Sujet de master recherche « Architectures logicielles distribuées » 2004–2005

# Intégration de l'interrogation de résumés de données dans le SGBD PostgreSQL

Encadrant principal : Laurent UGHETTO
courriel : `Laurent.Ughetto@univ-nantes.fr`
tél. : 02 51 12 58 37

Co-encadrant(s) : Noureddine MOUADDIBcourriel : `Noureddine.Mouaddib@univ-nantes.fr`
tél. : 02 40 68 32 02

## Objectif du stage

Les techniques du résumé de données sont aujourd'hui considérées comme un bon moyen de traiter les grandes masses de données, en particulier lorsque les valeurs précises de ces données ne sont pas nécessaires. Cependant, la plupart du temps, les résumés produits sont destinés à être exploités directement par un utilisateur humain, et il existe peu d'outils de traitement automatique. Or, résumer de très grandes masses de données conduit immanquablement à produire un grand nombre de résumés, nombre qui dépasse largement les facultés de traitement humaines.

Dans l'équipe ATLAS-GRIM, un premier outil d'interrogation de résumés a été proposé, pour exploiter de façon plus efficace les résumés produits par le modèle SAINTETIQ.

Grossièrement, le modèle SAINTETIQ résume les données en définissant, pour chaque attribut, un vocabulaire plus grossier que le domaine de définition initial qui permet, après réécriture, de regrouper les tuples devenus indiscernables. Ce vocabulaire plus grossier est défini par des « variables linguistiques », c'est-à-dire des partitions floues du domaine de définition. À un niveau supérieur, des regroupement successifs des valeurs d'attributs permettent d'obtenir une hiérarchie de résumés, dont la racine regroupe (résume) l'ensemble des données. Le résultat est donc une hiérarchie de résumés, du plus général (la racine) aux plus spécifiques (les feuilles). Chaque feuille contient aussi un lien vers les données qu'elle résume. Ainsi, à partir des résumés feuilles, on peut aussi retrouver les données initiales. De ce fait, la hiérarchie obtenue peut être considérée comme une sorte d'index multidimensionnel.

Des algorithmes d'interrogation de cette structure ont été développés et testés à l'aide d'un prototype, dans le cadre de la thèse d'Amenel Voglozin.

L'étape suivante consiste à essayer d'intégrer cet outil d'interrogation à un SGBD `classique`, comme PostgreSQL. En effet, les résumés produits par SAINTETIQ, s'ils peuvent être utilisés directement, ont vocation à être intégrés à un SGBD. Il faut donc fournir à ce SGBD les moyen d'interroger efficacement la structure de ces résumés.

Ainsi, les résumés proposent une représentation condensée de la base de données, et l'outil d'interrogation permet d'avoir une vision rapide et synthétique de son contenu. De plus, on a vu que l'organisation arborescente des résumés leur confère un caractère d'index multidimensionnel. Dans ce cadre, il sera aussi intéressant d'étudier l'utilisation des résumés dans le SGBD sous l'angle de l'optimisation des requêtes classiques. Au lieu d'interroger directement la base, on interroge d'abord les résumés pour ensuite retrouver les tuples correspondants. L'integration de l'interrogation des résumés dans PostgreSQL permettra d'effectuer des mesures d'optimisation.

## Travail à réaliser

Pour intégrer l'interrogation des résumés à PostgreSQL, deux pistes sont envisagées, qui pourront être explorées parallèment et comparées ensuite :

- La première consiste à utiliser le moteur de requêtes de PostgeSQL. Pour cela, il faudra d'une part déterminer une représentation adéquate de l'arborescence des résumés dans le SGBD, et d'autre part mettre en place des filtres pour réécrire la requête de l'utilisateur et reformater les réponses du SGBD.
- La seconde consiste à essayer de tirer partie des algorithmes d'interrogation développés dans l'équipe, qui exploitent la structure hiérarchique des résumés de données. Pour cela, il faudra étudier la possibilité d'intégrer les algorithmes *maison* au moteur de requêtes de PostgreSQL.

Ces deux approches pourront ensuite être comparées, avec l'objectif de déterminer l'efficacité apportée par les algorithmes d'interrogation développés dans l'équipe, et par la représentation de la structure de résumés dans le SGBD.

Pour chacune de ces méthodes, le gain apporté par l'interrogation préalable des résumés dans le cadre du requêtage classique sera aussi calculé, sur des benchmarks à déterminer.

Ce travail pourra être poursuivi en thèse, avec notamment l'étude de l'intégration complète de SaintEtiQ au SGBD PostgreSQL.

## Bibliographie

Le stagiaire s'appuiera principalement sur les publications de l'équipe, des ouvrages spécialisés sur le SGBD utilisé, et la documentation technique disponible sur internet.