



Haute École de Namur - Liège - Luxembourg  
Implantation IESN

## Business Intelligence and Data Analytics

À l'attention de Monsieur S. Scholtes

Henallux IESN - Informatique de Gestion Bloc 3

PHILIPIN Pierre

LUTGEN Jordan

Année Académique : 2019-2020



## Table of Contents

|   |   |
|---|---|
| Description de la solution.....                     | 4 |
| Diagrammes .....                                    | 5 |
| Diagramme de la base de données source.....         | 5 |
| Diagramme de la base de données dimensionnelle..... | 6 |
| Granularité de la table de faits.....               | 7 |
| Justification des choix.....                        | 7 |
| Clés primaires du modèle DWH.....                   | 7 |
| Modélisation des dimensions et faits.....           | 7 |
| Limites rencontrées.....                            | 8 |
| Stratégie d'exécution .....                         | 8 |

## Description de la solution

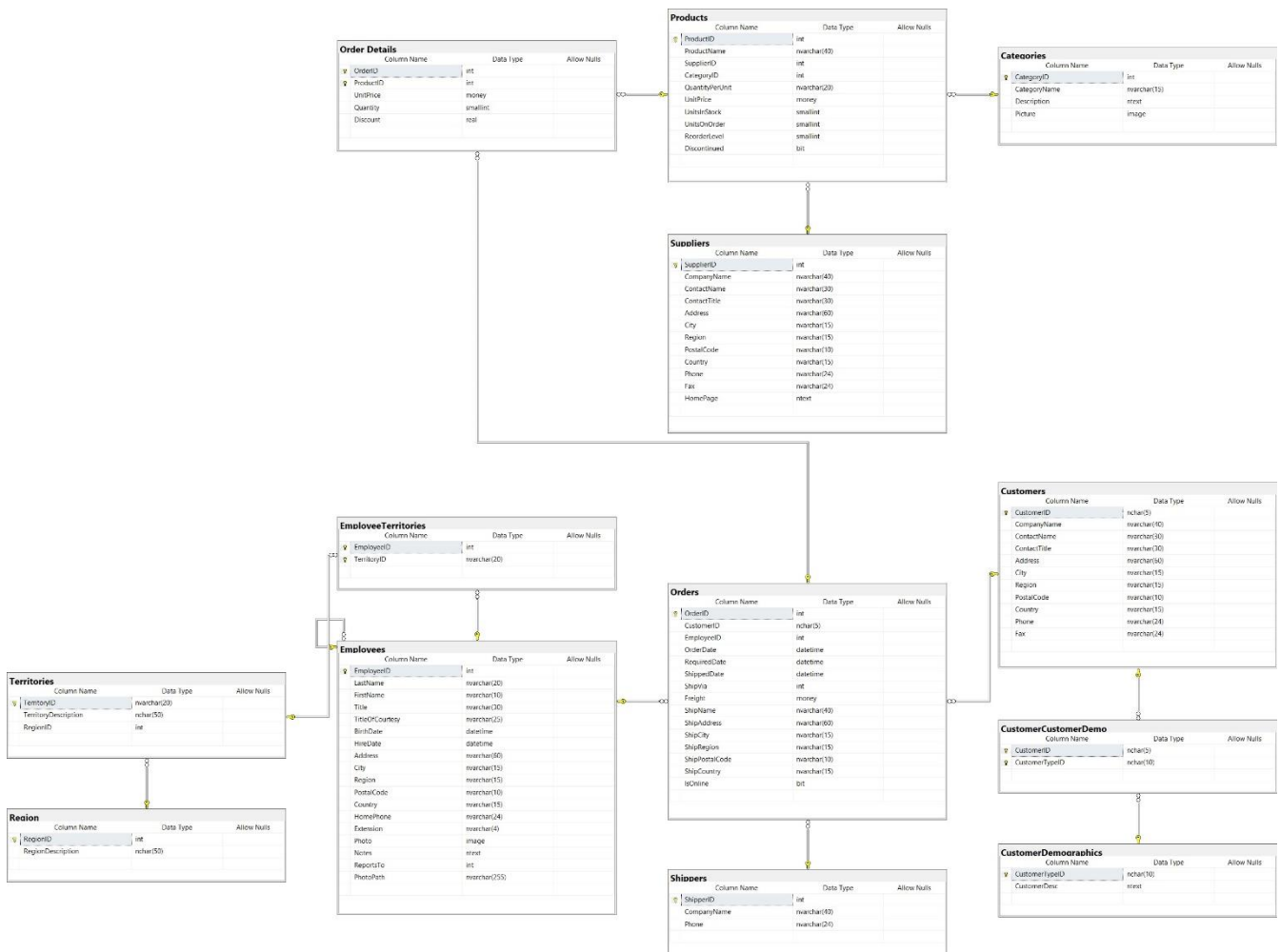
Le groupe commercial NutriFuture gère deux marques de produits alimentaires : Foodies997 et MirkaProShop. Ces marques enregistrent toutes les commandes de produits par le biais de ses employés, commerciaux déployés géographiquement. En plus de ses commerciaux, chaque enseigne dispose d'un webshop proposant le même catalogue de produits que ceux proposés par les commerciaux. Le groupe souhaite pouvoir tirer de l'information de toutes les données. Il souhaiterait disposer d'un outil simple à utiliser qui permettrait d'obtenir des réponses rapidement à des questions précises. Il souhaiterait pouvoir analyser les données sur plusieurs axes, faire des recoupements.

Afin de répondre à la demande de ce groupe, la technologie ETL, mise en place avec ce document, permet de transformer la base de données relationnelle "Foodies997" en base de dimensionnelle. Cette nouvelle base de données comprend six tables représentant des dimensions (Customer, Date, Employee, Product, Shipper et Supplier) et une table de fait (Sale).

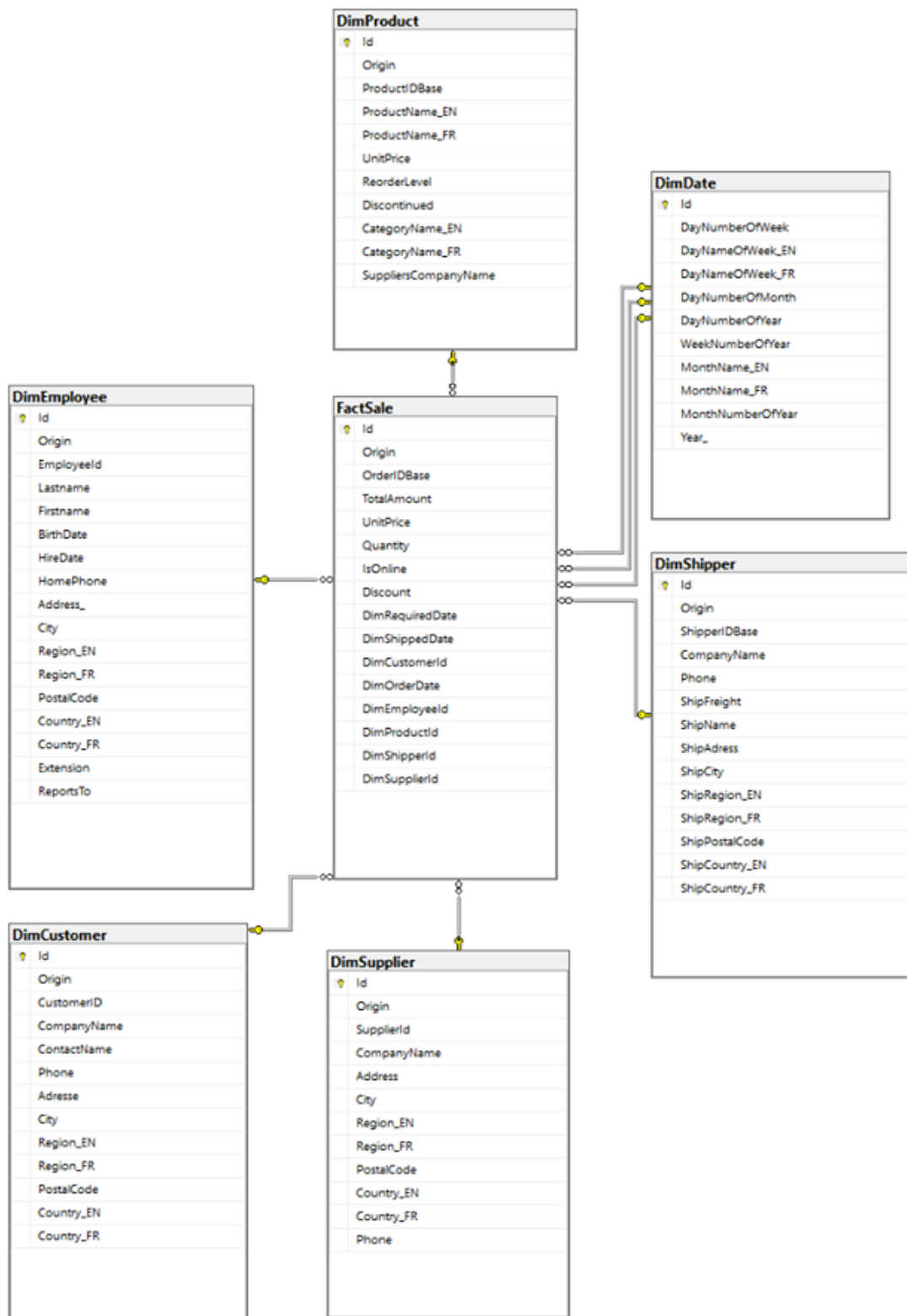
Le traitement débute par la suppression des données des différentes tables de la base de données dimensionnelle, ensuite les différentes dimensions sont remplies et enfin le fait. Toutes les erreurs sont envoyées dans des fichiers plats correspondant à leurs dimensions respectives.

# Diagrammes

## Diagramme de la base de données source



## Diagramme de la base de données dimensionnelle



## Granularité de la table de faits

Afin d'obtenir la meilleure précision possible, la granularité a été choisie au niveau de la ligne de commande (Order Details).

Ceci afin de permettre à l'utilisateur métier de creuser les données le plus loin possible (drill down).

## Justification des choix

### Clés primaires du modèle DWH

Afin que la base de données dimensionnelle soit indépendante de/des base(s) de données source(s), des clés primaires auto-incrémentées (Surogate Keys) sont attribuées à chaque table.

De cette manière, si des enregistrements proviennent de différentes bases de données (tel que Foodies997 et MirkaProShop), les identifiants originaux des différentes tables ne pourront entrer en collisions ou même poser problème s'ils changent de structures selon l'origine.

Les identifiants artificiels sont de types entiers pour des raisons de performances. À l'exception de la table DimDate qui utilise un identifiant naturel de type entier représentant la date au format 'YYYYMMDD' pour des raisons de stockage et de duplications.

### Modélisation des dimensions et faits

Le modèle en étoile a été choisi, car il permet de simplifier la modélisation, de la rendre plus compréhensible par un utilisateur métier et améliore la performance des requêtes en lecture.

Les champs dimensions reprennent, à quelques exceptions près listées ci-après, les champs de leurs tables d'origines respectives.

Exceptions :

- La dimension "DimShipper" reprend les champs de la table d'origine "Shipper" et certains champs de la table "Order"
- Les champs se terminant par "<préfixe>\_EN" reprennent les champs (préfixes) de tables d'origines.
- La dimension "DimDate" reprend les informations concernant un jour. Elle ne provient pas d'une table dans la base de données d'origine. Sa granularité s'arrête au jour. L'enregistrement du 01/01/1900 identifie les enregistrements sans valeur pour la date.
- Des enregistrements spéciaux sont ajoutés aux dimensions suivantes : "DimCustomer", "DimEmployee", "DimShipper" et "DimSupplier". Ces dimensions reprennent un identifiant "Id" à -1, concernant les FK nullable de la base de données d'origine.
- Le fait "FactSale" reprend les champs de la table d'origine "Order Detail" avec les Foreign Key (FK) correspondant aux différentes dimensions qui lui sont rattachées.

## Limites rencontrées

Ne disposant pas des traductions nécessaires à l'internationalisation, cette base de données dimensionnelle dispose déjà des champs en plusieurs langues (anglais et français), mais manque d'informations pour alimenter les champs spécifiques au français, identifié "<préfixe>\_FR".

Il est impossible, avec les données actuelles, de répondre à la question suivante : "Les points de vente disposant d'une grande superficie de vente vendent-ils plus de produits que les autres ?". Aucune des tables de la base de données d'origine ne contient d'information sur la superficie d'un point de vente.

Étant donné la granularité journalière de la dimension "DimDate", il n'est pas possible d'évaluer quelles heures d'une journée sont les plus productives. L'ajout d'une dimension "DimHeure" ne permettrait pas non plus d'y répondre, la précision des informations de la table d'origine "Order" se limite à la journée.

## Stratégie d'exécution

La procédure d'exécution de l'ETL commence par vider entièrement les différentes tables de la base de données dimensionnelle, ensuite chaque dimension est remplie via les données de la table d'origine. De ce fait, même si un crash survient durant l'exécution de l'ETL, le simple redémarrage suffit à recharger entièrement le Dataware House (DWH).

Les différentes erreurs rencontrées durant le traitement sont redirigées vers des fichiers plats au format "Fichier erreurs <tableDimensionConcernée>.txt". Cette redirection permet la continuité du traitement même si une erreur survient.