Rapport – Data Mining

Enseignant: SOUIDI Mohamed

FERREIRA Alicia – **KHALIFA Marina** – **NEJMI Manal**

ITS 2

Année 2021-2022

Introduction

Ce projet consiste à développer une application Python qui consiste à un tableau de bord interactifs en temps réel sur les flux de données et celles-ci seront visualisées sur Superset.

Host est notre machine Windows physique qui héberge différentes machines virtuelles. Chaque logiciel possède une machine virtuelle, donc en tout nous aurons 4 machines virtuelles.

Dans ce projet, nous allons réaliser différentes étapes :

- Nous allons dans un premier temps récupérer les données que nous souhaitons traiter
- Nous allons écrire un programme python qui récupère les données et qui les envoie sur Kafka
- Spark récupère les données sur Kafka pour les traiter puis les sauvegarde sur MySQL
- MySQL
- Superset se connecte à MySQL pour visualiser les données

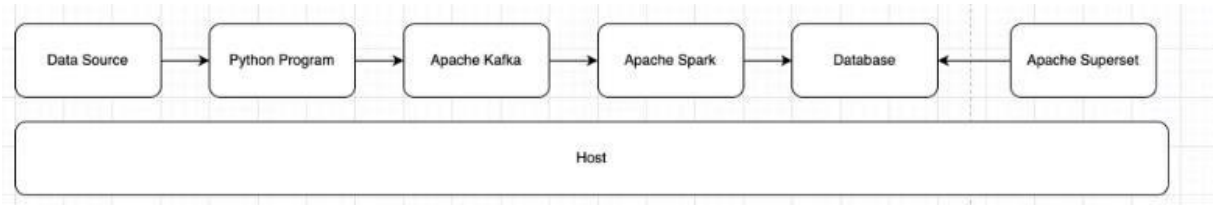


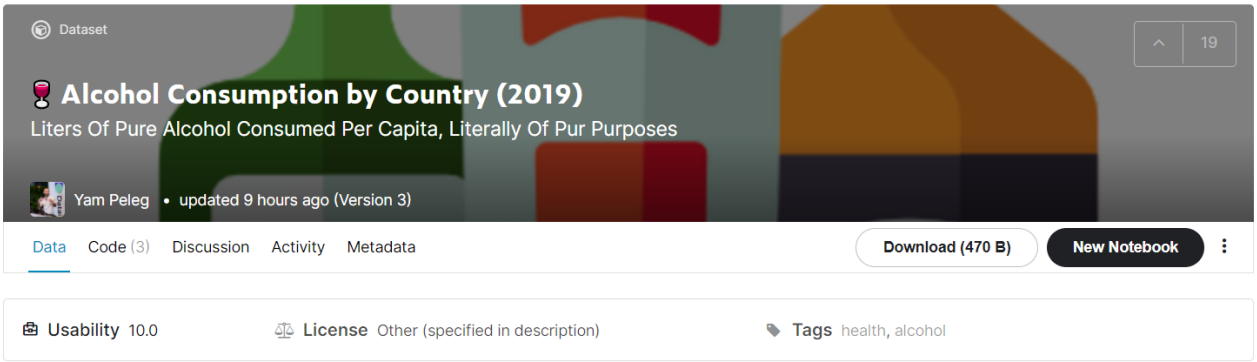
Figure 1 : Use Case du projet

Nous allons attribuer une adresse IP à chaque machine virtuelle afin que celles-ci puissent communiquer entre elles pour visualiser les données.

Data source

Nous avons prévu de traiter des données avec un fichier csv. Nous allons donc récupérer des données sur le site Kaggle sur Alcohol. Ce fichier comporte des données sur le litre d'alcool au total consommé par pays. De plus, les pays sont classés selon un rang : on peut constater que la Biélorussie est en première position avec une consommation d'alcool d'environ 17,5 contrairement à l'Angleterre.

Nous avons commencé à créer notre code python qui permet de récupérer les données et les streamers. Ce code sera similaire à celui de wordcount.py vu en classe.



Configuration des machines

Nous avons configuré 4 machines virtuelles pour chaque logiciel : Spark, Kafka, MySQL et Superset.

En effet, nous avons configuré une adresse IP différente pour chaque VM via un Vagrantfile.

Kafka

Kafka est une plateforme de streaming d'évènements qui capture des données en temps réel tel que des bases de données.

Kafka est un bus qui permet :

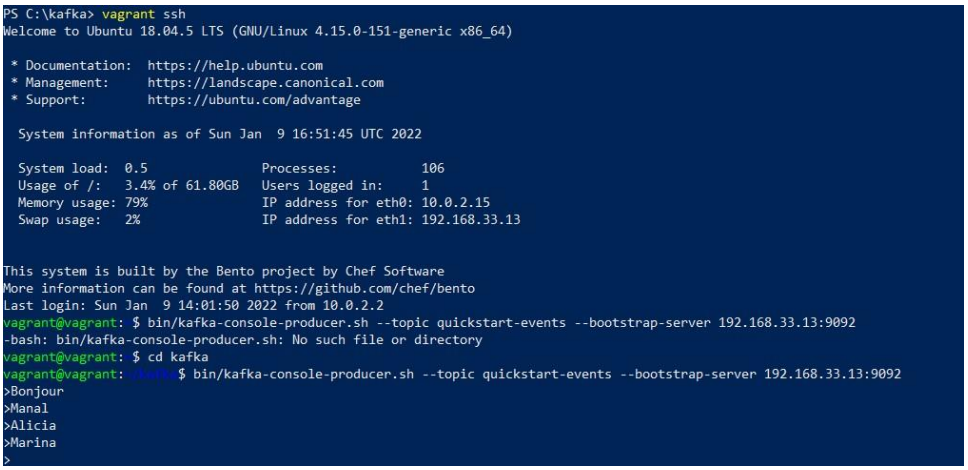
- Publier et s'abonner à des flux d'évènements en comprenant l'importation et l'exportation continue de nos données à partir d'autres systèmes
- Stocker les flux d'évènements de manière durable et fiable
- Traiter les flux d'évènements au fur et à mesure qu'ils se produisent

Kafka est reconnu comme Producer c'est-à-dire que c'est un client qui écrit des évènements.

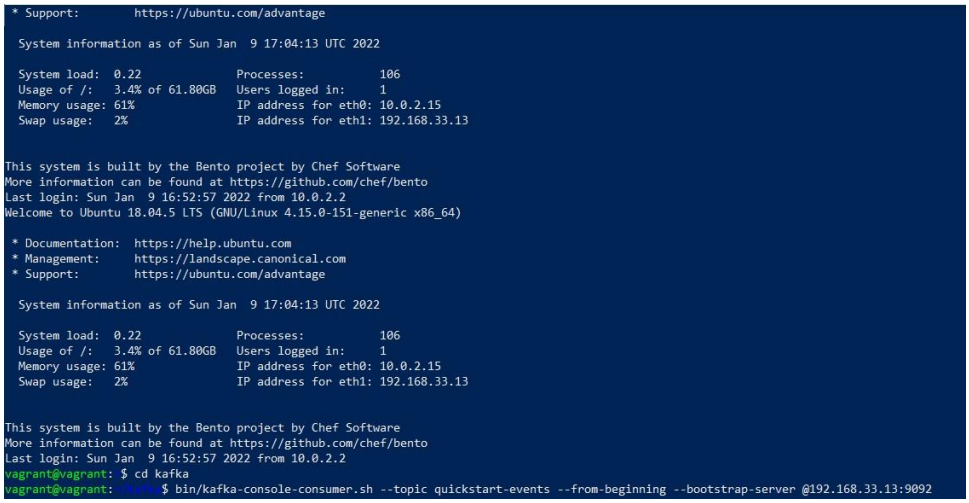
Nous avons attribué à Kafka l'adresse IP suivante : 192.168.33.13, puis nous avons installé et lancé les deux serveurs zookeeper et Kafka, finalement on a créé un topic avec les commandes vu durant le cours.

Au niveau de cette partie, nous avons envoyé des évènements (messages) via un terminal qui est le Producer et nous avons pu les visualiser via à un autre terminal qui agit en tant que consumer.

Producer sur Kafka



Consumer sur Kafka



Spark

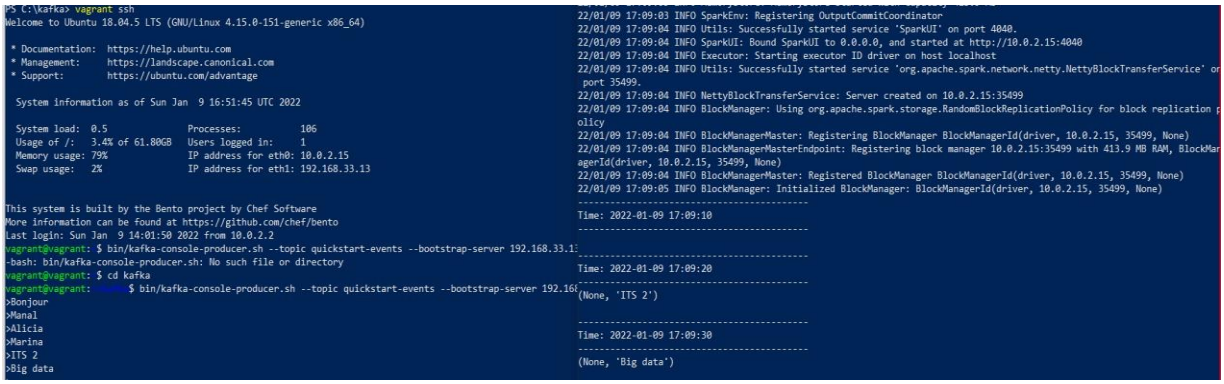
Spark est un Framework open source qui permet de traiter de grande quantité de données puis il récupère les informations dont on a besoin. La particularité de Spark est que les données s'exécutent en mémoire. Il a une autre fonction : c'est un consumer.

Spark streaming permet de décorréler un producer à un consumer. Le consumer a pour fonction de lire et de traiter ces évènements écrient par Kafka.

Nous avons attribué à Spark l'adresse IP suivante : 192.168.33.12

Pipeline spark avec kafka : Streaming en temps réel

Au niveau de cette partie, nous avons relié kafka avec spark. Le consumer (kafka) et le producer (spark) avec la configuration de l'adresse ip 193.168.33.13.



Mysql

MySQL est un système de gestion de bases de données relationnelles SQL (un langage informatique normalisé servant à exploiter des bases de données relationnelles).

Avec une base de données relationnelle, les données sont divisées en plusieurs zones de stockage séparées appelées tables, plutôt que de tout regrouper dans une seule grande unité de stockage. En utilisant ce qu'on appelle une « clé », on peut lier les données de ces deux tables entre elles afin de pouvoir les manipuler et les combiner dans différentes tables si nécessaire. Il est important de noter qu'une clé n'est pas le nom du client. Au lieu de cela, on utilise un nom unique, comme un numéro d'identification numérique.

Nous avons attribué à Spark l'adresse IP suivante : 192.168.33.14

```
vagrant@vagrant: $ sudo mysql
Welcome to the MySQL monitor.  Commands end with ; or \g.
Your MySQL connection id is 2
Server version: 5.7.36-0ubuntu0.18.04.1 (Ubuntu)

Copyright (c) 2000, 2021, Oracle and/or its affiliates.

Oracle is a registered trademark of Oracle Corporation and/or its
affiliates. Other names may be trademarks of their respective
owners.

Type 'help;' or '\h' for help. Type '\c' to clear the current input statement.

mysql>
mysql>
mysql>
mysql>
mysql> ls
    ->
    -> SELECT user,authentication_string,plugin,host FROM mysql.user;
SELECT user,authentication_string,plugin,host FROM mysql.user;
^C
mysql> SELECT user,authentication_string,plugin,host FROM mysql.user;
+------------------+-------------------------------------------+-----------------------+-----------+
| user             | authentication_string                     | plugin                | host      |
+------------------+-------------------------------------------+-----------------------+-----------+
| root             |                                           | auth_socket           | localhost |
| mysql.session    | *THISISNOTAVALIDPASSWORDTHATCANBEUSEDHERE | mysql_native_password | localhost |
| mysql.sys        | *THISISNOTAVALIDPASSWORDTHATCANBEUSEDHERE | mysql_native_password | localhost |
| debian-sys-maint | *3F7CD2E03C9E68782E5D59A643D5FDB6DC72BE4F | mysql_native_password | localhost |
+------------------+-------------------------------------------+-----------------------+-----------+
4 rows in set (0.00 sec)
```

Superset

Superset est un logiciel open source de datavisualisation et d'exploration des données. Ce logiciel permet de traiter des données massives.

Nous avons attribué à Spark l'adresse IP suivante : 192.168.33.10

```
PS C:\superset> vagrant ssh
Welcome to Ubuntu 18.04.5 LTS (GNU/Linux 4.15.0-151-generic x86_64)

 * Documentation:  https://help.ubuntu.com
 * Management:     https://landscape.canonical.com
 * Support:        https://ubuntu.com/advantage

  System information as of Sun Jan  9 17:13:38 UTC 2022

  System load:  0.0               Processes:           100
  Usage of /:   4.9% of 61.80GB   Users logged in:     0
  Memory usage: 16%               IP address for eth0: 10.0.2.15
  Swap usage:   0%                IP address for eth1: 192.168.33.10


This system is built by the Bento project by Chef Software
More information can be found at https://github.com/chef/bento
Last login: Sun Jan  9 16:42:54 2022 from 10.0.2.2
vagrant@vagrant: $
```

Réalisation du projet

En s'appuyant sur le fichier spark-streaming-kafka.py, nous avons affiché sur le consumer spark le contenu de notre fichier.csv.

Fichier spark-streaming-kafka.py

```python
# -*- coding: utf-8 -*-
"""
Created on Sun Feb  6 16:23:52 2022

@author: manal
"""

import json
from kafka import KafkaProducer
import time
#On se connecte à la machine Kafka
producer = KafkaProducer(bootstrap_servers='192.168.33.13:9092', value_serializer=lambda v: json.dumps(v).encode('utf-8'))#
with open('Alcohol.csv', 'r') as f:

    listusers= f.readlines()
    for i in listusers:
        user=i.strip()
        print(user)
        producer.send('quickstart-events', user)
        time.sleep(10)
```

Pour les transformations, le résultat affiche le rang, les capitales, et le litre d'alcool consommé par pays.

Voici les résultats :

Données affichées sur kafka

```
"Rank,Country,Liters of pure alcohol consumed per capita"
"1,Belarus,17.5"
"2,Moldova,16.8"
"3,Lithuania,15.4"
"4,Russia,15.1"
"5,Romania,14.4"
"6,Ukraine,13.9"
"7,Andorra,13.8"
"8,Hungary,13.3"
"9,Czech Republic,13.0"
"10,Slovakia,13.0"
"11,Portugal,12.9"
"12,Serbia,12.6"
"13,Grenada,12.5"
"14,Poland,12.5"
"15,Latvia,12.3"
"16,Finland,12.3"
"17,South Korea,12.3"
"18,France,12.2"
"19,Australia,12.2"
"20,Croatia,12.2"
"21,Ireland,11.9"
"22,Luxembourg,11.9"
"23,Germany,11.8"
"24,Slovenia,11.6"
"25,United Kingdom,11.6"
```

Données streamer en temps réelles sur spark

```
vagrant@vagrant: ~
(None, '"1,Belarus,17.5"')

-------------------------------------------
Time: 2022-02-08 23:46:20
-------------------------------------------
(None, '"2,Moldova,16.8"')

-------------------------------------------
Time: 2022-02-08 23:46:30
-------------------------------------------
(None, '"3,Lithuania,15.4"')

-------------------------------------------
Time: 2022-02-08 23:46:40
-------------------------------------------
(None, '"4,Russia,15.1"')

-------------------------------------------
Time: 2022-02-08 23:46:50
-------------------------------------------
(None, '"5,Romania,14.4"')

-------------------------------------------
Time: 2022-02-08 23:47:00
-------------------------------------------
(None, '"6,Ukraine,13.9"')

-------------------------------------------
Time: 2022-02-08 23:47:10
-------------------------------------------
(None, '"7,Andorra,13.8"')

-------------------------------------------
Time: 2022-02-08 23:47:20
-------------------------------------------
(None, '"8,Hungary,13.3"')

-------------------------------------------
Time: 2022-02-08 23:47:30
-------------------------------------------
(None, '"9,Czech Republic,13.0"')

-------------------------------------------
Time: 2022-02-08 23:47:40
-------------------------------------------
(None, '"10,Slovakia,13.0"')

-------------------------------------------
Time: 2022-02-08 23:47:50
-------------------------------------------
(None, '"11,Portugal,12.9"')

-------------------------------------------
Time: 2022-02-08 23:48:00
-------------------------------------------
```

Problèmes rencontrés

Afin de finir ce projet, il nous reste quelques étapes à réaliser : le traitement de notre base de données sur MySQL et puis la partie visualisation sur superset.