INFO5
Mamisoa RANDRIANARIMANANA & Xianxiang ZHANG

Qualité des données

Mamisoa RANDRIANARIMANANA & Xianxiang ZHANG

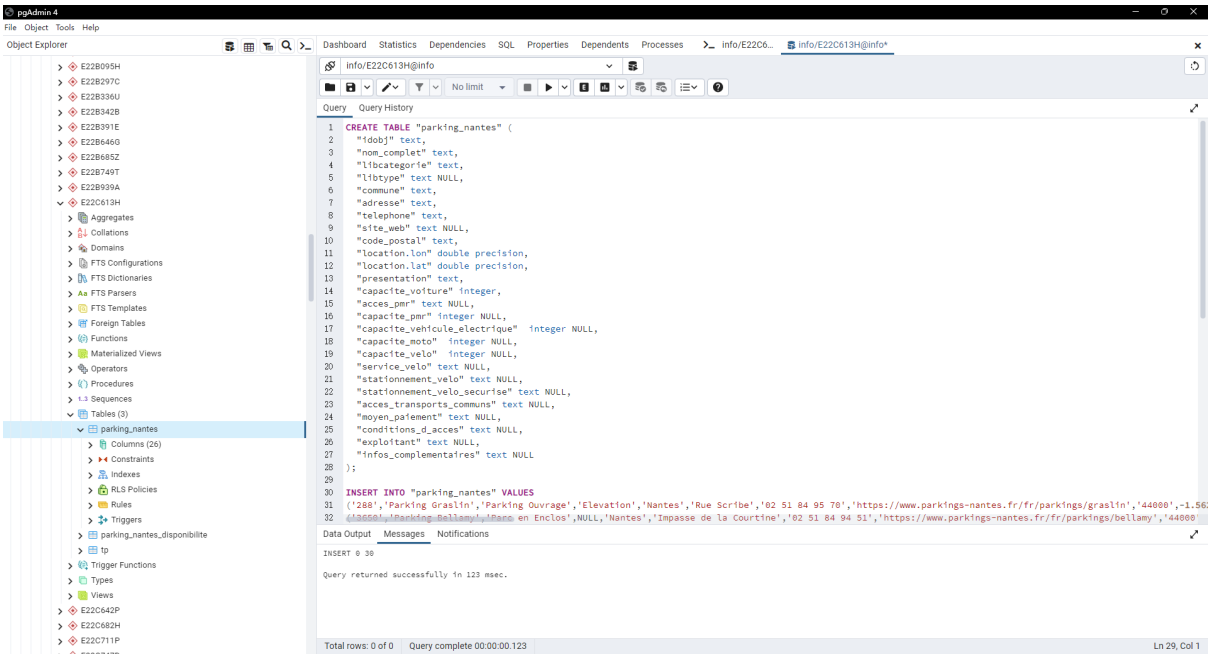
Jeux de données :
Pour notre étude, nous avons choisi de prendre les données sur la liste des parkings publics de Nantes et la liste pour les disponibilités de ces parkings.

Parkings publics de Nantes - Disponibilités :
https://data.nantesmetropole.fr/explore/dataset/244400404_parkings-publics-nantes-disponibilites/table/?disjunctive.grp_nom&disjunctive.grp_statut

Parkings publics de Nantes :
https://data.nantesmetropole.fr/explore/dataset/244400404_parkings-publics-nantes/table/?disjunctive.libcategorie&disjunctive.libtype&disjunctive.acces_pmr&disjunctive.service_velo&disjunctive.stationnement_velo&disjunctive.stationnement_velo_securise&disjunctive.moyen_paiement

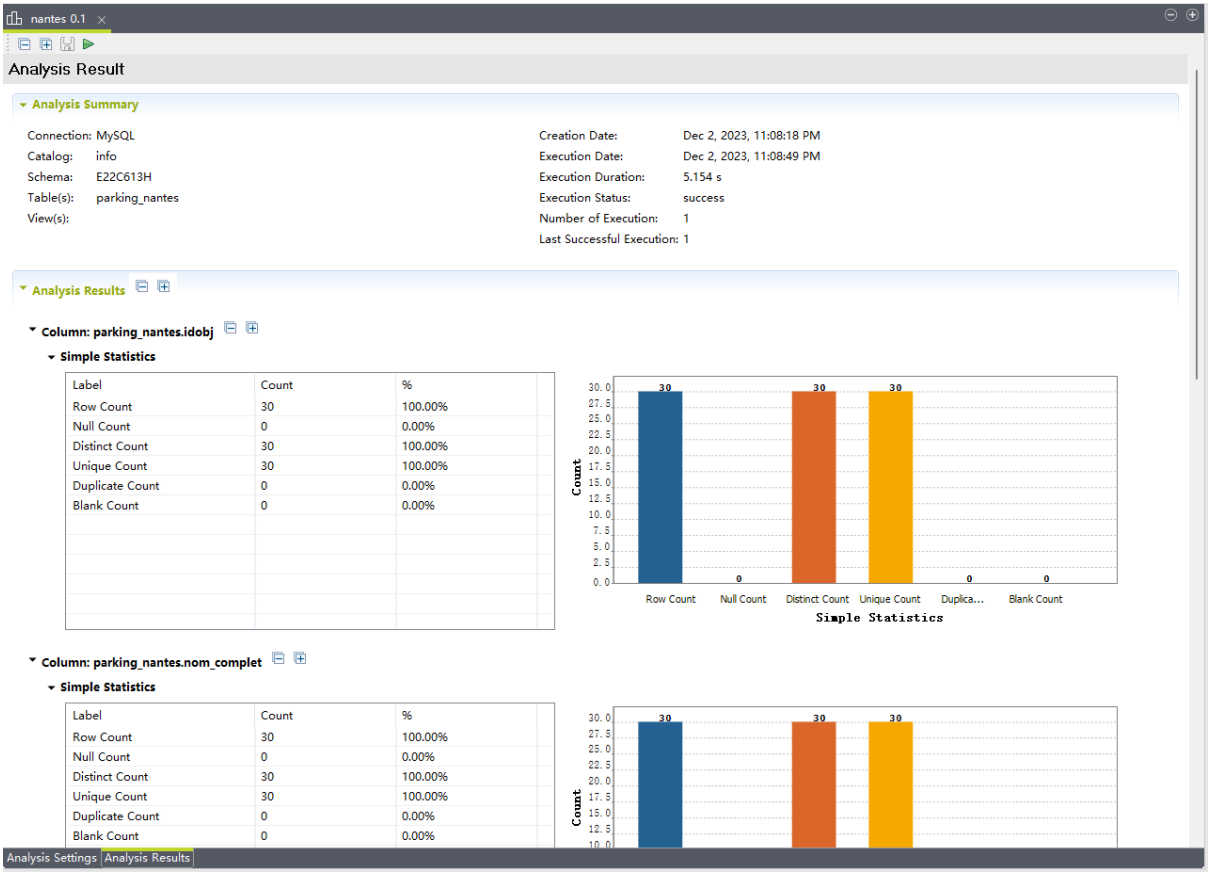
Préparation d'analyse :

Comme système de gestion des bases de données, nous avons décidé d'utiliser PostGreSql Nous utilisons pgAdmin 4 pour importer parking_nantes et parking_nantes_disponibilite dans la table via l'outil SQL.

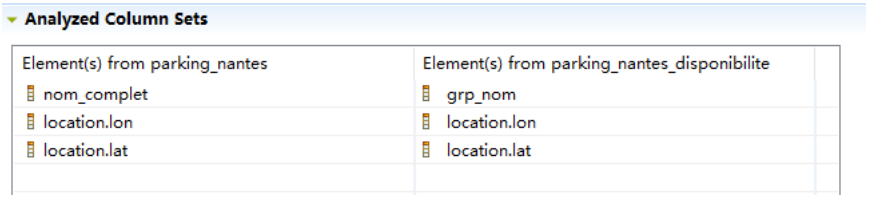


Le fait de pouvoir analyser la table après avoir ouvert Talend montre que nous avons réussi à importer les informations de la table.

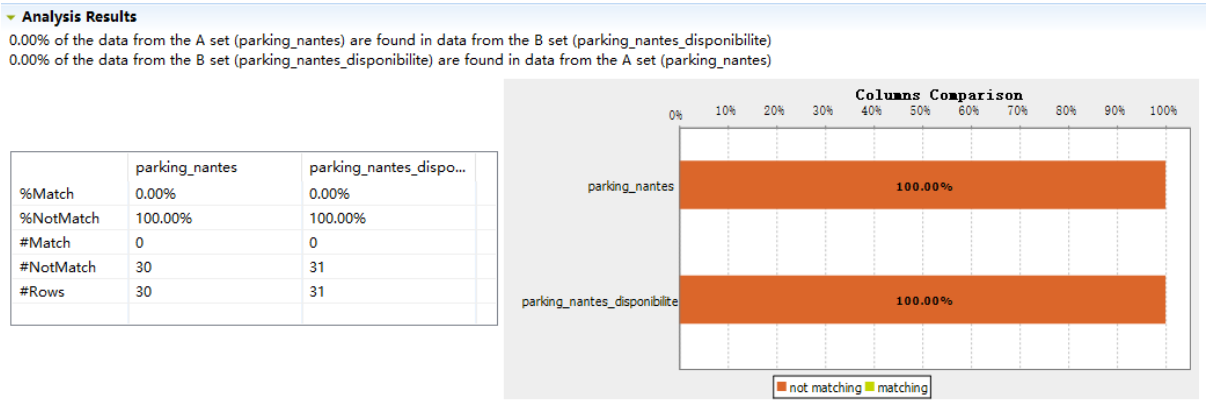
INFO5
Mamisoa RANDRIANARIMANANA & Xianxiang ZHANG



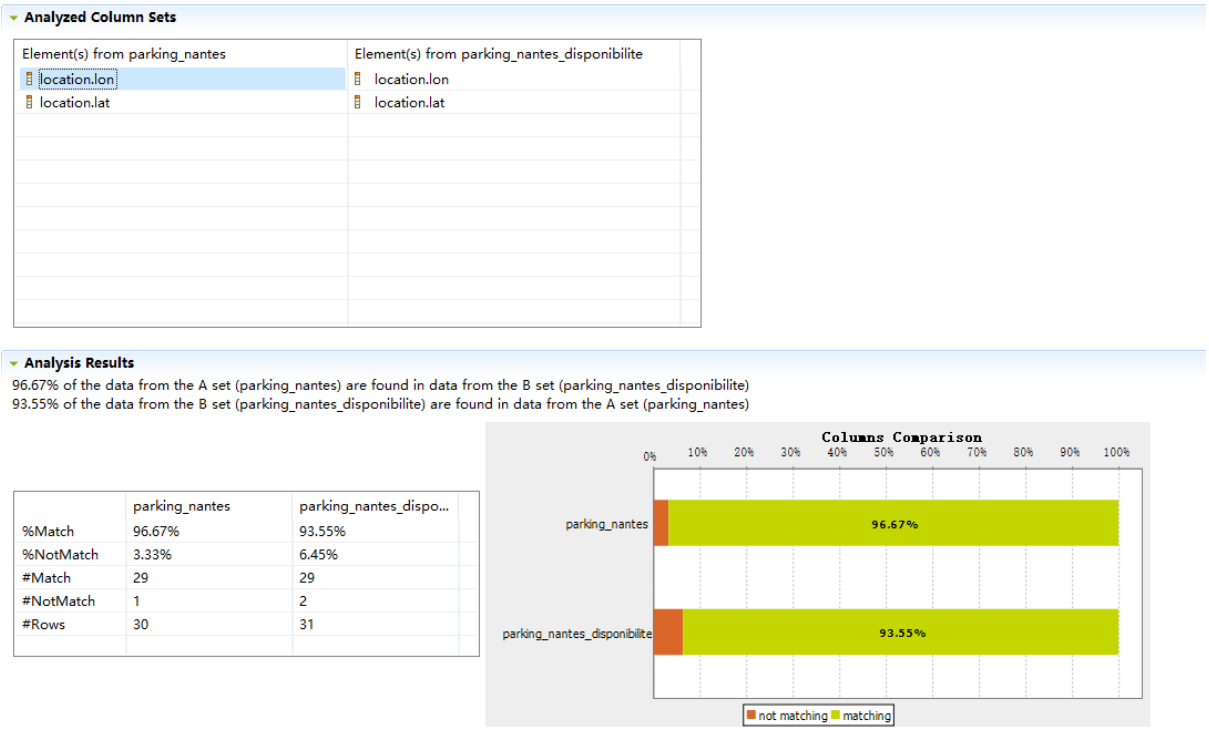
Analyse des redondances :



Nous avons utilisé le nom, la longitude et la latitude pour effectuer une première analyse de redondance. Nous avons remarqué qu'il n'y avait aucune correspondance. Ceci est dû à la dénomination différente des parkings.

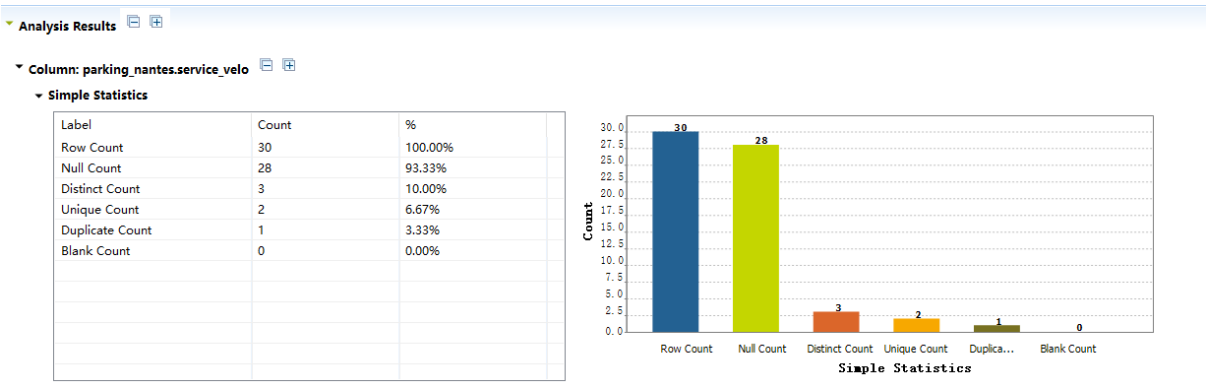


Mais lorsque nous basculons le paramètre de comparaison sur l'analyse des coordonnées de position, nous obtenons les résultats suivants :



En effet, les noms dans les deux tables ont des formes différentes, Talend ne peut donc pas rechercher le même nom (même si leurs noms sont identiques).
Par exemple, dans la table parking_nantes le nom est défini comme "Parking Feydeau", mais dans la table parking_nantes_disponibilite le nom est défini comme "Feydeau".
On est alors confronté à des problèmes de qualité de donnée liés à la cohérence dans la manière dont les noms sont enregistrés dans les deux tables des , car même s' ils parlent du même parking, leurs noms dans les différentes tables sont différents. Si l'objectif est de fusionner les données des deux tables en fonction du nom, ou faire des correspondances, ces variations peuvent entraîner des erreurs ou des résultats incomplets.

Une simple analyse des colonnes peut également révéler des informations intéressantes.
Par exemple, l'image ci-dessous montre un parking proposant des services de vélo.

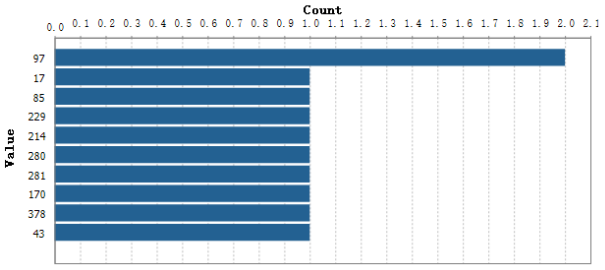


L'image ci-dessous montre le nombre de parkings avec le même nombre de places de stationnement.

INFO5
Mamisoa RANDRIANARIMANANA & Xianxiang ZHANG

Value Frequency

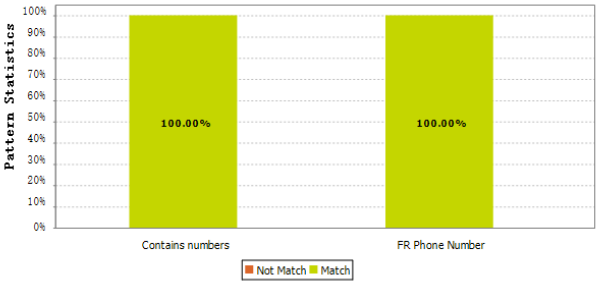
Value	Count	%
97	2	6.45%
17	1	3.23%
85	1	3.23%
229	1	3.23%
214	1	3.23%
280	1	3.23%
281	1	3.23%
170	1	3.23%
378	1	3.23%
43	1	3.23%



Nous pouvons également utiliser la correspondance de modèles pour vérifier si les numéros de téléphone du tableau sont conformes au format. (Vérification préliminaire de la colonne du numéro de téléphone pour les erreurs).
Si les numéros de téléphone ne sont pas conformes aux formats , on peut avoir un problème de qualité de donnée. Ici, tous les numéros ont des formats corrects. Cependant, on pourra toujours avoir de problème d'authenticité qu'on ne pourra pas vérifié ici si le numéro de téléphone renseigné n'est pas la bonne.

Pattern Matching

Label	Match%	Not Mat...	Match	Not Match
Contains numbers	100.00%	0.00%	30	0
FR Phone Number	100.00%	0.00%	30	0



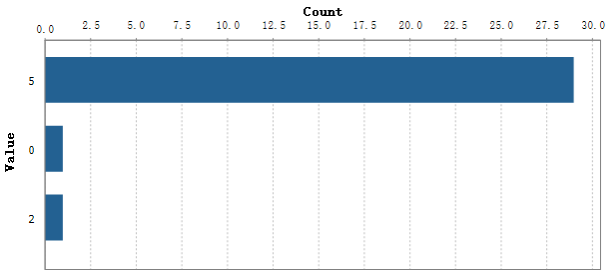
Nous pouvons également vérifier l'état du parking grâce à la fréquence des valeurs.

0 Invalide (comptage hors service)
1 Groupe parking fermé pour tous clients
2 Groupe parking fermé au client horaires et ouvert pour les abonnés
5 Groupe parking ouvert à tous les clients.

Column: parking_nantes_disponibilite.grp_statut

Value Frequency

Value	Count	%
5	29	N/A
0	1	N/A
2	1	N/A

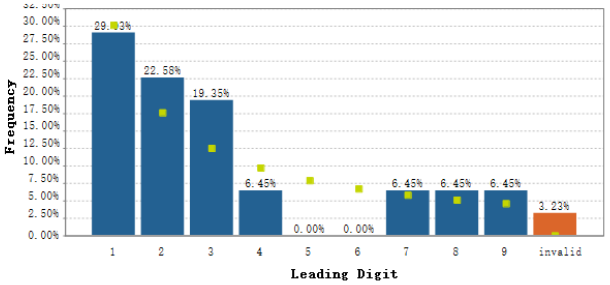


On peut également produire des graphiques pour comparer la distribution des données avec des lois mathématiques. Talend permet de comparer la distribution avec la loi de Benford.
L'état des parking pourrait être confronté à des problèmes de mise à jour, qui est un problème de qualité de donnée.

INFO5
Mamisoa RANDRIANARIMANANA & Xianxiang ZHANG

Benford Law

Leading Digit	Count	%
1	9	29.03%
2	7	22.58%
3	6	19.35%
4	2	6.45%
5	0	0.00%
6	0	0.00%
7	2	6.45%
8	2	6.45%
9	2	6.45%
invalid	1	3.23%



En conclusion, notre analyse des jeux de données sur les parkings publics de Nantes a révélé des défis intéressants liés à la qualité des données. Les variations dans la dénomination des parkings entre les deux tables ont présenté des obstacles lors de la recherche de correspondances. Les problèmes de cohérence dans l'enregistrement des noms peuvent entraîner des erreurs ou des résultats incomplets lors de la fusion des données.

De plus, l'analyse des colonnes a permis de mettre en évidence des informations pertinentes, telles que la présence de services de vélo dans certains parkings. La vérification de la conformité des numéros de téléphone a montré une bonne qualité des données, bien que des problèmes d'authenticité puissent subsister.

En utilisant des méthodes telles que l'analyse de fréquence des valeurs d'état des parkings et la comparaison avec la loi de Benford, nous avons exploré différentes perspectives pour évaluer la distribution des données. Ces approches fournissent des outils utiles pour détecter des anomalies ou des tendances dans les jeux de données.