Pipeline

Contents

Pipeline Configuration..... 3
What is a Pipeline?....3
Data Collector Console - Edit Mode....3
Merging Streams....3
Replicating Streams....4
Error Handling....4
Send to Error Stage Option....4
Reprocessing Error Records....4
Delivery Guarantee....5
SDC Record Data Format....5
Configuring a Pipeline....5

Pipeline Configuration

What is a Pipeline?

A pipeline describes the flow of data for the Data Collector.

Configure a pipeline on the *pipeline canvas*. To configure a pipeline, you add and configure an origin stage to represent source data, processor stages to perform processing and routing, and one or more destination stages to represent target data. Connect the stages to create the flow of data from origin to destinations.

The canvas automatically saves and validates your work. The pipeline canvas provides warnings for parts of the pipeline that are not yet complete or valid. You can view a full list of issues to help you configure the pipeline.

To help configure the pipeline, you can also perform a data preview to view how the pipeline transforms source data.

When you start a pipeline, the Data Collector runs the pipeline until you stop the pipeline or shut down the Data Collector. While the pipeline runs, you can monitor the pipeline and configure alerts to verify that the pipeline performs as expected.

Data Collector Console - Edit Mode

The following image shows the Data Collector console when you configure a pipeline:

Area / Icon	Name	Description
1	Pipeline canvas	Displays the pipeline. Use to configure the stages for the pipeline.
2	Properties panel	Displays the properties of the pipeline or selected stage when you configure a pipeline.
3	Stage list	Displays the name of the selected stage. To configure properties for a different stage, select the stage from the list. Or, click the stage in the pipeline.
	Add Stage button	Displays a list of stages. Use to add stages to the pipeline.
	Issues button	Displays the number of validation issues in the pipeline. Click to view a detailed list of issues.
	Error icon	Indicates one or more errors in stage configuration.

Merging Streams

You can merge streams of data in a pipeline by connecting two or more stages to the same downstream stage. When you merge streams of data, the Data Collector channels the data from all streams to the same stage, but does not perform a join of records in the stream.

For example, in the following pipeline, the Routes Data Stream Selector sends data with null values to Replaces Null Values Value Replacer.

The data from stream 2 of Routes Data and all data from Replaces Null Values pass to Field-Level Expressions for further processing, but in no particular order and with no record merging.

Note: The pipeline validation does not prevent duplicate data. To avoid writing duplicate data to destinations, configure pipeline logic to remove duplicate data or to prevent the generation of duplicate data.

Replicating Streams

When you can pass data from a stage to multiple stages, the source stage passes all data to all connected stages. You can configure the logic for connected stages to discard different records, as with Required Fields, but all records are passed by default.

Use a Stream Selector to separate and route data based on conditions.

Error Handling

You can configure the default error handling for a pipeline. Some stages include error handling properties that override the pipeline default. When a Data Collector encounters an unexpected error, it stops the pipeline.

You can configure a pipeline to discard error records or to save error records. When you save error records, you define the directory to use and when to create additional files. Error files are written to the directory with the following naming convention: records-<file number>.json.

Each stage in the pipeline includes built-in resilience. Some stages also include configurable error handling. When a stage without explicit error handling options encounters an error record, the Data Collector uses the default error handling for the pipeline.

For example, when a Kafka Consumer origin stage reads JSON data with a maximum object length of 4096 characters and the stage encounters an object with 5000 characters, the stage discards or saves the object based on the pipeline error handling configuration.

When a stage includes an error handling option, the stage configuration can override the pipeline configuration.

For example, a Field Splitter splits field values into three parts and discards records that cannot be split. Even if the pipeline is configured to save error records, the Field Splitter stage discards any record that cannot be split as requested.

Send to Error Stage Option

Any stage with an error handling property includes the **Send to Error** option. The Send to Error option passes the error record to the pipeline for error handling.

The record is either discarded or saved to an error record file based on how the pipeline is configured.

Reprocessing Error Records

You can use a Directory origin in a error pipeline to reprocess error record files. When you reprocess error record files, do not edit or rename the files. The Directory origin expects the files as generated by the original pipeline.

In the error pipeline, include the Directory origin and configure it to use the SDC Records data format. The SDC Records data format provides the following information:

- The original source record, as read by the pipeline that generated the error.
- The path the record took through the pipeline, including the stage that discarded the record.
- Information about the error, including the error code and message.

When you create the error pipeline, you can use error and record functions provided by the expression language to route different types of error records through pipelines that resolve the error and write the corrected record to destinations.

For example, your error files contain records with invalid product IDs discarded as a required field from the first processor in the pipeline, a Field Filter. They also include records discarded by a Field Converter for attempting an invalid data type conversion.

In the error pipeline, you can use a Stream Selector to route the records discarded from the Field Filter to a branch that corrects product IDs. It routes the records discarded by the Field Converter to a branch that performs a valid data type conversion.

Delivery Guarantee

When you configure a pipeline, you define how you want data to be treated: Do you want to prevent the loss of data, or do you want to prevent the duplication of data?

The Delivery Guarantee pipeline property offers the following choices:

At least once

The Data Collector ensures that the pipeline processes all data.

If a failure causes the Data Collector to stop while processing a batch of data, when restarts, it reprocesses the batch.

This option ensures that no data is lost, but can write duplicate data to targets. Up to one batch data might be duplicated in the target.

At most once

The Data Collector ensures that data is not processed more than once.

If a failure causes the Data Collector to stop while processing a batch of data, when it starts up, it begins processing the next batch of data.

This option ensures duplicate data is not written to the target due to reprocessing, but can result in the loss of data. Up to one batch of data might not be processed.

SDC Record Data Format

SDC Record is a data format used by the Data Collector to generate output and error files for the pipeline.

Use the SDC Record format to process files produced by another Data Collector pipeline, or to produce files that you might pass to another Data Collector pipeline.

The SDC Record Files destination writes all records using the SDC Record format. You can also use Hadoop FS and the Kafka Producer to write files that use the SDC Record format.

You can use the Directory or Kafka Consumer origins to read SDC Record files.

Configuring a Pipeline

Configure a pipeline to define the stream of data. After you configure the pipeline, you can start the pipeline.

A pipeline can include the following components:

- A single origin stage
- Multiple processor stages
- Multiple destination stages

1. If this is the first pipeline for the Data Collector, click **Create Pipeline**.

   If you have created other pipelines, if necessary, use the **Toggle Library** icon to display the Library, and click the **Create New Pipeline** icon.
2. In the **New Pipeline** window, enter a pipeline name and optional description, and click **Save**.
   The pipeline canvas displays the new name. The Properties panel displays the pipeline properties.
3. In the Properties panel, on the **General** tab, configure the **Delivery Guarantee** property:

Pipeline Property	Description
Delivery Guarantee	Determines how the Data Collector handles data after an unexpected event causes the pipeline to stop running. Select one of the following options: • **At least once**. Ensures all data is processed and written to the destination. Might result in duplicate rows. • **At most once**. Ensures that data is not reprocessed to prevent writing duplicate data to the destination. Might result in missing rows. Default is **At least once**.

4. Click the **Error Records** tab and configure the following error handling option:

Error Records Property	Description
Error Record Handling	Determines how to handle records that cannot be processed as expected. Use one of the following options: • **Trash** - Saves error records to a file. • **Records to File** - Discards error records.

5. On the **Bad Records - Records to File** tab, configure error file properties:

Error File Property	Description
Directory	Local directory for error record files.
File Wait Time (secs)	Number of seconds the Data Collector waits for error records. After that time, it creates a new error record file. You can enter a number of seconds or use the default expression to enter the time in minutes.
Max File Size (MB)	Maximum size for error files. Exceeding this size creates a new error file. Use 0 to write to avoid using this property.

6. Use the **Stage Library** icon to add an origin stage. In the Properties panel, configure the stage properties.
   For configuration details about origin stages, see *Origins*.
7. Use the **Stage Library** button to add the next stage that you want to use, connect the origin to the new stage, and configure the new stage.
   For configuration details about processors, see *Processors*.
   For configuration details about destinations, see *Destinations*.
   For configuration details about destinations, see *Destinations*.
8. Set additional stages as necessary.
9. When the pipeline is valid, you can use the **Preview** icon to preview data to fine-tune pipeline configuration.
   For more information, see *Data Preview*.
10. When the pipeline is complete, use the **Start** icon to run the pipeline.

The Data Collector starts the pipeline. The Monitor panel displays real-time statistics for the pipeline.

Index

D

delivery guarantee
    pipeline property *5*

E

error records
    reprocessing *4*

S

SDC Records
    data format *5*
Send to Error
    error handling option *4*