M1 INFO 2012/2013 – FDD – TP1

Entrepôts de données

Afin de pouvoir répondre aux questions suivantes, commencez par lire le tutoriel des commandes SQL et SQL étendues sous Oracle.

Exercice 1 : Rappels SQL et requêtes de classement sous Oracle

Les questions suivantes se basent sur les tables *Emp* et *Dept* créée à partir du script "exo*1.sql*".

Question 1.1 : Quel est le classement des salaires des employés par département pour les départements 10 et 30 ?

DEPTNO	ENAME	SAL	RANG
10	KING	5000	1
10	CLARK	2450	2
10	MILLER	1300	3
30	BLAKE	2850	1
30	ALLEN	1600	2
30	TURNER	1500	3
30	MARTIN	1250	4
30	WARD	1250	4
30	JAMES	950	6

Résultat souhaité

Question 1.2 : Idem en ôtant les trous dans le classement.

DEPTNO	ENAME	SAL	RANG
10	KING	5000	1
10	CLARK	2450	2
10	MILLER	1300	3
30	BLAKE	2850	1
30	ALLEN	1600	2
30	TURNER	1500	3
30	MARTIN	1250	4
30	WARD	1250	4
30	JAMES	950	5

Résultat souhaité

Question 1.3 : Quel est le classement décroissant des salaires différents par département pour les départements 10 et 20 ?

DEPTNO	SAL	RANG
10	5000	1
10	2450	2
10	1300	3
20	3000	1
20	2975	2
20	1100	3
20	800	4

Résultat souhaité

Question 1.4 : Quel est le salaire total versé par profession ? Répondre avec deux méthodes possibles, avec et sans *group by*.

JOB	TOT_SAL_JOB
CLERK	4150
SALESMAN	5600
PRESIDENT	5000
MANAGER	8275
ANALYST	6000

Résultat souhaité

Question 1.5 : Quel est la différence entre un *group by* et un *partition by* dans une requête SQL ?

Question 1.6 : Quel est le montant total des salaires versés, tout département et job confondus, par département et par département et job ?

DEPTNO	JOB	SUM(SAL)
10	CLERK	1300
10	MANAGER	2450
10	PRESIDENT	5000
10	(null)	8750
20	CLERK	1900
20	ANALYST	6000
20	MANAGER	2975
20	(null)	10875
30	CLERK	950
30	MANAGER	2850
30	SALESMAN	5600
30	(null)	9400
(null)	(null)	29025

Résultat souhaité

Question 1.7 : Idem en ôtant toute confusion avec les valeurs nulles. Répondre avec deux méthodes possibles.

DEPARTEMENT	JOB	SUM(SAL)
10	TousEmployés	8750
10	PRESIDENT	5000
10	MANAGER	2450
10	CLERK	1300
20	TousEmployés	10875
20	ANALYST	6000
20	MANAGER	2975
20	CLERK	1900
30	TousEmployés	9400
30	SALESMAN	5600
30	MANAGER	2850
30	CLERK	950
TousDep	TousEmployés	29025

Résultat souhaité

Exercice 2 : Conception et alimentation d'un entrepôt de données

Etat de l'existant: Une BD transactionnelle

Soit un extrait d'une base de données transactionnelle servant à la gestion de factures dans le cadre de la vente de produits alimentaires pour une enseigne donnée. Analyser cette BD à partir du modèle relationnel donné ci-dessous :

- **Client** (Num, Nom, Prenom, Adresse, Date_nais, Sexe)
- **Produit** (Num, Designation, Stock)
- **Prix_date** (Num, Produit=>Produit, date, prix, remise)
- **Facture** (Num, Client=>Client, date_etabli)
- **Ligne_facture** (Facture=>Facture, Produit=>Produit, Qte, Id_prix=>Prix_date)

Il est à noter que pour chacune des tables ci-dessus, la clé primaire est soulignée et les clés étrangères sont indiquées par des flèches de référencement. La définition de ces différentes tables est donnée dans le script "exo2.*sql*".

Modélisation : Etat des besoins

La première étape pour lancer un projet décisionnel est de faire l'inventaire des questions qui se trouvent sans réponse à travers l'ensemble de l'entreprise. En effet, il est fréquent qu'un projet décisionnel avec un périmètre très restreint, par exemple, uniquement pour la force de ventes, fonctionne bien dans un premier temps. Mais ses extensions n'ayant pas été prévues au départ, plusieurs choix techniques bons pour un petit projet, s'avèrent être aujourd'hui des handicapes majeurs à l'évolution du système. En moyenne, le volume des données double tous les deux ans. On parle donc souvent de «scalability », soit la capacité de monter en puissance d'une plateforme décisionnelle.

Une bonne phrase pour synthétiser la méthodologie d'un projet décisionnel est : « voir grand, mais commencer petit ».

Dans le cadre de ce projet, on cherche à analyser les ventes effectuées afin de les faire croître. Les caractéristiques intéressantes des ventes sont les prix et les quantités. On s'intéresse à des critères géographiques (où sont les clients qui achètent ?) afin de cibler des campagnes promotionnelles. La précision des analyses est variable, mais on notera que l'entreprise a à la fois une vocation locale (échelle de la ville) et internationale (échelle du pays). On s'intéresse à des critères temporels (quand se passent les achats ?) afin d'aider à

l'organisation de l'entreprise (période de vacances, embauches saisonnières supplémentaires, etc.). La précision de ces critères est peu exprimée, on cherchera à adopter des critères classiques.

Voici les principales questions auxquels nous essayons de répondre, sachant que dans le cas d'un Data Warehouse, l'enjeu est de proposer des tables sur les données qui dépassent les simples questions formulées, pour faire apparaître des relations non encore envisagées par les utilisateurs.

- Quels sont les produits les plus vendus, selon leurs désignations et catégories ?
- On s'intéresse aux caractéristiques des clients, pour analyser ce qu'ils achètent en fonction de leur âge, de leur groupe d'âge et de leur sexe
- Quels sont les chiffres d'affaire en fonction des jours, semaines et année (Quel est le chiffre d'affaire du jour 254, de la semaine 42 et de l'année 2003) ?
- Est ce qu'il existe une relation entre le chiffre d'affaire, les mois de l'année et les sexes des acheteurs (par exemple est-ce que les femmes achètent plus en novembre) ?
- Y a-t-il une relation entre le temps, l'espace et la vente de produit ?
- Quels sont les trois produits les plus vendus en général, et par catégorie ?
- Combien se classerait dans le top des ventes toutes catégories confondues un produit vendu à 50 exemplaires ?
- Quels sont les produits qui contribuent à moins de 0.05% du CA pour un Pays/une année donné(e)?
- Est-ce que ces produits peuvent être remisés ?
- Quels sont les produits qui sont achetés ensemble ? Par exemple afin de les rendre plus proches sur le site de vente ?
- Quelle est la tendance des ventes pour l'année à venir ?
- Est-ce que les remises font augmenter les ventes ?
- etc.

Parmi l'ensemble des questions récoltées, une première sélection devra spécifier quelles sont celles dont le ressort est bien de l'informatique décisionnelle. Dans notre cas, certains commerciaux souhaitaient savoir si une remise donnée a été prise en compte dans la les dernières commandes de leurs clients. C'est une problématique opérationnelle.

Remarque : Nous avons ci-dessus une liste de questions non structurées, mal posées, pas claires. Il serait un comble de les considérer comme un cahier des charges. C'est juste une petite liste de questions, ce n'est pas un cahier des charges.

La deuxième étape de notre projet décisionnel est d'identifier les sources d'informations, internes et externes à l'entreprise permettant de construire le Data Warehouse qui répondra aux questions précédemment soulevées.

Question 2.1 : Proposez un modèle dimensionnel (schéma en étoile avec une table de faits et des tables de dimensions) de Data Warehouse capable de répondre aux besoins des utilisateurs. Représenter, sur un diagramme conceptuel, les données de l'entrepôt sous la forme du schéma en étoile proposé. Enoncez les hiérarchies pertinentes pour chaque dimension.

Implémentation et alimentation

Une fois la conception de votre entrepôt de données terminée, vous allez travailler maintenant sur l'intégration des données des différentes sources dans votre entrepôt de

données centralisé ; fonction permise par les outils d'Extraction /Transformation / Loading (ETL).
L'intégration est en fait un pré-traitement ayant pour but de faciliter l'accès aux données centralisées aux outils d'analyse et de restitution par la suite pour la prise de décision. Ainsi, l'intégration consiste à concentrer les données collectées dans un espace unifié, dont le socle informatique essentiel est l'entrepôt de données. Ce dernier est l'élément central du dispositif dans le sens où il permet aux applications d'aide à la décision de bénéficier d'une source d'information homogène, commune et fiable. Cette centralisation permet surtout de s'abstraire de la diversité des sources de données. Lors de cette étape les données sont transformées et filtrées en vue du maintien de la cohérence d'ensemble. Enfin, c'est aussi durant cette étape que sont effectués les éventuels calculs communs à l'ensemble du système d'information décisionnel (SID).
Sur le plan conceptuel, un entrepôt de données (fait et dimensions) peut être vu comme un ensemble des tables provenant d'un ensemble des sources sous-jacentes. De ce fait, les changements au niveau des sources doivent être répercutés périodiquement au niveau de l'entrepôt. La maintenance de l'entrepôt entraîne donc la consultation des sources sous-jacentes. La périodicité de mise à jour des tables est fonction des besoins des données sur le serveur dédié à l'analyse. Dans ce projet et pour mieux comprendre les fonctionnalités du processus ETL, vous n'allez pas utiliser des outils dédiés pour ça mais vous allez plutôt développer votre propre script SQL pour y parvenir. La solution ETL à développer dans ce projet consiste donc à utiliser les tables sous oracle.

Indications
Les éléments suivants peuvent vous aider à l'implémentation du processus ETL :

- La désignation d'un produit comporte son nom, sa catégorie suivie de sa sous-catégorie séparées les 3 par des ".". Si la sous-catégorie est absente, il faut la mettre à NULL dans l'entrepôt.
- L'adresse d'un client comporte son pays, son code postal et sa ville. Les deux premières lettres du code postal donnent la département.
- La fonction substr(X, A, B) renvoie les B caractères à partir du caractère A dans la chaîne X.
- La fonction instr(chaîne, sous-chaîne [,début [,nombre occurrences]]) recherche la position d'une sous-chaîne dans une chaîne. (Plus de détails dans le tutoriel joint à ce TP)
- Pour la gestion des dates, utilisez les fonctions to_date et to_number.
- La requête suivante permet de créer une liste de dates entre date_a et date_b

```
select level + date_a- 1 as date
from dual
connect by level < (date_b – date_a + 2)
```

- Pour le calcul du groupe d'âge pour chaque client utilisez le format suivant : Si l'âge est inférieur à 30, il écrit la chaine de caractère « <30 ans ». Si l'âge est compris entre 30 et 45, il écrit la chaine de caractère « 30-45 ans ». Si l'âge est compris entre 45 et 60, il écrit la chaine de caractère « 46-60 ans ». Sinon, il écrit la chaine de caractère « >60 ans ». La structure de contrôle **case when…then…end** dans le select de votre requête peut faire l'affaire.

Question 2.2 : Ecrivez les requêtes SQL permettant le transfert/transformation des données depuis la base transactionnelle vers le Data Warehouse.

Question 2.3 : Créez si nécessaire les différentes clés primaires et étrangères pour les différentes tables.

Exercice 3 : OLAP/ROLAP sous Oracle

Dans le monde de l'informatique décisionnelle, le ROLAP (Relational On-Line Analytical Processing) est une solution OLAP qui voit un entrepôt de données comme un SGBD relationnel et qui permet de l'interroger en utilisant des requêtes SQL étendu (en général très complexes et très exigeantes en terme de ressources et de temps d'exécution).
Une requête ROLAP est en général exprimée comme suit.

- On calcule le joint de la table des faits et des relations dimensionnelles.
- On sélectionne des tuples en fonctions des données dimensionnelles.
- On groupe ces données suivant certaines dimensions.
- On calcule une valeur agrégée (le plus souvent une somme).

Ce type de requêtes est décrit d'avantage dans le tutoriel joint à ce TP.
Répondre aux questions suivantes par des requêtes. La définition des différentes tables est donnée dans le script "exo3.*sql*".

Question 3.1 : Est-il nécessaire de modifier le script ? Si oui, le modifier avant de l'exécuter.

Question 3.2 : Donner la moyenne des ventes pour les années 2009 et 2010

- par année, pays (du client), catégorie (du produit)
- par année, pays
- par année

Question 3.3 : Donner la moyenne des ventes par année pour les années 2009 et 2010, selon les dimensions pays et catégorie.

Question 3.4 : Quel est le produit le plus vendu par année et par catégorie (cf. *RANK*) ?

Question 3.5 : Pour chaque année, donner le total des ventes ainsi que le total des ventes par catégorie. On ne veut pas le résultat pour le total des années (cf. *GROUPING_ID*).

Question 3.6 : Donner le total des ventes par année selon la dimension catégorie, et selon la dimension client (cf. *GROUPING_SETS*).

Question 3.7 : Quel est le meilleur mois de vente du produit "Sirop d'érable" pour chacune des années ?

Question 3.8 : Pour chaque catégorie, quelle est la quantité de produits vendus les 5 premiers jours de chaque mois de l'année 2010 ?

Question 3.9 : Quelle est la répartition par tiers des catégories selon leurs quantités totales vendues en 2010 (cf. *NTILE*) ?