StreamSets Data Collector and Data Collector Edge 3.10.0 Release Notes

August 1, 2019

We're happy to announce new versions of StreamSets Data Collector and StreamSets Data Collector Edge. Version 3.10.0 contains several new features, enhancements, and some important bug fixes. This document contains important information about the following topics:

New Features and Enhancements in Version 3.10.x

Version 3.10.x includes several new features and enhancements for Data Collector and Data Collector Edge.

Data Collector New Features and Enhancements

This Data Collector version includes new features and enhancements in the following areas.

Enterprise Stage Libraries

Enterprise stage libraries are free for development purposes only. For information about purchasing an Enterprise stage library for use in production, contact StreamSets.

On August 15, 2019, StreamSets released new and updated Enterprise stage libraries. For a list of available Enterprise libraries, see Enterprise Stage Libraries in the Data Collector documentation. For more information about the new features, fixed issues, and known issues in an Enterprise stage library, see the release notes for the Enterprise stage library, available under Enterprise Libraries Documentation on the StreamSets Documentation page.

Origins

This release includes the following new origins:

- **Groovy Scripting** - Runs a Groovy script to create Data Collector records.
- **JavaScript Scripting** - Runs a JavaScript script to create Data Collector records.
- **Jython Scripting** - Runs a Jython script to create Data Collector records.
- **NiFi HTTP Server** - Listens for requests from a NiFi PutHTTP processor and processes NiFi FlowFiles.

This release includes enhancements to the following origins:

- **SQL Server CDC Client** - The origin now has two new record header attributes:
    - `jdbc.cdc.source_schema_name` - Stores the source schema.
    - `jdbc.cdc.source_name` - Stores the source table.

  Also, the origin no longer requires you to install a JDBC driver. Data Collector now includes the Microsoft SQL Server JDBC driver.

- **SQL Server Change Tracking** - The origin no longer requires you to install a JDBC driver. Data Collector now includes the Microsoft SQL Server JDBC driver.

Processors

This release includes enhancements to the following processors:

- **Groovy Evaluator**, **JavaScript Evaluator**, and **Jython Evaluator** - These processors now support the following:
    - User-defined parameters - On the Advanced tab, enter parameters and values. In the script, access the value with the `sdc.userParams` dictionary.
    - Full-screen script editing - With your cursor in the script field, press either F11 or Esc, depending on your operating system, to toggle full-screen editing.

Destinations

This release includes enhancements to the following destinations:

- **Cassandra** - The destination has four new properties to help you debug issues with the destination: Connection Timeout, Read Timeout, Consistency Level, and Log Slow Queries.

- **RabbitMQ Producer** - The destination has a new Set Expiration property, available when setting AMQP message properties on the RabbitMQ tab. Clear the Set Expiration property to disable expiration on messages that the destination sends.

Executors

This release includes enhancements to the following executor:

- **JDBC Query** - The executor can now run queries in parallel to improve throughput. On the Advanced tab, select the Enable Parallel Queries property to have the executor run queries simultaneously on each connection to the database.

Data Formats

This release includes enhancements to the following data formats:

- **Delimited data format** - You can now specify when Data Collector inserts quotes in generated delimited data. The Data Generator processor and destinations that write delimited data include a new Quote Mode property on the Data Format tab when you select a custom delimiter format for delimited data. Configure the Quote Mode property to generate data that quotes all fields, only fields that contain special characters, or no fields.

- **Excel data format** - In origins that read the Excel data format, you can now configure the origin to read either from all sheets in a workbook or from particular sheets in a workbook. Also, you can configure the origin to skip cells that do not have a corresponding header value.

Data Governance Tools

This release includes the following data governance tool enhancement:

- **Cloudera Navigator versions** - Data Collector can now publish metadata to Cloudera Navigator running on Cloudera Manager version 6.1.

Expression Language

This release includes the following new field functions:

- `f:index()` - Returns the index within the parent list field. Returns -1 if the field is not in a list.
- `f:parentPath()` - Returns the path of the parent field.
- `f:parent()` - Returns the parent field.
- `f:getSiblingWithName(<name>)` - Returns the sibling field with the name matching `<name>`, if the field exists.
- `f:hasSiblingWithName(<name>)` - Returns `true` if there is a sibling field with a name matching `<name>`.
- `f:hasSiblingWithValue(<name>, <value>)` - Returns `true` if there is a sibling field with a name matching `<name>` that has value matching `<value>`.

Data Collector Configuration

This release includes the following Data Collector configuration enhancement:

- The Data Collector configuration file `sdc.properties` contains a new stage-specific property, `stage.conf_com.streamsets.pipeline.stage.jdbc.drivers.load`, where you can list JDBC drivers that Data Collector automatically loads for all pipelines.

Stage Libraries

This release includes the following stage library enhancements:

- **New stage libraries** - This release includes the following new stage libraries:

Stage Library Name	Description
streamsets-datacollector-cdh_6_1-lib	For the Cloudera CDH version 6.1 distribution of Apache Hadoop.
streamsets-datacollector-cdh_6_2-lib	For the Cloudera CDH version 6.2 distribution of Apache Hadoop.
streamsets-datacollector-cdh_spark_2_3_r3-lib	For the Cloudera CDH cluster Kafka with CDS powered by Spark 2.3 release 3.

streamsets-datacollector-cdh_spark_2_3_r4-lib	For the Cloudera CDH cluster Kafka with CDS powered by Spark 2.3 release 4.

- **Legacy stage libraries** - The following stage libraries are now legacy stage libraries:

Stage Library Name	Description
streamsets-datacollector-hdp_2_6-lib	For the Hortonworks version 2.6.x distribution of Apache Hadoop.
streamsets-datacollector-hdp_2_6-flume-lib	For the Hortonworks version 2.6.x distribution of Apache Flume.
streamsets-datacollector-hdp_2_6-hive2-lib	For the Hortonworks version 2.6.x distribution of Apache Hive version 2.1.
streamsets-datacollector-hdp_2_6_1-hive1-lib	For the Hortonworks version 2.6.1 distribution of Apache Hive version 1.x.
streamsets-datacollector-hdp_2_6_2-hive1-lib	For the Hortonworks version 2.6.2 distribution of Apache Hive version 1.x.

Upgrading to Version 3.10.x

You can upgrade previous versions of Data Collector to version 3.10.0. For complete instructions on upgrading, see the Upgrade documentation.

Update Pipelines Using Legacy Stage Libraries

Starting with version 3.10.0, the following older stage libraries are now legacy stage libraries and are no longer included with Data Collector:

- streamsets-datacollector-hdp_2_6-lib
- streamsets-datacollector-hdp_2_6-flume-lib
- streamsets-datacollector-hdp_2_6-hive2-lib
- streamsets-datacollector-hdp_2_6_1-hive1-lib
- streamsets-datacollector-hdp_2_6_2-hive1-lib

Pipelines that use these legacy stage libraries will not run until you perform one of the following tasks:

Use a current stage library
We strongly recommend that you upgrade your system and use a current stage library in the pipeline:

1. Upgrade the system to a more current version.
2. Install the stage library for the upgraded system.
3. In the pipeline, edit the stage and select the appropriate stage library.

Install the legacy stage library
Though not recommended, you can still download and install the older stage libraries as custom stage libraries. For more information, see Legacy Stage Libraries.

Upgrade Enterprise Stage Libraries

When you upgrade Data Collector, you must determine whether to upgrade your Enterprise stage libraries. See Enterprise Stage Libraries in the Data Collector documentation for a list of available Enterprise stage libraries, the latest available versions, and links to the supported versions and the stage documentation. To view the release notes for Enterprise stage libraries, see the StreamSets Documentation page.

Note: Enterprise stage libraries are free for development purposes only. For information about purchasing an Enterprise stage library for use in production, contact StreamSets.

1. Uninstall the previous version of the Enterprise stage library.
    a. In Package Manager, select the installed version.
    b. Click the **Uninstall** icon.
    c. Restart Data Collector.
2. Follow the stage documentation to install the new version of the Enterprise stage library and restart Data Collector.

Fixed Issues in Version 3.10.0

The following table lists some of the known issues that are fixed with this release.

For the full list, click here.

JIRA	Description
SDC-11232	The MapR Multitopic Streams Consumer origin does not offer the option to upgrade the stage library.
SDC-9863	The Hadoop FS Standalone origin ignores files in the parent directory and only reads files from subdirectories when the origin reads files based on the last modified timestamp and processes subdirectories.

Known Issues in Version 3.10.0

Please note the following known issues with this release.

For a full list of known issues, click here.

JIRA	Description
SDC-9888	When record fields contain special characters, the InfluxDB destination writes invalid measurements and truncated values to the InfluxDB database.

SDC-9853	Running a cluster streaming mode pipeline using Spark 2.1 that includes the HTTP Client processor encounters a ClassCastException error. Workaround: Copy the `jersey-server-2.25.1.jar` file from the `$SDC_DIST/container-lib` directory into the `$SDC_DIST/streamsets-libs/streamsets-datacollector-basic-lib/lib` directory. Then, restart Data Collector and re-submit the cluster application.
SDC-9514	Runtime parameters are not supported in all configuration properties in cluster batch execution mode, such as Max Batch Size.
SDC-8855	The MySQL Binary Log origin does not start reading from the offset specified in the Initial Offset property after a pipeline restart.
SDC-8697	Starting multiple pipelines concurrently that run a Jython import call can lead to retry errors and cause some of the pipelines to fail.
SDC-8514	The Data Parser processor sends a record to the next stage for processing even when the record encounters an error. Workaround: Use a Stream Selector processor after the Data Parser. Define a condition for the Stream Selector that checks if the fields in the record were correctly parsed. If not parsed correctly, send the record to a stream that handles the error.
SDC-8474	The Data Parser processor loses the original record when the record encounters an error.
SDC-8320	Data Collector inaccurately calculates the Record Throughput statistics for cluster mode pipelines when some Data Collector workers have completed while others are still running.
SDC-8078	The HTTP Server origin does not release the ports that it uses after the pipeline stops. Releasing the ports requires restarting Data Collector.
SDC-7761	The Java keystore credential store implementation fails to work for a Data Collector installed through Cloudera Manager. The jks-cs command creates the Java keystore file in the Data Collector configuration directory defined for the parcel. However, for Data Collector to access the Java keystore file, the file must be outside of the parcel directory. The CyberArk and Vault credential store implementations do work with a Data Collector installed through Cloudera Manager.
SDC-7645	The Data Collector Docker image does not support processing data using another locale. Workaround: Install Data Collector from the tarball or RPM package.
SDC-6554	When converting Avro to Parquet on Impala, Decimal fields seem to be unreadable. Data Collector writes the Decimal data as variable-length byte arrays. And due to Impala issue IMPALA-2494, Impala cannot read the data.

SDC-5141	Due to a limitation in the Javascript engine, the Javascript Evaluator issues a null pointer exception when unable to compile a script.
SDC-4212	If you configure a UDP Source or UDP to Kafka origin to enable multithreading after you have already run the pipeline with the option disabled, the following validation error displays: `Multithreaded UDP server is not available on your platform.` Workaround: Restart Data Collector.
SDC-3944	The Hive Streaming destination using the MapR library cannot connect to a MapR cluster that uses Kerberos or username/password login authentication.
SDC-2374	A cluster mode pipeline can hang with a CONNECT_ERROR status. This can be a temporary connection problem that resolves, returning the pipeline to the RUNNING status. If the problem is not temporary, you might need to manually edit the pipeline state file to set the pipeline to STOPPED. Edit the file only after you confirm that the pipeline is no longer running on the cluster or that the cluster has been decommissioned. To manually change the pipeline state, edit the following file: `$SDC_DATA/runInfo/<cluster pipeline name>/<revision>/pipelineState.json` In the file, change CONNECT_ERROR to STOPPED and save the file.

Contact Information

For more information about StreamSets, visit our website: https://streamsets.com/.

Check out our Documentation page for doc highlights, what's new, and tutorials: streamsets.com/docs

Or you can go straight to our latest documentation here:
https://streamsets.com/documentation/datacollector/latest/help

To report an issue, to get help from our Google group, Slack channel, or Ask site, or to find out about our next meetup, check out our Community page: https://streamsets.com/community/.

For general inquiries, email us at info@streamsets.com.