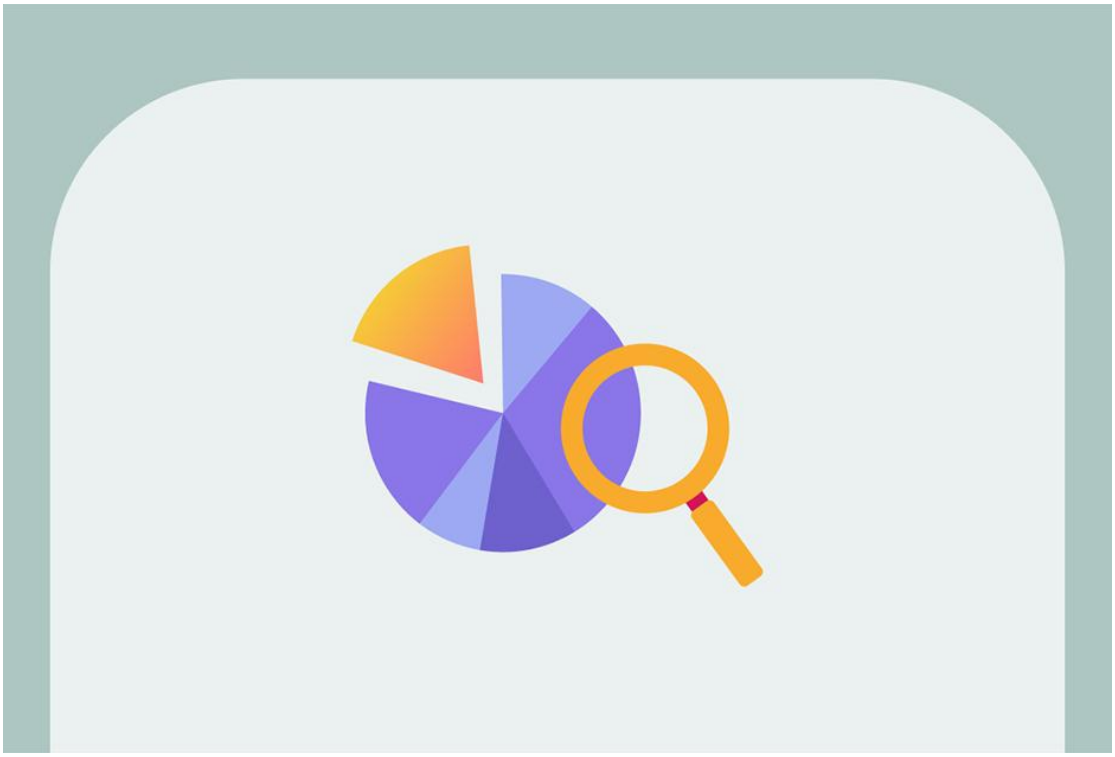


Requêtez une base de données avec SQL

Bootcamp Data Analyst – Zied Belhocine

Etape 1 – Découvrez les différents types de données

Nous disposons de trois documents relatifs aux contrats d'assurance habitation en France :

- Un fichier CSV contenant les données des contrats clients

Contrat_ID	No_voie	B_T_Q	Type_de_voie	Voie	Code_dep_code_commune	Code_postal	Surface	Type_local	Occupation	Type_contrat	Formule	Valeur_declaree_biens	Prix_cotisation_mensuel
100601	190	A	RUE	CENTRALE	1350	1370	50	Appartement	Locataire	Residence principale	Integral	0-25000	25
100602	347		RUE	DU CHATEAU	1103	1170	48	Appartement	Locataire	Residence principale	Classique	0-25000	30
100603	58		AV	DU MONT BLANC	1143	1220	131	Appartement	Proprietaire	Residence principale	Integral	25000-50000	57
100604	140		RUE	DE L'ABBE JOLIVET	1288	1630	109	Maison	Locataire	Residence principale	Integral	25000-50000	43
100605	39		RUE	BUFFON	1033	1200	109	Appartement	Locataire	Residence principale	Classique	0-25000	33
100606	8		RUE	DE GENEVE	1354	1630	53	Appartement	Proprietaire	Residence principale	Classique	0-25000	19
100607	2		RUE	DU RECULET	1354	1630	59	Appartement	Proprietaire	Residence principale	Integral	0-25000	15
100608	1403		RUE	JEAN DE GINGINS	1143	1220	93	Maison	Proprietaire	Mise en location	Integral	25000-50000	34

- Un fichier CSV contenant le référentiel géographique des régions françaises

Code_dep_code_commune	reg_code	reg_nom	aca_nom	dep_nom	com_nom_maj_court	dep_code	dep_nom_num
1001	84	Auvergne-Rhône-Alpes	Lyon	Ain	L ABERGEMENT CLEMENCIAT	1	Ain (01)
1002	84	Auvergne-Rhône-Alpes	Lyon	Ain	L ABERGEMENT DE VAREY	1	Ain (01)
1003	84	Auvergne-Rhône-Alpes	Lyon	Ain	AMAREINS	1	Ain (01)
1004	84	Auvergne-Rhône-Alpes	Lyon	Ain	AMBERIEU EN BUGEY	1	Ain (01)
1005	84	Auvergne-Rhône-Alpes	Lyon	Ain	AMBERIEUX EN DOMBES	1	Ain (01)
1006	84	Auvergne-Rhône-Alpes	Lyon	Ain	AMBLEON	1	Ain (01)

- Un dictionnaire de données recensant les champs utilisés, leur type, taille et définition

	Nom des colonnes	Type de données	Taille	Clé	Description
CONTRAT.CSV	Contrat_ID	INT		Clé primaire	Id unique pour les contrats
	No_voie	INT			Numéro dans la voie pour l'adresse du logement assuré
	B_T_Q	CHAR	1		Indicateur éventuel de répétition pour l'adresse du logement assuré sur un caractère
	Type_de_voie	CVARCHAR			Type de voie pour l'adresse du logement assuré: rue, av (Avenue), rte (Route), ...
	Voie				Libellé de la voie pour l'adresse du logement assuré
	Code_dep_code_commune			Clé secondaire	Concaténation du code département et code commune pour avoir une clé unique
	Code_postal				Code postal pour l'adresse du logement assuré
REGION.CSV	Code_dep_code_commune			Clé primaire	Concaténation du code département et code commune pour avoir une clé unique

La première étape consiste à établir une correspondance entre, d'une part, les noms et les types de données des colonnes de notre fichier CSV et, d'autre part, les spécifications définies dans le dictionnaire de données. Ce rapprochement permet de déterminer la composition des champs qui structureront les tables de notre future base de données. Il est également important de décrire et de contextualiser la signification de ces données afin que l'utilisateur qui exploitera le jeu de données puisse en saisir la nature exacte et le rôle.

Pour définir la taille, on peut utiliser la fonction MAX(NBCAR(([**colonne à évaluer**])) pour déterminer la taille à accorder à un champ (pour les VARCHAR par exemple).

Pour ce qui est du type et de la description des données, il faut parcourir les différentes données de la colonne à l'aide des filtres pour savoir si ce qu'elle contient. Par exemple :

- Prix_cotisation_mensuel ne contient que des nombres entiers d'une longueur maximale de 3. Ce sera donc un INTEGER qui donne le prix mensuel d'une cotisation sur un contrat.

- Type_local contient 2 valeurs possibles : Appartement / Maison. C'est donc un VARCHAR d'une longueur maximale de 11 qui renvoie la nature du logement assuré.

Pour la table **Contrat** :

- Voie : VARCHAR, taille max = 26 caractères
- Code_dep_code_commune : INTEGER
- Code_postal : INTEGER
- Surface : INTEGER, Description : Surface du logement assuré
- Type_local : VARCHAR, taille max = 11 caractères Description : Nature du logement assuré : Appartement, maison, …
- Occupation : VARCHAR, taille max = 12 caractères Description : Nature de l'occupant du logement : Locataire, Propriétaire
- Type_contrat : VARCHAR, taille max = 20 caractères Description : Type d'occupation du bien : résidence principale, mise en location, résidence secondaire.
- Formule : VARCHAR, taille max = 9 caractères Description : Formule du contrat affectée au logement : Integral, Classique
- Valeur_declaree_biens : VARCHAR, taille max = 12 caractères, Description : Tranche estimée de la valeur des biens présetns dans le logement
- Prix_cotisation_mensuel : INTEGER, Description : Prix payé chaque mois par le souscripteur

Pour la table **Region** :

- reg_code : INTEGER, Description : Code à un ou deux chiffre pour désigner la région
- reg_nom : VARCHAR, taille max = 26 caractères Description : Nom de la région où se situe la commune
- aca_nom : VARCHAR, taille max = 24 caractères Description : Nom de l'académie à laquelle est rattachée la commune
- dep_nom : VARCHAR, taille max = 43 caractères Description : Nom du département où se situe la commune
- com_nom_maj_court : VARCHAR, taille max = 32 caractères Description : Nom de la commune en majuscule avec les abréviations (ex : 'Saint' est noté 'ST')
- dep_code : INTEGER, Description : Code du département
- dep_nom_num : VARCHAR, taille max = 50 caractères Description : Concaténation du nom de département et du code département : Nom(code)

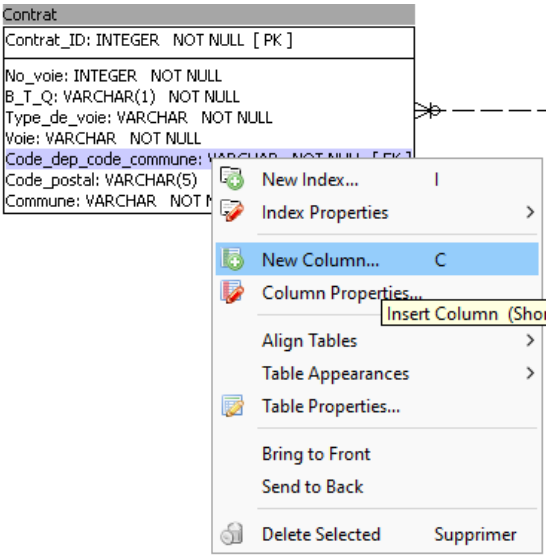
Notre Dictionnaire est maintenant rempli. Nous allons pouvoir réaliser le schéma relationnel normalisé de notre base de données.

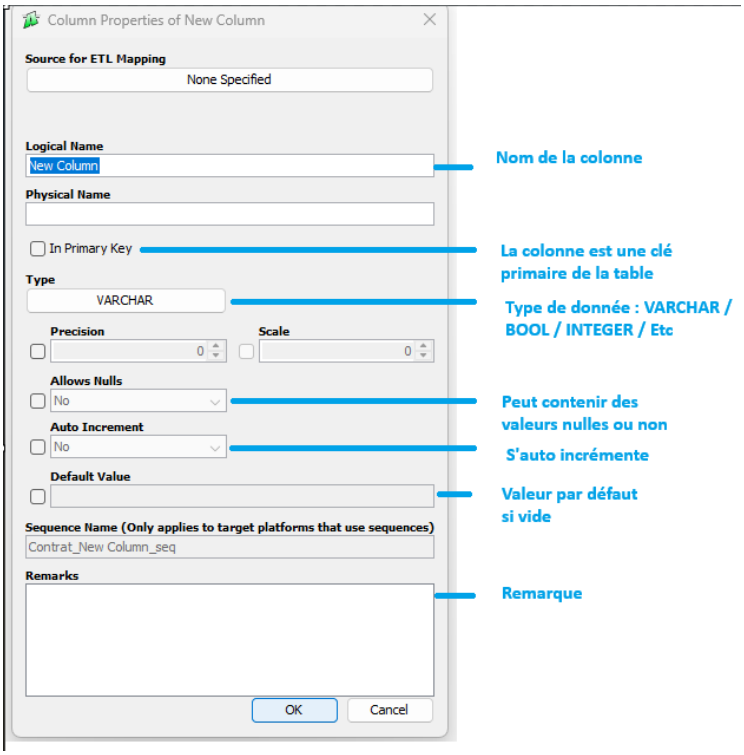
	Nom des colonnes	Type de données	Taille	Clé	Description
CONTRAT.CSV	Contrat_ID	INT	6	Clé primaire	Id unique pour les contrats
	No_voie	INT	4		Numéro dans la voie pour l'adresse du logement assuré
	B_T_Q	CHAR	1		Indicateur éventuel de répétition pour l'adresse du logement assuré sur un caractère
	Type_de_voie	VARCHAR	4		Type de voie pour l'adresse du logement assuré: rue, av (Avenue), rte (Route), …
	Voie	VARCHAR	26		Libellé de la voie pour l'adresse du logement assuré
	Code_dep_code_commune	INT	6	Clé secondaire	Concaténation du code département et code commune pour avoir une clé unique
	Code_postal	INT	5		Code postal pour l'adresse du logement assuré
	Surface	INT	3		Surface du logement assuré
	Type_local	VARCHAR	11		Nature du logement assuré : Appartement, maison, …
	Occupation	VARCHAR	12		Nature de l'occupant du logement : Locataire, Propriétaire
	Type_contrat	VARCHAR	20		Type d'occupation du bien : résidence principale, mise en location, résidence secondaire.
	Formule	VARCHAR	9		Formule du contrat affectée au logement : Integral, Classique
	Valeur_declaree_biens	VARCHAR	12		Tranche estimée de la valeur des biens présetns dans le logement
	Prix_cotisation_mensuel	INT	3		Prix payé chaque mois par le souscripteur
REGION.CSV	Code_dep_code_commune	INT	6	Clé primaire	Concaténation du code département et code commune pour avoir une clé unique
	reg_code	INT	2		Code à un ou deux chiffre pour désigner la région
	reg_nom	VARCHAR	26		Nom de la région où se situe la commune
	aca_nom	VARCHAR	24		Nom de l'académie à laquelle est rattachée la commune
	dep_nom	VARCHAR	43		Nom du département où se situe la commune
	com_nom_maj_court	VARCHAR	32		Nom de la commune en majuscule avec les abréviations (ex : 'Saint' est noté 'ST')
	dep_code	INT	3		Code du département
	dep_nom_num	VARCHAR	50		Concaténation du nom de département et du code département : Nom(code)

Etape 2 – Découvrez la conception de schéma relationnel

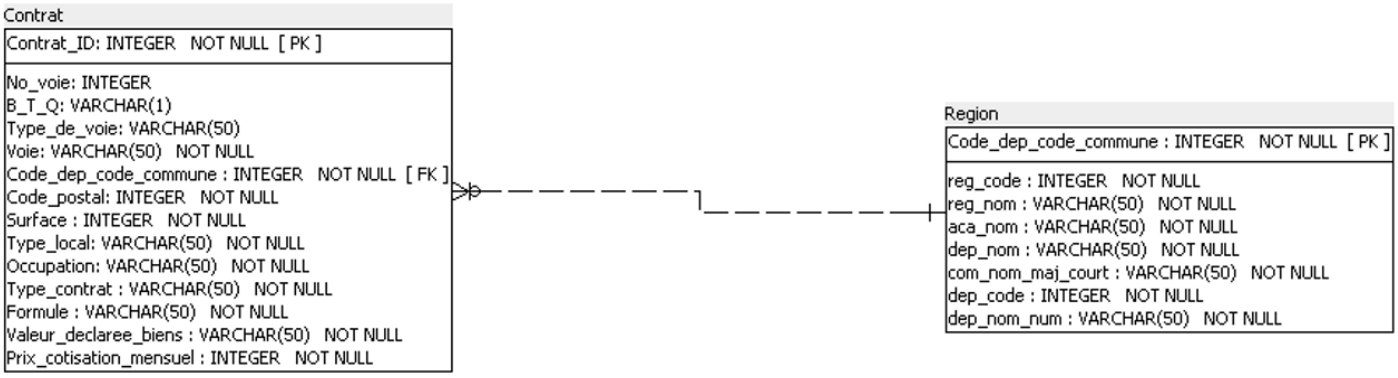
Maintenant que nous avons une idée plus claire de notre base de données, nous allons pouvoir la modéliser sous forme de schéma grâce au logiciel « SQL Power Architects ». On reçoit une ébauche :

Ici nous avons juste les données présentes dans le dictionnaire de données non rempli. Il faut donc ajouter les colonnes manquantes ainsi que les contraintes relevées dans notre dictionnaire. D'autres part, il faut aussi dire si ces colonnes peuvent être nulles ou non (NOT NULL). Seules No_voie et Type_de_voie contiennent des cellules vides. On ajoute une colonne en faisant clic droit sur la table => New Column :





Si on veut modifier une colonne existante, il suffit de double-cliquer dessus. Après avoir renseigner toute nos colonnes, on obtient ce schéma relationnel :



Note : J'ai mis VARCHAR(50) au cas où des données plus grandes venaient intégrer notre BDD.

D'autre part, il est important d'observer ici nos PK(Primary Key) et FK (Foreign Key).

Rappel :

- PK = On appelle « clé primaire » une ou plusieurs colonnes d'une table de base de données dont les valeurs identifient de manière unique chaque ligne ou enregistrement. Par exemple, une colonne contenant l'identifiant d'un employé peut être une clé primaire dans une table contenant des informations sur les employés.

- FK = Une clé étrangère, dans une base de données relationnelle, est une contrainte qui garantit l'intégrité référentielle entre deux tables. Une clé étrangère identifie une colonne ou un ensemble de colonnes d'une table comme référençant une colonne ou un ensemble de colonnes d'une autre table.

Dans la table Region, La colonne permettant d'identifier une ligne est « Code_dep_code_commune ». Dans la table Contrat, c'est « Contrat_ID ».

D'autre part, le champ « Code_dep_code_commune » est aussi présent dans la table Contrat. C'est la clé étrangère qui fait référence à la table Region. Cette relation va nous permettre de relier les contrats avec les informations géographiques dans nos futures requêtes.

Grâce à toutes ces informations, nous avons désormais les clés en main pour écrire le code SQL nécessaire à la création de nos tables Region et Contrat. La table Contrat a une clé étrangère qui fait référence à la table Region, il faut donc crée cette dernière en premier :

```
CREATE TABLE Region (
   Code_dep_code_commune  INTEGER PRIMARY KEY,
   reg_code            INTEGER,
   reg_nom             VARCHAR(50),
   aca_nom             VARCHAR(50),
   dep_nom             VARCHAR(50),
   com_nom_maj_court      VARCHAR(50),
   dep_code            INTEGER,
   dep_nom_num         VARCHAR(50)
);
```

On utilise la fonction CREATE TABLE suivie du nom de la table avec entre parenthèse les champs suivis du type et de la contrainte (s'il y en a une) ainsi que « primary key sur la colonne contenant la clé primaire.

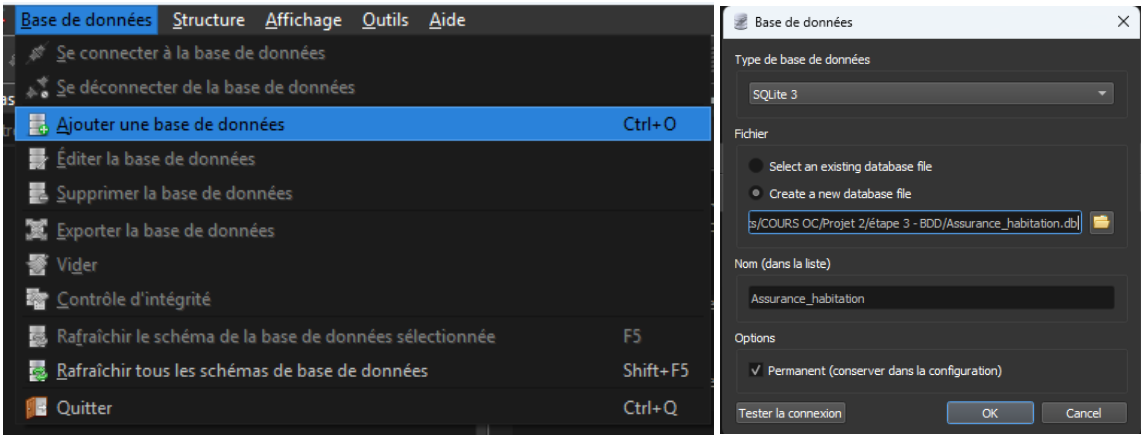
On fait la même chose avec la table Contrat, en ajoutant la ligne pour la clé étrangère :

```
CREATE TABLE Contrat (
   Contrat_ID                INTEGER PRIMARY KEY,
   No_voie                   INTEGER,
   B_T_Q                     VARCHAR(50),
   Type_de_voie              VARCHAR(50),
   Voie                      VARCHAR(50),
   Code_dep_code_commune     INTEGER,
   Code_postal               INTEGER,
   Surface                   INTEGER,
   Type_local                VARCHAR(50),
   Occupation                VARCHAR(50),
   Type_contrat              VARCHAR(50),
   Formule                   VARCHAR(50),
   Valeur_declaree_biens     VARCHAR(50),
   Prix_cotisation_mensuel   INTEGER,
   FOREIGN KEY (Code_dep_code_commune) REFERENCES
Region(Code_dep_code_commune)  ON DELETE RESTRICT
ON UPDATE CASCADE
);
```

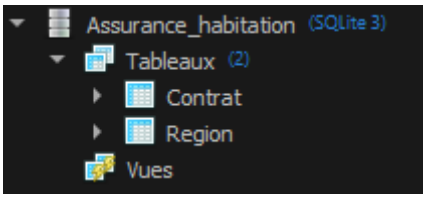
Voilà, nous sommes prêts à créer notre base de données.

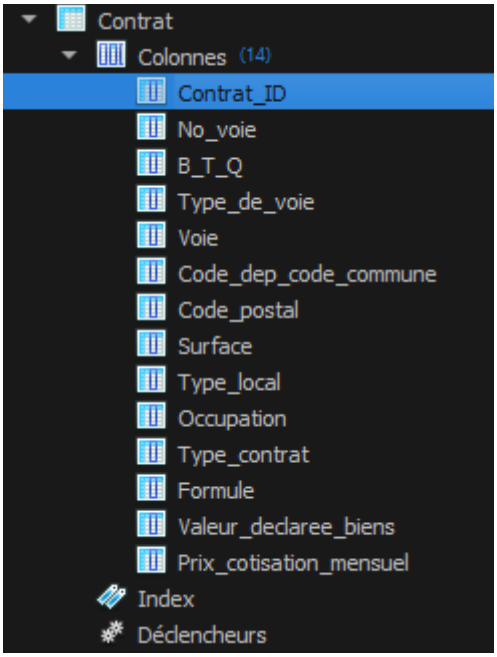
Etape 3 – Découvrez la création et le chargement d'une base de données.

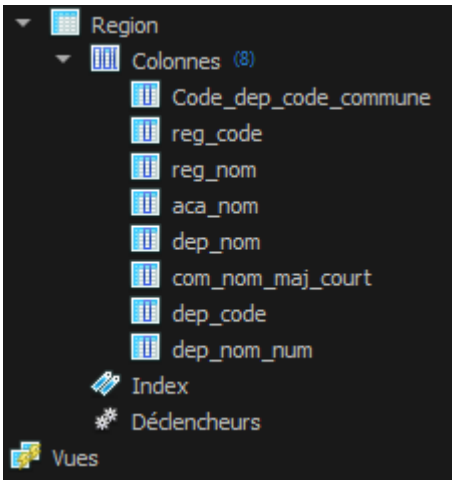
Comme dans le cours « **Requêtez une base de données avec SQL »,** j'ai choisi le logiciel SQLiteStudio (aussi pour son interface graphique). On crée une nouvelle BDD :



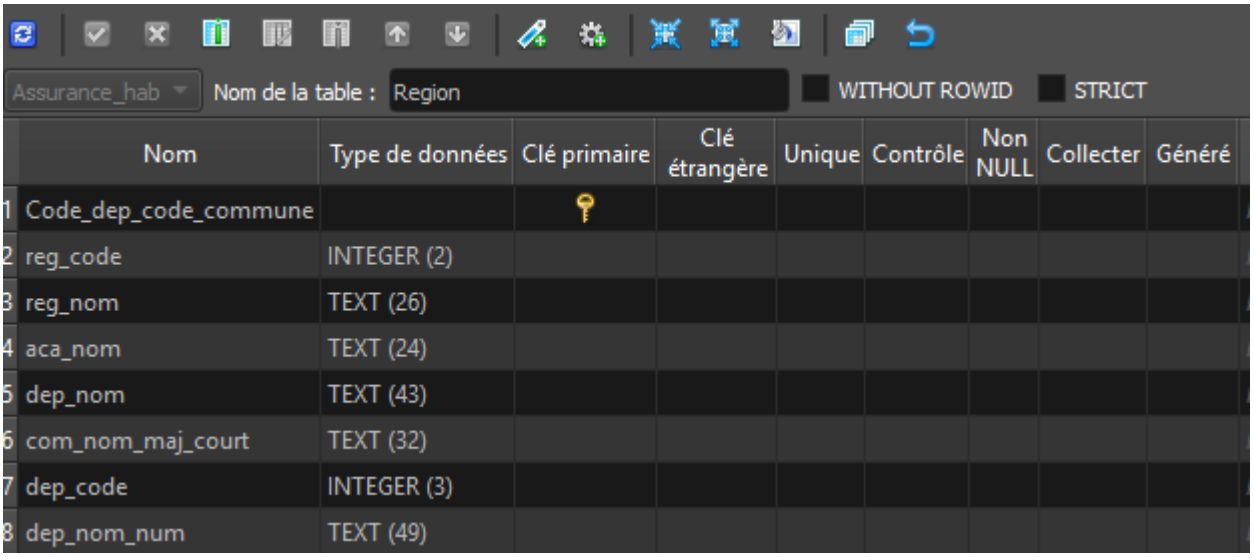
On crée nos tables en rentrant le Code SQL générant les tables.







Note : Il faut vérifier que nos PK et FK sont bien renseignées :

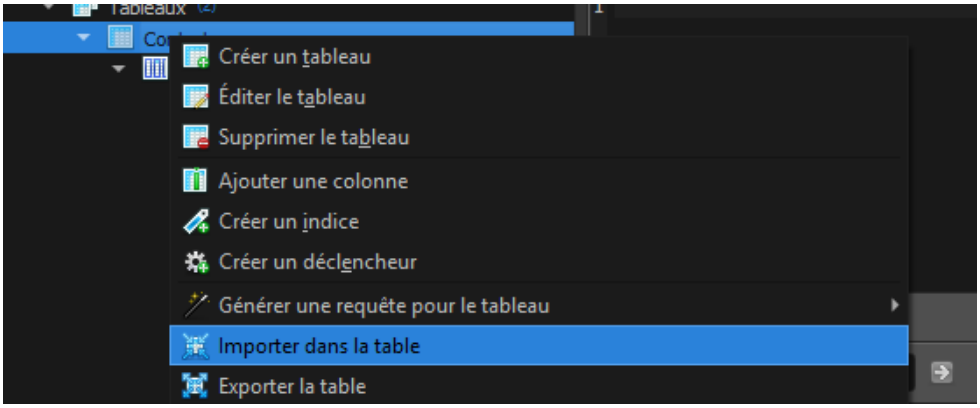


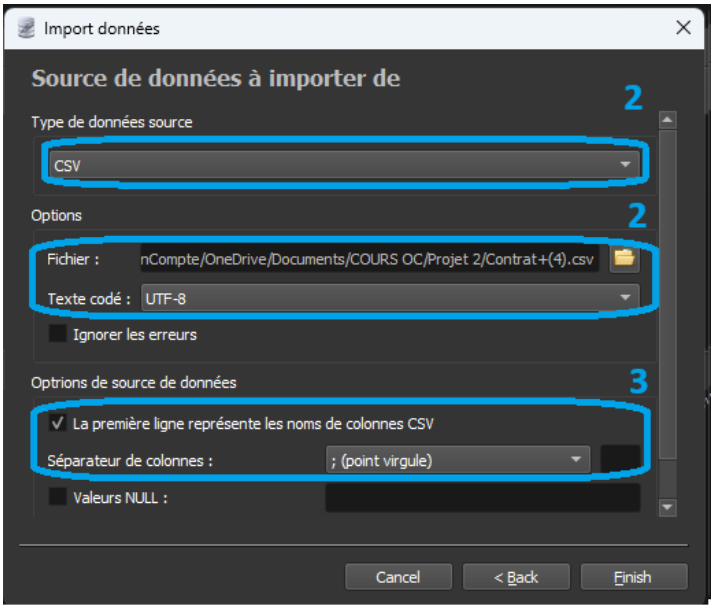
	Nom	Type de données	Clé primaire	Clé étrangère	Unique	Contrôle	Non NULL	Collecter	Généré
1	Code_dep_code_commune		✓						
2	reg_code	INTEGER (2)							
3	reg_nom	TEXT (26)							
4	aca_nom	TEXT (24)							
5	dep_nom	TEXT (43)							
6	com_nom_maj_court	TEXT (32)							
7	dep_code	INTEGER (3)							
8	dep_nom_num	TEXT (49)							

Assurance_hab | Nom de la table : Contrat | WITHOUT ROWID | STRICT

	Nom	Type de données	Clé primaire	Clé étrangère	Unique	Contrôle	Non NULL	Collecter	Généré
1	Contrat_ID	INTEGER							
2	No_voie	INTEGER							
3	B_T_Q	TEXT (1)							
4	Type_de_voie	TEXT (50)							
5	Voie	TEXT (50)							
6	Code_dep_code_commune	INTEGER							
7	Code_postal	INTEGER							
8	Surface	INTEGER							
9	Type_local	TEXT (50)							
10	Occupation	TEXT (50)							
11	Type_contrat	TEXT (50)							
12	Formule	TEXT (50)							
13	Valeur_declaree_biens	TEXT (50)							
14	Prix_cotisation_mensuel	INTEGER							

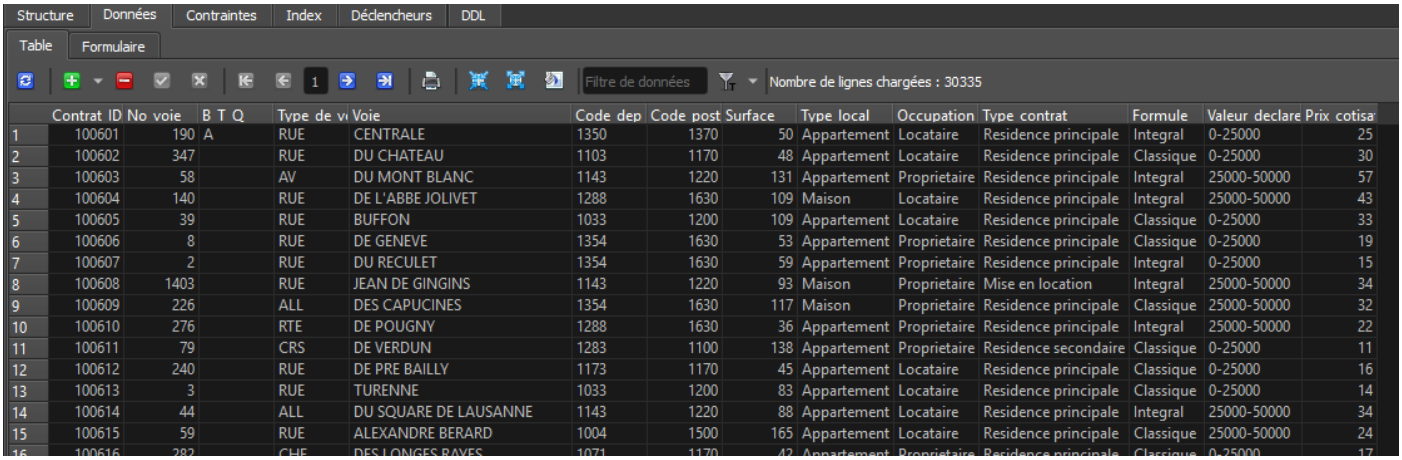
On peut maintenant importer les données dans notre base :



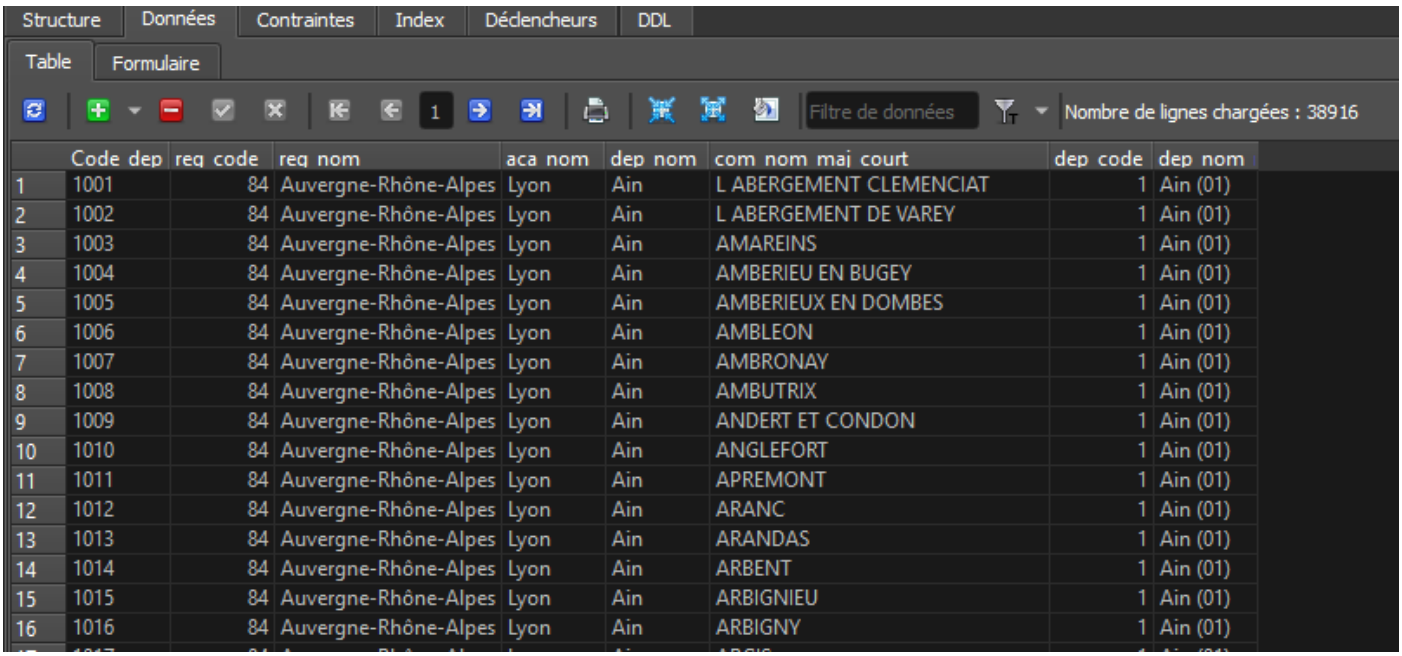


On sélectionne « CSV » comme type de données, on choisit le CSV associé, l'encodage UTF-8. Il faut aussi cocher « la première ligne représente les noms des colonnes CSV » et « ; » comme séparateur (si on ouvre notre CSV comme un fichier texte, on peut voir que c'est le « ; » qui est utilisé comme séparateur).

On a maintenant nos données chargées et prêtes à être utilisées.



	Contrat ID	No voie	B T Q	Type de vi	Voie	Code dep	Code post	Surface	Type local	Occupation	Type contrat	Formule	Valeur declare	Prix cotisa
1	100601	190	A	RUE	CENTRALE	1350	1370	50	Appartement	Locataire	Residence principale	Integral	0-25000	25
2	100602	347		RUE	DU CHATEAU	1103	1170	48	Appartement	Locataire	Residence principale	Classique	0-25000	30
3	100603	58		AV	DU MONT BLANC	1143	1220	131	Appartement	Proprietaire	Residence principale	Integral	25000-50000	57
4	100604	140		RUE	DE L'ABBE JOLIVET	1288	1630	109	Maison	Locataire	Residence principale	Integral	25000-50000	43
5	100605	39		RUE	BUFFON	1033	1200	109	Appartement	Locataire	Residence principale	Classique	0-25000	33
6	100606	8		RUE	DE GENEVE	1354	1630	53	Appartement	Proprietaire	Residence principale	Classique	0-25000	19
7	100607	2		RUE	DU RECULET	1354	1630	59	Appartement	Proprietaire	Residence principale	Integral	0-25000	15
8	100608	1403		RUE	JEAN DE GINGINS	1143	1220	93	Maison	Proprietaire	Mise en location	Integral	25000-50000	34
9	100609	226		ALL	DES CAPUCINES	1354	1630	117	Maison	Proprietaire	Residence principale	Classique	25000-50000	32
10	100610	276		RTE	DE POUGNY	1288	1630	36	Appartement	Proprietaire	Residence principale	Integral	25000-50000	22
11	100611	79		CRS	DE VERDUN	1283	1100	138	Appartement	Proprietaire	Residence secondaire	Classique	0-25000	11
12	100612	240		RUE	DE PRE BAILLY	1173	1170	45	Appartement	Locataire	Residence principale	Classique	0-25000	16
13	100613	3		RUE	TURENNE	1033	1200	83	Appartement	Locataire	Residence principale	Classique	0-25000	14
14	100614	44		ALL	DU SQUARE DE LAUSANNE	1143	1220	88	Appartement	Locataire	Residence principale	Integral	25000-50000	34
15	100615	59		RUE	ALEXANDRE BERARD	1004	1500	165	Appartement	Locataire	Residence principale	Classique	25000-50000	24



	Code dep	reg code	reg nom	aca nom	dep nom	com nom maj court	dep code	dep nom
1	1001	84	Auvergne-Rhône-Alpes	Lyon	Ain	L ABERGEMENT CLEMENCIAT	1	Ain (01)
2	1002	84	Auvergne-Rhône-Alpes	Lyon	Ain	L ABERGEMENT DE VAREY	1	Ain (01)
3	1003	84	Auvergne-Rhône-Alpes	Lyon	Ain	AMAREINS	1	Ain (01)
4	1004	84	Auvergne-Rhône-Alpes	Lyon	Ain	AMBERIEU EN BUGEY	1	Ain (01)
5	1005	84	Auvergne-Rhône-Alpes	Lyon	Ain	AMBERIEUX EN DOMBES	1	Ain (01)
6	1006	84	Auvergne-Rhône-Alpes	Lyon	Ain	AMBLEON	1	Ain (01)
7	1007	84	Auvergne-Rhône-Alpes	Lyon	Ain	AMBRONAY	1	Ain (01)
8	1008	84	Auvergne-Rhône-Alpes	Lyon	Ain	AMBUTRIX	1	Ain (01)
9	1009	84	Auvergne-Rhône-Alpes	Lyon	Ain	ANDERT ET CONDON	1	Ain (01)
10	1010	84	Auvergne-Rhône-Alpes	Lyon	Ain	ANGLEFORT	1	Ain (01)
11	1011	84	Auvergne-Rhône-Alpes	Lyon	Ain	APREMONT	1	Ain (01)
12	1012	84	Auvergne-Rhône-Alpes	Lyon	Ain	ARANC	1	Ain (01)
13	1013	84	Auvergne-Rhône-Alpes	Lyon	Ain	ARANDAS	1	Ain (01)
14	1014	84	Auvergne-Rhône-Alpes	Lyon	Ain	ARBENT	1	Ain (01)
15	1015	84	Auvergne-Rhône-Alpes	Lyon	Ain	ARBIGNIEU	1	Ain (01)
16	1016	84	Auvergne-Rhône-Alpes	Lyon	Ain	ARBIGNY	1	Ain (01)

Note : J'ai reçu un message d'erreur lors du chargement des données car des lignes de la table Contrat ont des Code_dep_code_commune qui n'ont pas de correspondance dans la table Region. Le SGBD renvoie donc une erreur et cela bloque l'import. Pour y remédier, il faut cocher « Ignorer les erreurs », les lignes concernées sont alors remontées dans la Barre d'état en bas.