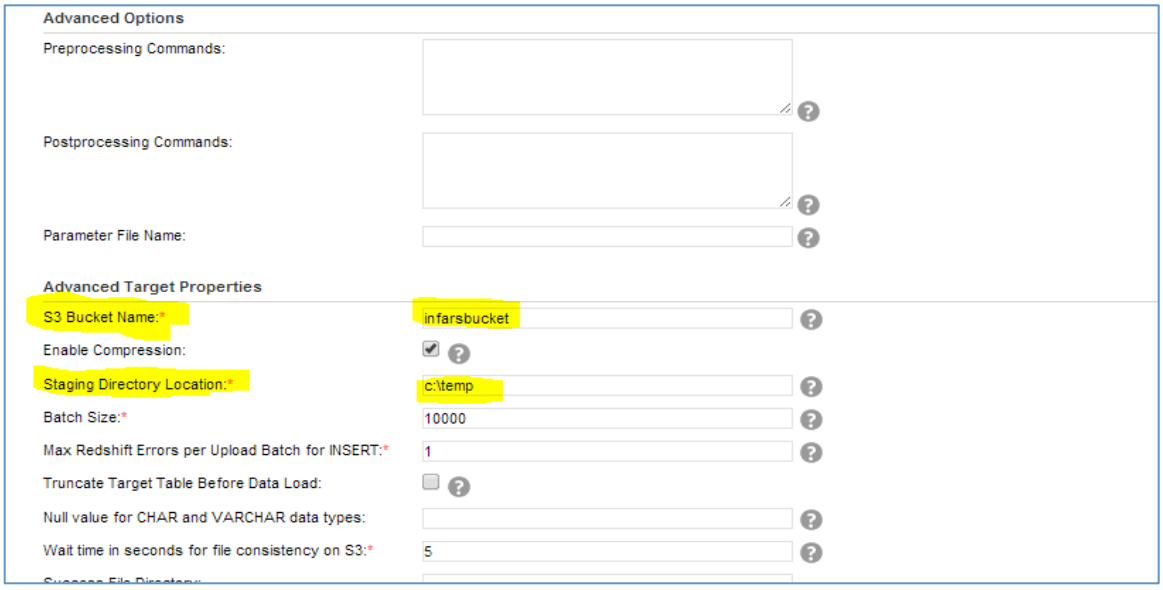
13. In Step 5, shown below, you specify the mapping via the drag and drop interface or by using the "Automatch" feature. You can also apply transformations or do lookups. You can get more information on how to do this by taking a look at the following training video: http://asdasd.asdasd.com

14. 

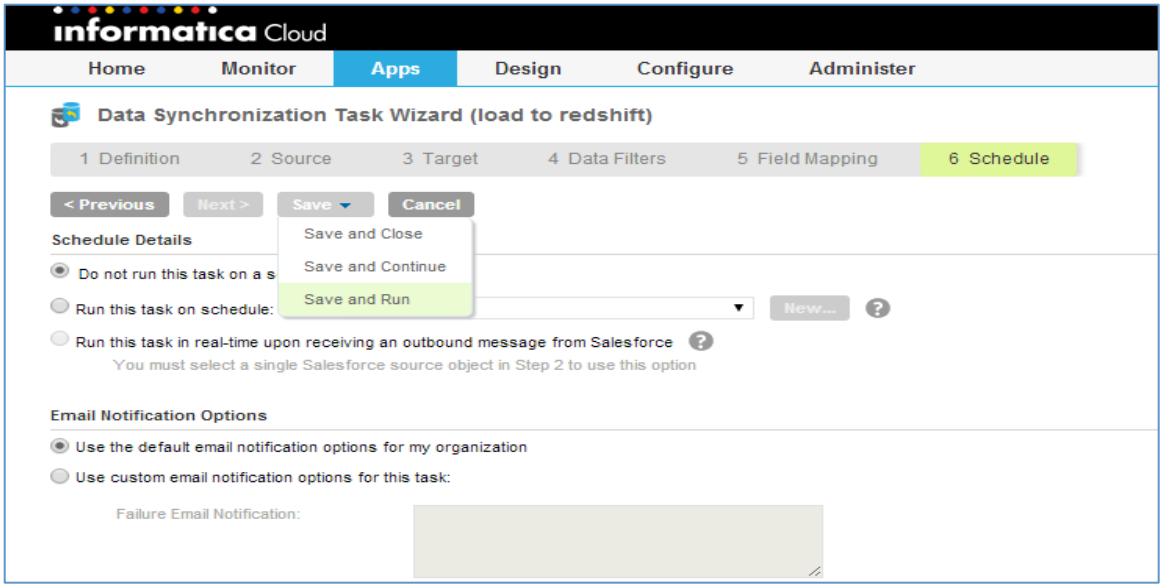
15. In the last step, Step 6, you can choose to run the task immediately or run it on a schedule.

16. 

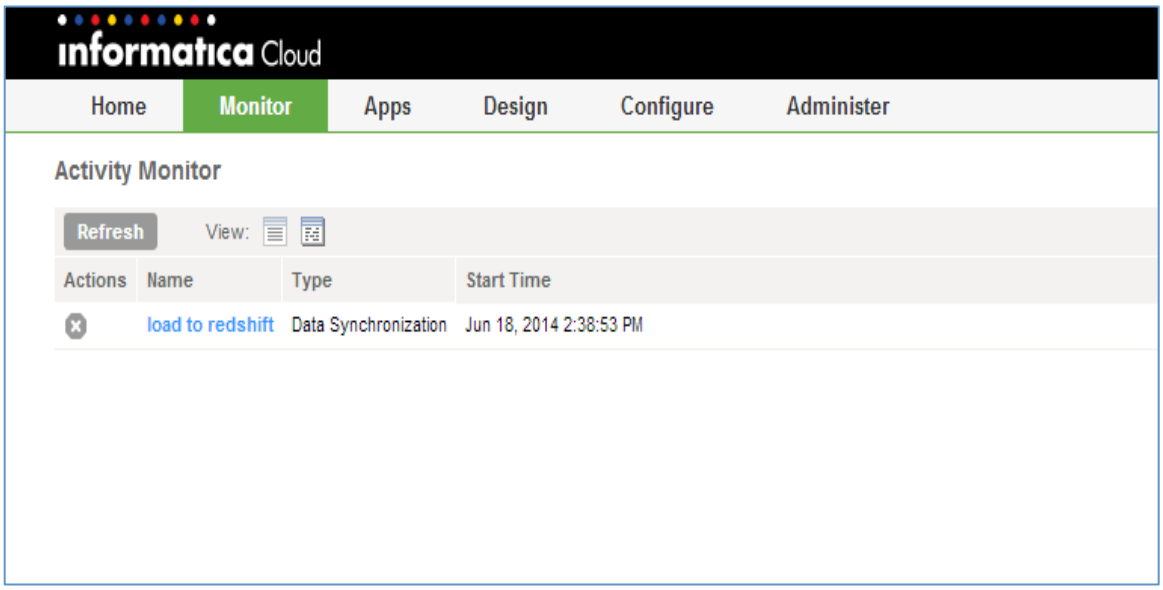
17. Before we run the task however, we need to enter some additional information specific to Redshift. Under the "Advanced Options" enter the S3 bucket name and the folder location for the Secure Agent to use to stage the files it will upload to S3.

18. 

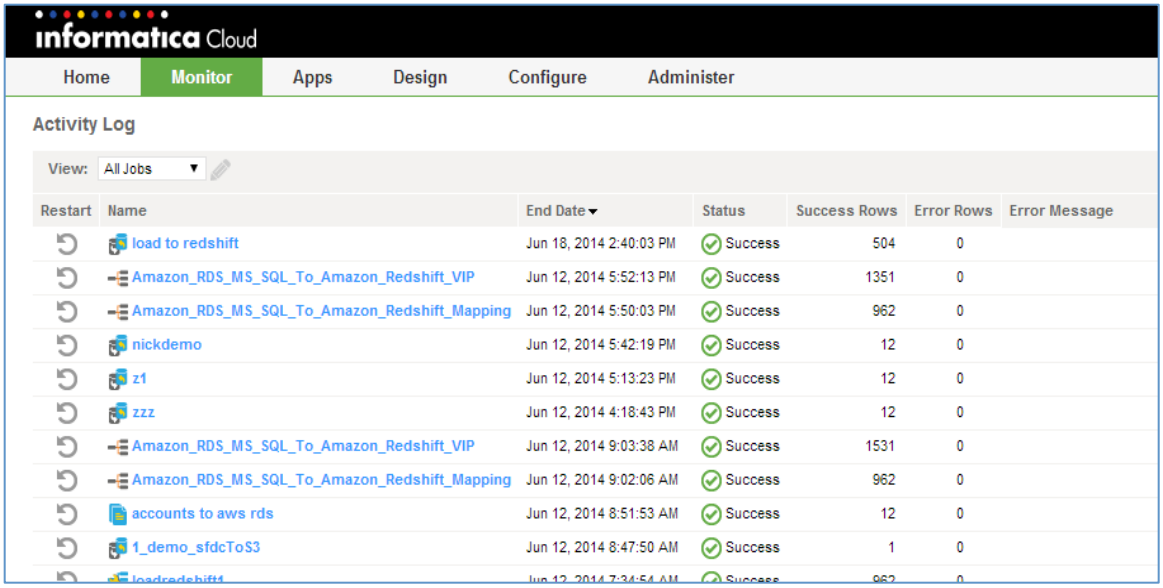
19. You can now run the task by selecting the "Save and Run" menu option from the "Save" menu.

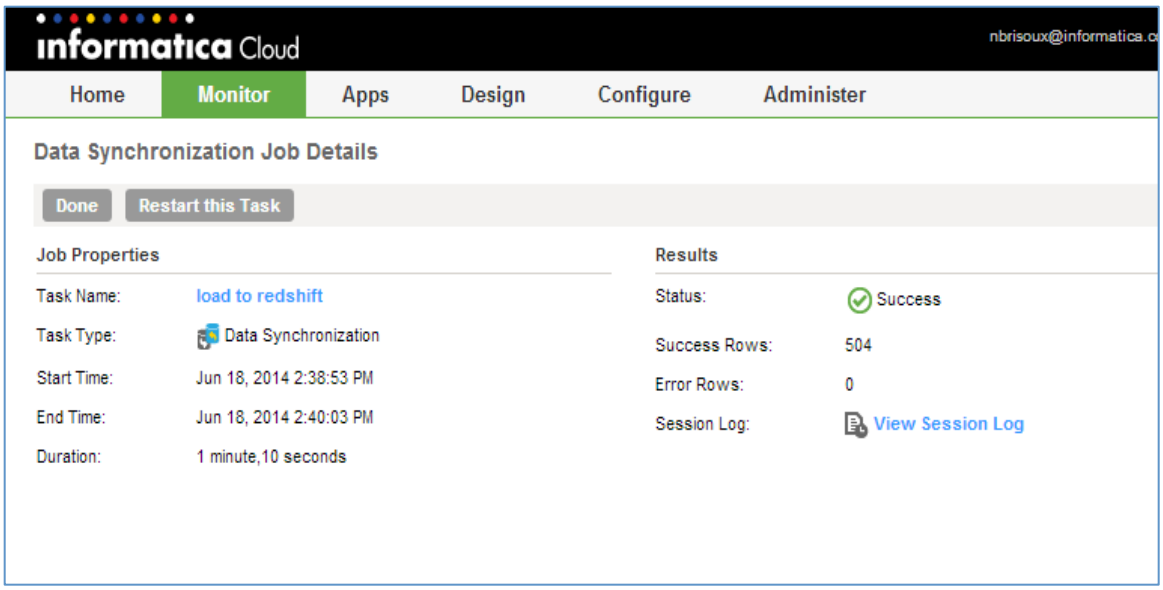
20. 

21. You will now be shown the Activity Monitor where you can see the running status of your task.

22. 

23. Once the tasks complete you will be shown the Activity Log. Click on your task to get detailed information about the task results as well view the session log.

24. 

25. 

Reading Data From Redshift

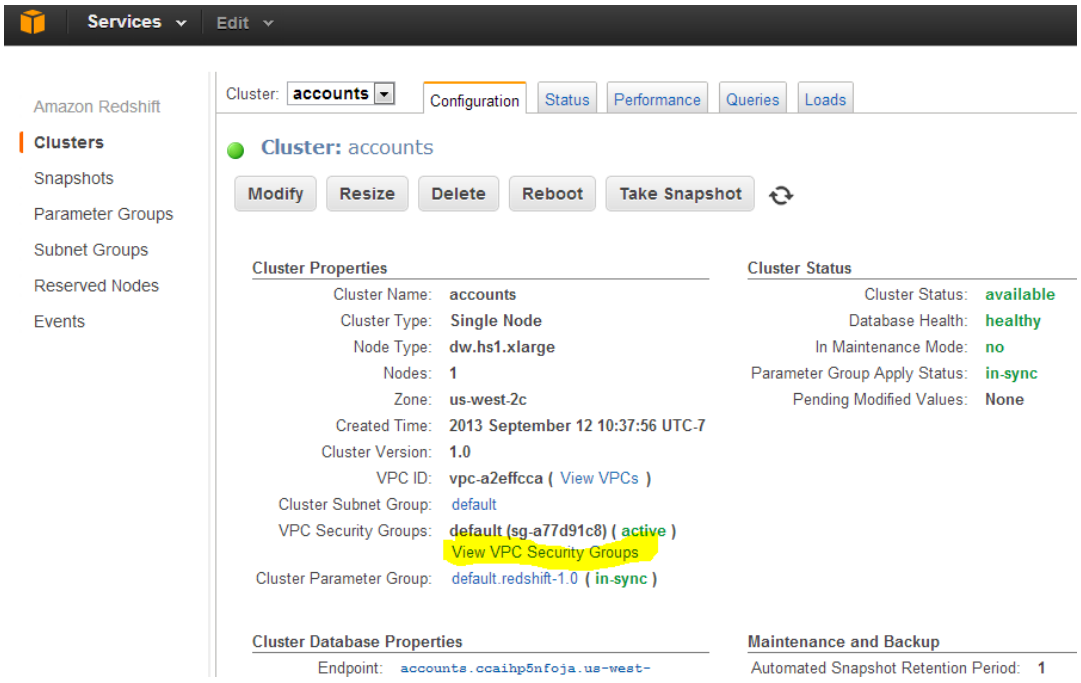
You can read data from using PostgreSQL JDBC or ODBC drivers (see the following Amazon documentation for detailed information: http://docs.aws.amazon.com/redshift/latest/mgmt/configuring-connections.html) In this section we will explain how to configure ODBC to work with Informatica Cloud. In these examples we will be using Windows. Refer to the PostgresSQL website (http://www.postgresql.org/) for how to configure these drivers for Linux.

Security Considerations

Configuring The Redshift Cluster VPC's Inbound IP Security

1. Go to the Redshift cluster you will be using with the Informatica Cloud Agent.
2. From the Redshift cluster management panel click on the name of your redshift cluster.
3. You can go through the next steps even if your cluster isnt active yet
4. On the following screen, click on **View VPC Security Groups**

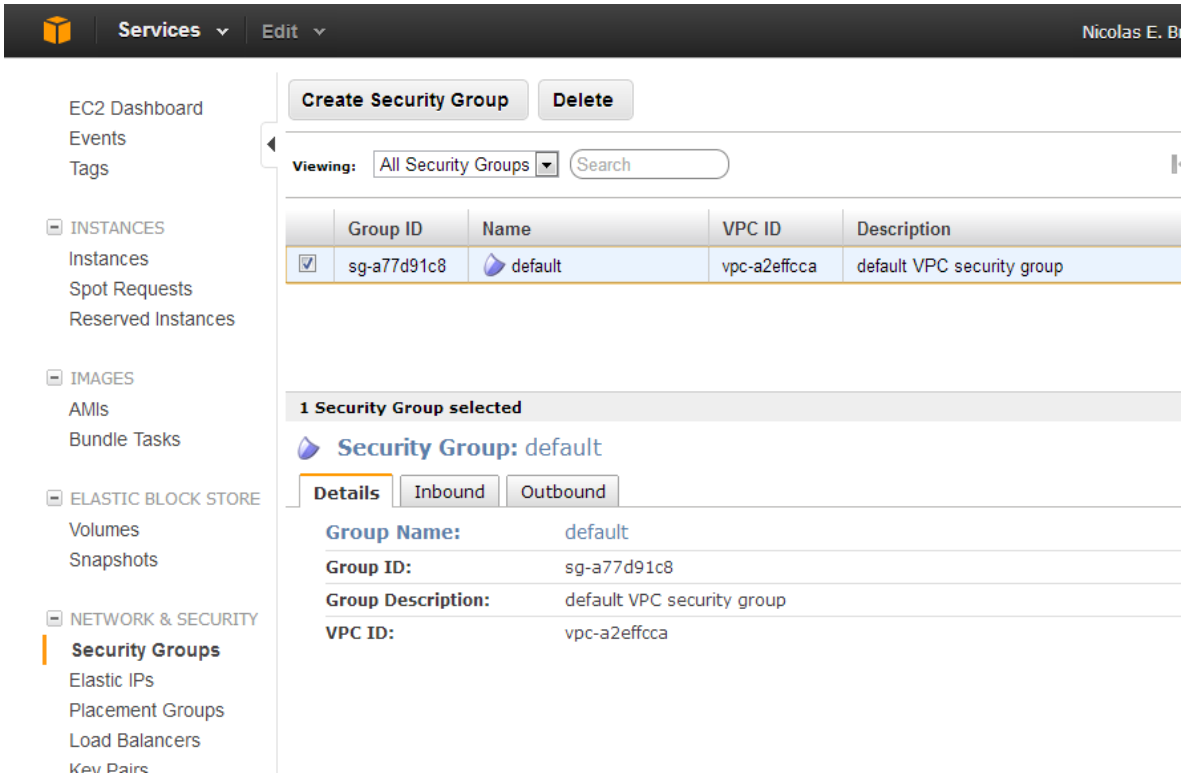
a.



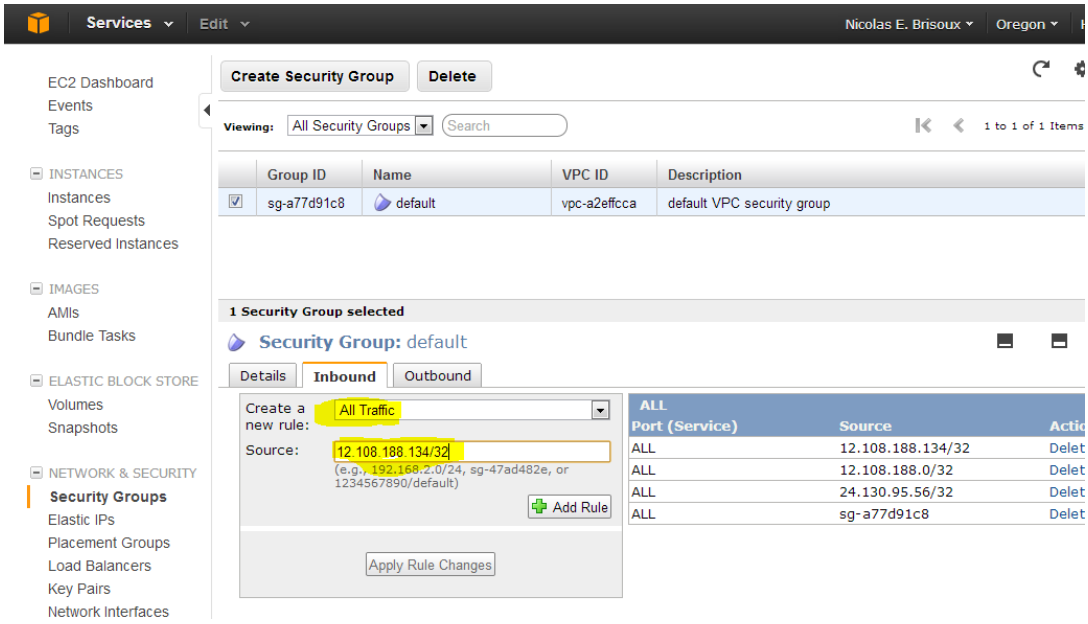
b. You should see your default VPC group listed

5. Select the default VPC group, and a panel will appear as below

a.



b. You will need to add any IP you are going to run the Cloud Agent from from to the Inbound list. In the example below, we use Informatica HQ's external IP.



i.

c. Apply the rule changes

Configuring For Redshift SSL

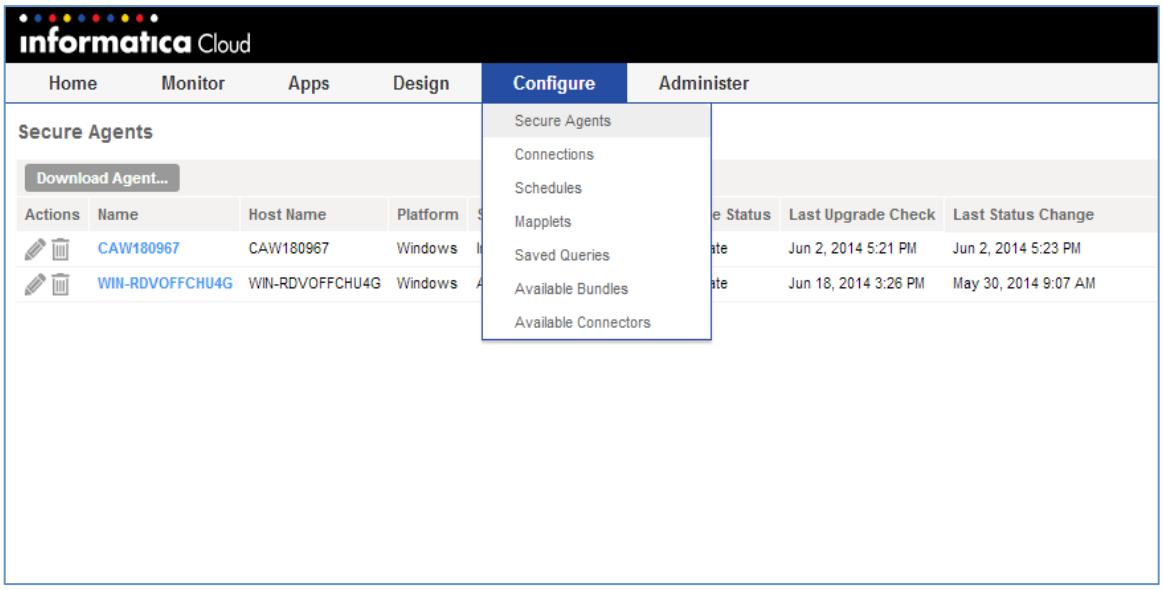
The Secure Agent can be configured to support an SSL connection to Redshift. We recommend consulting the Amazon Redshift documentation on this topic (see http://docs.aws.amazon.com/redshift/latest/mgmt/connecting-ssl-support.html#connecting-ssl-support-java). The following steps outline how to configure your Secure Agent to run with an SSL connection.

1. First you will need to add the Amazon Redshift certificate to the Java system truststoreDownload the certificate from https://s3.amazonaws.com/redshift-downloads/redshift-ssl-ca-cert.pem

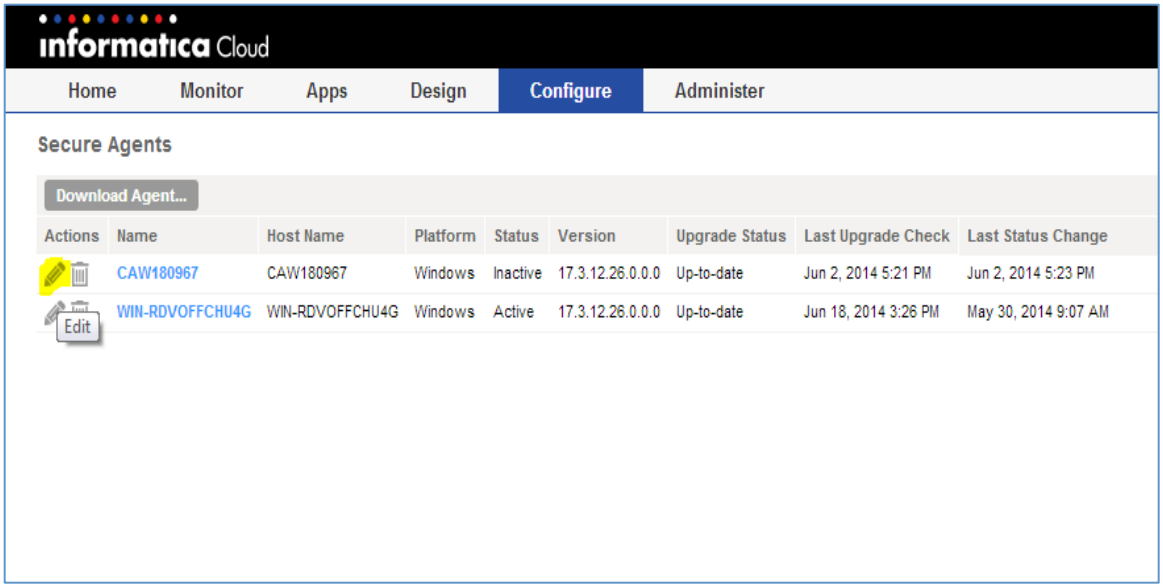
2. Add the certificate to the key store by executing the following command:${JAVA_HOME}/bin/keytool –keystore {JAVA_HOME}/lib/security/cacerts -import -alias <alias> -file <certificate_filename> Where <alias> is any user-provided string value and <certificate_filename> is the full path to the certificate file that you downloaded in Step 1.

3. You need to change the Secure Agent JVM startup properties to specify the keystore and password.

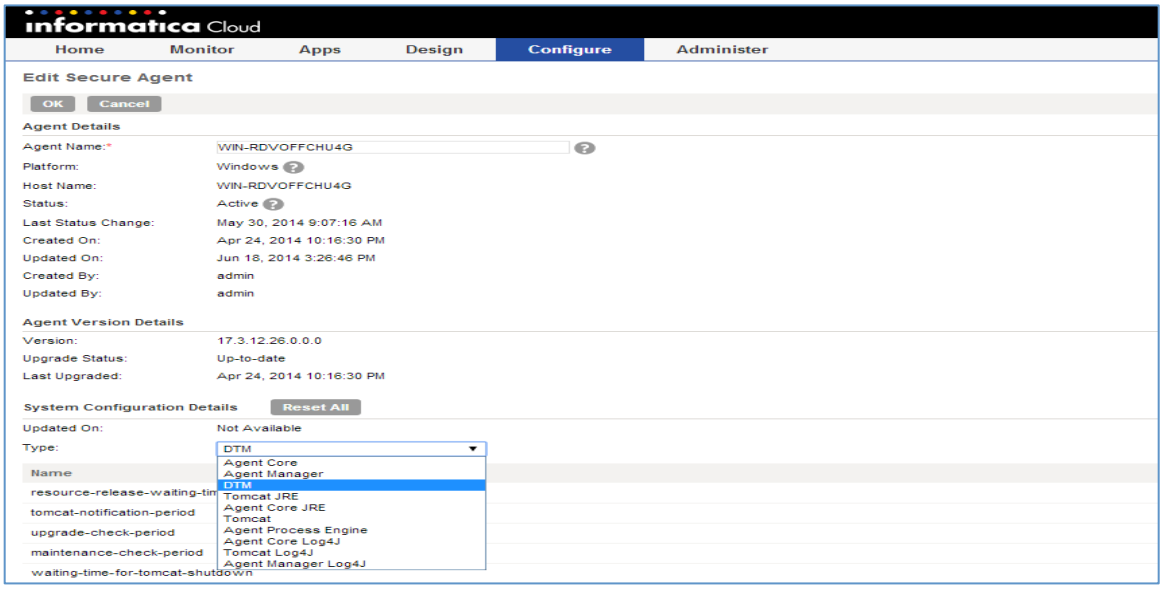
4. Go to your Secure Agents configuration page.

5. 

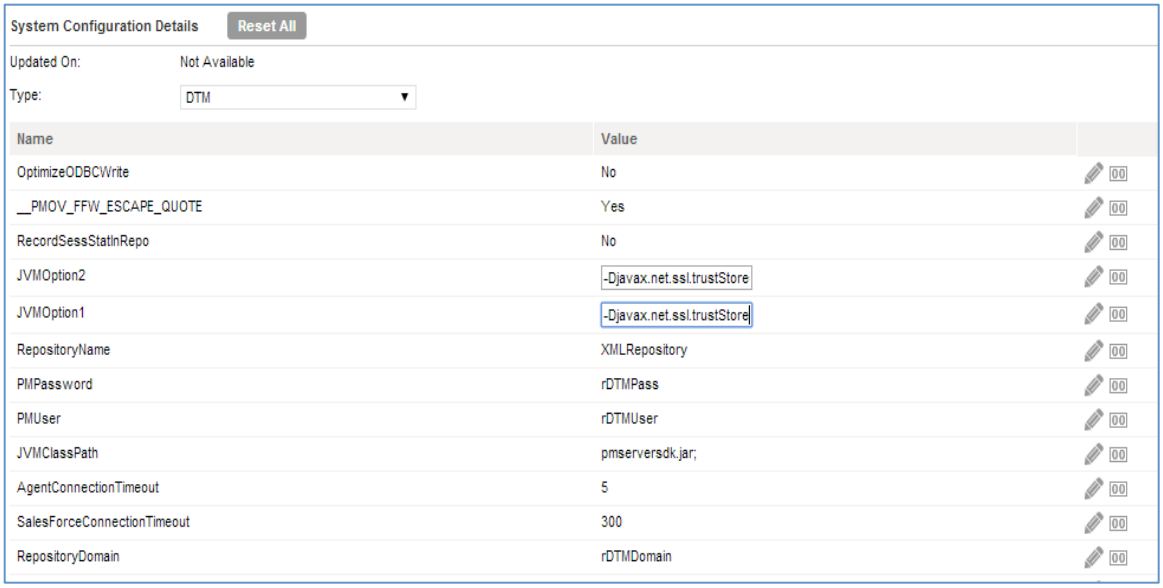
6. Next, click on the edit button to left of your Secure Agent.

7. 

8. In the System Configuration Details drop down box, change the Type to DTM

9. 

10. Add the following to JVMOption1 and JVMOption2: -Djavax.net.ssl.trustStore=<keystore_name> and -Djavax.net.ssl.trustStorePassword=<password>. Here <keystore_name> is cacerts or the keystore you have created manually.

11. 

12. Lastly, add a parameter to the JDBC URL you specified in your Redshift Connection properties, "ssl=true". See example below: jdbc:postgresql://mycluster.xyz789.us-west-2.redshift.amazonaws.com:5439/dev?ssl=true

Redshift Connector Best Practices

When working with the Redshift connector we recommend the following best practices.

1. Follow Amazon's best practices when designing your tables: http://docs.aws.amazon.com/redshift/latest/dg/c_designing-tables-best-practices.html

2. Choose a batch size where the number of batches matches the number of slices in your cluster. Each XL node has 2 slices, each 8XL node has 16. If you have a 2 node XL cluster and 40,000 rows of data, choose a batch size of 10,000. The Informatica Cloud Redshift connector can maximize Amazon's parallel processing capabilities this way.

3. Only use the "upsert" when you know you will be updating rows. Otherwise use the "insert" capability as it will load the data more efficiently.