CS443

Analyse de consommation d'énergie

2023 – 2024

> *Vous êtes recruté·e·s pour analyser la consommation d'énergie électrique et gaz en France. Pour cela, vous déployez une base de données et la requêtez pour répondre aux attentes de votre commanditaire.*

Ce projet a pour objectifs de vous entrainer à modéliser une base de données et de vous permettre de vous familiariser avec les outils SQL.

Pour cela, vous allez :

1. Créer un modèle conceptuel de données et le traduire en modèle relationnel ;
2. Déployer une base de données SQLite et la peupler à partir de données que l'on vous fournira ;
3. Requêter et instrumentaliser la base de données pour répondre à des questions ;

Jeu de Données

Un jeu de données vous sera fourni. Il est constitué d'une version réduite d'un jeu de données gouvernemental concernant la « consommation annuelle d'électricité et gaz par commune et par secteur d'activité » [1] ainsi qu'une version réduite du recensement de la population entre 2011 et 2020 [2–11].

Vous aurez à traiter des relevés comprenant :

- Un nom d'opérateur ;
- Une année ;
- Des consommations avec :
  - un type ;
  - une catégorie ;
  - une valeur ;
  - un nombre de Points de Livraison (PdL) ;
  - un indice qualité.
- Une commune avec :
  - un code de commune ;
  - un libellé de commune ;
  - un code postal.
- Un département avec :
  - un code de département ;
  - un libellé de département.
- Une région avec :
  - un code de région ;
  - un libellé de région.

Ces données sont régies par les contraintes suivantes :

- Les noms d'opérateur sont uniques et appartiennent à la liste donnée dans `files/operateurs.csv` ;

- Le type d'une consommation est soit `Gaz` soit `Électricité`;
- Un relevé comprend une ou zéro consommation agricole, une ou zéro consommation industrielle, une ou zéro consommation résidentielle et une ou zéro consommation tertiaire;
- La valeur d'une consommation est un nombre entier;
- Le nombre de points de livraison est un nombre entier;
- L'indice qualité est un nombre flottant compris entre 0 et 1;
- Les codes de commune, département, région et postaux sont des chaines de caractères de taille fixe;
- Les libellés sont des chaines de caractères à longueurs variables;
- Les codes de commune, de département et de région sont uniques.

De plus, une table comprenant le nombre d'habitants par commune vous est fournie. Cette table est sous la forme décrit en Code 1. Son `CodeCommune` correspond aux codes des communes précédemment cités et `NombreHabitants` est entier.

```
Habitants(CodeCommune, Année, NombreHabitants)
```

Code 1: Description de la table des habitants.

Modélisation Conceptuelle et Logique des Données

Définissez le modèle conceptuel des données, notamment par un schéma Entités-Associations et listez les possibles contraintes que vous ne pouvez pas modéliser dans ce schéma.

Transformez le modèle conceptuel est un modèle logique.

Outils :

- Papier & crayon.

Résultats :

- Un document décrivant le modèle conceptuel et le modèle logique ainsi que les choix pris pour les obtenir.
- Toute explication supplémentaire que vous jugez nécessaire.

Schéma Entités-Associations et Modèle Relationnel

Pour la suite du projet, nous utiliserons une base décrite par le schéma Entités-Associations et modèle relationnel ci-dessous.

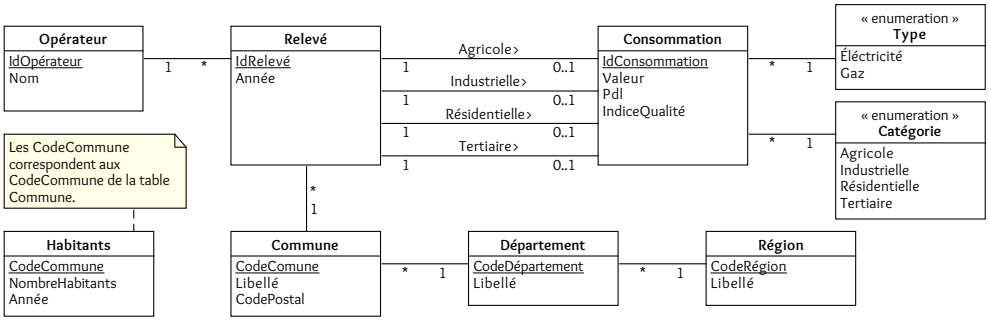

FIGURE 1 – Schéma Entités-Associations de la base de données du projet.

```
Habitants(CodeCommune, Année, NombreHabitants)
Opérateur(IdOpérateur, Nom)
Région(CodeRégion, Libellé)
Département(CodeDépartement, Libellé, #CodeRégion)
Commune(CodeCommune, Libellé, CodePostal, #CodeDépartement)
Catégorie(CatégorieId, Nom)
Type(TypeId, Nom)
Relevé(IdRelevé, Année, #IdOpérateur, #CodeCommune)
Consommation(IdConsommation, Valeur, Pdl, IndiceQualité, #IdRelevé,
             #TypeId, #CatégorieId)
```

Code 2: Modèle relationnel de la base de données du projet.

Création de la Base de Données

Créez la base à partir du modèle logique fourni.

Pour tester votre base, dans un premier temps, insérez des valeurs et tester des requêtes, puis, utilisez `files/test-db.db`.

Voir `doc/creation_tables.md` pour les informations nécessaires à la création de tables et l'insertion de valeurs.
Voir `doc/sqlite3.md` pour quelques informations sur l'utilisation de l'outil SQLite3.

Outils :

- `doc/creation_tables.md` sur Chamilo ;
- `doc/sqlite3.md` sur Chamilo ;
- `files/test-db.db` sur Chamilo ;
- `DB Browser for SQLite` installé sur les machines ;
- Ligne de commande `sqlite3` installée sur les machines.

Résultats :

- Un fichier reprenant les scripts de création de tables.
- Une liste de requêtes et leur résultats validant la bonne création de la base ainsi que sa correspondance avec le modèle logique.
- Toute explication supplémentaire que vous jugez nécessaire.

Requêtage SQL

Dans un premier temps, à partir de la base de données fournie, avancer le plus possible dans l'analyse en n'utilisant que le langage SQL et les outils SQLite.

Outils :

- `DB Browser for SQLite` installé sur les machines ;
- Ligne de commande `sqlite3` installée sur les machines ;
- Le fichier `files/create-db.sql`, utilisé pour la création de la base ;
- Une base de données fournie pour votre binôme.

Résultats :

- Une liste de réponses aux questions ainsi que les requêtes utilisées pour y répondre.
- Un retour d'expérience sur l'utilisation du langage SQL et le fait qu'il soit approprié, ou non, pour répondre aux questions posées.
- Toute explication supplémentaire que vous jugez nécessaire.

Manipulation de la Base de Données dans une application

Dans un deuxième temps, à partir de la base de données fournie, en utilisant un script Python et la librairie `sqlite3` [1], continuez l'analyse.

Durant cet exercice, vous aurez l'occasion de choisir la proportion d'algorithmie que vous écrivez en SQL et en Python pour répondre à chaque question. Posez vous la question de possibles coûts, notamment en temps d'exécution mais aussi en temps de développement, que ces choix engendre. Vous pouvez notamment mesurer le temps d'exécution en utilisant la librairie `timeit` [2].

Outils :

- Python et ses librairies standards `sqlite3` et `timeit` ;
- `files/example.db` et `files/example.py` ;
- Base de données fournie sur Chamilo.

Résultats :

- Une liste de réponses aux questions ainsi que le ou les script(s) utilisé(s) pour y répondre.
- Un retour d'expérience sur l'utilisation du langage Python pour s'interfacer avec une base de données SQLite et le fait que ce soit approprié, ou non, pour répondre aux questions posées.
- Un retour sur la réflexion et les possibles mesures quant à la répartition de l'algorithmie Python et SQL.
- Toute explication supplémentaire que vous jugez nécessaire.

Analyses

Consignes générales :

Pour l'ensemble des tâches :

- Si aucune année n'est précisée, la question concerne l'ensemble des années.
- Si aucune catégorie (agricole, industrielle, tertiaire, ou résidentielle) n'est précisée, la question concerne l'ensemble des catégories.

1. https://docs.python.org/3/library/sqlite3.html, visité le 05/10/2023
2. https://docs.python.org/3/library/timeit.html, visité le 05/10/2023

- Si aucun type d'énergie (électricité ou gaz) n'est précisé, la question concerne les deux types (électricité et gaz).
- Si aucun méthode de classement n'est précisée, garder les données ordonnées telles qu'elles le sont dans la base de données.

Pour les questions d'exploration et d'analyse :

- Chaque réponse doit consister en une table contenant la réponse et uniquement la réponse : s'il vous est demandé de donner les informations d'une commune, la table de réponse ne doit contenir qu'une ligne.

Exploration

> *Avant toute autre chose, vous explorez le jeu de données que vous allez manipuler lors du projet.*

Répondez aux plus de questions possibles tout en fournissant la réponse sous la forme d'une table unique ne contenant que les informations demandées.

1. À quelles régions appartiennent les communes de votre base de données ?
   Donnez le code de la région et son libellé.

Note Il n'y a pas de relevé sans consommation.

2. Il y a-t-il des consommations `NULL` ? Si oui, combien ?
3. Pour l'année 2011, combien de relevés votre base comprend elle ? Pour l'année 2012, combien de relevés votre base comprend elle ?
4. Pour l'année 2013, combien de communes ont au moins un relevé ? Pour l'année 2014, combien de communes ont au moins un relevé ?
5. Quelles sont les trois année avec le plus de communes mesurées ?
   Donnez l'année et le nombre de commune mesurées.
6. Pour l'année 2015, quels libellés de communes sont utilisés plusieurs fois ? Et combien de fois les sont-ils ?
7. Pour l'année 2016, quels sont les 16 opérateurs avec le plus de relevés ?
   Donnez l'identifiant, le nom et les nombres de relevés de ces opérateurs.
   Classez les opérateurs par nombre décroissant de relevés.

Première Analyse

> *Maintenant que vous êtes familiarisé·e·s avec le jeu de données, vous pouvez répondre aux questions de votre commanditaire.*

Répondez aux plus de questions possibles tout en fournissant la réponse sous la forme d'une table unique ne contenant que les informations demandées.

1. Pour l'année 2017, quelles sont les 17 communes avec la plus forte consommation d'électricité dans le secteur agricole ?
   Donnez le code, le libellé, le code postal des communes ainsi que la valeur de la consommation en question.
   Classez par consommation d'électricité dans le secteur agricole décroissante.
2. Pour l'année 2018, quelles sont les 18 communes avec la plus faible consommation tertiaire.
   Donnez le code, le libellé, le code postal des communes ainsi que la valeur de la consommation en question.
   Classez par consommation tertiaire croissante.

3. Pour l'année 2019, quelles sont les 19 communes par la plus forte consommation en gaz.
   Donnez le code, le libellé, le code postal des communes ainsi que la valeur de la consommation en question.
   Classez par consommation de gaz croissante.
4. Pour une commune que vous choisirez, quelle est la consommation d'électricité au cours des années pour lesquelles vous avez des mesures.
5. Pour l'année 2015, quelle est la consommation moyenne d'électricité par habitants de chaque commune ?
   Donnez le code, le libellé, le code postal des communes ainsi que la valeur de la consommation en question.
6. Pour l'année 2016, quelle est est la consommation moyenne d'électricité par habitants de chaque département ?
   Donnez le code, le libellé du département ainsi que la valeur de la consommation en question.
7. Pour l'année 2017, pour un département que vous choisirez, quelle est la proportion de la consommation d'électricité par rapport à la consommation totale ?
   Donnez le code, le libellé du département ainsi que la consommation demandée.
   Donnez aussi le nombre de relevés de consommation d'électricité et le nombre de relevés de consommation de gaz.
8. Pour l'année 2018, quelle est la proportion de consommation d'électricité de chaque secteur ? Donnez le code, le nom ainsi que les quatre ratios demandés.
9. Qu'est-ce que l'indice qualité d'une consommation ? Comment sont les indice qualités des consommations répartis entre les seuils 1, 0.9, 0.7, 0.5, 0.1 ?
10. Reprenez les dernières questions en ne conservant que les consommation d'indice qualité supérieur à un seuil que vous choisirez et dont vous argumenterez le choix. Les résultats changent-ils ?
11. Que représente le pdl d'une consommation ? Reprenez les dernières questions en pondérant les valeurs des consommation par leur pdl. Les résultats changent-ils ?

Complétion et Correction

> *Lors de l'analyse, vous vous rendez compte que de nombreuses valeurs manquent. Vous décidez donc d'ajouter des données pour corriger ces manques. Pour cela, vous extrapolez ces valeurs.*

1. Proposez au moins deux méthodes d'extrapolation de données pour remplacer les valeurs `NULL` dans `Consommation()`.
2. Remplacer le plus de valeurs `NULL` dans `Consommation()` par des valeurs extrapolées.
3. Appliquez la même méthodologie pour ajouter des relevés manquants.
4. Serait-il possible d'utiliser ces méthodes pour prédire les consommations futures ?

Seconde Analyse

> *Maintenant que le jeu de données est plus complet, vous rejouez la première analyse. De plus, votre succès impressionne votre commanditaire qui vous demande de collaborer avec d'autres équipes pour répondre à des requêtes portant sur plusieurs bases de données.*

0. Rejouez les requêtes de la première analyse et comparez les résultats.

1. À l'échelle du pays, par catégorie, quelle est la proportion de consommation d'électricité par rapport à la consommation globale d'énergie ?
2. Pour l'année 2019, quelle est est la consommation d'électricité moyenne par habitants de chaque région ?
   Donnez le code, le libellé de la région ainsi que la valeur de la consommation en question.
3. Pour l'année 2020, quelle est est la consommation d'électricité moyenne par habitants en France ?
4.
5.

Représentation graphique — Bonus

1. Dessinez les courbes de consommation d'électricité et de gaz par habitants en France au cours du temps.
2. Pour une année que vous choisirez, représentez consommation d'énergie en France par département sur une carte de France. Vous pouvez utilisez un outil tel que `pygam`[1] pour cela.

Conclusion & Retour d'expérience

En tirant parti de l'expérience acquise lors de ce projet, discutez des différentes technologies (SQL, Python, ORM) que vous avez manipulé et de leurs intérêts respectifs.

Une rapide présentation orale sera faite devant le chargé de TP.

Références

[1] AGENCE ORE & GESTIONNAIRES DE RÉSEAUX ÉLECTRICITÉ ET GAZ. *Consommation annuelle d'électricité et gaz par commune et par secteur d'activité*. Mis à jour le 3 novembre 2022. 2023. URL : https://www.data.gouv.fr/fr/datasets/consommation-annuelle-delectricite-et-gaz-par-commune-et-par-secteur-dactivite/ (visité le 19/10/2023).

[2] INSTITUT NATIONAL DE LA STATISTIQUE ET DES ÉTUDES ÉCONOMIQUES. *Évolution et structure de la population en 2011*. 2023. URL : https://www.insee.fr/fr/statistiques/2044745 (visité le 19/10/2023).

[3] INSTITUT NATIONAL DE LA STATISTIQUE ET DES ÉTUDES ÉCONOMIQUES. *Évolution et structure de la population en 2012*. 2023. URL : https://www.insee.fr/fr/statistiques/2044748 (visité le 19/10/2023).

[4] INSTITUT NATIONAL DE LA STATISTIQUE ET DES ÉTUDES ÉCONOMIQUES. *Évolution et structure de la population en 2013*. 2023. URL : https://www.insee.fr/fr/statistiques/2044751 (visité le 19/10/2023).

[5] INSTITUT NATIONAL DE LA STATISTIQUE ET DES ÉTUDES ÉCONOMIQUES. *Évolution et structure de la population en 2014*. 2023. URL : https://www.insee.fr/fr/statistiques/2862200 (visité le 19/10/2023).

[6] INSTITUT NATIONAL DE LA STATISTIQUE ET DES ÉTUDES ÉCONOMIQUES. *Évolution et structure de la population en 2015*. 2023. URL : https://www.insee.fr/fr/statistiques/3564100 (visité le 19/10/2023).

1. https://www.pygal.org/en/stable/documentation/types/maps/pygal_maps_fr.html, visité le 10-11-2023

[7] INSTITUT NATIONAL DE LA STATISTIQUE ET DES ÉTUDES ÉCONOMIQUES. *Évolution et structure de la population en 2016*. 2023. URL : https://www.insee.fr/fr/statistiques/4171334 (visité le 19/10/2023).

[8] INSTITUT NATIONAL DE LA STATISTIQUE ET DES ÉTUDES ÉCONOMIQUES. *Évolution et structure de la population en 2017*. 2023. URL : https://www.insee.fr/fr/statistiques/4515565 (visité le 19/10/2023).

[9] INSTITUT NATIONAL DE LA STATISTIQUE ET DES ÉTUDES ÉCONOMIQUES. *Évolution et structure de la population en 2018*. 2023. URL : https://www.insee.fr/fr/statistiques/5395875 (visité le 19/10/2023).

[10] INSTITUT NATIONAL DE LA STATISTIQUE ET DES ÉTUDES ÉCONOMIQUES. *Évolution et structure de la population en 2019*. 2023. URL : https://www.insee.fr/fr/statistiques/6456153 (visité le 19/10/2023).

[11] INSTITUT NATIONAL DE LA STATISTIQUE ET DES ÉTUDES ÉCONOMIQUES. *Évolution et structure de la population en 2020*. 2023. URL : https://www.insee.fr/fr/statistiques/7632446 (visité le 19/10/2023).