ANGOSTO Estelle ITS2

MASLIAH Anita

NINH Thi-Loa-Lan

Rapport de projet

Big Data

Année scolaire 2021-2022 Mr Mohamed Souidi

Table des matières

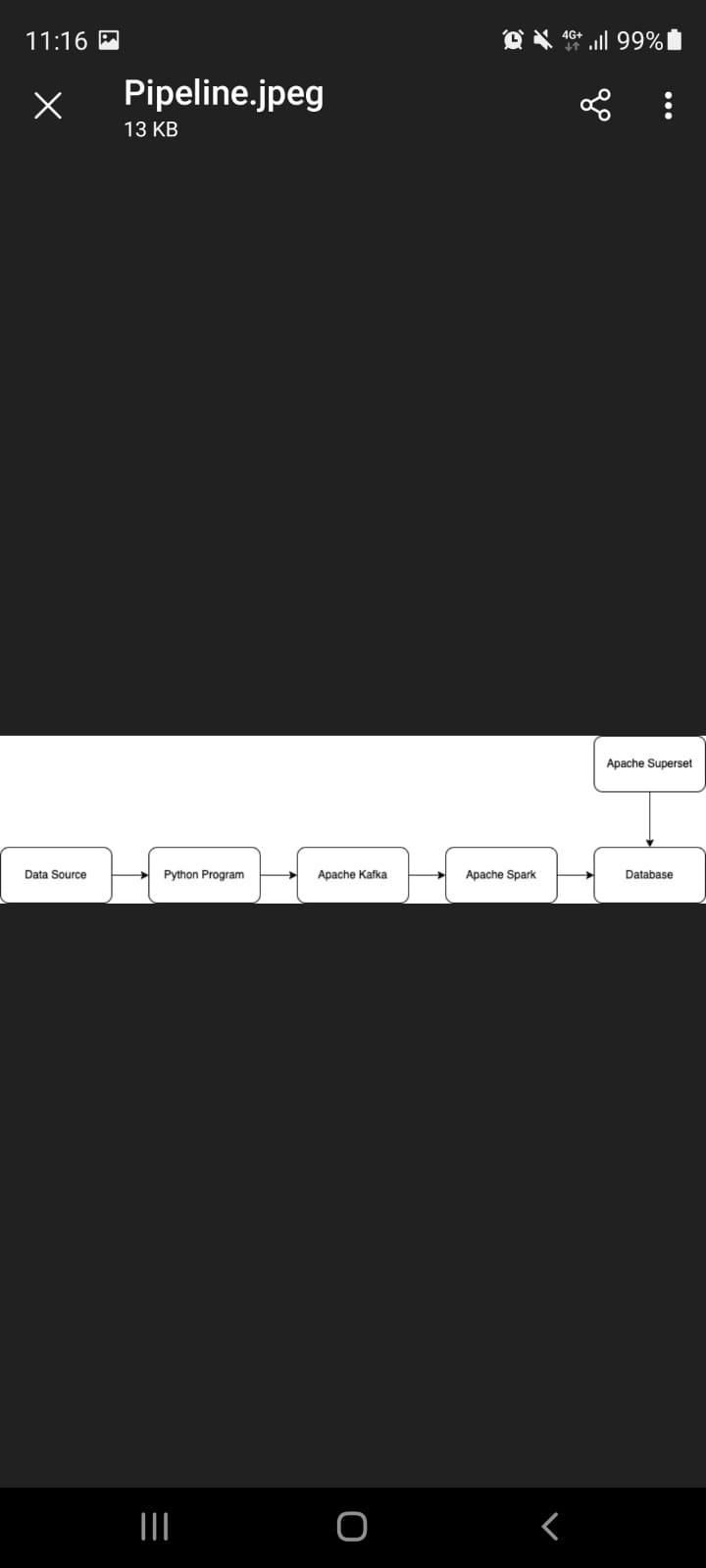
[I) Data source 3](#_Toc92370588)

[II) Structure du Pipeline de traitement 5](#_Toc92370589)

[III) Problèmes rencontrés 7](#_Toc92370590)

Introduction

Dans le cadre du module de Data Mining, nous devons réaliser un pipeline de traitement de données en utilisant les Spark, Kafka, Superset et des scripts en Python.



# Data source

Nous allons procéder à la récupération de données sur twitter.

Installation de la librairie tweepy :

**pip install tweepy**

Creation d’un stream listener

Une image contenant texte

Description générée automatiquement

Nous avons rencontré des problèmes avec l’utilisation de MyStreamListener. En effet ce code est disponible sur le site de la documentation de tweepy seulement une mise à jour a été effectué depuis octobre 2021 et la documentation n’est toujours pas à jours. Nous avons essayé d’utiliser ce qui semble être la nouvelle version de MyStreamListener à savoir Stream. Seulement des erreurs continuent d’être affiché par la console python. D’autre recherches sur des forums et/ou l’aide de camarade nous aideront sûrement à régler ce problème. Une autre solution est de télécharger une ancienne version de tweepy.

Creation d’un stream

Authentification à l’API tweeter

On va créer une application tweeter. On créer une nouvelle adresses mail pour le projet et un compte twitter.

Une image contenant texte

Description générée automatiquement

Ensuite on se connecte sur Twitter développer pour créer une application :

Une image contenant texte, capture d’écran, téléphone mobile

Description générée automatiquement

On récupère les API key, API Key Secret.

Ensuite on fait une demande d’Acces Token et d’Access Token Secret puis on récupère les clés. Ils nous serviront dans le code pour obtenir les tweets. En résumé on à 4 clés qui sont les suivantes :

Une image contenant table

Description générée automatiquement

4bkk45ZsZ6l0aM0br2Lg6ephm

JAoT2LCL2H21waPCKVg0htoprEYQC7FuOq06Em6kpCwTd1JYTH

1478316578592673798-Z8enGtaiTzACI1vmYvFr6KhXImbhIH

96aR8JbRl9HvWlJDp3V8hZESaoN5kI5mSNC6KXRnxYSNq

Voici l’authentification

Une image contenant texte

Description générée automatiquement

Récupération de tweet

Ensuite nous avons voulu tester la récupération des tweets avec l’exemple hello tweepy suivant.

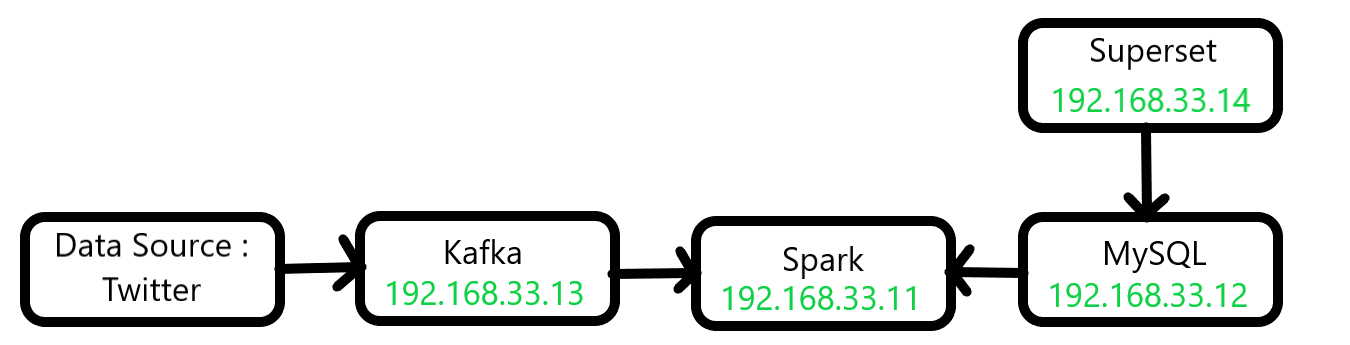
Une image contenant texte

Description générée automatiquement

Seulement une version d’accès « Elevated » est demandé pour accéder à ces fonctions. Nous avons donc effectué une demande auprès de Twitter et attendons encore la réponse.

# Structure du Pipeline de traitement

Notre projet prend la structure suivante :



Nous avons une source de donnée, ici Twitter, qui est relié à une machine virtuelle sur laquelle se trouve un serveur Kafka. La configuration de notre source de données a été présenté dans la partie précédente. Dans cette partie nous allons plutôt parler de la manière dont nous avons configuré les quatre machines virtuelles nécessaires à ce projet.

Machine virtuelle Kafka

Pour la configuration de cette machine virtuelle, nous avons utilisé les fichiers de configuration fournis. Ces fichiers sont un fichier Vagrantfile qui permet de configurer le système d’exploitation et l’adresse IP de la machine et deux autres fichiers permettant d’installer et de lancer les serveurs Kafka et Zookeper. Nous avons testé le bon fonctionnement de Kafka avec les commandes suivantes :

#Création d’un topic

bin/kafka-topics.sh --create --topic test --bootstrap-server 192.168.33.13:9092

#Création d’un producer

bin/kafka-console-producer.sh --topic test --bootstrap-server 192.168.33.13:9092

#Dans un autre terminal, on crée un consumer

bin/kafka-console-consumer.sh --topic test --from-beginning --bootstrap-server 192.168.33.13:9092

Machine virtuelle MySQL

Pour créer cette machine virtuelle, nous n’utilisons que le fichier Vagrantfile afin de la configurer comme une machine Ubuntu avec deux processeurs et l’adresse IP 192.168.33.12.

Nous avons ensuite installé un serveur MySQL avec les commandes suivantes :

#On réalise les mises à jour de la VM.

sudo apt update

#On installe le package MySQL

sudo apt install mysql-server

#On exécute cette commande afin de configurer le serveur

sudo mysql\_secure\_installation

Une fois le serveur installé et paramétré, nous avons créé un utilisateur pour notre base de données :

#Création de notre utilisateur

CREATE USER 'eal'@'localhost' IDENTIFIED BY 'password';

#Ajout des privilèges

GRANT ALL PRIVILEGES ON \*.\* TO ‘eal’@’localhost’ WITH GRANT OPTION;

Nous avons ensuite testé le bon fonctionnement de notre serveur MySQL :

#On regarde l’état de notre service

systemctl status mysql.service

#On essaye de réaliser des commandes avec mysqladmin qui fonctionne comme un client.

sudo mysqladmin -p -u root version

Machine virtuelle MySQL

La création de cette machine virtuelle se fait de la même façon que pour la machine virtuelle précédente. Nous lui attribuons l’adresse IP suivante : 192.168.33.14.

Nous allons commencer par l’installation des dépendances de Superset et notamment de Python 3.

sudo apt update

sudo apt upgrade

sudo apt-get install build-essential libssl-dev python3-dev python3-pip libsasl2-dev

On doit ensuite créer un environnement virtuel Python :

#Installation de l’environnement virtuel Python

sudo apt-get install python3-venv

#Création de l’environnement

python3 -m venv superset-env

#Activation de l’environnement

Source superset-env/bin/activate

On procède ensuite à l’installation et au paramétrage de Superset :

pip install apache-superset

#On initialize la base de données

superset db upgrade

#Création par défaut des rôles et des permissions

superset init

#Création d’un super utilisateur

superset fab create-admin

On teste notre serveur Superset :

#On charge les exemples de Superset

superset load\_examples

#Lancement du serveur sur le port 8080

superset run -p 8080 -h 192.168.33.14

# Problèmes rencontrés

Vagrant ne fonctionnait plus sur la machine de Lan avec le passage à Windows 11.

La demande d’accès au niveau élevé de Twitter prend du temps et nous nous sentions bloqués par le fait que nous n’avions pas de source de données.