

StreamSets Data Collector 1.3.0.0 Release Notes

April 13, 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 Kinesis Firehose destination
- New MongoDB destination
- New Kudu destination
- New Base64 Field Decoder and Base64 Field Encoder processors
- File Tail and Directory origins allow reading from late directories
- Additional metrics for the File Tail origin, viewable with external JMX tools
- Field Hasher allows hashing in place, hashing to a target field or header, and hashing the entire record
- The HBase destination supports implicit field mappings
- The Kinesis Consumer origin supports AWS proxy settings
- The JMS Consumer origin provides configurable custom JNDI properties
- Users with the Admin role can restart Data Collector from the console
- Configurable timeout for inactive user sessions
- REST API support for cross-origin resource sharing (CORS)
- Data Collector now uses webhelp to display the latest available documentation. When internet access is not available, the console uses the local help installed with the product.

Please feel free to check out the [Documentation](#) for this release.

Upgrade Notes

Reconcile the Data Collector Configuration Files

When you upgrade Data Collector, you need to manually update the new Data Collector configuration file, `$(SDC_CONF)/sdc.properties`.

A new release can include new properties required for Data Collector to start or function properly. When you install a new version of Data Collector, compare the existing and new versions of the file, and update the new file as needed.

Fully-Qualified Protobuf Message Type

When upgrading Data Collector from any version earlier than 1.2.1.0, note the following requirement if you process protobuf data.

Data Collector now requires a fully-qualified message type name, as follows:

```
<package name>.<message type name>
```

If you were processing protobuf data with a previous version of Data Collector, perform the following steps:

StreamSets Data Collector 1.3.0.0 Release Notes

1. Edit all pipelines that read or write protobuf data.
2. For the origin or destination, on the **Protobuf** tab, update the **Message Type** property to add the package name as follows: <package name>.<message type name>.

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-2340	When processing Avro files, leave the Compression Format property set to None in the origin. Typically compression is handled within Avro, and the Avro file does not need to be compressed.

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-2756	The Hadoop FS and Local FS destinations cannot write files larger than 2 GB at this time.
SDC-2716	Installations with Cloudera Manager cannot be configured for LDAP authentication at this time.
SDC-2586	<p>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, <code>SDC_CONF/sdc.properties</code>.</p> <p>Workaround: Set the value of DataFactoryBuilder.OverRunLimit property in the SDC_JAVA_OPTS environment variable in the Data Collector environment file, <code>SDC_DIST/libexec/sdc-env.sh</code> or <code>SDC_DIST/libexec/sdcd-env.sh</code>.</p> <p>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:</p> <pre>SDC_JAVA_OPTS="-DDataFactoryBuilder.OverRunLimit=2097152"</pre>

StreamSets Data Collector 1.3.0.0 Release Notes

JIRA	Description
SDC-2554	MapR FS does not support Kerberos authentication at this time.
SDC-2552	<p>When an invalid topic is specified for a MapR Streams Consumer or a MapR Streams Producer, the following incorrect message displays:</p> <pre>CONTAINER_0701 - Stage 'com_streamsets_pipeline_stage_origin_maprstreams_MapRStreamsD Source_1' initialization error: org.apache.kafka.common.config.ConfigException: No bootstrap urls given in bootstrap.servers</pre> <p>This message is misleading because MapR Streams does not support the <code>bootstrap.servers</code> option.</p>
SDC-2374	<p>A cluster mode pipeline can hang with a <code>CONNECT_ERROR</code> status. This can be a temporary connection problem that resolves, returning the pipeline to the <code>RUNNING</code> status.</p> <p>If the problem is not temporary, you might need to manually edit the pipeline state file to set the pipeline to <code>STOPPED</code>. Edit the file only after you confirm that the pipeline is no longer running on the cluster or that the cluster has been decommissioned.</p> <p>To manually change the pipeline state, edit the following file: <code>\$SDC_DATA/runInfo/<cluster pipeline name>/<revision>/pipelineState.json</code></p> <p>In the file, change <code>CONNECT_ERROR</code> to <code>STOPPED</code> and save the file.</p>
SDC-2359	<p>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.</p> <p>For more information about the Kafka JIRA, see https://issues.apache.org/jira/browse/KAFKA-2880.</p>
SDC-2226	When the Data Collector UI is configured for HTTPS, pipelines with the Hadoop FS origin fail to start.
SDC-2133	You cannot stop a pipeline in the middle of a long-running processor, such as a Jython Evaluator performing a complex script.

StreamSets Data Collector 1.3.0.0 Release Notes

SDC-1731	<p>When using the Kafka Consumer or Kafka Producer on HDP 2.3 with Kerberos enabled, set the Kafka broker configuration property <code>security.inter.broker.protocol</code> to PLAINTEXT.</p> <p>When enabling Kerberos, HDP 2.3 sets the <code>security.inter.broker.protocol</code> property to PLAINTEXTSASL, which is not supported.</p> <p>If the property is not set to PLAINTEXT, when the pipeline starts, validation errors indicate a problem connecting to Kafka.</p>
SDC-1567	<p>You cannot use cluster mode pipelines to read from HDP 2.3 due to a HDP integration issue with Kafka and Spark Streaming.</p> <p>For more information about the HDP issue, see http://hortonworks.com/community/forums/topic/kafka-and-spark-streaming-nosuch-methoderror-kafka-consumer-simpleconsumer/.</p>
SDC-891	<p>At this time, writing to error records to file is not supported for cluster mode pipelines.</p> <p>Workaround: Write error records to Kafka or to an SDC RPC pipeline.</p>
SDC-890	<p>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.</p> <p>Workaround: To stop all pipelines, use the Stop icon in the Data Collector console.</p>

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.