Rapport

# Résumé

# Abstract

# Table des matiers

# Table des figures

# Introduction

## Contexte

Jean-Edouard de la Motte Rouge, avec sa startup spécialisée dans le conseil en campagnes électorales, souhaite intégrer l'intelligence artificielle au sein de son entreprise. Plusieurs facteurs, qu'ils soient économiques, sociaux ou autres, influencent les choix des électeurs et impactent les résultats des élections. L'objectif est de développer un modèle d'apprentissage capable de prédire les tendances électorales au fil des années en prenant en compte ces différents contextes.

## Objectifs et résultat attendu

L'objectif du projet est d'entraîner une intelligence artificielle à prédire les tendances des élections présidentielles futures en fonction de divers facteurs contextuels. Pour cela, nous nous baserons sur les élections de 2017 et 2022 en France, en utilisant des données liées aux élections, aux candidats, ainsi qu'aux contextes concernant la sécurité, la population, l'inflation, le SMIC, le taux de chômage, le taux d'emploi, le taux d'activité et le nombre de demandeurs d'emploi.

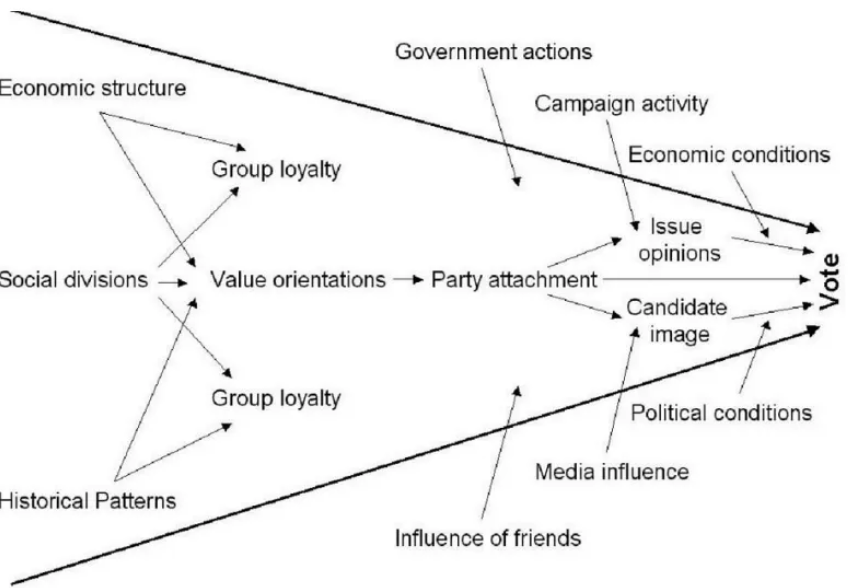
## Organisation

Nous avons identifié trois étapes majeures dans ce projet afin de le mener à bien :

* Maîtriser le sujet et comprendre les facteurs influençant les élections
* Collecte et compréhension des données
* Nettoyage des données
* Construction du model d’IA
* Visualisation des résultats de chaque model

# Maîtriser le sujet et comprendre les facteurs influençant les élections

Pour mieux comprendre le processus des élections, le "funnel of causality" (entonnoir de causalité) est un concept clé en sciences politiques. Il permet d'expliquer les influences complexes et variées qui affectent le comportement des électeurs. En France, ce modèle analyse comment différents facteurs interagissent pour façonner les décisions politiques au fil du temps. Des variables telles que l'identité sociale, les facteurs économiques (chômage, pouvoir d'achat, inégalités), les enjeux sécuritaires, et les tendances politiques globales se resserrent à mesure que l'on se rapproche du moment de vote. Ce processus aide à comprendre comment les contextes historiques, sociaux et économiques influencent progressivement les préférences électorales des citoyens. Le modèle permet ainsi de mieux prédire les résultats électoraux en prenant en compte cette combinaison de causes, toutes interdépendantes mais ayant des poids variables selon chaque élection.



# Collecte et compréhension des données

Nous commençons par collecter les données relatives aux élections et aux contextes socio-économiques. Sur la base de ces données, nous définissons la zone géographique et les années sur lesquelles nous allons entraîner notre modèle. Après nos analyses, nous avons choisi la région Île-de-France et les élections de 2017 et 2022 comme points de référence. Nous détaillerons dans la suite le processus de sélection et les raisons qui ont guidé ces choix.

## Choix de la zone

Histoire + la diversité il de France 🡪 representative de la France

## Choix de la période

Nous avons limité notre périmètre d’étude aux données relatives aux élections présidentielles de 2017 et 2022 en raison de la disponibilité et de la fiabilité des datasets représentant les contextes liés à ces élections.

## Compréhension et analyse structurelle

Nous disposans de ces datasets :

|  |  |  |  |
| --- | --- | --- | --- |
| Datasets | Periode | Par departement | Remarque |
| Dliquance et population | 2016 2023 | Oui |  |
|  | 2000 2013 | Oui | Mais chaque dpt dans une colonne |
|  | 2015 2018 | Oui |  |
|  | 2012 2023 | QUE PARIS |  |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |

Avec l’hétérogénéité des données que nous disposons, nous avons dans un premier temps essayé, de comprendre les données en faisant de graphe …..

## Distribution des données

## Nettoyage des données . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . .

## Analyse statistique . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . .

### Données relatives à la formation suivie . . . . . . . . . . . . . . . . . . . .

### 8 2.4.2 Notes aux épreuves anticipées du baccalauréat de français . . . . . . . . .

### 8 2.4.3 Notes au principales matières en terminale . . . . . . . . . . . . . . . . . .

### 9 2.4.4 Notes au principales matières en première . . . . . . . . . . . . . . . . . .

### 9 2.4.5 Avis des professeurs du lycée . . . . . . . . . . . . . . . . . . . . . . . . .

# Implémentation du modèle de prédiction

Kenza : en cours

Coorelatiion

Ahmed : model

**Features :**

Age candidat,

Sexe

Orientation

Gouvernemant\_sortant

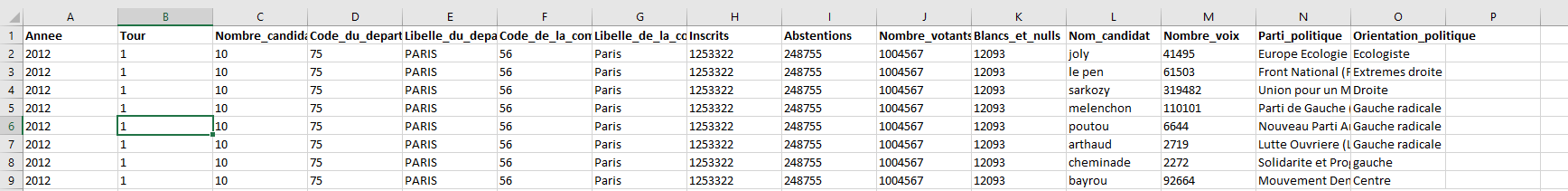
+ les contexts

**Objectif**:label : nombre de voix

# Schéma de données pour les élections

Objectif : créer un seul schéma de données et inclure les donnes des élections de tous les data sets 2012 , 2017 ,2022.

Structure de schéma visé :



## Méthodologie :

Choix de l’etl : Talend Open Studio pour sa facilité et ses composants intégré et personnalisé qui permets de transfomer la données comme on veut .

On créé un nouveau projet taled : mspr\_data puiis :

### Nouveau job pour les elections : elections : version 1

**Shema :**

* + **Code\_du\_departement**
  + **Libelle\_du\_departement**
  + **Code\_de\_la\_commune**
  + **Libelle\_de\_la\_commune**
  + **Votants**
  + **NicolasSarkozy**
  + **EvaJOLY**
  + **MarineLEPEN**
  + **JeanLucMELENCHON**
  + **PhilippePOUTOU**
  + **NathalieARTHAUD**
  + **JacquesCHEMINADE**
  + **FrancoisBAYROU**
  + **NicolasDUPONT\_AIGNAN**
  + **FrancoisHOLLAND**

Remarque : tous les fichiers des eletions téléchargés à partir de data.gouv.fr ont le meme shcema initiale :

#### Créer une métadonnées pour les elections 2012 Tour1 :

* + Fichier : 2012T1
  + Configuration de la métadonnée

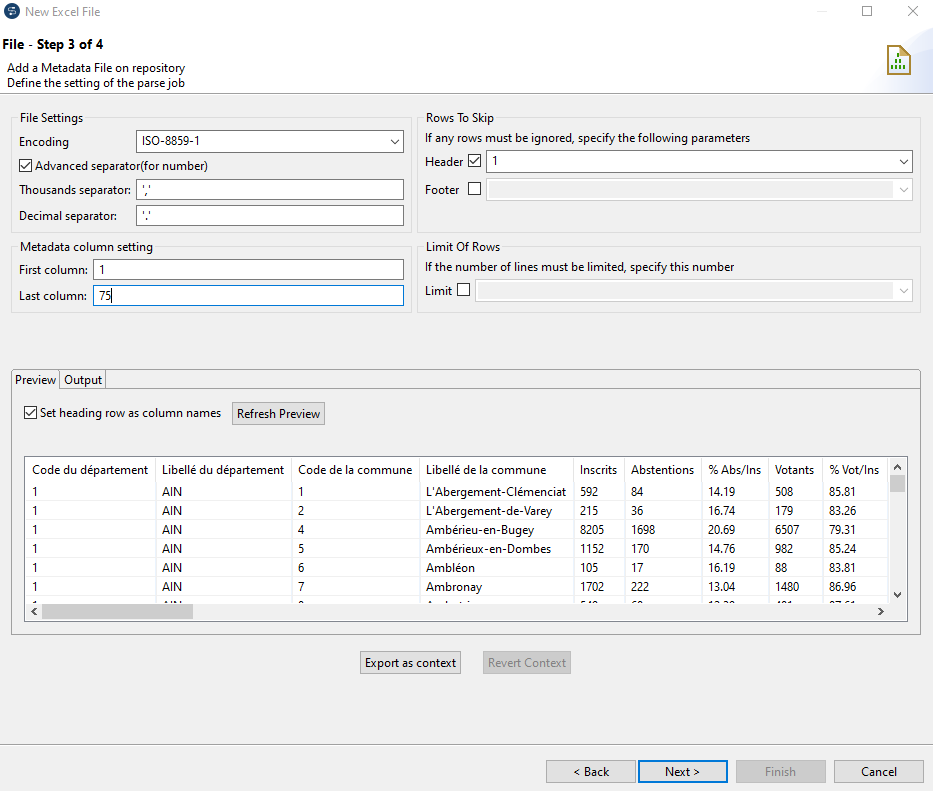


Figure 1Metadata elecion2012

#### Configurer le TFileInputExcel :

Il faut utiliser la Metadata définie ci-dessus pour lire le fichier

#### Ajout et configure tmap :

Source :

Liste des candidats : <https://www.france-politique.fr/election-presidentielle-2012.htm>

Utilise le Tmap pour defenir le schema de données qu’on souhaite utiliser pour le model d’apprentissage.

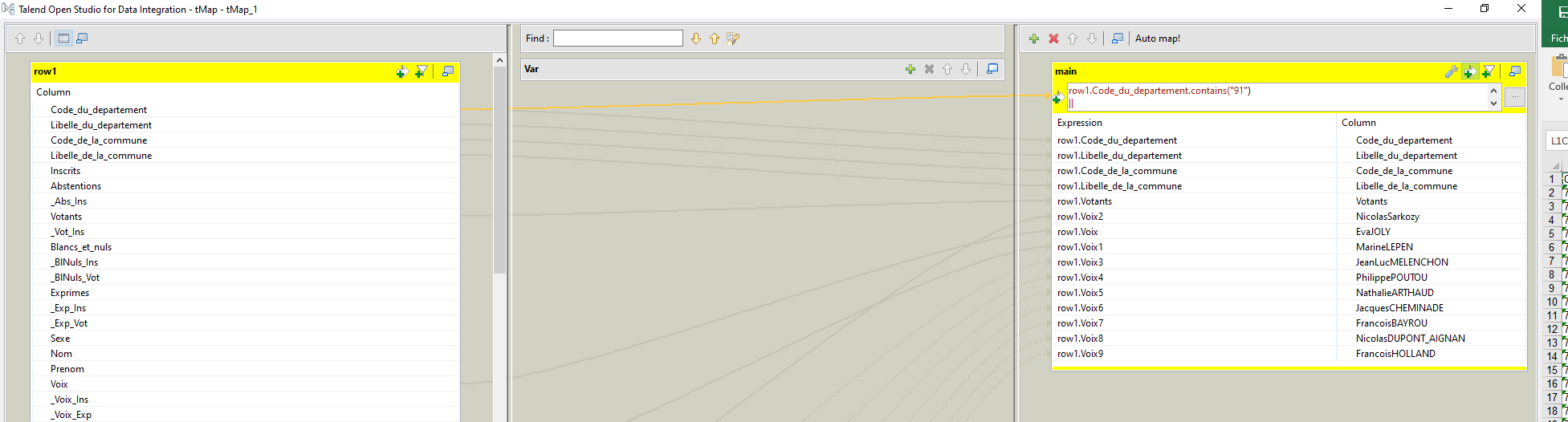
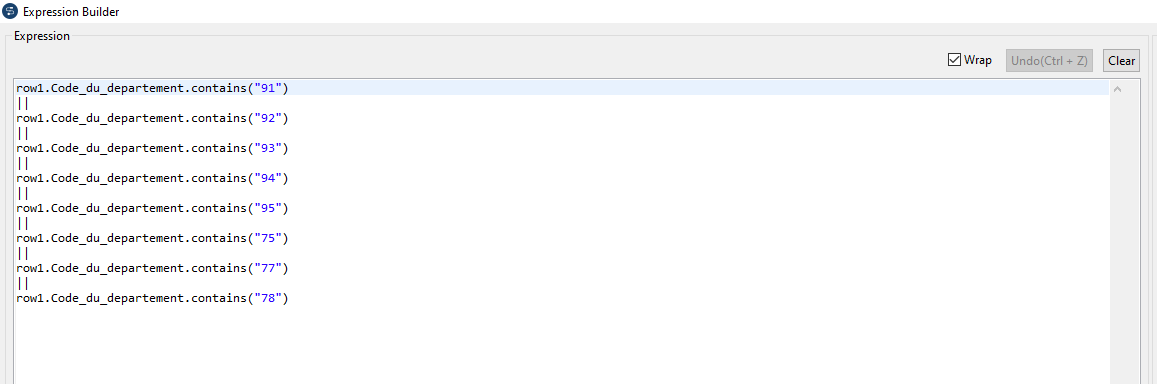
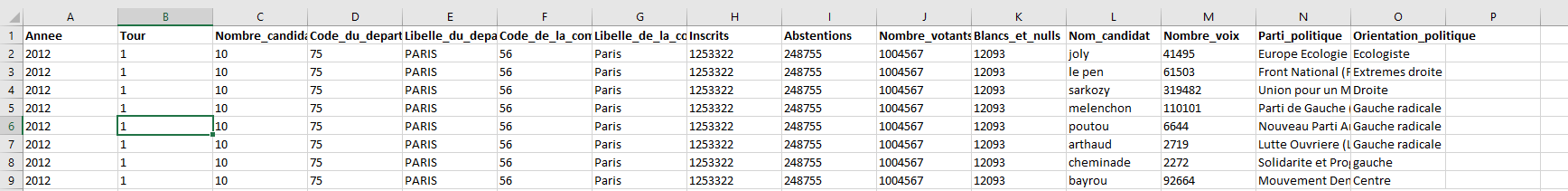
* Transfomer le shema sous le format
  + Code\_du\_departement
  + Libelle\_du\_departement
  + Code\_de\_la\_commune
  + Libelle\_de\_la\_commune
  + Votants
  + NicolasSarkozy
  + EvaJOLY
  + MarineLEPEN
  + JeanLucMELENCHON
  + PhilippePOUTOU
  + NathalieARTHAUD
  + JacquesCHEMINADE
  + FrancoisBAYROU
  + NicolasDUPONT\_AIGNAN
  + FrancoisHOLLAND
* Je veux que la région-il de France :
  + Filtrer la sortie main dans le tmap :

Figure 2Tmap :Epression pour garder que IlDeFrance

Figure 3Tmap

* Je veux transfomer le shéma :
  + Créer une colonne pour chaque candidat :
  + Source pour le nom des candidats :
    - Liste des candidats : <https://www.france-politique.fr/election-presidentielle-2012.htm>
  + Ajout des collonnes :
    - PartiPolitique
    - Orientation

### **Nouveau job pour les élections : élections : version 2 : un autre schéma de données**



Ce schéma vous permet de standardiser les données à travers plusieurs élections, même avec des candidats différents, On peut facilement intégrer les résultats de 2012, 2017, 2022 et au-delà, tout en rendant le modèle plus flexible et généralisable pour prédire des tendances futures.

Ressources et documentations : <https://bekwam.blogspot.com/2012/08/iterating-over-java-collection-with.html?m=1>

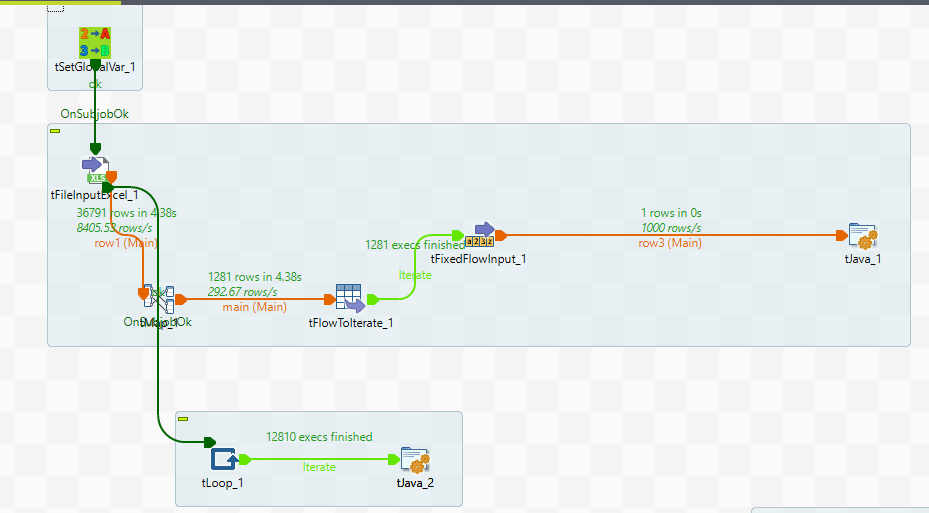
Source des Datasets :

Format : Fichier excel avec deux feuille : Tour1 et Tour2

Shema des données initiales  :

Shema de données visée

Processus du traitement :



Ajout des donnés candiats :

Parti et orientatin politique : on a un fihcer candidats

* Diret dans tjava en sauvegardnant dans la liste global avec code java avant la loop
* Direct dans la bd apres la loop

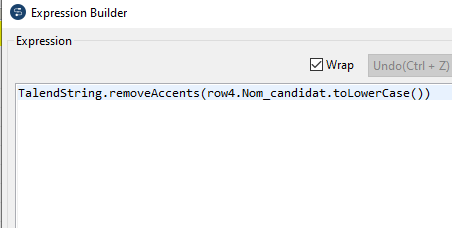
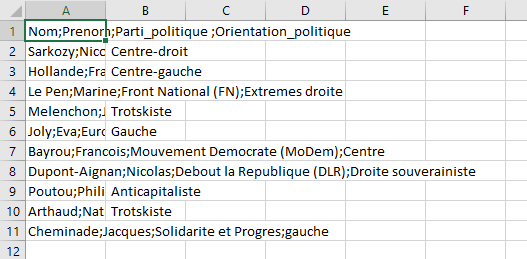
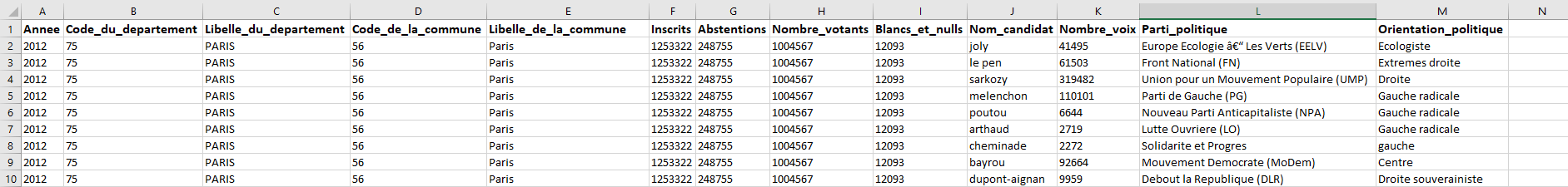
Log 12 09

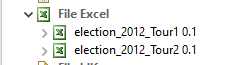
Job : test21heures ;---------objectif add data parties et orientation from candidats.csv file

Change in schema tmap5 after tllop and tjava2 : remove the columns : parti et orientation : if I keep this I have to change shema iinput of all elections : excel metadata

Problem dans la jointure : I get more rows then expected , after the join

Debug :

* tlog row de tmpa5 avant la jointure :
* TLOG si ok i have 12810 ROW
* Tlog pour tmap7 apres la jointure :natural joiin
* Tmap 7 : ok i get 12810 rows with data for Parti et orientation
* Link tmap7 to tjavaflex4 to create csv file and check data
  + Null and no values
  + Debug :
    - Tlogrow : Data are good before the last component that write to csv file
    - use only 75 departemnet ,change the first tmap that do filters
    - tMap7 : all data out are string but I’m using string value of in the next component wich create problem , solution d’ont use string.valueof in tjavaFlex4 and change row structure
  + Update done in tjavaflex4
* BUG : null value in parti et orientation for mélencon et cheminade
  + Cheminada jaques is not included in candidats.csv
    - Added to candidats.csv
  + Melencon : error in name not use accent
    - Tmap5 update
    - 
* Test the update : objectif : create csv file with correct data
  + BUG : cheminade has null for Parti and orientation
    - Refresh talend repo, Nom et prenom invesed in the candidats files : corrected :
* RUN OK
* 
* Supprimer tous les caractères spéciaux
  + Avant le composant de sortie : csv ou db
* Ajouter tour2 :
  + Lire tour1 et Tour2
    - Metadonnée pour chaque tour car on a pas le meme shema , tour 1 on a 10 candidat tour 2 on a 2

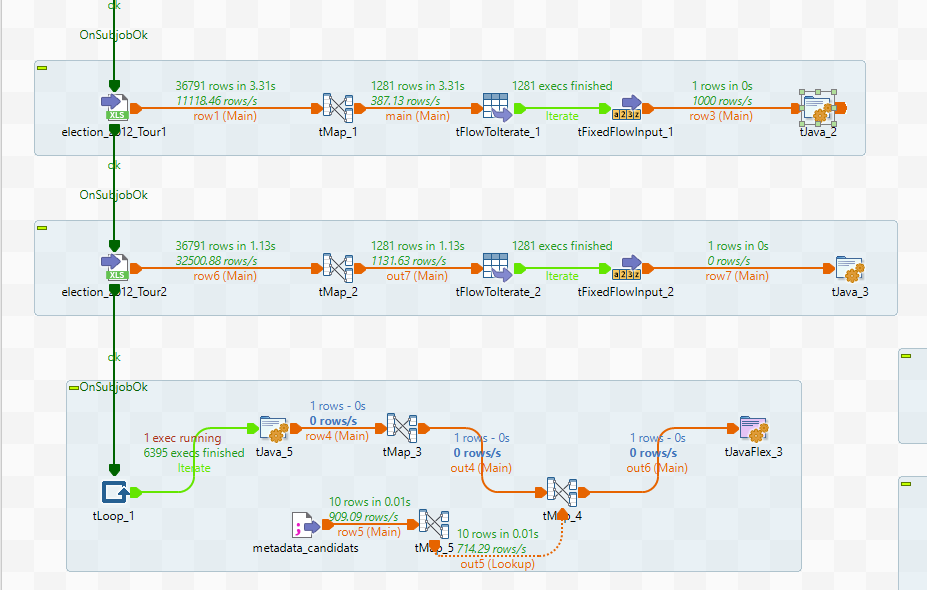


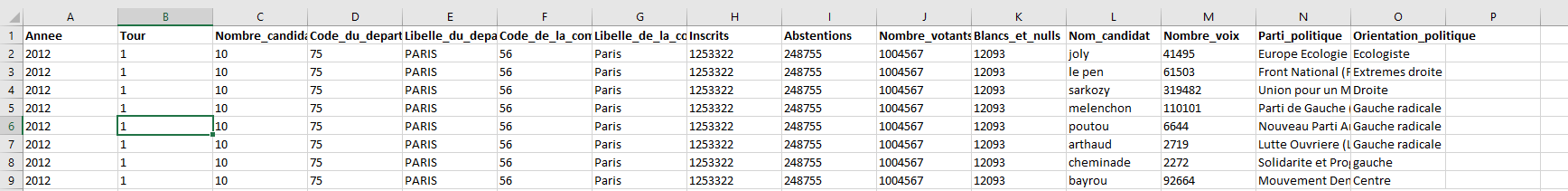
* + - AJOUTER LES COLLONES
      * Nombre de candidats
      * Tour
* Update the code of tjava2 and tjava3
  + Ces comopanst vont ecrire dans la liste global qui sera utilize apres pour ecrire dasn le csv ou la bd
    - Variable statique comme tour et année ,adaptation de la liste des noms ,pour tour1 on a 10 noms et tour2 on a 2.
* Shema de job

Transfomer le shema de données principale en un shema exploitable

Subjob 1 : pour tour 1

Subjob 2 pour tour 2

Subjob 3 : ajouter les information : partie poliqutge , orientation et ecrie vers une sortie : db ou file



Ajout dataset : 2017 TOUR1

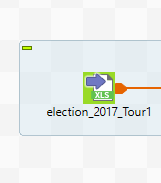
Source :

Un seul fichier excel , avec deux feuilles Tour1 et Tour2

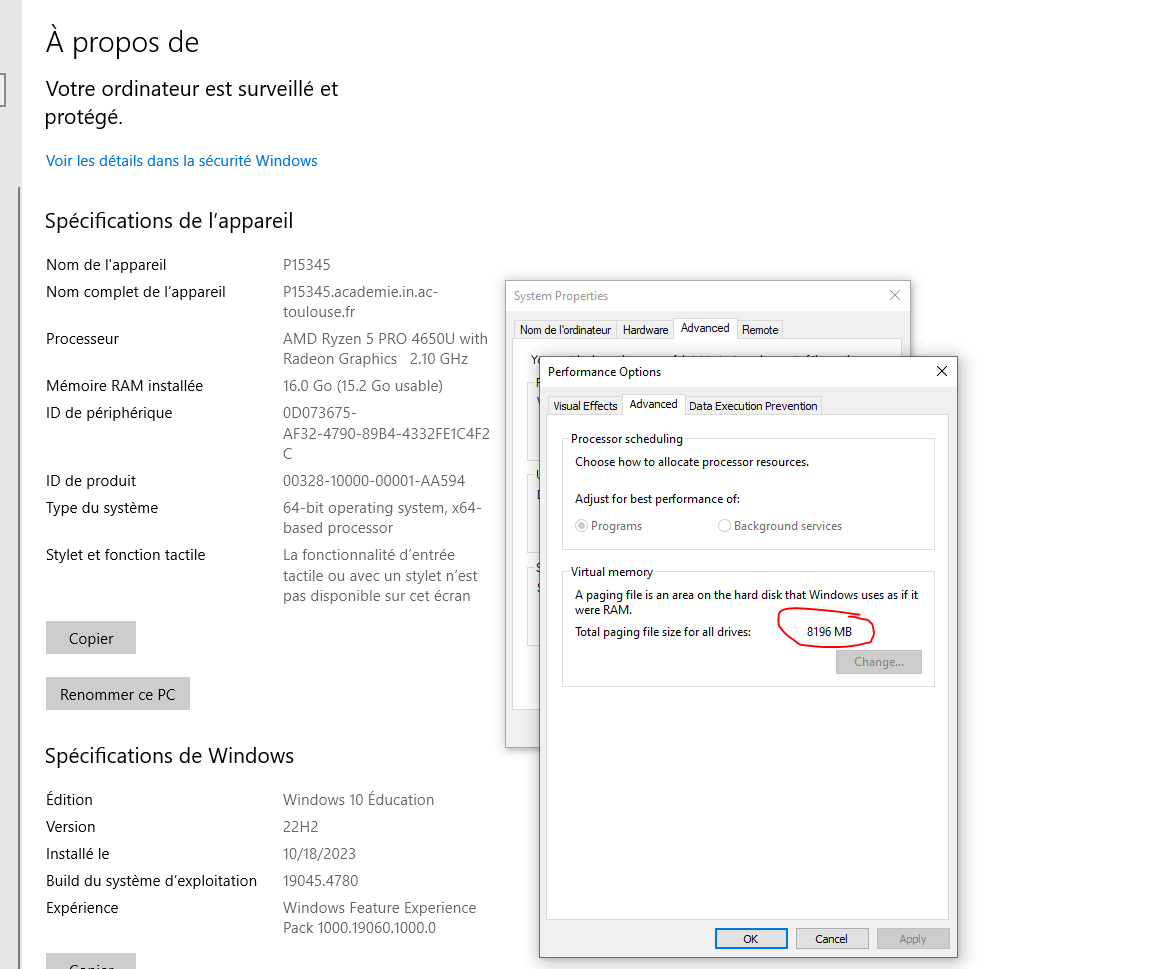
* Nouvelle metadoonées pour Tour1 de type excel : add number of rows limits equal to number of rows in the file
* Nouvelle metadonnées pour Tour2 de type excel

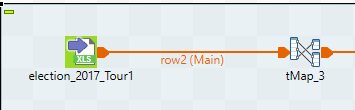
Tour1 :

* Lire les donnes :



* Filtrer uniquement Il de France :
  + Prerequis memory error 🡪 change the VM max size





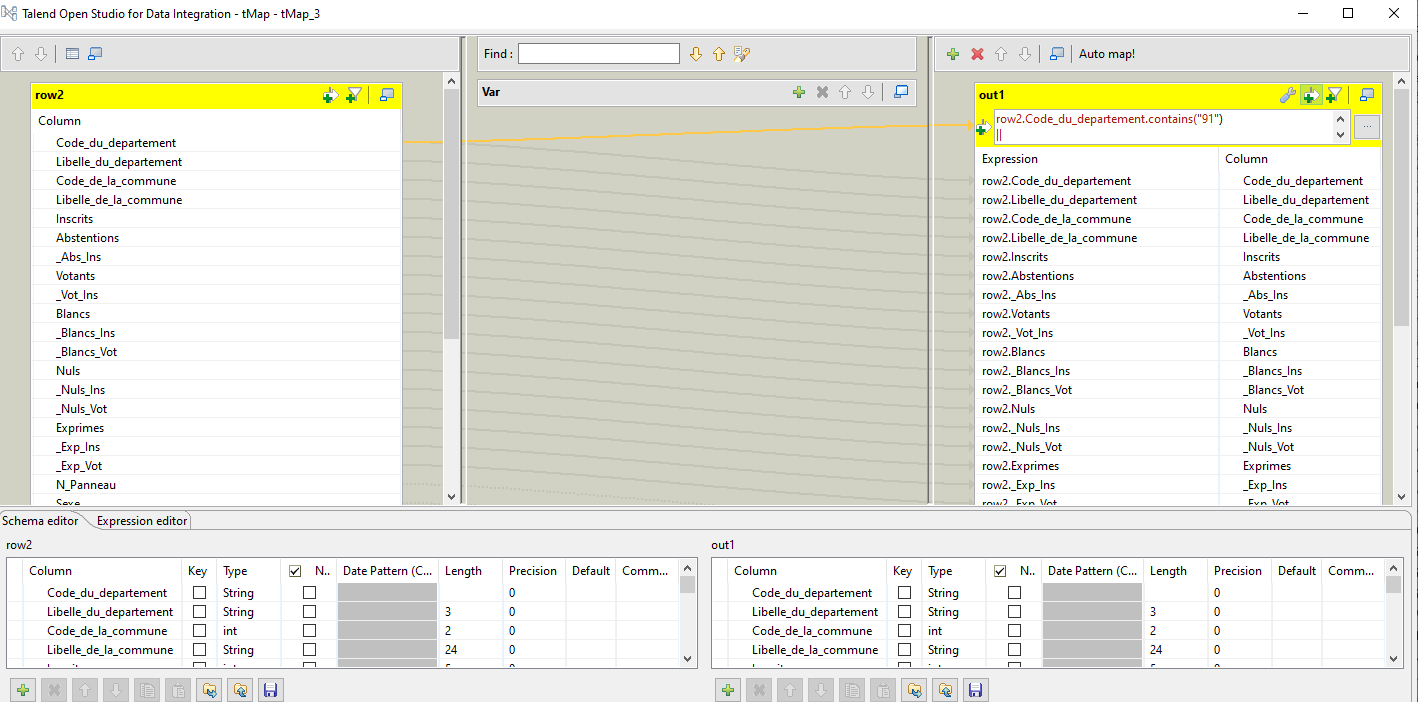
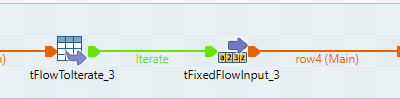
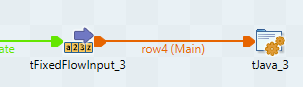


Figure 4tmap3

* Parcourir chaque lingne



* Traiter les donnes
  + Dans le tjava3 : on garde des donénes du fichier source et on ajoute des données statiques comme le tour et l’année le nombre du candidat , on mape chaque candidat avec le nombre de voix
    - On inser tous dans la liste global
  + 

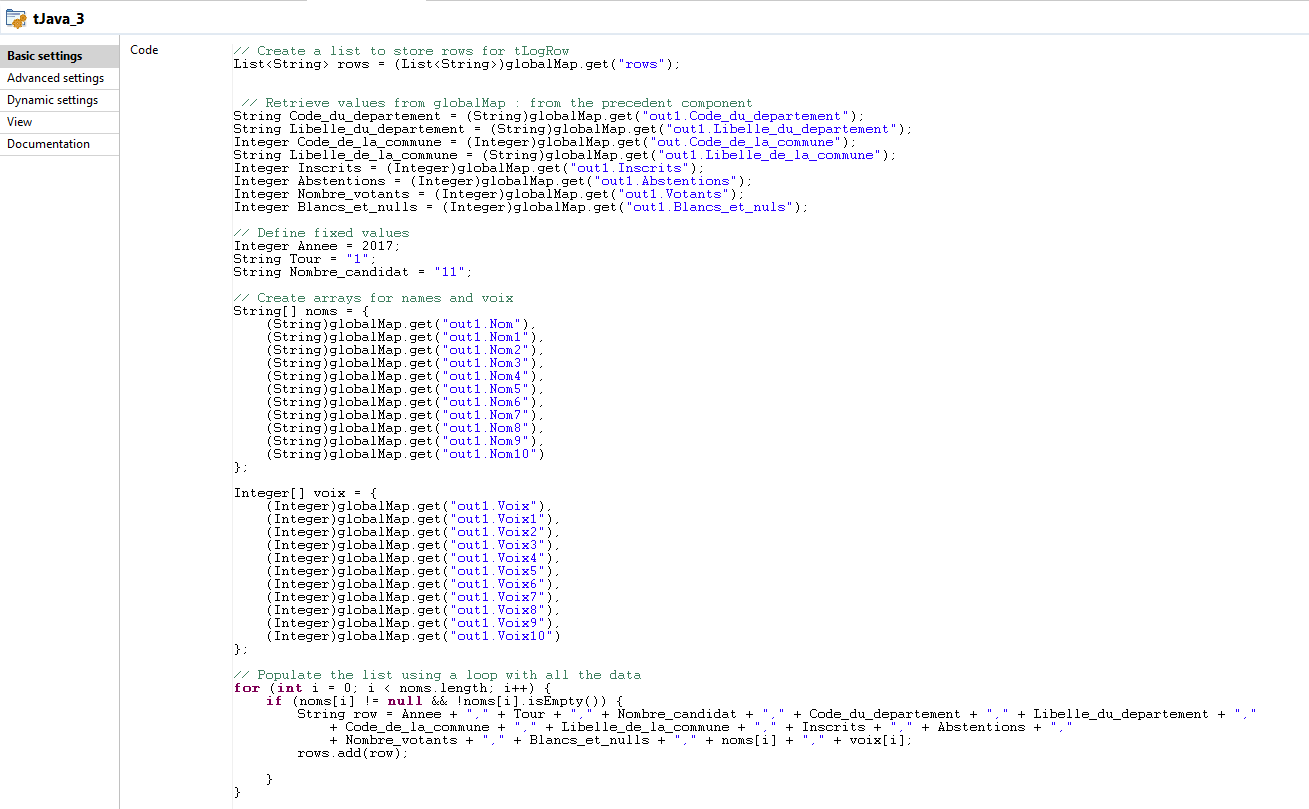
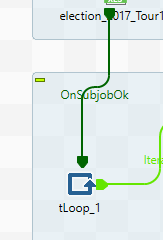
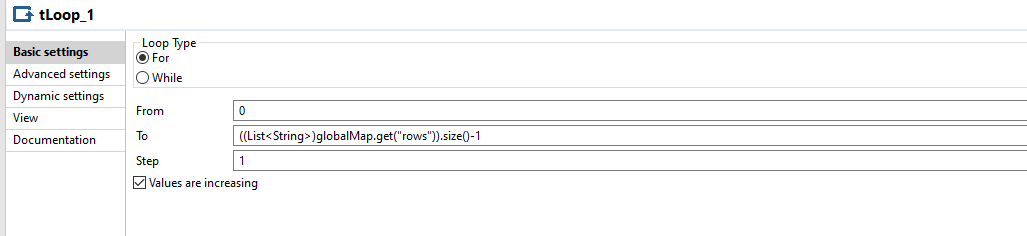


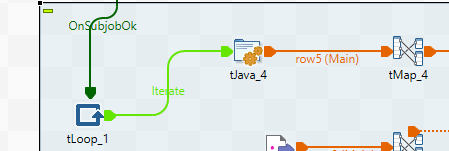
Figure 5Tjava3

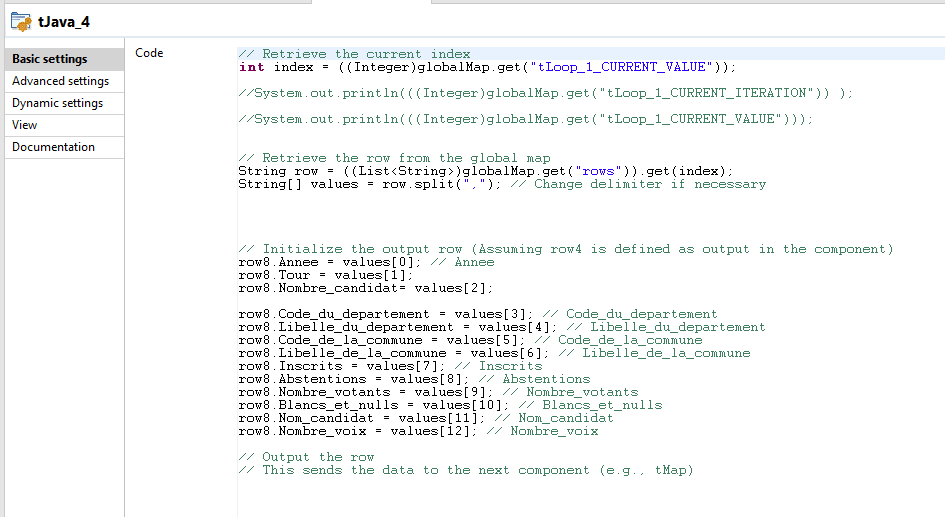
Ajouter les données sur le parti politique et sans orientation pour chaque canndat :

* Pour chaque ligne de la liste déjà créer : avec tloop



* 
* On passe les données ligne par ligne au composant suivant tmap4 :



* 

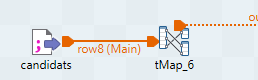
**Preparer pour faire la jointure**

**\*Source : la liste**

* Dans tmap4 on formate les collone nom et année pour avoir le meme format,pour faire une jointure

**\*Source : la liste :** Fichier candidats.csv

* + Nouvelle structure : ajout de la colonne Année
    - On peut avoir le meme candidat qui change de parti par année ,donc
    - On utilise annee et nom dans la jointure
  + Lire le fichier et pereparer les collones pour la jointure Anne et nom :



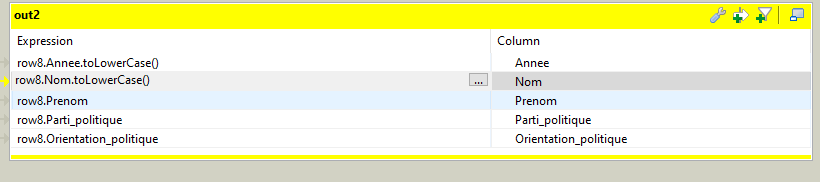
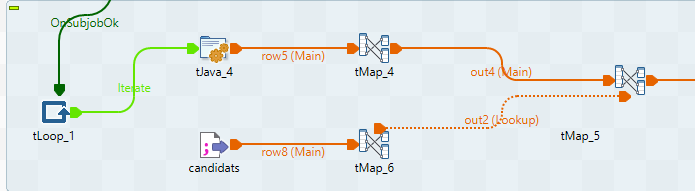


Figure 6tmap6

**Faire la jointure**

**Dans le tmap5 :**



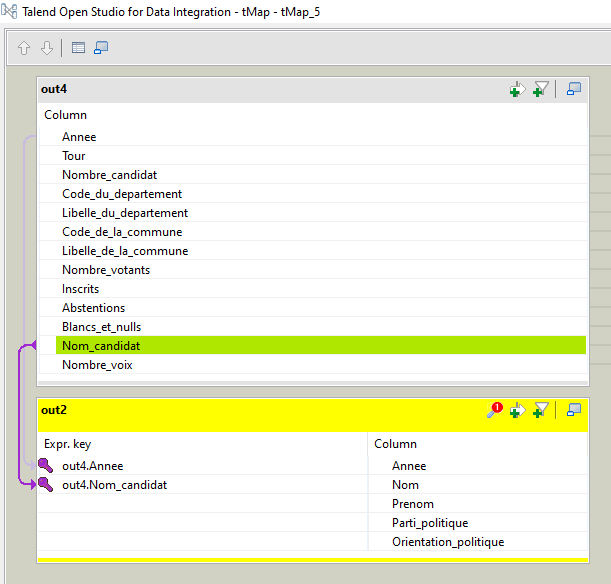


Figure 7tmap( Pour la jpointure

Bug dans le fichier resultat :

* Valeur null
  + Collonne : code de la commune , blancs et null
  + Que sur 2017 tour 1
* Debuge :
  + Fichier source : deux colonnes l : blanc ,nul elle sont sépraré .
  + Fichier resultat ,je fais reference blancs\_et\_null alors qu’il n’est pas dans le fichier source :
  + Solution :
    - On a une seul colonne blancs\_et\_nulls = somme des 2 colonnes
* 2017 TOUR2 : ajouté avec succes
* 2022 : on utilise le fichier source sous format texte car le excel est endommagé
  + Plusierus colonnes n’on pas de noms : on sait que entre deux nom on a 7 lignes donc on adapte le code dans Tjava3 ,pareil pour les voix.
  + Faire les meme étape que pour 2012 ET 2017
  + 2022 Tour 2