Apache NiFi 3

Managing a Data Flow

Date of Publish: 2020-04-28

https://docs.cloudera.com/

Contents

Command and Control of the DataFlow

When a component is added to the NiFi canvas, it is in the Stopped state. In order to cause the component to be triggered, the component must be started. Once started, the component can be stopped at any time. From a Stopped state, the component can be configured, started, or disabled.

Starting a Component

In order to start a component, the following conditions must be met:

- The component's configuration must be valid.
- All defined Relationships for the component must be connected to another component or auto-terminated.
- The component must be stopped.
- The component must be enabled.
- The component must have no active tasks. For more information about active tasks, see the "Anatomy of …" sections under Monitoring of DataFlow.

Components can be started by selecting all of the components to start and then clicking the "Start" button (

) in the Operate Palette or by right-clicking a single component and choosing Start from the context menu.

If starting a Process Group, all components within that Process Group (including child Process Groups) will be started, with the exception of those components that are invalid or disabled.

Once started, the status indicator of a Processor will change to a Play symbol (

).

Stopping a Component

A component can be stopped any time that it is running. A component is stopped by right-clicking on the component and clicking Stop from the context menu, or by selecting the component and clicking the "Stop" button (

) in the Operate Palette.

If a Process Group is stopped, all of the components within the Process Group (including child Process Groups) will be stopped.

Once stopped, the status indicator of a component will change to the Stop symbol (

).

Stopping a component does not interrupt its currently running tasks. Rather, it stops scheduling new tasks to be performed. The number of active tasks is shown in the top-right corner of the Processor (See Anatomy of a Processor for more information).

Enabling/Disabling a Component

When a component is enabled, it is able to be started. Users may choose to disable components when they are part of a dataflow that is still being assembled, for example. Typically, if a component is not intended to be run, the component is disabled, rather than being left in the Stopped state. This helps to distinguish between components that are intentionally not running and those that may have been stopped temporarily (for instance, to change the component's configuration) and inadvertently were never restarted.

When it is desirable to re-enable a component, it can be enabled by selecting the component and clicking the "Enable" button

(

) in the Operate Palette. This is available only when the selected component or components are disabled. Alternatively, a component can be enabled by checking the checkbox next to the "Enabled" option in the Settings tab of the Processor configuration dialog or the configuration dialog for a Port.

Once enabled, the component's status indicator will change to either Invalid (

) or Stopped (

), depending on whether or not the component is valid.

A component is then disabled by selecting the component and clicking the "Disable" button (

) in the Operate Palette, or by clearing the checkbox next to the "Enabled" option in the Settings tab of the Processor configuration dialog or the configuration dialog for a Port.

Only Ports and Processors can be enabled and disabled.

Remote Process Group Transmission

Remote Process Groups provide a mechanism for sending data to or retrieving data from a remote instance of NiFi. When a Remote Process Group (RPG) is added to the canvas, it is added with the Transmission Disabled, as indicated by the icon (

) in the top-left corner. When Transmission is Disabled, it can be enabled by right-clicking on the RPG and clicking the "Enable transmission" menu item. This will cause all ports for which there is a Connection to begin transmitting data. This will cause the status indicator to then change to the Transmission Enabled icon (

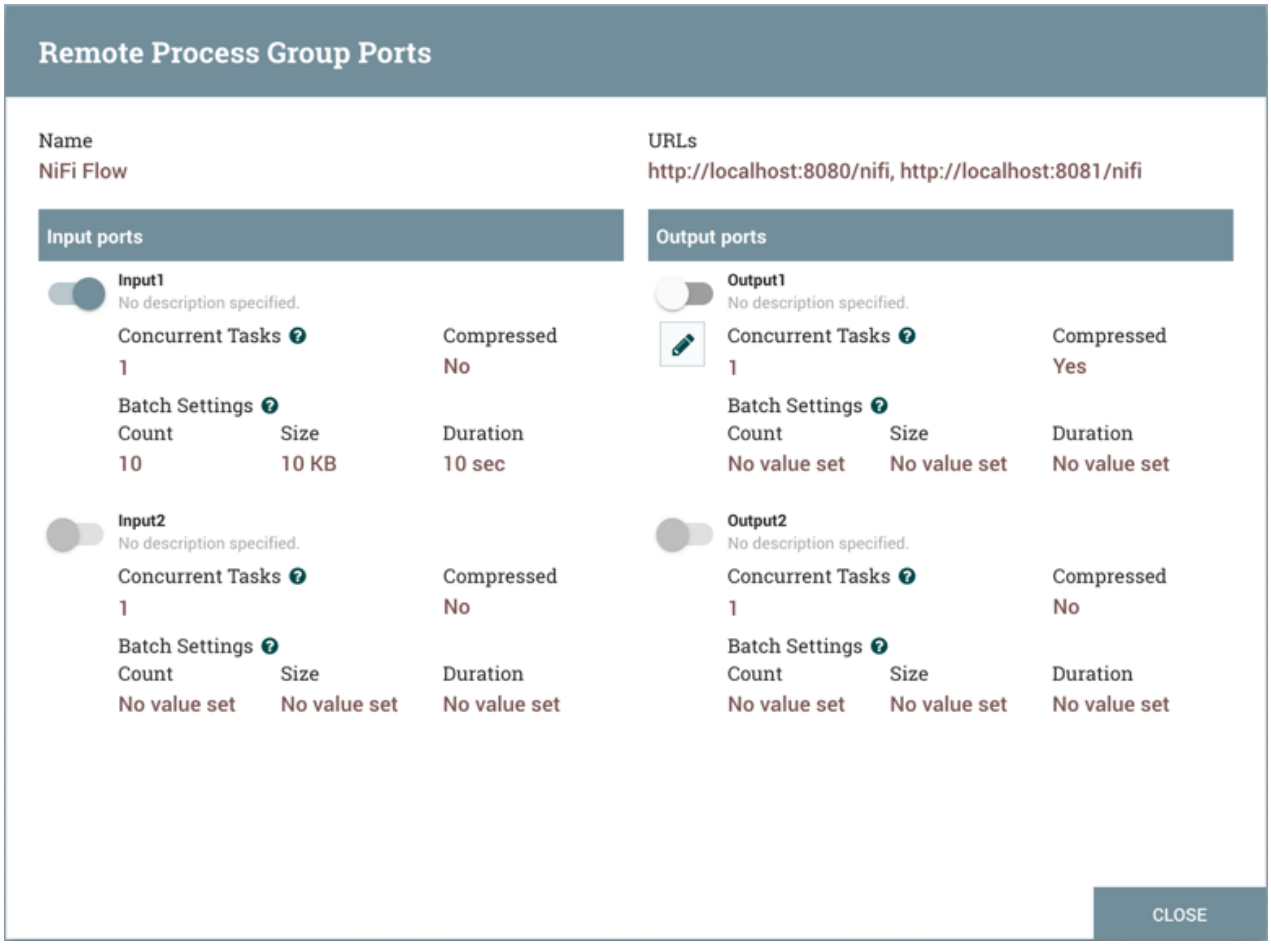
).

If there are problems communicating with the Remote Process Group, a Warning indicator (

) may instead be present in the top-left corner. Hovering over this Warning indicator with the mouse will provide more information about the problem.

Individual Port Transmission

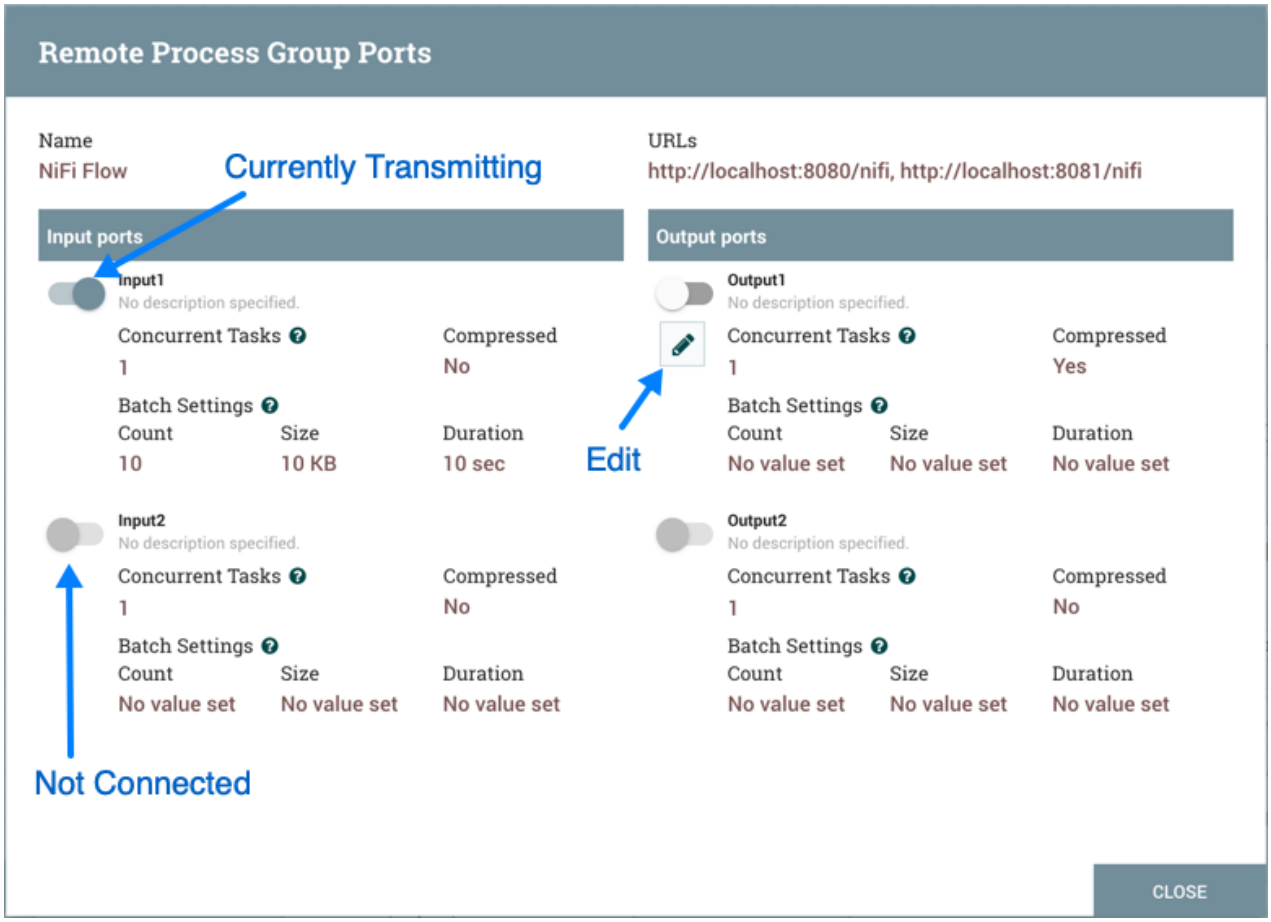
There are times when the DFM may want to either enable or disable transmission for only a specific port within the Remote Process Group. This can be accomplished by right-clicking on the Remote Process Group and choosing the "Manage remote ports" menu item. This provides a configuration dialog from which ports can be configured:



The left-hand side lists all of the Input Ports that the remote instance of NiFi allows data to be sent to. The right-hand side lists all of the Output Ports from which this instance is able to pull data. If the remote instance is using secure communications (the URL of the NiFi instance begins with https://, rather than http://), any ports that the remote instance has not made available to this instance will not be shown.

Note: If a port that is expected to be shown is not shown in this dialog, ensure that the instance has proper permissions and that the Remote Process Group's flow is current. This can be checked by closing the Remote Process Group Ports dialog and looking at the bottom-left corner of the Remote Process Group. The date and time when the flow was last refreshed is displayed. If the flow appears to be outdated, it can be updated by right-clicking on the Remote Process Group and selecting "Refresh remote". (See Anatomy of a Remote Process Group for more information).

Each port is shown with its Name, its Description, configured number of Concurrent Tasks, and whether or not data sent to this port will be Compressed. Additionally, the port's configured Batch Settings (Count, Size and Duration) are displayed. To the left of this information is a toggle switch to turn the port on or off. Ports that have no connections attached to them are grayed out:



The on/off toggle switch provides a mechanism to enable and disable transmission for each port in the Remote Process Group independently. Those ports that are connected but are not currently transmitting can be configured by clicking the pencil icon

()

below the on/off toggle switch. Clicking this icon will allow the DFM to change the number of Concurrent Tasks, whether or not compression should be used when transmitting data to or from this port, and Batch Settings.

For an Input Port, the batch settings control how NiFi sends data to the remote input port in a transaction. NiFi will transfer flow files, as they are queued in incoming relationships, until any of the limits (Count, Size, Duration) is met. If none of the settings are configured, a 500 milliseconds batch duration is used by default.

For an Output Port, the batch settings tells the remote NiFi how NiFi prefers to receive data from the remote output port in a transaction. The remote NiFi will use the specified settings (Count, Size, Duration) to control the transfer of flow files. If none of the settings are configured, a 5 seconds batch duration is used by default.

Encrypted Content Repository

While OS-level access control can offer some security over the flowfile content data written to the disk in a repository, there are scenarios where the data may be sensitive, compliance and regulatory requirements exist, or NiFi is running on hardware not under the direct control of the organization (cloud, etc.). In this case, the content repository allows for all data to be encrypted before being persisted to the disk. For more information on the internal workings of the content repository, see NiFi In-Depth - Content Repository.

What is it?

The EncryptedFileSystemRepository is a new implementation of the content repository which encrypts all content data before it is written to the repository. This allows for storage on systems where OS-level access controls are not sufficient to protect the data while still allowing querying and access to the data through the NiFi UI/API.

How does it work?

The FileSystemRepository was introduced in NiFi 0.2.1 and provided the only persistent content repository implementation. The encrypted version wraps that implementation with functionality to return to the Session (usually StandardProcessSession) a special OutputStream/InputStream which encrypt and decrypt the serialized bytes respectively. This allows all components to continue interacting with the content repository interface in the same way as before and continue operating on content data in a streaming manner, without requiring any changes to handle the data protection.

The fully qualified class org.apache.nifi.content.EncryptedFileSystemRepository is specified as the content repository implementation in nifi.properties as the value of nifi.content.repository.implementation. In addition, new properties must be populated to allow successful initialization.

StaticKeyProvider

The StaticKeyProvider implementation defines keys directly in nifi.properties. Individual keys are provided in hexadecimal encoding. The keys can also be encrypted like any other sensitive property in nifi.properties using the ./encrypt-config.sh tool in the NiFi Toolkit.

The following configuration section would result in a key provider with two available keys, "Key1" (active) and "AnotherKey".

```
nifi.content.repository.encryption.key.provider.implementation=org.apache.nifi.security
nifi.content.repository.encryption.key.id=Key1
nifi.content.repository.encryption.key=0123456789ABCDEFFEDCBA98765432100123456789ABCDEF
nifi.content.repository.encryption.key.id.AnotherKey=0101010101010101010101010101010101
```

FileBasedKeyProvider

The FileBasedKeyProvider implementation reads from an encrypted definition file of the format:

```
key1=NGCpDpxBZNN0DBodz0p1SDbTjC2FG5kp1pCmdUKJlxxtcMSo6GC4fMlTyy1mPeKOxzLut3DRX
+51j6PCO5SznA==
key2=GYxPbMMDbnraXs09eGJudAM5jTvVYp05XtImkAg4JY4rIbmHOiVUUI6OeOf7ZW
+hH42jtPgNW9pSkkQ9HWY/vQ==
key3=SFe11xuz7J89Y/IQ7YbJPOL0/YKZRFL/
VUxJgEHxxlXpd/8ELA7wwN59K1KTr3BURCcFP5YGmwrSKfr4OE4Vlg==
key4=kZprfcTSTH69UuOU3jMkZfrtiVR/eqWmmbdku3bQcUJ/
+UToecNB5lzOVEMBChyEXppyXXC35Wa6GEXFK6PMKw==
key5=c6FzfnKm7UR7xqI2NFpZ+fEKBfSU7+1NvRw
+XWQ9U39MONWqk5gvoyOCdFR1kUgeg46jrN5dGXk13sRqE0GETQ==
```

Each line defines a key ID and then the Base64-encoded cipher text of a 16 byte IV and wrapped AES-128, AES-192, or AES-256 key depending on the JCE policies available. The individual keys are wrapped by AES/GCM encryption using the master key defined by nifi.bootstrap.sensitive.key in conf/bootstrap.conf.

Data Protection vs. Key Protection

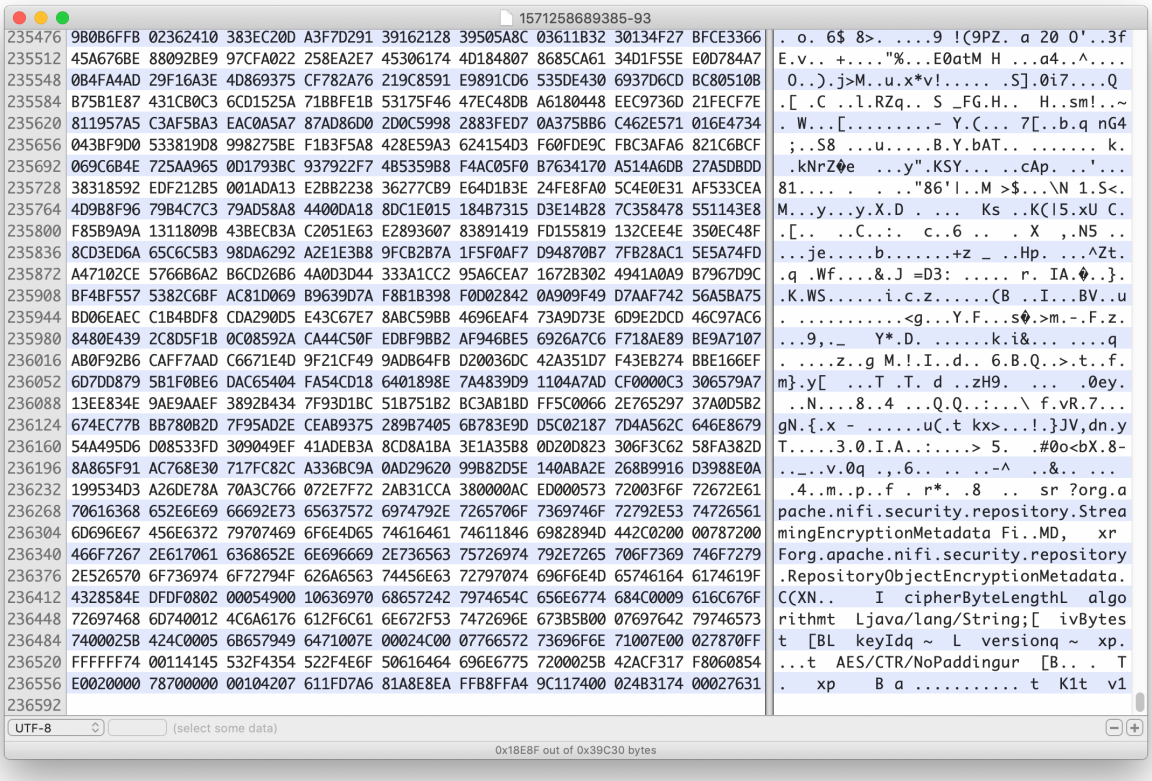
Even though the flowfile content is encrypted using AES/CTR to handle streaming data, if using the Config Encrypt Tool or FileBasedKeyProvider, those keys will be protected using AES/GCM to provide authenticated encryption over the key material.

Key Rotation

Simply update nifi.properties to reference a new key ID in nifi.content.repository.encryption.key.id. Previously-encrypted content claims can still be decrypted as long as that key is still available in the key definition file or nifi.content.repository.encryption.key.id.<OldKeyID> as the key ID is serialized alongside the encrypted content.

Writing and Reading Content Claims

Once the repository is initialized, all content claim write operations are serialized using RepositoryObjectStreamEncryptor (the only currently existing implementation is RepositoryObjectAESCTREncryptor) to an OutputStream. The actual implementation is EncryptedContentRepositoryOutputStream, which encrypts the data written by the component via StandardProcessSession inline and the encryption metadata (keyId, algorithm, version, IV) is serialized and prepended. The complete OutputStream is then written to the repository on disk as normal.



On content claim read, the process is reversed. The encryption metadata (RepositoryObjectEncryptionMetadata) is parsed and used to decrypt the serialized bytes, which are then deserialized into a CipherInputStream object. The delegation to the normal repository file system interaction allows for "random-access" (i.e. immediate seek without decryption of unnecessary content claims).

Within the NiFi UI/API, there is no detectable difference between an encrypted and unencrypted content repository. The Provenance Query operations to view content work as expected with no change to the process.

Potential Issues

- Switching between unencrypted and encrypted repositories

  - If a user has an existing repository (FileSystemRepository) that is not encrypted and switches their configuration to use an encrypted repository, the application writes an error to the log but starts up. However, previous content claims are not accessible through the provenance query interface and new content claims will overwrite the existing claims. The same behavior occurs if a user switches from an encrypted repository to an unencrypted repository. Automatic roll-over is a future effort (https://issues.apache.org/jira/browse/NIFI-6783) but NiFi is not intended for long-term storage of content claims so the impact should be minimal. There are two scenarios for roll-over:
    - Encrypted # unencrypted - if the previous repository implementation was encrypted, these claims should be handled seamlessly as long as the key provider available still has the keys used to encrypt the claims (see *Key Rotation*.)
    - Unencrypted # encrypted - if the previous repository implementation was unencrypted, these claims should be handled seamlessly as the previously written claims simply need to be read with a plaintext InputStream and then be written back with the EncryptedContentRepositoryOutputStream
  - There is also a future effort to provide a standalone tool in NiFi Toolkit to encrypt/decrypt an existing content repository to make the transition easier. The translation process could take a long time depending on the size of the existing repository, and being able to perform this task outside of application startup would be valuable (https://issues.apache.org/jira/browse/NIFI-6783).
- Multiple repositories - No additional effort or testing has been applied to multiple repositories at this time. It is possible/likely issues will occur with repositories on different physical devices. There is no option to provide a heterogenous environment (i.e. one encrypted, one plaintext repository).
- Corruption - when a disk is filled or corrupted, there have been reported issues with the repository becoming corrupted and recovery steps are necessary. This is likely to continue to be an issue with the encrypted repository, although still limited in scope to individual claims (i.e. an entire repository file won't be irrecoverable due to the encryption). Some testing has been performed on scenarios where disk space is exhausted. While the flow can no longer write additional content claims to the repository in that case, the NiFi application continues to function properly, and successfully written content claims are still available via the Provenance Query operations. Stopping NiFi and removing the content repository (or moving it to a larger disk) resolves the issue.

Encrypted FlowFile Repository

While OS-level access control can offer some security over the flowfile attribute and content claim data written to the disk in a repository, there are scenarios where the data may be sensitive, compliance and regulatory requirements exist, or NiFi is running on hardware not under the direct control of the organization (cloud, etc.). In this case, the flowfile repository allows for all data to be encrypted before being persisted to the disk. For more information on the internal workings of the flowfile repository, see NiFi In-Depth - FlowFile Repository.

What is it?

The EncryptedSequentialAccessWriteAheadLog is a new implementation of the flowfile write-ahead log which encrypts all flowfile attribute data before it is written to the repository. This allows for storage on systems where OS-level access controls are not sufficient to protect the data while still allowing querying and access to the data through the NiFi UI/API.

How does it work?

The SequentialAccessWriteAheadLog was introduced in NiFi 1.6.0 and provided a faster flowfile repository implementation. The encrypted version wraps that implementation with functionality to transparently encrypt and decrypt the serialized RepositoryRecord objects during file system interaction. During all writes to disk (swapping, snapshotting, journaling, and checkpointing), the flowfile containers are serialized to bytes based on a schema, and this serialized form is encrypted before writing. This allows the snapshot handler to continue interacting with the

flowfile repository interface in the same way as before and continue operating on flowfile data in a random access manner, without requiring any changes to handle the data protection.

The fully qualified class org.apache.nifi.wali.EncryptedSequentialAccessWriteAheadLog is specified as the flowfile repository write-ahead log implementation in nifi.properties as the value of nifi.flowfile.repository.wal.implementation. In addition, new properties must be populated to allow successful initialization.

StaticKeyProvider

The StaticKeyProvider implementation defines keys directly in nifi.properties. Individual keys are provided in hexadecimal encoding. The keys can also be encrypted like any other sensitive property in nifi.properties using the ./encrypt-config.sh tool in the NiFi Toolkit.

The following configuration section would result in a key provider with two available keys, "Key1" (active) and "AnotherKey".

```
nifi.flowfile.repository.encryption.key.provider.implementation=org.apache.nifi.security
nifi.flowfile.repository.encryption.key.id=Key1
nifi.flowfile.repository.encryption.key=0123456789ABCDEFFEDCBA98765432100123456789ABCDEF
nifi.flowfile.repository.encryption.key.id.AnotherKey=01010101010101010101010101010101
```

FileBasedKeyProvider

The FileBasedKeyProvider implementation reads from an encrypted definition file of the format:

```
key1=NGCpDpxBZNN0DBodz0p1SDbTjC2FG5kp1pCmdUKJlxxtcMSo6GC4fMlTyy1mPeKOxzLut3DRX
+51j6PCO5SznA==
key2=GYxPbMMDbnraXs09eGJudAM5jTvVYp05XtImkAg4JY4rIbmHOiVUUI6OeOf7ZW
+hH42jtPgNW9pSkkQ9HWY/vQ==
key3=SFe11xuz7J89Y/IQ7YbJPOL0/YKZRFL/
VUxJgEHxxlXpd/8ELA7wwN59K1KTr3BURCcFP5YGmwrSKfr4OE4Vlg==
key4=kZprfcTSTH69UuOU3jMkZfrtiVR/eqWmmbdku3bQcUJ/
+UToecNB5lzOVEMBChyEXppyXXC35Wa6GEXFK6PMKw==
key5=c6FzfnKm7UR7xqI2NFpZ+fEKBfSU7+1NvRw
+XWQ9U39MONWqk5gvoyOCdFR1kUgeg46jrN5dGXk13sRqE0GETQ==
```

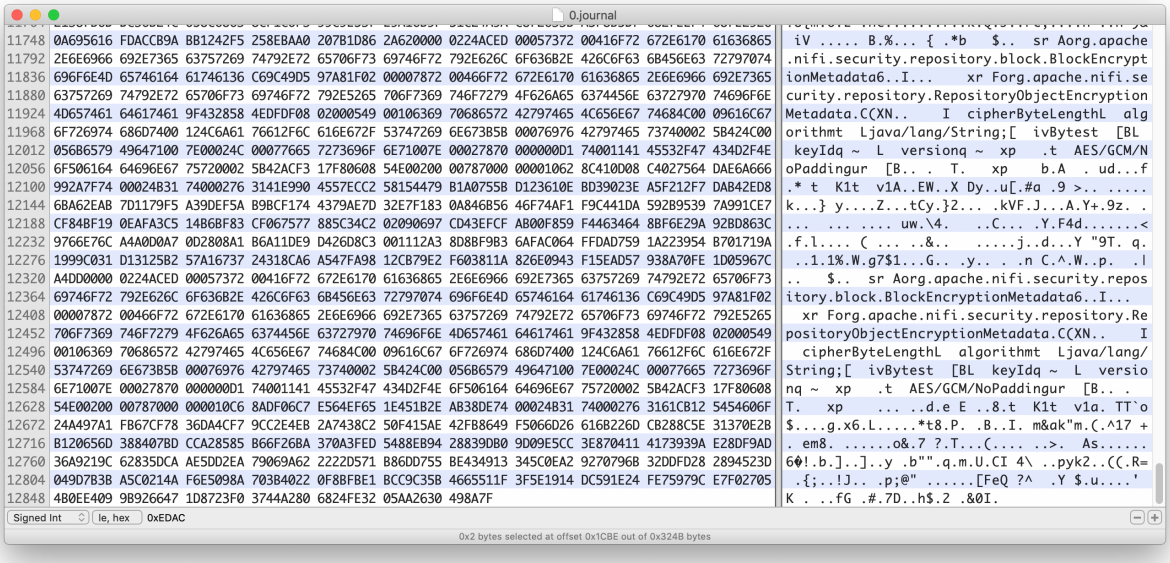
Each line defines a key ID and then the Base64-encoded cipher text of a 16 byte IV and wrapped AES-128, AES-192, or AES-256 key depending on the JCE policies available. The individual keys are wrapped by AES/GCM encryption using the master key defined by nifi.bootstrap.sensitive.key in conf/bootstrap.conf.

Key Rotation

Simply update nifi.properties to reference a new key ID in nifi.flowfile.repository.encryption.key.id. Previously-encrypted flowfile records can still be decrypted as long as that key is still available in the key definition file or nifi.flowfile.repository.encryption.key.id.<OldKeyID> as the key ID is serialized alongside the encrypted record.

Writing and Reading FlowFiles

Once the repository is initialized, all flowfile record write operations are serialized using RepositoryObjectBlockEncryptor (the only currently existing implementation is RepositoryObjectAESGCMEncryptor) to the provided DataOutputStream. The original stream is swapped with a temporary wrapped stream, which encrypts the data written by the wrapped serializer/deserializer via EncryptedSchemaRepositoryRecordSerde inline and the encryption metadata (keyId, algorithm, version, IV, cipherByteLength) is serialized and prepended. The complete length and encrypted bytes are then written to the original DataOutputStream on disk as normal.



On flowfile record read, the process is reversed. The encryption metadata (RepositoryObjectEncryptionMetadata) is parsed and used to decrypt the serialized bytes, which are then deserialized into a DataInputStream object.

During swaps and recoveries, the flowfile records are deserialized and reserialized, so if the active key has been changed, the flowfile records will be re-encrypted with the new active key.

Within the NiFi UI/API, there is no detectable difference between an encrypted and unencrypted flowfile repository. All framework interactions with flowfiles work as expected with no change to the process.

Potential Issues

- Switching between unencrypted and encrypted repositories
  - If a user has an existing write-ahead repository (WriteAheadFlowFileRepository) that is not encrypted (uses the SequentialAccessWriteAheadLog) and switches their configuration to use an encrypted repository, the application handles this and all flowfile records will be recovered on startup. Future writes (including re-serialization of these same flowfiles) will be encrypted. If a user switches from an encrypted repository to an unencrypted repository, the flowfiles cannot be recovered, and it is recommended to delete the existing flowfile repository before switching in this direction. Automatic roll-over is a future effort (https://issues.apache.org/jira/browse/NIFI-6994) but NiFi is not intended for long-term storage of flowfile records so the impact should be minimal. There are two scenarios for roll-over:
    - Encrypted # unencrypted - if the previous repository implementation was encrypted, these records should be handled seamlessly as long as the key provider available still has the keys used to encrypt the claims (see *Key Rotation*.)
    - Unencrypted # encrypted - currently handled seamlesssly for SequentialAccessWriteAheadLog but there are other initial implementations which could be handled
  - There is also a future effort to provide a standalone tool in NiFi Toolkit to encrypt/decrypt an existing flowfile repository to make the transition easier. The translation process could take a long time depending on the size of the existing repository, and being able to perform this task outside of application startup would be valuable (https://issues.apache.org/jira/browse/NIFI-6994).
- Multiple repositories - No additional effort or testing has been applied to multiple repositories at this time. Current implementations of the flowfile repository allow only for one repository, though it can reside across multiple volumes and partitions. It is possible/likely issues will occur with repositories on different physical devices. There is no option to provide a heterogenous environment (i.e. one encrypted, one plaintext partition/directory).

- Corruption - when a disk is filled or corrupted, there have been reported issues with the repository becoming corrupted and recovery steps are necessary. This is likely to continue to be an issue with the encrypted repository, although still limited in scope to individual records (i.e. an entire repository file won't be irrecoverable due to the encryption). It is important for the continued operation of NiFi to ensure that the disk storing the flowfile repository does not run out of available space.

Experimental Warning

While all Apache licensed code is provided "on an "AS IS" BASIS, WITHOUT WARRANTIES OR CONDITIONS OF ANY KIND, either express or implied" (see https://www.apache.org/licenses/LICENSE-2.0), some features of Apache NiFi may be marked experimental. Experimental features may:

- have undergone less extensive testing than is normal for standard NiFi features
- interact with unstable external dependencies
- be subject to change (any exposed APIs should not be considered covered under the minor release backward compatibility guidelines of https://semver.org)
- potentially cause data loss
- not be directly supported by the community in the event issues arise

Every attempt is made to provide more detailed and specific information around the nature of the experimental warning on a per-feature basis. Questions around specific experimental features should be directed to the dev@nifi.apache.org.

Other Management Features

In addition to the Summary Page, Data Provenance Page, Template Management Page, and Bulletin Board Page, there are other tools in the Global Menu (see NiFi User Interface) that are useful to the DFM. Select Flow Configuration History to view all the changes that have been made to the dataflow. The history can aid in troubleshooting, such as if a recent change to the dataflow has caused a problem and needs to be fixed. The DFM can see what changes have been made and adjust the flow as needed to fix the problem. While NiFi does not have an "undo" feature, the DFM can make new changes to the dataflow that will fix the problem.

Two other tools in the Global Menu are Controller Settings and Users. The Controller Settings page provides the ability to change the name of the NiFi instance, add comments describing the NiFi instance, and set the maximum number of threads that are available to the application. It also provides tabs where DFMs may add and configure Controller Services and Reporting Tasks. The Users page is used to manage user access, which is described in the System Administrator’s Guide.