

Alimentation avec la Plateforme Pentaho - Kettle

Une entreprise qui regroupe des magasins de ventes d'ouvrages littéraires répartis sur toute la France souhaite faire une étude sur les ventes d'une année.

L'objectif est d'**analyser les performances des ventes des magasins par la mesure du nombre de produits vendus en fonction d'un certain nombre de critères tels que les produits, les dates, les magasins, les départements, la population et les ratios significatifs de ces mesures.**

Elle vous charge dans ce cadre de mettre en place une solution logicielle permettant d'intégrer les données pertinentes via un ETL et de pouvoir générer des rapports d'analyse ou des cubes via un logiciel de diffusion pour répondre aux besoins résumés ci-dessus.

Données disponibles

Afin de réaliser votre travail, l'entreprise vous met à disposition les données suivantes :

- **Catalogue des livres** : une base Oracle contient le catalogue complet de l'entreprise que chaque magasin a à sa disposition.
 - Cette base, composée d'une seule table publique catalogue, est disponible sur le serveur Oracle **sme-oracle.sme.utc**, sous le schéma **nf26**.
- **Fichier des ventes** : un fichier contient une consolidation de l'ensemble des ventes de l'année passée réalisées dans chaque magasin.
 - Ces données sont disponibles sous la forme d'un fichier CSV dans un répertoire du serveur **sme-oracle.sme.utc:/home/nf26/data**.
 - La structure du fichier est : numéro de ticket, date de ticket, produit, magasin.
- **Fichier des magasins** : un fichier ODS géré par la direction marketing contient pour chaque magasin l'organisation des rayonnages : **marketing.ods**.
 - Le responsable des ventes de chaque département décide de l'organisation des rayonnages des magasins de son département.
 - Il existe 3 types de rayonnage : par Auteur (A), par Année (Y), par Éditeur (E).
 - Le fichier est déposé dans un répertoire du serveur sme-oracle.sme.utc : **/home/nf26/fantastic**.
- **Données géographique sur les départements** :
 - Un stagiaire a trouvé sur Internet un fichier CSV permettant de connaître la population de chaque département, un peu daté mais qui pourra suffire : **departementsInsee2003.txt**

Vous devez intégrer dans vos transformations tous les tests nécessaires de façon à rejeter les données qui ne sont pas conformes pour un traitement complet. Il y aura donc autant de tables de rejets que de tests nécessaires qui pourront ensuite être analysées et remontées à l'entreprise pour réaliser les corrections éventuelles.

Les critères que les données doivent valider pour ne pas être rejetées sont les suivants :

- ISBN : 13 caractères,
- Date : non vide et mise au format yyyy-MM-dd,
- Magasin : M suivi de chiffres
- Vous transformerez la colonne auteurs en deux colonnes, l'une contenant le nom du premier auteur, la seconde contenant le reste des informations
- On souhaite pouvoir exploiter la date de publication et la langue des ouvrages pour générer des rapports d'analyse.

1 Modèle conceptuel de données

Vous repartirez du modèle réalisé lors du premier TD.

2 Création de l'alimentation du datawarehouse

À l'aide de la plateforme Pentaho Kettle vous allez réaliser l'alimentation du datawarehouse. Pour cela vous devez créer chacune des dimensions du datamart et la table de faits à l'aide de « transformations ».

2.1 Présentation du logiciel utilisé

Pentaho est une plate-forme décisionnelle open source complète composée d'une série d'outils associés à chaque étape de la Business Intelligence :

- Pentaho Data Integration Community Edition (PDI-CE, appelée également Kettle) est un outil ETL de la suite Pentaho CE (intégration de données) .
- Pentaho Analysis est un outil OLAP (Online Analytical Processing).
- Pentaho Dashboards et Pentaho Reporting sont des outils qui permettent de produire, respectivement, des tableaux de bords et des rapports.
- Pentaho Data Mining (basé sur Weka) enfin permet d'approfondir l'analyse exploratoire en s'appuyant sur les techniques de fouille de données.

Pentaho permet d'adresser deux typologies d'utilisateurs :

- Les utilisateurs de base, consommateurs d'indicateurs prédéfinis
- Les utilisateurs avancés qui ont besoin d'outils d'analyse et d'exploration.

La Community Edition (Pentaho CE) est téléchargeable librement. : <http://community.pentaho.com/>

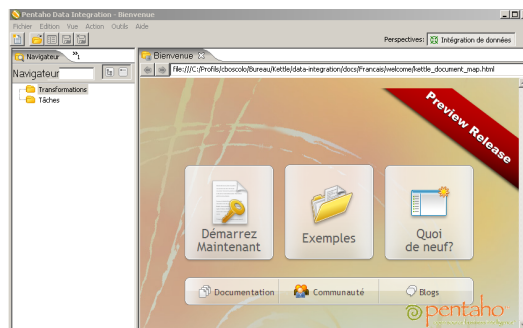
2.2 Lancement du logiciel

Depuis un terminal taper *pentaho*.

Fermer la fenêtre Firefox si elle s'ouvre.

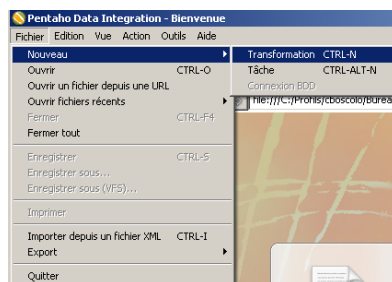
Nous n'utiliserons pas le repository, vous pouvez annuler la "Repository Connection".

Après avoir fermé "Astuces PDI..." vous arrivez sur la page d'accueil :



2.3 Création d'une transformation

Pour créer une nouvelle transformation, sélectionnez Fichier/Nouveau :



ou cliquez sur l'icône "Nouveau traitement"



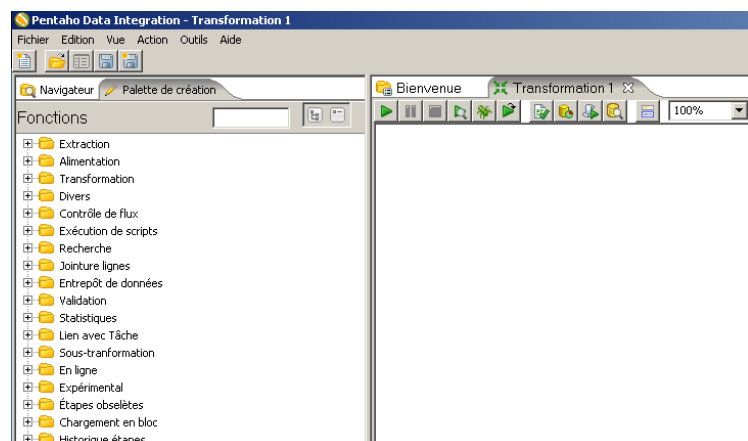
et choisissez



Transformation

La fenêtre principale prend un nouvel aspect :

- sur la gauche, dans la palette de création, vous disposez des outils de manipulation de données
- sur la droite un espace de travail vous permet de définir les séquences d'opérations sous forme de diagramme de traitements



2.4 Création d'une connexion à la base de données

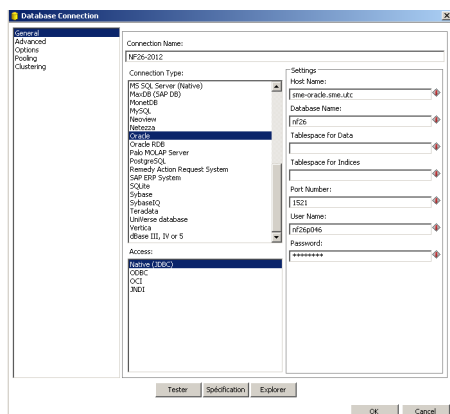
Depuis l'icône "Nouveau traitement"



créez une



Connexion base de données



- Nommez votre connexion
- Remplissez les paramètres de connexion :
 - **Host Name** : sme-oracle.sme.utc
 - **Database Name** : nf26
 - **Port Number** : 1521
 - **User Name** : nf26p46
 - **Password** : xxx
- Cliquez sur "Tester" puis OK.

Attention, une connexion n'est valable que pour une seule transformation.

2.5 Création de la dimension Magasins

A partir des deux fichiers « departements.txt » et « marketing.ods » vous allez créer la table « Magasins ».

2.5.1 Lecture du fichier « departements.txt »



Extraction depuis fichier CSV

Depuis le dossier *Extraction* de la Palette de création, glissez le composant « *Extraction depuis fichier CSV* » dans l'espace de travail. Double cliquez sur le composant.

Utilisez le bouton « **Parcourir** » pour sélectionner le fichier puis configurez le composant.

Séparateur champs	:
Champs mis entre	"
Taille tampon NIO	50000
Conversion de type repoussée	<input checked="" type="checkbox"/>
En-tête présent	<input type="checkbox"/>
Ajouter nom fichier au résultat	<input type="checkbox"/>
Champ N° ligne	
Traitement parallèle	<input type="checkbox"/>
New line possible in fields?	<input type="checkbox"/>
Encodage	UTF-8

^	#	Nom	Type	Format	Longueur
	1	Dpt	Integer	#	15
	2	DptName	String		23
	3	DptPop	Integer	#	15

Vérifiez vos données à l'aide du bouton « **Pré visualiser** » pour vérifier vos données.

2.5.2 Tri des départements



Tri lignes

Depuis le dossier *Transformation* de la Palette de création, glissez le composant « *Tri lignes* » dans l'espace de travail.

Reliez la sortie du composant « **Extraction depuis fichier CSV** » au composant « **Tri lignes** » en choisissant le lien « **Main** ».

Pour créer un lien entre les étapes, cliquez sur un des composants puis sur le lien nécessaire.

Les icônes correspondent à des opérateurs, les flèches qui les relient symbolisent les flux de données.

Configuration du tri des départements

Nommez l'étape (Par exemple : Tri dpt) et configurez le champ qui servira au tri.

Pour éviter les erreurs de saisie, utilisez le bouton **Récupérer champs** puis supprimez toutes les lignes inutiles.

^	#	Nom champ	Ascendant	Respecter la casse
	1	Dpt	O	N

2.5.3 Lecture du fichier « marketing.ods » et tri

Depuis le dossier **Extraction** de la Palette de création, glissez un nouveau composant « **Extraction depuis fichier MS Excel** » dans l'espace de travail. Sélectionnez le fichier dans l'onglet « Fichiers → Parcourir », puis cliquez sur « Ajouter ».

Configurez-le :

- dans l'onglet « Contenu » précisez la limite et le « Type de tableur »
- dans l'onglet « Champs » récupérez tous les champs et modifiez le type du n° de département en **Integer**.

^	#	Nom champ	Ascendant	Respecter la casse
	1	Dpt	O	N

Puis triez-le comme le fichier des départements.
Nommez l'étape (par exemple : Tri marketing).

2.5.4 Jointure des 2 tris



Jointure comparaison lignes


Depuis le dossier **Jointure lignes** de la Palette de création, glissez le composant « **Jointure comparaison lignes** » dans l'espace de travail.

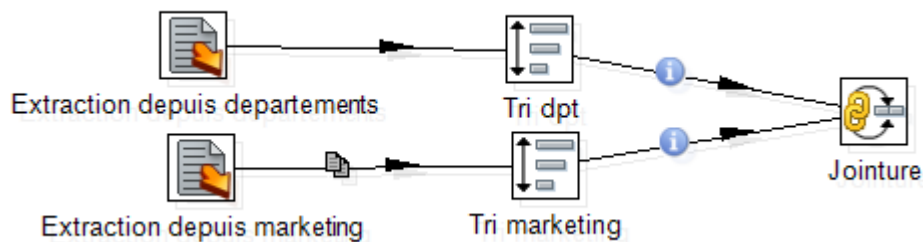
Reliez-le aux deux composants de tri et configurez-le.

#	Champ clé
1	DPT_No

#	champ clé
1	Dpt_No

2.5.5 Vérification du résultat obtenu

Exécutez votre transformation à l'aide du bouton  (**Exécuter cette transformation**).



Résultats exécution

Historique Trace Statistiques Performance

#	Nom étape	N°Copie	Lignes lues	Lignes écrites	Lignes en entrées
1	Extraction depuis departements	0	0	95	95
2	Tri dpt	0	95	95	0
3	Extraction depuis marketing	0	0	152	152
4	Tri marketing	0	152	152	0
5	Jointure	0	247	152	0

2.5.6 Sélection des champs



Altération structure flux

Avec un composant « *Altération structure flux* » sélectionnez les champs que vous désirez garder, en particulier ne gardez qu'un seul n° de département.

Altération structure flux

Nom étape: **Altération structure flux**

Sélectionner Retirer Méta-données

Champs

#	Nom champ	Renommer en	Longueur	Précision
1	Dpt			
2	DptName			
3	DptPop			
4	Mag			
5	Rayon			
6	RayonLibelle			
7	Best			
8	Recent			

Récupérer champs
Editer mapping

Inclure champs non spécifiés, ordonnées par nom ☐

OK Annuler

2.5.7 Création de la table de sortie



Insertion dans table

Depuis le dossier *Alimentation* de la Palette de création, glissez le composant « *Insertion dans table* » dans l'espace de travail.

Reliez-le au composant précédent et configurez-le, renseignez le schéma cible et la table cible (voir schéma page suivante). À chaque exécution les données sont insérées dans la table, pour vider la table, cochez « **Tronquer la table** ».

Insertion dans table

Nom étape: Insertion dans table

Connexion: nf26 [Editer...] [Nouvelle...] [Assistant...]

Schéma cible: NF26P055 [Parcourir...]

Table cible: Dim_Magasin [Parcourir...]

Valider transaction toutes les: 1000

☒ Tronquer la table

☒ Ignorer les erreur d'insertion

☐ Sélectionner champs

Général Champs table

Tables avec données partitionnées ☐

Champ partitionnement: [dropdown]

Partitionnement des données par mois ☒

Partitionnement données par jour ☐

Activer insertion groupée ☐

Nom table défini dans un champ ☐

Champ contenant le nom de la table: [dropdown]

Insérer champ nom table dans table ☒

Retourner une clé auto-générée ☐

Nom de la clé auto-générée: [dropdown]

[Help] [OK] [Annuler] [SQL]

Cliquez sur le bouton « SQL » pour créer la table puis sur « OK ».
Vous obtenez la nouvelle transformation ci-dessous, exécutez la.

Extraction depuis departements

Extraction depuis marketing

Tri dpt

Tri marketing

Jointure

Altération structure flux

P_dimension_magasin

Résultats exécution

Historique Trace Statistiques Performance

^	#	Nom étape	N°Copie	Lignes lues	Lignes écrites	Lignes en entrées	Lignes en sortie
	1	Tri dpt	0	95	95	0	0
	2	Jointure	0	247	152	0	0
	3	Tri marketing	0	152	152	0	0
	4	Altération structure flux	0	152	152	0	0
	5	P_dimension_magasin	0	152	152	0	152
	6	Extraction depuis departements	0	0	95	95	0
	7	Extraction depuis marketing	0	0	152	152	0

3 Création de la suite de votre datawarehouse.

Recommandation : utilisez une transformation différente pour chaque dimension.
Si vous créez une nouvelle transformation, il faut **penser à créer la connexion**.

3.1 Dimension date


Un exemple de génération de dates « **General - Populate date dimension.ktr** » se trouve avec d'autres exemples dans local2/data-integration/samples/transformations.

Vous pouvez utiliser cet exemple en l'ouvrant depuis Pentaho et en le modifiant pour obtenir la période et les champs dont vous avez besoin.

Pensez à la cohérence entre le format des clés étrangères de la table fait et celui des clés des dimensions.

Evitez les accents et les mots réservés, par exemple « Date », pour les noms de champ.

Vous obtenez la transformation suivante :



Résultats exécution

Historique | Trace | Statistiques | Performance

#	Nom étape	N°Copie	Lignes lues	Lignes écrites	Lignes en entrées	Lignes en sortie
1	365 days	0	0	365	0	0
2	Days_since	0	365	365	0	0
3	Calc Date	0	365	365	0	0
4	Select values	0	365	365	0	0
5	P_dimension_date	0	365	365	0	365

3.2 Créez la dernière dimension et la table des faits.

Attention : pour la table des faits, **limitez le nombre des lignes lues à 200 000 :**

Entrée fichier

Nom étape: Extraction depuis fichier

Fichier | Contenu | Gestion d'erreur | Filtres | Champs | Champs additionnels

Type fichier: CSV

Délimiteur: ;

Entouré par: "

Format: Unix

Encodage: UTF-8

Limite: 20000

Corriger dates: ☒

Format date locale: fr_FR

Nom fichier résultat:

Ajouter nom fichier au résultat: ☒

OK | Prévisualiser lignes | Annuler

Vous devez vous assurer pour la table de fait de la bonne correspondance des clés étrangères de la table des faits avec leurs correspondantes dans toutes les dimensions.

Alimentation avec la Plateforme Pentaho – Kettle

Présentation de quelques composants



Extraction depuis table

Dans le dossier *Extraction*.

Permet, par exemple, de lire la table Catalogue :

Extraction Table

Nom étape: Extraction depuis table

Connexion: NF26

SQL

```
SELECT  
  ISBN  
  , TITRE  
  , AUTEUR  
  , LANGUE  
  , PUBLICATION  
  , EDITEUR  
  , GENRE  
FROM NF26.CATALOGUE
```

Ligne 1 Colonne 0

Repousser conversion de type ☐

Remplacer les variables dans le script SQL ☐

Insérer données à partir de

Exécuter pour chaque ligne ☐

Limite: 0

OK Prévisualiser Annuler



Filtrage lignes

Dans le dossier *Contrôle de flux*.

Permet, par exemple, de vérifier que le champ Date est de la forme cccc-cc-cc :

Filtrage lignes

Nom étape: Vérification Dates

Envoyer les données 'VRAI' à l'étape: Vérification ISBN

Envoyer les données 'FAUX' à l'étape: Rejets Dates

Condition (VRAI)

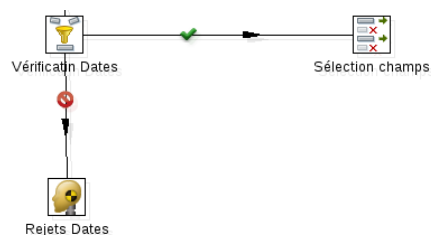
DAT REGEXP [0-9]{4}-[0-9]{2}-[0-9]{2}



Factice

Dans le dossier *Contrôle de flux*.

Cette étape peut être utile pour tester le résultat d'un composant, d'une transformation, tenir lieu de fichier de sortie.





Remplacer dans chaînes de caractères

Dans le dossier *Transformation*.

Permet, par exemple, de supprimer les chiffres du champ AUTHORS :

Remplacer dans chaînes de caractères

Nom étape

Champ à manipuler

^	#	Champ en entrée	Champ en sortie	RegEx	Rechercher	Remplacer par valeur
	1	AUTHORS		O	\d+-\d+,	



Décomposition Champs

Dans le dossier *Transformation*.

Permet, par exemple, de séparer en deux parties le champ AUTHORS :

Champ à décomposer

Nom étape

Champ à décomposer

Délimiteur

Champs

^	#	Nouveau champ	ID	Retirer ID?	Type	Longueur	Précision
	1	AUTEUR_PREMIER		N	String		
	2	AUTEURS_AUTRES		N	String		

OK Annuler



Agrégation valeurs

Dans le dossier *Transformation*.

Permet, par exemple, de calculer une quantité :

Agrégation

Nom étape

Inclure toutes les lignes ☐

Répertoire des fichiers temporaires Parcourir...

Préfixe fichier TMP

Ajouter numéro de ligne, initialiser à chaque vz ☐

Nom champ contenant numéro ligne

Toujours retourner une ligne ☐

Les champs de groupement:

^	#	Champ groupe
	1	MAG
	2	DAT
	3	ISBN

Récupérer champs

Agrégation :

^	#	Nouveau champ	Champ agrégé	Type
	1	QUANTITE	ISBN	Nombre de valeurs

Récupérer champs



Validation de données

Dans le dossier *Validation*.

Vérification des flux entrants grâce à la définition de règles

Suggestion d'expressions régulières de contrôle

Suppression des chiffres	<code>\d+ - \d+</code>
Normalisation de la date de publication	<code>T. +</code>
Vérification d'une date	<code>[0-9]{4} - [0-9]{2} - [0-9]{2}</code>
Vérification de la langue	<code>\w+</code>

Expressions régulières de contrôle pour la table des faits

Suppression des espaces	<code>(\s*)</code>
Vérification du n°ticket	<code>([0-9]{9})</code>
Vérification de la date	<code>[0-9]{4} - [0-9]{2} - [0-9]{2}</code>
Vérification du ISBN	<code>\d{13}\$</code>
Vérification de l'id_magasin	<code>M\d+</code>