Projet Qualité de données

Intégration de données de pollution

A base d'un ETL pour évaluer et améliorer

La qualité des données

2021-2022

Compte rendu

Etudiants :

- Toufik GUENANE
- Koussaila ARAB
- Nadjib RAHMANI

ENCADRANTE : **Zoubida KEDAD**

1- Introduction :

Le projet d'intégration de données de pollution a pour but d'intégrer des données de certains tables sources afin de définir des tables cibles, les évaluer et améliorer leur qualité, et cela passant par différentes étapes de qualités de données et en répondant aux critères de qualité imposé par le projet.

1.1 Les données sources :

Sources 1 : Mesures de stations fixes
Mesures (ID_Polluant, Date, ID_Station, Taux_relevé)
Station (ID_Station, Num, Rue, Ville, Tel, Contact_mail)

Sources 2 : Mesures de stations fixes
Mesures (ID_Polluant, Date, ID_Station, Taux_relevé)
Station (ID_Station, Num, Rue, Ville, Tel, Contact_mail)

Sources 3 : Polluants
Polluants (ID_Polluant, Description, Seuil_toléré)

Sources 4 : Mesures de captures mobiles
Mesures (ID_Polluant, Date, ID_Capteur, Localisation, Taux_relevé)

Sources 5 : Mesures de captures mobiles
Mesures (ID_Polluant, Date, ID_Capteur, Localisation, Taux_relevé)

1.2 Les données cibles :

Paris (ID_Polluant, Taux_moyen_jour, NBS, NBC, Statut)
Hauts_de_Seine (ID_Polluant, Taux_moyen_jour, NBS, NBC, Statut)
Yvelines (ID_Polluant, Taux_moyen_jour, NBS, NBC, Statut)
Stations (ID_Station, Adresse, Tel, Contact_mail)

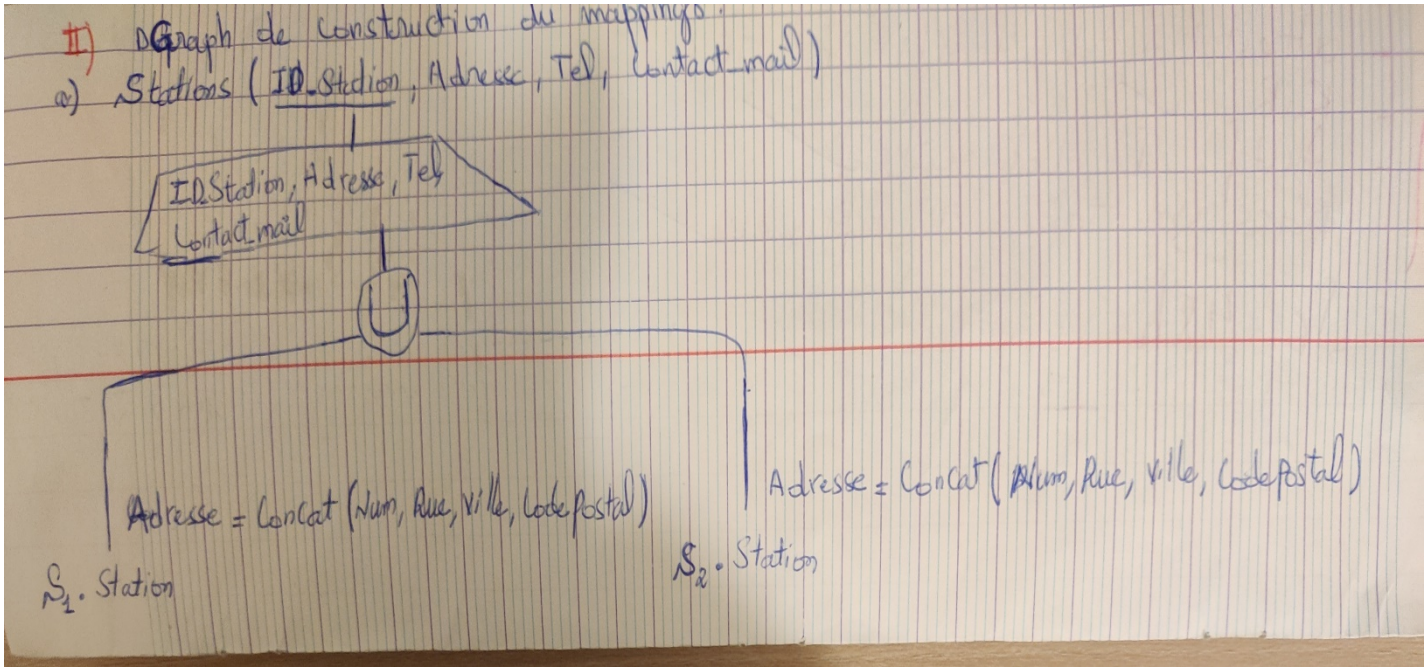
2- Conceptions des mappings des données sources et cibles :

Le projet se compose des tables de données sources qui devraient subir des transformations afin d'être intégré dans des tables cibles tout en respectant la conformité des critères de qualité.
Pour cela on a fait quelques suppositions de départ afin de créer les mappings initiaux qui réaliseront le but attendu, et parmi ceux :

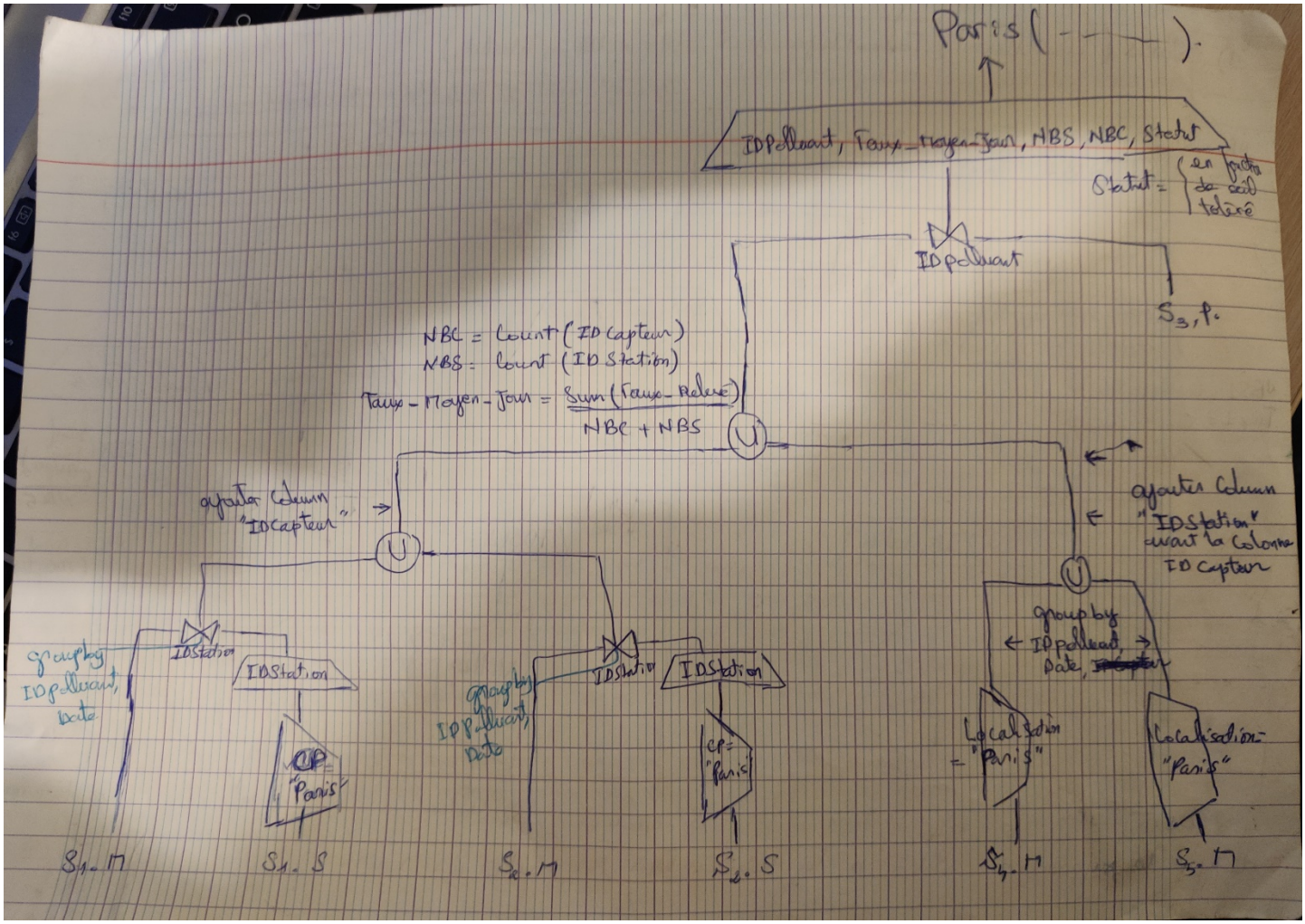
- Si deux attributs (colonne d'une table) de deux tables différentes ont la même syntaxe on peut conclure qu'ils sont sémantiquement identiques (**ex :** S1_Station.ID_Station est identique à S1_Mesures.ID_Station) ;
- L'attribut Ville des sources S1_Station et S2_Station correspondent (égaux) à l'attribut Localisation dans les sources S4_Mesures et S5_Mesures ;

Voici les pseudos_mappings (prototypes) des mappings crées par la suite :

Pseudo_Mapping: Pseudo_Stations_mapping



Pseudo_Mapping: Pseudo_Paris_mapping



Pour les deux autres pseudos_mappings : Yvelines_pseudo_mapping et HautsDeSeine_pseudo_mapping sont respectivement les même comme celui de Paris_pseudo_mapping, ce qui change pour ces deux-là c'est que sur S1.Station et S2.Station, on fait une restriction sur CP= «Yvelines » et CP= « HautsDeSeine » respectivement, et sur S4.Mesures et S5.Mesures on fait une restriction sur l'attribut Localisation, en posant Localisation= « Yvelines » et Localisation= « HautsDeSeine » respectivement.

3- Problèmes rencontrés lors de l'intégration des données Excel imposé par l'encadreur sur les mappings :

L'intégration de données est un processus qui consiste à mettre en œuvre un mécanisme qui permet de prendre des données de plusieurs sources et les intégrés par la suite dans une base de données cible, et cela en passant par les mappings. Malgré ce mécanisme a l'air d'une perfection, mais cela n'est pas le cas car on peut toujours se retrouver sous plusieurs problèmes pendant ce processus, et on a constaté par la suite les problèmes suivants :

Mapping : Stations_mapping (Sur la cible : Stations)

Dans ce mapping l'un des problèmes rencontrer est celui de l'incohérence de certaines données qui viennent avec la présence de doublons, et deux autres problèmes qui sont :

Conformité du format sur les colonnes adresse, tel et contact_mail ;

Complétude des données sur la colonne adresse car y'a des valeurs null qui proviennent des concaténations faites sur : Numéro et Ville.

NB : Sa résolution sera introduite dans la section 4 critères de qualité

Mapping: Paris_mapping (idem: pour Yvelines_mapping et HautsDeSeine_mapping)

Les problèmes rencontres sont les suivants :

Jointure entre S1_Station et S1_Mesures : On a eu un problème sur la colonne taux_releve.

Jointure entre S2_Station et S2_Mesures : nous renvoie une erreur de type NULL Pointer Exception quand on a des valeurs null sur la colonne ville quand on fait une restriction sur ville, et cela ne permet pas d'effectuer la jointure par la suite.

S4_Mesures, S5_Mesures : Cette source contient des données comme Localisation qui sont des cordonnées GPS qui permettent pas de faire une restriction sur une ville car ces des types différents.

NB : Les résolutions de ces problèmes seront introduite dans la section 4 critères de qualité

4- Critères de qualités :

1.1 – Conformité à un format, une codification :

Facteur de qualité : Conformité à un format, une codification.

Niveau : colonnes Contact_Mail, Téléphone, Localisation et Taux_Releve.

Source : Source2.Station(Contact_Mail), Source1.Station(Téléphone), Source4.Mesures (Localisation) et toutes les Mesures(Taux_releve) des Sources 1,2,4 et 5.

Métrique : pourcentage de valeurs non conformes (par rapport aux colonnes qu'on a)

Description de la détection :

Les emails doivent respecter le format : [Lettre + Lettres/numéros + '@' + Lettres + '.' + Lettres].

Les numéros de téléphone doivent respecte le format : [0+(numéros)] (tel que numéro doit avoir neuf chiffres).

Localisation doit respecter le format : (décimal, décimal) qui sont des coordonnées d'un point ou d'une localisation géographique spécifique.

Taux_Releve doit respecter le format : de type float au lieu du format trouver '(x,y)' tel que x et y sont des décimaux. ''(x ,y)'' au lieu de x.y

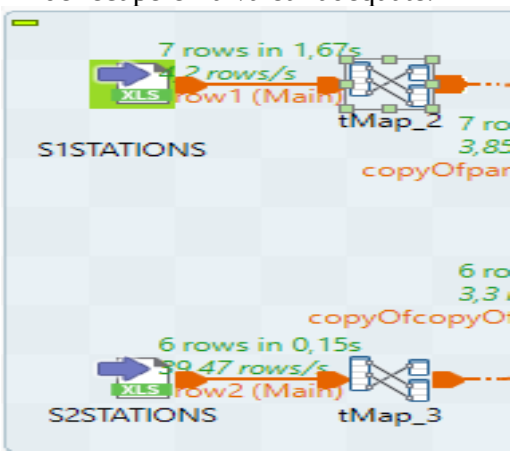
Solutions proposés :

- Contact_mail :

On fait appel à une routine «`MailTester(String mail)`» qui prend le mail de chaque tuple et il retourne vrai si le format est correct et on le garde, sinon on fait appelle à une autre routine «`mailCorrector(String mail,String ville)`» qui vas essayer de corriger le Conatct_mail du tuple en s'appuyant sur les colonnes Contact_mail d'autres sources pour récupérer la valeur adéquate, sinon on retourne la « ville+''@airparif.fr'' ».

- Téléphone :

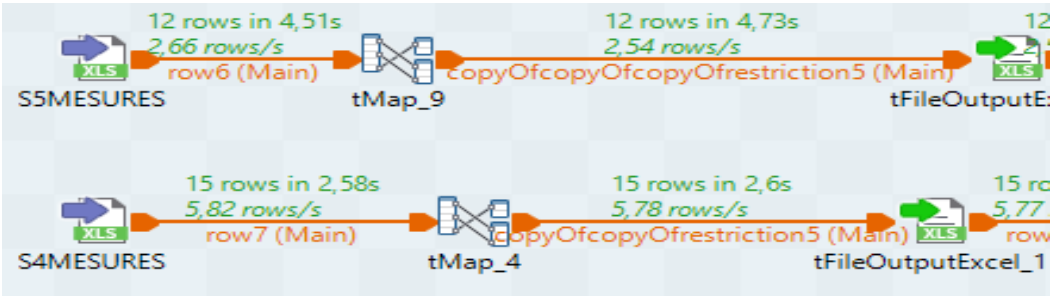
On fait appel à une routine «`NumberTester(String numTel)`» qui prend le téléphone de chaque tuple et il retourne vrai si le format est correct et on le garde, sinon on fait appelle à une autre routine «`numberCorrector(String numTel,String rue,String cp)`» qui vas essayer de corriger le numéro de téléphone du tuple en s'appuyant sur les colonnes rue et Code_postal d'autres sources stations afin de récupérer la valeur adéquate.



L'appel des routines se fait dans les tMap2,tMap3.

- Localisation :

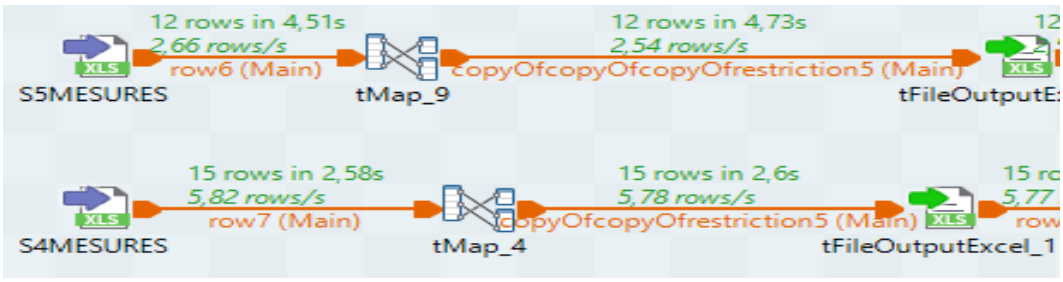
On fait appel à une routine «`getZIPcode(String Localisation)`» qui prend la localisation en (x,y) de chaque tuple et il retourne le code postal de la localisation et cela en appelant une api.

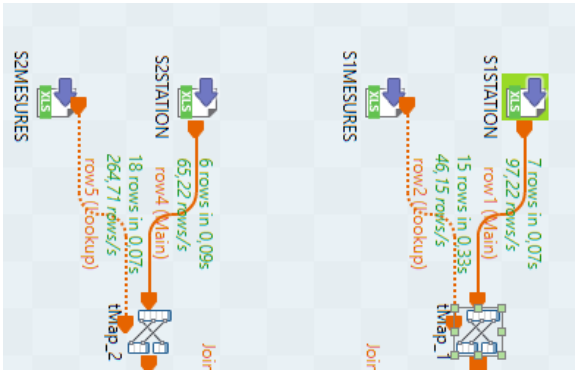


L'appel des routines se fait dans le tMap9 et tMap4.

- Taux_Releve:

On fait appel à une fonction replaceAll qui prend « x,y » et donne « x .y», puis on convertit le résultat en float avec la fonction Float.valueOf.

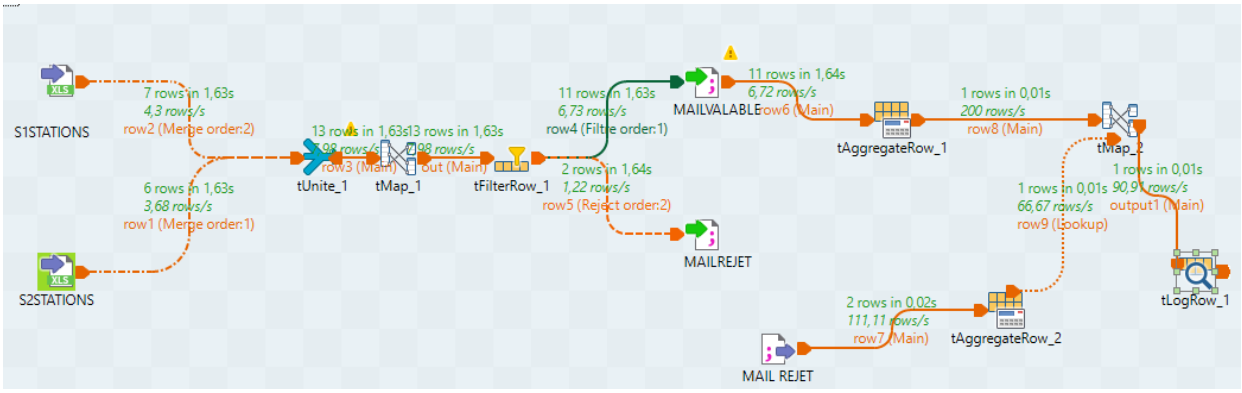




L'appel des routines se fait dans le tMap9, tMap4, tMap2 et tMap1.

Pour calculer le nombre de Contact_mail rejeté et accepté avant la correction totale des données on a fait çà :

(Idem pour les trois autres problèmes)



MailValable	PourcentageMailValable	NombreMailValable	MailRejet	PourcentageRejet	NombreMailRejet
mail validé	0.8461538461538461	11	mail rejeté	0.15384615384615385	2

1- 2– Granularité des données et hétérogénéité des échelles :

Facteur de qualité : Granularité des données et hétérogénéité des échelles.

Niveau : Table Mesures de toutes les sources de données.

Source : Source 1,2,4 et 5.

Métrique : Booléen, s'il y a un problème c'est 1 sinon 0

Description de la détection :

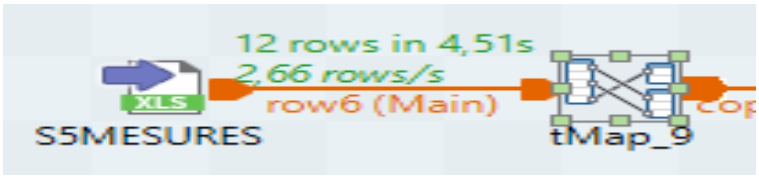
L'étude se fait sur la table Mesures.Taux_releve de toutes les sources 1,2,4 et 5, on calcule la moyenne de la colonne Taux_Releve de la source 1 et la moyenne des Taux_releve de la source 2 ainsi que la valeur max et min de ces deux colonnes, puis nous calculons les quotients : min(S2) /min(S1) ; max(S2) /max(S1) ; avg(S2) /avg(S1) (Idem pour S4. Mesures et S5. Mesures)

Source1	Table1	Source2	Table2	colomn1	colomn2	Moy(S4)/Moy(S5)	Min(S4)/Min(S5)	Max(S4)/Max(S5)
Soure4	Mesures	Source5	Mesures	Taux_Relevé	Taux_Relevé	550	23400	592
Source1	Mesures	Source2	Mesures	Taux_Relevé	Taux_Relevé	1,9	1	1

Si ces valeurs > seuil(Moyenne=500) alors on considère qu'il y'a un problème de granularité. (Cas de S5 et S4)

Solution proposée :

On multiplie le Taux_Releve de la colonne S5.Mesures par 500.



Le traitement a été effectué dans le tMap9.

1-3 – Complétude des données :

Facteur de qualité : Complétude des données.

Niveau : Numéro, Ville, Localisation et Taux_releve.

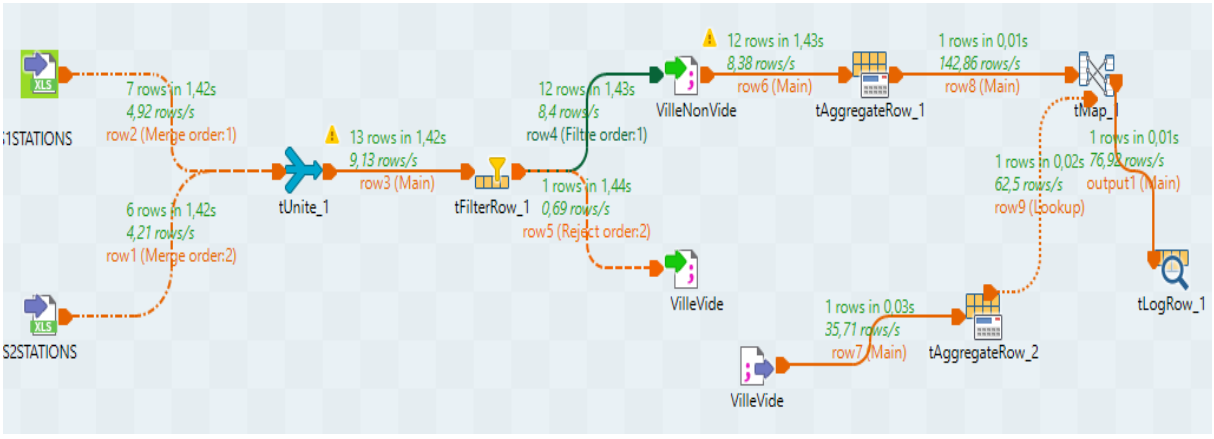
Source : Source1.Station. Numéro, Source2.Station. Numéro, Source2.Station. Ville, Source2.Station. Contact_mail, Source4.Localisation, Source4.Taux_releve et Source5.Taux_releve.

Métrique : Calcul de nombre de valeurs null pour chaque colonne existante

Description de la détection :

On calcule le taux de valeur null par colonne.

Schéma de détection : On prend pour Stations. Ville (idem pour les autres détections de problèmes)



VilleNonVide	PourcentageVilleNonVide	NombreVilleNonVide	VilleVide	PourcentageVilleVIde	NombreVilleVide
Ville non vide	0.9230769230769231	12	Ville Vide	0.07692307692307693	1

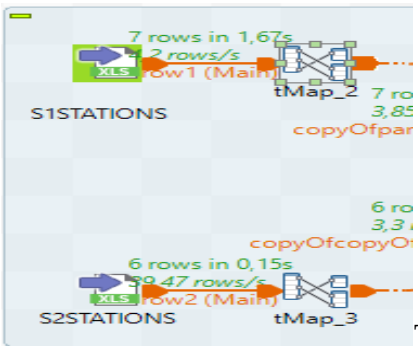
Résolutions des problèmes :

- Numéro :

Si numéro est null, on fait appelle la routine « `roadCorrector2("-1", row2.Rue, row2.Code_postal)` » `‘’-1’’ pour null` qui vas parcourir la station source et verifie si y’a correspondance sur Rue et Code_postal et par la suite il retourne le bon numéro, si y’a pas de correspondance on renvoie ‘‘ caractère vide.

- Ville :

Si ville est null, on regarde les deux premiers caractères de code postale et on retourne le nom de département qui correspond à ce code postale (75=Paris, 78=Yvelines et 92=Hauts de Seine)



Traitement effectué dans les tMap2 et tMap3(pour Ville et Numéro)

- Contact_mail et Localisation :

Même résolution comme le critère conformité à un format(même résolution plus même mapping)

- Taux_Releve :

Si taux_releve est null, on fait appelle à une routine « `getTauxReleveS5corrected("-1",row6.Localisation, row6.Date, row6.ID_Capteur, row6.ID_Polluant)` » `‘’-1’’ pour null,` afin de remplir les valeurs nulles on prend en considération la date, ID_Capteur, localisation et ID_Polluant.

Pour les mappings se retrouvent dans la section conformité à un format sur le point Taux_Releve.

1-4 –Détection et élimination des doublons :

Facteur de qualité : Détection et élimination des doublons.

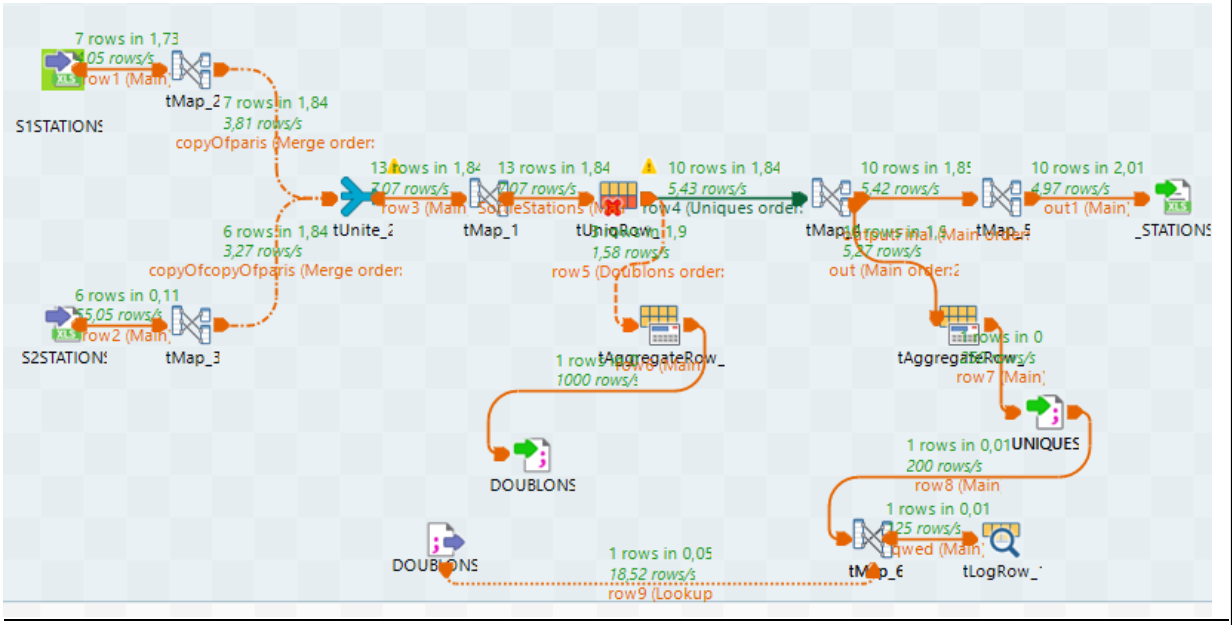
Niveau : Tables Cibles.

Source : Source1.Station et Source2.Station.

Description de la détection :

Après l'union des deux sources, si deux tuples ont le même nom, prénom et email alors on les considère des doublons.

Schéma de détection :



doublons	pourcentageDoublons	nombreTuplesDoublons	uniques	pourcentageUniques	tuplesUniques
Doublons	0.23076923076923078	3	UNIQUES	0.7692307692307693	10

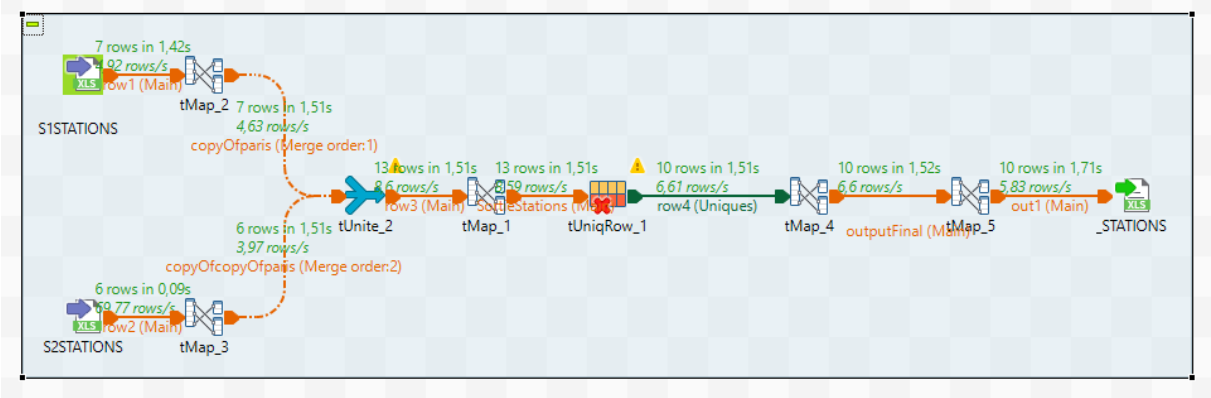
Résolutions des problèmes :

Supprimer les tuples doublant.

5- Mappings Finales :

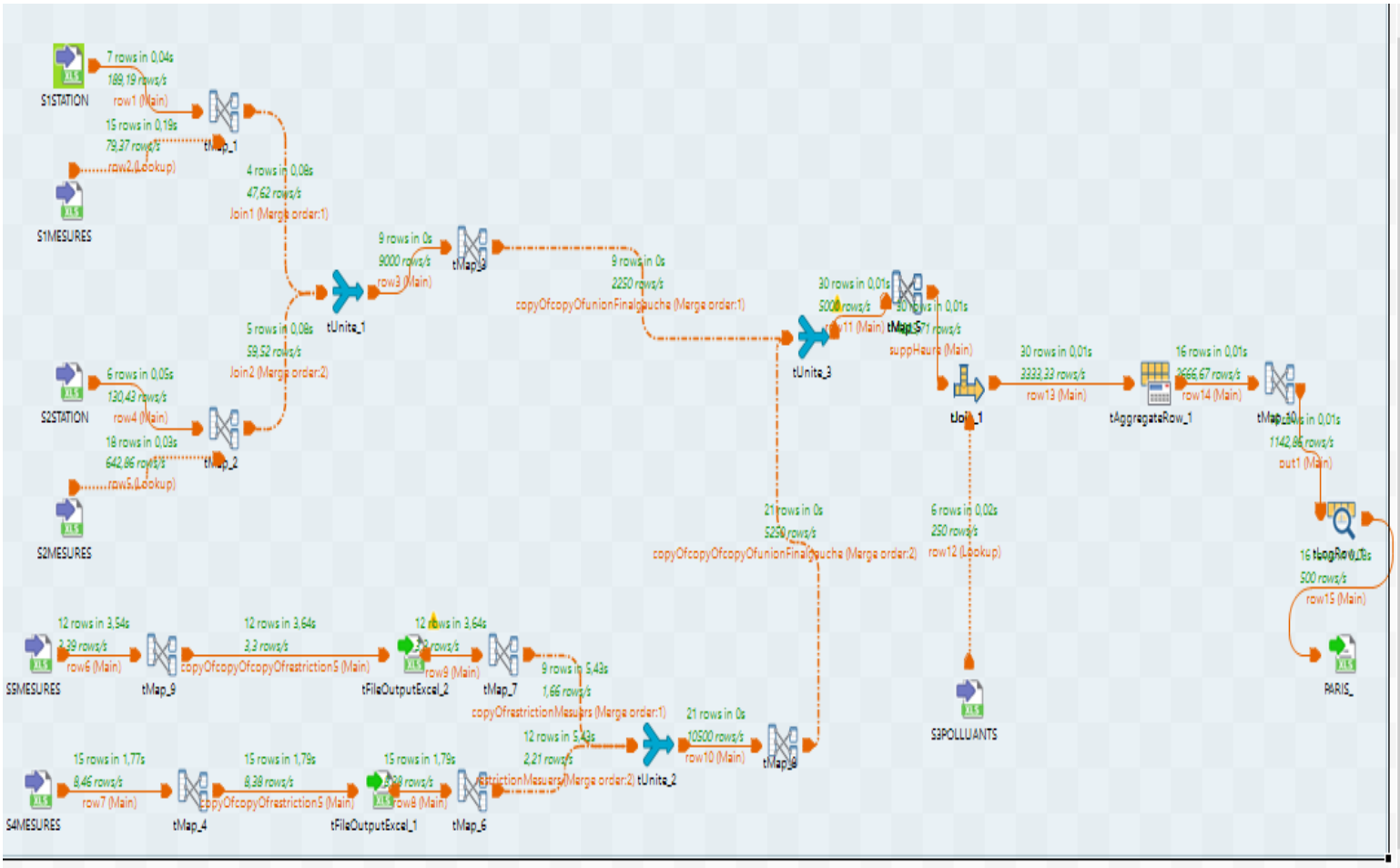
Après l'intégration totale des données et vérification de leur qualité on a eu les mappings finals qui peuvent crée les tables cibles, et les voilà :

Mapping Stations



ID_Station	Adresse	Tel	Contact_mail
1	5 Etant d'or 78120 Rambouillet	01 75 03 40 00	rambouillet@airparif.fr
2	8 Limoges 78000 Versailles	01 39 25 40 00	versailles@airparif.fr
3	1 Emile Zola 78200 Mantes-la-Jolie	01 34 78 81 00	mantes@airparif.fr
4	Parvis de la Défense 92800 Puteaux	01 46 92 92 92	ladefense@airparif.fr
5	Allée des Refusniks 75007 Paris	01 53 58 75 07	paris07@airparif.fr
6	2 bis Quai de la Mégisserie 75001 Paris	01 44 50 75 01	pari01@airparif.fr
7	60 Richelieu 92230 Gennevillier	01 40 85 66 66	gennevilier@airparif.fr
8	Château_princeloup 78120 Sonchamp	01 34 84 41 08	sonchamp@airparif.fr
9	11 Commandant Pilot 92200 Neuilly-sur-seine	01 40 88 88 88	Neuilly-sur-seine@airparif.fr
10	7 Ferdinand Flocon 75018 Paris	01 53 41 18 18	Paris@airparif.fr

Maping Paris : (idem pour mapping yvelines et mapping hauts de seine)



Valeur de sortie de mapping paris :

Avec la date :

ID_Polluant	Date	NBS	NBC	Taux_Moyen_Jour	Statut
O3	10/3/21	1	1	39.785	Normal
PM10	10/2/21	0	2	1209.2	Alerte pollution
CO	10/2/21	1	2	5162.447	Normal
CO	10/1/21	0	1	4523.0	Normal
CO	10/15/21	1	0	5678.34	Normal
NO2	10/2/21	1	1	4431.0	Alerte pollution
CO	10/9/21	0	1	98.4	Normal
NO2	10/1/21	0	1	98.4	Normal
NO2	10/6/21	0	1	95.8	Normal
NO2	10/3/21	2	2	836.93	Alerte pollution
CO	10/7/21	0	1	3428.9	Normal
PM10	10/7/21	1	2	29.296667	Normal
NO2	10/8/21	1	2	103.56667	Normal
NO2	10/7/21	1	2	21.366667	Normal
PM10	10/6/21	0	1	0.5	Normal
NO2	10/9/21	0	1	4048.9998	Alerte pollution

Sans la date : (Structure finale)

ID_Polluant	Taux_Moyen_Jour	NBS	NBC	Statut
O3	39.785	1	1	Normal
PM10	1209.2	0	2	Alerte pollution
CO	5162.447	1	2	Normal
CO	4523.0	0	1	Normal
CO	5678.34	1	0	Normal
NO2	4431.0	1	1	Alerte pollution
CO	98.4	0	1	Normal
NO2	98.4	0	1	Normal
NO2	95.8	0	1	Normal
NO2	836.93	2	2	Alerte pollution
CO	3428.9	0	1	Normal
PM10	29.296667	1	2	Normal
NO2	103.56667	1	2	Normal
NO2	21.366667	1	2	Normal
PM10	0.5	0	1	Normal
NO2	4048.9998	0	1	Alerte pollution

Valeur de sortie de mapping Yvelines :

ID_Polluant	Taux_Moyen_Jour	NBS	NBC	Statut
PM10	151.95	2	0	Alerte pollution
O3	190.175	2	2	Alerte pollution
O3	201.4	1	1	Alerte pollution
PM10	23.4	2	0	Normal
NO2	100.5	1	0	Normal
PM10	23.89	1	0	Normal
PM10	51.4	2	0	Alerte pollution
CO	7.0	1	0	Normal

Valeur de sortie de mapping hauts de seine :

ID_Polluant	Taux_Moyen_Jour	NBS	NBC	Statut
PM10	50.0	0	1	Alerte pollution
O3	77.15	1	1	Normal
PM10	4789.67	1	0	Alerte pollution
PM10	145.7	1	0	Alerte pollution
CO	4789.67	1	0	Normal
CO	8098.0	1	0	Normal
CO	4515.25	1	1	Normal
NO2	95.8	1	0	Normal
CO	2340.78	1	0	Normal
NO2	23.89	1	0	Normal
NO2	124.35	1	0	Normal
PM10	2340.78	1	0	Alerte pollution
NO2	8098.0	1	0	Alerte pollution

:

6- Bilan de l'audit :

6-1 Bilan avant l'amélioration de la qualité des données :

Source	NomTable	Facteur Qualité	Nom Column	ValeurQualité
Source1	Mesures	Conformité	Taux_Relevé	0,00 %
Source2	Mesures	Conformité	Taux_Relevé	0,00 %
Source1	Station	Conformité	Téléphone	86,00 %
Source2	Station	Conformité	Contact_Mail	67%
Source4	Mesures	Conformité	Localisation	85,00 %
Source4	Mesures	Conformité	Taux_Relevé	0,00 %
Source5	Mesures	Conformité	Taux_Relevé	0,00 %
Source1	Station	Complétude	Numéro	71,00 %
Source2	Station	Complétude	Numéro	50,00 %
Source2	Station	Complétude	Ville	83,00 %
Source2	Station	Complétude	Contact_Mail	67,00 %
Source4	Mesures	Complétude	Localisation	73,00 %
Source4	Mesures	Complétude	Taux_Relevé	87,00 %
Source5	Mesures	Complétude	Taux_Relevé	83,00 %

Source1	Table1	Coloumn1	Source2	Table2	Coloumn2	Facteur Qualité	valeur Qualité
Source4	Mesures	Taux_ relevé	Source5	Mesures	Taux_Relevé	Hétérogénéité des échelles	OUI

6-2 Bilan après l'amélioration de la qualité des données :

Source	NomTable	Facteur Qualité	Nom Column	ValeurQualité
Source1	Mesures	Conformité	Taux_Relevé	100,00 %
Source2	Mesures	Conformité	Taux_Relevé	100,00 %
Source1	Station	Conformité	Téléphone	100,00 %
Source2	Station	Conformité	Contact_Mail	100,00 %
Source4	Mesures	Conformité	Localisation	100,00 %
Source4	Mesures	Conformité	Taux_Relevé	100,00 %
Source5	Mesures	Conformité	Taux_Relevé	100,00 %
Source1	Station	Complétude	Numéro	86,00 %
Source2	Station	Complétude	Numéro	67,00 %
Source2	Station	Complétude	Ville	100,00 %
Source2	Station	Complétude	Contact_Mail	100,00 %
Source4	Mesures	Complétude	Localisation	100,00 %
Source4	Mesures	Complétude	Taux_Relevé	100,00 %
Source5	Mesures	Complétude	Taux_Relevé	100,00 %

Source1	Table1	Coloumn1	Source2	Table2	Coloumn2	Facteur Qualité	valeur Qualité
Source4	Mesures	Taux_ relevé	Source5	Mesures	Taux_Relevé	Hétérogénéité des échelles	NON

7- Conclusion :

L’étude de cas avait comme objectif d’intégrer des différentes sources de données dans un seul entrepôt de données cible. Ce travail est réalisé en deux grandes parties, une dédié à la détection des problèmes de qualité de données, et l’autre à l’amélioration de cette qualité de données.

Durant cette réalisation nous avons pris connaissance de l’ETL Talend qui nous a permet d’évaluer et améliorer la qualité des données sur des métriques que nous avons choisies.