

Contexte

Le projet s'inscrit dans le domaine de la santé, où les données sont nombreuses et complexes. L'objectif est de concevoir une base de données relationnelle à partir de données inspirées de la réalité, de l'implémenter, puis d'en réaliser une analyse exploratoire.

Méthodologie

- 1 Concevoir les modèles de base de données (MCD, MLD, MPD)
- 2 Implémenter la base de données (création des tables et insertion des données)
- 3 Explorer la base de données et générer un rapport Excel

Outils utilisés

 
 

Vocabulaire

- **Classe d'entité:** représente un ensemble d'objets ou de concepts partageant les mêmes caractéristiques (ex: Patients, Assurances).
- **Attribut:** propriété d'une classe d'entité (ex: nom, date de naissance).
- **Association:** lien sémantique entre deux ou plusieurs classes d'entités (ex: un patient est couvert par une assurance).
- **MCD (Modèle Conceptuel de Données):** représentation abstraite des classes d'entités, de leurs attributs et de leurs associations, sans contrainte technique.
- **MLD (Modèle Logique de Données):** représentation des données sous forme de tables relationnelles, avec des contraintes... prête à être traduite dans un SGBD (ici MySQL).

Nettoyage des données et conception de la base

FICHIER Informations sur les patients, docteurs, hôpitaux, assurances, médicaments,...

NETTOYAGE DU FICHIER

- correction des types de données
- traitement des valeurs manquantes
- correction des incohérences et erreurs de typographie
- suppression des doublons
- gestion des valeurs aberrantes
- création de nouvelles colonnes

Pandas(Python)

Des données propres pour des résultats valides !

BRUT

	Name	Doctor	Hospital	Insurance Provider	Billing Amount
100	mARcUS ZAmOrA	Jeremiah Wolf	Hernandez, Ritter and Huffman	Cigna	25425.73 €
132	ashLEY ERICKSON	Gerald Hooper	and Johnson Moore, Branch	Aetna	-502.51 €
799	CHRisTOPHer wEISS	Kelly Thompson	Hunter-Hughes	Aetna	-1018.25 €
1018	Ashley WaRNER	Andrea Bentley	Wagner, Lee Klein	Aetna	-306.36 €
1421	JAY galloWaY	Debra Everett	Group Peters	Blue Cross	-109.10 €

NETTOYÉ

	Name	Doctor	Hospital	Insurance Provider	Billing Amount
100	Marcus Zamora	Jeremiah Wolf	Hernandez, Ritter and Huffman	Cigna	25425.73
132	Ashley Erickson	Gerald Hooper	Johnson Moore, Branch	Aetna	502.51
799	Christopher Weiss	Kelly Thompson	Hunter-Hughes	Aetna	1018.25
1018	Ashley Warner	Andrea Bentley	Wagner, Lee Klein	Aetna	306.36
1421	Jay Galloway	Debra Everett	Group Peters	Blue Cross	109.10

Une fois les données nettoyées, corrigées et prêtes à l'emploi, elles peuvent être structurées de façon optimale pour l'analyse.

MODÉLISATION

MODÈLE CONCEPTUEL

Représente les entités et leurs relations pour clarifier la structure des données.

```

    graph TD
        Patients[Patients  
1,N  
id_pat  
nom_pat  
prenom_pat  
genre  
groupe_sanguin] -- couvrir --> Assurances[Assurances  
1,N  
id_assur  
nom_assur]
        Patients -- admettre --> Hospitalisations[Hospitalisations  
1,1  
id  
date_admission  
date_sortie  
type_admission  
montant  
age_pat]
        Hospitalisations -- presenter --> Patients
        Hospitalisations -- regrouper --> Patients
        Hospitalisations -- affecter --> Docteurs[Docteurs  
1,N  
id_doc  
nom_doc  
prenom_doc]
        Docteurs -- travailler --> Hopitaux[Hopitaux  
1,N  
id_hosp  
nom_hosp]
    
```

MODÈLE LOGIQUE

transforme le MCD en tables et attributs pour préparer la création de la base de données.

Patients ([id_pat](#), nom_pat, prenom_pat, genre, groupe_sanguin)
 Assurances ([id_assur](#), nom_assur)
 Assu_Pat (#[id_assur](#), #[id_pat](#))
 Docteurs ([id_doc](#), nom_doc, prenom_doc)
 Hopitaux ([id_hosp](#), nom_hosp)
 Doct_hosp (#[id_doc](#), #[id_hosp](#))
 Hospitalisations ([id](#), date_admission, date_sortie, type_admission, montant, age, #[id_doc](#), #[id_med](#), #[id_pat](#), #[id_test](#), #[id_cond](#))

Une fois la conception terminée, les tables du MLD sont traduites en tables physiques : c'est l'implémentation de la base de données.

Source : [Kaggle](#)

Implémentation de la base, exploration et visualisation

IMPLÉMENTATION DE LA BASE DE DONNÉES

MODÈLE PHYSIQUE

La base de donnée conçue a été implémentée automatiquement depuis Python via MySQL Workbench.

INSERTION DES DONNÉES

Les données nettoyées avec Pandas ont été insérées dans la base MySQL à l'aide de scripts Python.

```
try :
    curseur = connexion.cursor(dictionary=True)
    placeholder = ', '.join(['%s'] * len(valeurs))
    req = f""" INSERT IGNORE INTO {nom_table} {chaine_cols}
    VALUES ({placeholder}); """
    curseur.execute(req, valeurs)
except mysqlcon.Error as e :
    print(f"Erreur lors de l'insertion : {e}")
```

La structure `try...except` permet d'intercepter et de traiter les erreurs lors de l'insertion, assurant ainsi la robustesse du processus.

DONNÉES
54966 lignes
17 colonnes

Conseil : utiliser des requêtes paramétrées pour sécuriser les échanges avec la base et éviter les injections SQL

Temps d'insertion : 5 min

ANALYSE EXPLORATOIRE DES DONNÉES

NOMBRE DE SÉJOURS HOSPITALIERS

Répartition des séjours par année

Année	Nombre de séjours hospitaliers
2020	11500
2023	11000
2022	11000
2021	11000
2019	7500
2024	4000

Le nombre de séjours hospitaliers diminue au fil du temps.

TOP 5 DES HOPITAUX PAR BUDGET

Top 5 des Hôpitaux par Budget

Hôpital	Budget
Johnson PLC	1080000
LLC Smith	1060000
Smith PLC	1040000
Ltd Smith	980000
Smith Ltd	920000

Le plus grand budget d'hôpital pendant la période est de 1,080,000\$ et est mobilisé par l'hôpital Johnson PLC.

PERSPECTIVES D'AMÉLIORATION

- Insérer une période de validité pour les assurances
- Réduire le temps d'insertion des données
- Enregistrer les médecins suppléants du médecin principal
- Permettre l'insertion de plusieurs médicaments pour une même hospitalisation
- Associer une localisation géographique à chaque hôpital