ANNEE	2024
PROGRAMME	Data Ananlyst
CURSUS	Bootcamp
Date de remise	17 Mai 2024

Capstone 2

Nom, prénom des stagiaires : PHOTHISENH Guillaume

Sujet du Capstone 2 : Analyse et Gestion des Données des Employés pour DataTech Solutions

Table des matières

I. Contexte

En tant que Data Analyst indépendant, la société DataTech Solutions aux Etats-Unis a fait appel à vous pour analyser et sécuriser plusieurs jeux de données concernant ses employés actuels. Au total, elle met à votre disposition 6 fichiers à traiter.

La société attend de vous que vous puissiez nettoyer, préparer les données puis de les charger dans une base SQL que vous créerez. Un des objectifs est de s'assurer que l'on pourra facilement accéder et interroger ces données dans un futur proche

II. Objectifs

A. Data wrangling

Un rappel du data process :

1. Récupérer la donnée
2. Profiler
3. Préparer (= Nettoyer + Transformer)
4. Analyser (= Filtrer + Agréger)
5. Visualiser la donnée
6. Enrichir

Après avoir récupérer les données, il est nécessaire de répondre à différentes questions essentielles telles que :

- Sous quelle forme ? BDD, fichier plat, non structuré, etc.
- Quel format ? Du texte ou autre ("binary") ?
- Quelle structure ? Table, objet, etc.
- Quel volume ? Nombre de lignes, de colonnes
- A quoi correspond une observation / a record ? Ex. Une ligne = une transaction
- Quelle est la clef primaire (ID)
- Quel périmètre ? Temporel, spatial, organisationnel, etc.
- Qui est l'émetteur / le responsable de la donnée ?

Après avoir répondu à la majorité des questions il est également important de comprendre de quoi il s'agit.

Dans notre cas, cela concerne des données d'employés d'une entreprise (nom, prénom, adresse…). Ensuite, je dois vérifier la qualité des données et détecter les problèmes (doublons, valeurs manquantes, problèmes d'unités, nommage…) :

Pour ce faire, je vais utiliser Python avec Pandas afin de visualiser les données.

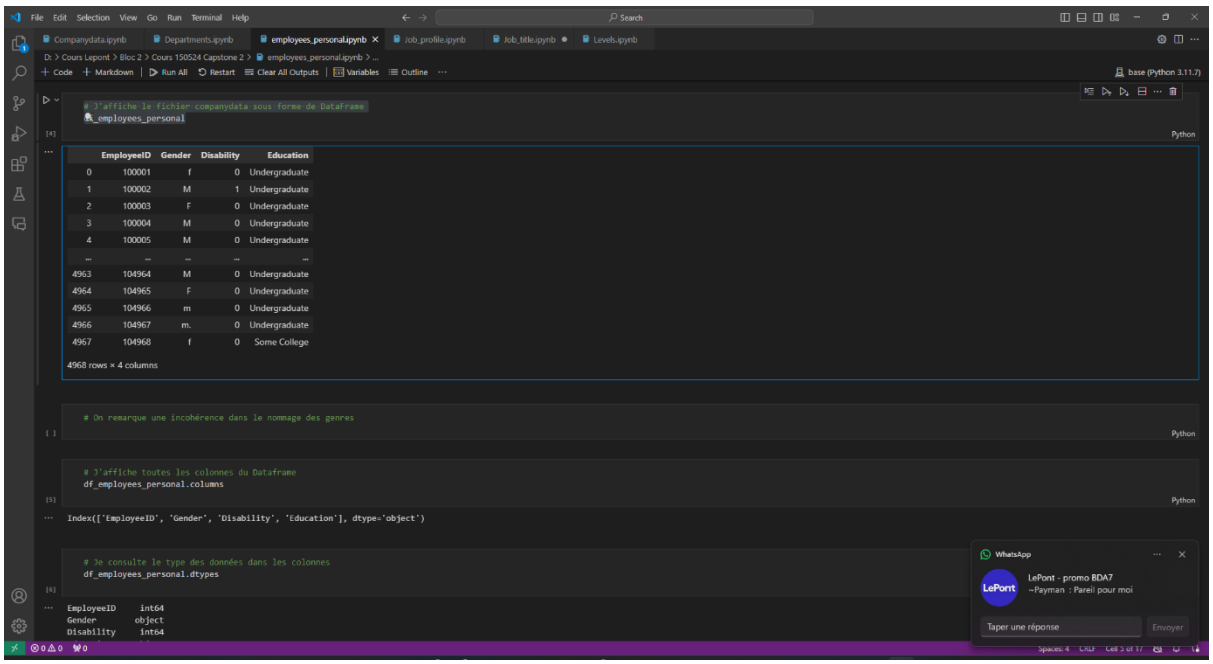
Ex :

	EmployeeID	First_Name	Surname	StreetAddress	Age	Office	Office_Type	Active Status	Notes	Start_Date	Termination_Date
0	100001	patrice	moore	1427 Buckhannan Avenue	35	NYC	Corporate	1	Changes for 2021.06:	05042009	December 12, 2100
1	100002	dAViD	rickards	4265 Graystone Lakes	49	NYC	Corporate	1	Changes for 2021.06:	May 04-2009	December 12, 2100
2	104964	grace	maldonado	1680 Hudson Street	32	NYC	Corporate	0	Changes for 2021.06:	05182009	06052013
3	100004	juSTIn	edgin	1262 Limer Street	25	Boulder	Corporate	0	Changes for 2021.06:	06/22/2009	October 16, 2013
4	100005	bENJAMiN	vargas	2431 Rainbow Road	49	NYC	Corporate	0	Changes for 2021.06:	07/13/2009	01/10/2011
...	...	...	...	...	...	...	...	...	...	...	...
4983	100115	heAtH	roland	2708 Masonic Hill Road	60	NYC	Corporate	0	Changes for 2021.06:	07262010	12/30/2010
4984	103482	bilLie	elliott	1558 Hummingbird Way	35	Boulder	Technology	1	Changes for 2021.06:	October 23, 2017	December 12, 2100
4985	100133	aSHLEY	orr	4215 Devils Hill Road	45	NYC	Corporate	0	Changes for 2021.06:	Aug 23-2010	08/04/2014
4986	103244	mariLyn	key	2420 Taylor Street	58	NYC	Corporate	1	Changes for 2021.06:	July 03, 2017	12122100
4987	102926	tHOmaS	mitchell	3099 Rose Street	43	NYC	Corporate	1	Changes for 2021.06:	Apr 17-2017	12/06/2021

4988 rows × 11 columns

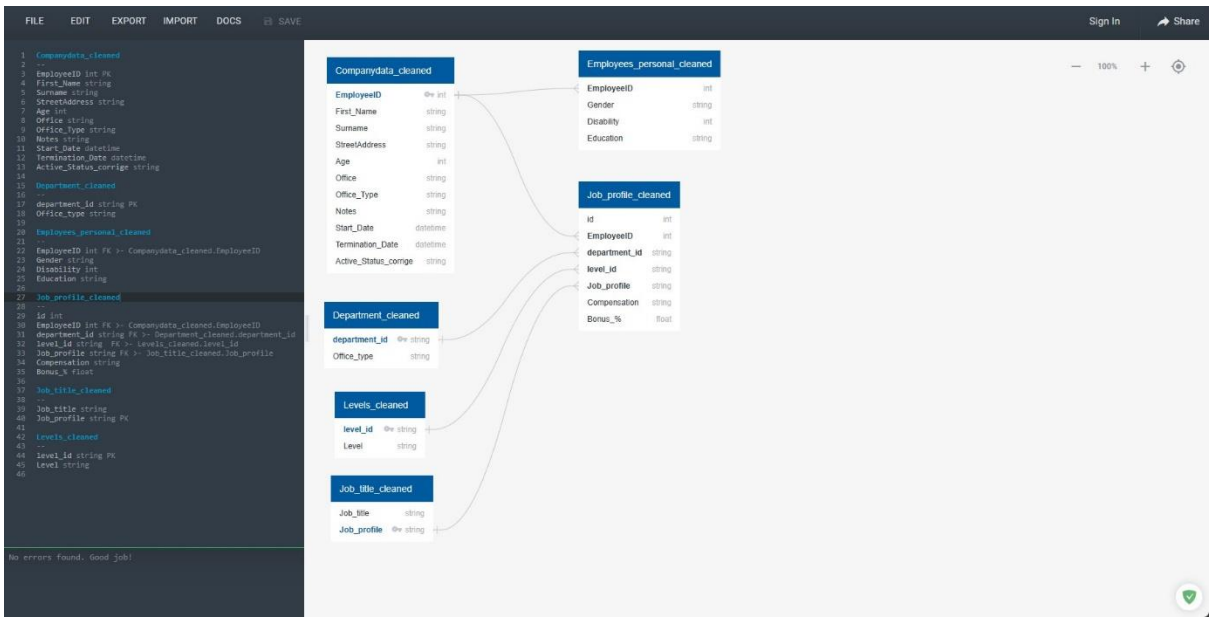
```
# on remarque que les données ne sont pas de qualité. En effet, il est nécéssaire de proceder à différents nettoyages:
```

J’ai pu identifier certaines anomalies que je vais préparer, nettoyer et analyser par la suite. On fait de même sur les autres fichiers remis afin d’avoir des données de qualité et propres pour pouvoir les exploiter dans SQL.



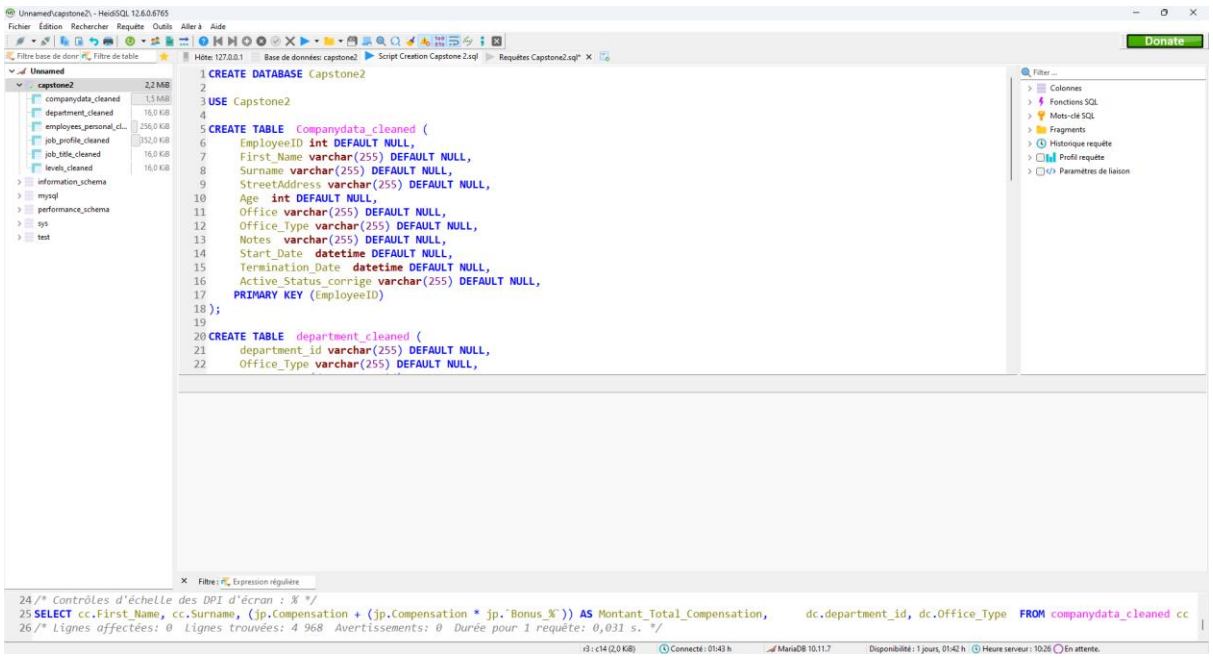
B. Une base de données SQL

Avant de créer une base de données, je conçois un schéma relationnel des différentes tables. J'ai utilisé le site recommandé : https://www.quickdatabasediagrams.com/.

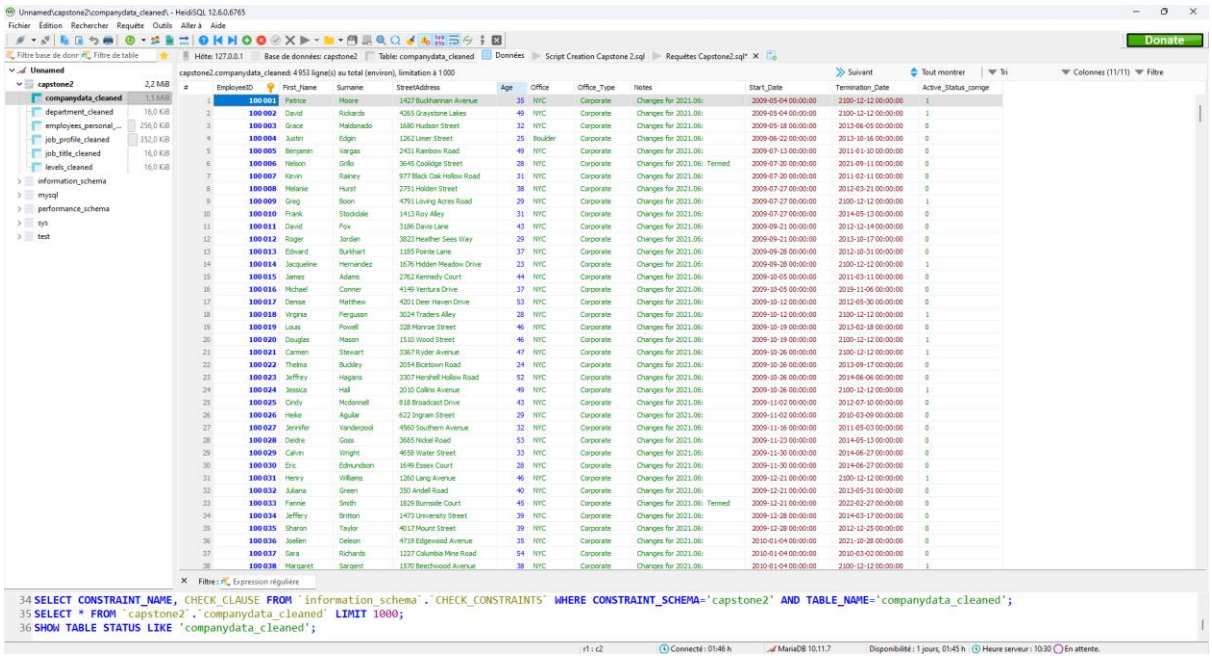


Dans cette partie, j'ai créé les relations de chaque table et définit les clés primaires et étrangères.

Je m'attaque à la création de la base de données ainsi que les tables sous Heidi SQL



Après la création, j'importe les données sous format csv afin d'alimenter les tables et exploiter ses données par des requêtes.

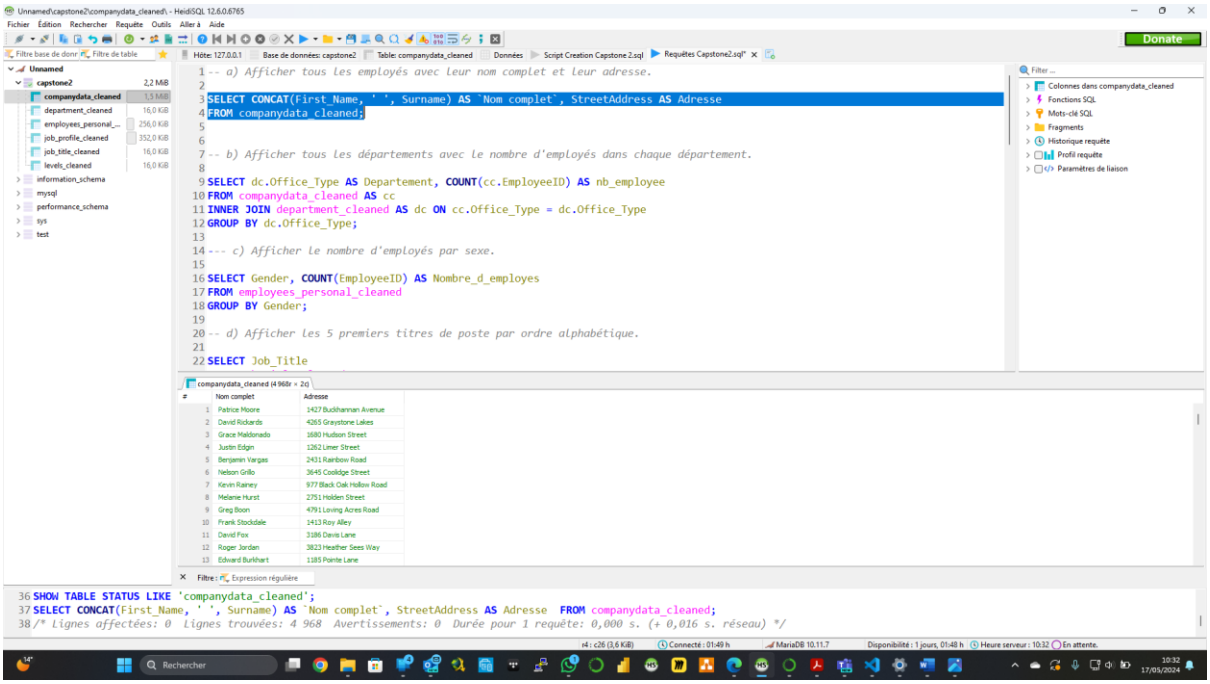


J'écris les requêtes afin de répondre aux questions suivantes

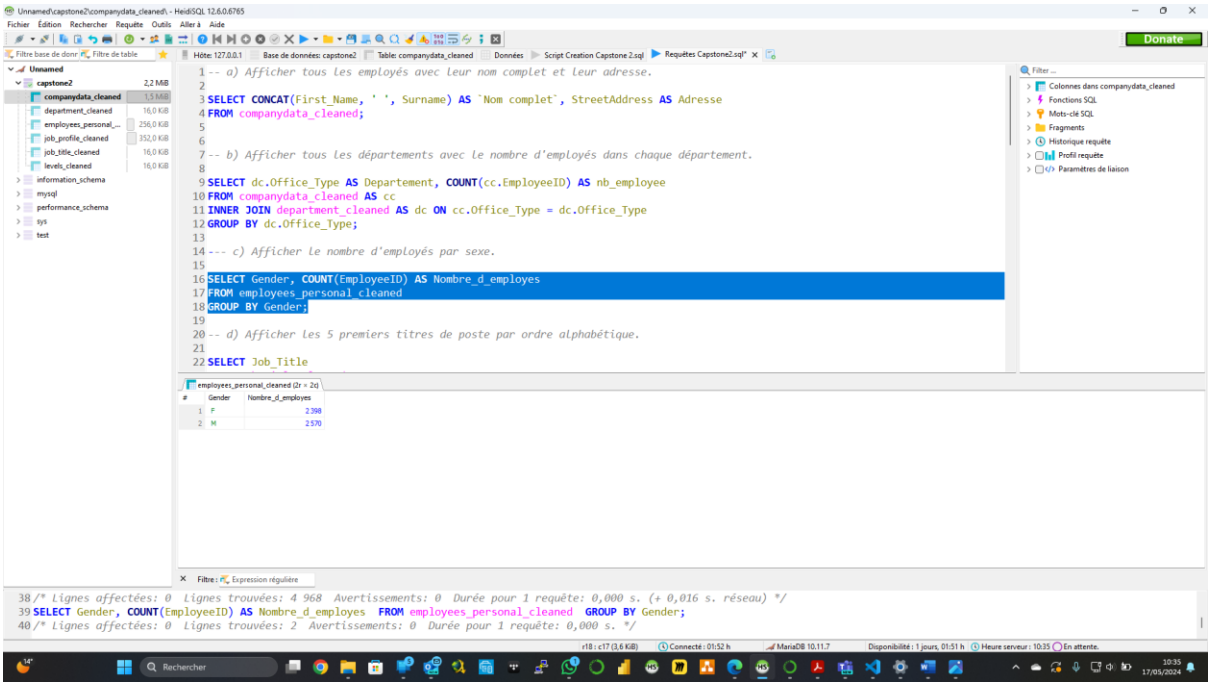
C. Une réponse aux questions suivantes à l'aide de requêtes SQL :

Exemples de requêtes :

- Afficher tous les employés avec leur nom complet et leur adresse.



- Afficher le nombre d'employés par sexe.



III. Le contenu attendu

Le travail délivré contient les éléments suivants :

- Le support de votre présentation orale au format pdf ou Powerpoint (juste quelques slides pour montrer votre travail).
- Un ou plusieurs Notebooks (format .ipynb) explicitant vos étapes dans la data « prep » .
- Un fichier image (png ou jpeg) de votre schéma de données (Entity-Relationship Diagram).
- Un script au format .sql qui permettra de créer votre base de données.
- L’ensemble de vos requêtes SQL dans un éditeur de texte.
- Les résultats de vos requêtes SQL placés dans tableur Excel.