StreamSets Data Collector and Data Collector Edge 3.12.0 Release Notes

December 6, 2019

We're happy to announce new versions of StreamSets Data Collector and StreamSets Data Collector Edge. Version 3.12.0 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 3.12.x

Version 3.12.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.

In December 2019, StreamSets released a new Enterprise stage library for SQL Server 2019 Big Data Cluster.

In January 2020, StreamSets released an updated Enterprise stage library for Snowflake.

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 enhancements to the following origins:

- **Groovy Scripting**, **JavaScript Scripting**, and **Jython Scripting** - The origins now include two new methods in the `batch` object:

    - `addError(<record>,<String message>` - Appends an error record to the batch. The appended error record contains the associated error message. This method replaces the `sdc.error.write(<record>, <String message>)` method.
    - `addEvent(<event record>)` - Appends an event to the batch.This method replaces the `toEvent(<event record>)` method in the `sdc` object.

- **RabbitMQ Consumer** - The origin now supports transport layer security (TLS) for connections with a RabbitMQ server. You can configure the required properties on the TLS tab.
- **Salesforce** - The origin has a new property to configure the streaming buffer size when subscribed to notifications. Configure this property to eliminate buffering capacity errors.
- **SQL Server CDC Client** - The origin can now be configured to convert unsupported data types into strings and continue processing data.
- **SQL Server Change Tracking Client** - The origin includes the following enhancements:
    - The origin now interprets a 0 in the Fetch Size property as an indication to use the database default fetch size in JDBC statements.
    - The origin can now be configured to convert unsupported data types into strings and continue processing data.

Destinations

This release includes enhancements to the following destinations:

- **Azure Event Hub Producer** - The destination can now write records as XML data.
- **Cassandra** - The destination includes the following enhancements:
    - The Disable Batch Insert property has been renamed Enable Batches, and the renamed property is enabled by default.
    - The Request Timeout property has been renamed Write Timeout.
- **RabbitMQ Producer** - The destination now supports transport layer security (TLS) for connections with a RabbitMQ server. You can configure the required properties on the TLS tab.

Data Formats

This release includes the following data formats enhancement:

- **Avro** - For schemas located in Confluent Schema Registry, Data Collector now includes a property for you to specify the user information needed to connect to Schema Registry through basic authentication.

Deprecated Features in Version 3.12.x

Version 3.12.x newly deprecates the following feature:

- **Uploading support bundles** - Starting from December 16th, 2019, the ability to upload support bundles directly from Data Collector is deprecated, and will be removed in a future release.

  Going forward, please use Data Collector to generate and download a support bundle to your local machine. Then upload the file to the appropriate support ticket in the StreamSets Zendesk Support portal (no file size limit).

Upgrading to Version 3.12.x

You can upgrade previous versions of Data Collector to version 3.12.0. For complete instructions on upgrading, see the Upgrade documentation.

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 3.12.0

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

For the full list, click here.

JIRA	Description
SDC-13044	When stopping a large number of pipelines in bulk, Data Collector might skip sending a heartbeat status to Control Hub.
SDC-13043	Race conditions might cause Data Collector to skip sending pipeline status updates to Control Hub.
SDC-12937	During concurrent requests, the single sign-on service does not use the specified timeout.

SDC-12740, SDC-12642	The Azure Data Lake Storage Gen1 and Gen2 origins encounter performance issues due to excessive requests when scanning directories to read.
SDC-12633	The PostgreSQL CDC Client origin does not generate a new batch after reaching the maximum number of attempts to read new CDC data.
SDC-12620	The SQL Server Change Tracking Client origin cannot read tables that have reserved words in the table name.
SDC-12579	Byte order mark (BOM) characters in source files cause incorrect processing.
SDC-12324	Data Collector uses a non-compliant Jetty version.
SDC-8738	When the Kafka destination is configured to stop the pipeline on a record-level error and the destination encounters a stage exception, Data Collector incorrectly stops the pipeline instead of attempting to retry the pipeline.

Known Issues in Version 3.12.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-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.

Document revised on January 2, 2020