StreamSets Data Collector and Data Collector Edge 3.11.0 Release Notes

October 8, 2019

We're happy to announce new versions of StreamSets Data Collector and StreamSets Data Collector Edge. Version 3.11.0 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 3.11.x

Version 3.11.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:

- **Amazon S3** - The origin now generates event records when it starts processing a new object and when it finishes processing an object.
- **Azure Data Lake Storage Gen1** - The origin is no longer considered a Technology Preview feature and is approved for use in production.
- **Azure Data Lake Storage Gen2** - The origin is no longer considered a Technology Preview feature and is approved for use in production.
- **Google Big Query** - The origin now supports JSON service-account credentials pasted directly into the UI.
- **Google Cloud Storage** - The origin now supports JSON service-account credentials pasted directly into the UI.
- **Google Pub/Sub Subscriber** - The origin now supports JSON service-account credentials pasted directly into the UI.
- **HTTP Client** - The origin now supports time functions in the Resource URL property.

- **Kafka Consumer** - The origin can now be configured to save the Kafka message key in the record. The origin can save the key in a record header attribute, a record field, or both.
- **Kafka Multitopic Consumer** - The origin can now be configured to save the Kafka message key in the record. The origin can save the key in a record header attribute, a record field, or both.
- **Salesforce** - The origin has a new Mismatched Types Behavior property, which specifies how to handle fields with types that do not match the schema.
- **SFTP/FTP/FTPS Client** - The origin has three new timeout properties: Socket Timeout, Connection Timeout, and Data Timeout.

Processors

This release includes enhancements to the following processors:

- **Field Type Converter** - The processor can now convert to the Zoned Datetime data type from the Datetime data type or the Date data type.
- **Groovy Evaluator** - The processor now supports the use of the `sdc` wrapper object to access the constants, methods, and objects available to each script type.
- **HTTP Client** - When responses to requests contain multiple values, the processor can now return the first matching value, all matching values in a list in a single record, or all matching values in separate records.
- **JavaScript Evaluator** - The processor now supports the use of the `sdc` wrapper object to access the constants, methods, and objects available to each script type.
- **Jython Evaluator** - The processor now supports the use of the `sdc` wrapper object to access the constants, methods, and objects available to each script type.

Destinations

This release includes enhancements to the following destinations:

- **Azure Data Lake Storage Gen1** - The destination is no longer considered a Technology Preview feature and is approved for use in production.
- **Azure Data Lake Storage Gen2** - The destination is no longer considered a Technology Preview feature and is approved for use in production.
- **Cassandra** - The destination has new properties to disable batches and to set a timeout for individual write requests.
- **Google Big Query** - The destination now supports JSON service-account credentials pasted directly into the UI.
- **Google Cloud Storage** - The destination now supports JSON service-account credentials pasted directly into the UI.
- **Google Pub/Sub Subscriber** - The destination now supports JSON service-account credentials pasted directly into the UI.

- **HTTP Client** - The destination now supports time functions in the Resource URL property.
- **Kafka Producer** - The destination can now read the Kafka message key stored in a record header. On the Data Format tab, you configure the expected format of the key.
- **Salesforce** - The destination now writes data to Salesforce objects by matching case-sensitive field names. You can override the default field mappings by continuing to define specific mappings.
- **SFTP/FTP/FTPS Client** - The destination has three new timeout properties: Socket Timeout, Connection Timeout, and Data Timeout.
- **Solr destination** - The destination has two new timeout properties: Connection Timeout and Socket Timeout.

Executors

This release includes enhancements to the following executors:

- **ADLS Gen1 File Metadata** - The executor is no longer considered a Technology Preview feature and is approved for use in production.
- **ADLS Gen2 File Metadata** - The executor is no longer considered a Technology Preview feature and is approved for use in production.
- **JDBC Query** - The executor can now generate events that you can use in an event stream. You can configure the executor to include the number of rows returned or affected by the query when generating events.
- **Spark** - The executor now includes the following:
    - Additional fields in generated event records to store the user who submitted the job and the time that the job started.
    - Additional JARs property for applications written in Python.

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 newly available in this release:

- **Cron Scheduler origin** - Generates a record with the current datetime as scheduled by a cron expression.
- **Start Pipeline origin** - Starts a Data Collector, Data Collector Edge, or Transformer pipeline.

- **Control Hub API processor** - Calls a Control Hub API.
- **Start Job processor** - Starts a Control Hub job.
- **Start Pipeline processor** - Starts a Data Collector, Data Collector Edge, or Transformer pipeline.

Pipelines

This release includes the following pipeline enhancement:

- You can now configure pipelines to write error records to Amazon S3.

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.hive.impersonate.current.user`. You can set the property to `true` to enable the Hive Metadata processor, the Hive Metastore destination, and the Hive Query executor to impersonate the current user when connecting to Hive.

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_3-lib	For the Cloudera CDH version 6.3 distribution of Apache Hadoop.
streamsets-datacollector-orchestrator-lib	For the orchestrator stages.

- **Updated stage libraries** - This release includes updates to the following stage library:

Stage Library Name	Description
streamsets-datacollector-hdp_3_1-lib	For Hortonworks 3.1, the library now includes two additional stages: • Spark Evaluator processor • Spark executor

Deprecated Features in Version 3.11.x

Version 3.11.x newly deprecates the following features:

- **Azure Data Lake Storage (Legacy) destination** - This destination is now deprecated and will be removed in a future release. StreamSets recommends using the Azure Data Lake Storage Gen1 destination to write data to Microsoft Azure Data Lake Storage Gen1.
- **Scripting processor object and names** - The `sdcFunctions` object and certain names used to access methods that evaluate or modify data are deprecated in the Groovy Evaluator, JavaScript Evaluator, and Jython Evaluator processors. The object and names will be removed in a future release. Instead, StreamSets recommends using the `sdc` wrapper object to access the same methods.

Upgrading to Version 3.11.x

You can upgrade previous versions of Data Collector to version 3.11.0. 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 3.11.0

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

For the full list, click here.

JIRA	Description
SDC-8697	Starting multiple pipelines concurrently that run a Jython import call can lead to retry errors and cause some of the pipelines to fail.

Known Issues in Version 3.11.0

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-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 October 16, 2019