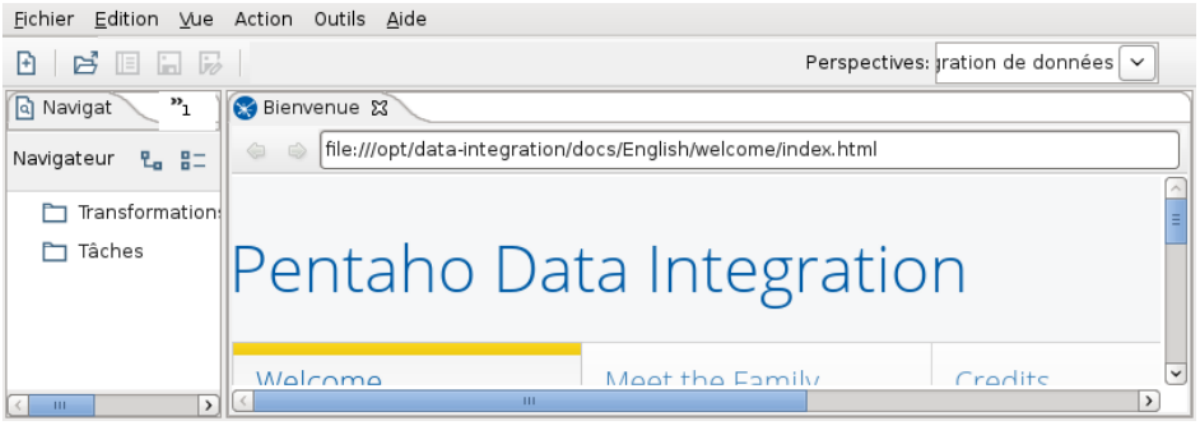
BI M2DAC – TP1 ETL

Configuration Pentaho PDI – uniquement en salle 14-15 307 :

- Disponible sous /usr/local/data-integration/
- Lancer Pentaho PDI : sh spoon.sh



Créer des ETL avec Pentaho PDI :

- Fichier → Nouveau → Transformation
- Vous avez un catalogue de chargement/transformation/exportation de données dans l'onglet « Palette de création »
- Pensez à sauvegarder vos transformations (nous en aurons besoin en fin de TME)

Exercice 1 :

- Générer 10 lignes avec un champs *test* de valeur = 1 – (**générer lignes**)
- Rajouter une colonne *gender* de valeur *male* (**Ajout constantes**)
- Prévisualiser le résultat -> (Bouton droit -> prévisualiser puis fenêtre du bas : prévisualiser)
- Rajouter une colonne *cola* dans la génération de lignes de valeur *test*
- Rajouter un checksum après la génération de ligne (**Ajout checksum**)
  - De type CRC32
  - Récupérer les champs en entrée
  - Le champ de sortie sera *checksum*
- Numérotez les lignes en utilisant **Ajout séquence**
- Exporter le résultat dans un fichier XML **(XML Output)**

Exercice 2 :

- Créez un fichier CSV contenant une colonne *cola* dont les 10 premières valeurs valent 1 et les 10 suivantes valent 2
- Numérotez les lignes dans un champs *index*
- Numérotez les lignes dont cola vaut 1 et les lignes dont cola vaut 2 indépendamment à l'aide de **Ajout séquence ré-initialisable** dans un compteur *colb*
- Rajouter une colonne à l'entrée de valeur *cola/colb* grâce à un **Calculateur** en utilisant la formule *[cola]/[colb]*

Exercice 3 :

- Importer le fichier noms-prenom.csv
- A l'aide de **Manipulation de chaînes de caractères,** écrire les noms en majuscule, les prénoms en minuscule et les pays avec la première lettre en majuscule
- Créer une troisième colonne qui contient les 3 premières lettres du nom de famille à l'aide de **Calculateur** ainsi que de **Extraction depuis chaînes de caractères**
- Trier par ordre lexicographique sur le nom
- Sauvegarder la sortie dans un fichier CSV
- Sauvegarder la sortie dans un fichier Excel et ouvrez le fichier

Exercice 4 :

- Téléchargez le fichier **Mobiliers de Stationnement** (*http://opendata.paris.fr)* au format CSV
- Créer un fichier qui ne contient que les coordonnées x et y des stations Velib
- Sauvegarder la sortie au format JSON

Exercice 5 :

Soit le fichier titanic.csv (le récupérer sur kaggle)

- Compter le nombre de survivants pour toutes les combinaisons des champs de classe et sexe
- Calculer la fréquence des survivants pour ces mêmes combinaisons

Exercice 6 :

Récupérer des données depuis http://opendata.paris.fr/api/records/1.0/search?dataset=stations-velib-disponibilites-en-temps-reel

- Pour cela, il faut créer une colonne url contenant l'URL
- Utiliser l'icône **Client REST**
- Sauvegardez la sortie dans un fichier

Exercice 7 :

- Importer le fichier noms-prenom.csv
- Calculer la moyenne d'âge par pays - Exportez le fichier afin de pouvoir l'utiliser dans Google Map : https ://developers.google.com/chart/interactive/docs/gallery/geochart

Exercice 8 :

Il existe de nombreuses autres fonctionnalités. Faites quelques tests sur les fonctionnalités suivantes :

- Lister les noms des sous-répertoires
- Lister les fichiers .txt dans un répertoires (joker= « .*txt »)
- Dédoublonner des valeurs (transformation – dédoublonnage)
- Répartir les données en fonction de valeurs : « Filtrages lignes » permet de séparer en deux sous-ensembles et « Branchement conditionnel » permet de construire plusieurs sous-ensembles

Exercice 9 :

On va s'intéresser à créer des jobs qui lancent successivement différentes transformations.

- Ouvrir une nouvelle tâche
- Définir un début de lancement (START – pas de planification)
- Appeler deux transformations créées dans les exercices précédents : (Exécution Transformation)
- Rajouter un délai d'attente de 10 secondes entre les deux transformations (Mise en place Temporisation)

Il est aussi possible d'exécuter plusieurs tâches en même temps.

- Créer une tâche (tâche 1) qui exécute la transformation de l'exercice 3
- Créer une tâche (tâche 2) qui :
    - Supprime le fichier créé par l'exercice 3
    - Attend qu'il réapparaisse (Attente apparition fichier)
    - Lance l'exercice 1 ensuite
- Exécuter la tâche 2, attendre 10 secondes, observer…
- Exécuter la tâche 1, observer ce qui se passe pour la tâche 2