StreamSets Data Collector
Cumulative 2.6.x.x Release Notes

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

****

StreamSets Data Collector 2.6.0.1 Release Notes

June 23, 2017

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

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

Upgrading to Version 2.6.0.1

You can upgrade previous versions of Data Collector to version 2.6.0.1. For complete instructions on upgrading, see the Upgrade Documentation.

Migrate to Java 8

As of version 2.5.0.0, Data Collector requires Java 8. If your previous Data Collector version ran on Java 7, you must migrate to Java 8 before upgrading to Data Collector version 2.6.0.0. For instructions, see Pre-Upgrade Tasks.

All services that use Data Collector JAR files also must run on Java 8. This means that your Hadoop cluster must run on Java 8 if you are using cluster pipelines, the Spark Executor, or the MapReduce Executor.

New Features and Enhancements

This version includes the following enhancement:

- **Kinesis Consumer origin** - You can now reset the origin for Kinesis Consumer pipelines. Resetting the origin for Kinesis Consumer differs from other origins, so please note the requirement and guidelines.

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-6452	Cluster streaming mode does not propagate all stage properties from the Kafka Consumer origin.
SDC-6446	The Hadoop FS, Local FS, and MapR FS destinations recover temporary files during data preview, generating file closure events that can be lost when using the event framework.
SDC-6439	When using the Salesforce Streaming API, the Salesforce origin can drop records when receiving large volumes of data.
SDC-6365	When the Hadoop FS, Local FS, or MapR FS destination encounters a stage error that causes it to fail to finish writing a whole file, it can change the name of the file from <filename>_tmp file to the final file name so that it looks like it is complete.
SDC-6360	In certain timeout conditions, the HTTP Client origin can enter a short loop and generate many stage errors before recovering.

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-6540	The Field Hasher appends a null character (0x00) after each field value when

	performing field-level hashing.
SDC-6521	The JDBC Tee processor cannot be used with Microsoft SQL Server databases.
SDC-6509	When using the Write to Another Pipeline pipeline error handling option in cluster batch pipeline, if the error handling pipeline encounters a problem and stops, the original pipeline stops with the message “Job has been finished” instead of indicating there was a problem with the error handling pipeline.
SDC-6438	The MapR distribution for Spark 2.x is not supported by cluster streaming pipelines.
SDC-6210	The show-vault-id command returns a NPE when Data Collector is installed on EC2 with IPv6 enabled. Workaround: If you can run Data Collector without IPv6, in the /etc/sysctl.conf file, disable IPv6 by configuring the disable IPv6 property as follows: `net.ipv6.conf.all.disable_ipv6 = 1`
SDC-6077	The Field Remover processor does not remove list fields from list-map data.
SDC-5871	Attempting to write a record to Kudu with a null primary key causes the pipeline to fail.
SDC-5758	Due to expected Kudu behavior, if Data Collector is configured to use Kerberos, but your Kudu installation does not, Data Collector cannot connect. This applies to Kudu version 1.2 and later. Workaround: If necessary, install a separate Data Collector that does not have Kerberos enabled.
SDC-5521	Kerberos-enabled pipelines that are valid using an Apache Kafka stage library can fail validation when using a Cloudera distribution of Kafka. The Apache Kafka client libraries have been modified by StreamSets to allow connectivity to Kerberized Kafka without requiring a JAAS configuration file. The Apache Kafka stage libraries (version 0.9 and greater) do not require a JAAS configuration file when enabling Kerberos. For CDH Kafka versions, a JAAS configuration is still required. Workaround: Include a JAAS configuration file on the classpath by modifying the `sdc-env.sh` or `sdcd-env.sh` file to include the following option in the SDC_JAVA_OPTS environment variable: `-Djava.security.auth.login.config=<path-to-jaas-config>` Or, you can use the matching version Apache Kafka stage library, instead of the Cloudera stage libraries.

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-5325	Cluster mode pipelines that read from a MapR cluster fail when the MapR cluster uses username/password login authentication. Workaround: On the Cluster tab for the pipeline, add the following Java property to the Worker Java Options property: `-Dmaprlogin.password.enabled=true`
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-5039	When you use the Hadoop FS origin to read files from all subdirectories, the origin cannot use the configured Hadoop FS User as a proxy user to read from HDFS. Workaround: If you need to use a proxy user to read from all subdirectories of the specified directories, set the HADOOP_PROXY_USER environment variable to the proxy user in libexec/_cluster-manager script, as follows: `export HADOOP_PROXY_USER = <proxy-user>`
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-3133	When you upgrade Data Collector from the RPM package, the environment configuration file `$SDC_DIST/libexec/sdcd-env.sh` is overwritten. Workaround: Back up the `sdcd-env.sh` file before you upgrade.
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, ask for help, or find out about our next meetup, check out our Community page:
https://streamsets.com/community/.

For general inquiries, email us at info@streamsets.com.

****

StreamSets Data Collector 2.6.0.0 Release Notes

June 9, 2017

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

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

Upgrading to Version 2.6.0.0

You can upgrade previous versions of Data Collector to version 2.6.0.0. For complete instructions on upgrading, see the Upgrade Documentation.

Migrate to Java 8

As of version 2.5.0.0, Data Collector requires Java 8. If your previous Data Collector version ran on Java 7, you must migrate to Java 8 before upgrading to Data Collector version 2.6.0.0. For instructions, see Pre-Upgrade Tasks.

All services that use Data Collector JAR files also must run on Java 8. This means that your Hadoop cluster must run on Java 8 if you are using cluster pipelines, the Spark Executor, or the MapReduce Executor.

New Features and Enhancements

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

Installation

- **MapR prerequisites** - You can now run the `setup-mapr` command in interactive or non-interactive mode. In interactive mode, the command prompts you for the MapR version and home directory. In non-interactive mode, you define the MapR version and home directory in environment variables before running the command.

Data Collector Configuration

- **New buffer size configuration** - You can now use a new `parser.limit` configuration property to increase the Data Collector parser buffer size. The parser buffer is used by the origin to process many data formats, including Delimited, JSON, and XML. The parser buffer size limits the size of the records that origins can process. The Data Collector parser buffer size is 1048576 bytes by default.

  For more information about how the default buffer size and `parser.limit` property affect the data format maximum record size properties, see "Maximum Record Size."

Stage Libraries

Data Collector now supports the following stage libraries:

- Hortonworks version 2.6 distribution of Apache Hadoop
- Cloudera distribution of Spark 2.1
- MapR distribution of Spark 2.1

Drift Synchronization Solution for Hive

- **Parquet support -** You can now use the Drift Synchronization Solution for Hive to generate Parquet files. Previously, the Data Synchronization Solution supported only Avro data. This enhancement includes the following updates:

    - **Hive Metadata processor data format property -** Use the new data format property to indicate the data format to use.
    - **Parquet support in the Hive Metastore destination -** The destination can now create and update Parquet tables in Hive. The destination no longer includes a data format property since that information is now configured in the Hive Metadata processor.

See the documentation for implementation details and a Parquet case study.

Multithreaded Pipelines

The multithreaded framework includes the following enhancements:

- **Origins for multithreaded pipelines** - You can now use the following origins to create multithreaded pipelines:
    - CoAP Server origin
    - TCP Server origin

- **Multithreaded origin icons** - The icons for multithreaded origins now include the following multithreaded indicator:

  For example, here's the updated Elasticsearch origin icon:

Dataflow Triggers

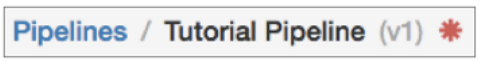
- **New executors** - You can now use the following executors to perform tasks upon receiving an event:
    - Email executor
    - Shell executor

Dataflow Performance Manager (DPM)

- **Pipeline statistics** - You can now configure a pipeline to write statistics directly to DPM. Write statistics directly to DPM when you run a job for the pipeline on a single Data Collector.

  When you run a job on multiple Data Collectors, a remote pipeline instance runs on each of the Data Collectors. To view aggregated statistics for the job within DPM, you must configure the pipeline to write the statistics to a Kafka cluster, Amazon Kinesis Streams, or SDC RPC.

- **Update published pipelines** - When you update a published pipeline, Data Collector now displays an asterisk next to the pipeline name to indicate that the pipeline has been updated since it was last published, as follows:

Origins

- **New CoAP Server origin** - An origin that listens on a CoAP endpoint and processes the contents of all authorized CoAP requests. The origin performs parallel processing and can generate multithreaded pipelines.

- **New TCP Server origin** - An origin that listens at the specified ports, establishes TCP sessions with clients that initiate TCP connections, and then processes the incoming data. The origin can process NetFlow, syslog, and most Data Collector data formats as separated records. You can configure custom acknowledgement messages and use a new batchSize variable, as well as other expressions, in the messages.

- **SFTP/FTP Client origin enhancement** - You can now specify the first file to process. This enables you to skip processing files with earlier timestamps.

Processors

- **Groovy, JavaScript, and Jython Evaluator processor enhancements**:
  - You can now include some methods of the sdcFunctions scripting object in the initialization and destroy scripts for the processors.
  - You can now use runtime parameters in the code developed for a Groovy Evaluator processor.

- **Hive Metadata processor enhancements:**
  - The Hive Metadata processor can now process Parquet data as part of the Drift Synchronization Solution for Hive.
  - You can now specify the data format to use: Avro or Parquet.
  - You can now configure an expression that defines comments for generated columns.

- **JDBC Lookup processor enhancements**:
  - The JDBC Lookup processor can now return multiple values. You can now configure the lookup to return the first value or to return all matches as separate records.
  - When you monitor a pipeline that includes the JDBC Lookup processor, you can now view stage statistics about the number of queries the processor makes and the average time of the queries.

- **Spark Evaluator processor enhancement** - The Spark Evaluator now supports Spark 2.x.

Destinations

- **New CoAP Client destination** - A destination that writes to a CoAP endpoint.

- **Hive Metastore destination enhancements**:
  - The destination can now create and update Parquet tables in Hive.
  - Also, the data format property has been removed. You now specify the data format in the Hive Metadata processor. Since the Hive Metastore previously supported only Avro data, there is no upgrade impact.

- **Kudu destination enhancement** - You can use the new Mutation Buffer Space property to set the buffer size that the Kudu client uses to write each batch.

Executors

- **New Email executor** - Use to send custom emails upon receiving an event. See the Dataflow Triggers chapter for a case study.

- **New Shell executor** - Use to execute shell scripts upon receiving an event.

- **JDBC Query executor enhancement** - A new Batch Commit property allows the executor to commit to the database after each batch. Previously, the executor did not call commits by default.
  For new pipelines, the property is enabled by default. For upgraded pipelines, the property is disabled to prevent changes in pipeline behavior.

- **Spark executor enhancement** - The executor now supports Spark 2.x.

REST API / Command Line Interface

- **Offset management** - Both the REST API and command line interface can now retrieve the last-saved offset for a pipeline and update the offset for a pipeline when it is not running. Use these commands to implement pipeline failover using an external storage system. Otherwise, pipeline offsets are managed by Data Collector and there is no need to update the offsets.

Expression Language

- **vault:read enhancement** - The vault:read function now supports returning the value for keys nested in maps.

General

- **Support bundles** - You can now use Data Collector to generate a support bundle. A support bundle is a ZIP file that includes Data Collector logs, environment and configuration information, pipeline JSON files, resource files, and pipeline snapshots. You upload the generated file to the StreamSets support team so that we can use the information to troubleshoot your support tickets.

- **TLS property enhancements** - Stages that support SSL/TLS now provide the following enhanced set of properties that enable more specific configuration:
  - Keystore and truststore type - You can now choose between Java Keystore (JKS) and PKCS-12 (p-12). Previously, Data Collector only supported JKS.
  - Transport protocols - You can now specify the transport protocols that you want to allow. By default, Data Collector allows only TLSv1.2.
  - Cipher suites - You can now specify the cipher suites to allow. Data Collector provides a modern set of default cipher suites. Previously, Data Collector always allowed the default cipher suites for the JRE.

  To avoid upgrade impact, all SSL/TLS/HTTPS properties in existing pipelines are preserved during upgrade.

- **Cluster mode enhancement** - Cluster streaming mode now supports Spark 2.x. For information about using Spark 2.x stages with cluster mode, see "Stage Limitations".

- **Precondition enhancement** - Stages with user-defined preconditions now process all preconditions before passing a record to error handling. This allows error records to include all

precondition failures in the error message.

- **Pipeline import/export enhancement** - When you export multiple pipelines, Data Collector now includes all pipelines in a single zip file. You can also import multiple pipelines from a single zip file.

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-6383	Oracle sometimes returns an SCN that is not yet in the redo logs. The following exception occurs when Oracle CDC Client origin cannot locate an SCN:
	`ERROR OracleCDCSource - Error while attempting to produce records`
	`java.lang.IllegalStateException: SCN: '<scn>' is not valid and cannot be found in LogMiner logs`
	This can occur when the Oracle CDC Client Initial Change property is set to Latest Change or From SCN.
	Workaround: Set the Initial Change property to From Date, and enter the start date to use.

SDC-6365	When the Hadoop FS or Local FS destination encounters an exception while streaming a whole file, the destination might close the file and thus commit a partially written file.
SDC-6348	The MapReduce executor encounters intermittent Kerberos failures when submitting MapReduce jobs.
SDC-6288	The Directory origin does not accurately track offsets for directories described in the origin with a trailing slash.
SDC-6240	The following stages should use TlsConfigBean cipher suites and protocols: ● HTTP Server origin ● SDC RPC origin and destination ● SDC RPC to Kafka origin ● WebSocket Server origin
SDC-6237	The Hadoop FS destination might attempt to rename an output file more than once.
SDC-6233	The XML Parser processor throws an XML object exceeded maximum length error message after encountering any XML parsing error.
SDC-6182	The Salesforce origin encounters a null pointer exception when it executes a query that uses the Bulk API and that follows relationships between Salesforce objects.

SDC-6146	Previewing a pipeline using the command line interface results in an error about a missing CREATED value.
SDC-6087	Data Collector might encounter a deadlock when private classloaders are disabled and multiple Hadoop pipelines start simultaneously.
SDC-5959	Data Collector should list all registered JDBC drivers in the log.
SDC-5904	When the Hadoop FS, Local FS, or MapR FS destinations rename temporary files to support recovery, the destinations expect that the configured directory template ends with a directory separator.
SDC-5902	When the Directory origin is configured to use the last-modified timestamp and the file post-processing option is set for Delete or Archive, the origin deletes or moves all existing files in the directory when you start the pipeline and fails to process those files.
SDC-5863	Validation and preview of a pipeline fails with a PipelineWebhookConfig class not found error.
SDC-5818	The UDP Source origin can generate inaccurate information for the Timestamp, First, and Last fields when reading Netflow messages.
SDC-5801	Cloudera Manager does not pass the modified value of the runner.thread.pool.size property to Data Collector.
SDC-5757	The JDBC Producer destination should quote column names only when the Enclose Object Names property is enabled.

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-6438	The MapR distribution for Spark 2.x is not supported by cluster streaming pipelines.
SDC-6210	The show-vault-id command returns a NPE when Data Collector is installed on EC2 with IPv6 enabled. Workaround: If you can run Data Collector without IPv6, in the /etc/sysctl.conf file, disable IPv6 by configuring the disable IPv6 property as follows: `net.ipv6.conf.all.disable_ipv6 = 1`
SDC-6077	The Field Remover processor does not remove list fields from list-map data.

SDC-5871	Attempting to write a record to Kudu with a null primary key causes the pipeline to fail.
SDC-5758	Due to expected Kudu behavior, if Data Collector is configured to use Kerberos, but your Kudu installation does not, Data Collector cannot connect. This applies to Kudu version 1.2 and later. Workaround: If necessary, install a separate Data Collector that does not have Kerberos enabled.
SDC-5521	Kerberos-enabled pipelines that are valid using an Apache Kafka stage library can fail validation when using a Cloudera distribution of Kafka. The Apache Kafka client libraries have been modified by StreamSets to allow connectivity to Kerberized Kafka without requiring a JAAS configuration file. The Apache Kafka stage libraries (version 0.9 and greater) do not require a JAAS configuration file when enabling Kerberos. For CDH Kafka versions, a JAAS configuration is still required. Workaround: Include a JAAS configuration file on the classpath by modifying the `sdc-env.sh` or `sdcd-env.sh` file to include the following option in the SDC_JAVA_OPTS environment variable: `-Djava.security.auth.login.config=<path-to-jaas-config>` Or, you can use the matching version Apache Kafka stage library, instead of the Cloudera stage libraries.
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-5325	Cluster mode pipelines that read from a MapR cluster fail when the MapR cluster uses username/password login authentication. Workaround: On the Cluster tab for the pipeline, add the following Java property to the Worker Java Options property: `-Dmaprlogin.password.enabled=true`

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-5039	When you use the Hadoop FS origin to read files from all subdirectories, the origin cannot use the configured Hadoop FS User as a proxy user to read from HDFS. Workaround: If you need to use a proxy user to read from all subdirectories of the specified directories, set the HADOOP_PROXY_USER environment variable to the proxy user in libexec/_cluster-manager script, as follows: `export HADOOP_PROXY_USER = <proxy-user>`
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-3133	When you upgrade Data Collector from the RPM package, the environment configuration file `$SDC_DIST/libexec/sdcd-env.sh` is overwritten. Workaround: Back up the `sdcd-env.sh` file before you upgrade.
SDC-2374	A cluster mode pipeline can hang with a CONNECT_ERROR status. This can be a temporary connection problem that resolves, returning the pipeline to the RUNNING status. If the problem is not temporary, you might need to manually edit the pipeline state file to set the pipeline to STOPPED. Edit the file only after you confirm that the pipeline is no longer running on the cluster or that the cluster has been decommissioned. To manually change the pipeline state, edit the following file: `$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, ask for help, or find out about our next meetup, check out our Community page: https://streamsets.com/community/.

For general inquiries, email us at info@streamsets.com.