Samuel Forcier, 536 793 028

Maxime Mainardi, 536 942 625

Marie-Charlotte Meese-Coutaud, 537 020 554

Fany Ingride Tido Fokou, 536 931 949

Nhut Tri Mach, 536 915 310

OPEN FOOD FACT

Remise et présentation écrite du produit minimum viable

Équipe 23

Travail présenté à Eloïse Prévot

GLO-4035 – Base de données avancée

Faculté des sciences et de génie

Université Laval

19 décembre 2025

Déclaration IA

L'utilisation de l'intelligence artificielle, tel que ChatGPT ou Microsoft Copilot, fut utilisé afin d'éclaircir certains éléments pour nous aider à prendre de meilleures décisions et afin d'accélérer certaines étapes répétitives de transformation de données. Nous avons utilisé l'intelligence artificielle pour faciliter la catégorisation. En effet, en recueillant une liste de différents mots-clés (*keywords*) uniques des bases de données d'OFF et les différentes catégories de Food Data Central, nous avons demandé à l'IA de séparer et de regrouper les différents termes avec leurs catégories basées sur le guide alimentaire canadien.

Table des matières

1- Introduction

L'alimentation joue un rôle central dans la santé des individus. L'accès à une information claire, fiable et compréhensible est essentiel pour permettre aux consommateurs de faire des choix éclairés. *Open Food Facts* constitue aujourd'hui la plus grande base de données ouverte sur les produits alimentaires à l'échelle mondiale. Elle permet notamment de fournir des indicateurs reconnus tels que le Nutri-Score, l'Eco-Score et la classification NOVA, qui aident à évaluer respectivement la qualité nutritionnelle, l'impact environnemental et le degré de transformation des produits.

Problématique

Cependant, ces indicateurs reposent fortement sur la complétude et la qualité des données nutritionnelles. Lorsqu'un seul composant est manquant, les calculs deviennent impossibles, ce qui entraîne l'absence de scores pour un grand nombre de produits. Cette problématique est particulièrement marquée pour les produits disponibles en Amérique du Nord, et plus spécifiquement au Québec, où la couverture des données dans *Open Food Facts* est nettement inférieure à celle observée en Europe.

De plus, la base de données actuelle est majoritairement composée de produits transformés disposant d'un code-barres, ce qui exclut de nombreux aliments de base tels que les fruits et légumes vendus à l'unité.

Ainsi, bien qu'*Open Food Facts* représente un socle de données extrêmement riche, elle ne répond pas entièrement aux besoins spécifiques du contexte québécois.

Solution

Afin de répondre à ces limitations, ce projet vise à démontrer la viabilité technique de la création d'une version québécoise de la base de données *Open Food Facts*. Cette version adaptée aurait pour objectif de mieux représenter les produits disponibles au Québec, d'uniformiser et de simplifier les données, et de compléter les informations manquantes à l'aide de la base de données *FoodDataCentral* en complément.

Dans ce cadre, une application de type preuve de concept est développée et s'adresse principalement à l'équipe d'*Open Food Facts* Québec. Elle permet de rechercher des produits alimentaires selon différents

critères de qualité (Nutri Score, Eco-Score, NOVA), ainsi que de proposer des recommandations de produits à partir des ingrédients d'une recette de cuisine.

Présentation

Ce rapport vise à justifier les choix technologiques et méthodologiques effectués, en mettant l'accent sur la fiabilité, la maintenabilité et l'extensibilité de la solution proposée.

Le présent rapport est structuré en plusieurs sections. Après l'introduction, la section 2 décrit la stratégie d'acquisition des données et les différentes sources utilisées. La section 3 est consacrée aux technologies retenues pour le développement de l'application. La section 4 détaille ensuite les processus d'extraction, de transformation et de chargement des données (ETL). La section 5 présente le pipeline de données ainsi que les algorithmes de recommandation de produits. Enfin, la section 6 expose un plan d'expansion traitant des enjeux de passage à l'échelle, incluant la réplication, le partitionnement et la sécurité des bases de données.

L'ensemble du document vise à démontrer que la création d'une base de données alimentaire adaptée au contexte québécois est non seulement pertinente, mais également réalisable sur le plan technique.

2 – Stratégie d'acquisition des données

Sources

Pour ce projet, nous avons choisi de combiner cinq sources principales de données ouvertes :

- Open Food Facts (OFF), une base de données collaborative mondiale répertoriant plus de 2 millions de produits alimentaires, incluant des informations sur les ingrédients, le Nutri-Score, l'Eco-Score, et la classification Nova.
- Le Fichier canadien sur les éléments nutritifs (FCEN)**,** publié par Santé Canada, qui fournit des informations nutritionnelles détaillées sur les aliments disponibles au Canada.
- FoodData Central (FDC) qui regroupe cinq bases de données américaines, incluant des produits bruts et transformés.

- Le Guide alimentaire canadien, publié par Santé Canada, qui fournit plusieurs recettes canadiennes complets ainsi qu'un tableau de catégorisation.
- *Love Food Hate Waste Canada* suggéré par Open Food Facts pour les recettes.

Méthode d'extraction

Les données d'OFF ont été extraites à l'aide de leur API REST publique et de leur dump JSON téléchargeable. Les produits ont été filtrés pour ne conserver que ceux vendus en Amérique du Nord : Canada, États-Unis et Mexique.

Les données du FCEN ont été obtenues via les fichiers CSV officiels disponible sur le portail de données du gouvernement du Canada. Nous dénormaliserons ces données, car nous voulons fusionner avec les données de Open Food Facts, qui sont déjà dans une BD de type document.

Les données de FoodData Central ont été obtenues via les fichiers JSON disponible sur le site officiel d'United States Departement of Agriculture (USDA). Nous dénormaliserons ces données, au même but que les données FCEN, soit de fusionner avec les données d'OFF. Les données du guide alimentaire canadien et de *Love Food Hate Waste Canada* sont obtenues par extraction de données web (*web scraping*) et constitueront les données initiales pour les recettes de la base de données. Voir annexe 1 pour un exemple de donnée source. La base de données de OFF contient un très grand nombre de catégories différentes, souvent trop détaillées. Nous avons donc catégorisé les produits selon le guide alimentaire canadien (GAC). Voir l'annexe 11 pour la structure complète de la catégorisation des aliments du GAC. Une analyse préliminaire des données a montré que les données FCEN et FDC sont normalisées ce qui rend la lecture moins intuitive. En explorant un peu plus les données, plusieurs champs comme *microbes*, *trade_channel*, *lab_method_description* , *lab_method_technique*, *subbrand_name* et bien plus dans FDC ne paraissent pas utiles à conserver dans ce contexte ainsi que les champs *yield_amount* et certains *serving_size* dans FCEN. Aussi, plusieurs produits d'Open Food Facts ne possèdent pas de Nutri-Score ou Eco-Score (voir un exemple à l'Annexe 2). Grâce au FCEN et FDC, il devient possible d'estimer ces indicateurs pour un plus grand nombre d'aliments, car elles ont des champs qui donnent les quantités de sodium, protéine, énergie (KJ), fibre, gras saturé et sucre par 100g. Ces champs seront utiles dans le calcul du nutri-score en particulier. Cette intégration de données permettra ainsi de créer une base de données plus complète, contextualisée et exploitable pour des fonctionnalités de recommandation de produits et d'analyse nutritionnelle.

3- Technologies utilisées

Langage de programmation

Python (backend)

Pourquoi Python est maintenable ? Il est souvent vu comme étant un choix populaire d'entrée dans le domaine et donc, n'importe quels développeurs rejoignant le projet plus tard ne seraient pas trop pris à dépourvus et ça permettra de faciliter leur intégration. Ce langage est reconnu pour sa syntaxe simple, sa forte communauté et la possibilité de structuration du code en modules pour la réutilisation. Ces qualités facilitent la maintenance à long terme.

Pourquoi Python est extensible ? Python peut être utilisé dans différent environnement, grâce à la possibilité d'avoir des modules de codes. Ces modules pourront s'évoluer indépendamment. Ce langage peut être utilisé pour différentes applications en micro services, API REST ou pipelines distribués. En termes de scalabilité horizontale, Python peut être déployé sur plusieurs serveurs ou conteneurs. Alors, ça permet d'utiliser des traitements ou des requêtes en parallélismes en gérant des millions d'utilisateurs, ce qui offre une possibilité intéressante pour toute application qui continuera de grandir.

Pourquoi Python est fiable ? C'est un langage mature et stable**,** utilisé dans de nombreux environnements de production. Ses outils intégrés de gestion d'erreurs garantissent une exécution fiable et cohérente. De plus, il s'intègre facilement dans des environnements modernes tels que Docker, ce qui renforce la portabilité et la fiabilité de l'application.

TypeScript, HTML et CSS (frontend)

Pour le frontend, nous avons choisi un stack moderne. HTML structure les composants de l'interface, CSS en assure le style et la responsivité, tandis que TypeScript gère la logique, les interactions et la sécurité du typage. Elle nous permet de bénéficier de la flexibilité de JavaScript tout en ajoutant une couche de sécurité grâce au typage statique. Cette combinaison garantit une interface maintenable, extensible et fiable. Le choix des langages Python et TypeScript nous permet de construire une application robuste, évolutive et bien structurée. Python garantit la fiabilité et la puissance du backend, tandis que TypeScript assure la maintenabilité et l'extensibilité du frontend.

Base de données

Dans ce projet, une approche polyglotte a été adoptée pour le stockage des données, c'est-à-dire l'utilisation de plusieurs types de bases de données selon la nature des données et les besoins fonctionnels. Cette stratégie permet de tirer parti des forces spécifiques de chaque technologie pour optimiser la performance, la flexibilité et la maintenabilité du système. Les bases de données utilisées sont MongoDB pour les produits et Neo4j pour les recettes.

Base de données de type document

Le choix d'une base de données orientée documents s'est imposé naturellement. En effet, notre application manipule des données hétérogènes et semi-structurées, comme des recettes aux formats variés (nombre d'étapes, ingrédients) ou des produits aux profils nutritionnels inégaux. Une base de données comme MongoDB nous permet de stocker ces entités sous forme de documents JSON, ce qui favorise une structure souple et adaptable. Chaque entité peut être pensée comme une fiche autonome, structurée en paires clé-valeur. Cette approche répond au principe d'extensibilité, car elle permet d'ajouter ou modifier des champs sans impacter les autres documents ni nécessiter de migration de schéma.

De plus, MongoDB offre une bonne scalabilité horizontale, ce qui signifie que nous pourrions gérer une croissance importante du volume de données (ex. : des milliers de recettes ou produits) sans compromettre les performances. Afin d'assurer la performance des requêtes dans MongoDB, nous avons également prévu l'utilisation d'index, car elle permet d'accélérer la recherche sur des champs spécifiques comme code_barre, nutri-score ou marque. Ces optimisations renforcent la fiabilité du système en assurant des temps de réponse constants, même avec un volume croissant de données.

Base de données de type graphe

Étant donné que notre application doit effectuer des recommandations de produits et de recettes, nous avons choisi d'utiliser une base de données de type graphe, en l'occurrence Neo4j, afin de pouvoir modéliser efficacement les relations entre les entités. Cette structure de données est particulièrement adaptée aux moteurs de recommandation, car elle permet d'exécuter des requêtes complexes sur les liens entre les entités de manière performante.

Neo4j respecte les propriétés ACID (Atomicité, Cohérence, Isolation, Durabilité), garantissant ainsi la fiabilité et les cohérences des opérations. Il est également extensible, car son modèle sans schéma

permet d'ajouter facilement de nouveaux types de nœuds et relations, ce qui soutient la croissance future de l'application.

Dans notre cas, le graphe est composé de nœuds (produits, recettes, auteur, type, temps_preparation et temps_cuisson) reliés par des relations ECRITE_PAR, EST_DE_TYPE, A_COMME_TEMPS_CUISSON, A_COMME_TEMPS_PREPARATION et UTILISE.

Ainsi, l'utilisation conjointe de ces deux bases de données améliore la fiabilité du stockage, la maintenabilité du modèle de données et l'extensibilité du système, en permettant d'ajouter facilement de nouvelles entités (par ex. bio, transformés, etc.) sans impacter la structure globale.

4- Les détails du processus d'extraction, de transformation et de conversion (ETL)

Processus d'acquisition initiale des données

L'acquisition initiale des données consistait à collecter et à préparer les informations nutritionnelles et culinaires issues de plusieurs sources fiables. Ce processus a permis de constituer une BD alimentaire adaptée au contexte québécois. Les sources utilisées : Open Food Facts (OFF), FoodData Central (FDC) et Fichier canadien sur les éléments nutritifs (FCEN).

OFF est la bd principale et nous l'avons obtenu par leur API REST publique et de leur dump JSON. FDC a été obtenu par des fichiers JSON sur le site et les données obtenues sont normalisées. FCEN a été extraite par téléchargement du fichier CSV et les données obtenues sont normalisées.

Et deux sources de données le Guide alimentaire canadien constitué de recettes équilibrées et *Love Food Hate Waste Canada* constitué de recettes anti-gaspillage, toutes obtenues par web scraping en format JSON. Les annexes 1 et 3 sont des exemples de données obtenues. L'annexe 7 représente un diagramme d'activité du processus d'acquisition initiale.

Processus d’acquisition incrémental des données

L’acquisition incrémentale permet de maintenir la BD à jour en intégrant les nouveaux produits régulièrement, tout en assurant la cohérence et la qualité des données déjà présents.

Grace au processus incrémental, il sera possible d’ajouter les nouveaux produits alimentaires d’Amérique du Nord, de mettre à jour les produits présents au cas où la composition d’un produit changerait ou que l’ingrédient d’une recette est modifié et de supprimer des produits dont les marques ne sont plus sur le marché.

Pour cela il faudra surveiller les sources de données utilisées en s’inspirant du fonctionnement du système de git. Le type de traitement proposé est par batch pour les produits c’est à dire une exécution du processus d’acquisition tous les sept jours et pour les recettes cette extraction est faite à chaque fois que le conteneur Docker est lancé. Initialement, il faudra extraire les données des sources de données, calculer le hash avec un algorithme de hachage avant de journaliser les changements après comparaison et enfin de mettre à jour les données dans la BD.

Extraction des données

Une récupération des données de bases de données utilisées se fera tous les sept jours.

Horodatage

C'est l’enregistrement de la date et l’heure de la dernière mise à jour d’un produit. Pour le faire, il faudra ajouter un champ « derniereMiseAJour » dans chaque produit. Ce champ nous permettra de savoir si un produit a été modifié depuis la dernière synchronisation, de faciliter la comparaison entre les produits et de gérer les doublons.

Journalisation (log)

Pour pouvoir garder une trace des différentes modifications effectuées dans la base de données, une nouvelle collection nommée « journal » sera créée pour stocker les différentes actions effectuées dans la base de données (insertion, suppression, mise à jour). Chaque entrée contiendra : l’identifiant du produit, le type d’action, la date, et les champs modifiés.

Détection de changement

Pour savoir quel produit a été modifié, il faudra faire une comparaison des versions des produits par un algorithme de hachage ce qui permettra une détection automatique, car on ne vérifiera pas, champ par champ. Le hash de chaque produit dans la BD sera conservé et recalculé après chaque extraction.

Mise à jour de la base de données

L'étape finale du processus incrémental consiste à appliquer les changements détectés à la BD québécoise. Elle se repose sur trois actions principales : ajout, mise à jour et suppression.

Une fois qu'un nouveau produit est ajouté, le champ « dernieremiseAJour » est ajouté, un calcul et stockage de son hash est effectué pour les futures comparaisons et cette nouvelle écriture est consigné dans la collection journal.

Pour un produit modifié, son hash est changé dans ce cas les champs différents sont mise à jour ainsi que le champ « dernieremiseAJour » et cela est également consigné dans le journal.

Pour un produit supprimé, c'est-à-dire non existant dans les données extraites par batch, il est également supprimé dans la base de données et cette suppression est écrite dans le journal.

L'annexe 8 représente un diagramme d'activité du processus d'acquisition incrémentale.

Processus de transformation des données

Le processus de transformation vise à rendre les données extraites exploitables, cohérentes et enrichies pour l'analyse, la recommandation alimentaire et la construction de recettes. Il intervient après l'extraction et avant la fusion des sources.

Les données extraites depuis OFF avaient plusieurs champ vide. Un filtrage géographique a été effectué afin de ne garder que les produits vendus en Amérique du Nord. Un nettoyage des champs jugés non pertinents (champs vides ou incomplètes) a également été effectué.

Les données FCEN et FDC étaient initialement normalisées, ce qui nécessitait minimale une dénormalisation afin de pouvoir les insérer dans la BD de MongoDB. Avant la normalisation plusieurs champs ont été supprimé, car certaines informations n'étaient pas utiles pour le projet. Les champs supprimés : (FDC) *microbes*, *conversion_factor*, *trade_channel*, *lab_method_technique*, *lab_method_description*, *discontinued_date;* etc et (FCEN) : *yield_amount*, *serving_size*, *unit_conversion*,

etc. Nous avons conservé (FDC) : *fdc_id, data_type, description, ingredients, food_category, publication_date, brand_owner, brand_name, serving_size, serving_size_unit, market_country* et les quantités de nutriments par 100g/100ml, et (FCEN) : tous les *nutrient_amount* par 100g/100ml uniquement, *food_code* et *product_name*. Les nutriments nous permettront de calculer le nutri-score du produit. Nous avons aussi ajouté plusieurs champs pour FCEN dont le *data_sources*, qui est toujours "*FCEN"* pour connaitre la provenance de ces données, un *ingredient_text* et un *food*_groups, qui sont des champs vides nous permettant éventuellement d'ajouter une liste d'ingrédients et une catégorisation. Le nom de ces champs est déjà conforme avec OFF. Après, une dénormalisation des données obtenues a été effectué à l'aide de commandes *aggregate*. Cette étape a permis de fusionner les tables nutritionnelles, de regrouper les informations par aliment et ainsi être formaté à la BD OFF. Ainsi nous permettant de tout fusionner. De plus, tous les champs ont été renommé pour s'uniformiser aux données OFF.

La dernière transformation a été de catégoriser les produits en se basant sur la catégorisation des produits du guide alimentaire canadien. Nous avons associé les différentes combinaisons de mots-clés OFF et les catégories initiales de FDC aux différentes catégorisations du GAC pour ainsi créer des champs *category_ca* et *category_code_ca.* Les annexes 4,5 et 6 présentent le résultat des données obtenues après transformation des données.

Pour insérer les recettes dans la base de données Neo4j, elles ont été converties en graphe via un script python qui effectuait initialement un traitement sur les ingrédients en enlevant les quantités, ensuite chaque produit de la liste d'ingrédients constituait un nœud. L'annexe 2 montre un exemple de nœuds et relations obtenus. L'insertion des nœuds et relations s'est fait à partir des merge pour éviter que des nœuds ou relations ne se répètent. L'annexe 9 représente un diagramme d'activité de transformation des données.

Schéma du pipeline d'ETL

Une fois les données extraites, nettoyées, enrichies et fusionnées, elles sont sauvegardées dans deux bases distinctes selon leur nature. Les produits issus des bases OFF, FDC et FCEN sont stockés dans MongoDB et les recettes dans la base Neo4j. L'annexe 10 présente la pipeline ETL complet.

5- Les détails du pipeline de données

Création de recommandations de produits

Le pipeline de recommandation de produits vise à proposer, pour chaque ingrédient d'une recette, une liste de produits alimentaires pertinents et adaptés aux préférences de l'utilisateur.
Elle repose sur l'exploitation conjointe de deux bases de données complémentaires :

- Neo4j pour la modélisation des recettes et leurs relations,
- MongoDB pour le stockage des produits alimentaires et leurs caractéristiques nutritionnelles.

Choix et sélection d'une recette (Neo4j)

La première étape consiste à sélectionner une recette à partir de la base Neo4j.
Les recettes peuvent être filtrées selon plusieurs critères fonctionnels, notamment : le type de plat (ex. Dessert), le temps de préparation ou de cuisson et l'auteur de la recette.

Une fois la recette sélectionnée, une requête permet d'extraire l'ensemble des ingrédients associés via les relations **UTILISE**.

Extraction et normalisation des ingrédients

Les ingrédients récupérés depuis Neo4j ont été préalablement normalisés lors de la phase ETL : suppression des quantités, noms descriptifs et unités, homogénéisation des noms d'ingrédients.

Chaque ingrédient est ensuite traité indépendamment, ce qui permet de générer des recommandations spécifiques pour chacun d'eux.

Algorithme de sélection des produits selon les ingrédients

Dans un premier temps, pour chaque ingrédient d'une recette, une recherche est effectuée dans la base de données MongoDB afin d'identifier les produits dont le nom présente la plus forte similarité textuelle avec l'ingrédient. Les identifiants des produits correspondants sont alors extraits.

Ensuite, ces identifiants sont utilisés pour établir des liens entre les ingrédients et les produits dans la base de données Neo4j. Ainsi, chaque ingrédient est relié aux produits susceptibles de le représenter via

des relations. Lors de la génération d'une recommandation, lorsqu'une recette est sélectionnée, le système récupère directement les identifiants des produits associés à ses ingrédients depuis Neo4j. Ces identifiants servent ensuite de clés de recherche dans MongoDB afin d'obtenir les informations détaillées des produits (profil nutritionnel, scores, marque, etc.).

Enfin, le processus de recommandation commence par le filtrage des produits respectant les préférences de l'utilisateur.

Obtention des recommandations de produits

Le résultat final est structuré sous la forme d'un objet JSON où chaque ingrédient de la recette est une clé et la valeur associée est une liste ordonnée de produits recommandés.

Le parcours de la requête suit le flux suivant :

1. Le client envoie une requête de recommandation avec une recette et des préférences ;
2. Le backend interroge Neo4j pour récupérer la recette et ses ingrédients ;
3. Pour chaque ingrédient, le backend interroge MongoDB afin d'obtenir les produits candidats ;
4. Les produits sont évalués, classés et agrégés ;
5. La réponse finale est retournée au client.

Ce découplage garantit une bonne maintenabilité et facilite l'évolution future du système. Ce pipeline démontre la faisabilité technique d'un système de recommandation alimentaire basé sur des recettes, intégrant des critères nutritionnels, environnementaux et des préférences utilisateurs.

Voir annexe 12 pour un diagramme de séquence circulaire du flux de données de la procédure de création et d'obtention de recommandations de produits.

6- Plan d’expansion

Description du scénario d’expansion

Pour un déploiement à l’échelle du Québec, il est impératif de penser à une augmentation progressive du volume de données et du nombre d’utilisateurs. Cette application pourrait être utilisé par des consommateurs individuels pour la recherche de produits plus sains ou par des amateurs de cuisine pour la recommandation de recettes.

Ainsi le scénario d’expansion implique :

- Une croissance du nombre de produits dans la base de données ;
- Une augmentation du nombre de requêtes de recommandation ;
- Un accès concurrent aux bases de données.

Deux métriques principales ont été retenues pour analyser l’impact de l’expansion : la vitesse et le nombre de requêtes de recommandation. Ces métriques sont cohérentes avec la nature de l’application qui est orientée vers la consultation et la recommandation de produits et recettes.

Stratégie de réplication

Le système repose sur une réplication de type *single leader* des bases de données. Le *leader* est responsable des écritures (ETL, mises à jour), tandis que les réplicas sont dédiés aux lectures. Cette approche est adaptée car les écritures sont peu fréquentes et les recommandations sont majoritaires.

L’impact de ce type de réplication sur le nombre de requêtes de recommandations par minute est qu’il permettra la distribution des requêtes de lecture sur plusieurs réplicas, réduisant ainsi la charge sur le leader et améliorant le temps de réponse.

Stratégie de partitionnement

Le partitionnement concerne principalement les produits alimentaires. Il se fera selon nos sources de données initiales transformées (OFF, FDC, FCEN) puis seront ensuite repartitionnées par tranche d’identifiants de produits et selon la région de disponibilité (Québec, Canada, Amérique du Nord). Cette stratégie permettra de réduire la taille des partitions interrogés lors des recommandations ce qui améliorera les temps de réponse des requêtes.

Le partitionnement aura un impact significatif sur la vitesse d'exécution des requêtes de recommandation. En fragmentant les données, le système est en mesure d'identifier rapidement la ou les partitions contenant les produits recherchés, ce qui permet de limiter le périmètre de recherche et d'éviter le balayage complet de la base de données. Cette approche réduit le temps d'accès aux données, améliore la scalabilité du système et garantit de meilleures performances, en particulier lorsque le volume de produits est large.

Sécurité des bases de données

Cas d'utilisation 1 : Processus ETL et mise à jour

Les processus ETL accèdent aux bases de données pour insérer ou mettre à jour des produits et des recettes. Les stratégies de mitigation à mettre en place sont : une séparation des rôles entre lecture et écriture, une journalisation des accès et des modifications ainsi qu'une configuration des mots de passe des bases de données.

Cas d'utilisation 2 : Utilisateur final

Un utilisateur consulte des recommandations de produits à partir d'une recette.
Les risques principaux sont : un accès non autorisé aux données et une exposition de préférences personnelles. Les stratégies de mitigation pourraient être : un contrôle d'accès basé sur les rôles ainsi qu'une anonymisation des informations personnelles des utilisateurs.

7-Fonctionnalités additionnelles

Enrichissement des données existantes

Notre base de données repose majoritairement sur des données issues d'Open Food Facts. Une grande partie d'entre elles ne possède pas de scores Nutri, Nova ou Éco, particulièrement celles provenant de Food Data Central, qui n'utilise aucun de ces systèmes. Notre objectif était donc d'enrichir ces données en leur attribuant les scores Nutri et Nova. L'Éco-score aurait aussi pu être intégré, mais sa complexité et le temps limité nous ont poussés à le remettre à une itération future. Par ailleurs, nous souhaitions également revoir et corriger notre classification par catégories, à la suite des constats faits lors de la remise précédente.

Retour sur la classification des catégories

Après analyse, nous avons constaté une anomalie majeure : un nombre excessif de produits étaient classés sous le code 9990, correspondant aux aliments et boissons non classés. Ce code regroupait 346 551 produits, souvent en raison d'un manque d'information ou d'un script initial trop limité. Pour corriger la situation, nous avons extrait et inspecté le champ categories. Cela a révélé la présence de nombreuses données fautives qui auraient dû être supprimées : catégories dans d'autres langues que le français ou l'anglais, chaînes contenant des caractères indésirables (_, $, /, %, etc.), ou encore des valeurs ne représentant que des unités de mesure (ml, g, etc.). Ces données ont donc été nettoyées. Ensuite, parmi les produits restants en 9990, plusieurs étaient mal classés en raison d'un lexique insuffisant ou d'un script trop restreint. Nous avons donc enrichi notre vocabulaire de classification et relancé le traitement. Finalement, nous sommes passés de 346 551 à 126 550 produits en 9990, soit une réduction d'environ 63 %. Toutes les opérations ont été effectuées par lots afin d'assurer un nettoyage précis et contrôlé.

Classification du Nova score

Données d'Open Food Facts

Parmi les données d'OFF, 59 542 produits ne possédaient pas de score Nova, soit environ 16 % du total. Pour les classifier, nous avons utilisé les champs *ingredients*, ingredients_text, *additives_tags*, *additives_n*, *categories* et *_keywords*.

Le calcul des nova scores c'est fait selon les critères officiel. Voici un court résumé de notre logique de classification : Un score de 1 pour les produits sans additifs, mais 2 pour les produits avec légère transformation (présence de *oil*, *salt*, *honey,* etc.). Un score de 3 pour les produits avec additifs, mais de 4 pour les produits ultra transformés (présence de *sweetener*, *starch*, *supplement*, *super greens*, *drink mix*, etc.).

Comme pour les autres phases du projet, nous avons procédé par lots afin d'identifier les mots les plus représentatifs et d'améliorer progressivement notre lexique. Finalement, 8 212 produits ont pu être classés, soit environ 14 % des données initialement sans score Nova.

Données de Food Data Central

La classification des produits de Food Data Central a été plus simple, leurs données étant beaucoup mieux structurées. Nous avons ciblé le champ *food_groups*, en suivant une démarche similaire à celle

utilisée pour la classification par catégories. En explorant les différentes occurrences de ce champ nous en avons retiré ceux qui sont trop court ou trop vague. Assisté de l'IA, nous avons regroupé les food_groups sélectionnés et nous les avons associés à un nova score. Cette approche a permis d'attribuer un score à environ 69 % des produits. 100 863 produits sont de nova score 1, ce qui constitue nos produits de bases. Il y a aussi une forte proportion de produits ayant un nutri score 3 ou 4 ce qui est cohérent avec la nature des produits américains, souvent plus transformés.

Classification du Nutri score

Nous avons attribué un Nutri score à nos 850 000 produits de *FoodDataCentral*. Pour se faire nous avons appliqué l'algorithme officiel d'un nutri score approuvé par *Santé Publique France*[1]. Nous avions besoin de classé chaque produit selon ces 3 catégories : "huile, noix ou gras", "boisson" ou "autre (cas général)". Nous avions aussi besoin de la quantité de certains nutriments (sodium, sucre, fibre, protéine, etc.) par 100g du produit. Nous avions aussi besoin du pourcentage de fruits et légumes de l'aliment qu'on a dû estimer selon leur champs *category_ca,* puisque c'est une information absente de la base FoodDataCentral. Avec ces informations nous avons pu faire le calcul et convertir le résultat numérique en une lettre (A-E) selon le barème. Chacune des 3 catégories de Nutri score ont un calcul et un barème différents. Nous avons réussi à calculer l'entièreté des données *FoodDataCentral* incluant leurs produits de bases et transformés.

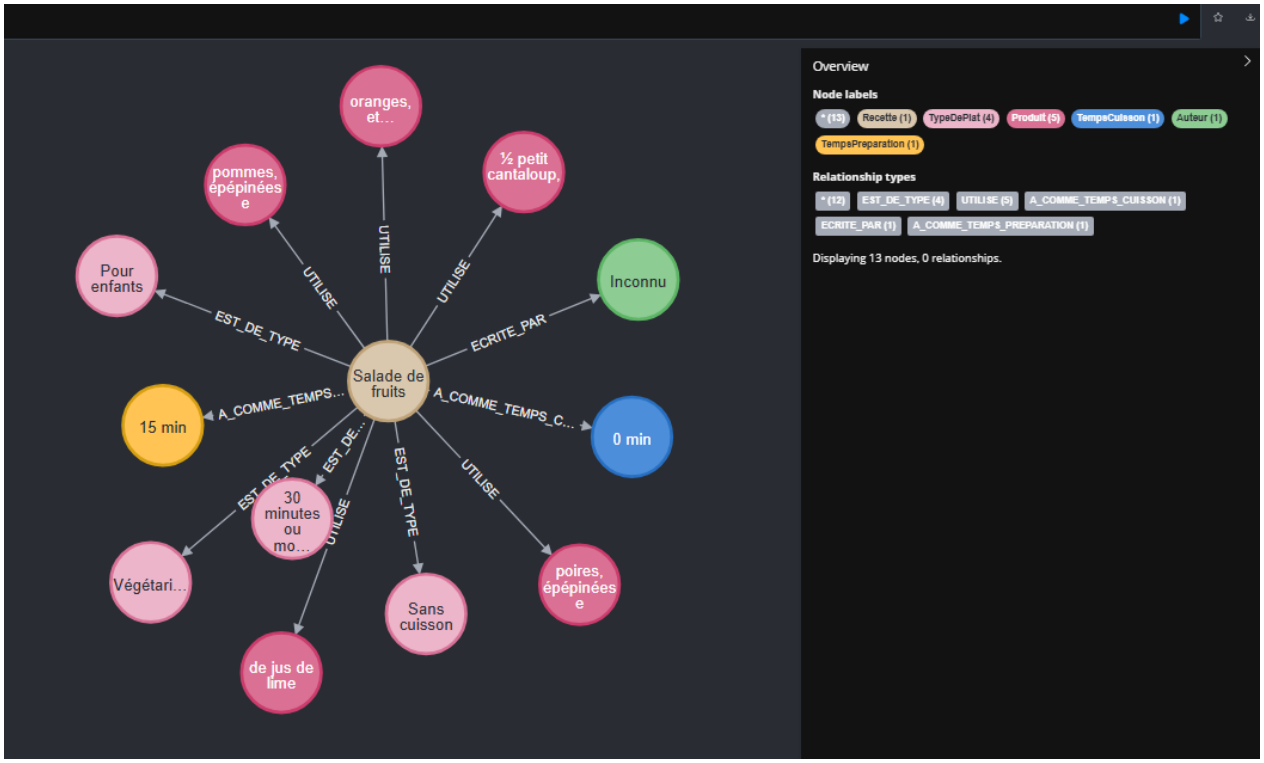
Pour les produits d'*Open Food Facts*, plusieurs avaient déjà un nutri score attribuer et ceux sans nutri score avaient très rarement tous les champs nécessaires pour le calcul. Nous avons décidé de le garder tel quel.

[1] https://www.santepubliquefrance.fr

Annexe 1 – Exemple de recette non transformée


```
{
  "titre": "Salade de fruits",
  "description": "Cette salade de fruits savoureuse satisfera à coup sûr votre envie de sucre et constitue un excellent dessert ou collation santé.",
  "auteur": null,
  "type_de_plat": [
    "30 minutes ou moins",
    "Pour enfants",
    "Sans cuisson",
    "Végétarien"
  ],
  "cote": null,
  "portions": "8",
  "temps_preparation": "15 min",
  "temps_cuisson": "0 min",
  "temps_total": null,
  "ingredients": [
    "2 pommes, épépinées et hachées",
    "2 oranges, pelées et hachées",
    "½ petit cantaloup, épépiné et haché",
    "2 poires, épépinées et hachées",
    "30ml (2 c. à table) de jus de lime (environ 1 lime)"
  ],
  "instructions": [
    "Déposer les pommes, les oranges, le cantaloup et les poires dans un bol. Presser le jus de lime sur les fruits. Mélanger et servir immédiatement ou réfrigérer jusqu’à consommation."
  ],
  "url": "https://guide-alimentaire.canada.ca/fr/recettes/amusante-salade-fruits/"
}
```

Annexe 2 – Exemple de recette transformée

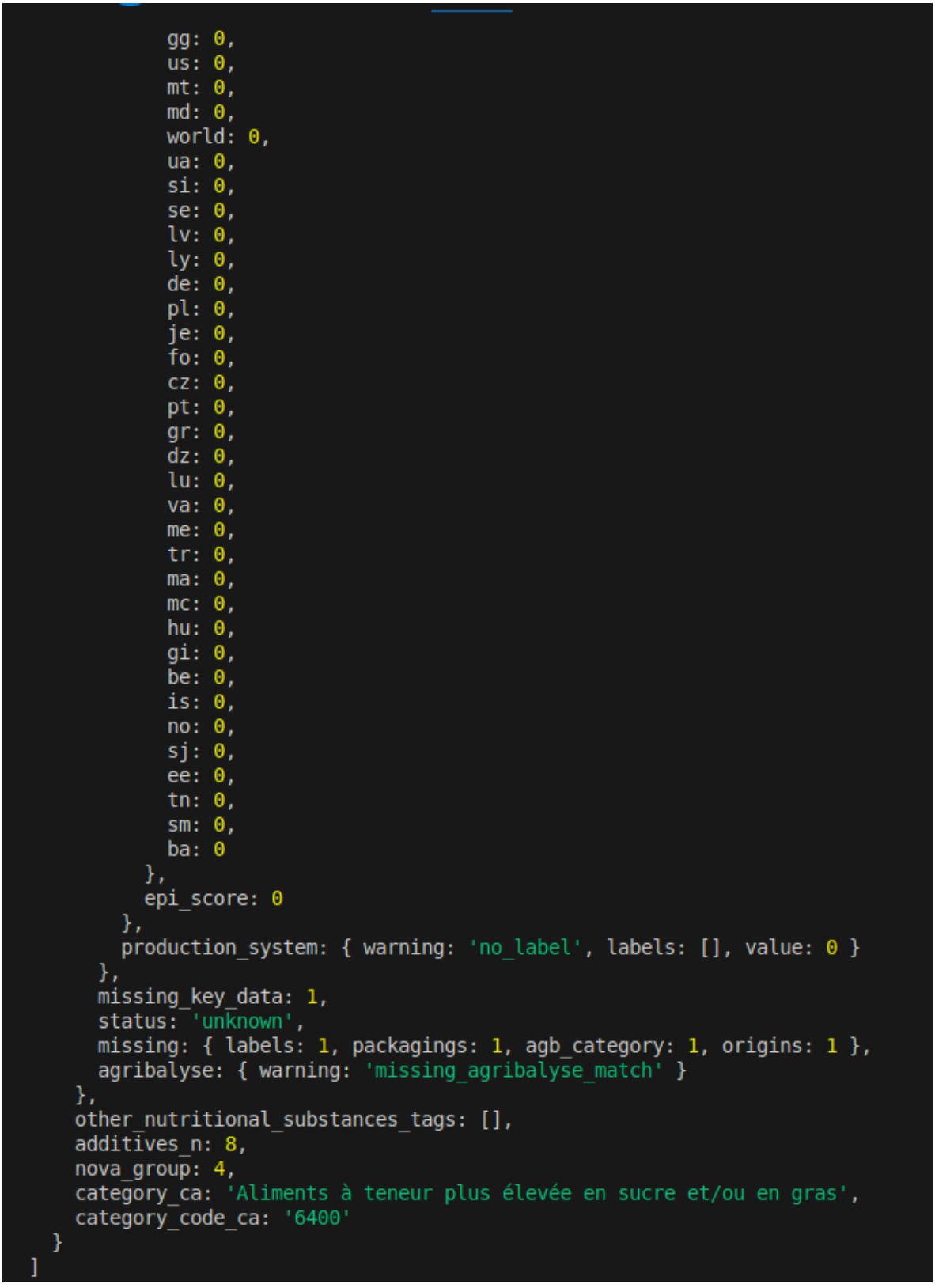


Annexe 3 – Exemple de donnée OFF non transformée

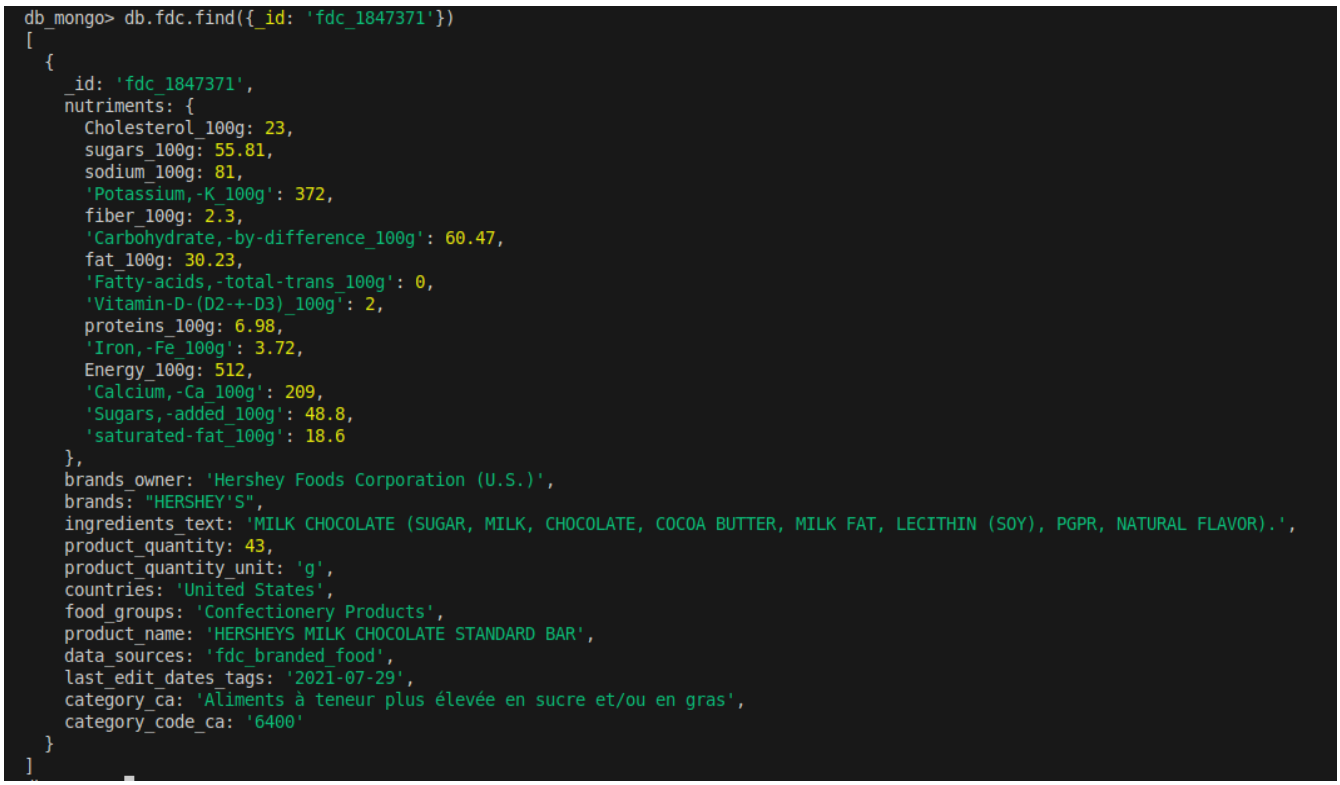
{"_id":"0000101209159","lc":"fr","correctors_tags":["openfoodfacts-contributors","tacite-mass-editor","sebleouf","moon-rabbit","roboto-app","timotheeberthault"],"data_sources_tags":["app-yuka","apps"],"data_quality_errors_tags":[],"countries_hierarchy":["en:france"],"categories_tags":["en:breakfasts","en:spreads","en:sweet-spreads","fr:pates-a-tartiner","en:hazelnut-spreads","en:chocolate-spreads","en:cocoa-and-hazelnuts-spreads"],"rev":18,"categories_old":"Petit-déjeuners, Produits à tartiner, Produits à tartiner sucrés, Pâtes à tartiner, Pâtes à tartiner aux noisettes, Pâtes à tartiner au chocolat, Pâtes à tartiner aux noisettes et au cacao","unknown_nutrients_tags":[],"removed_countries_tags":[],"packaging_old":"","nutriscore_tags":["e"],"nucleotides_prev_tags":[],"data_quality_info_tags":["en:packaging-data-incomplete","en:environmental-score-extended-data-not-computed","en:food-groups-1-known","en:food-groups-2-known","en:food-groups-3-unknown"],"nutriscore_grade":"e","brands_old":"Bovetti","categories_properties":{"agribalyse_food_code:en":"31032","ciqual_food_code:en":"31032","agribalyse_proxy_food_code:en":"31032"},"codes_tags":["code-13","conflict-with-upc-12","0000101209xxx","000010120xxxx","00001012xxxxx","0000101xxxxxx","000010xxxxxxx","00001xxxxxxxx","0000xxxxxxxxx","000xxxxxxxxxx","00xxxxxxxxxxx","0xxxxxxxxxxxx"],"compared_to_category":"en:cocoa-and-hazelnuts-spreads","minerals_prev_tags":[],"ingredients_text_fr":"","nutrition_grades":"e","ecoscore_tags":["d"],"debug_param_sorted_langs":["fr"],"ingredients_text":"","minerals_tags":[],"states_tags":["en:to-be-completed","en:nutrition-facts-completed","en:ingredients-to-be-completed","en:expiration-date-completed","en:packaging-code-to-be-completed","en:characteristics-to-be-completed","en:origins-to-be-completed","en:categories-completed","en:brands-completed","en:packaging-to-be-completed","en:quantity-completed","en:product-name-completed","en:photos-to-be-validated","en:packaging-photo-to-be-selected","en:nutrition-photo-selected","en:ingredients-photo-selected","en:front-photo-selected","en:photos-uploaded"],"product_quantity_unit":"g","product_name":"Véritable pâte à tartiner noisettes chocolat noir","other_nutritional_substances_tags":[],"nutrition_score_warning_no_fiber":1,"interface_version_created":"20150316.jqm2","origins":"","created_t":1519297017,"traces_hierarchy":[],"labels_hierarchy":["en:no-gluten","en:no-palm-oil"],"max_imgid":"3","allergens_hierarchy":["en:nuts"],"nutrition_score_beverage":0,"update_key":"brands","complete":0,"generic_name":"","languages_hierarchy":["en:french"],"emb_codes_orig":"","weighers_tags":[],"nutrient_levels_tags":["en:fat-in-high-quantity","en:saturated-fat-in-high-quantity","en:sugars-in-high-quantity","en:salt-in-low-quantity"],"brands_lc":"xx","purchase_places":"","environmental_score_data":{"adjustments":{"packaging":{"non_recyclable_and_non_biodegradable_materials":0,"score":61.0,"warning":"unscored_shape
… }

Annexe 4 – Exemple de donnée OFF transformée


```
{
  _id: '0000360055009',
  nutrition_data_prepared_per: '100g',
  serving_size_imported: '2 PIECES (2.5 g)',
  nutrient_levels: { sugars: 'low', fat: 'low', salt: 'moderate' },
  product_name: 'Dental gum',
  countries_tags: [ 'en:united-states' ],
  nutrition_grade_fr: 'unknown',
  pnns_groups_2_tags: [ 'sweets', 'known' ],
  brand_owner_imported: 'Lotus Brands, Inc.',
  pnns_groups_1_tags: [ 'sugary-snacks', 'known' ],
  countries_imported: 'United States',
  lc_imported: 'en',
  sources: [
    {
      images: [],
      id: 'database-usda',
      name: 'database-usda',
      import_t: 1587649513,
      fields: [
        'product_name_en',
        'categories',
        'countries',
        'brand_owner',
        'data_sources',
        'nutrition_data_per',
        'nutrition_data_prepared_per',
        'serving_size',
        'ingredients_text_en',
        'nutrients.carbohydrates_unit',
        'nutrients.carbohydrates_value',
        'nutrients.energy_unit',
        'nutrients.energy_value',
        'nutrients.energy-kcal_unit',
        'nutrients.energy-kcal_value',
        'nutrients.fat_unit',
        'nutrients.fat_value',
        'nutrients.salt_unit',
        'nutrients.salt_value',
        'nutrients.sugars_unit',
        'nutrients.sugars_value',
        'nutrients.vitamin-a_unit',
        'nutrients.vitamin-a_value',
        'nutrients.vitamin-c_unit',
        'nutrients.vitamin-c_value'
      ],
      url: null,
      manufacturer: null
    }
  ],
  additives_tags: [
    'en:e322',   'en:e322i',
    'en:e341',   'en:e341ii',
    'en:e420',   'en:e422',
    'en:e500',   'en:e500ii',
    'en:e903',   'en:e965',
    'en:e965ii', 'en:e967'
  ],
  brand_owner: 'Lotus Brands, Inc.',
```


gg: 0,
us: 0,
mt: 0,
md: 0,
world: 0,
ua: 0,
si: 0,
se: 0,
lv: 0,
ly: 0,
de: 0,
pl: 0,
je: 0,
fo: 0,
cz: 0,
pt: 0,
gr: 0,
dz: 0,
lu: 0,
va: 0,
me: 0,
tr: 0,
ma: 0,
mc: 0,
hu: 0,
gi: 0,
be: 0,
is: 0,
no: 0,
sj: 0,
ee: 0,
tn: 0,
sm: 0,
ba: 0
},
epi_score: 0
},
production_system: { warning: 'no_label', labels: [], value: 0 }
},
missing_key_data: 1,
status: 'unknown',
missing: { labels: 1, packagings: 1, agb_category: 1, origins: 1 },
agribalyse: { warning: 'missing_agribalyse_match' }
},
other_nutritional_substances_tags: [],
additives_n: 8,
nova_group: 4,
category_ca: 'Aliments à teneur plus élevée en sucre et/ou en gras',
category_code_ca: '6400'
}
]

Annexe 5 – Exemple de donnée FDC transformée



```
db_mongo> db.fdc.find({_id: 'fdc_1847371'})
[
  {
    _id: 'fdc_1847371',
    nutriments: {
      Cholesterol_100g: 23,
      sugars_100g: 55.81,
      sodium_100g: 81,
      'Potassium,-K_100g': 372,
      fiber_100g: 2.3,
      'Carbohydrate,-by-difference_100g': 60.47,
      fat_100g: 30.23,
      'Fatty-acids,-total-trans_100g': 0,
      'Vitamin-D-(D2-+-D3)_100g': 2,
      proteins_100g: 6.98,
      'Iron,-Fe_100g': 3.72,
      Energy_100g: 512,
      'Calcium,-Ca_100g': 209,
      'Sugars,-added_100g': 48.8,
      'saturated-fat_100g': 18.6
    },
    brands_owner: 'Hershey Foods Corporation (U.S.)',
    brands: "HERSHEY'S",
    ingredients_text: 'MILK CHOCOLATE (SUGAR, MILK, CHOCOLATE, COCOA BUTTER, MILK FAT, LECITHIN (SOY), PGPR, NATURAL FLAVOR).',
    product_quantity: 43,
    product_quantity_unit: 'g',
    countries: 'United States',
    food_groups: 'Confectionery Products',
    product_name: 'HERSHEYS MILK CHOCOLATE STANDARD BAR',
    data_sources: 'fdc_branded_food',
    last_edit_dates_tags: '2021-07-29',
    category_ca: 'Aliments à teneur plus élevée en sucre et/ou en gras',
    category_code_ca: '6400'
  }
]
```

Annexe 6 – Exemple de donnée FCEN transformée

_id: 'fcne_1493',

product_name: 'Pomme, congelée, tranchée, non sucrée, non chauffée',

nutriments: {

ALCOOL_100g: 0,

'CAFÉINE_100g': 0,

SODIUM_100g: 3,

'MAGNÉSIUM_100g': 3,

FER_100g: 0.18,

CUIVRE_100g: 0.058,

'ÉQUIVALENT EN NIACINE TOTALE_100g': 0.092,

CALCIUM_100g: 4,

[...]

POTASSIUM_100g: 77,

'ÉNERGIE (KILOCALORIES)_100g': 48,

PHOSPHORE_100g: 8,

EAU_100g: 86.85,

Energy_100g: 201,

sugars_100g: 12.31,

proteins_100g: 0.28,

'saturated-fat_100g': 0.053,

fiber_100g:1.3

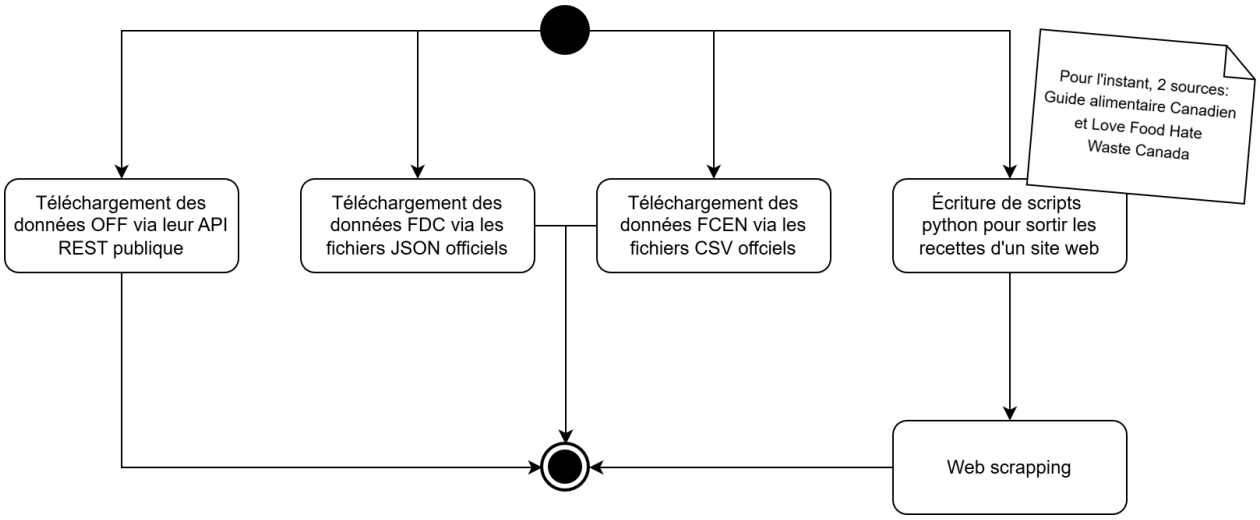
},

data_sources: 'FCNE',

ingredients_text: '' ,

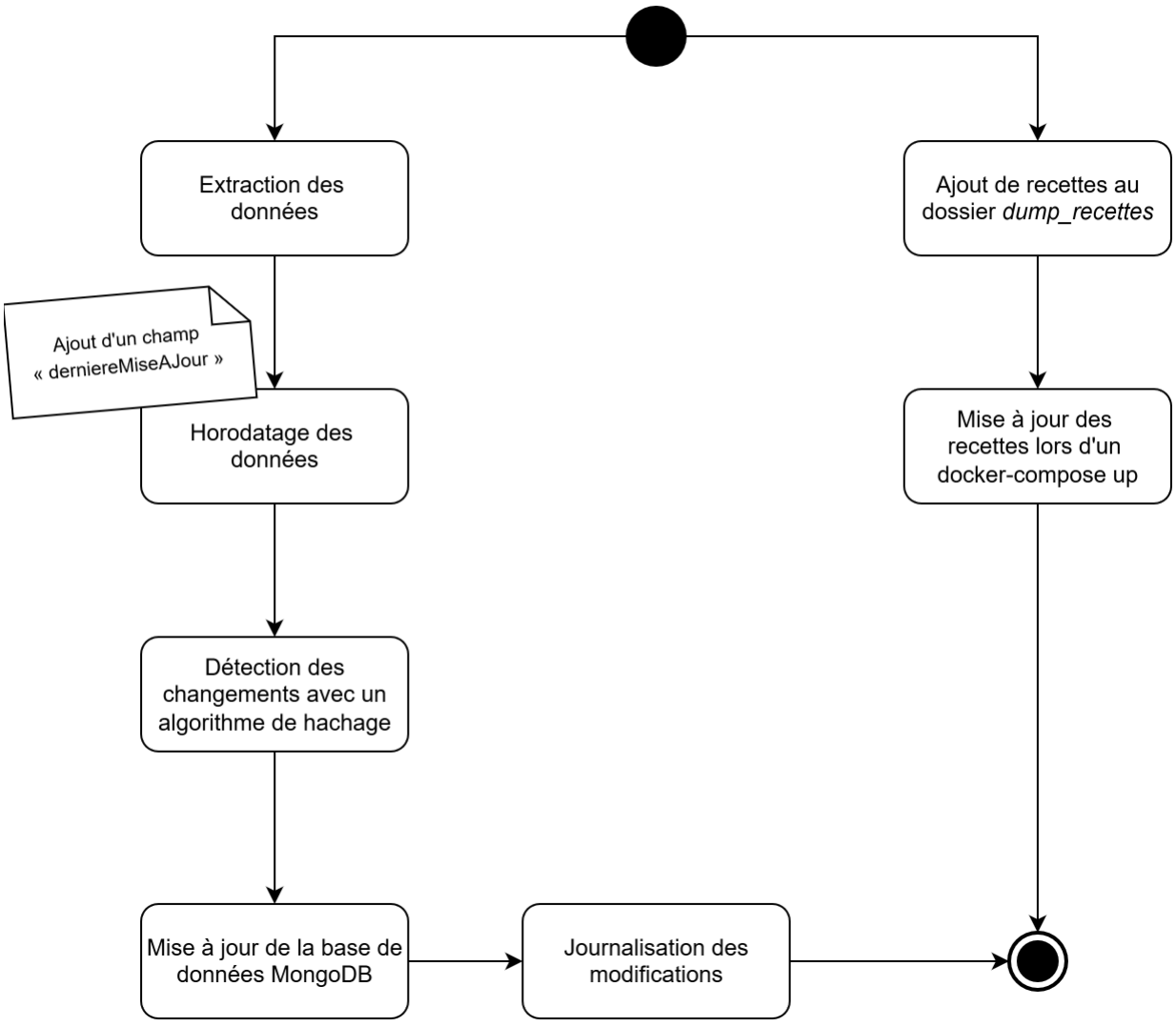
food_groups: ''

Annexe 7 – Diagramme d’acquisition des données



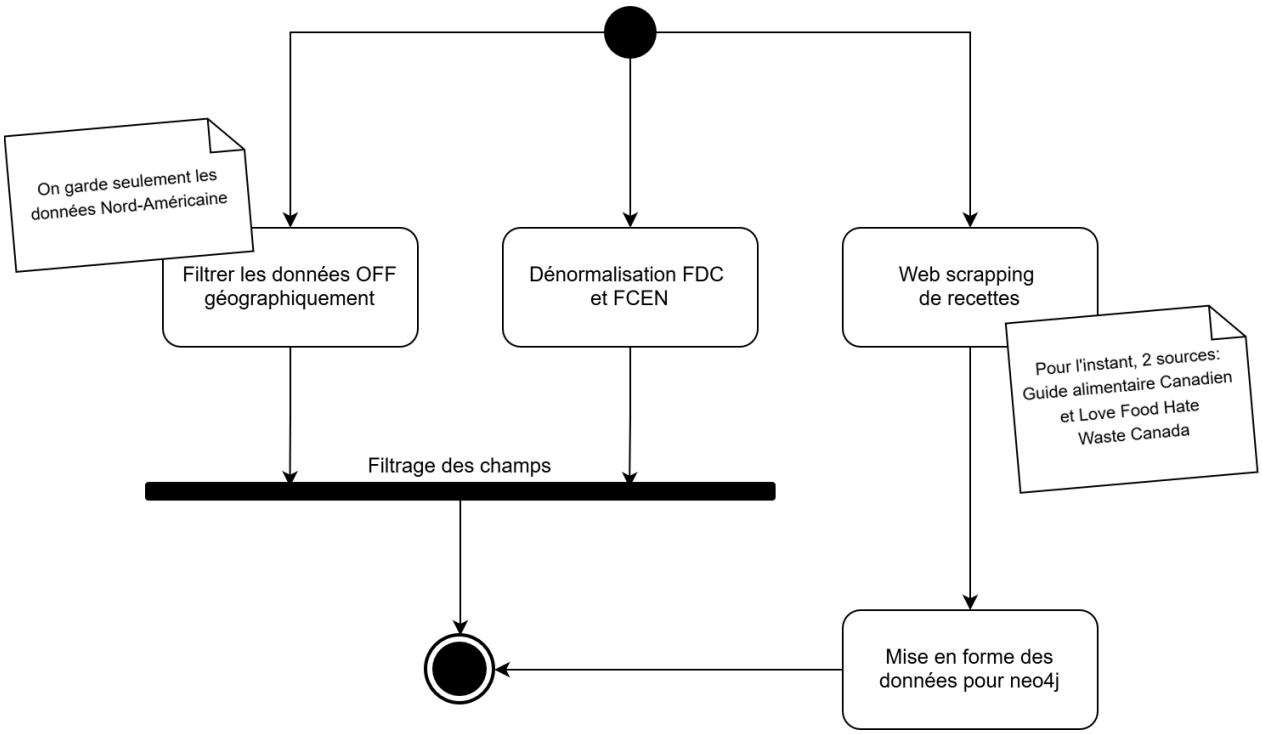
Étapes acquisition des données

Annexe 8 – Diagramme d’acquisition incrémentale des données



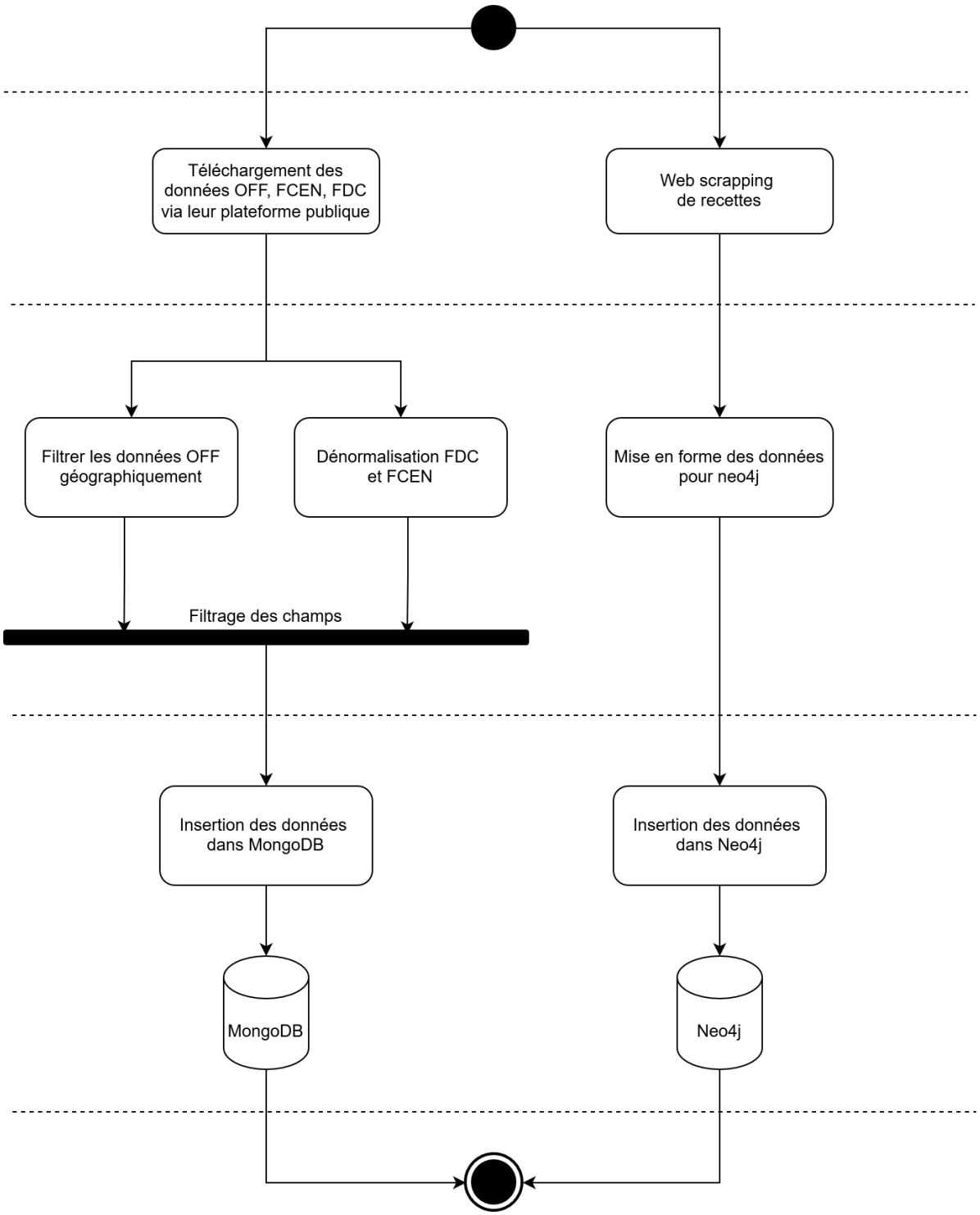
Étapes d'acquisition incrémentale des données

Annexe 9 – Diagramme de transformation des données



Étapes de transformation des données

Annexe 10 – Diagramme du processus ETL



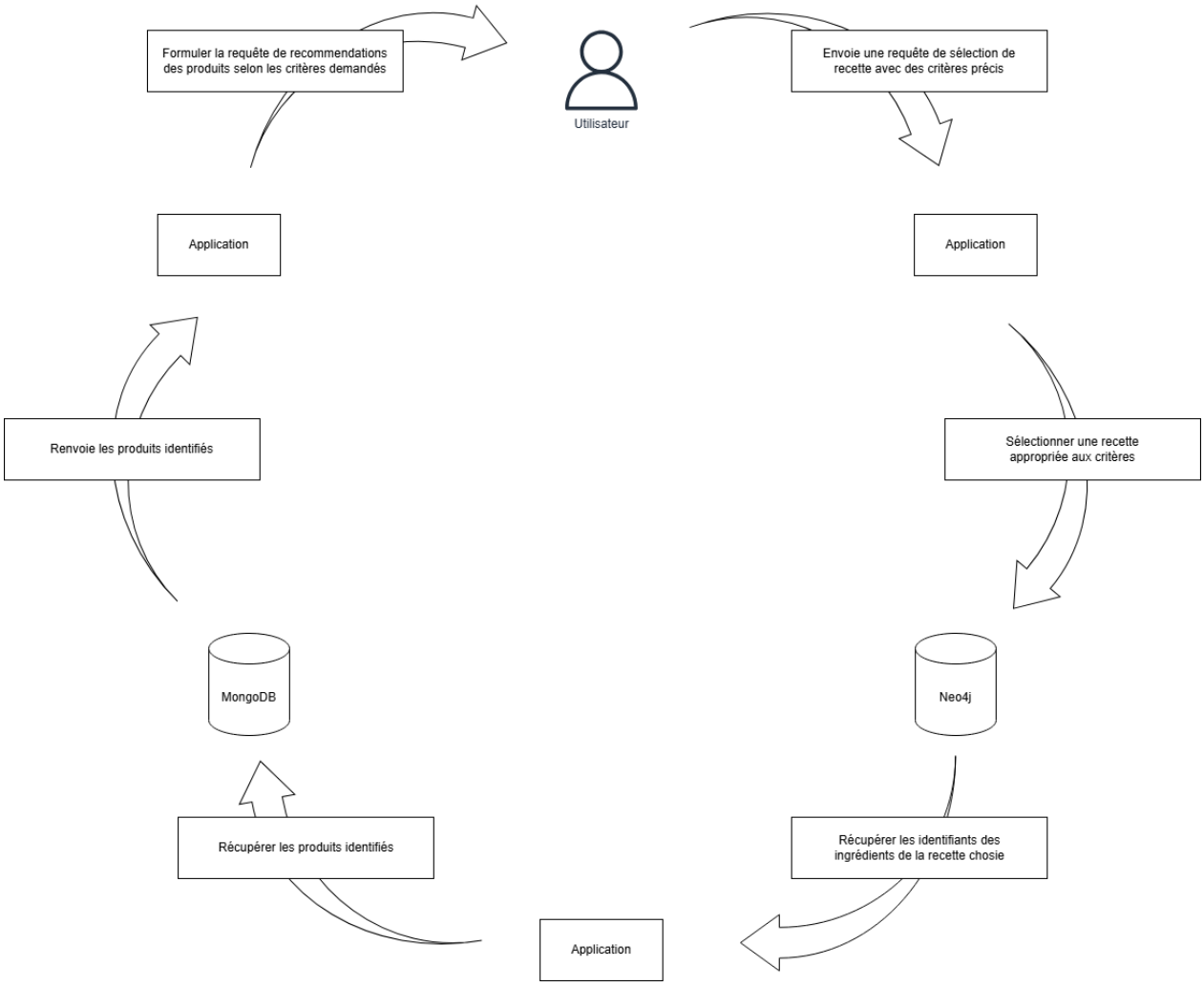
Étapes du processus ETL

Annexe 11 – Classification des aliments du GAC

Catégories Niveau 1	Catégories Niveau 2	Catégories Niveau 3
1 - Légumes et Fruits	11 - Fruits	1120 - Fruits
	12 - Légumes	1210 - Légumes vert foncé
		1220 - Légumes jaune ou orange
		1230 - Légumes féculents
		1240 - Autres légumes
2 - Grains entiers, aliments à grains entiers et de blé entier	21 - Grains entiers (100 %)	2100 - Grains entiers 100%
	22 - Aliments à grains entiers et aliments de blé entier	2210 - Aliments à grains entiers
		2220 - Aliments de blé entier
3 et 4 - Aliments protéinés	3 - Aliments protéinés d'origine végétale	3200 - Yogourts végétaux
		3300 - Fromages végétaux enrichis
		3400 - Légumineuse
		3500 - Similis-viandes
		3600 - Noix et graines
		3700 - Autres aliments végétales
	4 - Aliments protéinés d'origine animale	4200 - Yogourts et kéfir
		4300 - Fromages
		4400 - Autres aliments laitiers
		4500 - Viandes rouges
		4600 - Viandes de gibier
		4700 - Volaille et oiseaux sauvages
		4710 - Œufs
		4800 - Poisson et fruits de mer
		4900 - Abats
5 - Boissons	51 - Boissons	5110 - Eau
		5120 - Boissons végétales enrichies
		5130 - Boissons végétales non enrichies
		5140 - Boissons végétales enrichies (protéines
		5150 - Laits
		5160 - Jus de fruits
		5170 - Jus de légumes
		5180 - Autres boissons

6 - Autres aliments	61 - Autres aliments d'origine végétale (qui ne contiennent pas suffisamment de protéines)	6100 - Autres aliments d'origine végétale (qui ne contiennent pas suffisamment de protéines)
	62 - Condiments, sauces et vinaigrettes à faible teneur en matières grasses	6200 - Condiments, sauces et vinaigrettes à faible teneur en matières grasses
	63 - Autres grignotines	6300 - Autres grignotines
	64 - Aliments à teneur plus élevée en sucre et/ou en gras	6400 - Aliments à teneur plus élevée en sucre et/ou gras
	65 - Aliments qui ne sont pas à grains entiers et de blé entier	6510 - Aliments enrichis qui ne sont pas à grains entiers et de blé entier
		6520 - Aliments non enrichis qui ne sont pas à grains entiers et de blé entier
	66 - Viandes transformées	6600 - Viandes transformées
7 - Graisses et huiles	7110 - Graisses et huiles non saturées	7110 - Graisses et huiles non saturées
	7120 - Graisses et huiles saturées et trans	7120 - Graisses et huiles saturées et trans
8 - Autres aliments qui ne sont pas classés	81 - Aliments pour bébés	8100 - Aliments pour bébés
	82 - Substituts de repas et suppléments	8200 - Substituts de repas et suppléments
	83 - Boissons alcooliques	8300 - Boissons alcooliques
	84 - Recettes du FCÉN	8400 - Recettes du FCÉN
	85 - Aliments divers	8500 - Aliments divers
9 - Aliments et boissons pour lesquels il manque des données pour la classification	99 - Aliments et boissons non classés	9990 - Aliments et boissons non classés

Annexe 12 – Diagramme de création et d’obtention de recommandations de produits



Processus de création et d'obtentions de recommandations de produits