A picture containing text, clipart

Description automatically generatedLogo

Description automatically generated

* **NIFI**
* **KAFKA**
* **NIFI REGISTRY**
* **AIRFLOW**
* **SPARK**

**Projet de fin de module**

**Concevoir un Pipeline de Données**

**Proposé par :**

**Mr. Antoine POINTEAU**

**Rédigé par :**

**LOURHMATI RadWane**

**NZWENJA DJIAM Laurent**

Table of Contents

[CHAPITRE 1: NIFI, KAFKA et NIFI REGISTRY 1](#_Toc90674261)

[1.1. NIFI 1](#_Toc90674262)

[1.1.1. Présentation de Apache NIFI 1](#_Toc90674263)

[1.1.2. Composante et fonctionnalités 2](#_Toc90674264)

[1.2. La collecte des différentes sources de données avec Nifi 5](#_Toc90674265)

# NIFI, KAFKA et NIFI REGISTRY

## NIFI

### Présentation de Apache NIFI

Apache NIFI est un autre cadre pour la création de pipelines d’ingénierie de données, et il utilise des DAG. Il a été construit par la National Security Agency et est utilisée dans plusieurs agence fédérales. Apache est plus facile à configurer et est utile pour les nouvelles données ingénieurs. L’interface graphique est excellente et bien que vous puissiez utiliser, Jython, Clojure, Scala ou Groovy pour écrire des processeurs, vous pouvez accomplir beaucoup avec une simple configuration de processeurs.

### Composante et fonctionnalités

Les composantes principales de NIFI sur la JVM sont les suivants :

* Serveur Web qui héberge l’interface graphique,
* Un flow contrôler qui orchestre les opérations. Il fournit des taches à exécuter aux extensions et gère leur ordonnancement,
* FlowFile Repository dans lequel NIFI enregistre l’état d’un FlowFile,
* Content Repository ou les données d’un FlowFile sont stockées,
* Répertoire de la provenance ou toutes les données d’événements de provenance sont stockées.

Graphical user interface, application

Description automatically generated

Figure 1 : Architecture NiFi

En outre, NiFi possède **plus de 200 connecteurs ou processeurs** qui permettent de **collecter en temps réel des données** issues de plus de 80 sources : bases de données, messages, fichiers, flux Twitter, etc.

Graphical user interface, table

Description automatically generated

Figure 2 : Recherche / Ajout d’un processeur

La richesse de la bibliothèque des processeurs permet de réaliser :

* La data Transformation
* Les accès aux bases de données
* La Data Ingestion
* La Data Egress/Sending Data

Apache NiFi s’intègre avec différents outils de streaming et sa combinaison avec Apache Kafka est très robuste.

NiFi peut jouer le rôle du producteur pour Kafka.

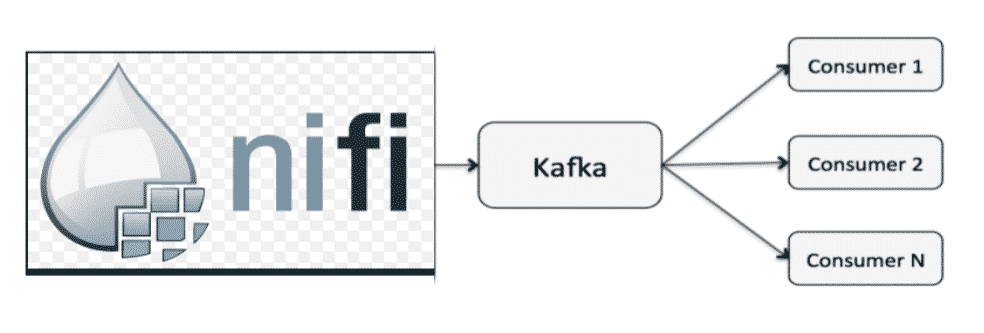


Figure 3 : **Apache NiFi en producteur de données pour Kafka**

* + 1. Apache NiFi en producteur de donnes pour KafKa

Diagram

Description automatically generated

Figure 4 : Apache NiFi en consommateur de données

**De nombreux cas d’usage** sont réalisables avec ces deux superbes outils, comme **construire un CDC** (Change Data Capture), **faire du streaming** avec Kafka ou encore **récupérer des logs de serveurs webs**, etc.

Diagram

Description automatically generated Figure 5 : Préparer les données pour les indexer dans Apache Solr

## La collecte des différentes sources de données avec Nifi

1. Configure Processor

Table

Description automatically generated

Figure 7 : la data qu’on a utlisé pour travailler le projet

A screenshot of a computer

Description automatically generated

Figure 6 : GetFile Processor pour recuréper la data du systéme fichier

1. Installation de KafKa

Text

Description automatically generated

Figure 8 : Kafka

Graphical user interface, application

Description automatically generated

Figure9 : Configuration de kafka en faisant la liason avec le topic examenspark

Graphical user interface, application

Description automatically generated

Figure 10 : L’ajout de csv writer à kafka

A screenshot of a computer

Description automatically generated with medium confidence

Figure 11 : csv writer ajouté

Text

Description automatically generated

Figure 12 : le lancement de kafka et zookeeper

Diagram

Description automatically generated with medium confidence

Figure 13 : data pipeline pour publier la data en kafka

Graphical user interface, text, application, email

Description automatically generated

Figure 14: Nifi Registry

En fait ici on a essayé de lier nifi avec nifi registry mais on n’a pas réussi de faire ça

Graphical user interface, text

Description automatically generated

Figure 15 : fichier de configuration nifi registry

# APACHE SPARK

## Presentation de Apache Spark

1. Définition

Apache Spark est une infrastructure de traitement de données pouvant effectuer rapidement des taches de traitement sur des grands ensembles de données.

1. Composants de Spark

Spark dans son ensemble se compose de diverses bibliothèques, API, bases de données, etc,

Ses principaux composants sont les suivants :

* Spark Core

Spark Core est le bloc de base, qui comprend tous les composants pour la planification des taches.

* Spark SQL

Apache Spark fonctionne avec les données non structurées à l’aide de son outil <<aller a >>, Spark SQL. Ce dernier permet d’interroger des données via SQL

* Spark Streaming

Cet outil traite les flux de données en direct. Des exemples de ces données incluent les fichiers journaux, les messages contenants des mises a jour de statut publiées par les utilisateurs, etc.

* Graphx
* MLlib

Apache Spark propose une bibliothèque contenant des services communs de Machine Leaning(ML) appelés MLlib.

1. Préparation des données

Graphical user interface, text, application

Description automatically generated

Figure 16 : Fichier pyspark pour faire le traitement et le nettoiyage de la data maires-candidates

Graphical user interface, text, application

Description automatically generated

Figure 17 : Le Fichier Dag qu’on a utilié pour nettoyer la data

CHAPITRE 3 AIRflOW

1. Presentation de Airflwo

Airflow est une plateforme créée par la communauté pour programmer la création, la programmation et le suivi des flux de travail. Le « machine learning » est un sujet d’actualité dans l’industrie. Ce ne sera pas aussi cool si ce n’est pour le traitement des données qu’il implique

Airflow est un outil d’orchestration de flux de travail ETL (Extraction, Transformation, Chargement), utilisé dans les pipelines de transformation de données.

1. Orchestration et automatisation des phases de collecte et de préparation des données

Graphical user interface, text, application, email

Description automatically generated

Figure 18 : execusion de tache de fichier dat

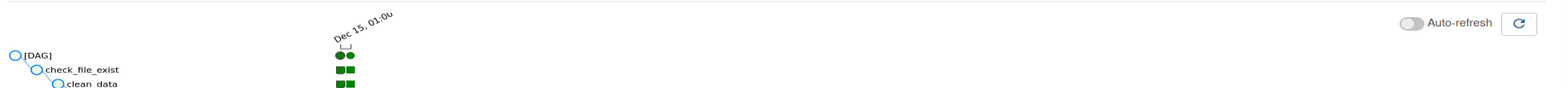


Figure 19 : execusion de tache de fichier dat

LE fichier dag contient 2 taches la première c’est tache c’est check file existe

C’est à dit on vérifier pour voir si le fichier dag existe ou pas

Et la deuxième tache c’est clean data. Ici on fait les opérations du traitement et du nettoyage de la data

Graphical user interface, text, application

Description automatically generated

Figure 20 : Vérification de l’existence du fichier Dag

Graphical user interface, text, application

Description automatically generated

Figure 21 : Les opérations du traitement et le traitement de la Data