Université Joseph Fourier
ISTG

Année Universitaire 2001-2002
RICM 3
Systèmes d'Information Décisionnels
Examen

Concepteurs : Ahlame DOUZAL, Didier DONSEZ
Durée : 2 heures
Remarques : Calculatrice et tout document autorisé
Conseil : Lire le sujet jusqu'au bout.

Exercice 1: Modélisation décisionnelle d'un Entrepôt de Données pour le suivi des cybernautes d'un site marchand

Un site marchant sur Internet souhaiterait mieux connaître sa clientèle afin de mieux cibler les offres promotionnelles en fonction du profil du client (visiteur) qui apparaissent dans les bannières des pages que consultent les clients. Un des buts ultimes de cette connaissance est le JIT (Just In Time) ECR (Efficient Customer Response) : c'est à dire l'adaptation temps réel du contenu des pages retournées au visiteur pour maximiser le terminaison d'une visite par un achat.
Suivi de lien sur un site Web après un mailing personnalisé

Pour cela, la société qui gère le site marchant souhaite mettre en place un entrepôt de données et sa réalisation vous est confiée.
L'entrepôt est alimenté (en information) à partir des journaux du serveur Web et du système de prise de commande.
Les journaux contiennent la liste des requêtes HTTP. Chaque entrée contient :

- La Date et heure de la requête de la requête
- L'Adresse IP du visiteur
- Le cookie ID identifiant une session d'un visiteur
- La page, le document ou le script demandé
- Le type de la requete (GET ou POST)
- L'URL de la page à partir de laquelle le visiteur est entré sur le site (par exemple depuis du page de résultat de recherche d'un moteur de recherche comme Google, d'une *newsletter*, …)
- Le navigateur (agent) utilisé (usuellement Netscape ou Internet Explorer).

Ces informations sont trop brutes pour être utilisé : elles sont complétées par les informations trouvées dans le système de prise de commande.

Le schéma de l'entrepôt est constitué des bases suivantes :

Session(SessionKey,TypeSession, ComportementVisiteur…)
TypeSession peut etre « SessionSansAchat », « SessionAvecPaiement »,…
Date(DateKey, Année, Mois, JourDeMois, JourDeSemaine, TrancheHoraire, DrapeauVacances, …)
Visiteur(VisiteurKey, AdresseIP, Nom, Prenom, FuseauHoraire, …)
Produit(ProduitKey, Designation, Couleur, TypeProduit, …)
Page(PageKey, ProfondeurDepuisLaRacine,TypePage…)
TypePage peut etre « Information », « Formulaire »,…
ProfondeurDepuisLaRacine représente le nombre minimum de pages à parcourir depuis la racine du site (www.sitemarchand.com) pour arriver jusqu'à la page.
Référent(ReferentKey, URL, TypeReferent, …)
TypeReferent peut etre « Moteur de recherche Public», « Magazine en ligne », « Bannière sur un site sponsorisé », « URL Entrée Manuellement Ou Bookmark», « Interne », « Moteur De Recherche Interne », « Newsletter »
Promotion(PromotionKey, TypePromotion, …)

Requete(SessionKey, Date, VisiteurKey, ProduitKey, PageKey, ReferentKey, PromotionKey, VisiteId, NombrePagesTraverséesAvantLaPage, NombreSecondesPasséesDansLaPage, MontantAchétéDansLaPage)

NombrePagesTraverséesAvantLaPage est le nombre de pages du site marchant que le visiteur a traversées avant cette page. NombrePagesTraverséesAvantLaPage=1 signifie que c'est la première page par laquelle le visiteur est arrivé sur le site marchand (il arrive à partir d'un référent qui peut être par exemple un moteur de recherche).

MontantAchétéDansLaPage peut être positif, négatif (retrait d'un produit du caddie), zéro (pas d'achat).

Rappel

Ex1: Rappelez le principe de chargement de l'entrepôt de données en commentant de 10 lignes un schéma.

- Voir cours

Ex2: Quelle est la table de fait dans cet entrepôt ?. Justifiez !

- Requete (contient des clés étrangères référençant les tables de dimension

Ex3: Pourquoi la visite n'est pas une dimension (on trouve VisiteId dans Requete) ?. Justifiez !

- c'est une dimension dégénérée !

Dimensionnement

Ex4: A partir des informations suivantes,

Nombre de visiteurs par jour	200 000
Nombre de requêtes par visite	10
Ration de visiteurs ayant déjà fréquenté le site	0,3
Ratio d'achats par visite	0,1
Nombre de jours	1200
Nombre de tranches horaires	48

Donnez le nombre d'enregistrements de la table de fait.

- Nombre d'enregistrements = 200000 * 10 * 1200 = 2 400 000 000

Donnez la taille des attributs et des clés ?

- 4 octets pour les clés et les attributs

Donnez la taille d'un enregistrement de la table de fait ?

- 44 octets

Donnez la taille (en Octets) de stockage de la table de fait?

- 89,40 Go

Donnez la taille d'un index bitmap de la table de fait sur la colonne « TypeReferent »?

- La cardinalité de TypeReferent est 7
- Taille = 2 400 000 000 * 7 / 8= 1,9 Go

Configuration Matérielle

Ex5: A partir des résultats du benchmark TPC/H (http://www.tpc.org/tpch/results/tpch_results.xls), choisissez la configuration matérielle et logicielle qui est la plus adaptée à votre infocentre pour une performance minimale de 1200 QphH ? Quels sont vos critères de choix ? Remarque : vous négligerez la taille des tables de dimensions.

- Compaq ProLiant 8000 à 1308 QphH (174 $QphH) 228104 US$ Microsoft SQL 2000 Microsoft Windows 2000 Intel Pentium III Xeon 550 MHz

Rapports

Ex6: Donnez la requête SQL qui donne la moyenne du nombre de pages parcourues lors d'une visite dans une session.

```
SELECT AVG(NbPagesVisite)
FROM(
  SELECT Count(*) AS NbPagesVisite
  FROM Requete
  GROUP BY VisiteId
)
```

Ex7: Donnez la requête SQL qui donne la montant moyenne des achats dans les sessions du type « SessionAvecAchat ».

```
SELECT AVG(MontantTotalVisite)
FROM(
  SELECT VisiteId, SUM (MontantAchétéDansLaPage) AS MontantTotalVisite
  FROM Requete NATURAL JOIN Session
  WHERE TypeSession=”SessionAvecAchat”
  GROUP BY VisiteId
)
```

Ex8: Donnez la requête SQL qui donne la répartition des ventes en fonction du type du référent de la première page de la visite.

```
SELECT TypeReferent, SUM(MontantTotalVisite)
FROM(
  SELECT VisiteId, SUM (MontantAchétéDansLaPage) AS MontantTotalVisite
  FROM Requete NATURAL JOIN Session
  WHERE TypeSession=”SessionAvecAchat”
  GROUP BY VisiteId
) AS M JOIN (
  SELECT VisiteId, TypeReferent
  FROM Requete NATURAL JOIN Referent
  WHERE NombrePagesTraverséesAvantLaPage =1
) AS T USING (VisiteId)
GROUP BY TypeReferent
```

Conception

Ex9: Donnez le schéma de la table de fait d’un second entrepôt dont un fait (enregistrement) représente une visite avec sa durée totale, le montant totale des achats et le nombre de requêtes effectuées

Ex10: Donnez la taille en octets de cette table.

Problème 2 : Analyse des profils de navigation d'un site marchand (10 pts)

On considère le tableau de données suivant issues de l'entrepôt pour le suivi des cybernautes d'un site marchand :

Visite	Nb Page	Nb Click	Produit	Durée	Action
S1	10	2	O	2	A
S2	4	9	M	7	N
S3	12	3	M	3	C
S4	5	7	F	10	N
S5	3	10	M	13	N
S6	7	3	O	10	C

Ce tableau donne la description des visites utilisateurs par : le nombre de pages visitées (Nb Page), le nombre de clicks par page (Nb Click), le produit consulté (Produit) de type 'O' pour ouvrage, 'M' pour musique ou 'F' pour film, on admet qu'il n'y a qu'un seul produit consulté par visite, la durée moyenne de navigation par page en minutes (durée), et l'action commerciale de type 'A' pour achat, 'C' pour commande ou 'N' pour annulation.

Q1- On se positionne dans l'espace tri-dimensionnel défini par les attributs « NbPage », « NbClick » et « Produit ». Lesquelles des visites S1, S2 et S3 sont les plus similaires. Donner la description, dans ce même espace, du nuplet S123 centre de S1, S2 et S3. Quel est le problème rencontré ?. Comment y remédier ?

Q2- On se positionne dans l'espace défini par les dimensions « Nb Page », « Produit », « Durée » et « Action ». Les architectes du sites marchand considèrent qu'un temps moyen de navigation supérieur à cinq minutes ou un nombre de pages visitées supérieur à six nuit à l'action d'achat ou de commande. Quels technique(s) et/ou codage(s) doit-on utiliser afin de vérifier la validité de cette assertion ?. Justifez.

Q3- On se positionne dans l'espace défini par les dimensions « Nb Click » et « Durée ».
On souhaite partitionner l'ensemble des visites en trois groupes, chacun caractérisant le profil de navigation au sein du site marchand. Utiliser la méthode des moyennes mobiles afin d'extraire ces trois groupes. On considère S1, S3 et S6 comme nuplets-centres de départ.

Q4- Afin d'augmenter la rentabilité du site marchand, on souhaite pouvoir prédire avec une probabilité p (à préciser) l'action d'achat, de commande ou d'annulation en fonction du nombre de clicks (seuil = 6), du type de produit consulté et de la durée moyenne de navigation par page(seuil = 5). Utiliser la méthode appropriée afin de répondre à cet objectif.