StreamSets Data Collector

1.5.x.x Cumulative Release Notes

Updated August 24, 2016

This document includes release notes from 1.5.1.3 to 1.5.0.0 in reverse chronological order. For information about earlier releases, read on…

++

StreamSets Data Collector 1.5.1.3 Release Notes

August 15, 2016

1.5.1.3 New Features and Enhancements

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

This version includes several bug fixes, described below.

1.5.1.3 Upgrade

You can upgrade a previous version of Data Collector to version 1.5.1.3. For instructions on upgrading, see the Upgrade Documentation.

1.5.1.3 Fixed Issues

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

JIRA	Description
SDC-3638	The Hadoop FS destination causes deadlock when idle close is enabled.
SDC-3634	The Hadoop FS destination throws InterruptedException during IdleClose.

1.5.1.3 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.
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.
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.
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.

StreamSets Data Collector 1.5.1.2 Release Notes

August 4, 2016

1.5.1.2 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.

1.5.1.2 Upgrade

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

1.5.1.2 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.

1.5.1.2 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.
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.
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.
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.

StreamSets Data Collector 1.5.1.1 Release Notes

July 27, 2016

1.5.1.1 New Features and Enhancements

We're happy to announce a new version of StreamSets Data Collector. This version includes several important bug fixes, described below.

As always, feel free to check out the Documentation for this release.

1.5.1.1 Upgrade

You can upgrade a previous version of Data Collector to version 1.5.1.1. For instructions on upgrading, see the Upgrade Documentation.

Update Vault Pipelines as Needed

Due to a known issue, in 1.5.0.0, you can use Vault functions to call Vault secrets from within any pipeline or stage property.

To protect the security of sensitive information, calling Vault is now restricted to the following properties:

- Usernames, passwords, and similar properties such as AWS Access Key ID and Secret Access Key.
- HTTP headers and bodies when using HTTPS.

If you are upgrading from 1.5.0.0 to 1.5.1.1, update any pipeline that uses Vault functions in other properties. Remove Vault functions from unsupported properties or the pipeline will fail validation when you validate or start the pipeline.

1.5.1.1 Fixed Issues

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

JIRA	Description
SDC-3553	The HDFS destination might encounter a ConcurrentModificationException when an idle timeout is set.
SDC-3552	The Redis origin returns null for the offset.
SDC-3547	The Redis origin fails after the first batch.

1.5.1.1 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.
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.
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.
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.

StreamSets Data Collector 1.5.1.0 Release Notes

July 22, 2016

1.5.1.0 New Features and Enhancements

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

This version features the following new features and enhancements:

- **Support for Apache Solr 6.1.0.** If Java 8 is installed on the Data Collector machine, you can configure the Solr destination to write data to an Apache Solr 6.1.0 node or cluster.
- **Support for Azure Blob storage using the WASB protocol.** You can now use Data Collector to write directly to Azure HDInsight.
- **Change in Hashicorp Vault support.** You can now access Hashicorp Vault for only connection information such as usernames and passwords, URLs and connection strings, and HTTPS request headers and bodies.

  To protect the security of sensitive information, you can no longer use Vault secrets in the Expression Evaluator or similar properties that might write the information to the data stream. This can require changes to existing pipelines. For more information, see Update Vault Pipelines as Needed.
- **Updates to counter metric rules.** You can now create a counter metric rule to provide an alert based on the number of input records, output records, error records, or stage errors.

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

1.5.1.0 Upgrade

You can upgrade a previous version of Data Collector to version 1.5.1.0. For instructions on upgrading, see the Upgrade Documentation.

Update Vault Pipelines as Needed

Due to a known issue, in 1.5.0.0, you can use Vault functions to call Vault secrets from within any pipeline or stage property.

To protect the security of sensitive information, calling Vault is now restricted to the following properties:

- Usernames, passwords, and similar properties such as AWS Access Key ID and Secret Access Key.
- HTTP headers and bodies when using HTTPS.

After upgrading to 1.5.1.0, update any pipeline that uses Vault functions in other properties. Remove Vault functions from unsupported properties or the pipeline will fail validation when you validate or start the pipeline.

1.5.1.0 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-3442	Remove VaultEL from default ELs. Note: If you used Vault functions in 1.5.0.0, you might need to update your pipelines. Please see the upgrade notes above.
SDC-3418	The Hadoop FS and Local FS destinations create an invalid empty output directory when the pipeline starts.
SDC-3110	When a File Tail origin is configured using the Active File with Reverse Counter naming convention, the origin throws a StringIndexOutOfBoundsException.

1.5.1.0 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.
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 ${SDC_JAVA_OPTS}"` Setting the property to true returns `java.sql.Timestamp` instead of `oracle.sql.TIMESTAMP` for the timestamp SQL type.
SDC-2950	When a pipeline writes error records to Elasticsearch, the record header information - error code, error message, and error stage - is not preserved.
SDC-2822	If you configure a Kafka Producer destination to write one message per batch, and then use a cluster pipeline to process that data from the Kafka cluster, the cluster pipeline might encounter an out of memory error.
SDC-2586	To process records larger than 1 MB, you must configure the DataFactoryBuilder.OverRunLimit property. However, this property is not configurable in the Data Collector configuration file, `$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.
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.

StreamSets Data Collector 1.5.0.0 Release Notes

June 30, 2016

1.5.0.0 New Features and Enhancements

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

This version features the following new features and enhancements:

- **The Hive Drift Solution** detects drift in incoming data and updates the corresponding Hive tables. The solution enables creating and updating Hive tables based on record requirements and writing data to HDFS based on record header attributes. The Hive Drift Solution supports writing only Avro data to HDFS at this time.

  You can use the full functionality of the solution or individual pieces, as needed.

  - **New Hive Metadata processor.** Evaluates record structures and generates Hive metadata to create and update Hive tables as needed. Also embeds write information in record header attributes that Hadoop FS can use to write data to HDFS.
  - **New Hive Metastore destination.** Creates and updates Hive tables as needed.
  - **Hadoop FS dynamic writes based on record headers.** The Hadoop FS destination can write records to HDFS based on record header attributes.

- **New Redis origin to read messages from Redis channels.**

- **New Redis destination to write data to Redis.** You can configure the destination to write data to Redis key-value pairs, or to publish data as messages to a Redis channel.

- **New HTTP Client processor.** Use to send requests to an HTTP resource URL and write the results to a field in the record.

- **Updates to the HTTP Client origin.** You can now configure headers, use a truststore and keystore.

- **JDBC Consumer origin can generate JDBC namespace header attributes.** The attributes provide the source table names for the record, original SQL data types for each field, and the precision and scale for numeric and decimal fields.

- **Kafka Consumer origin includes record header attributes.** The attributes provide the offset, partition, and topic for each record.

- **Amazon S3 destination and Server-Side Encryption (SSE).** You can configure the destination to use Amazon Web Services server-side encryption to encrypt data written to Amazon S3.

- **Support for the Cloudera distribution of Apache Kafka 2.0.1 (0.9.0).**

- **Integration with Hashicorp Vault.** You can now access sensitive information, such as user credentials, that you have stored in Vault.

- **Regular expressions enabled in the Field Renamer processor.** You can use regex to rename sets of fields.

- **Destinations that process Avro data allow loading schemas from record headers.**

- **New functions:**

    - **sdc:hostname()**. Returns the host name of the Data Collector machine.
    - **Base64 functions**. Encodes and decodes strings and byte array fields.
    - **Math functions.** Provides several math functions such as ABS, ROUND, MIN and MAX.
    - **record:attributeOrDefault.** Returns the attribute value or a default value when the attribute is missing.
- **New Time Data Collector data type.**
- **Dev Data Generator update.** You can now configure precision and scale for Decimal data in this development stage.

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

1.5.0.0 Upgrade

You can upgrade a previous version of Data Collector to version 1.5.0.0. For instructions on upgrading, see the Upgrade Documentation.

1.5.0.0 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-3218	The time portion of timestamps are dropped by the JDBC Producer.
SDC-3171	The virtual file system client in the SFTP/FTP Client origin is not thread-safe.
SDC-3166	Support explicit evaluation of ELs in the HTTP Client origin for Polling mode.
SDC-3160	Kinesis Producer destination should send records larger than 1MB to error.
SDC-3152	The JVM Metrics page does not display graphics under certain conditions.
SDC-3122	HTTP Client origin should provide explicit authentication settings.
SDC-3105	Elasticsearch destination does not allow HTTPS.

1.5.0.0 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-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-3110	When a File Tail origin is configured using the Active File with Reverse Counter naming convention, the origin throws a StringIndexOutOfBoundsException.
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.
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.
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.