# Projet 2 NF26 : entrepôt de données et analyse des élections législatives françaises

Antoine LE ROUZIC - Pierre-Louis GUILLOT

Juin 2018

## 1 Modélisation du problème

Ce projet consiste à conceptualiser un datawarehouse en bases de données orientées colonnes permettant d'analyser statistiquement les données relatives aux différentes élections législatives selon des facteurs temporels – comme le suffrage et l'année du suffrage –, géographiques selon la situation du bureau de vote, et enfin politiques selon l'opinion exprimée par le vote.
C'est donc tout naturellement que nous avons pensé à développer une modélisation qui représentera comme fait élementaire le vote et comme dimension le **temps**, l'**opinion exprimée** et le **lieu**. On a décidé de présenter cette modélisation par l'UML que l'on peut retrouver dans la figure 1.

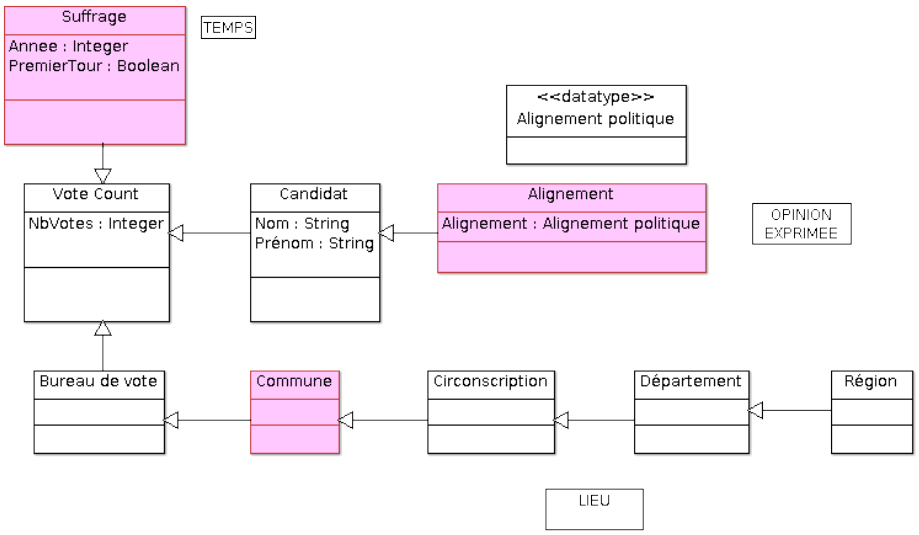


Figure 1: Modélisation du problème en UML

Le vote représenté sera une aggréation des *nb_votes* votes pour le même candidat lors du même suffrage dans un même bureau de vote.
Le temps sera matérialisé par un *suffrage* qui correspondra à la concaténation de l'année et du tour (par exemple "1er tour de l'élection 2017"). L'opinion exprimée sera représentée par le candidat pour lequel les votes ont été émis au niveau de granularité le plus fin mais aussi par l'alignement politique (la colonne *Nuance* fournie dans les jeux de données). Les lieux géographiques seront représentés par toutes les informations disponibles allant du bureau de vote pour le grain le plus fin aux régions pour le grain le plus épais.

On a agrégé tous les candidats du même alignement. Il y a en effet plusieurs candidats dans chacune des $\approx 577$ circonscriptions soit beaucoup trop pour permettre de faire une analyse lisible. Même en zoomant sur une circonscription particulière – pour laquelle se présentent les candidats – on a estimé qu'il n'y avait pas assez de communes différentes au sein de chacune pour tirer des analyses statistiques pertinentes des votes. On ne considérera donc dans notre modélisation que des *niveaux de zooms* plus faibles.
On a aussi agrégé tous les bureaux de vote de la même commune et pensé qu'il était beaucoup plus facile de trouver des statistiques pertinentes sur les communes, entités géographiques bien plus fortes que les bureaux de votes qui ne servent quasiment que pour les élections. De plus les données par bureaux de votes ne sont disponibles que depuis le suffrage de 2002.

**Note :** Les **données socio-économiques** n'ont pas été intégrées au modèle en UML pour la simple raison que celles-ci intéragissent à la fois avec des données géographiques et temporelles et qu'il est donc impossible de les modéliser rigoureusement sans créer de cycle entre différentes dimensions.

# 2 Matérialisation des données

## 2.1 Multiplicité des matérialisation

La philosophie générale de conception des entrepôts de données en NoSQL nous pousse à vouloir optimiser en priorité la **vitesse des analyses** plutôt que la **quantité de données stockées** pourvu que celles-ci soit bien distribuées.
On ne va donc pas hésiter à faire plusieurs matérialisation pour peu que celles-ci aient une certaine pertinence vis-à-vis de questions qu'on aurait pu se poser, et qu'elles respectent des contraintes qu'on va expliciter.

## 2.2 Clé de partitionnement et facilitation du traitement

Il est important de déterminer une façon de partitionner les données qui facilite le traitement qui pourra être fait lors d'analyses.
En effet si on décide de faire un traitement sur une sous-section particulière des données, il serait préférable que notre partitionnement nous permette d'accéder à la dite section sans devoir parcourir l'entièreté de la base. Selon ce principe on va déterminer les différents morceaux de l'information qui pourraient servir de critères de sélection lors de requêtes spécifiques et selon lesquels il serait intéressant de partitionner.
En premier lieu on remarque qu'il serait intéressant de pouvoir observer la répartition des votes, des taux d'absentions pour les différentes **nuances** et les différentes **communes** au sein d'un même département lors d'un ou plusieurs suffrages.
Dans un second temps il serait intéressant de faire les même analyses en séparant selon les **régions**, notamment pour pouvoir observer la répartition des votes d'un parti donné dans une même région selon le **département**.
Enfin il serait intéressant de répondre à des questions posées sur un **parti spécifique** selon des critères notamment géographiques.

## 2.3 Scalabilité horizontale

Par rapport à ce que l'on vient de dire, on aurait eu tendance à créer trois table avec les trois clés de partitionnement suivantes : Departement , Region et Nuance .
Cependant pour respecter la *scalabilité horizontale* et pouvoir augmenter le nombre de machines plutôt qu'améliorer les machines existantes, on va vouloir qu'au fil du temps ce soit **le nombre de partitions** qui augmente et pas **le nombre de lignes dans chaque partitions**.
Il est évident que si l'on choisi les clés de partitionnements évoquées ci-dessus la scalabilité horizontale

n'est pas respectée : à chaque nouveau suffrage de nouveaux votes seronts enregistrés dans chaque département, dans chaque région et pour chaque nuance.
Comme souvent on va donc vouloir rajouter une colonne issue d'une dimension temporelle dans les clés de partionnement. Plus précisément on a décidé, quitte à partitionner selon le temps, de séparer les données selon l'année de l'élection **et** le tour.
Il nous a en effet semblé que deux élections d'une même année pouvaient être analysées séparemment.
On a donc partitionné les données des différentes tables selon les clés suivantes :
| Departement | Année | Tour | , | Région | Année | Tour | et | Nuance | Année | Tour | .

## 2.4 Compromis nombre et tailles des partitions

Il est important pour choisir une clé de partitionnement de garder en tête le compromis qu'il peut y avoir entre la quantité de partitions et la taille des données dans chaque partition.
En effet, plus on augmente le **nombre de partitions**, plus il est facile de distribuer les données parmis différents noeuds du cluster. La granularité plus fine nous permet plus de souplesse dans le stockage.
Cependant, plus on augmente la **taille des partitions**, plus le traitement des données est facile : en effet les calculs effectués sur des très grandes partitions risquent d'être mieux distribués.
On est en outre certain que lorsqu'on traite un ensemble de données regroupées dans une même partition on se situe sur le même noeud du réseau. C'est une des façons dont la communication entre les partitions lors des calculs peut poser problème : réduire le nombre de partition facilite alors le traitement.
Il est nécessaire de trouver un compromis entre ces deux aspects, et donc de prendre une clé de partition qui produit un nombre raisonnable de partitions tout en gardant des tailles intéressantes.
Observons le nombre de partitions et leurs tailles pour les différentes façons de partitionner qu'on a évoqué dans la table 1.

| **Partitionnement selon** | **Nombre de partitions** | **Taille des partitions** |
|---|---|---|
| Année, Tour, Département | $100 \times \text{nb}_{\text{suffrages}}$ | $\frac{\text{nb}_{\text{communes}}}{100} \approx 3600$ |
| Année, Tour, Région | $27 \times \text{nb}_{\text{suffrages}}$ avant 2016<br>$14 \times \text{nb}_{\text{suffrages}}$ après 2016 | $\frac{\text{nb}_{\text{communes}}}{27} \approx 1333$ avant 2016<br>$\frac{\text{nb}_{\text{communes}}}{18} \approx 2000$ après 2016 |
| Année, Tour, Nuance | $\approx 20 \times \text{nb}_{\text{suffrages}}$ | $\text{nb}_{\text{communes}} \approx 36000$ |

Table 1: Nombres et tailles des partitions selon les différentes clés

Il nous semble que nous sommes parvenus à un bon compromis avec un nombre de partition qui reste faible devant la taille de celles-ci.
On remarque que les partitions seront beaucoup plus grandes pour la troisième table, puisque de manière générale celle-ci stocke beaucoup plus de lignes (une par commune **et** par nuance).
Un avis définitif dépendra cependant du contexte, des objectifs qu'on se fixe, et des contraintes matérielles sur lesquelles on aurait pu plus développer en situation concrète. Il nous est donc difficile de statuer définitivement sur ce point précis sans avoir fait d'implémentation réelle.

## 2.5 Expliciter les colonnes

Nous avons décidé de ne garder que les **nombres** entiers de votes et pas les **taux**.
En effet bien qu'on puisse déterminer l'un à partir de l'autre pour peu qu'on connaisse le nombres de votants total, les taux de votes sont présentés sous la forme de flottants plus difficiles à enregistrer. Ils risquent en outre de comporter des imprécisions et on n'est pas certain de retrouver le nombre de votes exact en multipliant le taux par le nombre de votant.

Pour les tables qui seront partitionnées selon | Departement | Année | Tour | et | Région | Année | Tour | nous avons décidé d'enregister **une seule ligne par commune** dans laquelle les votes pour les différents partis correspondront à différentes colonnes. L'objectif est de pouvoir répondre plus facilement aux questions portant soit sur les taux de votants **d'un parti précis** soit sur les **taux d'abstentions**. En faisant ainsi on aura besoin d'interroger moins de lignes à chaque fois.
Il nous sera cependant impossible de répondre **facilement** en **un seul map-reduce** à la question *"Quels sont les taux de vote pour les différents partis dans l'oise ?"* à moins de connaître à l'avance la liste des différents partis et de faire un mapping sur un vecteur de la forme (Nb_votes_parti_1, ..., Nb_votes_parti_N). On pourrait le faire en utilisant la table partitionnée selon | Nuance | Année | Tour | – qui elle contiendra un vote par commune **et** par parti, mais cela nous demandera d'interroger des données dans toutes les partitions alors que notre question portait seulement sur un département particulier.
Finalement, nous avons aussi pris le parti de rajouter "abstentions" et "blanc&nuls" comme des nuances avec leurs propres partition au sein de la troisième table selon l'idée que cela ne nous demandera de stocker que deux partis de plus parmi la vingtaine présente à chaque suffrage et que cela pourra nous permettra d'intégrer automatiquement les abstentions à des analyses qui regrouperaient tous les partis.

## 2.6 Indicateurs socio-économiques

On a décidé de choisir un nombre restreint d'indicateurs socio-économiques. On les a aussi choisi suffisamment simples pour qu'ils soient facilement exploitables dans les analyses.
Les indicateurs choisis sont le **revenu médian** et le **nombres d'habitants** des communes et des départements. On sait en effet qu'un vote dépendra énormément de la taille de la ville dans lequel il a été effectué ainsi que de son revenu médian et il en va de même pour le département d'où il provient.
On aurait pu intégrer un tel indicateur pour les régions. Celles-ci sont cependant peu nombreuses (27 ou 14 selon l'époque) et on s'est dit que cela ne valait pas le coup de faire des analyses statistiques à une échelle où on pourrait simplement comparer les élements entre eux *"à la main"*.
On imagine en effet que le fait de dire *"La région X parmi les 14 régions françaises a plus voté pour telle ou telle nuance"* pourra être assez directement intérprêté par l'utilisateur qui connaitra plus facilement toutes les régions que tous les départements.

## 2.7 Résultat

Finalement, pour répondre à toutes les contraintes qu'on a expliqué lors de cette partie, on a modélisé les trois matérialisations que l'on trouvera dans les tables 2, 3, et 4.

| **Clé de partitionnement** | **Clé de tri** | **Colonnes** |
|---|---|---|
| Année, Tour, Département | Commune | Inscrits, Abstentions, Votants, Blancs&Nuls, Exprimés, Nombre_habitants_commune, Revenu_median_commune, Vote_nuance_X, Vote_nuance_Y, ... |

Table 2: Première matérialisation "par département"

On gardera à l'esprit que toutes les données socio-économiques telles que Nombre_habitants_commune correspondent à des valeurs collectées **à un moment précis** qui ne seront pas constantes d'un suffrage à l'autre. On notera aussi que pour la table 3 on a choisi comme clé de tri | Département | Commune | et pas | Commune | étant donné qu'on enregistre le **numéro** de la commune qui est relatif au département auquel elle appartient.
Notons finalement qu'on aurait pu imaginer construire d'autres tables, notamment des tables par-

| Clé de partitionnement | Clé de tri | Colonnes |
|---|---|---|
| Année, Tour, Région | Département, Commune | Inscrits, Abstentions, Votants, Blancs&Nuls, Exprimés, Nombre_habitants_commune, Revenu_median_commune, Nombre_habitants_département, Revenu_median_département, Vote_nuance_X, Vote_nuance_Y, ... |

Table 3: Seconde matérialisation "par région"

| Clé de partitionnement | Clé de tri | Colonnes |
|---|---|---|
| Année, Tour, Nuance | Région, Département, Commune | Inscrits, Abstentions, Votants, Blancs&Nuls, Exprimés, Nombre_habitants_commune, Revenu_median_commune, Nombre_habitants_département, Revenu_median_département, Vote_nuance_X |

Table 4: Troisième matérialisation "par nuance"

titionnées autour de seuils sur nos indicateurs socio-économiques, mais qu'on a préféré garder dans un premier temps une matérialisation relativement simple.

# 3 Analyses des données en algorithmes distribués

## 3.1 Analyses pour la matérialisation 1 (table 1)

On peut calculer les taux de votes et d'abstentions pour chaque commune de France pour chaque tour de chaque legislative :

- $map : d \longmapsto \left( (\textbf{Annee, Tour, Département, Commune}), \left( \frac{\text{Votants}}{\text{Inscrits}}, \frac{\text{Abstentions}}{\text{Inscrits}} \right) \right)$

Ces résultats peuvent facilement être généralisés pour se restreindre aux communes d'un département précis.
On peut aussi calculer les taux de votes au sein d'un ou plusieurs départements :

- $map : d \longmapsto ((\textbf{Annee, Tour, Département}), (\text{Votants}, \text{Abstentions}, \text{Inscrits}))$
- $reduce : ((\text{V1}, \text{A1}, \text{I1}), (\text{V2}, \text{A2}, \text{I2})) \longmapsto (\text{V1} + \text{V2}, \text{A1} + \text{A2}, \text{I1} + \text{I2})$
- $map : d \longmapsto \left( (\textbf{Annee, Tour, Département}), \left( \frac{\text{Votants}}{\text{Inscrits}}, \frac{\text{Abstentions}}{\text{Inscrits}} \right) \right)$

**À noter:** Pour intégrer les indicateurs socio-économique à l'analyse on peut simplement rajouter dans la clé du mapping les données de la commune comme le nombre d'habitants et le revenu médian.

## 3.2 Analyses pour la matérialisation 2 (table 2)

On peut calculer le nombre de votes et d'abstentions pour une ou plusieurs régions pour chaque tour de chaque legislative :

- $map : d \longmapsto ((\textbf{Annee, Tour, Région}), (\text{Votants}, \text{Abstentions}, \text{Inscrits}))$
- $reduce : ((\text{V1}, \text{A1}, \text{I1}), (\text{V2}, \text{A2}, \text{I2})) \longmapsto (\text{V1} + \text{V2}, \text{A1} + \text{A2}, \text{I1} + \text{I2})$

- $map : d \longmapsto \left( (\textbf{Annee, Tour, Région}) , \left( \frac{\text{Votants}}{\text{Inscrits}}, \frac{\text{Abstentions}}{\text{Inscrits}} \right) \right)$

On peut aussi calculer les taux de vote et d'abstention un ou plusieurs département appartenant ou non à la même région, selon le map-reduce qu'on a déjà vu pour la matérialisatoin 2

De manière similaire nous pouvons obtenirs les taux de votes et d'abstentions pour **toute la France** pour un suffrage donné. Il faut pour cela faire des mapping qui ne prennent en comptes que l'année et le tour.
On oubliera pas pour tous les map-reduce où nous comptons les votes de parti de transformer le nombre de votes en 0 lorsque la colonne du parti n'est pas dans la ligne correspondante. Cela revient à évaluer l'absence d'une nuance dans une ville de la même façon que si tous les candidats du parti avaient cumulé 0 votes. Cette approximation nous semble acceptable : le simple fait qu'il n'y ait pas de candidat représente un échec pour la nuance.
On pourrait imaginer d'autres map-reduce comptant le taux de vote uniquement **dans les villes où le parti est représenté** en mettant dans le mapping le nombre d'inscrits à 0 dès lors que le parti n'y est pas.
Enfin on peut ajouter à l'analyse les données socio-économiques des départements en rajoutant leurs nombres d'habitants et leurs revenus médians à la liste des clés dans le mapping, ou en créant des seuils sur ces facteurs qui permettraient un mapping différent.

## 3.3 Analyses pour la matérialisation 3 (table 3)

On peut calculer le taux de votes pour une nuance pour chaque départements lors d'un suffrage donné :

- $map : d \longmapsto ((\textbf{Annee, Tour,Nuance, Département}) , (\text{Votes\_nuance}, \text{Inscrits}))$
- $reduce : ((\text{Votes\_1}, \text{Inscrits\_1}) , (\text{Votes\_2}, \text{Inscrits\_2})) \longmapsto (\text{Votes\_1} + \text{Votes\_2}, \text{Inscrits\_1} + \text{Inscrits\_2})$
- $map : d \longmapsto \left( (\textbf{Annee, Tour,Nuance, Département}) , \left( \frac{\text{Votes\_nuance}}{\text{Inscrits}} \right) \right)$

Soit la constante $\text{seuil}_{revenu}$ un seuil pour les revenus médian qu'on peut calculer par exemple en prenant le revenu médian médian au sein du pays ou des départements – selon la méthode qui permet de calculer les médianes par map-reduce qu'on a étudié en cours. On peut alors calculer le taux de votes pour une nuance selon si les départements ont un revenu médian inférieur ou supérieur au dit seuil :

- $map : d \longmapsto \left( \left( \textbf{Annee, Tour,Nuance}, \mathbb{1}_{\text{r\_m\_d} \leq \text{seuil}_{\text{nuance}}} \right) , (\text{Votes\_nuance}, \text{Inscrits}) \right)$
  Où r_m_d le revenu médian du département
- $reduce : ((\text{Votes\_1}, \text{Inscrits\_1}) , (\text{Votes\_2}, \text{Inscrits\_2})) \longmapsto (\text{Votes\_1} + \text{Votes\_2}, \text{Inscrits\_1} + \text{Inscrits\_2})$
- $map : d \longmapsto \left( (\textbf{Annee, Tour,Nuance, Departement\_inferieur\_seuil}) , \left( \frac{\text{Votes\_nuance}}{\text{Inscrits}} \right) \right)$

Cela nous permet d'observer la façon dont intéragit le fait d'être dans un département plus aisé avec le fait de voter pour tel ou tel parti.
On aurait ainsi pu concevoir moult opérations de map-reduce généralisant cette logique de seuil – à commencer par une qui ferait la même chose en prenant en compte les revenus médians **de la commune**.
Nous n'avons cependant présenté ici qu'un échantillon limité dans le but de donner un aperçu du type d'information que l'on pourrait extraire en matérialisant ainsi le jeu de données.