



## Formation PySpark (Atelier Data Profiling)

-  Mohammed ATTIK (Copyright)
-  [mohammed.attik@gmail.com](mailto:mohammed.attik@gmail.com)

## Le data profiling

---

Le data profiling, également appelé profilage des données, est un processus d'analyse des données qui vise à comprendre la structure, la qualité et le contenu des données. C'est une étape importante dans la préparation et la compréhension des données avant de les utiliser dans des analyses plus avancées ou des applications métiers.

Le data profiling fournit une vue d'ensemble des données, ce qui aide les analystes et les scientifiques des données à comprendre leur qualité et leur signification, et à prendre des décisions éclairées sur la manière de les manipuler pour répondre aux besoins spécifiques d'analyse ou d'application.

## Le data profiling

---

Le data profiling implique généralement l'examen systématique des caractéristiques des données, telles que :

1. **Types de données** : Identifier les types de données présents dans chaque colonne (texte, nombre, date, etc.).
2. **Valeurs manquantes** : Déterminer les données manquantes et évaluer leur impact sur l'analyse.
3. **Doublons** : Identifier les enregistrements en double dans les données.
4. **Modèles de données** : Rechercher des modèles ou des structures récurrents dans les données.
5. **Distributions** : Analyser la répartition des valeurs pour chaque colonne.
6. **Relations entre les colonnes** : Identifier les relations entre les différentes colonnes de données.

L'objectif est de découvrir des informations sur les données qui peuvent être utilisées pour prendre des décisions éclairées sur la manière de nettoyer, transformer et gérer les données.

## Le data profiling : indicateurs clés

---

Les résultats d'un profil de données capturent généralement des ensembles d'indicateurs clés associés aux valeurs de chaque colonne individuelle de données, tel que le nombre et le pourcentage de champs nuls ou remplis, le nombre de valeurs uniques, le nombre de fréquences pour chaque valeur et les patterns, les valeurs maximales ou minimales ; et l'information sur les types de données et la longueur des chaînes de caractères. Les applications de data profiling les plus sophistiquées offrent un niveau de détails supplémentaire sur les dépendances entre les colonnes et les relations entre tables, notamment. L'objectif est d'identifier les anomalies ou partager les spécificités des éléments des données.

# Le data profiling : problématiques

---

Voici quelques exemples de problèmes types que l'on peut rencontrer avec des techniques de data profiling :

- Des valeurs manquantes.
- Des valeurs présentes mais qui auraient dû être absentes.
- Des valeurs qui apparaissent à une fréquence imprévues, qu'elle soit basse ou haute.
- Des valeurs qui ne respectent pas un pattern ou un format donné.
- Des données aberrantes qui sont bien trop basses ou trop élevées pour la plage définie.

## Atelier de Data Profiling avec Apache Spark

---

- **jeu de données** : <https://static.openfoodfacts.org/data/en.openfoodfacts.org.products.csv>
- **Objectif** : Réaliser une analyse de profilage des données sur ce jeu de données en utilisant Apache Spark pour comprendre la qualité et la structure des données.
- **Tâche** :
  1. Téléchargez le jeu de données
  2. Utilisez Apache Spark pour charger le jeu de données dans un environnement Spark.
  3. Effectuez une analyse de profilage des données en utilisant les fonctionnalités de Spark (Spark SQL) pour comprendre la qualité et la structure des données. Ceci peut inclure :
    - Identification des types de données pour chaque colonne.
    - Analyse des valeurs manquantes.
    - Recherche de doublons.
    - Évaluation des distributions des valeurs pour chaque colonne.
    - Identification des relations potentielles entre les colonnes.
    - Autres analyses pertinentes pour comprendre les caractéristiques des données.
  4. Préparez un rapport sous Colab décrivant vos observations et conclusions à partir de l'analyse de profilage des données. Incluez des visualisations ou des statistiques descriptives pour illustrer vos résultats.
- **Livraison** :
  - Un notebook Jupyter sous Colab contenant le code Spark utilisé pour l'analyse de profilage des données.