M1 informatique - Bases de données

Partie Optimisation. Projet

1 Objectifs

Ce projet a pour buts de compléter le cours de base de données des deux manières suivantes :

1. collecter des informations sur différents SGBD (**objectif 1**), et
2. mettre en pratique les notions vues en cours sur une base de données de test, avec une charge de test (**objectif 2**).

Pour l'objectif 1 : les notions de cours sont vues de manière théorique, sans être liées à un SGBD particulier. Il s'agit donc d'étudier comment elles ont été mises en pratique dans le cas d'un SGBD particulier.

Pour l'objectif 2 : il s'agit d'utiliser un benchmark classique (en l'occurence le benchmark TPC-H, cf. http://www.tpc.org/tpch/default.asp) pour générer un base de test sur laquelle seront évaluées des requêtes considérées comme la charge de la base. Les notions du cours devront alors être appliquées pour minimiser le coût d'évaluation des requêtes de la charge (c'est à dire mettre en oeuvre une stratégie d'optimisation de requête).

2 Organisation

Le projet sera traité par groupe de 3 étudiants. Chaque groupe doit travailler sur un SGBD différent. Le choix du SGBD est laissé libre, la seule contrainte étant que le SGBD soit disponible gratuitement. Il sera possible que le groupe et que le SGBD étudié soient les mêmes que pour le projet de langages de requêtes du premier semestre.

3 Travail à faire

Pour l'objectif 1 : il s'agit d'établir une grille reprenant les différents points traités en cours et de remplir cette grille avec les informations collectées sur le SGBD choisi. Cette grille devra compléter la grille obtenue pour le projet du premier semestre, en ajoutant notamment les aspects :

- organisation des données,
- indexation,
- plan d'exécution et évaluation de coût,

– sélection de plan physique.

Pour l'objectif 2 : il s'agit de remettre un rapport détaillant une stratégie d'optimisation de requête sur la base de données de test. Le rapport devra en outre préciser le rôle de chaque membre du groupe dans le projet.

La **base de données de test** est celle du benchmark TCP-H. Son schéma est décrit dans les page 11 à 17 du benchmark (http://www.tpc.org/tpch/spec/tpch2.9.0.pdf).

L'**instance de test** sera générée avec l'outil dbgen (http://www.tpc.org/tpch/spec/tpch_2_8_0.zip), avec un facteur d'échelle (scale factor) de 1 (correspondant environ à 1 Go de données). Les tailles des différentes instances sont données page 90 du benchmark.

La **charge de la base de données** sera constituée de 3 requêtes du benchmark à l'exception des requêtes Q3, Q6, et Q19 du benchmark (décrites respectivement aux pages 32, 38 et 62 du benchmark). Le choix des 3 requêtes utilisées devra être motivé. Les paramètres à utiliser pour ces requêtes seront les paramètres de validation des requêtes (donné dans le benchmark dans la description des requêtes).

4 Déroulement

Le projet sera présenté et discuté lors de la première séance de cours du semestre. Le choix du SGBD par chaque groupe devra être fait et communiqué dans les deux jours qui suivent. Un point sera fait à chaque nouvelle séance pour constater l'avancement du projet.

Pour l'objectif 1, ce point prendra la forme de questions sur la récolte d'information. Voici quelques exemples de questions :

- Quelle(s) taille(s) de bloc est(sont) permise(s) ?
- Quelle taille mémoire est allouée au traitement des requêtes ?
- Quel(s) type(s) d'index est(sont) autorisé(s) ? Comment est-il préconisé d'utiliser les indexes ?
- Le SGBD autorise-t-il le hachage ?
- Quelles statistiques sont gardées pour l'évaluation du coût d'une requête ?

La grille construite en réponse au premier objectif devra être remise au plus tard le **10 mars 2010**.

Pour l'objectif 2, le planning **indicatif** ci-dessous pourra être utilisé :

date	travail accompli
18/01	étude du benchmark
25/01	génération des données avec dbgen essai de définition des schémas et alimentation des tables évaluation de la charge test sans optimisation
03/03	prise de connaissance de la documentation du SGBD (organisation des données, indexation, optimiseur, etc.)
19/03	test de différentes organisations test de différentes indexations test de différents plans d'exécution
26/03	remise du rapport

5 Evaluation

L'objectif 1 comptera pour 30% de la note de projet. Il sera notamment tenu compte du fait que la grille rendue couvre tous les aspects étudiés en cours.

L'objectif 2 comptera pour 70% de la note de projet. Le rapport devra présenter la solution d'optimisation retenue avec les coût d'évaluation obtenus, ainsi que les alternatives envisagées.