StreamSets Data Collector 3.1.0.0 Release Notes

February 9, 2018

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

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

Upgrading to Version 3.1.0.0

You can upgrade previous versions of Data Collector to version 3.1.0.0. For complete instructions on upgrading, see the Upgrade Documentation.

Update Value Replacer Pipelines

With version 3.1.0.0, Data Collector introduces a new Field Replacer processor and has deprecated the Value Replacer processor. The Field Replacer processor lets you define more complex conditions to replace values. For example, unlike the Value Replacer, the Field Replacer can replace values that fall within a specified range.

You can continue to use the deprecated Value Replacer processor in pipelines. However, the processor will be removed in a future release - so we recommend that you update pipelines to use the Field Replacer as soon as possible.

To update your pipelines, replace the Value Replacer processor with the Field Replacer processor. The Field Replacer replaces values in fields with nulls or with new values. In the Field Replacer, use field path expressions to replace values based on a condition.

For examples on how to use field path expressions in the Field Replacer to accomplish the same replacements as the Value Replacer, see Update Value Replacer Pipelines.

Update Einstein Analytics Pipelines

With version 3.1.0.0, the Einstein Analytics destination introduces a new append operation that lets you combine data into a single dataset. Configuring the destination to use dataflows to combine data into a single dataset has been deprecated.

You can continue to configure the destination to use dataflows. However, dataflows will be removed in a future release - so we recommend that you update pipelines to use the append operation as soon as possible.

New Features and Enhancements

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

Data Synchronization Solution for Postgres

This release includes a beta version of the Data Synchronization Solution for Postgres. The solution uses the new Postgres Metadata processor to detect drift in incoming data and automatically create or alter corresponding PostgreSQL tables as needed before the data is written. The solution also leverages the JDBC Producer destination to perform the writes.

As a beta feature, use the Data Synchronization Solution for Postgres for development or testing only. Do not use the solution in production environments.

Support for additional databases is planned for future releases. To state a preference, leave a comment on this issue.

Data Collector Edge (SDC Edge)

SDC Edge includes the following enhancements:

- Edge sending pipelines now support the following stages:
  - Dev Raw Data Source origin
  - Kafka Producer destination
- Edge pipelines now support the following functions:
  - emptyList()
  - emptyMap()
  - isEmptyMap()
  - isEmptyList()
  - length()
  - record:attribute()
  - record:attributeOrDefault()
  - size()
- When you start SDC Edge, you can now change the default log directory.

Origins

- **HTTP Client origin enhancement** - You can now configure the origin to use the Link in Response Field pagination type. After processing the current page, this pagination type uses a field in the response body to access the next page.

- **HTTP Server origin enhancement** - You can now use the origin to process the contents of all authorized HTTP PUT requests.

- **Kinesis Consumer origin enhancement** - You can now define tags to apply to the DynamoDB lease table that the origin creates to store offsets.

- **MQTT Subscriber origin enhancement** - The origin now includes a TOPIC_HEADER_NAME record header attribute that includes the topic information for each record.

- **MongoDB origin enhancement** - The origin now generates a no-more-data event when it has processed all available documents and the configured batch wait time has elapsed.

- **Oracle CDC Client origin enhancement** - You can now specify the tables to process by using SQL-like syntax in table inclusion patterns and exclusion patterns.

- **Salesforce origin enhancements** - The origin includes the following enhancements:
    - The origin can now subscribe to Salesforce platform events.
    - You can now configure the origin to use Salesforce PK Chunking.
    - When necessary, you can disable query validation.
    - You can now use Mutual Authentication to connect to Salesforce.

Processors

- **New Field Replacer processor** - A new processor that replaces values in fields with nulls or with new values.

  The Field Replacer processor replaces the Value Replacer processor which has been deprecated. The Field Replacer processor lets you define more complex conditions to replace values. For example, unlike the Value Replacer, the Field Replacer can replace values that fall within a specified range.

  StreamSets recommends that you update Value Replacer pipelines as soon as possible.

- **New Postgres Metadata processor** - A new processor that determines when changes in data structure occur and creates and alters PostgreSQL tables accordingly. Use as part of the Drift Synchronization Solution for Postgres in development or testing environments only.

- **Aggregator processor enhancements** - The processor includes the following enhancements:
    - Event records now include the results of the aggregation.
    - You can now configure the root field for event records: You can use a String or Map root field. Upgraded pipelines retain the previous behavior, writing aggregation data to a String root field.

- **JDBC Lookup processor enhancement** - The processor includes the following enhancements:
    - You can now configure a Missing Values Behavior property that defines processor behavior when a lookup returns no value. Upgraded pipelines continue to send records with no return value to error.
    - You can now enable the Retry on Cache Miss property so that the processor retries lookups for known missing values. By default, the processor always returns the default value for known missing values to avoid unnecessary lookups.

- **Kudu Lookup processor enhancement** - The processor no longer requires that you add a primary key column to the Key Columns Mapping. However, adding only non-primary keys can slow the performance of the lookup.

- **Salesforce Lookup processor enhancement** - You can now use Mutual Authentication to connect to Salesforce.

Destinations

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

- **New Named Pipe destination** - A new destination that writes data to a UNIX named pipe.

- **Einstein Analytics destination enhancements** - The destination includes the following enhancements:
    - You can specify the name of the Edgemart container that contains the dataset.
    - You can define the operation to perform: Append, Delete, Overwrite, or Upsert.
    - You can now use Mutual Authentication to connect to Salesforce.

- **Elasticsearch destination enhancement** - You can now configure the destination to merge data which performs an update with `doc_as_upsert`.

- **Salesforce destination enhancement** - The destination includes the following enhancements:
    - The destination can now publish Salesforce platform events.
    - You can now use Mutual Authentication to connect to Salesforce.

Data Formats

- **Log data format enhancement** - Data Collector can now process data using the following log format types:
    - Common Event Format (CEF)
    - Log Event Extended Format (LEEF)

Expression Language

- **Error record functions** - This release includes the following new function:
    - **record:errorStackTrace()** - Returns the error stack trace for the record.

- **Time functions** - This release includes the following new functions:
    - **time:dateTimeZoneOffset()** - Returns the time zone offset in milliseconds for the specified date and time zone.

    - **time:timeZoneOffset()** - Returns the time zone offset in milliseconds for the specified time zone.

- **Miscellaneous functions** - This release includes the following changed and new functions:
    - **runtime:loadResource()** - This function has been changed to trim any leading or trailing whitespace characters from the file before returning the value in the file. Previously, the function did not trim whitespace characters - you had to avoid including unnecessary characters in the file.

    - **runtime:loadResourceRaw()** - New function that returns the value in the specified file, including any leading or trailing whitespace characters in the file.

Additional Stage Libraries

This release includes the following additional stage libraries:

- Apache Kudu 1.6

- Cloudera 5.13 distribution of Apache Kafka 2.1
- Cloudera 5.14 distribution of Apache Kafka 2.1
- Cloudera CDH 5.14 distribution of Hadoop
- Kinetica 6.1

Miscellaneous

- **Data Collector classpath validation** - Data Collector now performs a classpath health check upon starting up. The results of the health check are written to the Data Collector log. When necessary, you can configure Data Collector to skip the health check or to stop upon errors.

- **Support bundle Data Collector property** - You can configure the bundle.upload.on_error property in the Data Collector configuration file to have Data Collector automatically upload support bundles when problems occur. The property is disabled by default.

- **Runtime properties enhancement** - You can now use environment variables in runtime properties.

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-8431	When configured to read files based on the last modified timestamp and to archive files after processing them, the Directory origin can encounter an exception if it attempts to compare the timestamp of an archived offset file with the remaining files.
SDC-8416	When a batch is empty, you cannot use the Pipeline is Idle default metric rule.
SDC-8414	The Directory origin can encounter a race condition when it is configured to read files based on the last modified timestamp and multiple files are added to the directory within a few seconds.
SDC-8407	The HTTP Client origin encounters the following error when the request times out on the first run and there is no prior saved offset: `'java.lang.IllegalArgumentException: Offset must have at least 8 parts'`
SDC-8340	The Oracle CDC origin returns null for a string with the value "null".
SDC-8321	When you delete a pipeline, the pipeline state cache information in memory is not cleared.
SDC-8292	The Solr destination cannot authenticate when Solr is configured with both Basic and Negotiate authentication.

SDC-8190	The Field Type Converter processor encounters a null pointer exception when converting the Zoned DateTime data type.
SDC-8095	The JMS Consumer origin fails to process SDC Record data with the following error: `PARSER_03 - Cannot parse record from message 'JMSTestQueue::0': com.streamsets.pipeline.lib.parser.DataParserException: SDC_RECORD_PARSER_00 - Could advance reader 'JMSTestQueue::0' to '0' offset`
SDC-8069	After upgrading to Data Collector 3.0.0.0 that is not enabled to work with Control Hub, cluster batch pipelines fail validation.
SDC-7987	The Hadoop FS origin and the MapReduce executor should not close shared instances of FileSystem objects.
SDC-7412	The HTTP Server origin does not correctly clean up its server socket when it fails unexpectedly.
SDC-6086	When processing Avro data using the schema definition included in the record header attribute, a destination fails to register the Avro schema with the Confluent Schema Registry.
SDC-5325	Cluster batch mode pipelines that read from a MapR cluster fail when the MapR cluster is secured with built-in security.

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-8480	When a Salesforce Lookup processor uses SOQL Query mode, the processor doesn't retrieve the object type from the query causing the pipeline to fail with the following error: `FORCE_21 - Can't get metadata for object:` Workaround: 1. In the **Lookup** tab for the processor, set **Lookup Mode** to **Retrieve**. 2. Set the **Object Type** property to the object type in your query - for example Account, or Customer__c. You can ignore the remaining properties. 3. Set **Lookup Mode** back to **SOQL Query** and run the pipeline.
SDC-8386	Many stages generate duplicate error records when a record does not meet all

	configured preconditions.
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-8257	When using multiple threads to process data and enabled for event generation, the Directory origin generates no-more-data events when each thread completes processing data, instead of when all threads complete processing available data.
SDC-8252	When overriding the ldap-login.conf file with the “Data Collector Advanced Configuration Snippet (Safety Valve) for ldap-login.conf” Cloudera Manager configuration option, the Data Collector CSD does not generate the ldap-bind-password.txt configuration file. Workarounds: Store the LDAP password separately, or include the password in plain text in the safety valve.
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-7872	Due to Oracle LogMiner returning an unsupported SQL format, Data Collector cannot parse changes to tables containing the XML type.
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-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-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-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-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-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-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 from our Google group or Slack channel, 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.