Projet de application. Synthèse et exploration

Organisation des données 2

octobre - novembre 2017

1 Objectifs du travail demandé

Dans ce travail vous allez travailler sur des données réelles acquises par des capteurs et soigneusement gardées dans une base de données dont on vous fournit une copie.

Le projet a une vocation multiple.

— D'une part c'est l'occasion de mettre en pratique directement les connaissances acquises en cours sur les BD relationnelles, sur les BD NoSQL et sur les entrepôts de données.

— Ensuite on demandera de trouver les réponses à des questions ouvertes qu'un observateur raisonnable pourrait se poser.

— En dernier lieu, on voudrait avoir des perspectives, pas forcément réalisables dans l'immédiat en termes techniques ou d'exploitation, de ce système.

Je reste à votre disposition pour toute information complémentaire et, naturellement, pour toute aide technique.

2 Description du fonctionnement des capteurs

Le but de ce projet est d'explorer le mieux possible des données issues des capteurs placés dans le bâtiment Espace Fauriel au 4-ème étage. Ces données réelles ont été collectées entre mars et septembre 2017. Les capteurs ont été placés dans divers emplacements : bureaux, couloirs, etc, des pièces ayant des fenêtres ou des pièces aveugles. Tous les capteurs sont identiques quant à leur mode de fonctionnement : ils émettent de l'information que dans le cas d'une variation de température, humidité et luminosité. Chaque capteur transmet le paramètre qui a varié estampillé avec la date d'émission et avec l'identifiant du capteur qui est unique.

Les données sont émises par les capteurs et stockées par la suite dans une base de données conçue à cette effet [1]. Ces données ont été importées depuis cette base et intégrées dans une base Oracle, la même que celle que nous avons utilisée en TP. L'accès aux tables est en lecture seule avec le droit de copier l'objet dans son espace ou exporter son contenu. Il y a uniquement trois tables :

— POSITION avec les identifiants des capteurs et une description de l'endroit où le capteur a été placé. Nous avons volontairement anonymisé cette donnée.

— CALIBRATION avec des données prises à certains moments qui ont servies à étudier l'écart entre le fonctionnement réel du capteur et la mesure réelle prise au même instant

1. C'est une base MySQL/MariaDB qui se trouve dans le réseau interne du département de recherche

— EVENTS avec toutes les données collectées après la mise en route des capteurs. Comme indiqué auparavant, chaque modification notable de température, humidité ou pression est capturée et ensuite transmise au serveur qui la transforme en enregistrement dans cette table. Il y a deux attributs de type DATE : le temps du capteur et le temps du système lors de l'insertion.

En ce moment le serveur ne fait que capter les données émises par les capteurs et écrire les enregistrements dans la table EVENTS. Au jour le jour on suit le fonctionnement du système grâce à des interrogations directes en SQL dans la base :

— le dernier enregistrement dans la table EVENTS

— l'état courant de tous les capteurs : pour chaque capteur on indique son identifiant, sa position et le nombre total d'enregistrements qui existent dans la table EVENTS le concernant.

3 Base Oracle mise à disposition

Dans la base Oracle d'instance `cdb1` qui est hébergée sur la machine IP `193.49.175.51` port `1521`, un utilisateur `c##capteurs` a été crée avec les mêmes tables que la base réelle attachée au serveur qui capte l'activité des capteurs. Ses tables ont aussi un synonyme public, donc on peut les accéder soit avec l'identifiant complet, comme `c##capteurs.events`, soit avec leur nom `cevents`. Nous avons mis une seule clé primaire sur la table EVENTS. Délibérément nous avons "oublié" les autres clés primaires, étrangères et même les index.

Vous pouvez exporter les tables, les consulter en lecture seulement ou crérr de vues basées sur ces tables. Ceux qui préfèrent monter eux-mêmes une autre base (Oracle ou un autre SGBDR de votre choix) vous avez à disposition les fichiers ayant servi à l'importation des données et le script de génération des tables.

Une attention particulière doit être portée au type `DATE` car la date de réception de l'information est très importante pour le traitement. Les données contiennent les attributs de ce type sont de format `25/03/2017 09:21:35` ce qui correspond à un motif de description `DD/MM/YYYY hh24:mi:ss`.

4 Outils logiciel à utiliser et autres outils et sources d'information. Rendu

L'usage de la base Oracle n'est pas obligatoire, si vous préférez un autre SGBD, comme, par exemple, MySQL ou PostgreSQL, vous êtes libres de le faire.

L'outil NoSQL à utiliser sera soit Cassandra, soit OracleNoSQL, selon votre préférence.

Pour la partie exploratoire, la partie intégration de données ou la partie rendu de résultats, vous avez la liberté d'utiliser tout autre outil ou source d'information, à condition de les nommer explicitement *(exemples : nous avons utiliser la commande* `wc` *pour vérifier que .. , à l'aide de la fonction* `kmeans` *du logiciel R nous avons obtenu ...)*.

Le projet sera rendu sous forme .pdf avec tout détail qui vous semble utile. Les parties intéressantes techniquement mais de taille conséquente seront mises en annexes.

5 Travail à effectuer

1. Sans mettre des index ou des autres clés sur la base relationnelle écrivez des requêtes pour trouver :

(a) combien d'enregistrements il y a dans chaque table ;

(b) quel est le dernier élément inséré dans la table EVENTS et quelle est sa référence /position ?

(c) pour chaque capteur indiquez son nom identifiant, sa position et le nombre d'enregistrements EVENTS le concernant ;

(d) quelle est l'intervalle des valeurs possibles pour chaque mesure ?

(e) notez pour chacune de requêtes le temps de réponse.

2. Faites une copie des tables dans votre espace (CREATE TABLE ...) ou travaillez directement sur les tables de l'utilisateur `c##capteurs` qui a accordé le droit de référencer et de créer des index sur ses tables. Si vous avez besoin d'autres tables, créez les.

   (a) détectez les valeurs hors rang (outliers) et éliminez les.

   (b) mettez les clés primaires, les clé étrangère et les index qui vous semblent utiles. Notez soigneusement le temps de création des index (s'il y en a) pour la table EVENTS.

   (c) reprenez les requêtes du point précédent et regardez s'il y a eu un gain de temps ;

   (d) quelles étaient pour un jour donné (exemple 25/05/2017) les valeurs réelles et moyennes des trois paramètres pour un capteur donné ?

   (e) détectez le capteur qui a la plus grande variation de valeurs pour la température, puis pour l'humidité, puis pour la luminosité.

3. Implémentez une base NoSQL de votre choix et faites les requêtes des points précédents. Est-ce qu'il y a un gain de temps ?

4. Dans une optique entrepôt de données proposez un fait lié aux valeurs mesurées par le capteur et décrivez un modèle en étoile ou en flocon pour les données. Vous pouvez supposer que des capteurs ont été mis dans divers bâtiments situés sur divers campus d'une même institution. Je conseille l'outil Indyco pour la représentation du modèle de data warehouse.

5. Quelle exploitation feriez vous des données de cet entrepôt ?

6. Essayez de trouver quels capteurs sont placés dans des bureaux et les quels dans des parties communes. Quels capteurs sont placé dans des pièces sans fenêtres ?

7. Tentez d'indiquer les jours des vacances scolaires, les jours de fermeture de l'Ecole, les jours sans étudiants.

8. Est-ce que vous constatez des mesures de températures qui dépassent les limites légales ?

9. Pouvez vous indiquer le nombre de personnes qui travaillent dans ces locaux ?

10. Pouvez vous indiquer le nombre de filles et de garçons ou le ratio ?

11. Proposez une autre exploitation possible des données capturées.