

Partitioning target S3 files in Informatica Cloud (IICS)

Contents

Introduction

Distribution Column

Steps to create partitioned output files using Distribution Column

Creating multiple target output files using Source Partitioning

Introduction

Informatica creating multiple target S3 files by configuring a single target transformation in a mapping. This can be achieved using the **Distribution Column** option available in the advanced target properties.

Distribution Column

This behaves similar to a transaction control transformation where you specify a column and any change in that column value commits the transaction and create an output target file.

This method is even simpler where you specify a column as Distribution Column in target advanced properties and the secure agent creates multiple files based on that column value.

Each target file name is appended with the value of the distribution column as shown below.

TargetFileName_DistributionColumnValue.csv

Note that this method is supported only for the flat file format.

Steps to create partitioned output files using Distribution Column

Consider a scenario where you have customer's information from 2019 till 2022 coming from a database which needs to be loaded into S3 bucket. Since the data is huge you wanted the output files created in S3 partitioned by year.

Follow below steps to create partitioned output files using Distribution Column

1. In the source transformation of the mapping, select the source object with customer data.
2. Since the requirement is to create the target files based on year, make sure there is a column which provides year information from source. Else derive the year value from other existing columns.
3. For example, there is field named Purchase_Date of data type date in the source. Derive the year value as shown below using expression transformation.

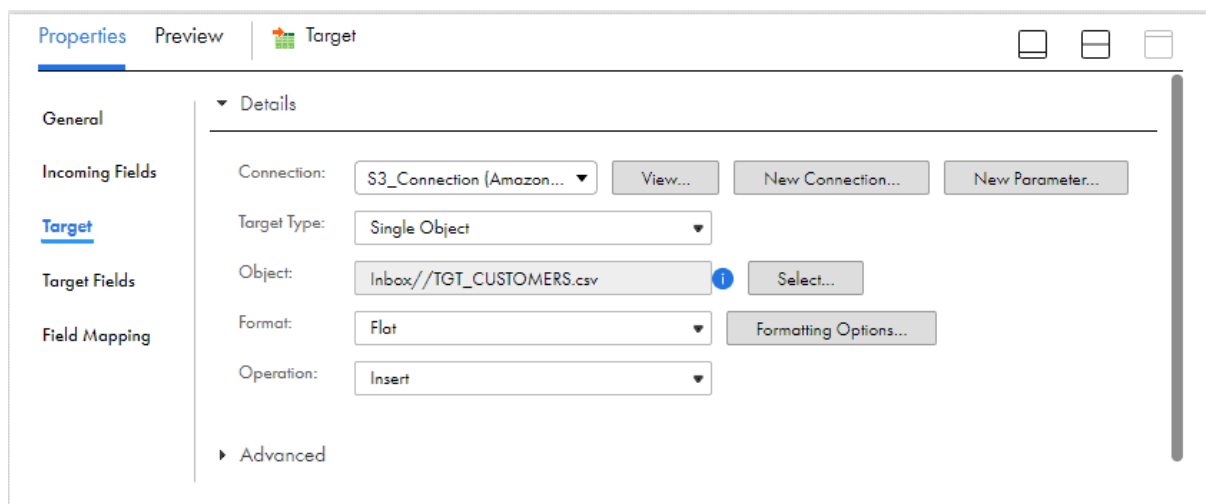
`YEAR = SUBSTR(TO_CHAR(PURCHASE_DATE),7,4)`

Example:

Purchase_Date = 12/25/2019

Year = 2019

4. In the target transformation, select the S3 connection. Under the Object, select Create New at Runtime and enter the target Object Name and Path where the files needed to be created as shown below.



Target transformation configured with S3 object

5. Next select the file format as Flat.
6. Under Formatting Options, after selecting the flat file properties as desired navigate to Distribution Column and enter the value as YEAR created earlier and click OK.

Formatting Options ↗ (X)

Flat File Type: ☒ Delimited ☐ Fixed Width

Schema Target: Derived from input fields when creating a new target.

Delimiter:

Escape Character:

Qualifier:

Qualifier Mode: ▼

Code Page: ▼

Header Line Number:

First Data Row:

Target Header: ▼

Distribution Column:

☐ Retain Escape Character in Data

Maximum Rows to Preview:





Row Delimiter:

?
OK
Cancel

Distribution Column configured in Formatting Options

7. Save and trigger the mapping.

Once the mapping is succeeded, the output files created in S3 will be as below.

| <input type="checkbox"/> | Name ▲ | Type ▼ | Size ▼ | Storage class ▼ |
|--------------------------|----------------------------------------------------------------------------------------------------------------------------|--------|---------|-----------------|
| <input type="checkbox"/> |  TGT_CUSTOMERS_2019.csv | csv | 20.1 MB | Standard |
| <input type="checkbox"/> |  TGT_CUSTOMERS_2020.csv | csv | 32.7 MB | Standard |
| <input type="checkbox"/> |  TGT_CUSTOMERS_2021.csv | csv | 32.8 MB | Standard |
| <input type="checkbox"/> |  TGT_CUSTOMERS_2022.csv | csv | 3.9 MB | Standard |

Target files generated using Distribution Column

Creating multiple target output files using Source Partitioning

As discussed creating multiple output files using distribution column is supported only for flat files in S3.

But what if you are working with a different file formats like parquet and wanted to generate multiple output files?

Well, we could still achieve this using several ways, let us only discuss what could be achieved purely from Informatica end.

One way of achieving this is by using source partitioning technique.

Though Partitioning is a performance tuning technique which enables parallel processing of data through separate pipelines, we could use it to our advantage to create multiple output files in S3.

If your source is a relational, the partitioning type supported is Key Range and the partitions for the example we discussed in earlier section should be configured as below.

The screenshot shows the 'Properties' window in Informatica, with the 'Partitions' tab selected. The 'Partitioning type' is set to 'Key Range'. Below this, the 'Partition key' is set to 'YEAR'. A table titled 'Key Ranges' displays four partitions with their respective start and end ranges. A plus icon is visible to the right of the table header.

| Partition | Start range | End range |
|-----------|-------------|-----------|
| #1 | 2019 | 2020 |
| #2 | 2020 | 2021 |
| #3 | 2021 | 2022 |
| #4 | 2022 | 2023 |

Configuring Key Range Partitioning for relational sources

If your source is a non-relational, the partitioning type supported is Fixed and the number of partitions for the example we discussed in earlier section should be configured as below.

The screenshot shows a software interface with a 'Properties' window. The 'Partitions' tab is selected, showing 'Partitioning type' as 'Fixed' and 'Number of partitions' as 4. The 'Source' tab is also visible, showing instructions to enter the number of partitions to process data in parallel.

Configuring Fixed Partitioning for non-relational sources

In order to generate a separate output file for each of the configured partition, make sure Merge Partition Files property is unchecked in Advanced properties section of target transformation.

The output files created in S3 using partitioning method will be as below.

| <input type="checkbox"/> | Name | Type | Size | Storage class |
|--------------------------|---------------------|------|---------|---------------|
| <input type="checkbox"/> | TGT_CUSTOMERS_0.csv | csv | 20.1 MB | Standard |
| <input type="checkbox"/> | TGT_CUSTOMERS_1.csv | csv | 32.7 MB | Standard |
| <input type="checkbox"/> | TGT_CUSTOMERS_2.csv | csv | 32.8 MB | Standard |
| <input type="checkbox"/> | TGT_CUSTOMERS_3.csv | csv | 3.9 MB | Standard |

Target files generated using Partitioning

The disadvantage using this method is that you need to know about the data you are processing (data size and contents) and also should configure the number of partitions ahead.