

RAPPORT PROJET - OLAP

Rapport de projet BD - OLAP

Papa Farba Ndour
Babacar Ndour
Fatim Toure

Github :

<https://github.com/farba07/bdd-warehouse>

Source des données :

<https://www.kaggle.com/etiennelq/french-employment-by-town>

<https://www.data.gouv.fr/s/resources/data-insee-sur-les-communes/20141212-105948/MDB-INSEE-V2.xls>

Plan du rapport

Introduction	2
1 Présentation des données	2
1.1 Données sur la structure de la population	2
1.2 Données sur les salaires	2
1.3 Données sur les zones géographiques	3
1.4 Des données sur les établissements ou entreprises	4
1.5 Données sur le chômage	4
1.6 Donnée sur les dates	4
2 Modèle de données	4
3 Création sur Oracle	9
4 Transformation des données	14
5 Intégration des données	14
6 Requêtes	22
6.1 Scripts	22
6.2 Résultats	26
Conclusion	33

Introduction

Ce travail entre dans le cadre du projet de base données évoluée.

L'objectif de ce projet est d'intégrer des données dans un entrepôt de données. Notre sujet d'analyse consiste en une étude du marché de travail en France, et nous avons choisi six jeux de données liés aux facteurs suivants : les salaires, la population, le taux de chômage, les zones géographiques et les entreprises.

1 Présentation des données

L'entrepôt est construit à partir de six jeux de données enregistrés dont deux sont au format xls (excel) et quatre sont au format csv.

1.1 Données sur la structure de la population

Ces données renseignent sur la structure de la population française regroupée par commune et sont contenues dans un fichier xls nommé "**population.xls**".

Il fournit des informations comme le nombre de femmes, d'hommes, d'étudiants etc.. au sein de chaque commune. Il est composé de 17 colonnes dont les plus importantes sont :

- l'identifiant de la commune (Population_id) : qui identifie chaque commune de manière unique et sera utilisée comme clé primaire.
- le nombre de personne de la commune (population).
- le nombre de femmes et le nombre d'hommes (Nb_Femmes et Nb_Hommes) : qui renseigne sur le nombre de femmes et d'hommes de la commune.
- le nombre de mineurs, majeurs, d'étudiants de la commune (Nb_mineurs, Nb_majeurs, Nb_etudiants).

1.2 Données sur les salaires

Ces données regroupent les salaires horaires net moyen pour chaque catégorie de population (les cadres, l'employé, les jeunes etc.. Les informations sont contenues dans un fichier csv nommé "**salaire.csv**".

Ce fichier est composé d'un identifiant pour chaque enregistrements et des colonnes suivantes :

- SNHM :salaire net moyen
- SNHMC :salaire horaire net moyen pour le cadre
- SNHMCi :salaire horaire net moyen pour le cadre intermédiaire

- SNHME : salaire horaire net moyen pour l'employé
- SNHMO : salaire horaire net moyen pour le travailleur
- SNHMF : salaire horaire net moyen des femmes
- SNHMO : salaire horaire net moyen pour le travailleur
- SNHMF : salaire horaire net moyen des femmes
- SNHMF C : salaire horaire net moyen pour un cadre féminin,
- SNHMF P : salaire horaire net moyen pour le middle manager féminin,
- SNHMF E : salaire horaire net moyen pour une employée féminine,
- SNHMF O : salaire horaire net moyen pour une travailleuse,
- SNHMH : salaire horaire net moyen pour l'homme,
- SNHMH C : salaire horaire net moyen pour un cadre masculin,
- SNHMH P : salaire horaire net moyen pour le middle manager masculin,
- SNHMH E : salaire horaire net moyen pour un employé masculin,
- SNHMH O : salaire horaire net moyen pour un travailleur masculin,
- SNHM18 : salaire horaire net moyen pour les 18-25 ans,
- SNHM26 : salaire horaire net moyen pour les 26-50 ans,
- SNHM50 : salaire horaire net moyen pour les > 50 ans,
- SNHMF18 : salaire horaire net moyen pour les femmes de 18 à 25 ans,
- SNHMF26 : salaire horaire net moyen pour les femmes de 26 à 50 ans,
- SNHMF50 : salaire horaire net moyen pour les femmes > 50 ans,
- SNHMH18 : salaire horaire net moyen pour les hommes de 18 à 25 ans,
- SNHMH26 : salaire horaire net moyen pour les hommes de 26 à 50 ans,
- SNHMH50 : salaire horaire net moyen pour les hommes > 50 ans,

1.3 Données sur les zones géographiques

Ce jeux de données nous donne des informations sur la géographie de la France. Les données sont stockées sur un fichier csv nommé "**zones.csv**".

Ce fichier contient un identifiant pour entrée qui correspond au code INSEE de la commune et de 6 autres colonnes qui apportent des précisions sur la position géographique :

- le nom de la commune (NOM_COMMUNE)
- le code postal (CODES_POSTAUX)
- le nom du département (NOM_DEPARTEMENT)
- le nom de la région (NOM_REGION)
- la colonne EU_CIRCO : qui indique le point cardinal de la commune (*Est - Ouest - Nord - Sud*)

1.4 Des données sur les établissements ou entreprises

Ce jeux de données donne des informations sur les entreprises en France regroupées par commune. Il est composé des colonnes suivantes :

- le nombre d'entreprises de la commune (NBRE_ENTREPRISE)
- les catégories d'entreprises en terme de nombre d'employés allant de 1 à 5 employés pour (CATEGORIE_1) à des entreprises ayant des nombres d'employés supérieurs à 500 (CATEGORIE_8).

1.5 Données sur le chômage

Ce jeu de données donne des informations sur le taux de chômage par trimestre et par commune pour l'année 2019. Il est donc composé de 3 colonnes (TAUX_T1, TAUX_T2, TAUX_T3) correspondent respectivement aux trimestres 1,2 et 3. L'identifiant constitue aussi le code postal d'une commune.

Chaque ligne présente les taux de chômage par trimestre d'une localité précise.

1.6 Donnée sur les dates

Cette table contient les informations sur les différentes dates d'enregistrements des informations que contiennent les autres données.

Elle a un identifiant nommé DATE_ID. Elle donne des informations sur les semestres, les trimestres, la date du jour, les jours, les mois et les années.

2 Modèle de données

Les données sont représentées en schéma en étoile. L'entrepot est composé de 6 dimensions reliés par un table des faits regroupant les identifiants de chaque dimensions et deux faits d'analyses : le taux de chomage moyen par zone et le taux d'emploi moyen par zone.

Les dimensions :

— Dimension chômage :

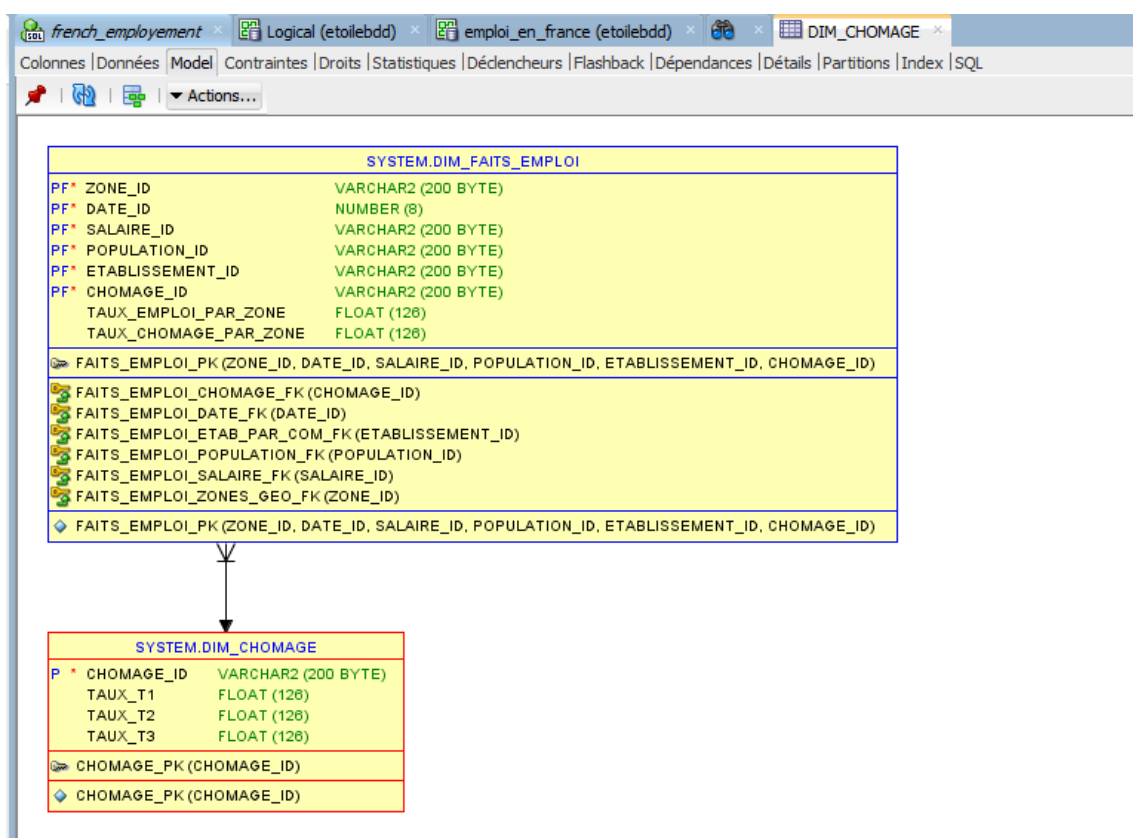


FIGURE 1 – Dimension chômage

— Dimension Date :

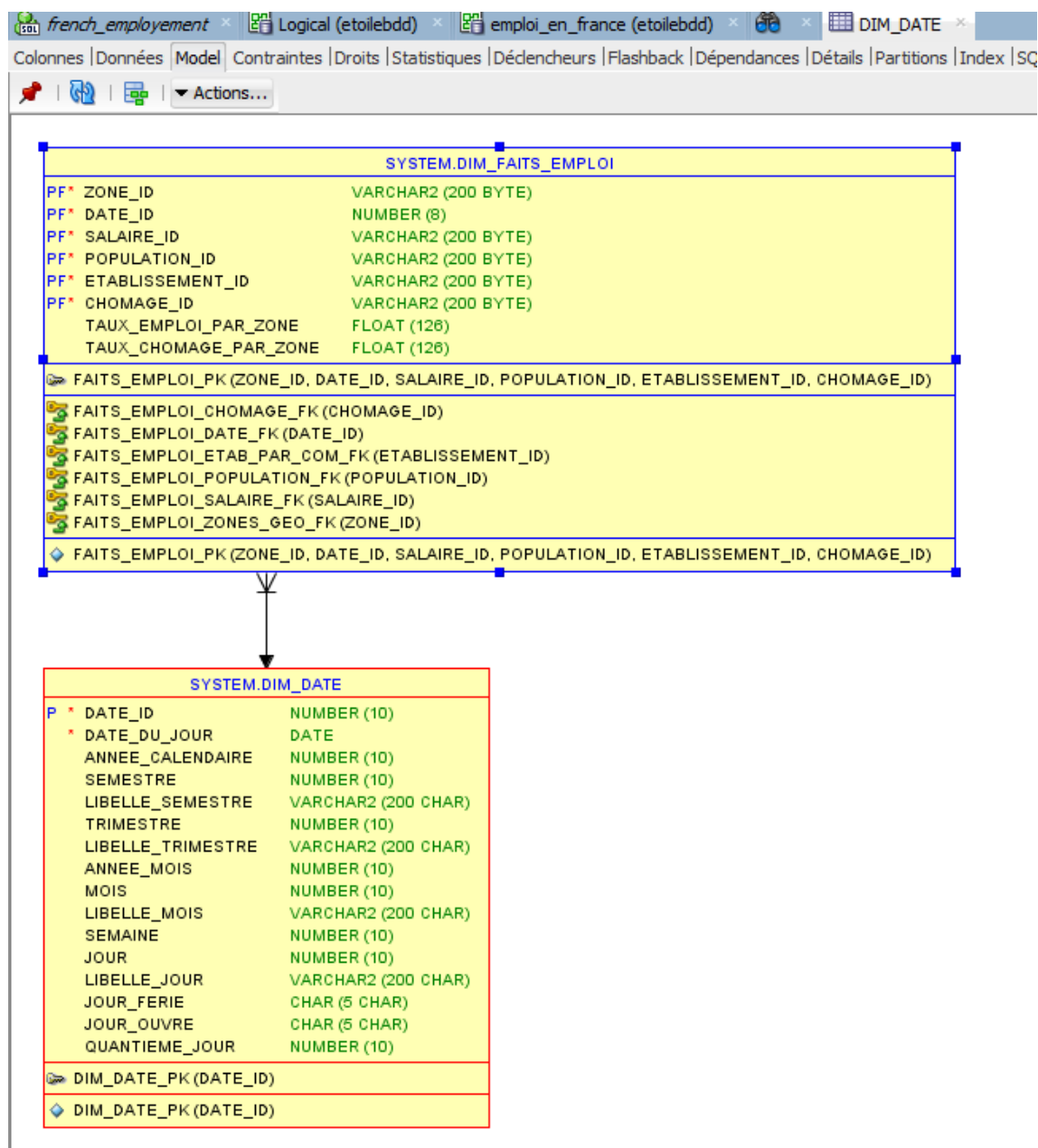


FIGURE 2 – Dimension date

— Dimension Etablissement :

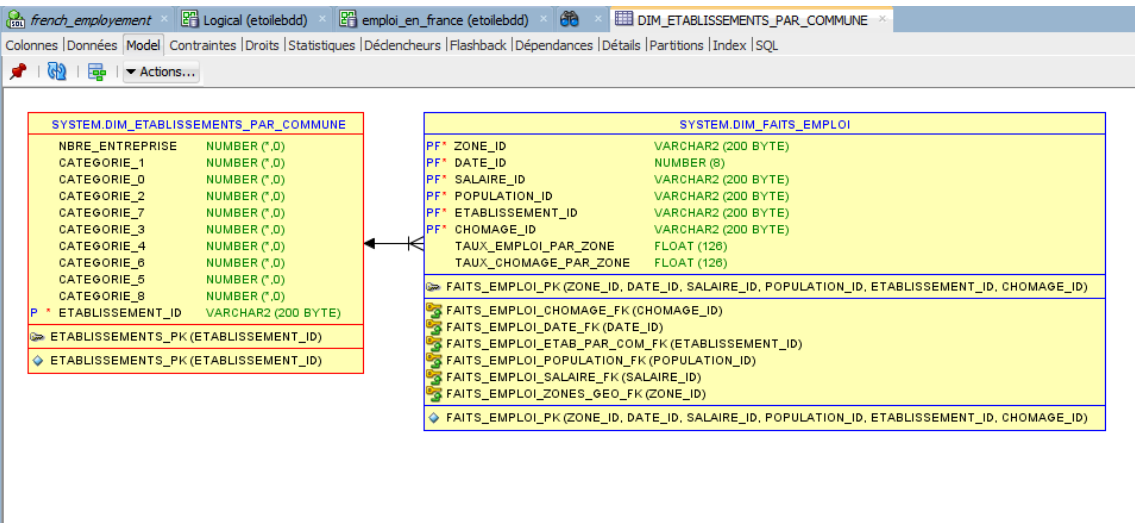


FIGURE 3 – Dimension Etablissement

— Dimension Salaire :

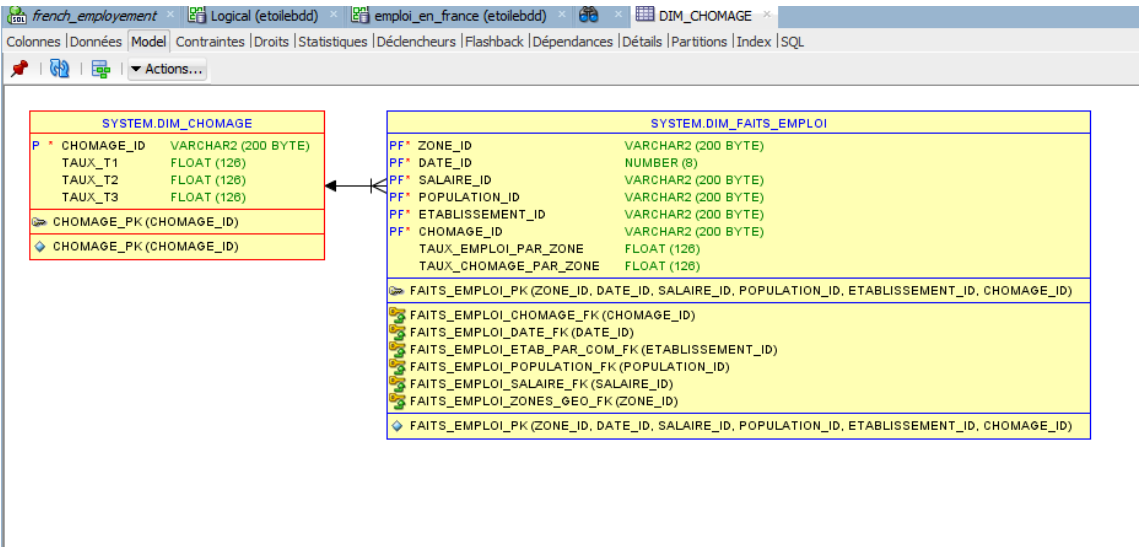


FIGURE 4 – Dimension Salaire

Table des faits : schéma en étoile de la base de données

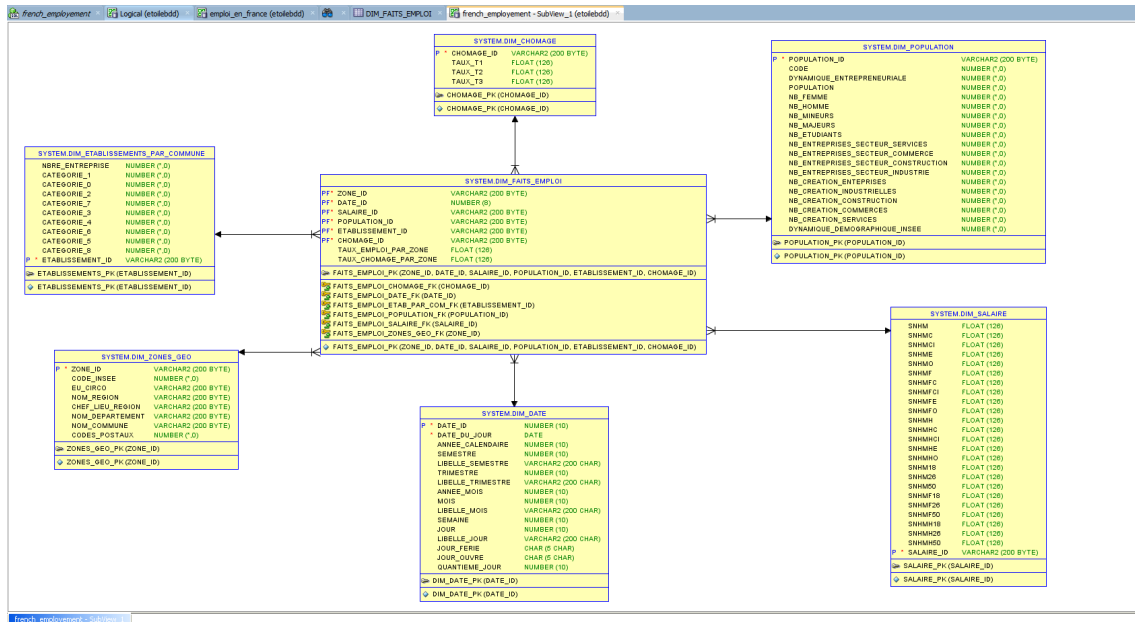


FIGURE 5 – Schéma en étoile

`SYSTEM.FAITS_EMPLOI` est la table des faits. Elle est constituée à partir des six tables qui lui sont rattachées par l'intermédiaire de leurs clés primaires. Les autres attributs de cette table constituent les faits d'analyses.

3 Création sur Oracle

La table des faits et les dimensions ont été créées à l'aide du script sql suivante :

— *Dimension chomage*

```
CREATE TABLE "SYSTEM"."DIM_CHOMAGE"
(
  "CHOMAGE_ID"          VARCHAR2(200 BYTE) ,
  "TAUX_T1"             FLOAT(126) ,
  "TAUX_T2"             FLOAT(126) ,
  "TAUX_T3"             FLOAT(126)
```

```
)  
ALTER TABLE SYSTEM.DIM_CHOMAGE ADD CONSTRAINT chomage_pk  
PRIMARY KEY ( chomage_id );  
  
— Dimension Date  
  
CREATE TABLE "SYSTEM"."DIM_DATE"  
(  
    "DATE_ID"          NUMBER(10,0) ,  
    "DATE_DU_JOUR"     DATE,  
    "ANNEE_CALENDRAIRE" NUMBER(10,0) ,  
    "SEMESTRE"         NUMBER(10,0) ,  
    "LIBELLE_SEMESTRE" VARCHAR2(200 CHAR) ,  
    "TRIMESTRE"        NUMBER(10,0) ,  
    "LIBELLE_TRIMESTRE" VARCHAR2(200 CHAR) ,  
    "ANNEE_MOIS"       NUMBER(10,0) ,  
    "MOIS"             NUMBER(10,0) ,  
    "LIBELLE_MOIS"     VARCHAR2(200 CHAR) ,  
    "SEMAINE"          NUMBER(10,0) ,  
    "JOUR"             NUMBER(10,0) ,  
    "LIBELLE_JOUR"     VARCHAR2(200 CHAR) ,  
    "JOUR_FERIE"       CHAR(5 CHAR) ,  
    "JOUR_OUVRE"       CHAR(5 CHAR) ,  
    "QUANTIEME_JOUR"   NUMBER(10,0)  
)  
ALTER TABLE SYSTEM.DIM_DATE ADD CONSTRAINT date_pk PRIMARY  
KEY ( date_id );  
  
— Dimension Etablissement  
  
CREATE TABLE "SYSTEM"."DIM_ETABLISSEMENTS_PAR_COMMUNE"  
(  
    "NBRE_ENTREPRISE"   NUMBER(*,0) ,  
    "CATEGORIE_1"       NUMBER(*,0) ,  
    "CATEGORIE_0"       NUMBER(*,0) ,  
    "CATEGORIE_2"       NUMBER(*,0) ,  
    "CATEGORIE_7"       NUMBER(*,0) ,  
    "CATEGORIE_3"       NUMBER(*,0) ,
```

```

"CATEGORIE_4"          NUMBER(*,0) ,
"CATEGORIE_6"          NUMBER(*,0) ,
"CATEGORIE_5"          NUMBER(*,0) ,
"CATEGORIE_8"          NUMBER(*,0) ,
"ETABLISSEMENT_ID"     VARCHAR2(200 BYTE)
)
ALTER TABLE SYSTEM.DIM_ETABLISSEMENTS_PAR_COMMUNE ADD
CONSTRAINT etablisements_pk PRIMARY KEY (
    etablisement_id );

```

— *Dimension Population*

```

CREATE TABLE "SYSTEM"."DIM_POPULATION"
(
    "POPULATION_ID"          VARCHAR2(200 BYTE) ,
    "CODE"                   NUMBER(*,0) ,
    "DYNAMIQUE_ENTREPRENEURIALE" NUMBER(*,0) ,
    "POPULATION"             NUMBER(*,0) ,
    "NB_FEMME"               NUMBER(*,0) ,
    "NB_HOMME"               NUMBER(*,0) ,
    "NB_MINEURS"             NUMBER(*,0) ,
    "NB_MAJEURS"             NUMBER(*,0) ,
    "NB_ETUDIANTS"           NUMBER(*,0) ,
    "NB_ENTREPRISES_SECTEUR_SERVICES" NUMBER(*,0) ,
    "NB_ENTREPRISES_SECTEUR_COMMERCE" NUMBER(*,0) ,
    "NB_ENTREPRISES_SECTEUR_CONSTRUCTION" NUMBER(*,0) ,
    "NB_ENTREPRISES_SECTEUR_INDUSTRIE" NUMBER(*,0) ,
    "NB_CREATION_ENTEPRSES"   NUMBER(*,0) ,
    "NB_CREATION_INDUSTRIELLES" NUMBER(*,0) ,
    "NB_CREATION_CONSTRUCTION" NUMBER(*,0) ,
    "NB_CREATION_COMMERCES"   NUMBER(*,0) ,
    "NB_CREATION_SERVICES"    NUMBER(*,0) ,
    "DYNAMIQUE_DEMOGRAPHIQUE_INSEE" NUMBER(*,0)
)
ALTER TABLE SYSTEM.DIM_POPULATION ADD CONSTRAINT
    population_pk PRIMARY KEY ( population_id );

```

— *Dimension Salaire*


```

"NOM_REGION"          VARCHAR2(200 BYTE) ,
"CHEF_LIEU_REGION"    VARCHAR2(200 BYTE) ,
"NOM_DEPARTEMENT"     VARCHAR2(200 BYTE) ,
"NOM_COMMUNE"         VARCHAR2(200 BYTE) ,
"CODES_POSTAUX"       NUMBER(*,0)
)
ALTER TABLE SYSTEM.DIM_ZONES_GEO ADD CONSTRAINT zones_geo_pk
    PRIMARY KEY ( zone_id );

— Table des faits

CREATE TABLE "SYSTEM"."DIM_FAITS_EMPLOI"
(
    "ZONE_ID"          VARCHAR2(200 BYTE) ,
    "DATE_ID"          NUMBER(8,0) ,
    "SALAIRE_ID"        VARCHAR2(200 BYTE) ,
    "POPULATION_ID"     VARCHAR2(200 BYTE) ,
    "ETABLISSEMENT_ID" VARCHAR2(200 BYTE) ,
    "CHOMAGE_ID"        VARCHAR2(200 BYTE) ,
    "TAUX_EMPLOI_PAR_ZONE" FLOAT(126) ,
    "TAUX_CHOMAGE_PAR_ZONE" FLOAT(126)
)
ALTER TABLE faits_emploi
    ADD CONSTRAINT faits_emploi_pk PRIMARY KEY ( code_insee
        , date_id , salaire_id , population_id ,
        etablissement_id );

ALTER TABLE faits_emploi
    ADD CONSTRAINT faits_emploi_chomage_fk FOREIGN KEY (
        chomage_id )
        REFERENCES dim_chomage ( chomage_id );

ALTER TABLE faits_emploi
    ADD CONSTRAINT faits_emploi_date_fk FOREIGN KEY (
        date_id )
        REFERENCES dim_date ( date_id );

ALTER TABLE faits_emploi

```

```
ADD CONSTRAINT faits_emploi_etab_par_com_fk FOREIGN KEY
( etablisement_id )
REFERENCES dim_etablisements_par_commune (
    etablisement_id );

ALTER TABLE faits_emploi
ADD CONSTRAINT faits_emploi_population_fk FOREIGN KEY (
    population_id )
REFERENCES dim_population ( population_id );

ALTER TABLE faits_emploi
ADD CONSTRAINT faits_emploi_salaire_fk FOREIGN KEY (
    salaire_id )
REFERENCES dim_salaire ( salaire_id );

ALTER TABLE faits_emploi
ADD CONSTRAINT faits_emploi_zones_geo_fk FOREIGN KEY (
    zone_id )
REFERENCES dim_zones_geo ( zone_id );
```

4 Transformation des données

Certaines de nos tables ont été modifiées avant de faire l'intégration des données. Des colonnes jugées inutiles ont été supprimées. Des identifiants ont été construits au besoin pour certaines tables grâce à de petits scripts écrits en python. Certains enregistrements redondants ou contenant des valeurs nulles pour la plupart des colonnes ont été également supprimés.

5 Intégration des données

L'intégration des données est faite à l'aide du logiciel de traitement de données Talend. La procédure d'intégration des données d'un fichier vers une table de la base de données consiste à créer d'abord un "job" pour chaque dimension de l'entrepôt et la table des faits. Ensuite, la connexion entre la base de données oracle et le logiciel Talend est établie pour permettre la communication entre les deux logiciels. Les schémas des différentes tables de la base de données sont récupérées depuis Talend. Puis les données sont récupérées à partir du fichier les contenant qui peut être de plusieurs types (csv ou xls). Un fichier est créé pour chaque dimension et est chargé dans talend. Enfin, l'outil tMap de talend nous a permis de mapper les colonnes

du fichier avec celles de la base de données. Un fichier d'entrée est spécifié grâce à l'outil "tFileInputDelimited" ou "tFileInputExcel" en fonction du type de fichier correspondant. En sortie, l'outil "tDBOutput" est utilisé pour représenter la table de la base oracle correspondante.

A la fin de l'exécution, le nombre de lignes récupérées depuis le fichier et insérées dans la table sont marqués.

Ci-dessous la structure du projet sous talend :

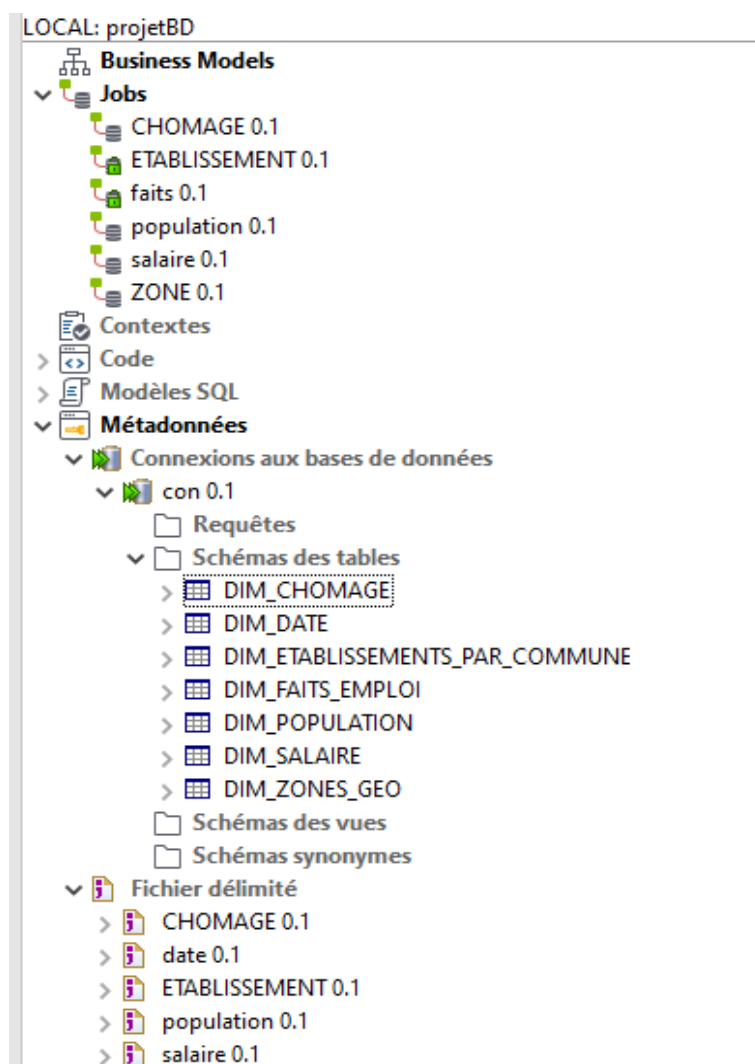


FIGURE 6 – Structure du projet Talend

L'intégration des données pour chaque dimension est illustrée avec les images suivantes :

— Dimension Chômage :

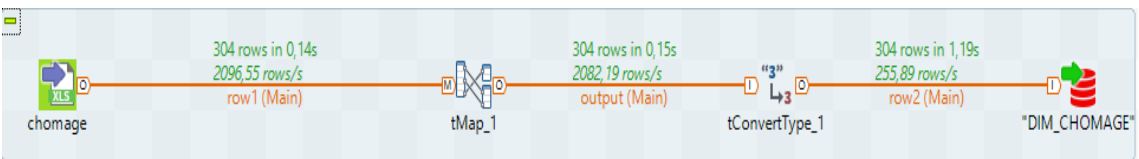


FIGURE 7 – Intégration dimension chômage

row1

Column
CHOMAGE_ID
taux_t1
taux_t2
taux_t3

Find:

Var

Mapping auto

Expression	Column
row1.CHOMAGE_ID	CHOMAGE_ID
row1.taux_t1	TAUX_T1
row1.taux_t2	TAUX_T2
row1.taux_t3	TAUX_T3

row1

Colonne	Clé	Type	N.	Modèle de date...	Length	Precision	Défaut	Comme...
CHOMAGE_ID	<input checked="" type="checkbox"/>	BigDeci...	<input type="checkbox"/>		38	0		
taux_t1	<input type="checkbox"/>	Float	<input checked="" type="checkbox"/>		126	0		
taux_t2	<input type="checkbox"/>	Float	<input checked="" type="checkbox"/>		126	0		
taux_t3	<input type="checkbox"/>	Float	<input checked="" type="checkbox"/>		126	0		

output

Colonne	Clé	Type	N.	Modèle de date...	Length	Precision	Défaut	Comme...
CHOMAGE_ID	<input checked="" type="checkbox"/>	BigDeci...	<input type="checkbox"/>		38	0		
TAUX_T1	<input type="checkbox"/>	Float	<input checked="" type="checkbox"/>		126	0		
TAUX_T2	<input type="checkbox"/>	Float	<input checked="" type="checkbox"/>		126	0		
TAUX_T3	<input type="checkbox"/>	Float	<input checked="" type="checkbox"/>		126	0		

FIGURE 8 – Mapping des colonnes de la dimension chômage

— Dimension Établissement :

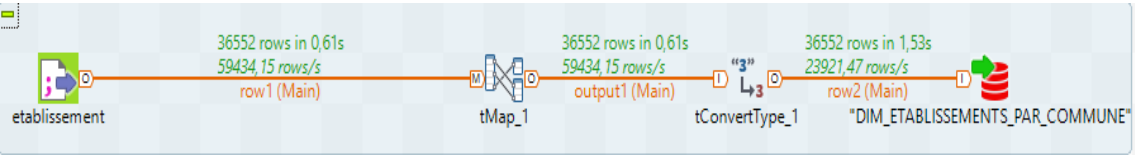


FIGURE 9 – Intégration dimension Établissement

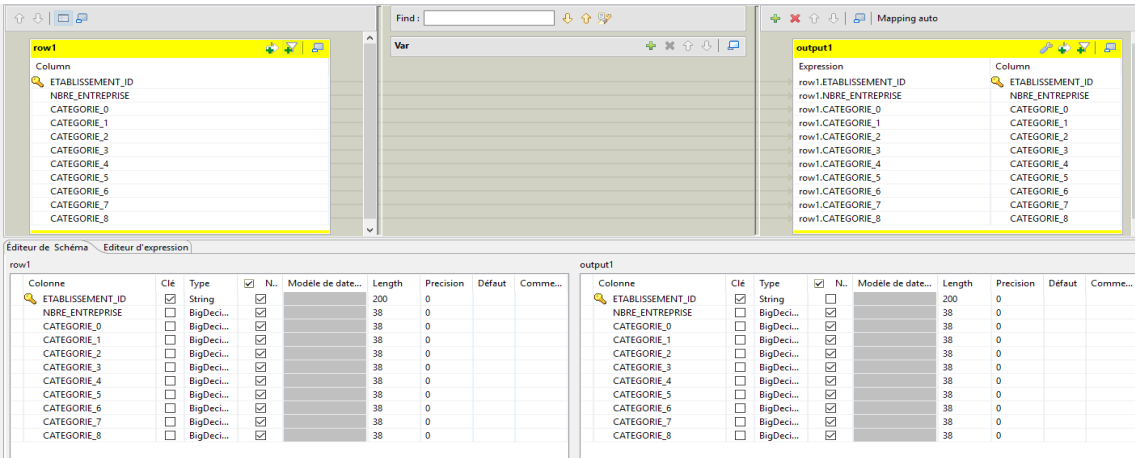


FIGURE 10 – Mapping des colonnes de la dimension établissement

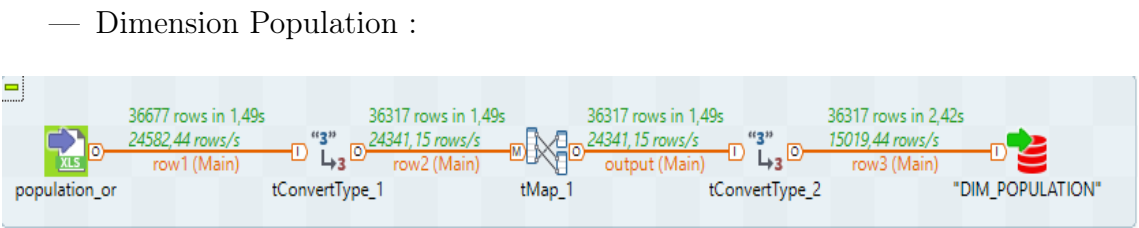


FIGURE 11 – Intégration dimension population

Figure 12 shows the mapping of columns from the 'population' table to the 'DIM_POPULATION' table. The mapping is as follows:

Column	Type	Length	Precision	Default	Commentaire
POPULATION_ID	String	20	0		
CODEGEO	Integer	38	0		
DYNAMIQUE_ENTREPRENEURIALE	Integer	38	0		
POPULATION	Integer	38	0		
NB_FEMME	Integer	38	0		
NB_HOMME	Integer	38	0		
NB_MINEURS	Integer	38	0		
NB_MAJEURS	Integer	38	0		
NB_ETUDIANTS	Integer	38	0		
NB_ENTREPRISES_SECTEUR_SERVICES	Integer	38	0		
NB_ENTREPRISES_SECTEUR_COMMERCE	Integer	38	0		
NB_ENTREPRISES_SECTEUR_CONSTRUCTION	Integer	38	0		
NB_ENTREPRISES_SECTEUR_INDUSTRIE	Integer	38	0		
NB_CREATION_ENTREPRISES	Integer	38	0		
NB_CREATION_INDUSTRIELLES	Integer	38	0		
NB_CREATION_CONSTRUCTION	Integer	38	0		
NB_CREATION_COMMERCE	Integer	38	0		
NB_CREATION_SERVICES	Integer	38	0		
DYNAMIQUE_DEMOGRAPHIQUE_INL	Integer	38	0		

FIGURE 12 – Mapping des colonnes de la table population

— Dimension Salaire :

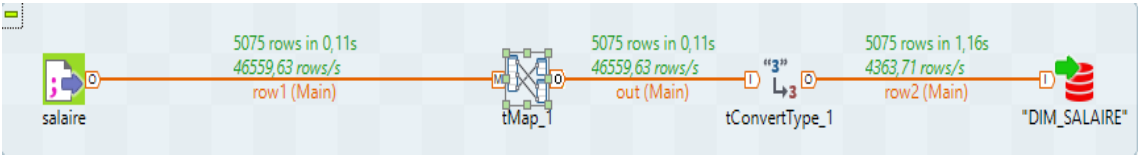


FIGURE 13 – Intégration dimension salaire

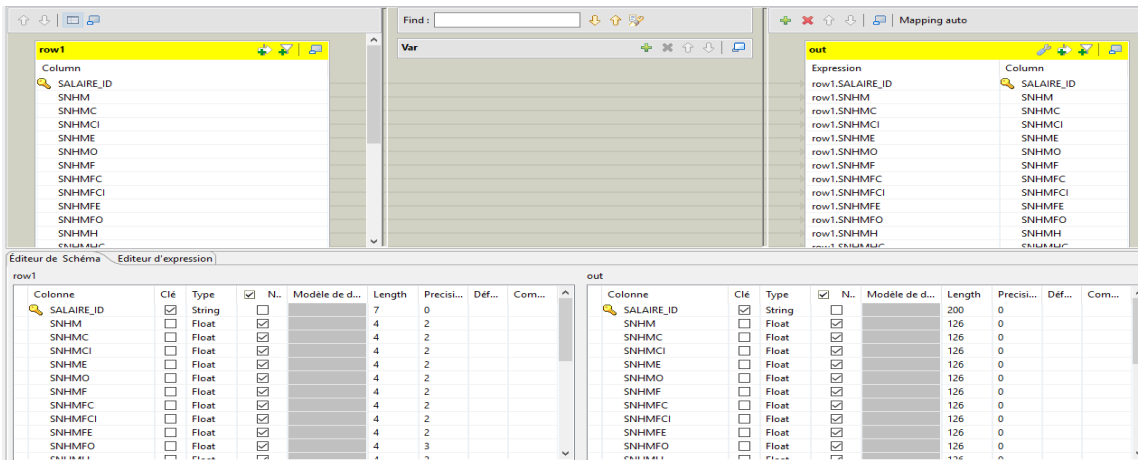


FIGURE 14 – Mapping des colonnes de la dimension salaire

— Dimension Zones :

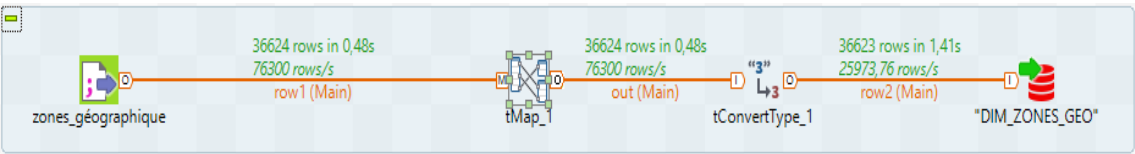


FIGURE 15 – Intégration dimension zones

The screenshot shows the 'Éditeur de Schéma' (Schema Editor) interface. It displays two tables: 'row1' on the left and 'out' on the right. The 'row1' table lists columns: ZONE_ID, EU_CIRCO, NOM_REGION, CHEF_LIEU_REGION, NOM_DEPARTEMENT, NOM_COMMUNE, CODES_POSTAUX, and CODE_INSEE. The 'out' table lists the same columns. The 'Mapping auto' tab is selected, showing the mapping between the two tables. Below the tables, there is a detailed view of the column properties, including 'Colonne', 'Clé', 'Type', 'N.', 'Modèle de date...', 'Length', 'Precision', 'Défaut', and 'Comme...'. The 'row1' table has columns with lengths ranging from 3 to 200. The 'out' table has columns with lengths ranging from 38 to 200.

Colonne	Clé	Type	N.	Modèle de date...	Length	Precision	Défaut	Comme...
ZONE_ID	<input checked="" type="checkbox"/>	String	<input type="checkbox"/>		5	0		
EU_CIRCO	<input type="checkbox"/>	String	<input checked="" type="checkbox"/>		7	0		
NOM_REGION	<input type="checkbox"/>	String	<input checked="" type="checkbox"/>		12	0		
CHEF_LIEU_REGION	<input type="checkbox"/>	String	<input checked="" type="checkbox"/>		4	0		
NOM_DEPARTEMENT	<input type="checkbox"/>	String	<input checked="" type="checkbox"/>		3	0		
NOM_COMMUNE	<input type="checkbox"/>	String	<input checked="" type="checkbox"/>		21	0		
CODES_POSTAUX	<input type="checkbox"/>	Integer	<input checked="" type="checkbox"/>		4	0		
CODE_INSEE	<input type="checkbox"/>	String	<input checked="" type="checkbox"/>		200	0		

Colonne	Clé	Type	N.	Modèle de date...	Length	Precision	Défaut	Comme...
ZONE_ID	<input checked="" type="checkbox"/>	String	<input type="checkbox"/>		200	0		
EU_CIRCO	<input type="checkbox"/>	String	<input checked="" type="checkbox"/>		200	0		
NOM_REGION	<input type="checkbox"/>	String	<input checked="" type="checkbox"/>		200	0		
CHEF_LIEU_REGION	<input type="checkbox"/>	String	<input checked="" type="checkbox"/>		200	0		
NOM_DEPARTEMENT	<input type="checkbox"/>	String	<input checked="" type="checkbox"/>		200	0		
NOM_COMMUNE	<input type="checkbox"/>	String	<input checked="" type="checkbox"/>		200	0		
CODES_POSTAUX	<input type="checkbox"/>	Integer	<input checked="" type="checkbox"/>		38	0		
CODE_INSEE	<input type="checkbox"/>	String	<input checked="" type="checkbox"/>		38	0		

FIGURE 16 – Mapping des colonnes de la dimension zone

— Table des faits :

La table des faits regroupe l'ensemble des jobs réalisés précédemment.

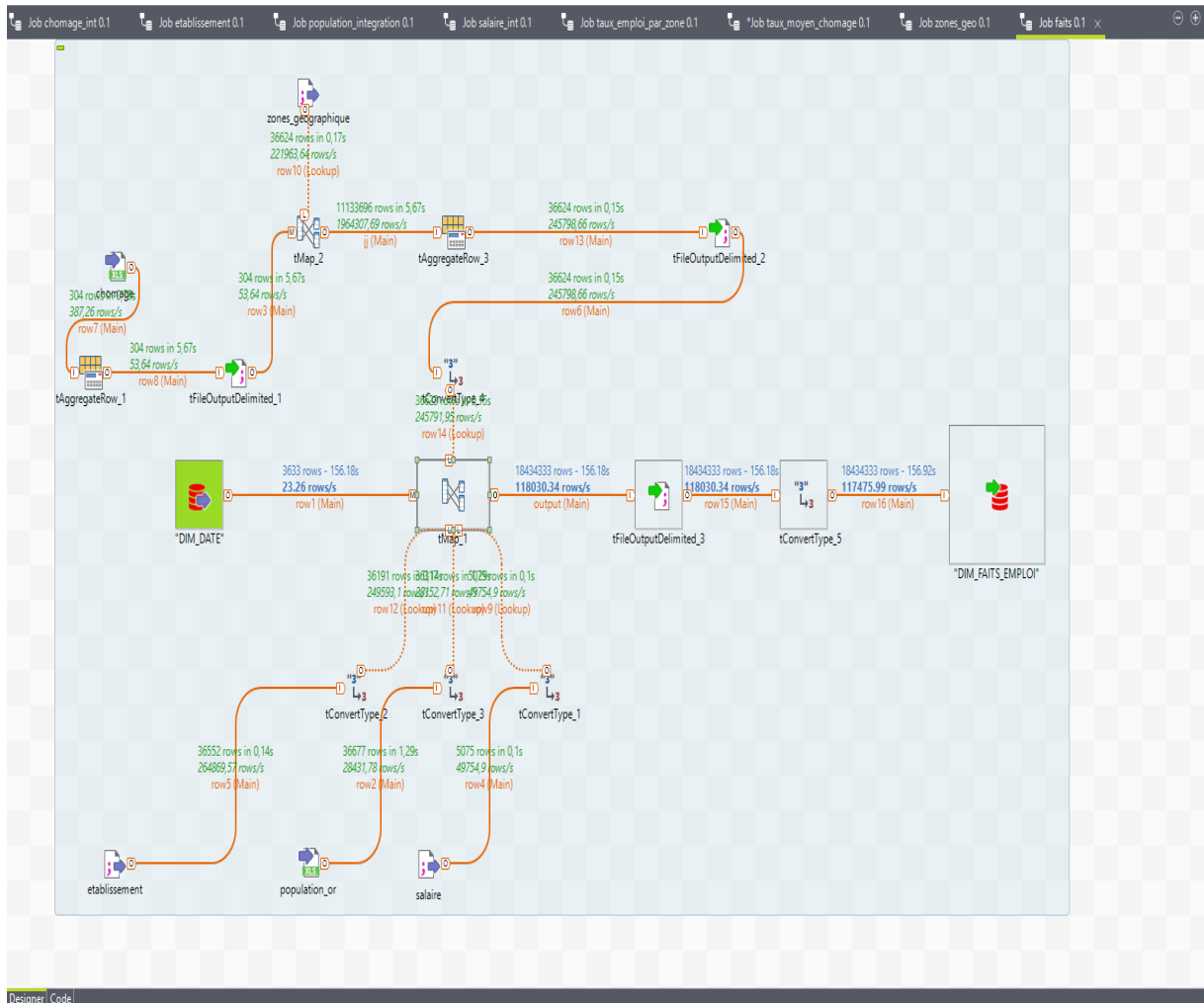


FIGURE 17 – Intégration de la table des faits

6 Requêtes

6.1 Scripts

Les requetes sont contenues dans le script suivant :

```
— 1 la requete sql qui regroupe le nombre d'
   entreprise et le semestre
— avec les sous totaux: nbre_entreprise,
   nbre_entreprise et semestre,
— plus un total general
```

```
SELECT nbre_entreprise, semestre
FROM dim_faits_emploi NATURAL JOIN
      dim_etablisements_par_commune, dim_faits_emploi
      NATURAL JOIN dim_date
GROUP BY ROLLUP(nbre_entreprise, semestre);
```

```
— 2 la requete sql qui regroupe
— le nom de chaque commune et le salaire net moyen
   par heure
— avec les sous totaux: nom de commune, snhm plus
   le total general
```

```
SELECT nom_commune, snhm
FROM dim_faits_emploi NATURAL JOIN dim_salaire,
      dim_faits_emploi NATURAL JOIN dim_zones_geo
GROUP BY CUBE(nom_commune, snhm);
```

```
— 3 la requete sql qui regroupe le nom departement
   et
— le nombre d'entreprises dans le secteur
   industriel
```

— avec les sous totaux: nom_departement,
 nb_entreprises_secteur_industrie ,
 — nom_departement et
 nb_entreprises_secteur_industrie , plus le total
 general ,
 — en ajoutant 1 s'il s'agit d'un agregat ou d'un
 sous total et 0 dans le cas contraire

```
SELECT nom_departement ,
       nb_entreprises_secteur_industrie ,
GROUPING (nom_departement) AS nom_dpt,
GROUPING (nb_entreprises_secteur_industrie) AS
       nb_industrie
FROM dim_faits_emploi NATURAL JOIN dim_zones_geo ,
     dim_faits_emploi NATURAL JOIN dim_population
GROUP BY CUBE (nom_departement ,
              nb_entreprises_secteur_industrie)
```

— 4 Le top des dix zones qui ont le meilleur taux
 de chomage moyen

```
SELECT * FROM( SELECT taux_chomage_moyen, zone_id
                FROM dim_faits_emploi)
WHERE ROWNUM<=10
```

— 5 la requete sql qui cumule le salaire net
 mensuel
 — par heure que gagne un homme par
 nb_entreprises_secteur_commerce
 — et nb_entreprises_secteur_construction avec les
 sous totaux :
 — nb_entreprises_secteur_commerce ,
 nb_entreprises_secteur_construction
 — et le total general, plus le total general

```
SELECT SUM(snhmh), nb_entreprises_secteur_commerce,
       nb_entreprises_construction
FROM dim_faits_emploi NATURAL JOIN dim_salaire,
     dim_faits_emploi NATURAL JOIN dim_population
GROUP BY GROUPING SETS ((
    nb_entreprises_secteur_commerce,
    nb_entreprises_secteur_construction), (
    nb_entreprises_secteur_commerce), ())
```

— 6 le taux de chômage moyen par semestre

```
SELECT semestre, AVG(taux_chomage_moyen)
FROM dim_faits_emploi NATURAL JOIN dim_date
GROUP BY semestre
```

*— 7 La requete sql qui partitionne les zones dont
le nombre
d'entreprises est superieur 39000
en dix groupes*

```
SELECT nbre_entreprise, zone_id, NTILE(10) OVER (
    ORDER BY nbre_entreprise DESC)
FROM dim_faits_emploi NATURAL JOIN
     dim_etablisements_par_commune
WHERE nbre_entreprise > 39000
```

*— 8 la requete sql qui affiche le nombre d'
entreprises
creees dans chaque region et son rang*

```
SELECT nom_region, SUM(nbre_entreprise), RANK() OVER
    (ORDER BY SUM(nbre_entreprise) DESC) rank
FROM dim_faits_emploi NATURAL JOIN
     dim_etablisements_par_commune, dim_faits_emploi
```

```
NATURAL JOIN dim_zones_geo  
GROUP BY nom_region
```

```
— 9 le meilleur taux d'emploi pour mois et par  
   annee
```

```
SELECT MAX(taux_emploi_moyen) AS max,  
       annee_calendaire , mois  
FROM dim_faits_emploi NATURAL JOIN dim_date  
GROUP BY annee_calendaire , mois
```

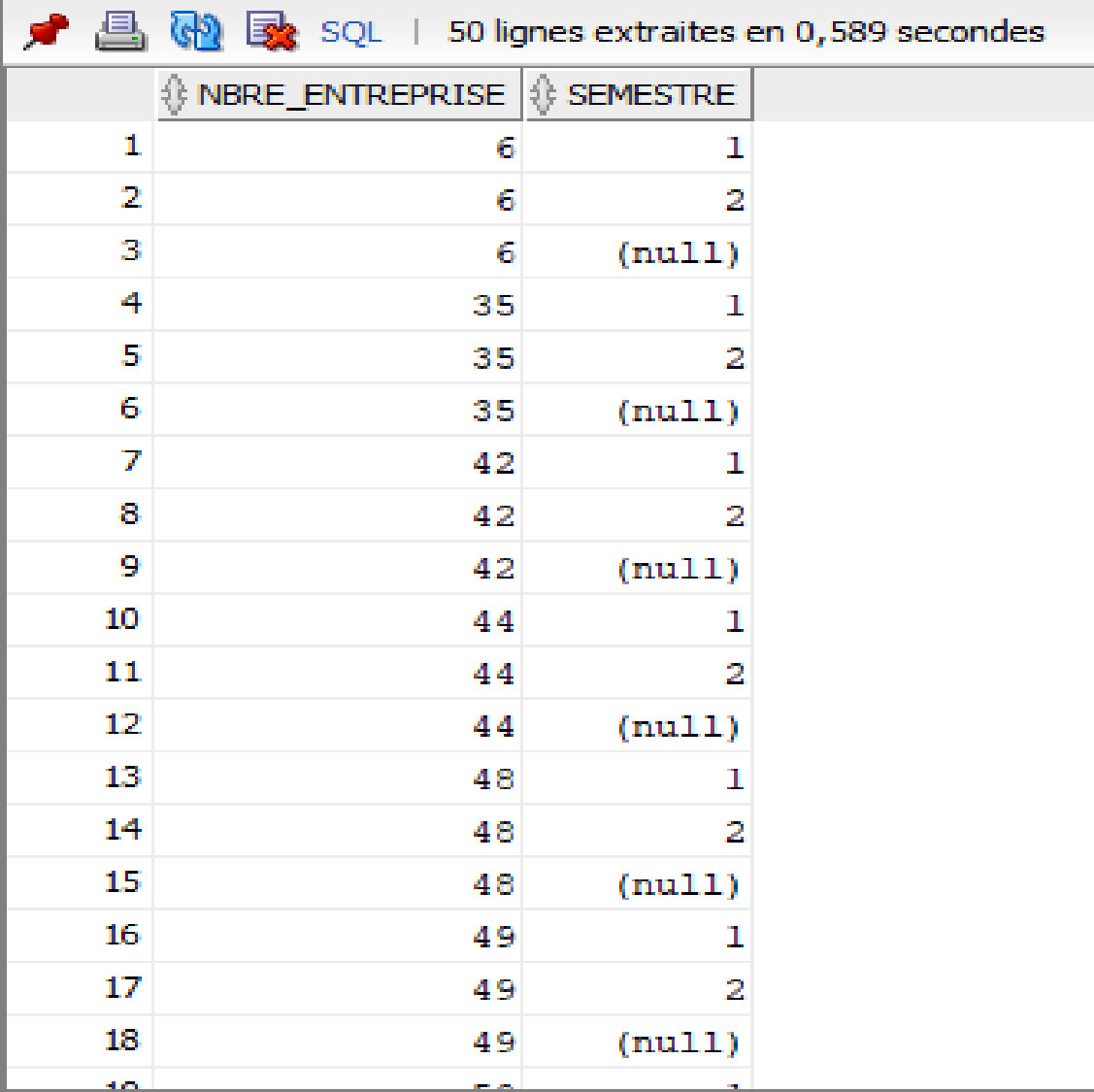
```
— 10 la region qui a le moins d'entreprise par annee
```

```
SELECT nom_region , MIN(nbre_entreprise) ,  
       annee_calendaire  
FROM dim_faits_emploi NATURAL JOIN  
     dim_etablisements_par_commune , dim_faits_emploi  
     NATURAL JOIN dim_date , dim_faits_emploi NATURAL  
     JOIN dim_zones_geo  
GROUP BY nom_region , annee_calendaire
```

6.2 Résultats

Certaines requêtes affichent plusieurs lignes, par conséquent une partie du résultat est montrée en image.

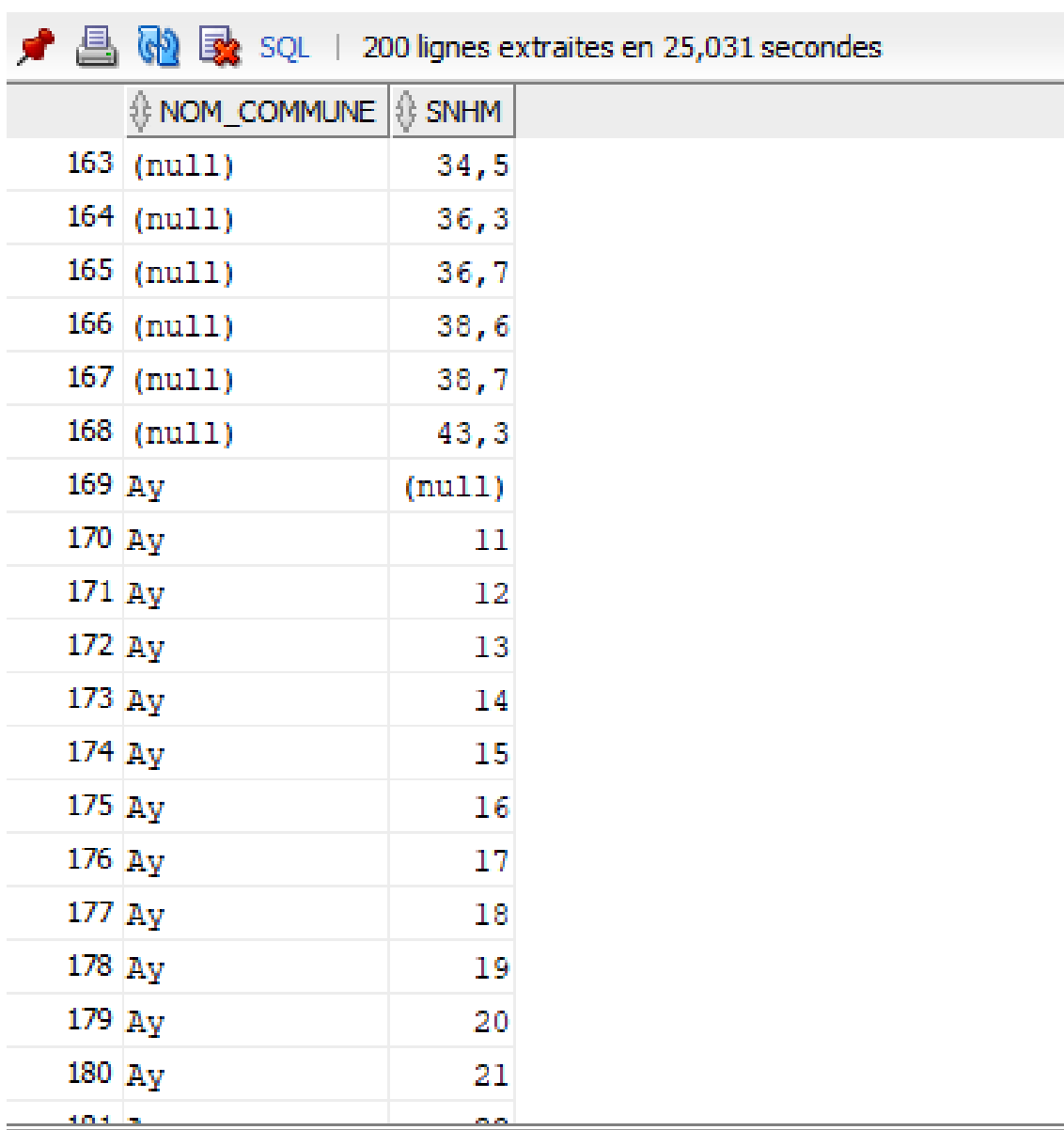
Résultat 1 : Nombre d'entreprises par semestre avec les sous totaux sur le nombre d'entreprises.



	NBRE_ENTREPRISE	SEMESTRE
1	6	1
2	6	2
3	6	(null)
4	35	1
5	35	2
6	35	(null)
7	42	1
8	42	2
9	42	(null)
10	44	1
11	44	2
12	44	(null)
13	48	1
14	48	2
15	48	(null)
16	49	1
17	49	2
18	49	(null)

FIGURE 18 – Résultat requête 1

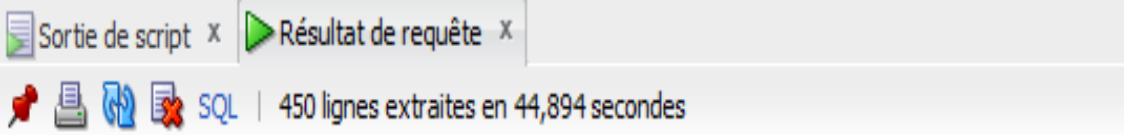
Résultat 2 : Nom de chaque commune et le salaire net moyen par heure avec les sous totaux sur le nom de commune, le snhm et le total général



	NOM_COMMUNE	SNHM
163	(null)	34,5
164	(null)	36,3
165	(null)	36,7
166	(null)	38,6
167	(null)	38,7
168	(null)	43,3
169	Ay	(null)
170	Ay	11
171	Ay	12
172	Ay	13
173	Ay	14
174	Ay	15
175	Ay	16
176	Ay	17
177	Ay	18
178	Ay	19
179	Ay	20
180	Ay	21
181	Ay	22

FIGURE 19 – Résultat requête 2

Résultat 3 : Nombre d'entreprises dans le secteur industriel avec les sous totaux sur le nom du département, le nombre d'entreprises par secteur industriel, plus le total général, en ajoutant 1 s'il s'agit d'un agrégat ou d'un sous total et 0 dans le cas contraire.



	NOM_DEPARTEMENT	NB_ENTREPRISES_SECTEUR_INDUSTRIE	NOM_DPT	NB_INDUSTRIE
388	(null)	3612	1	0
389	(null)	4687	1	0
390	(null)	6818	1	0
391	(null)	36275	1	0
392	Ain	(null)	0	1
393	Ain	1	0	0
394	Ain	2	0	0
395	Ain	3	0	0
396	Ain	4	0	0
397	Ain	5	0	0
398	Ain	6	0	0
399	Ain	7	0	0
400	Ain	8	0	0
401	Ain	9	0	0
402	Ain	10	0	0

FIGURE 20 – Résultat requête 3

Résultat 5 : Cumul du salaire net mensuel par heure que gagne un homme par secteur et nombre d'entreprises par secteur de construction avec les sous totaux sur le nombre d'entreprises par secteur et le total général.

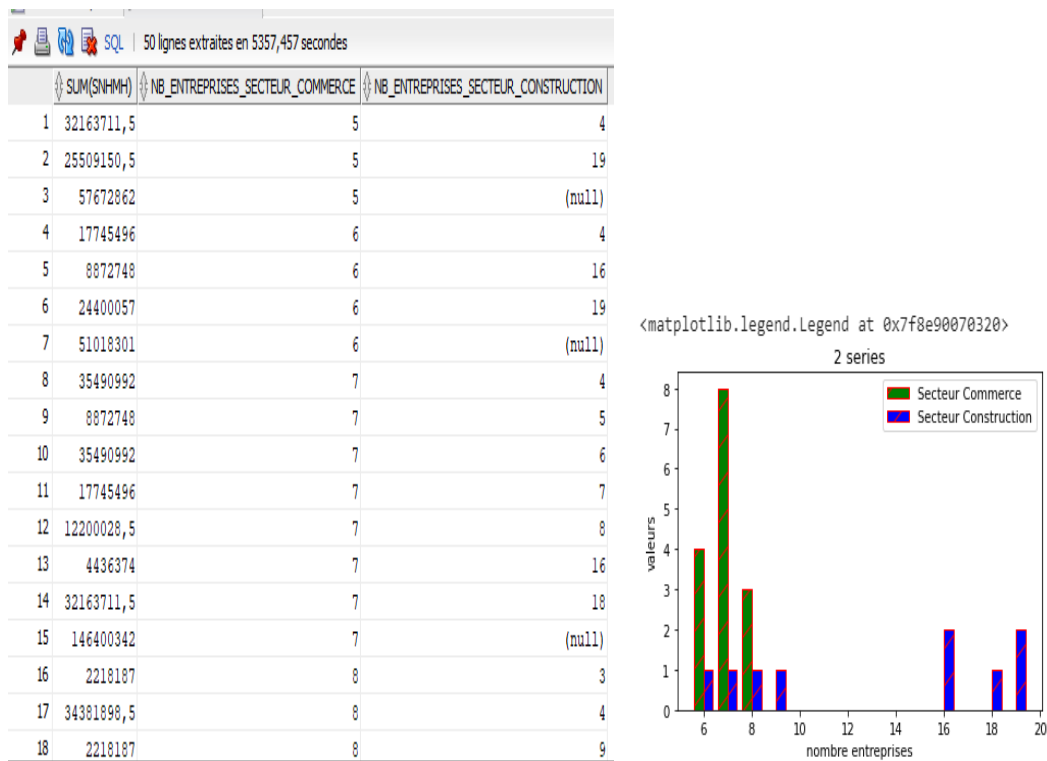


FIGURE 22 – Résultat requête 5

Résultat 6 : le taux de chômage moyen par semestre pour l'année 2019

SEMESTRE	TAUX_CHOMAGE_PAR_ZONE
1	9,866666
2	6,666665

FIGURE 23 – Résultat requête 6

Résultat 7 : Les zones dont le nombre d'entreprises est supérieur à 39000 en dix groupes.

	↕ NBRE_ENTREPRISE	↕ ZONE_ID	↕ NTILE(10)OVER(ORDERBYNBRE_ENTREPRISEDESC)	
1	427385	z31791		1
2	427385	z31791		1
3	427385	z31791		1
4	427385	z31791		2
5	427385	z31791		2
6	427385	z31791		2
7	427385	z31791		3
8	427385	z31791		3
9	427385	z31791		3
10	427385	z31791		4
11	427385	z31791		4
12	68332	z5764		4
13	68332	z5764		5
14	68332	z5764		5
15	68332	z5764		5
16	68332	z5764		6
17	49756	z29406		6
18	49756	z29406		6
19	49756	z29406		7
20	49756	z29406		7
21	49756	z29406		7
22	49756	z29406		8

FIGURE 24 – Résultat requête 7

Résultat 8 : Nombre d'entreprises créées dans chaque région et son rang

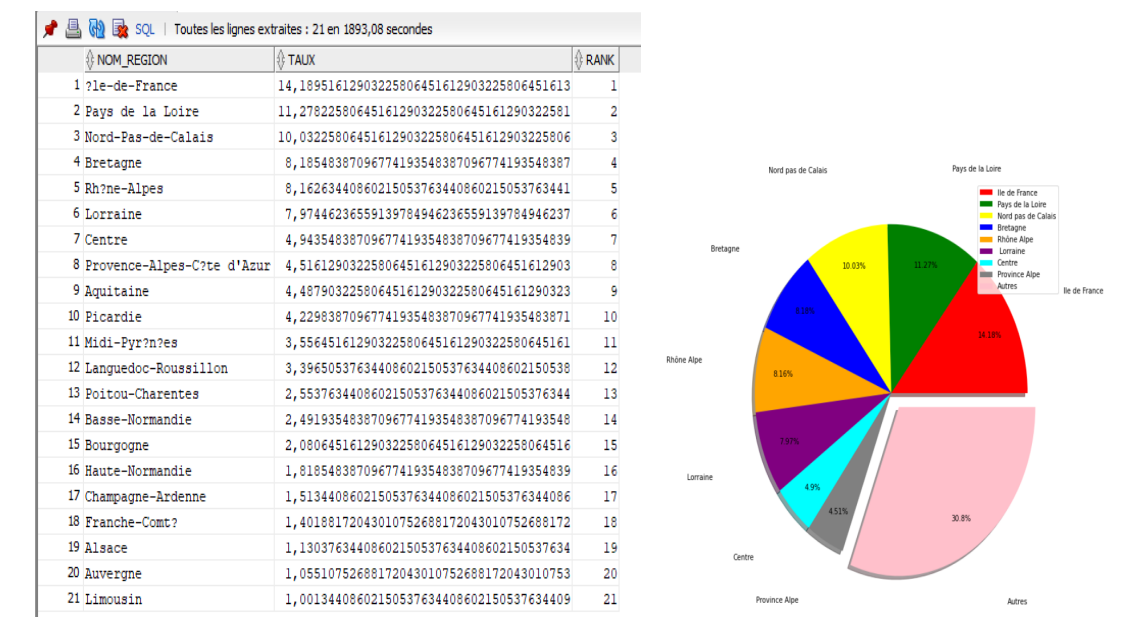


FIGURE 25 – Résultat requête 8 : Répartition des établissements par régions

Résultat 9 : Le meilleur taux d'emploi par mois pour l'année 2019

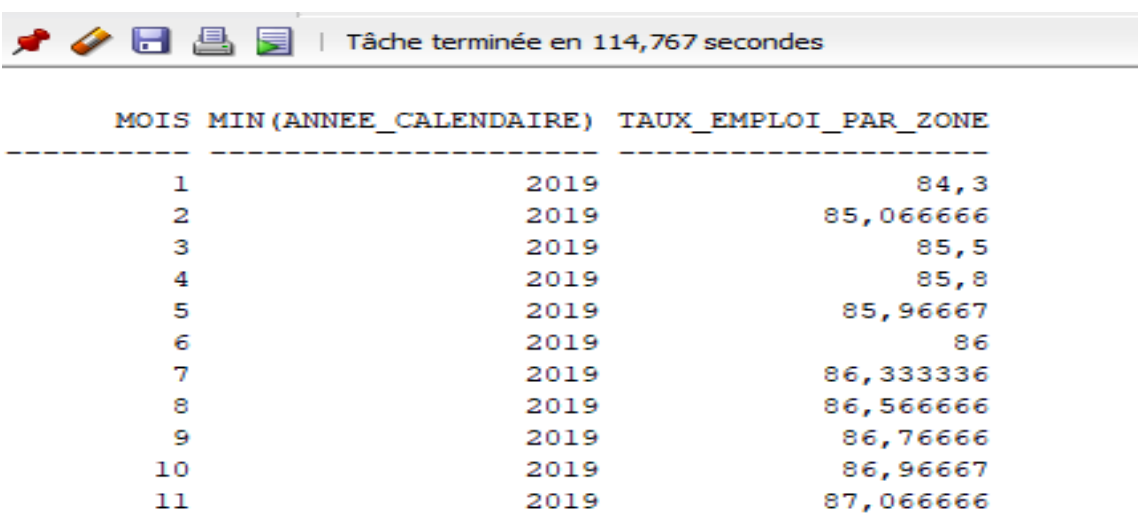
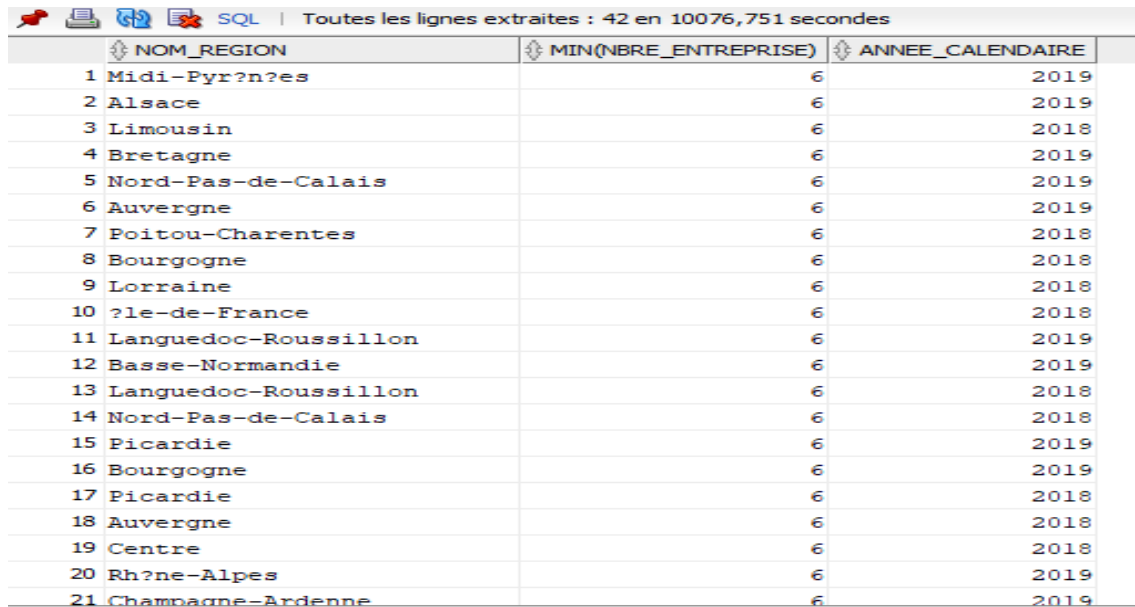


FIGURE 26 – Résultat requête 9

Résultat 10 : la région qui a le moins d'entreprises par année



	NOM_REGION	MIN(NBRE_ENTREPRISE)	ANNEE_CALENDRAIRE
1	Midi-Pyrénées	6	2019
2	Alsace	6	2019
3	Limousin	6	2018
4	Bretagne	6	2019
5	Nord-Pas-de-Calais	6	2019
6	Auvergne	6	2019
7	Poitou-Charentes	6	2018
8	Bourgogne	6	2018
9	Lorraine	6	2018
10	Île-de-France	6	2018
11	Languedoc-Roussillon	6	2019
12	Basse-Normandie	6	2019
13	Languedoc-Roussillon	6	2018
14	Nord-Pas-de-Calais	6	2018
15	Picardie	6	2019
16	Bourgogne	6	2019
17	Picardie	6	2018
18	Auvergne	6	2018
19	Centre	6	2018
20	Rhône-Alpes	6	2019
21	Champagne-Ardenne	6	2019

FIGURE 27 – Résultat requête 10

Conclusion

Ce projet fut une belle expérience. Il nous a permis de mettre en pratique les requêtes OLAP, de bien comprendre notre cours des bases de données évoluées et de découvrir l'outil Talend servant d'intégration avec les logiciels de SGBD comme oracle et mysql. Notre problème majeur était lié à la préparation des données, certaines colonnes contenaient des données de types différent et une quantité énorme de données dans la table des faits avec notamment plus de 9 millions de lignes.