# 0TP1 : RDD (Resilient Distributed Dataset)

## Objectifs du TP :

Le but de ce TP est d’apprendre les concepts de base d’Apache Spark, et de pouvoir exécuter des opérations sur les RDD.

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

## Exercice 1 :

Créer un RDD à partir de la liste des nombres suivante : L= 50,6,54,9,325,47,62.

1. Calculer la somme des nombres de l’RDD.
2. Filtrer les nombres pairs et les nombres impairs de la liste dans deux nouvelles RDD.
3. Carré chaque numéro (x2) de la liste initiale. Le résultat doit être mis dans une nouvelle RDD.

## Exercice 2 :

1. Créer un fichier texte contenant le texte suivant :

*Apache Spark est un Framework de traitement des données distribué. Il est largement utilisé pour l'analyse de données, le traitement de flux en temps réel et bien plus encore. Spark prend en charge plusieurs langages de programmation, notamment Scala, Python et Java.*

1. Créer un RDD qui permet de lire le fichier.
2. Combien de lignes contient le fichier.
3. Séparer le texte en mots.
4. Compter le nombre total de mots.
5. Compter le nombre d’occurrence de chaque mot du texte.
6. Sauvegarder le résultat dans un fichier texte « output ».

## Exercice 3 :

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 lorsque les élèves étaient dans l’autobus
4. Le nombre total d’incidents lorsque les élèves n’étaient pas dans l’autobus
5. L’année où les accidents ont été moins nombreux