StreamSets Data Collector 2.7.0.0 Release Notes

August 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:

Upgrading to Version 2.7.0.0

You can upgrade previous versions of Data Collector to version 2.7.0.0. For complete 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 2.7.0.0. 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.

Update Vault Pipelines

With version 2.7.0.0, Data Collector introduces a credential store API and credential expression language functions to access Hashicorp Vault secrets. The following JDBC stages can use the new credential functions:

- JDBC Multitable Consumer origin
- JDBC Query Consumer origin
- Oracle CDC Client origin
- SQL Server CDC Client origin
- SQL Server Change Tracking Client origin
- JDBC Lookup processor
- JDBC Tee processor
- JDBC Producer destination
- JDBC Query executor

In addition, the Data Collector Vault integration now relies on Vault's App Role authentication backend.

Previously, Data Collector used Vault expression language functions to access Vault secrets and relied on Vault's App ID authentication backend. StreamSets has deprecated the Vault functions, and Hashicorp has deprecated the App ID authentication backend.

After upgrading, update pipelines that use Vault functions in one of the following ways:

- Use the new credential expression language functions in pipelines that include JDBC stages - Recommended method.
- Continue to use the deprecated Vault expression language functions - You can continue to use the deprecated Vault functions after you make the required changes to the Vault configuration properties. However, the functions will be removed in a future release - so we recommend that you use the new credential functions as soon as possible.

For instructions, see Update Vault Pipelines.

New Features and Enhancements

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

Credential Stores

Data Collector now has a credential store API that integrates with the following credential store systems:

- Java keystore
- Hashicorp Vault

You define the credentials required by external systems - user names or passwords - in a Java keystore file or in Vault. Then you use credential expression language functions in JDBC stage properties to retrieve those values, instead of directly entering credential values in stage properties.

The following JDBC stages can use the new credential functions:

- JDBC Multitable Consumer origin
- JDBC Query Consumer origin
- Oracle CDC Client origin
- SQL Server CDC Client origin
- SQL Server Change Tracking origin
- JDBC Lookup processor
- JDBC Tee processor
- JDBC Producer destination
- JDBC Query executor

Publish Pipeline Metadata to Cloudera Navigator (Beta)

Data Collector now provides beta support for publishing metadata about running pipelines to Cloudera Navigator. You can then use Cloudera Navigator to explore the pipeline metadata, including viewing lineage diagrams of the metadata.

Feel free to try out this feature in a development or test Data Collector, and send us your feedback. We are continuing to refine metadata publishing as we gather input from the community and work with Cloudera.

Stage Libraries

Data Collector includes the following new stage libraries:

- Apache Kudu version 1.4.0
- Cloudera CDH version 5.12 distribution of Hadoop
- Cloudera version 5.12 distribution of Apache Kafka 2.1
- Google Cloud - Includes the Google BigQuery origin, Google Pub/Sub Subscriber origin, and Google Pub/Sub Publisher destination.
- Java keystore credential store - For use with credential stores.
- Vault credential store - For use with credential stores.

Data Collector Configuration

- **Access Hashicorp Vault secrets** - The Data Collector Vault integration now relies on Vault's App Role authentication backend. Previously, Data Collector relied on Vault's App ID authentication backend. Hashicorp has deprecated the App ID authentication backend.

- **New Hadoop user impersonation property** - When you enable Data Collector to impersonate the current Data Collector user when writing to Hadoop, you can now also configure Data Collector to make the username lowercase. This can be helpful with case-sensitive implementations of LDAP.

- **New Java security properties** - The Data Collector configuration file now includes properties with a "java.security." prefix, which you can use to configure Java security properties.

- **New property to define the amount of time to cache DNS lookups** - By default, the java.security.networkaddress.cache.ttl property is set to 0 so that the JVM uses the Domain Name Service (DNS) time to live value, instead of caching the lookups for the lifetime of the JVM.

- **SDC_HEAPDUMP_PATH enhancement** - The new default file name, $SDC_LOG/sdc_heapdump_${timestamp}.hprof, includes a timestamp so you can write multiple heap dump files to the specified directory.

Dataflow Triggers

- **Pipeline events** - The event framework now generates pipeline lifecycle events when the pipeline stops and starts. You can pass each pipeline event to an executor or to another pipeline for more complex processing. Use pipeline events to trigger tasks before pipeline processing begins or after it stops.

Origins

- **New Google BigQuery origin** - An origin that executes a query job and reads the result from Google BigQuery.

- **New Google Pub/Sub Subscriber origin** - A multithreaded origin that consumes messages from a Google Pub/Sub subscription.

- **New OPC UA Client origin** - An origin that processes data from an OPC UA server.

- **New SQL Server CDC Client origin** - A multithreaded origin that reads data from Microsoft SQL Server CDC tables.

- **New SQL Server Change Tracking origin** **-** A multithreaded origin that reads data from Microsoft SQL Server change tracking tables and generates the latest version of each record.

- **Directory origin event enhancements** - The Directory origin can now generate no-more-data events when it completes processing all available files and the batch wait time has elapsed without the arrival of new files. Also, the File Finished event now includes the number of records and files processed.

- **Hadoop FS origin enhancement** - The Hadoop FS origin now allows you to read data from other file systems using the Hadoop FileSystem interface. Use the Hadoop FS origin in cluster batch pipelines.

- **HTTP Client origin enhancement** - The HTTP Client origin now allows time functions and datetime variables in the request body. It also allows you to specify the time zone to use when evaluating the request body.

- **HTTP Server origin enhancement** - The HTTP Server origin can now process Avro files.

- **JDBC Query Consumer origin enhancement** - You can now configure the behavior for the origin when it encounters data of an unknown data type.

- **JDBC Multitable Consumer origin enhancements**:
    - You can now use the origin to perform multithreaded processing of partitions within a table. Use partition processing to handle even larger volumes of data. This enhancement also includes new JDBC header attributes.

      By default, all new pipelines use partition processing when possible. Upgraded pipelines use multithreaded table processing to preserve previous behavior.
    - You can now configure the behavior for the origin when it encounters data of an unknown data type.

- **Oracle CDC Client origin enhancements**:
    - The origin can now buffer data locally rather than utilizing Oracle LogMiner buffers.
    - You can now specify the behavior when the origin encounters an unsupported field type - send to the pipeline, send to error, or discard.
    - You can configure the origin to include null values passed from the LogMiner full supplemental logging. By default, the origin ignores null values.
    - You now must select the target server time zone for the origin.
    - You can now configure a query timeout for the origin.
    - The origin now includes the row ID in the oracle.cdc.rowId record header attribute.

- **RabbitMQ Consumer origin enhancement** - When available, the origin now provides attributes generated by RabbitMQ, such as contentType, contentEncoding, and deliveryMode, as record header attributes.

- **TCP Server origin enhancement** - The origin can now process character-based data that includes a length prefix.

- **UDP Source origin enhancement** **-** The origin can now process binary and character-based raw data.

- **New last-modified time record header attribute** - Directory, File Tail, and SFTP/FTP Client origins now include the last modified time for the originating file for a record in an mtime record header attribute.

Processors

- **New Data Parser processor** - A processor that extracts NetFlow or syslog messages as well as other supported data formats that are embedded in a field.
- **New JSON Generator processor** - A processor that serializes data from a record field to a JSON-encoded string.
- **New Kudu Lookup processor** - A processor that performs lookups in Kudu to enrich records with additional data.
- **Hive Metadata processor enhancement** - You can now configure custom record header attributes for metadata records.

Destinations

- **New Google Pub/Sub Publisher destination** - A destination that publishes messages to Google Pub/Sub.
- **New JMS Producer destination** **-** A destination that writes data to JMS.
- **Amazon S3 destination enhancements:**
  - You can now use expressions in the Bucket property for the Amazon S3 destination. This enables you to write records dynamically based expression evaluation.
  - The Amazon S3 object written event record now includes the number of records written to the object.
- **Azure Data Lake Store destination enhancement** **-** The Client ID and Client Key properties have been renamed Application ID and Application Key to align with the updated property names in the new Azure portal.
- **Cassandra destination enhancement** - The destination now supports Kerberos authentication if you have installed the DataStax Enterprise Java driver.
- **Elasticsearch destination enhancement** - The destination can now create parent-child relationships between documents in the same index.
- **Hive Metastore destination** - You can now configure the destination to create custom record header attributes.
- **Kafka Producer destination enhancement** - The destination can now write XML documents.
- **Solr destination enhancement** - You can now configure the destination to skip connection validation when the Solr configuration file, `solrconfig.xml`, does not define the default search field (“df”) parameter.

Executors

- **New Amazon S3 executor** - Use the Amazon S3 executor to create new Amazon S3 objects for the specified content or add tags to existing objects each time it receives an event.

- **HDFS File Metadata executor enhancement** - The executor can now remove a file or directory when it receives an event.

Dataflow Performance Manager

- **Revert changes to published pipelines** - If you update a published pipeline but decide not to publish the updates to DPM as a new version, you can revert the changes made to the pipeline configuration.

Pipelines

- **Pipeline error handling enhancements:**
    - Use the new Error Record Policy to specify the version of the record to include in error records.
    - You can now write error records to Amazon Kinesis Streams.

- **Error records enhancement** - Error records now include the user-defined stage label in the errorStageLabel header attribute.

- **Pipeline state enhancements -** Pipelines can now display the following new states: STARTING_ERROR, STOPPING_ERROR, and STOP_ERROR.

Data Formats

- **Writing XML** - You can now use the Google Pub/Sub Publisher, JMS Producer, and Kafka Producer destinations to write XML documents to destination systems. Note the record structure requirement before you use this data format.

- **Avro:**
    - Origins now write the Avro schema to an avroSchema record header attribute.
    - Origins now include precision and scale field attributes for every Decimal field.
    - Data Collector now supports the time-based logical types added to Avro in version 1.8.

- **Delimited** - Data Collector can now continue processing records with delimited data when a row has more fields than the header. Previously, rows with more fields than the header were sent to error.

Cluster Pipelines

This release includes the following Cluster Yarn Streaming enhancements:

- Use a new Worker Count property to limit the number of worker nodes used in Cluster Yarn Streaming pipelines. By default, a Data Collector worker is spawned for each partition in the topic.
- You can now define Spark configuration properties to pass to the spark-submit script.

Expression Language

This release includes the following new functions:

- **credential:get()** - Returns credential values from a credential store.
- **credential:getWithOptions()** - Returns credential values from a credential store using additional options to communicate with the credential store.
- **record:errorStageLabel()** - Returns the user-defined name of the stage that generated the error record.
- **list:join()** - Returns each element of a List field joined on the specified character sequence.
- **list:joinSkipNulls()** - Returns each element of a List field joined on the specified character sequence, skipping null values.
- **str:indexOf()** - Returns the index within a string of the first occurrence of the specified substring.

Miscellaneous

- **Global bulk edit mode** - In any property where you would previously click an Add icon to add additional configurations, you can now switch to bulk edit mode to enter a list of configurations in JSON format.

- **Snapshot enhancement** - Snapshots no longer produce empty batches when waiting for data.

- **Webhooks enhancement** - You can use several new pipeline state notification parameters in webhooks.

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-7007	When you upgrade a pipeline that includes the HTTP Server origin, you cannot configure additional cipher suites for the origin.
SDC-6963	When data is not received within the initial batch timeout, the TCP Server origin reports incorrect record metrics.
SDC-6835	The HTTP Client origin fails to read the response when the server response is over HTTPS and includes the `connection: close` header.
SDC-6758	A pipeline with the Kafka Consumer origin fails when the origin is configured to produce a single record and to send the record to the pipeline for error handling.
SDC-6725	The SFTP/FTP Client origin can re-read the same file if an error is encountered.

SDC-6593	When the Salesforce origin is configured to subscribe to notifications, the origin might encounter a FORCE_04 error with non-string data.
SDC-6551	Need to improve the Kudu destination error message when the data type of the input record is not the same as the data type of the output record.
SDC-6540	The Field Hasher appends a null character (0x00) after each field value when performing field-level hashing.
SDC-6521	The JDBC Tee processor cannot be used with Microsoft SQL Server databases.
SDC-6265	The SFTP/FTP Client origin encounters a null pointer exception when attempting to read a null file.
SDC-6396	An alert webhook in an upgraded pipeline throws an exception.
SDC-6361	The Expression Evaluator processor does not allow time functions to be used in record header attribute expressions.
SDC-6176	When the Kafka Consumer origin is configured to work with the Confluent Schema Registry, the origin continues to use the old schema until the pipeline is restarted.
SDC-6001	You cannot use a pipeline runtime parameter for the email ID when configuring emails for rules.
SDC-5871	Attempting to write a record to Kudu with a null primary key causes the pipeline to fail.

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-6554	When converting Avro to Parquet on Impala, Decimal fields seem to be unreadable. Data Collector writes the Decimal data as variable-length byte arrays. And due to Impala issue IMPALA-2494, Impala cannot read the data.
SDC-6509	When using the Write to Another Pipeline pipeline error handling option in cluster batch pipeline, if the error handling pipeline encounters a problem and stops, the original pipeline stops with the message “Job has been finished” instead of indicating there was a problem with the error handling pipeline.

SDC-6438	The MapR distribution for Spark 2.x is not supported by cluster streaming pipelines.
SDC-6210	The show-vault-id command returns a NPE when Data Collector is installed on EC2 with IPv6 enabled. Workaround: If you can run Data Collector without IPv6, in the /etc/sysctl.conf file, disable IPv6 by configuring the disable IPv6 property as follows: `net.ipv6.conf.all.disable_ipv6 = 1`
SDC-6077	The Field Remover processor does not remove list fields from list-map data.
SDC-5758	Due to expected Kudu behavior, if Data Collector is configured to use Kerberos, but your Kudu installation does not, Data Collector cannot connect. This applies to Kudu version 1.2 and later. Workaround: If necessary, install a separate Data Collector that does not have Kerberos 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 $SDC_DATA directory. This allows users to access pipelines that they might not have permission to access within Data Collector. Workaround: To secure your pipelines, complete the following tasks: 1. Remove the Jython stage library and use the Groovy Evaluator or JavaScript Evaluator processor instead of the Jython Evaluator. 2. Update the Data Collector security policy file, `$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.