# UA3  Évaluation sommative

Projet :  Analyse de Données Massives avec Spark

## Objectif

Réaliser un mini-projet d'analyse de données massives en utilisant Apache Spark, en appliquant les notions étudiées en UA3 : architecture Spark, RDD, DataFrames, transformations, actions et analyse exploratoire.

## Jeu de données

Choisir un jeu de données répondant aux critères suivants :

- Au moins 500 000 lignes
- Au moins 5 colonnes
- Format : CSV (fortement recommandé) ou JSON
- Source ouverte : Kaggle, Gouvernement du Canada, etc.

## Travail demandé

### Modélisation

- Utilisation principale des DataFrames
- Comparaison partielle avec les RDD (si pertinente)
- Justification des choix techniques (optimisation, performance, cas d'usage)

### Transformations et actions

Appliquer au minimum :

- 5 transformations (ex. : filter, select, groupBy, orderBy, withColumn)
- 3 actions (ex. : count, show, write)

### Analyse du dataset

- Formuler 3 à 4 questions métier
- Produire des résultats concrets : statistiques, agrégations, top N, distributions
- Appliquer un nettoyage de base si nécessaire : valeurs manquantes, types incorrects, incohérences

### Rapport (à Inclure)

( Présentation du dataset,explication de l'architecture Spark utilisée, choix de modélisation ,transformations et actions appliquées, résultats/interprétation,limites et pistes d'amélioration )