Implémentation des Plugins Logstash de Généralisation et de Chiffrement pour l'Anonymisation et la Pseudonymisation

Ali Hassan, Amine Mrabet, Patrice Darmon

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

Résumé. Ce papier présente une nouvelle implémentation des méthodes de protection des données à caractère personnel basées sur des algorithmes de chiffrement et de généralisation spécifiques mis en oeuvre dans des plugins Logstash. Notre algorithme de chiffrement est adapté aux données personnelles en considérant les différentes catégories : identifiants, quasi-identifiants et attributs sensibles. En outre, notre solution d'anonymisation propose plusieurs méthodes de généralisation, paramétrables selon les types des données. Afin de valider les résultats de ces méthodes, nous proposons également une étape de vérification des modèles de protection de la vie privée comme le k-anonymat et le l-diversité.

1 Introduction

Les besoins de collecte, de stockage et d'analyse des données sont en croissance constante notamment en raison de la révolution liée aux objets connectés. L'analyse de ces données est essentielle pour les entreprises avec différents enjeux : IA, statistique, publicité, etc... Cependant, le stockage et l'analyse de données personnelles posent des vrais problèmes de confidentialité. Les techniques de protection de la vie privée sont conçues et mises en oeuvre pour équilibrer les usages et la confidentialité des données à caractère personnel (DCP). La confidentialité et les principes de la protection des DCP doivent être garantis dans toutes les phases de collecte, stockage, traitement, analyse et partage des données.

Dans le cadre d'une démarche de protection des données, l'étape de découverte des données est indispensable. Une automatisation de cette étape est proposée dans (Mrabet et al., 2019). Les recommandations relatives à la pseudonymisation et à l'anonymisation visent à protéger l'individu et à renforcer la conformité au RGPD. Dans ce contexte, les techniques d'anonymisation sont une solution pour la protection, mais elles semblent inappropriées dans certaines circonstances. En outre, la pseudonymisation est utilisée à la fois pour réduire les risques de profilage et aider à respecter les obligations de protection des DCP. Donc, l'anonymisation et la pseudonymisation peuvent être utilisées de manière complémentaire ou séparée.

Motivation et cas d'étude

Dans le cadre d'un projet Big Data de smart territoire, une collectivité territoriale, qui collecte des données d'utilisation de wifi dans l'espace public, voudrait protéger le stockage (chez