# StreamSets Data Collector
# Cumulative 3.14.x Release Notes

## March 20, 2020

We're happy to announce a new version of StreamSets Data Collector. Version 3.14.x contains several new features, enhancements, and some important bug fixes in the following release:

- Version 3.14.0 - March 20, 2020

This document contains important information about the following topics:

# New Features and Enhancements

Version 3.14.x includes several new features and enhancements for Data Collector and Data Collector Edge.

## Data Collector New Features and Enhancements

Data Collector version 3.14.x includes new features and enhancements in the following areas.

### Origins

This release includes enhancements to the following origins:

- **Cron Scheduler origin** - The origin is no longer considered a Technology Preview feature and is approved for use in production.

- **HTTP Server origin** - You can now configure the origin either to not require an application ID or to accept requests from one of many application IDs.

- **JDBC Multitable Consumer origin** - The origin now converts SQL Server columns of the Datetimeoffset data type to the Data Collector Zoned_datetime data type.

- **JDBC Query Consumer origin** - The origin now converts SQL Server columns of the Datetimeoffset data type to the Data Collector Zoned_datetime data type.

- **Start Pipeline origin** - The origin is no longer considered a Technology Preview feature and is approved for use in production.

## Processors

This release includes enhancements to the following processors:

- **Control Hub API processor** - The processor is no longer considered a Technology Preview feature and is approved for use in production.

- **Start Job processor** - The processor is no longer considered a Technology Preview feature and is approved for use in production.

- **Start Pipeline processor** - The processor is no longer considered a Technology Preview feature and is approved for use in production.

## Executors

This release includes enhancements to the following executor:

- **JDBC Query executor** - The executor can now run one or more user-defined SQL queries each time it receives an event record. Previously, the executor could only run a single SQL query.

## Data Formats

This release includes the following data formats enhancement:

- **XML** - When an origin reads XML data, you can now configure the origin to include the root element in the generated records. Previously, origins did not include the root element in the generated records.

## Microservice Pipelines

This release includes the following microservice pipeline enhancement:

- **Send Response to Origin destination** - You can now specify response headers that the destination passes to the REST Service origin in a microservice pipeline.

## Working with Control Hub

This release includes the following enhancement when working with Control Hub:

- **Register Data Collector with Control Hub** - Data Collector instances that are deployed behind a reverse proxy using user-defined path-based routes can now successfully be registered with Control Hub.

### Stage Libraries

This release includes the following stage library enhancements:

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

| Stage Library Name | Description |
|---|---|
| streamsets-datacollector-cdh_kafka_4_1-lib | For the Cloudera distribution of Apache Kafka - CDK 4.1.0 (based on Apache Kafka version 2.2.1). |

## Data Collector Edge New Features and Enhancements

This Data Collector Edge (SDC Edge) version 3.14.x includes new features and enhancements in the following areas.

### Destinations in Edge Pipelines

Data Collector Edge pipelines now support the InfluxDB destination.

# Upgrade Information

You can upgrade previous versions of Data Collector to version 3.14.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, the latest available versions, and links to the supported versions and the stage documentation. To view the release notes for Enterprise stage libraries, see the StreamSets Documentation page.

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

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 3.14.0

The following table lists some of the known issues that are fixed in Data Collector version 3.14.0.

For the full list, click here.

| JIRA | Description |
|------|-------------|
| SDC-13770 | A runtime statistics gauge generates an error when the statistics aggregator is enabled. |
| SDC-13623 | The JDBC Multitable Consumer origin reads duplicate data when an initial offset is specified. |
| SDC-13559 | The Directory origin produces continuous batches when configured to read files in a late directory. |
| SDC-13556, SDC-13555 | The JDBC Producer destination and JDBC Tee processor generate empty batches. |
| SDC-13331 | The Kinesis Producer destination generates an error when running in Linux versions based on Debian due to an incorrect library file. |
| SDC-13288 | Error handling for DB2 causes pipelines to fail. |
| SDC-13278 | The Amazon S3 origin has poor performance when there is no data on startup. |
| SDC-13180 | The SFTP/FTP/FTPS Client origin does not create error records when reading JSON data and the number of characters  exceeds the value configured in the Max Object Length (chars) property, possibly leading to data loss.<br><br>Workaround: Configure a higher value in Max Object Length (chars) property. |
| SDC-13117 | The Salesforce destination cannot use the Bulk API to write data with an external ID that references more than one data type. |
| SDC-13098 | Some stage libraries contain security vulnerabilities. |
| SDC-13025 | The SFTP/FTP/FTPS Client origin cannot use read-ahead streaming when in a pipeline that writes to the Amazon S3 destination. |
| SDC-12989 | The Amazon S3 origin generates a data parser error upon network loss. |
| SDC-12953 | After a user forces a pipeline to stop, stages with connection threads sometimes remain running until Data Collector restarts. |

| SDC-12877 | The Azure Data Lake Storage origins do not correctly implement postprocessing. |
|-----------|-------------------------------------------------------------------------------|
| SDC-12842 | The scripting origins and processors do not properly close the scripting engines. |
| SDC-12793 | Data Collector fails to start after adding a Hashicorp Vault configuration using HTTPS. |
| SDC-12746 | The SFTP/FTP/FTPS origin cannot handle rare cases where offset is -1. |
| SDC-12704 | The Salesforce destination generates an error if you specify a blank string for a foreign key and try to upsert it with the SOAP API. |
| SDC-12516 | The Directory and Hadoop FS Standalone origins generate no-such-file exceptions due to a multithreading issue.. |
| SDC-12302 | The Azure stage library uses Hadoop files that generate issues when reading from Azure Data Lake Storage. |

## Known Issues

Please note the following known issues with Data Collector version 3.14.x.

For a full list of known issues, click here.

| JIRA | Description |
|------|-------------|
| SDC-13918 | In some cases, the PostgreSQL CDC Client origin incorrectly synchronizes with Postgres due to offset handling. |
| 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. <br><br> 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-8514 | The Data Parser processor sends a record to the next stage for processing even when the record encounters an error.<br><br>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-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.<br><br>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.`<br><br>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.