StreamSets Data Collector 3.5.0 Release Notes

October 1, 2018

We're happy to announce a new version of StreamSets Data Collector. This version contains several new features, enhancements, and some important bug fixes.

This document contains important information about the following topics for this release:

Upgrading to Version 3.5.0

You can upgrade previous versions of Data Collector to version 3.5.0. For complete instructions on upgrading, see the Upgrade Documentation.

Update Spark Executor with Databricks Pipelines

With version 3.5.0, Data Collector introduces a new Databricks executor and has removed the ability to use the Spark executor with Databricks.

If you upgrade pipelines that include the Spark executor with Databricks, you must update the pipeline to use the Databricks executor after you upgrade.

New Features and Enhancements

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

Origins

- **New Pulsar Consumer origin** - A new origin that reads messages from one or more topics in an Apache Pulsar cluster.
- **JDBC Multitable Consumer and JDBC Query Consumer origin enhancements** - These origins now include an option to convert timestamp data to the String data type instead of to the Datetime data type to ensure that the precision is maintained.
- **Salesforce origin enhancement** - When using the Bulk API, the origin can now execute an SOQL query that includes one or more subqueries.
- **WebSocket Client and WebSocket Server origin enhancement** - When included in a microservice pipeline, the origins can now send responses back to the originating REST API client when used with destinations that send records to the origin in the same microservice pipeline.

Processors

- **New Encrypt and Decrypt Fields processor** - A new processor that encrypts or decrypts individual field values.

- **New MongoDB Lookup processor** - A new processor that performs lookups in MongoDB and passes all values from the returned document to a new list-map field. Use the MongoDB Lookup to enrich records with additional data.

- **New HTTP Router processor** - A new processor that passes records to streams based on the HTTP method and URL path in record header attributes. Use the HTTP Router processor in pipelines with an origin that creates HTTP method and path record header attributes - including the HTTP Server origin and the REST Service origin.

- **Field Type Converter processor enhancement** - The processor can now convert the Boolean data type to the Integer, Long, or Short data type.

- **Salesforce Lookup processor enhancements** - The processor includes the following enhancements:
    - The processor can now return multiple values. You can configure the lookup to return the first value or to return all matches as separate records.
    - You can now configure how the processor handles a lookup that returns no value in fields with no default value defined. Upgraded pipelines continue to send records with no return value and no default value to error.

Destinations

- **New Pulsar Producer destination** - A new destination that writes data to topics in an Apache Pulsar cluster.

- **New Syslog destination** - A new destination that writes data to a Syslog server.

- **HTTP Client, Kafka Producer, and Kinesis Producer destination enhancement** - When included in a microservice pipeline, the destinations can now send records to the origin in the microservice pipeline with the specified response.

Executors

- **New Databricks executor** - A new executor that starts a Databricks job each time it receives an event.

  With the addition of this new executor, Data Collector has removed the ability to use the Spark executor with Databricks. If you upgrade pipelines that include the Spark executor with Databricks, you must update the pipeline to use the Databricks executor after you upgrade.

Hive Stages

- **JDBC Credentials** - The following Hive stages now allow you to enter credentials separately from the JDBC URL for Hive:
    - Hive Metadata processor
    - Hive Metastore destination
    - Hive Query executor

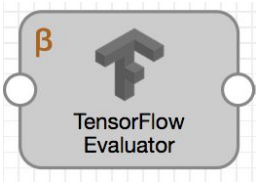
Salesforce Stages

- **API version** - Data Collector now ships with version 43.0 of the Salesforce Web Services Connector libraries used by the following Salesforce stages:
  - Salesforce origin
  - Salesforce Lookup processor
  - Einstein Analytics destination
  - Salesforce destination

Technology Preview Functionality

Data Collector now 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 display a Technology Preview icon on the upper left corner of the stage, as follows:



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:

- **Databricks ML Evaluator processor** - A new processor that uses Spark-trained machine learning models to generate evaluations, scoring, or classifications of data.
- **MLeap Evaluator processor** - A new processor that uses machine learning models stored in MLeap format to generate evaluations, scoring, or classifications of data.
- **PMML Evaluator processor** - A new processor that uses a machine learning model stored in the Predictive Model Markup Language (PMML) format to generate evaluations, scoring, or classifications of data.
- **TensorFlow Evaluator processor** - A new processor that uses TensorFlow machine learning models to generate predictions or classifications of data.

Data Formats

- **Delimited data format enhancement** - When reading delimited data that contains headers with empty values, Data Collector now replaces the empty values with the string “empty-” plus the column number starting from zero. For example, if the 3rd column header is empty, then the field name in Data Collector becomes “empty-2”. Previously, Data Collector retained the empty field name.
- **Excel data format enhancement** - When reading Excel data, Data Collector now processes the underlying raw values for numeric columns in a spreadsheet, rather than the displayed values. For example, if a cell contains 3.14159 but the display format is set to 2 decimals such that the spreadsheet displays 3.14, Data Collector still processes the full value of 3.14159.

Previously, Data Collector encountered errors when processing an Excel spreadsheet that contained displayed values.

Data Collector Edge (SDC Edge)

- **Download an installer for Windows** - You can now download a Microsoft installer to install SDC Edge on a Windows operating system.
- **Run SDC Edge as a service** - You can now register SDC Edge to run as a system service on Darwin, Linux, or Windows operating systems.
- **System Metrics origin enhancement** - The origin can now read metrics from specific processes running on the edge device.
- **Windows Event Log origin enhancement** - The origin can now read from a custom Windows log.
- **Dev Data Generator origin supported** - Edge pipelines now support the Dev Data Generator origin.
- **TensorFlow Evaluator processor supported** - Edge pipelines support the new TensorFlow Evaluator processor.
- **Functions** - Edge pipelines now support all job functions and the pipeline:startTime() function.
- **Disable the ability to manage production edge pipelines** - By default, you can use the Data Collector UI or REST API to manage edge pipelines deployed to an SDC Edge - including previewing, validating, starting, stopping, resetting the origin, and monitoring the pipelines. You can now disable the ability to manage edge pipelines in a production environment using the Data Collector UI or REST API. When disabled, you manage edge pipelines using Control Hub.
- **Skip verifying trusted certificates from a Control Hub on-premises installation** - If working with a Control Hub on-premises installation enabled for HTTPS in a test or development environment, you can now configure SDC Edge to skip verifying the Control Hub trusted certificates. StreamSets highly recommends that you configure SDC Edge to verify trusted certificates in a production environment.

Working with Control Hub

- **Automate registering and unregistering Data Collectors** - You can now use an automation tool such as Ansible, Chef, or Puppet to automate the registering and unregistering of Data Collectors using the following commands:

```
streamsets sch register
streamsets sch unregister
```

Microservice Pipelines

- **Origins for microservice pipelines** - The following origins can now send responses back to the originating REST API client when used with destinations that send records to the origin in the same microservice pipeline:
  - WebSocket Client origin

    - WebSocket Server origin

- **Destinations for microservice pipelines** - The following destinations can now send records to the origin in the microservice pipeline with the specified response:
    - HTTP Client destination
    - Kafka Producer destination
    - Kinesis Producer destination

- **Sample microservice pipeline** - When you create a microservice pipeline, the sample microservice pipeline now includes the new HTTP Router processor instead of the Stream Selector processor to route data to different streams based on the request method.

Data Governance Tools

- **Supported stages** - Data Collector can now publish metadata to data governance tools for the following stages:
    - Amazon S3 origin
    - Kafka Multitopic Consumer origin
    - SFTP/FTP Client origin
    - Kafka Producer destination

- **Cloudera Navigator versions** - Data Collector can now publish metadata to Cloudera Navigator running on Cloudera Manager versions 5.10 to 5.15.

  Previously, publishing metadata to Cloudera Navigator was supported only on Cloudera Manager version 5.10 or 5.11.

- **Secure connections to Cloudera Navigator** - If Cloudera Navigator is configured for TLS/SSL, Data Collector requires a local truststore file to verify the identity of the Cloudera Navigator Metadata Server. You now configure the truststore file location and password in the `$SDC_CONF/sdc.properties` file when you configure the connection to Cloudera Navigator.

Credential Stores

- **New Microsoft Azure Key Vault credential store** - You can now define credentials in Microsoft Azure Key Vault and then use the Data Collector credential functions in stage properties to retrieve those values.

- **Commands for a Java keystore credential store** - You now use the `stagelib-cli jks-credentialstore` command to add, list, and delete credentials in a Java keystore credential store. Previously you used the `jks-cs` command, which has now been deprecated.

Expression Language

- **String functions** - This release includes the following new function:
    - str:split() - Splits a string into a list of string values.

- **Pipeline functions** - This release includes the following new function:
    - pipeline:startTime() - Returns the start time of the pipeline as a Datetime value.

- **Job functions** - This release includes the following new functions:
    - job:id() - Returns the ID of the job if the pipeline was run from a Control Hub job. Otherwise, returns “UNDEFINED”.

- job:name() - Returns the name of the job if the pipeline was run from a Control Hub job. Otherwise, returns “UNDEFINED”.
- job:startTime() - Returns the start time of the job if the pipeline was run from a Control Hub job. Otherwise, returns the start time of the pipeline.
- job:user() - Returns the user who started the job if the pipeline was run from a Control Hub job. Otherwise, returns “UNDEFINED”.

Stage Libraries

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

Stage Library Name	Description
streamsets-datacollector-apache-kafka_1_1-lib	Apache Kafka version 1.1.x
streamsets-datacollector-apache-kafka_2_0-lib	Apache Kafka version 2.0.x
streamsets-datacollector-apache-pulsar_2-lib	Apache Pulsar version 2.1.0-incubating
streamsets-datacollector-azure-keyvault-credentialstore-lib	Microsoft Azure Key Vault credential store system
streamsets-datacollector-cdh_6_0-lib	Cloudera CDH version 6.0 distribution of Apache Hadoop **Note:** Does not include the following stages: ● HBase Lookup processor ● Spark Evaluator processor ● HBase destination
streamsets-datacollector-crypto-lib	For cryptography stages, including the Encrypt and Decrypt Fields processor
streamsets-datacollector-databricks-ml_2-lib	Databricks ML
streamsets-datacollector-mapr_6_0-mep5-lib	MapR Ecosystem Pack (MEP) version 5 for MapR 6.0.1
streamsets-datacollector-mleap-lib	MLeap
streamsets-datacollector-tensorflow-lib	TensorFlow

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

Stage Library Name	Description
streamsets-datacollector-apache-kafka_0_9-lib	Apache Kafka version 0.9.x
streamsets-datacollector-apache-kafka_0_10-lib	Apache Kafka version 0.10.x
streamsets-datacollector-cdh_5_8-lib	Cloudera CDH version 5.8 distribution of

	Apache Hadoop
streamsets-datacollector-cdh_5_9-lib	Cloudera CDH version 5.9 distribution of Apache Hadoop
streamsets-datacollector-cdh_kafka_2_0-lib	Cloudera distribution of Apache Kafka 2.0.x (0.9.0)
streamsets-datacollector-hdp_2_4-lib	Hortonworks version 2.4 distribution of Apache Hadoop
streamsets-datacollector-hdp_2_4-hive1-lib	Hortonworks version 2.4.x distribution of Apache Hive version 1.x
streamsets-datacollector-hdp_2_5-lib	Hortonworks version 2.5.x distribution of Apache Hadoop
streamsets-datacollector-hdp_2_5-flume-lib	Hortonworks version 2.5.x distribution of Apache Flume
streamsets-datacollector-mapr_5_1-lib	MapR version 5.1

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.

Miscellaneous

- **Import pipelines from an external HTTP URL** - You can now use Data Collector to import pipelines from an external HTTP URL. For example, you can import pipelines from the StreamSets GitHub repository.

- **Collection of usage statistics** - When you log in to Data Collector as the admin/admin user for the first time, you can now choose to improve Data Collector by sending anonymized usage data. Previously, the `ui.enable.usage.data.collection` property in the Data Collector configuration file determined whether usage data was collected. This property has been removed.

Fixed Issues

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

For the full list, click here.

JIRA	Description

SDC-10091	When using timestamp ordering, the Directory origin can encounter errors when checking for new files or can fail to completely process files.
SDC-10062	The JDBC Producer destination encounters a JDBC_23 error when field is string but the database column is not string.
SDC-9763	When a pipeline with a Hadoop FS Standalone origin runs from a Control Hub job, the pipeline fails with a null pointer exception.
SDC-9722	The JDBC Producer destination used with the MySQL Connector/J 8.0.X driver throws an unknown table error and does not write any data to the database.
SDC-9679	The URL for a state notification webhook cannot evaluate expression language functions.
SDC-9673	The Data Collector UI displays logs for unrelated pipelines when remote pipelines are run from Control Hub jobs.
SDC-9643	When Data Collector is enabled for LDAP authentication, you cannot use the API to export a pipeline as a user that belongs to an LDAP group with the admin privilege.
SDC-9480, SDC-8567	The SFTP/FTP Client origin encounters a null pointer exception if you specify an absolute path to the resource URL but select the Path Relative to User Home Directory property.
SDC-9444	The pipelineStateHistory.json is not closed before the pipeline is deleted.
SDC-9381	The Password property for a webhook does not allow the use of credential functions.
SDC-6371	Runtime properties are not evaluated in cluster streaming pipelines.

Known Issues

Please note the following known issues with this release.

For a full list of known issues, check out our JIRA.

JIRA	Description
SDC-10037	Origins that can read the Excel data format fail to read Microsoft Excel files with a large number of rows.
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-9798	When the SFTP/FTP Client origin fails to connect to the server, the pipeline silently fails - the pipeline keeps running but no longer processes data.
SDC-9514	Runtime parameters are not supported in all configuration properties in cluster batch execution mode, such as Max Batch Size.
SDC-8893	The Field Type Converter processor performs slower than it did in version 2.5.1.1.
SDC-8876	When you configure a metric rule with a meter metric type and a 1 minute rate for the metric element, the rule does not wait until 1 minute passes before evaluating the condition and sending the alert. Instead, the alert is sent when the pipeline starts and stops.
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-8598	Upon starting, Data Collector writes the following messages to the log file about the Cloudera CDH 5.14 stage library:
	`The following stages have invalid classpath: cdh_5_14_lib`
	`Detected colliding dependency versions: <additional information>`
	These messages are written in error and can be safely ignored.

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-5357	The Jython Evaluator and stages that can access the file system - such as the Directory and File Tail origins - can access all pipeline JSON files stored in the $SDC_DATA directory. This allows users to access pipelines that they might not have permission to access within Data Collector. Workaround: To secure your pipelines, complete the following tasks: 1. Remove the Jython stage library and use the Groovy Evaluator or JavaScript Evaluator processor instead of the Jython Evaluator. 2. Update the Data Collector security policy file, `$SDC_CONF/sdc-security.policy`, so that Data Collector stages do not have `AllPermission` access to the file system. Update the security policy for the following code bases: streamsets-libs-extras, streamsets-libs, and streamsets-datacollector-dev-lib. Use the policy file syntax to set the security policies.
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.