Informatica

Informatica® Cloud Data Integration

Amazon Athena Connector

Informatica Cloud Data Integration Amazon Athena Connector
April 2024

Publication Date: 2024-04-07

Table of Contents

Preface

Use *Amazon Athena Connector* to learn how to read from Amazon Athena by using Cloud Data Integration. Learn to create an Amazon Athena connection and develop and run mapping tasks and mappings in Cloud Data Integration.

Informatica Resources

Informatica provides you with a range of product resources through the Informatica Network and other online portals. Use the resources to get the most from your Informatica products and solutions and to learn from other Informatica users and subject matter experts.

Informatica Documentation

Use the Informatica Documentation Portal to explore an extensive library of documentation for current and recent product releases. To explore the Documentation Portal, visit https://docs.informatica.com.

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

Informatica Intelligent Cloud Services web site

You can access the Informatica Intelligent Cloud Services web site at http://www.informatica.com/cloud. This site contains information about Informatica Cloud integration services.

Informatica Intelligent Cloud Services Communities

Use the Informatica Intelligent Cloud Services Community to discuss and resolve technical issues. You can also find technical tips, documentation updates, and answers to frequently asked questions.

Access the Informatica Intelligent Cloud Services Community at:

https://network.informatica.com/community/informatica-network/products/cloud-integration

Developers can learn more and share tips at the Cloud Developer community:

https://network.informatica.com/community/informatica-network/products/cloud-integration/cloud-developers

Informatica Intelligent Cloud Services Marketplace

Visit the Informatica Marketplace to try and buy Data Integration Connectors, templates, and mapplets:

https://marketplace.informatica.com/

Data Integration connector documentation

You can access documentation for Data Integration Connectors at the Documentation Portal. To explore the Documentation Portal, visit https://docs.informatica.com.

Informatica Knowledge Base

Use the Informatica Knowledge Base to find product resources such as how-to articles, best practices, video tutorials, and answers to frequently asked questions.

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

Informatica Intelligent Cloud Services Trust Center

The Informatica Intelligent Cloud Services Trust Center provides information about Informatica security policies and real-time system availability.

You can access the trust center at https://www.informatica.com/trust-center.html.

Subscribe to the Informatica Intelligent Cloud Services Trust Center to receive upgrade, maintenance, and incident notifications. The Informatica Intelligent Cloud Services Status page displays the production status of all the Informatica cloud products. All maintenance updates are posted to this page, and during an outage, it will have the most current information. To ensure you are notified of updates and outages, you can subscribe to receive updates for a single component or all Informatica Intelligent Cloud Services components. Subscribing to all components is the best way to be certain you never miss an update.

To subscribe, on the Informatica Intelligent Cloud Services Status page, click **SUBSCRIBE TO UPDATES**. You can choose to receive notifications sent as emails, SMS text messages, webhooks, RSS feeds, or any combination of the four.

Informatica Global Customer Support

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

To find online support resources on the Informatica Network, click **Contact Support** in the Informatica Intelligent Cloud Services Help menu to go to the **Cloud Support** page. The **Cloud Support** page includes system status information and community discussions. Log in to Informatica Network and click **Need Help** to find additional resources and to contact Informatica Global Customer Support through email.

The telephone numbers for Informatica Global Customer Support are available from the Informatica web site at https://www.informatica.com/services-and-training/support-services/contact-us.html.

CHAPTER 1

Introduction to Amazon Athena Connector

Amazon Athena is a query service that you can use to analyze data in Amazon Simple Storage Service (Amazon S3) by using the standard SQL. Amazon Athena executes multiple queries simultaneously, and fetches the result quickly even with large data sets and complex queries.

Amazon Athena helps you analyze unstructured, semi-structured, and structured data stored in Amazon S3.

Amazon Athena stores the schema in a data catalog or AWS Glue data catalog and uses it when you run queries. Amazon Athena uses Apache Hive to create tables and databases.

Overview

You can use Amazon Athena Connector to read data from Amazon Athena by using Cloud Data Integration. Use Amazon Athena Connector to read flat files and Parquet files from Amazon S3 using Amazon Athena.

You can use Amazon Athena Connector to read data from views and external tables in the Athena data catalog and AWS Glue data catalog. Use Amazon Athena Connector to read and query Amazon S3 files that are mapped to the Amazon S3 location in an external table.

You can run queries in Amazon Athena on encrypted data in Amazon S3. You must run the Amazon Athena queries in the region where Amazon S3 is hosted. You can also encrypt the Amazon Athena query result stored on Amazon S3.

Amazon Athena Connector assets

Create assets in Data Integration to integrate data using Amazon Athena Connector.

When you use Amazon Athena Connector, you can include the following Data Integration assets:

- Data transfer task
- Mapping
- Mapping task

For more information about configuring assets and transformations, see *Mappings*, *Transformations*, and *Tasks* in the Data Integration documentation.

AWS Lake Formation use case

You can use Amazon S3 V2 Connector to write data to Lake Formation and use Amazon Athena Connector to read data from Lake Formation when Lake Formation is configured with specific permissions.

Lake Formation users can have specific permissions such as write access to Lake Formation, read with full access to external tables and columns, read with restricted access to external tables, and read with tag-based access to external tables.

For information on how to integrate data with Lake Formation, see Writing to and reading from AWS Lake Formation .

CHAPTER 2

Connections for Amazon Athena

You can use an Amazon Athena connection to read data from Amazon Athena. You can use the connections to specify sources in mappings and mapping tasks.

Prepare for authentication

You can configure permanent IAM credentials and EC2 instance profile authentication types to access Amazon Athena.

To use the permanent IAM credentials authentication, create an IAM user and generate the access and secret key in the AWS Console. Keep these details handy to use in the connection properties.

To use EC2 instance profile authentication, set up an EC2 instance and attach the EC2 role to the EC2 instance.

Before you configure the connection properties, create the minimal Amazon S3 policy, AWS Glue data catalog policy, and the Amazon Athena policies and define the required permissions for the IAM user or EC2 role in the policies.

Attach the policies to the IAM user or EC2 role based on the authentication type you want to configure.

Create an Amazon S3 policy

Create an Amazon S3 policy and define the permissions to store Amazon Athena results on Amazon S3.

Use the following minimum required permissions to store Amazon Athena results on Amazon S3:

- PutObject
- GetObject
- DeleteObject
- ListBucket
- GetBucketLocation
- ListAllMyBuckets
- GetBucketAcl

You can use the following sample Amazon S3 policy:

```
{
    "Version": "2012-10-17",
    "Statement": [
        {
            "Sid": "VisualEditor0",
```

```
                "Effect": "Allow",
                "Action": [
                    "s3:PutObject",
                    "s3:GetObject",
                    "s3:ListBucket",
                    "s3:DeleteObject"
                ],
                "Resource": "*"
            },
            {
                "Effect": "Allow",
                "Action": [
                    "s3:ListBucket",
                    "s3:GetBucketLocation",
                    "s3:ListAllMyBuckets",
                    "s3:GetBucketAcl"
                ],
                "Resource": [
                    "*"
                ]
            }
        ]
    }
```

Create an AWS Glue data catalog policy

You can use AWS IAM to define policies and roles to access resources used by AWS Glue.

Amazon Athena uses the AWS Glue Data Catalog to store and retrieve table metadata for the Amazon S3 data in your AWS account.

You can use the following sample policy for AWS Glue Data Catalog:

```
{
    "Version": "2012-10-17",
    "Statement": [
        {
            "Effect": "Allow",
            "Action": [
                "glue:*",
            ],
            "Resource": [
                "*"
            ]
        }
    ]
}
```

Create an Amazon Athena policy

Specify the minimum required permissions for Amazon Athena Connector to read data from views and external tables in the AWS Glue data catalog and to read and query Amazon S3 files.

You can use the following minimum required permissions:

- GetWorkGroup
- GetTableMetadata
- StartQueryExecution
- GetQueryResultsStream
- ListDatabases
- GetQueryExecution
- GetQueryResults

- GetDatabase
- ListTableMetadata
- GetDataCatalog
- CreatePreparedStatement
- DeletePreparedStatement

You can use the following sample policy for Amazon Athena:

```
{
    "Version": "2012-10-17",
    "Statement": [
        {
            "Effect": "Allow",
            "Action": [
                "athena:GetWorkGroup",
                "athena:GetTableMetadata",
                "athena:StartQueryExecution",
                "athena:GetQueryResultsStream",
                "athena:ListDatabases",
                "athena:GetQueryExecution",
                "athena:GetQueryResults",
                "athena:GetDatabase",
                "athena:ListTableMetadata",
                "athena:GetDataCatalog"
                "athena:CreatePreparedStatement",
                "athena:DeletePreparedStatement"
            ],
            "Resource": [
                "arn:aws:athena:*:*:workgroup/*",
                "arn:aws:athena:*:*:datacatalog/*"
            ]
        },
        {
            "Effect": "Allow",
            "Action": [
                "athena:ListDataCatalogs",
                "athena:ListWorkGroups"
            ],
            "Resource": "*"
        }
    ]
}
```

Connect to Amazon Athena

Let's configure the Amazon Athena connection properties to connect to Amazon Athena.

Before you begin

Before you get started, you'll need to get information from your Amazon Athena account based on the authentication type that you want to configure.

To configure permanent IAM credentials authentication, get the access key and secret key.

To use EC2 instance profile authentication, set up an EC2 instance and attach the EC2 role to the EC2 instance.

Attach the required policies to the IAM user or EC2 role based on the authentication type you want to configure.

Check out "Prepare for authentication" on page 8 to learn more about the authentication prerequisites.

Connection details

The following table describes the basic connection properties:

Property	Description
Connection Name	Name of the connection. Each connection name must be unique within the organization. Connection names can contain alphanumeric characters, spaces, and the following special characters: _ . + -, Maximum length is 255 characters.
Description	Description of the connection. Maximum length is 4000 characters.
Type	Amazon Athena
Use Secret Vault	Stores sensitive credentials for this connection in the secrets manager that is configured for your organization. This property appears only if secrets manager is set up for your organization. When you enable the secret vault in the connection, you can select which credentials that the Secure Agent retrieves from the secrets manager. If you don't enable this option, the credentials are stored in the repository or on a local Secure Agent, depending on how your organization is configured. For information about how to configure and use a secrets manager, see "Secrets manager configuration" in the Administrator help.
Runtime Environment	The name of the runtime environment where you want to run tasks. Select a Secure Agent, Hosted Agent, or serverless runtime environment.

Authentication types

You can configure permanent IAM credentials and EC2 instance profile authentication types to access Amazon Athena.

Permanent IAM credentials

Permanent IAM credentials authentication is the default type that requires the access key and secret key to connect to Amazon Athena.

The following table describes the basic connection properties for permanent IAM credentials authentication:

Property	Description
Access Key	The access key to connect to Amazon Athena.
Secret Key	The secret key to connect to Amazon Athena.

Property	Description
JDBC URL	The URL to connect to Amazon Athena. Enter the JDBC URL in the following format: `jdbc:awsathena://AwsRegion=<region_name>;S3OutputLocation=<S3_Output_Location>;`
Customer Master Key ID	The customer master key ID generated by AWS Key Management Service (AWS KMS) or the ARN of your custom key for cross-account access when you stage data in Amazon S3. The customer master key serves to encrypt your data at the destination before they are saved in Amazon S3.You can either enter the customer-generated customer master key ID or the default customer master key ID. Ensure that you generate the customer master key for the same region where your Amazon S3 bucket resides.

EC2 instance profile

You can configure AWS Identity and Access Management (IAM) authentication to connect to Amazon Athena when the Secure Agent is installed on an Amazon Elastic Compute Cloud (EC2) system.

The following table describes the basic connection properties for EC2 instance profile authentication:

Property	Description
JDBC URL	The URL of the Amazon Athena connection. Enter the JDBC URL in the following format: `jdbc:awsathena://AwsRegion=<region_name>;S3OutputLocation=<S3_Output_Location>;`
Customer Master Key ID	The customer master key ID generated by AWS Key Management Service (AWS KMS) or the ARN of your custom key for cross-account access when you stage data in Amazon S3. The customer master key serves to encrypt your data at the destination before they are saved in Amazon S3.You can either enter the customer-generated customer master key ID or the default customer master key ID. Ensure that you generate the customer master key for the same region where your Amazon S3 bucket resides.

Proxy server settings

If your organization uses an outgoing proxy server to connect to the Internet, you can use the serverless runtime environment to connect to Informatica Intelligent Cloud Services through the proxy server.

You can use the unauthenticated or authenticated proxy server.

To configure the proxy settings for the serverless runtime environment, see *Runtime Environments* in the Administrator help.

CHAPTER 3

Mappings and mapping tasks with Amazon Athena Connector

Use the Data Integration Mapping Designer to create a mapping. When you create a mapping, you can configure a source to represent the object.

Amazon Athena source features

When you configure the advanced source properties, configure properties specific to Amazon Athena.

The following list shows the Amazon Athena source features that you can use in mappings:

- Server-side encryption
- Server-side encryption with KMS
- Client-side encryption with KMS
- Filter
- Sort
- Data Preview
- Table name override
- Schema name override

Data encryption in Amazon Athena sources

You can run queries in Amazon Athena on encrypted data in Amazon S3. You must run the Amazon Athena queries in the region where Amazon S3 is hosted. You can also encrypt the query results in Amazon S3.

Select the type of the encryption in the **Encryption Type** field under the Amazon Athena advanced source properties.

Use the customer master key ID generated by AWS Key Management Service for server-side encryption.

You can configure the following types of encryption:

None

The data is not encrypted.

SSE-S3

If you select the **SSE-S3** encryption type, Amazon Athena encrypts the file using the Amazon S3-managed key for server-side encryption.

SSE-KMS

If you select the **SSE-KMS** encryption type, Amazon Athena encrypts the file using the AWS KMS-managed key or Amazon Resource Name (ARN) for server-side encryption.

CSE-KMS

If you select the **CSE-KMS** encryption type, Amazon Athena encrypts the file using the AWS KMS-managed key for client-side encryption.

Amazon Athena sources in mappings

In a mapping, you can configure a Source transformation to represent an Amazon Athena source.

The following table describes the Amazon Athena source properties that you can configure in a Source transformation:

Property	Description
Connection	Name of the source connection. Select a source connection, or click **New Parameter** to define a parameter for the source connection.
Source type	Type of the source object. Select any of the following source object types: - Single Object. Select to specify a single Amazon Athena object. - Query. When you select the source type as query, you must map all the fields selected in the query in the **Field Mapping** tab. - Parameter. Select to specify a parameter name.
Object	Name of the source object. Select a single source object.

Note: When you preview a column in a table that uses binary data type, the column displays blank values.

The following table describes the query options that you can configure in a Source transformation:

Property	Description
Filter	Filter value in a read operation. Click **Configure** to add conditions to filter records and reduce the number of rows that the Secure Agent reads from the source. You can specify the following filter conditions: - **Not parameterized**. Use a basic filter to specify the object, field, operator, and value to read specific records. - **Completely parameterized**. Use a parameter to represent the filter conditions. - **Advanced**. Use an advanced filter to define a complex filter condition.
Sort	Conditions to sort records. You can specify the following sort conditions: - **Not parameterized.** Select the fields and type of sorting to use. - **Parameterized.** Use a parameter to specify the sort condition. - **Sort Order.** Specify whether you want to sort data in ascending or descending order.

Note: When you configure an advanced filter on a Timestamp column, you must specify the value as `TIMESTAMP` in the filter condition. For example,

```
SELECT * FROM <dbname>.<tablename> where col_date < TIMESTAMP '2030-06-22 18:30:00.000';
```

The following table describes the Amazon Athena advanced source properties that you can configure in a Source transformation:

Property	Description
Retain Athena Query Result On S3 File	Specifies whether you want to retain the Amazon Athena query result on the Amazon S3 file. Select the check box to retain the Amazon Athena query result on the Amazon S3 file. The Amazon Athena query result in stored in the CSV file format. By default, the **Retain Athena Query Result on S3 File** check box is not selected.
S3OutputLocation	Specifies the location of the Amazon S3 file that stores the result of the Amazon Athena query. You can also specify the Amazon S3 file location in the `S3OutputLocation` parameter in the **JDBC URL** connection property. If you specify the Amazon S3 output location in both the connection and the advanced source properties, the Secure Agent uses the Amazon S3 output location specified in the advanced source properties.
Fetch Size	Determines the number of rows to read in one result set from Amazon Athena. Default is 10000.
Encryption Type	Encrypts the data in the Amazon S3 staging directory. You can select the following encryption types: - None - SSE-S3 - SSE-KMS - CSE-KMS Default is None.
Schema Name	Overrides the schema name of the source object.
Source Table Name	Overrides the table name used in the metadata import with the table name that you specify.
SQL Query	Overrides the default SQL query. Enclose column names in double quotes. The SQL query is case sensitive. Specify an SQL statement supported by the Amazon Athena database. When you specify the columns in the SQL query, ensure that the column name in the query matches the source column name in the mapping.

You can set the tracing level in the Amazon Athena advanced source properties to determine the amount of details that logs contain.

The following table describes the tracing levels that you can configure:

Tracing Level	Description
Terse	The Secure Agent logs initialization information, error messages, and notification of rejected data.
Normal	The Secure Agent logs initialization and status information, errors encountered, and skipped rows due to transformation row errors. Summarizes session results, but not at the level of individual rows.

Tracing Level	Description
Verbose Initialization	In addition to normal tracing, the Secure Agent logs additional initialization details, names of index and data files used, and detailed transformation statistics.
Verbose Data	In addition to verbose initialization tracing, the Secure Agent logs each row that passes into the mapping. Also notes where the Secure Agent truncates string data to fit the precision of a column and provides detailed transformation statistics. When you configure the tracing level to verbose data, the Secure Agent writes row data for all rows in a block when it processes a transformation.

Amazon Athena sources in mapping tasks

For Amazon Athena source connections used in template-based mapping tasks, you can configure advanced properties in the **Sources** page.

The **Sources** page appears in the mapping task wizard if you defined a parameter for the connection or source object in the associated mapping.

You can configure all the advanced source properties, except **Tracing Levels**, for a mapping task that you configured for a mapping. For more information, see "Amazon Athena sources in mappings" on page 14.

Note: If the mapping uses a specific source connection and a parameter for the source object, you can specify these advanced properties in the mapping and also specify them in the mapping task. In this case, the properties in the mapping task override those in the mapping.

CHAPTER 4

Data type reference

Data Integration uses the following data types in mappings and mapping tasks with Amazon Athena:

Amazon Athena native data types

Amazon Athena data types appear in the **Fields** tab of the Source and Target transformations when you edit metadata for the fields.

Transformation data types

Set of data types that appear in the transformations. They are internal data types based on ANSI SQL-92 generic data types, which the Secure Agent uses to move data across platforms. Transformation data types appear in all transformations in a mapping.

When Data Integration reads source data, it converts the native data types to the comparable transformation data types before transforming the data. When Data Integration writes to a target, it converts the transformation data types to the comparable native data types.

Amazon Athena and transformation data types

The following table lists the Amazon Athena native data types that Data Integration supports and the corresponding transformation data types:

Amazon Athena Data Type	Transformation Data Type	Description
Bigint	Bigint	Signed eight-byte integer
Binary	Binary	1 to 104,857,600 bytes
Boolean	Integer	Logical Boolean (true/false)
Char	String	Fixed-length character string
Date	Date/Time	Calendar date (year, month, day)
Decimal	Decimal	Exact numeric of selectable precision
Double	Double	Precision 15
Float	Double	Precision 15
Int	Integer	Signed four-byte integer

Amazon Athena Data Type	Transformation Data Type	Description
Smallint	Integer	Signed two-byte integer
String	String	-1 to 104,857,600 characters
Timestamp	Date/Time	Date and time (without time zone)
Tinyint	Integer	Signed one-byte integer
Varchar	String	Variable-length character string with a user-defined limit

Note: You can use the Binary data type only with Parquet files.

INDEX