IDOL Ingest

Software Version 12.13

Getting Started with IDOL Ingest on AWS Marketplace

Document Release Date: January 2023
Software Release Date: January 2023

Legal notices

The only warranties for products and services of Micro Focus and its affiliates and licensors (“Micro Focus”) are as may be set forth in the express warranty statements accompanying such products and services. Nothing herein should be construed as constituting an additional warranty. Micro Focus shall not be liable for technical or editorial errors or omissions contained herein. The information contained herein is subject to change without notice.

Documentation updates

The title page of this document contains the following identifying information:

- Software Version number, which indicates the software version.
- Document Release Date, which changes each time the document is updated.
- Software Release Date, which indicates the release date of this version of the software.

To check for updated documentation, visit https://www.microfocus.com/support-and-services/documentation/.

Support

Visit the MySupport portal to access contact information and details about the products, services, and support that Micro Focus offers.

This portal also provides customer self-solve capabilities. It gives you a fast and efficient way to access interactive technical support tools needed to manage your business. As a valued support customer, you can benefit by using the MySupport portal to:

- View information about all services that Support offers
- Submit and track service requests
- Contact customer support
- Search for knowledge documents of interest
- View software vulnerability alerts
- Enter into discussions with other software customers
- Download software patches
- Manage software licenses, downloads, and support contracts

Many areas of the portal require you to sign in. If you need an account, you can create one when prompted to sign in.

Contents

Introduction

IDOL Ingest is a set of components for data retrieval and enrichment, that run within an open-source framework called Apache NiFi.

Micro Focus has added **IDOL Ingest with Usage Billing** to the Amazon Web Services (AWS) Marketplace, so that you can easily deploy an IDOL Ingest pipeline on hardware managed by AWS. Micro Focus provides an Amazon Machine Image (AMI) with Apache NiFi, IDOL Ingest components, and a selection of IDOL Connectors pre-installed.

To use IDOL Ingest with Usage Billing, you do not need to purchase an IDOL license from Micro Focus. Instead, your AWS account is charged, based on the amount of data you process and the types of processing you perform. To give some examples, you are charged when a connector retrieves a file from a data repository, when you use KeyView to filter the text from a file, or when you use Optical Character Recognition to extract text from an image.

KeyView Filtering and Eduction are commonly used together, so in each metering period you are charged for whichever is greater - the amount of text output from filtering or the amount of text sent to Eduction. This means that after you have filtered text from your files, you can use IDOL Eduction to extract entities (including PII) for no extra cost. For complete pricing information, refer to the AWS Marketplace.

IDOL Ingest with Usage Billing is ideal if you want to ingest data from repositories in the cloud, for example Microsoft 365 services such as Exchange, OneDrive, or SharePoint. You can also use it for constructing proof-of-concept systems and for trying new IDOL features.

This document describes how to get started with IDOL Ingest on AWS, but does not attempt to describe all of the features that IDOL Ingest provides. If you are not familiar with IDOL Ingest, Micro Focus recommends using this guide in conjunction with other IDOL documentation and the documentation for Apache NiFi.

Which IDOL Components Are Available?

File System Connector and Web Connector are installed by default.

The following connectors are also available but you must install them using the Micro Focus IDOL Ingest utilities page (see Install Extensions, on page 18).

- Amazon S3 Connector
- Exchange OData Connector
- Microsoft Teams Connector
- OneDrive Connector
- SharePoint OData Connector

You can process your data with IDOL Ingest components, performing operations such as:

- KeyView Filtering
- KeyView HTML Export
- Eduction
- Optical Character Recognition
- Face detection
- PII image analysis (combined OCR and face detection)

IDOL Ingest with Usage Billing does not provide a way to run an IDOL index (Content engine) but you can write IDOL documents to the file system, or send them to an Amazon Kinesis Data Firehose and write them to an Amazon S3 bucket.

If you want to run a complete IDOL system in AWS, or use other components that are not available in the AMI, Micro Focus recommends that you purchase an IDOL license. You can then use the IDOL "Bring Your Own License" offering in the AWS Marketplace.

Start a New IDOL Ingest Instance

This section provides a brief overview of how to subscribe to the IDOL Ingest product from the Amazon Marketplace, and start your first instance of IDOL Ingest. After you have subscribed to IDOL Ingest you can create as many instances as you like through the AWS Management Console. Micro Focus recommends referring to the Amazon documentation for comprehensive information about using and configuring Amazon Web Services.

To subscribe to IDOL Ingest and Create a New Instance

1. Go to the Amazon Marketplace and search for **Micro Focus IDOL Ingest**.
2. On the Micro Focus IDOL Ingest product page, click **Continue to Subscribe**.

   The Subscribe to this software page opens.
3. Read the license agreement and, if you agree to the terms, click **Accept Terms**.

   Amazon processes your subscription request. The page is updated when the subscription becomes active. The Effective date and Expiration date are not applicable (N/A) because you are charged based on the amount of data you process and the types of processing you perform.
4. Click **Continue to Configuration**.

   The Configure this software page opens.
5. In the **Region** box, choose a region in which to deploy the software. For example, you might prefer to have the software deployed in an AWS datacenter that is close to your location.
6. Click **Continue to Launch**.

   The Launch this Software page opens.
7. In the **Choose Action** list, click **Launch through EC2** and click **Launch**.

   The Amazon Web Service management console opens and you are asked to choose an instance type, which determines the specifications of the machine that the software is deployed on. IDOL Ingest requires at least 4GB of memory, so some of the instance types are unavailable. Micro Focus recommends the **t3.large** instance type as a good starting point but you can choose a different instance type depending on your requirements.
8. Choose an instance type and click **Next: Configure Instance Details**.

   The Configure Instance Details page opens.

   You must assign an IAM role so that the software can report metering events to AWS. The default setting is to automatically create a new IAM role.
9. Click **Next: Add Storage**.

   The Add Storage page opens. You can choose the amount of storage that will be available to the EC2 instance. The minimum required for the IDOL Ingest installation is 16GB but additional space is required for storing the data that you are processing, so Micro Focus recommends increasing this value.

10. Choose an amount of storage and click **Next: Add Tags**.

    The Add Tags page opens.

11. Add a name tag to the EC2 instance, drive volumes, and network interfaces. This makes it easier to find the instance at a later time. To do this, click the hyperlink labeled **click to add a Name tag**, and then type a value.

12. Click **Next: Configure Security Group**.

    The Configure Security Group page opens. By default, the machine will accept inbound connections to port 22 (for SSH access) and port 443 (the HTTPS port for the Apache NiFi UI) from any IP address.

13. Accept the default settings, change them as required, or use an existing security group that you have already configured. For example, you can limit access to known IP addresses or address ranges.

14. Click **Review and Launch**.

    The Review Instance Launch page opens.

15. Review the confirmation page and click **Launch**.

    A dialog box opens so that you can create a new key pair, or choose an existing key pair to use to log into the EC2 instance. Create your private key in `.ppk` format, so that in a later step you can use the key with PuTTY and log on to the machine.

16. Create or select a key pair and then click **Launch Instances**.

    Amazon starts the new EC2 instance and the Launch Status page opens.

17. Review the information and click **View Instances**.

    The Instances page, in the AWS Management Console, opens.

18. In the list of instances, find the instance that you created. For example, you can search for the tag name that you applied earlier. Select the instance to view details such as its hostname.

19. When the IDOL Ingest instance has started, open the Public IPv4 DNS address in your web browser.

    Your browser is likely to display a security warning because the SSL certificate that is presented by the server is self-signed. After you acknowledge the warning, a web page opens with links to NiFi and the NiFi Registry.

20. If the Apache NiFi application has already started, you are redirected straight to NiFi. Otherwise, click **NiFi** to go to the Apache NiFi login page.

    > **NOTE:** The Apache NiFi application might take up to 10 minutes to start.

    The Apache NiFi instance has been configured with a default user name and random password. The user name is **`nifi`**. The following steps describe how to obtain the password.

21. Open an SSH client such as PuTTY, and log in to the EC2 instance.

- Obtain the hostname from the AWS management console, and use port 22.
- Specify the path of your private key file. In PuTTY you can do this from the configuration dialog by clicking **Connection** > **SSH** > **Auth**.
- When you start the session and are asked for the user name, type **`ec2-user`**.

22. In the home directory of `ec2-user` is a text file named `nifi-password` that contains the password. Log into NiFi with the user name **`nifi`** and the password contained in this file.

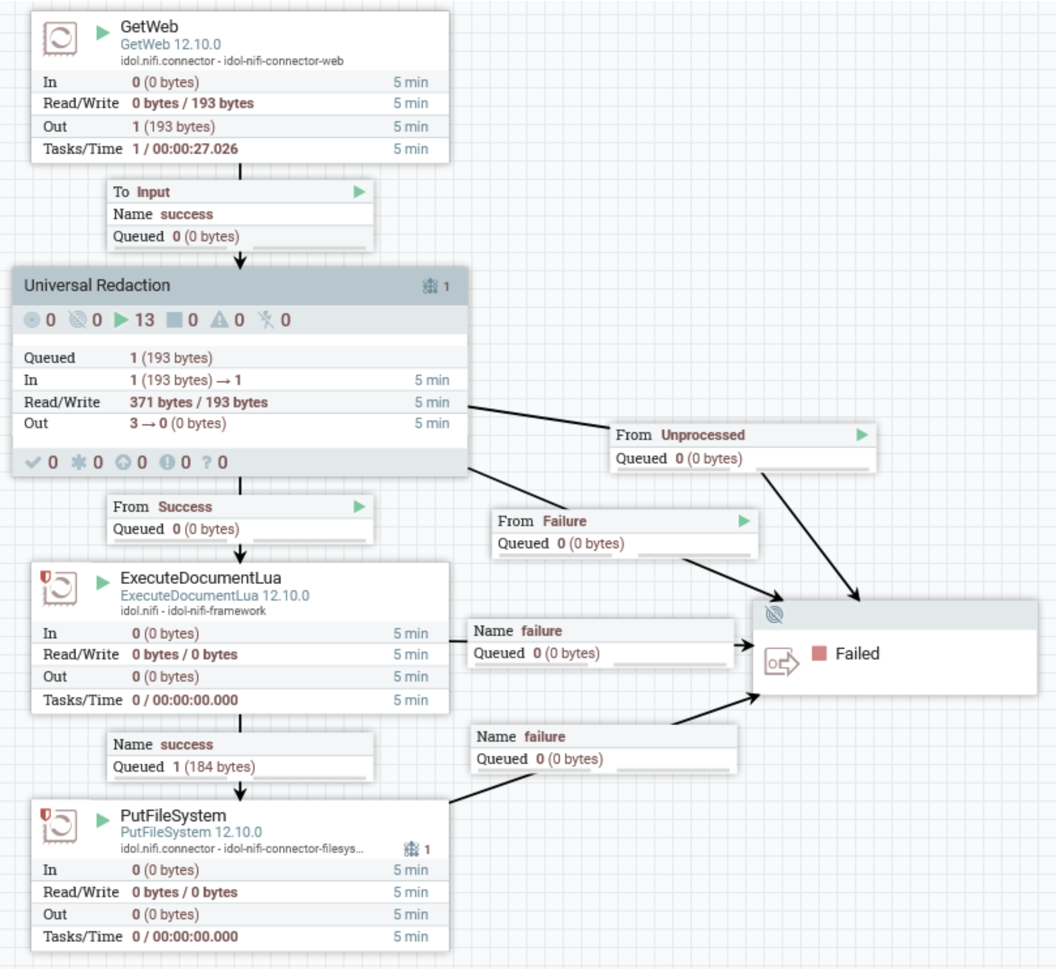
Create a Dataflow

After following the instructions in Start a New IDOL Ingest Instance you can log in to the Apache NiFi web interface. Initially the canvas is blank, but the NiFi instance has pre-defined controller services for KeyView filtering, KeyView export, and for licensing through AWS.

> **NOTE:** The components that you use in your dataflow must reference the pre-configured AWS licensing service, and you must not create any other licensing services.

The IDOL NiFi Ingest documentation provides detailed instructions about how to set up a dataflow using IDOL Ingest components in Apache NiFi. This section provides a short guided walkthrough that demonstrates how to perform universal redaction of personal information such as names. This is just one possible dataflow that you could construct.

After following all of the steps in this section, the complete dataflow should look like the one in the following image.



GetWeb
GetWeb 12.10.0
idol.nifi.connector - idol-nifi-connector-web

In	0 (0 bytes)	5 min
Read/Write	0 bytes / 193 bytes	5 min
Out	1 (193 bytes)	5 min
Tasks/Time	1 / 00:00:27.026	5 min

To Input
Name success
Queued 0 (0 bytes)

Universal Redaction 1
0 0 13 0 0 0

Queued	1 (193 bytes)	
In	1 (193 bytes) → 1	5 min
Read/Write	371 bytes / 193 bytes	5 min
Out	3 → 0 (0 bytes)	5 min

✓ 0 ✱ 0 0 0 ? 0

From Unprocessed
Queued 0 (0 bytes)

From Success
Queued 0 (0 bytes)

From Failure
Queued 0 (0 bytes)

ExecuteDocumentLua
ExecuteDocumentLua 12.10.0
idol.nifi - idol-nifi-framework

In	0 (0 bytes)	5 min
Read/Write	0 bytes / 0 bytes	5 min
Out	0 (0 bytes)	5 min
Tasks/Time	0 / 00:00:00.000	5 min

Name failure
Queued 0 (0 bytes)

Failed

Name success
Queued 1 (184 bytes)

Name failure
Queued 0 (0 bytes)

PutFileSystem
PutFileSystem 12.10.0
idol.nifi.connector - idol-nifi-connector-filesys... 1

In	0 (0 bytes)	5 min
Read/Write	0 bytes / 0 bytes	5 min
Out	0 (0 bytes)	5 min
Tasks/Time	0 / 00:00:00.000	5 min

The dataflow includes the following steps:

- First, the IDOL Web Connector (the GetWeb processor) retrieves a page from the web. The connector creates a FlowFile to which it attaches the HTML page it has retrieved. You could replace the Web Connector with a different connector if you prefer. See Which IDOL Components Are Available?
- The FlowFile is passed to the universal redaction process group. This is a pre-configured group of components provided by Micro Focus as a NiFi template. The universal redaction process group is capable of processing HTML pages, images, and any type of document that can be exported to HTML by KeyView HTML Export. In this case the connector retrieves HTML, so the process group renders the web page to produce an image, and then runs Eduction on the text to locate any personal information. The end result is that the HTML file, that was retrieved by the Web Connector, is replaced by a redacted image of the web page.
- An ExecuteDocumentLua processor adds an attribute to the FlowFile, named `idol.put.migrationuri`. IDOL Connectors can use the migration URI to write the document to a different repository. In this example, we retrieve a page from the Web but are writing the redacted image to a file system. The Lua script also adds the correct file extension to the output file path.
- The redacted image is written to disk by a File System Connector (the PutFileSystem processor).

Add a Connector to Retrieve Data

In most dataflows the first step is to retrieve the data that you want to analyze. In this example we deploy a Web Connector to retrieve a page from the web.

To add the connector

1. Add a processor, by dragging the processor icon from the components toolbar to the canvas.

   The Add Processor dialog box opens.

2. In the **Source** list, click **idol.nifi.connector**.

   The list of processors is filtered to show only IDOL connectors.

3. Select the **GetWeb** processor and click **ADD**.

   The processor is added to the canvas.

4. Right-click the processor and click **Configure**.

   The Configure Processor dialog box opens.

5. Click the **Properties** tab.
6. Click **ADVANCED**.

   The guided setup wizard opens.

7. Configure the connector, by setting the following properties. For all other properties, you can accept the default value.

IDOL License Service	You must use the pre-configured license service, which should be selected automatically.
Startpoint Type	Set this property to `URL`.
Start URLs	Choose a web page that you want to redact. You could choose a page that contains names, for example: https://en.wikipedia.org/wiki/2012_Summer_Olympics_medal_table
Maximum Crawl Depth	To begin with, for the purposes of this example, set this property to **0** (zero) so that the connector does not follow any links and only ingests the initial URL.
Spider Url Must Have Regex	`https:\/\/en\.wikipedia\.org\/[^:]*`
Content Type Cant Have Regex	`(application|text)/(javascript|xml|x-javascript|css)(;.*)?`
Max Links Per Page	Set this property to `0` (unlimited)
Remove Comment Tags	`TRUE`
Remove NoFrames Tags	`TRUE`
Remove NoScript Tags	`TRUE`
Remove Script Tags	`TRUE`

8. Close the configuration dialog.

Add Universal Redaction

Micro Focus provides a NiFi template that contains a process group for performing redaction.

To add the universal redaction process group

1. Drag the template icon from the components toolbar to the canvas.

   The Add Template dialog box opens.

2. In the **Choose Template** list, click **Universal Redaction**, and then click **ADD**.

   The process group appears on the canvas.

3. Create a connection between the GetWeb connector and the Universal Redaction process group. Hover the mouse over the connector until you see the connection icon - ↘ - and then drag the icon to the process group.

The Create Connection dialog box opens.

4. In the **For Relationships** area, select the **success** check box so that documents that were successfully retrieved are queued for processing. Then, click **ADD**.

   The connection appears on the canvas. NiFi automatically adds a queue between the connector and the process group.

5. Double click the process group to open it.

6. In the Operate Palette, click **Configuration** ⚙.

   The Universal Redaction Configuration dialog box opens.

7. The process group includes a pre-configured Media Service. Enable the service by clicking **Enable** ⚡.

8. Close the Universal Redaction Configuration dialog box, and exit the process group (right-click on the canvas and click **Leave Group**).

Write the Redacted Images to Disk

To write the redacted images to disk

1. Add a processor, by dragging the processor icon from the components toolbar to the canvas.

   The Add Processor dialog box opens.

2. Select the **ExecuteDocumentLua** processor and click **ADD**.

   The processor is added to the canvas.

3. Right-click the processor and click **Configure**.

   The Configure Processor dialog box opens.

4. Configure the ExecuteDocumentLua processor:

   - Set the **IDOL License Service** property to the pre-confgured AWS license service.
   - Set the **Lua script function arguments** property to **`LuaFlowFileDocument, LuaProcessorSession`**.
   - Click **ADVANCED** to open the advanced configuration interface, and paste the following Lua script into the code editor, replacing the empty `handler` function. Then, click **SAVE**.

```
function handler(flowfiledocument, session)
  flowfiledocument:modify({
    preAction = function(action)
      local meta = action:getXmlMetadata()
      local path = meta:getFirstChild("AUTN_MIGRATION_URI"):getValue()

      if meta:getElementsByXPath("//HTML_RENDER") ~= nil then
        -- RenderHTML generates .bmp files
        path = string.gsub(path, "?authority", ".bmp?authority", 1)
      end

      action:setAttribute("idol.put.migrationuri", path)

    end})
end
```

- On the **Settings** tab of the configuration dialog box, find the **Automatically Terminate Relationships** area, and select the **returned** check box.

5. Add a **PutFileSystem** processor to the canvas.

6. Configure the processor:

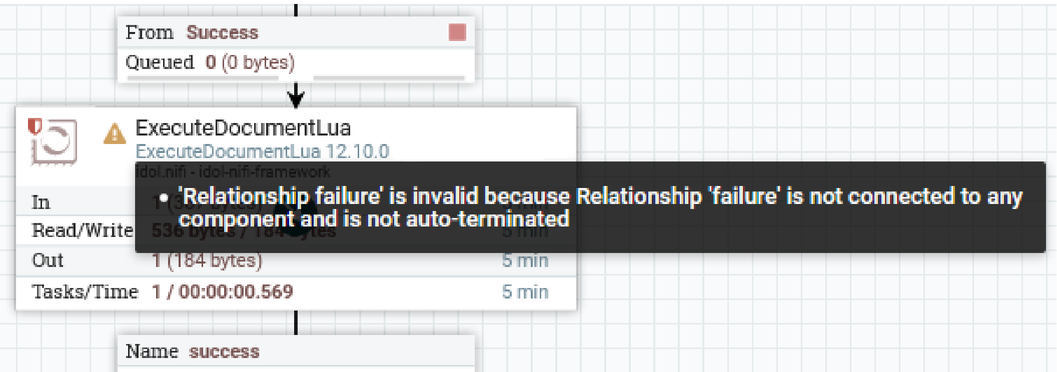
   - Set the **IDOL License Service** property to the pre-confgured AWS license service.

   - Click + and add a dynamic property named **`migration:rootDirectory`**. For the value of the property, use the path where you want to output your redacted images, for example **`/home/nifi/output/`**.

7. Create a connection between the Universal Redaction process group and the ExecuteDocumentLua processor. Select the **Success** output relationship, to process successfully redacted files.

8. Create a connection between the ExecuteDocumentLua processor and the PutFileSystem processor. Select the **success** output relationship.

   At this point the processors cannot be started. If you hover the mouse pointer over the warning triangle you will see the following message. This informs us that we cannot start the processor because FlowFiles that are not successfully processed are not routed anywhere.



From Success
Queued 0 (0 bytes)

ExecuteDocumentLua
ExecuteDocumentLua 12.10.0
idol.nifi - idol-nifi-framework

- 'Relationship failure' is invalid because Relationship 'failure' is not connected to any component and is not auto-terminated

In
Read/Write 536 bytes / 184 bytes 5 min
Out 1 (184 bytes) 5 min
Tasks/Time 1 / 00:00:00.569 5 min

Name success

In a production system you could route failed FlowFiles somewhere so that they could be re-processed. In this example, we can route the FlowFiles to an output port. Any FlowFiles that are sent to the output port will be visible in a queue, but in this example they are not routed anywhere.

9. Add an output port by dragging the output port icon from the components toolbar to the canvas.

   The Add Port dialog box opens.

10. In the **Output Port Name** box, type a name for the output port and click **ADD**.

    The output port is added to the canvas.

11. Create connections from the Universal Redaction process group, ExecuteDocumentLua processor, and PutFileSystem processor to the output port. In each case, select the **failure** relationship. If you wish, you can also connect the **unprocessed** relationship from the Universal Redaction process group to the output port, so that files that could not be processed are also routed there.

Start Processing

After you have finished building the dataflow, you can start processing.

To start processing

1. In the NiFi canvas, right-click your GetWeb processor and click **Start**.

   The connector begins fetching data. The remaining processors have not been started, so the document waits in the following queue. In this example, we configured the Web Connector to retrieve a single page, so the connector produces only one document.

2. You can start each processor and see the FlowFile move from queue to queue, or use the start button in the operate palette to start every processor in your dataflow.

3. After the FlowFile has been processed by the PutFileSystem processor, connect to the machine and look in the directory `/home/nifi/output`. You should see your redacted image.

   If you wish you can copy the image to the `ec2-user` home directory (`/home/ec2-user/`) and then use the command-line tool `pscp` (supplied with PuTTY) to copy the image to your local machine. For example:

   a. In PuTTY, while connected to the EC2 virtual machine through SSH:

   ```
   sudo cp -r /home/nifi/output ~
   ```

   b. Then, from the command-line on your local machine:

   ```
   pscp -i idol-ingest.ppk ec2-
   user@hostname.compute.amazonaws.com:/home/ec2-user/output/wiki/*.bmp .
   ```

where *idol-ingest.ppk* is the path of the private key for logging on to the EC2 virtual machine, and *hostname* is replaced by the correct name.

Eduction Resource Files

IDOL Ingest with Usage Billing includes the PII, PCI, and PHI Eduction grammars. These grammar files and associated resources are pre-installed. When you configure the Eduction processor, you do not need to specify the full path to a resource file. You can specify only the file name, with an appropriate prefix.

Eduction resources	Prefix	Example
PII grammars	`PII/`	`PII/address.ecr`
PII post-processing scripts	`PII/scripts/`	`PII/scripts/pii_postprocessing.lua`
PII pre-filter resources	`PII/prefilter/`	`PII/prefilter/address_street_markers.dpf`
PCI grammars	`PCI/`	`PCI/pci_numbers.ecr`
PCI post-processing scripts	`PCI/scripts/`	`PCI/scripts/pci_postprocessing.lua`
PHI grammars	`PHI/`	`PHI/medical_terms.ecr`
PHI post-processing scripts	`PHI/scripts/`	`PHI/scripts/phi_postprocessing.lua`
PHI pre-filter resources	`PHI/prefilter/`	`PHI/prefilter/medical_terms_prefilter_dict.dpf`
Standard grammars	No prefix	`team_baseball.ecr`

IDOL Ingest Utilities

You can use the IDOL Ingest utilities to:

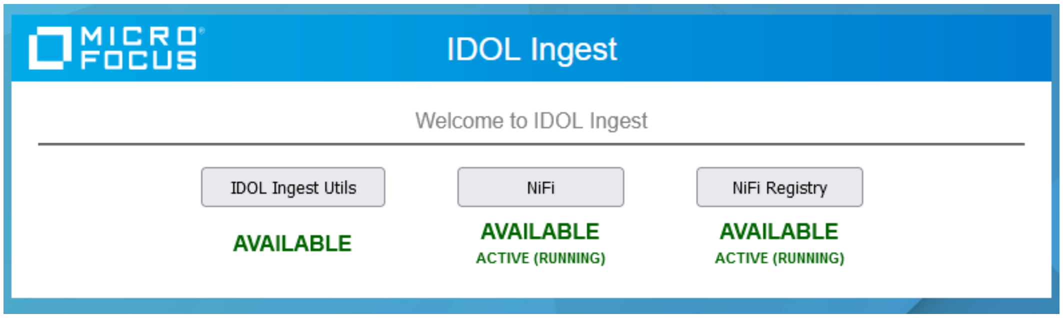
- Backup and restore the state of the NiFi instance.
- Install NiFi extensions that are not installed by default, for example additional connectors.
- Change the password that you use to log in to NiFi, the NiFi registry, and the IDOL Ingest utilities.
- Migrate from AWS Licensing (usage billing) to an IDOL license that you have purchased ("Bring Your Own License").

To access the IDOL Ingest utilities

1. Go to the following URL in a web browser, where `<ec2-hostname>` is the public DNS address for your IDOL Ingest instance:

   ```
   https://<ec2-hostname>
   ```

   The following web page opens. This page shows the status of NiFi and the NiFi Registry.

2. Click **IDOL Ingest Utils**.
3. Log in using the same user name and password that you would use to log in to NiFi. The default user name is **`nifi`** and the steps for obtaining the password are described in Start a New IDOL Ingest Instance, on page 6.

Backup and Restore the NiFi State

You can use the IDOL Ingest utilities to backup and restore the state of your NiFi instance. You can also use this feature to move your dataflow from one IDOL Ingest instance to another of the same version or later.

> **IMPORTANT:** The exported state information does not include any FlowFiles. In other words, your dataflow is saved but the data being processed is not.

To backup the NiFi state

1. In NiFi, stop all of your processors and ensure that there are no FlowFiles waiting in queues.
2. Open the IDOL Ingest utilities page.
3. In the **Export NiFi State** area, click **Export**.

   Your browser will ask you where to save the file (`nifi_export.tar.gz`).

To restore the NiFi state

1. In NiFi, stop all of your processors and ensure that there are no FlowFiles waiting in queues.

   > **CAUTION:** When you import state information, the imported state will overwrite any dataflow that you have constructed in NiFi.

2. Open the IDOL Ingest utilities page.
3. In the **Import NiFi State** area, click **Browse** and select the file that you exported.
4. Click **Import**.

Install Extensions

To ensure that IDOL Ingest starts quickly when used for the first time, it includes only some of the available extensions. You can install additional extensions, such as additional connectors that you want to use.

To install extensions

1. Open the IDOL Ingest utilities page.
2. In the **Install extension** area, choose a NAR file to install by selecting options from the filters.

   For example, to install the SharePoint OData connector, you could select **idol.nifi.connector** in the **Filter by group name** list, and then choose **idol-nifi-connector-sharepointodata** from the **Filter by artifact name** list.

3. Click **Install extension**.

   The extension is installed. The NiFi application does not restart during this process. It might take a minute for the new extension to become available in NiFi.

Change Your NiFi Password

To change the password that you use to log in to NiFi, the NiFi registry, and the IDOL Ingest utilities page, follow these steps.

To change your NiFi password

1. Open the IDOL Ingest utilities page.
2. In the **Change Password** area, type your old password, and choose a new password.
3. Click **Change Password**.

   Your password is updated.

Use Your Own IDOL License

To use an IDOL license that you have purchased from Micro Focus, rather than AWS Licensing (usage billing), follow these steps.

You remain responsible for the cost of running the Amazon EC2 instance. The IDOL license replaces the charges for using IDOL features such as filtering text from files or using Optical Character Recognition.

To migrate to your own IDOL license

1. Open the IDOL Ingest utilities page.
2. Find the **Migrate to BYOL** area.
3. Type the host name and port of your IDOL License Server (which must be accessible from the internet).
4. Click **Migrate**.

Troubleshooting

Where is the Apache NiFi instance installed?

On your Amazon EC2 virtual machine, the Apache NiFi application is installed in the directory `/opt/mf/nifi/`.

NiFi processors display licensing errors

The Amazon EC2 virtual machine that hosts the NiFi instance must have an associated IAM role, so that the Micro Focus software can report metering events to AWS. If you see your NiFi processors displaying licensing errors, such as "'IDOL License Service' is invalid because License Suspended", it is possible that the IAM role was not assigned correctly.

By default, an IAM role is created automatically when you start a new IDOL Ingest instance (see step 8 in Start a New IDOL Ingest Instance, on page 6). Ensure that this IAM role is associated with your EC2 virtual machine. For information about how to associate an IAM role with your EC2 instance, refer to the AWS documentation.

Documentation

The following documentation provides more information.

- For information about IDOL Ingest components, refer to the *IDOL NiFi Ingest Help*.
- For additional information about using connectors, refer to the IDOL Connector Documentation. A help system is available for each IDOL Connector.

IDOL documentation is available from https://www.microfocus.com/documentation/idol.