StreamSets Data Collector 2.1.0.1 Release Notes

October 17, 2016

We're happy to announce a new version of StreamSets Data Collector. This version features a few important bug fixes.

Upgrade

You can upgrade previous versions of Data Collector to version 2.1.0.1. For instructions on upgrading, see the Upgrade Documentation.

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-4224	Remove limit on the number of oldest SCNs selected in the Oracle CDC Client origin.
SDC-4087	Data preview fails for pipelines that use the Dev Raw Data Source origin when you refresh the data preview or run data preview with changes.
SDC-3911	The Field Renamer processor does not support quoting regex special characters such as the pipe symbol (|) in field names. If you followed the workaround in a previous release to manually quote the special character by wrapping it in \Q and \E, you can remove those characters. For example, to specify a field name of `/'tag|attr'` in a previous release, you had to use the workaround of `/'tag\Q|\Eattr'`. Now you can specify that field name as `/'tag|attr'`.

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-4954	The Cassandra destination encounters problems connecting to a Cassandra cluster because the Cassandra stage library directory contains a mixed version of netty JAR files. Workaround:

© 2016, StreamSets, Inc., Apache License, Version 2.0

StreamSets Data Collector 2.1.0.1 Release Notes

	1. Remove all netty* JAR files from the following directory: `$SDC_DIST/streamsets-libs/streamsets-datacollector-cassandra_3-lib/lib` 2. Download the following netty JAR file: `http://central.maven.org/maven2/io/netty/netty-all/4.0.41.Final/netty-all-4.0.41.Final.jar` 3. Add the netty-all-4.0.41.Final.jar file to the Cassandra stage library directory.
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-4172	Data Collector cannot access Vault secrets stored in Hashicorp Vault.
SDC-4128	In cluster mode, Data Collector does not generate log files for worker Data Collectors.
SDC-4090	The MapR FS destination does not support impersonating an HDFS user. Instead, the destination always uses the user account who started the Data Collector to connect to MapR FS.
SDC-4047	The XML Flattener processor fails to parse XML that contains whitespace after the XML prolog. Workaround: Use an Expression Evaluator or scripting processor to remove the whitespace before using the XML Flattener.
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-3357	If you run Data Collector from Docker, you cannot shut down Data Collector by running `docker stop` or pressing Ctrl+C from the Docker Quickstart Terminal. Workaround: In the Data Collector console, click **Administration** > **Shut Down**.
SDC-3356	Using the following commands to shut down or restart Data Collector does not properly complete the shutdown: • `service sdc stop` • `service sdc restart` Workaround: In the Data Collector console, click **Administration** > **Shut Down** or **Administration** > **Restart**.
SDC-3234	Cluster streaming pipelines that run on YARN use the YARN user instead of the Data Collector user to run executors.
SDC-3133	When you upgrade Data Collector from the RPM package, the environment configuration file `$SDC_DIST/libexec/sdcd-env.sh` is overwritten.

© 2016, StreamSets, Inc., Apache License, Version 2.0

StreamSets Data Collector 2.1.0.1 Release Notes

	Workaround: Back up the `sdcd-env.sh` file before you upgrade.
SDC-2950	When a pipeline writes error records to Elasticsearch, the record header information - error code, error message, and error stage - is not preserved.
SDC-2822	If you configure a Kafka Producer destination to write one message per batch, and then use a cluster pipeline to process that data from the Kafka cluster, the cluster pipeline might encounter an out of memory error.
SDC-2586	To process records larger than 1 MB, you must configure the DataFactoryBuilder.OverRunLimit property. However, this property is not configurable in the Data Collector configuration file, `$SDC_CONF/sdc.properties` Workaround: Set the value of DataFactoryBuilder.OverRunLimit property in the SDC_JAVA_OPTS environment variable in the Data Collector environment file, `$SDC_DIST/libexec/sdc-env.sh` or `$SDC_DIST/libexec/sdcd-env.sh` Set the property greater than the largest record you want to process. For example, to process records up to 2 MB, set the property to 2097152 as follows: `SDC_JAVA_OPTS="-DDataFactoryBuilder.OverRunLimit=2097152"`
SDC-2374	A cluster mode pipeline can hang with a CONNECT_ERROR status. This can be a temporary connection problem that resolves, returning the pipeline to the RUNNING status. If the problem is not temporary, you might need to manually edit the pipeline state file to set the pipeline to STOPPED. Edit the file only after you confirm that the pipeline is no longer running on the cluster or that the cluster has been decommissioned. To manually change the pipeline state, edit the following file: `$SDC_DATA/runInfo/<cluster pipeline name>/<revision>/pipelineState.json` In the file, change CONNECT_ERROR to STOPPED and save the file.
SDC-1731	When using the Kafka Consumer or Kafka Producer on HDP 2.3 with Kerberos enabled, set the Kafka broker configuration property `security.inter.broker.protocol` to PLAINTEXT. When enabling Kerberos, HDP 2.3 sets the `security.inter.broker.protocol` property to PLAINTEXTSASL, which is not supported. If the property is not set to PLAINTEXT, when the pipeline starts, validation errors indicate a problem connecting to Kafka.
SDC-891	At this time, writing to error records to file is not supported for cluster mode pipelines. Workaround: Write error records to Kafka or to an SDC RPC pipeline.
SDC-890	For cluster mode pipelines configured to stop on error or to stop upon reaching a memory limit, the Data Collector cannot stop all worker pipelines as expected. Workaround: To stop all pipelines, use the Stop icon in the Data Collector console.

© 2016, StreamSets, Inc., Apache License, Version 2.0

StreamSets Data Collector 2.1.0.1 Release Notes

Contact Information

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

To review the latest documentation or try out our tutorials, check out the following links:

- User Guide
- User Guide tutorial
- GitHub tutorials

To report an issue, ask for help, or find out about our next meetup, check out our Community page: http://streamsets.com/community/.

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

© 2016, StreamSets, Inc., Apache License, Version 2.0