StreamSets Data Collector 1.5.1.2 Release Notes

August 4, 2016

New Features and Enhancements

We're happy to announce a new version of StreamSets Data Collector.

This version includes several bug fixes, described below.

Feel free to check out the Documentation for this release.

Upgrade

You can upgrade a previous version of Data Collector to version 1.5.1.2. For instructions on upgrading, see the Upgrade Documentation.

Fixed Issues

The following table lists the known issues that are fixed with this release.

JIRA	Description
SDC-3616	The Hadoop FS destination incorrectly updates the idle timeout for files that are idle when the destination writes data to other files.
SDC-3602	The Hive Metadata processor and Hive Metastore destination do not correctly route to error when Hive query failures occur.
SDC-3600	The Hive Metadata processor and the Hive Metastore destination use the incorrect field types.
SDC-3588	The Hive Metastore destination uses the incorrect date format for Avro schema serialization.

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-3445	When multiple pipelines write JSON data to MapR FS, the pipelines might encounter an exception.

StreamSets Data Collector 1.5.1.2 Release Notes

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. Workaround: Back up the `sdcd-env.sh` file before you upgrade.
SDC-3017	Hadoop FS does not close and rename open files when the pipeline stops. Open files are named `prefix _tmp`. When the pipeline restarts, Hadoop FS continues processing the files, closing and renaming them as expected. Workaround: If your workflow immediately processes output files from Hadoop FS and your workflow requires synchronized data, pause downstream processing before you stop the pipeline.
SDC-2998	The JDBC Consumer origin does not correctly parse Oracle timestamp fields. Workaround: Set the JVM system property `oracle.jdbc.J2EE13Compliant` to true in the SDC_JAVA_OPTS environment variable in the `sdc-env.sh` or `sdcd-env.sh` file. For example: `export SDC_JAVA_OPTS="-Xmx1024m -Xms1024m -XX:PermSize=128m -XX:MaxPermSize=256m -Doracle.jdbc.J2EE13Compliant=true -server ${SDC_JAVA_OPTS}"` Setting the property to true returns `java.sql.Timestamp` instead of `oracle.sql.TIMESTAMP` for the timestamp SQL type.
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.

StreamSets Data Collector 1.5.1.2 Release Notes

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 https://issues.apache.org/jira/browse/KAFKA-2880.
SDC-2133	You cannot stop a pipeline in the middle of a long-running processor, such as a Jython Evaluator performing a complex script.

StreamSets Data Collector 1.5.1.2 Release Notes

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 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.