INSTITUT DE TECHNOLOGIE
UNIVERSITÉ Clermont Auvergne

SD Aurillac Science des données Cybersécurité

Préparation et intégration de données 2

Kilian GIRAUD | Marie CHALLET

Anonymisation de données
BUT Sciences des Données | Option VCOD

2023-2024

Professeur référent :
Clement JACQ

Table des matières

1 Introduction **2**

2 Remerciement **2**

3 Quels niveaux a-t-on choisi ? **3**

4 Application des niveaux aux jobs **3**
4.1 Données identifiantes 4
4.1.1 Nom et Genre 4
4.1.2 Age 5
4.1.3 Localisation 5
4.1.4 Date 6
4.2 Données sensibles 6
4.2.1 Maladie, Etat et Traitement 6

5 Résultats **7**
5.1 K-Anonymité 7
5.2 L-Diversité 7

6 Conclusion **8**

7 Références **9**

1 Introduction

Talend est utile pour extraire et transformer les données mais est tout aussi utile pour securiser et anonymiser une base de données.

Pour ce projet, une base de données "médicale" où sont disponibles plusieurs informations sur un patient.

Le but est d'identifier les variables identifiantes et sensibles afin d'effectuer plusieurs niveau d'anonymisation que l'on vérifiera avec la *K-Anonymité* et la *L-Diversité*, tout cela dans plusieurs jobs **Talend** (un pour chaque colonne / groupe de colonnes).

2 Remerciement

Nous tenons à remercier notre professeur **Clément JACQ** pour l'aide qu'il nous aura fournit lors de la rélisation de ce projet, nous aurions probablement pas réussi seuls. La reprise des bases de l'anonymisation aura été primordiale pour la compréhension totale du sujet et donc pour une bonne réalisation.

3 Quels niveaux a-t-on choisi ?

Il y a en tout 4 niveaux d'anonymisation dans notre procédure, allant du moins anonymisant au plus anonymisant (0 -> 3), nous allons les lister ci-dessous :

- **Niveau 0 :** Ce niveau n'est pas du tout anonymisant, il renvoie tout simplement la base de données telle qu'elle, c'est intéressant de l'avoir afin de tout de même calculer les valeur de K et L.
- **Niveau 1 et 2 :** Niveaux intermédiaires
- **Niveau 3 :** Ce niveau correspond au maximum de l'anonymisation, les valeurs de K et L seront bien evidemment élevées mais la perte d'informations sera également conséquentes.

4 Application des niveaux aux jobs

Nous utilisons différents jobs pour anonymiser les données selon un niveau choisi. Pour sélectionner le niveau, nous utilisons les contextes de Talend en utilisant une variable "niveau" pouvant prendre les valeurs 0, 1, 2 ou 3 selon le niveau choisi. Pour chaque job, le niveau 0 correspond à la table non anonymisée, tandis que le niveau 3 correspond à la table la plus anonymisée tout en conservant plusieurs lignes différentes. Les niveaux 1 et 2 sont des niveaux intermédiaires que nous avons choisis en fonction de nos préférences et de notre intuition.

À la fin de chaque job, nous exportons la table résultante afin d'utiliser un job pour joindre chaque attribut anonymisé dans une seule et même table, permettant ainsi d'anonymiser l'ensemble de la table. Une fois la table entièrement anonymisée, elle est ensuite exportée dans un fichier CSV.3.1 Données identifiantes

4.1 Données identifiantes

4.1.1 Nom et Genre

Nous avons créé une routine nommée "nomsetgenres" comprenant trois fonctions distinctes :

- La première fonction, *getPrenoms(int niveau)*, utilise la fonction *getFirstName()* de TalendDataGenerator tout en ajoutant des prénoms féminins pour améliorer le niveau d'anonymisation. Selon le niveau choisi, la liste de prénoms varie en longueur. Par exemple, dans le dernier niveau, la liste ne comprend qu'un prénom de chaque genre (Jane et John).
- La deuxième fonction, *getNoms(int niveau)*, génère un nom de famille en sélectionnant aléatoirement un nom dans une liste des noms les plus populaires en France, dont la longueur dépend du niveau choisi. Par exemple, le dernier niveau ne comporte qu'un seul nom, "Doe", pour correspondre aux prénoms du niveau trois, représentant le nom complet des personnes non identifiées aux États-Unis (Jane Doe ou John Doe).
- La dernière fonction est *getGenre(String prenom, int niveau)*, prenant en paramètre le prénom généré précédemment et le niveau d'anonymisation souhaité. Deux tMaps sont nécessaires pour anonymiser les prénoms, les noms et les genres.

Les différents niveaux d'anonymisation sont les suivants :

- **Niveau 0** : Aucun changement sur la base de données.
- **Niveau 1** : Génère aléatoirement un prénom et associe un genre correct à ce prénom.
- **Niveau 2** : Génère un genre aléatoire.
- **Niveau 3** : Ne renvoie aucun genre, juste un "*".

Pour décrire brièvement le processus du job, nous séparons d'abord les noms et prénoms *(tExtractDelimitedField)*, puis nous anonymisons (ou non) les noms et les genres *(2 tMap)*, ensuite nous réassocions les noms et prénoms dans une seule colonne avant de renvoyer le jeu de données *(tFileOutputDelimited)*.

À la fin de ce job, nous exportons la table dans un fichier nommé *"nomsgenre_ out.csv"* situé dans un dossier nommé "out".

4.1.2 Age

Les ages ont été anonymisés par des classes (intervalles) ayant aucun changement pour le **niveau 0**, des amplitudes de 10 pour le **niveau 1**, 50 pour le **niveau 2** et max pour le **niveau 3**.

Pour ce faire, nous avons créé une routine appelée *"getAge"*.

À la fin de ce job, nous exportons la table dans un fichier nommé *"age_ out.csv"* situé dans un dossier nommé *"out"*.

4.1.3 Localisation

Cette section anonymise deux attributs interdépendants : l'attribut de résidence et l'attribut du code postal. Pour ce faire, nous avons utilisé un fichier CSV externe comprenant différentes colonnes telles que le code postal, le département, la région, la préfecture du département, une colonne vide contenant des astérisques, et enfin une colonne indiquant si la région se situe au Nord, au Centre ou au Sud de la France. Étant donné que certaines villes non métropolitaines sont incluses dans les données, nous avons choisi de les prendre en compte en utilisant les premiers chiffres de leur code postal, commençant par 97 ou 98.

Nous utilisons un *TMap* pour réaliser une fusion entre les données et cette table de localisation afin de remplir les colonnes avec d'autres informations et maximiser l'anonymisation. Le niveau de confidentialité choisi détermine le degré d'anonymisation. Par exemple, au **niveau 1**, la colonne de résidence est convertie en nom du département et la colonne du code postal ne conserve que le numéro du département. Pour le niveau supérieur, nous avons hésité entre conserver la région ou effectuer une anonymisation plus poussée en utilisant la situation géographique *(Nord, Centre, Sud, Autres)*, en ne conservant qu'un astérisque dans la colonne du code postal. Finalement, nous avons opté pour la seconde option après avoir calculé le K-Anonymat et la L-Diversité.

Enfin, pour le dernier niveau, nous avons décidé de ne mentionner que le pays, ici la France. À la fin de ce job, nous exportons la table dans un fichier nommé *"localisation_ out.csv"* situé dans un dossier nommé *"out"*.

4.1.4 Date

Ici, nous avons anonymisé l'attribut date de consultation. Pour ce faire, nous avons décidé de fractionner cet attribut à partir des "/". Cela nous permet d'obtenir des colonnes pour le jour, le mois et l'année. Selon le niveau d'anonymisation, la table finale comportera soit le mois et l'année, soit la saison, soit la décennie. Pour le **niveau 1**, seuls le mois et l'année sont conservés ; pour le **niveau 2**, ce sont les saisons ; et enfin, pour le **niveau 3**, ce sont les décennies. Pour cela, nous avons créé une routine *"getSaison"* pour avoir la saison correspondant aux mois.

À la fin de ce job, nous exportons la table dans un fichier nommé *"date_out.csv"* situé dans un dossier nommé *"out"*.

4.2 Données sensibles

4.2.1 Maladie, Etat et Traitement

Pour cette partie, il nous était demandé de mélanger les données sensibles. Nous avons donc regroupé les trois attributs de données sensibles en une seule colonne, puis grâce à une routine *"shuffle"* que nous avons créée, nous avons mélangé les données avant de les insérer dans la table finale. Avant de les exporter vers un fichier CSV, nous avons extrait les attributs de la colonne créée afin que le schéma de la table reste le même que celui qu'elle avait initialement.

À la fin de ce job, nous exportons la table dans un fichier nommé *"maladie_out.csv"* situé dans un dossier nommé *"out"*.

5 Résultats

Les résultats d'une anonymisation de base de données se calculent de plusieurs, ici nous en verrons deux, la **K-Anonymité** et la **L-Diversité**.

5.1 K-Anonymité

La K-Anonymité est un concept de protection de la vie privée des données personnelles. Il stipule qu'une donnée ne doit pas être associée à moins de **K** individus dans un ensemble de données, rendant ainsi difficile l'identification d'une personne spécifique à partir de cette donnée, même en combinant différentes sources d'informations.

La K-Anonymité a été calculée dans un job à part récupérant les données de la base de données anonymisées, la procédure de calcul se fait en plusieurs étapes :

- **tMap** : Dans un **tMap** on concatène les colonnes de sorte à avoir une nouvelle base de données de sortie ayant deux colonnes *(identifiante et sensible.*
- **tAggregateRow** : On effectue ensuite un **group_by** sur la colonne *Identifiante* avec l'opération *Nombre* pour avoir la valeur du **K**.
- **tLogRow** : On affiche notre résultat

La valeur maximale de **K** que nous avons atteint est 27 avec l'anonymisation au **niveau 3**.

5.2 L-Diversité

La L-Diversité est un concept de protection de la vie privée des données personnelles qui vise à garantir que chaque groupe homogène de données sensibles contient au moins **L** différents types de sensibilités. En d'autres termes, cela empêche la divulgation involontaire d'informations en assurant qu'une donnée sensible est diversifiée au sein d'un groupe afin de rendre l'identification des individus plus difficile

Le calcul de la L-Diversité se fait dans le même job que la K-Anonymité, on doit effectuer un *group_ by* sur les données *Identifiante* avec l'opération *Nombre* et également sur les données *Sensible* mais cette fois-ci avec l'opération *Compte (distincts)*.

La valeur max de **L** obtenue est 15 avec le **niveau 3** d'anonymisation

6 Conclusion

Ce projet nous a permis de mettre en pratique nos connaissances en intégration de données et en anonymisation. La partie la plus difficile a été de prendre les bonnes décisions d'anonymisation pour chaque niveau, étant donné qu'il existe une multitude de configurations possibles. Nous devions également trouver le juste équilibre entre l'anonymisation et la préservation de l'information. Ce projet nous a également offert l'opportunité d'utiliser nos compétences récentes dans les fonctionnalités plus avancées du logiciel Talend, telles que les routines et les contextes.

Nous avons décidé de limiter le nombre de niveaux d'anonymisation à trois, mais au fur et à mesure que le projet avançait, de nouvelles idées émergeaient. Avec un peu plus de temps et d'organisation, nous aurions pu ajouter plusieurs niveaux pour éviter une disproportion apparente entre les niveaux existants.

Travailler sur ce projet en équipe a été enrichissant, tant au niveau des idées d'anonymisation que grâce à nos différents niveaux de compétences techniques dans les différentes fonctionnalités.

7 Références

- Documentation, *R 4-08-VCOD : Préparation/Intégration de Données*

- Aide du professeur, *Clément JACQ*