Projet chef d’œuvre

CollectSpirit

Certification Simplon 2021

Développer data

Apprenant :

Mayel Pèllé

Années :

2020-2021

SOMMAIRE

Introduction

* Enonciation du projet
* Genèse du projet
* Annonce du plan

Dev

1. Préparation du projet
   1. Phase de planification
      1. Phase macro
      2. Kanban
   2. Expression du besoin
      1. Etat de l’art
         1. Le marché du rhum
         2. Les collectionneurs
         3. Les applications existantes
      2. Le périmètre du projet

* Le choix du rhum pour le moment
* La granularité des données, pour trouver un maximum de données

1. Conception du projet
   1. L’application
      1. Schéma prévisionnel de l’app
      2. Schéma fonctionnel
      3. Mcd - MLd
      4. Choix des technos
   2. Sources des données
      1. Liste
      2. Collecte des données

* Fichier csv
* Scrapping
* Api ebay
  + 1. Rgpd
  1. Création de la BDD
  2. Préparation des données
     1. Pandas
     2. Regex
  3. Insertion des données
  4. Optimisation/ gestion des doublons

1. Exploitation des données
   1. Requête sql interrogation de la bdd
   2. Dash
2. **Conclusion**

# INTRODUCTION

Nous allons aborder dans ce rapport la thématique de la gestion d’une collection de spiritueux appliquée à un travail technique lié à la data. Le but est de concevoir une base de données permettant à un utilisateur de créer rapidement l’inventaire de sa collection.

L’idée du projet vient de mon histoire personnelle. J’ai travaillé pendant trois ans pour un producteur de rhum en Martinique appelé HSE. Cela m’a permis d’apprendre énormément sur l’univers des spiritueux. Et ce qui était d’abord un travail, c’est transformé petit à petit, de dégustation en dégustation, en passion. Au fil du temps j’ai commencé à acquérir des bouteilles de plus en plus prestigieuses et ainsi constituer une petite collection d’une centaine de références. Lorsqu’il a fallu déménager en France, j’ai dû établir « à la main » un inventaire de ma collection à l’aide d’un fichier Excel. La tâche c’est révélé plus ardu qu’anticipé. C’est à partir de cet événement qu’à germé l’idée de créer une application pour faciliter ce travail. Je profite de cette formation pour me lancer dans ce projet.

Dans un premier temps, rappelons que le rhum (anglais : rum, espagnol : ron) est une eau-de-vie originaire des Amériques, produite soit par distillation de sous-produits fermentés de l'industrie sucrière, aussi appelé mélasse : le rhum industriel ou traditionnel, soit à partir du jus de canne à sucre fermenté : le rhum agricole (Wikipédia).

Le secteur du rhum a beaucoup évolué ces dernières années, tant en volume de vente qu’en perception. Considéré pendant longtemps comme un alcool bas de gamme, il est désormais regardé comme un spiritueux aussi noble que le whisky ou le cognac.

Ce changement n’a pas échappé aux collectionneurs qui sont de plus en plus nombreux à partir à la découverte des joyaux des tropiques. Mais le monde du rhum excède largement les frontières des caraïbes. Disposer d’une base de données sur le rhum sera à la fois un moyen de réaliser l’inventaire de sa collection mais permettra aussi de découvrir de nouvelles références.

Pour ce projet, nous allons essayer de répondre à des problématiques tournant autour de la création d’une base de données d’un spiritueux et à la gestion d’une collection. Pour ce faire, nous allons analyser les envies et les attentes d’un collectionneur.

Ainsi, nous verrons dans ce rapport, les différentes étapes de la construction de ce projet. Elles sont au nombre de trois :

* La préparation du projet : la phase de planification et de recherche
* La conception du projet : construction de la base de données
* L’exploitation des données : l’applicatif.

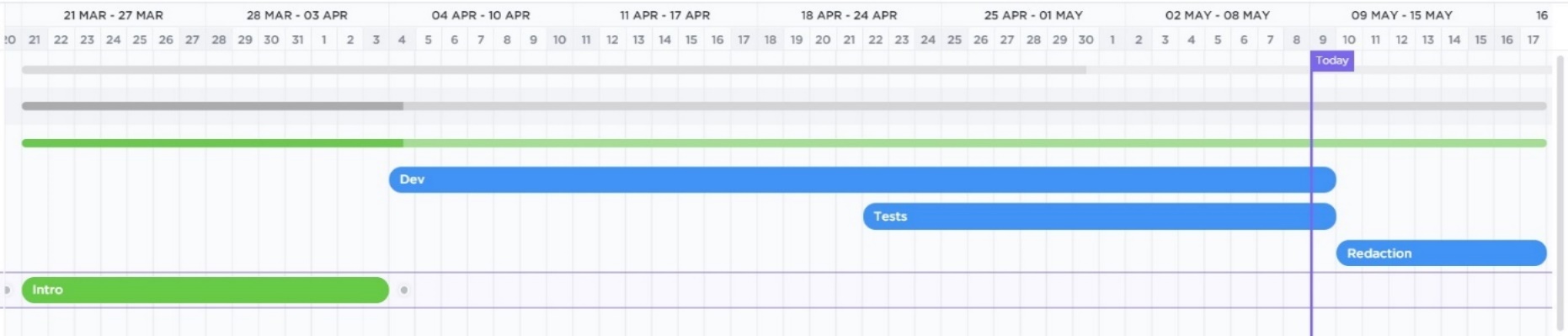
Nous allons ici nous focaliser essentiellement sur le rhum, car c’est mon domaine d’expertise. Mais je prévoie, pour des versions ultérieures, d’intégrer d’autres alcools.

1. Préparation du projet
2. La gestion du projet
   1. La phase de planification

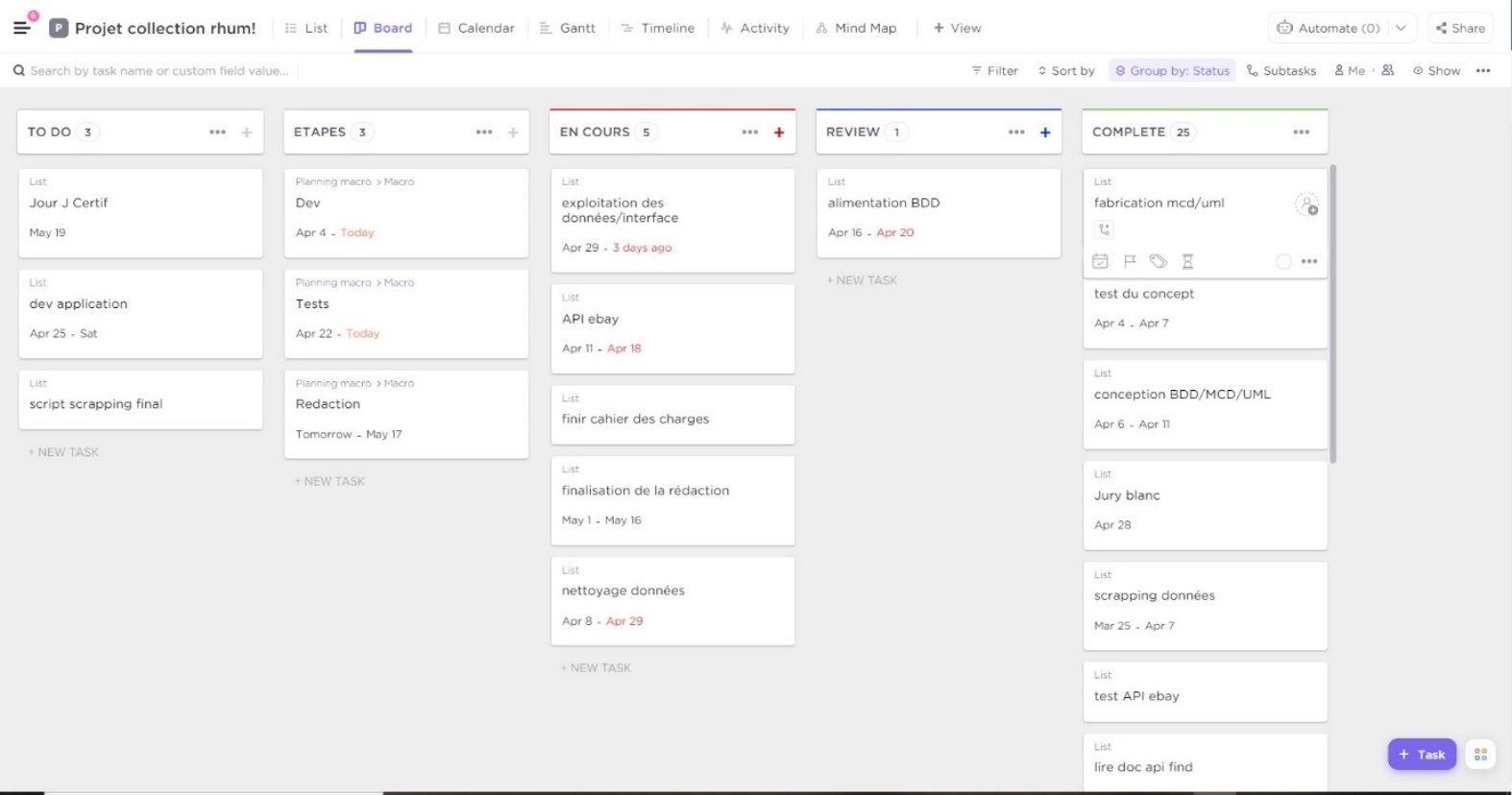
Notre projet de certification est un projet individuel dont le sujet est libre. La première étape a donc été de faire des recherches préliminaires pour déterminer le sujet et sa faisabilité.

Après la validation du sujet par nos formateurs, le premier challenge a été de déterminer le planning prévisionnel du projet. En tant que novice, il n’est pas évident d’évaluer le temps à consacrer à chaque tâche. Pour m’aider, j’ai utilisé l’application Clickup pour créer mon planning. Elle dispose de nombreuses fonctionnalités : calendrier, graphique de gant, todo-list… On a également implémenté des doses d’agilités dans le suivi du projet (Kanban, review).

J’ai commencé par définir les grandes étapes du projet en trois grandes phases modélisé dans un graphique de Gant : la phase de recherche, celle de développement et de tests et enfin celle de rédaction.



Dans un second temps, j’ai essayé de définir plus finement les tâches à accomplir. Et pour chacune d’entre elles j’ai estimé leurs priorités, leurs difficultés et le temps nécessaire à les accomplir. J’ai pu grâce à clickup, créer un tableau en m’inspirant de Kanban. Grâce à ce tableau je pouvais visualiser les tâches en cours ou celles qui restent à faire par exemple.



Cette approche, en deux temps, m’a permis d’avoir une vision globale du planning et une meilleure vision du temps à ma disposition pour réaliser ce projet.

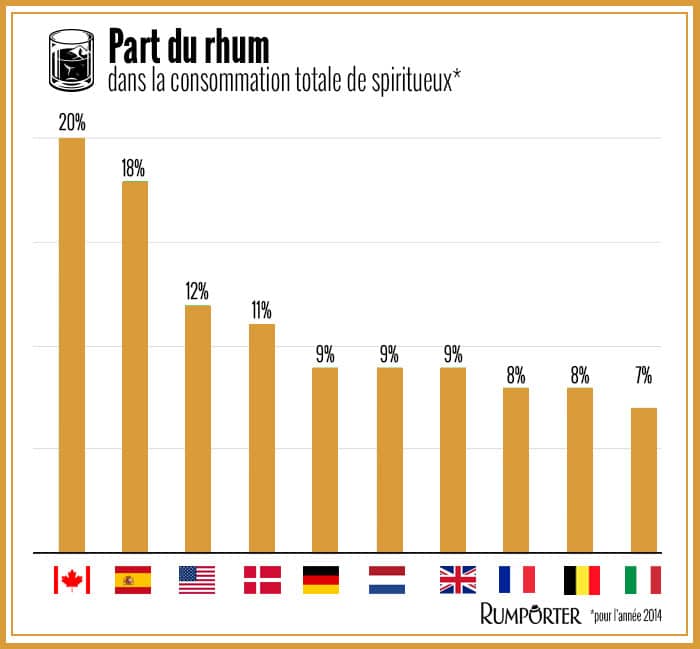
Pour la gestion quotidienne de mon projet, je notais simplement les tâches du jour sur un calepin.

Je contactais au moins deux fois par semaine mes formateurs, afin de leur faire un retour sur l’avancé de mon projet, les difficultés rencontrées et les solutions trouvées. Et je les interpellais quand je bloqué sur un sujet. Ce contact régulier était essentiel, car nous étions principalement en distanciel, pour s’assurer que l’on avance dans la bonne direction et aussi pour rester motivé.

Malgré ce travail, on s’aperçoit qu’on a toujours sous-estimés certaines tâches. Pour moi, j’ai passé plus de temps sur la préparation de mon projet et sur la partie de nettoyage de données qu’anticipé.

* 1. L’expression du besoin
     1. La progression du marché du rhum

Le rhum est le troisième spiritueux le plus consommé au monde, derrière le whisky et le brandy. C'est le spiritueux ayant connu la plus forte croissance au cours de ces dix dernières années (+ 40 %). 536 000 tonnes de rhum sont bues chaque année dans le monde, soit 17 litres par seconde.

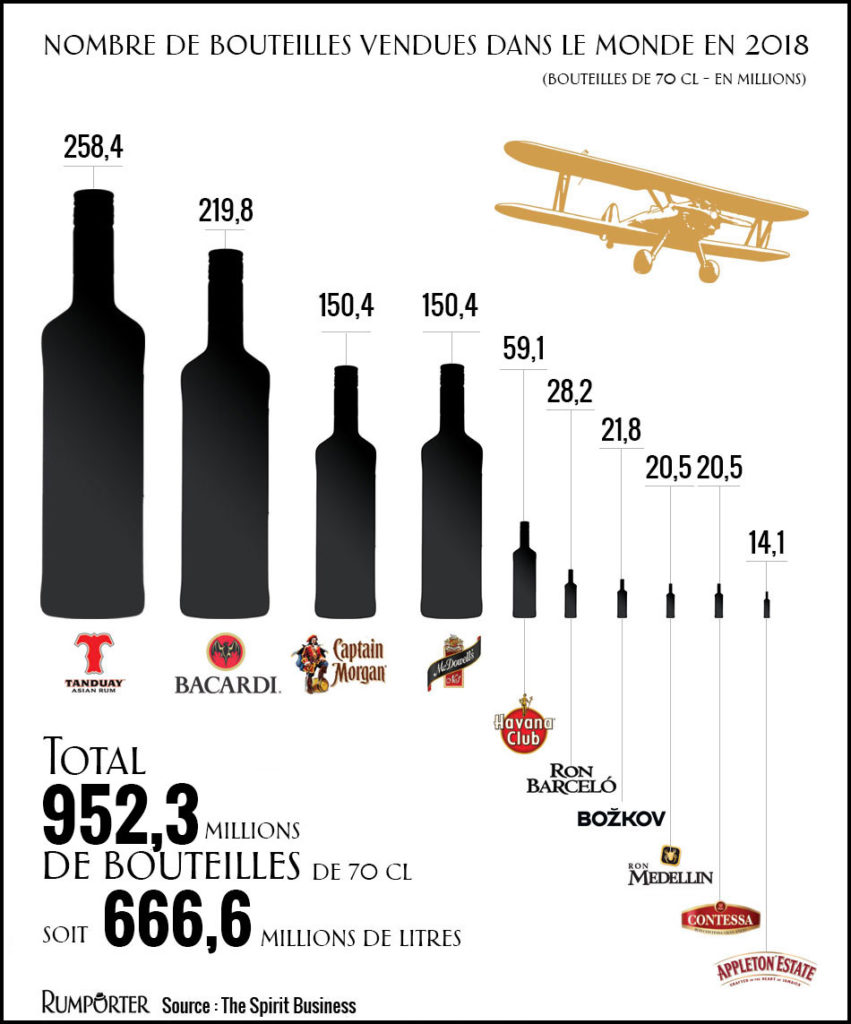


En France, le rhum représente 8 % de la consommation totale de spiritueux avec 34 millions de litres consommés et reste un des principaux moteurs de croissance de la catégorie alcools. La France est d'ailleurs le deuxième plus grand consommateur de rhum en Europe derrière l'Espagne.

La production de rhum français, exclusivement dans les départements français d’Outre-Mer représente plus de 55 millions de litres par an :

* 20 % de cette production est consommée sur place (DOM)
* 30 % du rhum est exporté en Europe et dans le Monde
* 50 % du rhum est consommé en Métropole

Le rhum industriel représente aujourd’hui environ 90% de la production mondiale de rhum. Les principaux producteurs de rhum sont Tanduay (philipine) et bacardi (bermudes).



Les chiffres le montrent, le rhum séduit de plus en plus de consommateurs. Il n’est plus cantonné au ti-punch et aux mojitos. Ils intéressent de plus en plus les mixologues ou les amateurs de spiritueux. Ils fédèrent une communauté d’amateur de plus en plus grande. Par exemple, les groupes français sur Facebook de la Confrérie du Rhum et du Rhum Arrangé comptent respectivement 45 000 et 150 000 membres.

Quelles sont les raisons de ce succès ? Premièrement, le goût, on trouve des rhums à la qualité de fabrication et la complexité gustative qui rivalise avec les plus grands cognacs ou scotch. Deuxièmement, on trouve une grande diversité dans la fabrication de rhum. Que ce soit au niveau de la production que du vieillissement. On utilise différent type d’alambic (colonne, pot still, double copper, colonne hybride, …), de levures (boulangères, œnologiques, indigènes, …), de types de fûts (style de bois, chauffe, …). Il y a beaucoup de dynamisme dans le secteur. La Martinique, productrice de rhum agricole, constitue un des fleurons actuels du secteur, étant la première au monde à posséder une AOC. Cette diversité permet de toucher un public plus large. Enfin, autre avantage de taille, le prix ! Le whisky est un marché mature qui subit une inflation galopante surtout les créneaux premium. Le rhum même pour ses bouteilles les plus prestigieux restes plus abordables que les whiskys de même gamme.

* + 1. Les collectionneurs de spiritueux

Ce projet s’adresse aux collectionneurs de spiritueux. Mais qu’est-ce qu’un collectionneur de spiritueux ? Quelles sont ses envies ? Sur quels critères se base-t-il pour acheter ?

Il serait difficile de définir le collectionneur type de rhum. Il y a diverses raisons de commencer à acquérir des bouteilles. On peut être naturellement amateur de spiritueux ou spéculateur ou juste attiré par l’esthétique d’un flacon. Personnellement, j’ai commencé à acquérir des bouteilles par curiosité technique (production et gustative).

Essayons de déterminer les attentes d’un collectionneur moyen qui cherche à agrandir sa collection. De manières générales, un collectionneur fera attention à la réputation de la marque et/ou de la bouteille, aux prix, à la qualité de conservation de la bouteille, à la rareté de la référence, à l’esthétique du flacon, …

J’ai pu rencontrer des collectionneurs plus exigeants qui vont faire attention au numéro de série, à la date de mise en fût et de mise en bouteille (surtout pour les références de rhum vieux millésimé car cela donne l’âge du spiritueux) ou qui vont collectionner toutes les références d’une seule marque.

*Une particularité intéressante des collectionneurs d’alcool, c’est qu’ils vont souvent acquérir deux exemplaires d’une référence qui leurs plaisent. L’une pour la garde, l’autre pour l’ouvrir.*

Le graal pour un collectionneur c’est d’acquérir une bouteille rare, ancienne qui jouit d’une excellente réputation gustative, bien conservé et à un prix raisonnable.

Donc, en conclusion, le collectionneur est une personne qui est avide de renseignements. La question est de savoir, si toutes ces informations sont pertinentes à être collectés pour notre projet.

La sécurité/discrétion, le bon prix, les informations.

* + 1. L’applicatif

L’importance et la granularité des informations nécessaire pour satisfaire la curiosité d’un collectionneur peut être importante. Nous allons voir ici qu’elles sont le type informations que nous sélectionneront pour notre projet.

Comme on l’a dit, l’un de nos buts est de créer une base de données de références. Notre objectif est la gestion d’une collection et pas forcément d’être un Wikipédia du rhum.

On pourrait intégrer beaucoup d’informations dans notre base de données. Mais beaucoup d’entre elles seraient difficiles et fastidieuses à obtenir (comme par exemple le numéro de la bouteille, le numéro de l’embouteillage ou batch pour une référence).

*Par exemple, la base de données la plus complète sur le rhum est celle de wikirum. Ils ont passé plus de deux ans à la constituer.*

Nous allons nous concentrer sur les informations générales. Des informations pas forcément exhaustives, mais suffisante pour permettre de constituer une première base de données.

Les informations seront : le nom de la référence, la marque/le producteur, le volume, le degré et l’origine. J’ai choisi ces informations car elles me permettent déjà de donner des renseignements précis sur une référence et elles sont plus facile à obtenir.

Des informations comme la réputation, la conservation ou la qualité d’un produit sont trop subjectives. Des informations comme le numéro de la bouteille, la date d’embouteillage ou de mise en fût sont-elles fastidieuses à trouver. C’est pour cela que je n’intégrerai pas directement ce type d’informations très précises. Par contre, on peut prévoir la possibilité à l’utilisateur d’ajouter des informations sur une référence.

Pour illustrer mes propos, prenons le cas d’usage suivant : nous avons un utilisateur qui cherche un rhum martiniquais extra vieux de 2003. Un collectionneur moyen voudra savoir s’il s’agit d’un rhum de JM (valeur estimé 200€) ou un HSE (valeur 150€ sur site e-commerce) ou un J. Bailly (valeur marchande 60€). Mais contrairement à un collectionneur avancé, il ne souhaitera pas forcément différencier les multiples embouteillages du 2003 d’HSE (3 par an, lot de 1000 bouteilles environ). Car le collectionneur n’est pas forcément un grand amateur éclairé du produit.

* + 1. Les applications existantes

Ils existent déjà plusieurs applications dans le domaine des vins et spiritueux. Celles-ci permettent de gérer une collection, d’obtenir des informations sur des références, d’écrire des notes de dégustations, … Pour les spiritueux, ce sont surtout des applications concernant le whisky que l’on trouve.

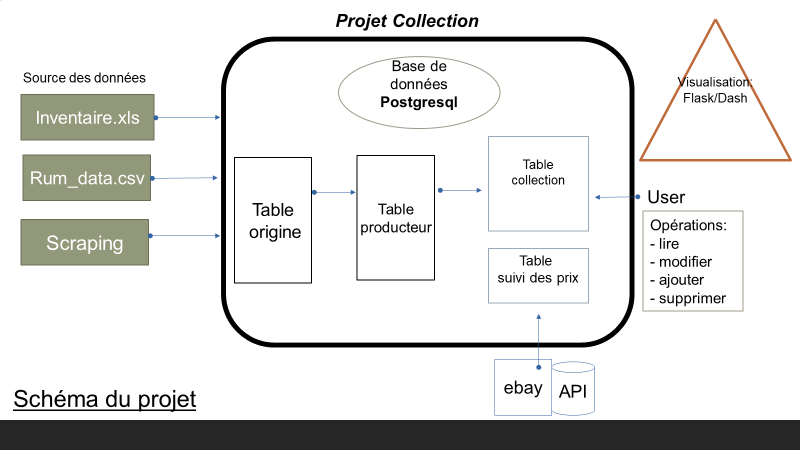
Il existe actuellement deux applications autour de la gestion de collection spécifique au rhum qui ont été lancé en 2020 :

* Wikirum : c’est une web application construit avec Angular qui est la base de données la plus complète sur le rhum. C’est un projet porté par des nantais de Saint Herblain. Ils travaillent directement avec les distilleries et les marques pour obtenir des données. C’est avant tout un wikipédia du rhum. Il permet également de se constituer une collection et d’acquérir des bouteilles.
* RumX (anciennement rum tasting note) : c’était en premier une application qui permettait de partager des notes de dégustation avec sa communauté. Dans sa nouvelle version, on peut également faire un inventaire de sa collection et acquérir de nouvelles bouteilles via un site d’enchère. Actuellement l’application qui offre le plus de fonctionnalités.
* Rumratings : c’est un site américain qui contient plus de 7000 références et qui permet de donner une note. Le site n’est actuellement plus activement maintenu. Ses deux créateurs cherchant à se diversifier leurs activités vers d’autres alcools.

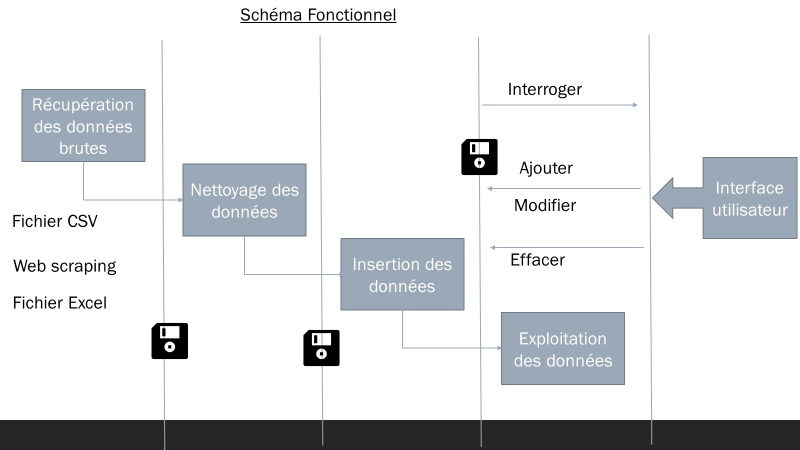
# CONCEPTION DU PROJET

1. L’application
   1. Traduction du besoin client

Afin d’avoir une meilleure vision du projet, j’ai réalisé un schéma du projet. Celui-ci à beaucoup évolué au fur et à mesure de l’avancement du projet. Ce schéma représente l’architecture du projet. Nous avons à gauche les sources des données, au milieu la modélisation de la base de données et à droite la partie exploitation de celle-ci.



J’ai également réalisé un schéma fonctionnel qui lui modélise les étapes de réalisation du projet qui seront vu dans la suite de ce rapport.



* 1. Choix des technologies

Voici un tableau représentant l’ensemble des technologies choisies pour la mise en œuvre du projet.

|  |  |  |
| --- | --- | --- |
| Langages | Bibliothèques | Outils |
| Python | Pandas | Jupyter lab |
| Postgresql | Psycopg2 | PgAdmin 4 V5 |
| Nosql (JSON) | Plotly Dash | Microsoft office |
|  |  | Exiftool |
|  |  | VS Code |
|  |  | Postman |
|  |  | Scrappy |

Après analyse du besoin, j’ai décidé de proposer cette solution :

* Création d’une base de données relationnelle sous PostgreSQL.
* Création d’un dashboard pour visualiser les informations de la base de données avec Dash.

1. Les sources de données
   1. La collecte des données

J’ai pu récolter des jeux de données auprès de différentes sources. En voici le descriptif :

* Test.json : fichier contenant 596 références issue du scrapping du site Christian de Montaguère avec la librairie scrappy.
* Pendant la phase de recherche, il était prévu que je récupère les données uniquement par scrapping. Je comptais scrapper des sites spécialisés ou des sites d’e-commerce. J’ai fait des essais avec les libraires python BeautifulSoup et Scrappy. Le scrapping c’est révélé moins fructueux que prévu. Soit la donnée collectée était peu exploitable (informations incomplètes et parcellaires) soit l’automatisation du scrapping se révélé fastidieux (code html mal structuré). Scrappy c’est révélé êre plus difficile à appréhender que BeautifulSoup, mais est plus performant. Il offre un véritable micro-framework pour le scrapping. J’ai suivi un tutoriel sur youtube qui m’a permis de scrapper le site e-commerce Christian de Montaguère.
* Inventaire.xls : l’inventaire de ma propre collection établis lors de mon déménagement dans un fichier excel. Etant des données privées je peux librement les utiliser. Cet inventaire contient 89 références.
* Rum\_data.csv : jeu de données en format csv trouvé sur kaggle (site qui contient des dataset opensource). Il contient une liste de 6000 rhums.
* Liste\_wikirum : fichier csv obtenu grâce à l’aimable participation du site wikirum. Il contient 1000 références.
* Api\_ebay : le projet prévoit de récolter l’évolution des prix grâce à l’API Find d’Ebay, mais faute de temps, cette foncionnalité ne sera intégré que dans les futurs versions.
  1. RGPD

L’autre difficulté, c’est l’aspect légal du scrapping est un procédé qui flirte avec la légalité. Certains sites prévoyaient directement dans leurs conditions d’utilisation d’interdire le scrapping. Le site de la Compagnie du rhum, site e-commerce de référence mondiale sur lequel je comptais beaucoup, m’a décliné la possibilité de scrapper leurs sites.

1. Création de la base de données
   1. Le modèle conceptuel des données (MCD)

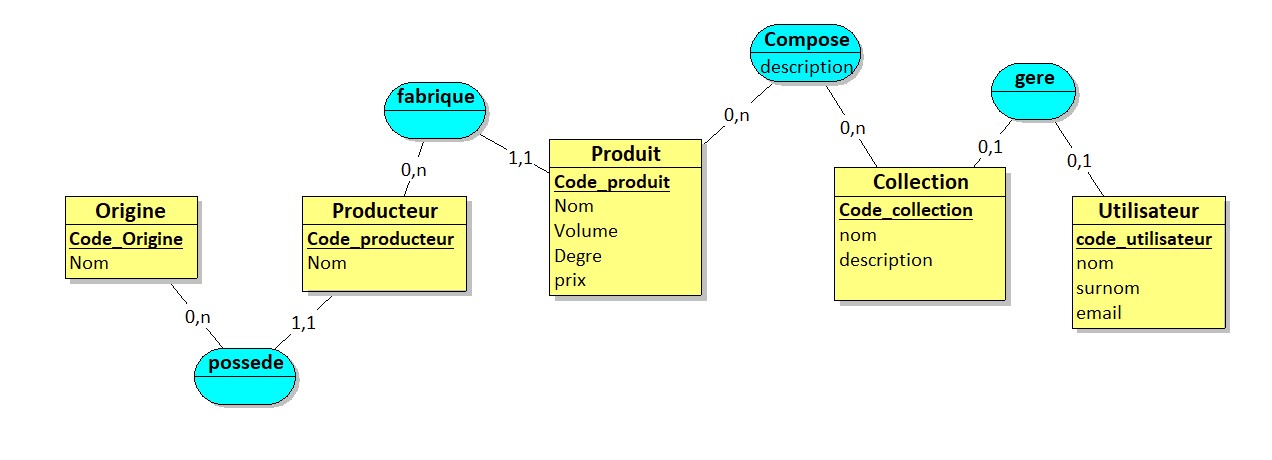
J’ai choisi d’utilisé Looping pour la création de mon MCD. Logiciel simple d’utilisation et complet.

Pour construire notre MCD nous devons définir nos entités, leurs attributs et les relations entre entités.

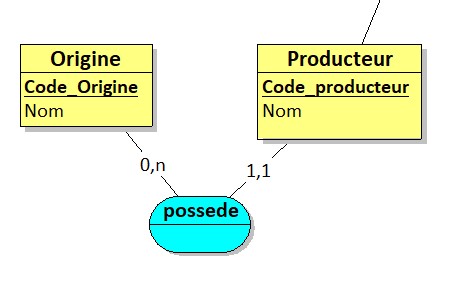
Nos entités seront :

* Origine : tel que données dans les ressources.
  + Il peut être difficile de déterminer l’origine d’un rhum. Car ils existent de nombreuses marques qui ne produisent pas de rhum. Elles achètent des lots de rhum de différents endroits du monde. Il arrive souvent que ces lots d’origines différentes soit assemblé par la suite pour ne former qu’une seule référence. Il arrive aussi que la provenance du rhum soit plus importante que la marque qui le distribue (surtout pour les rhums Caribéens
* Producteur : englobe les distillateurs, les sous-marques et les embouteilleurs indépendant.
* Produit : le nom du rhum.
* Collection : informations sur la collection d’un utilisateur.
* Utilisateur : les informations sur l’utilisateurs.

Voici la modélisation de notre MCD.



Maintenant analysons plus en détails notre MCD :

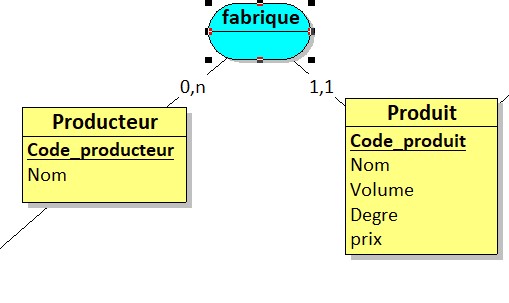


Les entités Origines et Producteur ont le même types d’attribues : un nom et un code spécifique qui fera office d’identifiant.

Elles sont associées par une relation « one to many » soit « un à plusieurs » en français.

Exemple de lecture :

* Un producteur possède une seule Origine.
* Une même origine peut être posséder par plusieurs producteurs

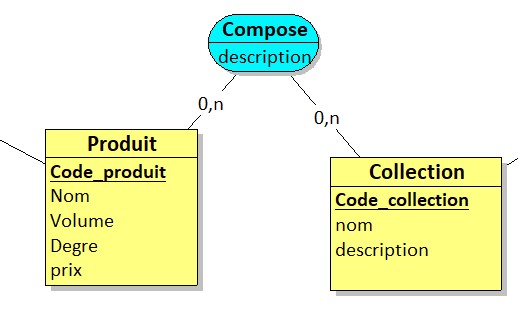


L’entité produit contient les informations nécessaires pour identifier une référence. C’est à partir de cette table que l’on pourra constituer les entités Collections des utilisateurs. Elle contient les attributs : code\_produit (identifiant), nom (de la référence), volume, degre, prix.

Les entités Produit et Producteur sont associées par une relation « one to many ».

Exemple de lecture :

* Un producteur fabrique un ou plusieurs produits.
* Un produit n’est fabriqué que par un seul producteur.



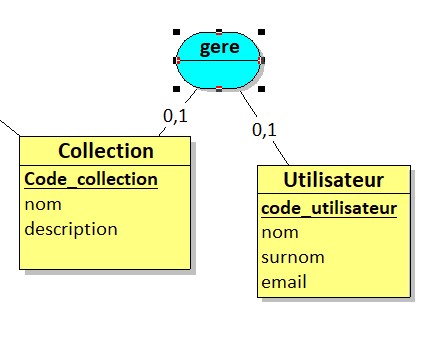
L’entité collection contient les informations nécessaires pour identifier une collection d’un utilisateur. Elle contient trois attributs : code\_collection (identifiant), nom, description

Les entités Produit et Producteur sont associées par une relation « many to many » (plusieurs à plusieurs en français).

L’association « compose » possède également un attribut « description » qui permettra d’ajouter des informations sur une référence.

Exemple de lecture :

* Un produit compose une ou plusieurs collections.
* Une collection est composée d’aucun ou plusieurs produits.



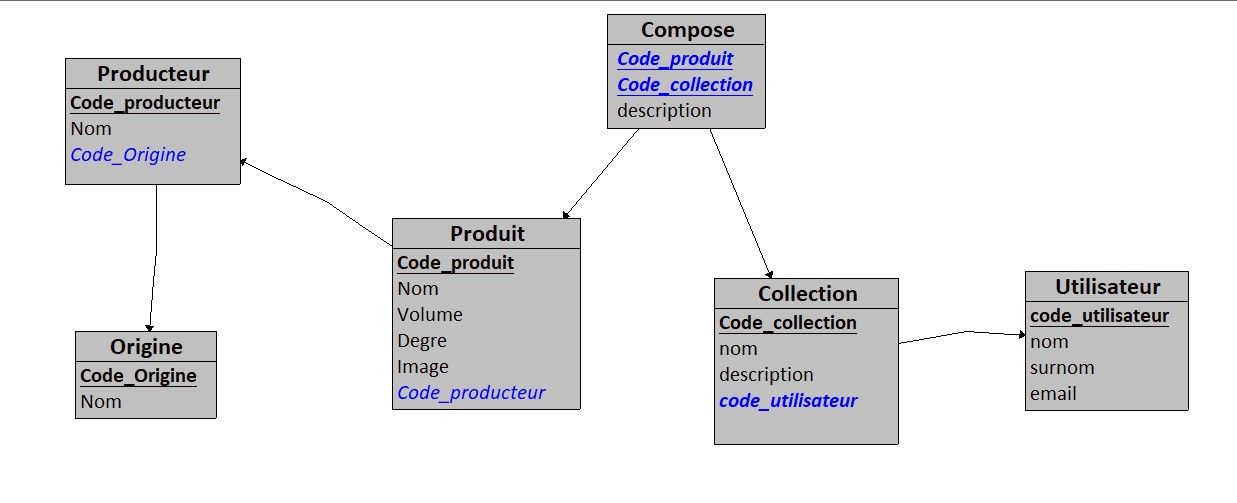
L’entité Utilisateur contient les informations d’identification d’un utilisateur. Elle contient les attributs suivants : code\_utilisateur (clé primaire), nom, surnom, email.

Les entités collection et utilisateurs sont associées par une relation « one to one » (un à un en français). C’est une relation forte entre deux entités.

Cela veut dire qu’un utilisateur ne gérer qu’une seule collection.

* 1. Le Modèle Logique des Données (MLD)

Le MLD nous permet de passer du MCD à la modélisation effective de notre base de données que l’on pourra implémenter dans notre SGBD.



* La clé primaire de la table origine devient une clé étrangère de la table producteur.
* La clé primaire de la table producteur devient une clé étrangère de la table produit
* Ajout de la table compose pour réaliser la relation many to many entre les tables produit et collection. Cette table à une clé primaire composé des clé primaires code\_produit et code\_collection.
* La clé primaire de la table utilisateur devient une clé étrangère de la table collection.

1. Nettoyage des données.

Le nettoyage de la donnée a été fait à partir d’un jupyter notebook avec l’utilisation de la bibliothèque Pandas, les modules json et re.

Opération :

Caractères spéciaux

Suppression de colonnes

Renommé des colonnes

Regex

Company inconnu

Traduction nom pays.

Suppression de lignes

Visualisation des données

Sauvegarde et métadonnées

Introduction

Collectionneur secret.

Nettoyage donné

Nettoyage fichiers

Transposition jupyter en scripts python

Changer nom des producteurs

* Nettoyage du fichier excel :
  + Utilisation de la bibliothèque pandas : changement nom de colonne, suppression de colonne, suppression référence non rhum, suppression de lignes producteur
  + Changement à la main : comme le fichier est petit certain changement ont été fait directement dans le fichier, par exemple changement nom producteur pour qu’il soit le même que le fichier rum\_data
* Nettoyage fichiers csv : remplacement des caractères spéciaux
* Nettoyage particulier rum\_data : dans la colonne company, beaucoup sont marqué inconnu. J’ai pris de la décision, de prendre le nom de la bouteille comme nom de company. Beaucoup de ses références contiennent de toute façon le nom de la company dans leurs noms. Beaucoup de ses company sont de petites marques américaines, des embouteilleurs indépendants qui n’ont pas d’activité de production.
* Nettoyage fichier json : utilsation de regex pour extraire les informations nom volume, degre, marque. Split pour origine.
* New data : changement type de colonne float/ int