M1 MIAGE - Sécurité des bases de données

Projet
Assainissement

Synthèse du projet

- Implémentation du mécanisme de perturbation de Laplace pour assainir des résultats de requêtes agrégats de type `COUNT` et `SUM`
- Extension à la publication d'histogrammes sur un attribut
- Analyse expérimentale de la qualité

Modalités

- 4 séances encadrées. Finalisations éventuelles à la maison.
- Rapport (entre 5p et 10p) : a minima l'architecture technique, les justifications de vos choix techniques, les analyses de sensibilité, l'analyse expérimentale de la qualité obtenue (avec graphes et commentaires)
- Code : commenté, avec readme pour l'installation et l'utilisation, ainsi qu'avec les bibliothèques nécessaires
- **Échéances : intermédiaire le dimanche 01/04/2018 23H59 et finale le dimanche 15/04/2018 23H59**

---

1 Objectif

Dans ce projet vous endossez le rôle d'une organisation qui collecte des données personnelles, et qui souhaite ouvrir sa BD à des utilisateurs authentifiés (*e.g.,* statisticiens). Vous avez heureusement été sensibilisé(e) aux techniques de publication de données respecteuses de la vie privée et avez décidé de mettre en place un mécanisme de publication interactive satisfaisant la differential privacy. Le mécanisme choisi est l'ajout d'une perturbation aléatoire qui suit la distribution de Laplace. L'objectif de ce projet est la conception de la preuve-de-concept sur des requêtes `COUNT` et `SUM`, qui permettra de valider l'approche avant la mise en production. Une requête `COUNT` suit la forme :

`SELECT COUNT(*) FROM TABLE WHERE` θ

Où θ est une condition logique arbitraire. Une requête suit le même format.

2 Choix techniques

Interface Vous êtes entièrement libres de choisir le langage de programmation que vous utiliserez pour implémenter l'interface utilisateur/données. Ce choix aura bien sûr un impact majeur sur les techniques que vous emploierez, soyez prudents ! Vous avez notamment a priori deux grandes orientations possibles : (1) utilisation de techniques internes au moteur du SGBD (vues, procédures stockées, déclencheurs), ou bien (2) utilisation d'un langage de programmation externe au SGBD qui s'interfacera avec le SGBD d'un côté (*e.g.,* JDBC) et avec l'utilisateur de l'autre (*e.g.,* une API)

SGBD Vous êtes libre de choisir votre SGBD parmi les deux suivants : PostgreSQL et Oracle. Votre choix dépendra notamment des techniques internes au SGBD dont vous aurez éventuellement besoin.

3 Fonctionnalités

Base : Perturbation de `COUNT` et de `SUM`
- Analyse de la sensibilité d'un `COUNT` et d'une `SUM`
- Paramétrage et gestion du budget de confidentialité ϵ
- Génération de perturbations suivant une distribution de Laplace bien paramétrée (*e.g.,* `https://en.wikipedia.org/wiki/Laplace_distribution#Generating_random_variables_according_to_the_Laplace_distribution`)

Mode DEBUG : Prévoyez un mode debug, vous permettant par exemple de désactiver la consommation de budget. Il vous sera utile pendant la validation expérimentale.

Bonus : Perturbation d'un `Histogramme` sur un attribut cible pour ensuite calculer les `COUNT` dessus
- Analyse de la sensibilité d'un histogramme
- Génération d'un histogramme sur une dimension (vous pouvez fixer le nombres de barres de l'histogramme et la largeur de chaque barre)
- Génération de perturbations pour `Histogramme` suivant une distribution de Laplace bien paramétrée

4 Validation expérimentale

Jeu de données Le jeu de données sur lequel vous lancerez votre validation est le `Adult Data Set` de l'UCI Machine Learning repository. Il est souvent utilisé pour valider les techniques d'assainissement de données.
- Téléchargez-le ici : `https://archive.ics.uci.edu/ml/datasets/Adult`.
- Stockez-le dans une table

Banc de tests Vous validerez expérimentalement cette approche en mesurant les variations de l'erreur moyenne. Vous effectuerez chaque mesure 100 fois (sans consommation de budget bien sûr) pour obtenir l'erreur absolue moyenne, sa variance, et sa médiane. Vous ferez varier les paramètres suivants :
- Valeur de ϵ (de $\epsilon = 0.01$ à $\epsilon = 1$)
- Taille du `COUNT` et de la `SUM`

Si vous vous attaquez à la génération d'histogramme, concentrez-vous sur la colonne `Age` et fixez à 10 le nombre de barres de l'histogramme (largeurs identiques)

Analyse Vous générerez les graphes de qualité en faisant varier le type de requêtes, leurs tailles, et la valeur de ϵ. Si vous faites le bonus, comparez l'approche par perturbation de `COUNT` à l'approche par perturbation d'histogramme. Vous conclurez en donnant votre opinion.