SORBONNE UNIVERSITÉ

03/01/2023

Rapport projet : Génie-Logiciel

Analyse du commerce extérieur de la France pour l'année 2019

Youcef CHORFI, Gontran GILLES

MASTER 2 - ISI

Introduction

Dans le cadre de l'unité d'enseignement Génie Logiciel, nous sommes chargés de construire une base de données à partir de données brutes récupérées sur internet, après avoir défini les modèles conceptuel et logique de données. Nous devons également exploiter cette base en réalisant des requêtes avec le langage SQL. Nous proposons de plus une application dotée d'une interface graphique pour faciliter l'analyse de notre base de données.

I. Présentation du sujet

Depuis le site *data.gouv.fr*, nous avons fait le choix de traiter notre sujet sur le commerce extérieur de la France pour l'année 2019 (seule année à notre disposition sur le site) [1]. Les données mises à notre disposition nous permettent, pour cette année uniquement, de connaître les différentes transactions qui ont eu lieues avec la France. Ces transactions décrivent tous les échanges de marchandises possibles, c'est-à-dire tous les biens matériels entrants et sortants du territoire ; cette collecte statistique est faite par les Douanes françaises.

1. Présentation des données brutes

Nous avons récupéré sur le site deux dossiers compressés : un pour les données d'importation et l'autre pour les données d'exportation. Chacun est organisé de la même manière : il comporte 5 fichiers textes contenant différentes données que nous allons décrire.

a) Fichier des transactions

Ce fichier contient toutes les transactions effectuées sur une année. Dans notre cas le fichier d'importation contient 1 687 753 entrées, et celui d'exportation 2 627 612 entrées.

La transaction est la donnée centrale de notre base de données. Elle décrit l'échange d'un bien à un instant donné. Nous y trouvons donc les informations suivantes :

- Flux : importation ou exportation
- Mois : mois de l'année auquel a été réalisée la transaction
- Année : année de la transaction (dans notre cas, une seule année : 2019)
- Code CPF6 : classification de produit française ; code à 6 chiffres qui décrit de manière précise un produit (ex : 011111, blé dur)
- Code A129 : classification d'un produit dans un groupe (nomenclature française) ; décrit un groupe de produits (ex : C27B, matériel électrique)
- Code NC8 : nomenclature combinée à 8 chiffres (commune aux Etats membres de l'Union Européenne) ; code à 8 chiffres qui décrit un produit (ex : 01012910, chevaux destinés à la boucherie)
- Code pays : code international à 2 lettres identifiant un pays (selon norme ISO 3166-1 alpha-2)
- Valeur : valeur d'échange exprimée en euros
- Masse : masse du (des) biens(s) échangé(s) en kilogramme
- USUP : unité supplémentaire ; donne la valeur du (des) biens(s) échangé(s) dans une unité autre que le kilogramme (ex : pour un produit de revêtement de sol, type parquet, l'unité supplémentaire sera le mètre carré)

[1] https://www.data.gouv.fr/fr/datasets/statistiques-nationales-du-commerce-exterieur/

b) Fichier des libelles

Les 4 autres fichiers contiennent les libelles des codes décrits ci-dessus. Ainsi pour le code NC8 nous avons toutes les correspondances entre un code et sa signification, idem pour les codes pays, A129, CPF6.

2. Données supplémentaires

Afin d'étoffer notre base de données et de répondre à certaines requêtes qui sont intéressantes pour l'analyse, nous avons récupérés d'autres données brutes sur internet.

a) Table des libelles NC8 avec sections et sous-sections

Le code NC8 permet de décrire très précisément un produit en faisant des distinctions très fines. Par exemple les codes 01022910, 01022921, 01022929 décrivent tous les trois des bovins mais font la différence en fonction du poids de la bête et sa destination (reproducteur ou pour la boucherie). Ainsi il existe 12 codes NC8 différents pour décrire des bovins. Cette finesse de description peut être utile dans certains cas, mais peut compliquer la recherche pour certaines requêtes. Toujours en suivant cet exemple, si nous souhaitons faire un bilan des importations et exportations concernant l'espèce bovine, il nous faudrait connaître toute la nomenclature s'y rapportant (les 12 codes). Pour éviter cette complication nous avons trouvé un tableau excel qui découpe toute la nomenclature NC8 en sous-sections et sections.

Ainsi, l'espèce bovine est regroupée avec un code de sous-section à 4 chiffres (ce sont les 4 chiffres de poids les plus forts des codes à 8 chiffres). De la même manière une section regroupe de nombreux codes sous une désignation plus générique, par exemple l'espèce bovine fait partie de la Section I : Animaux vivants et produits du règne animal.

Dans nos différentes requêtes nous utilisons ces notions pour faire des analyses plus générales des importations et exportations. Il nous a alors fallu créer les tables correspondantes à partir de ce tableau excel.

b) Table de passage NC2019 vers NC2020

La table décrit en *2.a)* utilise les codes NC8 selon la nomenclature 2020, or nos fichiers de transactions décrivent les produits selon la nomenclature de 2019. Nous avons alors dû convertir les codes NC8 2019 en codes NC8 2020. Nous l'avons fait avec un tableau excel qui donne ces correspondances.

c) Pays, continents et zones économiques

A partir des données brutes nous ne disposons que de l'information sur le pays. Nous avons trouvé intéressant d'affiner les résultats de certaines requêtes en fonction du continent et/ou de la zone économique d'un pays. Nous avons récupéré un tableau excel et un document pdf qui nous fournissait toutes ces informations. De plus le tableau excel nous donnait le code des pays à 3 lettres ; ce code nous est utile pour faire un affichage des données sur une carte avec le module plotly.

II. Modèles conceptuel et logique de données

La première étape avant de construire la base proprement dite est de la modéliser. Nous faisons apparaître les tables et les relations entre ces tables afin de répondre à des questions qui permettent l'analyse du commerce extérieur.

1. Modèle conceptuel de données (MCD)

Nous avons créé 6 tables qu'on peut décomposer en trois catégories :

- Les transactions : 1 table (*transaction*)
- Les dénominations des produits échangés : 3 tables (*produit*, *sections*, *sous-sections*)
- Les informations liées aux pays : 2 tables (*pays*, *zone-économique*)

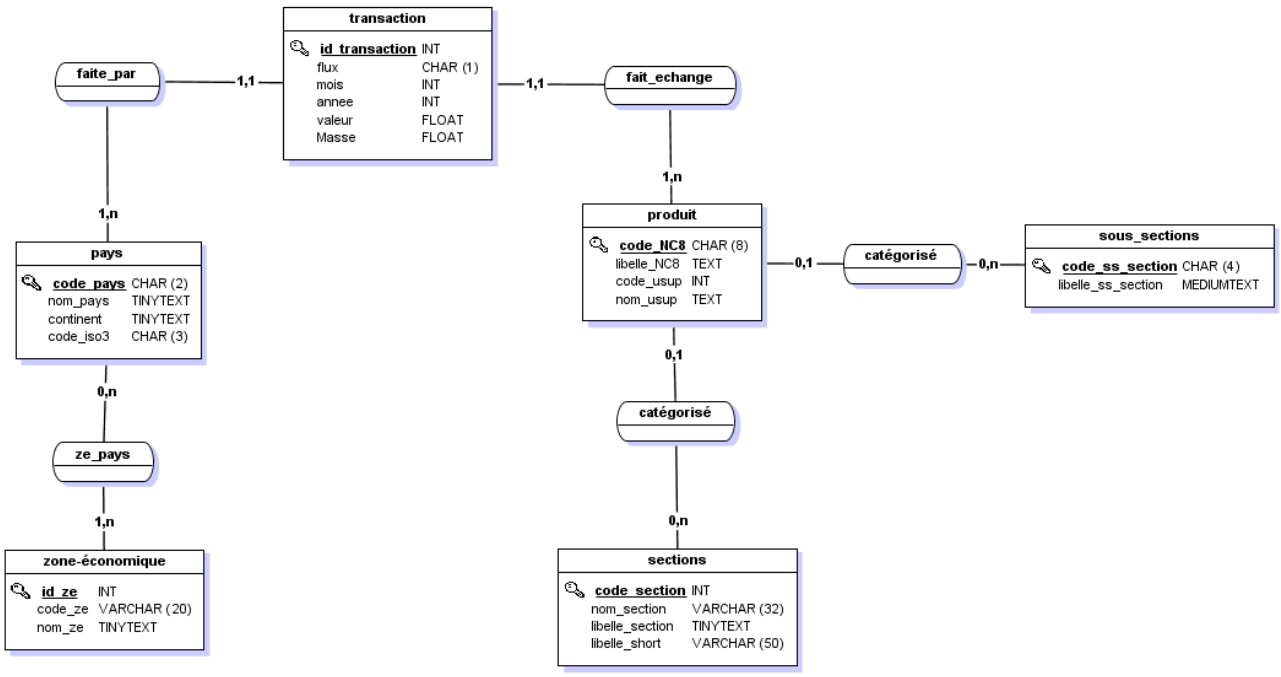


Figure 1 - Modèle conceptuel de données - Base de données Commerce extérieur 2019

a) Table transaction

La table *transaction* contient tous les échanges. Chaque transaction (ou échange) est identifiée de manière unique par un entier. Une ligne de cette table nous permet de connaître le flux (importation ou exportation), le mois, l'année, la valeur et la masse. Mais elle ne nous permet pas de connaître le produit qui est échangé ni le pays avec lequel l'échange est fait.

Pour cela nous avons créé deux autres tables : *produit* et *pays*. Ces tables sont associées à la table *transaction*. Ainsi, à chaque transaction est associée le produit échangé et le pays avec lequel l'échange est fait.

Les cardinalités sont les mêmes pour les deux relations : ***1,1 - 1,n***. Une transaction est faite par un pays, et un pays peut faire une ou plusieurs transactions. Une transaction fait l'échange d'un produit, et un produit peut être échangé lors d'une ou plusieurs transactions.

b) Table produit

Cette table contient toutes les descriptions des produits selon la nomenclature NC8. Nous avons fait le choix de ne garder que cette nomenclature alors que dans les données brutes nous en avions d'autres, car cette nomenclature est très précise et nous permet de faire des regroupements en sections et sous-sections. Cette table contient en plus l'éventuelle code USUP (unité supplémentaire) associé.

Même si le code NC8 est un code à 8 chiffres, nous avons fait le choix de le coder dans la table comme une chaîne de 8 caractères (et non un entier). En effet, il faut se rappeler que les premiers codes NC8 commencent par 0 (0xxxxxxx), ainsi si nous l'encodons comme un entier nous perdons le 0 et ne gardons qu'un code à 7 chiffres. Ce choix peut être discuté.

Nous avons associé à cette table deux autres tables : *sections* et *sous-sections*. Ces tables permettent de catégoriser un produit à un niveau supérieur d'identification, dans le sens où l'on gagne en généralité (et l'on perd en précision). Les cardinalités sont les mêmes pour chacune des tables : ***0,1 - 0,n***. Ainsi, un produit peut être **ou** ne pas être catégorisé par **une** section ou sous-section. Et une section ou sous-section peut catégoriser **ou** ne pas catégoriser **un ou plusieurs** produits.

c) Tables sections et sous-sections

Contiennent les différentes sections et sous-sections qui permettent de catégoriser les produits. On notera que pour la table *sections* nous avons ajouté un champ *libelle_short* de 50 caractères maximum. Il s'agit en fait du *libelle_section* tronqué à 40 caractères. Ce champ n'a pas d'utilité en soi, mais nous sert lors des affichages des graphes. En effet, les libelles originaux sont parfois trop longs pour l'affichage.

d) Table pays

Cette table contient le code pays sur 2 lettres, qui est la clé primaire, le nom du pays, le continent auquel il appartient et le code du pays sur 3 lettres. Ce code sur 3 lettres nous sert pour l'affichage sur une carte avec le module Python *plotly*.

e) Table zone économique

Contient les différentes zones économiques. Nous avons associé cette table à la table pays afin qu'on associe un pays avec, éventuellement, une zone économique. Les cardinalités sont ***0,n*** – ***1,n*** : un pays peut **ou** peut ne pas appartenir à **une** zone économique. Une zone économique contient un ou plusieurs pays (il n'existe pas de zone économique vide).

2. Modèle logique de données (MLD)

Le modèle logique de données est la traduction du MCD précédent.

On notera tout de même la création d'une nouvelle table *ze_pays* qui associe un pays avec une zone économique, cela est dû à la cardinalité ***x,n - x,n*** de la relation. On observe également que la table *transaction* contient deux champs supplémentaires : ce sont les clés étrangères associées aux tables *pays* et *produit*. On remarque alors que la table transaction reprend tous les champs du fichier des données brutes transaction (à l'exception de ceux que nous n'avons pas retenus : CPF6 et A129).

III. Peuplement des tables

Principalement le peuplement des tables s'est fait à l'aide de la librairie *Pandas* pour le prétraitement des données et *SQLAlchemy* pour la connexion avec le serveur afin de remplir les tables. Le peuplement des tables a nécessité un nombre assez important d'opérations diverses afin de transformer les données brutes récupérées.

Etant donné le MLD nous avons dû suivre un certain ordre pour remplir les tables. D'une manière générale il faut toujours commencer par les tables qui ne possèdent pas de clés étrangères, puis celles qui possèdent des clés étrangères.

Mais avant de procéder au peuplement il faut prétraiter les données.

1. Prétraitement des données

Pour chaque table il faut en général procéder aux prétraitements suivants :

- Suppression des colonnes qui n'ont pas d'intérêts
- Eventuellement suppression des lignes qui ont des valeurs nulles, cela dépend des cas
- Renommage des colonnes pour faire correspondre avec notre MLD
- Création ou modification des index

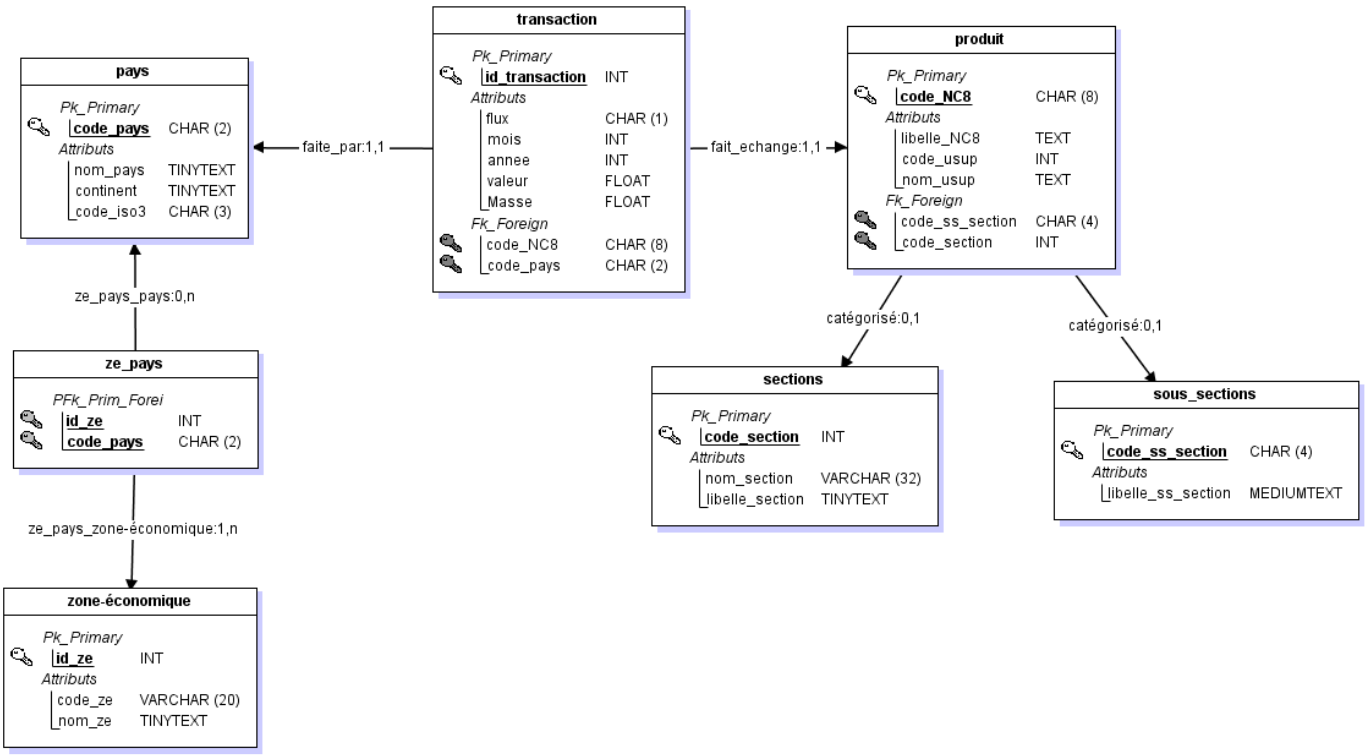


Figure 2 - Modèle logique de données - Base de données Commerce extérieur 2019

a) Passage des codes NC8 2019 vers 2020

Comme nous l'avons expliqué ci-dessus tous les produits de nos transactions sont identifiés avec le code NC8 2019, alors que notre tableau excel avec les sections et sous-sections de produits ont les codes de 2020. Ainsi, les données de toutes les lignes contenues dans les fichiers texte de transaction et le fichier libelle NC8 ont dû être mises à jour.

b) Création des tables *sections* et *sous-sections*

Ces données n'existent pas telles qu'elles. Elles sont noyées dans un fichier excel. Nous avons alors lu tout le fichier excel avec une méthode de *Pandas*, puis transformé et extrait toutes les informations nécessaires à l'aide de méthodes dédiées au traitement des chaînes de caractères (expressions régulières, remplacement de valeurs, extractions de valeurs…).

c) Création des tables pays et zone économique

La table *pays* peut être en partie remplie à l'aide du fichier texte contenu dans le jeu de données du site *data.gouv*, mais nous avons dû la compléter pour avoir les continents et les codes des pays à 3 lettres. Ces dernières informations ont été extraites d'un fichier excel récupéré sur internet.

La table *zone-économique* a également été créée de toutes pièces. Nous avons trouvé un document *pdf* dans lequel était décrit les différentes zones avec les pays associés. Nous avons copier-coller le texte brut dans un fichier texte, puis traité ce fichier à l'aide du module *re* de Python qui utilise les expressions régulières.

2. Peuplement des tables

Le peuplement des tables s'est fait dans l'ordre suivant :

1. Tables *sections* et *sous-sections*
2. *Table produit*
3. Table *pays* et *zone-économique*
4. Table *ze-pays*
5. Table *transaction*

Toutes les tables ont été peuplés à partir de *data frame Pandas*, sauf la table *ze-pays* qui a été fait avec une requête SQL de type 'INSERT INTO'.

IV. Exploitation de la base de données

Afin d'exploiter notre base de données nous présentons une série de questions avec les requêtes qui y répondent, puis une interface graphique sous forme d'une page web.

1. Requêtes SQL

Quelles sont les valeurs d'importation et d'exportation pour chaque pays ?

```
select code_iso3, nom_pays, continent, sum(valeur) as valeur_echanges, flux from transaction join pays using(code_pays)  group by flux, code_pays;
```

Quelles sont les valeurs d'importation et d'exportation pour chaque zone économique ?

```
select nom_ze, sum(valeur) as Valeur_echanges, flux from transaction join pays using(code_pays) join ze_pays using (code_pays) join zone_economique using (id_ze) group by nom_ze, flux ;
```

Quels sont les 10 secteurs dans lesquels la France exporte le plus ?

```
select sum(valeur) as valeur_exp, libelle_section from transaction join produit using(code_NC8) join sections using(code_section) where flux = 'E' group by flux, code_section order by valeur_exp desc limit 10;
```

Combien la France a exporté et importé pour l'année 2019 ?

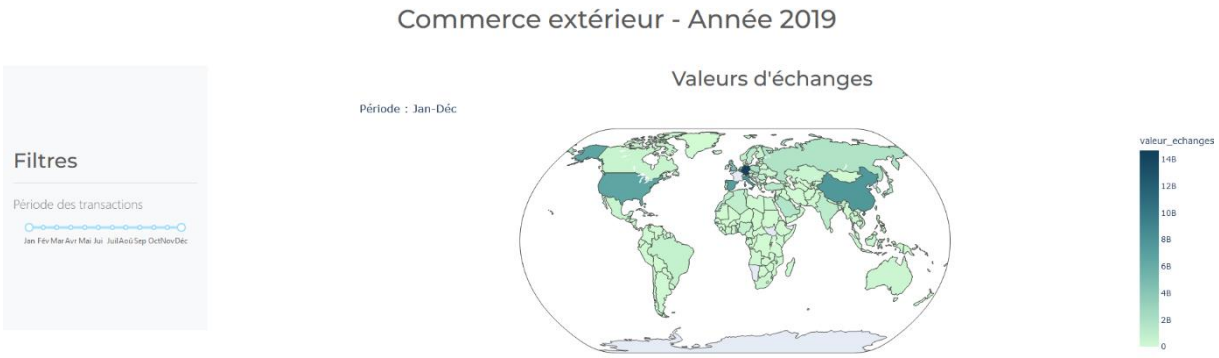
```
select sum(valeur) as valeur_exp, flux from transaction join produit using(code_NC8) where annee = 2019 group by flux;
```

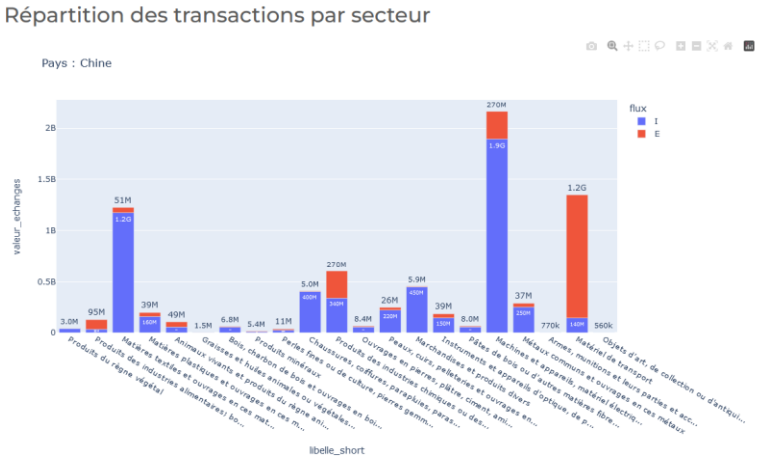
Quelles sont les valeurs d'échange (importation et exportation) avec la Chine entre mars et juillet ?

```
select nom_pays,  sum(valeur) as valeur_echanges, flux from transaction join pays using(code_pays) where mois between 3 and 7 and nom_pays = "Chine" group by flux, code_pays;
```

2. Interface graphique

L'interface graphique a été faite avec le module *dash* qui nous a permis d'intégrer l'affichage de graphes avec *plotly* et de gérer l'interaction.





Conclusion

La création des bases de données avec la récupération de données brutes, puis la conceptualisation et le peuplement est une étape très importante. Nous avons vu la difficulté de trouver et regrouper des informations diverses, provenant de sources différentes et de les assembler pour qu'on puisse en extraire des informations pertinentes lors de la phase d'exploitation de la base de données.

L'exploitation est intéressante à partir du moment où l'on dispose d'une interface rendant accessible les différentes informations contenues dans la base. Cela a été pour nous une tâche assez longue à mettre place car nous n'avions initialement aucune compétence dans la création d'une page web servant à visualiser des données. En conséquence, par manque de temps, nous avons assez peu exploité notre base ; mais il est possible de réaliser de nombreuses combinaisons entre toutes les données, même avec le peu de tables que nous avons.