StreamSets Data Collector 2.4.0.0 Release Notes

March 2, 2017

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

For important information about upgrading, new features, fixed issues, and known issues, read below.

Important: We highly recommend that you migrate to Java 8 with this version of Data Collector. Data Collector version 2.4.0.0 is the last release to support Java 7. **Data Collector will no longer run on Java 7 starting with version 2.5.0.0.**

Upgrading to Version 2.4.0.0

You can upgrade previous versions of Data Collector to version 2.4.0.0. For instructions on upgrading, see the Upgrade Documentation.

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 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.

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.

To retain pre-2.4.0.0 behavior, you can disable pipeline permissions by setting the pipeline.access.control.enabled property to false 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.

Enable Access to Data Collector Logs

With the 2.4.0.0 release, only users with the Admin role can view Data Collector log data. If necessary, grant the Admin role to users who require log access.

Note that the Admin role enables performing any Data Collector task and access to all pipelines, so avoid granting the Admin role unless necessary.

For more information about roles, see Roles.

Update Elasticsearch Pipelines

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

To upgrade pipelines that use the Elasticsearch destination, you must verify that Java 8 is installed on the Data Collector machine. Elasticsearch is no longer supported on Java 7.

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

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

Pipeline Sharing and Permissions

Data Collector now provides pipeline-level permissions. Permissions determine the access level that users and groups have on pipelines. To create a multitenant environment, create groups of users and then share pipelines with the groups to grant different levels of access.

With this change, only the pipeline owner and users with the Admin role can view a pipeline by default. If upgrading from a previous version of Data Collector, see the following post-upgrade task, Configure Pipeline Permissions.

This feature includes the following components:

- Pipeline permissions - Pipelines now have read, write, and execute permissions. Pipeline permissions overlay existing Data Collector roles to provide greater security. For information

about roles and permissions, see Roles and Permissions.

- Pipeline sharing - The pipeline owner and users with the Admin role can configure pipeline permissions for users and groups.

- Data Collector pipeline access control property - You can enable and disable the use of pipeline permissions with the pipeline.access.control.enabled configuration property. By default, this property is enabled.

- Permissions transfer - You can transfer all pipeline permissions associated with a user or group to a different user or group. Use pipeline transfer to easily migrate permissions after registering with DPM or after a user or group becomes obsolete.

Dataflow Performance Manager (DPM)

- Register Data Collectors with DPM - If Data Collector uses file-based authentication and if you register the Data Collector from the Data Collector UI, you can now create DPM user accounts and groups during the registration process.

- Aggregated statistics for DPM - When working with DPM, you can now configure a pipeline to write aggregated statistics to SDC RPC. Write statistics to SDC RPC for development purposes only. For a production environment, use a Kafka cluster or Amazon Kinesis Streams to aggregate statistics.

Origins

- Dev SDC RPC with Buffering origin - A new development stage that receives records from an SDC RPC destination, temporarily buffering the records to disk before passing the records to the next stage in the pipeline. Use as the origin in an SDC RPC destination pipeline.

- Amazon S3 origin enhancement - You can configure a new File Pool Size property to determine the maximum number of files that the origin stores in memory for processing after loading and sorting all files present on S3.

Other

- New stage libraries - This release supports the following new stage libraries:
    - Kudu versions 1.1 and 1.2
    - Cloudera CDH version 5.10 distribution of Hadoop
    - Cloudera version 5.10 distribution of Apache Kafka 2.1

- Install external libraries using the Data Collector user interface - You can now use the Data Collector user interface to install external libraries to make them available to stages. For example, you can install JDBC drivers for stages that use JDBC connections. Or, you can install external libraries to call external Java code from the Groovy, Java, and Jython Evaluator processors.

- Security enhancement - With this release, the ability to view log information and JVM metrics is restricted to users with the Admin role. For information about upgrade impact, see Enable Access to Data Collector Logs.

- Custom header enhancement - You can now use HTML in the ui.header.title configuration property to configure a custom header for the Data Collector UI. This allows you to specify the look and feel for any text that you use, and to include small images in the header.

- Groovy enhancement - You can configure the processor to use the invokedynamic bytecode instruction.

- Pipeline renaming - You can now rename a pipeline by clicking directly on the pipeline name when editing the pipeline, in addition to editing the Title general pipeline property.

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-5420	Kafka Consumer dies with the following error when trying to commit the offset in the middle of rebalancing: `CommitFailedException: Commit cannot be completed due to group rebalance`
SDC-5232	Upgrading to 2.3 causes the following error with pipelines that write to HDFS: `HADOOPFS_59 - Recovery failed to rename old _tmp_ files`

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-5514	Data Collector version 2.4.0.0 run from Docker encounters the following null pointer exception:
	`java.lang.NullPointerException`
	`at com.streamsets.datacollector.restapi.AclStoreResource.`
	`getPermissions(AclStoreResource.java:190)`
	
	Workaround: In the Data Collector configuration file, `$SDC_CONF/sdc.properties`, set the `http.authentication` property to `form`.

SDC-5478	Pipelines that write to Azure Data Lake Store encounter HTTP 401 Unauthorized errors and write all subsequent records to error when the Azure Active Directory access token expires. The access token lifetime can range from ten minutes to one day. The default is one hour.

	Workaround: Restart the pipeline before the access token expires.
SDC-5410	The MapReduce executor does not start MapReduce jobs on the MapR distribution of Hadoop FS.
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-5148	A pipeline with an Elasticsearch destination upgraded to the current release has 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 DocumentId. Workaround: Review all upgraded Elasticsearch pipelines to ensure the Default Operation is set to the correct operation.
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-4876	When running Data Collector on Java 8, you might receive “Cannot load driver” errors for stages that use JDBC drivers. Workaround: On the Legacy Drivers tab, configure the Driver Class Name property.

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-4128	In cluster mode, Data Collector does not generate log files for worker Data Collectors.
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-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 `$SDC_DIST/libexec/sdcd-env.sh` is overwritten. Workaround: Back up the `sdcd-env.sh` file before you upgrade.
SDC-2822	If you configure a Kafka Producer destination to write one message per batch, and then use a cluster pipeline to process that data from the Kafka cluster, the cluster pipeline might encounter an out of memory error.
SDC-2586	To process records larger than 1 MB, you must configure the DataFactoryBuilder.OverRunLimit property. However, this property is not configurable in the Data Collector configuration file, `$SDC_CONF/sdc.properties` Workaround: Set the value of DataFactoryBuilder.OverRunLimit property in the SDC_JAVA_OPTS environment variable in the Data Collector environment file, `$SDC_DIST/libexec/sdc-env.sh` or `$SDC_DIST/libexec/sdcd-env.sh` Set the property greater than the largest record you want to process. For example, to process records up to 2 MB, set the property to 2097152 as follows: `SDC_JAVA_OPTS="-DDataFactoryBuilder.OverRunLimit=2097152"`
SDC-2374	A cluster mode pipeline can hang with a CONNECT_ERROR status. This can be a temporary connection problem that resolves, returning the pipeline to the RUNNING status. If the problem is not temporary, you might need to manually edit the pipeline state file to set the pipeline to STOPPED. Edit the file only after you confirm that the pipeline is no longer running on the cluster or that the cluster has been decommissioned. To manually change the pipeline state, edit the following file: `$SDC_DATA/runInfo/<cluster pipeline name>/<revision>/pipelineState.json` In the file, change CONNECT_ERROR to STOPPED and save the file.
SDC-1731	When using the Kafka Consumer or Kafka Producer on HDP 2.3 with Kerberos enabled, set the Kafka broker configuration property

	`security.inter.broker.protocol` to PLAINTEXT. When enabling Kerberos, HDP 2.3 sets the `security.inter.broker.protocol` property to PLAINTEXTSASL, which is not supported. If the property is not set to PLAINTEXT, when the pipeline starts, validation errors indicate a problem connecting to Kafka.
SDC-891	At this time, writing to error records to file is not supported for cluster mode pipelines. Workaround: Write error records to Kafka or to an SDC RPC pipeline.
SDC-890	For cluster mode pipelines configured to stop on error or to stop upon reaching a memory limit, the Data Collector cannot stop all worker pipelines as expected. Workaround: To stop all pipelines, use the Stop icon in the Data Collector console.

Contact Information

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

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