UNIVERSITÉ DE TECHNOLOGIE DE COMPIÈGNE
NF26 : DATA WAREHOUSE ET DÉCISIONNEL

Compte-Rendu TD2 : Alimentation avec la Plateforme Pentaho – Kettle

Jingyi Hu & Julien Jerphanion

Printemps 2018

1 Présentation du sujet de TD

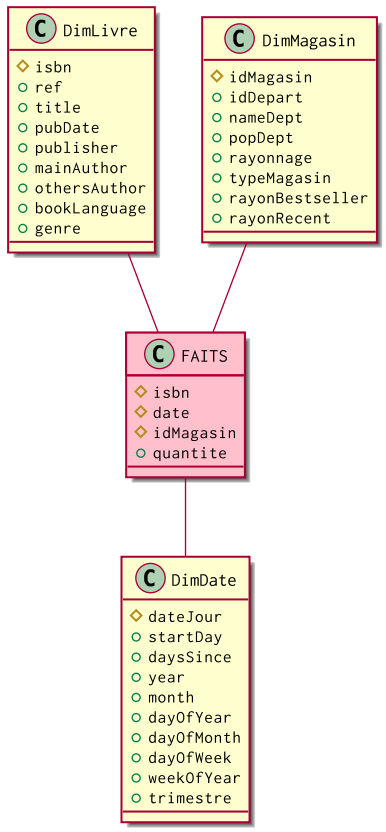

FIGURE 1 – Contexte, dimensions et table de faits

Lors du premier sujet de TD, nous avons pu modéliser les processus de l'entreprise *Fantastic*. Nous nous sommes sur cette partie concentrés sur l'alimentation de tables dans la bases de données à partir de différentes sources grâce au logiciel *Pentaho Kettle*[1] dédié à l'intégration de données.

Le schéma ci-contre (Figure 1) reprend le *contexte* développé en fin du premier TD dans une version affinée et améliorée.

Il a été question dans cette suite de TD de *réaliser l'intégration des différentes dimension afin de construire la table de faits*. Les sections suivantes reprennent de manière concise ce qui a été réalisé pour l'intégration de chacune des dimensions jusqu'aux jointures finales pour la table de faits.

2 Dimension Magasin – `DimMagasin`

Une jointure est réalisée entre le fichier `marketing.ods` et les départements avant d'alimenter une table dans la base de données Oracle.

Pour ce faire, une étape de validation est mise en place pour s'assurer que les identifiants des magasins sont bien formatée. Une autre validation est effectuée du côté des départements sur le fichier `departementsInsee2003-nf26.csv`.

1. Site officiel de *Pentaho Kettle* : `http://www.pentaho.com/product/data-integration`.

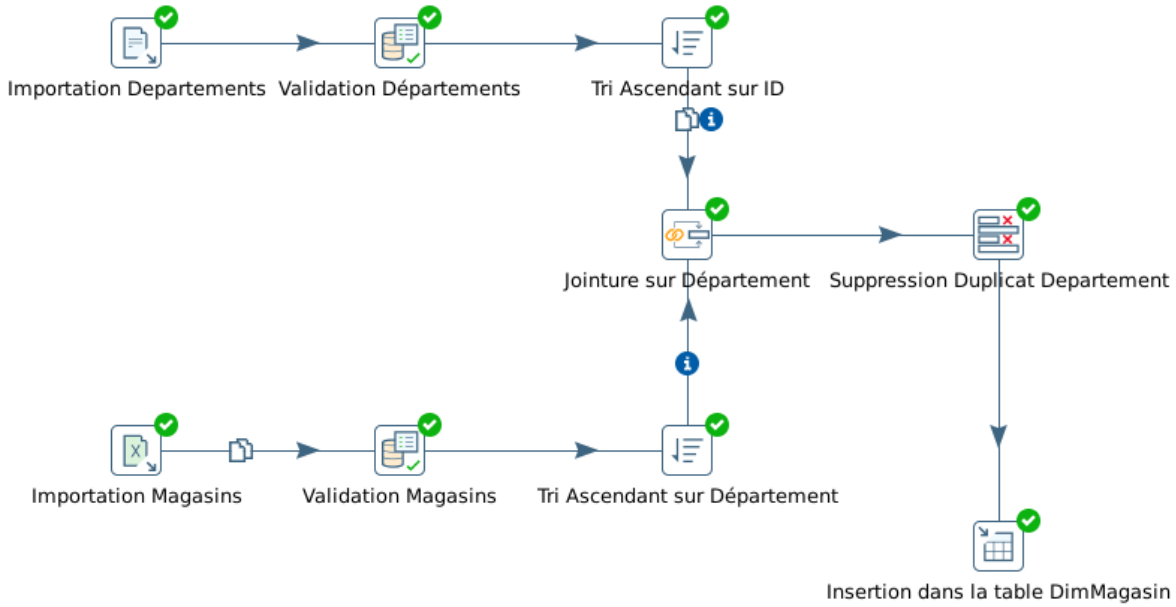

FIGURE 2 – Traitement de la Dimension Magasin

Le tout est inséré dans une table `DimDate` dans le schéma Oracle. La figure 2 reprend les traitements effectués pour cette dimension.

3 Dimension Date – `DimDate`

Pour la génération de la date, nous avons repris le template proposé par *Kettle*, c'est à dire le fichier `"Pentaho/design-tools/data-integration/samples/transformations/General - Populate date dimension.ktr"` où `Pentaho` est le répertoire d'installation du logiciel.

Nous avons adapté ce dernier fichier pour qu'il puisse générer des dates sur la même période que celle des ventes; en particulier, nous avons veiller à retirer les heures et à changer le formatage pour que celui-ci corresponde à ceux utiliser dans la suite. Les données générées ont été stockées dans la table une table `DimDate`. La figure 3 reprend les traitements effectués pour cette dimension.

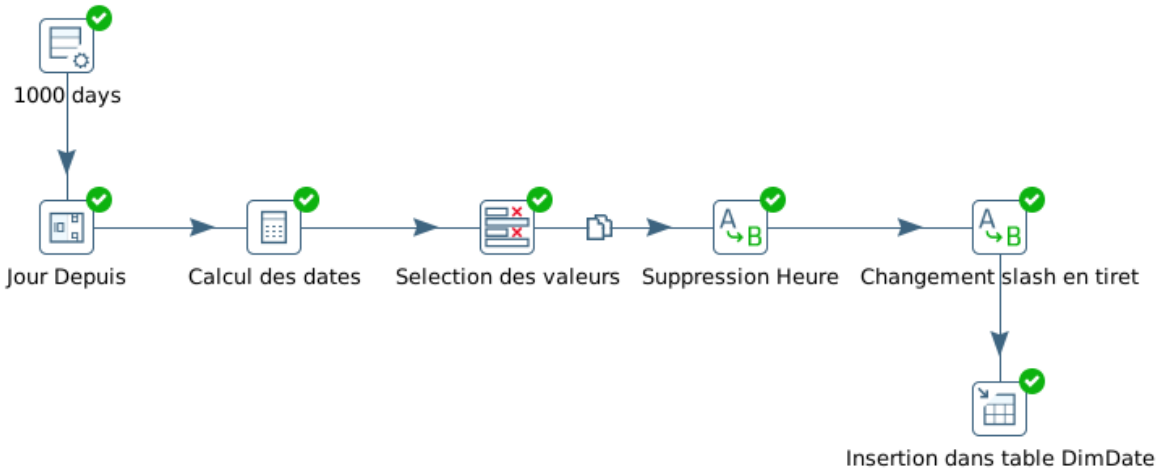

FIGURE 3 – Traitement de la Dimension Date

4 Dimension Livre – `DimLivre`

Pour les produits, il s'est principalement agit de récupérer les données depuis le catalogue disponible en lecture publique sur le schéma `NF26PROF2`. Les données ont été traitées pour extraire l'auteur principal et répertorier, s'il y en avait, les autres auteurs. La chaîne de caractère `"0000000000000"` a été choisie comme valeur par défaut pour les ISBN mal formatés. De mêmes, les autres champs ont été corrigés pour intégrer des valeurs par défaut sur les données manquantes. Les données ont été stockées dans la table une table `DimLivre`. La figure 4 reprend les traitements effectués pour cette dimension.

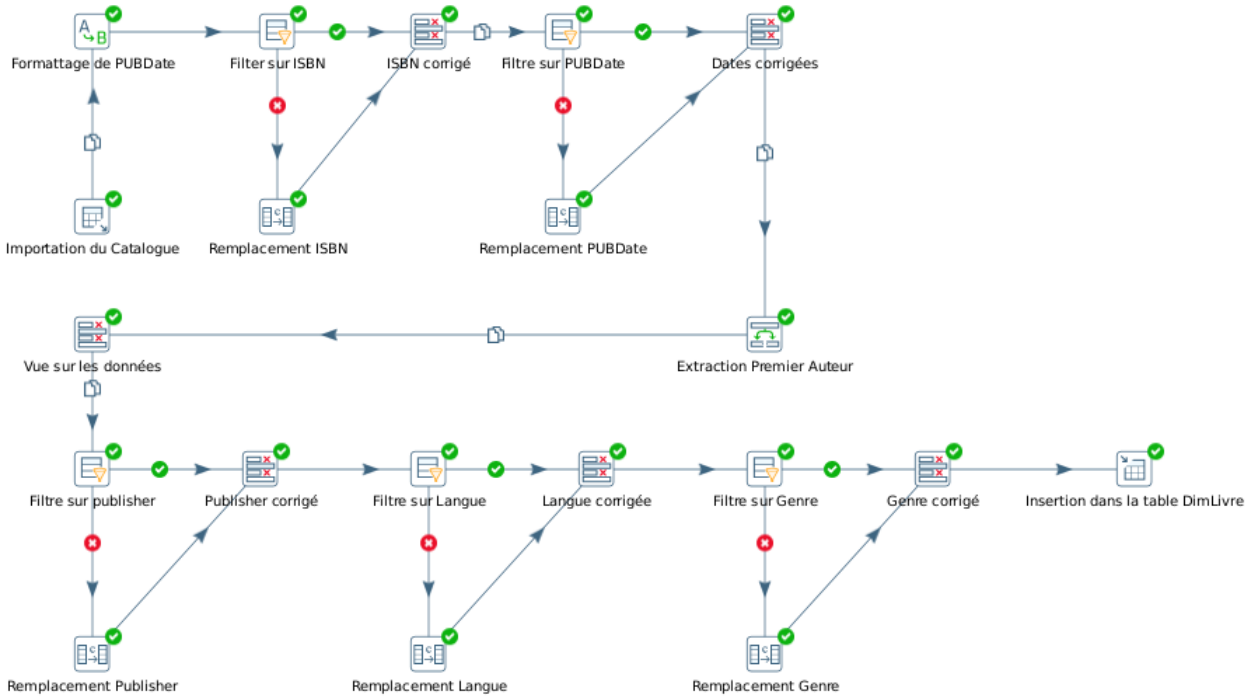

FIGURE 4 – Traitement de la Dimension Livre

5 Table des faits – `Faits`

Sur la table des faits, il s'est principalement agit de joindre les différentes informations grâce aux clés primaires de chacune des dimensions : `(isbn)` pour la dimension `DimLivre`; `(idMagasin)` pour la dimension `DimMagasin` et `(dateJour)` pour la dimension `DimDate`. Pour ce faire, nous avons renommé quelques champs pour que cela soit uniformisé en `camelCase`. Nous avons aussi pris soin de remplacer les données manquantes ainsi :

1. `"9999999999999"` a été choisie comme valeur par défaut pour les ISBN mal formatés du fichier de ventes `Fantastic`;
2. Une chaîne de caractère vide (`""`) a été choisie comme valeur par défaut pour les ID de magasin mal formatés;
3. `"0000-00-00"` a été choisie comme valeur par défaut pour les dates mal formatées;

Des tris ascendants ont été effectués avant d'effectuer les jointures. Après chaque jointure, nous avons veiller à supprimer une clé alors dupliquée puisqu'il s'est agit de faire des jointures internes – `"INNER JOIN"` en *SQL*. De même, il s'est agit de réaliser une agrégation de lignes pour compter le nombre de ventes réalisés d'un même produit, le même jour, et dans le même magasin. Les données agrégées ont ensuite été sauvegardées dans une table dans le schéma Orale.

Au niveau de la constitution de la base de faits, nous n'avons utilisé que des jointures internes aux données (`INNER JOIN` en SQL) mais aurions pu, en prenant du recul, utiliser des jointures externes, à

gauche ou à droite, (`LEFT JOIN` et `RIGHT JOIN` respectivement en SQL) pour conserver l'intégralité des données. Cela aurait nécessité un un plus de traitement – en particulier des remplacement de données nulle – mais aurait sûrement été profitable pour pouvoir ensuite effectuer un traitement sur les ventes pour lesquelles leur contexte n'est pas connu en intégralité. La capture d'écran suivante (Figure 5) reprend les différents traitements des lignes, de leurs importation à leur insertion dans la table de faits.

	Nom étape	N°Copie	Lignes lues	Lignes écrites	Lignes en entrées	Lignes en sortie	Lignes maj	Lignes rejetées	Lignes
1	Importation Dimension Magasin	0	0	152	152	0	0	0	
2	Importation Ventes (Fantastic)	0	0	200000	200000	0	0	0	
3	Filtre sur ISBN	0	200000	200000	0	0	0	0	
4	Importation Dimension Date	0	0	1000	1000	0	0	0	
5	Remplacement ISBN	0	44918	44918	0	0	0	0	
6	Renomage Champ Date	0	1000	1000	0	0	0	0	
7	Importation Dimension Livre	0	0	1443	1443	0	0	0	
8	ISBN corrigé	0	200000	200000	0	0	0	0	
9	Filtre sur Date	0	200000	200000	0	0	0	0	
10	Remplacement Date	0	7660	7660	0	0	0	0	
11	Date corrigée	0	200000	200000	0	0	0	0	
12	Renomage Champ Livre	0	1443	1443	0	0	0	0	
13	Filtre sur ID Magasin	0	200000	200000	0	0	0	0	
14	Renomage Champ Magasin	0	152	152	0	0	0	0	
15	Tri Dimension Date	0	1000	1000	0	0	0	0	
16	Tri Dimension Magasin	0	152	152	0	0	0	0	
17	Remplacement ID Magasin	0	6540	6540	0	0	0	0	
18	ID Magasin Corrigé	0	200000	200000	0	0	0	0	
19	Tri par ID Magasin	0	200000	200000	0	0	0	0	
20	Jointure sur ID Magasin	0	200152	193460	0	0	0	0	
21	Nettoyage Jointure Magasin	0	193460	193460	0	0	0	0	
22	Tri Date Vente	0	193460	193460	0	0	0	0	
23	Jointure sur Date	0	194460	186063	0	0	0	0	
24	Nettoyage Jointure Date	0	186063	186063	0	0	0	0	
25	Tri ISBN	0	186063	186063	0	0	0	0	
26	Tri Dimension Livre	0	1443	1443	0	0	0	0	
27	Jointure sur ISBN	0	187506	144302	0	0	0	0	
28	Nettoyage Resultats	0	144302	144302	0	0	0	0	
29	Tri Clés	0	144302	144302	0	0	0	0	
30	Agrégation sans tri	0	144302	132889	0	0	0	0	
31	Insertion Table des Faits	0	132889	132889	0	132889	0	0	

FIGURE 5 – Constitution de la base de faits – aperçu sur les lignes traitées

6 Difficultés rencontrées

Durant cette réalisation, nous avons rencontré quelques problèmes, en particulier lorsqu'il s'agissait d'effectuer des conversions. *Pentaho* n'arrivait pas à convertir les ISBN ainsi que les dates malformées – c'est à dire les ISBN qui ne consistaient pas en 13 chiffres et les dates sous un autre format que le format `"yyyy-m-dd"`. Cela était dû au fait que les champs étaient considérés comme des tableaux d'octets plutôt que comme des chaînse de caractères. Cela s'est réglé facilement en désactivant la *conversion repoussée* – « *lazy conversion* » en anglais – lors de l'importation du fichier CSV `Fantastic`.

Aussi il a été impossible dans certains cas de changer le nom de certaines colonnes car elles n'étaient alors pas reconnues dans la suite. Si ce problème est anodin, il n'a pas permis de finir notre travail de clarification des dénominations que nous voulions parfaire avec le formatage des noms en*camelCase*. Nous avons aussi dû convertir les dates en chaînes de caractère pour effectuer les jointures car nous rencontrions des erreurs.

Enfin, il fut difficile de travailler sur nos machines personnelles pour plusieurs raisons en particulier par ce que certaines dépendances ne sont pas installées – ce qui est normal pour les drivers propriétaires pour les connexions aux bases de données Oracle – ou parce que certaines sont dépréciées et donc non disponibles sur les dépôts officiel de certaines distributions – c'est le cas de la librairie `webkitgtk` qui n'est plus supportée sous *Fedora* 27. Néanmoins, nous avons réussi à résoudre cela avec un peu de débrouillardise.

⋆ ⋆ ⋆