StreamSets Data Collector

Cumulative 3.17.x Release Notes

August 14, 2020

We're happy to announce a new version of StreamSets Data Collector. Version 3.17.x contains several new features, enhancements, and some important bug fixes in the following release:

- Version 3.17.1 - August 14, 2020
- Version 3.17.0 - July 28, 2020

This document contains important information about the following topics:

New Features and Enhancements

Data Collector 3.17.x includes the following new features and enhancements:

New Stage

- **SAP HANA Query Consumer -** Use this origin to read data from an SAP HANA database with a user-defined query.

Stage Enhancements

- **Control Hub API processor** - The processor can process responses of any size. Previously the maximum response size was 50,000 characters.
- **Elasticsearch destination** - You can use record functions and delimited data record functions in the Additional Properties field.
- **Elasticsearch stages** - The Elasticsearch origin and destination include a User Name property and a Password property instead of a single Security Username/Password property.

  Existing pipelines are not affected by this change. During an upgrade, existing configurations for the Security Username/Password property are placed into the User Name property, which supports the `username:password` format.
- **HTTP Client processor** - You can configure the processor to use the following enhancements:

    - Actions to take based on the response status.
    - Pagination properties to enable processing large volumes of data from paginated APIs.
    - Action to take when the request times out because the HTTP service did not respond within the read timeout period.

- **JDBC MySQL data type conversions** - The JDBC origins and JDBC processors convert MySQL unsigned integers as follows:
    - Bigint Unsigned converts to Decimal.
    - Int Unsigned and Mediumint Unsigned convert to Long.
    - Smallint Unsigned converts to Integer.

  This change can require performing a post-upgrade task.

- **Kinesis stages** - The Kinesis Consumer origin, Kinesis Firehose destination, and Kinesis Producer destination provide an Authentication Method property that allows selecting either IAM Roles or AWS Keys.

  Previously, you used IAM roles by omitting AWS keys when configuring the stages. This change does not affect existing pipelines.

- **Orchestration stages**:
    - Some orchestration stages and properties have been renamed. These changes do not affect existing pipelines. This includes, but is not limited to, the following:
        - The Start Job origin and processor are now the Start Jobs origin and processor.
        - The Start Pipeline origin and processor are now the Start Pipelines origin and processor.
        - The Wait for Job Completion processor is now the Wait for Jobs processor.
        - The Wait for Pipeline Completion processor is now the Wait for Pipelines processor.
    - Records generated by the Start Jobs and Start Pipelines stages, and updated by the Wait for Jobs and Wait for Pipelines stages, include pipeline and stage metrics when available. This includes input record, output record, error record, and error message counts.

- **Scripting origins -** You can reset the origin for pipelines that include the Groovy Scripting, JavaScript Scripting, or Jython Scripting origin.

- **SFTP/FTP/FTPS origin** - The origin generates an error when it encounters a file that it does not have permission to read instead of stopping the pipeline.

- **TensorFlow Evaluator processor:**

    - The processor uses the 1.15 TensorFlow client library and supports all 1.x TensorFlow versions.

    - In the Fields to Convert property for each input configuration, you can configure a field type expression that defines a set of fields.

Pipeline Enhancements

- **Pipeline run history** - The pipeline run history displays the input, output, and error record count for each pipeline run.

- **Pipeline run summary** - Information about the most recent pipeline run remains available on the Summary tab of the pipeline after the pipeline stops. The summary includes run details such as the start time and duration.

- **Pipeline start and stop events** - The event records generated when a pipeline starts and stops include fields for the related Control Hub job ID and job name.

- **Stage library panel display and stage installation:**

    - The stage library panel displays all Data Collector stages, instead of only the installed stages. Stages that are not installed appear disabled, or greyed out.

    - When you click on a disabled stage, you can install the stage library that includes the stage.

Security Enhancements

- **File-based user authentication** - You can use the Data Collector UI to change your password when Data Collector is configured for file-based authentication.

- **Hashicorp Vault credential store** - You can enable the use of a namespace in Hashicorp Vault by configuring a namespace path for the credentialStore.vault.config.namespace property in the `$SDC_CONF/credential-stores.properties` file.

  For example, `credentialStore.vault.config.namespace=nspace1/nspace2/`.

- **Runtime:resourcesDirPath() function** - Returns the full path to the directory for runtime resource files.

- **SSL/TLS enhancement -** Stages that use SSL/TLS can load the contents of the keystore and truststore from a credential store.

Additional Enhancement

- **Data Collector production batch size** - The default value for the production.maxBatchSize property in the Data Collector configuration file has increased to 50,000 records.

  This change does not affect existing pipelines.

Deprecated Features

- **Databricks ML Evaluator processor** - This processor is deprecated and will be removed in a future release. Do not use the processor in new pipelines.

Upgrade Information

You can upgrade previous versions of Data Collector to version 3.17.x. For complete instructions on upgrading, see the Upgrade documentation.

Upgrade Enterprise Stage Libraries

When you upgrade Data Collector, you must determine whether to upgrade your Enterprise stage libraries. See Enterprise Stage Libraries in the Data Collector documentation for a list of available Enterprise stage libraries and links to the supported versions and the stage documentation. To view the release notes for Enterprise stage libraries, see the StreamSets Documentation page.

1. Uninstall the previous version of the Enterprise stage library.
    a. In Package Manager, select the installed version.
    b. Click the **Uninstall** icon.
    c. Restart Data Collector.
2. Follow the stage documentation to install the new version of the Enterprise stage library and restart Data Collector.

Review MySQL Data Processing

In Data Collector 3.17.0, JDBC origins and processors convert MySQL unsigned integer data types to different Data Collector types than in earlier Data Collector versions.

After you upgrade to 3.17.0, review pipelines that process MySQL database data to ensure that configured expressions provide the expected results.

The following table describes the data type conversion changes:

MySQL Data Type	Data Collector Data Type Conversion Before Version 3.17.0	Data Collector Data Type Conversion for Version 3.17.0 and Later
Bigint Unsigned	Long	Decimal
Int Unsigned	Integer	Long
Mediumint Unsigned	Integer	Long
Smallint Unsigned	Short	Short

Update Elasticsearch Security Properties (Optional)

In Data Collector 3.17.0, Elasticsearch stages provide a User Name property and a Password property. Elasticsearch stages in previous versions pass the credentials together in a single Security Username/Password property.

When you upgrade to version 3.17.0 or later, any configuration in the Security Username/Password properties is moved to the new User Name property, where the Security Username/Password format, `<username>:<password>`, remains valid.

Though not required, you can update Elasticsearch stages to use the new User Name and Password properties.

Fixed Issues in 3.17.1

The following table lists some of the known issues that are fixed in Data Collector 3.17.1.

For the full list, click here.

JIRA	Description
SDC-15332	Previewing a pipeline that includes the Oracle CDC Client origin generates an exception.
SDC-14645	An exception occurs when an expression references the field that the results of the expression are written to. This can occur in the Expression Evaluator processor, as well as other stages, such as the Field Replacer and Field Masker processors.

Fixed Issues in 3.17.0

The following table lists some of the known issues that are fixed in Data Collector 3.17.0.

For the full list, click here.

JIRA	Description
SDC-15066	When an Amazon S3 origin has a Common Prefix and Prefix Pattern defined without wildcards, the origin does not read the specified file.
SDC-15016	When reading from Oracle 12c, the Oracle CDC Client origin can encounter a PGA memory leak.
SDC-15012	When reading from Oracle 19c, the Oracle CDC Client origin encounters errors when trying to process a redo log that is not yet archived.
SDC-15006	The Kafka Multitopic Consumer origin fails when one or more topic names include a hyphen (-).

SDC-14997	The PostgreSQL CDC origin does not indicate when database connectivity issues occur.
SDC-14747	The Oracle CDC Client origin does not properly honor table exclusion patterns.
SDC-14701	The Azure Data Lake Storage Gen2 destination can sometimes cause out of memory errors.
SDC-14663	When Oracle CDC Client generates a CREATE event and the event does not include the schema, the origin creates a null event record instead of an empty event record.
SDC-13347	The Oracle CDC Client origin does not include new tables that should be tracked when you use wildcards in the Table Name Pattern property.
SDC-9278	The Kinesis Consumer origin does not use a region specified in the Other > Endpoint property.

Known Issues

Please note the following known issues with Data Collector 3.17.x.

For a full list of known issues, click here.

JIRA	Description

SDC-15193	Pipeline failures can occur when Data Collector is installed through a cloud service provider marketplace or from an RPM package on CentOS 7, Oracle Linux 7, or Red Hat Enterprise Linux 7, because the installation incorrectly points the SDC_RESOURCE environment variable to a directory that does not exist, `/opt/streamsets-datacollector/resources`.
	
	Pipelines failures due to this issue generate the following error message:
	`sdk.sdc_api.StartError: Unexpected error starting pipeline: java.lang.RuntimeException: ERROR: Serializing resources directory: 'resources': java.io.IOException: Failed to list contents of /opt/streamsets-datacollector/resources`
	
	Workaround: Use the following steps to set the SDC_RESOURCE environment variable:
	
	1. Use the following command to edit the systemd Data Collector configuration:
	`systemctl edit sdc`
	
	2. Enter the following lines, then save and exit the configuration:
	
	`[Service]`
	`Environment="SDC_RESOURCES=/var/lib/sdc-resources"`
	
	3. Restart Data Collector.

SDC-14338	When the Field Mapper processor encounters an error creating a new field, the error stops the pipeline.
SDC-13679	Pressing the Tab key while configuring a Field Remover processor can generate a null pointer exception.
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-9514	Runtime parameters are not supported in all configuration properties in cluster batch execution mode, such as Max Batch Size.
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-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-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-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.

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.