Ucu32u 91 MIAGE
9/3/2010

Examen de Informatique Décisionnelle
Master 1 Miage – 1ere session
Tout document autorisé - Durée 3 heures

Exercice 1 Datawarehouse

Le Ministère de la Santé et du Bien-Etre de Grolang veut construire un entrepôt de données afin de stocker les informations sur les consultations d'un pays. On veut notamment connaître le nombre de consultations, par rapport à différents critères (personnes, médecins, spécialités, etc). Ces informations sont stockées dans les relations suivantes :
PERSONNE (id, nom, tel, adresse, sexe)
MEDECIN (id, tel, adresse, spécialité)
CONSULTATION (id_med, id_pers, date, prix)
Question 1 : Proposer un schéma en étoile et les tables relationnelles correspondantes qui tienne compte de la date, du jour de la semaine, du mois, du trimestre et de l'année. Représenter le cube OLAP.

Question 2 : A partir de ce cube, indiquez quelles opérations OLAP (roll ul, drill down, slice, dice) il faut appliquer pour obtenir les informations suivantes :

a) le coût total des consultations par médecin en 2005
b) le nombre de consultations par jour de la semaine, par spécialité et par sexe du patient
c) le coût des consultations par patiente pour les mois d'octobre

Exercice 2 Datawarehouse

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, une autre société a déjà proposé un premier schéma. 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)
Patricien(CléPraticien, Spécialité, SousSpécialité, Région, Département, District, Ville, Quartier,MoisNaissance, AnnéeNaissance, DrapeauConventionné)
Acte(CléDate, CléAssuré, CléPraticien, CléPathologie, MontantActes, MontantPriseEnChargeCaissePrimaire, MontantPriseEnChargeCaisseComplémentaire, NombreMedicamentsPrescrits, MontantPharmacologieGenerique, MontantPharmacologieNonGenerique, MontantDesActesComplémentaires, DrapeauActesComplémentairesBiologie, DrapeauActesComplémentairesChirugie, 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 !
Q2: A votre avis, il y a t'il des dimensions douteuses dans cet entrepôt ? Justifiez.
Q3: Donnez les nouvelles tables si on décide de diminuer la taille de la table Assurée

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 20
- (Un praticien travaille 300 jours par an)
- Nombre de praticiens 300 000
- 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

Donnez la taille d'un enregistrement de la table de fait ?
Donnez la taille (en Octets) de stockage de la table de fait

Configuration Matérielle

Q5: A partir des résultats du benchmark TPC/H (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 ?

Exercice 3 Analyse des dépenses

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

Tableau 1 : assurés

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).

Les attributs sont :

- Age : entier sur [0-150]

- Nb-Praticien : Réel
- Mt-Dep : Réel
- Mt-Rem : Réel
- Stab-Eco : I (Instable), MS (Moyennement Stable), S (Stable) avec I < MS < S
- Cat-So-Pr : O (Ouvrier), C(Cadre), R (Retraité)
- 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. Evaluez la dissimilarité entre les assurés P1 et P2. Vous décrirez la notion de dissimilarité utilisée avant de faire les calculs.
Q2 - 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.
Q3 - 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 Pa2 Pa3	Pa1 Pa2 Pa4	Pa2 Pa3 Pa4	Pa1 Pa2 Pa3	Pa3 Pa4	Pa2 Pa3

Tableau 2 : Pathologie

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

Exercice 4 : Weka

On a réalisé sous weka une analyse des fichiers du tableau 1.

On obtient les résultats suivant : (à droite les numéros indiquent juste le numéro de la ligne)

```
=== Run information ===                                                    1
                                                                           2
Scheme:       XXXXXX                                                       3
Relation:     pathologie                                                   4
Instances:    7                                                            5
Attributes:   7                                                            6
              age                                                          7
              CatSoPr                                                      8
              StabEco                                                      9
              Caisse-Com                                                   10
              NbPraticien                                                  11
              MtDep                                                        12
              MtRem                                                        13
Test mode:    evaluate on training data                                    14
                                                                           15
=== Model and evaluation on training set ===                               16
                                                                           17
                                                                           18
Number of iterations: 2                                                    19
Within cluster sum of squared errors: 6.646703718232219                    20
                                                                           21
```

```
Cluster centroids:                                                          22
                                                                            23
Cluster 0                                                                   24
      Mean/Mode:    69.75   R S 1   5.5     837.5     787.5                 25
      Std Devs:    11.9548 N/A       N/A       N/A         3.3166 576.4475  26
356.7796
Cluster 1                                                                   27
      Mean/Mode:    33.3333 O I 0   2.6667 410       140                   28
      Std Devs:     9.0738 N/A       N/A       N/A         1.1547 297.1532  29
52.915
                                                                            30
Clustered Instances                                                         31
                                                                            32
0       4 ( 57%)                                                            33
1       3 ( 43%)                                                            34
```

Quel algorithme a été lancé sous weka ?
Commentez les résultats obtenus (vous pouvez vous servir des numéros de lignes pour plus facilement expliquer ce que vous observez).

ANNEXE

		TPC-H BENCHMARK RESULTS										
These results are valid as of date 1/14/2004 12:49:52 PM												
		TPC-H Results - Revision 1.X - 1000GB Scale Factor										
Company	System	QphH	Price Perf.($/	Total Sys. Co	Currency	Database Software	Operating System	CPU Type	# CPU's	Cluster	Date Submitte	Availability Da
Sun	SunFire V880	2241	104	232206,11	US $	Sybase IQ 12.5	Sun Solaris 9	Sun UltraSPA	8	N	02/09/2003	03/09/2003
HP	HP ProLiant DL760 G2 8P	3385	59	199475	US $	Microsoft SQL Server 2000 Enter	Microsoft Windows	Intel Xeon MP	8	N	12/11/2003	12/11/2003
Unisys	Unisys ES7000 Aries 420 Ente	5189	119,12	618309	US $	Microsoft SQL Server 2000 Enter	Microsoft Windows	Intel Itanium2	16	N	15/10/2003	15/10/2003
Legend Lt	Legend DeepComp 6800 Serv	9951	1321	13145728	China Yu	Oracle Database 10g Enterprise	Red Hat Linux Adva	Intel Itanium2	16	Y	06/11/2003	31/12/2003
IBM	IBM eServer p655 with DB2 U	20221	69	1403446	US $	IBM DB2 UDB 8.1	IBM AIX 5L V5.2	IBM Power 4 1	16	Y	08/12/2003	08/08/2004
HP	HPProLiant DL760 X900-128F	22362	253	5654579	US $	IBM DB2 UDB 7.2	Microsoft Windows	Intel Pentium I	128	Y	06/02/2002	20/06/2002
HP	HP 9000 Superdome Enterpris	25805	203	5249167	US $	Oracle 9i Database Enterprise Ed	HP UX 11i 64-bit	HP PA-RISC 8	64	N	29/10/2002	30/10/2002
FuSiemer	PRIMEPOWER 2500	34493	156	5380411	Euros	Oracle Database 10g Enterprise	Sun Solaris 9	Fujitsu SPARC	64	N	08/09/2003	08/03/2004
Fujitsu/ICL	PRIMEPOWER 2500	34493	141	4861985	US $	Oracle Database 10g Enterprise	Sun Solaris 9	Fujitsu SPARC	64	N	13/11/2003	08/03/2004
		TPC-H Results - Revision 1.X - 3000GB Scale Factor										
Company	System	QphH	Price Perf.($/	Total Sys. Co	Currency	Database Software	Operating System	CPU Type	# CPU's	Cluster	Date Submitte	Availability Da
HP	HP 9000 Superdome Enterpris	17909	476	8518084	US $	Oracle 9i Database Enterprise Ed	HP UX 11i 64-bit	HP PA-RISC 8	64	N	29/10/2002	15/05/2002
HP	HP ProLiant DL760 X900-128F	21054	283	5955754	US $	IBM DB2 UDB 7.2	Microsoft Windows	Intel Pentium I	128	Y	06/02/2002	20/06/2002
HP	HP 9000 Superdome Enterpris	27094	213	5781053	US $	Oracle 9i Database Enterprise Ed	HP UX 11i 64-bit	HP PA-RISC 8	64	N	29/10/2002	30/10/2002
Sun	Sun Fire[TM] 15K server	28946	184	5335742	US $	Oracle 9i R2 Enterprise Edition	Sun Solaris 9	Sun UltraSPA	72	N	07/04/2003	30/04/2003
FuSiemer	PRIMEPOWER 2500	34345	161	5541011	Euros	Oracle Database 10g Enterprise	Sun Solaris 9	Fujitsu SPARC	64	N	22/08/2003	22/02/2004
Fujitsu/ICL	PRIMEPOWER 2500	34345	147	5036936	US $	Oracle Database 10g Enterprise	Sun Solaris 9	Fujitsu SPARC	64	N	25/08/2003	22/02/2004
HP	HP Integrity Superdome Enter	45248	109	4922070	US $	Oracle Database 10g Enterprise	HP UX 11i 64-bit	Intel Itanium2	64	N	25/09/2003	25/03/2004
NCR	NCR 5350	79528	213	16937451	US $	Teradata V2R5.0	MP-RAS 3.02.00	Intel Xeon 2.6	128	Y	05/01/2003	20/12/2002
		TPC-H Results - Revision 1.X - 10000GB Scale Factor										
Company	System	QphH	Price Perf.($/	Total Sys. Co	Currency	Database Software	Operating System	CPU Type	# CPU's	Cluster	Date Submitte	Availability Da
HP	HP Integrity Superdome Enter	49105	118	5800875	US $	Oracle Database 10g Enterprise	HP UX 11i, 64-bit B	Intel Itanium2	64	N	05/01/2004	25/03/2004
IBM	IBM eServer p690 with DB2 U	62216	243	15112757	US $	IBM DB2 UDB 8.1	IBM AIX 5L V5.2	IBM Power 4 1	160	Y	05/12/2002	15/05/2003
NCR	NCR 5350	81502	243	19774904	US $	Teradata V2R5.0	Unix MP-RAS 3.02.(	Intel Xeon 2.8	128	Y	23/12/2002	20/12/2002