**Mini Projet**

Le dossier Mini-projet contient :

* Deux dataset (format csv)

Donos.csv : c’est une dataset américaine qui collecte les donateurs scolaire chaque donateur doit avoir un ID, la ville, le département où il habite, Est-ce qu’un professeur ou non et donors Zip)

Donations.csv : Contient les donations, le montant qui a été donnée dans cette donation, la date de donation…

* Le Jupyter-notebook à rendre

**Le but de ce Mini projet** :

1. Savoir comment configurer apache Spark sur votre machine virtuelle ou sur votre Windows (en suivant les étapes que nous avons abordées dans la séance 5 « configuration de Spark sur votre Windows »)
2. Savoir comment traiter une base de données volumineuse à l’aide de Spark

**Travaille à faire :**

Par l’utilisation de **l’Apache Spark** on vous demande de traiter les deux bases de données **Donars** et **Donations** format csv en répondant aux questions suivantes :

**Infrastructure**

1. Créer une session pyspark, c’est le tunnel de communication avec le cluster
2. Importer le dataset Donors.csv

**Affichage**

1. Afficher les 20 premières lignes de dataset Donors (Utilisez la fonction show())
2. Convertir le dataset en dataframe pandas (utilisez la fonction "toPandas()")
3. Trouver le nombre nul dans chaque colonne
4. Imprimer le schéma de dataset (pour imprimer le schéma, on utilise la fonction "printSchema")

**Filtrage**

1. Laissez que les enregistrements dont Donor City commence par A
2. Affichez les résultats on utilisant la fonction show()

**Transformation**

1. Construisez une nouvelle colonne Address en faisant une concaténation Donor\_City, Donor\_State, Donor\_Zip
2. Afficher les résultats

**Moteur SQL**

1. Persister le dataset de départ comme une Temporary View
2. Comptez le nombre de professeurs ayant participé à la donation
3. Afficher que les id des donateurs qui habite à California
4. Ingestion des données et publication en temporary view du fichier Donations.csv
5. Afficher le contenu de la base de données Donations.csv
6. Calculer le montant minimum, le montant maximum, le montant moyen en arrondissant à l'unité après la virgule de la colonne Donation\_Amount
7. Faites une jointure Entre le dataset des donneurs (Donors.csv), et le dataset des Donations (Donations.csv)
8. Calculez la somme de l'argent donnée par Les professeurs (Donor Is Teacher=Yes) et les non-professeurs utilisant seulement SQL

Vous avez le choix de travailler avec Pyspark sur Jupiter-Notebook ou bien sur Zeppelin Notebook.  
Si vous avez des questions/remarques merci de me contacter par mail [menyaniismail@gmail.com](mailto:menyaniismail@gmail.com)

Livrables :

* Notebook Jupyter ou Notebook Zeppelin avec le code et les résultats.
* Notebook jupyter ou Notebook Zeppelin format PDF (Nom\_Prenom.pdf)