# Projet Spark : Analyse et Prévision de la Pollution Urbaine

October 9, 2025

## Présentation du projet

L'objectif de ce projet est de concevoir une application capable de **mesurer, analyser et prédire la pollution** dans un réseau urbain, en utilisant **Apache Spark** pour le traitement de données massives et la **programmation fonctionnelle**. Ce projet mettra en évidence comment la **programmation fonctionnelle** est particulièrement adaptée au traitement de **big data**, en exploitant ses capacités pour la **parallélisation et la gestion efficace de grandes quantités de données**.

## Objectifs pédagogiques

- Maîtriser le traitement de grandes quantités de données avec Spark (RDD, DataFrames, SQL).
- Exploiter la programmation fonctionnelle avec des opérations telles que `map`, `flatMap`, `filter`, `reduce`, `groupBy` ou `aggregate`.
- Expérimenter le traitement **batch** et **streaming** des données.
- Découvrir l'usage de Spark MLlib pour des tâches de prédiction simples.

## Sources de données envisagées

- Mesures de pollution : $CO_2$, particules fines, bruit, humidité.
- Données temporelles : horaires, jours, semaines.
- Informations contextuelles : stations, lignes de transport, météo.

- Données additionnelles : trafic des passagers, événements ponctuels, jours fériés.

# Phases de réalisation

## 1. Ingestion et préparation des données

Vous êtes invités à **construire vous-mêmes le jeu de données**, soit en collectant des données publiques, soit en générant un dataset simulé adapté aux objectifs du projet.

- Lecture des fichiers CSV ou JSON avec Spark.
- Nettoyage des données : suppression des doublons et des valeurs manquantes.

## 2. Transformation et exploration fonctionnelle

- Utilisation de `map`, `filter` et `flatMap` pour transformer les données.
- Calcul de statistiques par station ou par ligne (moyenne, maximum, minimum).
- Extraction de variables temporelles pertinentes (heure, jour, mois).

## 3. Analyse approfondie

- Identification des stations les plus exposées à la pollution.
- Détection des pics horaires et périodes critiques.
- Création d'un indicateur global de pollution.
- Détection automatique des anomalies dans les données.

## 4. Modélisation des relations entre stations

- Représentation des stations et de leurs connexions sous forme de graphe avec **Spark GraphX**.
- Étude de la propagation de la pollution à travers le réseau.

### 5. Prédiction

- Construction d'un pipeline de transformation de données pour créer des features adaptées.
- Application de modèles de prédiction Spark MLlib (régression, arbres de décision, forêts aléatoires).

### 6. Traitement en temps réel (optionnel)

- Simulation de flux de données provenant de capteurs en continu.
- Application de transformations fonctionnelles sur ces flux.
- Détection et signalement automatique des anomalies.

## Livrables

- Un pdf, sous forme de slides, présentant les aspects et concepts de Spark que vous avez utilisés dans ce projet.
- Code Spark fonctionnel (Scala) utilisant la programmation fonctionnelle.
- Rapport décrivant la méthodologie et les résultats obtenus.
- Visualisations des analyses et prédictions (facultatif).

## Résumé

Ce projet permet de combiner traitement de données massives, programmation fonctionnelle, analyse statistique et prédiction, pour construire une application complète, robuste et évolutive.

## Références

### Apache Spark

- **Site officiel Apache Spark** : `https://spark.apache.org/`
  Documentation complète, guides rapides, API et tutoriels.

- Chambers, B., & Zaharia, M. (2018). *Spark: The Definitive Guide*. O'Reilly Media.
  Livre complet couvrant RDD, DataFrames, SQL, MLlib et Streaming.

- Damji, J. S., et al. (2020). *Learning Spark: Lightning-Fast Big Data Analysis*. O'Reilly Media.
  Idéal pour débutants et intermédiaires avec exemples en Scala et Python.

- **Databricks Spark Documentation** : `https://docs.databricks.com/`
  Tutoriels pratiques et notebooks concrets pour Spark.

## Scala

- Odersky, M., Spoon, L., & Venners, B. (2021). *Programming in Scala* (4ème édition). Artima.
  Livre écrit par le créateur de Scala, couvrant tout le langage et la programmation fonctionnelle.

- **Documentation officielle Scala** : `https://docs.scala-lang.org/`
  Guides, tutoriels, références API et bonnes pratiques.

- Chiusano, P., & Bjarnason, R. (2014). *Functional Programming in Scala*. Manning Publications.
  Pour comprendre la programmation fonctionnelle appliquée à Scala.

- **Scala Exercises** : `https://www.scala-exercises.org/`
  Plateforme interactive proposant des exercices sur Scala et la programmation fonctionnelle.