StreamSets Data Collector 3.4.0 Release Notes

July 30, 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 3.4.0

You can upgrade previous versions of Data Collector to version 3.4.0. For complete instructions on upgrading, see the Upgrade Documentation.

New Features and Enhancements

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

Origins

- **New PostgreSQL CDC Client origin** - Use the PostgreSQL CDC Client origin to process change data capture information for a PostgreSQL database.

- **New Test origin** - You can now configure a virtual test origin to provide test data for data preview to aid in pipeline development. In Control Hub, you can also use test origins when developing pipeline fragments.

- **Amazon S3, Directory, SFTP/FTP Client, Google Cloud Storage enhancements** - The listed origins can now process Microsoft Excel files.

- **Dev Data Generator origin enhancement** - The development origin can now generate additional types of data for testing purposes - such as sample address data, names, or prices.

- **Hadoop FS origin enhancements** - The origin includes the following enhancements:
  - Process Amazon S3 data in cluster EMR batch mode - Use the origin in a cluster EMR batch pipeline that runs on an Amazon EMR cluster to process data from Amazon S3.
  - Process Amazon S3 data in cluster batch mode - Use the origin in a cluster batch pipeline that runs on a Cloudera distribution of Hadoop (CDH) or Hortonworks Data Platform (HDP) cluster to process data from Amazon S3.

- **HTTP Client origin enhancements** - The origin includes the following changes and enhancements:
    - The origin now uses buffered request transfer encoding by default. Upgraded pipelines retain their previous configuration.
    - HEAD request responses create an empty record. Information returned from the HEAD appear in record header attributes.

- **HTTP Server origin enhancement** - The origin now includes the name of the the client or proxy that made the request in the `remoteHost` record header attribute.

- **MongoDB origin enhancement** - You can now use a date field as the offset field.

- **Oracle CDC Client origin enhancements** - The origin includes the following changes and enhancements:
    - Multithreaded parsing - When using local caching and parsing the SQL query, the origin can now use multiple threads to parse transactions.
    - PEG Parser - To improve performance for very wide tables, you can try our experimental PEG parser.
    - With this release, the Query Timeout property has been removed. You can no longer configure a query to timeout before the end of a LogMiner session. The existing LogMiner Session Window property defines how long the session lasts.

- **Salesforce origin enhancement** - When using the SOAP API, the origin can now execute an SOQL query that includes one or more subqueries. Support for subqueries using the Bulk API will be added in a future release.

Processors

- **New Whole File Transformer processor** - Use the Whole File Transformer processor to convert fully written Avro files to Parquet in a whole file pipeline.

- **Field Hasher processor enhancement** - The processor can now add a user-defined field separator to fields before hashing.

- **HTTP Client processor enhancements** - The processor includes the following changes and enhancements:
    - The processor now uses buffered request transfer encoding by default. Upgraded pipelines retain their previous configuration.
    - HEAD request responses create an empty record. Information returned from the HEAD appear in record header attributes.
    - The resolved request URL is now written to the Data Collector log when Data Collector logging is set to debug or higher.

- **JDBC Lookup processor enhancement** - When using local caching, the processor can now use additional cores to prepopulate the cache to enhance pipeline performance.

Destinations

- **New Couchbase destination** - A new destination that writes data to Couchbase.

- **New Splunk destination** - A new destination that writes data to Splunk using the Splunk HTTP Event Collector (HEC).

- **Cassandra destination enhancement** - You can now use SSL/TLS to connect to Cassandra.

- **HTTP Client destination enhancement** - The destination now uses buffered request transfer encoding by default. Upgraded pipelines retain their previous configuration.

Executors

- **Amazon S3 executor enhancements** - The executor includes the following enhancements:
    - The executor can now copy objects to a new location and optionally delete the original object.
    - The executor can now generate event records each time the executor creates a new object, adds tags to an existing object, or completes copying an object to a new location.

Data Collector Edge (SDC Edge)

- **New System Metrics origin** - A new origin that reads system metrics - such as CPU and memory usage - from the edge device where SDC Edge is installed.

- **HTTP Client origin supported** - Edge sending pipelines now support the HTTP Client origin. However, the origin does not currently support batch processing mode, pagination, or OAuth2 authorization in edge pipelines.

- **WebSocket Client origin supported** - Edge sending pipelines now support the WebSocket Client origin.

- **Pipeline functions** - Edge pipelines now support the following pipeline functions:
    - pipeline:id()
    - pipeline:title()
    - pipeline:user()

- **Preview and validate edge pipelines** **-** You can now use the Data Collector UI or the command line and REST API to preview and validate edge pipelines.

- **Publish multiple edge pipelines to SDC Edge** - You can now use the Data Collector Home page to directly publish multiple edge pipelines at one time to an SDC Edge that is running. Previously, you could only publish a single edge pipeline at a time.

- **Download edge pipelines from SDC Edge** - You can now use the Data Collector UI to download all edge pipelines deployed to an SDC Edge in addition to all sample edge pipelines included with SDC Edge.

- **Filter the Home page by edge pipelines** - You can now select Edge Pipelines as a category on the Data Collector Home page to view all available edge pipelines.

Microservice Pipelines

You can now create microservices using microservice pipelines. Use the following new stages in microservice pipelines:

- New REST Service origin - Listens on an HTTP endpoint, parses the contents of all authorized requests, and sends responses back to the originating REST API. Creates multiple threads to enable parallel processing in a multithreaded pipeline.

- Send Response to Origin destination - Sends records to the REST Service origin with the specified response.

Pipelines

- **Notifications** - You can now configure a pipeline to send an email or webhook when the pipeline changes to the Running_Error state.

- **Error records** - Error records now include an errorJobID internal header attribute when the pipeline that generated the error record was started by a Control Hub job.

- **Install external libraries from the properties panel** - You can now select a stage in the pipeline canvas and then install external libraries for that stage from the properties panel. Previously, you had to navigate to the Package Manager page to install external libraries.

Cluster Pipelines

- **New cluster EMR batch mode** - Data Collector can now use the cluster EMR batch mode to run on an Amazon EMR cluster to process data from Amazon S3. Data Collector runs as an application on top of MapReduce in the EMR cluster.

  Data Collector can run on an existing EMR cluster or on a new EMR cluster that is provisioned when the cluster pipeline starts. When you provision a new EMR cluster, you can configure whether the cluster remains active or terminates when the pipeline stops.

  Use the Hadoop FS origin to process data from Amazon S3 in cluster EMR batch mode.

- **Logs** - You can now configure the Data Collector on the master gateway node to use the log4j rolling file appender to write log messages to an sdc.log file. This configuration is propagated to the worker nodes such that each Data Collector worker writes log messages to an sdc.log file within the YARN application directory.

Data Formats

- **New Excel data format** - You can now use the following file-based origins to process Microsoft Excel files:
  - Amazon S3 origin
  - Directory origin
  - Google Cloud Storage origin
  - SFTP/FTP Client origin

- **Avro and Protobuf data formats** - To preserve the ordering of fields, the Avro and Protobuf data formats now use the list-map root field type instead of the map root field type.

Stage Libraries

This version of Data Collector includes the following new stage libraries:

- **streamsets-datacollector-cdh_5_15-lib** - The Cloudera CDH 5.15 distribution of Hadoop.

- **streamsets-datacollector-emr_hadoop_2_8_3-lib** - Includes the Hadoop FS origin for cluster EMR batch mode pipelines that run on an Amazon EMR cluster to process data from Amazon S3.

Miscellaneous

- **Cloudera Manager CSD enhancement** - The Cloudera Manager CSD now enables specifying a StreamSets Customer ID, used when generating support bundles. The customer ID is generated by the StreamSets Support team for users with a paid subscription.

- **Postgres changes** - Postgres CSV and Postgres Text delimited data format types are now known as PostgreSQL CSV and PostgreSQL Text, respectively. The Postgres Metadata processor is now known as the PostgreSQL Metadata processor. And the Drift Synchronization Solution for Postgres is now known as the Drift Synchronization Solution for PostgreSQL.

- **Documentation enhancement** - The online help has a new look and feel. All of the previous documentation remains exactly where you expect it, but it is now easier to view and navigate on smaller devices like your tablet or mobile phone.

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-9561	A Control Hub pipeline that is set up to aggregate statistics with an external system and that includes pipeline fragments can cause a null pointer exception in the system pipeline.
SDC-9401	When configured to delete files after processing, the Directory origin does not delete files in subdirectories.
SDC-9397	Control Hub pipelines need a default statistics aggregation option.
SDC-9303	Data Collector displays sensitive information such as passwords entered in the sdc.properties file.
SDC-9275	The Directory origin does not process files when the specified first file to process is empty and the directory path is lexically larger than the file path.
SDC-9401	The Oracle CDC Client origin stores the offset based on the timestamp of the last record, which can cause an error when restarting the pipeline when the logstash retention interval is shorter than the start time from the last record.
SDC-9185	The HTTP Client origin does not support the following types of OAuth 2 tokens: RFC-6749, Salesforce, and Google JWT.

SDC-9177	When configured to use a long batch wait time, the Directory origin waits the specified amount of time before processing a newly arrived file.
SDC-9132	JDBC destination cannot write to columns with the NUMERIC data type.
SDC-8959	The Hadoop FS Standalone origin does not successfully process the whole file data format.
SDC-8731	Pipelines with the Google Pub/Sub Subscriber origin might hang and display the following error in the Data Collector log: `io.grpc.StatusRuntimeException: UNAVAILABLE: The service was unable to fulfill your request. Please try again. [code=8a75]`
SDC-8680	The Azure Data Lake Store destination does not properly roll files based on the “roll” record header attribute.
SDC-8660	The MapR Streams destination cannot write to topics that are auto-created in MapR Streams when the destination is enabled for runtime resolution. Workaround: To write to an auto-created topic, disable runtime resolution.
SDC-7117	Data preview produces an error when data does not become available within the configured preview timeout window.

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-9514	Runtime parameters are not supported in all configuration properties in cluster batch execution mode, such as Max Batch Size.
SDC-9430	The HTTP Client stages encounter null pointer exceptions when receiving records that are not HEAD requests that contain no response body.
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.