# TP3 : Spark data frames

## Objectifs du TP :

Le but de ce TP est d’analyser des data sets en utilisant la notion des Data frames du Spark.

Sources utiles : <https://spark.apache.org/docs/latest/rdd-programming-guide.html>

## Exercice 1 (Exercice 2 du TP2) :

La police de Chicago compile des informations sur tous les crimes signalés chaque année. Vous êtes invité à analyser le jeu de données "crimes" disponible au format CSV en utilisant Spark Data frames.

Ce data set contient tous les cas signalés d’un crime dans la ville de Chicago du 1er janvier 2014 au 24 octobre 2016.

Au 1er novembre 2016, il y a eu 605 meurtres et 3003 fusillades depuis le début de l’année, dont 78 et 353, respectivement, en octobre.

Les données sont structurées comme suit :

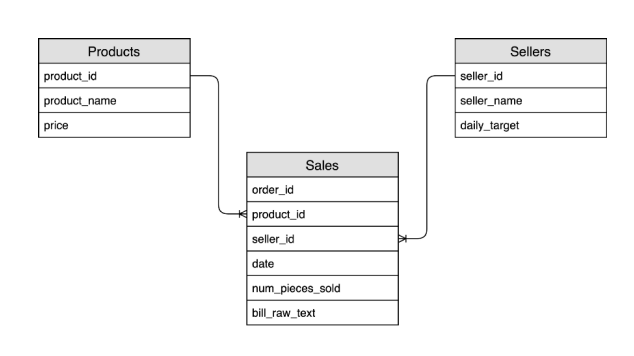
*ID,Case\_Number,Date,Block,IUCR,Primar\_yType,Description,Location,Arrest,Domestic,Beat,District,Ward,Community\_Area,FBI\_Code,Latitude,Longitude.*

Objectifs :

1. Déterminer les cinq principaux types de crimes ayant le plus grand nombre d’affaires dans l’ensemble.
2. Analyser les tendances des différents types de crimes au fil des mois.
3. Trouvez les trois principaux endroits (îlots ou secteurs) où les taux de crimes sont les plus élevés.
4. Obtenez les trois principaux types de crimes en fonction du nombre d’incidents dans la zone est « RESIDENCE » en utilisant « Location\_Description».
5. Calculez le pourcentage de crimes qui mènent à une arrestation.
6. Déterminez les jours de la semaine où les taux de crimes sont les plus élevés et les plus faibles.
7. Analyser le pourcentage d’affaires domestiques pour différents types de crimes.
8. Enquêter si certains types de crimes est plus courants au cours de saisons ou de mois précis.

**Exercice 2 :**

Soit le dataset se compose de trois tables issues de la base de données d'une boutique, avec les produits, les ventes et les vendeurs. La figure suivante montre comment les tables peuvent être connectées :



1. Créer 3 data frames (products, sales, sellers) à partir des 3 fichiers csv (products, sales, sellers) fournit en classe, puis afficher les 5 premières lignes de chaque data frame.
2. Afficher le schéma (les colonnes de chaque data frame).
3. Découvrez combien de commandes, combien de produits et combien de vendeurs sont dans les données.
4. Combien de produits ont été vendus au moins une fois ?
5. Quel est le produit qui a été vendu en plus de commandes ?
6. Combien de produits distincts ont été vendus chaque jour ?
7. Calculer le nombre de vendeurs qui ont vendu des produits.
8. Calculer la quantité totale des produits vendus.
9. Calculer la quantité des produits vendu par vendeur.
10. Calculer le nombre de pièces vendues par chaque vendeur pour chaque produit.
11. Quel est le revenu moyen des commandes ?

*En théorie, c'est simple : nous devons d'abord calculer le revenu de chaque commande, puis obtenir la moyenne. Rappelez-vous cela revenue = price \* quantity. Petit facile : le product\_price est dans le data frame products, tandis que le montant est dans le data frame sales.*

*Une approche pourrait être de simplement joindre les deux tables, de créer une nouvelle colonne et de faire la moyenne*