TP N°6 bis : Intégration des données avec Pentaho

Vous travaillez chez une société de transport. Vous recevez donc des commandes de transport de la part de vos clients et votre rôle est d'attribuer ces commandes à des transporteurs.

Un nouveau logiciel a été implanté dans la société pour suivre les commandes en temps réel et on souhaite donc pouvoir comparer les horaires de chargement et de livraison demandés par le client et les horaires réels obtenus grâce au nouveau logiciel. Ceci permettra d'évaluer les différents retards au niveau des usines (retard pour la préparation de la commande par exemple) mais également ceux qui concernent le transporteur (retard à la livraison par exemple).

Vous avez deux fichiers à disposition :

1) « orders.csv » : contient l'ensemble des commandes après affrètement, c'est-à-dire qu'il contient l'ensemble des commandes envoyées par le client ainsi que le transporteur que vous avez attribué à chaque commande. Les champs dans ce fichier sont les suivants :
   - OrderNumber : numéro de commande
   - LoadingPlace : ville de chargement
   - DeliveryPlace : ville de livraison
   - Carrier : le transporteur chargé d'effectuer le transport
   - LoadingDate : date de chargement demandée par le client
   - ETL (Estimated Time of Loading) : heure de chargement demandée par le client
   - DeliveryDate : date de livraison demandée par le client
   - ETA (Estimated Time of Arrival) : heure de livraison demandée par le client
2) « realtime.csv » : contient les commandes exploitées dans le logiciel de suivi. On y retrouve donc certaines commandes avec des horaires réels (pas toujours complets). Les champs dans ce fichier sont les suivants :
   - OrderNumber : numéro de commande
   - RealLoadingDate : date réelle de chargement
   - RealArrivalAtPlaceOfLoading : heure réelle de chargement
   - RealDeliveryDate : date réelle de livraison
   - RealArrivalAtPlaceOfDelivery : heure réelle de livraison

On veut procéder à la jointure des deux fichiers CSV en utilisant Pentaho Data Integrator. Pour cela, nous devons procéder en deux étapes car une jointure nécessite d'avoir trié les données au préalable.

Les étapes à suivre sont les suivantes :

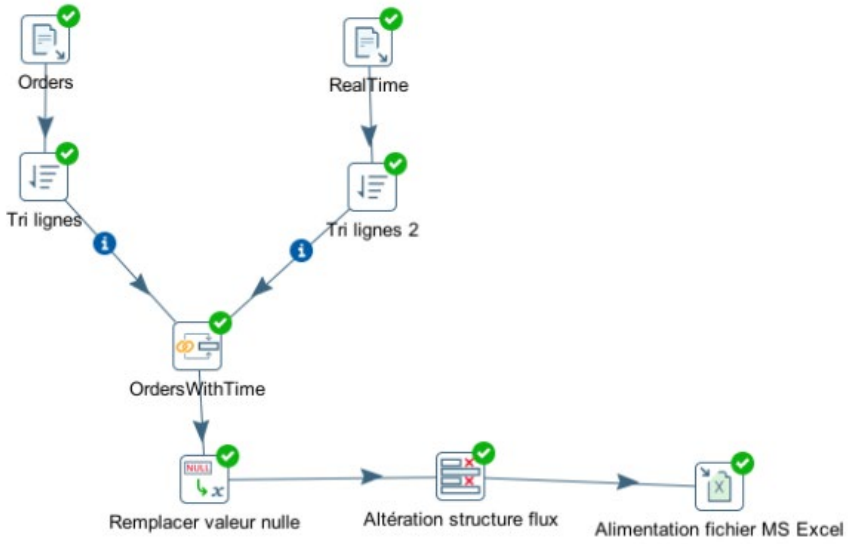
1) Créez une nouvelle transformation.
2) Ajoutez les deux fichiers sources et vérifier les champs dans les fichiers.
3) Faire un tri sur les deux fichiers (le tri est un préalable indispensable à la jointure). Configurer le tri pour qu'il se fasse sur le champ OrderNumber.
4) Faire la jointure. Choisissez le type de jointure « Left Outer ».

Pour l'analyse future de nos données, l'entreprise a souhaité que les horaires manquants soient considérés comme une "vraie" information.

Cela veut dire qu'on ne veut pas de valeurs nulles dans notre table mais plutôt une information du type "données inconnues" lorsqu'il manque une date ou un horaire. Donc pour continuer :

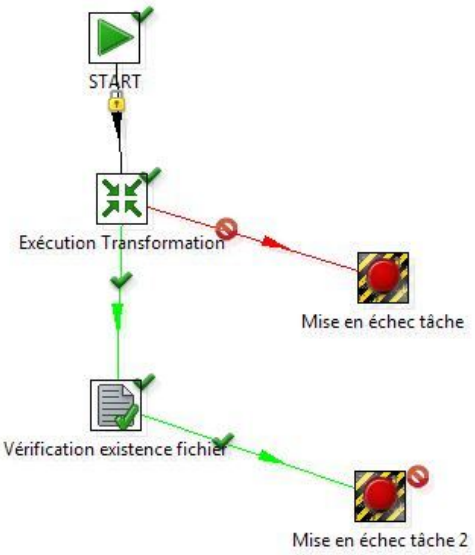
5) Remplacez les valeurs nulles qui concernent les dates et les horaires réels avec une information de type String.
6) Supprimez la colonne OrderNumber_1 de votre flux de données.
7) Exportez votre résultat dans un fichier Excel.
8) Exécutez votre transformation.

Une fois que la modélisation de votre transformation est terminée, vous devriez avoir le résultat suivant (cela dépend de la version de l'outil) :



Pour compléter votre travail, on vous demande de renvoyer un message en cas d'erreur et de vérifier que le fichier de sortie est bien créé en fin de transformation.

9) Créer une nouvelle tâche (**Attention, il s'agit d'une tâche et non d'une transformation**).
10) Glisser-déposer dans la partie droite un élément « Start », un élément « Exécution Transformation » et un élément « Vérification existence fichier ». Relier tous ces éléments dans l'ordre Start, Transformation, Vérification.
11) Insérer deux éléments « Mise en échec tâche ». Relier le premier avec l'étape de transformation et le second avec la vérification. Vous devriez avoir le résultat suivant :



Les liens n'ont pas tous la même couleur :

- Un lien bleu avec un cadenas indique que l'élément suivant sera toujours exécuté.
- Un lien rouge indique que l'élément suivant ne sera exécuté que s'il y a eu une erreur dans l'exécution de l'élément précédent.

- Un lien vert indique que l'élément suivant ne sera exécuté que si l'élément précédent s'est terminé avec succès.

12) On veut mettre la tâche en échec dans le cas où la transformation échoue d'une part mais également lorsque le fichier généré par la transformation n'existe pas. Établir les types de liens afin de répondre à ce besoin.
13) Configurer les éléments :
    a. étape de transformation ;
    b. fichier dont on veut vérifier l'existence ;
    c. messages d'erreur.
14) Exécuter la tâche.