# TP2 : RDD (Resilient Distributed Dataset)

## Objectifs du TP :

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

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

## Exercice 1 (exercice 3 du TP1) :

Énoncé du problème : L’autorité scolaire de New York recueille des données auprès de toutes les écoles. Ces données aident à comprendre si les autobus sont arrivés à l’heure ou non. Cela aide également à comprendre s’il y a un itinéraire spécifique où les autobus prennent plus de temps, afin qu’il puisse être amélioré.

On vous donne un ensemble de données sur les autobus qui sont tombés en panne ou qui étaient en retard. Vous avez eu reçu l’ensemble de données ci-dessous :

Nom de fichier : bus-breakdown-and-delays.txt

* School\_Year : Indique l’année à laquelle l’enregistrement fait référence.
* Run\_Type : Indique si une panne ou un retard s’est produit dans une catégorie précise de services d’autobus.
* Bus\_No : indique le numéro du bus.
* Route\_Number : indique le numéro d’itinéraire.
* Raison : Raison du retard indiquée par le personnel employé par l’autobus
* Occurred\_On : indique la date du retard
* Number\_Of\_Students\_On\_The\_Bus.

Vous êtes embauché en tant que consultant Big Data pour fournir des informations importantes. Vous devez rédiger une tâche Spark à l’aide des données ci-dessus et fournir ce qui suit :

1. Les raisons les plus courantes d’un retard ou d’une panne d’autobus
2. Les cinq principaux numéros d’itinéraire où l’autobus a été retardé ou en panne
3. Le nombre total d’incidents, en moyenne annuelle, lorsque les élèves étaient dans l’autobus
4. Le nombre total d’incidents, en moyenne annuelle, lorsque les élèves n’étaient pas dans l’autobus
5. L’année où les accidents ont été moins nombreux

Exercice 2 :

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 RDD.

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.