StreamSets Data Collector 2.1.0.0 Release Notes

October 13, 2016

New Features and Enhancements

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

Support for the Confluent Schema Registry

The Confluent Schema Registry is a distributed storage layer for Avro schemas. You can configure Data Collector stages that process Avro data to work with the Confluent Schema Registry in the following ways:

- Origins can look up Avro schemas in the Schema Registry by the specified schema ID or subject. The Kafka Consumer origin can also look up the Avro schema ID embedded in each Kafka message in the Schema Registry.
- Destinations can look up Avro schemas in the Schema Registry by the specified schema ID or subject. Or, destinations can register and store new Avro schemas in the Schema Registry. The Kafka Producer destination can also embed the Avro schema ID in each message that it writes.

Installation and Configuration

- **New stage libraries.** Data Collector now supports Elasticsearch version 2.4.
- **Install additional tarball libraries using the Data Collector user interface.** If you install the Data Collector core tarball, you can now use the Data Collector user interface to install individual stage libraries.

Amazon Web Services Stages

- **Connect to Amazon Web Services through endpoints.** In addition to connecting to the standard Amazon Web Services regions, you can select Other for the region and then specify the endpoint to connect to for the following stages:
    - Amazon S3 origin
    - Kinesis Consumer origin
    - Amazon S3 destination
    - Kinesis Firehose destination
    - Kinesis Producer destination

- **Renamed property for the Amazon S3 destination.** The File Name Prefix property has been renamed to the Object Name Prefix property.

Origins

- **New MapR FS origin**. Use the new MapR FS origin in a cluster mode pipeline to process files stored on MapR FS.

- **Directory origin enhancement.** The Directory origin can now create error records for delimited data with more than the expected number of fields.

© 2016, StreamSets, Inc., Apache License, Version 2.0

StreamSets Data Collector 2.1.0.0 Release Notes

- **Oracle CDC Client enhancement.** Improved performance by processing data based on the commit number in ascending order.

- **UDP Source and UDP to Kafka enhancement.** To improve performance when reading messages from UDP ports, you can configure the UDP Source and UDP to Kafka origins to use multiple receiver threads for each port. Because the multi-threading requires native libraries, it is available when Data Collector runs on 64-bit Linux.

Processors

- **Groovy Evaluator, JavaScript Evaluator, and Jython Evaluator enhancements.** You can now use the processors to create new records and to create list-map fields. In a pipeline that processes the whole file data format, you can now use the processors to access whole file data by creating an input stream to the file.

- **JDBC Lookup enhancement.** To improve pipeline performance, you can configure the JDBC Lookup processor to locally cache the lookup values returned from a database table.

Data Formats

- **Text data format enhancement.** When writing text data, you can now specify the record separator characters that you want to use.

- **Whole file data format enhancement.** You can now use the Groovy Evaluator, JavaScript Evaluator, and Jython Evaluator to access whole file data by creating an input stream to the file.

Cluster Mode

- **MapR support for cluster batch mode.** Use the new MapR FS origin in a cluster mode pipeline to process files from MapR FS.

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

Upgrade

You can upgrade previous versions of Data Collector to version 2.1.0.0. For instructions on upgrading, see the Upgrade Documentation.

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-4106	Add notification in the user interface and documentation for the Hive Metadata processor that Hive table names are created with lowercase letters.

© 2016, StreamSets, Inc., Apache License, Version 2.0

StreamSets Data Collector 2.1.0.0 Release Notes

SDC-4101	Memory leaks can occur when running thousands of pipelines because Data Collector does not purge the running pipeline cache at regular intervals.
SDC-4091	Expose additional functions for data rules in the Alert Text property.
SDC-4066	The RPM installation is missing the root-lib folder.
SDC-4046	Pipelines upgraded to SDC 2.0.0.0 that include the XML Flattener processor generate the following validation error: `CREATION_013 - Configuration value 'true' is not boolean, it is a '{}'`
SDC-4036	When the Directory origin encounters a line with an unexpected number of columns, it stops reading the rest of the file. Instead, it should generate an error record for the malformed line and then continue reading the rest of the file.
SDC-4033	Memory leaks can occur when Data Collector constantly evaluates different expressions because the commons-el library maintains a cache of all expressions that is not property evicted.
SDC-3988	When you stop a cluster mode pipeline, the _tmp file might not be renamed.
SDC-3712	The Hadoop FS origin incorrectly lists MapR as an available stage library.

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-4954	The Cassandra destination encounters problems connecting to a Cassandra cluster because the Cassandra stage library directory contains a mixed version of netty JAR files. Workaround: 1. Remove all netty* JAR files from the following directory: `$SDC_DIST/streamsets-libs/streamsets-datacollector-cassandra_3-lib/lib` 2. Download the following netty JAR file: `http://central.maven.org/maven2/io/netty/netty-all/4.0.41.Final/netty-all-4.0.41.Final.jar` 3. Add the netty-all-4.0.41.Final.jar file to the Cassandra stage library directory.

© 2016, StreamSets, Inc., Apache License, Version 2.0

StreamSets Data Collector 2.1.0.0 Release Notes

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-4172	Data Collector cannot access Vault secrets stored in Hashicorp Vault.
SDC-4128	In cluster mode, Data Collector does not generate log files for worker Data Collectors.
SDC-4090	The MapR FS destination does not support impersonating an HDFS user. Instead, the destination always uses the user account who started the Data Collector to connect to MapR FS.
SDC-4087	Data preview fails for pipelines that use the Dev Raw Data Source origin when you refresh the data preview or run data preview with changes.
SDC-4047	The XML Flattener processor fails to parse XML that contains whitespace after the XML prolog. Workaround: Use an Expression Evaluator or scripting processor to remove the whitespace before using the XML Flattener.
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-3911	The Field Renamer processor does not support quoting regex special characters in field names. For example, if you specify a field name of `/'tag|attr'`, the processor interprets the pipe symbol (|) as the regex OR and cannot find the field. Workaround: Manually quote the special character by wrapping it in \Q and \E like so: `/'tag\Q|\Eattr'`
SDC-3357	If you run Data Collector from Docker, you cannot shut down Data Collector by running `docker stop` or pressing Ctrl+C from the Docker Quickstart Terminal. Workaround: In the Data Collector console, click **Administration** > **Shut Down**.
SDC-3356	Using the following commands to shut down or restart Data Collector does not properly complete the shutdown: • `service sdc stop` • `service sdc restart` Workaround: In the Data Collector console, click **Administration** > **Shut Down** or **Administration** > **Restart**.
SDC-3234	Cluster streaming pipelines that run on YARN use the YARN user instead of the Data Collector user to run executors.

© 2016, StreamSets, Inc., Apache License, Version 2.0

StreamSets Data Collector 2.1.0.0 Release Notes

SDC-3133	When you upgrade Data Collector from the RPM package, the environment configuration file `$SDC_DIST/libexec/sdcd-env.sh` is overwritten. Workaround: Back up the `sdcd-env.sh` file before you upgrade.
SDC-2950	When a pipeline writes error records to Elasticsearch, the record header information - error code, error message, and error stage - is not preserved.
SDC-2822	If you configure a Kafka Producer destination to write one message per batch, and then use a cluster pipeline to process that data from the Kafka cluster, the cluster pipeline might encounter an out of memory error.
SDC-2586	To process records larger than 1 MB, you must configure the DataFactoryBuilder.OverRunLimit property. However, this property is not configurable in the Data Collector configuration file, `$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-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.

© 2016, StreamSets, Inc., Apache License, Version 2.0

StreamSets Data Collector 2.1.0.0 Release Notes

SDC-890	For cluster mode pipelines configured to stop on error or to stop upon reaching a memory limit, the Data Collector cannot stop all worker pipelines as expected. Workaround: To stop all pipelines, use the Stop icon in the Data Collector console.

Contact Information

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

To review the latest documentation or try out our tutorials, check out the following links:

- User Guide
- User Guide tutorial
- GitHub tutorials

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

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

© 2016, StreamSets, Inc., Apache License, Version 2.0