# Entrepôt de données

**(ED)**

Exercice traité en cours

**Enoncé**

Une grande entreprise à succursales multiples veut rassembler toutes les nuits dans un **entrepôt de données** des informations sur les ventes du jour afin de dresser des tableaux de bord sur les ventes.

L'entreprise dispose d'un système d'information complexe, constitué des éléments suivants :

- des applications et bases de données éparses et hétérogènes sur les produits qu'elle vend,
- des applications et BD, également variées, sur les clients,
- idem sur les personnels de l'entreprise.

**L'ED à modéliser doit pouvoir fournir le chiffre d'affaires des ventes d'un produit, par date, client, et vendeur, ainsi que toutes les sommations possibles de chiffre d'affaires.**

Une vente correspond à un produit et un seul,
Les produits sont regroupés par famille de produits.
La vente est effectuée par l'un des vendeurs du service de vente spécialisé dans le produit.
La semaine de vente est le numéro de semaine dans l'année.

## 1. Modélisation de base

La table principale de la base de données « entrepôt de données » sera alors la suivante :

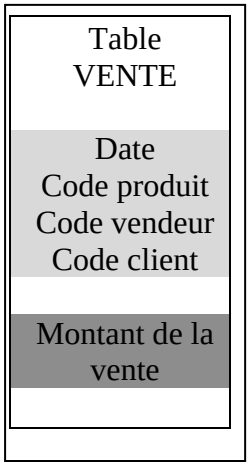


Figure 1 : Table VENTE

En grisé clair, apparaît la **clé multiple** de l'enregistrement, constitué de 4 éléments :
**date, code produit, code vendeur, code client**
En grisé foncé, figure la variable à mesurer, appelée **indicateur :**
**montant de la vente**

Cette table **VENTE** pourrait suffire à faire des sommations de chiffre d'affaires, si l'on se contentait des codes sur les éditions. C'est la table fondamentale d'un ED. On l'appelle :

**table de faits**

En fait, la plupart du temps, chacun des éléments de la clé multiple de la table de faits renvoie à un certain nombre d'attributs. Ici, par exemple, le code produit sera utilement complété par :

libellé du produit

code famille de produit

libellé famille

(nombreuses informations complémentaires possibles)

On fait alors une jointure entre la table de faits et une table dite table de dimension, selon le schéma suivant :

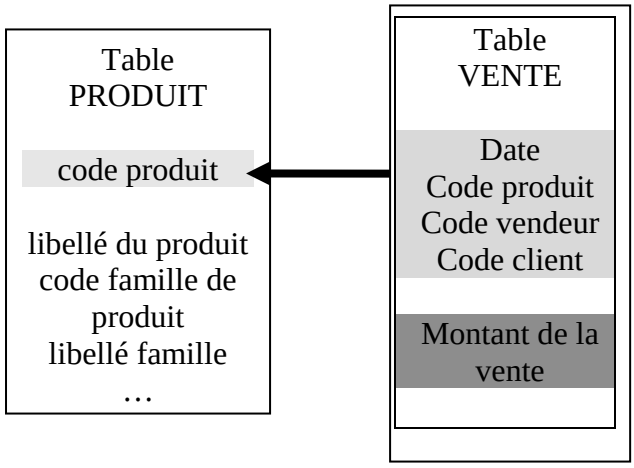


Figure 2 : Table de faits + 1 table de dimension

La table de dimension PRODUIT n'a pas d'indicateur, elle comporte seulement des attributs du produit. La clé « code produit » correspond à l'élément « code produit » de la clé multiple de VENTE.

De la même manière, les autres éléments de la clé multiple renvoient en général chacun à une table de dimension, selon le schéma suivant :

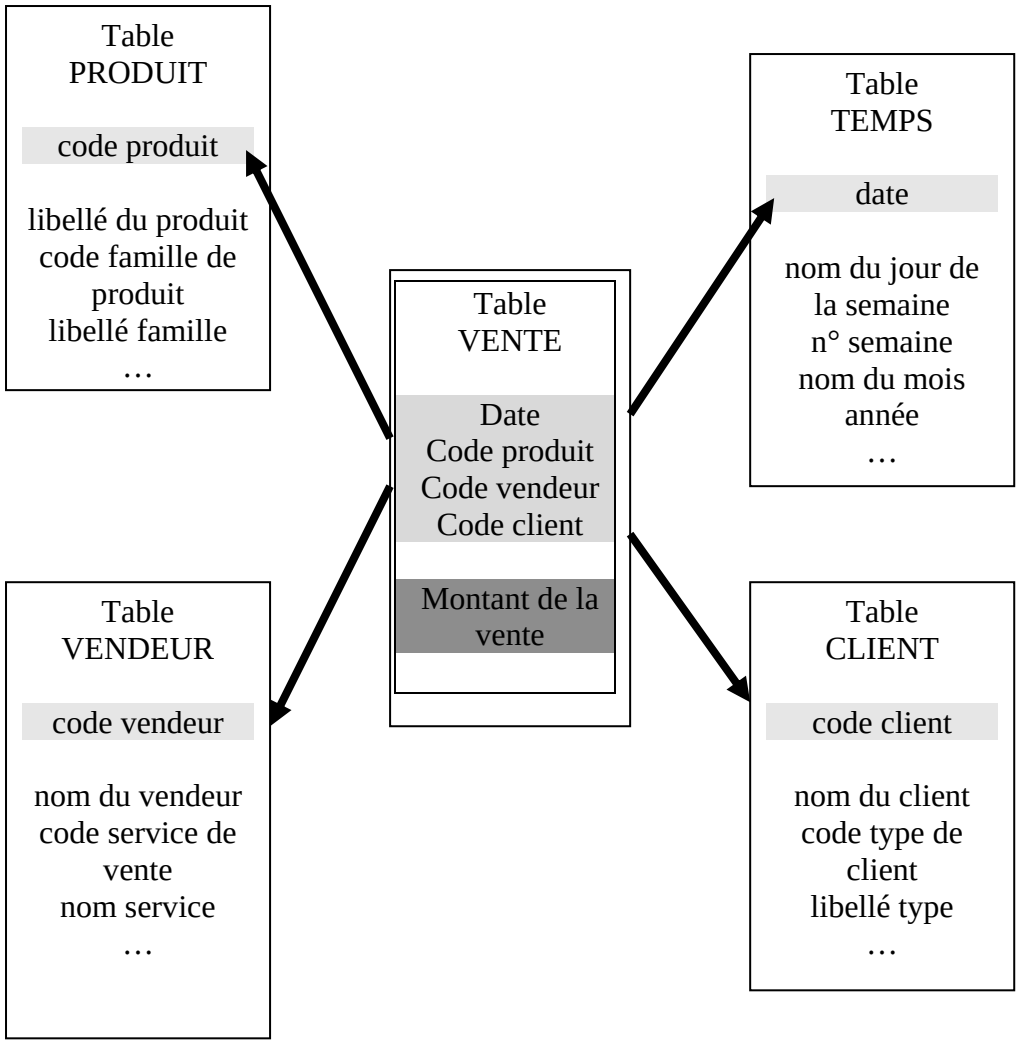


Figure 3 : Table de faits + 4 tables de dimensions

Ce schéma, avec au centre la table de faits, et autour les tables de dimension jointes, s'appelle :

**schéma en étoile**

Ce schéma est caractéristique de la **modélisation dimensionnelle** (du nom des dimensions) la plupart du temps mise en œuvre dans la conception d'un ED.

Un schéma en étoile peut également être représenté sous forme de

**cube de données**

A partir de la base de données relationnelle figurée par notre schéma en étoile, il est possible de développer un logiciel simple (à base de SQL par exemple) qui édite des sommes de « montant de la vente », ou chiffres d'affaires (CA).

Dans le tableau ci-dessous, les éléments sont les totaux des montants vendus toutes dates et clients confondus.

| Vendeur / produit | | | | Produit : ski | | |
|---|---|---|---|---|---|---|
| | | | | | | |
| Vendeur : Dupont | | | | 10500 | | |
| | | | | | | |

Le vendeur Dupont a vendu pour 10500 € de skis sur la période considérée

Figure 4 tableau à 2 dimensions

Ce tableau peut également être produit pour une date donnée (sélection sur la clé date). L'empilement de ces tableaux par date peut être figuré par le **cube** ci-dessous vu en perspective :

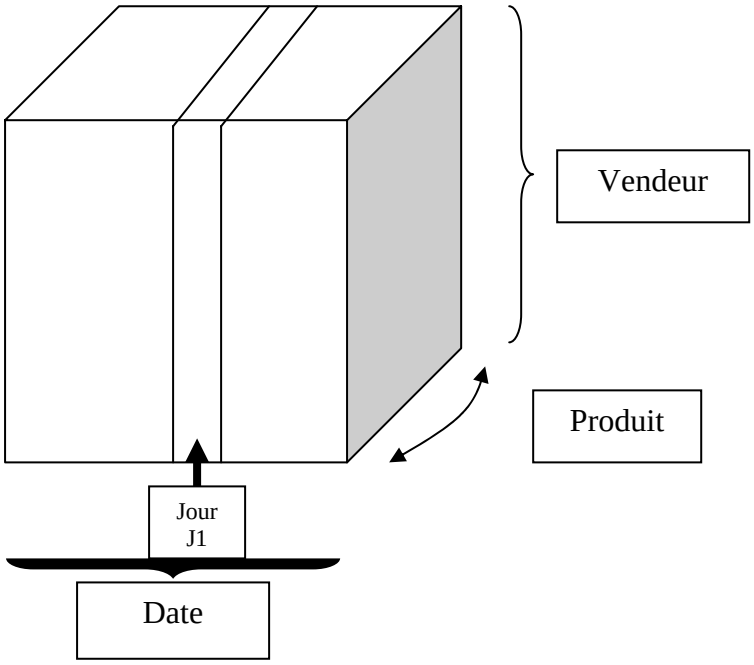


Figure 5 : cube de données à 3 dimensions

Lorsque le nombre de dimensions est de 3, la base de données peut être représentée par un **cube**. Lorsqu'il est supérieur à 3, comme c'est le cas dans l'exemple (qui comporte 4 dimensions), c'est un **hypercube**. Pour simplifier, on parle dans tous les cas d'un

**cube de données**

## 2. Coupes

Dans l'exemple traité, et représenté par le schéma en étoile (Figure 3), le cube de données est un hypercube à 4 dimensions : **produit, client, vendeur, date.**

Graphiquement, on peut dessiner en perspective 4 types de cubes à 3 dimensions :

| | | |
|---|---|---|
| A. | **client, vendeur, date** | (pour chaque valeur de produit) |
| B. | **produit, vendeur, date** | (pour chaque valeur de client) |
| C. | **produit, client, date** | (pour chaque valeur de vendeur) |
| D. | **produit, client, vendeur** | (pour chaque valeur de date) |

Dans chaque cube, l'élément de base est l'**indicateur** « montant de la vente ».

Tracé de cube D :

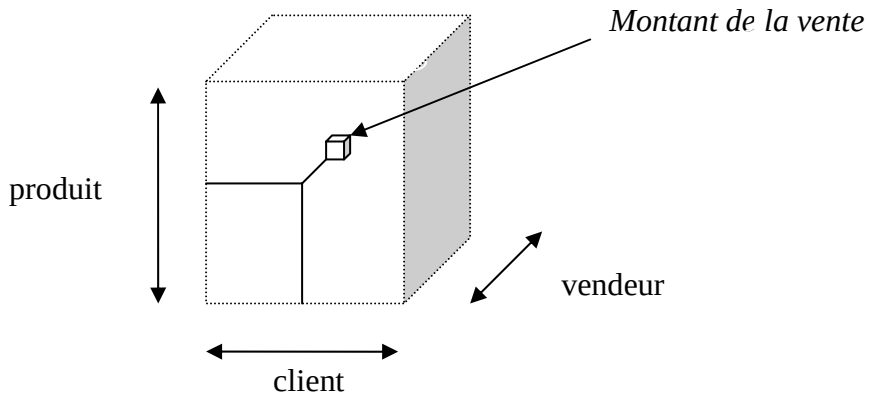


Figure 6 : un des types de cubes à 3 dimensions

On peut tracer autant de cubes D qu'il y a de valeurs pour la variable **date**.

Un cube D représente une **coupe** de l'hypercube à 4 dimensions, selon une valeur de la variable **date**.

De même, on peut faire des **coupes** du cube D pour toutes les valeurs de **produit**, par exemple. On obtient alors autant de tableaux à 2 dimensions (**client, vendeur**) qu'il y a de valeurs à **produit**.

A partir de D, on peut faire 3 types de tableaux à 2 dimensions :
(**client, vendeur**), (**produit, vendeur**), (**client, produit**)

A partir de l'ensemble A, B, C, D, on peut faire en plus les 3 coupes qui gardent **date**
(**client, date**), (**produit, date**), (**vendeur, date**),

**donc en tout 6 types de tableaux à 2 dimensions.**

# 3. Agrégats

Le but d'un entrepôt de données est la présentation de tableaux de bord. On a compris dans ce qui précède que lorsque le nombre de dimensions du cube de données est **n,** avec **n** supérieur à 2, il faut faire des **coupes** (en anglais **slice and dice**) en fixant les valeurs de **n-2** dimensions, pour se ramener à un tableau à 2 dimensions, donc affichable.

Plutôt que de couper, on peut aussi agréger les données, c'est-à-dire présenter un tableau à 2 dimensions en *sommant les valeurs* de certaines (voire toutes) des n-2 dimensions restantes.

C'est le cas du tableau vendeur – produit de la figure 4.

Lorsqu'on crée dans la base de données une table enregistrant ces sommations, on parle de **table d'agrégat**.

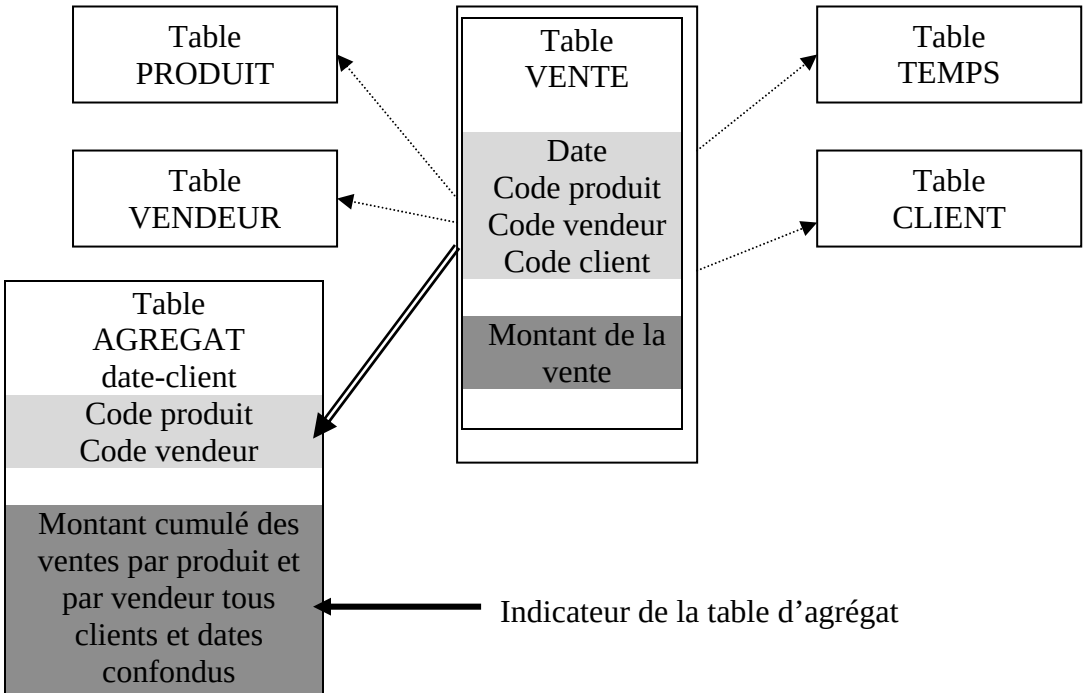


Figure 7 : création d'une table d'agrégat date-client

La création de tables d'agrégats a pour inconvénients :

- sur le plan conceptuel, de compliquer le modèle dimensionnel
- sur le plan technique, de multiplier l'espace de stockage sur disque

En revanche, le résultat d'une requête d'un cumul est obtenu plus rapidement lorsque la table d'agrégats existe.

# 4. Analyse multi-dimensionnelle

Reprenons la figure 4, où les éléments sont les totaux des montants vendus toutes dates et clients confondus, avec en lignes les vendeurs et en colonnes les produits.

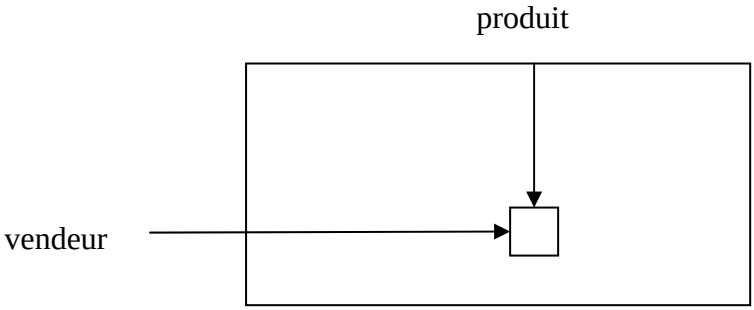


Figure 8 : tableau vendeur / produit

Il est indiqué dans l'énoncé qu'un **produit** appartient à une **famille de produits**, et un **vendeur** à un **service de vente**.

Il est donc légitime de vouloir éditer ce tableau en sommant les éléments par **famille et/ou service**

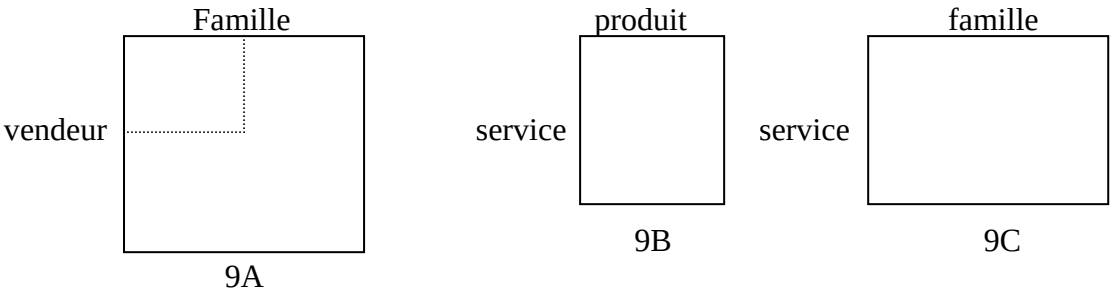


Figure 9 : sommations famille / service

Les dimensions **vendeur** et **produit** sont dites **hiérarchiques,** car elles peuvent se recomposer en **service** et **famille** respectivement.
Les variables **service** et **famille,** quant à elles, se décomposent.

Lorsqu'on va du tableau 8 au tableau **9A** (respectivement **9B**, **9C**), on fait de l'**analyse ascendante** sur la dimension **produit** (respectivement **vendeur**, **vendeur et produit**)

Losqu'on part d'un des tableaux 9 pour éditer le tableau 8, on fait de l'**analyse descendante**.

L'analyse ascendante/descendante est appelée **analyse multi-dimensionnelle**.

Un cube D comme celui de la figure 6 peut donc représenter une multitude de possibilités de sommations, compte tenu des dimensions hiérarchiques et des possibilités d'agrégats :

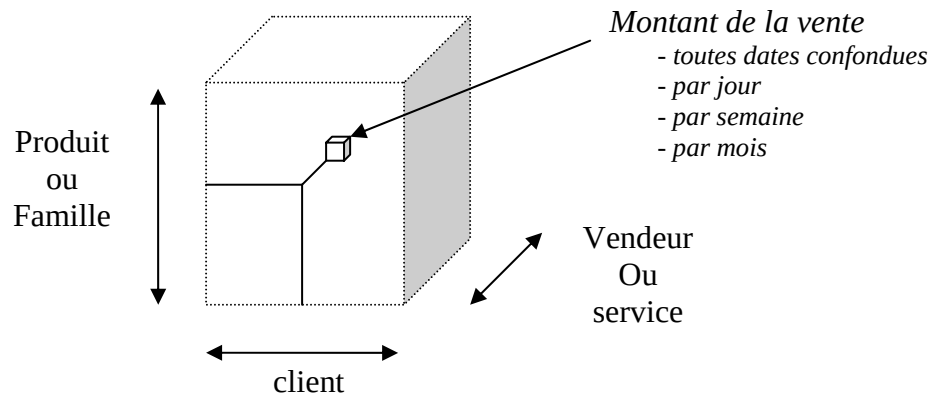


Figure 10 : autres possibilités de cubes D