LOG660 - Base de données de haute performance

Chapitre 19: Entrepôts de données

Exercices:

QUESTION 1

a) Donnez 3 différences entre une BD transactionnelle (OLTP) et un entrepôt de données (OLAP).

b) Décrivez le schéma en étoile que l'on retrouve dans les entrepôts de données. Précisez le rôle des différents types de tables dans un tel schéma.

c) Expliquez la différence entre les clauses CUBE et ROLLUP.

d) Identifiez un groupe d'attributs formant une hiérarchie dimensionnelle dans la table de dimension suivante :

Produit
idProduit (PK)
description
SKU
marque
sousCatégorie
catégorie
département
poids
taille
couleur
…

Comment serait modélisée cette hiérarchie dans un schéma normalisé ?

e) À quoi sert la pré-agrégation des faits et comment peut-on implémenter cette stratégie dans le contexte des BD relationnelles ?

f) Décrivez brièvement l'architecture *bus de magasins de données* pour les entrepôts de données. Quel type de modélisation est normalement employé pour les magasins de données ?

g) Dans quelle(s) situation(s) l'architecture d'entrepôts de données fédérés est-elle recommandée ?

h) Nommez deux avantages de la modélisation dimensionnelle par rapport au modèle entités-relations, dans le contexte analytique ?

i) Illustrez à l'aide d'un exemple chacune des opérations OLAP suivantes : *slice*, *rotate*, *roll-up*, et *drill-down*.

j) Quelle est la différence entre les approches ROLAP et MOLAP pour la gestion de données dimensionnelles ?

QUESTION 2

Considérez le schéma suivant:

- Film(idFilm, titre, annee, description);
- Personne(idPersonne, nom, dateNaissance, bio);
- RoleFilm(idFilm, idPersonne, personnage);
- GenreFilm(idFilm, genre);

a) Écrivez une requête SQL utilisant les *fonctions analytiques* d'Oracle permettant d'afficher, pour chaque genre et chaque année, le nombre de films produits depuis cette année ayant ce genre.

b) Quelle est la différence entre les résultats obtenus par les deux requêtes suivantes ?

Requête 1

```
SELECT F.idFilm, COUNT(*) as col
FROM Film F, GenreFilm GF
WHERE F.idFilm = GF.idFilm
GROUP BY F.idFilm
```

Requête 2

```
SELECT F.idFilm, COUNT(*) OVER (
  PARTITION BY F.idFilm) as col
FROM Film F, Genre G
WHERE F.idFilm = G.idFilm
```

QUESTION 3

Une agence de voyage aimerait pouvoir analyser ses données afin de planifier de meilleures campagnes de promotion auprès de ses clients. Plus particulièrement, elle aimerait analyser le nombre et le montant des ventes en fonction :

- De la destination: hôtel, ville, pays, région, catégorie de région (ex: bord de mer, alpine, etc.), catégorie de destination (ex: familial ou non), catégorie hôtel (ex: 1-4 étoiles) ;
- De la date d'achat: jour de l'année, jour de la semaine, mois, année, saison touristique (ex: basse ou haute saison);
- De la date de départ: jour de l'année, jour de la semaine, mois, année, saison touristique (ex: basse ou haute saison);

- Du forfait: nombre de personnes, nombre de nuits, type de forfait (ex: tout inclus, repas inclus, etc.), type de chambre (ex: standard, suite, penthouse, etc.) ;
- Du client: groupe d'âge, sexe, adresse, type d'acheteur (ex: nouveau, récurrent, etc.) ;
- Du canal de vente: catégorie (ex: magasin, internet, etc.) ;
- De la promotion: catégorie (ex: 2 pour 1, rabais 10%, rabais 25%, etc.), début et fin de validité ;
- Du mode de paiement: catégorie (ex: crédit, comptant, etc.) ;

a) Proposez un schéma en étoile permettant de faire ces analyses. Identifiez clairement les clés primaires et étrangères des tables de faits et de dimension;

b) Identifiez, pour chaque table de dimension, une hiérarchie de niveaux de granularité (e.g., attribut$_1$ ← attribut$_2$ ← …) ;

c) Proposez une stratégie d'agrégation ajoutant une nouvelle table de faits agrégés. Donnez le code SQL permettant de créer cette nouvelle table.

QUESTION 4

Source : http://community.mis.temple.edu/mis2502sec002s13/2013/03/06/in-class-exercise-star-schema-dimensional-modeling/

TU Hôtels est une petite chaîne d'hôtels ayant des propriétés dans plusieurs états américains. L'entreprise possède une base de données centralisée pour stocker et faire le suivi des réservations de ses clients. En 2008, ils ont installé des restaurants appelés *Café in the Hotel* dans plusieurs de leurs hôtels. Un système est employé pour faire le suivi des commandes et les relayer aux employés dans les cuisines.

TU Hôtels aimerait utiliser les données qu'ils ont emmagasinées pour mieux comprendre la performance de leurs hôtels et restaurants. Ils ont également accès à une base de données de critiques de clients provenant du site web *HotelComplainer.com*.

La tâche est de faire la conception de deux magasins de données (*data marts*) utilisant les données provenant des trois sources mentionnées ci-haut. Vous devrez faire un schéma en étoile pour chaque magasin de données en choisissant les dimensions, les faits, et les attributs à partir des sources, dont le schéma est fourni ci-dessous.

La table suivante présente les questions analytiques auxquelles devra répondre vos magasins de données :

Data mart 1: Performance des hôtels	Data mart 2: Performance des restaurants
• Durant quel mois y a-t-il le plus grand nombre de réservations de chambre? • Quelle est la saison morte pour les hôtels situés dans une région particulière? • Quels hôtels génèrent le plus de revenus (non attribuables aux restaurants)?	• Quels restaurants génèrent le plus de revenus? • Les restaurants les mieux cotés génèrent-ils plus de revenus? • Quel est l'item commandé le plus souvent dans une région particulière?

- Quel est la durée moyenne des séjours dans les hôtels de 4.5 étoiles ou plus?
- Les fumeurs restent-ils plus longtemps que les non-fumeurs?
- Pour un hôtel donné, combien y a-t-il de clients provenant d'un autre état?

Pour compléter l'exercice, vous devrez suivre les étapes suivantes :

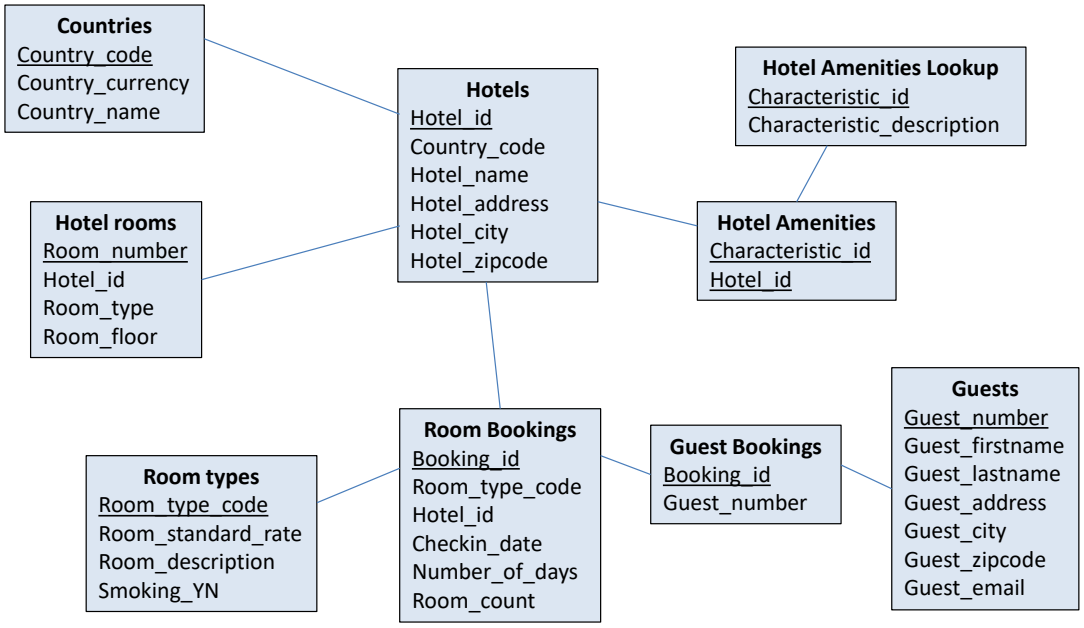
1) **Identifiez le principal évènement d'affaires pour chaque magasin de données;**
   Posez-vous la question suivante : "Quel est l'évènement d'affaires qui génère la (les) métrique(s) de performance?

2) **Identifiez les attributs associés aux faits.**
   Posez-vous la question suivante : "Comment l'évènement d'affaires est-il mesuré?"

3) **Identifiez les dimensions et leurs attributs.**
   Posez-vous la question suivante : "Quelles données caractérisent les différents aspects de l'évènement d'affaires?"

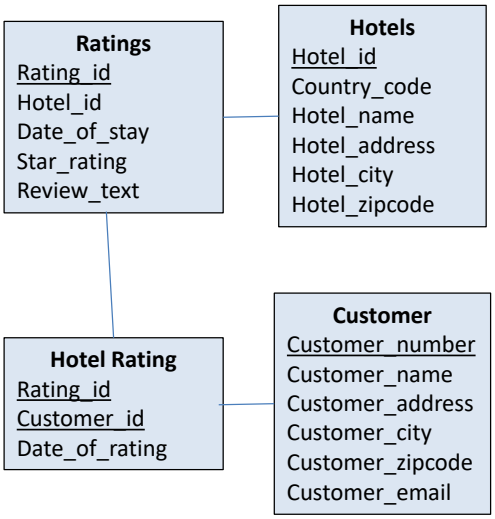
4) **Élaborez le schéma en étoile selon les principes vus en classe.**

Hotel Reservation Database



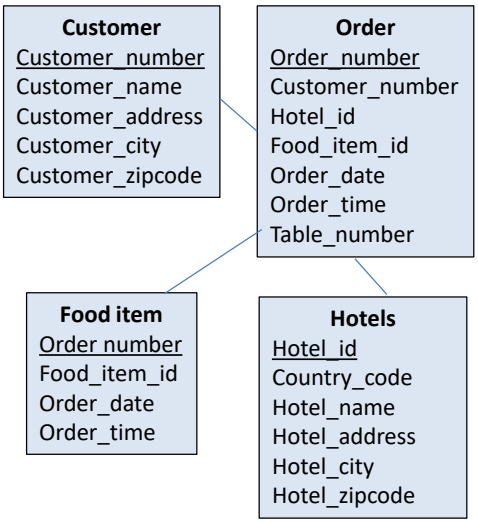
HotelComplainer Ratings Database

(totally external company)



Café in the Hotel Database

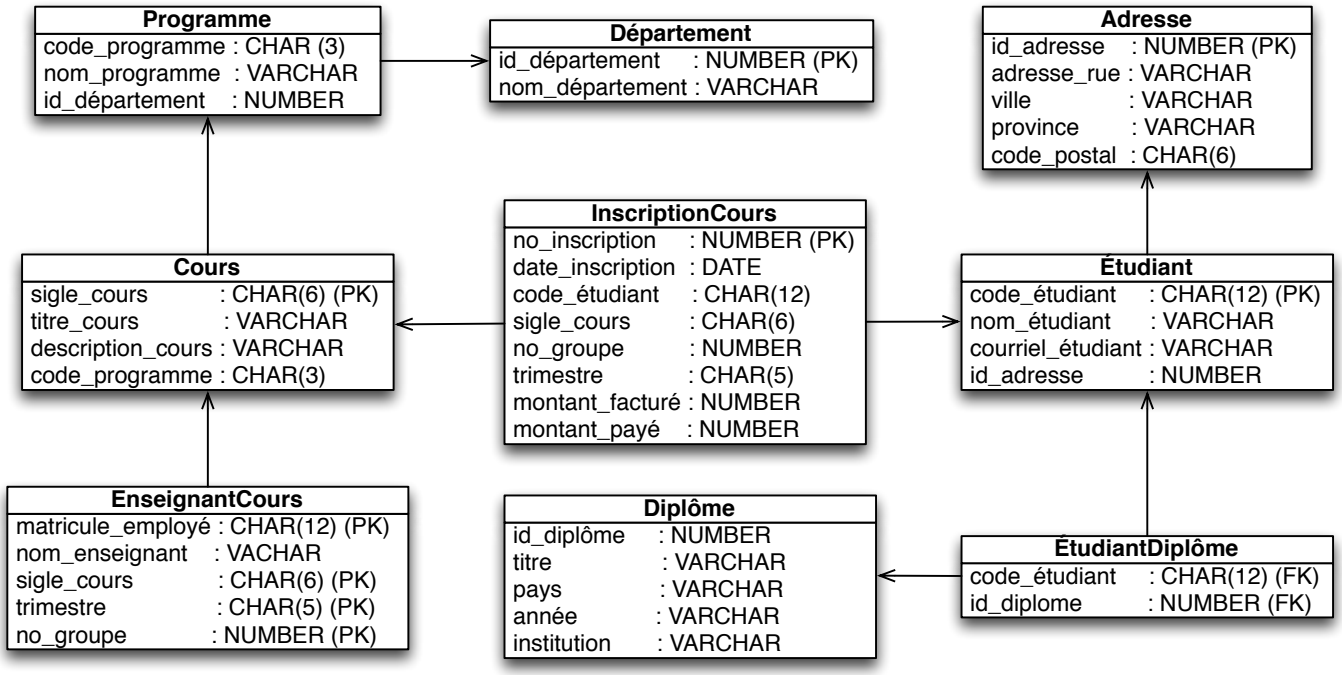
(same company but database is not connected to the hotel)



QUESTION 5

L'École de Technologie Infinie (ETI) désire développer un magasin de données afin d'améliorer la gestion de ses programmes d'enseignement. La principale source d'informations serait la base de données du registraire, dont le schéma relationnel est fourni ci-dessous. Le magasin de données devrait permettre de répondre, entre autres, aux questions analytiques suivantes :

- Quels sont les cours d'un certain programme ayant le plus (le moins) d'inscriptions ?
- Quel est le nombre moyen d'étudiants par groupe ?
- Quels enseignants ont eu les plus gros (plus petits) groupes d'étudiants ?
- Quelle est la proportion d'étudiants, ayant suivi les cours d'un certain programme, qui habitent à l'extérieur de Montréal ?


Programme
code_programme : CHAR (3)
nom_programme : VARCHAR
id_département : NUMBER
Département
id_département : NUMBER (PK)
nom_département : VARCHAR
Adresse
id_adresse : NUMBER (PK)
adresse_rue : VARCHAR
ville : VARCHAR
province : VARCHAR
code_postal : CHAR(6)
InscriptionCours
no_inscription : NUMBER (PK)
date_inscription : DATE
code_étudiant : CHAR(12)
sigle_cours : CHAR(6)
no_groupe : NUMBER
trimestre : CHAR(5)
montant_facturé : NUMBER
montant_payé : NUMBER
Cours
sigle_cours : CHAR(6) (PK)
titre_cours : VARCHAR
description_cours : VARCHAR
code_programme : CHAR(3)
Étudiant
code_étudiant : CHAR(12) (PK)
nom_étudiant : VARCHAR
courriel_étudiant : VARCHAR
id_adresse : NUMBER
EnseignantCours
matricule_employé : CHAR(12) (PK)
nom_enseignant : VACHAR
sigle_cours : CHAR(6) (PK)
trimestre : CHAR(5) (PK)
no_groupe : NUMBER (PK)
Diplôme
id_diplôme : NUMBER
titre : VARCHAR
pays : VARCHAR
année : VARCHAR
institution : VARCHAR
ÉtudiantDiplôme
code_étudiant : CHAR(12) (FK)
id_diplome : NUMBER (FK)