Université Joseph Fourier
PolyTech Grenoble

Année Universitaire 2004-2005
RICM 3
Systèmes d'Information Décisionnels
Examen

Concepteurs : Didier DONSEZ (Partie 1), Ahlame DOUZAL (Partie 2)
Date : Mars 2005
Durée : 2 heures
Remarques : Calculatrice **RECOMMANDEE** et tout document autorisé
Conseil : Lire le sujet jusqu'au bout.
Annexe : Résultats du benchmark TPC/H

Partie 1 : Modélisation décisionnelle des passagers à vols fréquents

La compagnie aérienne « Rasemotte » veut suivre tous les voyages effectués par chaque participant à son programme de vols fréquents. Le département marketing veut savoir quels sont les vols utilisés par les participants, quels avions prennent ils, quel tarif de base payent ils, combien de fois ils surclassent leur billet, comment paient ils leurs surclassements, comment utilisent ils les points (appelés « miles ») accumulés, s'ils réagissent aux promotions offrant des tarifs spéciaux, dans combien de temps sejournent ils au moins une nuit, quelle est la durée des séjours à ces destinations, quelles autres compagnies sont utilisées durant les mêmes voyages.

L'entrepôt de données est centré sur les voyages du client.
Le schéma de l'entrepôt est constitué des tables suivantes (les clés primaires sont soulignées)

Date(CléDate, Année, Mois, JourDeMois, Semaine JourDeSemaine, TrancheHoraire, Heure, Minute, DrapeauVacances, Drapeau Manisfestation, DrapeauAffluence)
Client(CléClient, Nom, AnnéeNaissance, Région, Département, District, Ville, Quartier, SituationFamille, RevenuFoyer, CatégorieSocioProfessionnel, SousCatégorieSocioProfessionnel, DomaineActivité, NombreEnfant, DateEntréeProgrammeFidélité).
CanalVente(CléCanal, TypeCanal, …)

- **TypeCanal** = « guichet agence, call center, site internet, site internet externe »

Vol(CléVol, ,TypeVol, Catégorie, …)
Promotion(CléPromotion, TypePromotion, CatégoriePromotion, ValeurPromotion, …)
Classe(CléClasse,TypeSiègeAchété, …)

- **TypeSiègeAchété** = » économie, classe affaire, première classe »

Aéroport(CléAéroport, NomAeroport, CodeAéroport, DistanceVille, …)
SegmentVoyage(CléAéroportDepart, CléAéroportArrivée, CléDateDepart, CléDateArrivée, CléDateVente, CléVol, CléClasse, CléPromotion, NumTicket, DuréeVol, DuréeRetard, SatisfactionClient, MilesGagnés, MilesDépensés)

- **SatisfactionClient** prends les valeurs du domaine « Très satisfait, Satisfait, Normal, Mécontent, Très mécontent, Non renseigné »
- **MilesGagnés** corresponds aux points « miles » gagnés grâce à ce voyage
- **MilesDépensés** corresponds aux points « miles » utilisé pour payer ce voyage

Rétro-Conception

Q1: Quelle est la table de fait dans cet entrepôt ?. Justifiez en 2 lignes !

Q2: Quelle est la clé de la table de fait de cet entrepôt ?

Q3: Que pensez vous de l'attribut NumTicket de Voyage ?

Q4: Que peut on dire de l'attribut Satisfaction de Voyage ? Quelles sont les fonctions d'agrégat qui peuvent être appliquées sur cette attributs ?

Q5: Donnez les nouvelles tables si on décide de diminuer la taille de la table Client par une mini-dimension démographique

Comme il est difficile de suivre les clients sur la totalité de leurs voyages (qui sont des suites de segment de voyage), nous introduisons une autre table dans l'entrepôt :

Voyage(CléAéroportDepart, CléAéroportArrivée, CléDateDepart, CléDateArrivée, CléDateVente, CléClasse, CléPromotion, NumTicket, DuréeVol, DuréeRetard, SatisfactionClient, MilesGagnés, MilesDépensés, PrixVoyage)

Q6: Quel est le type de cette table dans cet entrepôt ?. Quelle est la clé ?

Rapports

Q7: Donnez la requête SQL qui donne le top 10 des destinations (finales) les plus fréquentées par les clients quand ils utilisent leurs miles pour payer le billet (MilesDépensés>0) ?

Q8: Donnez la requête SQL qui donne le cumul des miles dépensés par semaine ?

Q9: Donnez le rapport précédent mais avec le ratio de cumul glissant (sur 3 semaines) des miles dépensés sur celui des miles gagnés.

Dimensionnement

Q10: Donnez le nombre de faits présents dans **toutes les tables** de fait.

Nombre de clients dans le programme grand voyageur	1 Million
Nombre de voyage par an	20
Nombre moyenne de vols par voyage	2
Nombre d'années	6
Taille des clés	4 octets
Taille des attributs numériques	4 octets
Taille des attributs discrets	1 octet
Taille des attributs booléens (comme les drapeaux !)	1 octet

Donnez la taille des enregistrements des tables de fait (vous ne tiendrez pas compte des modifications que vous avez proposées à la question Q5) ?
Donnez la taille totale (en Octets) de stockage des 2 tables de fait.

Conception physique

Q11 : Quel espace disque occupe les index bitmap (non compressé) créés sur les 2 tables de fait pour l'attribut TypeCanal de la dimension Canal ?

Q12 : Pourquoi est l'algorithme de jointure étoile (star-query) des SGBD H-OLAP est il plus performant que les algorithme de jointure binaire (2 tables en opérande) utilisés dans les SGBDs OLTP ? (justifiez en 8 lignes max)

Configuration Matérielle

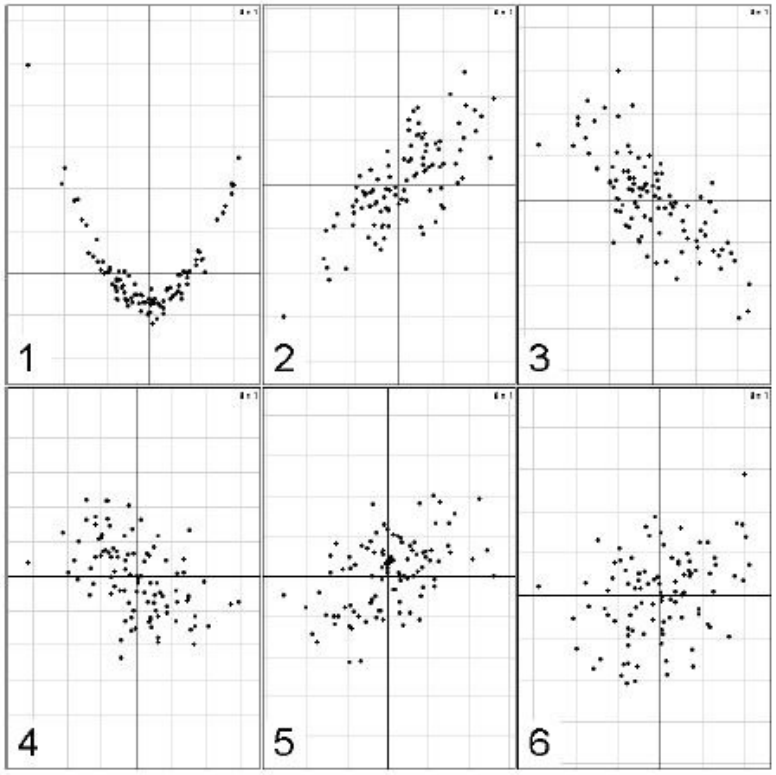
Q13: A partir des résultats du benchmark TPC/H (http://www.tpc.org/tpch/results/tpch_results.xls) donné en annexe, choisissez la configuration matérielle et logicielle (complète) qui est la plus adaptée à votre infocentre pour une performance minimale de 2500 QphH ? Quels sont vos critères de choix ?
Remarque : vous négligerez la taille des tables de dimensions.

Annexe : résultats du benchmark TPC/H (http://www.tpc.org/tpch/results/tpch_results.xls)

		TPC-H BENCHMARK RESULTS						
These results are valid as of date 12/24/2004 9:31:18 AM								
		TPC-H Results - Revision 1.X - 100GB Scale Factor						
Company	System	QphH	Price Perf.($/Q	Total Sys. Cost	Currency	Database Software	Operating System	CPU Type
Sun	SunFire V240 Server	1124	40	45021	US $	Sybase IQ 12.5	Sun Solaris 9	Sun UltraSPARC
HP	HP ProLiant ML370 G3 2P	1386	28	38683	US $	Microsoft SQL Server 2000 Enterp	Microsoft Windows Ser	Intel Xeon 3.06 G
Sun	SunFire V250	1443	23	33495	US $	Sybase Sybase IQ 12.5	Sun Solaris 9	Sun UltraSPARC
HP	HP Proliant DL580 G2	1695	66	111460	US $	Microsoft SQL Server 2000 Enterp	Microsoft Windows 200	Intel Pentium Xe
IBM	IBM eServer xSeries 346	1894	14	26536	US $	IBM DB2 UDB Express Edition v8.1	Suse Linux Enterprise	Intel Xeon 3.6GH
Dell	PowerEdge 6650/2.0/8GB	1984	45	89748	US $	Microsoft SQL Server 2000 Enterp	Microsoft Windows Ser	Intel Xeon MP 2.0
HP	HP ProLiant DL580G2 4P	2106	48	99545	US $	Microsoft SQL Server 2000 Enterp	Microsoft Windows Ser	Intel Xeon MP 2.0
Sun	SunFire V440	2429	28	67704	US $	Sybase Sybase IQ 12.5	Sun Solaris 9	Sun UltraSPARC
HP	HP ProLiant DL580 G2 4P	2606	38	97498	US $	Microsoft SQL Server 2000 Enterp	Microsoft Windows Ser	Intel Xeon MP 2.8
Sun	SunFire V440	2883	19	54277	US $	Sybase IQ 12.5	Sun Solaris 10	Sun UltraSPARC
HP	HP Proliant DL760 G2 8P	3347	65	216782	US $	Microsoft SQL Server 2000 Enterp	Microsoft Windows Ser	Intel Xeon MP 2.0
HP	HP ProLiant DL 760 G2 8P	4225	43	180721	US $	Microsoft SQL Server 2000 Enterp	Microsoft Windows Ser	Intel Xeon MP 2.8
IBM	IBM eServer xSeries 445 8P	5603	73	410229	US $	IBM DB2 UDB 8.1	Microsoft Windows Ser	Intel Xeon MP 2.8
IBM	IBM eServer OpenPower 720 w/D	6357	42	269383	US $	IBM DB2 UDB 8.2	SUSE LINUX Enterpris	IBM POWER5 1.
IBM	IBM eServer 325	12216	71	863410	US $	IBM DB2 UDB 8.1	Suse Linux Enterprise	AMD Opteron 2 0
		TPC-H Results - Revision 1.X - 300GB Scale Factor						
Company	System	QphH	Price Perf.($/Q	Total Sys. Cost	Currency	Database Software	Operating System	CPU Type
Sun	SunFire V240	1027	49	50297	US $	Sybase IQ 12.5	Sun Solaris 9	Sun UltraSPARC
Sun	SunFire V250	1283	27	34600	US $	Sybase Sybase IQ 12.5	Sun Solaris 9	Sun UltraSPARC
Sun	SunFire V440	3091	40	122244	US $	Sybase IQ 12.5	Sun Solaris 9	Sun UltraSPARC
HP	HP ProLiant DL760G2 8P	3335	71	235058	US $	Microsoft SQL Server 2000 Enterp	Microsoft Windows Ser	Intel Xeon MP 2.0
HP	HP ProLiant DL 760G2-8P	4064	43	172288	US $	Microsoft SQL Server 2000 Enterp	Microsoft Windows Ser	Intel Xeon 2.8GH
Unisys	Unisys ES7000 Orion 130 Enterp	4774	207	988328	US $	Microsoft SQL Server 2000 Enterp	Microsoft Windows .NE	Intel Itanium2 1 0
IBM	IBM eServer xSeries 365	5003	50	248778	US $	IBM DB2 UDB 8.1	Microsoft Windows Ser	Intel Xeon MP 3.0
IBM	IBM eServer xSeries 365	5090	45	222654	US $	IBM DB2 UDB 8.2	SUSE LINUX Enterpris	Intel Xeon MP 3.0
IBM	IBM eServer xSeries 445 8P	6552	66	431836	US $	IBM DB2 UDB 8.1	Microsoft Windows Ser	Intel Xeon MP 3.0
Dell	PowerEdge 6600/3.0Ghz/4MB Cl	6795	42	284858	US $	Oracle Database 10g Enterprise E	Red Hat Enterprise Lin	Intel Xeon MP 3.0
SGI	SGI Altix 3700 8P	8828	85	751939	US $	IBM DB2 UDB 8.2	SUSE LINUX Enterpris	Intel Itanium2 1.5
IBM	IBM eServer OpenPower 720 w/D	12007	40	481401	US $	IBM DB2 UDB 8.2	SUSE LINUX Enterpris	IBM POWER5 1.
IBM	IBM eServer 325	13195	65	863410	US $	IBM DB2 UDB 8.1	Suse Linux Enterprise	AMD Opteron 2 0

Problème 2 (10 points): Analyse de données de satisfaction clientèle

Question 1.
On considère six graphiques confrontant chacun les durées de voyage et de retard par compagnie aérienne.



Attribuez à chaque compagnie le degré de liaison (coefficient de corrélation) entre la durée des vols et les retards occasionnés figurant parmi l'ensemble {-1, -0.73, -0.49, -0.04, 0.33, 0.5, 0.74, 1}.
On considère le tableau suivant généré à partir de l'entrepôt de la partie 1:

Clients	Age	Cat-So-P	Canal d'achat	Prix du Voyage	Durée-Voyage	Durée Retard	Satisfaction Client
C1	25	E	I	600	360	120	NS
C2	32	C	C	300	180	20	S
C3	58	C	I	120	120	60	S
C4	62	R	I	800	300	65	MS
C5	75	R	A	1500	700	15	NS
C6	28	E	A	400	210	30	MS
C7	43	C	C	320	210	45	NS

Le rapport ci-dessus décrit les profils d'un ensemble de vols effectués ainsi que les degrés de satisfaction des clients vis-à-vis du service offert.

Type et domaine des attributs :

- *Attributs quantitatifs*
  Age : entier sur [0-150]
  Prix du voyage: Réel
  Durée-Voyage : Réel
  Durée-Retard: Réel

- *Attributs qualitatifs ordonnés*
  Satisfaction: NS (Non Satisfait), MS (Moyennement Satisfait), S (Satisfait)

- *Attributs qualitatifs non ordonnés*
  Cat-So-Pr : E (Etudiant), C(Cadre), R (Retraité)
  Canal d'achat : I (Internet), A (Agence), C (directement auprès de la compagnie)

Question 2.

a)- Après avoir procédé aux codages adéquats, évaluez la distance entre les profils des clients C1 et C4.
b)- Evaluez le profil central des clients C1 et C6. Quel problème cela pose. Proposez une solution pour l'évaluation du profil central.

Question 3.

On souhaite extraire les principaux profils de vols en se fondant uniquement sur les durées de voyage et les durées de retard. En procédant à une classification hiérarchique de lien Moyen (Mean-LINK) déterminez :

a) le nombre adéquat de profils,
b) l'ensemble des clients constituant chaque profil
c) la caractérisation d'un de ces profils (de cardinalité >1) dans l'espace de description initial.

Question 4.

Les responsables de la compagnie aérienne nous communiquent quelques informations concernant les statistiques des vols :

- un vol est dit de longue durée s'il porte sur une durée de plus de 3h (strictement),
- un retard par rapport à l'heure d'embarquement de plus de 45mn (strictement) est notifié long,
- le prix de base est dit élevé s'il est supérieur (strictement) à 400 euros.

Après avoir procédé aux codages adéquats, appliquez un arbre binaire de décision afin d'extraire les principales règles pouvant expliquer les degrés de satisfaction des clients.