StreamSets Data Collector

Cumulative 3.16.x Release Notes

August 12, 2020

We're happy to announce a new version of StreamSets Data Collector. Version 3.16.x contains several new features, enhancements, and some important bug fixes in the following releases:

- Version 3.16.2 - August 12, 2020
- Version 3.16.1 - June 17, 2020
- Version 3.16.0 - May 22, 2020

This document contains important information about the following topics:

New Features and Enhancements

Data Collector 3.16.x includes the following new features and enhancements:

New Stages

This release includes the following new stages:

- **SFTP/FTP/FTPS executor** - Use the executor to move or remove a file from an SFTP, FTP, or FTPS server upon receiving an event.
- **New Orchestrator stages:**
    - **Start Job origin** - Use this origin to start a Control Hub job.
    - **Wait for Job Completion processor** - Use this processor to wait for a Control Hub job to complete.
    - **Wait for Pipeline Completion processor** - Use this processor to wait for a Data Collector, Transformer, or Edge pipeline to complete.

Stage Enhancements

This release includes the following stage enhancements:

- **Amazon stages** - You can now configure an Authentication Method property to specify whether to connect with an IAM role, with AWS keys, or without authentication. Previously, you could not connect without authentication to a public bucket.
- **Amazon S3 stages** - The Amazon S3 origin, destination, and executor can use a virtual address model to access objects. Previously, the stages used a path address model.
- **HTTP Client stages:**
    - **HTTP status codes** - Error logging details generated by HTTP Client stages now include HTTP status codes.
    - **HTTP status header attribute** - The HTTP Client stages include an `HTTP-Status` header attribute that stores the HTTP status for each record.
    - **Timeout defaults** - The HTTP Client stages no longer allow 0 for the Connection Timeout or Read Timeout properties. The default value for Connect Timeout is now 250000 milliseconds. The default value for Read Timeout is 30000 milliseconds.
- **HTTP Client origin record generation** - You can configure the origin to generate records for all statuses that are not added to the Per-Status Actions list. You can also specify a field to write the error response body for those records.
- **HTTP Server origin authentication enhancements** - The origin can create more secure connections using Kerberos authentication. It can also use basic authentication when you enable TLS/SSL.
- **JDBC Multitable Consumer origin header attributes** - You can enable the origin to create JDBC header attributes.
- **Kafka stages allow Kerberos credentials** - You can specify Kerberos keytabs and principals in Kafka stages, including the Kafka Consumer, Kafka Multitopic Consumer, and Kafka Producer.
- **Kafka Multitopic Consumer origin** - You can include Kafka timestamps in the record header.
- **Oracle CDC Client origin enhancements:**
    - **19c support** - You can use the Oracle CDC Client origin to read changed data from Oracle 19c in addition to 11g, 12c, and 18c.
    - **Property removal** - With this release, initialization when Dictionary Source is set to Redo Logs has been improved. As a result, the Duration of Directory Extraction property is no longer needed and has been removed.
- **REST Service origin enhancements:**
    - **API Gateway -** You can configure the origin to use Data Collector as an API Gateway.
    - **Authentication** - The origin connects securely using Kerberos authentication. It can also use basic authentication when you enable TLS/SSL.

    - **Endpoint URL** - The microservice endpoint URL now displays in monitor mode.

- **Salesforce origin event enhancement** - The no-more-data event record generated by the origin now includes a `record-count` field that specifies the number of records that were successfully processed.

- **Salesforce Lookup processor:**

    - **Multiple results** - You can configure the processor to write multiple return values as a list in a single record instead of returning only the first value or creating a record for each value.

    - **Bulk API** - You can use the Salesforce Bulk API to lookup records in Salesforce based on a SOQL query.

- **SFTP/FTP/FTPS Client origin processing delay** - You can configure a File Processing Delay property when you want to allow time for a file to be completely written before processing it.

- **Start Pipeline processor:**

    - **Pipeline name support** - You can configure the Pipeline ID Type property to enable specifying the pipeline to start based on the pipeline name instead of the pipeline ID.

    - **Unique Task Name property** - You can specify a unique name for the processor that is included in the output record.

- **Start Job processor:**

    - **Job name support -** You can configure the Job ID Type property to enable specifying the Control Hub job to start based on the job name instead of the job ID.

    - **Unique Task Name property** - You can specify a unique name for the processor that is included in the output record.

Enterprise Stage Libraries

Enterprise stage libraries are free for use in both development and production.

In June 2020, StreamSets released an updated Enterprise stage library for SQL Server 2019 Big Data Cluster.

For a list of available Enterprise libraries, see Enterprise Stage Libraries. For more information about the new features, fixed issues, and known issues in an Enterprise stage library, see the release notes for the Enterprise stage library, available under Enterprise Libraries on the StreamSets Documentation page.

Additional Enhancements

This release includes the following additional enhancements:

- **Default users and groups for cloud service provider installations** - Data Collector installed through a cloud service provider marketplace now includes only a default `admin` user account and no default groups.

- **Group secrets in credential stores** - You can configure Data Collector to validate a user's group against a comma-separated list of groups allowed to access each secret.
- **MapR 5.x no longer supported** - With this release, MapR 5.x stage libraries are no longer supported and no longer available.
- **Monitor mode indicator** - Monitor mode displays the following real-time Running icon on the stage that is processing data: 

Upgrade Information

You can upgrade previous versions of Data Collector to version 3.16.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.

Fixed Issues in 3.16.2

The following table lists some of the issues that are fixed in Data Collector 3.16.2.

For the full list, click here.

JIRA	Description
SDC-15006	The Kafka Multitopic Consumer fails to process data in topics that include a hyphen in the name.
SDC-15281	When processing Avro data using the Avro Message Key Format, the Kafka Producer destination uses a string serializer for the expression partitioner when writing data, ignoring the setting for the Key Serializer property.
SDC-15174	When a pipeline that contains an Amazon S3 origin stops and resumes, the pipeline generates an error if the offset file was deleted.

Fixed Issues in 3.16.1

The following table lists some of the issues that are fixed in Data Collector 3.16.1.

For the full list, click here.

JIRA	Description
SDC-14882	The JDBC Query Consumer origin closes the connection to the database after each batch. This can cause the origin to indefinitely reprocess the records in the first batch when running in full mode.
SDC-14872	When using multiple characters as a delimiter, newline characters are not properly ignored when they appear between the configured quote characters.
SDC-14865	When processing Whole File data, stages do not honor the error handling configured in the stage.
SDC-14856	The Pulsar Consumer origin does not properly handle runtime parameters for topic names.
SDC-14835	The Amazon S3 origin generates errors when processing empty directories. With this fix, empty directories are not processed.
SDC-14808	The MapR Streams Consumer and MapR Streams Producer cannot connect to MapR.
SDC-14693 SDC-14683	Data loss can occur with the Kafka Multitopic Consumer.
SDC-14643	Reading incorrectly-formed JSON data with the RabbitMQ Consumer causes unexpected errors.

Fixed Issues in 3.16.0

The following table lists some of the known issues that are fixed in Data Collector 3.16.0.

For the full list, click here.

JIRA	Description
SDC-14763	When reading data, the PostgreSQL CDC Client origin stores the offset immediately, instead of waiting until the data is written to pipeline destinations. This can result in data loss in certain circumstances.
SDC-14752	When you restart a pipeline, the PostgreSQL CDC Client origin rereads the last record of the last batch of the previous pipeline run.

SDC-14525	When pipeline metadata is inconsistent, such as when pipeline run information is missing, Data Collector fails to provide pipeline details to Control Hub.
SDC-14253	Package Manager cannot install external libraries for custom stage libraries. Package Manager can only install external libraries for stage libraries that were installed using Package Manager. **Workaround:** Install external libraries manually for custom stage libraries. For example, to install a JDBC driver for a MemSQL Enterprise library that was installed as a custom stage library, copy the JDBC driver files to the following location: `<custom stage library dir>/streamsets-datacollector-memsql-lib/lib`
SDC-13918	In some cases, the PostgreSQL CDC Client origin incorrectly synchronizes with Postgres due to offset handling.

Known Issues

Please note the following known issues with Data Collector 3.16.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-14936	Jobs submitted by a MapReduce executor with the HDP 3.1.0 stage library can fail to complete in certain operating systems.
SDC-14701	The Azure Data Lake Storage Gen2 destination can sometimes cause out of memory errors.
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.

Additional Resources

Our Documentation page provides access to all StreamSets product documentation: streamsets.com/docs.

To report an issue, to get help from our Google group, Slack channel, or Ask site, or to find out about our next meetup, visit our Community page: https://streamsets.com/community/.

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

For more information about StreamSets, visit our website: https://streamsets.com/.