TP Extract Transform Load avec Talend

1. Présentation

La société Orion

Cette société fictive, présente au niveau mondial, est spécialisée dans la commercialisation d'articles de sport et d'extérieur. Les données disponibles regroupent des informations sur :

- les employés
- les produits
- les clients
- les commandes
- les fournisseurs

Le siège social aux États-Unis, gère des filiales en Belgique (depuis 1999), Pays Bas, Allemagne, Royaume-Uni, Danemark, France, Italie, Espagne et Australie. Les produits sont vendus en magasin, par catalogue et par internet. Une carte de fidélité : 'Orion Star Club', propose beaucoup d'avantages. L'historique d'information va du 1er janvier 1998 au 31 décembre 2002.

Structure de l'organisation

Le siège social héberge la majeure partie des fonctions administratives, soit un nombre important d'employés, entre 600 et 800. Le siège social centralise aussi la gestion des stocks, la vente par catalogue, la vente par internet et l'import - export. Néanmoins, certains employés gèrent aussi ces fonctions depuis les différentes filiales.

Les employés sont enregistrés dans la base de données selon cinq niveaux :

- Pays
- Compagnie
- Département
- Section
- Groupe

Les informations complémentaires sur les employés sont notamment :

- Date d'entrée et de départ de l'employé
- Date de début et de fin de contrat (pour certain contrat)
- Adresse
- Sexe
- Salaire
- Responsable hiérarchique

L'offre

La société propose environ 5500 références. Certaines ne sont pas vendues dans tous les pays, d'autres, de part les volumes commercialisés, reflètent certaines particularités régionales, certains sports nationaux. Tous les noms sont fictifs.

Les produits sont organisés selon 4 niveaux :

- Ligne de produit
- Catégorie de produit
- Groupe de produit
- Produit

Chaque produit a un coût et un prix de vente. Le système informatique gère tous les prix en dollars. En utilisant les dates de début et de fin, ces prix varient en fonction du temps. Cet historique est sauvegardé. Le système gère aussi les remises pour certains produits, à certaines périodes. Les prix sont généralement uniques de part le monde.

Les clients

Les clients sont repartis à travers le monde, notamment dans les pays où se trouvent des filiales, mais pas uniquement. Les noms et adresses sont fictifs, même si les villes, régions/comtés et pays, sont réels. La base de données enregistre environ 90 000 clients, pas tous actifs.

L'adresse des clients comprend tout ou partie des informations suivantes :

- Rue
- Code postal
- Ville
- Région / département / conté
- Etat
- Pays
- Continent

Les clients sont classés dans des groupes en fonction de leur activité d'achat.

Les commandes

Chaque commande pointe vers le commercial qui a enregistré la vente. Environ 980 000 commandes sont enregistrées, commandes qui reflètent notamment les saisonnalités. Chaque commande comprend une ou plusieurs lignes, une ligne par produit.

Les fournisseurs

Chaque produit provient d'un fournisseur qui est basé dans un pays, mais toutes les commandes sont passées par le siège social. Il y a 64 fournisseurs, mais un seul fournisseur par produit.

2. Mise en place d'un système décisionnel

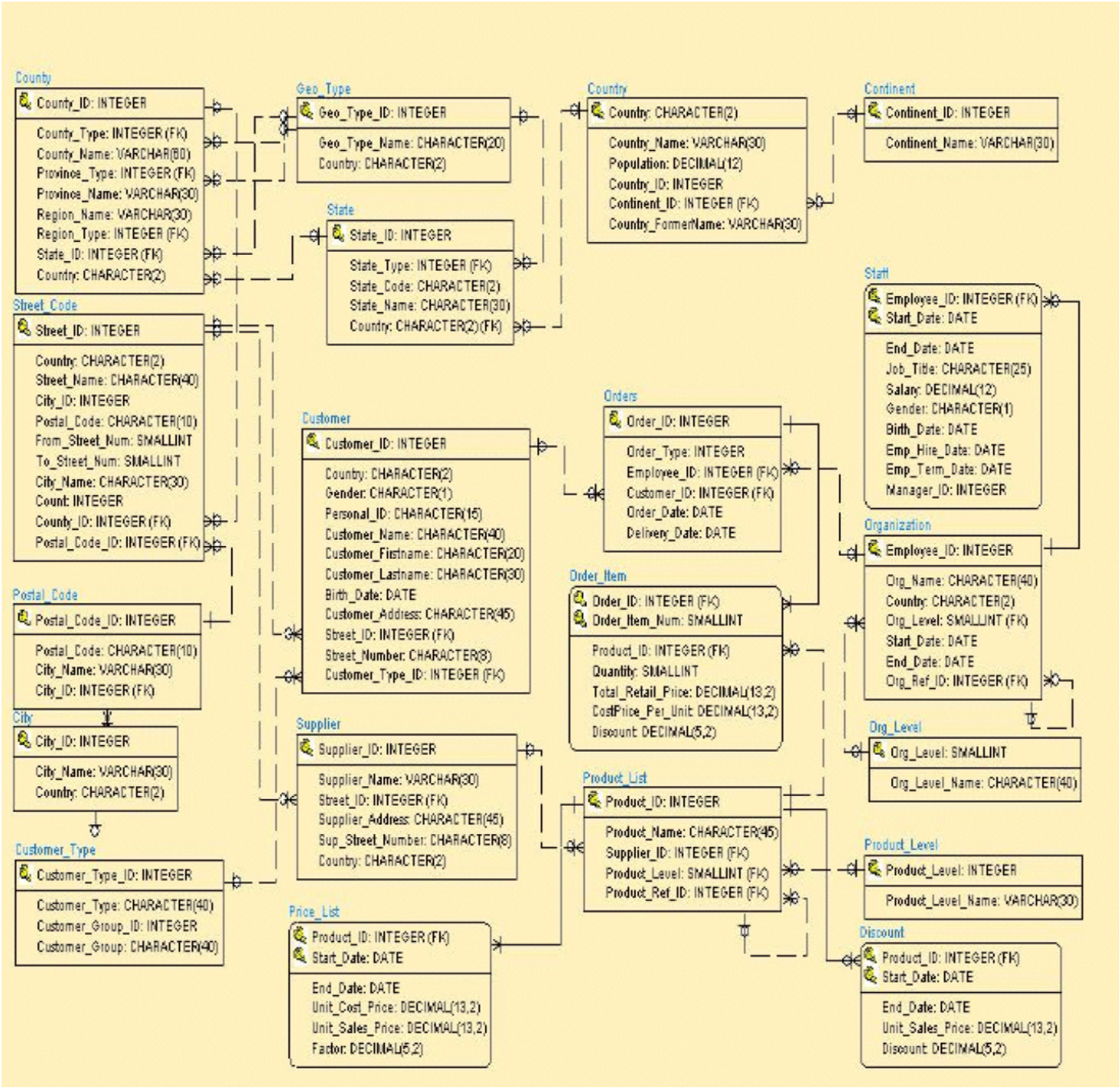
La société Orion souhaite améliorer sa performance à l'aide d'un système décisionnel. Voici quelques questions qui ont été recensées et auxquelles devrait répondre le système mis en place :

- Quels sont les produits qui se vendent le mieux ?
- Quels sont les produits en perte de vitesse ?
- Quels sont les produits qui contribuent très peu au chiffre d'affaire pour un pays et une année donnés ? Est-ce que ces produits peuvent être remisés ?
- Quelle est la marge générée par ce groupe de produit ?
- Est-ce que la marge dépend de la quantité vendue ?
- Est-ce que les remises font augmenter les ventes ?
- Est-ce que les remises font augmenter la marge ?
- Quels sont les commerciaux qui font le plus de ventes ?
- Quels sont les commerciaux qui performent le mieux par pays, sexe, âge, salaire ?
- Quels groupes de clients sont identifiés ?
- Quels sont les clients les plus rentables ?
- Quels fournisseurs proposent des produits rentables?
- Quelle est la moyenne et l'écart-type du chiffre d'affaire ?
- Quelles sont les variables qui expliquent le mieux l'importance du chiffre d'affaire ?
- Y-a-t'il une différence significative entre la moyenne de la somme du chiffre d'affaire géré par les commerciaux de sexe féminin et celle des commerciaux de sexe masculin ?

Il faut donc construire un entrepôt de données capable de répondre aux besoins de requête, de reporting, et d'analyses avancées.

3. Données sources

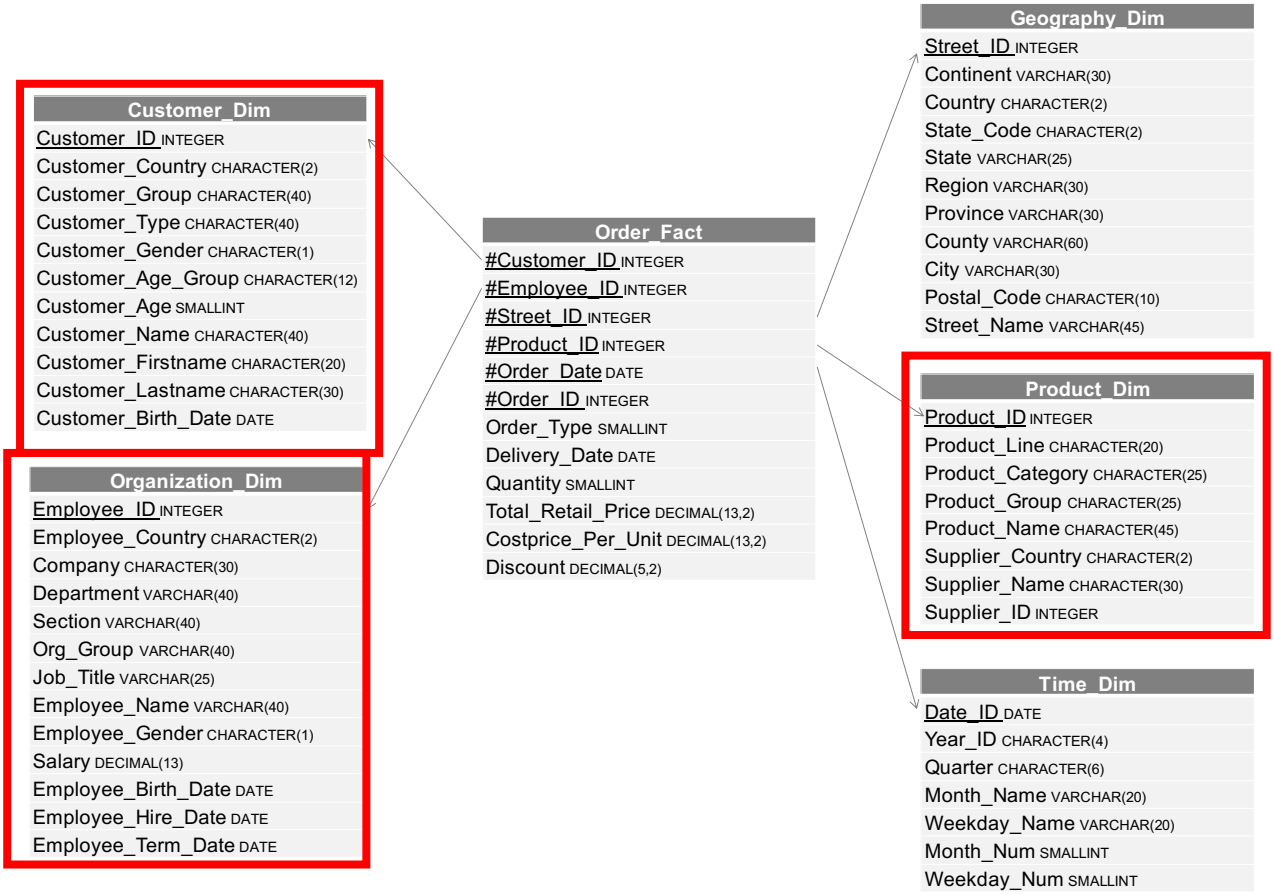
Voici le schéma relationnel de la base de données opérationnelle de l'entreprise d'où proviendront les données de l'entrepôt :



Ces tables ont été exportées sous format TXT hormis la table Staff stockée dans le fichier Microsoft Excel nommé staff.xls. La base de données est fournie avec le support de TP sur le site du cours.

4. Schéma de l'entrepôt

Voici le schéma en étoile de l'entrepôt de données :



Implémenter la création des tables de l'entrepôt sous MySQL. Commencer par les tables Customer_Dim et Product_Dim grâce aux scripts disponibles sur le site du cours TIW2.

Une bonne pratique est de créer un utilisateur spécifique pour votre ETL dans votre SGBD.

Maintenant que les tables de l'entrepôt sont créées, il faut réaliser les processus qui vont remplir ces tables à partir des données sources.

5. Connexion aux sources d'entrées et de sorties

- Etablir une connexion avec les fichiers nécessaires pour le projet :

Pour chaque fichier, il y a le type de chaque colonne dans la base de données sources (DB Type) et sa traduction dans Talend (Type). Dans un premier temps, utilisez seulement String, vous ferez les transformations dans un deuxième temps.

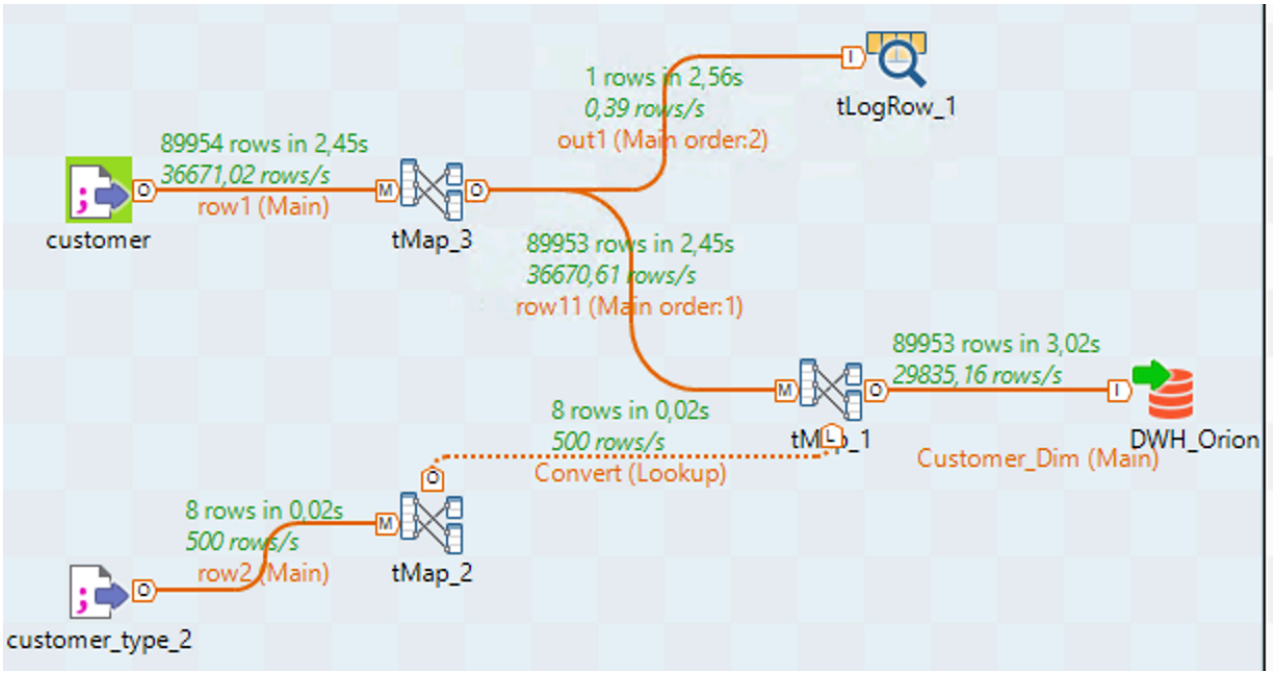
- Modifier les types des colonnes de façon à n'avoir que des Double, des String ou des Date.
- Etablir une connexion à votre entrepôt de données afin de pouvoir le remplir avec Talend (bien utiliser les types de la BDD : INT, NVARCHAR, SMALLINT, DATE).

6. Création du job de remplissage de la dimension Customer_Dim

- Pour chaque colonne de la table Customer_Dim, spécifier de quelle(s) donnée(s) source elle dépend.

Table cible	Colonne cible	Table(s) source	Colonne(s) source	Remarques

- Créer un job nommé Job01_Customer_Dim



- Choisir les fichiers sources (Customer puis Customer_Type) et les importer dans le Design Workspace
- Choisir la table cible (Customer_Dim) et l'importer dans le Design Workspace avec l'option tMySQLOutput ou TMSSqlOutput.
- A l'aide d'un tMap pour chaque fichier Customer et Customer_Type, modifier les types des colonnes :
  - Pour la colonne Customer_Type_ID du fichier Customer_Type, modifier son type String en Int avec les méthodes suivantes : Integer.parseInt, StringHandling.SUBSTR, indexOf(",")
  - Pour le fichier Customer,
    - Modifier la date d'anniversaire avec la méthode TalendDate.parseDate pour que celle-ci devienne une date
    - Modifier aussi le Customer_Type_ID pour que celui-ci soit aussi en INT
    - Modifier les colonnes Country et Gender avec la fonction StringHandling.SUBSTR pour que les données extraites soient bien présents dans la table du datawarehouse
- Après ces premières transformations, ajouter le composant « tMap » pour faire le lien entre les données sources et les données cibles.
  - Faire une jointure entre les deux tables sources (Customer_Type et Customer).
  - Relier les colonnes sources aux colonnes cibles en faisant les transformations nécessaires
  - Créer une nouvelle variable avec l'âge des clients :
    - Expression :
      Short.parseShort(String.valueOf(Mathematical.INT(TalendDate.formatDate("yyyy",TalendDate.getCurrentDate()))-Mathematical.INT(TalendDate.formatDate("yyyy",row11._Birth_Date_))))

        - Type : double
        - Variable : age
    - La colonne cible CUSTOMER_AGE est égale à cette variable age.
    - La colonne cible CUSTOMER_AGE_GROUP est définie de la façon suivante :
        - Var.age<30?"<30 years": Var.age<46?"30-45 years": Var.age<61?"46-60 years": Var.age<76?"61-75 years": ">75 years"
    - Utiliser les méthodes suivantes pour vérifier que les données ne dépassent pas dans la colonne de la table du datawarehouse : Integer.parseInt, StringHandling.SUBSTR, .indexOf(",")
    - Enfin, modifier la date d'anniversaire pour que celle-ci correspondent bien à la colonne attendue dans le datawarehouse avec les méthodes suivantes : TalendDate.TO_DATE, TalendDate.formatDate
- Dans le Design workspace, vous afficherez un petit commentaire pour décrire le job à l'aide du composant Misc / Note (à faire pour tous les jobs).
- Attention de bien supprimer les données dans le composant de sortie (option à sélectionner) afin de ne pas avoir de doublon à chaque lancement du job.
- Lancer le job en cliquant sur Run.
- Attention il se peut qu'une des lignes du fichier Customer ne soit pas bonne et déclenche des erreurs dans votre job. Dans le premier tMap, vous pouvez ajouter des filtres (Deuxième bouton)

- Ces filtres permettent de pouvoir filtrer en fonction de données attendues (les données ne correspondant pas ne sont pas prises en compte).
- Servez-vous du filtre pour enlever la ligne vous posant problème

7. Création du job de remplissage de la dimension Product_Dim

- Pour chaque colonne de la table Product_Dim, spécifier de quelle(s) donnée(s) source elle dépend.
- Créer un job nommé Job02_Product_Dim.
- Choisir les fichiers sources (Product List, Product Level, Supplier) et la table cible
- Modifier et nettoyer les données avec les méthodes vues dans le job précédent
- Ajouter le composant Processing / tMap puis ajouter les liens entre les différents composants (renommer les liens avec des noms pertinents).
- Dans le composant tMap :
    - Faire les jointures entre les différentes tables sources (utiliser les filtres).
    - Relier les colonnes sources aux colonnes cibles.
- Lancer le job