StreamSets Data Collector
Cumulative 2.5.x.x Release Notes

****

StreamSets Data Collector 2.5.1.1 Release Notes

May 11, 2017

We're happy to announce a new version of StreamSets Data Collector. This release includes an important fix to the previous release.

This document contains important information about the following topics for this release:

Important: You must migrate to Java 8 with this version of Data Collector. Data Collector no longer runs on Java 7 starting with version 2.5.0.0.

Upgrading to Version 2.5.1.1

You can upgrade previous versions of Data Collector to version 2.5.1.1. For instructions on upgrading, see the Upgrade Documentation.

Review Upgraded Pipelines that Write to CDH 5.11

Due to a Cloudera behavior change, when upgrading to CDH 5.11 from a previous version, you must update pipelines that set permissions on HDFS or Hive by modifying file mode bits with the minus or equals operators.

Pipelines can modify file mode bits on HDFS or Hive with the following stage properties:

- The HDFS File Metadata executor Set Permissions property
- The Hadoop FS destination whole file Permissions Expression whole file property

CDH 5.11 changes how the minus and equals operators are evaluated as follows:

- In previous CDH releases, the minus operator (-) grants the specified permissions. In the current release, it removes the specified permissions.

  For example, in previous releases, `a-rw` grants read and write permissions to all users. With CDH 5.11, it removes read and write permissions from all users.

- In earlier CDH releases, the equals operator (=) removes the specified permissions. In the current release, it grants the specified permissions.

  For example, in previous releases, `a=we` removes write and execute permission from all users. With CDH 5.11, it grants write and execute permission to all users.

To ensure that file permissions are set as expected, update all properties in upgraded pipelines that modify file mode bits with the minus or equals operators.

This behavior change is noted in the Cloudera documentation regarding the fix for HADOOP-13508.

Migrate to Java 8

As of version 2.5.0.0, Data Collector requires Java 8. If your previous Data Collector version ran on Java 7, you must migrate to Java 8 before upgrading to the latest Data Collector version. For instructions, see Pre-Upgrade Tasks.

All services that use Data Collector JAR files also must run on Java 8. This means that your Hadoop cluster must run on Java 8 if you are using cluster pipelines, the Spark Executor, or the MapReduce Executor.

Upgrade Cluster Streaming Pipelines

If you use cluster pipelines that run in cluster streaming mode and you are upgrading from a version earlier than 2.3.0.0, you must upgrade to Data Collector version 2.3.0.0 before upgrading to the latest version.

Prior to 2.3.0.0, Data Collector used the Spark checkpoint mechanism to recover cluster pipelines after a failure. Starting in version 2.3.0.0, Data Collector maintains the state of cluster pipelines without relying on Spark checkpoints.

For more information, see Pre-Upgrade Tasks.

Precondition Error Handling

With Data Collector version 2.5.0.0, precondition error handling has changed.

The Precondition stage property allows you to define conditions that must be met for a record to enter the stage. Previously, records that did not meet all specified preconditions were passed to the pipeline for error handling. That is, the records were processed based on the Error Records pipeline property.

With this release, records that do not meet the specified preconditions are handled by the error handling configured for the stage. Stage error handling occurs based on the On Record Error property on the General tab of the stage.

Review pipelines that use preconditions to verify that this change does not adversely affect the behavior of the pipelines.

Configure JDBC Producer Schema Names

With Data Collector version 2.5.0.0, you can use a Schema Name property to specify the database or schema name. In previous releases, you specified the database or schema name in the Table Name property.

Upgrading from a previous release does not require changing any existing configuration at this time. But we recommend using the new Schema Name property, since the ability to specify a database or schema name with the table name might be deprecated in the future.

Authentication for the Docker Image

As of version 2.4.1.0, the Docker image for Data Collector now uses the form type of file-based authentication by default. As a result, you must use a Data Collector user account to log in to the Data Collector. If you haven't set up custom user accounts, you can use the admin account shipped with the Data Collector. The default login is: admin / admin.

Earlier versions of the Docker image used no authentication.

Configure Pipeline Permissions

Data Collector version 2.4.0.0 is designed for multitenancy and enables you to share and grant permissions on pipelines. Permissions determine the access level that users and groups have on pipelines.

In earlier versions of Data Collector without pipeline permissions, pipeline access is determined by roles. For example, any user with the Creator role could edit any pipeline.

In version 2.4.0.0, roles are augmented with pipeline permissions. In addition to having the necessary role, users must also have the appropriate permissions for the given pipeline to perform pipeline tasks.

For example, to edit a pipeline in 2.4.0.0, a user with the Creator role must also have read and write permission on the pipeline. Without write permission, the user cannot edit the pipeline. Without read permission, the user cannot see the pipeline at all. It does not display in the list of available pipelines.

Note: With pipeline permissions enabled, all upgraded pipelines are initially visible only to users with the Admin role and the pipeline owner - the user who created the pipeline. To enable other users to work with pipelines, have an Admin user configure the appropriate permissions for each pipeline.

In Data Collector version 2.5.0.0, pipeline permissions are disabled by default. To enable pipeline permissions, set the pipeline.access.control.enabled property to true in the Data Collector configuration file.

Tip: You can configure pipeline permissions when permissions are disabled. Then, you can enable the pipeline permissions property after pipeline permissions are properly configured.

For more information about roles and permissions, see Roles and Permissions. For details about configuring pipeline permissions, see Sharing Pipelines.

Update Elasticsearch Pipelines

As of version 2.3.0.0, Data Collector includes an enhanced Elasticsearch destination that uses the Elasticsearch HTTP API.

Due to SDC-5148, upgraded Elasticsearch destinations have the Default Operation property set based on the configuration for the Enable Upsert property:
- With upsert enabled, the default operation is set to INDEX.
- With upsert not enabled, the default operation is set to CREATE which requires a Document ID.

Review all upgraded Elasticsearch pipelines to ensure the Default Operation is set to the correct operation.

For more information, see Update Elasticsearch Pipelines.

Update Kudu Pipelines

As of version 2.2.0.0, Data Collector supports Apache Kudu version 1.0.x. Earlier Kudu versions are no longer supported. To upgrade pipelines that contain a Kudu destination, upgrade your Kudu cluster to version 1.0.x and then add a stage alias for the earlier Kudu version to the Data Collector configuration file, `$SDC_CONF/sdc.properties`. For more information, see Update Kudu Pipelines.

Fixed Issues in 2.5.1.1

The following table lists the known issue fixed with this release.

JIRA	Description
SDC-6121	A problem with the Hadoop FS origin prevents using it in cluster batch execution mode.

Known Issues in 2.5.1.1

Please note the following known issues with this release.

For a full list of known issues, check out our JIRA.

JIRA	Description
SDC-6077	The Field Remover processor does not remove list fields from list-map data.
SDC-6058	Data Collector does not always track the offset correctly within compressed files. If the pipeline stops when an origin is reading from within a compressed file, upon restart, it might reprocess the entire file.
SDC-5902	When the Directory origin is configured to use the last-modified timestamp and the file post-processing option is set for Delete or Archive, the origin deletes or moves all existing files in the directory when you start the pipeline and fails to process those files. Workaround: Do not use the Delete or Archive post-processing option.
SDC-5871	Attempting to write a record to Kudu with a null primary key causes the pipeline to fail.
SDC-5818	The UDP Source origin can generate inaccurate information for the Timestamp, First, and Last fields when reading Netflow messages.
SDC-5758	When configured to use Kerberos authentication, Data Collector cannot connect to Kudu using the Kudu 1.1 or 1.2 stage libraries. This may be a Kudu issue. Possible workaround: In the pipeline, try using the Kudu 1.3 stage library.

SDC-5521	Kerberos-enabled pipelines that are valid using an Apache Kafka stage library can fail validation when using a Cloudera distribution of Kafka. The Apache Kafka client libraries have been modified by StreamSets to allow connectivity to Kerberized Kafka without requiring a JAAS configuration file. The Apache Kafka stage libraries (version 0.9 and greater) do not require a JAAS configuration file when enabling Kerberos. For CDH Kafka versions, a JAAS configuration is still required. Workaround: Include a JAAS configuration file on the classpath by modifying the `sdc-env.sh` or `sdcd-env.sh` file to include the following option in the SDC_JAVA_OPTS environment variable: `-Djava.security.auth.login.config=<path-to-jaas-config>` Or, you can use the matching version Apache Kafka stage library, instead of the Cloudera stage libraries.
SDC-5357	The Jython Evaluator and stages that can access the file system - such as the Directory and File Tail origins - can access all pipeline JSON files stored in the $SDC_DATA directory. This allows users to access pipelines that they might not have permission to access within Data Collector. Workaround: To secure your pipelines, complete the following tasks: 1. Remove the Jython stage library and use the Groovy Evaluator or JavaScript Evaluator processor instead of the Jython Evaluator. 2. Update the Data Collector security policy file, `$SDC_CONF/sdc-security.policy`, so that Data Collector stages do not have `AllPermission` access to the file system. Update the security policy for the following code bases: streamsets-libs-extras, streamsets-libs, and streamsets-datacollector-dev-lib. Use the policy file syntax to set the security policies.
SDC-5325	Cluster mode pipelines that read from a MapR cluster fail when the MapR cluster uses username/password login authentication. Workaround: On the Cluster tab for the pipeline, add the following Java property to the Worker Java Options property: `-Dmaprlogin.password.enabled=true`
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-5039	When you use the Hadoop FS origin to read files from all subdirectories, the origin cannot use the configured Hadoop FS User as a proxy user to read from HDFS. Workaround: If you need to use a proxy user to read from all subdirectories of the specified directories, set the HADOOP_PROXY_USER environment variable to the proxy user in libexec/_cluster-manager script, as follows: `export HADOOP_PROXY_USER = <proxy-user>`

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-3944	The Hive Streaming destination using the MapR library cannot connect to a MapR cluster that uses Kerberos or username/password login authentication.
SDC-3133	When you upgrade Data Collector from the RPM package, the environment configuration file `$SDC_DIST/libexec/sdcd-env.sh` is overwritten. Workaround: Back up the `sdcd-env.sh` file before you upgrade.
SDC-2374	A cluster mode pipeline can hang with a CONNECT_ERROR status. This can be a temporary connection problem that resolves, returning the pipeline to the RUNNING status. If the problem is not temporary, you might need to manually edit the pipeline state file to set the pipeline to STOPPED. Edit the file only after you confirm that the pipeline is no longer running on the cluster or that the cluster has been decommissioned. To manually change the pipeline state, edit the following file: `$SDC_DATA/runInfo/<cluster pipeline name>/<revision>/pipelineState.json` In the file, change CONNECT_ERROR to STOPPED and save the file.

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, ask for help, or find out about our next meetup, check out our Community page:
https://streamsets.com/community/.

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

StreamSets Data Collector 2.5.1.0 Release Notes

May 10, 2017

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

This document contains important information about the following topics for this release:

Important: You must migrate to Java 8 with this version of Data Collector. Data Collector no longer runs on Java 7 starting with version 2.5.0.0.

Upgrading to Version 2.5.1.0

You can upgrade previous versions of Data Collector to version 2.5.1.0. For instructions on upgrading, see the Upgrade Documentation.

Review Upgraded Pipelines that Write to CDH 5.11

Due to a Cloudera behavior change, when upgrading to CDH 5.11 from a previous version, you must update pipelines that set permissions on HDFS or Hive by modifying file mode bits with the minus or equals operators.

Pipelines can modify file mode bits on HDFS or Hive with the following stage properties:

- The HDFS File Metadata executor Set Permissions property
- The Hadoop FS destination whole file Permissions Expression whole file property

CDH 5.11 changes how the minus and equals operators are evaluated as follows:

- In previous CDH releases, the minus operator (-) grants the specified permissions. In the current release, it removes the specified permissions.

  For example, in previous releases, `a-rw` grants read and write permissions to all users. With CDH 5.11, it removes read and write permissions from all users.

- In earlier CDH releases, the equals operator (=) removes the specified permissions. In the current release, it grants the specified permissions.

  For example, in previous releases, `a=we` removes write and execute permission from all users. With CDH 5.11, it grants write and execute permission to all users.

To ensure that file permissions are set as expected, update all properties in upgraded pipelines that modify file mode bits with the minus or equals operators.

This behavior change is noted in the Cloudera documentation regarding the fix for HADOOP-13508.

Migrate to Java 8

As of version 2.5.0.0, Data Collector requires Java 8. If your previous Data Collector version ran on Java 7, you must migrate to Java 8 before upgrading to the latest Data Collector version. For instructions, see Pre-Upgrade Tasks.

All services that use Data Collector JAR files also must run on Java 8. This means that your Hadoop cluster must run on Java 8 if you are using cluster pipelines, the Spark Executor, or the MapReduce Executor.

Upgrade Cluster Streaming Pipelines

If you use cluster pipelines that run in cluster streaming mode and you are upgrading from a version earlier than 2.3.0.0, you must upgrade to Data Collector version 2.3.0.0 before upgrading to the latest version.

Prior to 2.3.0.0, Data Collector used the Spark checkpoint mechanism to recover cluster pipelines after a failure. Starting in version 2.3.0.0, Data Collector maintains the state of cluster pipelines without relying on Spark checkpoints.

For more information, see Pre-Upgrade Tasks.

Precondition Error Handling

With Data Collector version 2.5.0.0, precondition error handling has changed.

The Precondition stage property allows you to define conditions that must be met for a record to enter the stage. Previously, records that did not meet all specified preconditions were passed to the pipeline for error handling. That is, the records were processed based on the Error Records pipeline property.

With this release, records that do not meet the specified preconditions are handled by the error handling configured for the stage. Stage error handling occurs based on the On Record Error property on the General tab of the stage.

Review pipelines that use preconditions to verify that this change does not adversely affect the behavior of the pipelines.

Configure JDBC Producer Schema Names

With Data Collector version 2.5.0.0, you can use a Schema Name property to specify the database or schema name. In previous releases, you specified the database or schema name in the Table Name property.

Upgrading from a previous release does not require changing any existing configuration at this time. But we recommend using the new Schema Name property, since the ability to specify a database or schema name with the table name might be deprecated in the future.

Authentication for the Docker Image

As of version 2.4.1.0, the Docker image for Data Collector now uses the form type of file-based authentication by default. As a result, you must use a Data Collector user account to log in to the Data

Collector. If you haven't set up custom user accounts, you can use the admin account shipped with the Data Collector. The default login is: admin / admin.

Earlier versions of the Docker image used no authentication.

Configure Pipeline Permissions

Data Collector version 2.4.0.0 is designed for multitenancy and enables you to share and grant permissions on pipelines. Permissions determine the access level that users and groups have on pipelines.

In earlier versions of Data Collector without pipeline permissions, pipeline access is determined by roles. For example, any user with the Creator role could edit any pipeline.

In version 2.4.0.0, roles are augmented with pipeline permissions. In addition to having the necessary role, users must also have the appropriate permissions for the given pipeline to perform pipeline tasks.

For example, to edit a pipeline in 2.4.0.0, a user with the Creator role must also have read and write permission on the pipeline. Without write permission, the user cannot edit the pipeline. Without read permission, the user cannot see the pipeline at all. It does not display in the list of available pipelines.

Note: With pipeline permissions enabled, all upgraded pipelines are initially visible only to users with the Admin role and the pipeline owner - the user who created the pipeline. To enable other users to work with pipelines, have an Admin user configure the appropriate permissions for each pipeline.

In Data Collector version 2.5.0.0, pipeline permissions are disabled by default. To enable pipeline permissions, set the pipeline.access.control.enabled property to true in the Data Collector configuration file.

Tip: You can configure pipeline permissions when permissions are disabled. Then, you can enable the pipeline permissions property after pipeline permissions are properly configured.

For more information about roles and permissions, see Roles and Permissions. For details about configuring pipeline permissions, see Sharing Pipelines.

Update Elasticsearch Pipelines

As of version 2.3.0.0, Data Collector includes an enhanced Elasticsearch destination that uses the Elasticsearch HTTP API.

Due to SDC-5148, upgraded Elasticsearch destinations have the Default Operation property set based on the configuration for the Enable Upsert property:
- With upsert enabled, the default operation is set to INDEX.
- With upsert not enabled, the default operation is set to CREATE which requires a Document ID.

Review all upgraded Elasticsearch pipelines to ensure the Default Operation is set to the correct operation.

For more information, see Update Elasticsearch Pipelines.

Update Kudu Pipelines

As of version 2.2.0.0, Data Collector supports Apache Kudu version 1.0.x. Earlier Kudu versions are no longer supported. To upgrade pipelines that contain a Kudu destination, upgrade your Kudu cluster

to version 1.0.x and then add a stage alias for the earlier Kudu version to the Data Collector configuration file, `$SDC_CONF/sdc.properties`. For more information, see Update Kudu Pipelines.

New Features and Enhancements in 2.5.1.0

This version includes the following enhancement:

- **New stage library** - Data Collector now supports the Cloudera CDH version 5.11 distribution of Hadoop and the Cloudera version 5.11 distribution of Apache Kafka 2.1.

Fixed Issues in 2.5.1.0

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

For the full list, click here.

JIRA	Description
SDC-5910	The 2.5.0.0 version of the Cloudera Manager CSD does not include stage aliases. When upgrading from a previous release using Cloudera Manager, certain stages might not appear in existing pipelines. Similarly, importing pipelines from a previous release can fail.
SDC-5904	When the Hadoop FS, MapR FS, or Local FS destination tries to recover partially-written files after a pipeline unexpectedly stops, it expects a trailing delimiter at the end of the directory template.
SDC-5903	The HADOOPFS_59 error message does not include information about the root cause of the problem.
SDC-5757	JDBC Producer encloses all column names in quotation marks in queries. For MySQL databases without ANSI_QUOTES enabled, the destination generates JDBC-14 errors.

Known Issues in 2.5.1.0

Please note the following known issues with this release.

For a full list of known issues, check out our JIRA.

JIRA	Description
SDC-6077	The Field Remover processor does not remove list fields from list-map data.
SDC-6058	Data Collector does not always track the offset correctly within compressed files. If the pipeline stops when an origin is reading from within a compressed file, upon restart, it might reprocess the entire file.

SDC-5902	When the Directory origin is configured to use the last-modified timestamp and the file post-processing option is set for Delete or Archive, the origin deletes or moves all existing files in the directory when you start the pipeline and fails to process those files. Workaround: Do not use the Delete or Archive post-processing option.
SDC-5871	Attempting to write a record to Kudu with a null primary key causes the pipeline to fail.
SDC-5818	The UDP Source origin can generate inaccurate information for the Timestamp, First, and Last fields when reading Netflow messages.
SDC-5758	When configured to use Kerberos authentication, Data Collector cannot connect to Kudu using the Kudu 1.1 or 1.2 stage libraries. This may be a Kudu issue. Possible workaround: In the pipeline, try using the Kudu 1.3 stage library.
SDC-5521	Kerberos-enabled pipelines that are valid using an Apache Kafka stage library can fail validation when using a Cloudera distribution of Kafka. The Apache Kafka client libraries have been modified by StreamSets to allow connectivity to Kerberized Kafka without requiring a JAAS configuration file. The Apache Kafka stage libraries (version 0.9 and greater) do not require a JAAS configuration file when enabling Kerberos. For CDH Kafka versions, a JAAS configuration is still required. Workaround: Include a JAAS configuration file on the classpath by modifying the `sdc-env.sh` or `sdcd-env.sh` file to include the following option in the SDC_JAVA_OPTS environment variable: `-Djava.security.auth.login.config=<path-to-jaas-config>` Or, you can use the matching version Apache Kafka stage library, instead of the Cloudera stage libraries.
SDC-5357	The Jython Evaluator and stages that can access the file system - such as the Directory and File Tail origins - can access all pipeline JSON files stored in the $SDC_DATA directory. This allows users to access pipelines that they might not have permission to access within Data Collector. Workaround: To secure your pipelines, complete the following tasks: 3. Remove the Jython stage library and use the Groovy Evaluator or JavaScript Evaluator processor instead of the Jython Evaluator. 4. Update the Data Collector security policy file, `$SDC_CONF/sdc-security.policy`, so that Data Collector stages do not have `AllPermission` access to the file system. Update the security policy for the following code bases: streamsets-libs-extras, streamsets-libs, and streamsets-datacollector-dev-lib. Use the policy file syntax to set the security policies.
SDC-5325	Cluster mode pipelines that read from a MapR cluster fail when the MapR cluster uses username/password login authentication.

	Workaround: On the Cluster tab for the pipeline, add the following Java property to the Worker Java Options property: `-Dmaprlogin.password.enabled=true`
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-5039	When you use the Hadoop FS origin to read files from all subdirectories, the origin cannot use the configured Hadoop FS User as a proxy user to read from HDFS. Workaround: If you need to use a proxy user to read from all subdirectories of the specified directories, set the HADOOP_PROXY_USER environment variable to the proxy user in libexec/_cluster-manager script, as follows: `export HADOOP_PROXY_USER = <proxy-user>`
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-3944	The Hive Streaming destination using the MapR library cannot connect to a MapR cluster that uses Kerberos or username/password login authentication.
SDC-3133	When you upgrade Data Collector from the RPM package, the environment configuration file `$SDC_DIST/libexec/sdcd-env.sh` is overwritten. Workaround: Back up the `sdcd-env.sh` file before you upgrade.
SDC-2374	A cluster mode pipeline can hang with a CONNECT_ERROR status. This can be a temporary connection problem that resolves, returning the pipeline to the RUNNING status. If the problem is not temporary, you might need to manually edit the pipeline state file to set the pipeline to STOPPED. Edit the file only after you confirm that the pipeline is no longer running on the cluster or that the cluster has been decommissioned. To manually change the pipeline state, edit the following file: `$SDC_DATA/runInfo/<cluster pipeline name>/<revision>/pipelineState.json` In the file, change CONNECT_ERROR to STOPPED and save the file.

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, ask for help, or find out about our next meetup, check out our Community page:
https://streamsets.com/community/.

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

StreamSets Data Collector 2.5.0.0 Release Notes

April 18, 2017

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

This document contains important information about the following topics for this release:

Important: You must migrate to Java 8 with this version of Data Collector. Data Collector no longer runs on Java 7 starting with version 2.5.0.0.

Upgrading to Version 2.5.0.0

You can upgrade previous versions of Data Collector to version 2.5.0.0. For instructions on upgrading, see the Upgrade Documentation.

Migrate to Java 8

As of version 2.5.0.0, Data Collector requires Java 8. If your previous Data Collector version ran on Java 7, you must migrate to Java 8 before upgrading to Data Collector version 2.5.0.0. For instructions, see Pre-Upgrade Tasks.

All services that use Data Collector JAR files also must run on Java 8. This means that your Hadoop cluster must run on Java 8 if you are using cluster pipelines, the Spark Executor, or the MapReduce Executor.

Upgrade Cluster Streaming Pipelines

If you use cluster pipelines that run in cluster streaming mode and you are upgrading from a version earlier than 2.3.0.0, you must upgrade to Data Collector version 2.3.0.0 before upgrading to the latest version.

Prior to 2.3.0.0, Data Collector used the Spark checkpoint mechanism to recover cluster pipelines after a failure. Starting in version 2.3.0.0, Data Collector maintains the state of cluster pipelines without relying on Spark checkpoints.

For more information, see Pre-Upgrade Tasks.

Precondition Error Handling

With Data Collector version 2.5.0.0, precondition error handling has changed.

The Precondition stage property allows you to define conditions that must be met for a record to enter the stage. Previously, records that did not meet all specified preconditions were passed to the pipeline for error handling. That is, the records were processed based on the Error Records pipeline property.

With this release, records that do not meet the specified preconditions are handled by the error handling configured for the stage. Stage error handling occurs based on the On Record Error property on the General tab of the stage.

Review pipelines that use preconditions to verify that this change does not adversely affect the behavior of the pipelines.

Configure JDBC Producer Schema Names

With Data Collector version 2.5.0.0, you can use a Schema Name property to specify the database or schema name. In previous releases, you specified the database or schema name in the Table Name property.

Upgrading from a previous release does not require changing any existing configuration at this time. But we recommend using the new Schema Name property, since the ability to specify a database or schema name with the table name might be deprecated in the future.

Authentication for the Docker Image

As of version 2.4.1.0, the Docker image for Data Collector now uses the form type of file-based authentication by default. As a result, you must use a Data Collector user account to log in to the Data Collector. If you haven't set up custom user accounts, you can use the admin account shipped with the Data Collector. The default login is: admin / admin.

Earlier versions of the Docker image used no authentication.

Configure Pipeline Permissions

Data Collector version 2.4.0.0 is designed for multitenancy and enables you to share and grant permissions on pipelines. Permissions determine the access level that users and groups have on pipelines.

In earlier versions of Data Collector without pipeline permissions, pipeline access is determined by roles. For example, any user with the Creator role could edit any pipeline.

In version 2.4.0.0, roles are augmented with pipeline permissions. In addition to having the necessary role, users must also have the appropriate permissions for the given pipeline to perform pipeline tasks.

For example, to edit a pipeline in 2.4.0.0, a user with the Creator role must also have read and write permission on the pipeline. Without write permission, the user cannot edit the pipeline. Without read permission, the user cannot see the pipeline at all. It does not display in the list of available pipelines.

Note: With pipeline permissions enabled, all upgraded pipelines are initially visible only to users with the Admin role and the pipeline owner - the user who created the pipeline. To enable other users to work with pipelines, have an Admin user configure the appropriate permissions for each pipeline.

In Data Collector version 2.5.0.0, pipeline permissions are disabled by default. To enable pipeline permissions, set the pipeline.access.control.enabled property to true in the Data Collector configuration file.

Tip: You can configure pipeline permissions when permissions are disabled. Then, you can enable the pipeline permissions property after pipeline permissions are properly configured.

For more information about roles and permissions, see Roles and Permissions. For details about configuring pipeline permissions, see Sharing Pipelines.

Update Elasticsearch Pipelines

As of version 2.3.0.0, Data Collector includes an enhanced Elasticsearch destination that uses the Elasticsearch HTTP API.

Due to SDC-5148, upgraded Elasticsearch destinations have the Default Operation property set based on the configuration for the Enable Upsert property:
- With upsert enabled, the default operation is set to INDEX.
- With upsert not enabled, the default operation is set to CREATE which requires a Document ID.

Review all upgraded Elasticsearch pipelines to ensure the Default Operation is set to the correct operation.

For more information, see Update Elasticsearch Pipelines.

Update Kudu Pipelines

As of version 2.2.0.0, Data Collector supports Apache Kudu version 1.0.x. Earlier Kudu versions are no longer supported. To upgrade pipelines that contain a Kudu destination, upgrade your Kudu cluster to version 1.0.x and then add a stage alias for the earlier Kudu version to the Data Collector configuration file, `$SDC_CONF/sdc.properties`. For more information, see Update Kudu Pipelines.

New Features and Enhancements in 2.5.0.0

This version includes the following new features and enhancements in the following areas.

Multithreaded Pipelines

The multithreaded framework includes the following enhancements:

- **Origins for multithreaded pipelines** - You can now use the following origins to create multithreaded pipelines:
    - Elasticsearch origin
    - JDBC Multitable Consumer origin
    - Kinesis Consumer origin

    - WebSocket Server origin

- **Maximum pipeline runners** - You can now configure a maximum number of pipeline runners to use in a pipeline. Previously, Data Collector generated pipeline runners based on the number of threads created by the origin. This allows you to tune performance and resource usage. By default, Data Collector still generates runners based on the number of threads that the origin uses.

- **Record Deduplicator processor enhancement** - The processor can now deduplicate records across all pipeline runners in a multithreaded pipeline.

- **Pipeline validation enhancement** - The pipeline now displays duplicate errors generated by using multiple threads as one error message.

- **Log enhancement** - Multithreaded pipelines now include the runner ID in log information.

- **Monitoring** - Monitoring now displays a histogram of available pipeline runners, replacing the information previously included in the Runtime Statistics list.

Pipelines

- **Data Collector pipeline permissions change -** With this release, pipeline permissions are no longer enabled by default. To enable pipeline permissions, edit the pipeline.access.control.enabled Data Collector configuration property.

- **Stop pipeline execution** - You can configure pipelines to transfer data and automatically stop execution based on an event such as reaching the end of a table. The JDBC and Salesforce origins can generate events when they reach the end of available data that the Pipeline Finisher uses to stop the pipeline. Click here for a case study.

- **Pipeline runtime parameters** - You can now define runtime parameters when you configure a pipeline, and then call the parameters from within that pipeline. When you start the pipeline from the user interface, the command line, or the REST API, you specify the values to use for those parameters. Use pipeline parameters to represent any stage or pipeline property with a value that must change for each pipeline run - such as batch sizes and timeouts, directories, or URI.

  In previous versions, pipeline runtime parameters were named pipeline constants. You defined the constant values in the pipeline, and could not pass different values when you started the pipeline.

- **Pipeline ID enhancement** - Data Collector now prefixes the pipeline ID with the alphanumeric characters entered for the pipeline title. For example, if you enter “Oracle To HDFS” as the pipeline title, then the pipeline ID has the following value: OracleToHDFStad9f592-5f02-4695-bb10-127b2e41561c.

- **Webhooks for pipeline state changes and alerts** - You can now configure pipeline state changes and metric and data alerts to call webhooks in addition to sending email. For example, you can configure an incoming webhook in Slack so that an alert can be posted to a Slack channel. Or, you can configure a webhook to start another pipeline when the pipeline state is changed to Finished or Stopped.

- **Force a pipeline to stop from the command line** - If a pipeline remains in a Stopping state, you can now use the command line to force stop the pipeline immediately.

Stage Libraries

Data Collector now supports the Apache Kudu version 1.3.x. stage library.

Salesforce Stages

The following Salesforce stages include several enhancements:

- **Salesforce origin and Salesforce Lookup processor**
  - The origin and processor can use a proxy to connect to Salesforce.
  - You can now specify SELECT * FROM <object> in a SOQL query. The origin or processor expands * to all fields in the Salesforce object that are accessible to the configured user.
  - The origin and processor generate Salesforce field attributes that provide additional information about each field, such as the data type of the Salesforce field.
  - The origin and processor can now additionally retrieve deleted records from the Salesforce recycle bin.
  - The origin can now generate events when it completes processing all available data.

- **Salesforce destination** - The destination can now use a CRUD operation record header attribute to indicate the operation to perform for each record. You can also configure the destination to use a proxy to connect to Salesforce.

- **Wave Analytics destination** - You can now configure the authentication endpoint and the API version that the destination uses to connect to Salesforce Wave Analytics. You can also configure the destination to use a proxy to connect to Salesforce.

Origins

- **New Elasticsearch origin** - An origin that reads data from an Elasticsearch cluster. The origin uses the Elasticsearch scroll API to read documents using a user-defined Elasticsearch query. The origin performs parallel processing and can generate multithreaded pipelines.

- **New MQTT Subscriber origin** - An origin that subscribes to a topic on an MQTT broker to read messages from the broker.

- **New WebSocket Server origin** - An origin that listens on a WebSocket endpoint and processes the contents of all authorized WebSocket requests. The origin performs parallel processing and can generate multithreaded pipelines.

- **Dev Data Generator origin enhancement** - When you configure the origin to generate events to test event handling functionality, you can now specify the event type to use.

- **HTTP Client origin enhancements** - When using pagination, the origin can include all response fields in the resulting record in addition to the fields in the specified result field path. The origin can now also process the following new data formats: Binary, Delimited, Log, and SDC Record.

- **HTTP Server origin enhancement** - The origin requires that HTTP clients include the application ID in all requests. You can now configure HTTP clients to send data to a URL that

includes the application ID in a query parameter, rather than including the application ID in request headers.

- **JDBC Multitable Consumer origin enhancement** - The origin now performs parallel processing and can generate multithreaded pipelines. The origin can generate events when it completes processing all available data.
  You can also configure the quote character to use around table, schema, and column names in the query. And you can configure the number of times a thread tries to read a batch of data after receiving an SQL error.

- **JDBC Query Consumer origin enhancement** - The origin can now generate events when it completes processing all available data, and when it successfully completes or fails to complete a query.
  To handle transient connection or network errors, you can now specify how many times the origin should retry a query before stopping the pipeline.

- **Kinesis Consumer origin enhancement** - The origin now performs parallel processing and can generate multithreaded pipelines.

- **MongoDB origin and MongoDB Oplog origin enhancements** - The origins can now use LDAP authentication in addition to username/password authentication to connect to MongoDB. You can also now include credentials in the MongoDB connection string.

Processors

- **New Field Order processor** - A processor that orders fields in a map or list-map field and outputs the fields into a list-map or list root field.

- **Field Flattener enhancement** - You can now flatten a field in place to raise it to the parent level.

- **Groovy, JavaScript, and Jython Evaluator processor enhancement** - You can now develop an initialization script that the processor runs once when the pipeline starts. Use an initialization script to set up connections or resources required by the processor.
  You can also develop a destroy script that the processor runs once when the pipeline stops. Use a destroy script to close any connections or resources opened by the processor.

- **JDBC Lookup enhancement** - Default value date formats. When the default value data type is Date, use the following format: yyyy/MM/dd . When the default value data type is Datetime, use the following format: yyyy/MM/dd HH:mm:ss.

- **Record Deduplicator processor enhancement** - The processor can now deduplicate records across all pipeline runners in a multithreaded pipeline.

- **Spark Evaluator processor enhancements** - The processor is now included in the MapR 5.2 stage library.
  The processor also now provides beta support of cluster mode pipelines. In a development or test environment, you can use the processor in pipelines that process data from a Kafka or MapR cluster in cluster streaming mode. Do not use the Spark Evaluator processor in cluster mode pipelines in a production environment.

Destinations

- **New HTTP Client destination** - A destination that writes to an HTTP endpoint.
- **New MQTT Publisher destination** - A destination that publishes messages to a topic on an MQTT broker.
- **New WebSocket Client destination** - A destination that writes to a WebSocket endpoint.
- **Azure Data Lake Store destination enhancement** - You can now configure an idle timeout for output files.
- **Cassandra destination enhancements** - The destination now supports the Cassandra uuid and timeuuid data types. And you can now specify the Cassandra batch type to use: Logged or Unlogged. Previously, the destination used the Logged batch type.
- **JDBC Producer enhancements** - The origin now includes a Schema Name property for entering the schema name. For information about possible upgrade impact, see Configure JDBC Producer Schema Names.
  You can also use the Enclose Object Name property to enclose the database/schema, table, and column names in quotation marks when writing to the database.
- **MapR DB JSON destination enhancement** - You can now enter an expression that evaluates to the name of the MapR DB JSON table to write to.
- **MongoDB destination enhancements** - The destination can now use LDAP authentication in addition to username/password authentication to connect to MongoDB. You can also now include credentials in the MongoDB connection string.
- **SDC RPC destination enhancements** - The Back Off Period value that you enter now increases exponentially after each retry, until it reaches the maximum wait time of 5 minutes. Previously, there was no limit to the maximum wait time. The maximum value for the Retries per Batch property is now unlimited - previously it was 10 retries.
- **Solr destination enhancement** - You can now configure the action that the destination takes when it encounters missing fields in the record. The destination can discard the fields, send the record to error, or stop the pipeline.

Executors

- **New Spark executor** - The executor starts a Spark application on a YARN or Databricks cluster each time it receives an event.
- **New Pipeline Finisher executor** - The executor stops the pipeline and transitions it to a Finished state when it receives an event. Can be used with the JDBC Query Consumer, JDBC Multitable Consumer, and Salesforce origins to perform batch processing of available data.
- **HDFS File Metadata executor enhancement** - The executor can now create an empty file upon receiving an event. The executor can also generate a file-created event when generating events.

- **MapReduce executor enhancement** - When starting the provided Avro to Parquet job, the executor can now overwrite any temporary files created from a previous run of the job.

Functions

- **New escape XML functions** - Three new string functions enable you to escape and unescape XML.
- **New pipeline user function** - A new pipeline user function enables you to determine the user who started the pipeline.
- **New function to generate UUIDs** - A new function that enables you generate UUIDs.
- **New function returns the number of available processors** - The runtime:availableProcessors() function returns the number of processors available to the Java virtual machine.

General Enhancements

- **Data Collector Hadoop impersonation enhancement** - You can use the stage.conf_hadoop.always.impersonate.current.user Data Collector configuration property to ensure that Data Collector uses the current Data Collector user to read from or write to Hadoop systems.
  When enabled, you cannot configure alternate users in the following Hadoop-related stages:
  - Hadoop FS origin and destination
  - MapR FS origin and destination
  - HBase lookup and destination
  - MapR DB destination
  - HDFS File Metadata executor
  - MapReduce executor
- **Stage precondition property enhancement** - Records that do not meet all preconditions for a stage are now processed based on error handling configured in the stage. Previously, they were processed based on error handling configured for the pipeline. See Precondition Error Handling for information about upgrading.
- **XML parsing enhancement** - You can include field XPath expressions and namespaces in the record with the Include Field XPaths property. And use the new Output Field Attributes property to write XML attributes and namespace declarations to field attributes rather than including them in the record as fields.
- **Wrap long lines in properties** - You can now configure Data Collector to wrap long lines of text that you enter in properties, instead of displaying the text with a scroll bar.

Fixed Issues in 2.5.0.0

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

For the full list, click here.

JIRA	Description

SDC-5816	When the Kudu destination has the default operation set to UPSERT or UPDATE, it writes to all columns in the Kudu table. If a Kudu column is not mapped to a field in the record - the destination sets the column to null.
SDC-5785	The Directory origin processes a file name pattern as a glob when the File Name Pattern Mode property is set to regular expression.
SDC-5744	The pipeline action buttons take more than 30 minutes to display when you view a pipeline in a Data Collector that has more than 1,000 pipelines.
SDC-5686	The HTTP Client origin and HTTP Lookup processor fail to connect when using an authenticated proxy.
SDC-5683	The MongoDB destination no longer includes the Preconditions and On Record Error properties.
SDC-5648	The HTTP Client origin encounters a pagination error when the batch wait time expires.
SDC-5641	The Cassandra destination cannot write null values to Cassandra.
SDC-5632	The Cassandra destination does not support the Cassandra DATE type.
SDC-5605	The JDBC Producer destination cannot write to a case sensitive database table that has a name with all uppercase letters or a name with both upper and lowercase letters.

SDC-5527	When you complete the MapR prerequisites for MapR version 5.2.0, the `setup-mapr` command incorrectly removes version 5.2.0 from the blacklist property in the `sdc.properties` file by not removing the last comma and backslash, as follows:
	`system.stagelibs.blacklist=\`
	`  streamsets-datacollector-mapr_5_0-lib,\`
	`  streamsets-datacollector-mapr_5_1-lib,\`

SDC-5410	The MapReduce executor does not start MapReduce jobs on the MapR distribution of Hadoop FS.
SDC-5328	Data Collector cannot return the value of fields that include a backslash (\) or single quote (‘) in the field path.
SDC-3084	When the HBase Lookup processor performs a bulk lookup of all keys in a batch, it stops the pipeline when a row expression evaluates to an empty value.
SDC-2750	Data Collector keeps connections to Kafka when no pipeline is running.

Known Issues in 2.5.0.0

Please note the following known issues with this release.

For a full list of known issues, check out our JIRA.

JIRA	Description
SDC-5910	The 2.5.0.0 version of the Cloudera Manager CSD does not include stage aliases. When upgrading from a previous release using Cloudera Manager, certain stages might not appear in existing pipelines. Similarly, importing pipelines from a previous release can fail. **Workaround:** 1. In Cloudera Manager, select the **StreamSets** service and then click **Configuration**. 2. On the **Configuration** page, in the **Data Collector Advanced Configuration Snippet (Safety Valve) for sdc.properties property**, add the text from this gist. 3. Restart Data Collector.
SDC-5871	Attempting to write a record to Kudu with a null primary key causes the pipeline to fail.
SDC-5818	The UDP Source origin can generate inaccurate information for the Timestamp, First, and Last fields when reading Netflow messages.
SDC-5758	When configured to use Kerberos authentication, Data Collector cannot connect to Kudu using the Kudu 1.1 or 1.2 stage libraries. This may be a Kudu issue. Possible workaround: In the pipeline, try using the Kudu 1.3 stage library.
SDC-5757	JDBC Producer encloses all column names in quotation marks in queries. For MySQL databases without ANSI_QUOTES enabled, the destination generates JDBC-14 errors. MySQL workarounds: Use any one of the following workarounds: 1. In the JDBC Producer, add an Additional JDBC Configuration Property with the name set to “sessionVariables” and the value set to “sql_mode=ANSI_QUOTES”. 2. To enable ANSI_QUOTES mode for the MySQL database until the database instance is restarted, use the following MySQL command to set the mode: `SET @@GLOBAL.SQL_MODE =“ANSI_QUOTES”` Note: To ensure any other modes are maintained, append the current modes to the SET command. 3. To permanently enable ANSI_QUOTES mode for the entire database, modify the my.cnf file to append ANSI_QUOTES to any modes that are already enabled.
SDC-5521	Kerberos-enabled pipelines that are valid using an Apache Kafka stage library can fail validation when using a Cloudera distribution of Kafka. The Apache Kafka client libraries have been modified by StreamSets to allow connectivity to Kerberized Kafka without requiring a JAAS configuration file. The Apache Kafka stage libraries (version 0.9 and greater) do not require a JAAS configuration file when enabling Kerberos.

	For CDH Kafka versions, a JAAS configuration is still required. Workaround: Include a JAAS configuration file on the classpath by modifying the `sdc-env.sh` or `sdcd-env.sh` file to include the following option in the SDC_JAVA_OPTS environment variable: `-Djava.security.auth.login.config=<path-to-jaas-config>` Or, you can use the matching version Apache Kafka stage library, instead of the Cloudera stage libraries.
SDC-5357	The Jython Evaluator and stages that can access the file system - such as the Directory and File Tail origins - can access all pipeline JSON files stored in the $SDC_DATA directory. This allows users to access pipelines that they might not have permission to access within Data Collector. Workaround: To secure your pipelines, complete the following tasks: 5. Remove the Jython stage library and use the Groovy Evaluator or JavaScript Evaluator processor instead of the Jython Evaluator. 6. Update the Data Collector security policy file, `$SDC_CONF/sdc-security.policy`, so that Data Collector stages do not have `AllPermission` access to the file system. Update the security policy for the following code bases: streamsets-libs-extras, streamsets-libs, and streamsets-datacollector-dev-lib. Use the policy file syntax to set the security policies.
SDC-5325	Cluster mode pipelines that read from a MapR cluster fail when the MapR cluster uses username/password login authentication. Workaround: On the Cluster tab for the pipeline, add the following Java property to the Worker Java Options property: `-Dmaprlogin.password.enabled=true`
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-5039	When you use the Hadoop FS origin to read files from all subdirectories, the origin cannot use the configured Hadoop FS User as a proxy user to read from HDFS. Workaround: If you need to use a proxy user to read from all subdirectories of the specified directories, set the HADOOP_PROXY_USER environment variable to the proxy user in libexec/_cluster-manager script, as follows: `export HADOOP_PROXY_USER = <proxy-user>`
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-3944	The Hive Streaming destination using the MapR library cannot connect to a MapR cluster that uses Kerberos or username/password login authentication.
SDC-3133	When you upgrade Data Collector from the RPM package, the environment configuration file `$SDC_DIST/libexec/sdcd-env.sh` is overwritten. Workaround: Back up the `sdcd-env.sh` file before you upgrade.
SDC-2374	A cluster mode pipeline can hang with a CONNECT_ERROR status. This can be a temporary connection problem that resolves, returning the pipeline to the RUNNING status. If the problem is not temporary, you might need to manually edit the pipeline state file to set the pipeline to STOPPED. Edit the file only after you confirm that the pipeline is no longer running on the cluster or that the cluster has been decommissioned. To manually change the pipeline state, edit the following file: `$SDC_DATA/runInfo/<cluster pipeline name>/<revision>/pipelineState.json` In the file, change CONNECT_ERROR to STOPPED and save the file.

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, ask for help, or find out about our next meetup, check out our Community page:
https://streamsets.com/community/.

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