Université Abdelmalek Essaadi
Faculté ses Sciences et techniques de Tanger
Département Génie Informatique

A présenter : Lundi 20 Novembre 2023 par trinôme.

Mini Projet : Prédiction de désabonnement des clients en temps réel

Il s'agit de développer une application Web basée sur l'API **Apache KAFKA Stream**, permettant de faire l'analyse des données, dans le but de « **prédire en temps réel le désabonnement des clients d'une entreprise** »

Afin comprendre le fonctionnement de Kafka, il faut d'abord se familiariser avec le vocabulaire suivant :

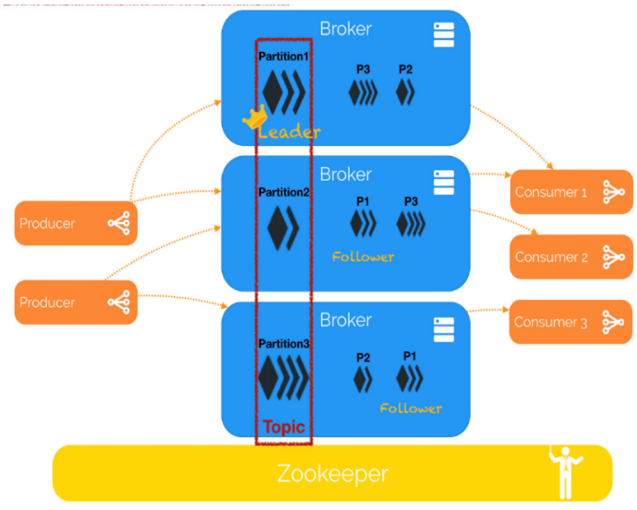
Topic: Un flux de messages appartenant à une catégorie particulière. Les données sont stockées dans des topics.

Partitions: Chaque topic est divisé en partitions. Pour chaque topic, Kafka conserve un minimum d'une partition. Chaque partition contient des messages dans une séquence ordonnée immuable. Une partition est implémentée comme un ensemble de segments de tailles égales.

Brokers: Les *brokers* (ou courtiers) sont de simples systèmes responsables de maintenir les données publiées. Chaque courtier peut avoir zéro ou plusieurs partitions par topic.

Producers: Les producteurs sont les éditeurs de messages à un ou plusieurs topics Kafka. Ils envoient des données aux courtiers Kafka. Chaque fois qu'un producteur publie un message à un courtier, ce dernier rattache le message au dernier sègment, ajouté ainsi à une partition. Un producteur peut également envoyer un message à une partition particulière.

Consumers: Les consommateurs lisent les données à partir des brokers. Ils souscrivent à un ou plusieurs topics, et consomment les messages publiés en extrayant les données à partir des brokers.



Architecture de Apache Kafka Stream

Université Abdelmalek Essaadi
Faculté ses Sciences et techniques de Tanger
Département Génie Informatique

A présenter : Lundi 20 Novembre 2023 par trinôme.

Démarche à suivre :

1. Lancer puis lire en temps réel les données du fichier **customer_churn.csv** en utilisant **Apache KAFKA Streams**
2. Faire des prétraitements nécessaires sur les données du fichier en utilisant les libraires **(Sklearn, Pyspark Mlib ou Pytorch ….)**
3. En utilisant des modèles Machine Learning supervisés (au moins 3 modèles), faire un entrainement sur la base d'apprentissage **customer_churn.csv**.
4. Enregistrer le meilleur modèle sous format **.pkl**
5. En utilisant le modèle préparé, entraîné et sauvegardé ; prédire en **temps réel** si le client quittera l'institution ou non sur les données de test : **new_customers.csv**
6. Présenter les résultats sous forme d'un tableau de bord d'une application Web
7. Uploader le projet sur le réseau **GitHub**

Outils de travail :

- **Librairies :** Apache Kafka Stream, PySpark Mlib, Sklearn, Pytorch, Pandas, Matplotlib .
- **Frameworks :** Flask, Django
- **Langages :** Python, Java, Java Script …
- **Editeurs :** IntelliJ IDEA, Eclipse, VsCode
- **Systèmes d'exploitation :** Unix, MacOS ou Windows …

Description de données :

Name : Name of the latest contact at Company

Age: Customer Age

Total_Purchase: Total Ads Purchased

Account_Manager: Binary 0=No manager, 1= Account manager assigned

Years: Totaly Years as a customer

Num_sites: Number of websites that use the service.

Onboard_date: Date that the name of the latest contact was onboarded

Location: Client HQ Address

Company: Name of Client Company

Churn : Target (label)

Source de données : https://github.com/Shantanu-Gupta-au16/Spark-Mini-Projects/blob/master/Customer%20Churn%20using%20Spark.ipynb