IBM InfoSphere Information Server
Version 11 Release 3

Connectivity Guide for Accessing Amazon S3

SC19-4249-00

IBM InfoSphere Information Server
Version 11 Release 3

Connectivity Guide for Accessing Amazon S3

SC19-4249-00

Note

Before using this information and the product that it supports, read the information in "Notices and trademarks" on page 29.

Contents

Chapter 1. Amazon S3 connector

Use the Amazon S3 connector to connect to Amazon Simple Storage Service (S3) and perform various read and write functions.

Designing jobs that use the Amazon S3 connector

You can use the Amazon S3 connector to develop jobs that read data from and write data to Amazon S3.

Procedure

1. Import metadata from Amazon S3.
2. Define a job that includes the Amazon S3 Connector stage.
3. Define a connection to Amazon S3.
4. To set up the Amazon S3 Connector stage to read data from Amazon S3, complete the following steps:
   a. Set up column definitions.
   b. Configure the Amazon S3 connector as a source of data.
   c. Optional: Create a reject link to manage rejected data.
   d. Optional: Read partitioned data.
5. To set up the Amazon S3 Connector stage to write data to Amazon S3, complete the following steps:
   a. Set up column definitions.
   b. Configure the Amazon S3 connector as a target.
   c. Optional: Write partitioned data.
6. Compile and run the job.

Importing Amazon S3 metadata

Before you use the Amazon S3 connector to read or write data, you can use InfoSphere® Metadata Asset Manager to import metadata about files and folders in Amazon S3. You can then use this metadata to create table definitions in InfoSphere DataStage®.

Before you begin

- Get an access key and secret key for your Amazon Web Services account.
- If you do not have metadata about files and folders in Amazon S3, specify column metadata and metadata about how a file is formatted by using one of the "Metadata formatting options" on page 9.

About this task

InfoSphere Metadata Asset Manager imports metadata that is specified in one of the following ways:

- As the first row of the file.
- In a `.osh` schema file that is in the same folder and is named *file*.osh or *folder*.osh, where *file* is the name of a file in the folder and *folder* is the name of the folder. For example, if `fileA.txt` is in the `sample` directory, metadata can be specified in the `fileA.txt.osh` or `sample.osh` files.

Procedure

Use InfoSphere Metadata Asset Manager to import the metadata.
For more information, see Importing metadata by using InfoSphere Metadata Asset Manager in IBM® Knowledge Center (http://www.ibm.com/support/knowledgecenter/SSZJPZ_11.3.0/com.ibm.swg.im.iis.mmi.doc/topics/t_importing_metadata_into_staging_area.html).

What to do next

Use the metadata that you imported to create table definitions in InfoSphere DataStage.

Defining a job that includes the Amazon S3 connector

To read data from or write data to Amazon S3, you can create a job that includes the Amazon S3 connector. Then, you add any additional stages that are required and create the necessary links.

Procedure

1. In the InfoSphere DataStage and QualityStage® Designer client, select **File** > **New** from the menu.
2. In the **New** window, select the **Parallel Job** icon, and then click **OK**.
3. Add the Amazon S3 connector to the job:
   a. In the palette, select the **File** category.
   b. Drag the Amazon S3 Connector stage to the canvas.
   c. Optional: Rename the Amazon S3 Connector stage. Choose a name that indicates the role of the stage in the job.
4. Create the necessary links and add additional stages for the job:
   - For a job that reads Amazon S3 data, create the next stage in the job, and then create an output link from the Amazon S3 connector to the next stage.
   - For a job that writes data to Amazon S3, create an input link from the previous stage in the job to the Amazon S3 connector.
5. Save the job.

Defining a connection to Amazon S3

To access data from Amazon S3, you must define a connection that specifies the access key and secret key.

Before you begin

Get an access key and secret key for your Amazon Web Services account.

Procedure

Specify the access key and secret key to use when the Amazon S3 connector connects to Amazon S3:

Table 1. Methods for specifying the access key and secret key

Option	Procedure	Best practices for security
Specify the values in the stage.	1. In the stage editor, specify values for the **Access key** and **Secret key** properties. 2. From the **Use credentials file** list, select **No**.	Use encrypted job parameters for the values that you specify in the stage editor.

Specify the values in a credentials file.	1. Create a file named `AwsCredentials.properties`.
	2. In the `AwsCredentials.properties` file, specify the `accessKey` and `secretKey` properties and values for the properties. For example, the file might contain the following lines:
	`accessKey=SampleName`
	`secretKey=SamplePassword`
	3. From the **Use credentials file** list, select **Yes**.
	4. In the **Credentials file** field, specify the fully qualified path for the `AwsCredentials.properties` file.
		Set the security permissions on the `AwsCredentials.properties` so that only the user who runs the job has read access to it.

Reading data from Amazon S3

You can configure the Amazon S3 connector to connect to Amazon S3 and read data from it.

Before you begin

- Define a job that contains the Amazon S3 Connector stage.
- Define a connection to Amazon S3.

About this task

The following figure shows an example of using the Amazon S3 connector to read data. In this example, the Amazon S3 connector reads data from Amazon S3 and then sends the data to a DB2 Connector stage. This job includes an optional reject link, on which the connector sends reject records to a Sequential File stage.

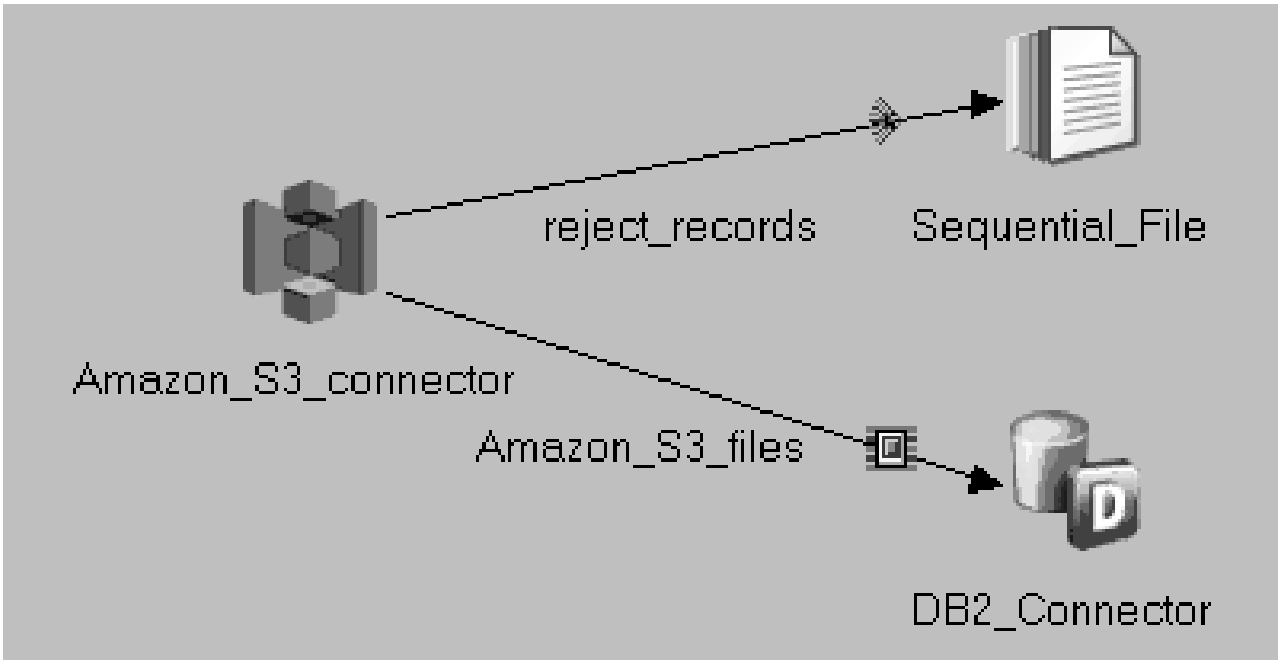

Figure 1. Example of reading data from Amazon S3

Setting up column definitions on a link

Column definitions, which you set on a link, specify the format of the data records that the Amazon S3 connector reads from or writes to Amazon S3.

About this task

Which column definitions you set depend on the read mode and the format of the Amazon S3 files. The following table shows how the read mode affects the requirements for column definitions.

Table 2. Requirements for column definitions based on read mode

Read mode	Requirement for column definitions
List buckets	Set up only one column definition for the output. Use a string data type such as VarChar.
List files	Set up one column definition for the file name and, optionally, one or more column definitions for metadata about the file. For information about the names and data types that are required for the columns that contain the file metadata, see "Column definitions for listing files" on page 17.

If the Amazon S3 files are in delimited file format, set up a column definition for each field in the file. Ensure that each column definition matches its respective field from the file in the following ways:

- Data type
- Attributes for data types, such as scale and precision
- File syntax, such as the file delimiter

Procedure

1. From the job design canvas, double-click the Amazon S3 connector icon.
2. Use one of the following methods to set up the column definitions:
   - Drag a table definition from the repository view to the link on the job canvas. Then, use the arrow buttons to move the columns between the **Available columns** and **Selected columns** lists.
   - On the **Columns** page, click **Load** and select a table definition from the metadata repository. Then, to choose which columns from the table definition apply to the link, move the columns from the **Available columns** list to the **Selected columns** list.
3. Configure the properties for the columns:
   a. Right-click within the columns grid, and select **Properties** from the menu.
   b. Select the properties to display, specify the order in which to display them, and then click **OK**.
4. Optional: Modify the column definitions. You can change the column names, data types, and other attributes. In addition, you can add, insert, or remove columns.
5. Optional: Save the new table definition in the metadata repository:
   a. On the **Columns** page, click **Save**, and then click **OK** to display the repository view.
   b. Navigate to an existing folder, or create a new folder in which to save the table definition.
   c. Select the folder, and then click **Save**.

Configuring the Amazon S3 connector as a source

To configure the connector to read Amazon S3 data or list Amazon S3 buckets and files, you must specify a read mode and configure properties for the read mode that you specified.

Procedure

1. From the job design canvas, double-click the Amazon S3 Connector stage.
2. Set the **Read mode** property to **Read single file**, **Read multiple files**, **List buckets**, or **List files**.
3. Configure the read process for the read mode that you specified.

Table 3. Reading data from Amazon S3

Read mode	Procedure
Read single file	1. Specify the name of the bucket that contains the files. 2. Specify the name of the file to read.
Read multiple files	1. Specify the name of the bucket that contains the files. 2. In the **File name** field, specify a prefix that the files that you want to read must have in their file path. For example, if you enter `transactions` as the prefix, the connector reads all of the files in the `transactions` folder, such as `transactions/january/day1.txt`, and a file named `transactions.txt`.
List buckets	No additional configuration is required.

Table 3. Reading data from Amazon S3 (continued)

Read mode	Procedure
List files	1. Specify the name of the bucket that contains the files. 2. **Optional:** In the **File name** field, specify a prefix that the files that you want to read must have in their file path. For example, if you enter `transactions` as the prefix, the connector reads all of the files in the `transactions` folder, such as `transactions/january/day1.txt`, and a file named `transactions.txt`. If you do not specify a file name prefix, all of the files in the bucket are listed.

4. Click **OK**, and then save the job.

Rejecting records that contain errors

When the Amazon S3 connector includes a reject link, you can configure the connector to send records that cannot be parsed to the target stage on the reject link. If you configure the connector to send records to the reject link, the job completes even if data is rejected.

Procedure

1. On the job design canvas, add and configure a target stage to receive the rejected records.
2. Right-click the Amazon S3 connector and drag to create a link from the Amazon S3 connector to the target stage
3. Double-click the connector to open the stage editor.
4. Set the **Reject mode** property to **Reject**.
5. Set up the reject link:
   a. On the Output page, select the link to the target stage for rejected records from the **Output name** list.
   b. On the Properties page, set the **Is reject link** property to **Yes**.
   c. If runtime column propagation is not enabled, set up column definitions for the link. Include one VarBinary column and, optionally, a column of a string data type to contain error messages for the rejected records.

Reading partitioned data

In a job that uses multiple nodes, each node that is specified for the stage reads a distinct subset of data from the source.

Before you begin

- Configure the stage on the output link to run in parallel.
- Define two or more processing nodes to run the job on.

About this task

When you configure the Amazon S3 connector to read in parallel, each node can read part of the same file or each node can read one or more different files.

By default, the Amazon S3 file that you specify in the **File name** property is range partitioned. Each node reads approximately the same number of rows from the file. For example, if a file has 1000 rows and you define four processing nodes for the job, each node reads approximately 250 rows from the file.

If you want each node to read one or more different files, the file names in Amazon S3 must contain a unique number that corresponds to a node number. Then, in the Amazon S3 Connector stage, you can specify `[[node-number]]` as part of the value for the **File name** property. For example, if you define two nodes for the job and specify `MyFile_[[node-number]].txt` for the **File name** property, node 0 reads `MyFile_0.txt` and node 1 reads `MyFile_1.txt`.

Procedure

1. On the job design canvas, double-click the Amazon S3 Connector stage, and then click the **Stage** tab.
2. On the Advanced page, set **Execution mode** to **Parallel**.
3. Specify a value for the **File name** property based on the number and names of the files to read.
4. Click **OK**, and then save the job.

Writing and deleting data in Amazon S3

You can configure the S3 connector to connect to Amazon S3 and write data to it or delete data from it.

Before you begin

- Define a job that contains the Amazon S3 Connector stage.
- Define a connection to Amazon S3.

About this task

The following figure shows an example of using the Amazon S3 connector to write data. In this example, the Amazon S3 connector reads data from a Sequential File stage and then writes the data to Amazon S3.

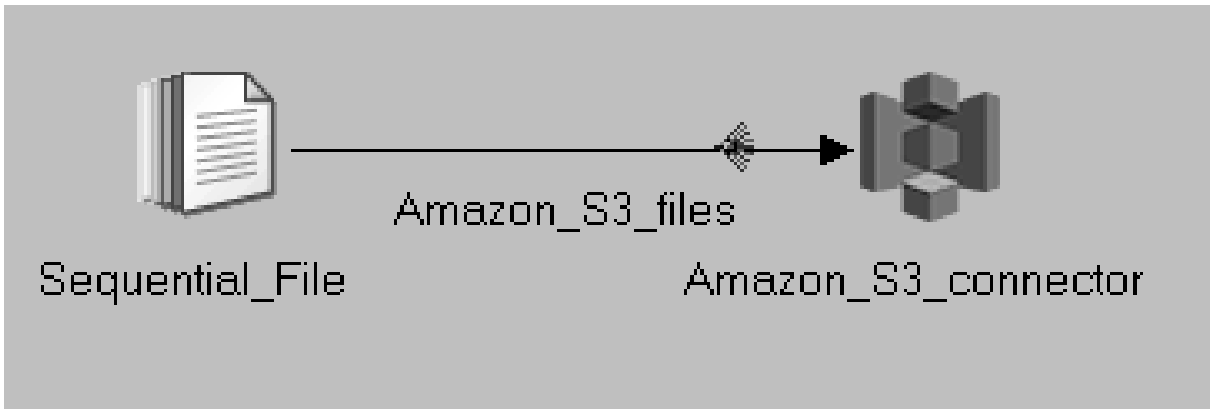

Figure 2. Example of writing data to Amazon S3

Setting up column definitions on a link

Column definitions, which you set on a link, specify the format of the data records that the Amazon S3 connector writes to Amazon S3.

About this task

If you plan to use the Amazon S3 connector to delete files from Amazon S3, set up one column definition of a string data type for the file name. The connector deletes the file that is named in each row of the column. If no file in the specified Amazon S3 bucket matches a file name in that column, no error is generated.

Optionally, you can also set up a column named Bucket to specify the bucket to delete files from.

Procedure

1. From the job design canvas, double-click the Amazon S3 connector icon.
2. Use one of the following methods to set up the column definitions:
   - Drag a table definition from the repository view to the link on the job canvas. Then, use the arrow buttons to move the columns between the **Available columns** and **Selected columns** lists.
   - On the **Columns** page, click **Load** and select a table definition from the metadata repository. Then, to choose which columns from the table definition apply to the link, move the columns from the **Available columns** list to the **Selected columns** list.
3. Configure the properties for the columns:
   a. Right-click within the columns grid, and select **Properties** from the menu.
   b. Select the properties to display, specify the order in which to display them, and then click **OK**.
4. Optional: Modify the column definitions. You can change the column names, data types, and other attributes. In addition, you can add, insert, or remove columns.
5. Optional: Save the new table definition in the metadata repository:
   a. On the **Columns** page, click **Save**, and then click **OK** to display the repository view.
   b. Navigate to an existing folder, or create a new folder in which to save the table definition.
   c. Select the folder, and then click **Save**.

Configuring the Amazon S3 connector as a target

You can configure the connector to write rows to or delete rows from Amazon S3.

Procedure

1. On the job design canvas, double-click the Amazon S3 Connector stage.
2. Select the input link to edit.
3. Configure the connector to write rows to or delete rows from Amazon S3.

Table 4. Procedures for writing or deleting data from Amazon S3

Action	Procedure
Write	1. Set the **Write mode** property to **Write**. 2. Specify the bucket to write the file to. 3. Specify the name of the file to write.

Table 4. Procedures for writing or deleting data from Amazon S3 (continued)

Action	Procedure
Delete	1. Set the **Write mode** property to **Delete**. 2. If you did not set up a column definition for a column named Bucket, specify the bucket to delete files from in the **Bucket** property.

4. Click **OK**, and then save the job.

Writing partitioned data

In a job that uses multiple nodes, records that arrive on the input link of the connector are distributed across multiple nodes. Then, the records are written in parallel from all of the nodes to Amazon S3.

Before you begin

- Configure the stage on the input link to run in parallel.
- Define two or more processing nodes to run the job on.

About this task

When you configure the Amazon S3 connector to write in parallel, each node writes a file to Amazon S3. By default, the name of the file is the name that is specified in the **File name** property with a node number appended to it.

Procedure

1. On the job design canvas, double-click the Amazon S3 Connector stage, and then click the **Stage** tab.
2. On the Advanced page, set **Execution mode** to **Parallel**.
3. Specify a value for the **File name** property based on the number and names of the files to write.
4. Click **OK**, and then save the job.

Reference

To use the Amazon S3 connector successfully, you might need detailed information, such as information about data type mappings, stage properties, and runtime column propagation.

Metadata formatting options

You can specify column metadata and metadata about how a file is formatted in a delimited string or in a `.osh` schema file. The metadata that you specify can be used for runtime column propagation and imported into the metadata repository.

Format for column metadata in delimited strings

When you specify column metadata in a delimited string, you can use a predefined set of data types and attributes in the column definitions.

If column metadata is specified as a delimited string, it must be specified in the format *column_name*:*data_type*. Metadata for each column is separated by the character that is specified for the **Field delimiter** property, and the column values can be enclosed in single or double quotation marks if a value is specified for the **Quotation mark** property.

For example, if the comma is specified as the field delimiter, you might specify the following columns:

```
C_BigInt:BigInt,C_Bit:SmallInt,C_Char:Char(10)
```

The following table shows the data types and attributes that you can specify for column metadata in a delimited string. You can specify the nullability attribute for all data types, for example, `MyColumn:VarChar(10) not nullable`. By default, columns are not nullable.

Table 5. Data types and attributes for column metadata in a delimited string.

Data type	Other attributes	Example specification
Binary data types: • Binary • LongVarBinary • VarBinary	length	VarBinary(10)
Bit	length	Bit(8)
Integer data types: • BigInt • Integer • SmallInt • TinyInt	signed or unsigned	SmallInt(unsigned)
Date		Date not nullable
Decimal	• precision • scale If a value is not specified for the scale attribute, the default scale is 0.	Decimal(10,2)
Double		Double nullable
Float		Float not nullable
Numeric	• precision • scale If a value is not specified for the scale attribute, the default scale is 0.	Numeric(5,1)
Real		Real nullable
String data types: • Char • LongNVarChar • LongVarChar • NChar • NVarChar • VarChar	length	NChar(100)
Time data types: • Time • Timestamp	microsecond If a value is not specified for the microsecond attribute, microsecond precision is not enabled.	Timestamp(microseconds)

Format for column metadata in `.osh` schema files

When you specify column metadata in a `.osh` schema file, you can use a predefined set of data types and options in the column definitions.

The following table shows the data types and attributes that you can specify for column metadata in a `.osh` schema file. Any OSH data types and attributes that are not in the table are not supported for the Amazon S3 connector. You can specify the nullability attribute for all of the OSH data types, for example, `SALES_DATE: not nullable date`. By default, columns are not nullable.

Table 6. Data types and attributes for column metadata.

OSH data type	InfoSphere DataStage data type	Other attributes
• int8 • uint8	TinyInt	
• int16 • uint16	SmallInt	
• int32 • uint32	Integer	
• int64 • uint64	BigInt	
sfloat	Float	
dfloat	Double	
string	• Char • VarChar	**No attribute** Unbounded length **[***length***]** Fixed length **max=[***length***]** Variable length
ustring	• Nchar • NVarChar	**No attribute** Unbounded length **[***length***]** Fixed length **max=[***length***]** Variable length
decimal	Decimal	• [*precision*] • [*precision*], [*scale*]
date	Date	
time	Time	[microseconds]
timestamp	TimeStamp	[microseconds]
raw	• Binary • VarBinary	**No attribute** Unbounded length **[***length***]** Fixed length **max=[***length***]** Variable length

OSH file formatting properties

When you use a `.osh` schema file to specify file formatting properties for the Amazon S3 connector, you can use a subset of the properties that are available in the OSH schema. In the `.osh` schema file, you can also specify properties for the Amazon S3 connector that are not part of the OSH schema in a comment line.

Properties and options from the OSH schema

The following table shows the file formatting properties and options from the OSH schema that the Amazon S3 connector supports.

Table 7. Properties from the OSH schema that you can use for the Amazon S3 connector.

Property	Supported values for the property
record_delim	• '*one-character_delimiter*' • null
record_delim_string	"*delimiter_string*"
final_delim	• end • '*one-character_delimiter*' • null
final_delim_string	"*delimiter_string*"
delim	• '*one-character_delimiter*' • null
null_field	• '*one-character_delimiter*' • "*delimiter_string*"
quote	• single • double • '*one-character_delimiter*'
charset	*character_set*
date_format	*format*
time_format	*format*
timestamp_format	*format*

Other properties and options that you can specify

The following table shows file formatting properties for the Amazon S3 connector that you can specify in a `.osh` schema file that are not part of the OSH schema.

Table 8. Properties that you can use for the Amazon S3 connector that are not part of the OSH schema.

Property	Supported values for the property	Description
file_format	• csv • delimited • redshift	Specify whether the file is a comma-separated value file, in delimited file format, or in a format that enables the file to be written to Amazon Redshift.

Table 8. Properties that you can use for the Amazon S3 connector that are not part of the OSH schema (continued).

Property	Supported values for the property	Description
header	• true • false	Specify whether the first row of the file contains field headers and is not part of the data.
escape	*escape_character*	Specify the character to use to escape field and row delimiters.

If you want to specify any of these properties, you must add them to a comment line that begins with *FileStructure:* and enclose the value for each property in single quotation marks. For example, you can specify the following line in the `.osh` schema file:

```
//  FileStructure:  file_format='delimited', header='true', escape ='\\'
```

Formatting options for Decimal, Time, Date, and Timestamp data types

When you read data from or write data to Amazon S3 that is of the Decimal data type, you can use the OSH schema format or the DecimalFormat Java class format. When the data is of the Time, Date, or Timestamp data types, you can use the OSH schema format or the SimpleDateFormat Java class format.

You can specify strings that define the format for fields of these data types in the `.osh` schema file or in the stage properties.

Time, Date, and Timestamp formatting options

If you use the SimpleDateFormat Java class format, add the letter *J* as the prefix for the format string. For example, the following strings define the same format:

String for the SimpleDateFormat Java class format
Jyyyy-MM-dd HH:mm-ss

String for the OSH schema format
%yyyy-%mm-%dd %hh-%nn-%ss

The following table shows the format strings from the OSH schema that are not supported.

Table 9. Format strings that are not supported

Data type	Format string	Description
Date	%NNNNyy	Cutoff year followed by year of the century
Date	%e	Integer in the range 1 - 7 representing the day of the week, beginning with Sunday
Date	%E	Integer in the range 1 - 7 representing the day of the week, beginning with Monday

Table 9. Format strings that are not supported (continued)

Data type	Format string	Description
Time	%SSSSSS	Microseconds
Time and Date	+N, -N	Option to left or right justify day or month names

Runtime column propagation

Use runtime column propagation to have the connector automatically add missing columns to the link schema when the job runs.

Usage

Before you can enable runtime column propagation in a stage, runtime column propagation must be enabled for parallel jobs at the project level from the InfoSphere DataStage Administrator client. To enable runtime column propagation for the output link of the stage, select the **Runtime column propagation** check box on the Columns page.

When runtime column propagation is enabled, the connector uses column metadata that is specified in one of the following ways:

- As the first row of the Amazon S3 file that is read
- As a delimited string that is specified as the value for the **Metadata source** property
- As a delimited string in the file that is specified as the value for the **Metadata source** property
- As a `.osh` schema file that is specified as the value for the **Metadata source** property

Properties for the Amazon S3 connector

Use these options to modify how the connector reads and writes data.

Field delimiter and row delimiter properties

Use the **Field delimiter** and **Row delimiter** properties to specify the characters that separate fields and rows in Amazon S3 files.

Usage

When you specify delimiters, you can use the strings that are shown in the following table in addition to individual characters.

Table 10. Strings for the delimiter properties

Delimiter value	String to specify
Carriage return	`<CR>`
Line feed	`<LF>`
New line	`<NL>`
Tab	`<TAB>`

Table 10. Strings for the delimiter properties (continued)

Delimiter value	String to specify
Unicode escape string	\u*NNNN* where *NNNN* is a set of four hexadecimal digits that represent the Unicode character code For example, to represent the uppercase letter *A*, specify `\u0041`.

Lifecycle rules

Lifecycle rules specify when an Amazon S3 file is set to expire or be archived.

Types of lifecycle rules

In the Amazon S3 Connector stage, you can specify the following types of lifecycle rules:

Expiration
When a file expires, it is deleted from Amazon Web Services.

Archival
When a file is archived, it is sent to the Amazon Glacier archive.

For both types of lifecycle rules, you can specify the date when the lifecycle rule is set to apply or the number of days that the file will exist in its current state before the lifecycle rule is set to apply. If you define lifecycle rules for archival and expiration for a file or folder, both lifecycle rules must use the same time period format.

Rule scope

A lifecycle rule can apply to only one file or to all of the files in the folder that contains the file. You cannot specify a lifecycle rule that applies to a file and a lifecycle rule that applies to the folder that contains the file.

When lifecycle rules are created in Amazon S3, the scope of the role is represented by a prefix. This prefix includes only the part of the file path for a file that the lifecycle rule applies to.

For example, suppose that you want to create a lifecycle rule for the `directory4/file1.txt` file. If you create a lifecycle rule and set the **Rule scope** property to **File**, the prefix is `directory4/file1.txt`. If you create a lifecycle rule and set the **Rule scope** property to **Folder**, the prefix is `directory4/`.

If the connector runs in parallel, the lifecycle rule must apply to all of the files in the folder. As a result, the **Rule scope** property must be set to **Folder**.

Examples

Suppose that on 1 January 2016, you want to create a lifecycle rule for all of the files in the `DepartmentA/Employees` folder. You want the files to be archived in three years and deleted in five years. Because you can specify lifecycle rules based on a date or a number of days but you must use the same format for both rules, you can configure the lifecycle rules in the following ways:

Table 11. Example of options for specifying lifecycle rules

Amazon S3 property	Value for specifying lifecycle rules based on a date	Value for specifying lifecycle rules based on a number of days
Rule scope	Folder	Folder
Time period format	Specific date	Days from creation date
Expiration	Yes	Yes
Expiration date	`"2021-01-01"`	Does not apply
Duration	Does not apply	`1825`
Archive	Yes	Yes
Date to archive	`"2019-01-01"`	Does not apply
Duration	Does not apply	`1095`

File encoding

You can specify the encoding of files that are read from or written to Amazon S3.

Methods for specifying the file encoding

You can specify the file encoding in the following ways, which are listed in their order of precedence:

1. As a value for the **Encoding** property in the stage editor.
2. As a value for the charset attribute in a `.osh` schema file. You can use this method only if runtime column propagation is enabled and the connector uses metadata from a `.osh` schema file.
3. As a value for the **`APT_IMPEXP_CHARSET`** environment variable.

The character set that you specify for the file encoding must be supported by the Java Virtual Machine (JVM).

Byte order marks

The Amazon S3 connector can match byte order marks (BOM) in files to the file encoding that you specify. If the BOM in the file specifies a different endian format than the file encoding, or the file encoding does not include an endian format, the endian format from the BOM is used. For example, if the encoding is specified as UTF-16 and the BOM indicates that the file is UTF-16 big endian, the encoding is changed to UTF-16BE.

When the Amazon S3 connector reads a file in parallel across multiple nodes, only the first node reads the BOM and adjusts the encoding automatically. As a result, if you configure the Amazon S3 Connector stage to read a file in parallel, ensure that you include the endian format in the encoding that you specify. For example, specify `UTF-16BE` instead of `UTF-16`.

When you use the Amazon S3 connector to write data, you can configure the connector to include byte order marks in files by setting **Include byte order mark** to **Yes**.

File name options for reading and writing partitioned data

When the Amazon S3 connector reads or writes partitioned data, you can configure the information that is included in the name of the file that is read from or written to Amazon S3.

The following table shows how you can configure the information that is included in the file name.

Table 12. Options for the names of the files that are read from or written to Amazon S3.

Purpose	Procedure	Example value for the File name property	Example file name in Amazon S3
Include the node number in the file name (default).	Specify the name of the file in the **File name** property.	`MyFile`	`MyFile.1`
Include the node number in a nondefault part of the file name.	Include the string `[[node-number]]` in the file name that you specify.	`MyFile_[[node-number]].txt`	`MyFile_1.txt`
Specify how the nodes are numbered. The first node can be node 0, 1, or 0 or 1 with leading zeroes. By default, the first node is node 0.	Include the string `[[node-number(`*`numbering_value`*`)]]` in the file name that you specify.	`MyFile_[[node-number(0001)]].txt`	`MyFile_0001.txt` In write mode, this file is the file that is written by the first node.
Include the total number of nodes in the file name.	Include the string `[[node-count]]` in the file name that you specify.	`MyFile.[[node-number(01)]].of.[[node-count]]`	`MyFile_02.of.05` In write mode, this file is the file that is written by the second node out of five total nodes.

Column definitions for listing files

When you use the Amazon S3 connector to list files that are stored on Amazon S3, you can set up column definitions for the metadata about the files.

The following table shows the names and data types that are required for the columns that contain the file metadata.

Table 13. Column requirements for file metadata.

Column name	Data type	Description
file_name	String	The name of the file. Do not use any of the column names that are designated for the other metadata columns.
Bucket	String	The name of the bucket that contains the file.
ContentType	String	The content type, for example, `text/plain`.

Table 13. Column requirements for file metadata (continued).

Column name	Data type	Description
Metadata	String	One or more name-value pairs for the metadata that was specified in Amazon S3 for the file. Each name-value pair is separated with a semicolon, for example, `Topic=News;Subtopic=Sports`.
ExpirationDate	Date	The date to set the file to be deleted from Amazon S3.
ExpirationDays	Integer	The number of days that the file will exist in Amazon S3 before it is set to expire.
ArchivalDate	Date	The date to set the file to be archived in Amazon Glacier.
ArchivalDays	Integer	The number of days that the file will exist in Amazon S3 before it is set to be archived in Amazon Glacier.
ExpirationEnabled	Integer	The status of the rules that specify when to set the file to expire or be archived. The column can have the following values: **0** Rules are not enabled. **1** Rules are enabled.
Encryption	String	The type of server-side encryption, for example, `AES-256`.
Version	String	The version number of the file.

If you include any of the following column definitions on the output link, the job that includes the Amazon S3 connector might take longer to complete:

- ContentType
- Metadata
- Encryption
- Version

Chapter 2. Environment variables: Amazon S3 connector

The Amazon S3 connector uses these environment variables.

APT_IMPEXP_CHARSET

Set this environment variable to control the character encoding of the files that are read from or written to Amazon S3.

If runtime column propagation is enabled and the connector uses metadata from a `.osh` schema file, the character set that is specified for the charset attribute is used instead of the value for this environment variable. If a value is specified for the **Encoding** property in the stage editor for the connector, that value is used instead of the value for the environment variable or the value in the `.osh` schema file.

CC_CHECK_INVALID_CHARS

Set this environment variable to 1 to validate the characters in a file based on the file encoding that is specified or the default file encoding. This environment variable applies only when a stage is configured to read data.

If the **CC_CHECK_INVALID_CHARS** environment variable is enabled, characters are validated when the job runs. Rows that contain invalid characters are marked as invalid, and the stage processes the invalid rows based on the setting for the **Reject mode** property. Because validating characters requires additional processing, enable this environment variable only if validation is required.

If this environment variable is not enabled, the stage replaces invalid characters with the replacement character that represents an invalid character for the specified character set.

CC_S3_LOG_FILE

Set this environment variable to the full file path for a log file that you want to contain messages from the Amazon Web Services logging system. If an invalid file path is specified, no messages are logged.

CC_S3_LOG_LEVEL

Set this environment variable to specify the level of logging for messages from the Amazon Web Services logging system. Any messages that are logged are written to the file that is specified for the **CC_S3_LOG_FILE** environment variable.

You can set the **CC_S3_LOG_LEVEL** environment variable to one of the following values. The values are listed by the amount of messages that are logged, from least to most. Each level includes all of the messages that are logged by the previous level.

OFF No messages are logged.

FATAL
Errors that cause a process to fail are logged.

ERROR
Errors that do not cause a process to fail are logged.

WARN
Messages about conditions that might cause errors or other issues are logged.

INFO General informational messages are logged.

DEBUG
Specific informational messages that might be useful for troubleshooting are logged.

TRACE
Very specific informational messages that might be useful for troubleshooting are logged.

ALL All messages are logged.

Appendix A. Product accessibility

You can get information about the accessibility status of IBM products.

The IBM InfoSphere Information Server product modules and user interfaces are not fully accessible.

For information about the accessibility status of IBM products, see the IBM product accessibility information at http://www.ibm.com/able/product_accessibility/index.html.

Accessible documentation

Accessible documentation for InfoSphere Information Server products is provided in an information center. The information center presents the documentation in XHTML 1.0 format, which is viewable in most web browsers. Because the information center uses XHTML, you can set display preferences in your browser. This also allows you to use screen readers and other assistive technologies to access the documentation.

The documentation that is in the information center is also provided in PDF files, which are not fully accessible.

IBM and accessibility

See the IBM Human Ability and Accessibility Center for more information about the commitment that IBM has to accessibility.

Appendix B. Contacting IBM

You can contact IBM for customer support, software services, product information, and general information. You also can provide feedback to IBM about products and documentation.

The following table lists resources for customer support, software services, training, and product and solutions information.

Table 14. IBM resources

Resource	Description and location
IBM Support Portal	You can customize support information by choosing the products and the topics that interest you at www.ibm.com/support/entry/portal/Software/Information_Management/InfoSphere_Information_Server
Software services	You can find information about software, IT, and business consulting services, on the solutions site at www.ibm.com/businesssolutions/
My IBM	You can manage links to IBM Web sites and information that meet your specific technical support needs by creating an account on the My IBM site at www.ibm.com/account/
Training and certification	You can learn about technical training and education services designed for individuals, companies, and public organizations to acquire, maintain, and optimize their IT skills at http://www.ibm.com/training
IBM representatives	You can contact an IBM representative to learn about solutions at www.ibm.com/connect/ibm/us/en/

Appendix C. Accessing the product documentation

Documentation is provided in a variety of formats: in the online IBM Knowledge Center, in an optional locally installed information center, and as PDF books. You can access the online or locally installed help directly from the product client interfaces.

IBM Knowledge Center is the best place to find the most up-to-date information for InfoSphere Information Server. IBM Knowledge Center contains help for most of the product interfaces, as well as complete documentation for all the product modules in the suite. You can open IBM Knowledge Center from the installed product or from a web browser.

Accessing IBM Knowledge Center

There are various ways to access the online documentation:

- Click the **Help** link in the upper right of the client interface.
- Press the F1 key. The F1 key typically opens the topic that describes the current context of the client interface.

  Note: The F1 key does not work in web clients.
- Type the address in a web browser, for example, when you are not logged in to the product.

  Enter the following address to access all versions of InfoSphere Information Server documentation:

  `http://www.ibm.com/support/knowledgecenter/SSZJPZ/`

  If you want to access a particular topic, specify the version number with the product identifier, the documentation plug-in name, and the topic path in the URL. For example, the URL for the 11.3 version of this topic is as follows. (The ⇒ symbol indicates a line continuation):

  ```
  http://www.ibm.com/support/knowledgecenter/SSZJPZ_11.3.0/⇒
  com.ibm.swg.im.iis.common.doc/common/accessingiidoc.html
  ```

  Tip:

  The knowledge center has a short URL as well:

  `http://ibm.biz/knowctr`

  To specify a short URL to a specific product page, version, or topic, use a hash character (#) between the short URL and the product identifier. For example, the short URL to all the InfoSphere Information Server documentation is the following URL:

  `http://ibm.biz/knowctr#SSZJPZ/`

  And, the short URL to the topic above to create a slightly shorter URL is the following URL (The ⇒ symbol indicates a line continuation):

  ```
  http://ibm.biz/knowctr#SSZJPZ_11.3.0/com.ibm.swg.im.iis.common.doc/⇒
  common/accessingiidoc.html
  ```

Changing help links to refer to locally installed documentation

IBM Knowledge Center contains the most up-to-date version of the documentation. However, you can install a local version of the documentation as an information center and configure your help links to point to it. A local information center is useful if your enterprise does not provide access to the internet.

Use the installation instructions that come with the information center installation package to install it on the computer of your choice. After you install and start the information center, you can use the **iisAdmin** command on the services tier computer to change the documentation location that the product F1 and help links refer to. (The ⇒ symbol indicates a line continuation):

Windows

```
IS_install_path\ASBServer\bin\iisAdmin.bat -set -key ⇒
com.ibm.iis.infocenter.url -value http://<host>:<port>/help/topic/
```

AIX® Linux

```
IS_install_path/ASBServer/bin/iisAdmin.sh -set -key ⇒
com.ibm.iis.infocenter.url -value http://<host>:<port>/help/topic/
```

Where <host> is the name of the computer where the information center is installed and <port> is the port number for the information center. The default port number is 8888. For example, on a computer named server1.example.com that uses the default port, the URL value would be http://server1.example.com:8888/help/topic/.

Obtaining PDF and hardcopy documentation

- The PDF file books are available online and can be accessed from this support document: https://www.ibm.com/support/docview.wss?uid=swg27008803&wv=1.
- You can also order IBM publications in hardcopy format online or through your local IBM representative. To order publications online, go to the IBM Publications Center at http://www.ibm.com/e-business/linkweb/publications/servlet/pbi.wss.

Appendix D. Providing feedback on the product documentation

You can provide helpful feedback regarding IBM documentation.

Your feedback helps IBM to provide quality information. You can use any of the following methods to provide comments:

- To provide a comment about a topic in IBM Knowledge Center that is hosted on the IBM website, sign in and add a comment by clicking **Add Comment** button at the bottom of the topic. Comments submitted this way are viewable by the public.
- To send a comment about the topic in IBM Knowledge Center to IBM that is not viewable by anyone else, sign in and click the **Feedback** link at the bottom of IBM Knowledge Center.
- Send your comments by using the online readers' comment form at www.ibm.com/software/awdtools/rcf/.
- Send your comments by e-mail to comments@us.ibm.com. Include the name of the product, the version number of the product, and the name and part number of the information (if applicable). If you are commenting on specific text, include the location of the text (for example, a title, a table number, or a page number).

Notices and trademarks

This information was developed for products and services offered in the U.S.A. This material may be available from IBM in other languages. However, you may be required to own a copy of the product or product version in that language in order to access it.

Notices

IBM may not offer the products, services, or features discussed in this document in other countries. Consult your local IBM representative for information on the products and services currently available in your area. Any reference to an IBM product, program, or service is not intended to state or imply that only that IBM product, program, or service may be used. Any functionally equivalent product, program, or service that does not infringe any IBM intellectual property right may be used instead. However, it is the user's responsibility to evaluate and verify the operation of any non-IBM product, program, or service.

IBM may have patents or pending patent applications covering subject matter described in this document. The furnishing of this document does not grant you any license to these patents. You can send license inquiries, in writing, to:

IBM Director of Licensing
IBM Corporation
North Castle Drive
Armonk, NY 10504-1785 U.S.A.

For license inquiries regarding double-byte character set (DBCS) information, contact the IBM Intellectual Property Department in your country or send inquiries, in writing, to:

Intellectual Property Licensing
Legal and Intellectual Property Law
IBM Japan Ltd.
19-21, Nihonbashi-Hakozakicho, Chuo-ku
Tokyo 103-8510, Japan

The following paragraph does not apply to the United Kingdom or any other country where such provisions are inconsistent with local law:
INTERNATIONAL BUSINESS MACHINES CORPORATION PROVIDES THIS PUBLICATION "AS IS" WITHOUT WARRANTY OF ANY KIND, EITHER EXPRESS OR IMPLIED, INCLUDING, BUT NOT LIMITED TO, THE IMPLIED WARRANTIES OF NON-INFRINGEMENT, MERCHANTABILITY OR FITNESS FOR A PARTICULAR PURPOSE. Some states do not allow disclaimer of express or implied warranties in certain transactions, therefore, this statement may not apply to you.

This information could include technical inaccuracies or typographical errors. Changes are periodically made to the information herein; these changes will be incorporated in new editions of the publication. IBM may make improvements and/or changes in the product(s) and/or the program(s) described in this publication at any time without notice.

Any references in this information to non-IBM Web sites are provided for convenience only and do not in any manner serve as an endorsement of those Web sites. The materials at those Web sites are not part of the materials for this IBM product and use of those Web sites is at your own risk.

IBM may use or distribute any of the information you supply in any way it believes appropriate without incurring any obligation to you.

Licensees of this program who wish to have information about it for the purpose of enabling: (i) the exchange of information between independently created programs and other programs (including this one) and (ii) the mutual use of the information which has been exchanged, should contact:

IBM Corporation
J46A/G4
555 Bailey Avenue
San Jose, CA 95141-1003 U.S.A.

Such information may be available, subject to appropriate terms and conditions, including in some cases, payment of a fee.

The licensed program described in this document and all licensed material available for it are provided by IBM under terms of the IBM Customer Agreement, IBM International Program License Agreement or any equivalent agreement between us.

Any performance data contained herein was determined in a controlled environment. Therefore, the results obtained in other operating environments may vary significantly. Some measurements may have been made on development-level systems and there is no guarantee that these measurements will be the same on generally available systems. Furthermore, some measurements may have been estimated through extrapolation. Actual results may vary. Users of this document should verify the applicable data for their specific environment.

Information concerning non-IBM products was obtained from the suppliers of those products, their published announcements or other publicly available sources. IBM has not tested those products and cannot confirm the accuracy of performance, compatibility or any other claims related to non-IBM products. Questions on the capabilities of non-IBM products should be addressed to the suppliers of those products.

All statements regarding IBM's future direction or intent are subject to change or withdrawal without notice, and represent goals and objectives only.

This information is for planning purposes only. The information herein is subject to change before the products described become available.

This information contains examples of data and reports used in daily business operations. To illustrate them as completely as possible, the examples include the names of individuals, companies, brands, and products. All of these names are fictitious and any similarity to the names and addresses used by an actual business enterprise is entirely coincidental.

COPYRIGHT LICENSE:

This information contains sample application programs in source language, which illustrate programming techniques on various operating platforms. You may copy, modify, and distribute these sample programs in any form without payment to IBM, for the purposes of developing, using, marketing or distributing application programs conforming to the application programming interface for the operating platform for which the sample programs are written. These examples have not been thoroughly tested under all conditions. IBM, therefore, cannot guarantee or imply reliability, serviceability, or function of these programs. The sample programs are provided "AS IS", without warranty of any kind. IBM shall not be liable for any damages arising out of your use of the sample programs.

Each copy or any portion of these sample programs or any derivative work, must include a copyright notice as follows:

© (your company name) (year). Portions of this code are derived from IBM Corp. Sample Programs. © Copyright IBM Corp. _enter the year or years_. All rights reserved.

If you are viewing this information softcopy, the photographs and color illustrations may not appear.

Privacy policy considerations

IBM Software products, including software as a service solutions, (“Software Offerings”) may use cookies or other technologies to collect product usage information, to help improve the end user experience, to tailor interactions with the end user or for other purposes. In many cases no personally identifiable information is collected by the Software Offerings. Some of our Software Offerings can help enable you to collect personally identifiable information. If this Software Offering uses cookies to collect personally identifiable information, specific information about this offering’s use of cookies is set forth below.

Depending upon the configurations deployed, this Software Offering may use session or persistent cookies. If a product or component is not listed, that product or component does not use cookies.

Table 15. Use of cookies by InfoSphere Information Server products and components

Product module	Component or feature	Type of cookie that is used	Collect this data	Purpose of data	Disabling the cookies
Any (part of InfoSphere Information Server installation)	InfoSphere Information Server web console	• Session • Persistent	User name	• Session management • Authentication	Cannot be disabled
Any (part of InfoSphere Information Server installation)	InfoSphere Metadata Asset Manager	• Session • Persistent	No personally identifiable information	• Session management • Authentication • Enhanced user usability • Single sign-on configuration	Cannot be disabled

Table 15. Use of cookies by InfoSphere Information Server products and components (continued)

Product module	Component or feature	Type of cookie that is used	Collect this data	Purpose of data	Disabling the cookies
InfoSphere DataStage	Big Data File stage	• Session • Persistent	• User name • Digital signature • Session ID	• Session management • Authentication • Single sign-on configuration	Cannot be disabled
InfoSphere DataStage	XML stage	Session	Internal identifiers	• Session management • Authentication	Cannot be disabled
InfoSphere DataStage	IBM InfoSphere DataStage and QualityStage Operations Console	Session	No personally identifiable information	• Session management • Authentication	Cannot be disabled
InfoSphere Data Click	InfoSphere Information Server web console	• Session • Persistent	User name	• Session management • Authentication	Cannot be disabled
InfoSphere Data Quality Console		Session	No personally identifiable information	• Session management • Authentication • Single sign-on configuration	Cannot be disabled
InfoSphere QualityStage Standardization Rules Designer	InfoSphere Information Server web console	• Session • Persistent	User name	• Session management • Authentication	Cannot be disabled
InfoSphere Information Governance Catalog		• Session • Persistent	• User name • Internal identifiers • State of the tree	• Session management • Authentication • Single sign-on configuration	Cannot be disabled
InfoSphere Information Analyzer	Data Rules stage in the InfoSphere DataStage and QualityStage Designer client	Session	Session ID	Session management	Cannot be disabled

If the configurations deployed for this Software Offering provide you as customer the ability to collect personally identifiable information from end users via cookies and other technologies, you should seek your own legal advice about any laws applicable to such data collection, including any requirements for notice and consent.

For more information about the use of various technologies, including cookies, for these purposes, see IBM's Privacy Policy at http://www.ibm.com/privacy and IBM's Online Privacy Statement at http://www.ibm.com/privacy/details the section entitled "Cookies, Web Beacons and Other Technologies" and the "IBM Software Products and Software-as-a-Service Privacy Statement" at http://www.ibm.com/software/info/product-privacy.

Trademarks

IBM, the IBM logo, and ibm.com® are trademarks or registered trademarks of International Business Machines Corp., registered in many jurisdictions worldwide. Other product and service names might be trademarks of IBM or other companies. A current list of IBM trademarks is available on the Web at www.ibm.com/legal/copytrade.shtml.

The following terms are trademarks or registered trademarks of other companies:

Adobe is a registered trademark of Adobe Systems Incorporated in the United States, and/or other countries.

Intel and Itanium are trademarks or registered trademarks of Intel Corporation or its subsidiaries in the United States and other countries.

Linux is a registered trademark of Linus Torvalds in the United States, other countries, or both.

Microsoft, Windows and Windows NT are trademarks of Microsoft Corporation in the United States, other countries, or both.

UNIX is a registered trademark of The Open Group in the United States and other countries.

Java™ and all Java-based trademarks and logos are trademarks or registered trademarks of Oracle and/or its affiliates.

The United States Postal Service owns the following trademarks: CASS, CASS Certified, DPV, LACS^{Link}, ZIP, ZIP + 4, ZIP Code, Post Office, Postal Service, USPS and United States Postal Service. IBM Corporation is a non-exclusive DPV and LACS^{Link} licensee of the United States Postal Service.

Other company, product or service names may be trademarks or service marks of others.

Index

Printed in USA

SC19-4249-00