

Architecture des différents jobs utilisés dans le processus de transformation et d’intégration des données commerciales

Projet Darties – Groupe2

Dossier ETL - Talend

**Création du document : Valentin BERNARD, Romain GIRARD, Sylvain LE QUANG**

**Date : 15/12/2010**

Dossier ETL - Talend

Projet Darties – Groupe2

# Contenu

[I. Contenu 1](#_Toc286674912)

[I. Introduction 2](#_Toc286674913)

[II. Aperçu global du projet 3](#_Toc286674914)

[III. Jobs de mise à jour du référentiel 4](#_Toc286674915)

[1. Jobs de suppressions du référentiel 5](#_Toc286674916)

[2. Jobs d’ajouts au référentiel 6](#_Toc286674917)

[3. Jobs de mises à jour du référentiel 6](#_Toc286674918)

[IV. Jobs de mise à jour annuelle 7](#_Toc286674919)

[V. Jobs de mise à jour mensuelle 8](#_Toc286674920)

# Introduction

Afin de mener à bien le projet commandé par le groupe Darties, nous avons développé un projet sous l’ETL libre Talend Open Studio. L’objectif était d’importer des données de fichiers Excel pour les insérer dans la base, après les vérifications et transformations de rigueur.

Nous allons passer en revue dans ce document l’architecture de notre projet, afin de permettre à de futurs développeurs d’intégrer le projet facilement afin de continuer à le maintenir et à le faire évoluer. Celui-ci consiste principalement en une arborescence de jobs que nous avons tenté de rendre aussi claire que possible.

# Aperçu global du projet

Dans l’état actuel, notre projet comporte 19 jobs, organisés selon une hiérarchie logique dont voici une copie :



En plus du job de contrôle de la connexion qui est appelé dans les autres procédures, nous voyons trois catégories principales de jobs :

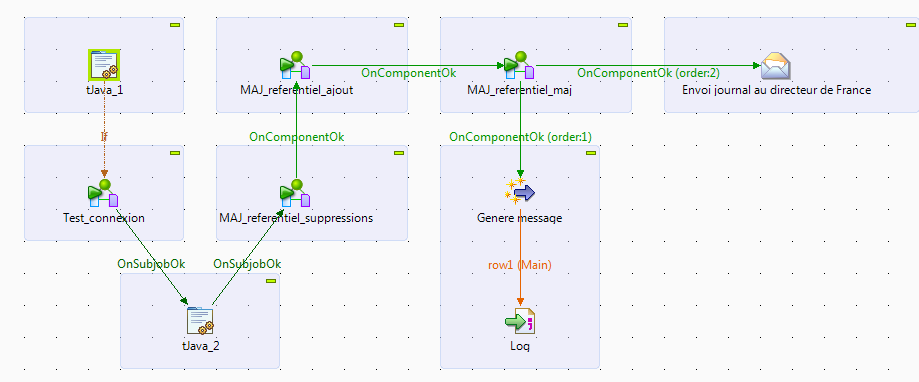
* Les jobs de mise à jour du référentiel ;
* Les jobs de mise à jour annuelle ;
* Les jobs de mise à jour mensuelle.

Nous allons passer en revue les jobs de chacune de ces catégories.

# Jobs de mise à jour du référentiel

Cette partie est constituée d’un job principal qui va appeler récursivement tous les autres jobs nécessaires à l’opération de mise à jour :

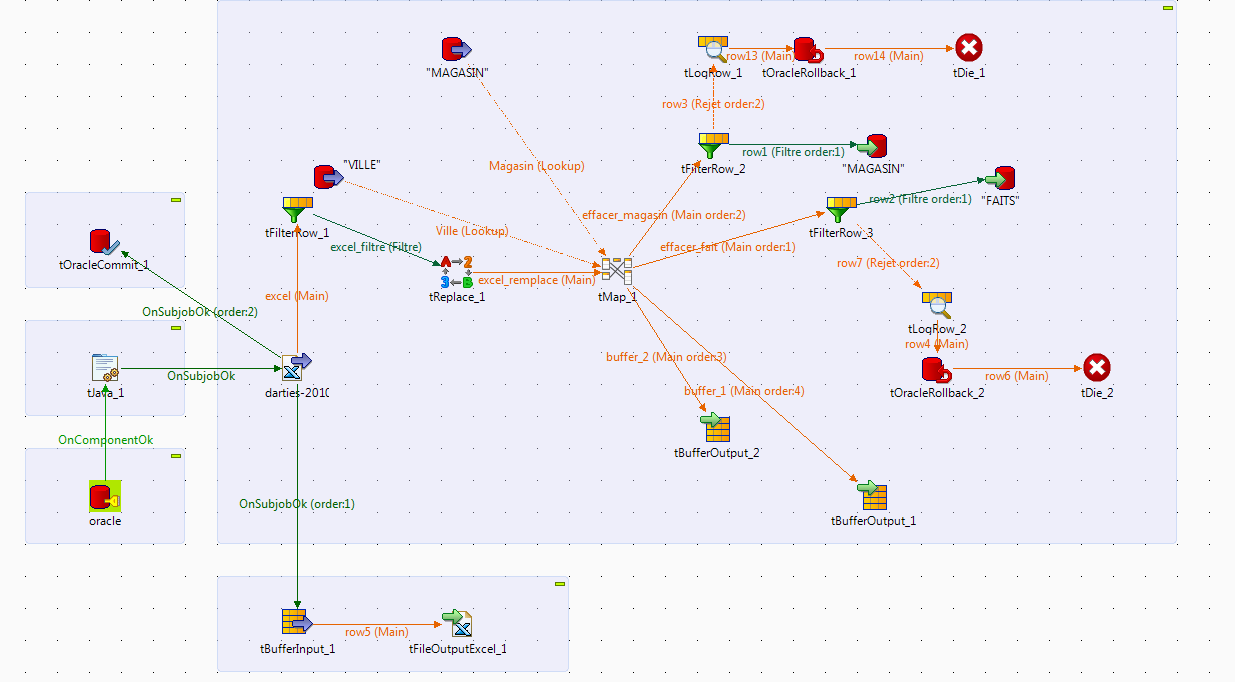
Le job MAJ\_referentiel se présente de la façon suivante sous Talend :



Le composant tJava du début sert simplement à appeler la routine de vérification du fichier Excel. Le job continue uniquement si cette routine ne détecte pas d’erreur. Le composant Test\_connexion vérifie la connexion à la base de données. Le tJava suivant affiche uniquement un message de statut.

Ensuite, les trois sous-jobs principaux sont appelés, afin d’effectuer respectivement les suppressions, ajouts et mises à jour requis. Enfin, les fichiers de logs sont générés et un e-mail est envoyé au directeur commercial.

## Jobs de suppressions du référentiel

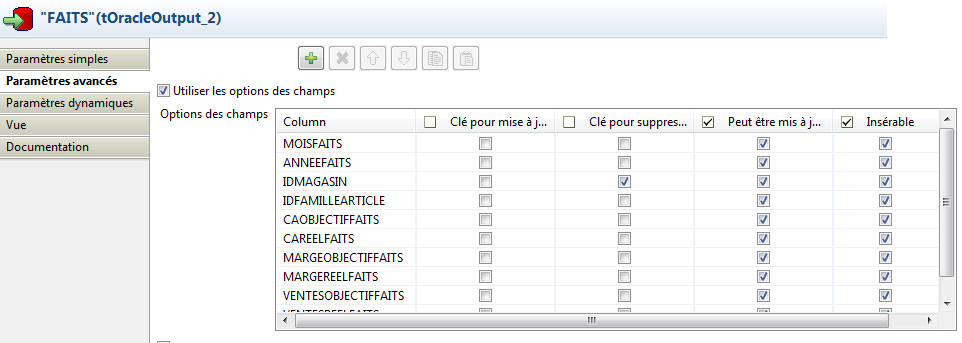
Ce job unique se présente de la façon suivante :

Le job commence par une ouverture de connexion (composant « oracle ») et par un affichage de message de statut.

Les données sont ensuite lues dans le fichier Excel, filtrées par un composant tFilterRow afin de ne conserver que les lignes dont l’action est « S » (« Supprimer »), légèrement transformées afin de transformer les underscores en espaces (composant tReplace), puis un mapping est effectué afin de récupérer l’identifiant du magasin qui correspond à la ville dont le nom a été donné.

Par la suite, tous les faits qui correspondent au magasin à supprimer sont eux-mêmes supprimés de la base, afin de garantir son intégrité référentielle. Enfin, le magasin lui-même est supprimé.

À noter que l’on veut que Talend effectue la suppression des faits sur la base de la colonne IDMAGASIN, alors que la clé primaire de cette table est normalement la combinaison (IDMAGASIN, IDFAMILLEARTICLE, ANNEEFAITS, MOISFAITS). Afin d’obtenir ce comportement, il nous faut préciser que la colonne IDMAGASIN est la seule clé de suppression :



À noter qu’avant la suppression proprement dite, un composant tFilter vérifie que le magasin à supprimer existait bien (IDMAGASIN différent de 0). Dans le cas contraire, un message d’erreur est généré et l’ensemble des opérations sont annulées (composant tOracleRollback).

## Jobs d’ajouts au référentiel

Le job principal d’ajouts au référentiel consiste principalement à appeler les cinq sous-jobs suivants :

* MAJ\_referentiel\_ajout\_continents\_devises\_enseignes, qui met à jour les données ne dépendant d’aucune autre (continents, devises et enseignes) ;
* MAJ\_referentiel\_ajout\_pays, qui met à jour les pays (qui dépendent des continents et des devises) ;
* MAJ\_referentiel\_ajout\_regions, qui met à jour les régions commerciales (qui dépendent des pays) ;
* MAJ\_referentiel\_ajout\_villes, qui met à jour les villes (qui dépendent des régions commerciales) ;
* MAJ\_referentiel\_ajout\_magasins, qui met à jour les magasins (qui dépendent des villes et des enseignes).

Ces jobs reposent sur la même structure, relativement simple ; seul le dernier job, qui ajoute les magasins, est plus complexe, puisqu’il nécessite la prise en compte de plus de cas d’erreurs.

## Jobs de mises à jour du référentiel

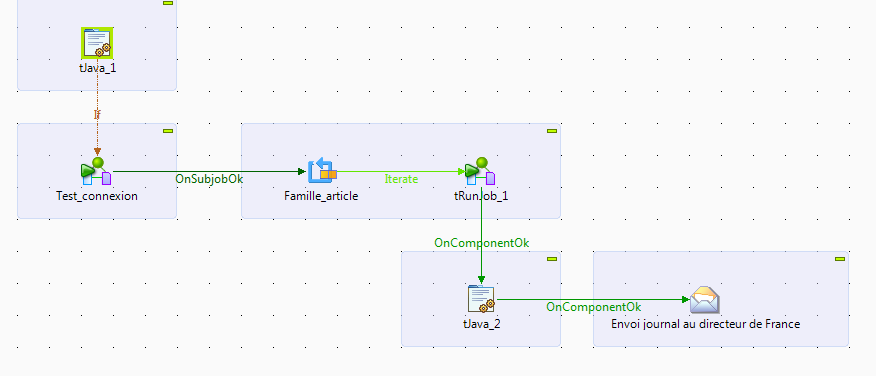
Les jobs de mises à jour du référentiel sont très proches de ceux d’ajouts : les principales différences tiennent au fait qu’ils s’occupent des lignes dont l’action à effectuer est « M » (« Mise à jour ») plutôt que « A » (« Ajout »), et qu’ils ne peuvent que mettre à jour des données existantes (la mise à jour d’une donnée inexistante provoque une erreur fatale).

Les sous-jobs sont les suivants :

* MAJ\_referentiel\_maj\_continents\_devises\_enseignes ;
* MAJ\_referentiel\_maj\_pays ;
* MAJ\_referentiel\_maj\_regions ;
* MAJ\_referentiel\_maj\_villes ;
* MAJ\_referentiel\_maj\_magasins.

# Jobs de mise à jour annuelle

Le job de mise à jour annuelle va appeler un sous-job pour chaque catégorie de famille d’articles possible. Il se présente de la façon suivante sous Talend :

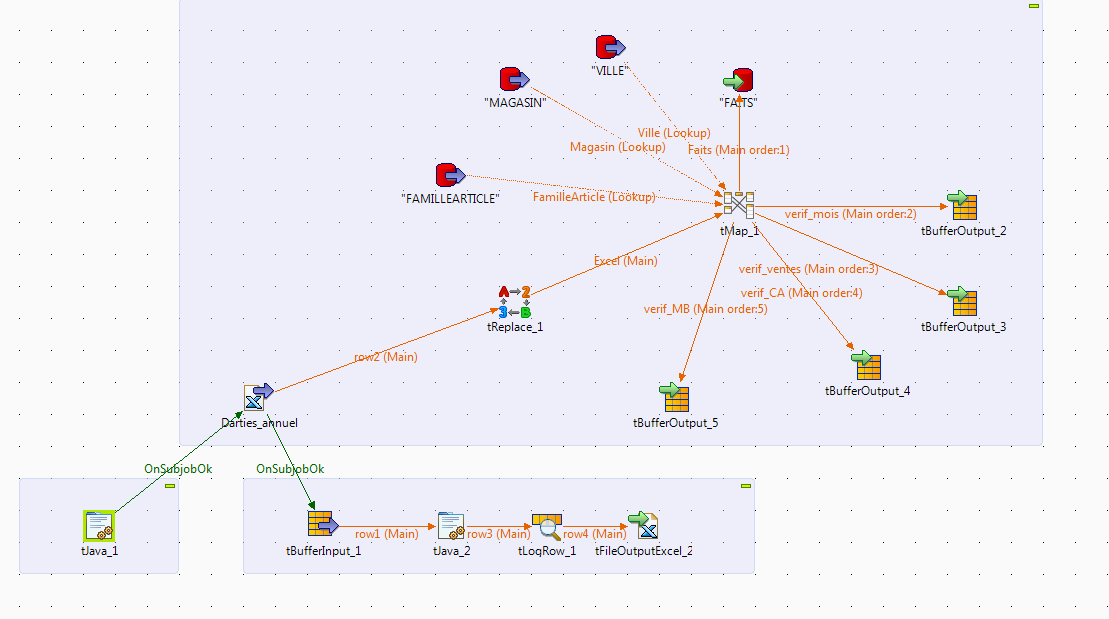


Le premier composant tJava appelle la routine de vérification du fichier Excel. Le test de connexion est ensuite effectué.

Le composant tForeach boucle sur chaque famille d’articles : à chaque itération, un composant tRunJob appelle le job sous-jacent sur lequel nous allons revenir.

Enfin, un fichier de log est généré puis envoyé au directeur commercial.

Le sous-job Maj\_Annuelle\_Famille\_article se base sur les données de contexte qui ont été définies par le job appelant ; celles-ci correspondent au nom de la famille d’articles à mettre à jour.



Le job consiste principalement en un gros mapping des différentes données, et une gestion d’erreurs plus souple que sur les jobs de mise à jour du référentiel (celles-ci ne sont plus rédhibitoires).

# Jobs de mise à jour mensuelle

Les jobs de mise à jour mensuelle sont très proches des jobs de mise à jour annuelle : la différence tient dans le fait que ces jobs ne cherchent plus les colonnes O\_Ventes, O\_CA et O\_MB mais R\_Ventes, R\_CA et R\_MB, puisque les données à mettre à jour sont maintenant les données réelles et non prévisionnelles.

De la même façon que précédemment, un sous-job est appelé pour chaque famille d’articles, puis un rapport est envoyé au directeur commercial.