StreamSets Data Collector 2.0.0.0 Release Notes

September 26, 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.

Integration with StreamSets Dataflow Performance Manager (DPM™)

You can enable Data Collector to work with our new product, StreamSets Dataflow Performance Manager (DPM). DPM is a management console for data in motion. With DPM, you can:

- Map multiple dataflows in a single visual topology and track changes to the dataflows over time.
- Measure dataflow performance across each topology, from end-to-end or point-to-point.
- Master your day-to-day operations by performing release and configuration management, and monitoring alerts to ensure incoming data meets business requirements for availability and accuracy.

Installation and Configuration

- **New stage libraries:**
  - Data Collector now supports MapR version 5.2.0.
  - HDP 2.4 library now provides a Kafka Consumer origin for cluster mode.

Origins

- **New Oracle CDC Client origin.** The Oracle CDC Client origin reads LogMiner redo logs to generate records with change data capture information.

- **Amazon S3 origin enhancements.**
  - **Error and post-processing options.** You can now copy as well as move error files or processed files to a different prefix or bucket.
  - **Read order.** You can now configure the read order for files. The origin can read based on timestamp or lexicographical key name order. Previously, the origin processed files based on timestamp.

- **SFTP/FTP Client origin enhancements.** You can now process whole files, process files in subdirectories, and process files that match a file name pattern. You can also reset the origin to process all available data instead of continuing from where the pipeline stopped.

- **Support for custom delimiters in text data.** When using the text data format, you can now generate records using custom delimiters instead of line breaks, and choose to include or exclude the delimiters from the resulting data.

Processors

- **New Field Flattener processor.** Use the Field Flattener to flatten all nested fields in a record. You can use this processor to flatten records for the Hive Drift Solution.

StreamSets Data Collector 2.0.0.0 Release Notes

- **Field Type Converter processor enhancement.** You can now convert fields with the Byte Array data type to the String data type.

- **Field Splitter processor enhancements.** You now can split fields based on a regular expression instead of a single character. You can also write additional data to a List field instead of the last split field in the record.

- **Geo IP processor enhancements.** You can perform lookups on multiple databases, which allows returning a wider range of data. IPv6 addresses are now supported. You can now use the following databases in addition to the original City and Country databases:
    - Anonymous IP
    - Connection Type
    - Domain
    - ISP

- **Hive Metadata processor enhancement.** Use the new Data Time Zone property to help determine how to evaluate datetime-based partition expressions.

- **Value Replacer processor enhancement.** You can now replace values with a constant based on a condition.

- **XML Flattener processor enhancement.** You can now keep all fields in the record.

Cluster Mode

- **MapR support for cluster streaming mode.** Use MapR Streams Consumer in a cluster mode pipeline to process data from MapR.

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

Upgrade

You can upgrade previous versions of Data Collector to version 2.0.0.0. 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-3946	The JDBC Consumer origin should not require an offset column for full mode.
SDC-3939	If the Expression Evaluator or Field Renamer processor cannot reach the specified target or output field because the full field path doesn't exist, the pipeline fails.

StreamSets Data Collector 2.0.0.0 Release Notes

SDC-3908	The core RPM installation does not include the Data Collector command line interface.
SDC-3899	The Geo IP processor should not attempt to validate IP addresses using regex.
SDC-3867	The HTTP Client origin errors out during preview due to an exception encountered by the JSON Parser processor.
SDC-3849	The Field Splitter processor does not correctly split on the ^ character.
SDC-3848	The HTTP Client origin running in streaming mode encounters an out of memory exception after a while.
SDC-3840	The Kinesis Consumer origin doesn't correctly update the Kinesis checkpointer when the records from multiple shards encounter errors.
SDC-3803	The HTTP Client origin silently fails to connect using an HTTP proxy if the protocol for the proxy server isn't specified.
SDC-3194	A Data Collector worker for a cluster mode pipeline displays an incorrect error message on pipeline failure.

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.

StreamSets Data Collector 2.0.0.0 Release Notes

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 '{}'` Workaround: In the XML Flattener, on the Flatten tab, select the Keep Original Fields property and the Overwrite Existing Fields property, and then clear them. Or to use the properties, simply select them.
SDC-3988	When you stop a cluster mode pipeline, the _tmp file might not be renamed.
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-3712	The Hadoop FS origin incorrectly lists MapR as an available stage library.
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.
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.

StreamSets Data Collector 2.0.0.0 Release Notes

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 `$SDC_DIST/libexec/sdcd-env.sh` Set the property greater than the largest record you want to process. For example, to process records up to 2 MB, set the property to 2097152 as follows: `SDC_JAVA_OPTS="-DDataFactoryBuilder.OverRunLimit=2097152"`
SDC-2552	When an invalid topic is specified for a MapR Streams Consumer or a MapR Streams Producer, the following incorrect message displays: `CONTAINER_0701 - Stage 'com_streamsets_pipeline_stage_origin_maprstreams_MapRStreamsDSource_1' initialization error: org.apache.kafka.common.config.ConfigException: No bootstrap urls given in bootstrap.servers` This message is misleading because MapR Streams does not support the `bootstrap.servers` option.
SDC-2374	A cluster mode pipeline can hang with a CONNECT_ERROR status. This can be a temporary connection problem that resolves, returning the pipeline to the RUNNING status. If the problem is not temporary, you might need to manually edit the pipeline state file to set the pipeline to STOPPED. Edit the file only after you confirm that the pipeline is no longer running on the cluster or that the cluster has been decommissioned. To manually change the pipeline state, edit the following file: `$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-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.

StreamSets Data Collector 2.0.0.0 Release Notes

	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/.

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.