StreamSets Data Collector and Data Collector Edge 3.9.1 Release Notes

June 27, 2019

We're happy to announce new versions of StreamSets Data Collector and StreamSets Data Collector Edge. Version 3.9.x contains several new features, enhancements, and some important bug fixes in the following versions of StreamSets Data Collector and Data Collector Edge:

- Version 3.9.1 - June 27, 2019
- Version 3.9.0 - June 6, 2019

This document contains important information about the following topics:

New Features and Enhancements in Version 3.9.x

Version 3.9.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.

Origins

This release includes enhancements to the following origins:

- **Hadoop FS Standalone** **and** **MapR FS Standalone** - These origins include the following tab and property name changes:
    - The Hadoop FS tab is now the Connection tab.
    - The Hadoop FS URI property is now the File System URI property.
    - The HDFS User property is now the Impersonation User property.
    - The Hadoop FS Configuration Directory property is now the Configuration Files Directory property.
    - The Hadoop FS Configuration property is now the Additional Configuration property.

  The functionality associated with these properties has not changed.

- **JDBC Multitable Consumer** - The origin now supports using multithreaded partition processing when the primary key or user-defined offset column is an Oracle Timestamp with time zone data type and each row has the same time zone.

- **JMS Consumer** - The origin now supports reading messages from durable topic subscriptions, which can have only one active subscriber at a time.

- **SFTP/FTP/FTPS Client** - The origin, formerly called SFTP/FTP Client, now supports FTPS (FTP over SSL). Post processing is now disabled after selecting the Whole File data format, which does not support post processing.

Processors

This release includes the following new processor:

- **Couchbase Lookup** - Performs key/value or N1QL lookups against a Couchbase bucket to enrich records with data.

This release includes enhancements to the following processors:

- **Groovy Evaluator**, **JavaScript Evaluator**, and **Jython Evaluator** - These scripting processors now support direct use of Data Collector records after you set the new Record Type property on the Advanced tab to Data Collector Records.

- **Hive Metadata** - The processor can now process datetime fields in their native format or can convert the fields to string before processing the data. By default, the processor processes datetime fields in their native format. Previously, the processor always converted datetime fields to string.

- **Log Parser** - The processor now has a new Data Format tab that contains properties related to format. These include new properties that configure the maximum line length, the character set, and the retention of the original line from the log. For the Grok Pattern format, the processor now supports entry of multiple grok patterns. For the Log4j format, the processor now has properties to configure the action taken on parse error and the size of the stack trace that can be included in a record for a log.

Destinations

This release includes the following new destination:

- **SFTP/FTP/FTPS Client** - Sends data to a URL using SFTP, FTP, or FTPS.

This release includes enhancements to the following destinations:

- **Aerospike** - The destination can now use CRUD operations defined in the sdc.operation.type record header attribute to upsert or delete data. You can define a default operation for records without the header attribute or value. You can also configure how to handle records with unsupported operations.

- **Azure Data Lake Storage (Legacy)** - The destination, formerly called Azure Data Lake Storage, has been renamed. Data Collector now includes the Azure Data Lake Storage Gen1 destination that also writes data to Microsoft Azure Data Lake Storage Gen1. The Azure Data Lake Storage Gen1 destination is a Technology Preview stage.

- **Couchbase** - The destination includes the following enhancements:
    - Supports CRUD operations defined in the sdc.operation.type record header attribute to write data. You can define a default operation for records without the header attribute or value. You can also configure how to handle records with unsupported operations.
    - Supports writing to sub-documents.
    - Supports writing data using the Avro, Binary, Delimited, JSON, Protobuf, SDC Record, and Text data formats.
- **Hadoop FS and MapR FS** - These destinations include the following tab and property name changes:
    - The Hadoop FS tab is now the Connection tab.
    - The Hadoop FS URI property is now the File System URI property.
    - The HDFS User property is now the Impersonation User property.
    - The Hadoop FS Configuration Directory property is now the Configuration Files Directory property.
    - The Hadoop FS Configuration property is now the Additional Configuration property.

  The functionality associated with these properties has not changed.
- **HBase** - The destination can now skip validating that a table exists in HBase before writing to the table. By default, the destination validates that a table exists, which requires that the HBase user that writes to HBase has HBase administrator rights.

  You might want to configure the destination to skip the validation when you do not want to grant HBase administrator rights to the HBase user. If you configure the destination to skip validation and a table does not exist, then the pipeline will enter an error state. Previously, the destination always validated that a table existed.
- **Solr** - The following destination properties are now enabled by default:
    - Map Fields Automatically
    - Ignore Optional Fields

  Previously, both properties were disabled by default.
- **Syslog** - The following destination properties have been removed from the Message tab:
    - Use Non-Text Message Format
    - Message Text

  You now configure the destination to use the text data format on the Data Format tab. If you upgrade pipelines that used the Syslog destination configured to use the text data format, you must complete the post upgrade task described in Update Syslog Pipelines.

Executors

This release includes the following enhancements to executors:

- **Pipeline Finisher** - The executor includes a new Reset Offset option that ensures that the pipeline processes all available data with each pipeline run.

Technology Preview Functionality

Data Collector includes certain new features and stages with the Technology Preview designation. Technology Preview functionality is available for use in development and testing, but is not meant for use in production.

Technology Preview stages include the following image on the stage icon:

When Technology Preview functionality becomes approved for use in production, the release notes and documentation reflect the change, and the Technology Preview icon is removed from the UI.

The following Technology Preview stages are available in this release:

- **Azure Data Lake Storage Gen1 origin** - Reads data from Microsoft Azure Data Lake Storage Gen1.
- **Azure Data Lake Storage Gen2 origin** - Reads data from Microsoft Azure Data Lake Storage Gen2.
- **Azure Data Lake Storage Gen1 destination** - Writes data to Microsoft Azure Data Lake Storage Gen1.
- **Azure Data Lake Storage Gen2 destination** - Writes data to Microsoft Azure Data Lake Storage Gen2.
- **ADLS Gen1 File Metadata executor** - Changes file metadata, creates an empty file, or removes a file or directory in Microsoft Azure Data Lake Storage Gen1 upon receipt of an event.
- **ADLS Gen2 File Metadata executor** - Changes file metadata, creates an empty file, or removes a file or directory in Microsoft Azure Data Lake Storage Gen2 upon receipt of an event.

Pipelines

This release includes the following pipeline enhancements:

- **Pipeline Start menu** - The Data Collector toolbar now includes a pipeline Start menu with the following options:
  - Start Pipeline
  - Reset Origin and Start
  - Start with Parameters

  Previously, the Reset Origin and Start option was not available. The Start with Parameters option was located under the More icon.

- **Generated events** - For stages that generate events, the properties panel now includes a Generated Events tab, which lists and describes the events that origins can generate.

Data Governance Tools

This release includes the following data governance tool enhancement:

- **Apache Atlas versions** - Data Collector can now publish metadata to Apache Atlas version 1.1.0.

Expression Language

This release includes the following new time function:

- `time:extractNanosecondsFromString(<string>)` – Converts a String date with nanoseconds precision to an epoch or UNIX time in milliseconds, and then adds the nanoseconds using the following format:
  `<milliseconds_from_epoch><n><nanoseconds>`

  For example, the string `'29/05/2019 10:12:09.123456789'` is converted to `1559124729123<n>456789`.

Data Collector Configuration

This release includes the following Data Collector configuration enhancements:

- **Antenna Doctor** - Data Collector now includes Antenna Doctor, which is a rule-based engine that suggests potential fixes and workarounds to common issues. When needed, you can edit the Data Collector configuration file, `sdc.properties`, to disable Antenna Doctor or to disable Antenna Doctor from periodically retrieving knowledge base updates from the internet.
- **Legacy stage libraries** - Package Manager can now install legacy stage libraries.
- **Thycotic Secret Server support** - Data Collector now integrates with the Thycotic Secret Server 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-cdh_5_16-lib	For the Cloudera CDH version 5.16 distribution of Apache Hadoop.
streamsets-datacollector-kinetica_7_0-lib	For Kinetica 7.0.
streamsets-datacollector-thycotic-credentialstore-lib	For the Thycotic Secret Server credential store system.

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

Stage Library Name	Description
streamsets-datacollector-apache-kafka_0_11-lib	For Kafka version 0.11.x.
streamsets-datacollector-cdh_5_12-lib	For the Cloudera CDH version 5.12 distribution of Apache Hadoop.
streamsets-datacollector-cdh_5_13-lib	For the Cloudera CDH version 5.13 distribution of Apache Hadoop.
streamsets-datacollector-cdh_kafka_2_1-lib	For the Cloudera distribution of Apache Kafka 2.1.x (0.9.0).
streamsets-datacollector-cdh_kafka_3_0-lib	For the Cloudera distribution of Apache Kafka 3.0.0 (0.11.0).
streamsets-datacollector-cdh-spark_2_1-lib	For the Cloudera CDH cluster Kafka with CDS powered by Spark 2.1.
streamsets-datacollector-mapr_5_2-lib	For MapR version 5.2.

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

When you enable SSL/TLS for the HTTP Server origin in a Data Collector Edge pipeline, the origin now supports using a keystore file in the PKCS #12 format.

Processors in Edge Pipelines

Data Collector Edge pipelines now support the HTTP Client processor.

Destinations in Edge Pipelines

Data Collector Edge pipelines now support the Azure Event Hub Producer and the Azure IOT Hub Producer destinations.

Upgrading to Version 3.9.x

You can upgrade previous versions of Data Collector to version 3.9.0. For complete instructions on upgrading, see the Upgrade documentation.

Update Pipelines Using Legacy Stage Libraries

Starting with version 3.9.0, the following older stage libraries are now legacy stage libraries and are no longer included with Data Collector:

- streamsets-datacollector-apache-kafka_0_11-lib
- streamsets-datacollector-cdh_5_12-lib
- streamsets-datacollector-cdh_5_13-lib
- streamsets-datacollector-cdh_kafka_2_1-lib
- streamsets-datacollector-cdh_kafka_3_0-lib
- streamsets-datacollector-cdh-spark_2_1-lib
- streamsets-datacollector-mapr_5_2-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.

Update Syslog Pipelines

Starting with version 3.9.0, the Syslog destination no longer includes the following properties on the Message tab:

- Use Non-Text Message Format
- Message Text

You now configure the destination to use the Text data format on the Data Format tab, just as you do with other destinations.

If pipelines created in a previous version include the Syslog destination configured to use text data, you must configure the Text data format properties on the Data Format tab after the upgrade.

Update JDBC Pipelines

If upgrading from Data Collector version 3.4 or earlier, you must update pipelines that use a JDBC connection. Beginning with version 3.5.0, Data Collector requires the maximum lifetime for a connection to be at least 30 minutes in stages that use a JDBC connection. Data Collector does not validate stages with lower non-zero values configured.

To update pipelines that include a stage that uses a JDBC connection, update the stage to set the maximum lifetime for a connection to be at least 30 minutes.

On the Advanced tab, set the Max Connection Lifetime property to be at least 30 minutes or 1800 seconds.

Upgrade Enterprise Stage Libraries

Starting with version 3.9.0, Data Collector supports the following versions of the Enterprise stage libraries:

- For the MemSQL and Teradata Enterprise stage libraries, use version 1.0.1.
- For the Oracle Enterprise stage library, use version 1.0.0
- For the Snowflake Enterprise stage library, use version 1.0.1, 1.0.2, or 1.1.0.

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.

Fixed Issues in Version 3.9.1

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

For the full list, click here.

JIRA	Description
SDC-11849	The Amazon S3 origin generates excessive warning messages when reading data in Whole File format.
SDC-11840	Creating or importing a pipeline fails with draft flag if pipeline ACL is enabled.
SDC-11767	Metrics are not published to Control Hub consistently with pipeline finisher.
SDC-11757	The PostgreSQL CDC Client origin does not honor the Max Batch Size property.
SDC-11734	The PostgreSQL CDC Client origin does not update status when polling.
SDC-11670	The Hive Metadata processor does not work with Hive 3.
SDC-10897	In the Kafka Multitopic Consumer origin, the Batch Wait Time and Max Batch Size properties do not work together.

Fixed Issues in Version 3.9.0

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

For the full list, click here.

JIRA	Description
SDC-11642	The PostgreSQL CDC Client origin doesn't work correctly for multiple tables.
SDC-10353	Unable to delete record headers in scripting evaluators.
SDC-10022	When the JDBC Multitable Consumer origin performs non-incremental processing, the origin does not correctly trigger the 'no-more-data' event.

Known Issues in Version 3.9.x

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.