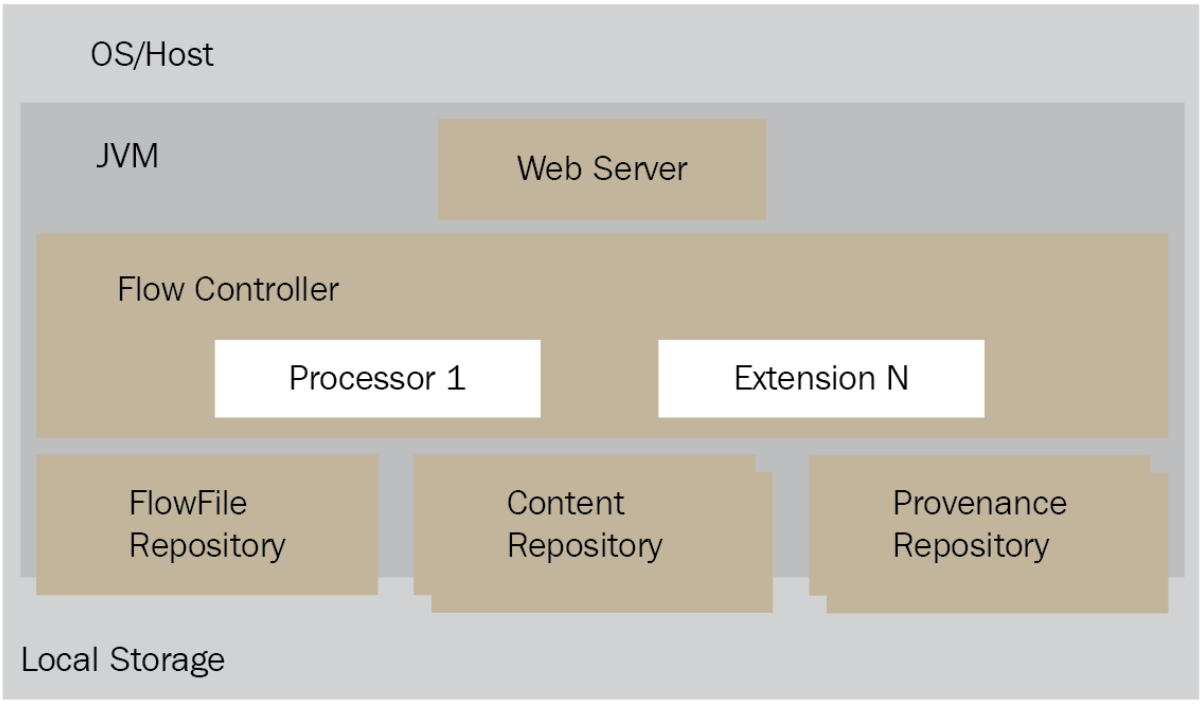
Apache NiFi Architecture

Architectural Overview

Apache NiFi is a highly valuable data flow system designed to automate and schedule various data operations across different processes **[1]**. It is capable of handling both simple and complex systems effectively. The performance of Apache NiFi, in terms of IO, CPU, and RAM utilisation, is spectacular as well. Apache NiFi has numerous characteristics and features, which can be highlighted in the following points:

- Flow management
- Ease of use
- Security
- Extensible architecture
- and flexible scaling model



Core Concepts

The most important concepts related to Apache NiFi are DataFlow Manager (DFM), FlowFile, Bulletin, Processor, Connection, Controller Service, Flow Controller, and Process Group. The significance of each concept is as follows:

- **DFM**

In simple words, a DFM is the user who has different permissions and complete management of Apache NiFi.

- **FlowFile**

The FlowFile concept refers to each object moving through the system. It has two main parts: attributes and content.

The first one stores metadata about the FlowFile, such as its filename, size, etc.

The standard attributes that each FlowFile contains are *Universally Unique Identifier (UUID)*, used to distinguish the FlowFile; *filename*, used to represent the

name of the output of the data when it is written to some path; and *path*, used to define the path of the filename.

On the other hand, the content part is responsible for storing the actual data of the FlowFile.

- **Bulletin**

It can be considered a service reporter tool that is available for each service. It provides critical information such as rolling statistics, current status, and severity levels such as Debug, Info, Warning, and Error, giving insights into the current situation of the component.

- **Processor**

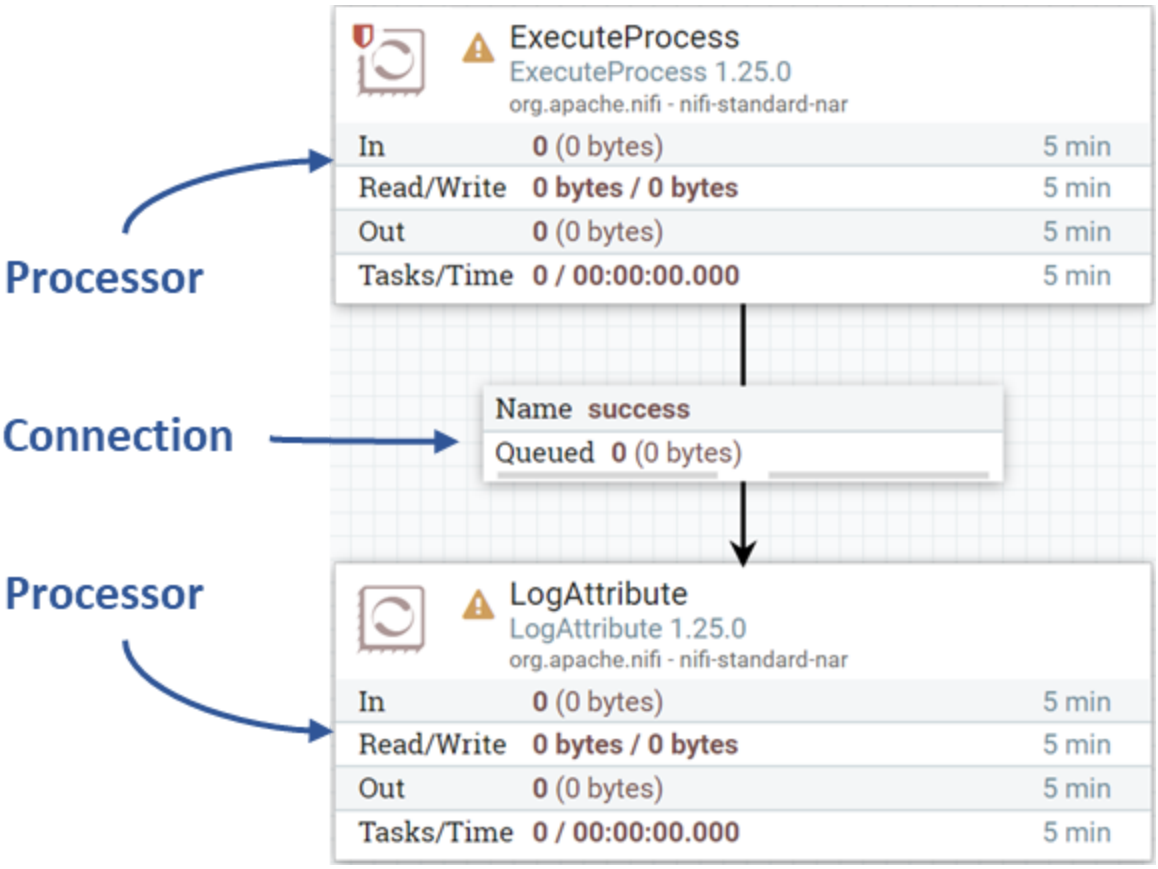
A Processor, as the name suggests, is responsible for performing the actual work, such as data routing and transformation, as it has direct access to the attributes and the content of the given FlowFile. For example, processor has the ability to read data files, execute different software languages such as Python scripts, and executing different SQL queries.

- **Connection**

A Connection can be regarded as the link between different processors.

It manages the interaction between the processors by establishing queues.

A queue in Apache NiFi refers to a buffer that has the ability to store FlowFiles. Therefore, the output of one processor, the FlowFile, is stored in a queue until the other processor is ready to process it.



- **Controller Service**

When a Controller Service is created, it will automatically start up every time Apache NiFi is initiated.

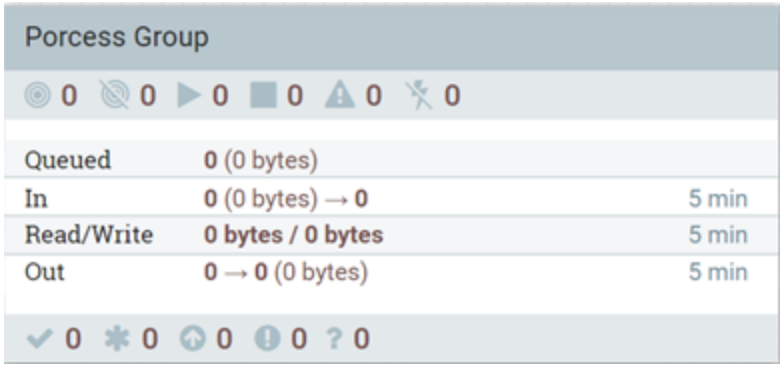
Several essential services can be specified in the Controller Service section. For instance, DBCPConnectionPool and CSVRecordSetWriter are utilised to establish connections between various database systems and Apache NiFi, and to enable writing CSV files to the local system, respectively. Once defined, these services can be utilised by any processors.

- **Flow Controller**

The Flow Controller, in short, manages the threads used by different processes. As a result, it can dynamically adjust the number of threads used for each flow, increasing or decreasing as necessary.

- **Process Group**

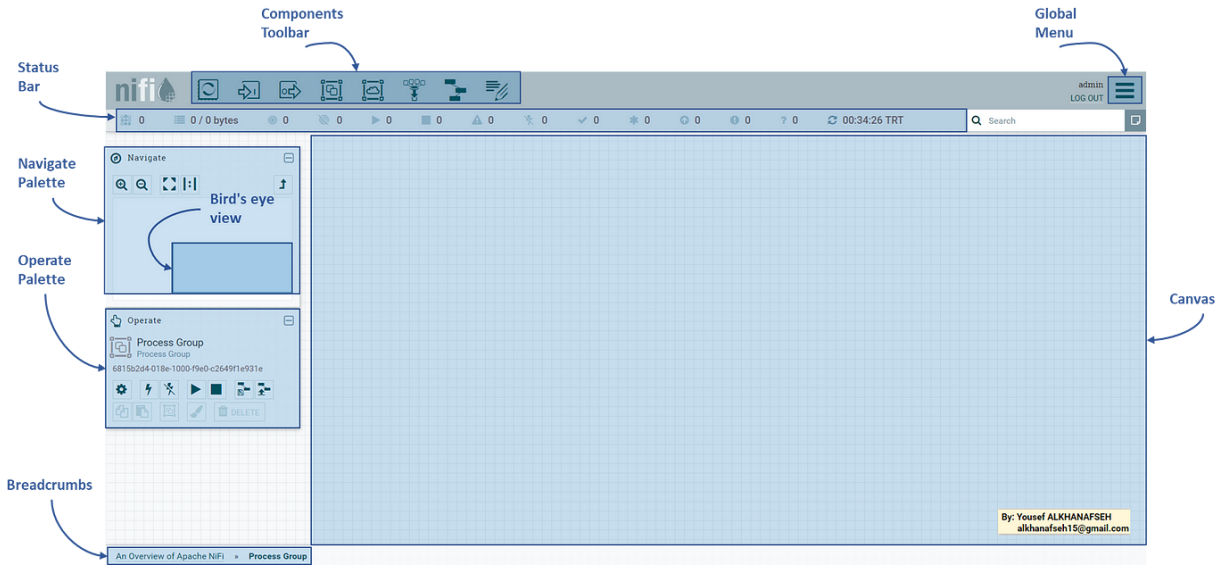
Grouping a set of processors and their connections is called a Process Group. The input and output data of this Process Group can be managed by Input Ports and Output Ports, respectively.



- **Input Ports**
- **Output Ports**

Apache NiFi - User Interface

Specifically, Apache NiFi UI comprises six main elements: the components toolbar, status bar, navigation palette, operation palette, global menu, and canvas.



Components Toolbar

The components toolbar comprises seven main elements:

1. Processor
2. Input & Output ports
3. Process Group
4. Remote Process Group
5. Funnel
6. Template
7. Label

The first three components have been discussed in the preceding section.

- **Funnel**

Generally, Funnel is utilised when data from several Connections need to be merged into a single Connection.

- **Remote Process Group**

A Remote Process Group in NiFi is used to facilitate the transfer of data between different NiFi systems.

- **Template**

Apache NiFi enables DFMs to create templates with unique names for their dataflows. These templates can be found inside the Template icon in the Components Toolbar. This feature allows them to easily reuse specific parts of the flow by simply selecting the corresponding template they have created.

- **Label**

In summary, Apache NiFi offers the capability to add extra labels to the canvas, enabling users to incorporate additional information, comments, or advices. This can be achieved by dragging the label icon onto the canvas.

Status Bar

Status Bar plays a crucial role in displaying vital information about the currently running dataflow. It consists of 15 different components.

The first one is *active threads*, which is the number of total threads being used by the current dataflow.

The second one is *total queued data*, which represents the amount of data waiting to be processed by other processors.

The third and fourth components are *transmitting and not transmitting remote process groups*, which show the count of the currently running groups for each state, respectively.

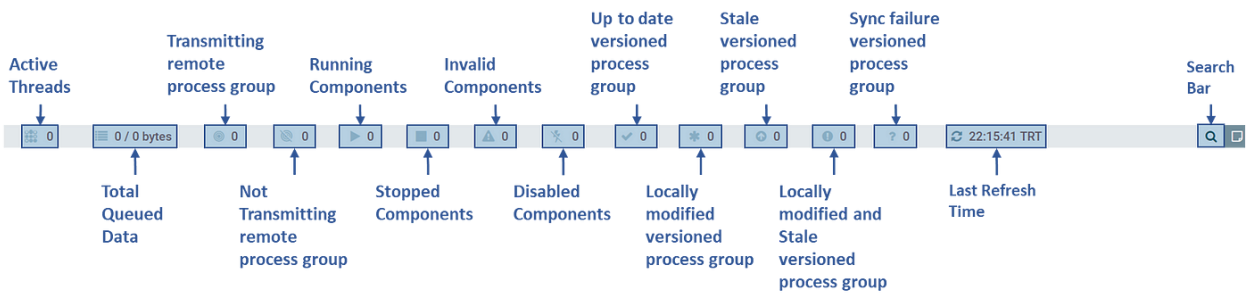
The next four components are related to the status of components, indicating if they are *running*, *stopped*, *invalid*, or *disabled*, respectively.

On the other hand, the subsequent five components are responsible for version states. The *Up to date versioned process group* shows the latest flow version.

Moreover, the *Locally modified versioned process group* indicates the number of local changes that have been made. The *Stale versioned process group* component informs if there is a new available version of the flow.

The *Locally modified and Stale versioned process group* occurs when the combination of the previous two components occurs simultaneously. The *Sync failure versioned process group* indicates if there is a conflict in synchronising the flow with the registry.

Additionally, the *Last refresh time* component displays the last refresh time of the current dataflow. Finally, the *search bar* is used to make general searches in Apache NiFi.

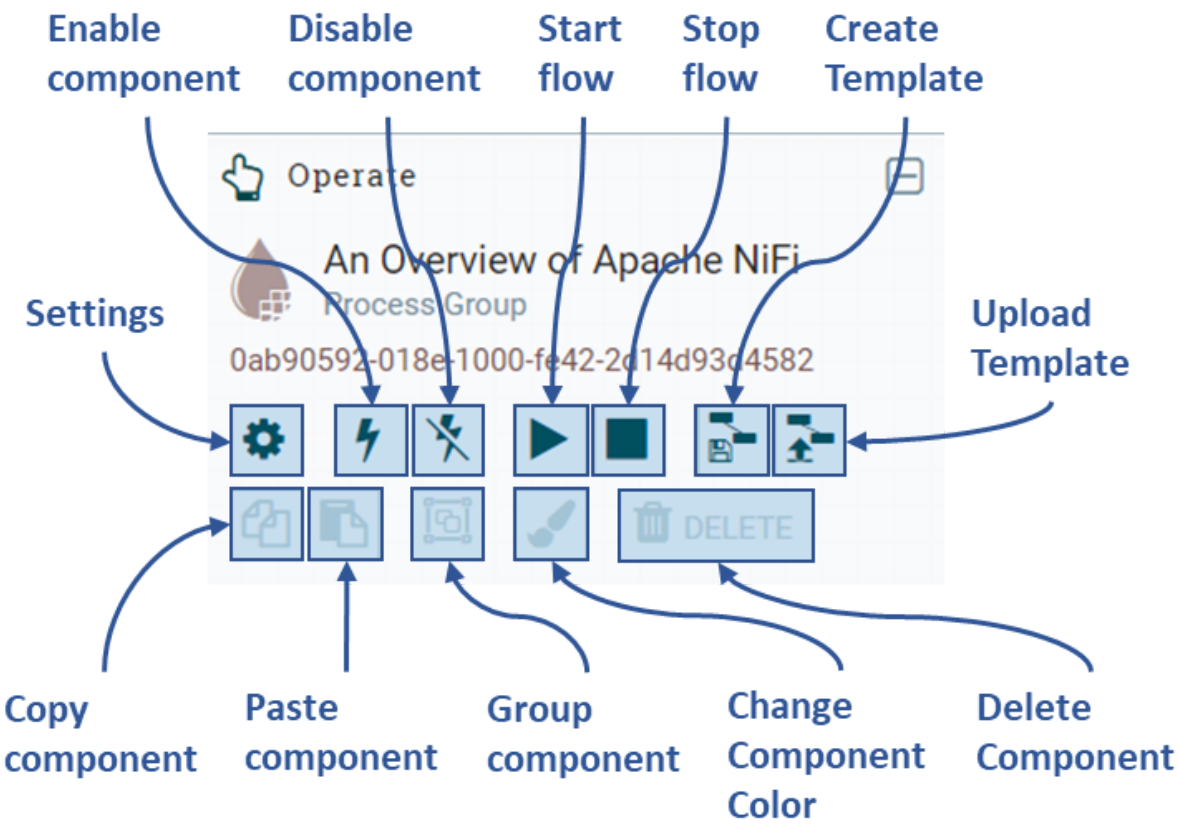


Navigate palette, Bird’s eye view, Breadcrumbs, and Canvas

The Navigate Palette enables DFMs to navigate around the canvas and to zoom in and out. The Bird’s Eye View is a part of the Navigate Palette that allows the required components to be displayed on the canvas within one frame. The Breadcrumbs section shows the depth of the current dataflow and process groups, which can be used to navigate around them. Eventually, Canvas is the home screen of Apache NiFi which is supposed to carry different services or processors.

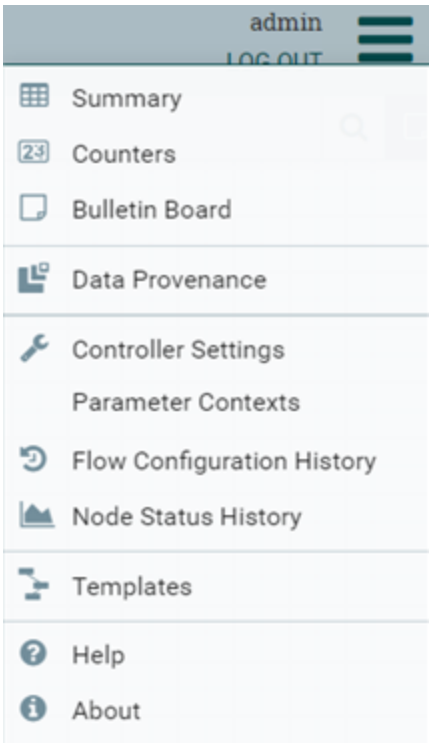
Operate Palette

The Operate Palette is home to the main buttons of the flow providing DFMs flexibility to execute different actions. These include configuring system settings, such as modifying general flow information or adding new Controller Services. Additionally, the Operate Palette allows for a range of operations on the current flow, like enabling, disabling, running, stopping, copying, pasting, grouping, changing colors, and deleting components. It also features two buttons dedicated to creating a template for the current flow or uploading a previously saved template.



Global Menu

Specifically, the Global Menu is employed to make modifications to the existing components in the dataflow. It also facilitates access to the flow configuration history, node status history, controller settings, bulletin board, and available templates. Additionally, the “Help” section within the Global Menu provides further information about the components.


admin
LOG OUT
Summary
Counters
Bulletin Board
Data Provenance
Controller Settings
Parameter Contexts
Flow Configuration History
Node Status History
Templates
Help
About