StreamSets Data Collector 2.2.0.0 Release Notes

December 1, 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.

Event Framework

The Data Collector *event framework* enables the pipeline to trigger tasks in external systems based on actions that occur in the pipeline, such as running a MapReduce job after the pipeline writes a file to HDFS. You can also use the event framework to store event information, such as when an origin starts or completes reading a file.

For details, see the Event Framework chapter.

The event framework includes the following new features and enhancements:

- **New executor stages.** A new type of stage that performs tasks in external systems upon receiving an event. This release includes the following executors:
  - **HDFS File Metadata executor** - Changes file metadata such as the name, location, permissions, and ACLs.
  - **Hive Query executor** - Runs a Hive or Impala query.
  - **JDBC Query executor** - Runs a SQL query.
  - **MapReduce executor** - Runs a custom MapReduce job or an Avro to Parquet MapReduce job.

- **Event generation.** The following stages now generate events that you can use in a pipeline:
  - **Directory and File Tail origins** - Generate events when they start and complete reading a file.
  - **Amazon S3 destination** - Generates events when it completes writing to an object or streaming a whole file.
  - **Hadoop FS, Local FS, and MapR FS destinations** - Generate events when they close an output file or complete streaming a whole file.
  - **Groovy Evaluator, JavaScript Evaluator, and Jython Evaluator processors** - Can run scripts that generate events.
  - **HDFS File Metadata executor** - Generates events when it changes file metadata.
  - **Hive Query executor** - Generates events when it runs a Hive or Impala query.

- **Dev stages.** You can use the following stages to develop and test event handling:
  - **Dev Data Generator enhancement** - You can now configure the Dev Data Generator to generate event records as well as data records. You can also specify the number of records in a batch.
  - **To Event** - Generates event records using the incoming record as the body of the event record.

Installation

- **Java requirement.** Oracle Java 7 is supported but now deprecated. Oracle announced the end of public updates for Java 7 in April 2015. StreamSets recommends migrating to Java 8, as Java 7 support will be removed in a future Data Collector release.

- **File descriptors requirement.** Data Collector now requires a minimum of 32,768 open file descriptors.

- **Core installation includes the basic stage library only.** The core RPM and tarball installations now include the basic stage library only, to allow Data Collector to use less disk space. Install additional stage libraries using the Package Manager for tarball installations or the command line for RPM and tarball installations.

  Previously, the core installation also included the Groovy, Jython, and statistics stage libraries.

Configuration

- **New stage libraries.** Data Collector now supports the following stage libraries:
  - Apache Kudu version 1.0.x. Earlier Kudu versions are no longer supported.
  - Cloudera CDH version 5.9 distribution of Apache Hadoop.
  - Cloudera version 5.9 distribution of Apache Kafka 2.0.
  - Elasticsearch version 5.0.x.
  - Google Cloud Bigtable.
  - Hortonworks HDP version 2.5 distribution of Apache Hadoop.
  - MySQL Binary Log.
  - Salesforce.

- **LDAP authentication.** If you use LDAP authentication, you can now configure Data Collector to connect to multiple LDAP servers. You can also configure Data Collector to support an LDAP deployment where members are defined by uid or by full DN.

- **Java garbage collector.** Data Collector now uses the Concurrent Mark Sweep (CMS) garbage collector by default. You can configure Data Collector to use a different garbage collector by modifying Java configuration options in the Data Collector environment configuration file.

- **Environment variables for Java configuration options.** Data Collector now uses three environment variables to define Java configuration options:
  - SDC_JAVA_OPTS - Includes configuration options for all Java versions.
  - SDC_JAVA7_OPTS - Includes configuration options used only when Data Collector is running Java 7.
  - SDC_JAVA8_OPTS - Includes configuration options used only when Data Collector is running Java 8.

- **New time zone property.** You can configure the Data Collector console to use UTC, the browser time zone, or the Data Collector time zone. The time zone property affects how dates and times display in the UI. The default is the browser time zone.

Origins

- **New MySQL Binary Log origin.** Reads MySQL binary logs to generate records with change data capture information.

- **New Salesforce origin.** Reads data from Salesforce. The origin can execute a SOQL query to read existing data from Salesforce. The origin can also subscribe to the Force.com Streaming API to receive notifications for changes to Salesforce data.

- **Directory origin enhancement.** You can configure the Directory origin to read files from all subdirectories when using the last-modified timestamp for the read order.

- **JDBC Consumer and Oracle CDC Client origin enhancement.** You can now configure the transaction isolation level that the JDBC Consumer and Oracle CDC Client origins use to connect to the database. Previously, the origins used the default transaction isolation level configured for the database.

Processors

- **New Spark Evaluator processor.** Processes data based on a Spark application that you develop. Use the Spark Evaluator processor to develop a Spark application that performs custom processing within a pipeline.

- **Field Flattener processor enhancements.** In addition to flattening the entire record, you can also now use the Field Flattener processor to flatten specific list or map fields in the record.

- **Field Type Converter processor enhancements.** You can now use the Field Type Converter processor to change the scale of a decimal field. Or, if you convert a field with another data type to the Decimal data type, you can configure the scale to use in the conversion.

- **Field Pivoter processor enhancements.** The List Pivoter processor has been renamed to the Field Pivoter processor. You can now use the processor to pivot data in a list, map, or list-map field. You can also use the processor to save the field name of the first-level item in the pivoted field.

- **JDBC Lookup and JDBC Tee processor enhancement.** You can now configure the transaction isolation level that the JDBC Lookup and JDBC Tee processors use to connect to the database. Previously, the origins used the default transaction isolation level configured for the database.

- **Scripting processor enhancements.** The Groovy Evaluator, JavaScript Evaluator, and Jython Evaluator processors can generate event records and work with record header attributes. The sample scripts now include examples of both and a new tip for generating unique record IDs.

- **XML Flattener processor enhancement.** You can now configure the XML Flattener processor to write the flattened data to a new output field. Previously, the processor wrote the flattened data to the same field.

- **XML Parser processor enhancement.** You can now generate records from XML documents using simplified XPath expressions. This enables reading records from deeper within XML documents.

Destinations

- **New Azure Data Lake Store destination.** Writes data to Microsoft Azure Data Lake Store.

- **New Google Bigtable destination.** Writes data to Google Cloud Bigtable.

- **New Salesforce destination.** Writes data to Salesforce.
- **New Wave Analytics destination.** Writes data to Salesforce Wave Analytics. The destination creates a dataset with external data.
- **Amazon S3 destination change.** The AWS KMS Key ID property has been renamed AWS KMS Key ARN. Data Collector upgrades existing pipelines seamlessly.
- **File suffix enhancement.** You can now configure a file suffix, such as txt or json, for output files generated by Hadoop FS, Local FS, MapR FS, and the Amazon S3 destinations.
- **JDBC Producer destination enhancement.** You can now configure the transaction isolation level that the JDBC Producer destination uses to connect to the database. Previously, the destination used the default transaction isolation level configured for the database.
- **Kudu destination enhancement.** You can now configure the destination to perform one of the following write operations: insert, update, delete, or upsert.

Data Formats

- **XML processing enhancement.** You can now generate records from XML documents using simplified XPath expressions with origins that process XML data and the XML Parser processor. This enables reading records from deeper within XML documents.
- **Consolidated data format properties.** You now configure the data format and related properties on a new Data Format tab. Previously, data formats had individual configuration tabs, e.g., Avro, Delimited, Log.

  Related properties, such as Charset, Compression Format, and Ignore Control Characters now appear on the Data Format tab as well.
- **Checksum generation for whole files.** Destinations that stream whole files can now generate checksums for the files so you can confirm the accurate transmission of the file.

Pipeline Maintenance

- **Add labels to pipelines from the Home page.** You can now add labels to multiple pipelines from the Data Collector Home page. Use labels to group similar pipelines. For example, you might want to group pipelines by database schema or by the test or production environment.
- **Reset the origin for multiple pipelines from the Home page.** You can now reset the origin for multiple pipelines at the same time from the Data Collector Home page.

Rules and Alerts

- **Metric rules and alerts enhancements.** The gauge metric type can now provide alerts based on the number of input, output, or error records for the last processed batch.

Expression Language Functions

- **New file functions.** You can use the following new file functions to work with file paths:
  - file:fileExtension(<filepath>) - Returns the file extension from a path.
  - file:fileName(<filepath>) - Returns a file name from a path.

    - file:parentPath(<filepath>) - Returns the parent path of the specified file or directory.
    - file:pathElement(<filepath>, <integer>) - Returns the portion of the file path specified by a positive or negative integer.
    - file:removeExtension(<filepath>) - Removes the file extension from a path.

- **New pipeline functions.** You can use the following new pipeline functions to determine information about a pipeline:
    - pipeline:name() - Returns the pipeline name.
    - pipeline:version() - Returns the pipeline version when the pipeline has been published to Dataflow Performance Manager (DPM).

- **New time functions.** You can use the following new time functions to transform datetime data:
    - time:extractLongFromDate(<Date object>, <string>) - Extracts a long value from a Date object, based on the specified date format.
    - time:extractStringFromDate(<Date object>, <string>) - Extracts a string value from a Date object, based on the specified date format.
    - time:millisecondsToDateTime(<long>) - Converts an epoch or UNIX time in milliseconds to a Date object.

Upgrade

You can upgrade previous versions of Data Collector to version 2.2.0.0. For instructions on upgrading, see the Upgrade Documentation.

Update Kudu Pipelines

Data Collector now 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 deprecated Kudu version to the Data Collector configuration file, `$SDC_CONF/sdc.properties` For more information, see Update Kudu Pipelines.

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-4555	Firefox truncates pipeline names that include a space during export of the pipelines.
SDC-4511	Data Collector cannot read Avro schemas that include arrays.
SDC-4497	Data Collector should not allow empty passwords for certain LDAP configurations.
SDC-4351	Exception occurs when refreshing preview after changing invalid data to valid data.

SDC-4204	When the Amazon S3 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-4172	Data Collector cannot access Vault secrets stored in Hashicorp Vault.
SDC-4047	The XML Flattener processor fails to parse XML that contains whitespace after the XML prolog.
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.

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.
SDC-4609	A pipeline stops unexpectedly when the Google Bigtable destination encounters non-numeric values for timestamps.
SDC-4608	The Google Bigtable destination encounters a null pointer exception when the stage uses the time associated with the record as the time basis and the specified field doesn't exist.
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/.

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.