

Équipe de travail - Équipe06- : Hiba Habri - 537066308 Hugo Mazéas - 537004098 Yassine Nouri - 537231375

Remise et présentation écrite du produit minimum viable

Travail présenté à Eloïse Prévôt Base de données avancés

> Faculté science et génie Université Laval 20 Décembre 2024

Table des matières

1.	Introduction	1
	1.1 Les produits alimentaires de base et d'Open Food Facts	1
	1.2 Problématiques	1
	1.3 Sommaire de la solution	1
	1.4 Présentation du rapport	2
	1.5 Public cible du rapport	2
2.	Stratégie d'acquisition des données	3
	2.1 Source et méthode d'extraction	3
	2.2 Exemples de données	4
3.	Technologies utilisées	5
	3.1 Langage de programmation	5
	3.1.1 NodeJS	5
	3.1.2 ExpressJS	5
	3.2 Bases de données	5
	3.2.1 MySQL	5
	3.2.2 MongoDB	6
4.	Les détails du processus d'extraction, de transformation et de conversion (ETL)	6
	4.1 Processus d'acquisition initiale des données	6
	4.2 Processus d'acquisition incrémental des données	7
	4.3 Processus de transformation des données	9
	4.3.1 MySQL :	9
	4.3.2 MongoDB	9
	4.4 Schéma du pipeline d'ETL	.10
5.	Les détails du pipeline de données	.10
	5.1 Création de recommandations de produits	.10
	5.1.1 Explication visuelle et textuelle	.10
	5.1.2 Algorithme	.10
	5.1.3 Diagramme de séquence	.11
	5.2 Obtention de recommandations de produits	.11
	5.2.1 Explication visuelle et textuelle	.11
	5.2.2 Algorithme :	.12

5.2.3 Diagramme de séquence circulaire1	2
6. Une explication du plan d'expansion	2
6.1 Description1	2
6.1.1 Scénario d'expansion réaliste	2
6.2.2 Deux métriques de charge cohérentes avec le scénario	3
6.2.3 Impact attendu de l'expansion sur les métriques	3
6.2 Réplication1	4
6.2.1 Type de réplication	4
6.2.2 Impact sur le volume de requêtes par seconde (RPS) :	4
6.3 Partitionnement1	4
6.3.1 Fragmentation de la base MySQL :1	4
6.3.2 Fragmentation de la base MongoDB :	4
6.3.3 Impact sur le volume de requêtes par seconde (RPS) :1	5
6.4 Sécurité1	5
6.4.1 : Cas d'utilisation 1 : Injection SQL dans la base MySQL1	5
6.4.2 Cas d'utilisation 2 : Modification illégitime des données de produits alimentaires1	5
7. Fonctionnalités additionnelles avancées	6
7.1 Interfaces web:1	6
7.2 Authentification:1	6
7.3 Script Bash pour l'ajout de recette :1	6
7.4 Recherche1	6
Annexe 0 : Résultats Remise 21	8
Annexe A: Extrait de données Open Food Facts1	9
Annexe B: Extrait de USDA Global Branded Food Products Database2	0
Annexe C: Extrait de données Recettes2	1
Annexe D: Fonctionnement NodeJS2	2
Annexe E : Fonctionnement ExpressJS2	3
Annexe F : UML filtrage JSON2	4
Annexe G: UML Scraping site web des recettes	5
Annexe H : Script Bash nettoyage liste ingrédients2	6
Annexe I : Transformation CSV-MySQL2	7
Annexe J : Script Bash mise à jour liste ingrédients2	8
Annexe K : Document OFF transformé2	9
Annexe L : Script de transformation de la BD USDA Branded	0

Annexe M : BD USDA Branded transformée	31
Annexe N : ETL général	32
Annexe O-1 : Diagramme de séquence création de recommandations	33
Annexe O-2 : Diagramme de séquence obtention de recommandations	33
Annexe P : Fragmentation BD MySQL	34
Annexe Q : Script Bash ajout recette BD MySQL	35
Annexe R : Page d'authentification	36
Annexe S : Page de recherche en action	37
Annexe T : Interfaces Web	38
Bibliographie	41

1. Introduction

1.1 Les produits alimentaires de base et d'Open Food Facts

Open Food Facts est une base de données collaborative qui répertorie des informations détaillées sur des milliers de produits alimentaires. Fondée en 2012, elle permet aux utilisateurs d'accéder à des données telles que la composition des produits, leur valeur nutritionnelle et leur impact écologique. Ces informations contribuent à sensibiliser les consommateurs aux choix alimentaires durables et sains.

Les produits alimentaires de base incluent des aliments essentiels à la préparation de recettes courantes, comme les fruits, légumes, céréales, et produits laitiers. Leur identification et classification dans une base de données permettent de mieux répondre aux besoins des consommateurs et des cuisiniers.

1.2 Problématiques

Dans un monde où les consommateurs recherchent des choix alimentaires éclairés et personnalisés, il est difficile de :

- Accéder à des informations fiables sur les produits alimentaires en temps réel.
- ldentifier des produits correspondant précisément aux besoins culinaires ou nutritionnels.
- ➤ Intégrer efficacement des données provenant de sources multiples pour proposer des recommandations pertinentes.

L'application vise à résoudre ces problématiques en fusionnant des bases de données variées (Open Food Facts, USDA, recettes) et en offrant des algorithmes de recommandation précis.

1.3 Sommaire de la solution

La solution proposée repose sur un pipeline de données ETL (Extract, Transform, Load) bien défini, capable de gérer des sources multiples et de répondre aux exigences d'un système polyglotte. Les principales étapes incluent :

1. Extraction des données depuis des sources hétérogènes.

- 2. Transformation des données pour les nettoyer, les normaliser et les enrichir.
- 3. Chargement des données transformées dans une base MySQL pour les recettes et une base MongoDB pour les produits alimentaires. De plus, des algorithmes dédiés sont conçus pour proposer des produits adaptés en fonction des ingrédients de recettes ou des besoins spécifiques des utilisateurs.

1.4 Présentation du rapport

Ce rapport présente les détails de l'implémentation de la solution. Les sections suivantes décrivent :

- Les stratégies d'acquisition des données.
- Les technologies utilisées pour gérer et intégrer les données.
- ❖ Les détails du pipeline ETL, incluant les processus incrémentaux et les transformations appliquées.
- Les algorithmes permettant de générer des recommandations basées sur les données collectées. Chaque section est accompagnée d'exemples de code pour faciliter la compréhension.

1.5 Public cible du rapport

Ce rapport s'adresse:

- ✓ Aux développeurs et ingénieurs souhaitant comprendre les aspects techniques du projet, notamment l'intégration de bases de données polyglottes et l'implémentation des algorithmes de recommandation.
- ✓ Aux décideurs et responsables produits intéressés par les applications pratiques des données alimentaires pour améliorer l'expérience utilisateur et promouvoir des choix alimentaires sains et durables.
- ✓ Aux enseignants et évaluateurs évaluant l'approche méthodologique et technique adoptée pour résoudre une problématique réelle.

2. Stratégie d'acquisition des données

2.1 Source et méthode d'extraction

Pour notre projet l'extraction des données s'est faite à partir de trois sources principales : Open Food Facts [1], USDA Global Branded Food Products Database [2] et le guide alimentaire canadien pour les recettes chacune ayant apporté des informations complémentaires pour répondre aux objectifs d'analyse des produits alimentaires au Québec.

- Open Food Fact "OFF": est une base de données collaborative et libre qui recense des informations sur des milliers de produits alimentaires à travers le monde. Fondée en 2012, elle permet aux consommateurs d'accéder à des données détaillées sur la composition des produits, leurs ingrédients, leur valeur nutritionnelle. Les utilisateurs peuvent contribuer en scannant les codes-barres des produits ou en ajoutant des informations manuellement, favorisant ainsi une transparence accrue dans l'industrie alimentaire. OFF s'engage également à offrir des outils pour évaluer la qualité nutritionnelle des produits à l'aide de scores comme le Nutri-Score, tout en soutenant des initiatives de sensibilisation aux choix alimentaires plus sains et durables. Grâce à sa nature ouverte et collaborative, OpenFoodFacts participe à l'amélioration de l'accès à des informations fiables, aidant ainsi les consommateurs à faire des choix éclairés pour leur santé et pour la planète. OFF viens avec plusieurs méthodes de téléchargements, parmi ces méthodes nous avons choisi de télécharger le fichier JSONL disponible ici.
- Le Guide alimentaire canadien, mis à jour régulièrement par Santé Canada, est un outil essentiel pour promouvoir une alimentation saine et équilibrée. Il recommande de privilégier la diversité alimentaire en incluant une grande variété d'aliments dans chaque repas, notamment des fruits et légumes, des protéines de qualité (comme les légumineuses, les viandes maigres et les poissons), ainsi que des grains entiers. Le Guide encourage également à limiter les aliments ultra-transformés, riches en sucres ajoutés, en sel et en graisses saturées. Lors de l'élaboration de recettes. Pour extraire ces recettes il va falloir scraper le site de guide alimentaire pour récupérer la liste de recette.

Le USDA Global Branded Food Products Database est une initiative de partenariat public-privé visant à améliorer la santé publique et le partage de données ouvertes. Ce projet vise à enrichir le USDA National Nutrient Database avec des informations sur la composition des nutriments et les ingrédients des produits alimentaires de marque et de marque privée fournis par l'industrie alimentaire. Le but est de fournir un accès public à ces données, permettant une évaluation précise de l'étendue et de la fluidité du système alimentaire. La base de données inclut des informations telles que le nom du produit, la taille des portions, les nutriments indiqués sur l'étiquette Nutrition Facts, la liste des ingrédients, et une date associée à la formulation la plus récente du produit. [10]

Cette source a été recommandée par notre point de contact avec OFF. Cette base de données existe sous deux formats CSV ou JSON, nous avons opté pour le téléchargement du format JSON pour sa facilité d'importation avec MongoDB puisque nous avons déjà une collection pour Open Food Facts dans MongoDB.

2.2 Exemples de données

Un exemple extrait de Open Food Facts est présenté dans la part <u>annexe A</u>.

Un exemple extrait de guide alimentaire de recette est présenté par <u>l'Annex C</u>. Les données n'existent pas sous forme de fichiers prêt à utiliser, on a dû utiliser la méthode de **scraping** pour obtenir un fichier CSV dont on a présenté une partie.

Un exemple extrait du USDA Global Branded Food Products Database est présenté par <u>l'annexe B</u>.

La base de données OFF ainsi que celle USDA Global Branded Food Products Database comprennent plusieurs informations non utiles pour notre contexte de projet, ainsi que plusieurs aliments ont plus de deux scores qui sont à nul ce qui rend l'existence de ces aliments dans notre base de données moins pertinentes. Nous allons essayer à appliquer des filtres pour réduire ces informations inutiles.

3. Technologies utilisées

Pour réaliser ce travail, nous utilisons un environnement Docker avec VSCode.

3.1 Langage de programmation

3.1.1 NodeJS

Node.js est un environnement d'exécution JavaScript open-source et multiplateforme. [3]

Node.js est une bibliothèque permettant d'exécuter des applications web en dehors du navigateur du client. Ryan Dahl l'a développé en 2009. Les développeurs utilisent Node.js pour créer des applications web côté serveur, et il est parfait pour les applications gourmandes en données puisqu'il utilise un modèle asynchrone, piloté par les événements. [4]

La figure de l'annexe 1 montre comment fonctionne NopdeJS. [5]

3.1.2 ExpressJS

Express.js est un framework web rapide, flexible et minimaliste pour Node.js. Il s'agit en fait d'un outil qui simplifie la création d'applications web et d'API en utilisant JavaScript côté serveur. Express est un logiciel libre développé et maintenu par la fondation Node.js.

Express.js offre un ensemble de fonctionnalités robustes qui améliorent la productivité et rationalisent les applications web. Il facilite l'organisation des fonctionnalités de l'application à l'aide d'intergiciels et de routages. Il ajoute des utilitaires utiles aux objets HTTP Node et facilite le rendu des objets HTTP dynamiques. [6]

L'annexe 2 montre l'utilité de l'utilisation de Expresse JS avec NodeJS. [7]

3.2 Bases de données

3.2.1 MySQL

Pour la base de données des recettes, nous avons choisi MySQL.

MySQL offre une **haute fiabilité** des données, minimisant les risques de corruption et garantissant la cohérence des informations en respectant les propriétés ACID. En termes de **maintenabilité**, MySQL propose des outils performants et une documentation exhaustive,

facilitant ainsi l'administration, la gestion des bases de données et le dépannage. Enfin, son architecture modulaire et sa compatibilité avec diverses extensions et plug-ins permettent une **grande extensibilité**, répondant aux besoins croissants de notre application. [8]

3.2.2 MongoDB

Pour notre base principale, celle d'Open Food Fact, nous avons opté pour l'utilisation de MongoDB. Tout d'abord, il s'agit du type de base de données déjà utilisé par OFF, bien qu'ils proposent divers formats de fichiers tels que CSV, mongodump ou JSONL. De plus, la base de données OFF est **très volumineuse**, et MongoDB facilite la gestion de grandes quantités de données sous format BSON, ce qui le rend à la fois extensible et maintenable.

Par ailleurs, notre objectif est de fusionner au moins deux sources différentes, impliquant des **structures de données variées**. MongoDB, étant une base de données **documentaire**, offre une **flexibilité accrue** dans ce contexte. C'est pourquoi les deux sources autres que les recettes seront transformées et importées dans MongoDB.

En outre, MongoDB propose nativement la réplication et le partitionnement des données, ce qui renforce les principes d'extensibilité et de fiabilité.

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

4.1 Processus d'acquisition initiale des données

L'acquisition initiale de données a était faite à partir des trois sites déjà mentionnés dans la section 2.

Pour l'OFF, nous avons téléchargé le fichier JSONL qui contient des données partout dans le monde c'est une base énorme de 10g. Chaque aliment est présenté par plusieurs clefs comme son ecosore_score et nutriscore_grade.

Pour l'USDA Global Branded Food Products Database, on a téléchargé un fichier JSON avec comme information : description, foodNutritients, foodNutritientSource, amount, dataType, etc...

Et enfin pour les recettes, on a gardé le nom, les ingrédients et l'image de chaque recette en utilisant le scraping.

On a présenté les 2 processus par un diagramme UML en <u>annexe F</u> et <u>annexe G</u>.

4.2 Processus d'acquisition incrémental des données

L'acquisition incrémentale, ou développement incrémental, est une approche où un système ou un produit est construit et amélioré par étapes successives, appelées incréments. Chaque incrément ajoute de nouvelles fonctionnalités ou améliore les fonctionnalités existantes sans interruption du service.

Pour notre application, nous proposons 2 méthodes pour chaque BD:

MySQL: Détection des Changements

Méthode 1 : Utilisation de la colonne last updated

- En ajoutant une colonne **last_updated**, elle peut être utilisée pour identifier les enregistrements ajoutés ou modifiés depuis le dernier import.
- Processus:
- 1. Stocker la date/heure du dernier import dans une variable (ex. : dernier_import).
- 2. Exécuter une requête SQL pour extraire uniquement les lignes mises à jour après cette date :

SELECT * FROM recettes WHERE last_updated > 'dernier_import';

3. Charger ces données dans le pipeline de synchronisation pour traitement.

Méthode 2 : Utilisation des journaux binaires (Binlogs)

- Les binlogs enregistrent toutes les modifications apportées aux tables MySQL (INSERT, UPDATE, DELETE).
- Avantages:
 - Permet un suivi des changements plus précis.
 - o Réduit le besoin de colonnes spécifiques dans la base de données.

• Processus:

- 1. Configurer MySQL pour activer les binlogs.
- 2. Utiliser un outil comme Debezium ou MySQL Binary Log Client pour lire ces journaux en temps réel.
- 3. Filtrer les modifications spécifiques à notre table de recettes.

MongoDB: Détection des Changements

Méthode 1 : Utilisation des Timestamps

- Nous ajouterons aux documents MongoDB un champ date_modified ou similaire, il peut être utilisé pour détecter les changements.
- Processus:
 - 1. Stocker la date/heure du dernier import dans une variable (ex. : lastImportDate).
 - Exécuter une requête pour récupérer les documents modifiés après cette date :
 db.aliments.find({ date_modified: { \$gt: lastImportDate } });
 - 3. Traiter les documents récupérés pour les synchroniser.

Méthode 2 : Utilisation de Change Streams

- MongoDB propose un mécanisme natif pour suivre les changements en temps réel via des Change Streams.
- Avantages :
 - Idéal pour des besoins en temps réel.
 - Permet de détecter les INSERT, UPDATE et DELETE.
- Processus :
- 1. Activez les Change Streams sur votre collection MongoDB.
- 2. Configurez un script pour écouter les événements et capturer les changements.

```
const changeStream = db.aliments.watch();
changeStream.on('change', (change) => {
  console.log(change); });
```

4.3 Processus de transformation des données

4.3.1 MySQL:

Pour la base de données des recettes après avoir faire le scraping, le résultat était un fichier CSV. Nous avons commencé par appliquer **un script Bash (annexe H)** pour nettoyer la liste des ingrédients pour chaque recette pour les préparer à la fonction de recommandation. Le script supprime toutes les unités de mesure telle que ml, l, boite etc. et les numéros. Il supprime aussi les espaces vides, les propositions non utile comme « d'» etc..

On a utilisé ce fichier transformé et le converti en SQL en utilisant le site web « https://convertcsv.com/csv-to-sql.htm » pour créer la BD via ce site web. Le résultat est illustré dans l'annexe I.

Après avoir créer la BD mysql, au fil de temps, bous avons constaté des ingrédients qui ne sont pas bien nettoyé, nous avons fais le nettoyage manuellement car se sont des petits changements uniques sans pattern global, mais nous avons automatiser la mise à jour de la BD par le script Bash de l'annexe J.

4.3.2 MongoDB

Pour les 2 base de données OFF et USDA Global Branded Food Products Database, nous avons appliqué quelques commandes « **jq** » (json query) pour filtrer les données.

Tout d'abord pour OFF, un premier filtre pour faire sortir les données concernant le Canada :

```
zcat openfoodfacts-products.jsonl.gz | jq -c '. | select(.countries_tags[]? == "en:canada")' > canada2.jsonl
```

Puis, garder les aliments qu'ont de Nutriscore. Et enfin garder quelques attributs utiles pour cette phase comme le nom de l'aliment, ses scores et sa catégorie :

```
cat canada2.jsonl | jq -c '. | select(.misc_tags[]? == "en:nutriscore-computed") | [.code,.product_name,.ecoscore_score,.nutriscore_grade,.nova_group,.categories,.brands] ' > canada-filtered.jsonl
```

Le résultat final est présenté dans <u>l'annexe K</u>.

Pour l'USDA Global Branded Food Products Database, nous avons gardé pour chaque aliment : sa description, et pour chaque foodNutrient, son nom, son rank, et sa valeur en utilisant le script de <u>l'annexe L</u>. Le résultat est illustré dans <u>l'annexe M</u>.

4.4 Schéma du pipeline d'ETL

ETL, pour Extract Transform Load, son but est de gérer et intégrer des données provenant de diverses sources pour les rendre exploitables dans un système cible. Dans notre projet, nous allons modifier deux fichiers JSON et scraper un site web pour avoir nos 3 bases de données.

Les deux fichiers JSON transformés avec la commande « jq » et un script JS puis chargés dans mongodb avec la commande mongoimport.

Pour la base de données recettes, nous avons l'importé directement à MySQL avec la commande **source** après avoir faire le web scraping et exécuter un script Bash.

Le schéma général est disponible en <u>annexe N</u>.

5. Les détails du pipeline de données

5.1 Création de recommandations de produits

5.1.1 Explication visuelle et textuelle

Les recommandations de produits sont générées à partir des ingrédients d'une recette. Le processus est divisé en deux étapes principales : la correspondance des ingrédients avec des aliments dans la base de données et la sélection des produits en fonction des préférences de l'utilisateur.

5.1.2 Algorithme

Étape 1 : Correspondance des ingrédients

- 1. Entrée : Liste d'ingrédients extraits d'une recette (issus de la base MySQL).
- 2. Traitement:

Chaque ingrédient est recherché dans la base MongoDB en utilisant une requête basée sur des expressions régulières.

Les correspondances incluent des aliments associés à chaque ingrédient dans la base Open Food Facts.

Une fois les correspondances trouvées, des métadonnées (comme le nutriscore, les catégories bio/locales) sont récupérées.

3. Sortie : Liste de correspondances entre les ingrédients et les aliments.

Étape 2 : Sélection des meilleurs produits

- 1. Entrée : Liste des aliments correspondants pour chaque ingrédient.
- 2. Traitement:

Les aliments sont filtrés selon les préférences de l'utilisateur (par exemple, bio, faible en gras, local).

Chaque aliment est trié en fonction de critères tels que le nutriscore, le prix ou les avis utilisateurs.

3. Sortie : Liste des produits recommandés pour la recette.

5.1.3 Diagramme de séquence

Veuillez trouver le diagramme de séquence dans l'annexe O-1.

Le diagramme illustre le flux de données suivant :

- 1. Extraction des ingrédients d'une recette depuis la base MySQL.
- 2. Recherche d'aliments correspondants dans MongoDB via des requêtes.
- 3. Filtrage et tri des aliments selon les critères utilisateur.

5.2 Obtention de recommandations de produits

5.2.1 Explication visuelle et textuelle

Les recommandations sont obtenues après la création des produits recommandés. Elle inclut une explication textuelle et un diagramme de séquence pour détailler les interactions entre les bases de données et l'application.

5.2.2 Algorithme:

Étape 1 : Recherche des aliments nécessaires pour une recette

1. Entrée : Identifiant de la recette.

2. Traitement:

Les ingrédients de la recette sont extraits de MySQL.

Les aliments correspondants sont recherchés dans MongoDB.

3. Sortie : Liste des aliments nécessaires pour réaliser la recette.

Étape 2 : Retour des produits optimaux

1. Entrée : Liste des aliments nécessaires.

2. Traitement:

Les aliments sont filtrés et triés pour ne conserver que les produits respectant les préférences utilisateur.

La liste finale est structurée et renvoyée au front-end.

3. Sortie : Liste des recommandations affichées à l'utilisateur.

5.2.3 Diagramme de séquence circulaire

Le diagramme de <u>l'annexe O-2</u> détaille les étapes suivantes :

- 1. Interaction utilisateur : sélection d'une recette.
- 2. Extraction des données depuis MySQL.
- 3. Correspondance et filtrage des aliments via MongoDB.
- 4. Renvoi des recommandations structurées au front-end.

6. Une explication du plan d'expansion

6.1 Description

6.1.1 Scénario d'expansion réaliste

Dans une optique d'expansion, l'application pourrait être déployée à l'échelle internationale pour intégrer les données d'OFF pour d'autres pays (nous avons commencé par les données de

l'USA) et d'autres données provenant de nouvelles sources alimentaires (bases locales ou régionales) et répondre à des besoins spécifiques des consommateurs dans différentes zones géographiques. Cette expansion inclurait l'ajout de nouvelles recettes (exemple <u>Ciqual</u> de la France), de produits locaux, et de critères de personnalisation plus avancés, comme les préférences culturelles ou les restrictions alimentaires.

6.2.2 Deux métriques de charge cohérentes avec le scénario

- 1. Volume de requêtes par seconde (RPS):
- Avec une expansion internationale, le nombre de requêtes simultanées augmentera considérablement.
 - Cette métrique permettra de suivre la capacité du système à gérer un flux de demandes accru.
 - 2. Temps moyen de réponse (RTT) :
 - Le temps nécessaire pour générer une recommandation, incluant l'extraction des données, leur transformation, et la livraison des résultats.
 - Une expansion nécessitera de maintenir un temps de réponse faible malgré l'augmentation du volume de données.

6.2.3 Impact attendu de l'expansion sur les métriques

- 1. Volume de requêtes par seconde (RPS):
- Augmentation significative due à un public plus large et des zones géographiques supplémentaires.
- Des solutions comme le scaling horizontal (ajout de serveurs) et la mise en cache des résultats fréquents seront nécessaires.
 - 2. Temps moyen de réponse (RTT) :
- Risque d'augmentation si les optimisations ne sont pas effectuées, comme le pré-chargement des données critiques ou l'utilisation de bases de données distribuées.

- Des optimisations d'algorithmes, comme la réduction du nombre de requêtes par cycle, seront mises en œuvre pour limiter cet impact.

6.2 Réplication

6.2.1 Type de réplication

Pour assurer une disponibilité élevée et une tolérance aux pannes, l'application utiliserait une réplication « maître-esclave » / « single leader », la plus répandue pour les bases de données MySQL et pour Mongodb et vu que nous n'avons pas trop d'écritures fréquentes. Dans ce modèle, un serveur principal (maître) gère toutes les écritures, tandis que les serveurs secondaires (esclaves) reçoivent une copie des données pour gérer les lectures.

6.2.2 Impact sur le volume de requêtes par seconde (RPS) :

- La réplication maître-esclave permettra de répartir la charge des requêtes en lecture sur les serveurs esclaves, augmentant ainsi la capacité du système à gérer un volume accru de requêtes simultanées.
- En cas de panne du serveur maître, un basculement rapide vers un esclave minimisera l'interruption du service, maintenant une expérience utilisateur fluide même en cas de défaillance.

6.3 Partitionnement

6.3.1 Fragmentation de la base MySQL:

Pour la BD des recettes, puisque les catégories sont assez réduites, nous optons pour une « fragmentation par Liste ». Cela permet de séparer les recettes en partitions distinctes basées sur leurs catégories. L'exemple de la création de la table avec fragmentation est démontré par l'annexe P.

6.3.2 Fragmentation de la base MongoDB:

Pour les produits alimentaires de MongoDB, chaque fragment contiendra des produits basés sur des critères géographiques ou catégoriels (exemple : produits locaux, bio, ou par région) qu'on va ajouter avec l'expansion de notre application, exemple :

{ name: "Carotte", category: "bio", region: "Ontario", nutriscore: "B" },

Cela sera fait avec l'activation de l'option « Sharding » du mongodb :

sh.enableSharding("off db")

Et l'affecter par exemple par région :

sh.shardCollection("off_db.aliments", { region: 1 })

6.3.3 Impact sur le volume de requêtes par seconde (RPS) :

- Le partitionnement permet de réduire la charge sur chaque nœud de base de données en dirigeant les requêtes vers les fragments pertinents uniquement.
- Cela améliore l'évolutivité en permettant des écritures et des lectures parallèles sur plusieurs fragments.
- En conséquence, le système pourra gérer un nombre beaucoup plus élevé de requêtes simultanées, avec un impact positif direct sur le RPS.

6.4 Sécurité

6.4.1 : Cas d'utilisation 1 : Injection SQL dans la base MySQL

- <u>Risque</u>: Un attaquant pourrait exploiter des failles dans les requêtes SQL pour exécuter des commandes malveillantes, compromettant ainsi la confidentialité et l'intégrité des données.

- Stratégie de mitigation :

- Utilisation de requêtes préparées pour éviter l'injection de commandes SQL malveillantes.
- Validation stricte des entrées utilisateur pour s'assurer qu'elles respectent les formats attendus.
- Implémentation d'un pare-feu d'applications web (WAF) pour détecter et bloquer les tentatives d'injection SQL.

6.4.2 Cas d'utilisation 2 : Modification illégitime des données de produits alimentaires

- **Risque** : Un attaquant pourrait altérer les informations des produits dans MongoDB, compromettant ainsi l'intégrité des recommandations.

- Stratégie de mitigation :

- Mise en place d'un système d'audit pour enregistrer toutes les modifications effectuées sur les données.
- Activation de la fonctionnalité `readOnly` pour les utilisateurs qui n'ont pas besoin d'accéder aux fonctionnalités d'écriture dans MongoDB.

7. Fonctionnalités additionnelles avancées

7.1 Interfaces web:

L'interface web permet de rechercher des recettes, explorer des ingrédients et accéder aux produits par marque via une navigation intuitive. Elle simplifie l'exploration des données transformées en proposant des ingrédients répertoriés pour chacune des recettes. L'interface peut être consulté à l'annexe T.

7.2 Authentification:

L'authentification est requise uniquement pour accéder aux pages de l'interface utilisateur. L'utilisateur unique (username : prof password : prof) permettent une connexion simple, idéale pour limiter l'accès au fonctionnalités avancées. L'utilisateur, bien qu'étant unique, est entreposé dans la base de données MySQL. La page d'authentification à l'application est montrée à <u>l'annexe</u> R.

7.3 Script Bash pour l'ajout de recette :

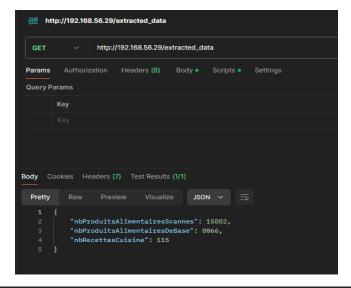
Nous avons préparé le script Bash précédemment mentionnée et l'adopter pour ajouter des recettes par la suite sans perturber le fonctionnement de notre application. Une partie du script est démontrée par <u>l'annexe Q</u>.

7.4 Recherche

La fonctionnalité de recherche dans les tableaux est assurée par une bibliothèque JavaScript nommée DataTables. Elle permet une recherche instantanée à travers toutes les colonnes, avec la

prise en charge des mots partiels, offrant une méthode efficace pour interagir avec les données présentées puis avoir un large éventail de résultat plus tôt que de risquer d'afficher aucun résultat. Une démonstration de la page recherche en action est présentée à <u>l'annexe S.</u>

Annexe 0: Résultats Remise 2



```
http://192.168.56.29/transformed_data
 Params Authorization Headers (8) Body • Scripts • Settings
 Query Params
Body Cookies Headers (7) Test Results (1/1)
  Pretty
                  "Nova": 9005,
"EcoScore": 9077
             },
"categoriesProduitAlimentaire": {
                  "70": 1,
"130": 1,
                  "140": 2,
                 "160": 1,
"210": 1,
                 "Pâtes farcies au fromage": 1,
"Dried-seaweed-unspecified": 1,
                 "Mackerels": 1,
                  "Non-alcoholic beverages": 27,
                 "Brochettes": 1,
"Ingemaakte Augurken": 1,
                  "Cappuccino instantané non sucré": 1,
                  "Spanish vegetables": 1,
                  "Thick Egg noodles": 1,
                  "Salades de légumes": 1,
"Miels de sapin": 1,
                  "Dark chocolates": 168,
                  "Chocolate eggs": 7,
```

Annexe A: Extrait de données Open Food Facts

```
"_id": "0008577002786",
"ecoscore score": 64,
"food_groups_tags": [],
"data_quality_bugs_tags": [],
"data_sources_imported": "Databases, database-usda",
"last image t": 1660213050,
"ingredients_analysis_tags": [
 "en:palm-oil-free",
  "en:vegan",
 "en:vegetarian"
"lc_imported": "en",
"correctors_tags": [
 "usda-ndb-import",
  "fix-missing-lang-bot",
 "org-database-usda",
 "org-label-non-gmo-project",
 "fighter-food-facts"
"allergens_from_ingredients": "",
"serving quantity unit": "ml",
"allergens_tags": [],
"product_name_en_imported": "100% pure vermont organic maple syrup, dark color robust taste",
"manufacturing_places": "",
"nutrition_grades_tags": [
"nutriscore_grade": "d",
"cities_tags": [],
"created_t": 1460234918,
"last_editor": "fighter-food-facts",
"removed_countries_tags": [],
"ingredients_non_nutritive_sweeteners_n": 0,
"ingredients_text": "Pure organic maple syrup",
"emb_codes": "",
```

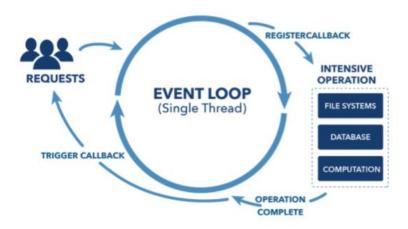
Annexe B: Extrait de USDA Global Branded Food Products Database

```
"foodClass": "Branded",
"description": "MANGO CHIA POD, MANGO",
"foodNutrients": [
       "type": "FoodNutrient",
       "id": 13796296,
       "nutrient": {
            "id": 1003,
            "number": "203",
           "name": "Protein",
           "rank": 600,
           "unitName": "g"
       "foodNutrientDerivation": {
            "code": "LCCS",
            "description": "Calculated from value per serving size measure",
            "foodNutrientSource": {
                "id": 9,
               "code": "12",
               "description": "Manufacturer's analytical; partial documentation"
        "amount": 1.76
```

Annexe C: Extrait de données Recettes

id	nomRecette	ingredientsRecette	imgUrl					
1	Salade de pois chiches et de carottes	Ingrédients 1 grand concombre, tranché 15 ml (1 c. à table) d'aneth frais haché ou 2 ml (½ c. à thé) d'aneth séché 2 carottes, épluchées et rāpées 1 boîte (540 ml/19 oz) de pois chiches, égouttés et rincés 500 ml (1 tasse) de tomates cerise, coupées en deux dans le sens de la longueur 60 ml (¾ tasse) de basilic frais haché 45 ml (3 c. à table) de vinaigre balsamique 15 ml (1 c. à table) de pesto de basilic 10 ml (2 c. à the) d'ilville d'olive extra vierge 1 gousse d'ail, émincée 1 ml (¼ c. à thé) de polvre moulu	https://guide-alimentaire.canada.ca/sites/default/f	iles/styles/cor	ntainer_width/	public/2020-07	/chickpea_carr	ot_salad.jpg
2	Pain aux bananes et aux noix de Grenoble	Ingrédients 2 œufs 85 ml (½ tasse) de sirop d'érable 5 ml (1c. à thé) d'extrait de vanille 125 ml (½ tasse) d'huile végétale 3 bananes, écrasées 250 ml (1 tasse) de farine de blé entier 190 ml (¾ tasse) de farine tout usage 5 ml (1 c. à thé) de bicarbonate de soude 5 ml (1 c. à thé) de cannelle 1 ml (¼ c. à thé) de sel 250 ml (1 tasse) de noix de Grenoble non salées, grillées et hachées	https://guide-alimentaire.canada.ca/sites/default/f	iles/styles/cor	ntainer_width/	public/2022-05	/Banana%20loa	af%203.jpg

Annexe D: Fonctionnement NodeJS

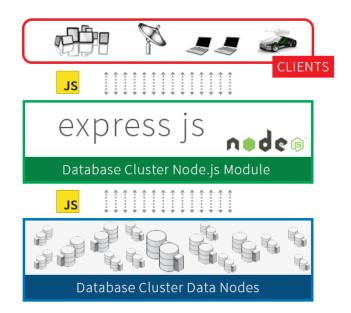


Fonctionnement NodeJS

Même s'il n'y a qu'une seule boucle d'événement, lorsqu'une demande est faite, la boucle transmet la demande à une fonction asynchrone qui effectue le travail. Lorsque cette fonction est terminée et qu'une réponse est renvoyée, elle peut alors être renvoyée à la boucle d'événement pour être exécutée par le callback et envoyée à l'utilisateur. Si les fonctions étaient synchrones, la boucle d'événements serait bloquée par la demande et la réponse d'un client, et tous les autres clients devraient attendre que ce client ait terminé. Grâce à la nature asynchrone de JavaScript, les applications utilisant Node peuvent gérer de nombreuses requêtes en même temps.

Cela signifie que lors de la programmation en Node.js, il est important de toujours garder à l'esprit que les fonctions écrites ne sont pas synchrones. Il est également très important d'attraper les erreurs sur le serveur avant qu'elles ne soient renvoyées au client. Cela permet d'éviter que des erreurs n'atteignent la boucle d'événements, ce qui pourrait faire planter le programme et tous les clients en souffriraient.

Annexe E: Fonctionnement ExpressJS

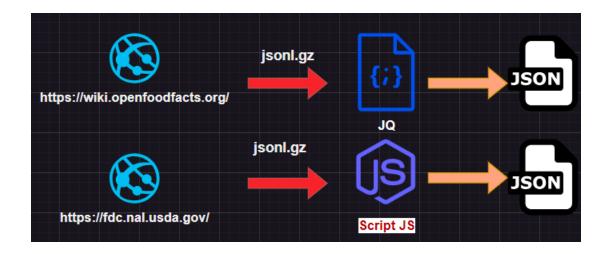


ExpressJS est un framework NodeJS préconstruit qui peut vous aider à créer des applications web côté serveur plus rapidement et plus intelligemment. Simplicité, minimalisme, flexibilité, évolutivité sont quelques-unes de ses caractéristiques et comme il est conçu en NodeJS, il a également hérité de ses performances.

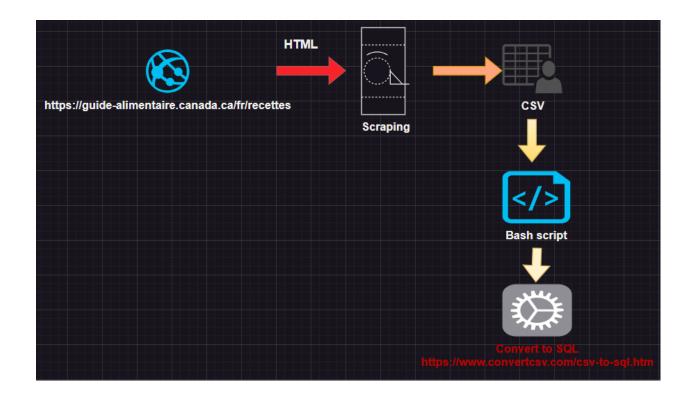
En bref, ExpressJS a fait pour NodeJS ce que Bootstrap a fait pour HTML/CSS et le responsive web design.

Il a fait du codage en NodeJS un jeu d'enfant et a donné aux programmeurs quelques fonctionnalités supplémentaires pour étendre leur codage côté serveur. ExpressJS est de loin le framework NodeJS le plus connu - à tel point que lorsque la plupart des gens parlent de NodeJS, ils veulent certainement dire NodeJS+ExpressJS.

Annexe F: UML filtrage JSON



Annexe G: UML Scraping site web des recettes



Annexe H: Script Bash nettoyage liste ingrédients

Annexe I: Transformation CSV-MySQL

Annexe J : Script Bash mise à jour liste ingrédients

```
$ mysql-inglistsh
1  #//bin/bash
2
3  # Paramètres de connexion MySQL
4  HOST="127.0.0.1"
5  USER="root"
6  PASSMORD="root_password"
7  DATABASE="recettes"
8  TABLE="recettes"
9
10  # Nom du fichier texte contenant les ingrédients
11  ##ILE="filtered_ingredients2.txt"
12  FILE="ingredients-comma.txt"
13  # Initialiser l'ID à 1
14  id=1
15
16  # tire le fichier ligne par ligne et mettre à jour la base de données
while IFS= read -r line
18  do
19  ingredients=$(echo "$line" | sed "s/'\\\'/g")
20  echo "Ingredients: $ingredients"
11  # Mise à jour de la colonne ingredientsList dans la table recettes
12  mysql -h "$HOST" -u "$USER" -p"$PASSMORD" "$DATABASE" -e "UPDATE $TABLE SET ingredientsRecette='$ingredients' WHERE id=$id;"
23  # Incrémenter l'ID
26  done < "$FILE"</pre>
```

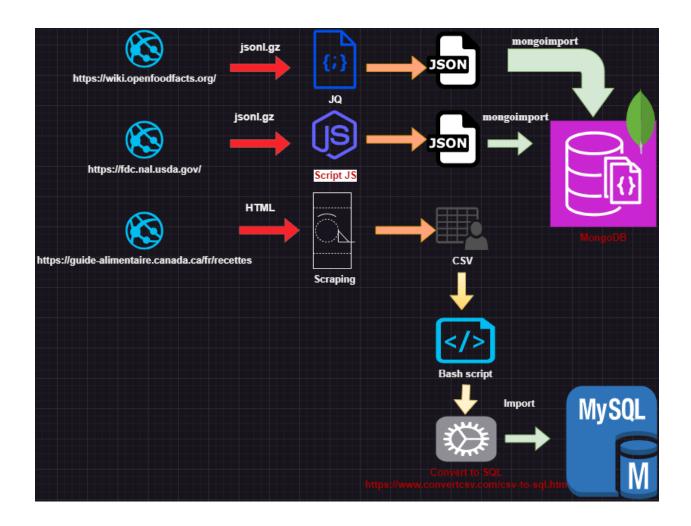
Annexe K: Document OFF transformé

```
{
   "_id": "0055872025019",
   "product_name": "Lait (2%)",
   "ecoscore_score": 48,
   "nutriscore_grade": "a",
   "nova_groups": "1",
   "categories": "Produits laitiers, Laits, Laits demi-écrémés",
   "brands": "Québon"
}
```

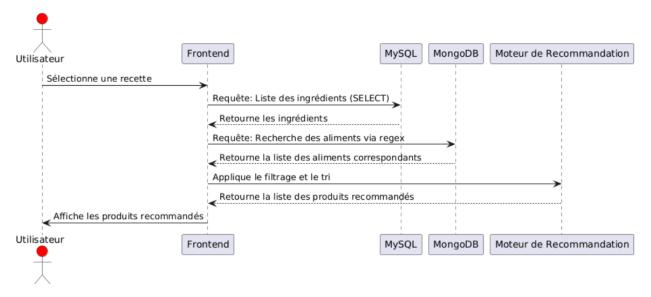
```
app > Js transform_branded_ingredients.js > ...
       const fs = require('fs');
       const readline = require('readline');
      // Paths for input and output files
       const inputFile = 'data/mongodb/brandedDownload.jsonl';
       const outputFile = 'data/mongodb/brandedDownload_transformed.jsonl';
      // Create a write stream for the output file
      const outputStream = fs.createWriteStream(outputFile);
 11
       const rl = readline.createInterface({
 12
           input: fs.createReadStream(inputFile),
           output: process.stdout,
           terminal: false
       });
 17
       rl.on('line', (line) => {
           const data = JSON.parse(line);
           const transformedData = {
               description: data.description,
               foodNutrients: data.foodNutrients.map(nutrient => ({
                   name: nutrient.nutrient?.name,
                   rank: nutrient.nutrient?.rank,
                   amount: nutrient.amount
               }))
           };
           // Write the transformed object as a JSON line to the output file
           outputStream.write(JSON.stringify(transformedData) + '\n');
       });
       rl.on('close', () => {
           outputStream.end();
           console.log('Transformation complete. Output saved to', outputFile);
       });
```

Annexe M: BD USDA Branded transformée

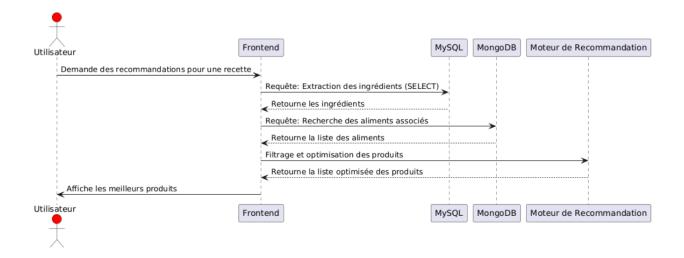
Annexe N : ETL général



Annexe O-1: Diagramme de séquence création de recommandations



Annexe O-2 : Diagramme de séquence obtention de recommandations



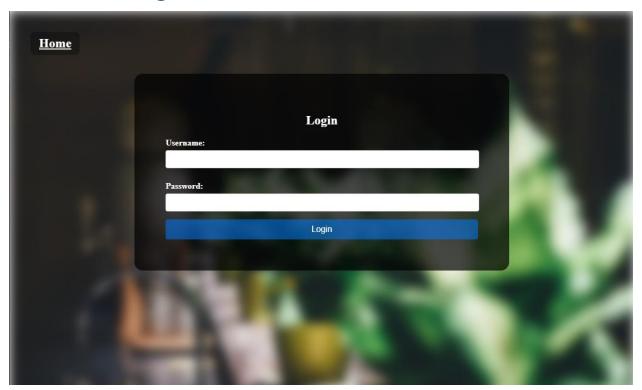
Annexe P: Fragmentation BD MySQL

```
CREATE TABLE recettes (
    id INTEGER NOT NULL PRIMARY KEY,
    nomRecette VARCHAR(350) NOT NULL,
    imgUrl VARCHAR(159) NOT NULL,
    categoriesRecette VARCHAR(150),
   tempsPrep VARCHAR(7) NOT NULL,
   tempsCuisson VARCHAR(7) NOT NULL,
    portions VARCHAR(50) NOT NULL,
    ingredientsRecette VARCHAR(1500),
    prepRecette VARCHAR(2500),
   descrRecette VARCHAR(2500)
PARTITION BY LIST (categoriesRecette) (
        PARTITION p 10ing VALUES IN ('10 ingrédients ou moins'),
        PARTITION p 30min VALUES IN ('30 min ou moins'),
        PARTITION p enfants VALUES IN ('Pour enfants'),
        PARTITION p végétarien VALUES IN ('Végétarien'),
        PARTITION p sanscuisson VALUES IN ('Sans cuisson'),
        PARTITION p congelables VALUES IN ('Congelable'),
        PARTITION p_autre VALUES IN ('autre', NULL)
```

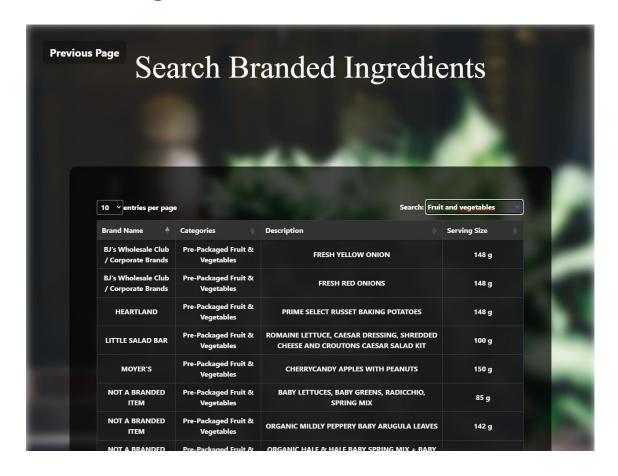
Annexe Q: Script Bash ajout recette BD MySQL

```
while IFS="," read -r nomRecette imgUrl categoriesRecette tempsPrep tempsCuisson \
                portions ingredientsRecette prepRecette ingredientsList descRecette
   # Échapper les ' dans chaque colonne
   nomRecette=$(echo "$nomRecette" | sed "s/'/\\'/g")
    imgUrl=$(echo "$imgUrl" | sed "s/'/\\'/g")
    categoriesRecette=$(echo "$categoriesRecette" | sed "s/'/\\'/g")
    ingredientsRecette=$(echo "$ingredientsRecette" | sed "s/'/\\'/g")
   prepRecette=$(echo "$prepRecette" | sed "s/'/\\'/g")
    ingredientsList=$(echo "$ingredientsList" | sed "s/'/\\'/g")
   descRecette=$(echo "$descRecette" | sed "s/'/\\'/g")
    # Insertion dans la table recettes
   mysql -h "$HOST" -u "$USER" -p"$PASSWORD" "$DATABASE" -e "
    INSERT INTO $TABLE (
       nomRecette,
        imgUrl,
        categoriesRecette,
        tempsPrep,
        tempsCuisson,
        portions,
        ingredientsRecette,
        prepRecette,
        ingredientsList,
        descRecette
     VALUES (
        '$nomRecette',
        '$imgUrl',
        '$categoriesRecette',
        $tempsPrep,
        $tempsCuisson,
        $portions,
        '$ingredientsRecette',
        '$prepRecette',
```

Annexe R : Page d'authentification



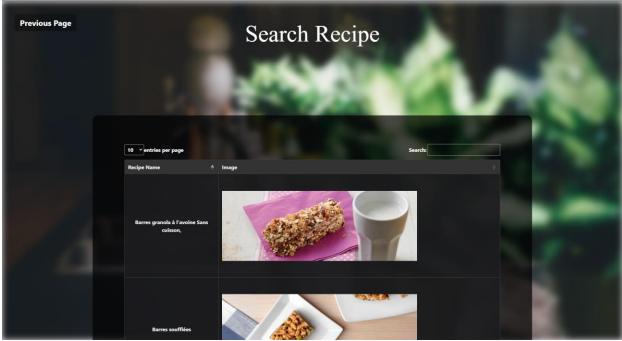
Annexe S: Page de recherche en action

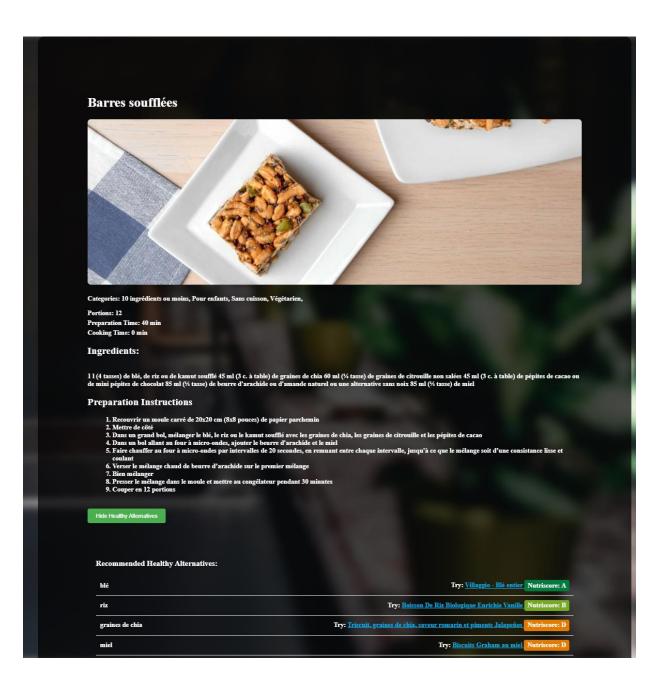


Ici nous voyons que la librairie a par elle-même choisi la colonne « Categories » pour faire la recherche puis a aussi détecté l'équivalence du symbole « & » pour « and » en anglais.

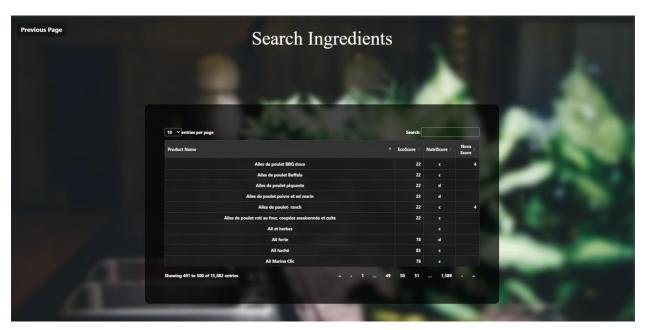
Annexe T: Interfaces Web

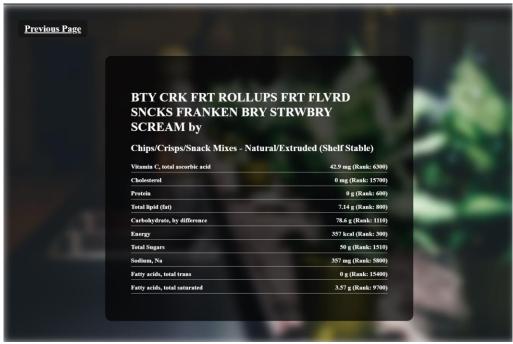












Bibliographie

- [1]: https://world.openfoodfacts.org/, consulté le 24 septembre 2024
- [2]: https://aliments-nutrition.canada.ca/cnf-fce/?lang=fre, consulté le 24 septembre 2024
- [3]: https://nodejs.org/en/about, consulté le 21 septembre 2024
- [4]: https://www.simplilearn.com/tutorials/nodejs-tutorial/what-is-nodejs#what_is_nodejs, consulté le 21 septembre 2024
- [5]: https://medium.com/@lukepierotti/node-js-what-is-it-and-how-does-it-work-e3e83d054662, consulté le 21 septembre 2024
- [6]: https://www.geeksforgeeks.org/express-js/, consulté le 21 septembre 2024
- [7]: https://www.algoworks.com/blog/why-use-expressjs-over-nodejs-for-server-side-coding/, consulté le 29 octobre 2024
- [8]: AI générative
- [9]: https://www.mongodb.com/company/what-is-mongodb, consulté le 29 septembre 2024
- [10]: https://fdc.nal.usda.gov/GBFPD Documentation.html, consulté le 29 septembre 2024