TP 2 – Intégration des données

Exercice 1

Ce TP décrit un Job d'intégration de données simple lisant des données relatives à des films, à partir d'un fichier CSV stocké localement et affiche les données dans la console de la vue Run. Au cours de ce TP, vous allez :

1. Créer un Job d'intégration de données.
2. Ajouter et relier des composants dans un Job d'intégration de données.
3. Créer une métadonnée de fichier dans le Repository.
4. Configurer et exécuter un Job d'intégration de données.

Créer le Job

Cette procédure décrit comment créer un dossier de Jobs nommé TP2 et un Job nommé movies, dans ce dossier.

Suivez les étapes ci-dessous pour créer un dossier nommé TP2:

1. Dans la vue Repository, cliquez-droit sur le nœud Job Designs et sélectionnez Create folder dans le menu contextuel.
2. Dans l'assistant [New Folder], nommez votre dossier de Jobs test puis cliquez sur Finish pour créer votre dossier.

Suivez les étapes ci-dessous pour créer un Job nommé movies dans le dossier TP2 :

1. Cliquez-droit sur le dossier TP2 et sélectionnez Create job dans le menu contextuel.
2. Dans l'assistant [New Job], saisissez un nom pour le Job à créer, ainsi que d'autres informations utiles. Saisissez movies dans le champ Name.
3. Cliquez sur Finish pour créer votre Job. Un Job vide s'ouvre dans le Studio.

Déposer et relier des composants

Cette procédure décrit comment ajouter et relier des composants dans le nouveau Job créé, pour lire un fichier CSV et afficher les données dans la console.

1. Déposez un **tFileInputDelimited** et un **tLogRow** de la Palette dans l'espace de modélisation graphique. Vous pouvez trouver le composant **tFileInputDelimited** dans le groupe Input de la famille **File** et le **tLogRow** dans la famille Logs & Errors, dans la Palette.
2. Cliquez sur le composant **tFileInputDelimited**, une icône représentant un **o** s'affiche, glissez-déposez l'icône **o** sur le composant **tLogRow**. Les deux composants sont reliés via un lien **Row > Main**.

Préparer la métadonnée relative aux films

Cette procédure décrit comment configurer la métadonnée du fichier source movies.csv dans le Repository. Les métadonnées stockées dans le référentiel peuvent être utilisées dans

plusieurs Jobs, vous permettant ainsi de configurer rapidement vos Jobs sans avoir à définir chaque paramètre et schéma manuellement.

Votre fichier source movies.csv doit être disponible dans le dossier C:….\
\workspace\PROJET_MMED\process\TP2

1. Dans la vue Repository, développez le nœud **Metadata**, cliquez-droit sur File **delimited** et sélectionnez **Create file delimited** dans le menu contextuel pour ouvrir l'assistant [New Delimited File].
2. Dans l'assistant [New Delimited File], saisissez un nom pour la métadonnée du fichier, movies, dans cet exemple et d'autres informations utiles permettant de décrire votre métadonnée, puis cliquez **Next** pour passer à l'étape suivante et définir les propriétés générales du fichier. Dans cette étape de l'assistant, **Name** est le seul champ obligatoire. Les informations fournies dans le champ Description s'affichent en tant qu'info-bulle lorsque vous placez votre curseur sur la métadonnée.
3. Dans le champ **File**, spécifiez le chemin du fichier source, ou cliquez sur **Browse** pour parcourir votre système jusqu'à ce fichier. La zone **File Viewer** affiche un aperçu du fichier, vous permettant de vérifier sa cohérence, la présence d'un en-tête et la structure du fichier.
4. Dans la liste **Format**, sélectionnez votre système d'exploitation et cliquez sur **Next** pour parser le fichier.
5. Dans l'onglet **Preview**, cochez la case **Set heading row as column names** pour récupérer les noms de colonnes de la première ligne, puis cliquez sur **Refresh Preview**. La case **Header** de la zone **Rows To Skip** est automatiquement cochée et le nombre de lignes d'en-tête à ignorer est incrémenté de **1**. Si le fichier contient plusieurs lignes d'en-tête devant être ignorées lors du parsage du fichier, spécifiez le nombre de lignes, dans ce champ, puis cliquez sur **Refresh Preview**.
6. Cliquez sur **Next** pour récupérer le schéma du fichier. La table Description of the Schema affiche le schéma généré du fichier.
7. Nommez le schéma movies_schema, vérifiez-le et modifiez-le selon vos besoins. Dans cet exemple, augmentez la valeur dans la colonne Length pour les lignes title et url.
8. Cliquez sur **Finish** pour valider le schéma et fermer l'assistant. La métadonnée de fichier créée s'affiche dans la vue **Repository**.

Configurer et exécuter votre Job

Cette procédure décrit comment configurer les composants à l'aide de la métadonnée créée dans la procédure précédente et comment exécuter votre Job.

1. Dans la vue **Repository**, double-cliquez sur le Job movies pour l'ouvrir dans l'espace de modélisation graphique. Vous pouvez ignorer cette étape si le Job est déjà ouvert dans l'espace de modélisation graphique.
2. Dans la vue **Repository**, développez **Metadata > File delimited** et glissez-déposez la métadonnée de fichier movies ou son schéma movies_schema sur le composant **tFileInputDelimited** dans l'espace de modélisation graphique. Lorsqu'une fenêtre vous propose de propager les modifications au composant de sortie, cliquez sur **Yes**. Dans l'onglet **Basic settings** de la vue **Component**, vous

pouvez voir que tous les paramètres du composant ont été automatiquement renseignés.
3. Double-cliquez sur le **tLogRow** pour ouvrir sa vue **Basic settings**.
4. Dans la zone **Mode**, sélectionnez l'option **Vertical (each row is a key/value list)** pour une meilleure lisibilité dans la console de la vue Run.
5. Appuyez sur **F6** ou cliquez sur le bouton **Run** de la vue **Run** pour exécuter votre Job. La console de la vue Run affiche les informations relatives aux films lues depuis le fichier source.

Exercice 2

Ce TP décrit comment filtrer le flux de données et obtenir uniquement les données des films ayant des informations valides relatives aux réalisateurs. Ce TP vous explique comment :

1. Dupliquer un Job.
2. Ajouter un composant en saisissant son nom sur un lien ou dans l'espace de modélisation graphique.
3. Déposer une métadonnée ou son schéma en tant que composant dans l'espace de modélisation graphique.
4. Effectuer un traitement simple sur des flux de données à l'aide du tMap.

Préparer la métadonnée relative aux réalisateurs

Cette procédure présente comment configurer la métadonnée du fichier de référence *directors.txt* dans le **Repository**. Cette métadonnée sera utilisée pour ajouter et configurer l'entrée de référence dans le TP.

Votre fichier source directors.txt doit être disponible dans le dossier C:….\ \workspace\PROJET_MMED\process\TP2

9. Dans la vue Repository, développez le nœud **Metadata**, cliquez-droit sur File **delimited** et sélectionnez **Create file delimited** dans le menu contextuel pour ouvrir l'assistant [New Delimited File].
10. Dans l'assistant [New Delimited File], saisissez un nom pour la métadonnée du fichier, directors.txt, dans cet exemple et d'autres informations utiles permettant de décrire votre métadonnée, puis cliquez **Next** pour passer à l'étape suivante et définir les propriétés générales du fichier.
11. Dans le champ **File**, spécifiez le chemin du fichier source, ou cliquez sur **Browse** pour parcourir votre système jusqu'à ce fichier.
12. La zone **File Viewer** affiche un aperçu du fichier, vous permettant de vérifier sa cohérence, la présence d'un en-tête et la structure du fichier.
13. Dans la liste **Format**, sélectionnez **Windows** et cliquez sur **Next** pour parser le fichier.
14. Dans la liste **Field Separator** de la zone **File Settings**, sélectionnez **Comma**.
15. Cliquez sur **Next** pour récupérer le schéma du fichier. La table Description of the Schema affiche le schéma généré du fichier.

16. Nommez le schéma *directors_schema* et renommez les colonnes en *directorID* et *directorName*, respectivement, puis modifiez le type de données de la colonne *directorID* d'**Integer** à **String**.
17. Cliquez sur **Finish** pour valider le schéma et fermer l'assistant. La métadonnée de fichier créée s'affiche dans la vue **Repository**.

Dupliquer le Job existant

Cette procédure vous présente comment créer un Job à partir d'un Job existant.

6. Dans la vue **Repository**, cliquez-droit sur le Job nommé *movies* et sélectionnez **Duplicate** dans le menu contextuel.
7. Dans la boîte de dialogue **[Duplicate]**, saisissez un nom pour le Job, *filter_movies* dans cet exemple et cliquez sur **OK** pour valider la création du Job et fermer la boîte de dialogue.

Ajouter un composant de mapping

La procédure ci-dessous présente comment ajouter un composant de mapping en saisissant le nom du composant directement sur le lien existant.

1. Dans le nouveau Job *filter_movies*, sélectionnez le lien **Row** reliant le **tFileInputDelimited** et le **tLogRow** et saisissez le nom du composant **tMap** ou une partie de son nom.
2. Double-cliquez sur le **tMap** dans la liste pour l'ajouter sur le lien. Le nouveau composant **tMap** est relié au composant d'entrée. Une boîte de dialogue s'ouvre et vous demande de saisir un nom pour le nouveau lien de sortie.
3. Saisissez un nom pour ce lien de sortie, *Valid_movies* dans cet exemple, puis cliquez sur **OK**. Lorsqu'il vous est proposé de propager le schéma d'entrée au composant cible de sortie, cliquez sur **Yes**. Le **tMap** est ajouté au Job et relié aux deux composants existants à l'aide d'un lien **Row** > **Main**.

Ajouter un composant lookup

La procédure ci-dessous vous explique comment ajouter un composant lookup (de référence) depuis le référentiel **Repository**, le relier au **tMap** et activer l'option supprimant les espaces.

1. Dans la vue **Repository**, développez **Metadata** > **File delimited**, glissez-déposez la métadonnée *directors* ou son schéma *directors_schema* dans l'espace de modélisation graphique. La boîte de dialogue **[Components]** s'ouvre, affichant une liste de composants que vous pouvez ajouter au Job à partir de cette métadonnée.
2. Sélectionnez le **tFileInputDelimited** et cliquez sur **OK**.
   Un **tFileInputDelimited** nommé *directors* est ajouté à l'espace de modélisation graphique et ses paramètres simples (onglet **Basic settings**) sont automatiquement renseignés.
3. Cliquez-droit sur le nouveau **tFileInputDelimited**, sélectionnez **Row** > **Main** dans le menu contextuel et cliquez sur le **tMap**. Le **tFileInputDelimited** est relié au **tMap** à l'aide d'un lien **Lookup**.
4. Dans l'onglet **Advanced settings** du nouveau **tFileInputDelimited** et cochez la case **Trim all columns**. Certains enregistrements du fichier d'entrée de

référence *directors.txt* contiennent des espaces blancs en début de champ. Cette option vous permet de supprimer ces espaces blancs du flux de référence lorsque le Job est exécuté.

Votre Job contient tous les composants nécessaires pour filtrer les informations relatives aux films. Vous allez ensuite configurer le mapping dans le composant **tMap** afin de filtrer le flux d'entrée principal par rapport au flux de référence et écrire en sortie les informations souhaitées.

Configurer le mapping et exécuter le Job

La procédure ci-dessous vous apprend à configurer les mappings et les jointures Inner Join pour écrire en sortie les informations relatives aux films ayant un ID de réalisateur valide.

1. Double-cliquez sur le composant **tMap** pour ouvrir son éditeur de mapping. L'éditeur de mapping affiche trois tables, nommées *row1*, *row2* et *Valid_movies* dans cet exemple, correspondant respectivement au schéma du fichier des films, au schéma du fichier des réalisateurs et au schéma de sortie des informations valides. Les colonnes de la table *row1* sont déjà mappées aux colonnes de la table *Valid_movies*.
2. Sélectionnez la colonne *directorID* de la table *row1* et glissez-la sur la colonne *directorID* dans la table *row2* afin de créer une jointure entre les deux ensembles de données basée sur l'ID des réalisateurs.
3. Cliquez sur le bouton **tMap settings** (représentant une clé anglaise), cliquez sur le champ **Value** pour **Join Model**, puis cliquez sur le bouton **[...]** qui s'affiche pour ouvrir la boîte de dialogue **[Options]**. Dans la boîte de dialogue, sélectionnez **Inner Join** et cliquez sur **OK** pour définir la jointure comme Inner Join. Grâce à ce paramètre, seuls les enregistrements de films dont l'ID du réalisateur correspond à ceux du fichier de référence seront passés au composant de sortie.
4. Dans la zone **Schema editor** au bas de l'éditeur de mapping, sélectionnez la colonne *directorID* du schéma de sortie, *Valid_movies* dans cet exemple et cliquez sur le bouton **[x]** afin de la supprimer.
5. Cliquez sur le bouton **[+]** sous la table de sortie pour ajouter une colonne, nommez-la *directedBy*, configurez sa longueur **Length** à *20*, puis déplacez-la pour la placer entre *title* et *releaseYear*.
6. Sélectionnez la colonne *directorName* de la table *row2* et glissez-la dans le champ **Expression** correspondant à la colonne *directedBy* dans la table de sortie. Un nouveau mapping est créé entre la table de référence et la table de sortie.
7. Cliquez sur **OK** pour valider les mappings et fermer l'éditeur, puis cliquez sur **Yes** lorsqu'il vous est proposé de propager les modifications. La configuration des mappings est sauvegardée et le schéma de sortie est synchronisé au composant de sortie **tLogRow**.
8. Appuyez sur **F6** ou cliquez sur le bouton **Run** de la vue **Run** pour exécuter le Job.

Seuls les enregistrements de films ayant des informations valides relatives aux réalisateurs sont affichées dans la console de la vue **Run**.

À partir du scénario décrit dans TP3, ce TP agrandit le Job afin de rassembler les données des films dans lesquelles manquent les informations des réalisateurs et afin d'écrire les données valides et invalides dans une base de données MySQL.

Exercice 4 :

Ce TP montre :

- Comment ajouter un composant en saisissant son nom dans l'espace de modélisation graphique ou en le glissant depuis un composant existant.
- Comment configurer les mappings pour les informations rejetées dans le **tMap**.
- Comment configurer les sorties de base de données.

Ajouter des composants de sortie de base de données à votre Job

Dans l'exemple ci-dessous, vous allez créer un nouveau Job à partir du Job **filter_movies** et ajouter deux composants **tMysqlOutput**. Ces composants seront utilisés pour écrire les informations des films traitées dans les tables de base de données spécifiées.

1. Créez un nouveau Job en dupliquant le Job créé dans le TP précédent et nommez le nouveau Job *write_movies_to_db*, puis double-cliquez sur le Job pour l'ouvrir dans l'espace de modélisation graphique.
2. Cliquez-droit sur le composant **tLogRow** et sélectionnez **Delete** dans le menu contextuel pour le supprimer.
3. Cliquez à l'ancien emplacement du **tLogRow** dans l'espace de modélisation graphique et saisissez le nom du **tMysqlOutput** ou une partie de celui-ci puis, sélectionnez et double-cliquez sur le **tMysqlOutput** dans la liste pour l'ajouter dans l'espace de modélisation graphique. Lorsque vous commencez à saisir un nom de composant, une liste de composants correspondant à votre recherche s'affiche. Vous pouvez en sélectionner un pour voir sa description, à côté de la liste.
4. Cliquez-droit sur le composant **tMap**, sélectionnez **Row** > **Valid_movies** dans le menu contextuel et cliquez sur le **tMysqlOutput** pour le relier au **tMap**. Le nom de la connexion **Valid_movies** correspond au nom de la table de sortie existante dans le **tMap**.
5. Cliquez sur le composant **tMap** et glissez-déposez l'icône **o** sur l'espace de modélisation graphique. Un champ textuel et une liste de composants suggérés s'affichent. Vous pouvez en sélectionner un pour voir sa description, à côté de la liste.
6. Dans le champ textuel, saisissez le nom du **tMysqlOutput**, sélectionnez le composant dans la liste et appuyez sur **Entrée** pour ajouter un autre composant **tMysqlOutput** dans l'espace de modélisation graphique. Une boîte de dialogue s'ouvre et vous demande de saisir un nom pour la connexion de sortie.
7. Dans la boîte de dialogue, saisissez *Invalid_movies* et cliquez sur **OK** pour relier le **tMap** au deuxième composant **tMysqlOutput**.

Vous avez ajouté et connecté les composants de sortie de base de données nécessaires pour écrire les informations des films traitées dans une base de données MySQL. Maintenant, vous devez configurer de nouveaux mappings dans le **tMap** et les paramètres de base de données dans les composants **tMysqlOutput**.

Configurer le mapping pour les données rejetées

La procédure ci-dessous vous explique comment configurer les mappings pour rassembler les informations rejetées.

1. Double-cliquez sur le composant **tMap** pour ouvrir l'éditeur **Map Editor**. Une deuxième table de sortie nommée *Invalid_movies* a été automatiquement créée.

2. Déposez les colonnes *movieID* et *title* à partir de la table *row1* vers la table *Invalid_movies*.
3. Cliquez sur le bouton **tMap settings** de la table *Invalid_movies* puis sur le champ **Value** pour **Catch lookup inner join reject** et cliquez sur le bouton **[...]** qui s'affiche pour ouvrir la boîte de dialogue **[Options]**. Dans la boîte de dialogue, sélectionnez **true** et cliquez sur **OK**. Grâce à ce paramètre, les enregistrements sans l'ID du réalisateur ou avec un ID ne correspondant pas à ceux du fichier de référence seront passés au composant de sortie.
4. Cliquez sur **OK** pour valider les mappings et fermez l'éditeur **Map Editor** puis, cliquez sur **Yes** lorsqu'il vous est proposé de propager les modifications. La configuration des mappings est sauvegardée. Le schéma de sortie et le composant de sortie sont synchronisés.

Vous avez configuré les mappings pour les sorties rejetées. Maintenant, vous devez configurer les composants de sortie pour écrire les flux de sortie des tables de base de données.

Configurer les sorties de base de données MySQL

La procédure ci-dessous vous explique comment configurer les composants de sortie de base de données pour écrire les informations des films des tables de base de données MySQL.

1. Double-cliquez sur le premier composant **tMysqlOutput** pour ouvrir sa vue **Component**.
2. Fournissez les détails de connexion nécessaires pour accéder à votre base de données, à savoir le nom d'hôte ou l'adresse IP, le numéro de port, le nom de la base de données, le nom et le mot de passe de l'utilisateur dans les champs correspondants.
3. Dans le champ **Table**, saisissez le nom de la table de base de données cible. Dans cet exemple, la table relative aux informations de films valides est *valid_movies*.
4. Dans les listes **Action on table** et **Action on data**, sélectionnez l'option répondant à vos besoins. Dans cet exemple, vous pouvez d'abord supprimer la table si elle existe déjà, en créer une nouvelle, vide et utilisez l'option par défaut de la liste **Action on data**.
5. Dans l'onglet **Basic settings** du deuxième composant **tMysqlOutput**, utilisez les mêmes paramètres que dans le premier **tMysqlOutput** sauf pour le nom de la table de base de données cible. Dans cet exemple, la table relative aux informations de films invalides est *invalid_movies*.
6. Appuyez sur **F6** ou cliquez sur le bouton **Run** de la vue **Run** pour exécuter votre Job.

Les enregistrements de films dont l'ID du réalisateur est valide sont sauvegardés dans la table de base de données nommée *valid_movies* et ceux dont l'ID du réalisateur est invalide sont sauvegardés dans la table de base de données nommée *invalid_movies*.