StreamSets Data Collector 1.4.0.0 Release Notes

June 2, 2016

New Features and Enhancements

We're happy to announce a new version of StreamSets Data Collector. This version has the following new features and enhancements:

- **New SFTP/FTP Client origin to read files from an SFTP or FTP server.**
- **New MapR DB destination to write data to MapR DB using the HBase API.**
- **Hive Streaming destination supports writing to MapR.**
- **New XML Flattener processor to flatten XML data in a string field.**
- **New HBase Lookup processor to perform key-value lookups from HBase.**
- **New Redis Lookup processor to perform key-value lookups from Redis.**
- **New Static Lookup processor to perform key-value lookups from local memory.**
- **Support for Elasticsearch 2.3 and HDP 2.4.0.**
- **Rate limiting for a pipeline.** You can limit the rate at which a pipeline processes records by defining the maximum number of records that the pipeline can read in a second.
- **Updated pipeline library on the Home page.** The pipeline library is now accessible on the Home page. You can use the pipeline library to filter pipelines by state or labels. You can use the pipeline list to perform the same action on multiple pipelines such as starting, stopping, or exporting the pipelines.
- **Pipeline labels used to group pipelines.** You can create and assign one or more labels to each pipeline. Use labels to group similar pipelines.
- **Pipeline duplication can create multiple copies of the pipeline.**
- **RabbitMQ Consumer origin generates a record for every object in a message.** By default, the origin now generates a record for each object in a RabbitMQ message. You can configure the origin to generate a single record for each RabbitMQ message instead.
- **Amazon S3 updates.**
    - The Amazon S3 origin can recurse through subfolders using glob patterns.
    - The Amazon S3 destination can write data to partitions based on expressions.
    - Both the origin and destination now use the terms "common prefix" to represent a base directory for files and "prefix pattern" for the directory and file name pattern.
- **Directory origin can read files based on last-modified timestamp.** You can configure the read order for the origin. The origin can read files based on the last-modified timestamp or on the file name.

  Upgraded pipelines continue to read files in order of file name as in previous releases.
- **Directory and File Tail origin record headers provide provenance.** The Directory and File Tail origins now provide file name, file path, and offset information in the record headers. You can use the record:attribute function to include this information in the body of the record.

StreamSets Data Collector 1.4.0.0 Release Notes

- **MongoDB origin allows resetting the origin.** You can now reset the origin to read all available data.
- **Hadoop FS can write records to directories specified in a record header attribute.** You can include the targetDirectory stage attribute in record headers to specify the directory to write the records to.
- **Configurable idle timeout for the Hadoop FS, Local FS, and MapR FS destinations.** You can configure the maximum time that an open output file written by the destination can remain idle.
- **Configurable time basis for the HBase destination.** You can configure whether the timestamp value added to each column written to HBase uses the processing time, record time, or system time.
- **Kudu destination can evaluate an expression to determine the existing table to write to.**
- **Unicode control character as a delimiter.** You can use a Unicode control character as the delimiter character when configuring an origin or destination to read or write delimited data or as the separator character when configuring a Field Splitter processor.
- **New line character replacement for delimited data format.** You can optionally enter a string to replace each newline character when a destination writes to a delimited data format.
- **record:valueOrDefault() function returns the default value if the field is null.** Previously, the function returned null if the field was null.
- **View stack traces for errors.** If a pipeline encounters errors, you can view the full stack trace if the error was produced by an exception.
- **Configurable thread pool size for running multiple standalone pipelines at the same time.**
- **Data Collector can be configured to bind to a specific host or IP address.**
- **Updated syntax to refer to files and environment variables in sdc.properties.** For more information, see Referencing Environment Variables and Values in Files in sdc.properties.
- **Stage-related custom metrics can be viewed in the Data Collector console.** Previously, you could view stage-related custom metrics only when you viewed Data Collector JMX metrics in external tools.

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

Upgrade

You can upgrade a previous version of Data Collector to version 1.4.0.0. For instructions on upgrading, see the Upgrade Documentation.

record:valueOrDefault() Function and Null Values

When upgrading Data Collector from any version earlier than 1.4.0.0, note the following change if upgraded pipelines use the record:valueOrDefault() function.

StreamSets Data Collector 1.4.0.0 Release Notes

The record:valueOrDefault() function now returns the default value if the field is null. Previously, the function returned null if the field was null.

Referencing Environment Variables and Values in Files in sdc.properties

When upgrading Data Collector from any version earlier than 1.4.0.0, note the following change if you reference sensitive values in files or reference environment variables in sdc.properties.

Use the following syntax to reference sensitive values in files:

```
${file("<filename>")}
```

Use the following syntax to reference environment variables:

```
${env("<environment variable name>")}
```

Previously, you used `@<filename>@` or `$<environment variable name>$`.

The previous syntax is supported for backward compatibility. However, when you update the configuration files during the upgrade, we recommend that you use the new syntax.

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-3023	Roles are not fetched when using LDAP authentication with an Active Directory server.
SDC-2930	Pipelines disappear when there is low disk space.
SDC-2917	Running multiple pipelines that use HDFS-related stages can cause a JVM deadlock.
SDC-2914	Cannot stop the Data Collector using the command line when it runs as a service.
SDC-2841	The HTTP origin cannot connect to HTTPS using a proxy.
SDC-2226	When the Data Collector UI is configured for HTTPS, cluster pipelines with the Hadoop FS origin fail to start.

StreamSets Data Collector 1.4.0.0 Release Notes

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-3110	When a File Tail origin is configured using the Active File with Reverse Counter naming convention, the origin throws a StringIndexOutOfBoundsException.
SDC-3017	Hadoop FS does not close and rename open files when the pipeline stops. Open files are named `prefix _tmp`. When the pipeline restarts, Hadoop FS continues processing the files, closing and renaming them as expected. Workaround: If your workflow immediately processes output files from Hadoop FS and your workflow requires synchronized data, pause downstream processing before you stop the pipeline.
SDC-2998	The JDBC Consumer origin does not correctly parse Oracle timestamp fields.
SDC-2950	When a pipeline writes error records to Elasticsearch, the record header information - error code, error message, and error stage - is not preserved.
SDC-2822	If you configure a Kafka Producer destination to write one message per batch, and then use a cluster pipeline to process that data from the Kafka cluster, the cluster pipeline might encounter an out of memory error.
SDC-2586	To process records larger than 1 MB, you must configure the DataFactoryBuilder.OverRunLimit property. However, this property is not configurable in the Data Collector configuration file, `$SDC_CONF/sdc.properties`. Workaround: Set the value of DataFactoryBuilder.OverRunLimit property in the SDC_JAVA_OPTS environment variable in the Data Collector environment file, `$SDC_DIST/libexec/sdc-env.sh` or `$SDC_DIST/libexec/sdcd-env.sh`. Set the property greater than the largest record you want to process. For example, to process records up to 2 MB, set the property to 2097152 as follows: `SDC_JAVA_OPTS="-DDataFactoryBuilder.OverRunLimit=2097152”`
SDC-2554	MapR FS does not support Kerberos authentication at this time.
SDC-2552	When an invalid topic is specified for a MapR Streams Consumer or a MapR Streams Producer, the following incorrect message displays: `CONTAINER_0701 - Stage 'com_streamsets_pipeline_stage_origin_maprstreams_MapRStreamsDS`

StreamSets Data Collector 1.4.0.0 Release Notes

	`ource_1' initialization error: org.apache.kafka.common.config.ConfigException: No bootstrap urls given in bootstrap.servers` This message is misleading because MapR Streams does not support the `bootstrap.servers` option.
SDC-2374	A cluster mode pipeline can hang with a CONNECT_ERROR status. This can be a temporary connection problem that resolves, returning the pipeline to the RUNNING status. If the problem is not temporary, you might need to manually edit the pipeline state file to set the pipeline to STOPPED. Edit the file only after you confirm that the pipeline is no longer running on the cluster or that the cluster has been decommissioned. To manually change the pipeline state, edit the following file: `$SDC_DATA/runInfo/<cluster pipeline name>/<revision>/pipelineState.json` In the file, change CONNECT_ERROR to STOPPED and save the file.
SDC-2359	Due to a Kafka issue, a pipeline with a Kafka Consumer or Kafka Producer can hang during validation or display initialization errors when unable to connect to a Kafka 0.9.0.0 broker. The Data Collector log indicates that the broker is unavailable. For more information about the Kafka JIRA, see https://issues.apache.org/jira/browse/KAFKA-2880.
SDC-2133	You cannot stop a pipeline in the middle of a long-running processor, such as a Jython Evaluator performing a complex script.
SDC-1731	When using the Kafka Consumer or Kafka Producer on HDP 2.3 with Kerberos enabled, set the Kafka broker configuration property `security.inter.broker.protocol` to PLAINTEXT. When enabling Kerberos, HDP 2.3 sets the `security.inter.broker.protocol` property to PLAINTEXTSASL, which is not supported. If the property is not set to PLAINTEXT, when the pipeline starts, validation errors indicate a problem connecting to Kafka.
SDC-1567	You cannot use cluster mode pipelines to read from HDP 2.3 due to a HDP integration issue with Kafka and Spark Streaming. For more information about the HDP issue, see http://hortonworks.com/community/forums/topic/kafka-and-spark-streaming-nosuchmethoderror-kafka-consumer-simpleconsumer/.

StreamSets Data Collector 1.4.0.0 Release Notes

SDC-891	At this time, writing to error records to file is not supported for cluster mode pipelines. Workaround: Write error records to Kafka or to an SDC RPC pipeline.
SDC-890	For cluster mode pipelines configured to stop on error or to stop upon reaching a memory limit, the Data Collector cannot stop all worker pipelines as expected. Workaround: To stop all pipelines, use the Stop icon in the Data Collector console.

Contact Information

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

To report an issue, ask for help, or find out about our next meetup, check out our Community page: http://streamsets.com/community/.

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