Université de Versailles Saint-Quentin-en-Yvelines
UFR des sciences
Département d'Informatique

Compte rendu

04 Novembre 2022

Projet Qualité des données

Encadré par
M. Zoubida KEDAD

Réalisé par
AOUCHICHE Salah
HADJ MAHFOUD Kenza
BOUDJEMAI Ali
BELHANNACHI Moundji Hocine

2022/2023

Table des matières

1 Introduction

Dans le cadre de nos études en deuxième année de Master DataScale et de l'UE de Qualité des données,Nous sommes amenés à étudier la qualité des données d'un jeu de données pour cela nous allons utilisé dans ce projet l'ETL **Talend**.

Nous avons à notre disposition 5 sources de données suivantes :

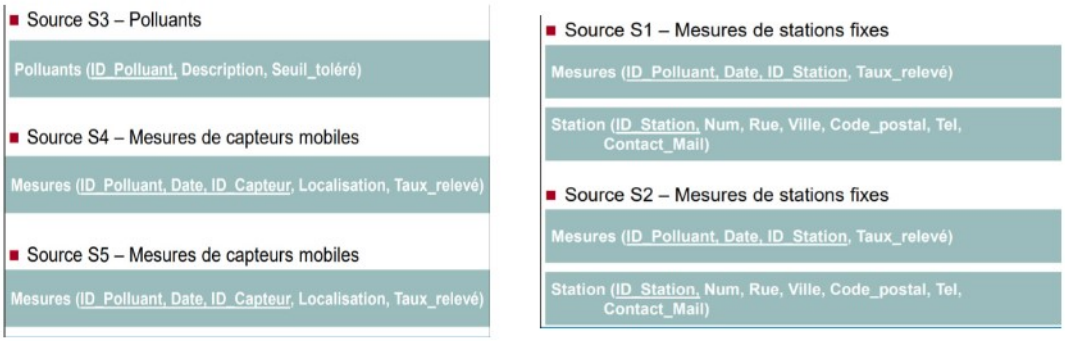

FIGURE 1 – Les sources

Nous avons aussi le schéma cible suivant :

FIGURE 2 – Les tables cibles

2 Définition des mapping :

Nous allons définir les mapping des tables cible.

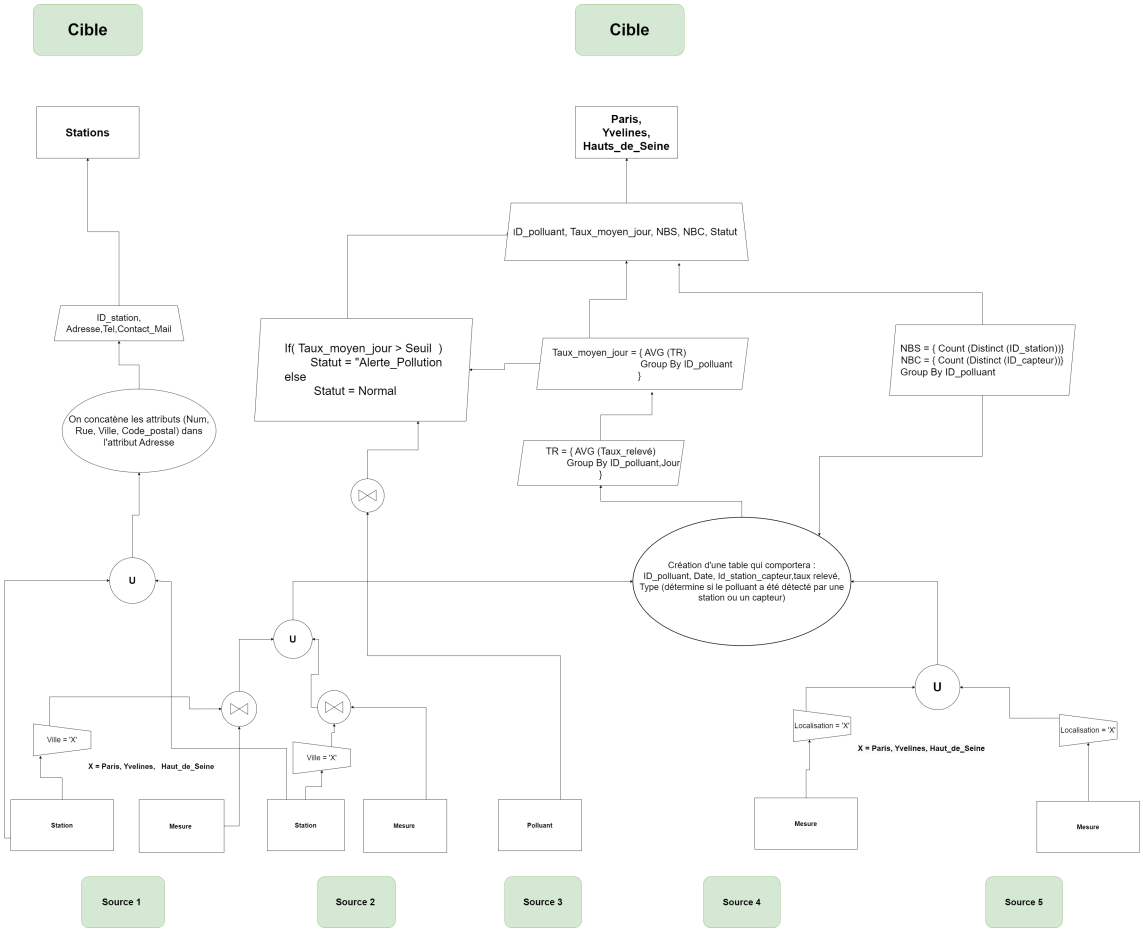

FIGURE 3 – Schéma des mapping des tables cibles

3 Implémention des mapping

Nous allons dans cette partie implémenté les mapping définit sur Talend.Nous avons à notre disposition deux cibles Stations et Département ainsi qu'un ensemble de sources. Dans cette section, nos données sont stockés dans des fichiers csv.

3.1 La cible Stations :

Nous avons implémenté le mapping de la table cible Stations.Pour cela nous avons fait l'union de deux tables stations de la source 1 et 2.

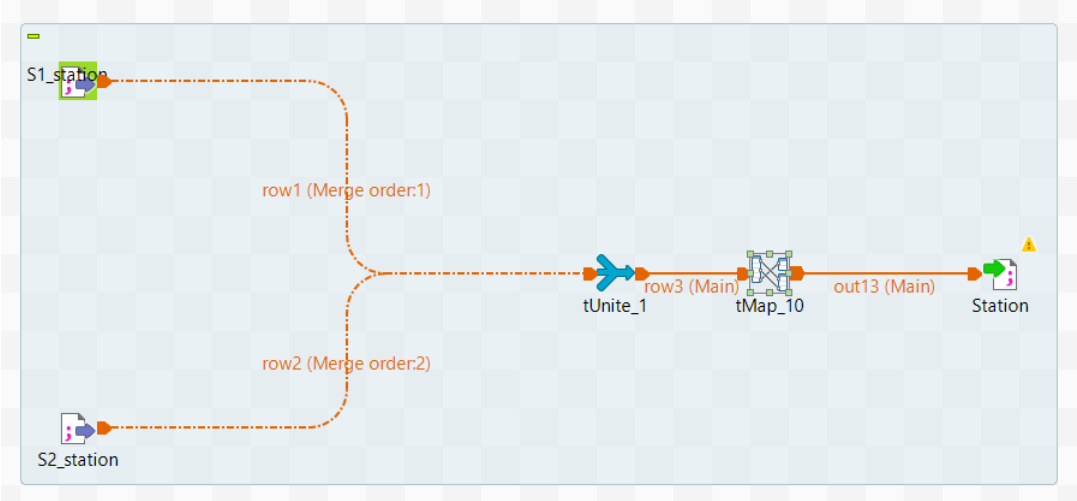

FIGURE 4 – Mapping de la table Stations

3.2 La cible Département :

Nous avons le mapping des tables cibles Paris, Yvelines, Haut-de-Seine ainsi que son implémention sur Talend. Pour cela nous avons suivi les étapes suivantes :

3.2.1 Étape 1 :

1. **Au niveau de la source 1 et 2** :
   Nous avons effectué des restrictions sur l'attribut Ville pour chaque table Station des sources. Ensuite, Nous avons fait une jointure pour chacune de ces tables avec la table Mesure.Ainsi, nous avons fait l'union des deux tables obtenues.Notons le résultat obtenu **table1**.

2. **Au niveau de la source 4 et 5** :
   Nous avons effectué des restrictions sur l'attribut localisation pour chaque table Mesure des sources.Ensuite, On fait l'union des deux résultats obtenues.Notons le résultat obtenu **table2**.

3. **Création de la table 3** :
   Cette table contiendra tous les tuples de la table 1 et 2,et comportera ID_Polluant, Date et Taux-Relevé, id_capteur_station et un attribut Type qui va différencier chaque tuple si il est détecté par un capteur ou une station (1 pour station et 0 pour capteur).

4. **À partir de table 3** :
   - Nous avons calculé le NBS et le NBC qui correspondent respectivement aux nombres de station utilisées et aux nombres de capteurs mobiles utilisés par ID_polluant.Ensuite on stocke les résultats de NBS dans un fichier "NBS.csv" et les résultats de NBC dans un fichier "NBC.csv".

   - Nous avons ensuite calculé la moyenne des Taux relevé par polluant puis par Jour qu'on va appelé TR, ensuite nous avons calculé la moyenne de TR par ID_polluant. Notons **Taux_moyen_jour** le résultat obtenu.

5. **À partir de Taux_moyen_jour** :
   En comparant le **Taux_moyen_jour** de chaque polluant avec son seuil récupéré à partir de la source 3, nous avons pu définir la valeur l'attribut **Statut** (Alerte Pollution, Normal). Ensuite, on stocke les résultats de Taux_moyen_jour et le statut de chaque polluant dans un fichier "taux_statut.csv".

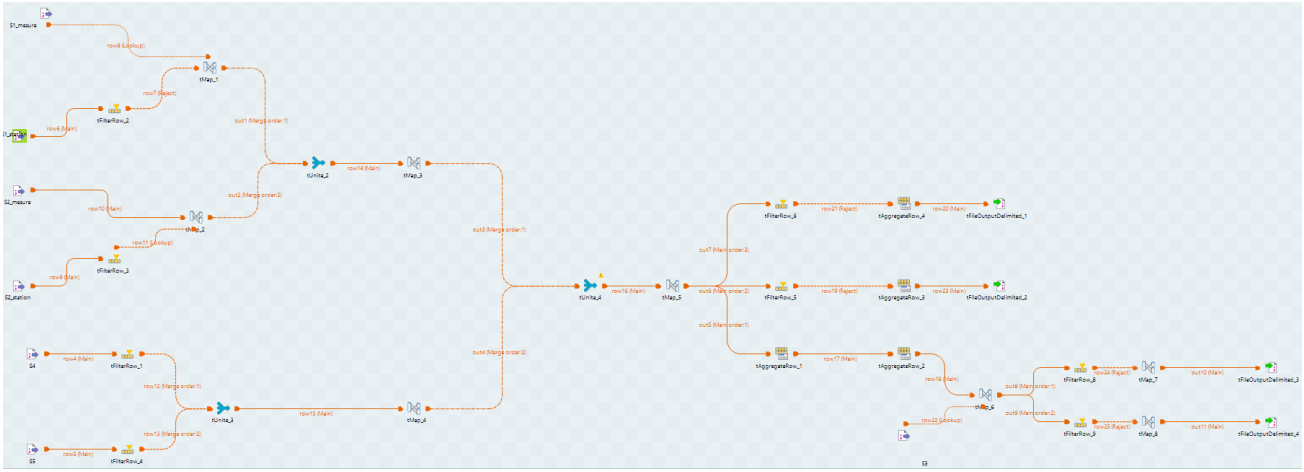

FIGURE 5 – Étape 1 de la création des Mapping (la cible département)

3.2.2 Étape 2 :

À partir des fichiers **NBS.csv**, **NBC.csv** et **taux_statut.csv**, Nous avons effectué une jointure de ces trois fichiers pour construire la cible finale (département).

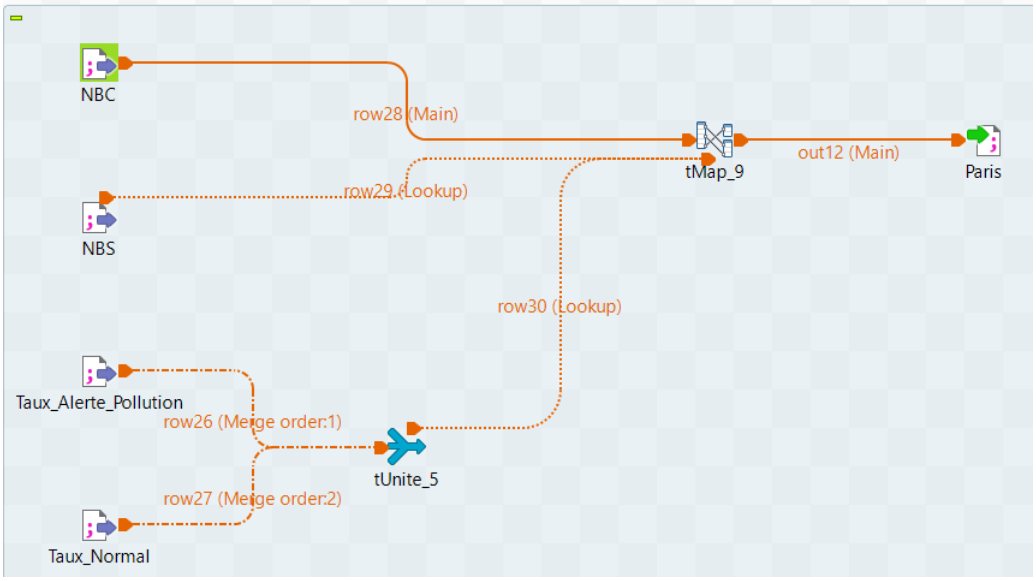

FIGURE 6 – Étape 2 de la création des Mapping (la cible département)

4 Facteurs de qualité de données et leurs métrique de mesure

4.1 Conformité à un format, une codification

Lors de notre implémentation sur Talend nous avons adopté la métrique qui consistait à calculer le rapport entre le nombre de tuples dont le format est incorrecte sur le nombre total de tuple, cela nous a permis d'avoir le taux de tuples correctes. La formule est la suivante :

Taux_evalué = 1 - [(Nbr_tuple_format_incorrect) / (nbr_tuple_total)]

4.1.1 Format d'un mail

Au niveau de la source **S2.station** on va définir une expression régulière sur l'attribut **Contact_Mail** pour exprimer le format d'un mail (exemple : xxxx@airparif.fr) et vérifier si l'adresse est écrite en minuscule.

4.1.2 Format du numéro de téléphone

Au niveau de la source **S1.station** et **S2.station** on va définir une expression régulière sur l'attribut **Téléphone** pour exprimer le format d'un numéro de téléphone (exemple : 01 xx xx xx xx).

4.1.3 Format de la localisation

Au niveau de la source **S4.station** et **S5.station**, l'attribut **Localisation**, doit contenir deux paramètres **longitude** et **latitude**. Par conséquent, ce dernier doit contenir deux float (exemple : A,B).

FIGURE 7 – Job d'évaluation de la conformité

4.2 Hétérogénéité des échelles, de la granularité

4.2.1 Hétérogénéité des échelles :

Au niveau des sources S1.mesure, S2.mesure et S4, nous avons fixé un intervalle pour chaque polluant comme hypothése de base :

- PM10 [250,20]
- PM2.5 [35,5]
- O3 [250,20]
- CO [12000,3000]

- NO2 [280,20]

Nous avons vérifié chaque valeur de Taux_releve si elle appartient à son intervalle ou pas, cela nous a permis d'avoir le taux de tuples correctes. La formule est la suivante :

Taux_evalué = 1 - [(Nbr_valeurs_en_dehors_intervalle) / (nbr_tuple_total

4.2.2 Granularité :

Au niveau de la source S5, on dit qu'il y a un problème de granularité dans le cas où la majorité des valeurs sont en dehors de leurs intervalle. Donc il faut normaliser les données pour qu'il prenne des valeurs qui vont avoir des ordres de grandeur similaires, cela nous a permis d'avoir le taux de tuples correctes. La formule est la suivante :

Taux_evalué = 1 - [(Nbr_valeurs_en_dehors_intervalle) / (nbr_tuple_total

- Si le taux des tuples correctes est inférieur à 60%, donc surement il y a un problème de granularité.

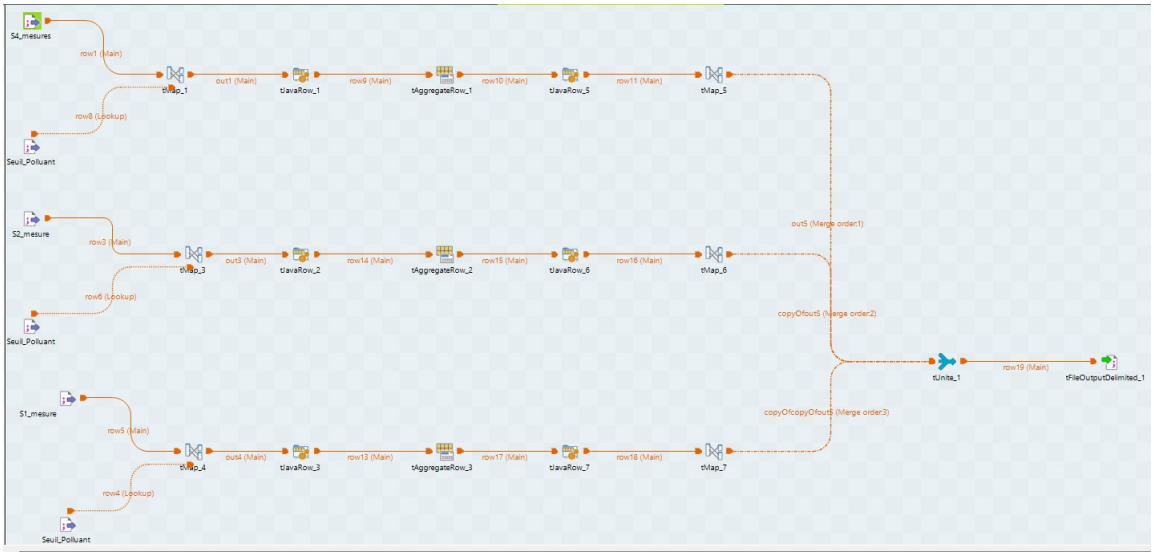

FIGURE 8 – Job d'évaluation de l'hétérogénéité des échelles, de la granularité

4.3 Complétude des données

La complétude des données se fera au niveau des tables sources suivantes :

- **S1.Station(Numero)**
- **S2.Station(Numero, Ville, Contact_Mail)**
- **S4.Mesure(Localisation, Taux_Releve)**
- **S5.Mesure(Taux_Releve)**

Nous allons essayer d'augmenter le taux de complétude de chaque attribut, en corrigeant les valeurs nulles.

- Lors de notre implémentation sur Talend nous avons adopté la métrique qui consistait à évaluer la complétude en calculant le rapport entre le nombre de tuples dont certaines valeurs d'attributs sont nulles sur le nombre total de tuple, cela nous a permis d'avoir le taux de tuples correctes. La formule est la suivante :

Taux_evalué = 1 - [(Nbr_tuple_attribut_null) / (nbr_tuple_total)]

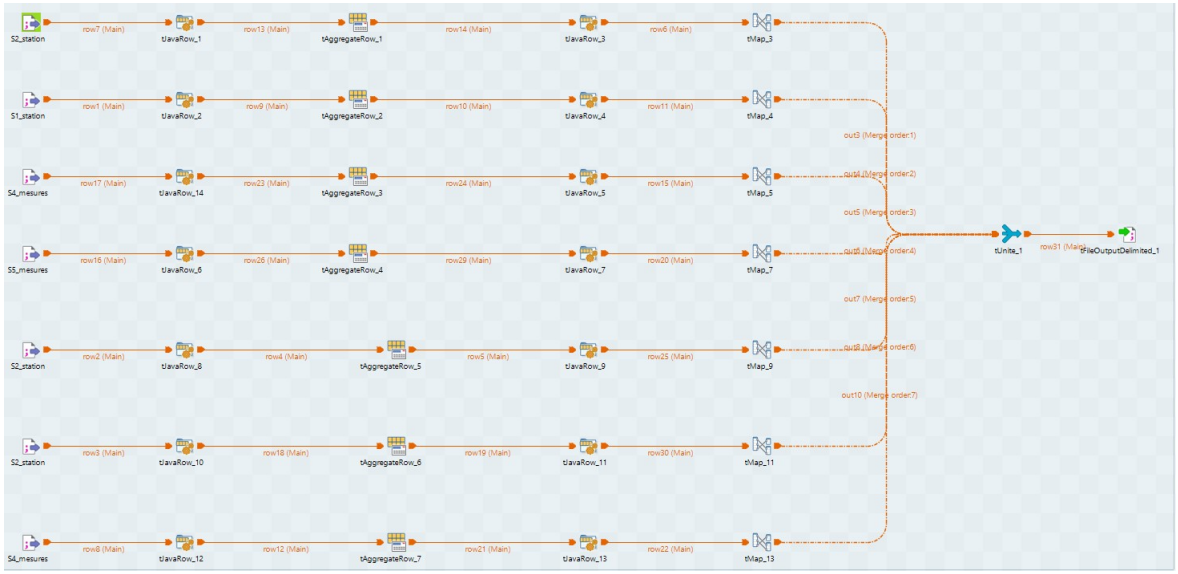

FIGURE 9 – Job d'évaluation de la complétude

4.4 Détection et élimination de doublons

Au niveau des mappings, on va éliminer les doublons obtenus dans les résultats des unions et des jointures à l'intérieur des mappings.

- Dans le cas de l'union des deux station, on peut avoir des même stations mais avec des ID_station différents vu qu'elles viennent de deux BD différentes. Donc pour détecter et supprimer les stations identiques il suffit juste de vérifier le numéro de téléphone ou bien l'adresse mail de la station).

- Nous avons suivis la métrique suivante : Tout d'abord nous avons fait l'union entre les tables S1.Station et S2.Sation puis nous avons evalué le nombre de tuple sans doublons en utilisant la formule suivante :

  Taux_evalué = 1 - [(Nbr_tuple_avec_doublons) / (nbr_tuple_total)]

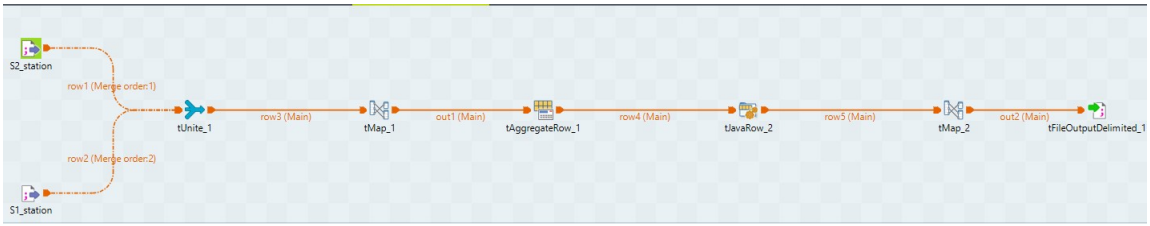

FIGURE 10 – Job de détection de doublons

- **Résumé des différentes valeurs des différentes métriques pour les facteurs de qualité** :
  Nous avons stockés sous forme d'historique les pourcentages obtenus lors de l'étude des facteurs de qualité (complétude, doublons, conformité, granularité et hétérogeinité).
  - La figure suivante montre l'historique de l'évaluation de la complétude.

Date	Source	Attribut	Type	Taux
27/10/2022 23:24:27	S2 Station	Numero	Completude	50.0
27/10/2022 23:24:27	S1 Station	Numero	Completude	71.42857
27/10/2022 23:24:27	S4	Localisation	Completude	86.666664
27/10/2022 23:24:27	S5	Taux releve	Completude	83.33333
27/10/2022 23:24:27	S2 Station	Ville	Completude	83.33333
27/10/2022 23:24:27	S2 Station	Mail contact	Completude	83.33333
27/10/2022 23:24:27	S4	Taux releve	Completude	86.666664
27/10/2022 23:24:35	S2 Station	Numero	Completude	50.0
27/10/2022 23:24:35	S1 Station	Numero	Completude	71.42857
27/10/2022 23:24:35	S4	Localisation	Completude	86.666664
27/10/2022 23:24:35	S5	Taux releve	Completude	83.33333
27/10/2022 23:24:35	S2 Station	Ville	Completude	83.33333
27/10/2022 23:24:35	S2 Station	Mail contact	Completude	83.33333
27/10/2022 23:24:35	S4	Taux releve	Completude	86.666664
28/10/2022 00:07:16	S2 Station	Numero	Completude	50.0
28/10/2022 00:07:16	S1 Station	Numero	Completude	71.42857
28/10/2022 00:07:17	S4	Localisation	Completude	86.666664
28/10/2022 00:07:17	S5	Taux releve	Completude	83.33333
28/10/2022 00:07:17	S2 Station	Ville	Completude	83.33333

FIGURE 11 – Historique l'étude de la complètude

5 Amélioration de la qualité des données

5.1 Complètude des données :

Dans cette partie nous allons étudier la complétude de nos données sur certains attributs.

5.1.1 Localisation :

Nous avons opté pour la solution qui consiste remplacer la valeur nulle par la localisation d'une mesure faite **à quelques minutes prés (la différence des deux timestamp ne dépasse pas les 60min)** par le même capteur sur un autre polluant.
Si on trouve aucune localisation remplaçante, on supprime le tuple concerné de

notre table.

Exemple : dans la table Mesure4 on a une valeur nulle pour la localisation faite par Can45 à "02/10/2021 15 :00" sur le polluant PM10. Le but est donc de chercher dans la table une autre localisation faite par ce même capteur pour un polluant quelconque à quelques minutes prés ou au même moment.

5.1.2 Taux relevé :

En Essayant de chercher d'autres valeurs de taux_relevé pour le même polluant et le même capteur avec une date à quelques minutes prés d'un taux manquant sur les sources S4 et S5, on remplace la valeur nulle par un des taux trouvés, et dans le cas échéant on supprime le tuple.

5.1.3 Numéro :

Au niveau de la station 1 et 2, lorsque nous détectons qu'un numéro de rue d'une station est nul, nous allons chercher la station qui a le même mail ou bien le même le numéro de téléphone que cette dernière et récupérer son Numéro de rue (Nous corrigeons par rapport aux doublons)

5.1.4 Contact_Mail :

Nous allons étudier la complétude de l'attribut Contact_Mail au niveau de Station2 et Station1.
les mails suivent le formats suivant "xxx@airparif.fr" ou "xxx" correspond à la ville donc pour compléter les champs vides. Nous avons deux cas :

- Premier cas : Si la ville c'est **Paris**, nous allons regarder les deux derniers chiffres du code postal par exemple **75019**, ensuite nous allons suivre le format "@airparif.fr" donc la valeur nulle sera remplacé par "paris19"@airparif.fr".

- Deuxième cas : Si la ville est différente de **Paris**, nous allons récupérer la ville et faire respecter le format imposé.
  Exemple : Pour le tuple 6 nous avons Contact_Mail = nul donc nous allons récupérer ici la ville qui est "putaux" ensuite nous allons suivre le format "@airparif.fr" donc la valeur nulle sera remplacé par "puteaux"@airparif.fr".

5.1.5 Ville :

Nous allons étudier la complétude de l'attribut Ville au niveau de Station2 et Station1.
Pour cela nous allons récupérer le code postal, nous avons 3 cas, si le code postal commence par :

- 75 : On affecte la valeur **Paris** à l'attribut Ville.

- 78 : On affecte la valeur **Yvelines** à l'attribut Ville.
- 92 : On affecte la valeur **Hauts-de-seine** à l'attribut Ville.

5.2 Conformité à un format, une codification

Nous allons dans cette partie amélioré la qualité de nos données en respectant le critère de conformité. Ce dernier est nécessaire pour avoir des données cohérentes au niveau des sources.

5.2.1 Format d'un mail :

Au niveau de la station 1 et 2, nous allons mettre tous les mails sous le format suivant "xxx@airparif.fr" ou "xxx" correspond à la ville donc pour corriger le format du mail. Nous avons deux cas :

- Premier cas : Si la ville c'est **Paris**, nous allons regarder les deux derniers chiffres du code postal par exemple **75019**, ensuite nous allons suivre le format "@airparif.fr" donc la valeur nulle sera remplacé par "paris19"@airparif.fr".

- Deuxième cas : Si la ville est différente de **Paris**, nous allons récupérer la ville et faire respecter le format imposé.
  Exemple : Pour le tuple 6 nous avons Contact_Mail = nul donc nous allons récupérer ici la ville qui est "putaux" ensuite nous allons suivre le format "@airparif.fr" donc la valeur nulle sera remplacé par "puteaux"@airparif.fr".

5.2.2 Format du numéro de téléphone :

Au niveau de la station 1 et 2, lorsque nous détectons qu'un numéro d'une station n'est pas dans le bon format, nous regardons la station qui a le même mail que cette dernière, pour prendre son numéro. (Nous corrigeons par rapport aux doublons)

5.2.3 Localisation :

Nous avons donc opté pour la solution qui consiste à remplacer la valeur nulle par la localisation d'une mesure de la même table faite **à quelques minutes prés (la différence des deux timestamp ne dépasse pas les 60min)** par le même capteur sur un autre polluant.

Si on trouve aucune localisation remplaçante, on supprime le tuple concerné de notre table.

5.3 Détection et élimination de doublons :

Cette amélioration doit être éxécuter aprés l'amélioration de la completude et la conformité, pour qu'on puisse détecter tous les doublons possibles (Pour éviter le cas (**t1(1, Null, Null, 3) et t2(2, a, b, Null)**).

- Au niveau de l'union des stations 1 et 2,On va d'abord faire la jointure entre les deux tables s1 et s2 puis regarder si dans le tableau on a des doublons pour ensuite les supprimer.Nous avons comparé chaque valeur des deux tables afin de garder un seul tuple avec les valeurs les plus cohérents possibles.
  Le critère pour détecter les doublons de prendre les tuples qui ont soit la même adresse mail ou le même numéro de téléphone.Pour cela nous avons utilisé le composant **tUniqRow**.

- Si par exemple, un tuple a son champ nul pour un attribut donné et l'autre tuple détecté comme son doublon a une valeur numérique non nulle pour le même attribut, nous allons garder la valeur non nulle.
  Et si les deux champs sont non nuls, on garde celui qui est conforme à la règle de constitution des valeurs de l'attribut. Et si les deux ont respecté les règles de conformité,on privilégie arbitrairement la valeur du tuple provenant de la source S1.

5.4 Hétérogénéité des échelles, de la granularité :

5.4.1 Hétérogénéité des échelles :

Au niveau des sources S1.mesure, S2.mesure et S4, lorsqu'on detecte qu'une valeur de Taux_releve est en dehors de **l'intervalle (qui a été définit pour chaque polluant comme hypothése de base)**, on va la corriger par :

- **((MAX(Intervalle)) + (MIN(Intervalle))) / 2**.

5.4.2 Granularité :

Au niveau de la source S5, nous avons la majorité des valeurs sont en dehors de leurs intervalle, ce qui nous permet de dire que y a un problème de granularité. Pour mettre sur la même échelle les valeurs de S5 nous avons utilisé la méthode suivante :

- Facteur = AVG(S4.mesure) / AVG(S5.mesure)

Pour obtenir un facteur F qui sera utilisé pour la mise à l'échelle (ou on va multiplier les valeurs de s5 par ce facteur).