

# RAPPORT DE TRAVAUX PRATIQUE

## Informatique décisionnelle et big data

---

**Conception et chargement d'un Entrepôt de Données :**  
**Schéma en étoile et ETL**

---

**Enseignant :**

Stéphane JEAN

**Réalisé par :**

Nguyen Quynh Nhu HO

Cursus : 1<sup>ère</sup> année de Master Informatique

Parcours Logiciel et Big Data

Année universitaire 2023/2024

# Table des matières

<b>1. Introduction et Description générale</b>	3
1.1 Introduction	3
1.2 Description d'architecture du système décisionnel	3
<b>2. Modélisation de l'entrepôt de données</b>	4
2.1 Livrable 1 - La capture d'écran du modèle conceptuel de l'entrepôt de données dans le formalisme Data Fact Model.	4
2.2 Livrable 2 - La capture d'écran des tables relationnelles de l'entrepôt de données	5
2.3 Livrable 3 - Le script SQL du schéma en étoile schéma_étoile.sql	5
<b>3. Peuplement de l'entrepôt</b>	6
3.1 La phrase de ELT en Java	6
3.1.1 Hiérarchie du projet	6
3.1.2 Configuration de la structure du projet	7
3.1.3 Livrable 4 - Chargement de la dimension PRODUIT et de la dimension CLIENT	8
3.1.4 Livrable 5 - Mise à jour la dimension PRODUIT	9
3.2 La phrase de ELT utilisant de l'outil Talend Studio	9
3.2.1 Livrable 6 – Job Talend réalisé	9
3.2.2 Livrable 7 – Un export du projet Talend	11

# 1. Introduction et Description générale

## 1.1 Introduction

Les systèmes décisionnels au sens des entrepôts de données sont nés d'un besoin exprimé par les entreprises qui n'était pas satisfait par les systèmes traditionnels de la base de données. En intégrant la technologie des entrepôts de données (data warehouses), le processus décisionnel apporte une réponse au problème de la croissance continue des données pouvant être de formats différents.

Il supporte efficacement les processus d'analyse en ligne (On-Line Analytical Processing – OLAP), se sont développés des systèmes de gestion de bases de données multidimensionnelles.

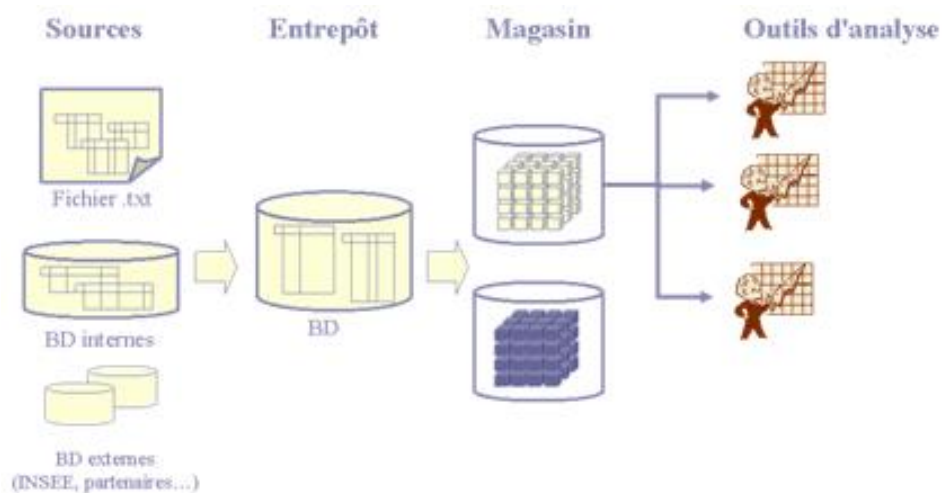


Figure 1 : Architecture général d'un entrepôt du système décisionnel

## 1.2 Description d'architecture du système décisionnel

- Les sources d'information
  - Une base de données SQLite (**comptoir.db**) servant à enregistrer les ventes effectuées.
  - Un fichier Excel (**Clients.xls**) recueillant des informations sur les clients de la société.
  - Un fichier XML (**vendeurs.xml**) exporté de l'outil RH permettant d'obtenir des informations sur les commerciaux de la société.
- L'entrepôt de données :

L'espace de stockage de la base de données d'Oracle centralisé regroupant dans un format homogène les données issues de différentes sources, qui peuvent faire l'objet de transformations et d'historisation, à des fins d'analyse pour répondre aux questions posées et ainsi pour la prise de décision.
- Le magasin de données :

Il peut constituer un extrait de l'entrepôt, où les données sont préparées de manière spécifique pour faciliter l'analyse et l'exploitation.

## 2. Modélisation de l'entrepôt de données

Les modèles multidimensionnels ont pour objectif de proposer un accès aux données intuitif et performant. Notre modèle correspond au modèle en étoile, la table des faits « VENTE » qui contient des informations à analyser, des tables de dimensions qui contiennent les informations sur les dimensions à l'analyser (le temps, le lieu, la description du produit, etc.)

### 2.1 Livrable 1 - La capture d'écran du modèle conceptuel de l'entrepôt de données dans le formalisme Data Fact Model.

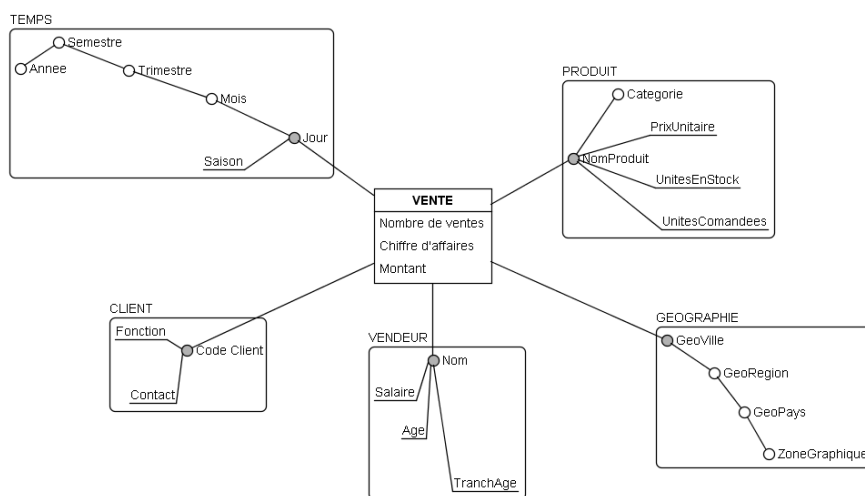


Figure 2 : Data fact modèle de l'entrepôt de données

Le modèle conceptuel est caractérisé par une table centrale de faits entourée de plusieurs tables de dimensions. Chaque table de dimension est reliée à la table de faits par une clé étrangère correspondante. Les clés étrangères dans la table VENTE (ID\_PERIODE, ID\_CLIENT, ID\_VENDEUR, ID\_PRODUIT, ID\_GEOGRAPHIE) établissent les liens avec les tables de dimensions TEMPS, CLIENT, VENDEUR, PRODUIT et GEOGRAPHIE, respectivement.

#### Table de Faits :

**VENTE** contient des mesures telles que le nombre de ventes, le chiffre d'affaires et le montant.

#### Tables de Dimensions :

**TEMPS** fournit des informations sur la période de vente, comme le jour, le mois, le trimestre, le semestre et la saison.

**CLIENT** fournit des détails sur les clients, tels que le code client, le contact principal et la fonction du contact.

**VENDEUR** contient des informations sur les commerciaux, comme leur nom, leur âge, leur tranche d'âge et leur salaire.

**PRODUIT** donne des détails sur les produits vendus, tels que la référence, le nom, la catégorie, le prix unitaire, etc.

**GEOGRAPHIE** fournit des informations sur la localisation géographique des ventes, y compris la ville, la région, le pays et la zone géographique.

## 2.2 Livrable 2 - La capture d'écran des tables relationnelles de l'entrepôt de données

Notre modèle correspond au **modèle en étoile**. La table des faits « VENTE » contient les clés étrangères vers les tables de dimension. L'ensemble de ces clés étrangères forment la clé primaire, la table des faits étant en BCNF.

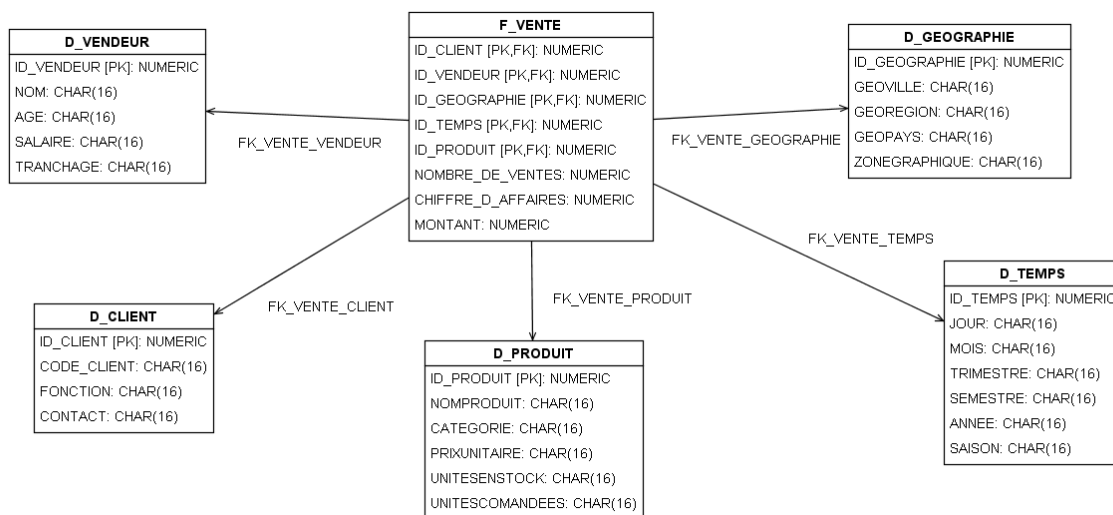


Figure 3 : Des tables relationnelles de l'entrepôt de données

Les **avantages** de ce modèle sont la facilité de navigation, grâce à la table des faits centrale, de **bonnes performances** en raison du **faible nombre de jointures** à effectuer pour l'analyse sur une dimension donnée et des agrégations faciles des mesures.

En comparant avec la modélisation en flocons de neige - une première variante du modèle en étoile, il consiste à décomposer les dimensions d'un modèle en étoile en hiérarchies explicites, chacun des niveaux de la hiérarchie correspondant à une table dans une implémentation ROLAP. Cette modélisation permet de réduire le volume de stockage et autorise des analyses par paliers sur la dimension hiérarchisée. En revanche, les jointures nécessaires pour accéder aux données entraînent une dégradation des performances.

## 2.3 Livrable 3 - Le script SQL du schéma en étoile schéma\_étoile.sql

## 3. Peuplement de l'entrepôt

L'alimentation d'un entrepôt de données est une phrase essentielle dans le processus d'entreposage. Elle se déroule en plusieurs étapes : extraction, transformation, chargement et rafraîchissement des données, qui sont prises en charge par le processus d'ETL (Extracting, Transforming and Loading).

**Extract** : accès aux données sources, on fait le programme en Java pour se connecter à la base de données SQLite, ainsi pour lire le fichier Excel.

**Transform** : transformation pour homogénéiser (formatage, nettoyage des données).

**Load** : chargement dans l'entrepôt. La mise à jour de l'entrepôt est coûteuse et se fait de manière régulière, par exemple toutes les nuits. Elle peut être incrémentale ou nécessiter le recalcul de tout l'entrepôt.

### 3.1 La phrase de ELT en Java

#### 3.1.1 Hiérarchie du projet

Dans le dossier src, nous avons deux classes : `ChargementED.java` et `MiseAJour.java`.

La première classe permet de charger les dimensions de l'entrepôt à partir des fichiers sources, tandis que la seconde permet de mettre à jour la dimension Produit.

Pour que le projet fonctionne correctement, nous devons ajouter les archives suivantes au classpath :

- jxl-2.6.jar
- ojdbc11.jar
- sqlite-jdbc-3.8.11.2.jar

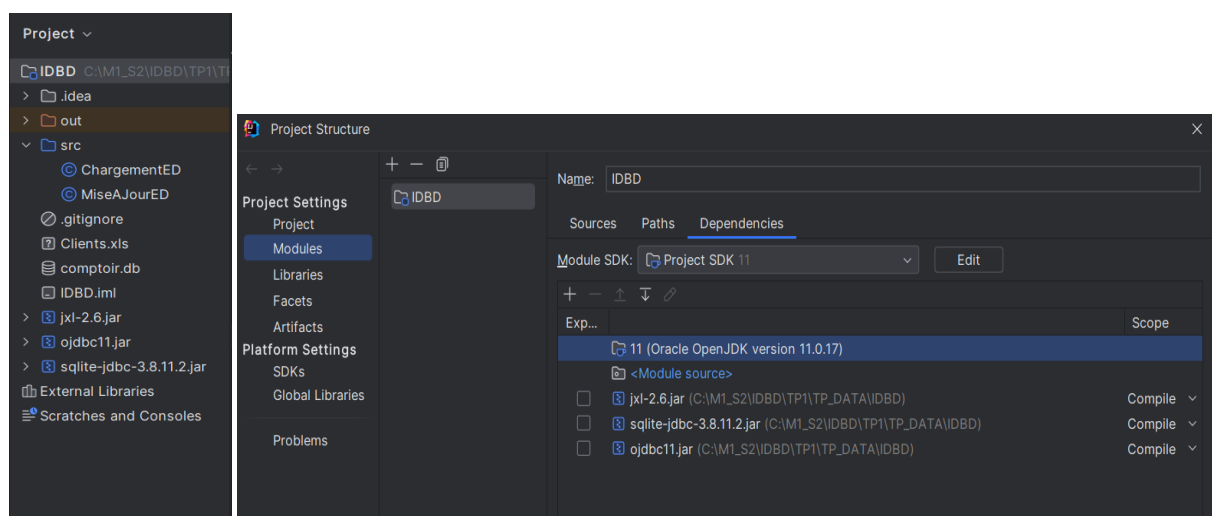
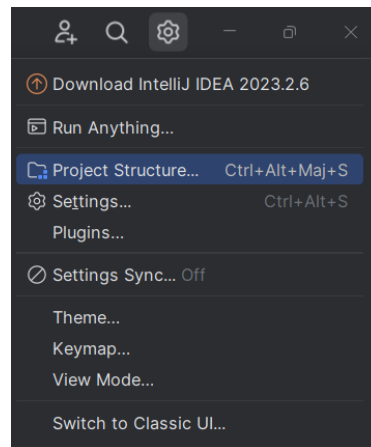


Figure 4 : Capture d'écran de la structure de projet sur IntelliJ

### 3.1.2 Configuration de la structure du projet

Dans IntelliJ IDEA, pour configurer la structure du projet :

- Dans Project Settings, sélectionner Modules.
- Sélectionner l'onglet Dependencies.
- Cliquer sur le signe "+" et choisir "JARs or directories".
- Ajouter les archives à partir de l'emplacement correspondant sur le système.



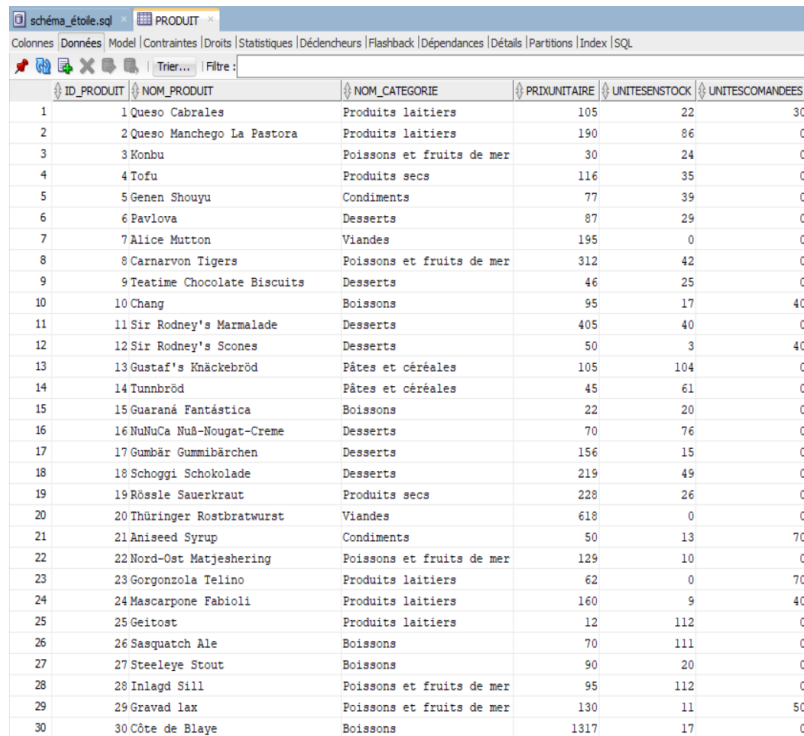
Une fois les archives ajoutées, le fichier IDBD.iml devrait comporter les bibliothèques ajoutées comme suit :

```
MiseAJourED.java  IDBD.iml  ChargementED.java
2  <module type="JAVA_MODULE" version="4">
3    <component name="NewModuleRootManager" inherit-compiler-output="true">
4      <exclude-output />
5      <content url="file://$MODULE_DIR$">
6        <sourceFolder url="file://$MODULE_DIR$/src" isTestSource="false" />
7      </content>
8      <orderEntry type="inheritedJdk" />
9      <orderEntry type="sourceFolder" forTests="false" />
10     <orderEntry type="module-library">
11       <library>
12         <CLASSES>
13           <root url="jar://$MODULE_DIR$/jxl-2.6.jar!/" />
14         </CLASSES>
15         <JAVADOC />
16         <SOURCES />
17       </library>
18     </orderEntry>
19     <orderEntry type="module-library">
20       <library>
21         <CLASSES>
22           <root url="jar://$MODULE_DIR$/sqlite-jdbc-3.8.11.2.jar!/" />
23         </CLASSES>
24         <JAVADOC />
25         <SOURCES>
26           <root url="jar://$MODULE_DIR$/sqlite-jdbc-3.8.11.2.jar!/" />
27         </SOURCES>
28       </library>
29     </orderEntry>
30     <orderEntry type="module-library">
31       <library>
32         <CLASSES>
33           <root url="jar://$MODULE_DIR$/ojdbc11.jar!/" />
34         </CLASSES>
35       </library>
36     </orderEntry>
37   </component>
38 </module>
```

Figure 5 : Extrait de code du fichier iml après avoir ajouté les archives

### 3.1.3 Livrable 4 - Chargement de la dimension PRODUIT et de la dimension CLIENT


Le code de chargement est réalisé dans le fichier « **ChargementED.java** »



ID_PRODUIT	NOM_PRODUIT	NOM_CATEGORIE	PRXUNITAIRE	UNITESENSTOCK	UNITESCOMANDEES
1	1 Queso Cabrales	Produits laitiers	105	22	30
2	2 Queso Manchego La Pastora	Produits laitiers	190	86	0
3	3 Konbu	Poissons et fruits de mer	30	24	0
4	4 Tofu	Produits secs	116	35	0
5	5 Genen Shouyu	Condiments	77	39	0
6	6 Pavlova	Desserts	87	29	0
7	7 Alice Mutton	Viandes	195	0	0
8	8 Carnarvon Tigers	Poissons et fruits de mer	312	42	0
9	9 Teatime Chocolate Biscuits	Desserts	46	25	0
10	10 Chang	Boissons	95	17	40
11	11 Sir Rodney's Marmalade	Desserts	405	40	0
12	12 Sir Rodney's Scones	Desserts	50	3	40
13	13 Gustaf's Knäckebröd	Pâtes et céréales	105	104	0
14	14 Tunnbröd	Pâtes et céréales	45	61	0
15	15 Guaraná Fantástica	Boissons	22	20	0
16	16 NuNuCa Nuß-Nougat-Creme	Desserts	70	76	0
17	17 Gumbär Gummibärchen	Desserts	156	15	0
18	18 Schoggi Schokolade	Desserts	219	49	0
19	19 Rössle Sauerkraut	Produits secs	228	26	0
20	20 Thüringer Rostbratwurst	Viandes	618	0	0
21	21 Aniseed Syrup	Condiments	50	13	70
22	22 Nord-Ost Matjeshering	Poissons et fruits de mer	129	10	0
23	23 Gorgonzola Telino	Produits laitiers	62	0	70
24	24 Mascarpone Fabioli	Produits laitiers	160	9	40
25	25 Geitost	Produits laitiers	12	112	0
26	26 Sasquatch Ale	Boissons	70	111	0
27	27 Steeleye Stout	Boissons	90	20	0
28	28 Inlagd Sill	Poissons et fruits de mer	95	112	0
29	29 Gravd lax	Poissons et fruits de mer	130	11	50
30	30 Côte de Blaye	Boissons	1317	17	0

Figure 6 :

Extrait du résultat de chargement de la dimension « **PRODUIT** » à partir de la base de données SQLite dans l'entrepôt de l'Oracle



ID_CLIENT	CODE_CLIENT	FONCTION	CONTACT
1	1 ALFKI	Repräsentant(e)	Maria Anders
2	2 ANATR	Propriétaire	Ana Trujillo
3	3 ANTON	Propriétaire	Antonio Moreno
4	4 AROUT	Repräsentant(e)	Thomas Hardy
5	5 BERGS	Acheteur	Christina Berglund
6	6 BLAUS	Repräsentant(e)	Hanna Moos
7	7 BLONP	Directeur du...	Fredrick Citeaux
8	8 BOLID	Propriétaire	Martin Sommer
9	9 BONAP	Propriétaire	Laurence Leblan
10	10 BOTTM	Chef comptable	Elizabeth Lincoln
11	11 BSBEV	Repräsentant(e)	Victoria Ashworth
12	12 CACTU	Assistant(e)...	Patricio Simpson
13	13 CENTC	Directeur du...	Francisco Chang
14	14 CHOPS	Propriétaire	Yang Wang
15	15 COMMI	Assistant(e)...	Pedro Afonso
16	16 CONSH	Repräsentant(e)	Elizabeth Brown
17	17 DRACD	Acheteur	Sven Ottilieb
18	18 DUMON	Propriétaire	Janine Labrun
19	19 EASTC	Assistant(e)...	Ann Devon
20	20 ERNISH	Chef des ventes	Roland Mendel
21	21 FAMILA	Assistant(e)...	Aria Cruz
22	22 FISSA	Chef comptable	Diego Roel
23	23 FOLIG	Repräsentant(e)	Martine RancD
24	24 FOLKO	Propriétaire	Maria Larsson
25	25 FRANK	Directeur du...	Peter Franken
26	26 FRANR	Directeur du...	Carine Schmitt
27	27 FRANS	Repräsentant(e)	Paolo Accorti
28	28 FURIB	Chef des ventes	Lino Rodriguez
29	29 GALED	Directeur du...	Eduardo Saavedra
30	30 GODOS	Chef des ventes	JosD Pedro Freyre

Figure 7 :

Extrait du résultat de chargement de la dimension « **CLIENT** » à partir d'un fichier Excel dans l'entrepôt de l'Oracle



## 3.1.4 Livrable 5 - Mise à jour la dimension PRODUIT

Le code de la mise à jour est réalisé dans le fichier « **MiseAJourED.java** »

The screenshot displays two software interfaces. On the left, Oracle SQL Developer shows a schema diagram for 'TP' with tables like CLIENT, PRODUIT, and UNITESENTOUR. On the right, SQLiteStudio shows a table named 'Produits' with columns: RefProduit, NomDuProduit, CodeCategorie, QuantiteParUnite, and PrixUnite. The table contains 80 rows of product data, including items like 'Escargots de B...', 'Raclette Cour...', and 'Grandma's Boya...'. A status bar at the bottom indicates a database connection error.

RefProduit	NomDuProduit	CodeCategorie	QuantiteParUnite	PrixUnite	UnitesED
57	48 tortes	3	48 tortes	246	17
58	Vegete-spread	2	15 pots (625 g)	219	24
59	Wimmers gûte	5	20 sacs x 4 p...	166	22
60	Louisiana Flery Ho...	2	32 bouteilles (8 onces)	105	76
61	Louisiana Hot ...	2	24 pots (8 onces)	85	4
62	Laughing ...	1	24 bouteilles (12 ...)	70	52
63	Scottish Longbreads	3	10 sacs x 8 p...	62	6
64	Gudbrandsdalsost	4	1 carton (10 kg)	180	26
65	Uncle Bob's Organ...	7	12 cartons (1 kg)	150	15
66	Outback Lager	1	24 bouteilles (355 ml)	75	15
67	Hilarymont	4	10 cartons (500 g)	107	26
68	Mozzarella di ...	4	24 cartons (200 g)	174	14
69	Pod Kavir	8	24 pots (150 g)	75	101
70	Longlife Tofu	7	1 carton (5 kg)	50	4
71	Pod Kavir	1	24 bouteilles (0,5 litre)	38	125
72	Lakikiki	1	1 bouteille (500 ml)	90	57
73	Original Frankfurte...	2	12 boîtes	65	32
74	Northwoods ...	2	12 pots (12 onces)	200	6
75	Mishi Kobe Niku	6	18 cartons (500 g)	485	29
76	Chai	1	10 boîtes x 20 sacs	90	39
77	Ikara	8	12 pots (200 g)	155	31
78	Chou à la crème	3	12 pièces par boîte	24	98
79	Canelle	3	1 carton (12 boîtes)	250	76
80	Eclat	3	5 pièces par boîte	30	98

Figure 8 : Extrait du résultat la mise à jour de la dimension « **PRODUIT** » à partir d'un fichier Excel dans l'entrepôt de l'Oracle lorsque de nouvelles lignes sont ajoutées dans la table « **Produits** » de la base de données SQLite

## 3.2 La phrase de ELT utilisant de l'outil Talend Studio

### 3.2.1 Livrable 6 – Job Talend réalisé

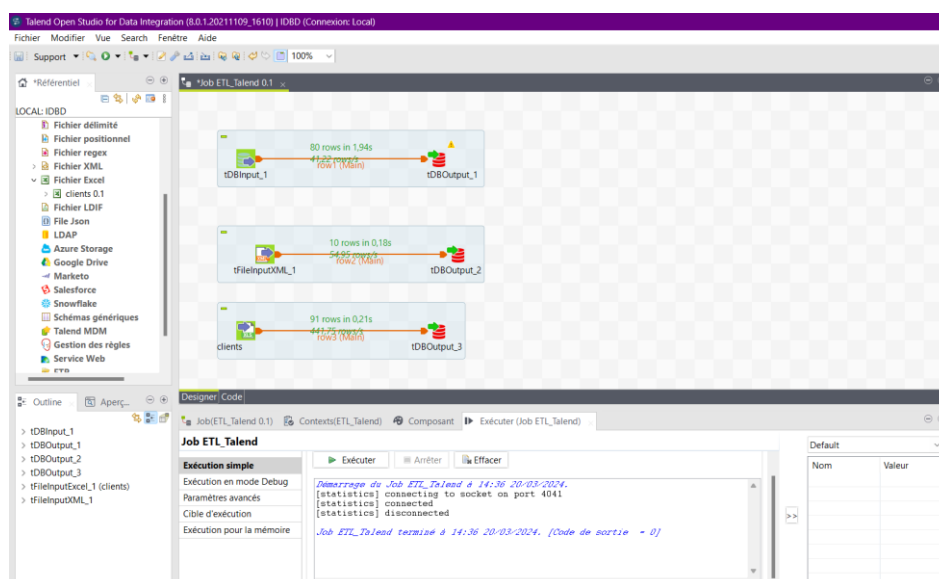


Figure 9 : Capture d'écran de Talend pour présenter le résultat de l'extraction des données depuis des sources dans la sortie Oracle

Oracle SQL Developer : Table TP.CLIENT@TP

Colonne Données Model Contraintes Droits Statistiques Déclencheurs Flashback Dépendances

1 92 ALFRI ReprOeantant(e) Maria Anders

2 93 AJATRI PropriOeaire Ana Trujillo

3 94 AJTON PropriOeaire Antonio Moreno

4 95 ARDIT ReprOeantant(e) Thomas Hardy

5 96 BERGS Acheteur Christina Berglund

6 97 BLAUS ReprOeantant(e) Hanna Moos

7 98 BLOPF Directeur du... FrOeDrique Citeaux

8 99 BOLID PropriOeaire MartCh Sommer

9 100 BODAP PropriOeaire Laurence Leblan

10 101 BOTTH Chef comptable Elizabeth Lincoln

11 102 BSREY ReprOeantant(e) Victoria Ashworth

12 103 CACTO Assistant(e)... Francisco Simpson

13 104 CEMTC Directeur du... Francisco Chang

14 105 CROPS PropriOeaire Yang Wang

15 106 COMEI Assistant(e)... Pedro Afonso

16 107 CONSH ReprOeantant(e) Elizabeth Brown

17 108 DBRAC Acheteur Sven Ottilieb

18 109 DONOW PropriOeaire Janine Labruno

19 110 EASCT Assistant(e)... Ann Devon

20 111 EBNSE Chef des ventes Roland Mendel

21 112 FEMIA Assistant(e)... Aris Crus

22 113 FTISSA Chef comptable Diego Roel

23 114 FOLIG ReprOeantant(e) Martine RancO

24 115 FOLIO PropriOeaire Maria Larsson

25 116 FRANKI Directeur du... Peter Franken

26 117 FRANKI ReprOeantant(e) Carine Schmitt

27 118 FRANS ReprOeantant(e) Paolo Accorti

28 119 FURIB Chef des ventes Lino Rodrigues

29 120 GALED Directeur du... Eduardo Saavedra

30 121 GARDI

Oracle SQL Developer

Colonne Données Model Contraintes Droits Statistiques Déclencheurs Flashback Dépendances

1 81 Queso C... Produits ... 105 22 30 11

2 82 Queso M... Produits ... 190 86 0 12

3 83 Ronbu Poissons ... 30 24 0 13

4 84 ToFu Produits ... 116 35 0 14

5 85 Genen S... Condiments 77 39 0 15

6 86 Pavlova Desserts 87 29 0 16

7 87 Alice M... Viandes 195 0 0 17

8 88 Carnarv... Poissons ... 312 42 0 18

9 89 Teatime... Desserts 46 25 0 19

10 90 Chang Boissons 95 17 40 2

11 91 Sir Rod... Desserts 405 40 0 20

12 92 Sir Rod... Desserts 50 3 40 21

13 93 Gustaf... FDes et ... 105 104 0 22

14 94 Tunnbrd... FDes et ... 45 61 0 23

15 95 Guarand... Boissons 22 20 0 24

16 96 NuluCa... Desserts 70 76 0 25

17 97 GumbGr... Desserts 156 15 0 26

18 98 Schoggi... Desserts 219 49 0 27

19 99 Rasle... Produits ... 228 26 0 28

20 100 ThDring... Viandes 618 0 0 29

21 101 Aniseed... Condiments 50 13 70 3

22 102 Nord-Os... Poissons ... 129 10 0 30

23 103 Gorgoni... Produits ... 62 0 70 31

24 104 Mascarp... Produits ... 160 9 40 32

25 105 Geitost... Produits ... 12 112 0 33

26 106 Saagat... Boissons 70 111 0 34

27 107 Steeley... Boissons 90 20 0 35

28 108 Inlagd... Poissons ... 95 112 0 36

29 109 Graveld... Poissons ... 130 11 50 37

30 110 FDes et ...

Talend Open Studio for Data Integration (8.0.1.20211109.1610) | IDBD (Connexion: Local)

Job ETL\_Talend 0.1

80 rows in 1.46s

10 rows in 0.23s

91 rows in 0.21s

Job ETL\_Talend

Exécution simple

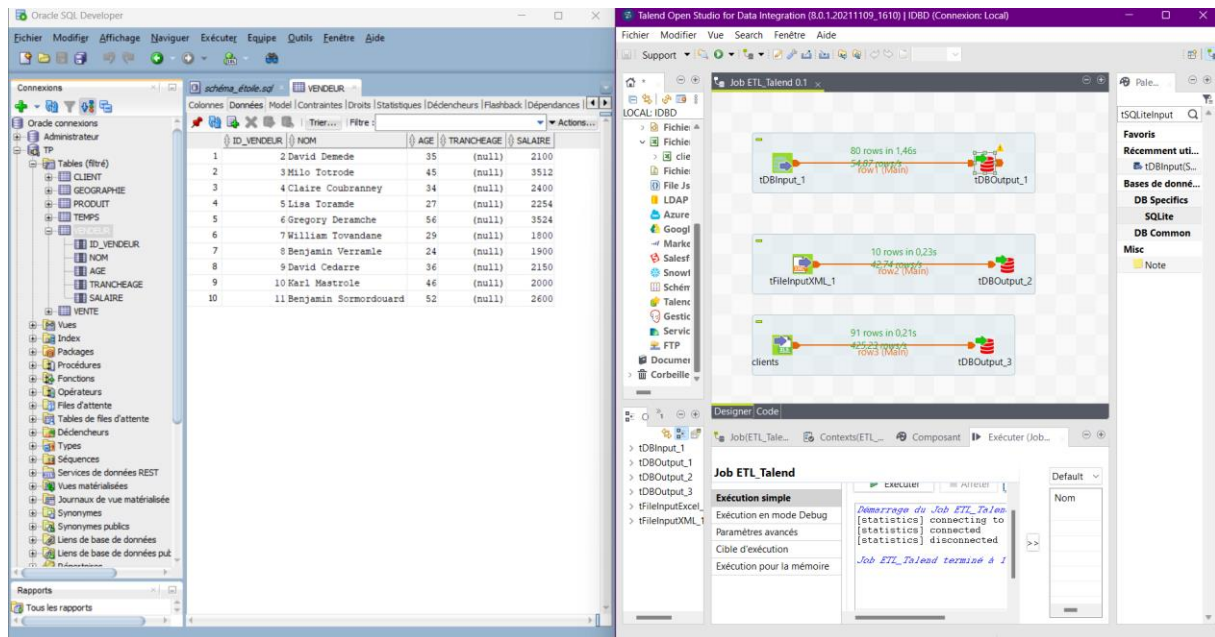
Exécution en mode Debug

Paramètres avancés

Cible d'exécution

Exécution pour la mémoire

Job ETL\_Talend termine à 1



### 3.2.2 Livrable 7 – Un export du projet Talend

L'export du projet Talend sous forme d'un fichier ChargementED.zip permettant de charger les dimensions.

*Fin de rapport. Merci pour votre lecture.*