StreamSets Data Collector 3.2.0.0 Release Notes

April 25, 2018

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

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

Upgrading to Version 3.2.0.0

You can upgrade previous versions of Data Collector to version 3.2.0.0. For complete instructions on upgrading, see the Upgrade Documentation.

Upgrade to Spark 2.1 or Later

When Data Collector first introduced cluster streaming mode with Spark 1.3 and Kafka 0.8, Kafka didn't support security features such as SSL/TLS and Kerberos authentication. With Spark 2.1, Data Collector introduced updated Kafka stages with support for these security features when used with Kafka 0.10 or later.

In Data Collector version 3.3.0.0, we will be introducing cluster streaming mode with support for Kafka security features using Spark 2.1 or later and Kafka 0.10 or later.

However, this means that cluster streaming mode and the Spark Evaluator processor using Spark 1.x are deprecated as of version 3.2.0.0. Support for Spark 1.x will be removed in version 3.3.0.0. If you are using cluster streaming mode or the Spark Evaluator processor, we highly recommend that you upgrade to Spark 2.1 or later in version 3.2.0.0.

Note: In Data Collector version 3.3.0.0, you will be able to use Spark 2.1 and Kafka 0.9.0.0 for cluster streaming mode. However, you cannot use Kafka security features unless you also upgrade to Kafka 0.10 or later.

Since Spark 1.x is now deprecated, the following stage libraries are also deprecated since they only support Spark 1.x:

- Cloudera CDH version 5.5 or earlier distribution of Apache Hadoop
- Hortonworks version 2.4 or earlier distribution of Apache Hadoop

These stage libraries will also be removed in version 3.3.0.0. We highly recommend that you upgrade to a newer Cloudera CDH or Hortonworks Hadoop distribution in version 3.2.0.0.

The major Hadoop distribution vendors provide a means for Spark 1.x and Spark 2.x to coexist on the same cluster, so you can use both versions in your clusters.

Data Collector supports the following Spark 2.x versions for the Hadoop distribution vendors:

- **Cloudera** - Cloudera Distribution of Spark 2.1 release 1 or later is supported. For more information, see Spark 2 Requirements.
- **Hortonworks** - Hortonworks Data Platform (HDP) 2.6 or later includes Spark 2.2.0. For more information, see the HDP 2.6 Release Notes.
- **MapR** - MapR with MapR Expansion Pack 3.0 or later is supported. For more information, see MEP Support by MapR Core Version.

New Features and Enhancements

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

Origins

- **New Hadoop FS Standalone origin** - Similar to the Directory origin, the Hadoop FS Standalone origin can use multiple threads to read fully-written files. Use this origin in standalone execution mode pipelines to read files in HDFS.

- **New MapR FS Standalone origin** - Similar to the Directory origin, the MapR FS Standalone origin can use multiple threads to read fully-written files. Use this origin in standalone execution mode pipelines to read files in MapR FS.

- **New Dev Snapshot Replaying origin** - The Dev Snapshot Replaying origin is a development stage that reads records from a downloaded snapshot file.

- **HTTP Client origin enhancement** - You can now configure the origin to process JSON files that include multiple JSON objects or a single JSON array.

- **JDBC Multitable Consumer origin enhancement** - The origin can now generate table-finished and schema-finished events when it completes processing all rows in a table or schema. You can also configure the number of seconds that the origin waits before generating the no-more-data event. You might want to configure a delay if you want the table-finished or schema-finished events to appear in the event stream before the no-more-data event.

- **Oracle CDC Client origin enhancements** - The origin includes the following enhancements:
  - You can set a new Parse SQL Query property to false to skip parsing the SQL queries. Instead, the origin writes the SQL query to a “sql” field that can be parsed later. Default is true, which retains the previous behavior of parsing the SQL queries.

  - The Send Redo Query property has been renamed. The new name is Send Redo Query in Headers.

- **TCP Server origin enhancement** - You can now use the origin to read Flume Avro data.

Processors

- **HTTP Client processor enhancement** - You can now use the PATCH method with the processor.

- **JDBC Lookup processor enhancement** - The Retry on Cache Miss property has been renamed to Retry on Missing Value.

- **Kudu Lookup processor enhancement** - You can now configure the processor behavior when a lookup returns no value.

Destinations

- **Hadoop FS destination enhancement** - The destination now supports writing records using the SDC Record format.

- **HTTP Client destination enhancement** - You can now use the PATCH method with the destination.

Executors

- **MapReduce executor enhancement** - You can now use the new Avro to ORC job to convert Avro files to ORC files.

Data Collector Edge (SDC Edge)

SDC Edge includes the following enhancements:

- **JavaScript Evaluator processor supported** - Both edge sending pipelines and edge receiving pipelines now support the JavaScript Evaluator processor.

- **Publish edge pipelines to SDC Edge** - You can now use the Data Collector UI to directly publish edge pipelines to an SDC Edge that is running. Previously, you had to first export edge pipelines from Data Collector, and then move them to the SDC Edge installed on the edge device.

- **Manage edge pipelines from the Data Collector UI** - You can now use the Data Collector UI to start, monitor, stop, and reset the origin for edge pipelines running on a remote SDC Edge. Previously, you had to use the command line and REST API to manage edge pipelines on SDC Edge.

Miscellaneous

- **Pipeline error handling enhancement** - You can now configure pipelines to write error records to Azure Event Hub.

- **Pipeline runner idle time enhancement** - You can configure the number of seconds that a pipeline runner waits before sending an empty batch.

- **Runtime statistics enhancement** - Runtime statistics now include the number of empty or idle batches that are generated by the pipeline.

- **Snapshot enhancement** - Snapshots now include record header attributes for error records. Previously, snapshots included only the record fields in an error record.

Stage Libraries

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

- Apache Kudu version 1.7

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-8847	Kinesis Consumer origin continues using threads after the pipeline is stopped.
SDC-8760	If a remote pipeline instance fails to start on a registered Data Collector due to an unexpected runtime error, the Data Collector might not consume any more commands from Control Hub.
SDC-8706	Control Hub cannot force stop a remote pipeline instance.
SDC-8656	On certain occasions, the Directory origin can add files to the processing queue multiple times.
SDC-8628	Multithreaded pipeline fails when it contains a Directory origin that reads whole files.
SDC-8592	Logging out from a Data Collector configured for LDAP authentication causes a null pointer exception when the forceBindingLogin LDAP property is set to true.
SDC-8508	Data Collector doesn't display the value of the STREAMSETS_LIBRARIES_EXTRA_DIR environment variable at startup.
SDC-8491	The following line is displayed too often in the Data Collector log: `WARN  DirectorySpooler - File '<file name>.json' already in queue, ignoring`
SDC-8479	When cluster mode pipelines fail to start, YARN application logs are not included in the Data Collector log.
SDC-8388	User with the Manager role cannot see the Preview or Validate icons in the UI.
SDC-7852	Using Cloudera Manager to register Data Collector with Control Hub no longer works.

SDC-6549	When Control Hub stops a remote pipeline instance, false messages are logged at 'ERROR' level to the Data Collector log.
SDCE-232	Control Hub jobs cannot start remote pipeline instances on a registered SDC Edge on Windows.

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-8808	If a pipeline encounters parsing errors when converting the SDC Record data format to Avro due to incompatible data types or invalid values, the pipeline fails.
SDC-8793	For the Vault credential store, if you update the secret ID value in the file configured for the `credentialStore.vault.config.secret.id` property, Data Collector does not use the updated value for currently running pipelines. Workaround: Restart Data Collector after updating the secret ID in the file.
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-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-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-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-7903	Do not use Cluster Streaming pipelines with MapR and Spark 2.x.
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.