**Apache NiFi :**

Apache NiFi is an open source software for automating and managing the flow of data between systems. It is a powerful and reliable system to process and distribute data. It provides a web-based User Interface for creating, monitoring, & controlling data flows. It has a highly configurable and modifiable data flow process that can modify data at runtime. It is easily extensible through the development of custom components.

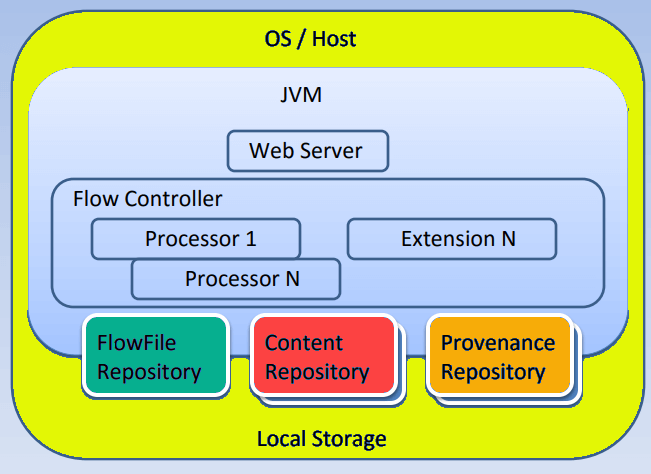
## Apache NIfi Uses :

* Allows you to do data ingestion to pull data into NiFi, from numerous data sources and create flow files
* It offers real-time control which helps you to manage the movement of data between any source & destination
* Visualize DataFlow at the enterprise level
* Provide common tooling and extensions
* Allows you to take advantage of existing libraries and Java ecosystem functionality
* Helps organizations to integrate Nifi with their existing infrastructure
* NiFi is designed to scale-out in clusters which offer guaranteed delivery of data
* Visualize and Monitor performance, behavior in a flow bulletin which offers insight and inline documentation
* Helps you to start and stop components separately or at the group level
* It helps you to listen, fetch, split, aggregate, route, transform and drag & drop Dataflow

## History of Apache NiFi

* Developed at NSA for over eight years
* 2014- It was donated to the Apache Software Foundation
* 2015- NiFi became an official part of the Apache Project Suite
* Since then every 6-8 weeks, Apache NiFi releases a new update

## NiFi Architecture



**Key components of NiFi architecture**

|  |  |
| --- | --- |
| **Nifi Component** | **Description** |
| FlowFile | FlowFile is original data with meta-information attached to it. It allows you to process not only CSV or other record-based data, but also pictures, videos, audio, or any other binary data. |
| Flowfile processor | Performs the work which acts as a building block of data flow in NiFi. |
| Flow controller | Keeps a record of how processes are connected. It manages the threads and allocations thereof which all processes use. |
| Web Server | Web server hosts NiFi's HTTP-based commands and API. |
| Extension | There are many types of NiFi extensions which operate and execute within the JVM. |
| Connection | Acts as a linkage between processors that contain a queue and relationship(s) which affects where data is routed. |
| Back Pressure | Stop the system of becoming overrun by controlling the quantity or data size of flow files that can be stored in the queue. |
| Process Group | A process group is a set of processes and their connections, which receives and send data with the help of ports. |
| Flowfile Repository | In the FlowFile Repository, NiFi keeps track of the state of what details it has about a given FlowFile which is active in the flow. |
| Content Repository | The Content Repository is an area where the actual content bytes of a given FlowFile exist. |
| Provenance Repository | The Provenance Repository is an area where all provenance event data is gathered. |

## Apache NiFi Features

 NiFi supports buffering of all queued data and offers an ability of back pressure as those queues may reach specified limits

 NiFi allows the setting of one or more prioritization schemes

 Provides connection processors for many data sources

 Support any device which runs Java

 Ideal for limited connectivity places

 Support for troubleshooting and flow optimization

 Offers role-based authentication/authorization

 Allows download, recovery, and replay of individual files

 Build your processors, controller services, and more

 Provide content encryption, communication over secure protocols

 Enables rapid development and effective testing

 Allows for the development of simple single-function components that can be reused and combined to make more complex flows

 Allows classloader isolation for easier management of dependencies.

## Nifi Use Cases

|  |  |
| --- | --- |
| **Industry** | **Usage** |
| Insurance | * Risk & underwriting analysis * Claims Analytics * Usage-based Insurance * New product development |
| HealthCare | * Single view of Patient * Real-time vital sign monitoring * EMR optimization * Supply Chain Optimization |
| Telecommunication | * Single view of the customer * CDR analysis * Dynamic Bandwidth allocation |
| Manufacturing | * Preventative Maintenance * Supply Chain Optimization * Quality Control |
| Oil & Gas- Industry | * Real-time monitoring * Single view of the Operation * Predictive Maintenance * Archive & Analytics * Unstructured data classification |
| Financial Services | * Anti-money laundering * Fraud- Detection * Risk- data management |

## Disadvantage of Nifi

* Need precise security and compliance controls
* You need to know the underlying system very well while working with Apache NiFi
* Must maintain chain of custody for data
* Transport / Messaging may not prove enough
* Data access needs exceed available resources to transport
* Not all data is created equally
* SSL and topic level authorization may not be sufficient

## (1)ListHDFS

Retrieves a listing of files from HDFS. For each file that is listed in HDFS, creates a FlowFile that represents the HDFS file so that it can be fetched in conjunction with ListHDFS. This Processor is designed to run on Primary Node only in a cluster. If the primary node changes, the new Primary Node will pick up where the previous node left off without duplicating all of the data. Unlike GetHDFS, this Processor does not delete any data from HDFS.

**Tags:**

hadoop, HDFS, get, list, ingest, source, filesystem

**Properties:**

|  |  |  |  |
| --- | --- | --- | --- |
| **Name** | **Default Value** | **Allowable Values** | **Description** |
| Hadoop Configuration Resources |  |  | A file or comma separated list of files which contains the Hadoop file system configuration. Without this, Hadoop will search the classpath for a 'core-site.xml' and 'hdfs-site.xml' file or will revert to a default configuration. |
| Kerberos Principal |  |  | Kerberos principal to authenticate as. Requires nifi.kerberos.krb5.file to be set in your nifi.properties |
| Kerberos Keytab |  |  | Kerberos keytab associated with the principal. Requires nifi.kerberos.krb5.file to be set in your nifi.properties |
| Kerberos Relogin Period | 4 hours |  | Period of time which should pass before attempting a kerberos relogin |
| Distributed Cache Service |  | **Controller Service API:** DistributedMapCacheClient **Implementation:** DistributedMapCacheClientService | Specifies the Controller Service that should be used to maintain state about what has been pulled from HDFS so that if a new node begins pulling data, it won't duplicate all of the work that has been done. |
| **Directory** |  |  | The HDFS directory from which files should be read |
| **Recurse Subdirectories** | true | \* true</br> \* false | Indicates whether to list files from subdirectories of the HDFS directory |

**Relationships:**

|  |  |
| --- | --- |
| **Name** | **Description** |
| success | All FlowFiles are transferred to this relationship |

**Reads Attributes:**

None specified.

**Writes Attributes:**

|  |  |
| --- | --- |
| **Name** | **Description** |
| filename | The name of the file that was read from HDFS. |
| path | The path is set to the absolute path of the file's directory on HDFS. For example, if the Directory property is set to /tmp, then files picked up from /tmp will have the path attribute set to "./". If the Recurse Subdirectories property is set to true and a file is picked up from /tmp/abc/1/2/3, then the path attribute will be set to "/tmp/abc/1/2/3". |
| hdfs.owner | The user that owns the file in HDFS |
| hdfs.group | The group that owns the file in HDFS |
| hdfs.lastModified | The timestamp of when the file in HDFS was last modified, as milliseconds since midnight Jan 1, 1970 UTC |
| hdfs.length | The number of bytes in the file in HDFS |
| hdfs.replication | The number of HDFS replicas for hte file |
| hdfs.permissions | The permissions for the file in HDFS. This is formatted as 3 characters for the owner, 3 for the group, and 3 for other users. For example rw-rw-r-- |

# (2)FetchHDFS

Retrieves a file from HDFS. The content of the incoming FlowFile is replaced by the content of the file in HDFS. The file in HDFS is left intact without any changes being made to it.

### Tags:

hadoop, hdfs, get, ingest, fetch, source, restricted

### Properties:

|  |  |  |  |
| --- | --- | --- | --- |
| **Name** | **Default Value** | **Allowable Values** | **Description** |
| Hadoop Configuration Resources |  |  | A file or comma separated list of files which contains the Hadoop file system configuration. Without this, Hadoop will search the classpath for a 'core-site.xml' and 'hdfs-site.xml' file or will revert to a default configuration. To use swebhdfs, see 'Additional Details' section of PutHDFS's documentation. **Supports Expression Language: true** |
| Kerberos Principal |  |  | Kerberos principal to authenticate as. Requires nifi.kerberos.krb5.file to be set in your nifi.properties **Supports Expression Language: true** |
| Kerberos Keytab |  |  | Kerberos keytab associated with the principal. Requires nifi.kerberos.krb5.file to be set in your nifi.properties **Supports Expression Language: true** |
| Kerberos Relogin Period | 4 hours |  | Period of time which should pass before attempting a kerberos relogin. This property has been deprecated, and has no effect on processing. Relogins now occur automatically. **Supports Expression Language: true** |
| Additional Classpath Resources |  |  | A comma-separated list of paths to files and/or directories that will be added to the classpath. When specifying a directory, all files with in the directory will be added to the classpath, but further sub-directories will not be included. |
| **HDFS Filename** | ${path}/${filename} |  | The name of the HDFS file to retrieve **Supports Expression Language: true** |
| **Compression codec** | NONE | * NONE No compression * DEFAULT Default ZLIB compression * BZIP BZIP compression * GZIP GZIP compression * LZ4 LZ4 compression * LZO LZO compression - it assumes LD_LIBRARY_PATH has been set and jar is available * SNAPPY Snappy compression * AUTOMATIC Will attempt to automatically detect the compression codec. | No Description Provided. |

### Relationships:

|  |  |
| --- | --- |
| **Name** | **Description** |
| success | FlowFiles will be routed to this relationship once they have been updated with the content of the HDFS file |
| comms.failure | FlowFiles will be routed to this relationship if the content of the HDFS file cannot be retrieve due to a communications failure. This generally indicates that the Fetch should be tried again. |
| failure | FlowFiles will be routed to this relationship if the content of the HDFS file cannot be retrieved and trying again will likely not be helpful. This would occur, for instance, if the file is not found or if there is a permissions issue |

### Reads Attributes:

None specified.

### Writes Attributes:

|  |  |
| --- | --- |
| **Name** | **Description** |
| hdfs.failure.reason | When a FlowFile is routed to 'failure', this attribute is added indicating why the file could not be fetched from HDFS |

# (3)PutHDFS

Write FlowFile data to Hadoop Distributed File System (HDFS)

### Tags:

hadoop, HDFS, put, copy, filesystem, restricted

### Properties:

|  |  |  |  |
| --- | --- | --- | --- |
| **Name** | **Default Value** | **Allowable Values** | **Description** |
| Hadoop Configuration Resources |  |  | A file or comma separated list of files which contains the Hadoop file system configuration. Without this, Hadoop will search the classpath for a 'core-site.xml' and 'hdfs-site.xml' file or will revert to a default configuration. To use swebhdfs, see 'Additional Details' section of PutHDFS's documentation. **Supports Expression Language: true** |
| Kerberos Principal |  |  | Kerberos principal to authenticate as. Requires nifi.kerberos.krb5.file to be set in your nifi.properties **Supports Expression Language: true** |
| Kerberos Keytab |  |  | Kerberos keytab associated with the principal. Requires nifi.kerberos.krb5.file to be set in your nifi.properties **Supports Expression Language: true** |
| Kerberos Relogin Period | 4 hours |  | Period of time which should pass before attempting a kerberos relogin. This property has been deprecated, and has no effect on processing. Relogins now occur automatically. **Supports Expression Language: true** |
| Additional Classpath Resources |  |  | A comma-separated list of paths to files and/or directories that will be added to the classpath. When specifying a directory, all files with in the directory will be added to the classpath, but further sub-directories will not be included. |
| **Directory** |  |  | The parent HDFS directory to which files should be written. The directory will be created if it doesn't exist. **Supports Expression Language: true** |
| **Conflict Resolution Strategy** | fail | * replace Replaces the existing file if any. * ignore Ignores the flow file and routes it to success. * fail Penalizes the flow file and routes it to failure. * append Appends to the existing file if any, creates a new file otherwise. | Indicates what should happen when a file with the same name already exists in the output directory |
| Block Size |  |  | Size of each block as written to HDFS. This overrides the Hadoop Configuration |
| IO Buffer Size |  |  | Amount of memory to use to buffer file contents during IO. This overrides the Hadoop Configuration |
| Replication |  |  | Number of times that HDFS will replicate each file. This overrides the Hadoop Configuration |
| Permissions umask |  |  | A umask represented as an octal number which determines the permissions of files written to HDFS. This overrides the Hadoop Configuration dfs.umaskmode |
| Remote Owner |  |  | Changes the owner of the HDFS file to this value after it is written. This only works if NiFi is running as a user that has HDFS super user privilege to change owner **Supports Expression Language: true** |
| Remote Group |  |  | Changes the group of the HDFS file to this value after it is written. This only works if NiFi is running as a user that has HDFS super user privilege to change group **Supports Expression Language: true** |
| **Compression codec** | NONE | * NONE No compression * DEFAULT Default ZLIB compression * BZIP BZIP compression * GZIP GZIP compression * LZ4 LZ4 compression * LZO LZO compression - it assumes LD_LIBRARY_PATH has been set and jar is available * SNAPPY Snappy compression * AUTOMATIC Will attempt to automatically detect the compression codec. | No Description Provided. |

**Relationships:**

|  |  |
| --- | --- |
| **Name** | **Description** |
| success | Files that have been successfully written to HDFS are transferred to this relationship |
| failure | Files that could not be written to HDFS for some reason are transferred to this relationship |

### Reads Attributes:

|  |  |
| --- | --- |
| **Name** | **Description** |
| filename | The name of the file written to HDFS comes from the value of this attribute. |

### Writes Attributes:

|  |  |
| --- | --- |
| **Name** | **Description** |
| filename | The name of the file written to HDFS is stored in this attribute. |
| absolute.hdfs.path | The absolute path to the file on HDFS is stored in this attribute. |

# (4)PublishKafka

Sends the contents of a FlowFile as a message to Apache Kafka using the Kafka 0.9.x Producer. The messages to send may be individual FlowFiles or may be delimited, using a user-specified delimiter, such as a new-line. Please note there are cases where the publisher can get into an indefinite stuck state. We are closely monitoring how this evolves in the Kafka community and will take advantage of those fixes as soon as we can. In the mean time it is possible to enter states where the only resolution will be to restart the JVM NiFi runs on. The complementary NiFi processor for fetching messages is ConsumeKafka.

### Tags:

Apache, Kafka, Put, Send, Message, PubSub, 0.9.x

### Properties:

|  |  |  |  |
| --- | --- | --- | --- |
| **Name** | **Default Value** | **Allowable Values** | **Description** |
| **Kafka Brokers** | localhost:9092 |  | A comma-separated list of known Kafka Brokers in the format <host>:<port> **Supports Expression Language: true** |
| **Security Protocol** | PLAINTEXT | * PLAINTEXT PLAINTEXT * SSL SSL * SASL\_PLAINTEXT SASL_PLAINTEXT * SASL\_SSL SASL_SSL | Protocol used to communicate with brokers. Corresponds to Kafka's 'security.protocol' property. |
| Kerberos Service Name |  |  | The Kerberos principal name that Kafka runs as. This can be defined either in Kafka's JAAS config or in Kafka's config. Corresponds to Kafka's 'security.protocol' property.It is ignored unless one of the SASL options of the <Security Protocol> are selected. |
| SSL Context Service |  | **Controller Service API:** SSLContextService **Implementations:**[StandardSSLContextService](https://nifi.apache.org/docs/nifi-docs/components/org.apache.nifi/nifi-ssl-context-service-nar/1.5.0/org.apache.nifi.ssl.StandardSSLContextService/index.html) [StandardRestrictedSSLContextService](https://nifi.apache.org/docs/nifi-docs/components/org.apache.nifi/nifi-ssl-context-service-nar/1.5.0/org.apache.nifi.ssl.StandardRestrictedSSLContextService/index.html) | Specifies the SSL Context Service to use for communicating with Kafka. |
| **Topic Name** |  |  | The name of the Kafka Topic to publish to. **Supports Expression Language: true** |
| **Delivery Guarantee** | 0 | * Best Effort FlowFile will be routed to success after successfully writing the content to a Kafka node, without waiting for a response. This provides the best performance but may result in data loss. * Guarantee Single Node Delivery FlowFile will be routed to success if the message is received by a single Kafka node, whether or not it is replicated. This is faster than <Guarantee Replicated Delivery> but can result in data loss if a Kafka node crashes * Guarantee Replicated Delivery FlowFile will be routed to failure unless the message is replicated to the appropriate number of Kafka Nodes according to the Topic configuration | Specifies the requirement for guaranteeing that a message is sent to Kafka. Corresponds to Kafka's 'acks' property. |
| Kafka Key |  |  | The Key to use for the Message. If not specified, the flow file attribute 'kafka.key' is used as the message key, if it is present.Beware that setting Kafka key and demarcating at the same time may potentially lead to many Kafka messages with the same key.Normally this is not a problem as Kafka does not enforce or assume message and key uniqueness. Still, setting the demarcator and Kafka key at the same time poses a risk of data loss on Kafka. During a topic compaction on Kafka, messages will be deduplicated based on this key. **Supports Expression Language: true** |
| **Key Attribute Encoding** | utf-8 | * UTF-8 Encoded The key is interpreted as a UTF-8 Encoded string. * Hex Encoded The key is interpreted as arbitrary binary data that is encoded using hexadecimal characters with uppercase letters. | FlowFiles that are emitted have an attribute named 'kafka.key'. This property dictates how the value of the attribute should be encoded. |
| Message Demarcator |  |  | Specifies the string (interpreted as UTF-8) to use for demarcating multiple messages within a single FlowFile. If not specified, the entire content of the FlowFile will be used as a single message. If specified, the contents of the FlowFile will be split on this delimiter and each section sent as a separate Kafka message. To enter special character such as 'new line' use CTRL+Enter or Shift+Enter, depending on your OS. **Supports Expression Language: true** |
| **Max Request Size** | 1 MB |  | The maximum size of a request in bytes. Corresponds to Kafka's 'max.request.size' property and defaults to 1 MB (1048576). |
| **Acknowledgment Wait Time** | 5 secs |  | After sending a message to Kafka, this indicates the amount of time that we are willing to wait for a response from Kafka. If Kafka does not acknowledge the message within this time period, the FlowFile will be routed to 'failure'. |
| **Max Metadata Wait Time** | 5 sec |  | The amount of time publisher will wait to obtain metadata or wait for the buffer to flush during the 'send' call before failing the entire 'send' call. Corresponds to Kafka's 'max.block.ms' property **Supports Expression Language: true** |
| Partitioner class | org.apache.kafka.clients.producer.internals.DefaultPartitioner | * RoundRobinPartitioner Messages will be assigned partitions in a round-robin fashion, sending the first message to Partition 1, the next Partition to Partition 2, and so on, wrapping as necessary. * DefaultPartitioner Messages will be assigned to random partitions. | Specifies which class to use to compute a partition id for a message. Corresponds to Kafka's 'partitioner.class' property. |
| **Compression Type** | none | * none * gzip * snappy * lz4 | This parameter allows you to specify the compression codec for all data generated by this producer. |

### Dynamic Properties:

|  |  |  |
| --- | --- | --- |
| **Name** | **Value** | **Description** |
| The name of a Kafka configuration property. | The value of a given Kafka configuration property. | These properties will be added on the Kafka configuration after loading any provided configuration properties. In the event a dynamic property represents a property that was already set, its value will be ignored and WARN message logged. For the list of available Kafka properties please refer to: http://kafka.apache.org/documentation.html#configuration. |

### Relationships:

|  |  |
| --- | --- |
| **Name** | **Description** |
| success | FlowFiles for which all content was sent to Kafka. |
| failure | Any FlowFile that cannot be sent to Kafka will be routed to this Relationship |

### Reads Attributes:

None specified.

### Writes Attributes:

|  |  |
| --- | --- |
| **Name** | **Description** |
| msg.count | The number of messages that were sent to Kafka for this FlowFile. This attribute is added only to FlowFiles that are routed to success. If the <Message Demarcator> Property is not set, this will always be 1, but if the Property is set, it may be greater than 1. |

# (5)UpdateAttribute

Updates the Attributes for a FlowFile by using the Attribute Expression Language and/or deletes the attributes based on a regular expression

### Tags:

attributes, modification, update, delete, Attribute Expression Language, state

### Properties:

|  |  |  |  |
| --- | --- | --- | --- |
| **Name** | **Default Value** | **Allowable Values** | **Description** |
| Delete Attributes Expression |  |  | Regular expression for attributes to be deleted from FlowFiles. Existing attributes that match will be deleted regardless of whether they are updated by this processor. **Supports Expression Language: true (will be evaluated using flow file attributes and variable registry)** |
| **Store State** | Do not store state | * Do not store state * Store state locally | Select whether or not state will be stored. Selecting 'Stateless' will offer the default functionality of purely updating the attributes on a FlowFile in a stateless manner. Selecting a stateful option will not only store the attributes on the FlowFile but also in the Processors state. See the 'Stateful Usage' topic of the 'Additional Details' section of this processor's documentation for more information |
| Stateful Variables Initial Value |  |  | If using state to set/reference variables then this value is used to set the initial value of the stateful variable. This will only be used in the @OnScheduled method when state does not contain a value for the variable. This is required if running statefully but can be empty if needed. |
| **Cache Value Lookup Cache Size** | 100 |  | Specifies how many canonical lookup values should be stored in the cache |

### Dynamic Properties:

|  |  |  |
| --- | --- | --- |
| **Name** | **Value** | **Description** |
| A FlowFile attribute to update | The value to set it to | Updates a FlowFile attribute specified by the Dynamic Property's key with the value specified by the Dynamic Property's value **Supports Expression Language: true (will be evaluated using flow file attributes and variable registry)** |

### Relationships:

|  |  |
| --- | --- |
| **Name** | **Description** |
| success | All successful FlowFiles are routed to this relationship |

### Reads Attributes:

None specified.

### Writes Attributes:

|  |  |
| --- | --- |
| **Name** | **Description** |
| See additional details | This processor may write or remove zero or more attributes as described in additional details |

### State management:

|  |  |
| --- | --- |
| **Scope** | **Description** |
| LOCAL | Gives the option to store values not only on the FlowFile but as stateful variables to be referenced in a recursive manner. |

Refered Link - <http://nifi.apache.org/>