21. Quand la France a-t-elle commencé à investir ?

TP Final : Utiliser spark streaming pour traiter des données d'une API

Nous allons streamer l'API Mastodon et traiter les threads d'un sujet donné: #IA.

Pour celà vous devrez :

- créer un programme `produce_stream_thread.py` pour produire des messages kafka correspondant à un sujet donné (#IA).
- créer un programme spark qui lit dans le topic spark et est capable de donner les informations suivantes sur les threads :
  - Quel est le nombre de threads produits par intervalle de 6h sur le hashtag #IA avec un chevauchement de ce nombre sur 30 minutes ?
- Pour la sortie:
  - reprendre les colonnes : date de début de la fenêtre, date de fin et compteur de status (threads)
  - utilisez un fichier csv
  - utiliser un mode append

Bonus :

- Comment pourriez-vous faire pour n'avoir qu'un seul fichier ?
- Peut-on trier les résultats ?

Voici quelques aides :

- La documentation de l'API Mastodon est disponible ici : https://docs.joinmastodon.org/
- Pour le getting started : https://docs.joinmastodon.org/client/intro/
- Nous utiliserons l'API publique des timelines : https://docs.joinmastodon.org/methods/timelines/#tag
- N'oubliez pas de démarrer un cluster Kafka avec un topic permettant d'écrire et lire des messages
- Spark Structured Streaming permet de se connecter à topic Kafka : https://spark.apache.org/docs/latest/structured-streaming-kafka-integration.html