StreamSets Data Collector 1.6.0.0 Release Notes

August 31, 2016

New Features and Enhancements

We're happy to announce a new version of StreamSets Data Collector. This version features new features and enhancements in the following areas.

Installation and Configuration

- **RPM installation enhancements.** You can now use the RPM package to perform a core installation of Data Collector. You can download and install the core RPM package, and then install individual stage libraries as needed.

- **New stage libraries.** You can now use the following new stage libraries:
  - Apache Kafka 0.10
  - Cassandra 3.x
  - Cloudera CDH 5.8
  - Elasticsearch 2.3.5

Origins

- **New SDC RPC to Kafka origin**. Processes large volumes of data from SDC RPC destinations in other pipelines and writes it immediately to Kafka.

- **New UDP to Kafka origin.** Processes large volumes of data from multiple UDP ports and writes it immediately to Kafka.

- **Amazon S3 origin enhancements.** You can include object metadata in record header attributes and use the origin to transfer whole files with the new whole file data format.

- **HTTP Client origin enhancements.** You can now read data from paginated APIs. The new batch mode enables reading available data and stopping the pipeline, and you can now process compressed and archived files. The origin writes the response header to the record header attributes and allows you to configure the request transfer encoding, connection timeout, and read timeout.

- **MongoDB origin enhancements.** You can now configure advanced options that determine how the origin connects to MongoDB, including enabling SSL.

- **JDBC Consumer origin enhancements.** When you configure the origin to perform a full query, the SQL query no longer requires a WHERE and ORDER BY clause.

Processors

- **New JDBC Lookup processor.** Performs lookups in a database table.

- **New JDBC Tee processor.** Writes data to a database table, and enriches records with data from generated database columns.

StreamSets Data Collector 1.6.0.0 Release Notes

- **New List Pivoter processor.** Pivots a list in a field, generating a new record for each item in the list.

- **Field Type Converter processor enhancements.** The Field Converter processor has been renamed to the Field Type Converter processor. You can now convert the data type of all fields with the specified type.

- **Groovy, JavaScript, and Jython Evaluators can associate nulls with a data type.** When the scripting processors process null values, they return the null value to the pipeline as the original data type. You can use constants in the scripting code to create a new field of a specific data type with a null value.

- **HTTP Client processor enhancements.** You can configure the processor to include the response header in the record as a field or as a set of record header attributes. You can also configure the request transfer encoding to use.

Destinations

- **Amazon S3 destination enhancements.** The Amazon S3 destination can write data asynchronously to improve performance when writing to multiple prefixes. You can tune performance with new advanced properties.

  You can configure the time basis and data time zone used by the Amazon S3 destination to write records to a time-based partition prefix. You can also use the destination to transfer whole files with the new whole file data format.

- **Hadoop FS, Local FS, and MapR FS destination enhancements.** You can use the whole file data format to move whole files. You can also use the binary data format to write binary data in a single field to a file. You can configure the stage to validate permissions when starting the pipeline.

- **MongoDB destination enhancements.** You can now configure advanced options that determine how the destination connects to MongoDB, including enabling SSL and entering the user credentials.

- **Solr destination supports Kerberos authentication.** You can now use Kerberos authentication to connect to a Solr node or cluster.

Data Formats

- **New whole file data format.** The whole file data format enables moving entire files from an origin system to a destination system. You can use the data format with the Directory and Amazon S3 origins and the Amazon S3, Hadoop FS, Local FS, and MapR FS destinations.

- **New datagram data format for Kafka Consumer.** The Kafka Consumer origin can now process datagram data.

- **Null string configuration for delimited data.** You can now replace a string constant with null values in delimited data.

StreamSets Data Collector 1.6.0.0 Release Notes

MapR Support

- **MapR prerequisite enhancements.** You can now run a MapR setup command that modifies configuration files and creates the required symbolic links for you.
- **Hive Drift Solution can write to MapR FS.** You can use the Hive Drift Solution with the MapR FS destination to write data to MapR FS.

Functions

- **New time functions.** Trim the date or time portion of a datetime value.

Please feel free to check out the Documentation for this release.

Upgrade

You can upgrade previous versions of Data Collector to version 1.6.0.0. 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-3700	Disable the TRACE HTTP method in the Data Collector web interface.
SDC-3666	The Geo IP processor does not allow you to configure the action to take when IP addresses are missing in the database file.
SDC-3644	The USER_LIBRARIES_DIR environment variable cannot be set outside of the $SDC_HOME directory.
SDC-3606	The command line interface does not correctly parse the SDC_JAVA_OPTS environment variable.
SDC-3599	The Geo IP processor does not allow city databases to be used to extract country information.
SDC-3551	The Directory origin might skip processing some files when they all arrive at the same time.
SDC-3511	The JDBC Consumer origin does not correctly parse Oracle timestamp fields.

StreamSets Data Collector 1.6.0.0 Release Notes

SDC-3445	When multiple pipelines write JSON data to MapR FS, the pipelines might encounter an exception.
SDC-3139	If a scripting processor includes a sleep, timed wait, or infinite loop, you cannot force the pipeline to stop.

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: 1. Remove all netty* JAR files from the following directory: `$SDC_DIST/streamsets-libs/streamsets-datacollector-cassandra_3-lib/lib` 2. Download the following netty JAR file: `http://central.maven.org/maven2/io/netty/netty-all/4.0.41.Final/netty-all-4.0.41.Final.jar` 3. Add the netty-all-4.0.41.Final.jar file to the Cassandra stage library directory.
SDC-3712	The Hadoop FS origin incorrectly lists MapR as an available stage library.
SDC-3357	If you run Data Collector from Docker, you cannot shut down Data Collector by running `docker stop` or pressing Ctrl+C from the Docker Quickstart Terminal. Workaround: In the Data Collector console, click **Administration** > **Shut Down**.
SDC-3356	Using the following commands to shut down or restart Data Collector does not properly complete the shutdown: • `service sdc stop` • `service sdc restart` Workaround: In the Data Collector console, click **Administration** > **Shut Down** or **Administration** > **Restart**.
SDC-3234	Cluster streaming pipelines that run on YARN use the YARN user instead of the Data Collector user to run executors.
SDC-3133	When you upgrade Data Collector from the RPM package, the environment configuration file `$SDC_DIST/libexec/sdcd-env.sh` is overwritten.

StreamSets Data Collector 1.6.0.0 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 `$SDC_DIST/libexec/sdcd-env.sh` Set the property greater than the largest record you want to process. For example, to process records up to 2 MB, set the property to 2097152 as follows: `SDC_JAVA_OPTS="-DDataFactoryBuilder.OverRunLimit=2097152”`
SDC-2552	When an invalid topic is specified for a MapR Streams Consumer or a MapR Streams Producer, the following incorrect message displays: `CONTAINER_0701 - Stage 'com_streamsets_pipeline_stage_origin_maprstreams_MapRStreamsDSource_1' initialization error: org.apache.kafka.common.config.ConfigException: No bootstrap urls given in bootstrap.servers` This message is misleading because MapR Streams does not support the `bootstrap.servers` option.
SDC-2374	A cluster mode pipeline can hang with a CONNECT_ERROR status. This can be a temporary connection problem that resolves, returning the pipeline to the RUNNING status. If the problem is not temporary, you might need to manually edit the pipeline state file to set the pipeline to STOPPED. Edit the file only after you confirm that the pipeline is no longer running on the cluster or that the cluster has been decommissioned. To manually change the pipeline state, edit the following file: `$SDC_DATA/runInfo/<cluster pipeline name>/<revision>/pipelineState.json` In the file, change CONNECT_ERROR to STOPPED and save the file.
SDC-2359	Due to a Kafka issue, a pipeline with a Kafka Consumer or Kafka Producer can hang during validation or display initialization errors when unable to connect to a Kafka 0.9.0.0 broker. The Data Collector log indicates that the broker is unavailable. For more information about the Kafka JIRA, see

StreamSets Data Collector 1.6.0.0 Release Notes

	https://issues.apache.org/jira/browse/KAFKA-2880.
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-1567	You cannot use cluster mode pipelines to read from HDP 2.3 due to a HDP integration issue with Kafka and Spark Streaming. For more information about the HDP issue, see http://hortonworks.com/community/forums/topic/kafka-and-spark-streaming-nosuchmethoderror-kafka-consumer-simpleconsumer/.
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.

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.