iykra Data Fellowship

**IYKRA**
Data Fellowship Program

# High-Volume Data Ingestion with NiFi

by Fariz Wakan

April 16th, 2024


# Intro to Apache NiFi

# Apache NiFi

- Apache NiFi is an open-source data integration tool designed to automate the flow of data between systems in a scalable, efficient, and secure manner.
- It provides a user-friendly web-based interface to design data flows and manage the movement of data between various sources and destinations.
- NiFi is part of the Apache Software Foundation and is built to handle real-time data streaming, large-scale data transfers, and data transformation tasks.

# Features

- Apache NiFi provides a web-based user interface, which provides seamless experience between design, control, feedback, and monitoring.
- It is highly configurable. This helps users with guaranteed delivery, low latency, high throughput, dynamic prioritization, back pressure and modify flows on runtime.
- It also provides data provenance module to track and monitor data from the start to the end of the flow.
- Developers can create their own custom processors and reporting tasks according to their needs.
- NiFi also provides support to secure protocols like SSL, HTTPS, SSH and other encryptions.
- It also supports user and role management and also can be configured with LDAP for authorization.

# Key Concepts

- **Process Group** – It is a group of NiFi flows, which helps a user to manage and keep flows in hierarchical manner.
- **Flow** – It is created connecting different processors to transfer and modify data if required from one data source or sources to another destination data sources.
- **Processor** – A processor is a java module responsible for either fetching data from sourcing system or storing it in destination system. Other processors are also used to add attributes or change content in flowfile.
- **Flowfile** – It is the basic usage of NiFi, which represents the single object of the data picked from source system in NiFi. NiFi processor makes changes to flowfile while it moves from the source processor to the destination. Different events like CREATE, CLONE, RECEIVE, etc. are performed on flowfile by different processors in a flow.

# Key Concepts

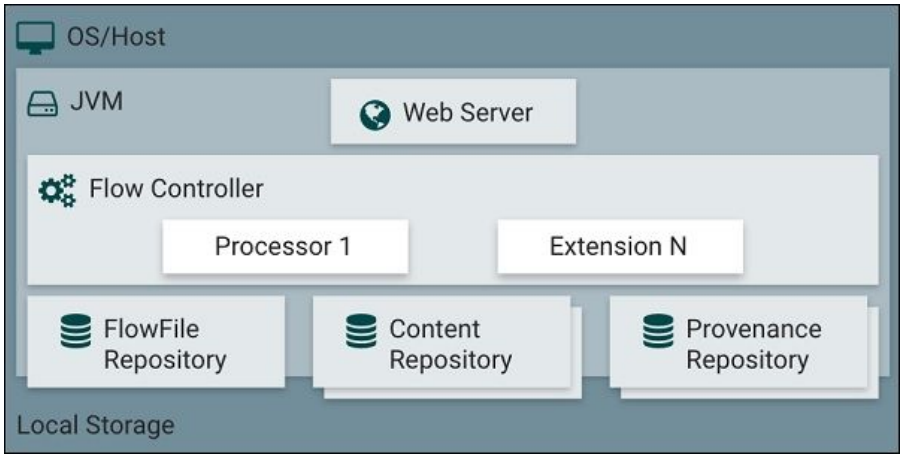
- **Event** – Events represent the change in flowfile while traversing through a NiFi Flow. These events are tracked in data provenance.
- **Data provenance** – It is a repository.It also has a UI, which enables users to check the information about a flowfile and helps in troubleshooting if any issues that arise during the processing of a flowfile.

# Advantages

- Apache NiFi enables data fetching from remote machines by using SFTP and guarantees data lineage.
- Apache NiFi supports clustering, so it can work on multiple nodes with same flow processing different data, which increase the performance of data processing.
- It also provides security policies on user level, process group level and other modules too.
- Its UI can also run on HTTPS, which makes the interaction of users with NiFi secure.
- NiFi supports around 188 processors and a user can also create custom plugins to support a wide variety of data systems.

# Basic Concepts

Apache NiFi consist of a web server, flow controller and a processor, which runs on Java Virtual Machine. It also has 3 repositories Flowfile Repository, Content Repository, and Provenance Repository.



# Flowfile Repository

This repository stores the current state and attributes of every flowfile that goes through the data flows of apache NiFi. The default location of this repository is in the root directory of apache NiFi. The location of this repository can be changed by changing the property named "nifi.flowfile.repository.directory".

# Content Repository

This repository contains all the content present in all the flowfiles of NiFi. Its default directory is also in the root directory of NiFi and it can be changed using "org.apache.nifi.controller.repository.FileSystemRepository" property. This directory uses large space in disk so it is advisable to have enough space in the installation disk.

# Provenance Repository

The repository tracks and stores all the events of all the flowfiles that flow in NiFi. There are two provenance repositories - volatile provenance repository (in this repository all the provenance data get lost after restart) and persistent provenance repository. Its default directory is also in the root directory of NiFi and it can be changed using "org.apache.nifi.provenance.PersistentProvenanceRepository" and "org.apache.nifi.provenance.VolatileProvenanceRepositor" property for the respective repositories.

# Play with Apache NiFi

Demo Session

iykra Data Fellowship

# Thank you!