

Université de Versailles  
Saint-Quentin-en-Yvelines



## M2 Datascale

---

# Evaluation et amélioration de la qualité de donnée de pollution atmosphérique

---

Par :

Sarra BRAHEM : 22302668

Oussama MAHDJOUR : 22410221

Katia BELGHERBI : 22407837

Soukaina TABAT : 22407847

---

# 1 Introduction

- Ce document synthétise notre démarche pour améliorer la qualité des données de pollution via Talend. Il décrit les étapes de traitement, d'intégration et d'optimisation, assurant ainsi une base fiable pour des analyses futures et une meilleure prise de décision.

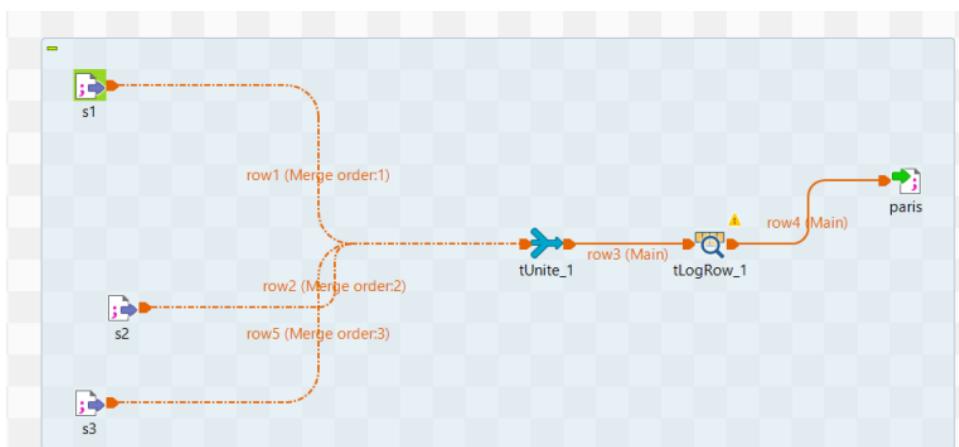
## 1.1 Exemple de Talend-Job pour s'entainer:

Dans cet exemple, nous avons conçu trois sources de données distinctes, chacune ayant un schéma différent. Le processus débute par une union des tables provenant de deux de ces sources, permettant ainsi de combiner les informations avant de générer une table de sortie consolidée.

Voici les étapes clés de l'ETL (Extract, Transform, Load) réalisé avec Talend :

1. **Création des sources de données** : Trois jeux de données distincts ont été définis, incluant des mesures des stations fixes, des données des capteurs mobiles, et des descriptions de polluants. Chacune de ces sources possède un schéma unique adapté à son type de données.
2. **Fusion des données** : Une opération d'union est effectuée entre les tables des sources pour regrouper et harmoniser les données pertinentes, facilitant ainsi une analyse globale.
3. **Génération de la table consolidée** : Une table de sortie est créée à partir des données fusionnées, constituant une base de données homogène et prête à l'exploitation pour des analyses futures.

Ce flux de travail montre les étapes essentielles pour établir un pipeline ETL dans Talend, allant de la collecte à l'intégration des données, avec pour objectif d'assurer une préparation efficace des données.



## 2 Sources de Données et Description

Dans cette section, nous allons présenter le schéma de nos sources de données et discuter des défis théoriques auxquels nous pourrions être confrontés lors de leur intégration et

traitement. Nous allons également expliquer comment nous avons anticipé ces problèmes et les solutions que nous avons envisagées pour les gérer. Les données proviennent de cinq sources distinctes (**S1**, **S2**, **S3**, **S4** et **S5**), chacune contenant des informations spécifiques et nécessitant des traitements adaptés pour assurer une harmonisation et une qualité optimale.

## 2.1 Source 1

La Source 1 contient deux tables principales : **Mesures** et **Stations**, représentant les relevés de pollution et les informations sur les stations fixes de surveillance.

### 2.1.1 Table Mesures

La table **Mesures** de la Source 1 comprend les relevés des polluants collectés par les capteurs des stations fixes. Les principales colonnes incluent :

- **ID\_Polluant** : Identifiant unique du polluant mesuré (par exemple, NO2, PM10, O3, CO).
- **Date** : Date et heure de la mesure.
- **ID\_Station** : Identifiant de la station ayant collecté la mesure.
- **Taux\_Relevé** : Concentration mesurée du polluant.

**Problème théorique** : La table **Mesures** peut contenir des valeurs manquantes ou des doublons, ce qui risque de fausser les analyses. Nous avons prévu d'appliquer des méthodes de nettoyage et de filtrage pour éliminer ces anomalies.

### 2.1.2 Table Stations

La table **Stations** fournit des informations détaillées sur chaque station fixe, incluant :

- **ID\_Station** : Identifiant unique de la station.
- **Numéro** : Numéro de la rue.
- **Rue** : Nom de la rue.
- **Code\_postal** : Code postal de la localisation de la station.
- **Ville** : Ville où se trouve la station.
- **Téléphone** : Numéro de téléphone de contact pour la station.
- **Contact\_Mail** : Adresse e-mail de contact pour la station.

**Problème théorique** : Des valeurs manquantes sont observées pour certains attributs, notamment le **Numéro** de la rue et le **Contact\_Mail**, ce qui complique l'identification des stations. Nous avons envisagé des transformations pour compléter les données et garantir la cohérence.

## 2.2 Source 2

La Source 2 contient des données similaires à celles de la Source 1, avec une structure comparable pour les tables **Mesures** et **Stations**.

### 2.2.1 Table Mesures

Les mesures collectées dans cette source suivent le même format que celles de la Source 1, mais proviennent de capteurs différents, nécessitant une harmonisation des données avant l'intégration.

**Problème théorique :** La diversité des capteurs et des unités de mesure peut entraîner une hétérogénéité des données. Nous avons normalisé les mesures avant l'union des tables pour éviter les incohérences.

### 2.2.2 Table Stations

Les informations sur les stations dans la Source 2 incluent des données de localisation et de contact similaires à celles de la Source 1.

**Problème théorique :** Comme pour la Source 1, il peut y avoir des valeurs manquantes ou des incohérences dans les adresses. Nous appliquerons des transformations pour garantir la cohérence des données.

## 2.3 Source 3

La Source 3 contient des informations sur les seuils de pollution pour chaque type de polluant. Cette table est utilisée pour classifier le statut des polluants (Normal ou Alerte Pollution).

- **ID\_Polluant** : Identifiant unique du polluant.
- **Désignation** : Nom du polluant (exemple : Dioxyde d'azote, Particules).
- **Seuil\_Toléré** : Valeur seuil tolérée pour le polluant (exprimée en  $\mu\text{g}/\text{m}^3$ ).

**Problème théorique :** Les valeurs seuils peuvent différer en fonction des réglementations locales. Une validation des données est nécessaire pour s'assurer qu'elles sont conformes aux standards utilisés.

## 2.4 Source 4

La Source 4 regroupe des mesures collectées par des capteurs mobiles. Les données incluent des coordonnées géographiques pour chaque relevé, facilitant ainsi l'analyse spatiale.

- **ID\_Polluant** : Identifiant unique du polluant.
- **Date** : Date et heure de la mesure.
- **ID\_Capteur** : Identifiant du capteur mobile.
- **Localisation** : Coordonnées géographiques (latitude, longitude).
- **Taux\_Relevé** : Concentration mesurée du polluant.

**Problème théorique :** La précision des coordonnées GPS peut varier, affectant l’analyse de la localisation des polluants. Nous utilisons une **API de géolocalisation** pour obtenir des données GPS précises et filtrons les valeurs aberrantes pour garantir la qualité des données.

## 2.5 Source 5

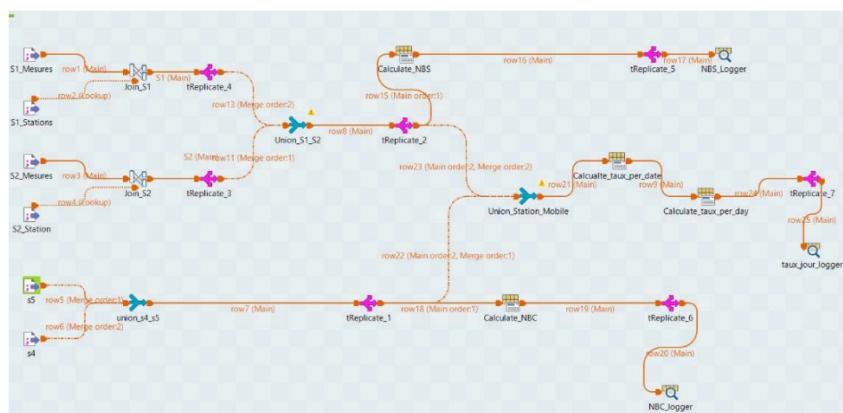
La Source 5 est similaire à la Source 4, mais provient d’une campagne de collecte de données différente. Les données nécessitent une harmonisation pour correspondre au format des mesures de la Source 4.

- **ID\_Polluant** : Identifiant unique du polluant.
- **Date** : Date et heure de la mesure.
- **ID\_Capteur** : Identifiant du capteur mobile.
- **Localisation** : Coordonnées géographiques (latitude, longitude).
- **Taux\_Relevé** : Concentration mesurée du polluant.

**Problème théorique :** La fusion des données de la Source 5 avec celles de la Source 4 peut entraîner des doublons et des incohérences. Nous appliquerons des règles de déduplication et des transformations pour assurer une intégration fluide.

## 3 Chemin initial du workflow dans Talend

Afin de structurer et d’organiser le flux de données, nous avons d’abord conçu un chemin de traitement dans Talend, représenté par le schéma ci-dessous. Il est important de noter que cette démarche a été réalisée de manière autonome, en utilisant nos propres données et en appliquant une approche intelligente pour élaborer un processus ETL optimisé. Nous n’avons pas simplement utilisé les données fournies par le professeur, mais avons créé et testé notre propre jeu de données pour mieux maîtriser le flux de travail et anticiper les éventuels problèmes de qualité des données.



Ce flux commence par l’intégration des sources de données **S1**, **S2**, **S4**, et **S5**, suivie de plusieurs opérations de jointures et d’unions.

Ces transformations permettent de combiner les données des stations fixes et des capteurs

mobiles, tout en calculant les indicateurs clés tels que **NBS** (Nombre de Stations Fixes) et **NBC** (Nombre de Capteurs Mobiles). Ce cheminement sert de base pour les étapes d'agrégation et de calcul des taux de pollution journaliers, assurant ainsi une collecte et une analyse efficaces des données.

| taux_jour_logger |                 |                |
|------------------|-----------------|----------------|
| ID_Polluant      | taux_moyen_jour | Ville          |
| NO2              | 60.17           | Paris          |
| NO2              | 48.21           | Hauts_de_Seine |
| PM10             | 43.845          | Yveline        |
| CO2              | 46.81           | null           |
| SO2              | 71.655          | Paris          |
| CO2              | 56.2175         | Yveline        |
| CO2              | 41.705          | Hauts_de_Seine |
| SO2              | 54.96           | null           |
| PM10             | 67.42           | null           |
| PM10             | 35.408          | Paris          |
| CO2              | 46.71           | Paris          |
| NO2              | 62.836666       | Yveline        |
| SO2              | 60.843334       | Yveline        |
| SO2              | 48.69           | Hauts_de_Seine |
| PM10             | 59.786667       | Hauts_de_Seine |

Ce tableau présente les taux moyens journaliers de pollution (**taux\_moyen\_jour**) par type de polluant (**ID\_Polluant**) et par ville (**Ville**).

| NBS_Logger  |     |                | NBC_logger  |     |                |
|-------------|-----|----------------|-------------|-----|----------------|
| ID_Polluant | NBS | Ville          | ID_Polluant | NBC | Ville          |
| NO2         | 7   | Paris          | PM10        | 2   | Paris          |
| CO2         | 1   | null           | CO2         | 2   | Paris          |
| SO2         | 2   | Paris          | NO2         | 1   | Paris          |
| CO2         | 2   | Yveline        | NO2         | 2   | Hauts_de_Seine |
| SO2         | 3   | null           | PM10        | 2   | Yveline        |
| CO2         | 1   | Hauts_de_Seine | NO2         | 2   | Yveline        |
| PM10        | 2   | null           | SO2         | 2   | Hauts_de_Seine |
| PM10        | 3   | Paris          | SO2         | 2   | Yveline        |
| CO2         | 4   | Paris          | PM10        | 1   | Hauts_de_Seine |
| NO2         | 1   | Yveline        | SO2         | 1   | Paris          |
| SO2         | 1   | Hauts_de_Seine | CO2         | 2   | Yveline        |
| SO2         | 1   | Yveline        | CO2         | 1   | Hauts_de_Seine |
| PM10        | 2   | Hauts_de_Seine |             |     |                |

Ce résultat montre le calcul du nombre de stations fixes (**NBS**) et du nombre de capteurs mobiles (**NBC**) par type de polluant (**ID\_Polluant**) et par ville (**Ville**).

## Étapes Réalisées

- Jointure sur les sources S1 :

Une jointure a été réalisée entre les tables de la source S1. Après cette jointure, une requête a été envoyée pour optimiser le traitement dans le SGBD, ce qui nous a permis de générer une table complète des mesures des stations pour S1.

- **Même traitement pour S2 :**

La même procédure a été appliquée à la source S2, en exécutant une jointure similaire et en récupérant les mesures correspondantes pour les stations de cette source.

- **Union entre les sources 4 et 5 :**

Une union a été réalisée entre les sources S4 et S5 pour combiner les données provenant des capteurs mobiles.

- **Fusion des sorties de jointure (S1 et S2) :**

Ensuite, une autre union a été faite entre la sortie de la jointure de S1 et celle de S2 pour fusionner les données des deux sources.

- **Agrégation pour calculer NBS et NBC :**

Une agrégation a été réalisée pour calculer le **NBS** (Nombre de Stations Fixes), en le groupant par ville et par **ID\_Polluant**. Le même traitement a été effectué pour calculer le **NBC** (Nombre de Capteurs Mobiles), en groupant également par ville et **ID\_Polluant**.

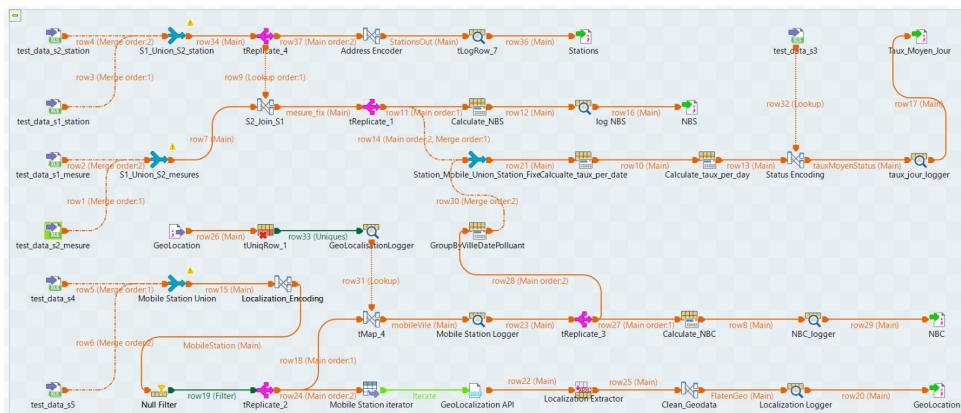
## 4 Implémentation ETL (Talend)

### 4.1 Aperçu du Job ETL

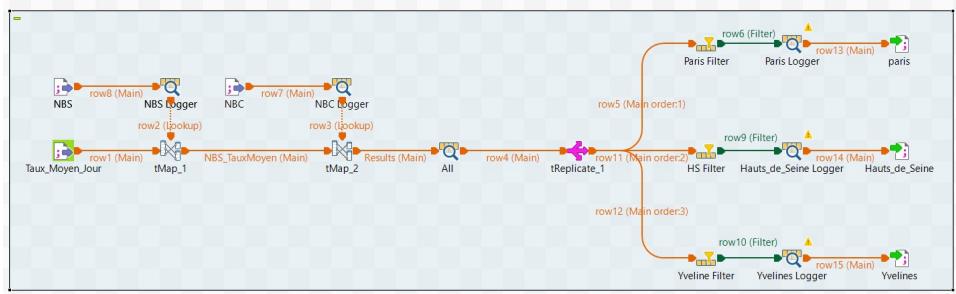
Dans ce projet, notre objectif est de rassembler des données provenant de plusieurs sources et de les intégrer dans une nouvelle table unifiée qui contient des informations importantes pour aider à la prise de décision. Dans cette section, nous allons vous donner un aperçu de notre travail et vous expliquer le processus ETL étape par étape.

Le processus ETL est divisé en deux jobs principaux dans Talend :

- **Job de Transformation :** Ce job s'occupe de transformer les données brutes provenant des différentes sources.



- **Job de Groupement et d’Agrégation :** Ce job regroupe les données transformées et effectue des calculs pour créer les tables finales.



#### 4.1.1 Job de Transformation

Dans le Job de Transformation, nous commençons par extraire les données des stations de surveillance de la pollution et les traiter pour créer une table appelée **Stations**. Nous réalisons également plusieurs transformations, notamment :

- Calculer le nombre de stations (**NBS**) par région.
- Calculer le nombre de capteurs (**NBC**) par région.
- Calculer la concentration moyenne de polluants par jour (**Taux moyen/jour**) pour chaque région.
- Classifier le statut des polluants pour chaque région comme étant soit ”**Alerte Pollution**” soit ”**Normal**”.

#### 4.1.2 Job de Groupement et d’Agrégation

Dans le Job de Groupement et d’Agrégation, nous utilisons les résultats du job précédent pour organiser les données dans une structure unifiée. Cette étape consiste à regrouper les données par région et à effectuer les agrégations nécessaires afin de créer les tables finales, qui serviront ensuite de base pour l’analyse et la prise de décision.

### 4.2 Processus ETL

#### 4.2.1 Job de Transformation

##### 1. Extraction des Stations Existantes

Dans ce processus, nous extrayons les données pour créer une table appelée **Stations**, comme le montre le schéma ci-dessous.



Nous avons pour objectif de créer une table **Stations** avec les champs suivants :

- **ID\_Station**
- **Adresse**
- **Contact**
- **Téléphone**

Pour atteindre cet objectif, nous commençons par combiner les données des deux tables **Stations**, l'une provenant de la **Source 1** et l'autre de la **Source 2**. Nous utilisons une opération d'Union (le composant **tUnite** dans Talend), qui fusionne les données des deux sources en un seul ensemble. Cela nous permet de consolider les informations des stations des deux sources dans un format unifié.

Ensuite, nous appliquons une transformation pour créer la colonne **Adresse**. L'adresse est constituée de plusieurs colonnes individuelles provenant des données sources :

- **Numero** (numéro de rue)
- **Rue** (nom de la rue)
- **Ville** (nom de la ville)

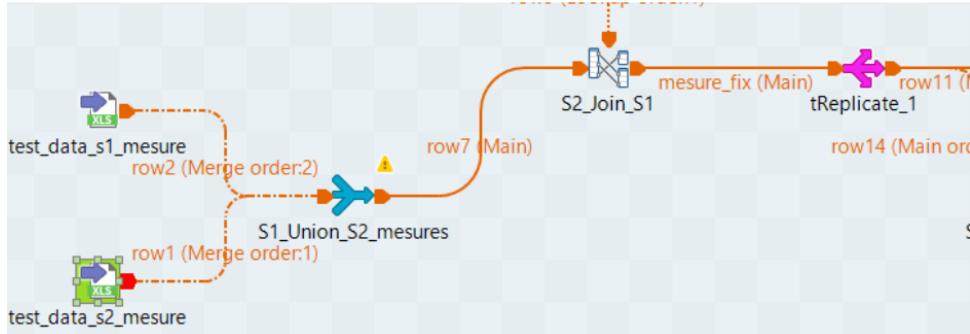
En concaténant ces trois colonnes, nous formons l'adresse complète pour chaque station.

Enfin, les données transformées sont sauvegardées dans une nouvelle table appelée **Station**, qui contient les informations consolidées des stations dans le schéma cible. Voici un exemple des résultats :

| Stations Logger |  |  |              |
|-----------------|--|--|--------------|
| ID_Station      | Address  | Telephone                                | Contact_Mail |
| 1               | 5 Etant d'or, Rambouillet  | 01 75 03 40 00 rambouillet@airparif.fr   |              |
| 2               | 8 Limoges, Versailles  | 01 39 25 40 00  versailles@airparif.fr   |              |
| 3               | 1 Emile Zola, Mantes-la-Jolie  | 01 34 78 81 00  mantes@airparif.fr       |              |
| 4               | null Parvis de la Défense, Puteaux   | 01 46 92 92 92  ladefense@airparif.fr    |              |
| 5               | null Allée des Refusniks, Paris  | 01 53 58 75 07  paris07@airparif.fr      |              |
| 6               | 2 bis Quai de la Mégisserie, Paris   | 01 44 50 75  pari01@airparif.fr          |              |
| 7               | 60 Richelieu, Gennevillier   | 01 40 85 66 66  gennevillier@airparif.fr |              |
| 1               | null Château_princeloup, Sonchamp  | 01 34 84 41 08  sonchamp@                |              |
| 2               | 60 Richelieu, Gennevillier   | 01 40 85 66 66  gennevillier@airparif.fr |              |
| 3               | 11 Commandant Pilot, Neuilly-sur-seine 01 40 88 88 88  neuilly@airparif.fr |  |              |
| 4               | null Quai de la Mégisserie, null   | 01 44 50 75 01 Paris01@airparif.fr       |              |
| 5               | 7 Ferdinand Flocon, Paris  | 01 53 41 18 18  paris18@airparif.fr      |              |
| 6               | null Parvis de la Défense, Puteaux   | 01 46 92 92 92  null                     |              |

## 2. Extraction des Mesures des Stations

Le processus est illustré ci-dessous, où nous nous concentrons sur l'extraction des données de mesure des polluants à partir des tables **Mesures** trouvées dans **Source 1** et **Source 2**.



Pour commencer, nous combinons les deux tables **Mesures** en utilisant une opération d'Union (le composant **tUnite** dans Talend). Cela nous permet de rassembler les données des deux sources dans un seul ensemble de données.

Une fois les données combinées, nous effectuons une opération de Jointure sur le champ **ID\_Station** (avec le composant **tMap** dans Talend). Cette jointure relie les données de mesure fusionnées à la table **Station**, créée lors de l'étape précédente (Extraction des Stations Existantes).

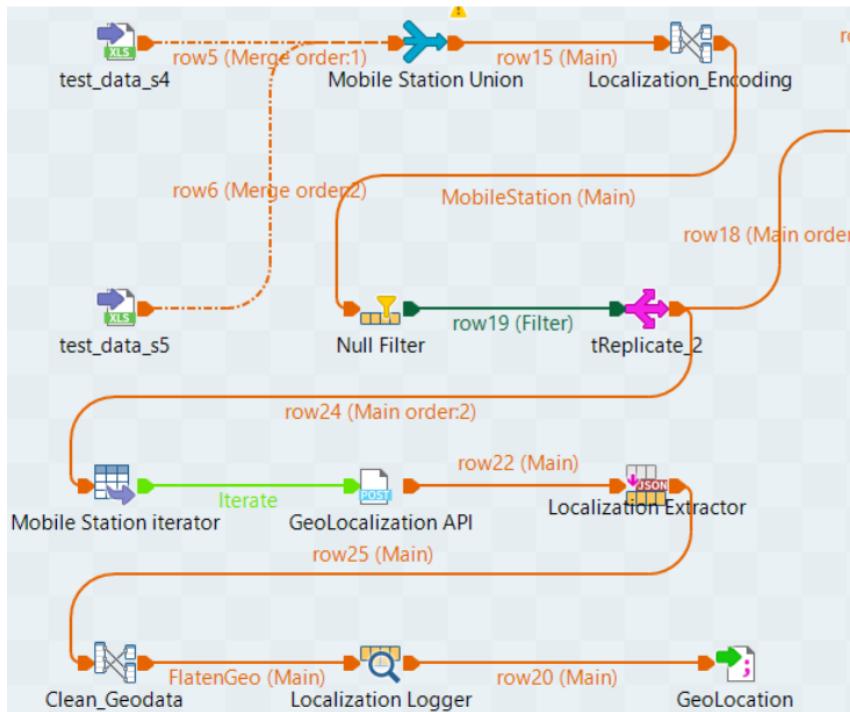
Le résultat de cette jointure est une nouvelle table appelée **Station\_Mesures**, avec le schéma suivant :

- **ID\_Polluant** (Identifiant du polluant)
- **Date** (Date de la mesure)
- **Taux\_releve** (Concentration mesurée)
- **Ville** (Nom de la ville/région)

Cette étape garantit que nous disposons de mesures de polluants liées à des stations et des régions spécifiques, ce qui nous permet de travailler avec des données plus détaillées et exploitables.

## 3. Extraction des Mesures des Capteurs Mobiles

Le processus est illustré dans la figure ci-dessous. Ici, nous nous concentrons sur l'extraction des données des tables **Mesures** provenant des **Sources 4** et **5**.



Pour commencer, nous utilisons une opération d’Union (avec le composant **tUnité** dans Talend) afin de combiner les données des deux sources en un seul ensemble de données.

Ensuite, nous transformons la colonne **Location** en deux colonnes distinctes : **Latitude** et **Longitude**. Cela nous permet de travailler avec des coordonnées géographiques nécessaires pour le traitement de la géolocalisation.

Pour garantir la qualité des données, nous filtrons les **valeurs nulles** en utilisant le composant **tFilterRow** dans Talend. Cette étape assure que seules les données valides sont transmises à l'**API de Géolocalisation**.

Après le filtrage, nous itérons sur chaque ligne de l’ensemble de données à l'aide du composant **tFlowToIterator**. Cette itération nous permet de faire un appel API pour chaque point de données. Nous utilisons ensuite le composant **tHttpRequest** pour envoyer les données à l'API de Géolocalisation.

Lorsque nous recevons la réponse de l'API, nous extrayons les champs pertinents : **Ville** (nom de la ville), **Latitude**, et **Longitude** en utilisant le composant **tExtractJsonFields** dans Talend. Cela nous permet de créer une nouvelle table appelée **Geolocalisation**, qui contient les coordonnées géographiques et les informations sur la ville pour chaque point de données.

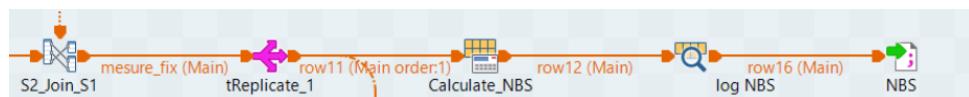
Enfin, nous enregistrons les résultats dans la table **Geolocalisation**, qui sera utilisée pour des analyses ultérieures.

Le tableau résultant est présenté dans la figure ci-dessous.

| GeoLocalisationLogger |           |          |
|-----------------------|-----------|----------|
| Ville                 | Lat       | Lon      |
| Paris                 | 48.886933 | 2.334836 |
| Paris                 | 48.833868 | 2.321821 |
| Versailles            | 48.803413 | 2.147692 |
| Paris                 | 48.859066 | 2.279786 |
|                       | 48.792784 | 2.108038 |
| Versailles            | 48.81144  | 2.138852 |
| Paris                 | 48.875044 | 2.319998 |
| Paris                 | 48.841419 | 2.307009 |
| Paris                 | 48.836561 | 2.291216 |
| Paris                 | 48.885797 | 2.351305 |
| Paris                 | 48.830952 | 2.33128  |
| Paris                 | 48.827336 | 2.32716  |
| Boulogne-Billancourt  | 48.837013 | 2.239375 |
| Paris                 | 48.82598  | 2.343639 |
| Paris                 | 48.831304 | 2.320748 |
| Paris                 | 48.826501 | 2.345124 |
| Paris                 | 48.821372 | 2.345562 |
| Boulogne-Billancourt  | 48.844595 | 2.234076 |
| Paris                 | 48.83338  | 2.359467 |
| Paris                 | 48.830329 | 2.360883 |

#### 4. Calcul du Nombre de Stations par Ville

Dans ce processus, nous utilisons la table **Station\_Measures**, qui a été créée lors de l'étape précédente. L'objectif est de calculer le nombre de stations pour chaque **ID\_Polluant** et chaque **Ville**.



Pour y parvenir, nous regroupons les données selon **ID\_Polluant** (identifiant du polluant) et **Ville** (nom de la ville/région). Pour chaque groupe, nous appliquons la fonction **COUNT** afin de calculer le nombre de stations, que nous enregistrons dans une nouvelle colonne appelée **NBS** (Nombre de Stations).

Les données résultantes sont ensuite sauvegardées dans une nouvelle table appelée **NBS**, avec le schéma suivant :

- **ID\_Polluant** : Identifiant du polluant

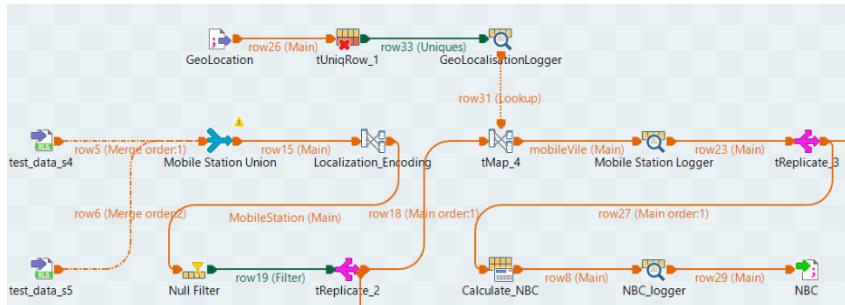
- **Ville** : Nom de la ville/région
- **NBS** : Nombre de stations
- **Taux\_releve** : Concentration mesurée

Les résultats de cette étape sont présentés dans la figure ci-dessous.

| NBS Logger  |     |                   |
|-------------|-----|-------------------|
| ID_Polluant | NBS | Ville             |
| CO          | 1   | Neuilly-sur-seine |
| PM10        | 4   | Puteaux           |
| IO3         | 1   | Sonchamp          |
| IO3         | 2   | Gennevillier      |
| IO3         | 1   | Puteaux           |
| PM10        | 6   | Sonchamp          |
| IO3         | 1   | Neuilly-sur-seine |
| NO2         | 1   | Gennevillier      |
| NO2         | 4   | Puteaux           |
| CO          | 1   | Gennevillier      |
| NO2         | 5   |                   |
| CO          | 3   |                   |
| CO          | 2   | Paris             |

## 5. Calcul du Nombre de Capteurs par Ville

Dans ce processus, illustré dans la figure ci-dessous, nous avons pour objectif de calculer le **NBC** (Nombre de Capteurs par Ville). Les étapes sont les suivantes :



Nous commençons par effectuer une jointure entre l'union des tables de mesures provenant des Sources 4 et 5 et la table **Geolocalisation** créée dans le processus précédent. Cette jointure permet d'associer chaque mesure à sa localisation géographique correspondante (**Ville**).

Ensuite, nous calculons le **NBC** en utilisant le composant **tAggregateRow** de Talend. L'agrégation est réalisée en regroupant les données par **ID\_Polluant** (Identifiant du Polluant) et **Ville**. Pour chaque groupe, nous comptons le nombre d'occurrences afin

de déterminer le nombre de capteurs, et nous enregistrons ce résultat dans une nouvelle colonne appelée **NBC**.

Les données résultantes sont ensuite sauvegardées dans une nouvelle table appelée **NBC**, avec le schéma suivant :

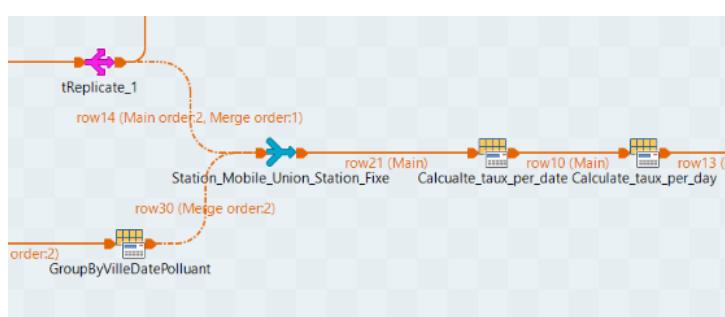
- **ID\_Polluant** : Identifiant du Polluant
- **Ville** : Nom de la Ville/Région
- **NBC** : Nombre de Capteurs
- **Taux\_Releve** : Concentration Mesurée

Les résultats des données de test sont également fournis dans la figure ci-dessous.

| NBC Logger  |                      |     |
|-------------|----------------------|-----|
| ID_Polluant | Ville                | NBC |
| PM10        | Paris                | 4   |
| NO2         | Paris                | 4   |
| O3          | Versailles           | 1   |
| NO2         |                      | 3   |
| PM10        | Boulogne-Billancourt | 1   |
| CO          | Paris                | 2   |
| O3          | Boulogne-Billancourt | 1   |
| O3          |                      | 1   |
| O3          |                      | 1   |
| O3          | Paris                | 1   |
| PM10        |                      | 1   |

## 6. Calcul du Taux Moyen/Jour par Ville

Dans ce processus, nous calculons le **Taux Moyen/Jour** (concentration moyenne quotidienne) pour chaque ville (région).



Nous commençons par effectuer une jointure entre la table **Station\_Mesures** (créeée dans un processus précédent) et la table **Capteur\_Mesures** (issue du processus précédent). Cette jointure est réalisée avec le composant **tMap** de Talend, en utilisant l'**ID\_Polluant**

comme clé de correspondance. Cette étape permet de combiner les mesures des stations et des capteurs pour créer un ensemble de données complet.

Les données jointes sont ensuite agrégées par **Date**, **Ville**, et **ID\_Polluant** à l'aide du composant **tAggregateRow**. Pendant cette étape, toutes les mesures pour une journée donnée dans une ville spécifique sont regroupées, et leur concentration moyenne est calculée pour obtenir le **Taux Moyen/Jour**.

Les résultats de cette première agrégation passent ensuite par une deuxième agrégation, où les données sont regroupées par **Ville** et **ID\_Polluant**. Cette étape calcule le **Taux Moyen/Jour** global pour chaque polluant dans chaque ville.

Les données finales sont enregistrées dans une table avec le schéma suivant :

- **Ville** : Nom de la Ville/Région
- **ID\_Polluant** : Identifiant du Polluant
- **Taux\_Moyen\_Jour** : Concentration Moyenne Quotidienne

## 7. Classification des Niveaux de Pollution par Ville

Dans ce processus, nous classifions les niveaux de pollution pour chaque ville (région) en fonction des seuils de tolérance des polluants.



Nous commençons par effectuer une jointure entre la table de sortie du processus précédent (qui contient les valeurs de **Taux\_Moyen\_Jour**) et la table **Polluants** provenant de la Source 3. Cette jointure utilise l'**ID\_Polluant** pour accéder aux valeurs de seuil (**Seuil\_Toléré**) pour chaque polluant.

En utilisant le composant **tMap** de Talend, nous comparons les valeurs de **Taux\_Moyen\_Jour** avec le **Seuil\_Toléré** correspondant. Selon le résultat de cette comparaison, nous classons chaque enregistrement dans l'un des deux statuts :

- **Alerte\_Pollution** : Si le **Taux\_Moyen\_Jour** dépasse le seuil.
- **Normal** : Si le **Taux\_Moyen\_Jour** est dans la limite acceptable.

La table finale, incluant le statut calculé, est enregistrée dans une table appelée **Taux\_Moyen\_Jour**, avec le schéma suivant :

- **Ville** : Nom de la Ville/Région
- **ID\_Polluant** : Identifiant du Polluant
- **Taux\_Moyen\_Jour** : Concentration Moyenne Quotidienne
- **Statut** : Statut de Pollution ("Alerte\_Pollution" ou "Normal")

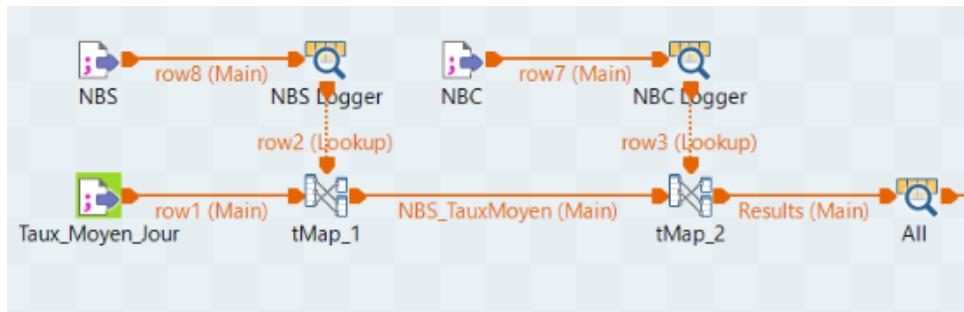
| taux_jour_logger |                 |                      |                  |  |
|------------------|-----------------|----------------------|------------------|--|
| ID_Polluant      | taux_moyen_jour | Ville                | status           |  |
| NO2              | 44.198746       | Paris                | Normal           |  |
| CO               | 8987.0          | Neuilly-sur-seine    | Normal           |  |
| O3               | 199.4           | Versailles           | Alerte_pollution |  |
| PM10             | 0.1             | Boulogne-Billancourt | Normal           |  |
| PM10             | 1827.6875       | Puteaux              | Alerte_pollution |  |
| O3               | 234.65          | Sonchamp             | Alerte_pollution |  |
| O3               | 0.087           | Boulogne-Billancourt | Normal           |  |
| O3               | 157.1           | Gennévillier         | Alerte_pollution |  |
| O3               | 41.57           | Puteaux              | Normal           |  |
| O3               | 145.7           | Neuilly-sur-seine    | Alerte_pollution |  |
| PM10             | 86.02           | Sonchamp             | Alerte_pollution |  |
| NO2              | 100.5           | Gennévillier         | Normal           |  |
| PM10             | 4.78967         | null                 | Normal           |  |
| NO2              | 2748.7466       | Puteaux              | Alerte_pollution |  |
| CO               | 7.0             | Gennévillier         | Normal           |  |
| PM10             | 25.563667       | Paris                | Normal           |  |
| NO2              | 645.8167        | null                 | Alerte_pollution |  |
| CO               | 5076.15         | null                 | Normal           |  |
| CO               | 4027.335        | Paris                | Normal           |  |
| O3               | 145.7           |                      | Alerte_pollution |  |
| O3               | 234.65          | null                 | Alerte_pollution |  |
| O3               | 38.0            | Paris                | Normal           |  |

#### 4.2.2 Job de Groupement

Le **Job de Groupement** est la dernière étape de notre processus ETL. À cette étape, nous consolidons et structurons les données transformées dans le schéma cible, et produisons les tables de sortie nécessaires pour l'analyse.

##### 1. Consolidation des Données

En utilisant les résultats du **Job de Transformation**, nous commençons par faire une jointure et un regroupement des données avec le composant **tMap** dans Talend. Cette étape permet de créer une table unifiée appelée **All\_Measurements**, qui contient toutes les mesures pertinentes selon le schéma cible suivant :



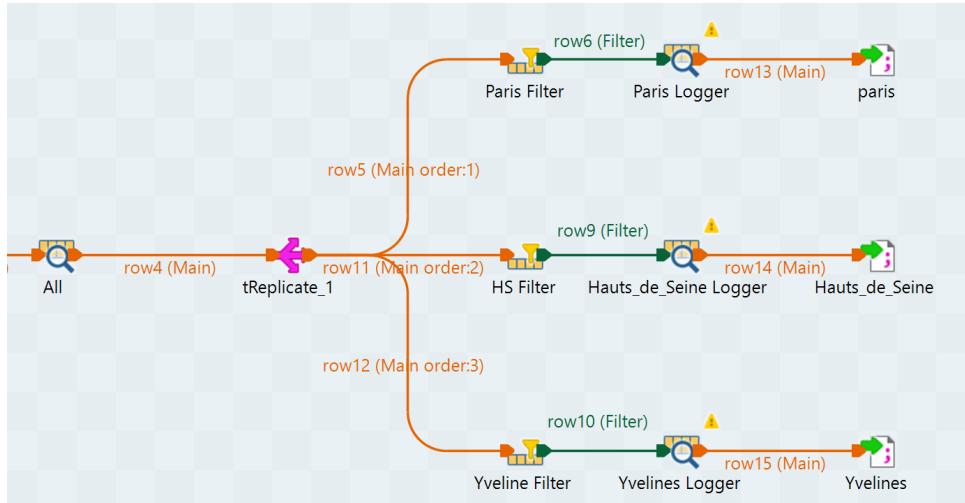
- **ID\_Polluant** (Identifiant du polluant)
- **Ville** (Ville/Région)
- **Taux\_Moyen** (Concentration moyenne quotidienne)
- **Statut** (Statut de pollution : "Alerte\_Pollution" ou "Normal")
- **NBS** (Nombre de Stations Fixes)
- **NBC** (Nombre de Capteurs Mobiles)

| All         |                 |                      |     |     |                  |
|-------------|-----------------|----------------------|-----|-----|------------------|
| ID_Polluant | taux_moyen_jour | Ville                | NBS | NBC | status           |
| NO2         | 44.198746       | Paris                | 0   | 4   | Normal           |
| CO          | 8987.0          | Neuilly-sur-seine    | 1   | 0   | Normal           |
| O3          | 199.4           | Versailles           | 0   | 1   | Alerte_pollution |
| PM10        | 0.1             | Boulogne-Billancourt | 0   | 1   | Normal           |
| PM10        | 1827.6875       | Puteaux              | 4   | 0   | Alerte_pollution |
| O3          | 234.65          | Sonchamp             | 1   | 0   | Alerte_pollution |
| O3          | 0.087           | Boulogne-Billancourt | 0   | 1   | Normal           |
| O3          | 157.1           | Gennevillier         | 2   | 0   | Alerte_pollution |
| O3          | 41.57           | Puteaux              | 1   | 0   | Normal           |
| O3          | 145.7           | Neuilly-sur-seine    | 1   | 0   | Alerte_pollution |
| PM10        | 86.02           | Sonchamp             | 6   | 0   | Alerte_pollution |
| NO2         | 100.5           | Gennevillier         | 1   | 0   | Normal           |
| PM10        | 4.78967         |                      | 0   | 1   | Normal           |
| NO2         | 2748.7466       | Puteaux              | 4   | 0   | Alerte_pollution |
| CO          | 7.0             | Gennevillier         | 1   | 0   | Normal           |
| PM10        | 25.563667       | Paris                | 0   | 4   | Normal           |
| NO2         | 645.8167        |                      | 5   | 3   | Alerte_pollution |
| CO          | 5076.15         |                      | 3   | 0   | Normal           |
| CO          | 4027.335        | Paris                | 2   | 2   | Normal           |
| O3          | 145.7           |                      | 0   | 1   | Alerte_pollution |
| O3          | 234.65          |                      | 0   | 1   | Alerte_pollution |
| O3          | 38.0            | Paris                | 0   | 1   | Normal           |

## 2. Crédation des Tables Cibles

Pour créer les tables cibles pour des départements spécifiques (Paris, Yvelines, Hauts-de-Seine), nous filtrons les données à l'aide du composant **tFilterRow**. Ce composant garantit que seuls les enregistrements correspondant aux villes de chaque département

sont inclus. Chaque table de département est ensuite agrégée et sauvegardée séparément.



### 3. Résultats

Le résultat final consiste en les tables cibles suivantes :

- Paris

| Paris Logger |                 |     |     |        |
|--------------|-----------------|-----|-----|--------|
| ID_Polluant  | taux_moyen_jour | NBS | NBC | status |
| NO2          | 44.198746       | 0   | 4   | Normal |
| PM10         | 25.563667       | 0   | 4   | Normal |
| CO           | 4027.335        | 2   | 2   | Normal |
| O3           | 38.0            | 0   | 1   | Normal |

- Yvelines

| Yvelines Logger |                 |     |     |                  |
|-----------------|-----------------|-----|-----|------------------|
| ID_Polluant     | taux_moyen_jour | NBS | NBC | status           |
| O3              | 199.4           | 0   | 1   | Alerte_pollution |

- Hauts-de-Seine

| Hauts_de_Seine Logger |                 |     |     |                  |  |  |
|-----------------------|-----------------|-----|-----|------------------|--|--|
| ID_Polluant           | taux_moyen_jour | NBS | NBC | status           |  |  |
| PM10                  | 0.1             | 0   | 1   | Normal           |  |  |
| PM10                  | 1827.6875       | 4   | 0   | Alerte_pollution |  |  |
| O3                    | 0.087           | 0   | 1   | Normal           |  |  |
| O3                    | 41.57           | 1   | 0   | Normal           |  |  |
| NO2                   | 2748.7466       | 4   | 0   | Alerte_pollution |  |  |

Ces tables sont structurées et prêtes pour des analyses ultérieures ou des processus de prise de décision.

### 4.3 Bilan sur les Données Utilisées

L'exécution de nos mappings repose sur des données provenant de cinq sources différentes (**S1**, **S2**, **S3**, **S4**, **S5**). Dans cette section, nous analysons ces données en détail et mettons en évidence les problématiques potentielles.

- **Présence de Valeurs Manquantes :**

Dans les tables **S1.Station** et **S2.Station**, des valeurs manquantes sont observées pour l'attribut **numéro de la rue**. L'attribut **ville** dans **S2.Station** comporte également des valeurs manquantes, compliquant ainsi l'identification précise de l'emplacement des stations. De plus, l'attribut **contact\_mail** dans **S2.Station** présente des valeurs manquantes.

Dans la table **S4.Mesure**, des valeurs nulles sont observées pour les coordonnées de géolocalisation (absence de valeurs ou présence d'une seule coordonnée X ou Y), limitant ainsi la précision de l'identification de l'emplacement de la mesure. Par ailleurs, les tables **S4.Mesure** et **S5.Mesure** affichent des valeurs manquantes pour l'attribut **Taux\_relevé**, rendant difficile l'obtention de mesures précises pour un polluant spécifique à un emplacement donné.

- **Différence des Échelles :**

Les données des stations utilisent des attributs tels que **numéro de la rue**, **nom de la rue**, **ville**, et **adresse postale** pour spécifier l'emplacement. En revanche, dans **S4.Mesure** et **S5.Mesure**, l'emplacement est défini par des coordonnées géographiques (X, Y). Cette différence d'échelle nécessite une transformation pour harmoniser les données et faciliter leur utilisation.

- **Disparité des Valeurs :**

Une analyse des données révèle une disparité notable dans certaines valeurs de l'attribut **Taux\_relevé** dans **S4.Taux\_relevé** et **S5.Taux\_relevé**, avec des écarts significatifs par rapport à la majorité des valeurs collectées. Ces disparités peuvent indiquer des anomalies ou des erreurs de mesure, nécessitant une attention particulière lors de l'analyse des résultats.

## 5 Problèmes de Qualité

Dans le cadre de notre étude, nous nous intéressons à plusieurs problématiques liées à la qualité des données, notamment la conformité à un format ou à une codification, l'hétérogénéité des échelles et de la granularité, la complétude des données, ainsi que la détection et l'élimination des doublons. Pour chaque facteur, nous préciserons notre objectif, la table et les colonnes ciblées, la métrique utilisée, la méthodologie adoptée et les résultats obtenus. De plus, nous discuterons des améliorations proposées afin d'optimiser davantage la qualité des données.

### 5.1 Complétude

#### 5.1.1 Objectif

L'étude de la complétude vise à garantir la présence de toutes les données nécessaires pour des analyses fiables et cohérentes. Elle permet d'identifier et de corriger les valeurs manquantes, réduisant ainsi les biais et erreurs dans les résultats. La complétude assure l'intégrité des analyses, facilite la prise de décisions justes, et respecte les normes de qualité.

#### 5.1.2 Table et Attribut Cible

La table cible est **S4\_Mesures**, et l'attribut étudié est **Taux\_releve**. Ce choix s'explique par l'importance de cet attribut pour analyser les niveaux de pollution mesurés par différents capteurs. La complétude de cette donnée est essentielle pour garantir l'exactitude des analyses environnementales. Les valeurs manquantes ou incohérentes dans **Taux\_releve** pourraient compromettre les calculs de moyennes ou les rapports statistiques. Améliorer cet attribut permet d'assurer des décisions fiables basées sur des données complètes et cohérentes.

#### 5.1.3 Description

##### a. Évaluation de la Complétude:

L'évaluation de la complétude vérifie si toutes les valeurs de **Taux\_releve** dans la table **S4\_Mesures** sont présentes. Cela signifie repérer les données manquantes ou nulles. Si une valeur est absente pour un type de polluant donné (**ID\_Polluant**), cela est considéré comme une donnée incomplète.

##### b. Métrique Utilisée

Le choix de cette métrique repose sur sa capacité à mesurer la proportion de lignes contenant une valeur pour l'attribut **Taux\_releve** par rapport au nombre total de lignes dans la table. Cette méthode offre une évaluation précise et quantitative du niveau de complétude des données.

$$\text{Pourcentage de Complétude} = \left( \frac{\text{nombre d'enregistrements complets}}{\text{nombre total d'enregistrements}} \right) \times 100 \quad (1)$$

Cette formule permet d'identifier clairement le pourcentage de complétude, constituant ainsi une base essentielle pour proposer et mettre en œuvre des améliorations adaptées.

### c. Amélioration Proposée

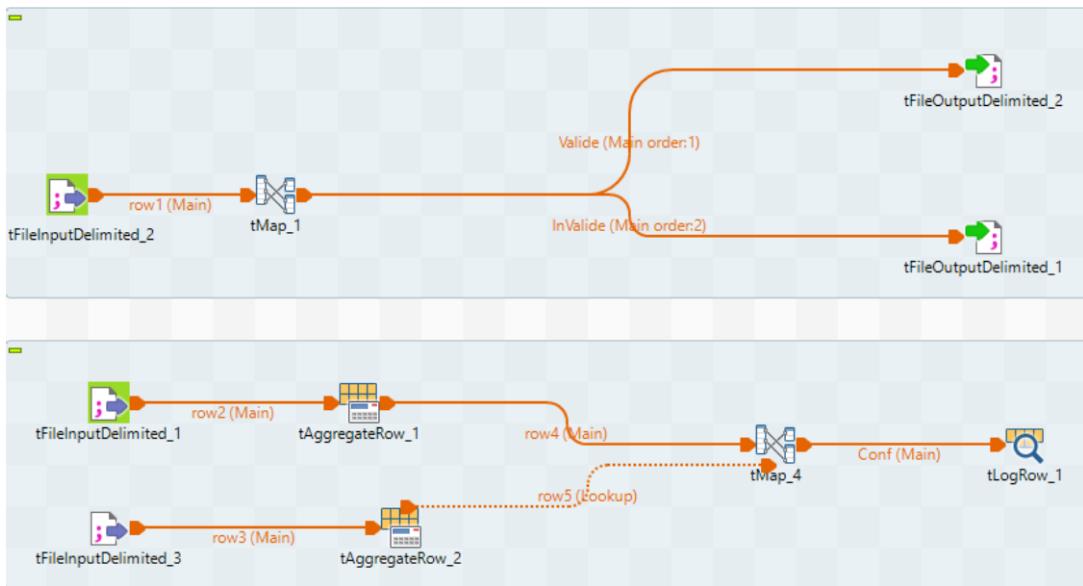
Pour améliorer la complétude, les valeurs manquantes de **Taux\_releve** dans la table **S4\_Mesures** sont remplacées par la moyenne des valeurs existantes pour chaque **ID\_Polluant**. La moyenne est utilisée car elle offre une valeur représentative de l'ensemble des données disponibles, garantissant ainsi une cohérence statistique dans la colonne. Cette méthode permet de préserver les tendances observées pour chaque type de polluant, tout en réduisant l'impact des données manquantes. Elle améliore la qualité globale des données, rendant celles-ci plus fiables et exploitables pour des analyses futures.

#### 5.1.4 Implémentation de la Complétude

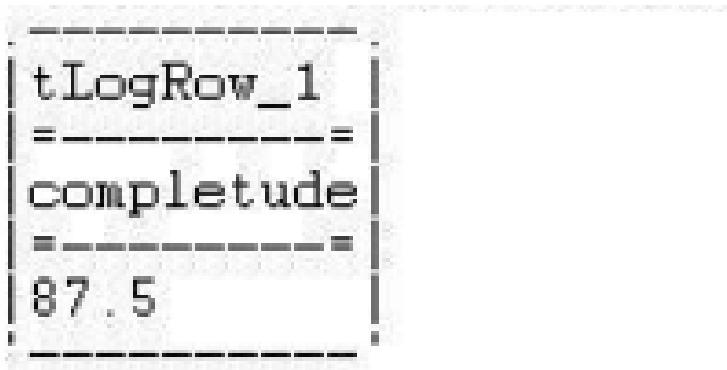
a. **Implémentation de l'Évaluation** L'évaluation de la complétude dans Talend commence par :

- Chargement des données de la table **S4\_Mesures** avec le composant **tFileInput-Delimited**, incluant l'attribut **Taux\_releve** et l'identifiant **ID\_Polluant**.
- Utilisation du composant **tMap** pour identifier les enregistrements où **Taux\_releve** est manquant ou vide.
- Séparation des enregistrements en deux flux distincts : un flux pour les enregistrements complets et un autre pour les enregistrements incomplets.
- Calcul de la proportion des enregistrements complets à l'aide du composant **tAggregateRow**, en divisant le nombre d'enregistrements complets par le nombre total d'enregistrements.
- Génération d'une métrique quantitative représentant le pourcentage de complétude des données.

La figure ci-dessous montre l'implémentation de ce facteur :



La complétude des données de la table **S4\_Mesures** a été évaluée à **87,5 %**, indiquant que la majorité des enregistrements de l'attribut **Taux\_releve** sont présents. Une telle complétude reflète un bon niveau de qualité des données, bien qu'il reste encore des valeurs manquantes à traiter. Cela souligne l'importance d'une amélioration ciblée pour maximiser l'utilité des données pour les analyses futures.



**[statistics] disconnected**

## b. Implémentation de l'Amélioration

Pour améliorer la complétude des taux relevés dans la table **S4\_Mesures**, nous avons envisagé de remplacer les valeurs nulles par la moyenne des valeurs connues pour le même polluant.

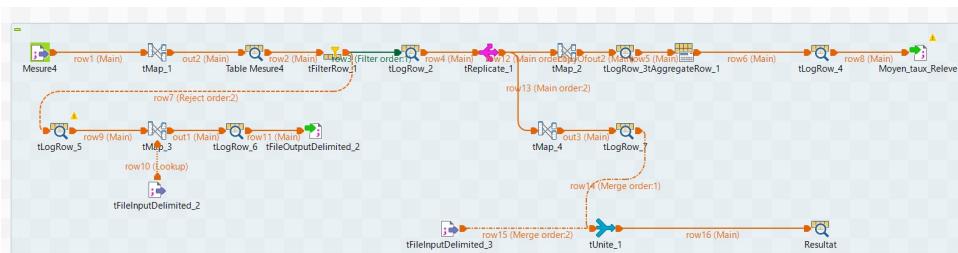
### Problème rencontré :

Lors de l'implémentation de cette solution, nous avons rencontré des difficultés liées au type des valeurs de **Taux\_releve** et à la conversion entre les types **String** et **Float**.

## Processus d'amélioration :

- **Extraction des données valides :** Dans un premier temps, nous avons extrait les tuples ne contenant pas de valeurs nulles à l'aide d'un **tFilterRow**. Ensuite, nous avons calculé la moyenne des taux relevés par polluant à partir des données filtrées.
- **Traitement des données incomplètes :** Nous avons extrait les tuples contenant des valeurs nulles et utilisé un **tMap** pour les compléter avec les moyennes calculées lors de la phase précédente.
- **Union des résultats intermédiaires :** Enfin, nous avons effectué une union entre les deux résultats intermédiaires pour obtenir le résultat final.

La figure suivante illustre le processus complet :



## Résultat :

- La première figure montre le contenu de la table **S4\_Mesures** avant le processus d'amélioration. Nous remarquons la présence de valeurs nulles pour le polluant **CO**.

| Table Mesure4 |                  |            |                     |             |        |
|---------------|------------------|------------|---------------------|-------------|--------|
| ID_Polluant   | Date             | ID_Capteur | Localisation        | Taux_Releve | ISNull |
| NO2           | 01/10/2021 12:00 | Can21      | 12.29               | 98.4        | 0      |
| PM10          | 07/10/2021 12:00 | Can21      | 48.886933, 2.334836 | 29.4        | 0      |
| CO            | 07/10/2021 12:01 | Can21      |                     | 3428.9      | 0      |
| CO            | 09/10/2021 12:00 | Can21      | 48.832914, 2.365659 |             | 1      |
| CO            | 01/10/2021 12:00 | Can23      | 48.833868, 2.321821 | 4523        | 0      |
| O3            | 01/10/2021 12:00 | Can23      | 48.803413, 2.147692 | 199.4       | 0      |
| O3            | 03/10/2021 18:00 | Can23      | 48.859066, 2.279786 | 38          | 0      |
| O3            | 03/10/2021 21:00 | Can23      | 48.792784, 2.108038 | 145.7       | 0      |
| O3            | 03/10/2021 23:00 | Can23      | 48.811440, 2.138852 | 234.65      | 0      |
| PM10          | 07/10/2021 12:00 | Can23      | 48.875044, 2.319998 | 23.89       | 0      |
| NO2           | 06/10/2021 12:00 | Can34      |                     | 95.8        | 0      |
| CO            | 02/10/2021 15:00 | Can45      | 67                  | 4789.67     | 0      |
| PM10          | 02/10/2021 15:00 | Can45      | 48.841419, 2.307009 | 23.4        | 0      |
| CO            | 02/10/2021 18:00 | Can45      | 48.836561, 2.291216 |             | 1      |
| NO2           | 08/10/2021 12:00 | Can45      | 48.885797, 2.351305 | 124.35      | 0      |

- La deuxième figure montre le contenu de la table **S4\_Mesures** après le processus d'amélioration, avec les valeurs nulles remplacées par les moyennes calculées.

| Résultat    |                  |            |                     |             |
|-------------|------------------|------------|---------------------|-------------|
| ID_Polluant | Date             | ID_Capteur | Localisation        | Taux_Relevé |
| NO2         | 01/10/2021 12:00 | Can21      | 2.29                | 98.4        |
| PM10        | 07/10/2021 12:00 | Can21      | 48.886933, 2.334836 | 29.4        |
| CO          | 07/10/2021 12:01 | Can21      |                     | 3428.9      |
| CO          | 01/10/2021 12:00 | Can23      | 48.833868, 2.321821 | 4523.0      |
| O3          | 01/10/2021 12:00 | Can23      | 48.803413, 2.147692 | 199.4       |
| O3          | 03/10/2021 18:00 | Can23      | 48.859066, 2.279786 | 38.0        |
| O3          | 03/10/2021 21:00 | Can23      | 48.792784, 2.108038 | 145.7       |
| O3          | 03/10/2021 23:00 | Can23      | 48.811440, 2.138852 | 234.65      |
| PM10        | 07/10/2021 12:00 | Can23      | 48.875044, 2.319998 | 23.89       |
| NO2         | 06/10/2021 12:00 | Can34      |                     | 95.8        |
| CO          | 02/10/2021 15:00 | Can45      | 67                  | 4789.67     |
| PM10        | 02/10/2021 15:00 | Can45      | 48.841419, 2.307009 | 23.4        |
| NO2         | 08/10/2021 12:00 | Can45      | 48.885797, 2.351305 | 124.35      |
| CO          | 09/10/2021 12:00 | Can21      | 48.832914, 2.365659 | 4247.19     |
| CO          | 02/10/2021 18:00 | Can45      | 48.836561, 2.291216 | 4247.19     |

## 5.2 Détection et Élimination des Doublons

### 5.2.1 L'Objectif

Il est essentiel dans notre étude d'évaluer la détection et l'élimination des doublons afin de garantir l'unicité et la fiabilité des données utilisées. Les doublons peuvent entraîner des analyses biaisées, fausser les résultats et compliquer les processus décisionnels. En identifiant et en supprimant les doublons, nous visons à améliorer la cohérence et la précision des données, tout en optimisant leur exploitation dans les étapes ultérieures de notre projet.

### 5.2.2 Tableau et Colonnes Cibles

Pour évaluer ce facteur, nous avons choisi d'étudier les tuples résultant de l'union des fichiers **S1.Station** et **S2.Station**. Cette table cible a été sélectionnée car l'évaluation des doublons est généralement réalisée lors de l'union de tables, où les données provenant de différentes sources doivent être consolidées. Nous nous concentrerons sur l'attribut **Contact\_Mail** pour identifier les doublons, car cet attribut est pertinent pour vérifier si des enregistrements distincts dans les deux fichiers font référence à la même entité.

### 5.2.3 Méthodologie

#### a. Description de l'Évaluation

L'approche adoptée repose sur une logique basée sur la connaissance du domaine, utilisant des règles prédéfinies. La règle consiste à considérer que deux tuples sont des doublons s'ils partagent le même **Contact\_Mail** et le même **Numéro de Téléphone**, avec l'identité comme fonction de comparaison.

Notre méthodologie se décompose en trois étapes principales :

- **Calcul du Nombre de Doublons** : Après l'union des tables **S1.Station** et **S2.Station**, nous avons supprimé les tuples où **Contact\_Mail** est vide. Ensuite, les tuples ont été regroupés par **Contact\_Mail**, et le nombre d'occurrences pour chaque email a été compté. Les tuples avec un compte supérieur à 1 ont été identifiés comme doublons.
- **Calcul du Nombre Total de Lignes** : Le nombre total de tuples issus de l'union des deux tables a été comptabilisé.
- **Calcul du Pourcentage de Doublons** : Le pourcentage de doublons a été déterminé en utilisant la formule suivante :

$$\text{Pourcentage de Doublons} = \left( \frac{\text{Nombre de Doublons DéTECTÉS}}{\text{Nombre Total d'Enregistrements}} \right) \times 100 \quad (2)$$

#### b. Métrique Utilisée

La métrique utilisée est le pourcentage de doublons. Cette mesure permet de quantifier l'impact des doublons sur la qualité des données et de fournir une base pour les améliorations à apporter.

#### c. Amélioration Proposée

Après l'union des tables **Station1** et **Station2**, plusieurs doublons ont été identifiés. Nous avons initialement envisagé de conserver l'enregistrement contenant le moins de valeurs nulles. Cependant, les doublons détectés ne contenaient pas de valeurs nulles. Par conséquent, nous avons décidé de ne conserver que les enregistrements de la source **S1**, jugée plus fiable.

### 5.2.4 Implémentation

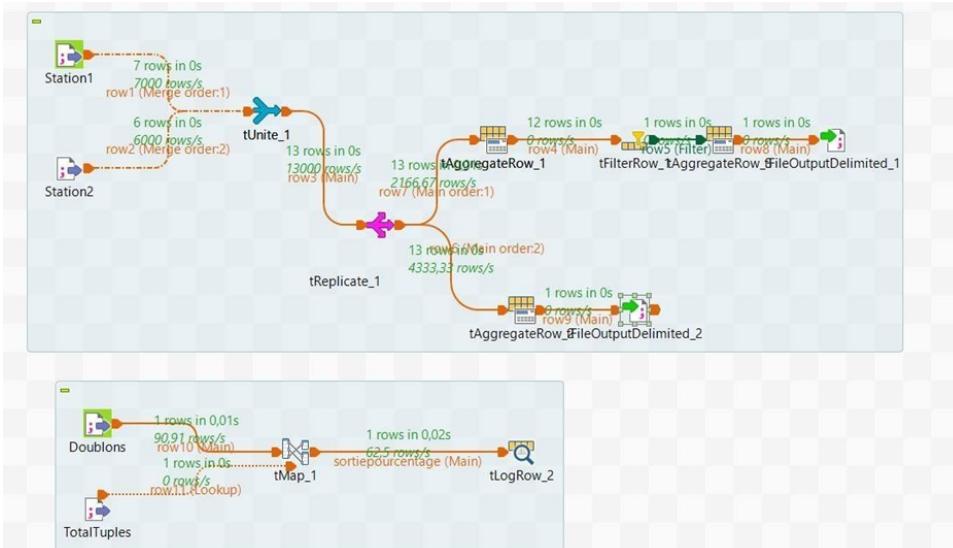
#### a. Implémentation de l'Évaluation

L'implémentation de la détection des doublons a été réalisée de la manière suivante:

- Nous avons utilisé le composant **tUnite** pour effectuer l'union des tables **Station1** et **Station2**.
- Pour calculer le nombre de doublons, nous avons employé un **tAggregateRow** afin de compter le nombre de lignes ayant le même **Contact\_Mail** et de les regrouper par **Contact\_Mail**. Ensuite, nous avons appliqué un **tFilterRow** pour ne conserver que les doublons (c'est-à-dire les lignes dont le comptage est supérieur à 1), puis nous avons calculé la somme des tuples résultants avec un autre **tAggregateRow**.

- Par ailleurs, nous avons utilisé un autre **tAggregateRow** pour calculer le nombre total de tuples provenant de l'union des deux tables.
- Enfin, la dernière étape de l'implémentation a consisté à utiliser un **tMap** pour calculer le pourcentage de doublons selon la formule précédemment définie.

La figure ci dessous montre l'implémentation de ce facteur :



Le pourcentage de doublons obtenu après l'union des tables **S1.Station** et **S2.Station** est de **7,69%**, indiquant une faible présence de doublons, ce qui laisse place à une optimisation.

Exécution

▶ Exécuter ■ Arrêter ✖ Effacer

```

Démarrage du Job Evaluation_Doublon à 16:59 10/11/2024.
[statistics] connecting to socket on port 3945
[statistics] connected
-----
| tLogRow_2 |
|-----|
Pourcentage
|-----|
7.692308
|-----|
[statistics] disconnected
Job Evaluation_Doublon terminé à 16:59 10/11/2024. [Code de sortie = 0]

```

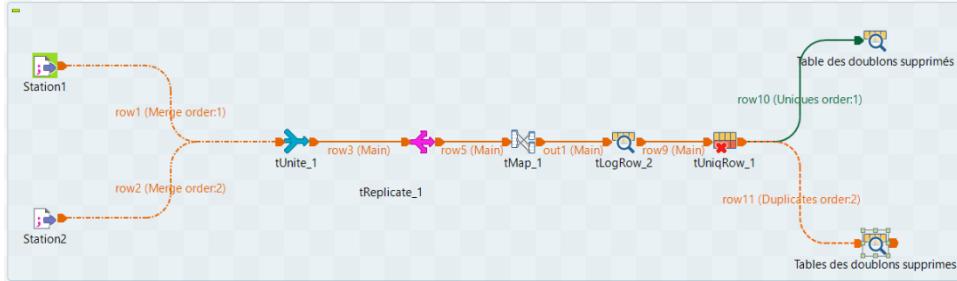
## b. Implémentation de l'Amélioration

L'implémentation de l'amélioration a suivi le processus suivant :

- Nous avons utilisé un **tMap** pour ajouter, à chaque ligne de l'union des tables Station1 et Station2, une colonne indiquant le nombre de valeurs nulles présentes dans chaque ligne.

- Ensuite, le composant **tUniqRow** a permis de supprimer les doublons de la source 2, générant ainsi deux sorties : la première est la table sans doublons, et la deuxième contient toutes les lignes de doublons supprimées.

La figure ci dessous montre l'implémentation de ce facteur :



### Résultat :

- La première figure montre la table **Station1 union Station2** après suppression des doublons.
- La deuxième figure illustre les enregistrements supprimés.

| Table des doublons supprimés |        |                        |             |  |
|------------------------------|--------|------------------------|-------------|--|
| ID_Station                   | Numero | Rue                    | Code_postal | Ville  |
| 1                            | 5      | Etant d'or             | 78120       | Rambouillet  À 01 75 03 40 00 rambouillet@airparif.fr 0  |
| 2                            | 8      | Limoges                | 78000       | Versailles  01 39 25 40 00  versailles@airparif.fr 0     |
| 3                            | 1      | Emile Zola             | 78200       | Mantes-la-Jolie  01 34 78 81 00  mantes@airparif.fr 0    |
| 4                            |        | Parvis de la Défense   | 92800       | Puteaux  01 46 92 92 92  ladefense@airparif.fr 1         |
| 5                            |        | Allée des Refusniks    | 75007       | Paris  01 53 58 75 07  paris07@airparif.fr 1             |
| 6                            | 2 bis  | Quai de la Magistrerie | 75001       | Paris  À 01 44 50 75  pari01@airparif.fr 0               |
| 7                            | 60     | Richelieu              | 92230       | Gennevillier  01 40 85 66 66  gennevillier@airparif.fr 0 |
| 1                            |        | Château_princeloup     | 78120       | Sonchamp  01 34 84 41 08  sonchamp@ 1                    |
| 3                            | 11     | Commandant Pilot       | 92200       | Neuilly-sur-seine 01 40 88 88 88  neuilly@airparif.fr 0  |
| 4                            |        | Quai de la Magistrerie | 75001       | À 01 44 50 75 01 paris01@airparif.fr 2                   |
| 5                            | 7      | Ferdinand Flocon       | 75018       | Paris  01 53 41 18 18  paris18@airparif.fr 0             |
| 6                            |        | Parvis de la Défense   | 92800       | Puteaux  01 46 92 92 92    2                             |

| Tables des doublons supprimés |        |           |             |  |
|-------------------------------|--------|-----------|-------------|--|
| ID_Station                    | Numero | Rue       | Code_postal | Ville  |
| 2                             | 60     | Richelieu | 92230       | Gennevillier 01 40 85 66 66 gennevillier@airparif.fr 0 |

## 5.3 Hétérogénéité des Échelles et de la Granularité

### 5.3.1 L'objectif

Il est crucial dans notre étude d'évaluer le facteur de l'**hétérogénéité des échelles** afin de garantir que les données soient représentées sur une échelle uniforme. Cela permet d'éviter d'introduire des erreurs de calcul ou d'interprétation ultérieures, qui pourraient

fauisser les résultats des analyses. En harmonisant les échelles des différentes variables, nous assurons une meilleure comparabilité et cohérence des données, ce qui est essentiel pour obtenir des conclusions fiables et précises dans les étapes suivantes de notre projet.

### 5.3.2 Table cible et colonnes

L'hétérogénéité des échelles se produit lorsque le même attribut est représenté différemment dans deux tables. Dans notre étude, nous avons choisi d'évaluer l'attribut **Taux\_Relevé** dans les tables Mesures des sources **S4** et **S5**, car il constitue la mesure clé de notre projet. Il est donc essentiel d'assurer l'homogénéité de cet attribut entre les deux sources afin d'éviter toute incohérence dans nos analyses et garantir la fiabilité des résultats.

### 5.3.3 Méthodologie

Pour étudier l'hétérogénéité des échelles, nous décrirons la méthodologie utilisée pour identifier et évaluer les différences d'échelle. Nous aborderons également la métrique choisie pour quantifier l'impact de ces différences et les améliorations proposées pour harmoniser les données et améliorer leur qualité.

#### a. Description de l'Évaluation

Notre approche de calcul se divise en trois étapes :

- **Calcul de la moyenne des Taux\_Relevé dans la table Mesure4 :**

Pour calculer la moyenne de Taux\_Releve dans la table Mesure4, sélectionnez les lignes valides (non nulles et non vides), additionnez les valeurs valides, puis divisez la somme par le nombre de lignes valides.

- **Calcul de la moyenne des Taux\_Relevé dans la table Mesure5 :**

Comme pour Mesure4, sélectionnez les lignes valides dans la table Mesure5 (où Taux\_Releve est non nul et non vide), additionnez les valeurs valides de Taux\_Releve, puis divisez cette somme par le nombre de lignes valides pour obtenir la moyenne des Taux\_Releve dans Mesure5.

- **Calcul de la différence et du pourcentage de différence :**

Pour évaluer l'hétérogénéité entre les tables MesureS4 et MesureS5, il convient de calculer la différence entre les moyennes des Taux\_Releve de chaque table. Ensuite, pour mieux appréhender cette différence en termes relatifs, il est nécessaire de calculer un pourcentage de différence entre ces moyennes. Ce pourcentage permet d'obtenir une estimation de l'écart entre les deux tables et d'évaluer si la différence est significative.

#### b. Métrique utilisée

La métrique pour évaluer l'hétérogénéité des données se résume par les étapes suivantes : d'abord, nous calculons la moyenne des 'Taux relevé' pour les tables S4 et S5, respectivement notées Moyenne\_Taux\_ReleveS4 et Moyenne\_Taux\_ReleveS5. Ensuite, nous calculons la différence entre ces deux moyennes :

$$\text{difference} = \text{Moyenne\_Taux\_ReleveS4} - \text{Moyenne\_Taux\_ReleveS5} \quad (3)$$

Enfin, le pourcentage de cette différence est déterminé à l'aide de la formule suivante :

$$\text{pourcentage\_diff} = \left| \frac{\text{difference}}{\text{Moyenne\_Taux\_ReleveS5}} \right| \times 100 \quad (4)$$

Cette approche nous permet de quantifier l'écart entre les deux échelles et d'évaluer l'impact de cette hétérogénéité.

### c. Description d'amélioration

Nous avons constaté que les ensembles de données Mesure4 et Mesure5 mesurent des valeurs similaires, mais sur des échelles différentes. Pour harmoniser ces données, nous avons appliqué un processus de normalisation aux valeurs de Taux\_Relevé dans Mesure5, les ajustant ainsi à l'échelle de Mesure4. Cette normalisation permet une comparaison cohérente entre les deux ensembles de données.

#### 5.3.4 Implémentation

Dans cette partie, nous aborderons deux aspects de l'implémentation : d'une part, l'implémentation de l'évaluation de ce facteur pour le mesurer, et d'autre part, l'implémentation des améliorations proposées pour optimiser la qualité des données.

##### a. Implementation d'évaluation

L'implémentation pour mesurer de lhétéroginité entre Mesure4 et Mesure5 a été réalisée de la manière suivante :

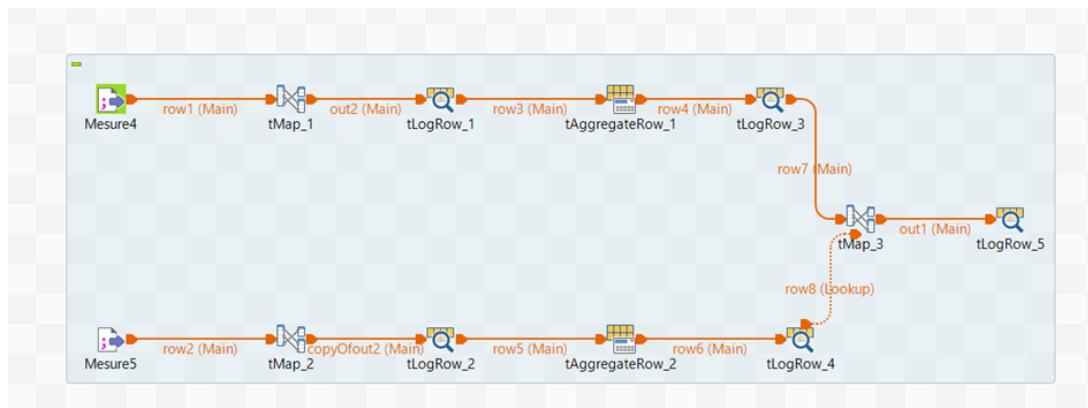
- Utiliser un **tAggregate** pour calculer la moyenne des taux relevés dans la table Mesure4 pour chaque id\_polluant, en excluant les valeurs nulles.
- Utiliser un tAggregate pour calculer la moyenne des taux relevés dans la table Mesure5 pour chaque id\_polluant, en excluant les valeurs nulles.
- Utilisation d'un **tmap** pour Calcul de la différence entre les moyennes ( les resultat intermidiaire calculer précédemment:

$$\text{difference} = \text{Moyenne\_Taux\_ReleveS4} - \text{Moyenne\_Taux\_ReleveS5} \quad (5)$$

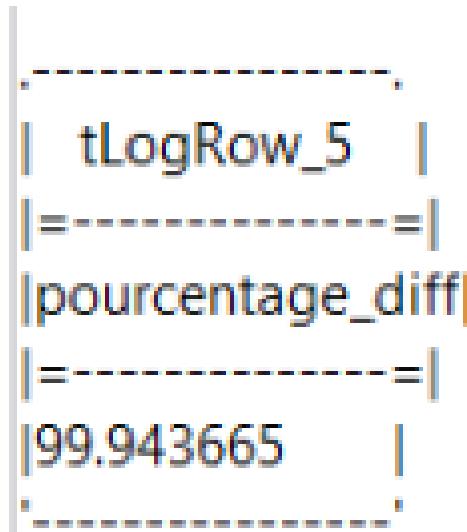
et du pourcentage de différence

$$\text{pourcentage\_diff} = \left| \frac{\text{difference}}{\text{Moyenne\_Taux\_ReleveS5}} \right| \times 100 \quad (6)$$

La figure ci dessous montre l'implémentation de ce facteur :

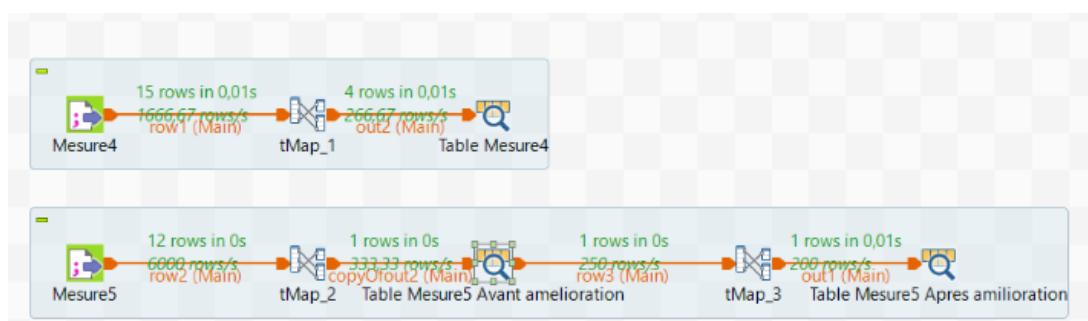


Une différence observée de 99,9 entre les moyennes des Taux\_Releve dans les deux tables (MesureS4 et MesureS5) est extrêmement élevée, ce qui suggère une grande hétérogénéité entre les deux ensembles de données.



## b. Implémentation de l'Amélioration

Pour améliorer l'hétérogénéité des données, nous avons appliqué un processus de normalisation visant à ajuster les valeurs de taux relevé de la table Mesure5 afin qu'elles soient sur la même échelle que celles de Mesure4.



## Résultats

Le résultat ci-dessous montre la table Mesure5 avant et après amélioration. Comme on peut le voir, les valeurs de Taux\_Releve sont maintenant plus proches de celles de la table Mesure4.

Nous pouvons donc conclure que l'hétérogénéité des échelles a été améliorée.

| [statistics] connected |                  |            |                     |             |
|------------------------|------------------|------------|---------------------|-------------|
| Table Mesure4          |                  |            |                     |             |
| ID_Polluant            | Date             | ID_Capteur | Localisation        | Taux_Releve |
| O3                     | 01/10/2021 12:00 | Can23      | 48.803413, 2.147692 | 199.4       |
| O3                     | 03/10/2021 18:00 | Can23      | 48.859066, 2.279786 | 38.0        |
| O3                     | 03/10/2021 21:00 | Can23      | 48.792784, 2.108038 | 145.7       |
| O3                     | 03/10/2021 23:00 | Can23      | 48.811440, 2.138852 | 234.65      |

| Table Mesure5 Avant amelioration |                  |            |                     |             |
|----------------------------------|------------------|------------|---------------------|-------------|
| ID_Polluant                      | Date             | ID_Capteur | Localisation        | Taux_Releve |
| O3                               | 02/10/2021 15:00 | Cairs22    | 48.837013, 2.239375 | 0.087       |

| Table Mesure5 Apres amilioration |                  |            |                     |             |
|----------------------------------|------------------|------------|---------------------|-------------|
| ID_Polluant                      | Date             | ID_Capteur | Localisation        | Taux_Releve |
| O3                               | 02/10/2021 15:00 | Cairs22    | 48.837013, 2.239375 | 87.0        |

## 5.4 Conformité

### 5.4.1 L'objectif

L'étude de la conformité vise à s'assurer que les données respectent un format ou une codification prédéfinis. Elle permet d'identifier et de corriger les valeurs non conformes, réduisant ainsi les erreurs et incohérences dans les données. La conformité garantit l'exactitude des informations, facilite les analyses fiables et renforce la confiance dans la qualité des données utilisées.

### 5.4.2 Table et colonnes cibles

La conformité est analysée sur l'attribut **Contact\_Mail** de la table **Station2**. L'objectif est de vérifier que les valeurs de cette colonne respectent le format défini par l'expression régulière '**[a-zA-Z0-9]+@airparif.fr\$**'. Les adresses email non conformes sont identifiées et remplacées par des valeurs corrigées respectant le format attendu.

### 5.4.3 Description

#### a. Description d'évaluation

L'évaluation de la conformité vise à vérifier si les valeurs de la colonne **Contact\_Mail** de la table **Station2** respectent un format défini pour les adresses e-mail.

Les adresses e-mail doivent respecter l'expression régulière '**[a-zA-Z0-9]+@airparif.fr\$**', indiquant un identifiant alphanumérique suivi du domaine **@airparif.fr**. À l'aide de cette expression, chaque valeur de la colonne **Contact\_Mail** est évaluée pour déterminer si elle respecte le format attendu. Les adresses conformes sont séparées des non conformes pour une meilleure analyse et correction. Le pourcentage d'adresses incorrectes est calculé pour mesurer l'ampleur du problème.

#### b. Métrique utilisée

Pour évaluer la conformité des adresses e-mail dans la table **Station2**, la métrique choisie est le pourcentage de conformité, qui se calcule comme suit :

$$\text{Pourcentage de conformité} = \left( \frac{\text{nombre d'e-mails conformes}}{\text{nombre total d'e-mails}} \right) \times 100 \quad (7)$$

Cette métrique offre une mesure claire et compréhensible du niveau de conformité des données. Elle permet de quantifier précisément la proportion des e-mails respectant le format attendu.

#### c. Description d'amélioration

Pour améliorer la conformité des adresses e-mail dans la table **Station2**, l'approche adoptée consiste à corriger les e-mails invalides en générant une version conforme. À partir de l'attribut **Ville**, la première partie de l'e-mail est récupérée. Par exemple, pour une ville "Paris01", la valeur "Paris01" sera utilisée comme base. Le domaine fixe **@airparif.fr** est ajouté à la partie extraite, formant ainsi une adresse e-mail conforme au format attendu (exemple : **Paris01@airparif.fr**). Les adresses e-mail qui ne respectent pas le format requis sont remplacées par ces nouvelles adresses générées.

#### Avantages de cette approche :

- **Automatisation** : Elle permet de corriger les adresses non conformes sans intervention manuelle.
- **Homogénéité** : Toutes les adresses e-mail suivent désormais un format uniforme, assurant une meilleure qualité des données.

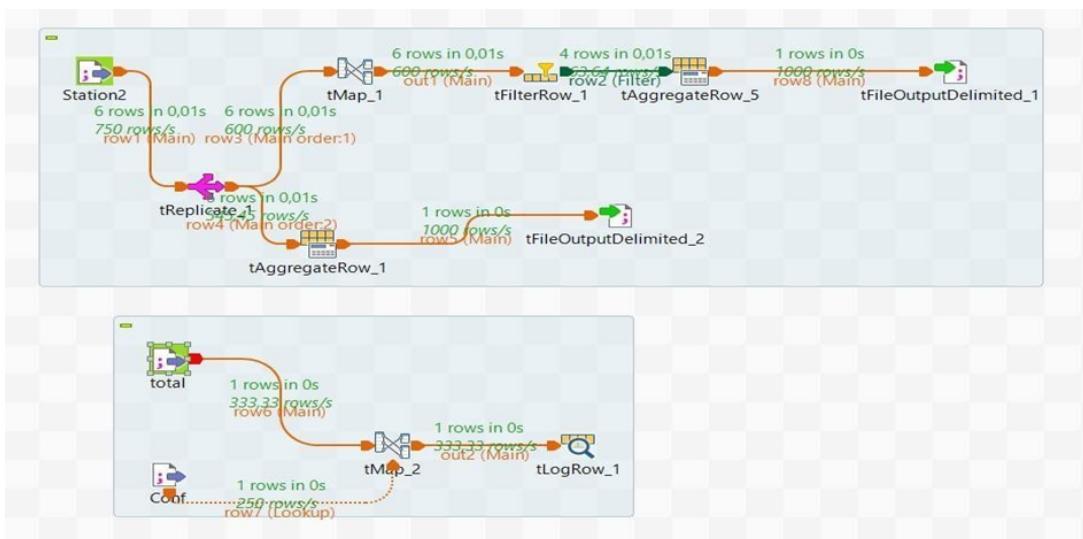
### 5.4.4 Implémentation

#### a. Implémentation d'évaluation

L'évaluation de la conformité a été appliquée aux données de la table **S2\_Station**, en utilisant l'expression régulière '**[a-zA-Z0-9]+@airparif.fr\$**' pour vérifier si l'e-mail est dans le bon format ou non.

- **Détermination du nombre de tuples conformes** : Nous utilisons un **tMap** pour appliquer une expression régulière. Cette expression vérifie la conformité des tuples en renvoyant 1 si le tuple est valide et 0 dans le cas contraire. Le **tMap** est ensuite relié à un **tFilterRow** qui filtre les enregistrements en ne conservant que ceux dont la conformité est égale à 1. Enfin, un **tAggregate** permet de calculer le nombre total de tuples conformes.
- **Calcul du nombre total de lignes dans la table S2\_Station** : Le nombre total de tuples est déterminé à l'aide d'un **tAggregate**.
- **Calcul du pourcentage de conformité** : Le pourcentage de conformité est calculé grâce à un **tMap**.

La figure ci-dessous montre l'implémentation de ce facteur :



Le pourcentage de conformité de 66,66 % obtenu sur la table **S2\_Station** signifie qu'environ 66,66 % des enregistrements dans cette table respectent les critères de conformité définis. Un pourcentage de conformité de 66,66 % peut être considéré comme un résultat mitigé. Cela signifie que plus d'un tiers des enregistrements ne sont pas conformes, ce qui pourrait indiquer qu'il y a encore des problèmes de qualité des données à résoudre.

```
Démarrage du Job Conformite
[statistics] connecting to ...
[statistics] connected
66.66667
[statistics] disconnected

Job Conformite terminé à l'
```

## b. Implémentation d'amélioration

La solution proposée consiste à identifier les e-mails qui ne respectent pas l'expression régulière '[a-zA-Z0-9]+@airparif.fr\$', puis à les remplacer par le bon **Contact\_Mail**. Pour cela, il faut récupérer la première partie de l'e-mail à partir de l'attribut **Ville**, puis la concaténer avec le domaine fixe @airparif.fr (soit **row2.Ville + '@airparif.fr'**). La figure ci-dessous montre la solution mise en œuvre.



Voici la table avant l'amélioration : on remarque que le premier et le dernier enregistrement contiennent des valeurs incorrectes pour l'attribut **Contact\_Mail**.

| tLogRow_1  |                                  |                          |                                      |                          |
|------------|----------------------------------|--------------------------|--------------------------------------|--------------------------|
| ID_Station | Numero Rue                       | Code_postal Ville        | Telephone                            | Contact_Mail             |
| 1          | Château_princeloup               | 78120  Sonchamp          | 01 34 84 41 08                       | sonchamp@airparif.fr     |
| 2          | 60  Richelieu                    | 92230  Gennevillier      | 01 40 85 66 66                       | gennevillier@airparif.fr |
| 3          | 11  Commandant Pilot             | 92200  Neuilly-sur-seine | 01 40 88 88 88                       | neuilly@airparif.fr      |
| 4          | Quai de la M&#233;gisserie 75001 |                          | À 01 44 50 75 01 Paris01@airparif.fr |                          |
| 5          | 7  Ferdinand Flocon              | 75018  Paris             | 01 53 41 18 18                       | paris18@airparif.fr      |
| 6          | Parvis de la D&#233;fense 92800  | Puteaux                  | 01 46 92 92 92                       |                          |

Après l'amélioration, toutes les adresses e-mail ont été correctement mises au bon format, comme le montre la figure suivante :

| tLogRow_2  |                                  |                          |                                      |                          |       |
|------------|----------------------------------|--------------------------|--------------------------------------|--------------------------|-------|
| ID_Station | Numero Rue                       | Code_postal Ville        | Telephone                            | Contact_Mail             | count |
| 1          | Château_princeloup               | 78120  Sonchamp          | 01 34 84 41 08                       | Sonchamp@airparif.fr     | 0     |
| 2          | 60  Richelieu                    | 92230  Gennevillier      | 01 40 85 66 66                       | gennevillier@airparif.fr | 1     |
| 3          | 11  Commandant Pilot             | 92200  Neuilly-sur-seine | 01 40 88 88 88                       | neuilly@airparif.fr      | 1     |
| 4          | Quai de la M&#233;gisserie 75001 |                          | À 01 44 50 75 01 Paris01@airparif.fr |                          | 1     |
| 5          | 7  Ferdinand Flocon              | 75018  Paris             | 01 53 41 18 18                       | paris18@airparif.fr      | 1     |
| 6          | Parvis de la D&#233;fense 92800  | Puteaux                  | 01 46 92 92 92                       | Puteaux@airparif.fr      | 0     |

## 6 Conclusion

Ce projet nous a permis d'évaluer et d'améliorer efficacement la qualité des données de pollution atmosphérique à travers différentes étapes de traitement dans Talend. Nous avons identifié des problèmes cruciaux tels que l'hétérogénéité des échelles, les valeurs manquantes, les doublons et la non-conformité des formats. Grâce à des méthodes rigoureuses d'évaluation et à des solutions de correction appropriées, nous avons harmonisé les données, augmenté leur complétude à 100%, éliminé les doublons et normalisé les valeurs pour garantir une meilleure cohérence.

Ces améliorations ont renforcé la fiabilité des données, posant une base solide pour des analyses environnementales futures. Elles permettent désormais une meilleure prévision des niveaux de pollution et facilitent des initiatives de réduction des émissions de polluants, contribuant ainsi à une meilleure prise de décision dans la gestion de la qualité de l'air.

L'utilisation de Talend a permis d'optimiser le flux de données et d'automatiser des tâches critiques, ce qui a conduit à une gestion plus efficace des données complexes et hétérogènes. En conclusion, ce projet a non seulement amélioré la qualité des données actuelles, mais a aussi ouvert la voie à des perspectives futures, telles que l'intégration de nouvelles sources de données et l'utilisation de techniques d'analyse avancées.