Prétraitement - Importation et nettoyage de données

Amina, KACIMI Valentin, DEHAINAULT Minh-Hoang, DANG

December 7, 2018

1 Importer les donées brutes dans le SGBD

On a la possibilité d'importer les données à partir d'un fichier CSV avec la requête `COPY ... FROM`
L'import des données va invoquer les règles définies sur la table cible.

```
COPY FROM will invoke any triggers and check constraints on the destination table.
However, it will not invoke rules.
```

Nous allons donc préparer deux tables vides, avec les attributs de types adéquats, puis les règles pour les données entrantes. La première table, nommée `CorrectionTemp` servira à la redirection des données vers la vraie table `Correction`, via une multitude de triggers.

2 Nettoyage de données

Comme nous l'avons discuté précédemment, nous allons écrire un `TRIGGER` pour chaque 'problème' ci-dessous. Afin d'obtenir le résultat le plus satisfaisant, nous avons besoins d'une librairie de fonctions qui permettent de traiter les données entrantes sous forme de `string`.

Fonction	But
split_string(str,delim1, delim2)	Séparer un string délimité par deux délimiteurs en un ensemble de strings
traitement_date(str)format d'entrée : année / mois année	Retourner un string contenant uniquement le mois si précisé et l'année de prise d'une photo
traitement_n_v(str)	Harmoniser la colonne négative ou inversible de manière à ce qu'elle contienne NEG pour négatif et INV pour inversible
array_expand(array, maxlength, fill)	Étendre le array jusqu'à certaine taille en ajoutant les éléments. Cette fonction est particulièrement utile pour combler les vides après la séparation des données.

2.1 Éliminer les incohérences dans les données

Nous avons remarqué une certaine incohérence dans les données: plusieures syntaxes pour exprimer une information (négatif/verre négatif), plusieurs types de tâches ou d'usure référencées, ... Ce traitement est chargé dans le trigger qui fait la séparation des lignes.

Pour supprimer les doublons, on procède de la manière suivante: On vérifie si deux lignes sont identiques, plus précisément si toutes les valeurs contenues dans une ligne sont identiques à celles d'une autre si c'est le cas on supprime.

2.2 Séparer les lignes combinées

Pour séparer les lignes combinées, nous avons crée un trigger qui détecte la présence des séparateurs (virgules, pipe, slash) selon chaque attribut du fichier CSV (donc la présence de multiples informations dans une même case). Avec les délimiteurs, on sépare les informations combinées dans un array correspondant à l'attribut concerné. Finalement, on passe les attributs dans la clause `VALUES` de `INSERT`.

2.3 Les Cantons associés aux communes

Après avoir récupéré un fichier csv d'une opendata contenant les coordonnées lambert 93 (lambertX,lambertY) de toutes les villes de france. - on mets à jour notre table pour ajouter deux colonnes LambertX et LambertY. - on crée une nouvelle table Villes qui contiendra les noms ainsi que les coordonnées Lambert de chaque ville. - on update notre table pour insérer les coordonnées.

2.4 Conclusion:

Pour résumer, le prétraitement se divise en sous-tâches, exécutées dans l'ordre suivant:
- Créer deux tables:
- Une table pour accueillir les données brutes
- Une table pour accueillir les données traitées - Séparer les lignes combinées:
- Éliminer des incohérences (regexp + fonctions)
- Transférer les données dans la bonne table (INSERT)
- Compléter les informations - Suppression des informations insignifiantes
- Ajouter les coordonnées

```
In [ ]:
```