

Concepteurs : Ahlame DOUZAL, Didier DONSEZ

Durée : 2 heures

Remarques : Calculatrice et tout document autorisé

Conseil : Lire le sujet jusqu'au bout.

Annexe : Résultats du benchmark TPC/H

**Rendre les réponses aux 2 problèmes sur des copies séparées**

**Problème 1 (10 points): Modélisation décisionnelle d'un Entrepôt de Données sur l'assurance maladie**

|| + CORRECTION

Le Ministère de la Santé et du Bien-Etre de Grolang vous sous-traite la réalisation d'un entrepôt de données pour réaliser des études sur les dépenses de santé dans son beau-pays. Les bases de production de cet entrepôt sont les systèmes d'information des centres de sécurité sociale et des assurances santé complémentaire de Groland qui gèrent les dossiers (électroniques) des assurés.

Le schéma de l'entrepôt est constitué des tables suivantes (les clés primaires sont soulignées)

**Date**(CléDate, Année, Mois, JourDeMois, JourDeSemaine, TrancheHoraire, DrapeauVacances)

**Assuré**(CléAssuré, MoisNaissance, AnnéeNaissance, MoisDécès, AnnéeDécès, Région, Département, District, Ville, Quartier, RevenuAssuré, RevenuFoyer, CatégorieSocioProfessionnel, SousCatégorieSocioProfessionnel, DomaineActivité, CaissePrimaire, CaisseComplémentaire, DrapeauAssuréPrincipal)

**Pratitien**(CléPratitien, Spécialité, SousSpécialité, Région, Département, District, Ville, Quartier, MoisNaissance, AnnéeNaissance, DrapeauConventionné)

**Acte**(CléDate, CléAssuré, CléPratitien, CléPathologie, MontantActes, MontantPriseEnChargeCaissePrimaire, MontantPriseEnChargeCaisseComplémentaire, NombreMedicamentsPrescrits, MontantPharmacologieGenerique, MontantPharmacologieNonGenerique, MontantDesActesComplémentaires, DrapeauActesComplémentairesBiologie, DrapeauActesComplémentairesChirurgie, DrapeauActesComplémentairesKinésithérapie, DrapeauActesComplémentairesRadiologie, NombreDeJoursDArrêtDeTravail, CoutJoursDArret).

**Pathologie**(CléPathologie, DesignationNormalisé, Spécialité, SousSpécialité, TauxDIncapacité, DuréeTraitement, Chronicité, DrapeauMaladieProfessionnelle)

Rétro-Conception

Q1: Quelle est la table de fait dans cet entrepôt ?. Justifiez !

|| Acte (car au centre des dimensions, attributs additifs ou numériques)

Q2: A votre avis, il y a t'il des dimensions douteuses dans cet entrepôt ? Rappelez la définition et justifiez.

|| Les assurés et patriciens sont très bien identifiés dans les bases de production (Num INSEE et Numero Ordre Medecin).

|| Par contre, la codification de la pathologie peut être ambiguë donc douteuse !.

Q3: Donnez les nouvelles tables si on décide de diminuer la taille de la table Assurée par une mini-dimension démographique

|| On crée une table de dimension Demographie avec les attributs d'Assuré (Région, Département, District, Ville, Quartier, RevenuAssuré, RevenuFoyer, CatégorieSocioProfessionnel, SousCatégorieSocioProfessionnel, DomaineActivité ou d'autres éventuellement)

- On supprime ces attributs d'Assuré
- On ajoute une clé de mini-dimension à la table de fait et à la table de dimension client

### Dimensionnement

Q4: Donnez le nombre de faits présents dans la table de fait.

Nombre d'assurés	60 Millions
Nombre de actes par praticien et par jour (Un praticien travaille 300 jours par an)	20
Nombre de praticiens	300 000
Montant total moyen d'un acte	100 Euros
Nombre d'actes supplémentaires prescrit par acte	0,1
Nombre d'années	6
Coûts annuel des actes	180 Milliard d'Euro
Taille des clés	4 octets
Taille des attributs numériques	4 octets
Taille des attributs booléens (comme les drapeaux !)	1 octet

Nombre d'actes=300000\*20\*300\*6=10 800 000 000 actes ou enregistrements  
 Donnez la taille d'un enregistrement de la table de fait ?

(4\*4 clés + 4\*9 attributs numériques + 1\*4 attributs booléens)=56 octets par fait  
 Donnez la taille (en Octets) de stockage de la table de fait.

Taille de la table de fait= 60480000000 octets soit 563,3 Go

### Configuration Matérielle

Q5: A partir des résultats du benchmark TPC/H ([http://www.tpc.org/tpch/results/tpch\\_results.xls](http://www.tpc.org/tpch/results/tpch_results.xls)) donné en annexe, choisissez la configuration matérielle et logicielle (complète) qui est la plus adaptée à votre infocentre pour une performance minimale de 12000 QphH ? Quels sont vos critères de choix ?

**Remarque : vous négligerez la taille des tables de dimensions.**

Pour un SF=1000Go  
 Sun Sun Fire[™] 15K server 1000 **18802.1** 256 **4815969**  
 US \$ Oracle 9i Database Enterprise Edition Sun Solaris 8 Sun UltraSPARC III  
 Cu 900 MHz  
 On pouvait aussi choisi le HP Super 2.0 pour une poignée de dollars de plus.

### Rapports

Q6: Donnez la requête SQL qui donne le top 10 des sous-spécialités des pathologies qui ont entraîné le plus de dépenses (montant des actes + montant pharmacologies)?

```
SELECT P.Spécialité, P.SousSpécialité, SUM(MontantActes+ MontantPharmacologieGenerique
+ MontantPharmacologieNonGenerique) AS MontantTotalActes
FROM Pathologie P JOIN Acte A USING (CléPathologie)
GROUP BY P.Spécialité, P.SousSpécialité
SORT BY MontantTotalActes DESC
TOP(10)
```

Q7: Donnez le rapport mensuel de progression du ratio des montants des médicament génériques par rapport aux médicaments non génériques.

```
SELECT D.Année, D.Mois,
SUM(A.MontantPharmacologieGenerique)/SUM(A.MontantPharmacologieNonGenerique)
AS Ratio
FROM Acte A JOIN Date D USING (CléDate)
GROUP BY D.Année, D.Mois
```

---

|| SORT BY D.Année ASC, D.Mois ASC

Q8: Donnez le rapport précédent mais avec une moyenne glissante sur les 3 mois précédents.

```
|| SELECT BY D.Année D.Mois
|| ( SUM(A.MontantPharmacologieGenerique) OVER (
||     ORDER BY D.Année D.Mois ASC
||     ROWS 2 PRECEDING)
|| / SUM(A.MontantPharmacologieNonGenerique) OVER (
||     ORDER BY t.Annee, t.Mois ASC
||     ROWS 2 PRECEDING)
|| ) AS Moyenne_Mouvante_Ratio
|| FROM Acte A JOIN Date D USING (CléDate)
|| GROUP BY D.Année, D.Mois
|| SORT BY D.Année ASC, D.Mois ASC
```

**Problème 2 (10 points): Analyse des dépenses de santé**

On considère le tableau suivant généré à partir de l'entrepôt sur l'assurance maladie :

Assurés	Age	Cat-So-Pr	Stab-Eco	Caisse-Com	Nb-Praticien	Mt-Dep	Mt-Rem
P1	25	O	I	0	2	200	100
P2	32	C	MS	0	4	750	200
P3	58	C	S	1	6	800	700
P4	62	R	S	1	10	1500	1200
P5	75	R	I	1	3	1000	350
P6	84	R	MS	1	3	950	900
P7	43	O	I	0	2	280	120

Le rapport annuel ci-dessus fournit la description des assurés de la caisse d'assurance maladie par leur Age, leur catégorie socio-professionnelle (Cat-So-Pr), la situation économique de leur foyer d'appartenance (Stab-Eco), s'ils bénéficient ou pas d'une caisse complémentaire (Caisse-Com), le nombre de praticiens différents visités durant l'année (Nb-Praticien), le montant total des dépenses (Mt-Dep) ainsi que le montant total des remboursements (Mt-Rem).

Type et domaine des attributs :

- Attributs quantitatifs

Age : entier sur [0-150]

Nb-Praticien : Réel

Mt-Dep : Réel

Mt-Rem : Réel

- Attributs qualitatifs ordonnés

Stab-Eco : I (Instable), MS (Moyennement Stable), S (Stable) avec  $I < MS < S$

- Attributs qualitatifs non ordonnés

Cat-So-Pr : O (Ouvrier), C(Cadre), R (Retraité)

- Attributs binaires

Caisse-Com : 1 (l'assuré bénéficie d'une assurance complémentaire partenaire de la caisse d'assurance maladie)

0 (l'assuré ne bénéficie pas d'une assurance complémentaire ou celle-ci n'est pas partenaire de la caisse d'assurance maladie).

Q1- On se situe dans l'espace de description défini par les attributs ci-dessus. Évaluez la dissimilarité entre les assurés P1 et P2.

On considère les matrices symétriques suivantes donnant le schéma d'agrégation des modalités des attributs Cat-So-Pr et Stab-Eco :

$$\begin{matrix}
 & O & C & R \\
 O & \begin{pmatrix} O & O & R \end{pmatrix} \\
 C & \begin{pmatrix} - & C & C \end{pmatrix} \\
 R & \begin{pmatrix} - & - & R \end{pmatrix}
 \end{matrix}
 \qquad
 \begin{matrix}
 & I & MS & S \\
 I & \begin{pmatrix} I & I & MS \end{pmatrix} \\
 MS & \begin{pmatrix} - & MS & S \end{pmatrix} \\
 S & \begin{pmatrix} - & - & S \end{pmatrix}
 \end{matrix}$$

Donnez la description du nuplet central de P1 et P2.

---

Q2- La caisse d'assurance maladie considère comme dépense élevée toute dépense annuelle supérieure (au sens strict) à 800 euros. Par ailleurs, les experts en assurance maladie considèrent qu'un pourcentage de remboursement supérieur à 70%, la catégorie socio-professionnelle des assurés ainsi que la situation économique de leur foyer d'appartenance ont un effet sur l'augmentation des dépenses.

Après avoir procédé aux codages adéquats, appliquez un arbre binaires de décision afin d'extraire les principales règles expliquant les dépenses élevées en fonction de la catégorie socio-professionnelle, de la stabilité économique des assurés ainsi que du pourcentage des remboursements. L'hypothèse énoncée ci-dessus est-elle vraie ? Justifiez votre réponse.

Q3- On souhaite partitionner l'ensemble de nos assurés en trois principaux profils. Pour cela, on restreint l'espace de description aux deux attributs Age et Nb-Praticien. Appliquez la méthode de votre choix afin d'extraire ces trois principaux profils. Donnez pour chaque profil extrait sa description par la totalité des attributs.

Q4- On considère le rapport suivant donnant pour chaque assuré la liste des pathologies pour lesquelles il y a eu prescription durant l'année.

Assurés	P1	P2	P3	P4	P5	P6
Pathologies	Pa1	Pa1	Pa2	Pa1	Pa3	Pa2
	Pa2	Pa2	Pa3	Pa2	Pa4	Pa3
	Pa3	Pa4	Pa4	Pa3		

Extraire les associations de pathologies les plus pertinentes en précisant les degrés de support et de confiance associés.