



Informatica® Data Integration - Free & PayGo

# Microsoft Azure Data Lake Storage Gen2 Connector

© Copyright Informatica LLC 2019, 2023

This software and documentation are provided only under a separate license agreement containing restrictions on use and disclosure. No part of this document may be reproduced or transmitted in any form, by any means (electronic, photocopying, recording or otherwise) without prior consent of Informatica LLC.

U.S. GOVERNMENT RIGHTS Programs, software, databases, and related documentation and technical data delivered to U.S. Government customers are "commercial computer software" or "commercial technical data" pursuant to the applicable Federal Acquisition Regulation and agency-specific supplemental regulations. As such, the use, duplication, disclosure, modification, and adaptation is subject to the restrictions and license terms set forth in the applicable Government contract, and, to the extent applicable by the terms of the Government contract, the additional rights set forth in FAR 52.227-19, Commercial Computer Software License.

Informatica, the Informatica logo, Informatica Cloud, and PowerCenter 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.

Portions of this software and/or documentation are subject to copyright held by third parties. Required third party notices are included with the product.

See patents at <https://www.informatica.com/legal/patents.html>.

DISCLAIMER: Informatica LLC provides this documentation "as is" without warranty of any kind, either express or implied, including, but not limited to, the implied warranties of noninfringement, merchantability, or use for a particular purpose. Informatica LLC does not warrant that this software or documentation is error free. The information provided in this software or documentation may include technical inaccuracies or typographical errors. The information in this software and documentation is subject to change at any time without notice.

#### NOTICES

This Informatica product (the "Software") includes certain drivers (the "DataDirect Drivers") from DataDirect Technologies, an operating company of Progress Software Corporation ("DataDirect") which are subject to the following terms and conditions:

1. THE DATADIRECT DRIVERS ARE PROVIDED "AS IS" WITHOUT WARRANTY OF ANY KIND, EITHER EXPRESSED OR IMPLIED, INCLUDING BUT NOT LIMITED TO, THE IMPLIED WARRANTIES OF MERCHANTABILITY, FITNESS FOR A PARTICULAR PURPOSE AND NON-INFRINGEMENT.
2. IN NO EVENT WILL DATADIRECT OR ITS THIRD PARTY SUPPLIERS BE LIABLE TO THE END-USER CUSTOMER FOR ANY DIRECT, INDIRECT, INCIDENTAL, SPECIAL, CONSEQUENTIAL OR OTHER DAMAGES ARISING OUT OF THE USE OF THE ODBC DRIVERS, WHETHER OR NOT INFORMED OF THE POSSIBILITIES OF DAMAGES IN ADVANCE. THESE LIMITATIONS APPLY TO ALL CAUSES OF ACTION, INCLUDING, WITHOUT LIMITATION, BREACH OF CONTRACT, BREACH OF WARRANTY, NEGLIGENCE, STRICT LIABILITY, MISREPRESENTATION AND OTHER TORTS.

The information in this documentation is subject to change without notice. If you find any problems in this documentation, report them to us at [infa\\_documentation@informatica.com](mailto:infa_documentation@informatica.com).

Informatica products are warranted according to the terms and conditions of the agreements under which they are provided. INFORMATICA 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.

Publication Date: 2023-04-04

# Table of Contents

<b>Preface .....</b>	<b>5</b>
Informatica Resources. ....	5
Informatica Documentation. ....	5
Informatica Intelligent Cloud Services web site. ....	5
Informatica Intelligent Cloud Services Communities. ....	5
Informatica Intelligent Cloud Services Marketplace. ....	5
Simple Data Integration connector documentation. ....	6
Informatica Knowledge Base. ....	6
Informatica Intelligent Cloud Services Trust Center. ....	6
Informatica Global Customer Support. ....	6
 <b>Chapter 1: Introduction to Microsoft Azure Data Lake Storage Gen2 Connector .....</b>	 <b>7</b>
Microsoft Azure Data Lake Storage Gen2 Connector assets. ....	7
Administration of Microsoft Azure Data Lake Storage Gen2 Connector. ....	8
Managed identity authentication. ....	9
Access control lists. ....	9
 <b>Chapter 2: Connections for Microsoft Azure Data Lake Storage Gen2.....</b>	 <b>11</b>
Microsoft Azure Data Lake Storage Gen2 connection properties. ....	11
Configuring the proxy server. ....	12
Configuring proxy server settings on Windows. ....	13
Configuring proxy server settings on Linux. ....	14
Bypassing the proxy server. ....	14
 <b>Chapter 3: Mappings for Microsoft Azure Data Lake Storage Gen2.....</b>	 <b>16</b>
Microsoft Azure Data Lake Storage Gen2 sources in mappings. ....	16
Directory source in Microsoft Azure Data Lake Storage Gen2 sources. ....	19
Wildcard characters. ....	19
Reading files from subdirectories. ....	20
Pushdown optimization. ....	20
File formatting options. ....	20
Fixed-width file formats. ....	22
FileName field. ....	22
Reading source objects path. ....	23
Parameterization. ....	23
 <b>Chapter 4: Data type reference .....</b>	 <b>24</b>
Flat file data types and transformation data types. ....	24
Avro data types and transformation data types. ....	25

JSON data types and transformation data types. . . . .	26
ORC data types and transformation data types. . . . .	26
Parquet data types and transformation data types. . . . .	27
<b>Chapter 5: Troubleshooting. . . . .</b>	<b>29</b>
Troubleshooting a mapping. . . . .	29
<b>Index. . . . .</b>	<b>31</b>

# Preface

Use *Microsoft Azure Data Lake Storage Gen2 Connector* to learn how to read from Microsoft Azure Data Lake Storage Gen2. Learn to create a connection, develop and run mappings, mapping tasks, and data transfer tasks in 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](mailto: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/>

## Simple Data Integration connector documentation

You can access documentation for Simple 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](mailto: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, go to <https://status.informatica.com/> and click **SUBSCRIBE TO UPDATES**. You can then 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 Customer Support Center by telephone or online.

For online support, click **Submit Support Request** in Informatica Intelligent Cloud Services. You can also use Online Support to log a case. Online Support requires a login. You can request a login at <https://network.informatica.com/welcome>.

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 Microsoft Azure Data Lake Storage Gen2 Connector

You can use Microsoft Azure Data Lake Storage Gen2 Connector to securely read data from Microsoft Azure Data Lake Storage Gen2.

Use Microsoft Azure Data Lake Storage Gen2 Connector to read flat files and flat structured complex files such as Avro, JSON, ORC, and Parquet. You can use Microsoft Azure Data Lake Storage Gen2 objects as sources in mappings and mapping tasks.

## Microsoft Azure Data Lake Storage Gen2 Connector assets

Create assets in Data Integration to integrate data using Microsoft Azure Data Lake Storage Gen2 Connector.

When you use Microsoft Azure Data Lake Storage Gen2 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.

# Administration of Microsoft Azure Data Lake Storage Gen2 Connector

Before you use Microsoft Azure Data Lake Storage Gen2 objects in tasks, an administrator must perform the following tasks:

- Create a storage account to use with Microsoft Azure Data Lake Storage Gen2 and enable **Hierarchical namespace** in the Azure portal.  
You can use role-based access control or access control lists to authorize the users to access the resources in the storage account.
  - **Role-based access control**  
If you use role-based access control, assign the Contributor role or Reader role to the users.  
  
The contributor role grants you full access to manage all resources in the storage account, but does not allow you to assign roles.  
  
The reader role allows you to view all resources in the storage account, but does not allow you to make any changes.  
  
**Note:** To add or remove role assignments, you must have write and delete permissions, such as an Owner role.
  - **Access control lists**  
If you use access control lists, you can provide read, write, and execute permissions to each directory and file for users.
- Register an application in Azure Active Directory to authenticate users to access the Microsoft Azure Data Lake Storage Gen2 account.  
You can use role-based access control or access control lists to authorize the application.
  - **Role-based access control**  
If you use role-based access control, assign the Storage Blob Data Contributor or Storage Blob Data Reader role to the application.  
  
The Storage Blob Data Contributor role lets you read, write, and delete Azure Storage containers and blobs in the storage account.  
  
The Storage Blob Data Reader role lets you only read and list Azure Storage containers and blobs in the storage account.
  - **Access control lists**  
If you use access control lists, you can provide read, write, and execute permissions to each directory and file in the container.
- Create a Blob container in the storage account.
- Create an Azure Active Directory web application for service-to-service authentication with Microsoft Azure Data Lake Storage Gen2.  
**Note:** Ensure that you have superuser privileges to access the folders or files created in the application using the connector.
- To read complex files, set the JVM options for type DTM to increase the -Xms and -Xmx values in the system configuration details of the Secure Agent to avoid java heap space error. The recommended -Xms and -Xmx values are 512 MB and 1024 MB respectively.
- When you read from Microsoft Azure Data Lake Storage Gen2, an interim directory in the agent machine is used to stage the files.  
Ensure that the interim directory is created in the agent machine and you have the read permissions to the interim directory.



For more information about configuring a Microsoft Azure Data Lake Storage Gen2 connection, see the Informatica How-To Library article, [Prerequisites to create a Microsoft Azure Data Lake Storage Gen2 connection](#).

## Managed identity authentication

Before you use managed identity authentication to connect to Microsoft Azure Data Lake Storage Gen2, complete the following prerequisites:

1. Create an Azure virtual machine.
2. Install the Secure Agent on the Azure virtual machine.
3. Enable system assigned identity or user assigned identity for the Azure virtual machine.  
If you enable both and do not specify the client ID, the system assigned identity is used for authentication.
4. After you add or remove a managed identity, restart the Azure virtual machine.

### Rules and guidelines for managed identity authentication

Consider the following rules and guidelines for managed identity authentication:

- When you create a Microsoft Azure Data Lake Storage Gen2 connection, select the Azure virtual machine that you configured as the runtime environment.
- If you enable system assigned identity, assign the required role or permissions to the Azure virtual machine to run the mappings and tasks.  
If you enable user assigned identity, assign the required role or permissions to the user assigned identity.

For example, if you use role-based access control, assign the Storage Blob Data Contributor role and if you use access control lists, assign the read, write, and execute permissions.

- You cannot use a proxy server with managed identity authentication.

## Access control lists

You can use access control lists to grant different levels of permissions to access directories and files to each user and service. If you do not want to use role-based access control to grant access to all of the data in a storage account, you can use access control lists to grant read and execute permissions to a specific directory or file.

**Note:** When you grant access to a file, you must assign the *execute* permission to the root folder of the container and to each folder in the hierarchy of folders that lead to the file.

The following table lists the permissions required when you create a mapping to access Microsoft Azure Data Lake Storage Gen2 using access control lists:

Action	Permissions
Test a connection	Read and execute permissions for the source.
Access an object	Read and execute permissions for the source folder that contains the object.
Select a file format	Read permission for the source file.
Preview data	Read permission for the source file.

Action	Permissions
Read data from a file or directory as source type	Read permission for the source file.
Use wildcard characters or recursive directory read	<ul style="list-style-type: none"> <li>- Read and execute permissions for the folders for which you want to use wildcard characters or recursive read.</li> <li>- Read permission for the source file.</li> </ul>

## CHAPTER 2

# Connections for Microsoft Azure Data Lake Storage Gen2

Create a Microsoft Azure Data Lake Storage Gen2 connection to securely read data from Microsoft Azure Data Lake Storage Gen2. You can use a Microsoft Azure Data Lake Storage Gen2 connection to specify sources in mappings and mapping tasks.

## Microsoft Azure Data Lake Storage Gen2 connection properties

When you set up a Microsoft Azure Data Lake Storage Gen2 connection, configure the connection properties.

The following table describes the Microsoft Azure Data Lake Storage Gen2 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	The Microsoft Azure Data Lake Storage Gen2 connection type.
Runtime Environment	The name of the runtime environment where you want to run the tasks. Specify a Secure Agent or a Hosted Agent.
Account Name	Microsoft Azure Data Lake Storage Gen2 account name or the service name.
Authentication Type	Authentication type to access the Microsoft Azure Data Lake Storage Gen2 account. Select one of the following options: <ul style="list-style-type: none"><li>- Service Principal Authentication. Uses the client ID, client secret, and tenant ID to connect to Microsoft Azure Data Lake Storage Gen2.</li><li>- Shared Key Authentication. Uses the account key to connect to Microsoft Azure Data Lake Storage Gen2.</li><li>- Managed Identity Authentication. Select to authenticate using identities that are assigned to applications in Azure to access Azure resources in Microsoft Azure Data Lake Storage Gen2.</li></ul>

Property	Description
Client ID	<p>Applies to Service Principal Authentication and Managed Identity Authentication.</p> <p>The client ID of your application.</p> <p>To use service principal authentication, specify the application ID or client ID for your application registered in the Azure Active Directory.</p> <p>To use managed identity authentication, specify the client ID for the user-assigned managed identity. If the permission is provided by system-assigned managed identity, leave the field empty. If there is no system-assigned identity but only a single user-assigned managed identity, you may also leave the field empty.</p>
Client Secret	<p>Applies to Service Principal Authentication.</p> <p>The client secret key to complete the OAuth authentication in the Azure Active Directory.</p>
Tenant ID	<p>Applies to Service Principal Authentication.</p> <p>The directory ID of the Azure Active Directory.</p>
Account Key	<p>Applies to Shared Key Authentication.</p> <p>The account key for the Microsoft Azure Data Lake Storage Gen2 account.</p>
File System Name	The name of the file system in the Microsoft Azure Data Lake Storage Gen2 account.
Directory Path	<p>The path of an existing directory without the file system name.</p> <p>You can select one of the following syntax:</p> <ul style="list-style-type: none"> <li>- / for root directory</li> <li>- /dir1</li> <li>- dir1/dir2</li> </ul> <p>There is no default directory.</p>
Adls Gen2 End-point	<p>The type of Microsoft Azure endpoints.</p> <p>Select one of the following endpoints:</p> <ul style="list-style-type: none"> <li>- core.windows.net. Connects to Azure endpoints.</li> <li>- core.usgovcloudapi.net. Connects to US government Microsoft Azure Data Lake storage Gen2 endpoints.</li> <li>- core.chinacloudapi.cn. Connects to Microsoft Azure Data Lake storage Gen2 endpoints in the China region.</li> </ul> <p>Default is core.windows.net.</p>

## Configuring the proxy server

If your organization uses an outgoing proxy server to connect to the internet, the agent connects to Informatica Intelligent Cloud Services and the Microsoft Azure Data Lake Storage Gen2 endpoints through the proxy server.

You can configure the Secure Agent to use the proxy server on Windows and Linux. You can use the unauthenticated or authenticated proxy server.

You can configure the Secure Agent to use the proxy server on Windows and Linux. Contact your network administrator for the correct proxy settings.

## Configuring proxy server settings on Windows

To configure the proxy server settings for the Secure Agent on a Windows machine, you can configure the proxy server settings through the Secure Agent or the JVM options of the Secure Agent.

### Configuring proxy server settings through the Secure Agent Manager

To configure the proxy server settings through the Secure Agent Manager, perform the following steps:

1. Click **Start > All Programs > Informatica Cloud Secure Agent > Informatica Cloud Secure Agent** to launch the Secure Agent Manager.

The Secure Agent Manager displays the Secure Agent status.

2. Click **Proxy** in the Secure Agent Manager page.
3. Click **Use a Proxy Server** to enter proxy server settings.
4. Configure the following proxy server details:

Field	Description
Proxy Host	Required. Host name of the outgoing proxy server that the Secure Agent uses.
Proxy Port	Required. Port number of the outgoing proxy server.

5. Click **OK**.

The Secure Agent Manager restarts the Secure Agent to apply the settings.

### Configuring proxy server settings through the JVMOptions

1. Log in to Informatica Intelligent Cloud Services.
2. Open Administrator and select **Runtime Environments**.
3. Select the Secure Agent for which you want to configure a proxy server.
4. On the upper-right corner of the page, click **Edit**.
5. In the **System Configuration Details** section, select the **Type** as **DTM** for the Data Integration Service.
  - Add the following parameters in any **JVMOption** field and specify appropriate values for each parameter:

Parameter	Description
-Dhttp.proxyHost=	Host name of the outgoing HTTP proxy server.
-Dhttp.proxyPort=	Port number of the outgoing HTTP proxy server.
-Dhttp.proxyUser=	Authenticated user name for the HTTP proxy server. This is required if the proxy server requires authentication.
-Dhttp.proxyPassword=	Password for the authenticated user. This is required if the proxy server requires authentication.
-Dhttps.proxyHost=	Host name of the outgoing HTTPS proxy server.

Parameter	Description
-Dhttps.proxyPort=	Port number of the outgoing HTTPS proxy server.
-Dhttps.proxyUser=	Authenticated user name for the HTTPS proxy server. This is required if the proxy server requires authentication.
-Dhttps.proxyPassword=	Password for the authenticated user. This is required if the proxy server requires authentication.

Example for HTTP:

```
JVMOption1=-Dhttp.proxyHost=<proxy_server_hostname>
JVMOption2=-Dhttp.proxyPort=8081
JVMOption3=-Dhttp.proxyUser=<proxy_user_name>
JVMOption4=-Dhttp.proxyPassword=<proxy_password>
```

Example for HTTPS,

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

6. Click **Save**.

The Secure Agent restarts to apply the settings.

## Configuring proxy server settings on Linux

The Secure Agent installer configures the proxy server settings for the Secure Agent based on settings configured in the browser. You can update the proxy server settings defined for the Secure Agent from the command line.

To configure the proxy server settings for the Secure Agent on a Linux machine, use a shell command that updates the `proxy.ini` file. Contact the network administrator to determine the proxy settings.

1. Navigate to the following directory:  
`<Secure Agent installation directory>/apps/agentcore`
2. Update the `proxy.ini` file.
  - To update the `proxy.ini` file for an unauthenticated proxy, enter the following command:  
`./consoleAgentManager.sh configureProxy <proxy host> <proxy port>`
3. Restart the Secure Agent.

## Bypassing the proxy server

You can bypass the proxy server settings configured for the Secure Agent from the command line.

Perform the following steps to bypass the proxy server:

1. Navigate to the following directory:  
`<Secure Agent installation directory>/apps/agentcore`
2. Specify the following command in the `proxy.ini` file:

```
InfaAgent.NonProxyHost=localhost|{*}core.windows.net|127.|[\\:1]*
```

To bypass proxy server for service principal authentication, append `login.microsoftonline.com` to the command.

To bypass proxy server for managed identity authentication, append `169.254.169.254` to the command.

For example,

```
InfaAgent.NonProxyHost=localhost|127.*|[\\:1]|<accountname>.blob.core.windows.net|  
<accountname>.dfs.core.windows.net|<accountname>.blob.core.windows.net|  
login.microsoftonline.com|169.254.169.254
```

3. Restart the Secure Agent.

## CHAPTER 3

# Mappings for Microsoft Azure Data Lake Storage Gen2

When you configure a mapping, you describe the flow of data from the source to the target.

A mapping defines reusable data flow logic that you can use in mapping tasks.

When you create a mapping, you define the Source transformation to represent a Microsoft Azure Data Lake Storage Gen2 object. Use the Mapping Designer in Data Integration to add the Source or Target transformations in the mapping canvas and configure the Microsoft Azure Data Lake Storage Gen2 source properties.

You can use Monitor to monitor the jobs.

## Microsoft Azure Data Lake Storage Gen2 sources in mappings

In a mapping, you can configure a source transformation to represent a single Microsoft Azure Data Lake Storage Gen2 object.

The following table describes the Microsoft Azure Data Lake Storage Gen2 source properties that you can configure in a source transformation:

Property	Description
Connection	Name of the source connection. Select a source connection or click <b>New Parameter</b> to define a new parameter for the source connection.  When you select the Allow parameter to be overridden at run time option, ensure that you provide a parameter file and in the correct format.  When you switch between a non-parameterized and a parameterized Microsoft Azure Data Lake Storage Gen2 connection, the advanced property values are retained.
Source Type	Select Single Object or Parameter.
Object	Name of the source object.  Ensure that the headers or file data does not contain special characters.



Property	Description
Parameter	<p>Select an existing parameter for the source object or click <b>New Parameter</b> to define a new parameter for the source object. The <b>Parameter</b> property appears only if you select Parameter as the source type.</p> <p>When you parameterize the source object, specify the complete object path including the file system in the default value of the parameter.</p>
Format	<p>Specifies the file format that the Microsoft Azure Data Lake Storage Gen2 Connector uses to read data from Microsoft Azure Data Lake Storage Gen2.</p> <p>You can select the following file format types:</p> <ul style="list-style-type: none"> <li>- Flat</li> <li>- Avro</li> <li>- Parquet</li> <li>- JSON</li> <li>- ORC</li> </ul> <p>Default is <b>None</b>. If you select <b>None</b> as the format type, Microsoft Azure Data Lake Storage Gen2 Connector reads data from Microsoft Azure Data Lake Storage Gen2 files in binary format.</p> <p>You cannot read a JSON file that exceeds 1 GB.</p> <p><b>Note:</b> Ensure that the source file is not empty.</p> <p>For more information, see <a href="#">"File formatting options" on page 20</a></p>

The following table describes the Microsoft Azure Data Lake Storage Gen2 source advance properties:

Property	Description
Concurrent Threads	<p>Number of concurrent connections to extract data from the Microsoft Azure Data Lake Storage Gen2. When reading a large file or object, you can spawn multiple threads to process data. Configure <b>Block Size</b> to divide a large file into smaller parts.</p> <p>Default is 4. Maximum is 10.</p>
Filesystem Name Override	<p>Overrides the default file system name.</p>
Source Type	<p>Select the type of source from which you want to read data. You can select the following source types:</p> <ul style="list-style-type: none"> <li>- File</li> <li>- Directory</li> </ul> <p>Default is File.</p>
Allow Wildcard Characters	<p>Indicates whether you want to use wildcard characters for the directory source type.</p> <p>For more information, see <a href="#">"Wildcard characters" on page 19</a>.</p>

Property	Description
Directory Override	<p>Microsoft Azure Data Lake Storage Gen2 directory that you use to read data. Default is root directory. The directory path specified at run time overrides the path specified while creating a connection.</p> <p>You can specify an absolute or a relative directory path:</p> <ul style="list-style-type: none"> <li>- Absolute path. The Secure Agent searches this directory path in the specified file system. Example of absolute path: <code>Dir1/Dir2</code></li> <li>- Relative path. The Secure Agent searches this directory path in the native directory path of the object. Example of relative path: <code>/Dir1/Dir2</code></li> </ul> <p>When you use the relative path, the imported object path is added to the file path used during the metadata fetch at runtime.</p> <p>Do not specify a root directory (<code>/</code>) to override the directory.</p>
File Name Override	Source object. Select the file from which you want to read data. The file specified at run time overrides the file specified in Object.
Block Size	<p>Applicable to flat file format. Divides a large file into smaller specified block size. When you read a large file, divide the file into smaller parts and configure concurrent connections to spawn the required number of threads to process data in parallel.</p> <p>Specify an integer value for the block size.</p> <p>Default value in bytes is 8388608.</p>
Timeout Interval	Not applicable.
Recursive Directory Read	<p>Indicates whether you want to read objects stored in subdirectories in mappings.</p> <p>For more information, see <a href="#">"Reading files from subdirectories" on page 20</a></p>
Incremental File Load	Not applicable.
Compression Format	<p>Reads compressed data from the source.</p> <p>Select one of the following options:</p> <ul style="list-style-type: none"> <li>- None. Select to read Avro, ORC, and Parquet files that use Snappy compression. The compressed files must have the <code>.snappy</code> extension. You cannot read compressed JSON files.</li> <li>- Gzip. Select to read flat files and Parquet files that use Gzip compression. The compressed files must have the <code>.gz</code> extension.</li> </ul> <p>You cannot preview data for a compressed flat file.</p>
Interim Directory	<p>Optional. Applicable to flat files and JSON files.</p> <p>Path to the staging directory in the Secure Agent machine.</p> <p>Specify the staging directory where you want to stage the files when you read data from Microsoft Azure Data Lake Storage Gen2. Ensure that the directory has sufficient space and you have write permissions to the directory.</p> <p>Default staging directory is <code>/tmp</code>.</p> <p>You cannot specify an interim directory when you use the Hosted Agent.</p>
Tracing Level	Sets the amount of detail that appears in the log file. You can choose terse, normal, verbose initialization or verbose data. Default is normal.

## Directory source in Microsoft Azure Data Lake Storage Gen2 sources

You can select the type of source from which you want to read data.

You can select the following type of sources from the **Source Type** option under the advanced source properties:

- File
- Directory

Use the following rules and guidelines to select **Directory** as the source type:

- All the source files in the directory must contain the same metadata.
- All the files must have data in the same format. For example, delimiters, header fields, and escape characters must be same.
- All the files under a specified directory are parsed. To parse the files in the subdirectories, use recursive read.

For more information, see [Reading files from subdirectories](#).

## Wildcard characters

When you read data from an Avro, flat, JSON, ORC, or Parquet file, you can use wildcard characters to specify the source file name.

To use wildcard characters for the source file name, select the source type as **Directory** and enable the **Allow Wildcard Characters** option in the advanced source properties.

When you read an Avro, JSON, ORC, Parquet, or flat file, you can use the ? and \* wildcard characters to define one or more characters in a search.

You can use the following wildcard characters:

### ? (Question mark)

The question mark character (?) allows one occurrence of any character. For example, if you enter the source file name as `a?b.txt`, the Secure Agent reads data from files with the following names:

- `a1b.txt`
- `a2b.txt`
- `aab.txt`
- `acb.txt`

### \* (Asterisk)

The asterisk mark character (\*) allows zero or more than one occurrence of any character. If you enter the source file name as `a*b.txt`, the Secure Agent reads data from files with the following names:

- `aab.txt`
- `a1b.txt`
- `ab.txt`
- `abc11b.txt`

## Rules and guidelines for wildcard characters

Consider the following rules and guidelines when you use wildcard characters:

- When you read a complex file in a mapping, do not use a tilde (~) in the sub-directory name or file name.
- When you use wildcard characters in directory override, the Secure Agent reads data from the folders as well as the files that match the name pattern.

## Reading files from subdirectories

You can read objects stored in subdirectories in Microsoft Azure Data Lake Storage Gen2 in mappings.

You can use recursive read for flat files and complex files in mappings.

To enable recursive read, select the source type as **Directory** in the advanced source properties. Enable the **Recursive Directory Read** advanced source property to read objects stored in subdirectories.

## Rules and guidelines for reading from subdirectories

Consider the following rules and guidelines when you read objects stored in subdirectories:

- When you read from a flat file in Microsoft Azure Data Lake Storage Gen2, ensure that the directory or subdirectory name does not contain the percentage (%) character. Else, the mapping fails.
- When you read a complex file in a mapping, do not use a tilde (~) in the subdirectory name or file name.
- When you read a flat file with only headers and no data and map the FileName field, the expected directory structure is not created with the FileName field.

## Pushdown optimization

You can enable full pushdown optimization when you want to load data from Microsoft Azure Data Lake Storage Gen2 sources to your data warehouse in Microsoft Azure Synapse SQL. While loading the data to Microsoft Azure Synapse SQL, you can transform the data as per your data warehouse model and requirements. When you enable full pushdown on a mapping task, the mapping logic is pushed to the Azure environment to leverage Azure commands. For more information, see the help for Microsoft Azure Synapse SQL Connector.

If you need to load data to any other supported cloud data warehouse, see the connector help for the applicable cloud data warehouse.

## File formatting options

Select the format of the Microsoft Azure Data Lake Storage Gen2 file and configure the formatting options.

The following table describes the formatting options for Avro, Parquet, JSON, ORC, and delimited flat files:

Property	Description
Schema Source	<p>The schema of the source file.</p> <p>Select one of the following options to specify a schema:</p> <ul style="list-style-type: none"> <li>- Read from data file. Imports the schema from a file in Microsoft Azure Data Lake Storage Gen2.</li> <li>- Import from schema file. Imports the schema from a schema definition file in the agent machine.</li> </ul>
Schema File	The schema definition file in the agent machine from where you want to upload the schema.

The following table describes the formatting options for flat files:

Property	Description
Flat File Type	<p>The type of flat file.</p> <p>Select one of the following options:</p> <ul style="list-style-type: none"> <li>- Delimited. Reads a flat file that contains column delimiters.</li> <li>- Fixed Width. Reads a flat file with fields that have a fixed length.</li> </ul> <p>You must select the file format in the <b>Fixed Width File Format</b> option.</p> <p>If you do not have a fixed-width file format, click <b>New &gt; Components &gt; Fixed Width File Format</b> to create one.</p>
Delimiter	<p>Character used to separate columns of data in a delimited flat file. You can set values as comma, tab, colon, semicolon, or others.</p> <p>You can use a single character or multi-character delimiter.</p> <p>You cannot set a tab as a delimiter directly in the <b>Delimiter</b> field. To set a tab as a delimiter, you must type the tab character in any text editor. Then, copy and paste the tab character in the <b>Delimiter</b> field.</p>
EscapeChar	Character immediately preceding a column delimiter character embedded in an unquoted string, or immediately preceding the quote character in a quoted string data in a delimited flat file.
Qualifier	Quote character that defines the boundaries of data in a delimited flat file. You can set qualifier as single quote or double quote.
Qualifier Mode	Not applicable.
Code Page	<p>Select the code page that the Secure Agent must use to read data from a delimited flat file.</p> <p>Select UTF-8 for mappings.</p>
Header Line Number	<p>Specify the line number that you want to use as the header when you read data from a delimited flat file.</p> <p>Specify the value as 0 or 1.</p> <p>To read data from a file with no header, specify the value as 0.</p>
First Data Row	<p>Specify the line number from where you want the Secure Agent to read data in a delimited flat file. You must enter a value that is greater or equal to one.</p> <p>To read data from the header, the value of the <b>Header Line Number</b> and the <b>First Data Row</b> fields should be the same. Default is 1.</p>
Target Header	Not applicable.
Distribution Column	Not applicable.

Property	Description
Max Rows To Preview	Not applicable.
Row Delimiter	Not applicable.

The following table describes the formatting options for JSON files:

Property	Description
Data elements to sample	Not applicable.
Memory available to process data	Not applicable.
Read multiple-line JSON files	Not applicable.

## Fixed-width file formats

You can use a fixed-width flat file as a source in mappings and mapping tasks.

When you configure a Source transformation and select the fixed-width flat file type, you must select the most appropriate fixed-width file format to use based on the data in the fixed-width flat file. Ensure that the sample flat file only uses UTF-8 character set encoding.

Consider the following rules and guidelines for a fixed-width flat file:

- You cannot use a fixed-width flat file as a source in data transfer tasks.
- When you create the fixed-width file format, ensure that the sample file uses the following character as the new line symbol, based on the operating system where the Secure Agent is installed:
  - For Linux, use `\n` character.
  - For Windows, use `\r\n` character.

The source file must also use the same character as defined in the sample file.

- When you use a fixed-width flat file as a source, you cannot edit the metadata for the fields.
- When you read data of the date data type, you can read the date only up to milliseconds.

## FileName field

A FileName field is a string field that contains the source path of a file. The default precision for a FileName field is 255 characters for a flat file and 1024 characters for a complex file.

You cannot configure the FileName field. You can delete the FileName field if you do not want to read the data in the FileName field. You cannot create a folder name with more than 255 characters for a flat file and 1024 characters for a complex file.

FileName is a reserved keyword. Avoid using FileName as the column name in the source data. The name is case sensitive.

The FileName field is applicable to the following file formats:

- Flat file
- Avro
- Parquet
- ORC

## Reading source objects path

When you import source objects, the Secure Agent appends a FileName field to the imported source object. The FileName field stores the absolute path of the source file from which the Secure Agent reads the data at run time.

For example, a directory contains a number of files and each file contains multiple records that you want to read. You select the directory as source type in the Microsoft Azure Data Lake Storage Gen2 source advanced properties. When you run the mapping, the Secure Agent reads each record and stores the absolute path of the respective source file in the FileName field.

# Parameterization

You can parameterize the connection, objects, and the advanced runtime properties in mappings.

To parameterize the connection, objects, and the advanced runtime properties using a parameter file, create the parameters in the Parameters panel when you create a mapping. Then, define the parameters in the parameter file, place the parameter file in the following location, and run the mapping task:

```
<Informatica Cloud Secure Agent\apps\Data_Integration_Server\data\userparameters>
```

You can also save the parameter file in a cloud-hosted directory in Microsoft Azure Data Lake Storage Gen2.

Consider the following rules and guidelines when you use parameterization:

### General guidelines

- When you create a mapping with a parameterized target that you want to create at runtime, set the target field mapping to automatic.
- You cannot parameterize the field mapping.

### Mappings

- When you use input parameters, specify the parameter name in the following format:
  - Format in a mapping task: \$name\$
  - Format in a parameter file: \$name
- When you use in-out parameters, specify the parameter name in the following format in a mapping task or a parameter file: \$\$name.
- You cannot parameterize a Microsoft Azure Data Lake Storage Gen2 target created at runtime. Instead, you can specify the parameter in the Directory Override to parameterize the target using a parameter file. Specify the parameter in the following format: \$\$name or \$name.

## CHAPTER 4

# Data type reference

Data Integration uses the following data types in Microsoft Azure Data Lake Storage Gen2 mappings and mapping tasks:

- Microsoft Azure Data Lake Storage Gen2 native data types appear in the Source transformation when you choose to edit metadata for the fields.
- Transformation data types. Set of data types that appear in the transformations. These are internal data types based on ANSI SQL-92 generic data types, which the Secure Agent uses to move data across platforms. They appear in all transformations in a mapping.

When the Secure Agent reads source data, it converts the native data types to the comparable transformation data types before transforming the data.

The following table lists the Microsoft Azure Data Lake Storage Gen2 data types that Data Integration supports and the corresponding transformation data types:

Microsoft Azure Data Lake Storage Gen2 Native Data Type	Transformation Data Type	Description
String	String	1 to 104,857,600 characters

## Flat file data types and transformation data types

Flat file data types map to transformation data types that the Secure Agent uses to move data across platforms.

The following table lists the flat file data types that the Secure Agent supports and the corresponding transformation data types:

Flat file data type	Transformation data type for mappings	Range and description
BigInt	Not applicable	-9,223,372,036,854,775,808 to 9,223,372,036,854,775,807 characters; precision 19, scale 0
Nstring*	Text	1 to 104,857,600 characters
Number*	Decimal	Precision from 1 through 28 digits, scale from 0 through 28 digits



Flat file data type	Transformation data type for mappings	Range and description
String*	String	1 to 104,857,600 characters. Precision 256.
*You must select the <b>Schema Source</b> as <b>Import from schema file</b> to read data of Number, String, or Nstring data type.		

## Avro data types and transformation data types

Avro file data types map to transformation data types that the Secure Agent uses to move data across platforms.

The following table lists the Avro file data types that the Secure Agent supports and the corresponding transformation data types:

Avro Data Type	Transformation Data Type	Range and Description
Boolean	Integer	1 or 0 True is equivalent to the integer 1 and False is equivalent to the integer 0.
Bytes	Binary	Precision 4000
Double	Double	Precision 15
Float	Double	Precision 15
Int	Integer	-2,147,483,648 to 2,147,483,647 Precision 10, scale 0
Long	Bigint	-9,223,372,036,854,775,808 to 9,223,372,036,854,775,807 Precision 19, scale 0
Null	Integer	-2,147,483,648 to 2,147,483,647 Precision 10, scale 0
String	String	1 to 104,857,600 characters Precision 4000

## JSON data types and transformation data types

JSON file data types map to transformation data types that the Secure Agent uses to move data across platforms.

The following table lists the JSON file data types that the Secure Agent supports and the corresponding transformation data types:

JSON Data Type	Transformation Data Type	Range and Description
boolean	integer	The default transformation type for boolean is integer. You can specify string data type with values of True and False. True is equivalent to the integer 1 and False is equivalent to the integer 0.
Number (double)	double	-1.79769313486231570E+308 to +1.79769313486231570E+308. Precision 15.
Number (float)	double	-1.79769313486231570E+308 to +1.79769313486231570E+308. Precision 15.
Number (int)	integer	-2,147,483,648 to 2,147,483,647 Precision 10, scale 0
Number (long)	bigint	-9,223,372,036,854,775,808 to 9,223,372,036,854,775,807 Precision 19, scale 0.
string	string	1 to 104,857,600 characters. Precision 4000

## ORC data types and transformation data types

ORC file data types map to transformation data types that the Secure Agent uses to move data across platforms.

The following table lists the ORC file data types that the Secure Agent supports and the corresponding transformation data types:

ORC File Data Type	Transformation Data Type	Range and Description
BigInt	BigInt	-9223372036854775808 to 9,223,372,036,854,775,807
Boolean	Integer	1 or 0 True is equivalent to the integer 1 and False is equivalent to the integer 0.
Char	String	1 to 104,857,600 characters

ORC File Data Type	Transformation Data Type	Range and Description
Date	Date/Time	Jan 1, 1753 A.D. to Dec 31, 4712 A.D. (precision to microsecond)
Double	Double	Precision of 15 digits
Float	Double	Precision of 15 digits
Integer	Integer	-2,147,483,648 to 2,147,483,647
SmallInt	Integer	-32,768 to 32,767
String	String	1 to 104,857,600 characters Precision 4000
Timestamp	Date/Time	1 to 19 characters Precision 19 to 26, scale 0 to 6
TinyInt	Integer	-128 to 127
Varchar	String	1 to 104,857,600 characters

## Parquet data types and transformation data types

Parquet file data types map to transformation data types that the Secure Agent uses to move data across platforms.

The following table lists the Parquet file data types that the Secure Agent supports and the corresponding transformation data types:

Parquet data type	Transformation data type	Range and description
Boolean	Integer	1 or 0 True is equivalent to the integer 1 and False is equivalent to the integer 0.
Byte_Array	Binary	Arbitrarily long byte array
Date	Date/Time	January 1, 0001 to December 31, 9999.
Decimal	Decimal	Precision 1 to 28 digits, scale 0 to 28. You cannot use decimal values with precision greater than 28.

Parquet data type	Transformation data type	Range and description
Double	Double	Precision 15
Float	Double	Precision 15
Int32	Integer	-2,147,483,648 to +2,147,483,647
Int64	Bigint	-9,223,372,036,854,775,808 to +9,223,372,036,854,775,807 8-byte signed integer
Int96	Binary	12-byte signed integer
String	String	1 to 104,857,600 characters Precision 4000
Time	Date/Time	Time of the day. Precision to microsecond.
Timestamp	Date/Time	January 1, 0001 00:00:00 to December 31, 9999 23:59:59.997. Precision to microsecond. You cannot set the precision to nanoseconds.

The Parquet schema that you specify for the Parquet file must be in smaller case. Parquet does not support case-sensitive schema.

### Parquet timestamp data type support

You can use the following Timestamp data types for Parquet file format:

- Timestamp\_micros
- Timestamp\_millis
- Time\_millis
- Time\_micros
- int96

You cannot use the following Timestamp data types for Parquet file format:

- Timestamp\_nanos
- Time\_nanos
- Timestamp\_tz

## CHAPTER 5

# Troubleshooting

Use the following sections to troubleshoot errors in mappings.

## Troubleshooting a mapping

**Time zone for the Date and Timestamp data type fields in Parquet or Avro file formats defaults to the Secure Agent host machine time zone.**

When you run a mapping to read from fields of the Date and Timestamp data types in the Parquet or Avro file formats, the time zone defaults to the Secure Agent host machine time zone.

To change the Date and Timestamp to the UTC time zone, configure the JVMOptions in the Secure Agent.

Perform the following steps to configure the JVM options in the Secure Agent:

1. Select **Administrator > Runtime Environments**.
2. On the **Runtime Environments** page, select the Secure Agent for which you want to configure the JVMOptions.
3. In the upper-right corner, click **Edit**.
4. In the **System Configuration Details** section, select **Data Integration Server** as the service and **DTM** as the type.
5. Edit the **JVMOption** field and set the value to `-Duser.timezone=UTC`.
6. Click **Save**.

### Mapping failed with a Java heap space error

When you read from large data sets in Microsoft Azure Data Lake Storage Gen2, certain mappings might fail with the following error:

```
[ERROR] java.lang.OutOfMemoryError: Java heap space
```

You must increase the heap size to run the mappings successfully. The recommended heap size is 1 GB.

Perform the following steps to configure the JVM options in the Secure Agent to increase the memory for the Java heap size:

1. Select **Administrator > Runtime Environments**.
2. On the **Runtime Environments** page, select the Secure Agent for which you want to increase memory from the list of available Secure Agents.
3. In the upper-right corner, click **Edit**.

4. In the **System Configuration Details** section, select **Data Integration Server** as the service and **DTM** as the type.
5. Edit the **JVMOption** field and set the value to **-Xmx1024m**.  
**Note:** The recommended heap size is 1 GB. You can increase the heap size based on the data you want to process.
6. Click **Save**.

#### **Mapping fails if the directory name or subdirectory name contains Unicode characters**

When you read from a flat file in Microsoft Azure Data Lake Storage Gen2 and if the directory name or subdirectory name contains Unicode characters, the mapping fails.

To resolve this issue, set the environment variable `LC_ALL="en_US.UTF-8"` in the Secure Agent, and restart the Secure Agent.

#### **Non-English characters in the source are incorrectly written to the target when you use the Append write strategy**

When you append data to a flat file in a Microsoft Azure Data Lake Storage Gen2 target, the non-English characters in the source are incorrectly written to the target.

To write the non-English characters correctly to the target, configure the JVMOptions in the Secure Agent.

Perform the following steps to configure the JVM options in the Secure Agent:

1. Select **Administrator > Runtime Environments**.
2. On the **Runtime Environments** page, select the Secure Agent for which you want to configure the JVMOptions.
3. In the upper-right corner, click **Edit**.
4. In the **System Configuration Details** section, select **Data Integration Server** as the service and **DTM** as the type.
5. Edit the **JVMOption** field and set the value to `-Dfile.encoding=UTF-8`.
6. Click **Save**.

#### **Mapping fails due to failure to access the /tmp directory**

When you run a Microsoft Azure Data Lake Storage Gen2 mapping, the Secure Agent stages the files in a temporary staging folder. By default, the folder for staging data is `/tmp`.

Ensure that you have the read and write permissions to the `/tmp` folder.

Perform the following steps to change the temporary staging folder:

1. Select **Administrator > Runtime Environments**.
2. On the **Runtime Environments** page, select the Secure Agent for which you want to configure the JRE\_OPTS field.
3. In the upper-right corner, click **Edit**.
4. In the **System Configuration Details** section, select **Data Integration Server** as the service and **DTM** as the type.
5. Edit the **JRE\_OPTS** field and set the value to `-Djava.io.tmpdir=<DIR>`.
6. Click **Save**.

# INDEX

## A

administration [8](#)  
Azure Data Lake Storage Gen2  
  connection properties [11](#)

## C

Cloud Application Integration community  
  URL [5](#)  
Cloud Developer community  
  URL [5](#)  
connections  
  Azure Data Lake Storage Gen2 [11](#)  
  Microsoft Azure Data Lake Storage Gen2 [11](#)

## D

Data Integration community  
  URL [5](#)  
data type reference  
  overview [24](#)  
data types  
  avro [25](#)  
  parquet [27](#)  
directory source  
  Microsoft Azure Blob Storage sources [19](#)

## I

Informatica Global Customer Support  
  contact information [6](#)  
Informatica Intelligent Cloud Services  
  web site [5](#)

## J

JSON file data types  
  transformation data types [26](#)

## L

Linux  
  configuring proxy settings [14](#)

## M

maintenance outages [6](#)

mappings  
  Microsoft Azure Data Lake Storage Gen2 Source properties [16](#)  
Microsoft Azure Data Lake Storage Gen2  
  Source transformation [16](#)  
  Sources in mappings [16](#)  
Microsoft Azure Data Lake Storage Gen2 Connection  
  overview [11](#)  
Microsoft Azure Data Lake Storage Gen2 Connector  
  introduction [7](#)

## O

ORC file data types  
  transformation data types [26](#)

## P

proxy settings  
  configuring on Linux [14](#)  
  configuring on Windows [13](#)  
  JVMOptions [13](#)

## S

Source transformation  
  Microsoft Azure Data Lake Storage Gen2 properties [16](#)  
Sources  
  Microsoft Azure Data Lake Storage Gen2 in mappings [16](#)  
status  
  Informatica Intelligent Cloud Services [6](#)  
system status [6](#)

## T

trust site  
  description [6](#)

## U

upgrade notifications [6](#)

## W

web site [5](#)  
wildcard character  
  overview [19](#)  
Windows  
  configuring proxy settings [13](#)