Implementing a Data Warehouse on AWS in a Hybrid Environment

INFORMATICA CLOUD AND AMAZON REDSHIFT
Contents
Abstract ................................................................................................................................. 3
Before You Get Started ....................................................................................................... 4
Overview of Informatica Cloud .......................................................................................... 5
Amazon Redshift Overview ................................................................................................. 7
Scenario and Solution Overview ........................................................................................ 9
  Getting Started with Redshift and Analytics ................................................................. 9
  Scenario 1: Loading Data from Relational Databases .................................................. 10
  Scenario 2: Integrating Cloud and Transactional Data Sources ............................... 10
Security Considerations .................................................................................................... 12
  Informatica Cloud Agent on premise connecting to Amazon Redshift using https .... 12
  Informatica Cloud Agent on premise connecting to Amazon Redshift using VPN ... 13
  Informatica Cloud Agent on AWS connecting to on premise sources using VPN .... 13
  Informatica Cloud Agent running in AWS connecting to cloud data sources ......... 14
Instance Type Overview .................................................................................................... 15
Hardware and Software Setup for Informatica Cloud Secure Agent ......................... 16
  Hardware and Network Configuration ........................................................................ 16
    Creating the tutorial environment in AWS ............................................................... 16
    Removing the environment from AWS .................................................................... 21
  Informatica Cloud Services Setup ............................................................................. 22
    Registering for Informatica Cloud Amazon Redshift Free Trial ......................... 22
    Installing Informatica Cloud Secure Agent .......................................................... 23
    Reviewing Informatica Cloud Licenses ................................................................... 25
Implementing a Solution ..................................................................................................... 26
  Tutorials Sample Data ................................................................................................. 26
  Tutorials Prerequisites ............................................................................................... 27
  Getting Started with Redshift and Analytics with Informatica Cloud .................. 28
    Tutorial #1 ............................................................................................................... 28
  Scenario 1: Load Data from Relational Databases ..................................................... 44
    Tutorial #2 ............................................................................................................... 44
    Tutorial #3 ............................................................................................................... 53
  Scenario 2: Integrating Hybrid Data Sources ............................................................. 57
    Tutorial #4 ............................................................................................................... 57
Operating Considerations and Best Practices.................................................................68
Informatica Cloud Service Monitoring........................................................................68
Logs and Monitoring ..................................................................................................68
Upgrades & Patching .................................................................................................68
Troubleshooting .......................................................................................................68
Appendix A – SQL Queries ..........................................................................................70
  CREATE TABLE for Oracle ......................................................................................70
  CREATE TABLE for MySQL ....................................................................................71
  CREATE TABLE for SQL Server ..............................................................................72
  CREATE TABLE for Redshift ....................................................................................73
Appendix B – Manual Steps for AWS Environment Creation .......................................76
  Set up the VPC environment ..................................................................................76
  Create Private Subnets ..........................................................................................77
  Create Security Groups .........................................................................................78
  Launch your EC2 instance ......................................................................................79
  Create Redshift Cluster .........................................................................................81
  Create RDS MySQL Instance ..................................................................................82
Abstract

Organizations migrating workloads to the cloud need to solve how to synchronize and move data between on-premise and cloud deployments. This is especially true when migrating or augmenting on-premise business intelligence solutions with cloud based services. Many organizations operate in a hybrid on-premise / cloud mode for extended periods of time for many reasons. A robust design for migrating data, ensuring data quality, and continuous synchronization is essential to success.

This guide provides design and implementation recommendations, an architectural reference, quick start guidance, and tutorials to help get you going with operating in on-premise/cloud environment using Informatica Cloud with AWS services.

Informatica PowerCenter and Informatica Cloud Services provide a single solution for integration data between AWS data stores and a wide variety of traditional on-premise resources and popular cloud-based services like Salesforce, Workday, and Marketo. In this document, we focus on Informatica Cloud capabilities for data integration.

This guide targets IT administrators, data engineers, and business analysts. This document discusses planning topics, architectural considerations and options, and configuration steps for dealing with a hybrid environment that uses Informatica Cloud Services (ICS) and AWS services, such as Amazon Redshift, Amazon Elastic Compute Cloud (Amazon EC2), Amazon Virtual Private Cloud (Amazon VPC), Amazon Simple Storage Service (Amazon S3), and Amazon Relational Database Services (RDS) to run a scalable and reliable managed data warehousing solution.

This guide discusses the planning and architectural considerations in the context of several deployment scenarios. This guide considers several approaches that have tradeoffs in terms of latency, cost, complexity, and security.

This guide also provides sample AWS CloudFormation templates that are designed to help you deploy the necessary and correctly configured infrastructure for the sample scenario in a repeatable and reliable manner.

This document offers step-by-step instructions for the following tasks:

- Creating an Amazon Virtual Private Cloud (VPC) instance.
- Launching a Windows virtual instance with Amazon Elastic Cloud Compute (EC2) service.
- Launching an Amazon Redshift data warehouse cluster.
- Launching an Amazon MySQL Relational Data Service (RDS) instance.
- Setting up Informatica Cloud.
- Loading data from various sources into Amazon Redshift with Informatica Cloud.
- Integrating hybrid data sources with Informatica Cloud.
Before You Get Started

Implementing a warehouse in the AWS cloud with hybrid environment is an advanced topic. If you are new to AWS, see the Getting Started section of the AWS documentation at http://docs.aws.amazon.com/gettingstarted/latest/awsgsg-intro/intro.html. The reference implementation will provide high-level references and links to specific areas of documentation. Your understanding of this guide will be easier if you are already familiar with the following topics:

- Amazon EC2, Amazon S3, and Amazon VPC
- Amazon Redshift
- Informatica Cloud
- Windows Server 2008 or higher or Linux Red Hat
- Windows Server Active Directory and DNS

This document briefly discusses Informatica Cloud Services setup tasks as part of Tutorial #1. It focuses on configuration topics that require careful consideration when you are planning and deploying a warehouse in the AWS cloud with hybrid environment using Informatica Cloud. For Informatica Cloud configuration guidance, functionality, and best practices consult the Informatica Cloud product documentation: https://community.informatica.com/docs/.
Overview of Informatica Cloud

Informatica Cloud is a multi-tenant on-demand subscription service. It delivers a set of cloud data integration and management applications built on the Informatica Vibe platform.

Informatica Cloud includes the following components:

- **Informatica Cloud Service** – hosted application service that stores all task and organization information (metadata). The application metadata is configured through a wizard-based web browser interface.
- **Informatica Cloud Applications** – purpose-built data integration applications, such as Data Synchronization, Data Replication, Contact Validation, Data Assessment, Data Masking, and Mapping Configuration.
- **Informatica Cloud Secure Agent powered by Vibe™** – a small footprint application that enables secure communication across the firewall between the client organization and Informatica Cloud. It is installed on a local machine or on an Amazon EC2 instance. It is a functionally equivalent, run-time version of the enterprise-class Informatica PowerCenter execution component. All Informatica Cloud data integration services use the Secure Agent to get through the firewall to access application, relational database and file sources and targets in the client local area network.
Informatica Cloud provides connectivity to various data sources through their native or generic API. It provides connectivity to several Amazon AWS services through the following AWS native APIs:

- Amazon S3
- Amazon DynamoDB
- Amazon Redshift
- Amazon Relational Database Services
Amazon Redshift Overview

Amazon Redshift is a petabyte-scale clustered data warehouse managed as a service. Redshift delivers fast query performance by using columnar storage technology to improve I/O efficiency and parallelizing queries across multiple nodes. You access Redshift by using standard SQL technology, which allows you to use a wide range of familiar SQL clients and analysis solutions. Most of the common administrative tasks associated with provisioning, configuring, and monitoring a data warehouse are automated with Redshift. Backups to Amazon S3 are continuous, incremental, and automatic. Restores are fast - you can start querying in minutes while your data is spooled down in the background. Enabling disaster recovery across regions takes just a few clicks. For more information about Redshift, see http://aws.amazon.com/redshift/.

You load data into Amazon Redshift in parallel in order to achieve high speed loading of large data sets. The most common way to achieve parallel loading is to place files into the Simple Storage Service (S3), and then issue a copy command. For the best performance, split your load data into multiple files based upon the size of the cluster. Amazon Redshift then assigns each file to nodes in the cluster to operate upon in parallel.

Informatica Cloud automatically manages the complexity of these operations for you when you choose Redshift as a target. Behind the scenes, Informatica Cloud optimizes the load for the cluster size, stages data files to S3 on your behalf, and then issues SQL commands to complete the loading process. The following diagram depicts the load process steps:
For more technical details on Redshift, see
Scenario and Solution Overview

Many customers are migrating applications to the cloud and implementing new functionality first in the cloud for cost savings and agility. These applications are either operated by the customer (typically by using models commonly called IaaS or PaaS), or operated as a service by third parties (commonly called SaaS). Often both on-premise and SaaS applications share common data like customer, product, and order information.

With the adoption of SaaS applications, a lot of data needs to be analyzed very frequently, in conjunction with data from the on-premise systems. Traditional data warehouses have long development cycles because of the set up and provisioning of the on-premise databases. Agile data warehousing in the cloud provide an easy, cost-effective means of augmenting existing environments, deploying development, test, and production environments.

This implementation guide provides insight on managing a hybrid data warehouse with Amazon Redshift and Informatica Cloud for data integration. It focuses on the specific considerations of operating in the hybrid cloud environment using AWS, and is not a general guide on implementing Informatica Cloud solutions. Many of the techniques and topologies described in this guide closely align with other types of problems, like initial and incremental load into a data warehouse, although they are not specific targets for this guide, and therefore the guidance may not be complete.

This guide is meant to help IT administrators and data information developers to get started with Amazon AWS and Informatica Cloud products. Even though we focus on Informatica Cloud as data integration tool, the same implementation concepts apply to Informatica PowerCenter.

Getting Started with Redshift and Analytics

The first scenario is a simple scenario to help get you going. Loading data from flat files into a data warehouse or other systems is a common operation. Very often when you run data transformations on premise, you will read files from a drop location in a file share to perform a load. A similar pattern when using the cloud is to read files out of a location in Amazon Simple Storage Service (S3). In this scenario, you will read data from flat files in S3 and load the data into various targets including Redshift, Relational Database Services, and SaaS applications like Salesforce.
Scenario 1: Loading Data from Relational Databases

Commonly, customers use Informatica Cloud to move data from a transactional system that reside on premise or in AWS into an Amazon Redshift data warehouse. The scenario shows how to pull data from multiple tables and apply filters. Considerations for sizing and topology for running Informatica Cloud appliance on premise or in AWS are also covered.

Scenario 2: Integrating Cloud and Transactional Data Sources

This scenario extends Scenario 1 by integrating data from a SaaS based source. In the example, data originating in Salesforce is merged with data from transactional systems and then pushed into a Redshift data warehouse. Data transformations are also applied.
Security Considerations

Security is of paramount importance to virtually every enterprise. If you run workloads in AWS, then you have already considered security implications. If this is your first venture into the AWS, then security is probably a top consideration. You can run workloads as securely or even more securely in the cloud than you would on premise. AWS addresses many of the common concerns and considerations in the AWS Security Center at http://aws.amazon.com/security and resources listed at http://aws.amazon.com/security/security-resources. Whitepapers papers are available on security processes, best practices, compliance, Microsoft platforms, and HIPAA. The AWS compliance center http://aws.amazon.com/compliance/ lists the commercial and federal certifications that AWS obtained.

There are many different possibilities when configuring your network security. You can choose to either run the Informatica Cloud Agent on AWS, or on premise. When extracting data from systems that reside on premise, systems need to be analyzed. You have several options for how you configure your environment. The following sections outline several common approaches.

Informatica Cloud uses 128-bit Secure Sockets Layer (SSL) technology to protect data. It uses authentication and encryption to ensure that data is secure and available only to users within the organization. When you log in to Informatica Cloud, https precedes the URL in the address field of the web browser to indicate that the connection is secure. See http://trust.informaticacloud.com/security for details.

Informatica Cloud Agent on premise connecting to Amazon Redshift using https

The simplest configuration is to deploy the agent on premise alongside your transactional systems and push the data across the public internet. You run your Amazon Redshift cluster in a Virtual Private Cloud (VPC) that can access the internet. The data from the Secure Agent to the Amazon Redshift data warehouse is encrypted in transit by using https for communications. In order to use this configuration, the Amazon Redshift cluster must also be accessible over the public internet. For higher security, you can configure the security group (firewall) for the Amazon Redshift cluster to only allow connections from the public IP address range of your company. You must allow public IP address ranges originating from the Informatica Cloud to connect to the Secure Agent through your firewall on premise. The public IP address ranges required for access are specified in the Informatica Cloud Users Guide, available at https://community.informatica.com/docs/.
Informatica Cloud Agent on premise connecting to Amazon Redshift using VPN

This is similar to the first case. Only you use a VPN connection from on premise to an AWS Virtual Private Cloud for additional security. You deploy the agent on premise alongside your transactional systems, and the agent pushes data across a virtual private connection established between your on premise data center and your VPC in AWS. The Virtual private connection encrypts data in transit. You can make your Virtual Private Cloud only accessible from your corporate network, and you can also make your Amazon Redshift cluster only accessible using a private IP address. You can also choose to use https for additional encryption across the VPN tunnel.

Informatica Cloud Agent on AWS connecting to on premise sources using VPN

You can deploy the Secure Agent in AWS, and pull the data across a private VPN connection established between your corporate network and the virtual private cloud, which encrypts all data in transit. You
can choose to make your Amazon Redshift data warehouse only privately accessible. You must have public connectivity in order for the Secure Agent to communicate with the Informatica Cloud, and you can restrict the public IP address ranges to the ranges specified in the Informatica Cloud user guide.

Informatica Cloud Agent running in AWS connecting to cloud data sources
When all of the data sources being analyzed originate in the cloud, then often you will run the Secure Agent in the cloud as well. In this case you might choose to omit a VPN connection back to your corporate network, although often this is still a requirement since users analyzing the data still need a secure connection.

For more on AWS security, see https://aws.amazon.com/documentation/vpc/.
**Instance Type Overview**

When you purchase virtual instances from AWS to run your workloads, you specify an instance type. Instance types determine the amount of compute, memory, network capacity, and ephemeral storage you will have available for your workload. Each of these aspects impacts your ability to process and transport data when using EC2 to host the Secure Agent. There are many options for instance types available in AWS for a wide variety of workloads. The most important aspects for Secure Agent performance are computing and network capacity. The C3 instance type family on AWS is tuned for compute intensive workloads while the m3 type provides a balance of performance, memory, and network performance. For smaller workloads, the m3.medium is recommended. The following table describes the minimal and recommended instance types for each type of workload. For information concerning Amazon instance types, see: [http://aws.amazon.com/ec2/instance-types/](http://aws.amazon.com/ec2/instance-types/).

<table>
<thead>
<tr>
<th>Workload Type</th>
<th>Storage Type</th>
<th>Recommended Instance Type</th>
<th>Minimum Instance Type</th>
</tr>
</thead>
<tbody>
<tr>
<td>Tutorials</td>
<td>any</td>
<td>T2.small</td>
<td>T2.small</td>
</tr>
<tr>
<td>Thousands of rows, some transformations</td>
<td>any</td>
<td>T2.medium or M3.medium</td>
<td>T2.medium</td>
</tr>
<tr>
<td>Hundreds of thousands of rows with complex</td>
<td>SSD</td>
<td>C3.large</td>
<td>M3.medium</td>
</tr>
<tr>
<td>transformations, joining of multiple data sources</td>
<td>SSD</td>
<td>C3.2Xlarge</td>
<td>C3.large</td>
</tr>
<tr>
<td>Replication or synchronization of existing data</td>
<td>SSD</td>
<td>C3.2Xlarge</td>
<td>C3.large</td>
</tr>
<tr>
<td>sources, millions of rows, few transformations,</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>lookups, or joins.</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Hardware and Software Setup for Informatica Cloud Secure Agent

The Secure Agent operates either on premise or running in AWS. The Secure Agent runs on either a Windows or Linux operating system. For the purposes of the tutorials, this guide outlines setting up the Secure Agent on Windows. The Informatica Cloud user guide describes how to set up the Secure Agent on a Linux operating system.

There are network and software considerations for using the Secure Agent. How you address the network considerations issues depends upon whether you run the agent on premise or in the cloud. The software considerations are the same in both cases.

Hardware and Network Configuration
There are many topology possibilities, four of which are outlined in the security considerations section. In this tutorial, we will run through how to setup a configuration on Amazon EC2, and provide automation that will help you create this environment. Automation refers to the ability within AWS to describe and script a network and server environment that can be created on your behalf by defining templates and using scripts.

The generic considerations for any hybrid approach are:

1) Providing connectivity from the Secure Agent to the Informatica Cloud
2) Providing connectivity from the Secure Agent to source and target systems
3) Provisioning connectivity between on premise and an AWS Virtual Private Cloud (VPC)
4) Creating a server instance to host the Informatica Cloud Secure Agent. In some simple non-production or limited-production cases, you might choose to run this from your desktop. That scenario is not considered herein.

Often you will want to setup a VPN. For this more advanced setup scenario, go to: http://docs.aws.amazon.com/AmazonVPC/latest/NetworkAdminGuide/Welcome.html.

Creating the tutorial environment in AWS
For information on the basics of setting up a new account with AWS, see http://aws.amazon.com/getting-started/. There are two options to set up your environment. The easiest way to set up the environment is to use a CloudFormation template provided with the tutorial. This template creates an Amazon Virtual Private Cloud (VPC), launches a Windows virtual instance using Elastic Cloud Compute (EC2), launches an Amazon Redshift data warehouse cluster, and then launches a MySQL RDS database. This is the easiest way to offset up the AWS environment since it is all automated.

If you want to perform the setup manually, see Appendix B. For simplicity, this setup does not configure a VPN connection.
In order to run the Cloud Formation script, you will need to have a key pair. A key pair is used to either access a Linux instance using a secure shell (SSH) or to retrieve a password for a Windows instance. Follow the instructions at [http://docs.aws.amazon.com/gettingstarted/latest/wah/getting-started-create-key-pair.html](http://docs.aws.amazon.com/gettingstarted/latest/wah/getting-started-create-key-pair.html) to create a key pair if you don’t already have one available.

1) Launch the Cloud Formation template: Tutorial Template.
2) Click “Next.”
3) Enter in the availability zones and your key pair name. You can optionally change the default user name and password for the MySql and Redshift instances (master/Password123). Availability zones are associated to a region. For more information, see [http://docs.aws.amazon.com/AWSEC2/latest/UserGuide/using-regions-availability-zones.html](http://docs.aws.amazon.com/AWSEC2/latest/UserGuide/using-regions-availability-zones.html)

4) Click “Next”. Click “Next” again for the options page.
5) Click “Create”.
6) Go to the CloudFormation dashboard to view the status of stack creation. When completed the status will change from “CREATE_IN_PROGRESS” to “CREATE_COMPLETE” for:
   a. “Informatica-Tutorial” stack
   b. “Informatica-Tutorial-NetworkStack” stack
   c. “Informatica-Tutorial-Secure Agent” stack

When completed, the CloudFormation template will have created the environment described in the following diagram:
This environment has configured the Secure Agent Security Group to permit access for Informatica Cloud to control the Secure Agent that you will install on the Windows instance for you. Instances running in the public DMZ are accessible from the Internet through the Internet Gateway. The private subnet containing the RDS MySQL database and Amazon Redshift cluster cannot be reached from the public internet directly. You can reach these instances from the Secure Agent windows instance. Typically you will also configure a VPN. Once the VPN is configured, you can reach the instances in the private subnet from your corporate network. If you are using the approach described in the section Informatica Cloud Agent on premise connecting to Amazon Redshift using https, then the Amazon Redshift cluster and RDS Database should be created in the public subnet.

To log into the Secure Agent, perform the following steps:

1) From the EC2 console, select the instance named “Secure-Agent-Tutorial” in your instance list.
2) Before logging into your instance using a Remote Desktop, select the instance from the EC2 dashboard, and then click “Connect”:

3) Click on “Get Password”. You will need to provide the location of the key file that you provided when you created the environment, which is used to decrypt your password.
4) After you have retrieved your password, click “Download Remote Desktop File”, and click on the downloaded file.

5) Start Remote Desktop Connection to log in to the instance using the decrypted password.

The EC2 instance has SQL Server 2012 Express installed. You need to change its Server Authentication mode:

1. Start the Microsoft SQL Server Management Studio (C:\Program Files (x86)\Microsoft SQL Server\110\Tools\Binn\ManagementStudio\Ssms.exe).
2. In the login window, enter “localhost” for the “Server Name”.
4. Click “Connect”.
5. In the Object Explorer panel, right-click on “localhost” and then click “Security”.
6. Change Server Authentication option to “SQL Server and Windows Authentication mode”.
7. Go to Security > Logins and right-click on “sa”.
8. Select “Properties” and then “Status”.
10. Go to the “General” page and create a new password for the “sa” login.
11. Restart the SQL Server. From “localhost” list select “Restart”.
12. Test your changes. Connect to the SQL Server as “sa” with “SQL Server Authentication” option.

Install SQL clients for MySQL and PostgreSQL onto the EC2 instance for the tutorial. Three free options available are:

- SQL Workbench (see http://docs.aws.amazon.com/redshift/latest/mgmt/connecting-using-workbench.html).
- Aginity workbench (see http://www.aginity.com/workbench/redshift/).
- PremiumSoft Navicat Lite 10 (see http://premiumsoft-navicat-lite.software.informer.com/10.0/).

To view the MySQL RDS instance launched by the stack and get its connection details, go to Amazon RDS dashboard. Make a note of the following database properties:

1. Host – from the RDS endpoint. For example, “ics-tutorial-db.xxxxxx.us-west-2.rds.amazonaws.com”.
2. Port – from the RDS endpoint. Default is “3306”.
3. User name and password. Default is master/Password123.
4. Database Name. Default is “icstutorial.”
5. Schema. Default is “public.”

Test the connection from your SQL client installed on your EC2 instance.

To view the Redshift database instance launched by the stack and get its connection details, go to the Amazon Redshift dashboard. Make a note of the following database properties:
1. Number of Nodes
2. Endpoint
3. Port – default is “5439”
4. Database Name – default is “dev”
5. JDBC URL

To complete the tutorials, you also need:

1. The Redshift database user name and password – default is master/Password123
2. AWS account Access Key
3. AWS account Secret Key

Test the connection from your SQL client installed on your EC2 instance.

You can now proceed to Informatica Cloud Services Setup to finish the setup.

Removing the environment from AWS

To clean up the tutorial environment in AWS created from the CloudFormation tutorial script:

1. From AWS Console, click on “CloudFormation” on the left under “History”.
2. Select “Informatica-Tutorial” stack.
3. Click “Delete Stack”.
4. Click “Yes, Delete”.
5. The status changes from “CREATE_COMPLETE” to “DELETE_IN_PROGRESS”.
6. The “Informatica-Tutorial” stack deletion process will automatically delete two other child stacks.
7. When the deletion process completes, the three tutorial stacks will disappear from the list.

**Note:** If you made any modifications to the Tutorial VPC properties (for example, the tutorial Security Group), you need to remove the resources manually.
Informatica Cloud Services Setup
Use the following instructions to set up Informatica Cloud on Windows. The setup process for Linux is quite similar. See details on Linux in *Informatica Cloud User Guide (Informatica Cloud Secure Agent chapter)*.

Registering for Informatica Cloud Amazon Redshift Free Trial
To register for a free 60-day trial of Amazon Redshift with Informatica Cloud, go to [https://community.informatica.com/solutions/infa_cloud_trial_for_amazon_redshift](https://community.informatica.com/solutions/infa_cloud_trial_for_amazon_redshift). Click on “CLICK HERE TO START 60 DAY TRIAL”. You will be prompted to log in to Informatica Passport.

If you already have an account with Informatica Cloud, the following message appears:

![Informatica Marketplace Image]

To contact Informatica Support team for the free trial license activation, click on “Click here”.

To sign up for a new account, click on “Sign up now”:  

![Login Image]
After successful Informatica Passport account creation, you will be prompted for Informatica Cloud registration:

Log in to Informatica Cloud by clicking on “Click here”:

Or log into Informatica Cloud from https://app.informaticaondemand.com/ma.

Installing Informatica Cloud Secure Agent

1. Log into your ICS account and install Informatica Cloud Secure Agent from Configure -> Secure Agents page. Select “Windows” platform:

2. When the Secure Agent installation is complete, you will be prompted to register it. Enter your ICS account name and password.

3. Verify that the secure agent has Status set to “Active”. The Secure Agent name will display your computer (EC2 instance) name:

![Secure Agents](image)

4. To access Informatica Cloud online help, click on the question mark button on the right-side of the window:

![Informatica Cloud Online Help](image)

5. To view the navigation panel, click on the folder icon in the upper-left corner of the online help page:
Reviewing Informatica Cloud Licenses

With a new Informatica Cloud account, you have a 60-day trial license for Amazon Redshift connector and a 30-day trial license for other ICS functionality.

Review Informatica Cloud license list (Administer -> Licenses). Amazon Redshift connection license is listed under Connector Licenses:

The expiration date is 60 days from your registration date.
Implementing a Solution
This chapter will guide you through developing data warehouse integration solutions with Informatica Cloud. It has four step-by-step tutorials:

**Tutorial #1** – Design and execute data synchronization tasks reading from a single data source:
1. Create data connections.
2. Read data from CSV files, RDMBS tables, and Redshift tables.
3. Load data into RDBMS tables, Redshift tables, and Salesforce objects.
4. Dynamically create tables in Redshift database.

**Tutorial #2** – Design and execute data synchronization tasks to read data from multiple tables in the same source.

**Tutorial #3** – Design and execute a task flow.

**Tutorial #4** – Implement a mapping task to read from data sources on-premise and in the cloud.

**Tutorials Sample Data**
The tutorials are based on Amazon Redshift TICKIT sample database described in the Redshift Database Developer Guide at [http://docs.aws.amazon.com/redshift/latest/dg/c_sampledb.html](http://docs.aws.amazon.com/redshift/latest/dg/c_sampledb.html). The tutorial consists of a subset of the TICKIT tables and few tables derived from them:

- ARS_VENUE – based on TICKIT VENUE
- ARS_DATE – based on TICKIT DATE
- ARS_CATEGORY – based on TICKIT CATEGORY
- ARS_EVENT – based on TICKIT EVENT
- ARS_EVENT_WIDE – join of ARS_VENUE, ARS_DATE, ARS_CATEGORY, and ARS_EVENT tables
- ARS_EVENT_SUMMARY – aggregation table
Tutorials Prerequisites

The tutorials assume that you have:

1. Access to Amazon Redshift data warehouse service.
2. Your S3 and Amazon Redshift services are in the same region.
3. Access to at least one RDBMS system (Oracle, SQL Server, or MySQL) with Amazon RDS service or installed locally.
4. Active Informatica Cloud account.
5. Installed and registered Informatica Cloud Secure Agent.

See Hardware and Software Setup chapter for details.

Before you start the tutorials:

1. Download and install Postgres JDBC driver from:
   The JDBC driver is for reading data from your Amazon Redshift database with Informatica Cloud.

2. With the SQL Client tools, connect to your SQL Server database, Amazon Redshift, or Amazon MySQL RDS and run the DDL scripts (see Appendix A). It will create the following tables:
   - ARS_VENUE
   - ARS_DATE
   - ARS_CATEGORY
   - ARS_EVENT
   - ARS_EVENT_WIDE
   - ARS_EVENT_SUMMARY

3. Log in to your Salesforce account and create a new custom object. See https://www.salesforce.com/us/developer/docs/apexcode/Content/apex_qs_customobject.htm for details on how to create a custom object:

   Object Name = ars_category

   API Name = ars_category__c

   Custom Fields:

<table>
<thead>
<tr>
<th>Field Label</th>
<th>API Name</th>
<th>Data Type</th>
<th>Is External ID</th>
<th>Is Unique</th>
</tr>
</thead>
<tbody>
<tr>
<td>catdesc</td>
<td>catdesc__c</td>
<td>Text(50)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>catgroup</td>
<td>catgroup__c</td>
<td>Text(10)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>catid</td>
<td>catid__c</td>
<td>Number(18,0)</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>catname</td>
<td>catname__c</td>
<td>Text(10)</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Getting Started with Redshift and Analytics with Informatica Cloud

Several Informatica Cloud data integration services support Amazon Redshift database as end point.

Informatica Cloud Data Synchronization application (provides a 6-step wizard to quickly configure data synchronization tasks between various sources and your Redshift database:

In this section, we describe the steps to start using Informatica Cloud. We focus on the Data Synchronization application and its capabilities to implement basic data movement requirements.

Tutorial #1
Time to complete: 90 min.

One of common data migration requirements is to replicate data from a source A to a target B. It is also common that you do not have your target objects defined. With the Informatica Cloud Data Synchronization application, you can dynamically create a target table in Redshift database based on the source of data.

Step 1: Creating Connections in Informatica Cloud
Informatica Cloud has numerous connection types, such as Oracle, SQL Server, MySQL, Redshift, ODBC, JDBC, file, Salesforce, SAP, and NetSuite. See a list of all supported connections in Informatica Cloud User Guide. In Step 1 of the tutorial, we will create the following ICS connections:

- MySQL database
- SQL Server database
- Amazon Redshift database
- Amazon S3
- JDBC
- Salesforce

Create MySQL connection.

The database can be your local database or an Amazon RDS instance. In the tutorials, we use MySQL connection type:
1. Log in to your Informatica Cloud account.
2. Go to Configure -> Connections page and click “New”.
3. Select “MySQL” for “Type”.
4. Fill in the rest of the form. See Creating the tutorial environment in AWS section about how to get connection details for your Amazon MySQL RDS:

5. Test the connection:

6. Click OK to save and exit.

Create a connection for SQL Server database.

It can be your local database or an Amazon RDS instance:

1. Go to Configure -> Connections page and click “New”.
2. Select “SQL Server” for “Type”.
3. If you have more than one Secure Agent associated with your ICS account, select the Secure Agent.
4. Fill in the rest of the form. See Creating the tutorial environment in AWS section about your EC2 instance SQL Server database connection details:
5. Test the connection:

6. Click “OK” to save and exit.

Create a Redshift connection:

1. Log into your AWS account and go to Redshift service page.
2. Go to your cluster configuration page and make a note of the cluster and cluster database properties:
   a. Number of Nodes
   b. Endpoint
   c. Port
   d. Database Name
   e. JDBC URL

You also will need:
• The Redshift database user name and password. It is different from your AWS account.
• AWS account Access Key
• AWS account Secret Key

3. Exit AWS console.
4. Log into your Informatica Cloud account.
5. Go to Configure -> Connections page and click “New”.
6. Select “AWS Redshift (Informatica)” for “Type” and fill in the rest of the form:

   ![Informatica Cloud](image)

   **New Connection**

   **Connection Details**
   - **Connection Name**: aws_redshift
   - **Description**: 
   - **Type**: AWS Redshift (Informatica)

   **AWS Redshift Connection Properties**
   - **Secure Agent**: CAW181232
   - **Username**: sa
   - **Password**: ****
   - **Schema**: public
   - **AWS Access Key ID**: ****
   - **AWS Secret Access Key**: ****
   - **Cluster Node Type**: XL
   - **Number of Nodes in the Cluster**: 2
   - **Jdbc URL**: jdbc:postgresql://redshift.ics.customer.com:5432/sa

7. Test the connection:

   ![Informatica Cloud](image)

   **New Connection**

   **Test**

   **The test for this connection was successful.**

8. Click “OK” to save and exit.

Create an Amazon S3 connection:

1. Go to Configure > Connections page and click “New”.
2. Enter “Connection Details”:
   a. Select “AmazonS3 (Informatica Cloud)” for “Type”.
   b. If you have more than one Secure Agent associated with your ICS account, select the Secure Agent.
   c. Enter your AWS account access key and secret key.
d. For “Regions” select “US_WEST_2”.

e. For “Folder Path” enter “informatica-sample” (S3 bucket name).

f. Type comma (“,“) as “Delimiter” and fill in the rest of the form:

3. Test the connection:

4. Click “OK” to save and exit.

Create a JDBC connection:

1. Go to Configure >Connections page and click “New”.
2. Select “JDBC_IC (Informatica Cloud)” for “Type”.
3. If you have more than one Secure Agent associated with your ICS account, select the Secure Agent.
4. Enter “JDBC Connection URL”, which is obtained when reviewing your Redshift database properties.
5. For “Jdbc Driver”, enter the location of JDBC jar file you downloaded.
6. Fill in the rest of the form:

![Edit Connection](image)

- **Connection Name**: ars_JDBC
- **Description**: 
- **Type**: JDBC_IC (Informatica Cloud)
- **Secure Agent**: WIN-550P8NH3RM
- **JDBC Connection URL**: jdbc:postgresql://cloud-vip-redshift.ogbrom2wvhc.us
- **JDBC Jar Directory**: ojdbc6
- **Username**: user_name
- **Password**: ********

7. Test the connection:

![Test the connection](image)

8. Click “OK” to save and exit.

Create a Salesforce connection:

1. Go to **Configure > Connections** page and click “New”.
2. Select “Salesforce” for “Type”.
3. Fill in the rest of the form:

![Edit Connection](image)

- **Connection Name**: ars_Salesforce
- **Username**: your Salesforce login
- **Password**: ********
- **Security Token**: ************
- **Service URL**: https://www.salesforce.com/services/SoapUI/38.0
Note: Salesforce connection does not require association with the Secure Agent.

Step 2: Creating Your First Informatica Cloud Data Synchronization Tasks

Create tasks to load an RDBMS table from a sample CSV file:

1. Log in to your ICS account.
2. Go to Apps > Data Synchronization:

3. Definition. Enter task name and select “Insert” for “Task Operation” and click “Next”:
4. **Source.** Select “ars_S3” for “Connection” and “ars_category__csv” as “Source Object”. In “Data Preview”, you will see the file content. Click “Next”:

5. **Target.** Select your RDBMS connection name for “Connection”. In our example, we use MySQL database with “ars_MySQL” connection name. Select “ars_category” as “Target Object”. Select “True” from “Truncate Target” drop-down list. Click “Next”:
6. **Data Filters.** You have a choice to define a filter for the source data. Click “Next”:

7. **Field Mapping.** Define source-to-target mapping rules. The wizard automatically maps ports with the same name. Click “Next”:

8. **Schedule.** Configure the task execution options. Leave it as-is:
9. From the “Save” drop-down list, select “Save and Run”. Click “OK” on “Message from webpage”:

10. After the task execution starts, “Activity Monitor” page opens. As soon as the execution ends, you will see the status execution:

11. When the task execution completes, you will see “Activity Log” page:

12. Click on the task name to review task execution statistics. Click “Done” when you are done:

Create more data synchronization tasks.
1. Go to *Apps > Data Synchronization* and click on “Copy” icon under “Actions” header next to your first task name. It will create a copy of your task. Click “OK” on “Message from webpage”:

![Message from webpage](image)

2. Enable the copied task edit mode by clicking on “Pencil” icon next to the copy name:

![Data Synchronization Tasks](image)

3. The wizard **Definition** page appears. Rename the task as “ars_Load_RDBMS_Date” and click “Next”:
4. **Source.** Select “ars_date__csv” from “Source Object” drop-down list. Ignore the warnings and click “Next”:

![Source Object Selection](image)

5. **Target.** Select “ars_date” from “Target Object” drop-down list. Ignore the warnings and click “Next”:

![Target Object Selection](image)

6. **Data Filters.** Keep default settings (no filters).

7. **Field Mapping.** Click on “5 Field Mapping” header to view port mapping. Notice that both source and target ports are refreshed as per new source and target. Click “Clear Mappings” and
then “Automatch”:

8. From the “Save” drop-down list, select “Save and Run”. Wait till task execution completes and review the task execution details:

9. Create a task to load Event data from S3 into the second RDBMS (SQL Server):
   a. Task Name = ars_Load_RDBMS_Event
   b. S3 Source Object = ars_event
   c. SQL Server Target Object = ars_event

10. Create a task to load Venue data from S3 into the second RDBMS (SQL Server):
    a. Task Name = ars_Load_RDBMS_Venue
    b. S3 Source Object = ars_event
    c. SQL Server Target Object = ars_venue

11. Create a task to load Category data from S3 into Salesforce:
    a. Task Name = ars_Load_Salesforce_Category
    b. S3 Source Object = ars_category
    c. Salesforce Target Object = ars_category (label = ars_category)

**Step 3: Dynamic table creation and load with Informatica Cloud DSS Tasks**

When your source is an RDBMS table or an application object, for example, Salesforce, you can configure a DSS task to dynamically create a target object and load it.

Create a data synchronization task to dynamically create and load a Redshift database table from the RDBMS table:

1. Log into your ICS account.
2. Go to Apps -> Data Synchronization > New:

3. **Definition.** Enter task name and select Insert for “Task Operation” and click “Next”:

4. **Source.** Select `ars_MySQL` for “Connection” and “`ars_category`” as “Source Object”. Leave “Source Type” default selection of “Single”. In Data Preview, you will see the table content. Click “Next”:
5. **Target.** Select “ars_Redshift” for “Connection”. Click on “Create Target” to create a Redshift table target. Name the new table “new_category”. By default, all the source object fields are selected for the target creation:

![Create Target window](image)

Specify the name of the target object to be created. You can also select fields from the source object to include in the target object. All source fields are included by default.

- **Connection:** ars_Redshift
- **Source Object:** ars_category
- **Target Object Name:** new_category

- **Source Fields:**
  - catid
  - catgroup
  - catname
  - catdesc

- **Target Fields:**
  - catid
  - catgroup
  - catname
  - catdesc

6. **Field Mapping.** Notice that all the ports are automatically mapped. Click “Next”:

![Field Mapping window](image)

7. **Schedule.** Scroll down to “Advanced Target Properties”. Enter your Amazon S3 account bucket name. Enter a staging location full path name. The Informatica Cloud Secure Agent will use this location to split the task data load into multiple files (per node per node slice) for performance
optimization. Check the box next to “Truncate Target table Before Data Load”:

8. From the “Save” drop-down list select “Save and Run”.

9. After the task execution completes go to “Activity Log” page and click on the task name:

10. Download the session log by clicking on “View Session Log” link. Open the log and search for “*****START LOAD SESSION*****” line:

11. Review the syntax for COPY command generated by the Secure Agent.

**Step 4: Reading data from your Redshift tables**

Occasionally, you will need to extract data from your Redshift tables. It can be accomplished with Informatica Cloud data synchronization tasks. The Informatica Cloud Redshift connector is designed to
write into a Redshift database. For reading from the Redshift database tables, you will use an Informatica Cloud JDBC or ODBC connection to do the job.

Create a data synchronization task to read from a Redshift table and write to a Salesforce object:

- Task Name = ars_ReadFromRedshift
- Operation = Insert
- Source Connection = ars_JDBC
- Source Object: = new_category
- Target Connection = ars_Salesforce
- Target Object = ars_category
- Field Mapping:

Select “Save and Run” to save and execute the task.

**Scenario 1: Load Data from Relational Databases**

In this chapter, we will guide you through creating more complex data integration processes with the Informatica Cloud Data Synchronization application. We also introduce you to Informatica Cloud task flow functionality.

**Tutorial #2**

Time to complete:  40 min.

Data integration processes frequently require data from multiple objects that reside in a single or multiple data sources. In this tutorial we learn how to configure Informatica Cloud data synchronization tasks to read from multiple tables of a single database.

With Informatica Cloud you have a choice to load data into Redshift tables as:

- Insert
- Insert Else Update (Upsert)
For high volume initial data migration Insert operation is appropriate. For incremental data synchronization, we recommend selecting `Upsert` operation.

You must complete Tutorial #1 to proceed.

**Step 1: Create a DSS task for the initial data load**

1. Log in to your Informatica Cloud account.
2. Go to `Apps -> Data Synchronization -> New` to create a new task.
3. **Definition**. View “Task Operation” options:

   ![Image of Data Synchronization Task Wizard](image)

   **Note:** Only “Insert” and “Upsert” operations are currently supported for Redshift tables. If you select “Update” or “Delete” the task execution will fail.

4. Select “Insert” as “Task Operation” and click “Next”.
5. **Source.** As “Connection” select your RDBMS connection. We use “ars_MySQL” as an example. Select “Multiple Source Type”. The wizard page will change:

   ![Image of Data Synchronization Task Wizard](image)
6. Click on “Add” and select all four tables:

7. Click on “User Defined Join” to create ”Join Condition”. Select individual objects and then choose the join fields:

```sql
ars_event.venueid = ars_venue.venueid and 
ars_event.catid = ars_category.catid and 
ars_event.dateid = ars_date.dateid
```
8. **Target.** As “Connection” select “ars_Redshift” and as “Target Object” select “ars_event_wide”:

![Data Synchronization Task Wizard (ars_MultiTableSource)](image)

9. **Data Filters.** Click on “New” to create a data filter. Choose “caldate” from “ars_date” and “Less Than” operator. Type in “2008-07-01 00:00:00” (without quotes) as condition value:

![Data Filter](image)
10. **Field Mapping.** Edit auto-matched mapping for “eventname” target field. Click on “Function” action button next to the field name. Explore Informatica Cloud available functions:

11. Apply a string function of your choice to “eventname” source field:

12. **Schedule.** Enter values for S3 bucket name and the staging location.
13. Click on “Save” and select “Save and Run”.


14. Review the task execution log. The Informatica Cloud Secure Agent generates an optimized SQL statement for the source that includes both join and filter conditions:

```
SELECT ARS.ars_category.catid, ARS.ars_category.catgroup, ARS.ars_category.catname, ARS.ars_category.catdesc, ARS.ars_date.holiday, ARS.ars_date.caldate, ARS.ars_date.day, ARS.ars_date.month, ARS.ars_date.qtr, ARS.ars_date.dateid, ARS.ars_date.year, ARS.ars_date.week, ARS.ars_event.starttime, ARS.ars_event.eventname, ARS.ars_event.eventid, ARS.ars_event.venueid, ARS.ars_venue.venuecity, ARS.ars_venue.venuesats, ARS.ars_venue.venuestate, ARS.ars_venue.venuename FROM ARS.ars_category, ARS.ars_date, ARS.ars_event, ARS.ars_venue WHERE ars_event.venueid = ars_venue.venueid and ars_event.catid = ars_category.catid and ars_event.dateid = ars_date.dateid AND ('ars_date'.'caldate' < str_to_date('2008-07-01 00:00:00', '%Y-%m-%d %H:%i:%s'))
```

Step 2: Create a data synchronization task for the incremental data load

1. Go to Apps -> Data Synchronization. On the task list page, make a copy of the task created in Step #1 (“ars_MultiTableSource”) and click on “Edit” icon (Pencil) next to the copy task name.
2. **Definition.** Rename the task to “ars_MultiTableSource_Delta” and choose “Upsert” as “Task Operation”:

![Image of Data Synchronization Task Wizard](image)

3. **Data Filters.** Click on the filter definition under “Filter by” to edit the filter and then click on “Advanced”. Click “OK” on the warning message. Edit the filter expression:

```sql
"ars_date"."caldate" >= str_to_date('2008-07-01 00:00:00', '%Y-%m-%d %H:%i:%s') and "ars_date"."caldate" < str_to_date('2008-10-01 00:00:00', '%Y-%m-%d %H:%i:%s')
```
4. Click on “Save” and choose “Save and Run”.
5. Review the task execution log. When loading into Redshift tables with Upsert operation the Secure Agent generates recommended by Redshift merge approach. First it creates and loads a temporary table in the Redshift database first and then it executes update on the target table:

```sql
CREATE TABLE public.ars_eventwide_table100001 (
    eventid int4 NOT NULL,
    venuid int4 NOT NULL,
    catid int4 NOT NULL,
    datid int4 NOT NULL,
    eventname varchar(100),
    venuename varchar(100),
    Ars_category varchar(100),
    venuestate char(2),
    venuestate_int4,
    catgroupvarchar(10),
    catname varchar(10),
    catdesc varchar(50),
    datadate NOT NULL,
    daychar(3) NOT NULL,
    monchar(3) NOT NULL,
    yearchar(3) NOT NULL,
    hollydaybool
);
```

**Step 3: Create Saved Query object**
To automate your daily incremental data synchronization tasks, you need to define a dynamic date calculation as a filter. Every database has its own set of date manipulation functions. You define your custom filtering condition with DSS task Saved Query source option.

1. Go to Configure -> Saved Queries and click on “New”.
2. Select “MySQL” as “Database Type” and copy/paste SQL query from Step #1 (task execution log).
3. Edit the SQL Query filter to include the database date manipulation functions so that start and end date of the extraction period is dynamically calculated. For MySQL database, the syntax to get records created since yesterday:

   ```sql
   <timestamp column> >= DATE_FORMAT(ADDDATE(SYSDATE(), interval -1 DAY), '%Y-%m-%d') AND
   <timestamp column> <= DATE_FORMAT(SYSDATE(), '%Y-%m-%d')
   ```

Since our static data is static, we will calculate ADDDATE () interval dynamically based on “2008-10-01” as extraction period start date:

```sql
SELECT ARS.ars_category.catid, ARS.ars_category.catgroup,
ARS.ars_category.catname, ARS.ars_category.catdesc,
ARS.ars_date.holiday, ARS.ars_date.caldate, ARS.ars_date.day,
ARS.ars_date.month, ARS.ars_date.qtr, ARS.ars_date.dateid,
```
ARS.ars_date.year, ARS.ars_date.week, ARS.ars_event.starttime, ARS.ars_event.eventname, ARS.ars_event.eventid, ARS.ars_event.venueid, ARS.ars_venue.venuecity, ARS.ars_venue.venueseats, ARS.ars_venue.venuestate, ARS.ars_venue.venuename FROM ARS.ars_category, ARS.ars_date, ARS.ars_event, ARS.ars_venue WHERE ars_event.venueid = ars_venue.venueid and ars_event.catid = ars_category.catid AND ars_event.dateid = ars_date.dateid AND ars_date.caldate >= DATE_FORMAT(ADDDATE(SYSDATE(), interval - DATEDIFF(SYSDATE(), '2008-10-01') DAY), '%Y-%m-%d') AND ars_date.caldate <= DATE_FORMAT(SYSDATE(), '%Y-%m-%d')

4. Click on “Get Columns” to test the query and create the metadata for the query:
Step 4: Create a data synchronization task with ‘Saved Query’ source option

1. Go to Apps -> Data Synchronization. On the task list page, create a copy of task from Step #2.
2. **Definition.** Rename the task as “ars_MultiTableSource_Delta_Query”.
3. **Source.** Select “Saved Query”:

   ![Data Synchronization Task Wizard](image)

   **Source Details**
   - **Connection**: ars_mysql
   - **Source Type**: Saved Query
   - **Source Object**: ars_SavedQuery

4. **Data Filters.** No filters.
5. **Field Mapping.** Click on “Clear Mapping” and then on “Automatch”.
6. Edit “eventname” field mapping. Click on Function icon next to the field name. Enter “INITCAP(eventname)”.
7. Select “Save and Run”.

**Tutorial #3**

Time to complete: 20 min.

Data warehouses require periodic updates to be in sync with its data sources. After you develop individual data synchronization tasks, you need to automate their periodic execution.

You can enable Informatica Cloud task scheduling option for individual tasks. When there is data load dependency use Informatica Cloud task flows.

You will create a task flow with two dependent data synchronization tasks:

1. The “ars_MultiTableSource_Delta” task must run first.
2. The “ars_MultiTableSource_Delta_Query” task must run after the first task completes successfully.

You must complete Tutorial #1 and Tutorial #2 to proceed.

Step 1: Create a task flow
1. Log in to your Informatica Cloud account.
2. Go to Design -> Task Flows and click on “New”:

3. The “Task Flow” page “Scheduling” section is similar to Schedule page of the DSS task wizard. Select “Run this task on schedule” and then “New”. Explore Informatica Cloud scheduler features:
4. Click on “Cancel” and then select “Do not run this task on a schedule”.
5. Click on “Add Task” button. Select “Data Synchronization” as “Task Type” and “ars_MultiTableSource_Delta” as “Task”: 
6. Click OK to add the task and then check “Stop on Error” task option:

7. Add the “ars_MultiTableSource_Delta_Query” task. Name the task flow “ars_MyFirstTaskFlow”:

8. Click “OK” and start the task flow clicking on “Run Now” icon next to the task name:

9. “Activity Monitor” page will appear. Click on “Detail View” button to see the task execution progress details:
10. After the task flow completes, review its execution details:

11. To view task execution individual logs, click on “Log” icon to the left of “Status” column.

Scenario 2: Integrating Hybrid Data Sources
In the Scenario #2 tutorials, we developed several simple data integration processes reading data from a single source type. A typical data warehouse might require combining data from heterogeneous sources and support of complex data transformations.

To access data from various sources Informatica Cloud offers:

- Template Integration Configuration – develop integration tasks with Informatica Cloud Integration Template Designer.
- Power Center – develop PowerCenter tasks with PowerCenter Designer.

See Informatica Cloud User Guide for details on these services.

Tutorial #4
Time to complete: 90 min.

In this tutorial we will develop and execute a mapping with Informatica Cloud Mapping Designer. We will implement the following data warehouse requirement:

1. Maintain an event summary table in the Redshift data warehouse (ARS_VENUE_SUMMARY).
2. Data sources for aggregation are tables from two heterogeneous RDBMS instances (SQL Server and MySQL) and a Salesforce custom object.
3. Calculated attributes in the summary table are: venue size, year-month name, and year-quarter name.

You must complete Tutorial #1 to proceed.

Step 1: Design a mapping with Informatica Cloud Mapping Designer
1. Log in to your Informatica Cloud account.
2. Go to Design -> Mappings and click on “New Mapping”. Name a mapping as “ars_EventSummary_Mapping” and press “OK”:

![Mapping Designer](image)

3. Drag and drop the following transformation objects from the Transformation palette into the canvas:
   a. 2 Sources
   b. 1 Joiner
   c. 1 Lookup
   d. 1 Aggregator
   e. 2 Expressions
   f. 1 Target

![Transformation Objects](image)

4. Connect the objects as follows:
5. Rename the objects:

Note: Always connect a source with smaller data set to the Master group of a Joiner transformation. In our example, `ars_event` with the most of data residing in the SQL Server database.

Click on “Save” and select “Save and Continue”.

Step 2: Configure the mapping

1. Source Object “src_Salesforce”:
   a. Connection = `ars_Salesforce`
   b. Source Type = `Object`
   c. Leave Allow Multiple unchecked
   d. Object -> Select Object = `ars_category__c`

2. Source Object “src_SQLServer”:
   a. Connection = `ars_SQLServer`
   b. Source Type = `Object`
   c. Check “Allow Multiple”
   d. Add Object = `arc_event`
   e. Add Related Objects for Custom Relationships = `ars_venue`

When you click on “Allow Multiple” checkbox, the page will change to:
To add an object, click on the down arrow on the left and feel the fields in the “Select Related Objects” form:

f. Your source properties should look like the following:

![Source Type: Object](image)

- **Source Object**: ars_event
- **Source Key**: venuemld
- **Join Type**: Inner Join
- **Join Operator**: = (Equals)
- **Related Object**: ars_venue
- **Object Key**: venuemld

![Select Related Objects](image)

![OK Cancel](image)

g. Change “dateid” field name to “dateid1”: 
Note: This step is necessary to ensure that multiple source ports have unique names.

3. Expression transformation “exp_DTConversion”:
   a. In the Expression dialog box, click on “Actions” drop-down list (on the right) and select “New Field”. Create a new field “sf_catid” of data type “integer”:

   ![Expression dialog box with new field creation](image)

   b. Configure the new field. Map it to “catid__c”:

   ![Expression configuration](image)

   Note: This expression converts “catid__c” of data type “decimal(18,0)” to data type “integer”.

4. Joiner transformation “jnr_InnerJoin”:
   a. Create Join Condition as “sf_condid = catid”:

   ![Joiner configuration](image)
5. Lookup transformation “lkp_MySQL”:
   a. Connection = ars_MySQL
   b. Source Type = Object
   c. Lookup Object –> Select Object = ars_date
   d. Lookup Condition = dateid = dateid1

6. Aggregator transformation “agg_EventCount”:
   a. Incoming Fields – Named Fields + Configure:
      b. Select the following fields:
         • catdesc__c
         • catgroup__c
         •catid
         • catname__c
         • eventid
         • month
         • qtr
         • venuecity
         • venueid
         • venuename
         • venues
         • venuestate
c. Add all selected fields except “eventid” to “Group By”:

```
- Field Name: caddese_c, caddes_c, caddes__c, eventid, eventname__c, year
```

```
<table>
<thead>
<tr>
<th>Field Name</th>
<th>Origin</th>
</tr>
</thead>
<tbody>
<tr>
<td>caddese_c</td>
<td>caddes_c</td>
</tr>
<tr>
<td>caddes_c</td>
<td>caddes__c</td>
</tr>
<tr>
<td>caddes__c</td>
<td>caddes__c</td>
</tr>
<tr>
<td>eventid</td>
<td>eventid</td>
</tr>
<tr>
<td>eventname__c</td>
<td>eventname__c</td>
</tr>
<tr>
<td>year</td>
<td>year</td>
</tr>
</tbody>
</table>
```


d. Create aggregate field “event_count” by selecting “New Field” from “Actions” list (drop-down list on the right):
e. Define the aggregate field as “COUNT(eventid)”:

7. Expression transformation “exp_AddFields”:
   a. Create the following expression fields (Actions -> New Field):
      - venue_size – string(100)
      - yearmonthname – string (50)
      - yearqtrname – string(50)

   b. Configure the fields with “Expression” editor:
      - venue_size = IIF(venueseats = 0, 'N/A',IIF(venueseats < 20000, 'SMALL',IIF(venueseats >= 20000 and venueseats < 50000, 'MEDIUM','LARGE')))
      - yearmonthname = year || '-' || month
      - yearname = CONCAT(year,qtr)
8. Target “ars_Redshift”:
   a. Enter “Target Details”:
      - Connection = ars_Redshift
      - Target Type = Object
      - Object = ars_event_summary
      - Operation = Insert
      - Advanced -> S3 Bucket Name = <your S3 bucket>
      - Check Enable Compression checkbox
      - Staging Directlry Location = <a directory on your system where Secure Agent is installed>
      - Truncate Target Table Before Data Load = Yes
   b. Map the fields. Click on “Automatch” to map the fields my name. Manually map the rest of the fields:

   Click on “Save” and select “Save and Continue”.

Step 3: Validate the mapping

1. Validate the mapping by clicking the “Validation” icon in the top right corner:
2. Any transformation that has an error will be highlighted in pink:

3. If you have any errors, go to the highlighted transformation and fix it.
4. Click on *Refresh* button. The fix will make the object color to change to green:

   Click on “*Save*” and select “*Save and Continue*”.

**Step 4: Run the mapping**

1. Close the Validation dialog.
2. Click on “*Save*” and select “*Save and Run*”.
3. If you have more than one Secure Agent associated with your ICS account you will be prompted to select the agent.
4. Monitor the mapping execution in “*Activity Monitor*” page:
5. Review the mapping task execution log after the task completes:

Congratulations! You have arrived to the end of tutorial materials.
Operating Considerations and Best Practices

Informatica Cloud Service Monitoring
The Informatica Cloud trust site, found at http://trust.informaticacloud.com, provides real time information about Informatica Cloud system availability, current and historical data about system performance, and details about Informatica Cloud security policies.

Logs and Monitoring
The Informatica Cloud provides the following tools to help monitor your Secure Agents, schedules, log files, and your currently running tasks.

- Activity log – Accessed from the Monitor menu. View the activity log for all completed tasks. You can drill down for more information and perform all related tasks, such as download the related log file.
- Activity monitor - Accessed from the Monitor menu. View and refresh the activity monitor. Displays the number of rows transferred so far.
- Schedules - Accessed from the Configure menu. View schedule details and create, edit, or delete schedules. You can also configure object-level permissions.
- Audit log – Accessed from the Administer menu. Contains information about user activity.
- Secure Agents – Accessed from the Configure menu. View the status of all of your Secure Agents. The information provided includes the current state (active or inactive) and the last time the agent was upgraded. You can also click on an agent name to drill down to additional information, such as connector and engine versions.

Upgrades & Patching
Informatica Cloud performs 3 major upgrades a year, typically Spring (April or May), Summer (July or August), and Winter (November or December). Notifications are sent out at least a month prior to the upgrade and all customers have access to a prerelease site to try out new features and bug fixes. In addition, new Informatica Cloud Connectors are constantly being released and existing ones upgraded. These deployments occur during the 3rd week of every month.

You do not need to take any action to upgrade your agent. If any updates need to be applied to the Secure Agent, such as adding new functionality, applying bug fixes, or upgrading a connector you are licensed for, this occurs automatically in the background. The agent will restart itself if that is required.

Troubleshooting
If you are having trouble using a specific connector or functionality, please consult the online help documentation as well as additional reference material available here: https://community.informatica.com/community/products/informatica_cloud?view=documents .
In addition, you can find more information about errors and help with troubleshooting here: https://community.informatica.com/docs/DOC-1619.
Appendix A – SQL Queries

Execute CREATE TABLE commands with your database client tool. Some of the SQL clients (like SQL Workbench) require explicit ‘commit’ to create tables on the database.

CREATE TABLE for Oracle

```sql
CREATE TABLE ars_category
(
    catid integer NOT NULL,
    catgroup varchar2(10) NULL,
    catname varchar2(10) NULL,
    catdesc varchar2(50) NULL,
    CONSTRAINT XPKars_category PRIMARY KEY (catid ASC)
);

CREATE TABLE ars_date
(
    dateid integer NOT NULL,
    caldate timestamp NOT NULL,
    day char(3) NOT NULL,
    week integer NOT NULL,
    month char(5) NOT NULL,
    qtr char(5) NOT NULL,
    year integer NOT NULL,
    holiday char(1) NOT NULL,
    CONSTRAINT XPKars_date PRIMARY KEY (dateid ASC)
);

CREATE TABLE ars_event
(
    eventid integer NOT NULL,
    venueid integer NOT NULL,
    catid integer NOT NULL,
    dateid integer NOT NULL,
    eventname varchar2(200) NULL,
    starttime timestamp NULL,
    CONSTRAINT XPKars_event PRIMARY KEY (eventid ASC)
);

CREATE INDEX XIE1ars_event ON ars_event
(
    dateid ASC
);

CREATE TABLE ars_venue
(

```
CREATE TABLE for MySQL

CREATE TABLE ars_category
(
    catid int NOT NULL,
    catgroup varchar(10) NULL,
    catname varchar(10) NULL,
    catdesc varchar(50) NULL,
    CONSTRAINT XPKars_category PRIMARY KEY (catid ASC)
);

CREATE TABLE ars_date
(
    dateid int NOT NULL,
    caldate datetime NOT NULL,
    day char(3) NOT NULL,
    week int NOT NULL,
    month char(5) NOT NULL,
    qtr char(5) NOT NULL,
    year int NOT NULL,
    holiday char(1) NOT NULL,
    CONSTRAINT XPKars_date PRIMARY KEY (dateid ASC)
);

CREATE TABLE ars_event
(
    eventid int NOT NULL,
    venueid int NOT NULL,
    catid int NOT NULL,
    dateid int NOT NULL,
    eventname varchar(200) NULL,
    starttime datetime NULL,
    CONSTRAINT XPKars_event PRIMARY KEY (eventid ASC)
);

CREATE INDEX XIE1ars_event ON ars_event
(}
CREATE TABLE ars_venue
(
    venueid int NOT NULL,
    venuename varchar(100) NULL,
    venuecity varchar(30) NULL,
    venuestate char(2) NULL,
    venueseats int NULL,
    CONSTRAINT XPKars_venue PRIMARY KEY (venueid ASC)
);
CONSTRAINT XPKars_event PRIMARY KEY (eventid ASC)
)
go

CREATE INDEX XIEars_event ON ars_event
{
dateid ASC
}
go

CREATE TABLE ars_venue
{
venueid integer NOT NULL,
venuename varchar(100) NULL,
venuecity varchar(30) NULL,
venuestate char(2) NULL,
venueseats integer NULL,
CONSTRAINT XPKars_venue PRIMARY KEY (venueid ASC)
}
go

CREATE TABLE for Redshift

create table ars_category
{
catid smallint not null distkey sortkey,
catgroup varchar(10),
catname varchar(10),
catdesc varchar(50),
CONSTRAINT XPKars_category PRIMARY KEY (catid)
};

CREATE TABLE ars_date
{
dateid smallint not null distkey sortkey,
caldate date not null,
day character(3) not null,
week smallint not null,
month character(5) not null,
qtr character(5) not null,
year smallint not null,
holiday boolean default('N') ,
CONSTRAINT XPKars_date PRIMARY KEY (dateid)
};

CREATE TABLE ars_event
(  
  eventid integer not null distkey,
  venueid smallint not null,
  catid smallint not null,
  dateid smallint not null sortkey,
  eventname varchar(200),
  starttime timestamp ,
  CONSTRAINT XPKars_event PRIMARY KEY (eventid)
);  

CREATE TABLE ars_venue  
(  
  venueid smallint not null distkey sortkey,
  venuename varchar(100),
  venuecity varchar(30),
  venuestate char(2),
  venueseats integer ,
  CONSTRAINT XPKars_venue PRIMARY KEY (venueid)
);  

CREATE TABLE ars_event_wide  
(  
  eventid integer not null distkey,
  venueid smallint not null,
  catid smallint not null,
  dateid smallint not null sortkey,
  eventname varchar(200),
  starttime timestamp ,
  venuename varchar(100),
  venuecity varchar(30),
  venuestate char(2),
  venueseats integer ,
  catgroup varchar(10),
  catname varchar(10),
  catdesc varchar(50),
  caldate date not null,
  day character(3) not null,
  week smallint not null,
  month character(5) not null,
  qtr character(5) not null,
  year smallint not null,
  holiday boolean default('N') ,
  CONSTRAINT XPKars_eventwide PRIMARY KEY (eventid)
);  

CREATE TABLE ars_event_summary  
(  
  venueid smallint not null distkey,
  catid smallint not null,
  event_count integer not null ,
  CONSTRAINT XPKars_event_summary PRIMARY KEY (venueid)
);
venuename varchar(100),
venuecity varchar(30),
venuestate char(2),
venue_size varchar(100) not null ,
catgroup varchar(10),
catname varchar(10),
catdesc varchar(50),
month character(5) not null,
qutr character(5) not null,
year smallint not null,
yearmonthname varchar(50) not null sortkey,
yearqtrname varchar(50) not null,
CONSTRAINT XPKars_eventsummary PRIMARY KEY (venueid)
);

CREATE TABLE ars_performance
(
    catid smallint not null distkey sortkey,
catgroup varchar(10),
catname varchar(10),
catdesc varchar(50),
CONSTRAINT XPKars_performance PRIMARY KEY (catid)
);
Appendix B – Manual Steps for AWS Environment Creation

Set up the VPC environment

Once you have your account set up, the first action will be to create a secure “container” that you fully control that for all of the cloud resources associated with the solution, like your Redshift Data Warehouse and possibly your Informatica Cloud Secure Agent. This secure container is called a Virtual Private Cloud (VPC). If you relatively new to AWS, then you will have a default VPC already created for you. If you signed up for AWS a while ago, then you may have an account that uses “classic” mode, in which case you will need to configure your own VPC. Customers frequently will want to segment their environments based upon organization, or functional area, by creating multiple VPC. For further details on VPCs, see http://aws.amazon.com/vpc/. The tutorial walkthrough will not setup a VPN connection as that is a more advanced scenario. See Appendix C for a walkthrough and template demonstrating how to setup a VPN connection.

1) Log in to the AWS console.
2) Under the “Compute & Networking” section, select VPC.
3) Make sure to select the region in the upper right for the region you want your solution deployed within.

4) Advanced users can create each element of the VPC. For new users, there is a wizard available for common configurations. For the purposes of tutorial we will use the wizard. Select “Start VPC Wizard”.

5) Select “VPC with a Single Public Subnet”. This is the simplest VPC configuration.

6) On the next wizard page, you will be asked to enter CIDR address ranges. A CIDR address range is a mechanism to define the private IP address ranges for your VPC and subnets. If you plan to connect the VPC back to your corporate network, then you should confer with your network administrator since conflicts in IP address ranges with your corporate network will cause issues later when setting up your VPN. In this case we are not connecting through a VPN, so just leave the defaults. Provide a name for your VPC, and then click “Create VPC”.


Your VPC will now be created. Under the covers, the wizard is creating a few additional pieces like an internet gateway as well that lets traffic get out onto the internet, and configures a route table that allows traffic from the internet to reach the public subnet. The final configuration if you use the defaults for the VPC is represented in the following diagram:

Create Private Subnets
1) From the VPC dashboard, select “Your VPCs” in the left column.
2) Select the ICS-Tutorial VPC you just created in the “Filter By VPC” drop-down in in the left upper corner of the dashboard.
3) Select “Subnet” from the list of dashboard options in the left, and then “Create Subnet”.
4) Select your VPC in the drop-down list.
5) Name the subnet “Private Subnet”, and enter “10.0.1.0/24” for the “CIDR block”.
6) Select the same “Availability Zone” as your public subnet.
7) Click “Yes, Create”.
8) In order to launch an RDS instance, we will need a second subnet (for failover). Select “Create Subnet” again.
9) Enter “Private Subnet Failover” for the subnet name, and enter “10.0.2.0/24” for the CIDR block.
10) Select a different Availability Zone for your subnet than the “Private Subnet” previously created.
11) Click “Yes, Create”.
Create Security Groups

1) Ensure you VPC is selected in the “Filter by VPC” drop-down.
2) Select “Security Groups” from the VPC Dashboard left menu.
3) Click “Create Security Group”.
4) Enter “Secure Agent SG” for “Name tag” and “Description”.
5) Select your VPC in the drop-down list.
6) Click “Yes, Create”.
7) Select the “Secure Agent SG” in the list.
8) In the lower section, click the “Inbound Rules” tab.
9) Click “Edit”.
10) Add a rule permitting Remote Desktop access (RDP) by selecting “RDP” under the “Type” column.
11) Under the “Source” column enter “0.0.0.0/0”. Normally you will restrict that address to just a range, like your corporation IP address range. For the purposes of the tutorial, we will permit RDP from any location.
12) You will now add rules that permit Informatica Cloud to access the instance. You should check the Informatica Cloud User Guide for the latest ranges, as they may change.
   a. Click “Add another rule”.
   b. Select “Custom TCP Rule” for the “Type” column.
   c. Enter “0-65535” for the “Port Range” column.
   d. Enter “209.34.91.0/24” for the “Source”.
13) Repeat step 12 for “206.80.61.0/24” and “206.80.52.0/23” values as “Source”.
14) Click “Save”.
15) Click “Create Security Group”.
16) Enter “ICS DB SG” for “Name tag” and “Description”.
17) Select “ICS Tutorial” VPC.
18) Click “Yes, Create”.
19) Select “ICS DB SG”, and select the “Inbound Rules” tab.
20) Click “Edit”.
21) Under the “Type” column, select “MySql”.
22) For Source, enter “10.0.0.0/24”. This is the IP address range for virtual instances launched in the public subnet.
23) Click “Add another rule”.
24) Under type column, select “Custom TCP Rule”.
25) Enter “5439” for the port (this is the default port for Redshift).
26) For source, enter “10.0.0.0/24”.
27) Click “Save”.

You’ve now created the network infrastructure required for the tutorial.
Launch your EC2 instance

Now that your Virtual Private Cloud is created, you will need to create a few more resources. First we will create a virtual instance for hosting the Informatica Cloud secure agent by using Amazon Elastic Cloud Compute (EC2).

1) From the AWS console, select EC2. Make sure that the console has selected the same region as your VPC in step 3 above.

2) Select “Instances” from the EC2 dashboard options on the left and then click on “Launch Instance”. This will bring you into the first step of a wizard that lists Amazon Machine Images (AMI’s) supported by AWS.

3) Select “Microsoft Windows Server 2012 with SQL Server Express” from the list with the following description: “Microsoft Windows Server 2012 Standard edition, 64-bit architecture, Microsoft SQL Server 2012 Express”.

4) The following example is from the US West region. The AMI ID may be different.

5) Select “m3.medium” instance type and click on “Next: Configure Instance Details”.

6) Select the VPC that you configured in the previous section. You will see the name of the VPC in the drop-down list.

7) Under the “Subnet” drop-down, select the “Public subnet”.

8) Under “auto-assign public IP”, select “Enable”.

9) Click “Next: Add Storage”.

10) Accept the defaults, and click “Next: Tag Instance”.

11) Enter “ICS-Tutorial” for the “Name” tag. Tags help you identify attributes of the instance. You can add additional tags if you like.

12) Click on “Next: Configure Security Group”.

13) Click on “Select an existing security group”.
14) Select the “Secure Agent SG” you previously created.
15) Select “Review and Launch”.
16) Click “Launch”.
17) If you have used AWS to launch instances in the region before, you will see a security key available for you. If you have not, you will need to create a security key. Select “Create a new key pair as shown below”.

![Create a new key pair](image)

18) If you created a new key, then name the key, and then select “download key pair”. Once the key pair is downloaded, move it into a location you can easily access. This key will be used to retrieve the password once the instance is launched, so make sure that you save it off, otherwise your instance will not be accessible.
19) If you are using a pre-existing key, then check “I acknowledge that I have access to the selected private key file...”.
20) Click “Launch Instances”. You can now see your instance launching. When the status checks shown in the console have completed, your instance will be accessible. There is sometimes a delay as the process for setting the password completes.
21) Log in to your instance by using a Remote Desktop, select the instance from the EC2 dashboard, and then select the connect button:

![Connect to instance](image)

22) Click on “Get Password”. You will need to provide the location of the key file you previously created, which is then used to decrypt your password. After you have retrieved your password, click “Download Remote Desktop File”, and click on the downloaded file. Log into the instance using the decrypted password.
23) Login into the EC2 instance via Remote Desktop Connection.
24) Install SQL clients for MySQL and PostgreSQL onto the EC2 instance for the tutorial. Three free options available are:
   - SQL Workbench (see http://docs.aws.amazon.com/redshift/latest/mgmt/connecting-using-workbench.html).
   - Aginity workbench (see http://www.aginity.com/workbench/redshift/).
   - PremiumSoft Navicat Lite 10 (see http://premiumsoft-navicat-lite.software.informer.com/10.0/).

Create Redshift Cluster
1) From “Services” in the upper left, select “Redshift”.
2) Click on “Security” in the left column.
3) Click on the “Subnet Groups” tab.
4) Click “Create Cluster Subnet Group”.
5) Enter “ics-tutorial” for “Name” and “Description”.
6) Under “VPC ID”, select the VPC ID for the ICS-Tutorial VPC. (You might need to go back to the VPC tab to identify this VPC if you have multiple VPCs created.)
7) Under “Availability Zone”, select the availability zone with your private subnet (not failover).
8) Select the Subnet ID for this subnet. (You might need to go back to the VPC tab to identify).
9) Click “Add”.
10) Click “Create”.
11) Click on “Clusters” in the left menu.
12) Click on “launch cluster”.
13) Enter “ics-tutorial” for the cluster name.
14) Enter “master” for the user name (or another name of your choosing).
15) Enter a password. The default used by the template “Password123”.
16) Leave “dw2.large” selected for node type.
17) Select “Multi-node” for cluster type.
18) Enter 2 for number of compute nodes.
19) Select “Continue”.
20) Under Choose a VPC, select the VPC you just configured the subnet group for.
21) Under Publicly Accessible, choose “No”. You could also choose to make your cluster publicly accessible or not. By choosing to make it publicly accessible, you will have a public IP address associated with it. The security groups and route tables still need to be configured to allow access. Since we haven’t allowed that, the cluster still will not be publicly accessible.

22) Select “ICS DB SG” under VPC security groups.

23) Select “No” for create cloudwatch alarms. This is a great feature for production clusters as it will automatically notify a topic of issues or changes. However for the tutorial, we will not need it.

24) Select “Continue”.

25) Select “Launch Cluster”.

Create RDS MySQL Instance

1) Select the RDS service from the Services menu.
2) Select “subnet groups” in left menu.
3) Select “Create DB Subnet Group”.
4) Name the subnet group “ics-tutorial”.
5) Select your VPC ID.
6) Select the AZ that you created the “Private Subnet” in.
7) Click “Add”.
8) Select the AZ that you created the “Private Subnet Failure” subnet in.
9) Click “Add”.
10) Click “Yes, Create”.
11) Click “Instances” in the left menu.
12) Click “Launch DB Instance”.
13) Select “MySQL”.
14) Select “No, This Instance is intended for use outside of production...”
15) Select “db.t2.small” for “DB Instance Class”.
16) Select “No” for Multi-AZ deployment.
17) Enter 5 for Allocated Storage.
18) Set DB Instance Identifier to “ics-tutorial”.
19) Enter the master user name and a password.
20) Click “Next”.
21) Make sure your VPC is selected, and the ics-tutorial subnet group.
22) For availability zone, select the AZ where you launched your Secure Agent.
23) For VPC security group, select “ICS DB SG”.
24) For Database name, enter “tutorial”.
25) Leave all other settings.
26) Click “Launch DB”.

82
At this point you have created the configuration for the tutorial.