Mustapha NEZZARI

Dans ce compte rendu, vous trouverez les réponses au TP2.

Compte rendu du TP2

SQL OLAP

Sommaire

[Modélisation dimensionnelle 4](#_Toc467424200)

[L’entrepôt de données 5](#_Toc467424201)

[La table des faits Vente 5](#_Toc467424202)

[Les mesures 5](#_Toc467424203)

[Script SQL de la vue matérialisée 5](#_Toc467424204)

[La dimension Produit 6](#_Toc467424205)

[Script SQL de la vue matérialisée 6](#_Toc467424206)

[Script SQL de la dimension 6](#_Toc467424207)

[Hiérarchie 6](#_Toc467424208)

[La dimension Client 8](#_Toc467424209)

[Script SQL de la vue matérialisée 8](#_Toc467424210)

[Explications 8](#_Toc467424211)

[Script SQL de la dimension 8](#_Toc467424212)

[Hiérarchie 8](#_Toc467424213)

[La dimension Lieu 9](#_Toc467424214)

[Script SQL de la vue matérialisée 9](#_Toc467424215)

[Explications 9](#_Toc467424216)

[Script SQL de la dimension 9](#_Toc467424217)

[Hiérarchie 9](#_Toc467424218)

[La dimension Temps 11](#_Toc467424219)

[Script SQL de la vue matérialisée 11](#_Toc467424220)

[Explications 11](#_Toc467424221)

[Script SQL de la dimension 12](#_Toc467424222)

[Hiérarchie 12](#_Toc467424223)

[Les vues matérialisées 13](#_Toc467424224)

[Mode de rafraichissement 13](#_Toc467424225)

[Tests 13](#_Toc467424226)

[Les clés primaires et étrangères 14](#_Toc467424227)

[Optimisations 15](#_Toc467424228)

[Les index 15](#_Toc467424229)

[Script SQL 15](#_Toc467424230)

[Explications 15](#_Toc467424231)

[Questions d’exploitation 16](#_Toc467424232)

[Question 1 16](#_Toc467424233)

[Script SQL 16](#_Toc467424234)

[Résultat 16](#_Toc467424235)

[Explications 16](#_Toc467424236)

[Question 2 17](#_Toc467424237)

[Script SQL 17](#_Toc467424238)

[Résultat 17](#_Toc467424239)

[Explications 17](#_Toc467424240)

[Question 3 18](#_Toc467424241)

[Script SQL 18](#_Toc467424242)

[Résultat 18](#_Toc467424243)

[Explications 18](#_Toc467424244)

[Question 4 18](#_Toc467424245)

[Script SQL 18](#_Toc467424246)

[Résultat 18](#_Toc467424247)

[Explications 19](#_Toc467424248)

# Modélisation dimensionnelle

Dans le cas présenté, l’entreprise cherche à analyser ses ventes selon des mesures différentes. Elle s’intéresse aux **quantités** vendues, au **prix** de ventes (avec éventuellement des **remises**) et aux **chiffres d’affaire**.

Elle souhaite aussi pouvoir avoir des informations sur les produits vendus comme leur **désignation**, leur **catégorie** et leur **sous-catégorie**.

Elle s’intéresse aussi aux caractéristiques de ses clients. Les questions qu’elle se pose montrent un intérêt pour leur **âge**, leur **tranche d’âge** et leur **sexe**. Afin de cibler ses campagnes promotionnelles, elle s’intéresse à des critères géographiques locaux et internationaux comme le **pays**, la **ville** ou encore le **code postal**.

Enfin, elle veut archiver le tout afin de pouvoir effectuer des analyses temporelles sur certaines **années**, **mois**, **semaines** et **jours**.

En tenant compte de toutes ces informations, j’ai réalisé le schéma en étoile Figure 1.

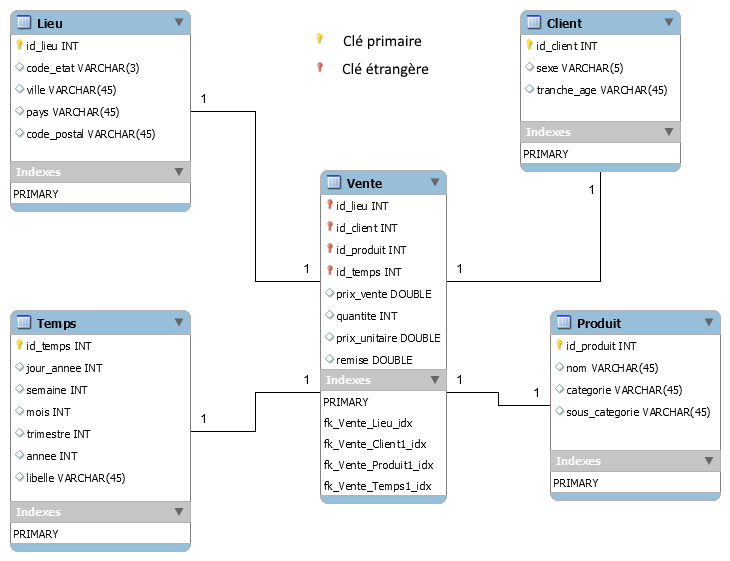


Figure 1

On peut y voir la table de fait Vente et les différentes dimensions qui représentent les axes possibles d’analyse à savoir Client, Lieu, Produit et Temps.

# L’entrepôt de données

Afin de créer cet entrepôt de données, j’ai utilisé les vues matérialisées. Il s’agit ici d’expliquer certains choix ainsi que certaines méthodes utilisées.

## La table des faits Vente

La table des faits Ventes va contenir une entrée par ventes d’un produit en stockant les références du client, du produit vendu, du lieu de vente ainsi que de la date.

### Les mesures

En plus des références précédentes, elle va permettre de stocker différentes mesures qui sont :

* La quantité du produit vendu
* Le prix du produit au moment de la vente
* La remise s’il y en a eu une
* Le prix total de ventes qui correspond à la formule : Quantité **x** Prix unitaire de vente à la date de vente **x** remise en pourcentage.
* L’âge du client au moment de la vente. J’ai décidé de ne pas mettre l’âge directement dans la dimension Client car on aurait perdu cette notion d’archivage. De plus je pense qu’il est plus pertinent de connaître l’âge de la personne au moment de la vente car l’achat peut être influencé par l’âge et donc la mentalité et le contexte. Par exemple, on pourrait croire qu’une publicité a plus touché les personnes jeunes plutôt que les personnes âgées et l’on pourrait alors le savoir en mettant cette donnée dans la table des ventes.

### Script SQL de la vue matérialisée



Dans cette requête, la jointure avec la vue Temps se fait en reconstruisant de la même manière qu’il que dans la dimension temps.

L’âge correspond à la formule : date de la facture **–** date de naissance du client.

## La dimension Produit

Cette dimension représente un produit à vendre.

### Script SQL de la vue matérialisée

Explications

Dans le SIO on dispose de la table Produit qui possède trois attributs :

* Sa référence (qui ne changera pas dans le SID)
* Sa désignation qui est composé du nom du produit, de sa catégorie et, s’il y en a une, de sa sous-catégorie.
* Son stock

Il a donc fallu découper la désignation pour séparer le nom, la catégorie et la sous-catégorie grâce aux regex. Ici j’ai utilisé la regex '[^.]+' qui permet de dire « un ou plusieurs caractères excepté le point.

La fonction NVL () a permis d’insérer une valeur nulle lorsqu’il n’y a pas de sous-catégorie.

Dans la dimension produit je n’ai pas mis le stock car je pense que pour les questions qui permettent de savoir comment améliorer les ventes, on n’a pas besoin de connaître le stock du produit dans le temps. On pourrait le rajouter plus tard si l’on veut mais je pense que c’est une donnée « inutile » puisqu’on a déjà les quantités vendues.

### Script SQL de la dimension



### Hiérarchie

La hiérarchie choisi (du plus grand au plus petit niveau) est : Catégorie > Sous-catégorie > Nom > id

## La dimension Client

Cette dimension représente un client qui a acheté au moins une fois.

### Script SQL de la vue matérialisée



### Explications

Les champs retenus dans la vue sont :

* La référence du client (la même que dans la table du SIO)
* Le sexe (idem)
* La tranche d’âge calculée à partir de la date maximum de vente pour le client. Il pourrait être plus performant d’utiliser la du jour (SYSDATE) plutôt que la date maximum des factures du client.

Le reste des champs de la table Client du SIO ont été utilisé dans la dimension Lieu car la localité est un axe d’analyse.

### Script SQL de la dimension



### Hiérarchie

La hiérarchie choisi (du plus grand au plus petit niveau) est : sexe > tranche d’âge > id

## La dimension Lieu

Cette dimension représente le lieu où une (ou plusieurs) vente a été effectuée.

### Script SQL de la vue matérialisée



### Explications

Cette vue utilise les informations de la table Client du SIO. Elle est composée :

* D’un id qui correspond à la référence client car chaque client a une référence différente et est associé à une adresse. Même en utilisant l’adresse complète pour l’id, on n’aurait pas pu garantir l’unicité des clés car si deux clients sont deux personnes mariées et habitent à la même adresse, il y aurait des conflits de clés. C’est pourquoi il est plus judicieux de prendre l’id du client comme id ici.
* Un code d’état
* Le pays
* La ville
* Le code postal

Il a fallu découper l’adresse pour séparer les trois derniers champs de la même manière que la désignation des produits.

### Script SQL de la dimension



### Hiérarchie

La hiérarchie choisi (du plus grand au plus petit niveau) est : Pays > Ville > Code postal > id

## La dimension Temps

Cette dimension est importante car elle permet l’historisation des données.

### Script SQL de la vue matérialisée



### Explications

Cette vue est constitué des champs suivants :

* L’id du temps qui est la concaténation du numéro du jour de l’année et de l’année (Par exemple : 365-2017)
* Du numéro du jour de l’année (de 1 à 365)
* Du mois (de 1 à 12)
* De l’année
* Du trimestre (de 1 à 4) que j’ai choisi de rajouter car on pourrait plus tard en avoir besoin
* Du libellé (Du Lundi au Dimanche) que j’ai aussi choisi de rajouter

Pour générer cette table, on utilise la sous requête avec la clause CONNECT BY qui permet de générer des dates entre deux dates. Le minimum et le maximum sont les dates de la plus ancienne et la plus récentes facture.

### Script SQL de la dimension



### Hiérarchie

La hiérarchie choisi (du plus grand au plus petit niveau) est : Année > Trimestre > Mois > Semaine > Jour de l’année > Libellé > id

## Les vues matérialisées

### Mode de rafraichissement

Selon moi, il n’est pas nécessaire de rafraichir les vues à chaque fois qu’il y a une entrée de fait dans le SIO car les données du SID seront utilisées ponctuellement et non de façon continue. J’ai donc décidé d’utiliser la méthode REFRESH FORCE ON DEMAND sur toutes mes vues, qui me permet de manuellement rafraîchir mes vues de façon incrémentale (si c’est possible, sinon toute la vue est rafraîchie).

Il faut donc utiliser les méthodes suivantes pour rafraîchir les vues à chaque fois que l’on veut s’en servir pour prendre des décisions :



Figure 2

### Tests

Afin de tester si mes vues fonctionnent, j’ai créé une procédure qui permet d’insérer 4 ventes dans les tables du SIO. Le script de cette procédure est disponible dans le fichier *procedure\_insert.sql*.

Il suffit d’appeler la procédure avec la commande « exec INSERER\_DONNEES; » et de rafraichir les vues grâce au commande de la Figure 2.

# Les clés primaires et étrangères

Dans notre cas, nous utilisons des vues matérialisée, il n’est donc pas nécessaires de créer des clés primaires puisqu’elles utilisent les clés des tables du SIO à part pour la table Temps où l’on génère nous même une clé. La création d’un index pourra apporter cette notion d’unicité (si l’on utilise un index unique).

# Optimisations

## Les index

### Script SQL



### Explications

Les index sont en général utilisés pour indexer des données (à l’aide de hashmaps par exemple) afin de les retrouver plus rapidement. On les utilise par exemple dans le cas où l’on a des jointures sur plusieurs tables sur lesquelles on a besoin de moins de 15% des données ou lorsqu’on utilise certaine colonnes dans des clauses WHERE.

Normalement, les vue matérialisées créent automatiquement des index sur les clés primaires du SIO. Cependant dans mon cas, seule celle sur la table produit est créé automatiquement. Je pense que c’est à cause des requêtes complexes pour créer les vues matérialisées que les index ne sont pas créés automatiquement pour les autres tables. J’ai donc dû les créer moi-même.

De plus j’ai créé un bitmap index pour le sexe du client car c’est un champ qui n’a que deux valeur possible ce qui satisfait les prérequis d’utilisation des bitmap index. J’ai fait de même pour la tranche d’âge qui n’a que quatre valeurs possibles.

Enfin je pense que comme nous voudrons faire des recherches selon toutes les granularités, on peut rajouter les champs suivants :

* Dans lieu : Ville, pays et code postal
* Dans produit : le nom, la catégorie et la sous-catégorie
* Dans temps : Tous les champs.

Cependant je ne le fait pas car je n’ai pas un jeu de données assez grand pour faire mes tests.

# Questions d’exploitation

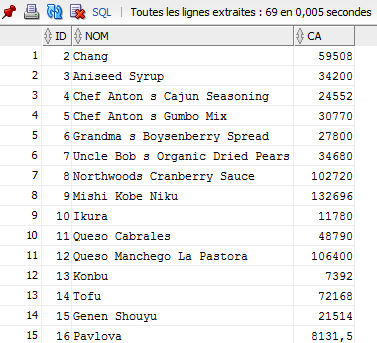
Les captures d’écran ont été prises en local car le VPN n’a pas voulu fonctionner lorsque j’ai voulu les faire sur ma base filora.

## Question 1

### Script SQL



### Résultat



### Explications

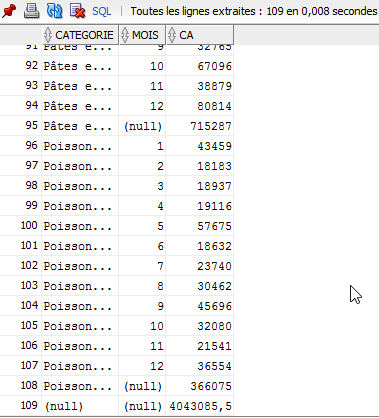
Cette requête donne le chiffre d’affaire par produit. On sélectionne simplement la somme des prix de ventes en groupant sur les prix produits. Bien sûr comme sur toutes les requêtes écrites jusqu’à présent, on fait les jointures nécessaires afin d’avoir les informations comme le nom etc ….

## Question 2

### Script SQL



### Résultat



### Explications

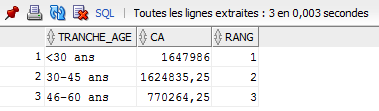
Cette requête donne le chiffre d’affaire par catégorie et mois, par catégorie et toute catégorie et mois confondus. Ceci est fait grâce à un rollup, comme on a maintenant l’habitude de faire.

## Question 3

### Script SQL



### Résultat



### Explications

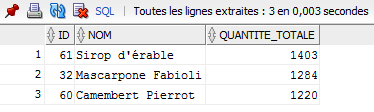
Cette requête donne le chiffre d’affaire par tranche d’âge, en donnant le rang de chaque tranche (1 pour celle qui a le plus grand chiffre d’affaire). Ici on utilise la fonction rank() qui va classer les chiffre d’affaire dans l’ordre décroissant. On affiche aussi le chiffre d’affaire grâce à la fonction SUM () et la clause GROUP BY.

## Question 4

### Script SQL



### Résultat



### Explications

Cette requête donne les 3 produits les plus vendus en quantité. Pour ce faire, on fait une sous requête ou l’on utilise la fonction rank () en classant les quantités vendues et en groupant ces quantités par produit. Ensuite la requête principale sélectionne les 3 premiers produits en filtrant sur les rangs entre 1 et 3.