Cloudera DataFlow for Data Hub 7.2.10

Cloudera DataFlow for Data Hub Release Notes

Date published: 2019-12-16
Date modified: 2021-06-14

https://docs.cloudera.com/

Legal Notice

The documentation is and contains Cloudera proprietary information protected by copyright and other intellectual property rights. No license under copyright or any other intellectual property right is granted herein.

Copyright information for Cloudera software may be found within the documentation accompanying each component in a particular release.

Cloudera software includes software from various open source or other third party projects, and may be released under the Apache Software License 2.0 (“ASLv2”), the Affero General Public License version 3 (AGPLv3), or other license terms. Other software included may be released under the terms of alternative open source licenses. Please review the license and notice files accompanying the software for additional licensing information.

Please visit the Cloudera software product page for more information on Cloudera software. For more information on Cloudera support services, please visit either the Support or Sales page. Feel free to contact us directly to discuss your specific needs.

Cloudera reserves the right to change any products at any time, and without notice. Cloudera assumes no responsibility nor liability arising from the use of products, except as expressly agreed to in writing by Cloudera.

Cloudera, Cloudera Altus, HUE, Impala, Cloudera Impala, and other Cloudera marks are registered or unregistered trademarks in the United States and other countries. All other trademarks are the property of their respective owners.

Disclaimer: EXCEPT AS EXPRESSLY PROVIDED IN A WRITTEN AGREEMENT WITH CLOUDERA, CLOUDERA DOES NOT MAKE NOR GIVE ANY REPRESENTATION, WARRANTY, NOR COVENANT OF ANY KIND, WHETHER EXPRESS OR IMPLIED, IN CONNECTION WITH CLOUDERA TECHNOLOGY OR RELATED SUPPORT PROVIDED IN CONNECTION THEREWITH. CLOUDERA DOES NOT WARRANT THAT CLOUDERA PRODUCTS NOR SOFTWARE WILL OPERATE UNINTERRUPTED NOR THAT IT WILL BE FREE FROM DEFECTS NOR ERRORS, THAT IT WILL PROTECT YOUR DATA FROM LOSS, CORRUPTION NOR UNAVAILABILITY, NOR THAT IT WILL MEET ALL OF CUSTOMER’S BUSINESS REQUIREMENTS. WITHOUT LIMITING THE FOREGOING, AND TO THE MAXIMUM EXTENT PERMITTED BY APPLICABLE LAW, CLOUDERA EXPRESSLY DISCLAIMS ANY AND ALL IMPLIED WARRANTIES, INCLUDING, BUT NOT LIMITED TO IMPLIED WARRANTIES OF MERCHANTABILITY, QUALITY, NON-INFRINGEMENT, TITLE, AND FITNESS FOR A PARTICULAR PURPOSE AND ANY REPRESENTATION, WARRANTY, OR COVENANT BASED ON COURSE OF DEALING OR USAGE IN TRADE.

Contents

What's New in Cloudera DataFlow for Data Hub 7.2.10....................................4
What's New in Flow Management...4
What's New in Streams Messaging...4

Component Support in Cloudera DataFlow for Data Hub 7.2.10....................... 5

Supported NiFi Extensions... 6
Supported NiFi Processors.. 6
Supported NiFi Controller Services.. 8
Supported NiFi Reporting Tasks..9

Unsupported Features in Cloudera DataFlow for Data Hub 7.2.10.................. 10
Unsupported Flow Management features...10
Unsupported Streams Messaging features.. 10

Apache Patch Information in Cloudera DataFlow for Data Hub 7.2.10........... 11
NiFi patches... 11
NiFi Registry patches... 11

Known Issues In Cloudera DataFlow for Data Hub 7.2.10................................12
Known Issues in Flow Management..12
Known Issues in Streams Messaging...13

Fixed Issues in Cloudera DataFlow for Data Hub 7.2.10.................................. 19
Fixed Issues in Flow Management.. 19
Fixed Issues in Streams Messaging.. 20

Fixed Issues in Cloudera DataFlow for Data Hub 7.2.10.1............................... 21

What's New in Cloudera DataFlow for Data Hub 7.2.10

Cloudera DataFlow for Data Hub 7.2.10 includes components for Flow Management, Streaming Analytics, and Streams Messaging. Learn about the new features and improvements in each of these components.

What's New in Flow Management

Learn about the new Flow Management features in Cloudera DataFlow for Data Hub 7.2.10.

The Flow Management clusters in CDP Public Cloud introduces the following features:

Encrypt and Decrypt PGP Processors

Flow Management clusters in CDP Public Cloud 7.2.10 include Encrypt and Decrypt PGP Processors and Services.

Apache NiFi Jira: NIFI-8251

ReplaceText improvements

`ReplaceText` includes a variable replacement strategy in CDP Public Cloud 7.2.10.

Apache NiFi Jira: NIFI-8474

PublishKafka processor improvements

`PublishKafka` and `PublishKafkaRecord` processors now include support for Rollback on Failure.

Flow Management cluster upgrade

You can use the Software Only upgrade procedure to upgrade Flow Management clusters in CDP Public cloud from the following CDP Public Cloud versions to CDP Public Cloud 7.2.10.

- CDP Public Cloud 7.2.9
- CDP Public Cloud 7.2.8
- CDP Public Cloud 7.2.7
- CDP Public Cloud 7.2.6
- CDP Public Cloud 7.2.2
- CDP Public Cloud 7.2.1
- CDP Public Cloud 7.2.0

Note:

This feature is in technical preview and customers should request access to it by reaching out to Cloudera via the support portal.

What's New in Streams Messaging

Learn about the new Streams Messaging features in Cloudera DataFlow for Data Hub 7.2.10.

Kafka

There are no new features for Kafka in this release.

Schema Registry

There are no new features for Schema Registry in this release.

Streams Messaging Manager

There are no new features for Streams Messaging Manager in this release.

Streams Replication Manager

Sensitive cluster connection information is now stored securely

SRM is now capable of storing all sensitive data in a secure manner. As a result of this improvement, the recommended method of how you configure clusters and cluster connection information for the SRM service (Driver and Service roles) and the srm-control tool has changed.

Previously, cluster connection information (aliases, bootstrap servers, security properties) was configured through the Streams Replication Manager's Replication Configs Cloudera Manager property. From now on, both external and co-located clusters can be defined using a new configuration pane in Cloudera Manager. In addition, co-located clusters can also be configured with a service dependency. The new configuration pane is called Kafka credentials and can be found in AdministrationExternal Accounts>Kafka credentials.

Additionally, an intermediary keystore that stores connection related sensitive data called SRM client's secure storage can be set up and configured in Cloudera Manager. This secure storage acts as an extension to the srm-control tool’s default configuration and must be set up for the tool if SRM is replicating a secure cluster.

Using the new configuration options and methods makes it possible to securely store all sensitive data that is added to SRM’s configuration.

⚠ **Important:** The old method of configuring connection related information with Streams Replication Manager's Replication Configs is still supported. However, Cloudera does not recommend that you use this property to specify cluster connection information.

New documentation is introduced that walks users through the new configuration workflows. For more information, on how to configure clusters using the new configuration options and workflow, see Defining and adding clusters for replication. For more information regarding the new configuration workflow for the srm-control tool, see Configuring srm-control. Additionally, all existing documentation affected by this change is also updated.

Related Information

Kafka Properties in Cloudera Runtime

Importing Confluent Schema Registry schemas into Cloudera Schema Registry

Component Support in Cloudera DataFlow for Data Hub 7.2.10

Cloudera DataFlow for Data Hub 7.2.10 includes the following components.

Flow Management clusters

- Apache NiFi 1.13.2
- Apache NiFi Registry 0.8.0

Streams Messaging clusters

- Apache Kafka 2.5.0
- Schema Registry 0.10.0
- Streams Messaging Manager 2.1.0
- Streams Replication Manager 1.0.0

Supported NiFi Extensions

Apache NiFi 1.13.2 ships with a set of Processors, Controller Services, and Reporting Tasks, most of which are supported by Cloudera Support. Review the supported extensions and avoid using any unsupported extensions in your production environments.

Supported NiFi Processors

This release ships with Apache NiFi 1.13.2 and includes a set of Processors, most of which are supported by Cloudera Support. You should be familiar with the available supported Processors, and avoid using any unsupported Processors in production environments.

Additional Processors are developed and tested by the Cloudera community but are not officially supported by Cloudera. Processors are excluded for a variety of reasons, including insufficient reliability or incomplete test case coverage, declaration of non-production readiness by the community at large, and feature deviation from Cloudera best practices.

- AttributesToCSV
- AttributesToJSON
- Base64EncodeContent
- CalculateRecordStats
- CaptureChangeMySQL
- CompressContent

  Memory intensive

  CPU intensive
- ConnectWebSocket
- ConsumeAMQP
- ConsumeAzureEventHub
- ConsumeEWS
- ConsumeGCPubSub
- ConsumeJMS
- ConsumeKafka
- ConsumeKafka_0_10
- ConsumeKafka_1_0
- ConsumeKafka_2_0
- ConsumeKafka_2_6
- ConsumeKafkaRecord_0_10
- ConsumeKafkaRecord_1_0
- ConsumeKafkaRecord_2_0
- ConsumeKafkaRecord_2_6
- ConsumeMQTT

  Memory intensive
- ConsumeWindowsEventLog
- ControlRate
- ConvertAvroSchema
- ConvertAvroToJSON
- ConvertAvroToORC
- ConvertAvroToParquet
- GetFile
- GetFTP
- GetHBase
- GetHDFS
- GetHDFSFileInfo
- GetHDFSSequenceFile
- GetHTMLElement
- GetHTTP
- GetIgniteCache
- GetJMSQueue
- GetJMSTopic
- GetKafka
- GetMongoRecord
- GetSFTP
- GetSolr
- GetSplunk
- GetSQS
- GetTCP
- GetTwitter
- HandleHttpRequest
- HandleHttpResponse
- HashAttribute
- HashContent
- IdentifyMimeType
- InvokeAWSGatewayApi
- InvokeGRPC
- InvokeHTTP
- InvokeScriptedProcessor
- JoltTransformJSON
- JoltTransformRecord
- JsonQueryElasticsearch
- ListAzureBlobStorage
- PutDistributedMapCache
- PutDynamoDB

  Memory intensive
- PutElasticsearch

  Memory intensive
- PutElasticsearchHttp

  Memory intensive
- PutElasticsearchHttpRecord
- PutElasticsearchRecord
- PutEmail

  Memory intensive
- PutFile
- PutFTP
- PutGCSObject
- PutGridFS
- PutHBaseCell

  Memory intensive
- PutHBaseJSON
- PutHBaseRecord
- PutHDFS
- PutHive3QL
- PutHive3Streaming
- PutHiveQL
- PutHiveStreaming
- PutHTMLElement
- PutInfluxDB
- PutJMS
- PutKafka
- PutKinesisFirehose
- PutKinesisStream
- PutKudu

- ConvertCharacterSet
- ConvertCSVToAvro
- ConvertJSONToAvro
- ConvertJSONToSQL
- ConvertRecord
- CreateHadoopSequenceFile
- CryptographicHashAttribute
- CryptographicHashContent
- DeleteAzureBlobStorage
- DeleteAzureDataLakeStorage
- DeleteByQueryElasticsearch
- DeleteDynamoDB
- DeleteGCSObject
- DeleteGridFS
- DeleteHBaseCells
- DeleteHBaseRow
- DeleteHDFS
- DeleteS3Object
- DeleteSQS
- DetectDuplicate
- DistributeLoad
- DuplicateFlowFile
- EncryptContent

  Memory intensive
- EnforceOrder
- EvaluateJsonPath
- EvaluateXPath
- EvaluateXQuery
- ExecuteGroovyScript
- ExecuteInfluxDBQuery
- ExecuteProcess
- ExecuteScript
- ExecuteSQL
- ExecuteSQLRecord
- ExecuteStreamCommand
- ExtractAvroMetadata
- ExtractGrok
- ExtractHL7Attributes
- ExtractImageMetadata
- ExtractText
- FetchAzureBlobStorage
- FetchAzureDataLakeStorage
- FetchDistributedMapCache
- FetchElasticsearch
- FetchElasticsearchHttp
- FetchFile
- FetchFTP
- FetchGCSObject
- FetchGridFS
- ListAzureDataLakeStorage
- ListDatabaseTables
- ListenFTP
- ListenGRPC
- ListenHTTP
- ListenRELP
- ListenSyslog
- ListenTCP
- ListenTCPRecord
- ListenUDP
- ListenUDPRecord
- ListenWebSocket
- ListFile
- ListFTP
- ListGCSBucket
- ListHDFS
- ListS3
- ListSFTP
- LogAttribute
- LogMessage
- LookupAttribute
- LookupRecord
- MergeContent

  Memory intensive
- MergeRecord
- ModifyHTMLElement
- MonitorActivity
- Notify
- ParseCEF
- ParseEvtx
- ParseSyslog
- PartitionRecord
- PostHTTP
- PublishAMQP

  Memory intensive
- PublishGCPubSub

  Memory intensive
- PublishJMS

  Memory intensive
- PublishKafka
- PublishKafka_0_10
- PublishKafka_1_0
- PublishKafka_2_0
- PublishKafka_2_6
- PublishKafkaRecord_0_10
- PublishKafkaRecord_1_0
- PublishKafkaRecord_2_0
- PublishKafkaRecord_2_6
- PutLambda
- PutMongoRecord
- PutORC
- PutParquet
- PutRecord
- PutRiemann
- PutS3Object
- PutSFTP
- PutSNS
- PutSolrContentStream
- PutSolrRecord
- PutSplunk
- PutSplunkHTTP

  Memory intensive
- PutSQL
- PutSQS
- PutSyslog
- PutTCP
- PutUDP
- PutWebSocket

  Memory intensive
- QueryCassandra
- QueryDatabaseTable
- QueryDatabaseTableRecord
- QueryElasticsearchHttp
- QueryRecord
- QuerySolr
- QuerySplunkIndexingStatus
- QueryWhois
- ReplaceText

  Memory intensive
- ReplaceTextWithMapping
- ResizeImage
- RetryFlowFile
- RouteHL7
- RouteOnAttribute
- RouteOnContent
- RouteText
- SampleRecord

  Memory intensive
- ScanAccumulo
- ScanAttribute
- ScanContent
- ScanHBase
- ScriptedTransformRecord
- ScrollElasticsearchHttp
- SegmentContent
- SelectHive3QL

- FetchHBaseRow
- FetchHDFS
- FetchParquet
- FetchS3Object
- FetchSFTP
- FlattenJson
- ForkRecord
- GenerateFlowFile
- GenerateTableFetch
- GeoEnrichIP
- GeoEnrichIPRecord
- GetAzureEventHub
- GetAzureQueueStorage
- GetCouchbaseKey

  Memory intensive
- PublishMQTT

  Memory intensive
- PutAccumuloRecord
- PutAzureBlobStorage
- PutAzureCosmosDBRecord

  Memory intensive
- PutAzureDataLakeStorage
- PutAzureEventHub

  Memory intensive
- PutAzureQueueStorage
- PutBigQueryBatch
- PutBigQueryStreaming

  Memory intensive
- PutCassandraQL

  Memory intensive
- PutCassandraRecord
- PutCloudWatchMetric
- PutCouchbaseKey

  Memory intensive
- PutDatabaseRecord
- SelectHiveQL
- SplitAvro

  Memory intensive
- SplitContent

  Memory intensive
- SplitJson

  Memory intensive
- SplitRecord
- SplitText

  Memory intensive
- SplitXml

  Memory intensive
- TagS3Object
- TailFile
- TransformXml
- UnpackContent
- UpdateAttribute
- UpdateCounter
- UpdateHive3Table
- UpdateHiveTable
- UpdateRecord
- ValidateCsv
- ValidateRecord
- ValidateXml
- Wait
- YandexTranslate

Supported NiFi Controller Services

This release ships with Apache NiFi 1.13.2 and includes a set of Controller Services, most of which are supported by Cloudera Support. You should be familiar with the available supported Controller Services, and avoid using any unsupported Controller Services in production environments.

Additional Controller Services are developed and tested by the Cloudera community but are not officially supported by Cloudera. Controller Services are excluded for a variety of reasons, including insufficient reliability or incomplete test case coverage, declaration of non-production readiness by the community at large, and feature deviation from Cloudera best practices.

AccumuloService
ActionHandlerLookup
ADLSCredentialsControllerService
ADLSIDBrokerCloudCredentialsProviderControllerService
AlertHandler
AvroReader
AvroRecordSetWriter
AvroSchemaRegistry
AWSCredentialsProviderControllerService
AWSIDBrokerCloudCredentialsProviderControllerService
AzureBlobIDBrokerCloudCredentialsProviderControllerService
ElasticSearchClientServiceImpl
ElasticSearchLookupService
ElasticSearchStringLookupService
EmbeddedHazelcastCacheManager
ExpressionHandler
ExternalHazelcastCacheManager
FreeFormTextRecordSetWriter
GCPCredentialsControllerService
GrokReader
HadoopDBCPConnectionPool
HazelcastMapCacheClient
LogHandler
MongoDBControllerService
MongoDBLookupService
ParquetReader
ParquetRecordSetWriter
PrometheusRecordSink
ReaderLookup
RecordSetWriterLookup
RecordSinkHandler
RecordSinkServiceLookup
RedisConnectionPoolService

AzureCosmosDBClientService
AzureStorageCredentialsControllerService
AzureStorageCredentialsControllerServiceLookup
CassandraDistributedMapCache
CassandraSessionProvider
CouchbaseClusterService
CouchbaseKeyValueLookupService
CouchbaseMapCacheClient
CouchbaseRecordLookupService
CSVReader
CSVRecordLookupService
CSVRecordSetWriter
DatabaseRecordLookupService
DatabaseRecordSink
DBCPConnectionPool
DBCPConnectionPoolLookup
DistributedMapCacheClientService
DistributedMapCacheLookupService
DistributedMapCacheServer
DistributedSetCacheClientService
DistributedSetCacheServer
EasyRulesEngineProvider
EasyRulesEngineService
HBase_1_1_2_ClientMapCacheService
HBase_1_1_2_ClientService
HBase_1_1_2_ListLookupService
HBase_1_1_2_RecordLookupService
HBase_2_ClientMapCacheService
HBase_2_ClientService
HBase_2_RecordLookupService
Hive3ConnectionPool
HiveConnectionPool
HortonworksSchemaRegistry
IPFIXReader
IPLookupService
JMSConnectionFactoryProvider
JndiJmsConnectionFactoryProvider
JsonPathReader
JsonRecordSetWriter
JsonTreeReader
KafkaRecordSink_1_0
KafkaRecordSink_2_0
KafkaRecordSink_2_6
KeytabCredentialsService
KuduLookupService
LoggingRecordSink
RedisDistributedMapCacheClientService
RestLookupService
ScriptedActionHandler
ScriptedLookupService
ScriptedReader
ScriptedRecordSetWriter
ScriptedRecordSink
ScriptedRulesEngine
SimpleDatabaseLookupService
SimpleKeyValueLookupService
SimpleScriptedLookupService
SiteToSiteReportingRecordSink
StandardHttpContextMap
StandardProxyConfigurationService
StandardRestrictedSSLContextService
StandardS3EncryptionService
StandardSSLContextService
Syslog5424Reader
SyslogReader
VolatileSchemaCache
WindowsEventLogReader
XMLReader
XMLRecordSetWriter

Supported NiFi Reporting Tasks

This release ships with Apache NiFi 1.13.2 and includes a set of Reporting Tasks, most of which are supported by Cloudera Support. You should be familiar with the available supported Reporting Tasks, and avoid using any unsupported Reporting Tasks in production environments.

Additional Reporting Tasks are developed and tested by the Cloudera community but are not officially supported by Cloudera. Reporting Tasks are excluded for a variety of reasons, including insufficient reliability or incomplete test case coverage, declaration of non-production readiness by the community at large, and feature deviation from Cloudera best practices. Do not use these features in your production environments.

- AmbariReportingTask
- ControllerStatusReportingTask
- MetricsEventReportingTask
- MonitorDiskUsage
- MonitorMemory
- PrometheusReportingTask
- QueryNiFiReportingTask
- ReportLineageToAtlas
- ScriptedReportingTask
- SiteToSiteBulletinReportingTask
- SiteToSiteMetricsReportingTask
- SiteToSiteProvenanceReportingTask
- SiteToSiteStatusReportingTask

Unsupported Features in Cloudera DataFlow for Data Hub 7.2.10

Some features exist within Cloudera DataFlow for Data Hub 7.2.10 components, but are not supported by Cloudera.

Note: The Streaming Analytics 7.2.10 templates were removed from Data Hub due to high security vulnerabilities exposed by Apache Log4j2 (CVE-2021-45105). Cloudera recommends upgrading to the latest available version of the Streaming Analytics templates.

Unsupported Flow Management features

There are no unsupported Flow Management features in Cloudera DataFlow for Data Hub 7.2.10

NiFi

There are no updates for this release.

NiFi Registry

There are no updates for this release.

Related Information

Cloudera Community Forum

Unsupported Streams Messaging features

Some Streams Messaging features exist in Cloudera DataFlow for Data Hub 7.2.10, but are not supported by Cloudera.

Kafka

The following Kafka features are not ready for production deployment. Cloudera encourages you to explore these features in non-production environments and provide feedback on your experiences through the *Cloudera Community Forums*.

- Only Java and .Net based clients are supported. Clients developed with C, C++, Python, and other languages are currently not supported.
- While Kafka Connect is available as part of Runtime, it is currently not supported in CDP Public Cloud. NiFi is a proven solution for batch and real time data loading that complement Kafka's message broker capability. For more information, see Creating your first Flow Management cluster.
- The Kafka default authorizer is not supported. This includes setting ACLs and all related APIs, broker functionality, and command-line tools.

Schema Registry

There are no updates for this release.

Streams Messaging Manager

There are no updates for this release.

Streams Replication Manager

There are no updates for this release.

Related Information

Cloudera Community Forum

Creating your first Streams Messaging cluster

Apache Patch Information in Cloudera DataFlow for Data Hub 7.2.10

The following sections list patches in each Cloudera DataFlow in Data Hub component, beyond what was fixed in the base version of the Apache component.

NiFi patches

This release provides Apache NiFi 1.13.2 and these additional Apache patches.

- NIFI-8360 - SplitContent does not find any 'splits' that occur after about 2 GB into FlowFile
- NIFI-8357 - ConsumeKafka(Record)_2_0, ConsumeKafka(Record)_2_6 do not reconnect if using statically assigned partitions
- NIFI-8353 - When node is offloaded, it may still receive data from load-balanced connections
- NIFI-8346 - PutAzureBlobStorage doesn't route to failure despite the exception during upload
- NIFI-8344 - Allow TailFile to continue tailing a file for some time after it has been rolled over
- NIFI-8319 - EncryptContent should support decrypting AES/CBC/NoPadding
- NIFI-8314 - Generate warning for any long-running tasks
- NIFI-8313 - Upgrade zip4j to 2.7.0
- NIFI-8312 - Support PKCS12 and BCFKS truststores in Atlas reporting task
- NIFI-8307 - Controller Services not fully enabling on startup, preventing NiFi from completing startup
- NIFI-8302 - Correct Sensitive Value Encoding in FingerprintFactory
- NIFI-8296 - Integration with API to retrieve all subjects associated with a schema id for Confluent Schema Registry v5.3.1+
- NIFI-8289 - EmbeddedQuestDbRolloverHandlerTest tests fail when local date and UTC date differ
- NIFI-8286 - CertificateUtils do not support embedded emailAddress in CN
- NIFI-8283 - Value handling in ScanAccumulo processor
- NIFI-8263 - ListenHTTP - thread pool size
- NIFI-8260 - Process Group Import JSON file
- NIFI-8258 - Add support for Service Principal authentication in ADLS processors
- NIFI-8224 - Add LoggingRecordSink controller service
- NIFI-8212 - Improve startup times for Stateless
- NIFI-8188 - Processors: right click / run once
- NIFI-8132 - Replace Framework Uses of MD5 with Modern Algorithm
- NIFI-8113 - Persisting status history
- NIFI-8030 - Improve Atlas lineage when using PutHDFS to push data accessed through Hive
- NIFI-7668 - Add configurable PBE AEAD algorithms to flow encryption
- NIFI-7127 - Allow injection of SecureHasher into FingerprintFactory
- NIFI-6752 - Create ASN.1 RecordReader

For more information on fixed Apache NiFi patches, see the *Apache NiFi Release Notes*.

NiFi Registry patches

This release provides Apache NiFi Registry 0.8.0 and these additional Apache patches.

- NIFIREG-434 - Support BCFKS Keystore Type
- NIFIREG-429 - Flyway error when upgrading from older release to 0.8.0

For more information on fixed Apache NiFi Registry patches, see the *Apache NiFi Registry Release Notes*.

Known Issues In Cloudera DataFlow for Data Hub 7.2.10

You must be aware of the known issues and limitations, the areas of impact, and workaround in Cloudera DataFlow for Data Hub 7.2.10.

Known Issues in Flow Management

Learn about the known issues in Flow Management clusters, the impact or changes to the functionality, and the workaround.

NiFi cannot connect to NiFi Registry

By default, NiFi is configured with a NiFi Registry client to interact with the NiFi Registry instance. The URL used to configure the Registry client may not be correct depending on your deployment model for CDP Public Cloud. For example:

```
https://***gateway***/.../.../cdp-proxy/nifi-registry-app/nifi-registry/
```

If the URL is not correct, you may face "connect timed out" errors when interacting with NiFi Registry from the NiFi UI.

You can manually change the configuration of the client and provide the right FQDN of the management node of the DataHub cluster where the NiFi Registry instance is installed. To update the NiFi Registry client, go into the top right Actions menu, and select Controller Settings | Registry Clients. A correct URL will look similar to:

```
https://***management0***/.../cdp-proxy/nifi-registry-app/nifi-registry/
```

JDK versions mismatch

If doing a software only upgrade for your Flow Management DataHub clusters and if repairing one of the NiFi nodes after the upgrade, you may be in a situation where the JDK used by NiFi is not the same across the nodes. In such a case, this may cause issues in the NiFi UI and you may get an "Unexpected error" message.

Ensure that the same JDK is used across the NiFi nodes and if there is a JDK versions mismatch, manually upgrade the JDK to match the JDK version being installed on the node that has been repaired.

NiFi UI Performance considerations

A known issue in Chrome 92.x causes significant slowness in the NiFi UI and may lead to high CPU consumption. For more information, see the Chrome Known Issues documentation at 1235045.

Use another version of Chrome or a different browser.

Technical Service Bulletins

TSB 2022-580: NiFi Processors cannot write to content repository

If the content repository disk is filled more than 50% (or any other value that is set in nifi.properties for nifi.content.repository.archive.max.usage.percentage), and if there is no data in the content repository archive, the following warning message can be found in the logs: "Unable to write flowfile content to content repository container default due to archive file size constraints; waiting for archive cleanup". This would block the processors and no more data is processed.

This appears to only happen if there is already data in the content repository on startup that needs to be archived, or if the following message is logged: “Found unknown file XYZ in the File System Repository; archiving file”.

Upstream JIRA

- NIFI-10023
- NIFI-9993

Knowledge article

For the latest update on this issue see the corresponding Knowledge article: TSB 2022-580: NiFi Processors cannot write to content repository

TSB 2022-589: CVE-2022-33140 Apache NiFi ShellUserGroupProvider Vulnerability

The optional ShellUserGroupProvider in Apache NiFi 1.10.0 to 1.16.2 and Apache NiFi Registry 0.6.0 to 1.16.2 does not neutralize arguments for group resolution commands, allowing injection of operating system commands on Linux and macOS platforms. The ShellUserGroupProvider is not included in the default configuration. Command injection requires ShellUserGroupProvider to be one of the enabled User Group Providers (UGP) in the Authorizers configuration. Command injection also requires an authenticated user with elevated privileges. Apache NiFi requires an authenticated user with authorization to modify access policies in order to execute the command. Apache NiFi Registry requires an authenticated user with authorization to read user groups in order to execute the command. The resolution removes command formatting based on user-provided arguments.

Knowledge article

For the latest update on this issue see the corresponding Knowledge article: TSB 2022-589: CVE-2022-33140 Apache NiFi ShellUserGroupProvider Vulnerability

Known Issues in Streams Messaging

Learn about the known issues in Streams Messaging clusters, the impact or changes to the functionality, and the workaround.

Kafka

Learn about the known issues and limitations in Kafka in this release:

Known Issues

Topics created with the kafka-topics tool are only accessible by the user who created them when the deprecated --zookeeper option is used

By default all created topics are secured. However, when topic creation and deletion is done with the kafka-topics tool using the --zookeeper option, the tool talks directly to Zookeeper. Because security is the responsibility of ZooKeeper authorization and authentication, Kafka cannot prevent users from making ZooKeeper changes. As a result, if the --zookeeper option is used, only the user who created the topic will be able to carry out administrative actions on it. In this scenario Kafka will not have permissions to perform tasks on topics created this way.

Use kafka-topics with the --bootstrap-server option that does not require direct access to Zookeeper.

Certain Kafka command line tools require direct access to Zookeeper

The following command line tools talk directly to ZooKeeper and therefore are not secured via Kafka:

- kafka-reassign-partitions

None

The offsets.topic.replication.factor property must be less than or equal to the number of live brokers

The offsets.topic.replication.factor broker configuration is now enforced upon auto topic creation. Internal auto topic creation will fail with a GROUP_COORDINATOR_NOT_AVAILABLE error until the cluster size meets this replication factor requirement.

None

Requests fail when sending to a nonexistent topic with auto.create.topics.enable set to true

The first few produce requests fail when sending to a nonexistent topic with auto.create.topics.e nable set to true.

Increase the number of retries in the producer configuration setting retries.

Custom Kerberos principal names cannot be used for kerberized ZooKeeper and Kafka instances

When using ZooKeeper authentication and a custom Kerberos principal, Kerberos-enabled Kafka does not start. You must disable ZooKeeper authentication for Kafka or use the default Kerberos principals for ZooKeeper and Kafka.

None

KAFKA-2561: Performance degradation when SSL Is enabled

In some configuration scenarios, significant performance degradation can occur when SSL is enabled. The impact varies depending on your CPU, JVM version, Kafka configuration, and message size. Consumers are typically more affected than producers.

Configure brokers and clients with ssl.secure.random.implementation = SHA1PRNG. It often reduces this degradation drastically, but its effect is CPU and JVM dependent.

OPSAPS-43236: Kafka garbage collection logs are written to the process directory

By default Kafka garbage collection logs are written to the agent process directory. Changing the default path for these log files is currently unsupported.

None

Limitations

Collection of Partition Level Metrics May Cause Cloudera Manager’s Performance to Degrade

If the Kafka service operates with a large number of partitions, collection of partition level metrics may cause Cloudera Manager's performance to degrade.

If you are observing performance degradation and your cluster is operating with a high number of partitions, you can choose to disable the collection of partition level metrics.

⚠ **Important:** If you are using SMM to monitor Kafka or Cruise Control for rebalancing Kafka partitions, be aware that both SMM and Cruise Control rely on partition level metrics. If partition level metric collection is disabled, SMM will not be able to display information about partitions. In addition, Cruise Control will not operate properly.

Complete the following steps to turn off the collection of partition level metrics:

1. Obtain the Kafka service name:

   a. In Cloudera Manager, Select the Kafka service.
   b. Select any available chart, and select Open in Chart Builder from the configuration icon drop-down.
   c. Find $SERVICENAME= near the top of the display.

      The Kafka service name is the value of $SERVICENAME.

2. Turn off the collection of partition level metrics:

   a. Go to HostsHosts Configuration.
   b. Find and configure the Cloudera Manager Agent Monitoring Advanced Configuration Snippet (Safety Valve) configuration property.

      Enter the following to turn off the collection of partition level metrics:

      ```
      [KAFKA_SERVICE_NAME]_feature_send_broker_topic_partition_ent
      ity_update_enabled=false
      ```

      Replace [KAFKA_SERVICE_NAME] with the service name of Kafka obtained in step 1. The service name should always be in lower case.
   c. Click Save Changes.

Schema Registry

There are no known issues in Schema Registry in this release.

Streams Messaging Manager

Learn about the known issues in Streams Messaging Manager in this release.

OPSAPS-59553: SMM's bootstrap server config should be updated based on Kafka's listeners

SMM does not show any metrics for Kafka or Kafka Connect when multiple listeners are set in Kafka.

Workaround: SMM cannot identify multiple listeners and still points to bootstrap server using the default broker port (9093 for SASL_SSL). You would have to override bootstrap server URL (hostname:port as set in the listeners for broker) in the following path:

Cloudera Manager > SMM > Configuration > Streams Messaging Manager Rest Admin Server Advanced Configuration Snippet (Safety Valve) for streams-messaging-manager.yaml > Save Changes > Restart SMM.

OPSAPS-59597: SMM UI logs are not supported by Cloudera Manager

Cloudera Manager does not support the log type used by SMM UI.

Workaround: View the SMM UI logs on the host.

OPSAPS-59828: SMM cannot connect to Schema Registry when TLS is enabled

When TLS is enabled, SMM by default cannot properly connect to Schema Registry.

As a result, when viewing topics in the SMM Data Explorer with the deserializer key or value set to Avro, the following error messages are shown:

- Error deserializing key/value for partition [***PARTITION***] at offset [***OFFSET***]. If needed, please seek past the record to continue consumption.
- Failed to fetch value schema versions for topic : '[***TOPIC**]'.

In addition, the following certificate error will also be present the SMM log:

- javax.net.ssl.SSLHandshakeException: PKIX path building failed:...

Workaround: Additional security properties must be set for SMM.

1. In Cloudera Manager, select the SMM service.
2. Go to Configuration.
3. Find and configure the SMM_JMX_OPTS property.Add the following JVM SSL properties:

   - Djavax.net.ssl.trustStore=[***SMM TRUSTSTORE LOCATION***]
   - Djavax.net.ssl.trustStorePassword=[***PASSWORD***]

Streams Replication Manager

Learn about the known issues and limitations in Streams Replication Manager in this release:

Known Issues

CDPD-22089: SRM does not sync re-created source topics until the offsets have caught up with target topic

Messages written to topics that were deleted and re-created are not replicated until the source topic reaches the same offset as the target topic. For example, if at the time of deletion and re-creation there are a 100 messages on the source and target clusters, new messages will only get replicated once the re-created source topic has 100 messages. This leads to messages being lost.

None

CDPD-14019: SRM may automatically re-create deleted topics

If auto.create.topics.enable is enabled, deleted topics are automatically recreated on source clusters.

Prior to deletion, remove the topic from the topic whitelist with the srm-control tool. This prevents topics from being re-created.

```
srm-control topics --source [SOURCE_CLUSTER] --target [TARGET_CL
USTER] --remove [TOPIC1][TOPIC2]
```

CDPD-13864 and CDPD-15327: Replication stops after the network configuration of a source or target cluster is changed

If the network configuration of a cluster which is taking part in a replication flow is changed, for example, port numbers are changed as a result of enabling or disabling TLS, SRM will not update its internal configuration even if SRM is reconfigured and restarted. From SRM's perspective, it is the cluster identity that has changed. SRM cannot determine whether the new identity corresponds to the same cluster or not, only the owner or administrator of that cluster can know. In this case, SRM tries to use the last known configuration of that cluster which might not be valid, resulting in the halt of replication.

The internal topic storing the configuration of SRM can be deleted. After a restart SRM will re-create and re-populate it with the configuration data loaded from its property file. The topic is hosted on the target cluster of the replication flow. The topic name is: mm2-configs.[*SOURCE_ALIAS*].internal. However, changing a replicated cluster's identity is generally not recommended.

CDPD-60823: Configuring the SRM Client's secure storage is mandatory for unsecured environments

In an unsecured environment the srm-control tool should not need any additional configuration to run. However, due to an issue with the automatic generation of the default configuration, configuring the SRM Client's secure storage is mandatory for the srm-control tool. This is true even if none of the clusters that the tool connects to are secured.

If a secure storage is not configured, the tool will fail with the following NullPointerException:

```
java.lang.NullPointerException
at com.cloudera.dim.mirror.SecureConfigProvider.retrievePassword(
SecureConfigProvider.java:99)
at com.cloudera.dim.mirror.SecureConfigProvider.configure(Secu
reConfigProvider.java:113)
at org.apache.kafka.common.config.AbstractConfig.instantiateConfi
gProviders(AbstractConfig.java:533)
at org.apache.kafka.common.config.AbstractConfig.resolveConfigVa
riables(AbstractConfig.java:477)
at org.apache.kafka.common.config.AbstractConfig.<init>(Abstrac
tConfig.java:107)
at org.apache.kafka.common.config.AbstractConfig.<init>(Abstra
ctConfig.java:142)
```

```
at org.apache.kafka.connect.mirror.MirrorMakerConfig.<init>(Mirro
rMakerConfig.java:88)
at com.cloudera.dim.mirror.MirrorControlCommand$SourceTargetCo
mmand.init(MirrorControlCommand.java:97)
at com.cloudera.dim.mirror.MirrorControlCommand.issueCommand(Mi
rrorControlCommand.java:369)
at com.cloudera.dim.mirror.MirrorControlCommand.main(MirrorCont
rolCommand.java:346)
```

Configure a secure storage password and set it as an environment variable in your CLI session before running the srm-control tool.

1. In Cloudera Manager, select the Streams Replication Manager service.
2. Go to Configuration.
3. Find and configure the SRM Client's Secure Storage Password property.

   Take note of the password that you configure.
4. Click Save changes.
5. Restart the SRM service
6. SSH into one of the SRM hosts in your cluster.
7. Set the secure storage password as an environment variable.

   ```
   export [***SECURE STORAGE ENV VAR***]=”[***SECURE STORAGE PA
   SSWORD***]”
   ```

   Replace *[***SECURE STORAGE ENV VAR***]* with the name of the environment variable you specified in Environment Variable Holding SRM Client's Secure Storage Password. Replace *[***SRM SECURE STORAGE PASSWORD***]* with the password you specified in SRM Client's Secure Storage Password. For example:

   ```
   export SECURESTOREPASS=”mypassword”
   ```

OPSAPS-61001: Saving configuration changes for SRM is not possible

Cloudera Manager incorrectly labels the SRM Client's Secure Storage Password property as mandatory. Moreover, it does not offer this property for configuration when SRM is installed with the Add Service Wizard.

As a result, it is possible to install and start SRM without configuring this property. However, in a case like this, making changes to SRM's configuration is not possible until the SRM Client's Secure Storage Password property is set.

Configure the SRM Client's Secure Storage Password property.

⚠ **Important:** Once the SRM Client's Secure Storage Password property is configured, you must set the password configured with the property as an environment variable in your CLI session before running the srm-control tool. The tool will fail to run if the password is not set as an environment variable. For more information see Configuring srm-control.

OPSAPS-60601: The SRM client's secure storage might become corrupted if the JAAS Secret properties are used

Cloudera Manager generates a secure storage for SRM clients that stores the sensitive data (security related properties) needed to access the clusters that SRM replicates. The sensitive data that the secure storage contains is sourced from the Kafka credentials created by the user in the AdministrationExternal AccountsKafka CredentialsAdd Kafka credentials modal window in Cloudera Manager. If the JAAS Secret properties available in this modal window are used, the generated secure storage can become corrupted. In a case like this, the JAAS configuration is only partially saved to the configuration.

Specify the JAAS configuration using the Streams Replication Manager’s Replication Configs Cloudera Manager property.

This is done by adding the sasl.jaas.config property to Streams Replication Manager’s Replication Configs with an appropriate prefix. For example:

```
[***ALIAS***].sasl.jaas.config=[***JAAS CONFIG***]
```

Replace [***ALIAS***] with a cluster alias specified in Streams Replication Manager Cluster alias. Replace [***JAAS CONFIG***] with a valid JAAS configuration. Repeat this process for all clusters that require a JAAS configuration.

OPSAPS-60601: Replication does not start when the target cluster of the replication is unsecured

When replicating data into an unsecured cluster, the configuration generated for SRM will contain references to defined, but otherwise empty environment variables related to TLS/SSL properties (keystore or truststore locations). The values of these variables cannot be processed by SRM. As a result, replication does not start.

Create and use a placeholder truststore file for the unsecured cluster:

1. Create a placeholder truststore with the keytool utility. For example:

```
keytool -genkeypair -alias placeholder -storepass secret -ke
ypass secret -keystore placeholder.jks -dname "CN=Placeholder,
 OU=Department, O=Company, L=City, ST=State, C=CA"
```

2. Copy the resulting placeholder.jks file to the same location on all SRM driver hosts.
3. Configure SRM to use the keystore for the unsecured cluster.

   This can be done by adding the ssl.truststore.location and ssl.truststore.password properties to the Streams Replication Manager’s Replication Configs Cloudera Manager property with an appropriate prefix. For example:

```
[***ALIAS***].ssl.truststore.location=[***TRUSTSTORE
 LOCATION***]
[***ALIAS***].ssl.truststore.password=[***TRUSTSTORE
 PASSWORD***]
```

   Replace [***ALIAS***] with the unsecured cluster's alias. You can find the alias in Streams Replication Manager Cluster alias. Replace [***TRUSTSTORE LOCATION**] with the location you copied the placeholder.jks file to in Step 2. Replace [***TRUSTSTORE PASSWORD***] with the password you specified when creating the keystore. Repeat this step for all unsecured clusters.

OPSAPS-61814: Using the service dependency method to configure Kerberos enabled co-located clusters is not supported

Using the Streams Replication Manager Co-located Kafka Cluster Alias property to auto-configure the connection to a Kerberos enabled co-located Kafka cluster is not supported. In a case like this, the generated JAAS configuration contains host-specific configuration. This causes SRM to fail to connect to the co-located Kafka cluster on other hosts.

Define your co-located Kafka clusters using Kafka credentials. For more information, see Defining co-located Kafka clusters using Kafka credentials. Alternatively, use the Streams Replication Manager's Replication Configs property to configure the connection to the co-located Kafka clusters.

OPSAPS-60775: Kafka External Accounts configurations are not generated for the SRM Service

Kafka External Account configurations are not generated for SRM Service, making it unable to target clusters defined through External Accounts.

Use the co-located cluster auto-configuration, or the legacy array configuration (Streams Replication Manager’s Replication Configs) to configure the target cluster of SRM Service.

OPSAPS-62546: Kafka External Account SSL keypassword configuration is used incorrectly by SRM

When a Kafka External Account specifies a keystore that uses an SSL key password, SRM uses it as the ssl.keystore.key configuration. Due to using the incorrect ssl.keystore.key configuration, SRM will fail to load the keystore in certain cases.

Workaround: For the keystores used by the Kafka External Accounts, the SSL key password should match the SSL keystore password, and the SSL keystore key password should not be provided. Alternatively, you can use the legacy connection configurations based on the streams.replication.manager.configs to specify the SSL key password.

Limitations

SRM cannot replicate Ranger authorization policies to or from Kafka clusters

Due to a limitation in the Kafka-Ranger plugin, SRM cannot replicate Ranger policies to or from clusters that are configured to use Ranger for authorization. If you are using SRM to replicate data to or from a cluster that uses Ranger, disable authorization policy synchronization in SRM. This can be achieved by clearing the Sync Topic Acls Enabled (sync.topic.acls.enabled) checkbox.

SRM cannot ensure the exactly-once semantics of transactional source topics

SRM data replication uses at-least-once guarantees, and as a result cannot ensure the exactly-once semantics (EOS) of transactional topics in the backup/target cluster.

Note: Even though EOS is not guaranteed, you can still replicate the data of a transactional source, but you must set isolation.level to read_committed for SRM's internal consumers. This can be done by adding *[***SOURCE CLUSTER ALIAS***]->[***TARGET CLUSTER ALIAS***]*.consumer.isolation.level=read_committed to the Streams Replication Manager's Replication Configs SRM service property in Cloudera Manger.

SRM checkpointing is not supported for transactional source topics

SRM does not correctly translate checkpoints (committed consumer group offsets) for transactional topics. Checkpointing assumes that the offset mapping function is always increasing, but with transactional source topics this is violated. Transactional topics have control messages in them, which take up an offset in the log, but they are never returned on the consumer API. This causes the mappings to decrease, causing issues in the checkpointing feature. As a result of this limitation, consumer failover operations for transactional topics is not possible.

Fixed Issues in Cloudera DataFlow for Data Hub 7.2.10

Fixed issues represent selected issues that were previously logged through Cloudera Support, but are addressed in the current release. These issues may have been reported in previous versions within the Known Issues section; meaning they were reported by customers or identified by Cloudera Quality Engineering team.

Review the list of issues that are resolved in Cloudera DataFlow for Data Hub 7.2.10.

Fixed Issues in Flow Management

Review the list of Flow Management issues that are resolved in Cloudera DataFlow for Data Hub 7.2.10.

`AzureBlobIDBrokerCloudCredentialsProvdider` Controller Service improvement

An issue has been fixed with the `AzureBlobIDBrokerCloudCredentialsProvider` Controller Service when using a Medium Duty Data Lake where the IDBroker is Highly Available.

NIFI-8390

Fixed the handling of HBase namespaces in the Atlas reporting task for the HBase processors.

NIFI-8419

NPE when updating parameter context in a secure instance/cluster.

NIFI-8458 & NIFI-8344

CDP Public Cloud 7.2.10 introduces improvements to the `TailFile` processors.

NIFI-8429

DBCPConnectionPool leaks registered drivers

NIFI-8368

Avro decimal logical type fails if scale > precision

NIFI-7912

Fixed Issues in Streams Messaging

Review the list of Streams Messaging issues that are resolved in Cloudera DataFlow for Data Hub 7.2.10.

Kafka

CDPD-24828: AvroConnectTranslator does not handle optional nested records properly

Avro to Connect schema conversion no longer fails when optional nested records are used in the Avro document.

Streams Messaging Manager

CDPD-24173: Restrict the allowed HTTP methods for SMM REST API

The following http methods are allowed in Streams Messaging Manager: GET, POST, PUT, DELETE, HEAD, and OPTIONS. TRACE method is disabled.

CDPD-24698: Configurable ConsumerEmission timeout

New SMM config: cm.metrics.emit.consumer.metrics.timeout.

SMM pushes the Consumer metrics into Cloudera Manager (only when Cloudera Manager is used as a metricsStore). This configuration allows you to configure the timeout of the emission API calls. By default the timeout is 10 seconds.

Schema Registry

CDPD-24415: Schema Registry RAZ NoClassDefFoundError

Due to changes in Ranger RAZ, Schema Registry was unable to upload serdes files to ABFS and S3. We traced the issue to Hadoop - via Ranger - bringing in transitive dependencies which were conflicting with Schema Registry's own classes. As a solution we isolated Hadoop's classes into a separate classpath and load them only when needed. This way future changes in Hadoop and Ranger classpaths will not affect Schema Registry.

CDPD-21913: Rename properties in registry yaml file

Schema Registry uses the Dropwizard framework which allows overriding configuration properties from the command line. Due to implementation specifics, some of the properties could not be overridden. We have fixed this issue now.

In Schema Registry's configuration file, FileStorageConfiguration properties can be directory, fsUrl, kerberosPrincipal and keytabLocation.

In Schema Registry's configuration file, StorageProviderConfiguration properties can be dbtype, queryTimeoutInSecs and properties that can be dataSourceClassName, dataSourceUrl, dataSourceUser, dataSourcePassword and connectionProperties that can be oracleNetSslVersion, oracleNetSslServerDnMatch, trustStore, trustStoreType, keyStore, keyStoreType.

Streams Replication Manager

There are no fixed Streams Replication Manager issues in this release.

Fixed Issues in Cloudera DataFlow for Data Hub 7.2.10.1

Fixed issues represent selected issues that were previously logged through Cloudera Support, but are addressed in the current release. These issues may have been reported in previous versions within the Known Issues section; meaning they were reported by customers or identified by Cloudera Quality Engineering team.

There are no fixed issues in Cloudera DataFlow for Data Hub 7.2.10.1.