# Exercice Atelier Intégration des Données

Niveau : I1

Spécificité de cet exercice : utilisation de données massives (Big Data)

Source de données : Openfoodfacts (Datalake (MongoDB) ou CSV)
ETL : Apache Spark ou Talend
Langage : Java ou Python (sachant que Spark est dispo sous Python avec PySpark)
Datawarehouse : BDD relationnelle SQL (MySQL, Postgresql…)

## Objectif pédagogique

Mettre en place une solution ETL chargée de collecter les données depuis un Datalake ou une source de données massives permettant de répondre à une problématique métier exposée.

## Cahier des charges

Mettre en place une solution ETL distribuée chargée de collecter les données depuis Openfoodfacts permettant de répondre à la problématique métier exposée.

Le client met en place une solution permettant de répondre aux préoccupations grandissantes de nombreux consommateurs, à savoir la mise en place d'un programme alimentaire adapté à leurs besoins (santé, sport, bien-être…)

Le client souhaite générer aléatoirement un menu sur une semaine composé de produits disponibles depuis le datalake openfoodfacts. Sachant que le menu hebdomadaire est le seul résultat observable par l'utilisateur final, seules les informations composant ce menu et celles le liant à l'utilisateur final seront stockées dans le DWH. Attention à la structure du modèle de données dans le (ou les) datamart(s).

Exemples de programmes alimentaires : FODMAP, méditerranéen, DASH, à index glycémique bas, végétarien ou végétalien, régime de la plaque, Paléo, cétogène, etc.

Le menu doit être équilibré sur la semaine mais également sur chaque jour de la semaine. Par "équilibré", on entend qu'il doit correspondre aux besoins alimentaires nutritionnels sélectionnés par l'utilisateur (programme alimentaire, âge, sexe, poids). Par exemple : le menu doit respecter le nombre de calories maximum fixé (s'il y en a) mais également respecter les proportions de sodium, lipides, glucides… etc. imposées par le programme alimentaire.

## Spécificités techniques

Source des données OpenFoodFacts : https://fr.openfoodfacts.org/data

Le Datalake OpenFoodFacts ne suffit évidemment pas à répondre seul au besoin. Il faut donc créer deux sources de données supplémentaires :

- une pour les régimes alimentaires que vous alimentez vous-même avec les seuils pour les différentes catégories nutritionnelles (exemple : glucides < 5 g par jour). Commencez avec un ou deux régimes alimentaires pour tester.
- une pour les données des utilisateurs qui souhaitent obtenir les menus générés, sachant que chaque utilisateur a un régime alimentaire spécifique.

Qualité des données : Bien entendu, il n'est pas possible de prendre en compte

- les produits qui ne peuvent être évalués dans le cadre du programme alimentaire (exemple : produits dont des données importantes ne sont pas renseignées)
- les produits sans nom ou non disponibles dans la région de l'utilisateur
- les produits avec valeurs aberrantes pour des composantes nutritionnelles importantes
- autres critères de qualité des données à évaluer…

Ne pas réaliser la partie Datavisualisation qui sera effectuée dans le cadre d'un prochain cours. Vous devez uniquement vous concentrer sur la collecte des données, la transformation de celles-ci et la dépose des résultats dans un Datawarehouse.

# Rendu

La restitution se fera sous la forme d'une remise de rapport, incluant le code source (il peut s'agir d'un repo git) et d'une présentation orale du projet.