Cahier des Charges pour le Projet ETL des Produits Alimentaires

1. Introduction

Le projet consiste à développer un système ETL (Extract, Transform, Load) pour gérer un catalogue de **100 000 produits alimentaires (80 000 déjà collectés)**. Ce système devra permettre de collecter des données à partir de sites externes via des robots de scraping, nettoyer et transformer ces données, puis les stocker dans une base de données NoSQL. L'ETL sera automatisé et hébergé dans le cloud, et utilisera des algorithmes d'intelligence artificielle (IA) et de machine learning pour compléter et enrichir les informations des produits (titre, description, ingrédients, valeurs nutritionnelles).

2. Objectifs du Projet

- **Scraping** de données depuis des sources externes pour les **marques**, **produits**, et leurs **détails**.
- **Nettoyage et transformation** des données dans une base NoSQL.
- Utilisation d'**intelligence artificielle (IA)** pour nettoyer et valider les données, et pour compléter les informations manquantes.
- **Automatisation** du processus ETL pour assurer une mise à jour continue des produits.
- **Hébergement cloud** du pipeline ETL pour assurer la haute disponibilité et la scalabilité.
- Mise en place d'**algorithmes de machine learning** pour prédire et générer des informations manquantes, basées sur des produits similaires.
- Analyse des produits à l'aide d'**indicateurs clés (KPIs)**.

3. Description du Projet

3.1. Collecte des Données (Extraction)

- **Robots de Scraping** : Développement de robots qui scrappent les informations des produits sur des sites externes tels que les e-commerce, les pages de fabricants, ou les bases de données publiques.
  - Les **données à scraper** :
    - **Marques** : Nom, description, logo, catégorie.
    - **Produits** : Nom, image, catégorie, prix.
    - **Détails des produits** : Titre, description, ingrédients, valeurs nutritionnelles, certifications, instructions d'utilisation.

3.2. Nettoyage et Transformation des Données

- **Transformation NoSQL** : Les données collectées seront nettoyées et structurées sous un format **NoSQL** (par exemple MongoDB, Cassandra) pour une flexibilité maximale dans la gestion des données non structurées.
  - **Processus de nettoyage** :
    - Normalisation des données (formats de dates, unités de mesure, etc.)
    - Gestion des données manquantes ou erronées.
    - Fusion des données provenant de différentes sources.

3.3. Nettoyage et Validation des Données avec IA

- **Intelligence Artificielle** : Utilisation de modèles IA pour détecter et corriger les erreurs dans les données, ainsi que pour valider la pertinence des informations collectées.
    - Utilisation de techniques comme le **traitement du langage naturel (NLP)** pour analyser les descriptions des produits.
    - Détection automatique des produits similaires pour améliorer l'intégrité des données et vérifier les correspondances.

3.4. Automatisation et Hébergement Cloud

- **Automatisation du Pipeline ETL** : Le processus ETL sera automatisé pour que les données soient mises à jour régulièrement sans intervention manuelle.
    - Planification des tâches à intervalles réguliers (par exemple, chaque nuit ou chaque semaine).

3.5. Prédiction des Informations Manquantes avec Machine Learning

- **Modèles de Machine Learning** : Développement de modèles de machine learning pour prédire les informations manquantes des produits en utilisant des **produits similaires**.
    - **Données à prédire** :
        - **Titre** : Création de titres pertinents en fonction des produits similaires.
        - **Description** : Génération de descriptions en analysant les attributs des produits similaires.
        - **Ingrédients** : Prédiction des ingrédients basée sur des produits ayant une composition similaire.
        - **Valeurs nutritionnelles** : Estimation des valeurs nutritionnelles en fonction de produits similaires et de leurs ingrédients.

3.6. Analyse des KPIs des Produits

- **Indicateurs clés (KPIs)** : Développement d'un tableau de bord pour analyser les produits et extraire des KPIs comme :
    - **Prix moyen**.
    - **Ingrédients principaux**.
    - **Valeur nutritionnelle**.
    - **Tendances de consommation**.
    - **Popularité de la marque et des produits**.

7. Conclusion

Le projet ETL des produits alimentaires permettra de collecter, nettoyer, enrichir et analyser des informations détaillées sur une large gamme de produits alimentaires. Grâce à l'utilisation d'IA et de machine learning, le projet offrira une solution robuste et évolutive pour gérer des données complexes et fournir des informations pertinentes aux utilisateurs finaux.

32, Rue De Perroquets 94350
Villiers Sur Marne, France

contact@onrtech.fr
www.onrtech.fr

+33 7 50 21 83 96