StreamSets Data Collector and Data Collector Edge Cumulative 3.8.x Release Notes

May 2, 2019

We're happy to announce new versions of StreamSets Data Collector and StreamSets Data Collector Edge. Version 3.8.x contains several new features, enhancements, and some important bug fixes in the following versions of StreamSets Data Collector and Data Collector Edge:

- Version 3.8.2 - May 2, 2019
- Version 3.8.1 - April 4, 2019
- Version 3.8.0 - March 14, 2019

This document contains important information about the following topics:

Upgrading to Version 3.8.x

You can upgrade previous versions of Data Collector to version 3.8.x. For complete instructions on upgrading, see the Upgrade Documentation.

Update Pipelines Using Legacy Stage Libraries

With version 3.8.0, the following older stage libraries are now legacy stage libraries and are no longer included with Data Collector:

- streamsets-datacollector-cdh_5_10-lib
- streamsets-datacollector-cdh_5_11-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

To use Enterprise stage libraries with Data Collector 3.8.0, you must install the latest version of the Enterprise stage library, as follows:

- For the MemSQL and Teradata Enterprise stage libraries, use version 1.0.1.
- For the Snowflake Enterprise stage library, use version 1.0.1 or 1.0.2.

For more information, see “Supported Versions” in the stage documentation. To view the release notes for the Enterprise stage libraries, see the StreamSets Documentation page.

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

Pipeline Export

With version 3.8.0, Data Collector has changed the behavior of the pipeline **Export** option. Data Collector now strips all plain text credentials from exported pipelines. Previously, Data Collector included plain text credentials in exported pipelines.

To use the previous behavior and include credentials in the export, choose the new **Export with Plain Text Credentials** option when exporting a pipeline.

New Features and Enhancements in Version 3.8.x

Version 3.8.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.

Memory Monitoring

With version 3.8.0, Data Collector memory monitoring has been removed. Memory monitoring was disabled by default and was to be used only in development for troubleshooting specific issues. StreamSets recommends using JMX or operating system monitoring for memory consumption.

If memory monitoring is enabled for Data Collector after upgrading to 3.8.0, a message appears in the log indicating that memory monitoring is no longer supported.

As part of this feature removal, the following changes have been made:

- The `monitor.memory` Data Collector configuration property has been removed from the Data Collector configuration file.
- Two related pipeline configuration properties have also been removed: Max Pipeline Memory and On Memory Exceeded.
- Two related counter statistics are no longer available: Heap Memory Usage and Stage Heap Memory Usage.

Enterprise Stage Libraries

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

With this release, you can use the following new Enterprise stage library:

Stage Library Name	Description
streamsets-datacollector-oracle-lib	For bulk loading from static Oracle tables. Includes the Oracle Bulkload origin.

Origins

This release includes enhancements to the following origins:

- **Dev Raw Data Source** - The development origin can now generate events.
- **Hadoop FS Standalone** - The origin can now read files from multiple directories specified with a glob pattern.
- **Oracle CDC Client** - JDBC Fetch Size property has been replaced by the following new properties:
    - JDBC Fetch Size for Current Window
    - JDBC Fetch Size for Past Windows

  To enable expected behavior, upgraded pipelines use the previous JDBC Fetch Size configuration for the new properties.
- **REST Service** - The origin can now generate responses in XML format in addition to JSON format.
- **Salesforce** - The origin now supports aggregate functions in SOQL queries.
- **SFTP/FTP Client** - The origin includes the following enhancements:
    - The origin now supports using an email address for the user name in the resource URL.
    - The origin now supports using a glob pattern or a regular expression to define the file name pattern. Previously, the origin supported only a glob pattern.
    - When configured for private key authentication, the origin now supports entering the full path to the private key file or entering the private key in plain text. Previously, the origin supported only entering the full path to the file.
    - After processing a file, the origin can now keep, archive, or delete the file.

- **SQL Server CDC Client** - The default for the Maximum Transaction Length property has changed from `${1*HOUR}` to `-1` to opt out of using the property. Upgraded pipelines are not affected.
- **WebSocket Client** - The origin can now generate responses in XML format in addition to JSON format.
- **WebSocket Server** - The origin can now generate responses in XML format in addition to JSON format.

Processors

This release includes the following new processor:

- **Field Mapper** - Maps an expression to a set of fields to alter field paths, field names, or field values.

This release includes enhancements to the following processors:

- **Field Flattener** - When flattening specific fields, the processor now supports selecting fields using preview data in addition to entering the path to each field.
- **Salesforce Lookup** - The processor now supports aggregate functions in SOQL queries.
- **Windowing Aggregator** - The Aggregator processor has been renamed to the Windowing Aggregator processor.

Destinations

This release includes enhancements to the following destinations:

- **Google Pub/Sub Publisher** - The destination now includes properties to configure batches.
- **Solr** - The destination can now directly map record fields to Solr schema fields.

Pipelines

This release includes the following pipeline enhancements:

- **Pipeline export with or without plain text credentials** - Data Collector now provides the following pipeline export options:
  - Export - Strips all plain text credentials from the exported pipeline.
  - Export with Plain Text Credentials - Includes all plain text credentials in the exported pipeline.

  Previously, Data Collector always included plain text credentials in exported pipelines.
- **New microservice raw responses** - Origins in microservice pipelines can now send raw responses - passing responses to the origin system without an envelope.
- **Pipeline labels enhancement** - You can now configure pipeline labels in the New Pipeline dialog box when you create a pipeline. As in earlier releases, you can also configure labels on the General tab of pipeline properties.

Data Formats

This release includes the following data formats enhancement:

- **Delimited** - Data Collector now supports multi-character field separators in delimited data.

Data Collector Configuration

This release includes the following Data Collector configuration enhancements:

- **Protect sensitive data in configuration files** - You can now protect sensitive data in Data Collector configuration files by storing the data in an external location and then using the `exec` function to call a script or executable that retrieves the data. For example, you can develop a script that decrypts an encrypted file containing a password. Or you can develop a script that calls an external REST API to retrieve a password from a remote vault system.

  After developing the script, use the `exec` function in the Data Collector configuration file to call the script or executable as follows:
  `${exec("<script name>")}`

- **AWS Secrets Manager support** - Data Collector now integrates with the AWS Secrets Manager credential store system.

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-aws-secrets-manager-credentialstore-lib	For the AWS Secrets Manager credential store system.
streamsets-datacollector-hdp_3_1-lib	For Hortonworks version 3.1.
streamsets-datacollector-mapr_6_1-lib	For MapR version 6.1.0.
streamsets-datacollector-mapr_6_1-mep6-lib	For MapR 6.1.0, MapR Ecosystem Pack (MEP) version 6.

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

Stage Library Name	Description
streamsets-datacollector-cdh_5_10-lib	For the Cloudera CDH version 5.10 distribution of Apache Hadoop.
streamsets-datacollector-cdh_5_11-lib	For the Cloudera CDH version 5.11 distribution of Apache Hadoop.

Legacy stage libraries that are more than two years old are not included with Data Collector. Though not recommended, you can still download and install the older stage libraries as custom stage libraries.

If you have pipelines that use these legacy stage libraries, you will need to update the pipelines to use a more current stage library or install the legacy stage library manually. For more information see Update Pipelines using Legacy Stage Libraries.

Data Collector Edge New Features and Enhancements

This Data Collector Edge (SDC Edge) version includes new features and enhancements in the following areas.

Origins in Edge Pipelines

In edge pipelines, the Directory origin now supports processing compressed files and supports file post-processing. You can now configure an error directory, and you can have the origin archive or delete files after processing.

Destinations in Edge Pipelines

This release includes the following destination enhancements:

- Edge pipelines now support the Amazon S3 destination.
- In edge pipelines, you can now configure the Kafka Producer destination to connect securely to Kafka through SSL/TLS.

Data Formats in Edge Pipelines

This release includes the following data formats enhancements:

- Stages included in edge pipelines now list only the data formats that are supported in edge pipelines.
- The following stages can now process the binary data format when they are included in edge pipelines:
  - Amazon S3 destination
  - CoAP Client destination
  - HTTP Client origin and destination
  - HTTP Server origin
  - Kafka Producer destination
  - Kinesis Producer destination
  - MQTT Subscriber origin and MQTT Publisher destination
  - WebSocket Client origin and destination
- The following stages can now process the whole file data format when they are included in edge pipelines:
  - Amazon S3 destination
  - Directory origin

Fixed Issues in Version 3.8.2

The following table lists some of the known issues that are fixed with version 3.8.2.

For the full list, click here.

JIRA	Description
SDC-11353	Data Collector does not retry after the Kerberos server becomes unavailable during ticket renewal.
SDC-11314	The JDBC Multitable Consumer and JDBC Query Consumer origins encounter a null pointer exception when processing null values in an Oracle Timestamp with time zone column.
SDC-11283	The Salesforce origin, Salesforce Lookup processor, and Salesforce destination do not automatically handle expired sessions for the Bulk API.
SDC-11271	Pipeline metric rules might issue larger values for idle pipelines.
SDC-11267	Connections to RabbitMQ time out and cause the pipeline to fail when pipeline processing takes too long.
SDC-11261	The MongoDB and MongoDB Oplog origins do not support the Decimal type.
SDC-11250	The Hive Query executor occasionally leaks connections.
SDC-11249	When configured to map fields automatically, the Solr destination incorrectly attempts to verify that fields that can be generated by Solr exist in the record.
SDC-11233	When run from a Control Hub job, the Elasticsearch origin might encounter a null-pointer exception because it does not correctly read from the last saved offset.
SDC-11142	The pipelineStateHistory.json is still not closed before the pipeline is deleted.

Fixed Issues in Version 3.8.1

The following table lists some of the known issues that are fixed with version 3.8.1.

For the full list, click here.

JIRA	Description
SDC-11221	Invalid stage names prevent Data Collector from writing statistics to Control Hub.
SDC-11206	The TCP Server origin pool size is too small.
SDC-11195	The HDP 3.1.0 stage library does not include Hive stages.
SDC-11176	The Amazon S3 origin only generates a no-more-data event for the first object read.
SDC-11168	The Kinesis Consumer origin does not start reading messages from a specified timestamp.
SDC-11163	In pipelines upgraded from an earlier version of Data Collector, the Amazon S3 origin

	might not correctly read from the last saved offset.
SDC-11162	The Amazon SQS Consumer origin does not work when the region is set to Other.
SDC-11135	The Hadoop FS Standalone origin does not delete files during post processing.
SDC-11129	The PostgreSQL CDC Client origin only processes changes in a Write-Ahead Logging (WAL) record when all changes in the record match the configured schema and table name pattern.
SDC-11102	Some versions of CDH might not read Avro with logical types.
SDC-11092	Destinations that use JDBC drivers do not always cover nonstandard SQL error codes related to data.
SDC-11086	The Salesforce origin does not retrieve fields with null values when querying with the Bulk API.
SDC-11082	Origins that use the Oracle JDBC driver do not read Timestamp With Time Zone data types from Oracle.
SDC-11024	The Kinesis Consumer origin might incorrectly read data when multiple pipelines read from the same AWS Kinesis stream and exceptions occur.
SDC-10865	The Hbase Lookup processor fails when the row and timestamp lookup parameters are defined or when only the row lookup parameter is defined.
SDC-10800	The JDBC Producer destination might write an update before the insert.
SDC-10501	The Kafka Multitopic Consumer origin does not check the delivery guarantee configured for the pipeline.

Fixed Issues in Version 3.8.0

The following table lists some of the known issues that are fixed with version 3.8.0.

For the full list, click here.

JIRA	Description
SDC-11022	Data Collector reuses the Avro union index from the origin when writing Avro files.
SDC-11018	Decimal values written in Avro format are not rescaled.
SDC-11006	Logs do not include the context of stages throwing exceptions.
SDC-10987	When the offset configuration of a JDBC Multitable Consumer origin results in the first batch returning no records, the origin reads the whole table.
SDC-10846	Connection issues result in excessive trace messages when reporting usage statistics.
SDC-10836	When reading large numbers of records, the Google BigQuery origin duplicates records.

SDC-10808	The Amazon S3 origin fails to generate no-more-data events, which enable the Pipeline Finisher executor to transition pipelines to a Finished state.
SDC-10732	The Whole File Transformer processor does not convert Avro files from an Amazon S3 origin to Parquet files for an Amazon S3 destination.
SDC-10719	The JDBC Producer destination does not support the expression language in the Schema Name property.
SDC-10562	When the JDBC property Use Multi-Row Operation is enabled, row-level errors are not processed.
SDC-10538	Pipelines do not send an email notification when transitioning to Running_Error state.
SDC-10425	Pipelines with the TCP Server origin do not release the TCP port after stopping due to an exception.
SDC-10384	When using test origin as the preview source, previewing a pipeline with an event-producing stage throws a nonreversible exception.
SDC-10073	The SQL Server CDC Client origin generates excessive no-more-data events when processing update or delete operations.

Known Issues in Version 3.8.x

Please note the following known issues with this release.

For a full list of known issues, click here.

JIRA	Description
SDC-10022	When the JDBC Multitable Consumer origin performs non-incremental processing, the origin does not correctly trigger the 'no-more-data' event.
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.