Compte Rendu
Qualité des données

Koussaila HAMMOUCHE, Said MOHAMMED SEGHIR,
Rachida OUCHENE, Lylia TOUAZI,
Mohamed Amokrane SELMI.

Décembre 2020

Superviseur : Mme Zoubida KEDAD-COINTOT

Année : 2020-2021

1 Introduction

2 Scénario

Le scénario choisi est basé sur la base de données du site IMDb https ://www.imdb.com/interfaces/ ainsi que sur les films et séries disponibles sur Netflix, nous souhaitons créer une nouvelle base qui peut regrouper tous les films qui ne sont pas disponibles sur Netflix et répondre aux besoins des utilisateurs.

2.1 Les Sources

- Source 1 :
  - **Acteurs.csv**(nconst, primaryName, birthYear, deathYear ,#knownForTitles, email, age).
  - **FilmSource1.csv**(tconst, titleType, originalTitle, isAdult, startYear, endYear, runtimeMinutes, genres).
  - **NoteFilm.xml**(#tconst, averagerating, numVotes).
- Source 2 :
  - **FilmNetflix.json**(show_id, type, title, director, cast, country, date_added, release_year, rating, duration, listed_in).
- Source 3
  - **FilmSource3.csv**(tconst, titleType, primaryTitle, isAdult, startYear, endYear, runtTimeSecondes, genres).

2.2 Objectif

L'objectif principal est d'avoir un catalogue de films non disponibles sur Netflix, ainsi que plusieurs autres possibilités :

- Rechercher des films bien notés.
- Trouver les films où un acteur particulier a joué.
- Rechercher tous les films sortis en une année donnée.
- Rechercher des films par genre.

2.3 Les Cibles

- **Film**(IdFilm, Title, Annee, runtimeMin, Note).
- **Acteur**(IdActeur, Nom, BirthYear, Age, Email).
- **Genres**(NomGenre).
- **Posseder**(#NomGenre,#IdFilm).
- **Jouer**(#IdFilm, #IdActeur).

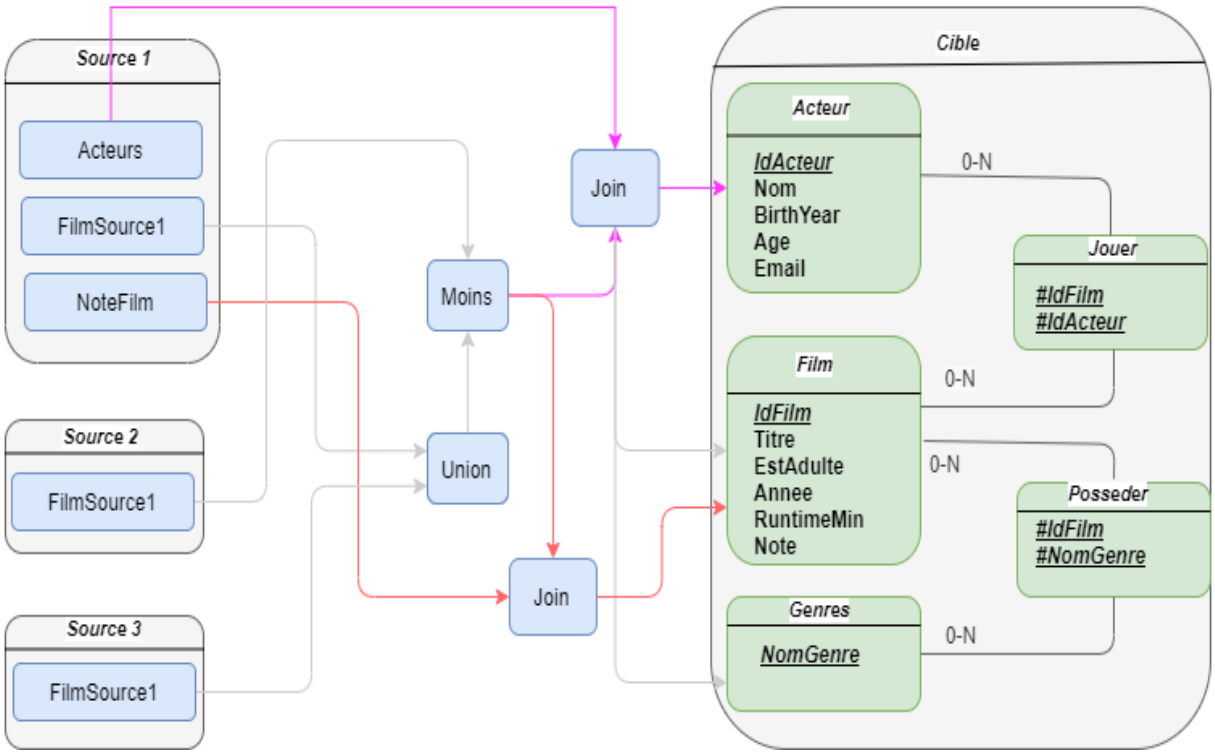


FIGURE 1 – Processus de construction des tables cible.

3 Facteurs de qualité

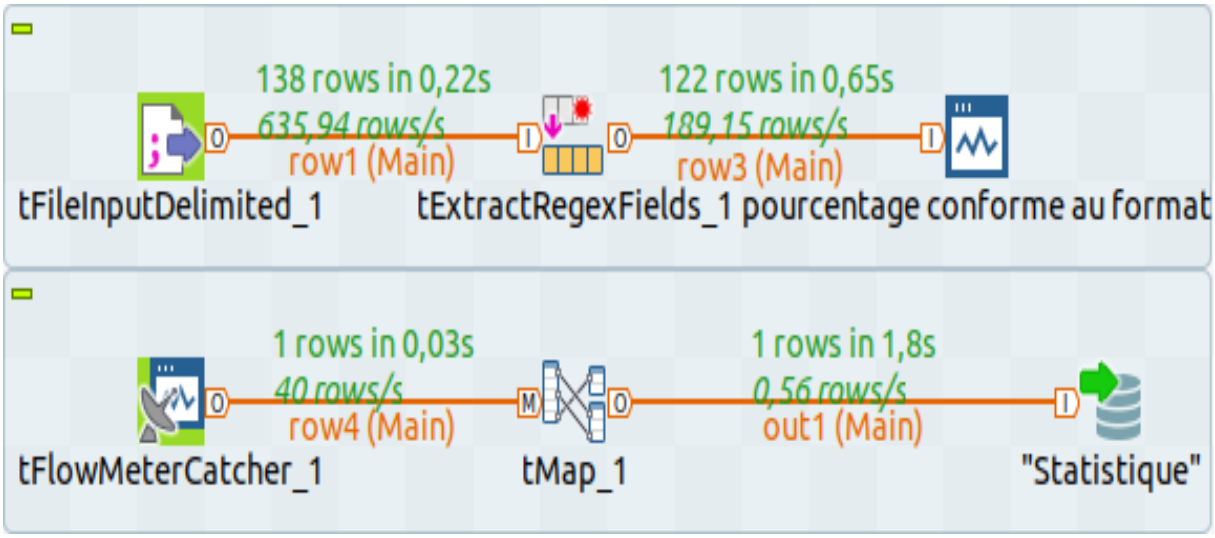
3.1 Conformité à un format, une codification

1. **Niveau** : Colonne email.
2. **Source** : Fichier Acteurs de la source 1.
3. **Métrique** : Le taux des adresses mail fausses dans la colonne.

3.1.1 Méthode de détection

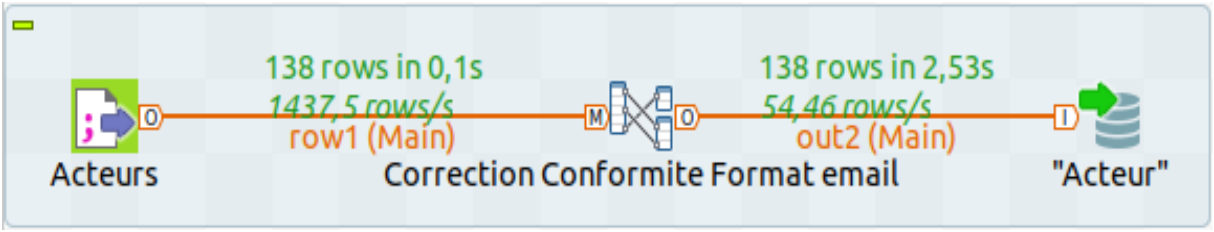
Certains champs ne respectent pas le format standard d'une adresse email donc on utilise le tExtractRegexFields pour les détecter à l'aide de l'expression régulière suivante :
"$[a-zA-Z][a-zA-Z0-9-\_.]+@[a-z]+.[a-z]$".

On stocke dans la table **Facteur_Qualite** le pourcentage des adresses email conformes.



3.2 Processus d'amélioration

On crée une routine qui permet de détecter les emails non conformes au format, puis on les remplace par Acteurs.Nom@gmail.com

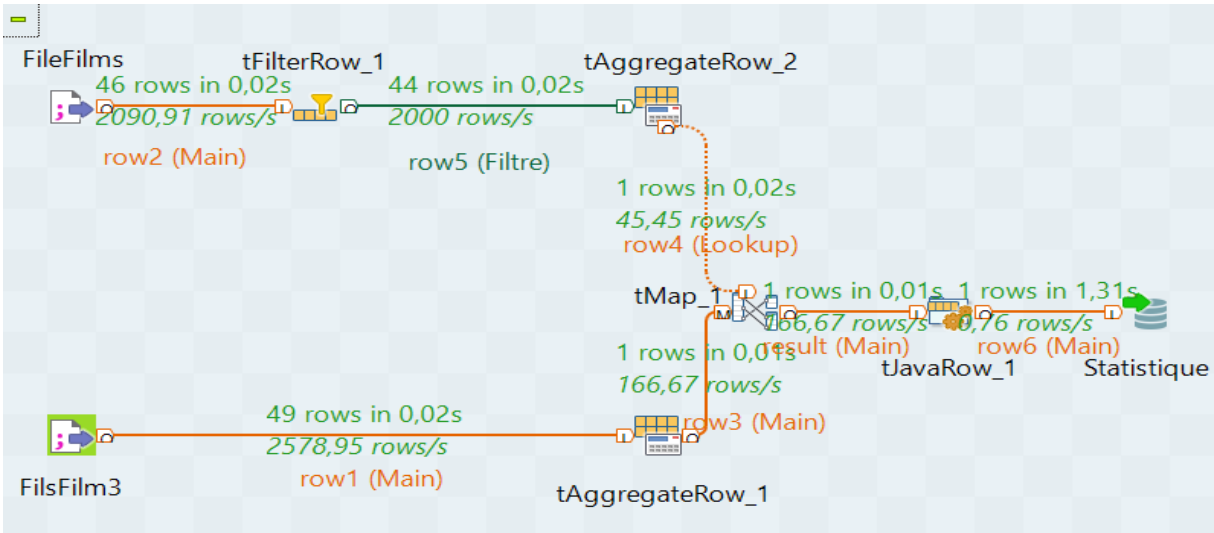


3.3 Hétérogénéité des échelles et de la granularité

1. **Niveau** : Colonne runTimeMinutes et Attribut runTimeSecondes.
2. **Source** : Source 1 FilmSource1 et source 3 FilmSource3.
3. **Métrique** : Boolean vrai si les valeurs ont la même échelle sinon faux.

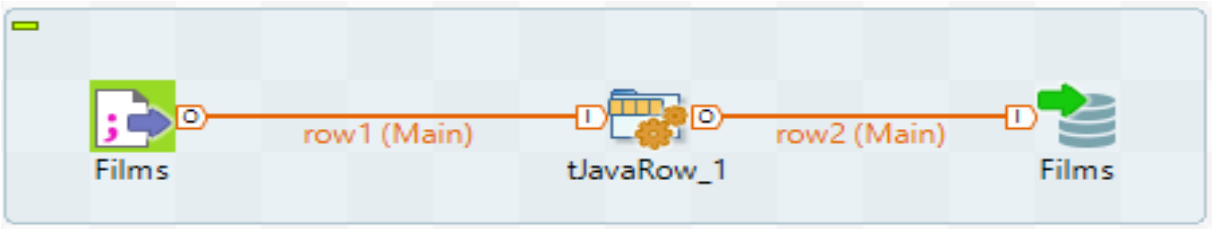
3.3.1 Méthode de détection

On calcule la moyenne sur la colonne RunTime des deux tables FilmSource1 et FilmSource3 puis on compare les résultats : Si le résultat de AVG(FilmSource1.runTimeSecondes) DIV AVG(FilmSource3.runTimeMinutes) $>$ 30 alors on peut conclure que la granularité des deux sources est différente.



3.3.2 Processus d'amélioration

La base données d'évaluation garde le facteur de granularité et la table qui n'est pas conforme au facteur (minutes). On récupère la table non conforme au facteur. Ensuite on divise sa colonne Runtime par le facteur de granularité.(on change le nom de l'attribut runTimeSecondes par runTimeMin).

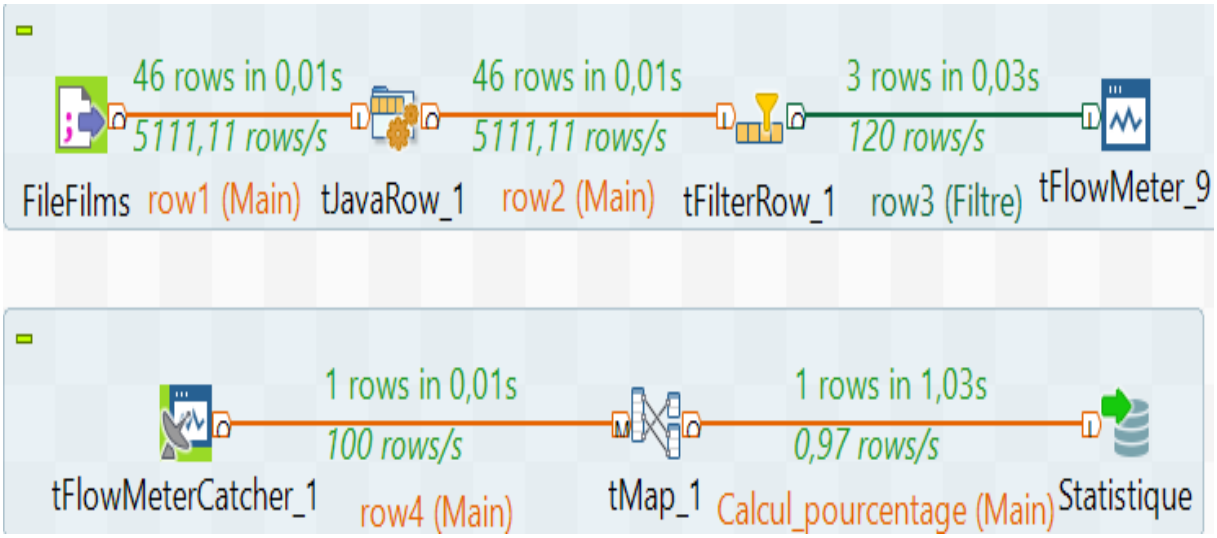


3.4 Complétude des données

1. **Complétude Tuples :**
   (a) **Niveau** : Table.
   (b) **Source** : Source 1 Table FilmSource1.
   (c) **Métrique** : Le pourcentage des tuples null.

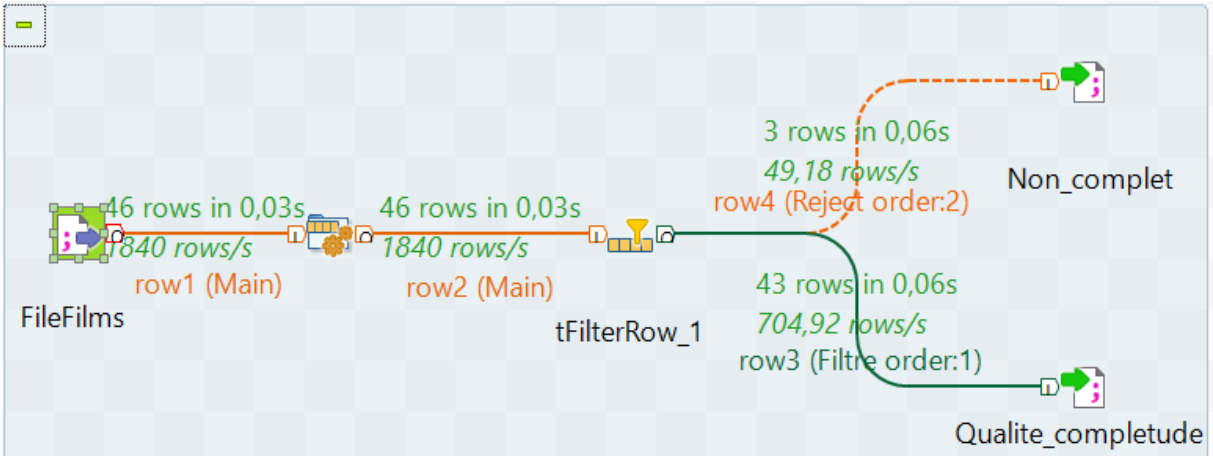
3.4.1 Méthode de détection

D'abord, on calcule le pourcentage des attributs null dans chaque tuple, et pour cela on calcule le nombre d'attributs null, ensuite on divise par le nombre total d'attributs dans le tuple. Si le résultat est supérieur à 50%, le tuple est considéré incomplet. A la fin on calcule le pourcentage des tuples considérés incomplets dans le fichier **FilmSource1.csv**.



3.4.2 Processus d'amélioration

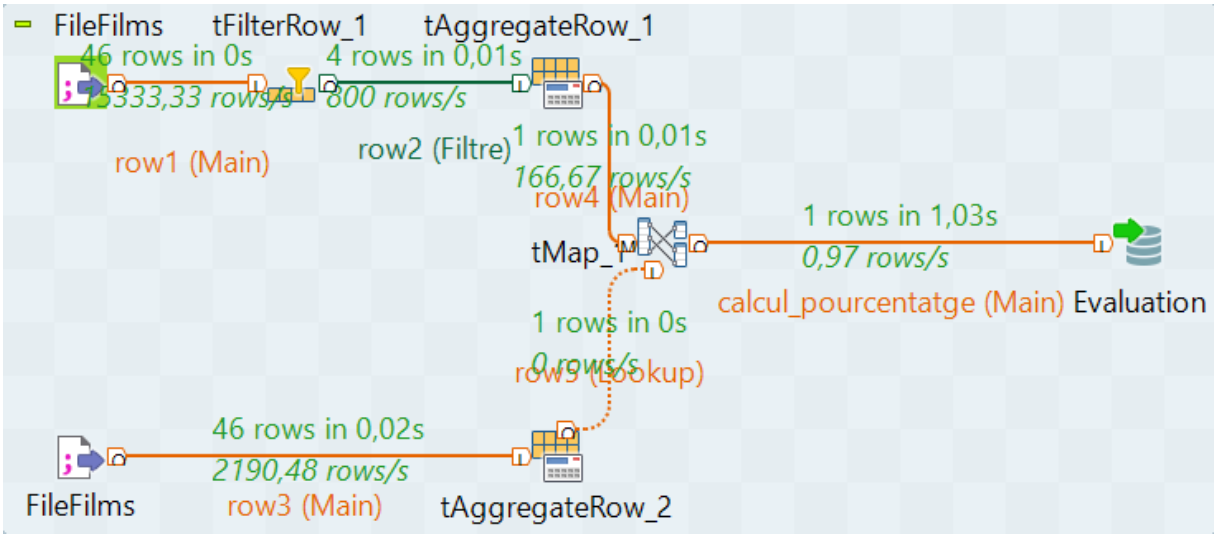
Nous avons décidé de mettre les tuples qui sont considérés incomplèts dans une base de données différente de celle de la cible.



2. **Complétude Colonne :**
   (a) **Niveau** : Colonne.
   (b) **Source** : Colonne "originalTitle" de la Table FilmSource1.
   (c) **Métrique** : Le pourcentage des colonnes null.

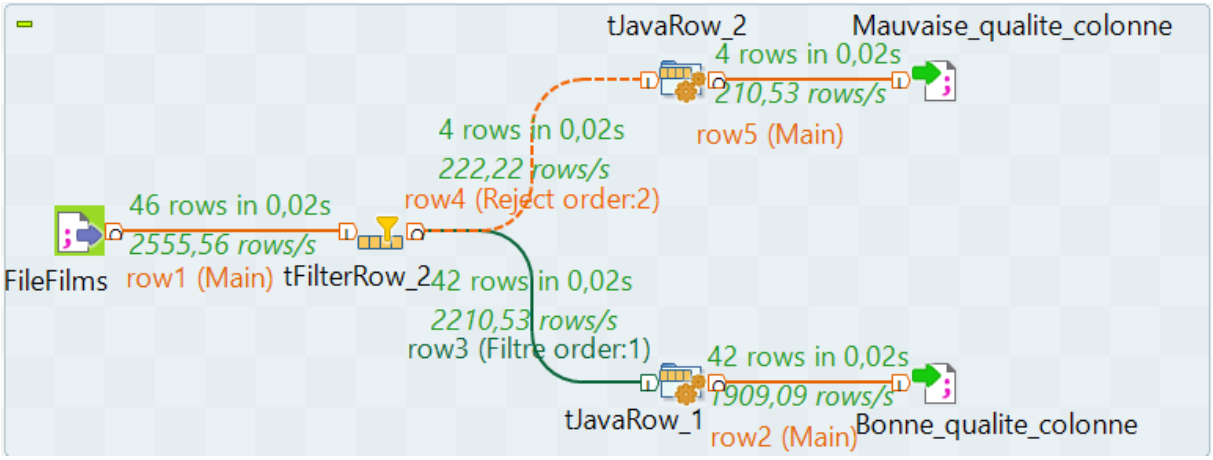
3.4.3 Méthode de détection

Pour calculer le pourcentage des colonnes null, on compte le nombre de valeurs non renseignées pour la colonne "originalTitle" et on divise à la fin sur le nombre total de tuples.



3.4.4 Processus d'amélioration

Comme le titre du film est un attribut important dans la table "Film", nous avons décidé de mettre les tuples qui n'ont pas de titre dans une base de données différente de celle de la cible.

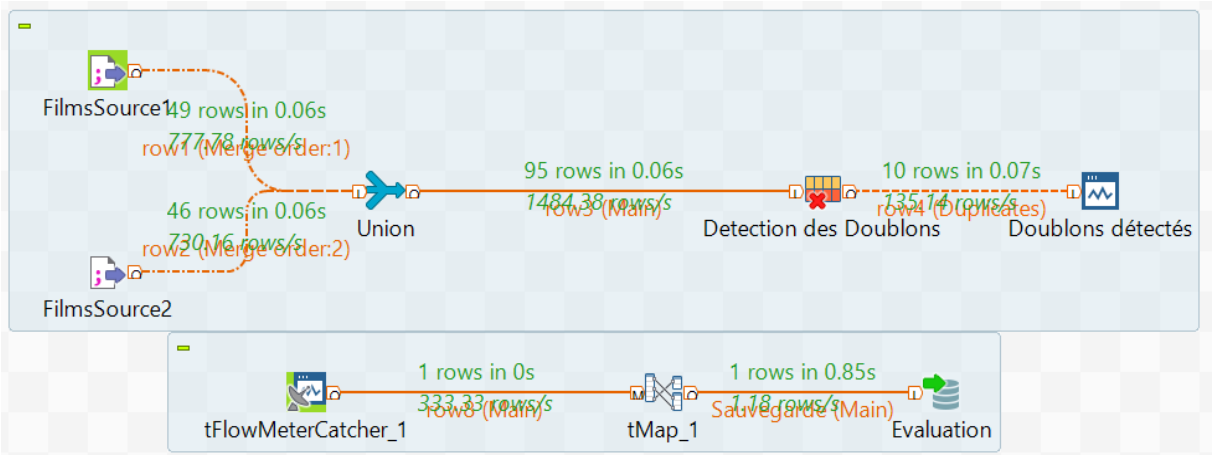


3.5 Détection et élimination des doublons

1. **Niveau** : Colonne Title.
2. **Source** : FilmSource1 et FilmSource3.
3. **Métrique** : Proportion des doublons parmi tous les films.

3.5.1 Méthode de détection

Après avoir fait l'union des sources 1 et 3, les films ayant le même titre sont considérés comme des doublons.



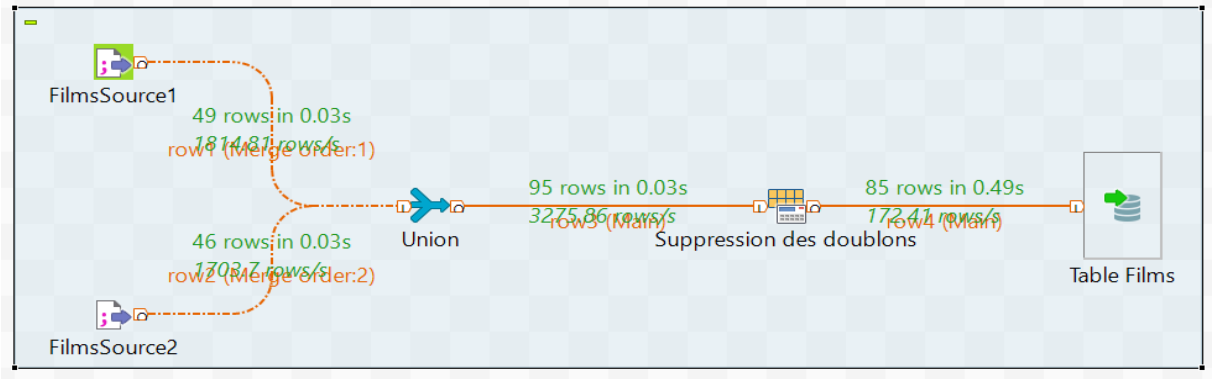
3.5.2 Processus d'amélioration

Pour gérer les doublons, nous considérons 3 situations :

1. Tuples identiques .
2. Différences dans certains attributs.
3. Valeurs absentes pour certains attributs dans l'un ou les deux tuples.

La méthode d'élimination du doublon repose sur un choix arbitraire, mais non moins réfléchi qui est de mettre en priorité la Source 1. IMDb étant la plus grande base de données de contenu visuel du monde, elle est donc plus fiable sur l'exactitude des données.

1. En cas de tuples identiques, le choix est facultatif.
2. Nous gardons les valeurs de la Source 1 en cas de différences avec la source 3
3. Si des valeurs manquent à la source 1, nous les complétons avec celles de la source 3 si elles existent.

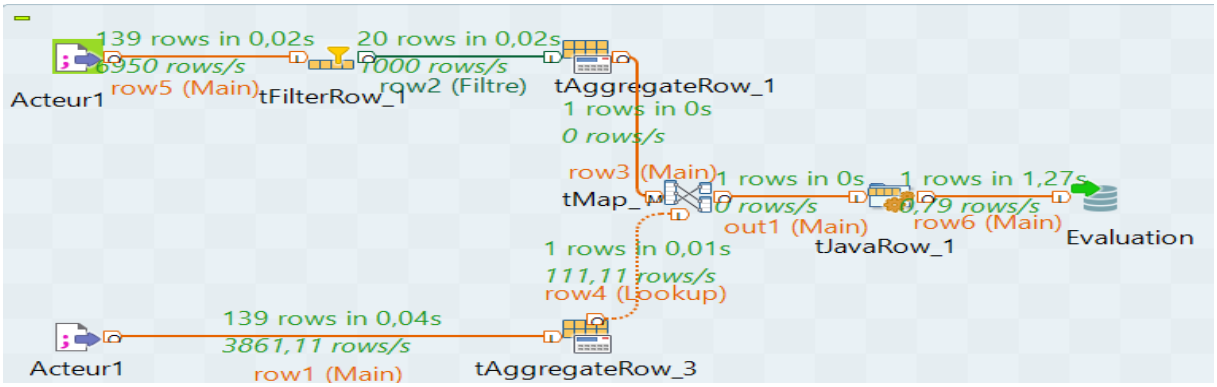


3.6 Cohérence des données.

1. **Niveau** : Colonnes birthYear et âge.
2. **Source** : Source 1 fichier Acteurs.csv.
3. **Métrique** : Pourcentage de valeurs cohérentes.

3.6.1 Méthode de détection

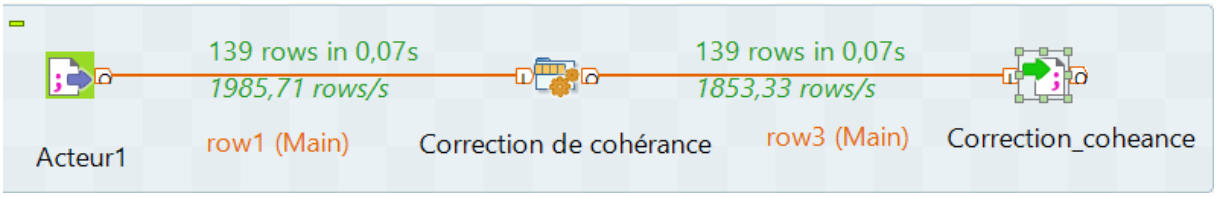
On compare la date de naissance de l'acteur avec son âge (qui sont dans la Source 1), donc si (date_actuelle - birthYear = age) on peut affirmer que la cohérence est respectée sinon un problème est détecté. On compte le nombre de valeurs incohérentes.



3.6.2 Processus d'amélioration

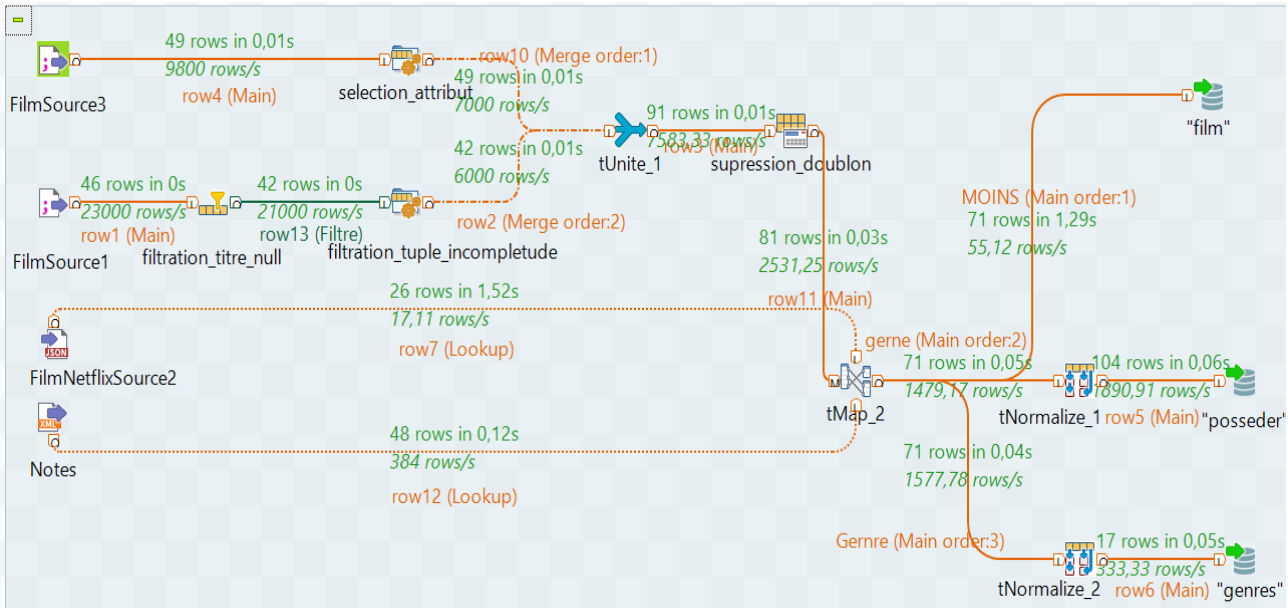
Nous considérons 4 situations :

1. ((birthYear = null ou birthYear > date_actuelle) et age = null).
   Dans ce cas on fait rien.
2. (birthYear != null et birthYear < dateactuelle et age = null).
   Alors on calcule l'âge comme suit : age = date_actuelle - birthYear
3. ((birthYear = null ou birthYear > date_actuelle) et age != null).
   Alors on calcule birthYear comme suit : birthYear = date_actuelle - age
4. (birthYear != null et birthYear < date_actuelle et age != null).
   Dans ce cas on suppose que les valeurs de l'attribut date de naissance ont un pourcentage d'exactitude plus élevé que celle de l'attribut âge. Par conséquent nous calculons la valeur d'âge comme suit : age=(date_actuelle - birthYear).



4 Création des tables cible

4.1 Création de la table Film, Genres et Posseder



Pour la table Film on commence d'abord par filtrer les tuples qui ont FilmTitre=null dans la table **FilmSource1**. Puis on filtre les tuples qui ont plus de 50% d'attributs null. Ensuite, on a vu lors de l'évaluation que runtime dans FilmSource1 est en secondes et dans FilmsSource3 il est en minutes, On traite donc cela en divisant la colonne runTime de la source 1 par 60, cette tache a été effectuée dans l'élément tjavaRow. Une fois les problèmes de complétude et d'hétérogénéité résolus, on fait l'union avec les films de la source 3 **FilmSource3** avec le composant tUnite où on va rencontrer le problème de doublons.
Une fois que les doublons sont traités comme décrit dans le processus d'amélioration des doublons, on fait la différence du résultat obtenu par cette dernière avec **FilmNetflix** pour ne garder que les films présents dans IMDb mais pas sur Netflix. Enfin, pour tous les films qui ne sont pas sur netflix, on récupère leur note qui se trouve dans le fichier NoteFilm.xml. Le résultat obtenu est **Film**(IdFilm,Title, Annee, runtimeMin, Note, genre).

4.1.1 Film

Pour la table film on prend le résultat obtenu dans la dernière étape (IdFilm, Titre, Annee, runtimeMin, Note, genre) on ne garde que (IdFilm, Titre, Annee, runtimeMin, Note) car l'attribut genre peut être multivalué. pour la création de la table cible des films. Cette séparation est effectuée dans tmap.

4.1.2 Genres

Pour la création de la table Genres on utilise le résultat de la dernière étape (IdFilm, Titre, Annee, runtimeMin, Note, genre) on ne récupére que l'attribut genre.
Comme chaque film peut avoir plusieurs genres donc l'attribut genre peut être composé de plusieurs valeurs différentes, ce qui justifie l'utilisation de tNormalize pour récupérer tous les genres de film. Exemple :
"(g1, g2, g3, g4), (g3, g5)" => "(g1, g2, g3, g4, g5)". Après on stocke le résultat dans la table Genres.

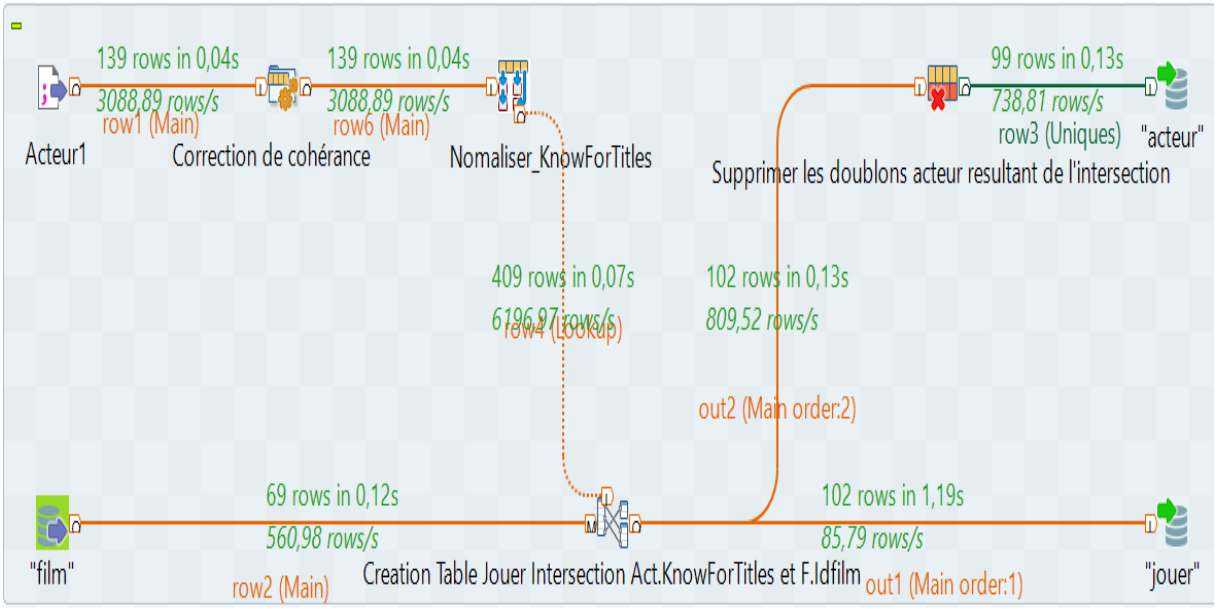
4.1.3 Posseder

On a crée la table **Posseder** à partir du résultat de la dernière étape. On récupère les deux attributs (#genre,#IdFilm) de tous les films. Mais il existe des films qui possèdent plusieurs genres donc on a utilisé **tNormalize** pour avoir chaque paire (#genre,#IdFilm), par exemple :
("genre1,genre2,genre3"IdFilm) => (#genre1, #IdFilm), (#genre2,#IdFilm), (#genre3,#IdFilm).
Le résultat est stocké dans la table Posseder.

4.2 Création de la table Jouer et acteur

On a crée la Table **Jouer(IdActeur,IdFilm)** à partir de la table Acteur et la table cible film.
Pour se faire on corrige d'abord la cohérence entre âge et date de naissance des acteurs avec le composant tjavarow, ensuite on normalise la table Acteur sur l'attribut *Acteur.knownForTitles*, puis on fait l'intersection entre *film.Idfilm* qui est dejà traitée et *Acteur.knownForTitle*, afin de n'avoir dans la table **Jouer** que les Idfilm qui figurent dans film.

Le tuniqueRow élimine les doublons acteur, car par l'intersection qui permet de construire la table jouer un acteur peut jouer dans plusieurs films, donc il sera dupliqué dans la table Jouer, ce qui est normal, mais on ne veut pas dupliquer le tuple dans acteur.



5 Echantillon des tables cible

5.1 Table Film

IdFilm	Titre	Annee	RuntimeMin	Note
tt0137818	Housesitter: The Night They Saved Siegfried's Brai...	2018	24.31	4
tt8962486	Thattum Purath Achuthan	2018	142	4.9
tt10001870	Disrupted Land	2019	77	0
tt8961950	Tampere Sinfonia	2018	58	0
tt0446792	Surviving in L.A.	2020	31.55	8.1
tt0846775	Cine Manifest	2019	75	7.2
tt0062336	El Tango del Viudo y Su Espejo Deformante	2020	70	6.6
tt10016254	Skvoznyye Shagi	2018	120	0
tt10003806	Jadugar 2	2019	80	0
tt0368133	The Promise of Perfume	2020	57	6.6

5.2 Table Genre

NomGenre
Action
Adventure
Biography
Comedy
Crime
Documentary
Drama

5.3 Table Posseder

IdFilm ▲ 1	NomGenre
tt0062336	Drama
tt0065392	Documentary
tt0111414	Comedy
tt0120589	Drama
tt0137818	Fantasy
tt0170651	Documentary
tt0192528	Drama
tt0276568	Action
tt0276568	Drama
tt0328810	Drama

5.4 Table Acteur

IdActeur	Nom	BirthYear	Age	Email
nm4401641	Brianna Noyes	1993	27	Brianna.Noyes838@hotmail.com
nm3616989	Shebin Backer	1993	27	Shebin.Backer880@hotmail.com
nm5875158	Deepankuran	1985	35	Deepankuran518@gmail.com
nm4132509	Louis Gagiano	1980	40	Louis.Gagiano684@hotmail.com
nm6330281	Elrich Yssel	1993	27	Elrich.Yssel953@gmail.com
nm7476471	Nuhan Yssel	1975	45	Nuhan.Yssel671@gmail.com
nm7112863	Lasse Heikkilä	1985	35	Lasse.Heikkilä552@hotmail.com
nm9583327	Mauri Valkeasuo	1998	22	Mauri.Valkeasuo592@hotmail.com
nm0730575	Connie Roberson	1987	33	Connie.Roberson312@hotmail.com
nm0861449	David Thorngren	1975	45	David.Thorngren768@hotmail.com
nm0905839	Christine Wagner	1963	57	Christine.Wagner159@gmail.com
nm0815612	Rubén Sotoconil	1975	45	Rubén.Sotoconil271@gmail.com
nm1471575	Galut Alarcón	1976	44	Galut.Alarcón748@hotmail.com
nm2065080	Luis Vilches	1985	35	Luis.Vilches991@gmail.com
nm1131208	Chamila Rodríguez	1977	43	Chamila.Rodríguez555@gmail.com
nm10539242	Bishesh Panta	1993	27	Bishesh.Panta605@gmail.com
nm10539243	Ganesh Sherestha	1985	35	Ganesh.Sherestha430@hotmail.com
nm10539244	Karna Bahadur Tamang	1987	33	Karna.Bahadur321@gmail.com

5.5 Table Jouer

IdFilm	IdActeur
tt0062336	nm0815612
tt0062336	nm1131208
tt0062336	nm1471575
tt0062336	nm2065080
tt0137818	nm4401641
tt0446792	nm0730575
tt0446792	nm0861449
tt0446792	nm0905839

6 Tables Statistiques des évaluations

Id	Facteur_Qualite	Metrique	Niveau	Valeur	Date
1	Complétude	le pourcentage des tuples null	Table FilmSource1	6.52	2020-12-10 01:52:21
2	doublons	Proportion des doublons	table cible film	11.58	2020-12-10 01:52:30
8	Complétude	le pourcentage des tuples null	Table FilmSource1	6.52	2020-12-16 23:29:50
4	Complétude	le pourcentage des tuples null	Table FilmSource1	6.52	2020-12-12 02:34:59
7	Complétude	le pourcentage des tuples null	Table FilmSource1	6.52	2020-12-16 23:21:02
6	doublons	Proportion des doublons	table cible film	11.58	2020-12-12 02:48:58
9	Complétude	le pourcentage des tuples null	Table FilmSource1	6.52	2020-12-16 23:31:00

Id	Facteur_Qualite	Metrique	Niveau	Colonne1	Colonne2	Valeur	Date
3	Cohérance	Pourcentage de valeurs non cohérentes	Table Acteur	Age	birthYear	14.29	2020-12-09 23:12:37
4	Cohérance	Pourcentage de valeurs non cohérentes	Table Acteur	Age	birthYear	14.39	2020-12-12 02:34:34
5	Cohérance	Pourcentage de valeurs non cohérentes	Table Acteur	Age	birthYear	14.39	2020-12-16 23:11:07
6	Cohérance	Pourcentage de valeurs non cohérentes	Table Acteur	Age	birthYear	14.39	2020-12-16 23:13:40
7	Cohérance	Pourcentage de valeurs non cohérentes	Table Acteur	Age	birthYear	14.39	2020-12-16 23:14:31

Facteur_Qualite	Metrique	Niveau	Colonne	Valeur	Date
Conformite Format	Pourcentage conforme au format	Table Acteurs	email	88	2020-12-14 02:17:03

Id	Facteur_Qualite	Metrique	Niveau1	Niveau2	Colonne1	Colonne2	Valeur	Date	Facteur
8	héteroeniété	si les valeurs ont la même échelle.	Table FilmSource1	Table FilmSource3	runTimeMinutes	runTimeSecondes	0	2020-12-12 02:40:07	60

Id	Facteur_Qualite	Metrique	Niveau	Colonne	Valeur	Date
1	Completude	le poucentage des valeur	FilmSource1	originalTitle	0.09	2020-12-09 21:20:14
2	Completude	le poucentage des valeur	FilmSource1	originalTitle	8.70	2020-12-09 21:37:36
3	Completude	le poucentage des valeur	FilmSource1	originalTitle	8.70	2020-12-10 01:51:00
4	confirmité au format	Le taux d'email faux	table acteur	colonne email	0.09	2020-12-10 02:35:24
5	Completude	le poucentage des valeur	FilmSource1	originalTitle	8.70	2020-12-10 17:54:59
6	Completude	le poucentage des valeur	FilmSource1	originalTitle	8.70	2020-12-12 02:34:45
7	confirmité au format	Le taux d'email faux	table acteur	colonne email	0.09	2020-12-12 02:35:10
8	Completude	le poucentage des valeur	FilmSource1	originalTitle	8.70	2020-12-16 23:22:28
9	Completude	le poucentage des valeur	FilmSource1	originalTitle	8.70	2020-12-16 23:22:53