Reference Architecture Guide: Deploying Informatica© PowerCenter© on Amazon Web Services

© Copyright Informatica LLC 2016

Informatica, the Informatica logo *[and any other trademarks appearing in the document]* are trademarks or registered trademarks of Informatica LLC in the United States and many jurisdictions throughout the world. A current list of Informatica trademarks is available on the web at https://www.informatica.com/trademarks.html. Other company and product names may be trade names or trademarks of their respective owners.

The information in this documentation is subject to change without notice. If you find any problems in this documentation, please report them to us in writing at Informatica LLC 2100 Seaport Blvd. Redwood City, CA 94063.

INFORMATICA LLC PROVIDES THE INFORMATION IN THIS DOCUMENT "AS IS" WITHOUT WARRANTY OF ANY KIND, EXPRESS OR IMPLIED, INCLUDING WITHOUT ANY WARRANTIES OF MERCHANTABILITY, FITNESS FOR A PARTICULAR PURPOSE AND ANY WARRANTY OR CONDITION OF NON-INFRINGEMENT.

Contents

Introduction

Organizations are looking for opportunities to reduce their on-premises datacenter footprint by offloading or extending on-premises applications and data warehouses to the cloud. Cloud deployments increase agility as they allow organizations to rapidly add new capabilities and scale up and down as their needs change. Cloud solutions free up IT resources from supporting commoditized infrastructure and allow them to focus on building differentiating technical capabilities.

A robust data integration solution will greatly increase the success of your organization's journey to the cloud, helping you implement hybrid cloud use cases such as hybrid data warehousing or application migration to the cloud. A successful data integration solution should enable your organization to focus on their current and future data management needs to address both current and future state.

Informatica PowerCenter is a proven data integration solution that transforms fragmented, raw data from any source, any technology, at any latency into complete, high-quality, actionable information.

Customers of Amazon Web Services (AWS) and Informatica can now deploy PowerCenter in the AWS public cloud, leveraging the data management power of PowerCenter and the flexibility of the AWS cloud. Customers, with investment in Informatica PowerCenter, who are migrating their environment to Amazon Web Services (AWS), can fully leverage this investment by deploying PowerCenter on AWS. Customers who are interested in a fully managed, multi-tenant iPaaS (Integration Platform as a Service) option, can also explore Informatica Cloud as an alternative.

This document provides technical guidance on how you can seamlessly expand the data management experience of PowerCenter to Amazon Web Services by migrating PowerCenter to AWS.

PowerCenter for all Your Data Integration Needs

PowerCenter is certified to run in the AWS environment, which offers a great option for current PowerCenter customers considering moving or offloading their applications and/or data warehouses to AWS. This allows them to realize cloud benefits while leveraging their existing data management investment in PowerCenter.

PowerCenter in the AWS public cloud provides several benefits:

- **Enterprise class data integration.** PowerCenter is an enterprise proven dataintegration solution that can process billions of records in the AWS cloud.
- **Connects to existing data sources and quickly onboards new data sources and data types**. PowerCenter offers a vast array of connectors, whether you want to connect to on- premises data sources or AWS services such as Amazon Redshift, Amazon RDS, or Amazon S3.
- **Faster time to insight**. By leveraging your existing on-premises PowerCentermappings, metadata, and workflows, you can get rapidly load data into AWS data services such as Amazon Redshift, delivering the right analytical data to your businessstakeholders.
- **Accelerates data architecture modernization.** If you are planning to modernize your data warehousing initiatives on AWS, PowerCenter's rich functionalities such as metadata driven data integration, dynamic mappings, SQL conversion mapping, and automatic datavalidation will help you to shorten development cycles and reduce time tomarket.
- **Delivers clean, complete and trustworthy data**. Whether you are offloading or extending on-premises applications to the cloud or fully embracing the cloud, delivering, complete,high- quality data is critical. PowerCenter has a long history of helping organizations empowertheir users with complete, high-quality, actionable data.
- **Meets business demands by improving developer productivity**. IT departments struggle to deliver trusted data as the amount of data and the demand from business keeps growing. To succeed, your organization's data management environment must continue to leverage greater speed and agility in order to delight your customers and outsmart your competitors. PowerCenter's highly visual, easy to use metadata driven environment not only enhances the ability to execute on today's projects, but also helps your organization cost-effectively scale future initiatives.

AWS Overview

Amazon Web Services offers the basic building blocks of storage, networking and compute, as well as services such as managed database, big data, and messaging services. PowerCenter can help you get the most out of the following:

Amazon Elastic Compute Cloud (EC2)

Amazon Elastic Compute Cloud (Amazon EC2) provides scalable computing capacity in the Amazon Web Services (AWS) cloud. Using Amazon EC2 eliminates your need to invest in hardware up front, so you can develop and deploy applications faster. You can use Amazon EC2 to launch as many or as few virtual servers as you need, configure security and networking, and manage storage. Amazon EC2 enables you to scale up or down to handle changes in requirements or spikes in popularity, reducing your need to forecast traffic.

Amazon Simple Storage Service (S3)

Amazon Simple Storage Service (Amazon S3) provides developers and IT teams with secure, durable, highly-scalable cloud storage. Amazon S3 is easy to use object storage, with a simple web service interface to store and retrieve any amount of data from anywhere on the web. With Amazon S3, you pay only for the storage you actually use. There is no minimum fee and no setup cost.

Amazon Relational Database Service (RDS)

Amazon Relational Database Service (Amazon RDS) makes it easy to set up, operate, and scale a relational database in the cloud. It provides cost-efficient and resizable capacity while managing time-consuming database administration tasks, freeing you up to focus on your applications and business. Amazon RDS provides you with several familiar database engines to choose from, including Amazon Aurora, Oracle, Microsoft SQL Server, PostgreSQL, MySQL and MariaDB. Please refer to Informatica Product Availability Matrix (PAM) for a complete support information.

Amazon Elastic Block Store (EBS)

Amazon Elastic Block Store (Amazon EBS) provides persistent block level storage volumes for use with Amazon EC2 instances in the AWS Cloud. Each Amazon EBS volume is automatically replicated within its Availability Zone to protect you from component failure, offering high availability and durability. Amazon EBS volumes offer the consistent and low-latencyperformance needed to run your workloads. With Amazon EBS, you can scale your usage up or down within minutes – all while paying a low price for only what you provision.

Amazon Virtual Private Cloud

Amazon Virtual Private Cloud (Amazon VPC) lets you provision a logically isolated section of the Amazon Web Services (AWS) cloud where you can launch AWS resources in a virtual network that you define. You have complete control over your virtual networking environment, including selection of your own IP address range, creation of subnets, and configuration of route tables and network gateways.

You can easily customize the network configuration for your Amazon Virtual Private Cloud. For example, you can create a public-facing subnet for your webservers that has access to the Internet, and place your backend systems such as databases or application servers in a private- facing subnet with no Internet access. You can leverage multiple layers of security, including security groups and network access control lists, to help control access to Amazon EC2 instances in each subnet.

PowerCenter Deployment Options

Starting with Informatica version 9.6.1 HotFix 3, PowerCenter customers can choose to seamlessly extend the data integration and data management experience to AWS. As a PowerCenter customer, you can execute the full Informatica data pipeline on AWS and take advantage of multiple AWS services. Informatica products can be deployed on AWS with new or existing on-premises licenses.

PowerCenter installed on an EC2 instance can leverage all standard PowerCenter connection types and PowerExchange© adapter connections to talk to on-premises applications. When PowerCenter is installed on EC2, you can take advantage of the economies of scale of AWS, while reducing your total cost of ownership. You can leverage lower latency to services like Amazon Redshift and capitalize on the security and availability features built-in to the AWS platform.

	PowerCenter On-premises	**PowerCenter on EC2**
Scalability	Horizontal scalability involves lead time to procure physical servers. Vertical scalability often involves guesswork to predict future load.	Scale up in minutes, not weeks. The infrastructure is easy to configure and can be highly automated. If your needs change, simply scale back and only pay for what you use.
Point in time snapshots	Managing snapshots in an on- premises environment can be costly and complex.	Easily automate your backup strategy and only pay for what you use, when you use it.
Back-up strategy	An often-costly effort that involves multiple vendors and media, with different management planes.	Back up your data to S3 for a durable, low cost approach and utilize the built-in data lifecycle policies to get the right storage at the right price.
Amazon RDS and Redshift connectivity	Connecting to AWS services through the corporate firewall adds complexity and latency.	Leverage a secure, low latency connection to popular services like Redshift.

Informatica fully supports its products running on Amazon EC2. Informatica does not provide general support for cloud computing specific issues. For general cloud computing support, we recommend maintain a support relationship with your cloud computing vendor. Please refer to the Informatica Support Statement related to Usage of Informatica Products on a Cloud Computing Platform for supported editions of PowerCenter on AWS.

Deployment Architecture

To install PowerCenter on the AWS Cloud Infrastructure, use one of the following installation methods: Marketplace Deployment (recommended) and Conventional and Manual Installation.

PowerCenter is available on AWS Marketplace. You can subscribe to PowerCenter listings and deploy PowerCenter in AWS Cloud Infrastructure with simple and quick configurations. Use the PowerCenter listing on AWS Marketplace for an optimal configuration of the Informatica domain, domain database and AWS infrastructure settings. You can also install PowerCenter on AWS Cloud Infrastructure with Manual configuration where you must configure AWS infrastructure settings such as Amazon EC2 Instance configuration, Networking settings (VPC, Security group, Inbound and Outbound rules etc.) and Domain Database configurations.

When you run PowerCenter on AWS Cloud Infrastructure, you experience the same product features as running PowerCenter on-premises. Whether you are installing PowerCenter for the first time or you are planning to migrate from an on-premises to AWS platform, the steps involved in running Informatica services inside AWS follow a similar deployment lifecycle.

Selecting the Appropriate Amazon EC2 Instance

When you configure an Amazon EC2 instance, the instance type that you specify determines the hardware configuration of the host computer used for the instance. Each instance type offers different compute, memory, and storage capabilities.

Recommended EC2 Instance types for

PowerCenter Nodes –

- c4.xlarge - c4.4xlarge
- m4.large - m4.4xlarge
- r3.large - r3.4xlarge

Informatica Domain Database –

- db.t2.large & db.t2.xlarge
- db.m3.large & db.m3.xlarge
- Oracle RDS configuration with 100 GB of Storage

Perform these steps to configure the Amazon EC2 instance:

- You are free to select any configuration for Manual installation on AWS. In such cases of manual configuration, you need to make sure your instance choice fulfills the minimum system requirements of a standard PowerCenter services installation. For more information about PowerCenter minimum system requirements, see the *Informatica Installation and Configuration Guide* in the Informatica Network.
- The PowerCenter listings on marketplace provides wide range of AWS infrastructure configuration and helps to deploy PowerCenter in most optimal way in AWS Cloud.
- Set the file descriptor limit to greater than 50,000 or unlimited.
- Choose an Amazon EBS volume as your root device, so you can easily resize your instance by changing its instance type. For example, you can change the instance type from M3.large to M3.xlarge. Always select HVM instances that are EBS optimized.

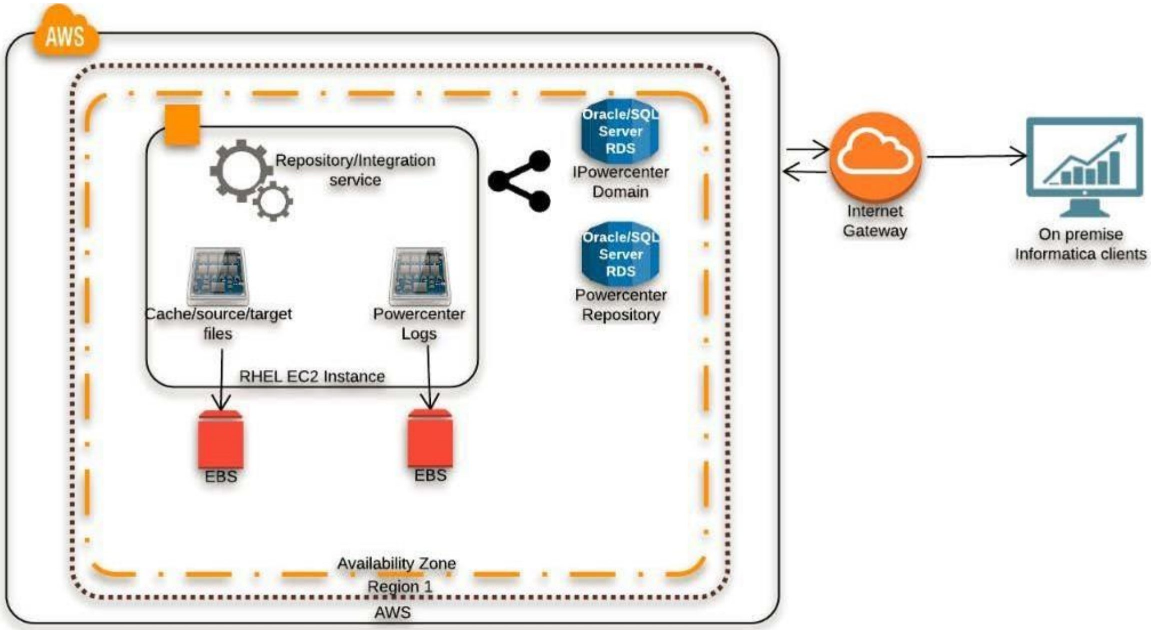


Figure 1: PowerCenter Services Running on AWS

Hosting the Domain Database and the Repository Database

For hosting the Informatica domain database and repository database tables there are three choices:

- **Existing "on-premises" repository.** This requires least effort but is not recommended because it often costlier and has higher latency. If you do choose this option, be sure to use a secured communications channel such as a VPN or Amazon Direct Connect in conjunction with an Amazon VPC to improve security.
- **Manage your own database on an EC2 instance.** This minimizes the latency, but all of the heavy lifting of database backups, management, and maintenance is still required. Use this option if you want a relational database on cloud that you can manage on your own.
- **Amazon RDS.** Apart from lowest network latency, Amazon RDS also handles time-consuming database management tasks, such as backups, patch management, and replication. If you want a relational database with minimal administration choose Oracle RDS. You also have the option to choose other popular databases such as Oracle and MS SQL Server. Please refer to Informatica Product Availability Matrix (PAM) for a complete support information.

Make sure to enable database snapshots. These full database backups will be stored by Amazon RDS until you explicitly delete them thus preserving your repository contents in case they need to be restored.

Note: Allocated storage in RDS cannot be scaled down. RDS backups cannot be used for database restore and recovery outside of AWS. Unlike EC2 instances, database users in RDS cannot be managed through AWS management console.

Security Group Settings

Configure a security group to allow traffic on the following ports:

Port Name	**Default Port Numbers**
Node Port	6005
Service Manager Port	6006
Service Manager Shutdown Port	6007
Informatica Administrator Port	HTTP: 6008 HTTPS:8443
Informatica Administrator Shutdown Port	6009

For more information about Informatica port administration, see https://kb.informatica.com/h2l/HowTo%20Library/1/0519-Informatica_Port_Adminstration-H2L.pdf

Storage Segregation

PowerCenter stores a variety of types of data when installed on EC2: workflow and session logs, repository backups, and both persistent and non-persistent cache files generated by transformations. For maximum performance on EC2, use EBS volumes for persistent data and ephemeral instance storage for temporary cache data. Examples of data suitable for EBS include: $INFA_HOME, repository backup, and the session log directory.

Network and Storage Prerequisites

If the client is on an on-premises Windows machine, at least 32Mbps of sustained bandwidth is recommended. If your instance supports an EBS-optimized flag, enable it to add up to 500 mbps of bandwidth to EBS. The amount of bandwidth available depends on type of instance chosen.

PowerCenter to Redshift Connectivity

PowerExchange for Amazon Redshift is a popular PowerExchange adapter that connects PowerCenter and Amazon Redshift. PowerExchange for Amazon Redshift facilitates bulk data movement from data source inside traditional on-premises to Amazon Redshift.

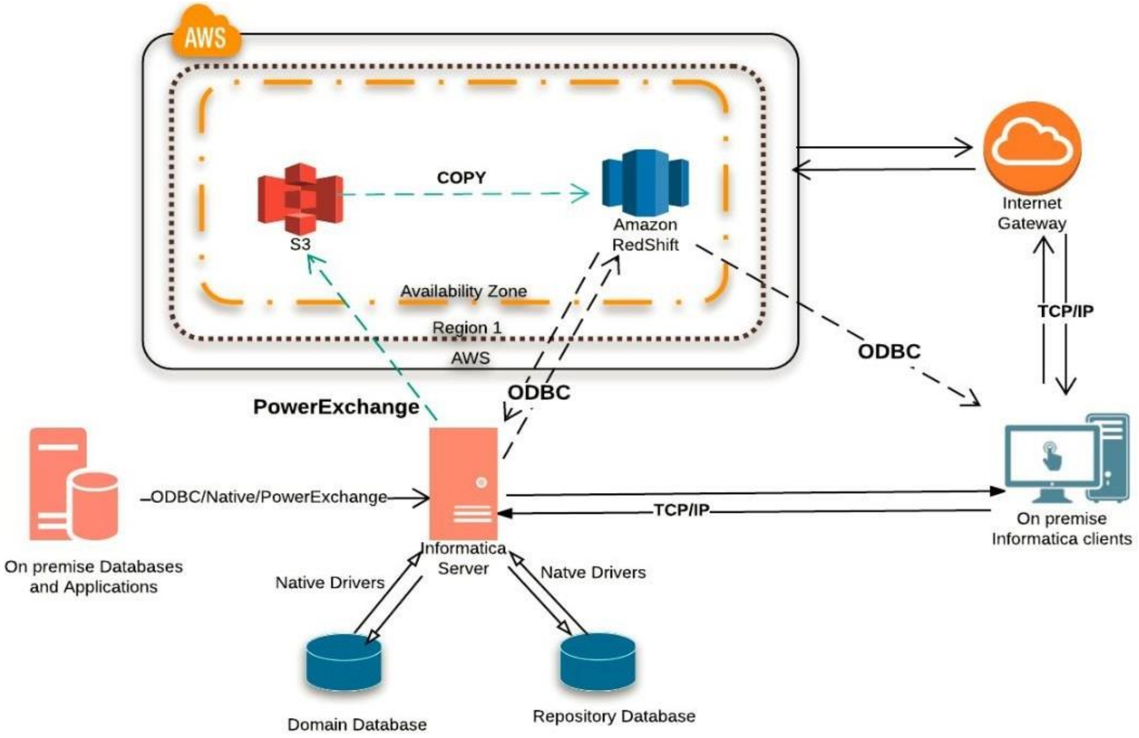


Figure 2: PowerCenter Services in an On-Premises Setup and Redshift Cluster

The PowerCenter Integration Service uses the Redshift driver to communicate with Amazon Redshift. The PowerCenter Integration Service writes data to Amazon Redshift based on the workflow and Amazon Redshift connection configuration.

The PowerCenter Integration Service first writes data to Amazon S3, and then initiates a copy of data into Redshift. This leverages the Amazon Redshift massively parallel processing (MPP) architecture to read and load data in parallel from files in an Amazon S3 bucket.

Make sure that the S3 bucket that is specified in the session properties has the correct permissions and is in the same region as Redshift so that Informatica can successfully upload the source data.

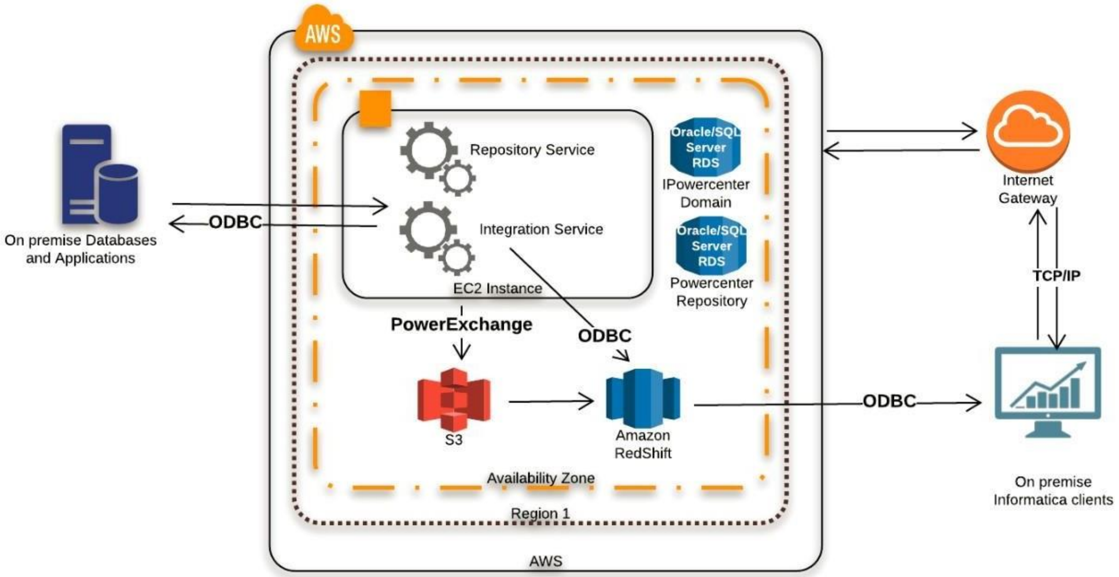


Figure 3: PowerCenter Services on AWS and with Redshift Connectivity

Amazon Redshift Overview

Cloud computing in general and Amazon Web Services (AWS) in particular provide an easy way to provision hardware required for running application services. Using AWS, customers can increase or decrease hardware capacities based on their needs and convert capital expenses needed for maintaining the physical data centers into variable expenses by provisioning resources on demand.

The Amazon Redshift ensemble consists of a group of machines called nodes. The group is called a Redshift cluster. A cluster can be comprised of a single node, which is a single machine. However, a cluster generally consists of more than one node.

In a “distributed mode”, a special machine called the leader node coordinates the incoming data traffic for the cluster. In a “pseudo distributed” mode, the same machine acts as a leader and compute node.

A compute node is where data actually resides and leader node is the gateway for any client requests, it parses the client requests, creates an execution plan for the query, compiles the code and dispatches the compiled code to compute node.

The compute nodes execute their portion of code, sent by a leader node, and responds back with an intermediate result. The control passes back to the leader node, which takes the individual results from the compute nodes, aggregates the results, and sends them back to the requesting client.

The compute nodes are the machines where the actual data resides. They are an individual unit of a cluster with their own CPU, memory, and disk storage. A compute node is partitioned into node slices. A node slice controls a portion of compute nodes memory and disk space. Node slices enable parallel execution of queries inside compute node by taking care of subset of data traffic flowing through the respective compute node.

Redshift has several salient features that make it suitable for a petabyte scale data warehouse:

Massively Parallel Processing Architecture

By employing a distributed architecture on multiple compute nodes, Redshift is able to process a single request using multiple threads. The distribution of the workload to allow parallel processing is called massively parallel processing (MPP) architecture.

Query Engine

Redshift makes use of a database engine (query optimizer) that's MPP aware and exploits the parallel processing capabilities to the fullest.

Columnar storage

In OLTP systems you typically query entire rows, but in a data warehouse architecture, data access patterns often return many rows of a fewer number of columns. Columnar storage stores this column data contiguously on disk allowing the database to process billions of rows lightning fast.

To access Redshift using PowerCenter, PowerCenter does not need to be inside the AWS ecosystem. PowerCenter can access the Redshift cluster using PowerExchange for Redshift installed on the Informatica on-premises server.

Informatica guidance: Use ODBC only for pushdown optimization when moving data within Redshift. PowerExchange for Redshift is much faster than ODBC when bringing data from outside AWS.

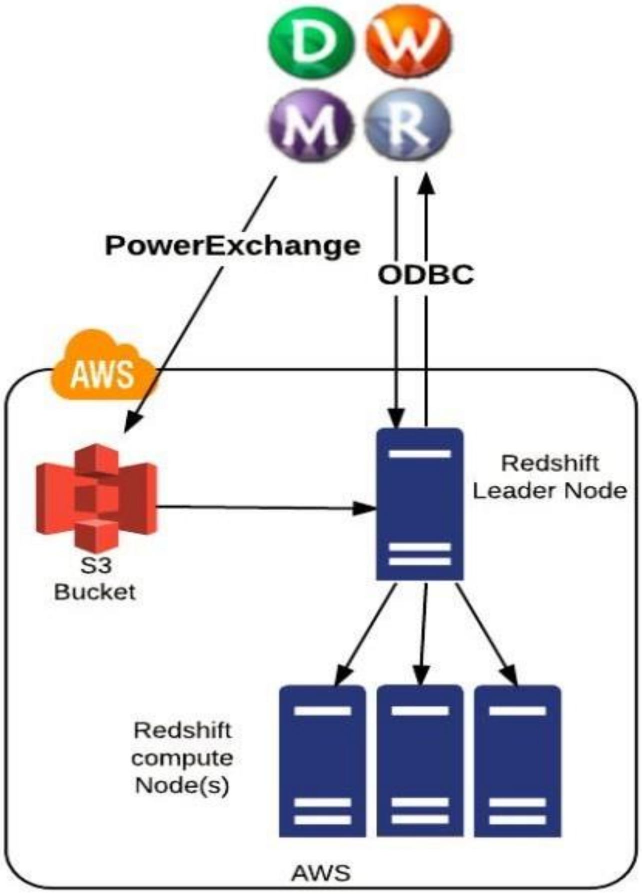


Figure 4: PowerCenter Clients and Amazon Redshift

The matrix below provides a comparison of capabilities PowerExchange for Redshift and ODBC provides when used inside the PowerCenter sessions:

Session Property	**Power Exchange**	**ODBC**
Read from Redshift	Yes	Yes
Write to Redshift	Yes	Yes
Redshift lookup connections	No	Yes
Pre/Post SQL to Redshift	No	Yes
Session Push Down Optimization (PDO)	No	Yes
MPP Optimized writes via S3 buckets	Yes	No
Connection for Redshift lookups tables	No	Yes
Insert operations	Yes	Yes
Delete operations	No	Yes
Update operations	Yes	Yes

ODBC facilitates pushdown optimization in PowerCenter. PowerExchange for Redshift, on the other hand, optimizes the bulk loads by first writing data to an S3 bucket and using a COPY command thereafter to load data into Redshift. The COPY command uses Redshift's MPP architecture to read and load data in parallel from multiple data sources and is faster and efficient than INSERT commands.

Configuring PowerExchange for Redshift

Schema Details

A Redshift database contains one or more schemas. Database schemas group database objects under a common namespace. To configure PowerExchange for Redshift, you need to provide the schema name where the Redshift connection will connect to the cluster.

PowerExchange for Redshift, validates that the table exists in the described schema.

```
SELECT tablename FROM pg_tables WHERE schemaname=<User Provided Schema>
```

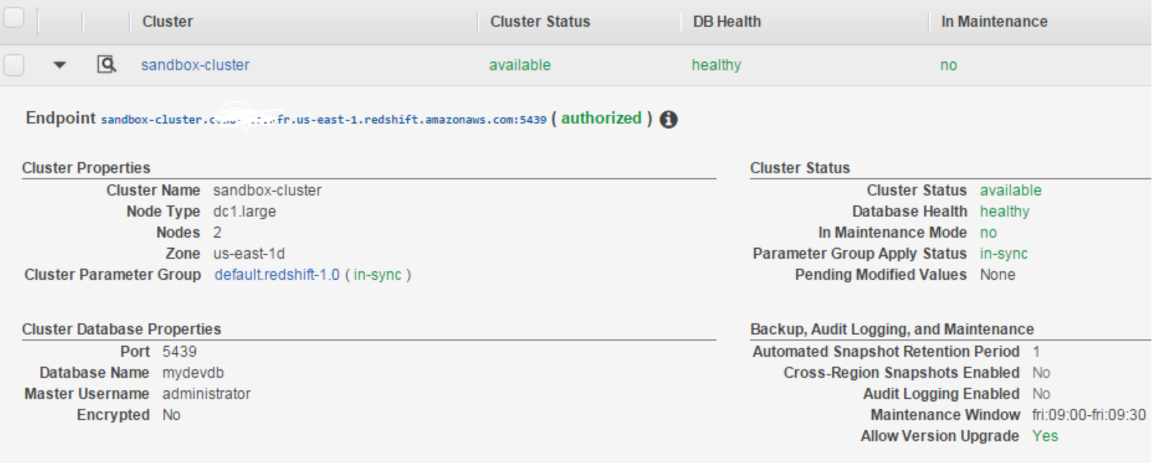
AWS Access Key ID and AWS Secret Access Key

PowerExchange for Redshift uses S3 buckets to host the data and use a copy command to load data into Redshift. Before using the PowerExchange for Redshift create a new S3 bucket or designate an existing one. You can create a new S3 bucket by using AWS console at https://console.aws.amazon.com/s3/.

To access the S3 bucket, you need an AWS key ID and a secret key. The access key ID isan alphanumeric text string. It uniquely identifies the user who owns the account that has privileges on the S3 bucket. The secret access key serves as the password to validate the credentials when the PowerCenter session tries to connect to S3 bucket. Never share your secret key with anyone!

Number of Nodes in the Cluster

Define the right number of nodes when you create your Amazon Redshift cluster. You can view this property in the Redshift console under Cluster Properties.



Cluster	Cluster Status	DB Health	In Maintenance
sandbox-cluster	available	healthy	no

Endpoint sandbox-cluster.[redacted]fr.us-east-1.redshift.amazonaws.com:5439 (authorized)

Cluster Properties

Cluster Name	sandbox-cluster
Node Type	dc1.large
Nodes	2
Zone	us-east-1d
Cluster Parameter Group	default.redshift-1.0 (in-sync)

Cluster Status

Cluster Status	available
Database Health	healthy
In Maintenance Mode	no
Parameter Group Apply Status	in-sync
Pending Modified Values	None

Cluster Database Properties

Port	5439
Database Name	mydevdb
Master Username	administrator
Encrypted	No

Backup, Audit Logging, and Maintenance

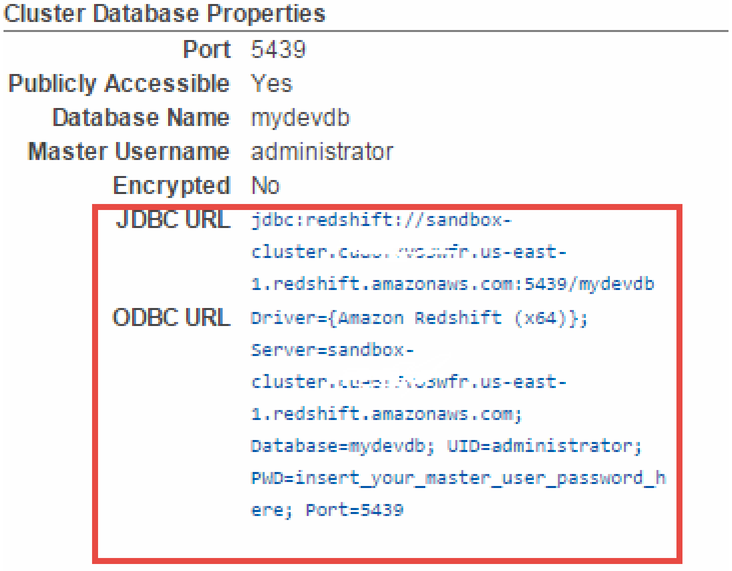
Automated Snapshot Retention Period	1
Cross-Region Snapshots Enabled	No
Audit Logging Enabled	No
Maintenance Window	fri:09:00-fri:09:30
Allow Version Upgrade	Yes

Cluster Node Type

The Cluster Properties provide the Node type of the Amazon Redshift cluster. For more information about Redshift pricing for Amazon Redshift node types in current generation and pricing details, visit https://aws.amazon.com/redshift/pricing/.

JDBC URL

The JDBC URL is your connection URL. Click on the cluster name on Amazon Redshift console. A window appears on the Configuration tab. Under Cluster Database Properties, there are two URLs: one each for JDBC and ODBC. Use the JDBC URL for your configuration in PowerCenter.



Cluster Database Properties

Port	5439
Publicly Accessible	Yes
Database Name	mydevdb
Master Username	administrator
Encrypted	No
JDBC URL	jdbc:redshift://sandbox-cluster.[redacted]fr.us-east-1.redshift.amazonaws.com:5439/mydevdb
ODBC URL	Driver={Amazon Redshift (x64)}; Server=sandbox-cluster.[redacted]wfr.us-east-1.redshift.amazonaws.com; Database=mydevdb; UID=administrator; PWD=insert_your_master_user_password_here; Port=5439

PowerCenter Session with a Redshift Connection

To read and write data with Amazon Redshift as a source or target:

- Create a mapping with any source and a relational target to write data to an Amazon Redshift target. Starting in PowerCenter 9.6.1 HotFix 3, you can import a Redshift source and target.
- To write data to an Amazon Redshift table, you must configure an Amazon Redshift connection in the Workflow Manager. Create a session and associate it with the mapping that you created to move data to an Amazon Redshift table. Define the session properties to write data to Amazon Redshift.
- The PowerCenter Integration Service writes the data to a staging directory and then to an Amazon S3 bucket before it writes the data to Amazon Redshift. You must specify the location of the staging directory in the session properties. You must also specify an Amazon S3 bucket name in the session properties. You must have write access to the Amazon S3 bucket.

Session Configuration

S3 Bucket Name

Use the bucket created for Redshift sessions. Create a bucket in the same region as the Redshift cluster.

Enable Compression

Improves the session performance. Enable compression property is enabled by default. The property compresses the staged files before the files are written to Amazon Redshift.

Staging Directory Location

A location on the node where the PowerCenter Integration Service is running. The PowerCenter Integration Service creates a staging file in the location provided in the session properties. This is a volatile directory that PowerCenter Integration Service writes the data to prior to writing the data to Amazon Redshift. The PowerCenter Integration Service writes the data from the staging directory into an Amazon S3 bucket and then deletes the staged files from the staging directory.

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

<staging directory>/infaRedShiftStaging<MMDDHHmmSS>

Truncate Target Table Before Data Load

Allows for the truncation of an Amazon Redshift target before writing data to the target. When this option is enable during the load, the Redshift query analyzer runs the following statements:

```
padb_fetch_sample: select * from <Target table name>
padb_fetch_sample: select count(*) from <Target table name>
```

Batch Size

A critical component in overall performance of the system. Use a batch size high enough to limit the number of batches created to 4 or 5.

Recommended Batch Size = Total number of rows on input source / 5

Success File Directory

A directory on the node where the PowerCenter Integration Service is running. The directory serves as the location for the success file in the session properties. By default, the PowerCenter Integration Service generates the success file with the following naming convention: <sessionName>_<timestamp>_success.csv.

The PowerCenter Integration Service generates a success file after each session execution and has an entry for each record that's successfully written into Amazon Redshift. Each entry describes the values written for all the fields of the record.

Error File Directory

A directory in the node where the PowerCenter Integration Service is running. The directory serves as the location for the error file in the session properties. The error file is used to debug issues when data isn't successfully written to Redshift. The file contains an entry for each data error. Each entry in the file contains the values for all fields of the record and the error message.

By default, the PowerCenter Integration Service writes a blank file to $PMBadFileDir, and the PowerCenter Integration Service generates errors file with the name <sessionName>_<timestamp>_error.csv.

S3 Encryption properties

- *Turn on S3 Server Side Encryption.* Use if server side encryption is already enabled on the S3 buckets, PowerCenter sessions maintain the encryption if this feature is turned on. This is recommended unless you want to use your own encryption keys.
- *Turn on S3 Client Side Encryption.* Use if a private encryption key needs to be used.Provide the master Symmetric Key in the Redshift application connection and turn on S3 client-side encryption.

For more information about S3 encryption, see server side encryption and client side encryption.

Integration Best Practices

Pushdown Optimization

In traditional PowerCenter mappings that use cache transformations, the PowerCenter Integration Service processes the data. RAM becomes a bottleneck when dealing with mappings using Lookup, Sorter, or Aggregator transformations to process large datasets.

Use pushdown optimization to take advantage of Redshift's MPP architecture to quickly do such SQL operations. It's easy to turn on pushdown optimization functionality in a mapping without any major design changes to the mapping.

In this example, we will create a mapping, m_agg_event, using sample tables from a TICKIT database. The mapping uses an EVENT table as lookup to populate an aggregate table AGG_VENUE. The mapping uses a VENUE table as the source.

Use the following link to access a TICKIT database, table DDLs, and sample data:

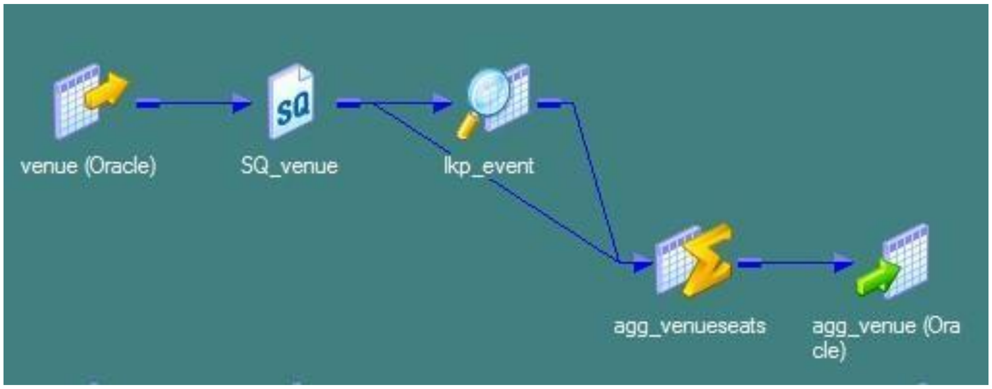


Figure 5: The mapping uses venueid as join column between venue and event tables.

http://docs.aws.amazon.com/redshift/latest/dg/t_creating_database.html.

	Source table	Lookup Table	Target Table
Pipeline 1	create table venue (venueid smallint not null distkey sortkey, venuename varchar(100), venuecity varchar(30), venuestate char(2), venueseats integer);	create table event (eventid integer not null distkey, venueid smallint not null, catid smallint not null, dateid smallint not null sortkey, eventname varchar(200), starttime timestamp);	create table agg_venue (eventname varchar(200), venuename varchar(100), count_venueseats integer);

Aggregator groups perform a count of *venueseats* using *eventname* and *venuename* as group by columns. The full pushdown option in the session is turned on.

The session pushes the following insert DML to the Redshift cluster:

```
INSERT INTO tickit.agg_venue(eventname, venuename, count_venueseats)
SELECT PM_Alkp_event_1.eventname, venue.venuename,
CAST(COUNT(venue.venueseats) AS FLOAT)
FROM (tickit.venue venue LEFT OUTER JOIN tickit.event PM_Alkp_event_1
ON (PM_Alkp_event_1.venueid = venue.venueid))
WHERE ((venue.venueid = (SELECT PM_Alkp_event_1.venueid FROM tickit.event PM_Alkp_event_1
WHERE s(PM_Alkp_event_1.venueid = venue.venueid))) OR (0=0))
GROUP BY PM_Alkp_event_1.eventname, venue.venuename.
```

Compression

- **Database level.** At the table level inside Redshift, column compression is used as a space reduction strategy. Compression saves disk space by compressing column values. Space reduction also minimizes I/O as compressed data is loaded into server memory before being uncompressed. Redshift supports multiple compression encodings and will automatically choose the most efficient based upon your data.

  See the Redshift documentation for a full list of compression encodings allowed in Redshift: http://docs.aws.amazon.com/redshift/latest/dg/c_Compression_encodings.html.

- **Session level.** Use compression at the PowerCenter session level to further reduce space occupied at intermediate stage of ETL and improve performance. PowerCenter supports loading compressed data to Amazon Redshift.

When loading large data sets, enable compression in the session properties of the Redshift target. This allows compression of staged files before writing the files to Amazon S3 bucket. The PowerCenter Integration Service issues a COPY command that copies compressed data from Amazon S3 to the Amazon Redshift target table using the GZIP option.

Here is a sample copy command that gets fired from the PowerCenter Integration Service:

```
copy sample_tbl (a) from 's3://sampleredshiftbucket/0b0ad503-1c2c-4514-95ac-
85a5adb71b3b1441213218371/INSERT_sample_tbl.batch_0.csv.' credentials
'aws_access_key_id=********;aws_secret_access_key=********' MAXERROR 1 DELIMITER ','
QUOTE '"' GZIP NULL '' IGNOREHEADER 1 CSV ROUNDEC ;
```

During the load execution, a COPY command similar to the above example is visible in the Redshift console on https://console.aws.amazon.com/redshift/.

After the Load execution, a COPY command similar to the above example becomes visible in the Redshift console.

Vacuum

Amazon Redshift does not reclaim and reuse space that is freed when you delete rows and update rows, unless specifically instructed by the vacuum command.

Vacuum is important from space as well as performance considerations. Redshift does a soft delete during a delete operation. The rows are marked for delete but not physically removed. Any query running on the table with deleted records will still scan the deleted records as they are not physically removed from the database blocks.

Vacuum performs housekeeping on the database by reclaiming the empty space left by deleted records in a table. Then, it performs a re-sort of the remaining records. A PowerCenter session provides three VACUUM options based on application needs:

- **Full**. When full vacuum is turned on, Amazon Redshift reclaims the space left void by any previous update or delete operation. This is also the default vacuum in Redshift. Another feature of vacuum is the resorting that it does after reclaiming all the unused space.

- **Sort only**. Sorts the new rows after an update or delete, but does not reclaim the disk space left open due to deletes. This option is less resource intensive compared to a full vacuum and allows for optimizer to take advantage of resorted order for query plans.

- **Delete only**. Allows for reclaiming any disk space left open by a previous delete or update operation. Use this option when disk space optimization is the primary goal. This option will not assist in any query optimization.

Analyze

The analyze command does a refresh of the table statistics to help the query optimizer create the most updated plan when it run the query next time. If analyze is not done after considerable records are added or deleted from a table, the optimizer will generate a query plan based on outdated table statistics.

When the analyze option is turned on while executing the load using PowerExchange for Redshift, the session will file a COPY ANALYZE <Target table Name> statement. COPY ANALYZE works on the input data and automatically applies optimal compression encodings to the target table. PowerExchange for Redshift can perform a vacuum and analyze on the whole database or a particular table based on need.

Update else Insert

Update else inserts (upserts) on Redshift perform best when source data coming for update and target table is already distributed on the same node slice. A multiphase ELT for upserts is recommended for upserts or update operations. In a multi-phase approach, the incoming data is staged into a work table before being used for upserts. The work table is distributed on the same column as target table for data locality and pushdown functionality is used to do any resource intensive joins inside of the Redshift engine.

Regions and Availability Zones

Regions are self-contained geographical locations where AWS services are deployed. Regions have their own deployment of each service. Each service within a region has its own endpoint that you can interact with to use the service. Regions contain availability zones, which are isolated fault domain within a general geographical location. Some regions have more availability zones than others. While provisioning, you can choose specific availability zones or let AWS select for you.

Networking, Connectivity and Security

Amazon Virtual Private Cloud (VPC)

VPC has several different configuration options. See the VPC documentation for a detailed explanation of the options and choose based on your networking requirements. You can deploy PowerCenter in either public or private subnets.

Connectivity to the Internet and Other AWS Services

Deploying the instances in a public subnet allows them to have access to the Internet for outgoing traffic as well as to other AWS services, such as S3 and RDS.

Private Data Center Connectivity

You can establish connectivity between your datacenter and the VPC hosting your PowerCenter by using a VPN or Direct Connect. We recommend using Direct Connect so that there is a dedicated link between the two networks with lower latency, higher bandwidth, and enhanced security. You can also connect to EC2 through the Internet via VPN tunnel if you prefer.

Security Groups

Security Groups are analogous to firewalls. You can define rules for EC2 instances and define allowable traffic, IP addresses, and port ranges. Instances can belong to multiple security groups.

About Informatica.

Digital transformation is changing our world. As the leader in enterprise cloud data management, we're prepared to help you intelligently lead the way. To provide you with the foresight to become more agile, realize new growth opportunities or even invent new things. We invite you to explore all that Informatica has to offer—and unleash the power of data to drive your next intelligent disruption. Not just once, but again and again.

About AWS.

For 10 years, Amazon Web Services has been the world's most comprehensive and broadly adopted cloud platform. AWS offers over 70 fully featured services for compute, storage, databases, analytics, mobile, Internet of Things (IoT) and enterprise applications from 33 Availability Zones (AZs) across 12 geographic regions in the U.S., Australia, Brazil, China, Germany, Ireland, Japan, Korea, and Singapore. AWS services are trusted by more than a million active customers around the world – including the fastest growing startups, largest enterprises, and leading government agencies – to power their infrastructure, make them more agile, and lower costs. To learn more about AWS, visit http://aws.amazon.com.

Informatica

Worldwide Headquarters, 2100 Seaport Blvd, Redwood City, CA 94063, USA Phone: 650.385.5000 Fax: 650.385.5500
Toll-free in the US: 1.800.653.3871 informatica.com linkedin.com/company/informatica twitter.com/ Informatica

© 2017 Informatica LLC. All rights reserved. Informatica® and Put potential to work™ are trademarks or registered trademarks of Informatica in the United States and in jurisdictions throughout the world. All other company and product names may be trade names or trademarks.
© Amazon Web Services, Inc or its afiliates. All rights reserved.