StreamSets Data Collector and Data Collector Edge Cumulative 3.6.x Release Notes

This document contains release information for the following versions of StreamSets Data Collector:

StreamSets Data Collector and Data Collector Edge 3.6.1 Release Notes

December 10, 2018

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

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

Upgrading to Version 3.6.1

You can upgrade previous versions of Data Collector to version 3.6.1. For complete instructions on upgrading, see the Upgrade Documentation.

Fixed Issues in 3.6.1

The following table lists the known issues that are fixed with this release:

JIRA	Description
SDC-10506	Pipelines run from Control Hub jobs fail to write aggregated statistics to SDC RPC.
SDC-10352	In Data Collector 3.5.0, the Salesforce origin did not allow non-String offset fields.

Known Issues in 3.6.1

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-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.

StreamSets Data Collector and Data Collector Edge 3.6.0 Release Notes

November 26, 2018

We're happy to announce a new version of StreamSets Data Collector and StreamSets Data Collector Edge. 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 3.6.0

You can upgrade previous versions of Data Collector to version 3.6.0. For complete instructions on upgrading, see the Upgrade Documentation.

New Features and Enhancements in 3.6.0

This version includes the following enhancements for Data Collector Edge (SDC Edge):

- **Register SDC Edge with Control Hub** - You can now use the command line to register SDC Edge with Control Hub.

- **Delimited data format** - Stages in edge pipelines can now process the delimited data format.
- **Functions** - The `sdc:hostname()` function can now return the host name of a Data Collector or SDC Edge machine and can be used within edge pipelines.

Fixed Issues in 3.6.0

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

For the full list, click here.

JIRA	Description
SDC-10424	When the Directory origin is configured to delete files post processing and another stage in the pipeline encounters an error that stops the pipeline, the origin might incorrectly remove the file that it is currently processing.

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-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 $SDC_DATA directory. This allows users to access pipelines that they might not have permission to access within Data Collector. Workaround: To secure your pipelines, complete the following tasks: 3. Remove the Jython stage library and use the Groovy Evaluator or JavaScript Evaluator processor instead of the Jython Evaluator. 4. Update the Data Collector security policy file, `$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.