StreamSets Data Collector and Data Collector Edge Cumulative 3.7.x Release Notes

January 11, 2019

We're happy to announce new versions of StreamSets Data Collector and StreamSets Data Collector Edge. Version 3.7.x contains several new features, enhancements, and some important bug fixes for the following versions of StreamSets Data Collector and Data Collector Edge:

- Version 3.7.1 - January 11, 2019
- Version 3.7.0 - January 9, 2019

This document contains important information about the following topics:

Upgrading to Version 3.7.x

You can upgrade previous versions of Data Collector to version 3.7.x. For complete instructions on upgrading, see the Upgrade Documentation.

Update Cluster Pipelines

With version 3.7.0, Data Collector now requires that the Java temporary directory on the gateway node in the cluster is writable.

The Java temporary directory is specified by the Java system property `java.io.tmpdir`. On UNIX, the default value of this property is typically `/tmp` and is writable.

Previous Data Collector versions did not have this requirement. Before running upgraded cluster pipelines, verify that the Java temporary directory on the gateway node is writable.

Update Pipelines that Use Kafka Consumer or Kafka Multitopic Consumer Origins

With version 3.7.0, Data Collector no longer uses the auto.offset.reset value set in the Kafka Configuration property to determine the initial offset. Instead, Data Collector uses the new Auto Offset Reset property to determine the initial offset. With the default setting of the new property, the origin reads all existing messages in a topic. In previous versions, the origin read only new messages by default. This setting only affects pipelines that have never ran. Running a pipeline sets an offset value.

After upgrading to 3.7.0, complete the following steps to update any pipelines that have not run and use these origins:

1. On the **Kafka** tab, set the value of the **Auto Offset Reset** property:
    - **Earliest** - Select to have the origin read messages starting with the first message in the topic (same behavior as if you configured **auto.offset.reset** to **earliest** in previous versions).
    - **Latest** - Select to have the origin read messages starting with the last message in the topic (same behavior as if you did not configure **auto.offset.reset** in previous versions).
    - **Timestamp** - Select to have origin read messages starting with messages at a particular timestamp, which you specify in the **Auto Offset Reset Timestamp** property.

2. If configured in the **Kafka Configuration** property, delete the **auto.offset.reset** property.

Duplicate Data in Oracle CDC Client Pipelines

After upgrading to version 3.7.0, pipelines that use the Oracle CDC Client origin can produce some duplicate data.

Due to a change in offset format, when the pipeline restarts, the Oracle CDC Client origin reprocesses all transactions with the commit SCN from the last offset to prevent skipping unread records. This issue occurs only for the last SCN that was processed before the upgrade, and only once, upon upgrading to Data Collector version 3.7.0.

When possible, remove the duplicate records from the destination system.

New Features and Enhancements in Version 3.7.x

Version 3.7.x includes several new features and enhancements for Data Collector and Data Collector Edge.

Data Collector New Features and Enhancements

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

New Microsoft Azure Support

With this release, you can now use the Hadoop FS Standalone origin to read data from Azure Data Lake Storage. You can also use the Hadoop FS destination to write to Azure Data Lake Storage.

Use the Hadoop FS destination when you need to use Azure Active Directory refresh token authentication to connect to Azure Data Lake Storage, or when you want to write to Azure Data Lake Storage in a cluster streaming pipeline. For all other cases, use the existing Azure Data Lake Storage destination.

New Enterprise Stage Libraries

Enterprise stage libraries are free for development purposes only. For information about purchasing the stage library for use in production, contact StreamSets.

This release includes the following new Enterprise stage libraries:

Stage Library Name	Description

streamsets-datacollector-memsql-lib	For MemSQL. Includes the MemSQL Fast Loader destination.
streamsets-datacollector-snowflake-lib	For Snowflake. Includes the Snowflake destination.
streamsets-datacollector-teradata-lib	For Teradata. Includes the Teradata Consumer origin.

Installation

- **Install Data Collector on Microsoft Azure** - The process to install Data Collector on Microsoft Azure has been enhanced. Data Collector now automatically starts as a service after the resource is deployed. You no longer need to use SSH to log in to the virtual machine to run the Data Collector installation script and then start Data Collector.

Origins

This release includes the following new origin:

- **Teradata Consumer origin** - Reads data from multiple Teradata Database tables. To use this origin, you must install the Teradata stage library. This is an Enterprise stage library.

This release includes enhancements to the following origins:

- **Amazon S3** - The origin can now create multiple threads to enable parallel processing in a multithreaded pipeline.
- **Dev Data Generator** - The development origin now generates fake data for email, race, sex, and social security numbers.
- **Elasticsearch** - The origin now supports authentication with AWS credentials when using Amazon Elasticsearch Service.
- **Hadoop FS Standalone** - The origin now supports reading data from Microsoft Azure Data Lake Storage.
- **Kafka Consumer** - The origin includes the following enhancements:
    - A new Auto Offset Reset property determines the first message read in a topic when a consumer group and topic has no previous offset stored. The origin can read from the earliest message, the latest message, or a particular timestamp. The default setting causes the origin to read all existing messages in a topic.

      In previous versions, the origin read only new messages by default. For information on upgrading pipelines that use the Kafka Consumer origin, see Update Pipelines that Use Kafka Consumer or Kafka Multitopic Consumer Origins.

    - A new Include Timestamps property enables you to include Kafka timestamps in the record header.
- **Kafka Multitopic Consumer** - The origin includes a new Auto Offset Reset property that determines the first message read in a topic when a consumer group and topic has no previous offset stored. The origin can read from the earliest message, the latest message, or a

particular timestamp. The default setting causes the origin to read all existing messages in a topic.

In previous versions, the origin read only new messages by default. For information on upgrading pipelines that use the Kafka Multitopic Consumer origin, see Update Pipelines that Use Kafka Consumer or Kafka Multitopic Consumer Origins.

- **PostgreSQL CDC Client** - The origin now has a new default value for the Replication Slot property: `sdc`. This property must contain only lowercase letters and numbers.
- **REST Service** - This microservice origin now supports SSL mutual authentication.
- **Salesforce** - The origin now includes a new subscription type: Change Data Capture.
- **SQL Server CDC Client** - The origin now includes the Use Direct Table Query property to enable direct table queries and the Maximum Transaction Length property to specify the amount of time to check for changes to a record before committing the data.
- **TCP Server** - The origin now includes a Read Timeout property that sets the amount of time that the origin waits to receive data before Data Collector closes the connection. The default is 5 minutes. In previous releases, the connection remained opened indefinitely.

Processors

This release includes enhancements to the following processors:

- **Databricks ML Evaluator** - With this release, this processor is no longer considered a Technology Preview feature and is approved for use in production. Also, you can now specify a model path relative to the Data Collector resources directory.
- **Field Hasher** - The processor now supports hashing with the SHA512 cryptographic hash function.
- **Field Remover** - In addition to removing fields, and removing fields with null values, the processor now supports removing fields under the following conditions:
    - When the values are empty strings.
    - When the values are null or empty strings.
    - When the values are a specified value.
- **Field Renamer** - The processor now supports StreamSets expression language in target field paths. With this feature, you can use string functions to change field names to be all uppercase or lowercase.
- **Kudu Lookup** - The processor now supports the Decimal data type available with Apache Kudu 1.7 and later.
- **JDBC Lookup** - The processor now supports returning multiple matching values as a list in a single record.
- **MLeap Evaluator** - With this release, this processor is no longer considered a Technology Preview feature and is approved for use in production. Also, you can now specify a model path relative to the Data Collector resources directory.
- **MongoDB Lookup** - The processor includes the following updates to property and tab names:
    - Several properties have moved from the MongoDB tab to a new Lookup tab.

    - The SDC Field to Document Field Mapping property is now known as Document to SDC Field Mappings.
    - The Field Name in Document property is now known as Document Field.
    - The New Field to Save Lookup Result property is now known as Result Field.

- **PMML Evaluator** - The processor is approved for use in production. This release removes the Technology Preview designation. Also, you can now specify a model path relative to the Data Collector resources directory.

- **Salesforce Lookup** - The processor now supports using time functions in SOQL queries.

- **TensorFlow Evaluator** - With this release, this processor is no longer considered a Technology Preview feature and is approved for use in production. Also, you can now specify a model path relative to the Data Collector resources directory.

Destinations

This release includes the following new destinations:

- **MemSQL Fast Loader destination** - Inserts data into a MemSQL or MySQL database table with a LOAD statement. To use this destination, you must install the MemSQL stage library. This is an Enterprise stage library.

- **Snowflake destination** - Writes new or CDC data to tables in a Snowflake database schema. To use this destination, you must install the Snowflake stage library. This is an Enterprise stage library.

This release includes enhancements to the following destinations:

- **Azure Data Lake Storage** - Due to Microsoft rebranding, the Azure Data Lake Store destination is now known as the Azure Data Lake Storage destination.

- **Elasticsearch** - The destination now includes:
    - Support for authentication with AWS credentials when using Amazon Elasticsearch Service.
    - A new Additional Properties property to specify an extra field in an action statement.

- **Hadoop FS** - The destination now supports writing data to Microsoft Azure Data Lake Storage.

- **Kudu** - The destination now supports the Decimal data type if using the Apache Kudu 1.7.0 stage library.

Data Governance Tools

- **Pipeline metadata** - Data Collector now publishes additional pipeline metadata to Cloudera Navigator and Apache Atlas, including the pipeline description, labels, parameters, version, and the user who started the pipeline.

Pipeline Parameters

- **Parameters for checkboxes and drop-down menus** - You can now call pipeline parameters for properties that display as checkboxes and drop-down menus. The parameters must evaluate to a valid option for the property.

Cluster Pipelines

- **Gateway node requires writable temporary directory** - When you run a cluster pipeline, Data Collector now requires that the Java temporary directory on the gateway node is writable. The Java temporary directory is specified by the Java system property `java.io.tmpdir`. On UNIX, the default value of this property is typically `/tmp` and is writable.

  For information about upgrading cluster pipelines that previously ran on gateway nodes without a writable temporary directory, see Update Cluster Pipelines.

Expression Language

- **String functions** - This release includes the following new function:
  - `str:lastIndexOf(<string>,<subset>)` - Returns the index within a string of the last occurrence of the specified subset of characters.

Data Collector Configuration

- **Data Collector Security Manager** - For enhanced security, the Data Collector configuration file now provides a property to enable the Data Collector Security Manager, instead of the Java Security Manager. The Data Collector Security Manager does not allow stages to access files in the following directories:
  - Configuration directory defined in the SDC_CONF environment variable.
  - Data directory defined in the SDC_DATA environment variable.

  In addition, the Data Collector Security Manager does not allow stages to write to files in the resources directory defined in the SDC_RESOURCES environment variable. Stages can only read files in the resources directory.

  By default, Data Collector uses the Java Security Manager which allows stages to access files in all Data Collector directories.

- **HTTP/2 support** - Data Collector now provides a property in the Data Collector configuration file to enable support of the HTTP/2 protocol for the UI and API. Because HTTP/2 requires TLS, to enable HTTP/2, configure both the `http2.enable` and the `https.port` properties.

- **Package Manager** - The Package Manager includes the following enhancements:
  - Data Collector now provides a `package.manager.repository.links` property to enable specifying alternate locations for the Package Manager repository.
  - The Package Manager now displays the list of stages associated with each stage library.

- **Data Collector logging** - Data Collector now logs the stage instance that generates each log line.

Stage Libraries

- **New Stage Libraries** - This release includes the following new stage libraries:

Stage Library Name	Description

streamsets-datacollector-cdh_kafka_3_1-lib	Cloudera distribution of Apache Kafka 3.1.0 (1.0.1).
streamsets-datacollector-kinetica_6_2-lib	For Kinetica 6.2. Includes the KineticaDB destination.

- **Updated Stage Libraries** - This release includes updates to the following stage libraries:

Stage Library Name	Description
streamsets-datacollector-apache-pulsar_2-lib	Apache Pulsar version 2.x.
streamsets-datacollector-cdh_6_0-lib	Cloudera CDH version 6.0.x distribution of Apache Hadoop. Now includes the following stages: ● HBase Lookup processor ● Spark Evaluator processor ● HBase destination The stage library no longer includes the following stage: ● Solr destination

Data Collector Edge New Features and Enhancements

This Data Collector Edge (SDC Edge) version includes new features and enhancements in the following areas.

Technology Preview Functionality

The following Technology Preview stage is available for edge pipelines in this release:

- **gRPC Client origin** - A new origin that processes data from a gRPC server by calling gRPC server methods. The origin can call unary and server streaming RPC methods. Use this origin only in pipelines configured for edge execution mode.

Origins in Edge Pipelines

This release includes enhancements to the following origins that are supported in edge pipelines:

- **File Tail** - The origin can now read multiple sets of files when it is included in edge pipelines.

- **Windows Event Log** - The origin can now use the Event Logging API or the Windows Event Log API to read data from a Microsoft Windows event log. Microsoft recommends using the newer Windows Event Log API.

  Previously, the origin used the Event Logging API only. Upgraded pipelines continue to use the Event Logging API.

Processors in Edge Pipelines

Edge pipelines now support the Dev Random Error processor.

Destinations in Edge Pipelines

Edge pipelines now support the following destinations:

- To Error
- To Event
- Kinesis Firehose
- Kinesis Producer

SDC Edge as a System Service

If you register SDC Edge to run as a system service, you can now run a command as an administrator to display the status of the service.

SDC Edge Configuration

- **Log file enhancement** - You can now modify the default log file directory in the SDC Edge configuration file, `<SDCEdge_home>/etc/edge.conf`. You can no longer modify the default log file directory when you manually start SDC Edge.

- **SDC Edge host name** - You can now configure the host name where SDC Edge runs by defining the `base-http-url` property in the SDC Edge configuration file.

Fixed Issues in Version 3.7.1

The following table lists some of the known issues that are fixed with version 3.7.1.

For the full list, click here.

JIRA	Description
SNOWFLAKE-68	An exception can occur when the Snowflake destination does not automatically create tables correctly.
SNOWFLAKE-67	The Snowflake destination does not work in multithreaded pipelines.
SNOWFLAKE-61	Improve the performance of the Snowflake destination when writing to multiple tables.
SDC-10788	When using Package Manager to install external stage libraries, the list of stage libraries reads “undefined”.
SDC-10743	When a Salesforce origin uses a date offset, preview of the pipeline shows duplicate records.

Fixed Issues in Version 3.7.0

The following table lists some of the known issues that are fixed with version 3.7.0.

For the full list, click here.

JIRA	Description
SDC-10695	The JDBC Lookup processor was opening additional connections unnecessarily.
SDC-10534	In certain cases, the SQL Server Change Tracking origin allowed committing a null offset.
SDC-10479	When the Kafka Consumer encounters an exception while polling for data, the origin can silently fail to process data.
SDC-10455	Incompatible characters in pipeline IDs do not allow running the pipelines with Control Hub.
SDC-10217	Data Collector throws interceptors registered for stage exceptions.
SDC-10162	The MapR Multitopic Streams Consumer ignores the batch size property configured in the stage.
SDC-10153	The HBase Lookup processor's batch lookup mode is inefficient.
SDC-10148	Improve performance for the HBase destination.
SDC-10062	The JDBC Producer did not convert string data to all compatible data types when writing to the database.
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-8876	Metric alerts send alerts when the pipeline starts instead of waiting for the specified condition.

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-7997	The `record:valueOrDefault` function should return the data type of data being evaluated when not returning the default.
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.

Known Issues in Version 3.7.x

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-9514	Runtime parameters are not supported in all configuration properties in cluster batch execution mode, such as Max Batch Size.
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-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-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.