Détection des Données à Caractère Personnel dans les Bases Multidimensionnelles

Amine Mrabet, Ali Hassan, Patrice Darmon

Research & Innovation - Umanis
7, Rue Paul Vaillant Couturier, 92300 Levallois-Perret, France
{amrabet, ahassan, pdarmon} @umanis.com

Résumé. Cartographier les données à caractère personnel dans la démarche de la de-identification est toujours un vrai pré-requis. Aujourd'hui avec les bases de données de grandes masses, nous sommes dans l'obligation d'automatiser l'étape de la détection dans cette démarche. Ce qui permet d'éviter les étapes chronophages, d'augmenter la précision de la détection et essentiellement permet de garantir la confidentialité. Pour toutes ces raisons nous avons proposé une nouvelle approche pour détecter les données personnelles. L'approche détaillée dans ce travail est adaptée aux bases de données multidimensionnelles. Nos méthodes utilisées dans cette approche sont sur deux niveaux. Nous proposons deux solutions de détection au niveau des données, et une solution au niveau des méta-données. Après détection des données à caractère personnel dans une base en utilisant les scores d'identifications, nous utilisons les scores de sensibilité afin d'évaluer la sensibilité totale de la base multidimensionnelle avant et après anonymisation.

1 Introduction

La protection de la vie privée est un droit humain fondamental. Ceci est reconnu par l'article 8 de la convention de sauvegarde des droits de l'homme et des libertés fondamentales (Conseil de l'Europe, 1948) qui assure le droit de chacun au respect de *"sa vie privée et familiale, de son domicile et de sa correspondance"*. De même, la Charte des droits fondamentaux de l'union européenne (European Commission for Justice, 2009) définit le *"respect de la vie privée et familiale"* (article 7) et ajoute un article spécifique sur la *"protection des données personnelles"* (article 8).

L'expression "données personnelles" a été définie dans la loi n 78-17 du 6 janvier 1978 relative à l'informatique, aux fichiers et aux libertés. La définition de ces données était : *"toute information relative à une personne physique identifiée ou qui peut être identifiée, directement ou indirectement, par référence à un numéro d'identification ou à un ou plusieurs éléments qui lui sont propres."*. La Commission nationale de l'informatique et des libertés (CNIL) reprend la même définition et donne les exemples suivants d'une donnée à caractère personnel : *"un nom, une photo, une empreinte, une adresse postale, une adresse mail, un numéro de téléphone, un numéro de sécurité sociale, un matricule interne, une adresse IP, un identifiant de connexion informatique, un enregistrement vocal"*.

La politique du règlement général sur la protection des données (RGPD) (PARLEMENT EUROPEEN, 2016) nous informe sur les catégories et les sensibilités de données à caractère personnel que les entreprises traitent, la façon d'utiliser ces données, les destinataires auxquels ils les communiquent ainsi que les droits dont les entreprises et la personne concernée disposent. Le RGPD exige l'adoption de mesures qui respectent les principes de protection des données dès la conception et de protection des données par défaut.

Trois objectifs spécifiques pour la protection de la vie privée (Hansen, 2012) : l'intraçabilité, la transparence et la possibilité d'intervention :

— **L'intraçabilité** garantit que les données pertinentes à la vie privée ne peuvent pas être liées entre les domaines. L'intraçabilité est liée au principe de la minimisation maximum des données.
— **La transparence** garantit que tous les processus de traitement de données de la vie privée, y compris le cadre juridique, technique et organisationnel, peuvent être compris et reconstruits à tout moment. L'information doit être disponible avant (traitement planifié), pendant (traitement actuel) et après le traitement (pour savoir ce qui s'est passé exactement). La quantité d'informations à fournir et la manière dont elles doivent être communiquées doivent être adaptées en fonction des capacités du public cible.
— **La possibilité d'intervention** permet d'intervenir sur tout traitement de données en cours ou prévu concernant les données personnelles, en particulier par les personnes dont les données sont traitées.

L'utilisation de BI en libre-service "self-service BI", proposée par les différentes solutions "Tableau" [1], "Power BI" [2], "Qlik Sense" [3], donne le pouvoir à une population non-informaticien de créer son propre reporting, lancer ses propres requêtes en toute autonomie et publier ses rapports sans passer par la case DSI. Cela pourrait augmenter le risque de construire et partager des rapports qui contiennent des données personnelles. La détermination du niveau de sensibilité des informations dans les rapports et les bases de données est sous la responsabilité des utilisateurs eux-mêmes. Cette tâche est effectuée manuellement et elle dépend des compétences de l'utilisateur.

Notre objectif dans ce papier donc est de proposer une approche qui permet de détecter automatiquement les données à caractère personnel (DCP) et de calculer le niveau de sensibilité de ces données dans les bases de données multidimensionnelles (BDM). Notre solution permet d'évaluer la sensibilité des informations publiées avant et après anonymisation. Nous proposons une nouvelle approche qui se base sur trois méthodes : deux méthodes détectent les DCP en analysant les valeurs de données en s'appuyant sur les expressions régulières et les bases de référence et une méthode qui détecte les DCP en s'appuyant sur la métadonnée (par exemple, le nom de l'attribut et celui de la dimension à laquelle il appartient). Cette dernière méthode se base sur une architecture d'ontologie.

L'article est structuré de la façon suivante. La Section 2 présente l'état de l'art de la protection et la détection de DCP. La Section 3 décrit le modèle multidimensionnel et elle présente le cas d'étude. La Section 4 est consacrée à notre méthode de détection de DCP proposée. Nous concluons dans la Section 5.

1. https://www.tableau.com
2. https://powerbi.microsoft.com
3. https://www.qlik.com

2 Positionnement et État de l'art

Dans cette section, nous présentons de manière générale les stratégies et les méthodes de la protection des données personnelles et nous détaillons ensuite les méthodes utilisées pour détecter les données personnelles dans les bases de données.

2.1 Protection de Données Personnelles

Afin d'atteindre les objectifs précédents de la protection de la vie privée, le développement de méthodologies appropriées pour la protection de la vie privée dès la conception a été discuté dans plusieurs travaux de recherche (Tschantz et Wing, 2009; Gürses et al., 2011; Hoepman, 2014; Kerschbaum, 2014; Antignac et Le Métayer, 2014). Un état de l'art détaillé se trouve dans (George et al., 2014; Lazaro et Le Métayer, 2015; Sobati Moghadam et al., 2017). (Hoepman, 2014) résume les différentes stratégies pour réaliser la protection de la vie privée dès la conception. Il les classe dans deux catégories : les stratégies orientées données et les stratégies orientées processus.

2.1.1 Les stratégies orientées données

Ces stratégies supportent l'intraçabilité. Les stratégies suivantes sont orientées données :

1. **Minimiser :** la quantité de données personnelles traitées devrait être limitée au minimum possible. Ce qui signifie qu'aucune donnée inutile n'est collectée.
2. **Cacher :** les données personnelles, et leurs interrelations, doivent être cachées. Plusieurs techniques peuvent être utilisées pour réaliser cette stratégie : le chiffrement des données, l'anonymisation et la pseudonymisation.
3. **Séparer :** cette stratégie indique que les données personnelles doivent être stockées dans différentes sources (bases) de données et traitées de manière distribuée.
4. **Agréger :** les DCP doivent être traitées à un niveau d'agrégation le plus haut possible (ayant le moins de détails) dans lequel elles sont encore utiles.

2.1.2 Les stratégies orientées processus

Ces stratégies supportent la transparence et la possibilité d'intervention. Les stratégies suivantes sont orientées processus :

1. **Informer :** les personnes concernées doivent être informées lorsque leurs données personnelles sont traitées. Elles devraient être informées de l'information qui a été traitée, dans quel but et par quel moyen.
2. **Contrôler :** les personnes concernées devraient avoir le pouvoir de voir, de mettre à jour et même de demander la suppression des données personnelles collectées.
3. **Renforcer :** cette stratégie garantit la mise en place d'une politique de la protection de la vie privée. Cette politique devrait être compatible avec les exigences légales. Un état de l'art montre les méthodes utilisées pour mettre en œuvre cette stratégie concernant les bases de données dans le cloud (Sobati Moghadam et al., 2017).
4. **Démontrer :** cette stratégie exige qu'un contrôleur de données puisse démontrer la conformité à la politique de la protection de la vie privée et aux exigences légales.

2.2 Détection des données personnelles

Le problème de la protection des données personnelles dans l'OLAP est largement traité dans l'état de l'art dans des travaux qui ont une stratégie orientée donnée (Lingyu Wang et al., 2004; Agrawal et al., 2005; Hua et al., 2005; Sung et al., 2006; Cuzzocrea et Saccà, 2010; Fessant et al., 2017; Cuzzocrea et al., 2018). Par contre, la question de la détection de ces données dans les bases de données est très peu discutée dans l'état de l'art.

Kamakshi et Babu (2012) proposent une méthode pour détecter le niveau de la sensibilité de données demandées dans une requête. Si la requête demande de données sensibles, un "swapping" est appliqué sur les données avant d'envoyer la réponse. Sinon les données sont envoyées sans modification. En revanche, cette proposition se base sur le fait que la sensibilité de données est déjà identifiée. Un poids (niveau de sensibilité) est déjà associé à chaque attribut. La valeur du poids est définie en fonction du facteur déterminant le rôle que joue l'attribut dans la révélation de l'identité d'un individu.

La proposition de (d. Mouza et al., 2010) n'a pas cette limite. Les auteurs proposent d'utiliser des règles génériques qui peuvent être spécifiques à chaque domaine, table, attribut et valeur d'attribut. Selon ces règles, un score de sensibilité est calculé pour chaque attribut. Cette proposition utilise les liens sémantiques entre les attributs et les techniques du traitement automatique du langage naturel pour trouver ces liens. Autrement dit, elle suppose que les noms des attributs sont des mots. Donc, l'utilisation des abréviations ou des noms qui ne sont pas pertinents pourrait perturber cette méthode.

L'outil commercial "DgSECURE DETECT" [4] propose une solution pour détecter les données personnelles. Il analyse les données elles-mêmes (les valeurs). Il effectue une inspection approfondie du contenu de données à l'aide de techniques intégrant des correspondances basées sur des dictionnaires et de mots-clés pondérés, et l'apprentissage automatique. Par contre, la méthode utilisée est trop générique et elle ne prend pas en compte la spécification de données de chaque domaine.

Par ailleurs, toutes ces méthodes ont été proposées pour détecter les DCP dans les bases de données relationnelles classiques. À notre connaissance, il n'y a aucun travail qui prend en compte la particularité des bases de données multidimensionnelles, par exemple, l'organisation de données en plusieurs niveaux de granularité.

3 Préliminaires : modèle multidimensionnel conceptuel

Dans cette section, nous présentons la définition du modèle multidimensionnel (Hassan et al., 2013, 2015).

Soient $\mathcal{N}$= $\{n_1, n_2, ...\}$ un ensemble fini de noms.

Définition 1. Un *fait* F_i est défini par (n^{Fi}, M^i) :

— $n^{Fi} \in \mathcal{N}$est le nom du fait,

— $M^i = \{m_1, ..., m_{pi}\}$ est un ensemble de *mesures*.

Définition 2. Une *dimension* D_i est définie par (n^{Di}, A^i, H^i) :

— $n^{Di} \in \mathcal{N}$est le nom de la dimension,

— $A^i = \{a_1^i, ..., a_{ri}^i\} \bigcup \{Id^i, All^i\}$ est l'ensemble de *attributs de la dimension*,

— $H^i = \{H_1^i, ..., H_{si}^i\}$ est un ensemble de *hiérarchies de la dimension*.

4. https://www.dataguise.com/detect/

Les attributs d'une dimension sont organisés en hiérarchies allant de la granularité la plus détaillée (Id^i) à la plus générale (All^i).

Définition 3. Une *hiérarchie* H_j est définie par (n^{Hj}, P^j, $\prec^{Hj}$) :

- $n^{Hj} \in \mathcal{N}$est le nom de la hiérarchie,
- $P^j = \{p_1^j, ..., p_{q_j}^j\}$ est un ensemble d'*attributs de la dimension* ($P^j \subseteq A^i$) appelés *paramètres*,
- $\prec^{Hj} = \{(p_x^j, p_y^j) \mid p_x^j \in P^j \wedge p_y^j \in P^j\}$ est une relation binaire antisymétrique et transitive définissant un chemin de navigation sur la dimension,
- $\text{Weak}^{Hj} : P^j \rightarrow 2^{A^i \setminus P^j}$ est une application qui associe à chaque paramètre un ensemble d'attributs de dimension, appelés *attributs faibles*.

Définition 4. A *schéma multidimensionnel* S est définie par (F, D, Star) :

- F = $\{F_1, ..., F_n\}$ est l'ensemble des faits,
- D = $\{D_1, ..., D_m\}$ est l'ensemble des dimensions,
- Star : F $\rightarrow 2^D$ est une fonction qui associe chaque fait à ses axes d'analyse (dimensions).

On pose : $M = \bigcup_{i=1}^{n} M^i$, $A = \bigcup_{i=1}^{m} A^i$, $P^i = \bigcup_{j=1}^{s_i} P^j$, $P = \bigcup_{i=1}^{m} P^i$, $W^i = \bigcup_{j=1}^{s_i} \bigcup_{k=1}^{q_j} \text{Weak}^{Hj}(p_k^j)$ and $W = \bigcup_{i=1}^{m} W^i$.

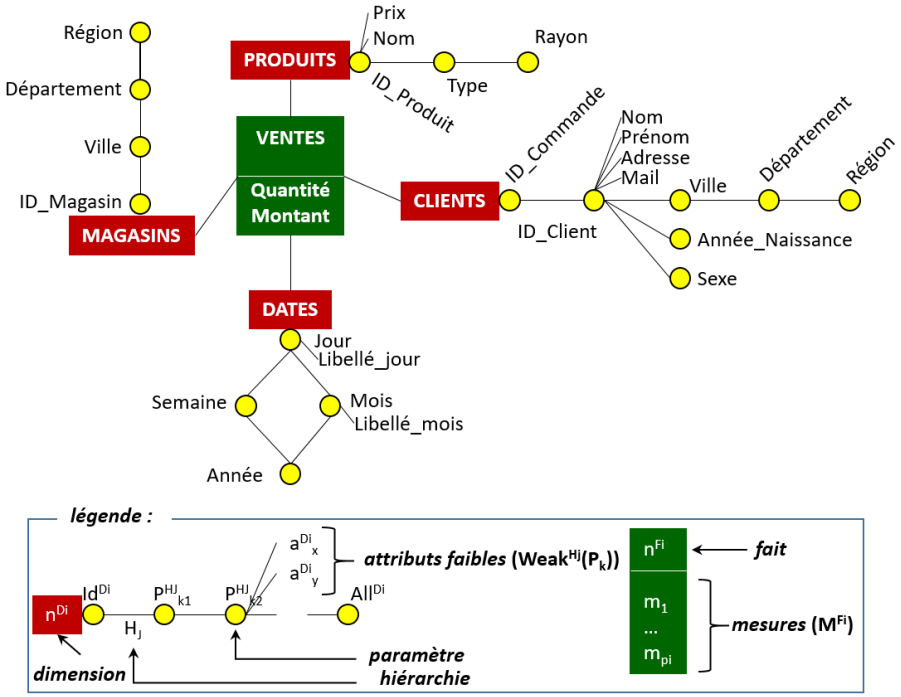

FIG. 1 – *Exemple de schéma conceptuel en étoile*

Exemple (cas d'étude) : La Figure 1 montre un exemple d'un schéma conceptuel multidimensionnel en étoile. Ce schéma permet d'analyser la quantité et le montant (mesures) de

ventes (fait) selon quatre dimensions : "Magasins", "Clients", "Dates" et "Produits". La dimension "Dates" organise les granularités temporelles en deux hiérarchies : une pour les semaines et l'autre pour les mois. Les dimensions "Produits" et "Magasins" comprennent une seule hiérarchie. La dimension "Clients" organise les informations de commandes des clients en trois hiérarchies : (1) selon la distribution géographique de clients (2) selon leur année de naissance (3) selon leur sexe. Sur la dimension "Clients", plusieurs attributs faibles ("Nom", "Prénom", "Adresse" et "Mail") sont associés au paramètre "ID_Client". Ils peuvent être considérés comme des données à caractère personnel qu'il faut protéger.

Afin de bien répondre au besoin d'anonymisation et de protection des données personnelles, nous proposons dans la suite de cet article une méthode qui permet d'automatiser la détection de données personnelles dans les bases multidimensionnelles et qui prend en compte la particularité de la structure de ces bases.

4 Méthode de détection de DCP dans les bases de données multidimensionnelles

Dans notre solution, nous proposons deux étapes dont la première est pour entraîner notre ontologie. La deuxième étape est la détection des DCP et de calculer le niveau de sensibilité d'une base de données multidimensionnelles utilisant des scores.

4.1 Entraînement d'ontologie

Nous commençons cette section par décrire la structure de l'ontologie que nous avons proposé. Par la suite, nous montrons les méthodes utilisées pour alimenter cette ontologie. Nous avons choisi d'utiliser une ontologie (base de connaissance sémantique) pour optimiser la détection des données personnelles. La détection via une ontologie est appliquée au niveau des méta-données. L'avantage de travailler à ce niveau de détection est d'augmenter la confidentialité et la performance. Cette solution est sécurisée car nous évitons la détection au niveau des données. Cette solution fonctionne bien car la détection au niveau des méta-données est plus rapide que la détection au niveau des données.

4.1.1 description d'ontologie

Dans l'état de l'art, l'ontologie est souvent associée au schéma multidimensionnel pour mieux comprendre ce dernier (Guizzardi, 2019). Par contre, nous utilisons une ontologie comme base de connaissance sémantique pour décrire le lien entre les attributs des schémas multidimensionnels et les données à caractère personnel.

Notre ontologie automatique est composée de deux parties. Une partie statique composée par une liste des entités des DCP prédéfinie dans notre méthode. Et une partie dynamique qui sera entraînée par la détection des DCP via deux méthodes. Une méthode utilise des standards de référence et l'autre utilise un standard des expressions régulières. Dans la partie statique, pour chaque domaine nous avons classifié ces entités de DCP sur plusieurs catégories proposées par le CNIL et nous avons affecté un niveau (un score) de sensibilité à chacune de ces entités (partie droite de la Figure 2). Ces scores sont prédéfinis et paramétrables selon le contexte de la base de données.

Afin d'adapter notre méthode de détection des données à caractère personnel aux bases de données multidimensionnelles, nous proposons une ontologie (cf. Figure 2) à quatre niveaux :

1. Niveau "Domaine" : ce niveau représente le secteur ou le domaine de données. Dans la partie gauche de la Figure 2, le nom du domaine correspond au nom du fait du schéma de Figure 1.
2. Niveau "Dimension" : chaque balise de ce niveau regroupe tous les attributs (paramètres et attributs faibles) qui appartiennent à une dimension du schéma multidimensionnel.
3. Niveau "Paramètre" : chaque balise de ce niveau regroupe tous les attributs faibles associés au paramètre concerné.
4. Niveau "Attribut faible".

Si on prend l'exemple de la Figure 1, nous trouvons que grâce à cette structure à quatre niveaux, notre ontologie est capable de différencier entre l'attribut "Nom" d'un client sur la dimension "Clients" et l'attribut "Nom" d'un produit sur la dimension "Produits".

Dans la partie dynamique, nous détectons à quelle entité de DCP correspond chaque attribut de la BDM. Par exemple l'attribut "nom" dans la dimension "Clients" correspond à l'entité "nom_de_personne". Le nom de l'entité de DCP est associé à la balise du paramètre ou de l'attribut faible (cf. ligne 11 de la partie gauche de la Figure 2). Les noms des attributs de la BDM qui correspondent à cette entité sont ajoutés dans une balise "Champs" (cf. lignes 12 à 15 de la partie gauche de la Figure 2). Après chaque détection, nous mettons à jour cette partie dynamique de l'ontologie.

```
2  <Domaine name="VENTES">
3    <Dimension name="Measure">
6    <Dimension name="CLIENT">
7      <Paramètre name="CLIENT_N">
8        <Champs>
9          <Champ name="ID_Client"/>
10       </Champs>
11       <WeakAttribut name="Nom_de_personne">
12         <Champs>
13           <Champ name="NOM_DE_FAMILLe"/>
14           <Champ name="NOM"/>
15         </Champs>
16       </WeakAttribut>
17       <WeakAttribut name="Prénom_de_personne">
18         <Champs>
19           <Champ name="PRENOM"/>
20         </Champs>
21       </WeakAttribut>
22     </Paramètre>
23   </Dimension>
```

```
1  <Domaine name="VENTES">
2    <DCP name="CLIENT_N">
3      <Catégorie>Identité</Catégorie>
4      <Score>20%</Score>
5    </DCP>
6    <DCP name="Nom_de_personne">
7      <Catégorie>Identité</Catégorie>
8      <Score>10%</Score>
9    </DCP>
10   <DCP name="Prénom_de_personne">
11     <Catégorie>Identité</Catégorie>
12     <Score>10%</Score>
13   </DCP>
14 </Domaine>
```

FIG. 2 – *Extrait de l'ontologie*

4.1.2 Extraction du schéma multidimensionnel (Méta-données)

Afin d'alimenter notre ontologie nous suivons un processus d'entraînement en se basant sur des bases d'entraînement. La figure 3 présente l'étape de détection de la méta-donnée dans une base de données multidimensionnelles. Après détection de la structure nous enchaînons les étapes de nettoyage, de transformation et de stockage afin d'avoir une base de structure correcte et prête pour l'entraînement. Ces opérations concernent les noms des attributs de la

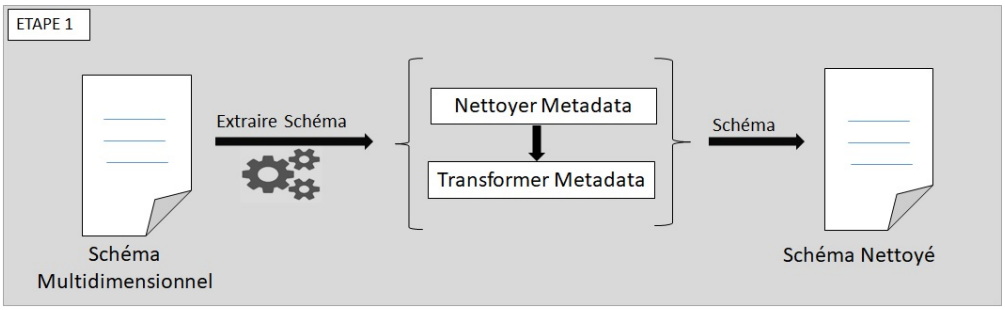

FIG. 3 – *Extraire Métadonnées*

BDM afin d'éviter les problèmes qui proviennent de la différence entre majuscule/minuscule et singulier/pluriel et des caractères spécifiques.

Suite à l'extraction de la méta-donnée nous proposons deux méthodes pour mettre à jour notre ontologie. Les deux méthodes qu'on utilise servent à calculer des scores d'identification pour chaque attribut. La première méthode se base sur un standard des expressions régulières prédéfini. La deuxième méthode repose sur des standards de référence (données ouvertes).

4.1.3 Mise à jour de l'ontologie

Comme détaillé ci-dessus, l'architecture proposée pour l'ontologie dans ce travail nécessite deux méthodes pour l'entraîner. Pour exécuter notre approche nous avons besoin de travailler au niveau données. En utilisant les deux méthodes présentées ci-dessus.

La figure 4 présente notre première méthode utilisée pour alimenter notre ontologie. Dans cette méthode, nous utilisons une base des expressions régulières. Dans cette méthode nous injectons des expressions régulières dans nos requêtes de scoring. Chaque requête retourne un score de correspondance de l'attribut en question avec le caractère personnel cherché. Le score calculé est servi pour mettre à jour la partie dynamique de l'ontologie.

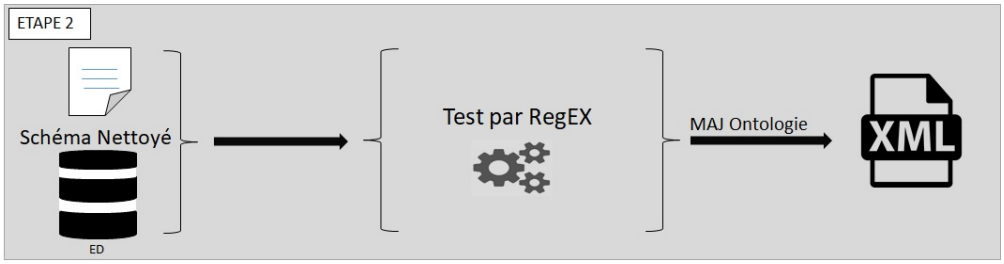

FIG. 4 – *Mise à jour via des RegEx*

La figure 5 présente notre deuxième méthode qu'on utilise pour alimenter notre ontologie. Dans cette méthode, nous avons préparé plusieurs bases de référence afin d'effectuer des comparaisons pour la détection. Afin de garantir la confidentialité des données, nous travaillons avec des données hachées de nos standards de référence. Et pour cette raison, nous utilisons la

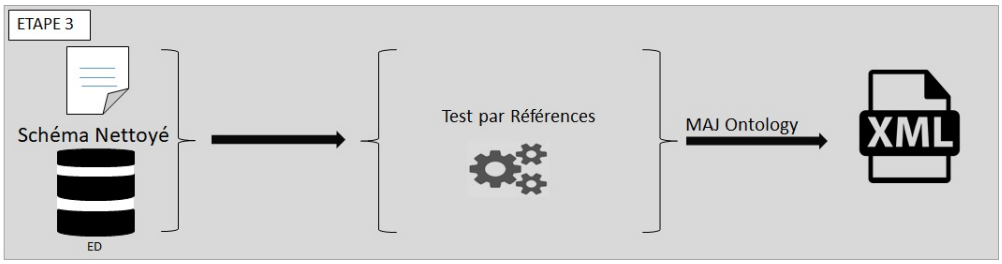

FIG. 5 – *Mise à jour via des bases de référence*

même fonction de hachage utilisée pour hacher les bases de référence et pour hacher les données à détecter. Cette méthode retourne aussi un score de correspondance entre l'attribut en question avec le caractère personnel cherché. Et de la même manière, nous utilisons ce score afin de mettre à jour la partie dynamique de l'ontologie.

4.2 Détection des données personnelles

Dans cette section, nous détaillons nos méthodes de détection des données personnelles. L'algorithme 1 détaille notre procédure de détection. Cet algorithme prend six entrées : le schéma multidimensionnel (S) représentant le niveau méta, l'entrepôt de données ($Base\_ED$) qui est la base de données (les tables avec les données), les parties statique et dynamique de notre ontologie ($Ontologie\_S$, $Ontologie\_D$), nos bases de référence et d'expressions régulières ($Base\_ref$, $Base\_RegEx$). Il retourne la liste des entités de DCP ($DCPs$).

Les différentes étapes que nous proposons dans cet algorithme sont :

1. Détection de la structure d'un schéma multidimensionnel (ligne 1). Pour effectuer cette étape nous utilisons la méthode "Extraire_Meta".
2. Détecter les données personnelles (pour tous les attributs et mesures dans le schéma multidimensionnel) via un matching avec notre ontologie entraînée (lignes 2 à 5). Dans cette étape nous utilisons la méthode "Matching_Ontologie" pour chercher les attributs sensibles dans la partie dynamique "Ontologie_D," de notre ontologie. Par la suite nous récupérons les informations dans l'ontologie statique "Ontologie_S," via la fonction "Chercher_DCP".
3. Utiliser notre base des expressions régulières afin de lancer aussi d'autre vérification des attributs non détectés dans l'étape précédente (lignes 6 et 7). Pour cette étape nous utilisons la méthode "Matching_S_RegEx".
4. Utiliser nos bases de référence afin de lancer des vérifications pour chaque attribut non détecté dans l'étape précédente (lignes 8 et 9). Pour cette étape nous utilisons la méthode "Matching_S_Ref".

Les étapes 3 et 4 de notre algorithme pourraient être inversées ou exécutées en parallèle afin d'optimiser les performances en temps d'exécution (des analyses de performance sont encours d'implémentation dans le cadre de ce projet.

Algorithm 1: Détection des données à caractère personnel

Input: S, $Base\_ED$, $Ontologie\_S$, $Ontologie\_D$, $Base\_ref$, $Base\_RegEx$
Output: $DCPs$

```
Schema ←Extraire_Meta(S);
foreach x ∈ Schema.A ∪ Schema.M do
    Result ←Matching_Ontologie(x, Ontologie_D);
    if (Result ≠ Null) then
        DCP ←Chercher_DCP (Result, Ontologie_S);
    else
        DCP ←Matching_S_RegEx(x, Base_RegEx);
        if DCP = Null then
            DCP ←Matching_S_Ref(x, Base_ref);
    DCPs.add(DCP);
return DCPs;
```

La figure 6 résume les méthodes de détection des DCP dans un modèle OLAP détaillé dans l'algorithme 1. Pour détecter les données, nous commençons avec la méthode ontologie, puis la méthode des expressions régulières et nous terminons avec la méthode de base des références.

FIG. 6 – *La procédure de détection des données à caractère personnel dans l'OLAP*

4.3 Score de sensibilité

Pour confirmer la protection des données à caractère personnel, il faut valider avec un score d'évaluation de sensibilité pour garantir le niveau de sécurité nécessaire. Pour cette raison, nous proposons un calcul de taux de sensibilité. Ce dernier est calculé en fonction des niveaux de sensibilité prédéfinis par caractère personnel dans la partie statique de notre ontologie. Pour définir ces niveaux nous nous sommes basés sur des recommandations de la CNIL.

Le score total de sensibilité dans une base multidimensionnelle est la somme des niveaux de sensibilité des mesures et des attributs (paramètre et attribut faible) du schéma. Certains attributs peuvent être calculés à partir d'autre attributs dans la même hiérarchie. Autrement

dit, un attribut peut être inclus dans un autre. Par exemple, le jour '08/03/1981' inclue le mois '03/1981' et l'année '1981'. Dans ce cas on dit que 'Jour' $\triangleright$ 'Mois' $\triangleright$ 'Année'.

$$x \triangleright y \Leftrightarrow x \prec^{Hj} y \wedge x.value \Rightarrow y.value$$

Dans ce genre de cas, les attributs inclus dans un autre ne seront pas comptés dans le calcul du score de sensibilité.

L'algorithme 2 détaille le calcul du score de sensibilité. Il prend trois entrées : la liste des DCP ($DCPs$), l'ontologie statique ($Ontologie\_S$) et le seuil de sensibilité ($Seuil$) à respecter. Ce dernier est paramétrable selon le cas d'utilisation. Cet algorithme calcule la somme de sensibilité en respectant la contrainte d'inclusion des attributs, et il alerte l'utilisateur que son système enfreint les règles RGPD.

Algorithm 2: Calcul de sensibilité dans ED

Input: $DCPs, Ontologie\_S, Seuil$
Output: $Sensibilte\_ED$

$Sensibilite\_ED \leftarrow Null$;
foreach $x \in DCPs \mid \nexists y \in DCPs \mid y \triangleright x$ **do**
    $Sensibilite\_ED \leftarrow Sensibilite\_ED + Ontologie\_S.findScore(x)$);
if $(Sensibilite\_ED > Seuil)$ **then**
    Alert();
return $Sensibilite\_ED$;

5 Conclusion

Nous avons développé dans ce papier une approche de détection automatique des DCP. Notre solution se repose sur une ontologie dynamique et entraînée. Afin d'entraîner cette ontologie et détecter les DCP, nous utilisons deux méthodes de détection en se basant sur des standards de référence et des expressions régulières. L'ontologie est mise à jour à chaque utilisation, ce qui nous permet d'avoir une ontologie avancée. Une telle ontologie garantit une haute précision de détection. Nous proposons également un scoring qui prend en compte la particularité des schémas multidimensionnels (l'organisation des attributs en plusieurs niveaux de granularité).

Dans la perspective de ce travail, nous avons commencé à développer une autre méthode basée sur des algorithmes d'intelligence artificielle. Cette méthode est la reconnaissance d'entités nommées. Nous envisageons de proposer un module de validation sémantique avant de mettre à jour l'ontologie.

Références

Agrawal, R., R. Srikant, et D. Thomas (2005). Privacy preserving OLAP. pp. 251.

Antignac, T. et D. Le Métayer (2014). Privacy by Design : From Technologies to Architectures (Position Paper). In B. Preneel et D. Ikonomou (Eds.), *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, Volume 8450 LNCS, pp. 1–17. Cham : Springer International Publishing.

Conseil de l'Europe (1948). Convention de sauvegarde des Droits de l'Homme et des Libertés fondamentales. *1974*, 1–22.

Cuzzocrea, A. et D. Saccà (2010). Balancing accuracy and privacy of OLAP aggregations on data cubes. pp. 93.

Cuzzocrea, A., A. Schuster, et G. Vercelli (2018). PP-OMDS : An Effective and Efficient Framework for Supporting Privacy-Preserving OLAP-based Monitoring of Data Streams. *1*(Iceis), 282–292.

d. Mouza, C., E. Métais, N. Lammari, J. Akoka, T. Aubonnet, I. Comyn-Wattiau, H. Fadili, et S. S. Cherfi (2010). Towards an Automatic Detection of Sensitive Information in a Database. In *2010 Second International Conference on Advances in Databases, Knowledge, and Data Applications*, pp. 247–252.

European Commission for Justice (2009). EU Charter of Fundamental Rights.

Fessant, F., T. Benkhelif, et F. Clérot (2017). Anonymiser des données multidimensionnelles à l'aide du coclustering. *Revue des Nouvelles Technologies de l'Information Extraction*, 153–164.

George, D., D.-F. Josep, H. Marit, J.-H. Hoepman, D. L. Métayer, T. Rodica, et S. Stefan (2014). Privacy and Data Protection by Design– from policy to engineering. *CoRR*.

Guizzardi, G. (2019). On the Application of Ontological Patterns for Conceptual Modeling in Multidimensional Models On the Application of Ontological Patterns for Conceptual Modeling in Multidimensional Models. In *ADBIS 2019 (à apparaître)*.

Gürses, S., C. Troncoso, et C. Diaz (2011). Engineering privacy by design.

Hansen, M. (2012). Top 10 mistakes in system design from a privacy perspective and privacy protection goals. In J. Camenisch, B. Crispo, S. Fischer-Hübner, R. Leenes, et G. Russello (Eds.), *IFIP Advances in Information and Communication Technology*, Volume 375 AICT, pp. 14–31. Berlin, Heidelberg : Springer Berlin Heidelberg.

Hassan, A., F. Ravat, O. Teste, R. Tournier, et G. Zurfluh (2013). OLAP in Multifunction Multidimensional Databases. In *Advances in Databases and Information Systems*, pp. 190–203.

Hassan, A., F. Ravat, O. Teste, R. Tournier, et G. Zurfluh (2015). Differentiated Multiple Aggregations in Multidimensional Databases. *TLDKS XXI 9260*, 20–47.

Hoepman, J.-H. (2014). Privacy Design Strategies. *ICT-System Security and Privacy Protection–29th IFIP TC 11*, 446–459.

Hua, M., S. Zhang, W. Wang, H. Zhou, et B. Shi (2005). FMC : An Approach for Privacy Preserving OLAP. pp. 408–417.

Kamakshi, P. et A. V. Babu (2012). Automatic detection of sensitive attribute in PPDM. In *2012 IEEE International Conference on Computational Intelligence and Computing Research*, pp. 1–5.

Kerschbaum, F. (2014). Privacy-Preserving Computation. In B. Preneel et D. Ikonomou (Eds.), *Privacy Technologies and Policy : First Annual Privacy Forum, APF 2012, Limassol, Cyprus, October 10-11, 2012, Revised Selected Papers*, pp. 41–54. Berlin, Heidelberg : Springer Berlin Heidelberg.

Lazaro, C. et D. Le Métayer (2015). The Control over personal data : True remedy or fairytale ? *SCRIPTed 12*(1).

Lingyu Wang, S. Jajodia, et D. Wijesekera (2004). Securing OLAP data cubes against privacy breaches. pp. 161–175.

PARLEMENT EUROPEEN (2016). Règlement (UE) 2016/679 du parlement européen et du conseil du 27 avril 2016. *Journal officiel de l'Union européenne L 119/1*(3).

Sobati Moghadam, S., J. Darmont, et G. Gavin (2017). Enforcing Privacy in Cloud Databases. In *DaWaK 2017*, Volume 10440, pp. 53–73. Springer.

Sung, S. Y., Y. Liu, H. Xiong, et P. A. Ng (2006). Privacy preservation for data cubes. *Knowledge and Information Systems 9*(1), 38–61.

Tschantz, M. C. et J. M. Wing (2009). Formal methods for privacy. In *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, Volume 5850 LNCS of *FM '09*, Berlin, Heidelberg, pp. 1–15. Springer-Verlag.

Summary

Mapping personal data in the process for de-identification is always a prerequisite. Today with the big data, we are obliged to automate the detection step in this process. That avoids time-consuming, increases the accuracy of detection and allow to ensure confidentiality. For all these reasons we have proposed a new approach to exploit the data. Our approach in this work is adapted to multidimensional databases. Our techniques used in this approach are based on two levels. We propose two detection solutions at the data level, and a solution at the metadata level. After detecting personal data in a database using the identification scores, we use the sensitivity scores to assess the total sensitivity of the multidimensional database before and after anonymization.