Modern American Journal of Engineering, Technology, and Innovation
ISSN(E): 3067-7939
Volume 2, Issue 5, May, 2026
Website: usajournals.org

USE INFORMATION RETRIEVAL TECHNIQUES TO CLASSIFY THE FILE TYPE WITH APACHE NiFi

Hasanain M. J. Alfouadi
University of Al-Qadisiyah, Computer Center
hasanain.alfouadi@qu.edu.iq

Abstract
When creating big data pipelines to carry data and deal with the nature of the data type, we need to think about how to accommodate the size, variety, and speed of the data.
All primary considerations such as scalability, reliability, adaptability, cost in terms of development time, etc. play out when determining which tools should be adapted to meet our requirements. In this article, we'll focus briefly on Apache View Tools: Apache NiFi, use information retrieval techniques to classify file types by file type, and store each file type in its own folder.

Keywords: Information Retrieval, Apache NiFi, Data Pipelines.

Introduction

Information Retrieval (IR) is the activity of obtaining information system resources related to the need for information from a group of those resources. Searches can be based on full text or other content-based indexing. Information retrieval is the science of searching for information in a document, searching for the documents themselves, as well as searching for metadata describing data, databases of text, images, or sounds and categorizing them according to their type. [1]
So Apache NiFi provides users with the ability to create and manipulate large, complex data streams with NiFi. This is accomplished using the basic components: processor, funnel, I / O port, process set, and remote set of operations. These can be considered building blocks of DataFlow. Sometimes,

Modern American Journal of Engineering, Technology, and Innovation
ISSN(E): 3067-7939
Volume 2, Issue 5, May, 2026
Website: usajournals.org

though, using these small building blocks can become tedious if the same reasoning needs to be repeated multiple times. To solve this problem, NiFi introduces the concept of the model. A template is a way of incorporating these basic blocks into larger building blocks. [2]

2 Data flow, Data pipeline and ETL 2-1 Data flow

Data flow - Transfer of data/content from source to destination, data can be in CSV, JSON, XML, HTTP data, image, video clips, telemetry data as in Figure 1, etc. in directed communication Communication, the data stream is around a coherent series of digitally encoded signals (data packets or data packets) used to send or receive information that is in the process of being transmitted [3] A data stream is a set of information extracted from a data provider.

[4] It contains raw data collected from users' browser behavior from websites, in which custom pixels are placed. Data flows are useful for data scientists to run big data and AI algorithms. The main data flow providers are data technology companies.

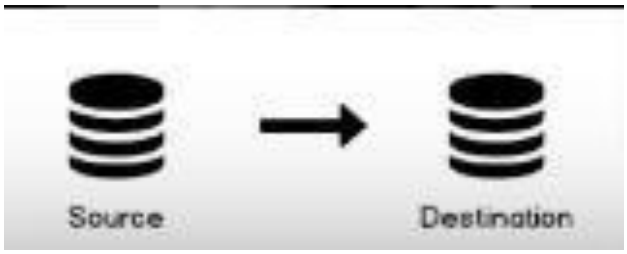

Figure 1 Data flow

2-2 Data Pipeline

The data pipeline is the movement and transformation of data/content from source to destination. Unequivocally, the information pipeline is an evolution of information-processing steps. If the information is not now stacked in the information stage, it will be bound at the beginning of the pipeline. In this stage, there is a progression of the steps as each progress gives a return talking about the contribution to the next stage. This continues until the pipeline is completed. At times, the disposition may be on par with independent developments.

Modern American Journal of Engineering, Technology, and Innovation
ISSN(E): 3067-7939
Volume 2, Issue 5, May, 2026
Website: usajournals.org

Information pipelines consist of three major components: source, stage or processing steps, and target. In some information pipelines, the target may be known as a sink. Information pipelines allow information to leak from the application to the information store, from the information lake to the investigation information base, or to the batch preparation window, for example. Likewise, information pipelines may contain similar sources and inventory, so the pipeline is only suitable for changing information gathering. When time information is prepared between point A and Point B (or focus B, C, and D, there is an information line between these points. As in Figure 2)

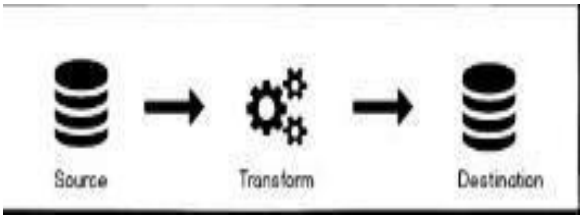

Figure 2 Data Pipeline

2-3 ETL

It is an acronym for extracting, converting, and uploading three capabilities built into one tool to pull out information/source material and locate it in a target.

Focus is the way to see information from the source. In this stage, information is gathered, often from various different types of sources. Change is the path towards changing the information separate from its previous structure to the one in which it should be in order to be well set in the goal. Change occurs by using rules, query tables, or by combining the information with other information. The heap is the way to create information in the target. As in Figure 3 [5]

Figure 3 ETL

Modern American Journal of Engineering, Technology, and Innovation
ISSN(E): 3067-7939
Volume 2, Issue 5, May, 2026
Website: usajournals.org

3 Introduction To Apache NiFi

Apache NiFi is a ground-breaking, simple to utilize, and dependable framework to measure and appropriate information between unique frameworks. It depends on Niagara File's innovation created by NSA and afterward following 8 years gave to the Apache Software Foundation. It is conveyed under Apache License Version 2.0, January 2004. The most recent adaptation for Apache NiFi is 1.9.2

Apache NiFi is a real-time data ingestion platform, which can transfer and manage data transfer between different sources and destination systems. It supports a wide variety of data formats like logs, geolocation data, social feeds, etc. It also supports many protocols like SFTP, HDFS, and KAFKA, etc. This supports a wide variety of data sources and protocols making this platform popular in many IT organizations. [6]

What is Apache NiFi?

An Open Source Data Distribution and Processing System, apache NiFi provides a way to move data from one place to another, making routing decisions and transformations as necessary along the way [7].

Why Use Apache NiFi?

- Easy to use
- Powerful
- Reliable
- Secure
- Scalable

Concepts of Apache NiFi

The concepts of Apache NiFi are as follows: [8]

1- Process Group: It is a group of NiFi flows, which helps a user to manage and keep flows in hierarchical manner. [11]

2- 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.

3-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.

Modern American Journal of Engineering, Technology, and Innovation
ISSN(E): 3067-7939
Volume 2, Issue 5, May, 2026
Website: usajournals.org

4-Flowfile: It is the fundamental use of NiFi, which speaks to the single object of the information picked from the source framework in NiFi. The NiFi processor makes changes to the stream document while it moves from the source processor to the objective. Various occasions like CREATE, CLONE, RECEIVE, and so on are performed on stream documents by various processors in a stream. [9]
5-Event: Events represent the change in flowfile while traversing through a NiFi Flow. These events are tracked in data provenance.
6-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.
[10]

4 Create and process data flow

At this stage of the work, as a first step we create new themes, and then dynamically search the accessibility properties, followed by the advanced theme rules/settings.

Creating a file(send, receive)

Create /home/Ubuntu/data flow/send directory.

Create /home/Ubuntu/data flow/receive/ directory

Add the processors in the User Interface

At this stage of the work, we add the processing elements on the graphical interface and they are prepared according to the work that we do and from these elements are as shown in Figure 4

- Getfile
- UpdateAttribute
- Putfile
- LogAttribute

Modern American Journal of Engineering, Technology, and Innovation
ISSN(E): 3067-7939
Volume 2, Issue 5, May, 2026
Website: usajournals.org

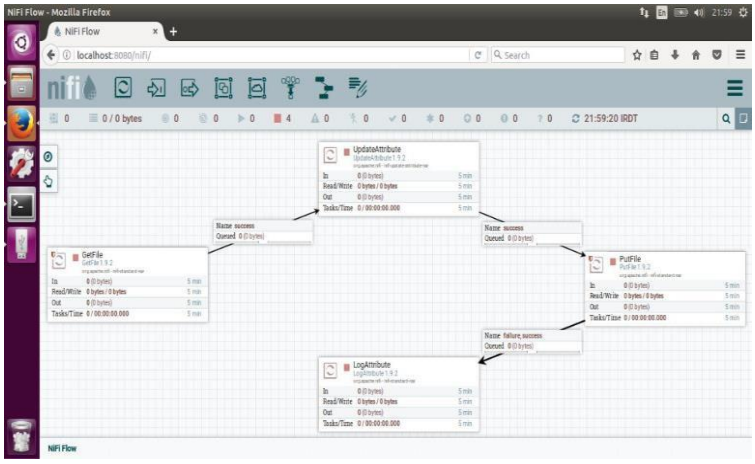

Figure 4 Add the processors

4-2-1 Getfile processors

GetFile offers multiple properties as shown in the image below, the formidable coercive properties such as input directory and file filter to optional properties like path filter and maximum file size. The user can manage the file fetching process using these features, as shown in Figure 5

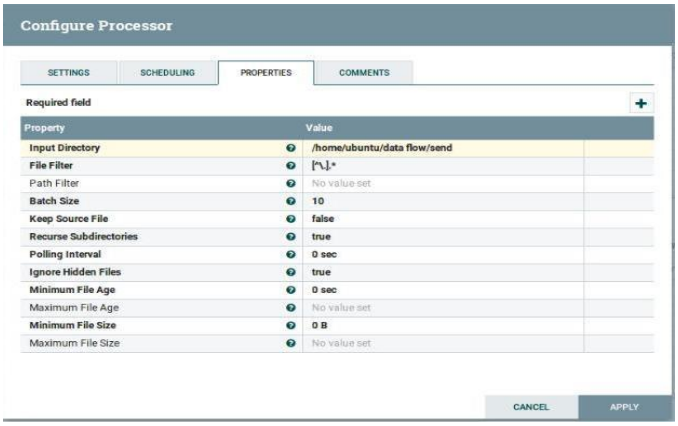

Figure 5 Getfile processors

4-2-2 Update Attribute processors

The feature extraction processors are responsible for extracting, analyzing, and changing the processing of stream file attributes in the NiFi data stream. Some of the processors in this category are Update Attribute, Evaluate JSON Path, Extract Text, Attributes To JSON, etc. As shown in Figure 6

Modern American Journal of Engineering, Technology, and Innovation
ISSN(E): 3067-7939
Volume 2, Issue 5, May, 2026
Website: usajournals.org

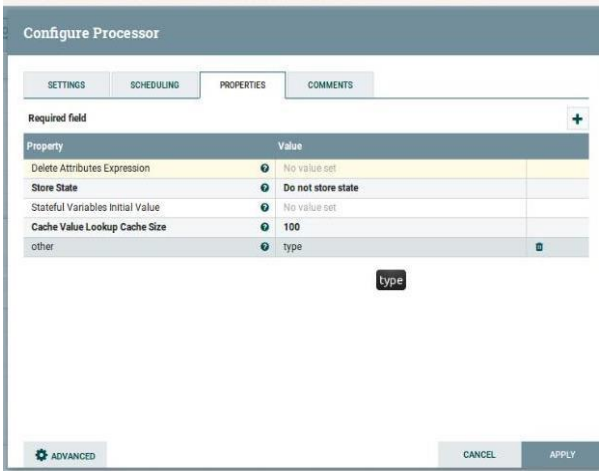

Figure 6 UpdateAttribute processors

4-2-3 Putfile processors

The PutFile Wizard provides properties like a directory to specify the output directory for file transfer purposes and others to manage the transfer as shown in the image below. As shown in Figure 7

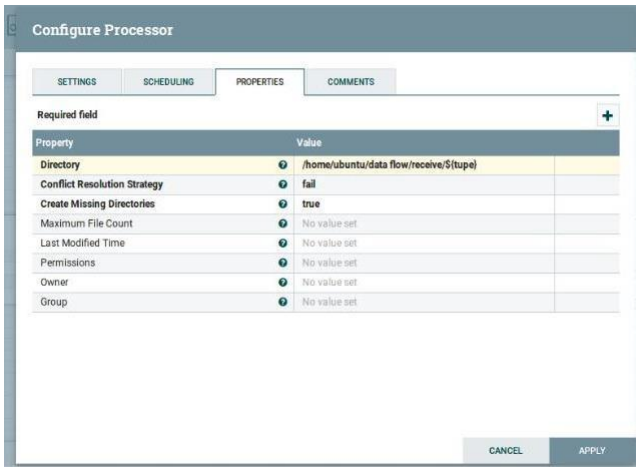

Figure 7 Putfile processors

5 CONCLUSIONS

Through the statement of responsibility, the method of work, and adjusting the elements on the graphic interface, we have added a large group of files of different types and formats in the sending folder, and when we use Information Retrieval

Modern American Journal of Engineering, Technology, and Innovation
ISSN(E): 3067-7939
Volume 2, Issue 5, May, 2026
Website: usajournals.org

techniques in processing this data, we obtainfiles arranged and classified according to their type and format in the receiving folder From this, we can use this method, classify and process large amounts of files and data very quickly.

References

1. W. B. F. a. R. Baeza-Yates, Information Retrieval Data Structures & Algorithms. Prentice-Hall, Inc., 2013.
2. W.-R. L. S.-S. Kim, "A Study on Utilization of Spatial Information in Heterogeneous System Based on Apache NiFi," in in 2019 International Conference on Information and Communication Technology, 2019.
3. A. Ha, "Lotame pitches an 'unstacked' approach to selling data tools," techcrunch.com.
4. A. Ha, "Lotame pitches an 'unstacked' approach to selling data tools," techcrunch.com.
5. S. Zhao, " What is ETL? (Extract, Transform, Load) | Experian. Experian Data Quality," 2018.
6. P. Kostakos, "Privacy preserving sentiment analysis on multiple edge data streams with Apache NiFi," in European Intelligence and Security Informatics Conference (EISIC), 2019.
7. G. Hohpe, "Enterprise Integration Patterns [online]. Retrieved:," in Patterns and Best Practices for Enterprise Integration, 2014.
8. A. Duvander, " The rise of the API economy and consumer-led ecosystems," 2014.
9. Y. Wang,"AIFSP: An adaptive.