

Les entrepôts de données (Data Warehouses)

Généralités Décisionnel ?

Qu'est ce que le Data Warehousing

Modélisation des DW

Construction d'un DW :

Alimentation d'un DW

Administration d'un DW

La restitution :

Les cubes de données, OLAP

Les entreprises passent à l' **ère de l'information**.

Défi : Transformer leur système d'information qui avait une vocation de production à un SI décisionnel dont la vocation de pilotage devient majeure.

Qu'est ce qu'un DW ?

D'après BILL Inmon :

*“Un DW est une collection de données **thématiques**, **intégrées**, **non volatiles** et **historisées**, organisées pour la prise de décision.”*

Thématiques : thèmes par activités majeures ;

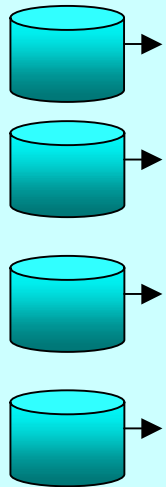
Intégrées : divers sources de données ;

Non volatiles : ne pas supprimer les données du DW ;

Historisées : trace des données, suivre l'évolution des indicateurs.

Pb de volumétrie, de stockage, d'accès.

Architecture Décisionnelle



**Bases
de
production**

Les bases de production : toutes les sources de données
(légales, juridiques, fiscales, politiques, techniques, marketing)

Comment organiser ces différentes données dans un ensemble cohérent afin de procéder à toutes les analyses nécessaires pour construire les **indicateurs** indispensables au pilotage de l'entreprise ?

Par un processus d'entreposage de données

Data Warehousing

Les différentes phases du Data warehousing



Conception



Construction



Administration



Restitution

- **Conception**

Il s'agit de définir la finalité du DW :

- ♠ Piloter quelle activité de l'entreprise ;
- ♠ Déterminer et recenser les données à entreposer ;
- ♠ Définir les aspects techniques de la réalisation ;
- ♠ modèle de données ;
- ♠ démarches d'alimentation ;
- ♠ stratégies d'administration ;
- ♠ définition des espaces d'analyse ;
- ♠ mode de restitution, ...

- **Construction**

Travail technique.

- ♠ Extraction des données des différentes BD de production (internes ou externes)
- ♠ Nettoyage des données, règles d'homogénéisation des données sous formes **de méta données**.
- ♠ Techniques d'alimentation :
 - ✕ Chargement des données dans le DW ;
 - ✕ Fréquences de rafraîchissement :
 - ✓ par applications d'interfaces entre les sources de données et le DW ;
 - ✓ par serveurs de réplication du SGBD ou par outils spécialisés.

• Administration

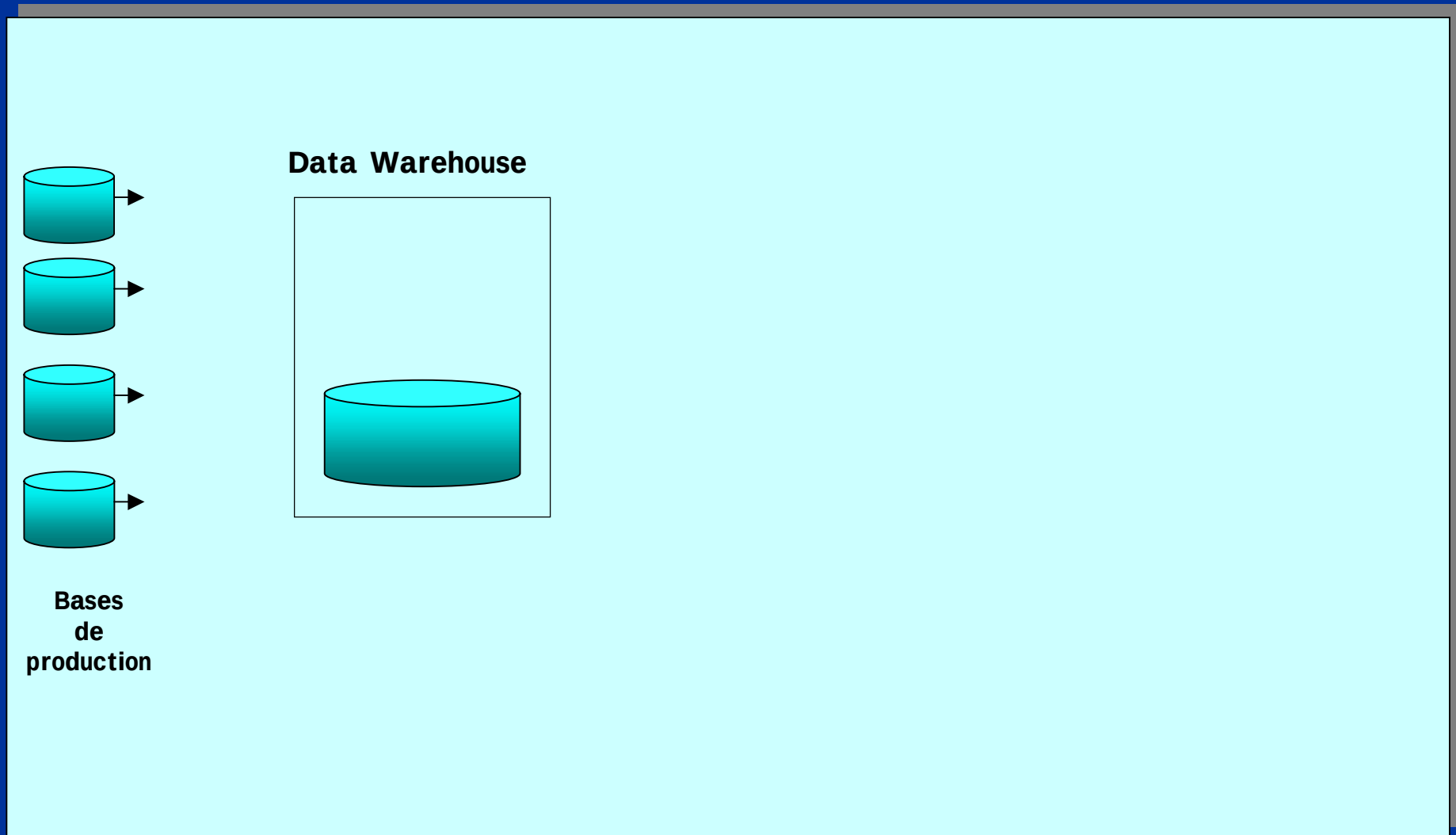
Elle est constituée de plusieurs tâches pour assurer :

- ♠ la qualité et la pérennité des données aux différents applicatifs ;
- ♠ la maintenance ;
- ♠ la gestion de configuration ;
- ♠ les mises à jour ;
- ♠ l'organisation, l'optimisation du SI ;
- ♠ la mise en sécurité du SI.

- **Restitution**

- ➡ C'est le but du processus d'entreposage des données.
- ➡ Elle conditionne le choix de l'architecture du DW et de sa construction.
- ➡ Elle doit permettre toutes la analyses nécessaires pour la construction des indicateurs recherchés.

Architecture Décisionnelle

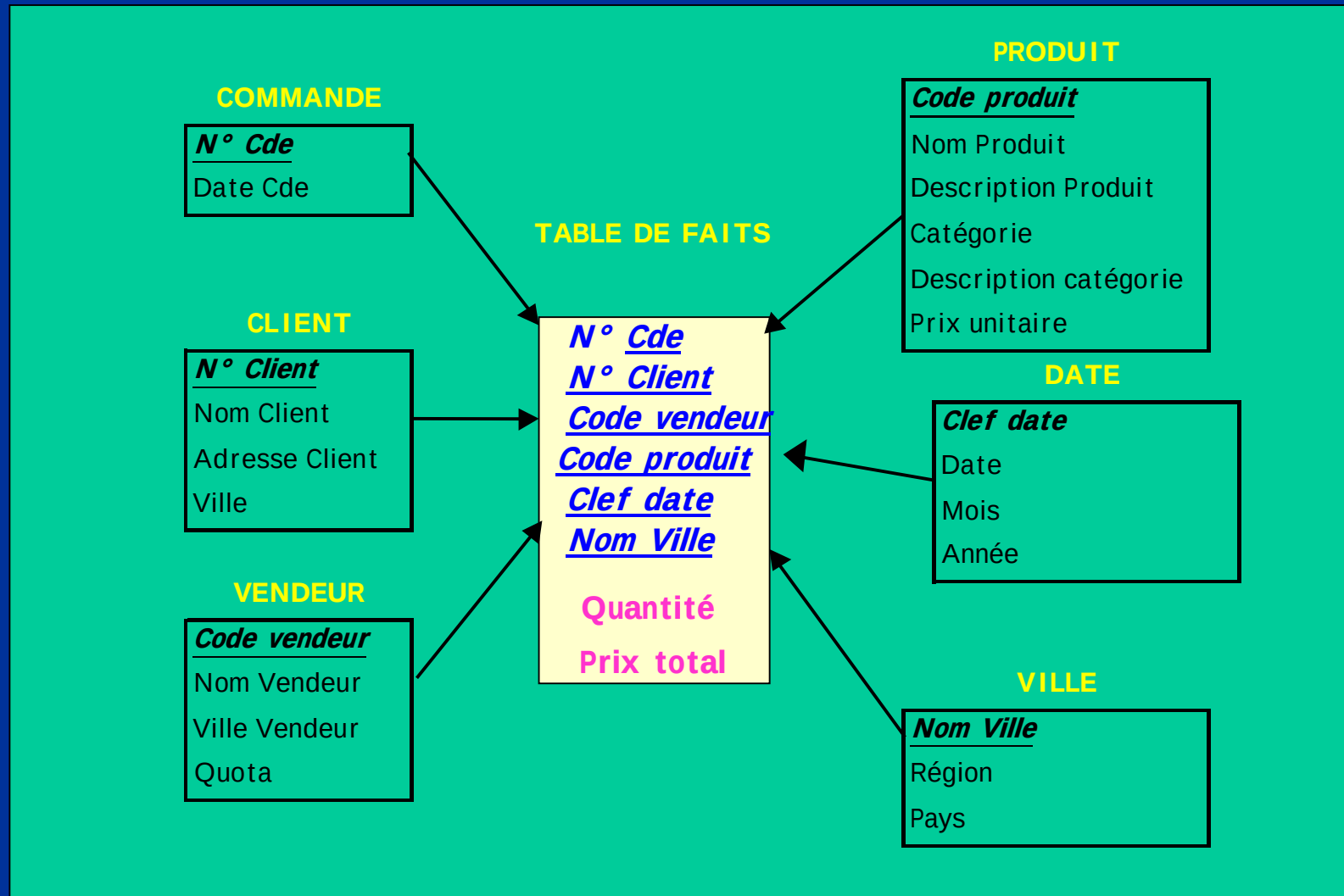


- **Conception**

Il s'agit de définir la finalité du DW :

- ♠ modèle de données ;

Modélisation des DW : Le modèle en étoile



Le modèle en étoile :

- ➡ Une (ou plusieurs) **table(s) de faits** : identifiants des tables de dimension ; une ou plusieurs mesures .
- ➡ Plusieurs **tables de dimension** : descripteurs des dimensions.
- ➡ Une **granularité** définie par les identifiants dans la table des faits.

Avantages :

- ◆ Facilité de navigation
- ◆ Performances : nombre de jointures limité ; gestion des données creuses.
- ◆ Gestion des agrégats
- ◆ Fiabilité des résultats

Inconvénients :

- ◆ Toutes les dimensions ne concernent pas les mesures
- ◆ Redondances dans les dimensions
- ◆ Alimentation complexe.

Propriétés des mesures

Additivité : somme sur toutes les mesures

Exemple : CA ; Quantité vendue, ...

Semi-additivité : somme sur certaine mesures :

Exemple : nbre de contacts clients, Etats des stocks, ...

Non-additivité : pas de somme , recalculer

Exemple : encours moyen fin de mois, plus grand CA pour l'ensemble des magasins

Quelques exemples de modèles de DW

Dans la grande distribution :

Quelques tables de faits : détaillées et volumineuses

Tables de dimensions :

Classiques : produit, fournisseur, temps, établissement
(structure géographique, fonctionnelle), ...

Stratégiques : Client, Promotions, ...

Rq : Obtenir le plus d'enregistrements possibles.

Dans le secteur des banques :

Tables de faits : nombreuses, dédiées à chaque produit , peu détaillées et peu volumineuses.

Tables de dimensions :

Classiques : produit, temps, établissement (structure géographique, fonctionnelle), ...

Stratégiques : Client, ...

Rq : Obtenir le plus de données (champs) possibles.

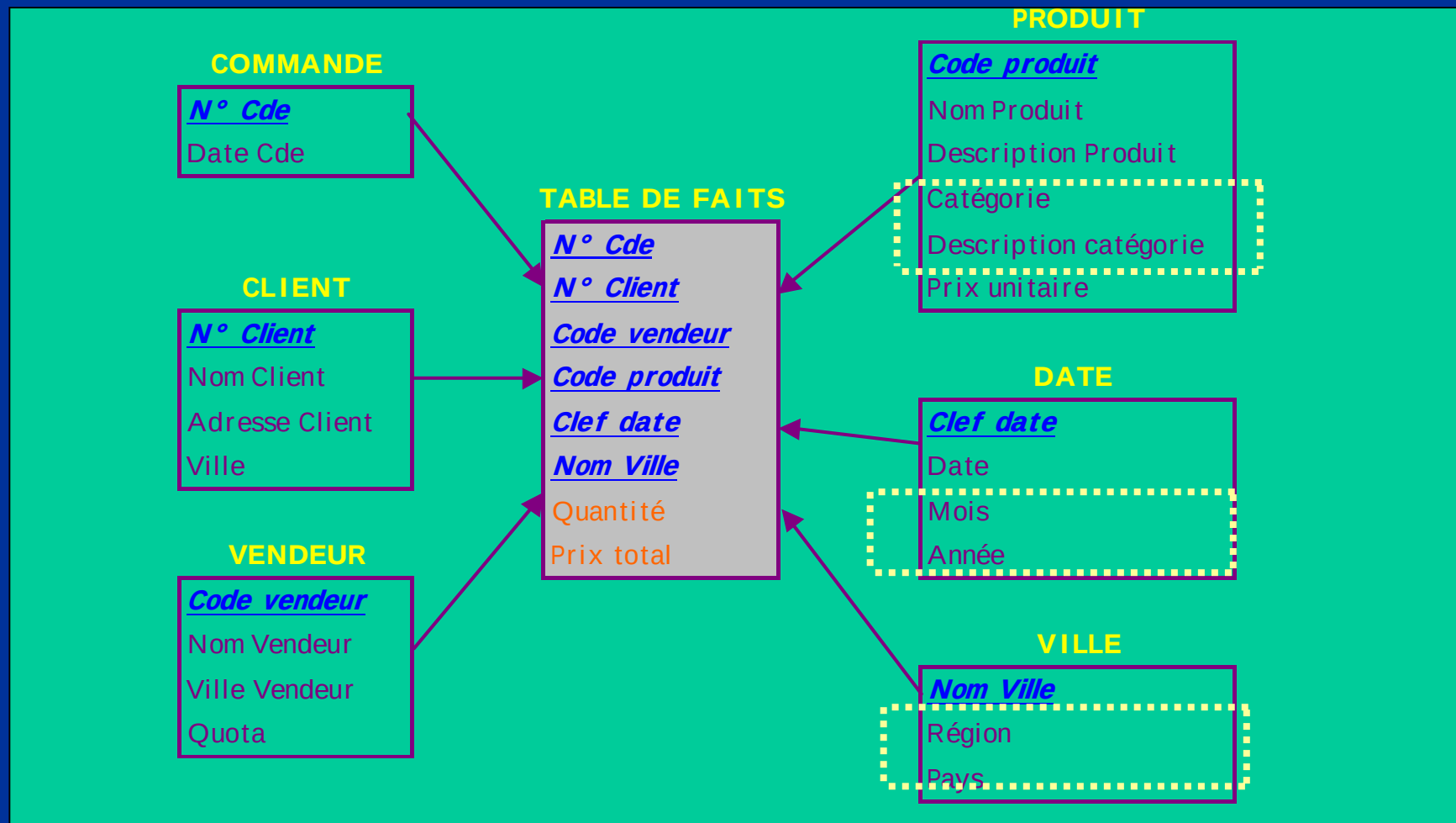
Le modèle en flocons de neige

Le modèle du DW doit être simple à comprendre.
On peut augmenter sa lisibilité en regroupant certaines dimensions.
On définit ainsi des **hiérarchies**.

Celles-ci peuvent être géographiques ou organisationnelles.

Exemple : Commune, Département, Région, Pays, Continent

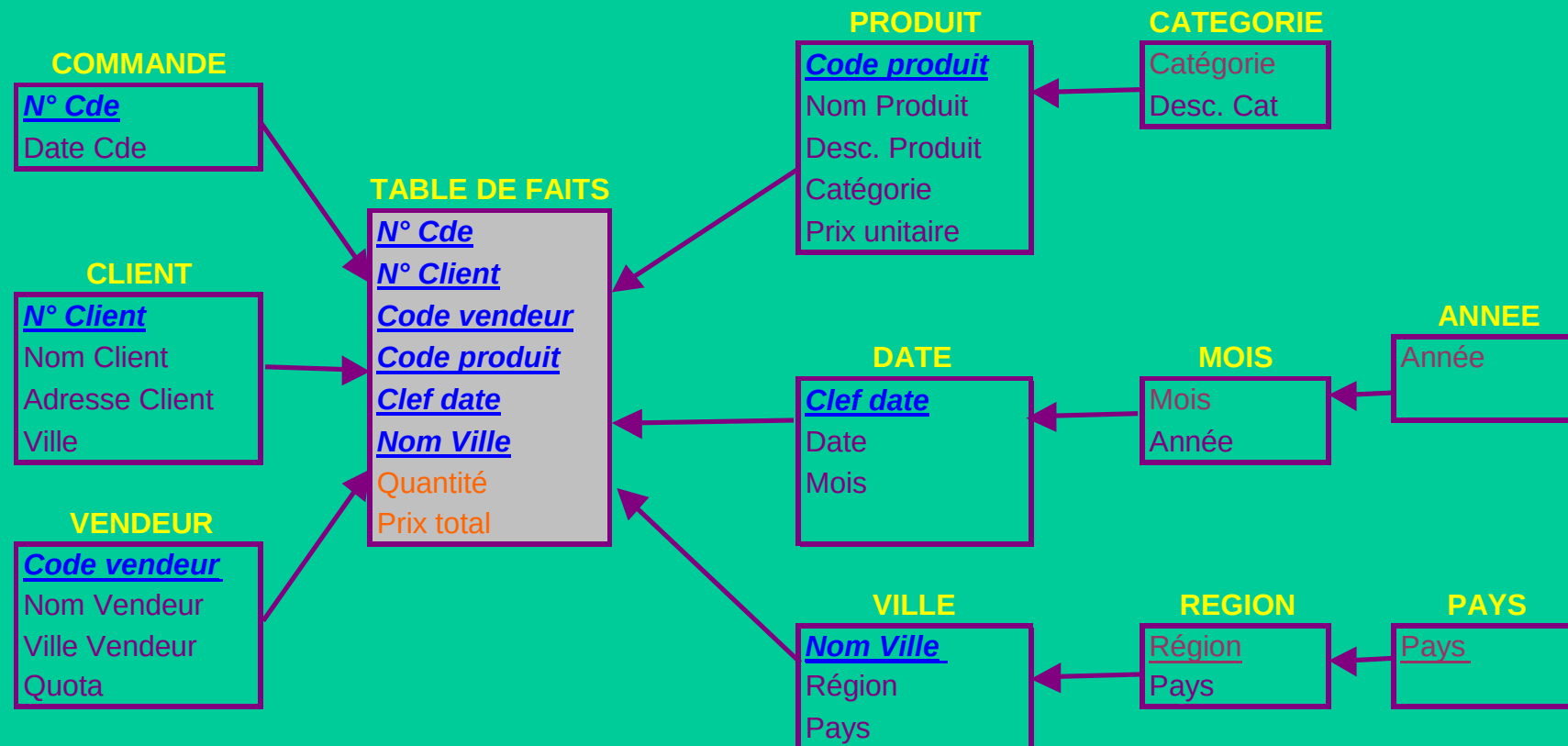
Client	Continent	Pays	Region	Département	Commune
Pepone	Europe	France	RhôneAlpes	Rhône	Lyon 1°
Testut	Europe	France	RhôneAlpes	Rhône	Lyon 2°
Soinin	Europe	France	RhôneAlpes	Rhône	Lyon 3°
Vepont	Europe	France	Ile de France	Paris	Paris 1°
Martin	Europe	France	Ile de France	Paris	Paris 2°
Elvert	Europe	France	Ile de France	Yvelines	Versailles



Les valeurs des données des dimensions de niveau supérieur se répètent :
dénormalisation (→ lien fonctionnel).

Lien fonctionnel entre plusieurs champs d'une dimension.
 Ces champs sont séparés dans une même table (taille limitée) pour avoir un temps de réponses rapide.

Le modèle floconné



Modèle floconné = Modèle en étoile + normalisation des dimension

Lorsque les tables sont trop volumineuses

Avantages :

- réduction du volume,
- permettre des analyses par pallier (drill down) sur la dimension hiérarchisée.

Inconvénients :

- navigation difficile ;
- nombreuses jointures.

Estimer le volume du DW

Calculer ou estimer le nombre d'enregistrements

Prendre en compte :

- La table des faits
- Les dimensions significatives
- Les agrégats
- Les index
- Saisonnalité des ventes
- Croissance du CA, des encours, du nombre de points de ventes

Exemples :

Grandes distribution :

CA annuel : 80 000 M\$

Prix moyen d'un article d'un ticket : 5\$

Nbre d'articles vendus pour un an : $80 * 10^9 / 5 = 16 * 10^9$

Volume du DW :

$16 * 10^9 * 3 \text{ ans} * 24 \text{ octets} = \underline{1,54 \text{ To}}$ ($1,54 * 10^{12} = 1\,540 \text{ Go}$)

Téléphonie :

Nbre d'appels quotidiens : 100 millions

Historique : 3 ans * 365 jours = 1 095 jours

Volume du DW :

$100 \text{ millions} * 1\,095 \text{ jours} * 24 \text{ octets} = \underline{3,94 \text{ To}}$

Cartes de crédit :

Nbre de clients : 50 millions

Nbre moyen mensuel de transactions : 30

Volume :

$50 \text{ millions} * 26 \text{ mois} * 30 \text{ transactions} * 24 \text{ octets} = \underline{1,73 \text{ To}}$

Alimenter un DW : (Outils ETL)

Le principe de l'entreposage des données est de rassembler de multiples données sources qui souvent sont hétérogènes en les rendant *homogènes* afin de les analyser.

Ce travail d'homogénéisation nécessite des règles précises servant de **dictionnaire** (ou de **référentiel**) et qui seront mémorisées sous forme de **méta-données** (information sur les données).

Ces règles permettent d'assurer des tâches d'administration et de gestion des données entreposées.

Le dictionnaire (ou référentiel) de données est constitué de l'ensemble des méta-données.

Il renferme des informations sur toutes les données du DW.
Il renferme également des informations sur chaque étape lors de la construction du DW ; sur le passage d'un niveau de données à un autre lors de l'exploitation du DW.

Le rôle des méta-données est de permettre :

- ♣ La définition des données
- ♣ La fabrication des données
- ♣ Le stockage des données
- ♣ L'accès aux données
- ♣ La présentation des données.

L'alimentation d'un DW est une procédure qui s'effectue en plusieurs étapes :

- **Sélection des données sources**
- **Extraction des données**
- **Transformation**
- **Chargement**
- **Administration**

- **Sélection des données sources**

Quelles sont les données de productions qu'il faut sélectionner pour alimenter le DW ?

Toutes les données sources ne sont forcément pas utiles.

Doit-on prendre l'adresse complète ou séparer le code postal.

Les données sélectionnées seront réorganisées pour devenir **des informations**.

La synthèse de ces données sources a pour but de les enrichir.

La **dénormalisation** des données crée des liens entre les données et permet des accès différents

- **Sélection des données sources (suite)**

La sélection des données utiles à partir des BD de production n'est pas simple à faire .

Les données sont :

- hétérogènes** (différents SGBD et différentes méthodes d'accès);
- diffuses** (différents environnements matériels et différents réseaux interconnectés ou non);
- complexes** (différents modèles logiques et physiques principalement orientés vers les traitements transactionnels).

La définition de la granularité dépend du niveau de raffinement de l'information qu'on veut obtenir.

- **Sélection des données sources (suite)**

Il existe plusieurs niveaux de données :

- Les données sont parfois assemblées avant d'être injectées dans le DW permettant une vision intégrée et transversale de l'entreprise.

Cette forme de données constitue **le niveau le plus fin** au niveau du DW : ceux sont les données de détail. Elles peuvent être agrégées et constituent ainsi un autre niveau de détail.

Elles seront par la suite **structurées** dans des espaces d'analyse (soit des cubes de données, soit des data marts).

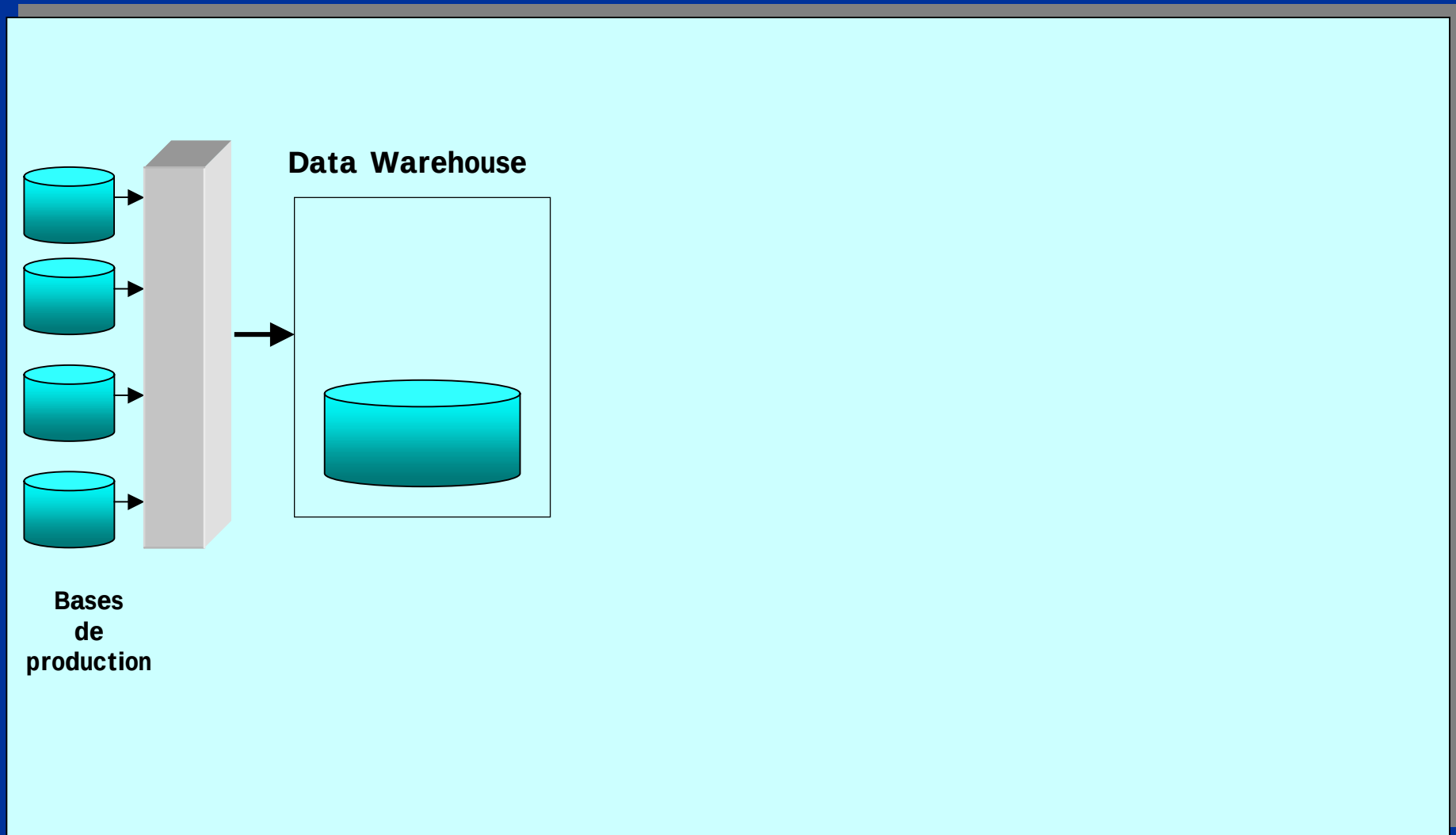
Elles seront finalement à **un niveau de présentation**, où elles peuvent avoir plusieurs formes (tableaux, graphiques, tableaux de bord, règles de connaissances...).

- **Extraction des données**

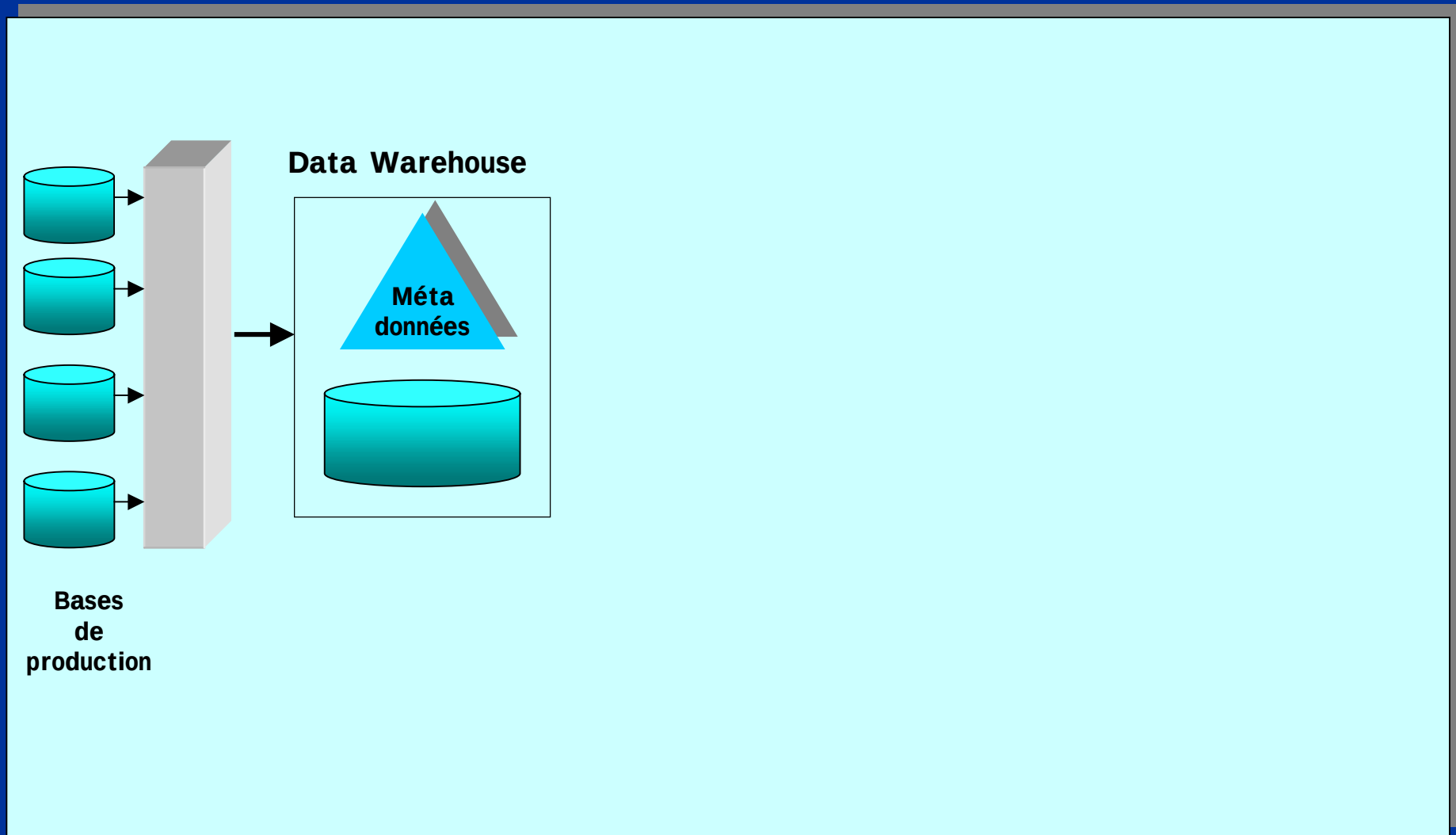
- L'extraction peut se faire à travers un outil d'alimentation qui doit travailler de façon native avec les SGBD qui gèrent les données sources.
- Ou alors créer des pgms extracteurs. L'inconvénient de cette approche est le risque de faire des extractions erronées, incomplètes et qui peuvent biaiser le DW.
- Il faut gérer les anomalies en les traitant et en gardant une trace

- **Extraction des données (suite)**
- L'extraction doit se faire conformément aux règles précises du référentiel.
- Elle ne doit non plus perturber les activités de production.
- Il faut faire attention aux données cycliques. Celles qu'on doit calculer à chaque période, pour pouvoir les prendre en considération.
- L'extraction peut se faire en interne selon l'horloge interne ou par un planificateur ou par la détection d'une donnée cible (du DW) ; ou en externe par des planificateurs externes.
- Les données extraites doivent être marquées par “ *horodatage* ” afin qu'elles puissent être pistées.

Architecture Décisionnelle



Architecture Décisionnelle



- **Transformation**

C'est une suite d'opérations qui a pour but de rendre les données cibles **homogènes** et puissent être traitées de façon **cohérente**.

Exemple

Donnés sources

Appli 1 : m, f

Appli 2 : 1, 0

Appli 3 : Homme, femme

données cibles

m,f

m,f

m,f

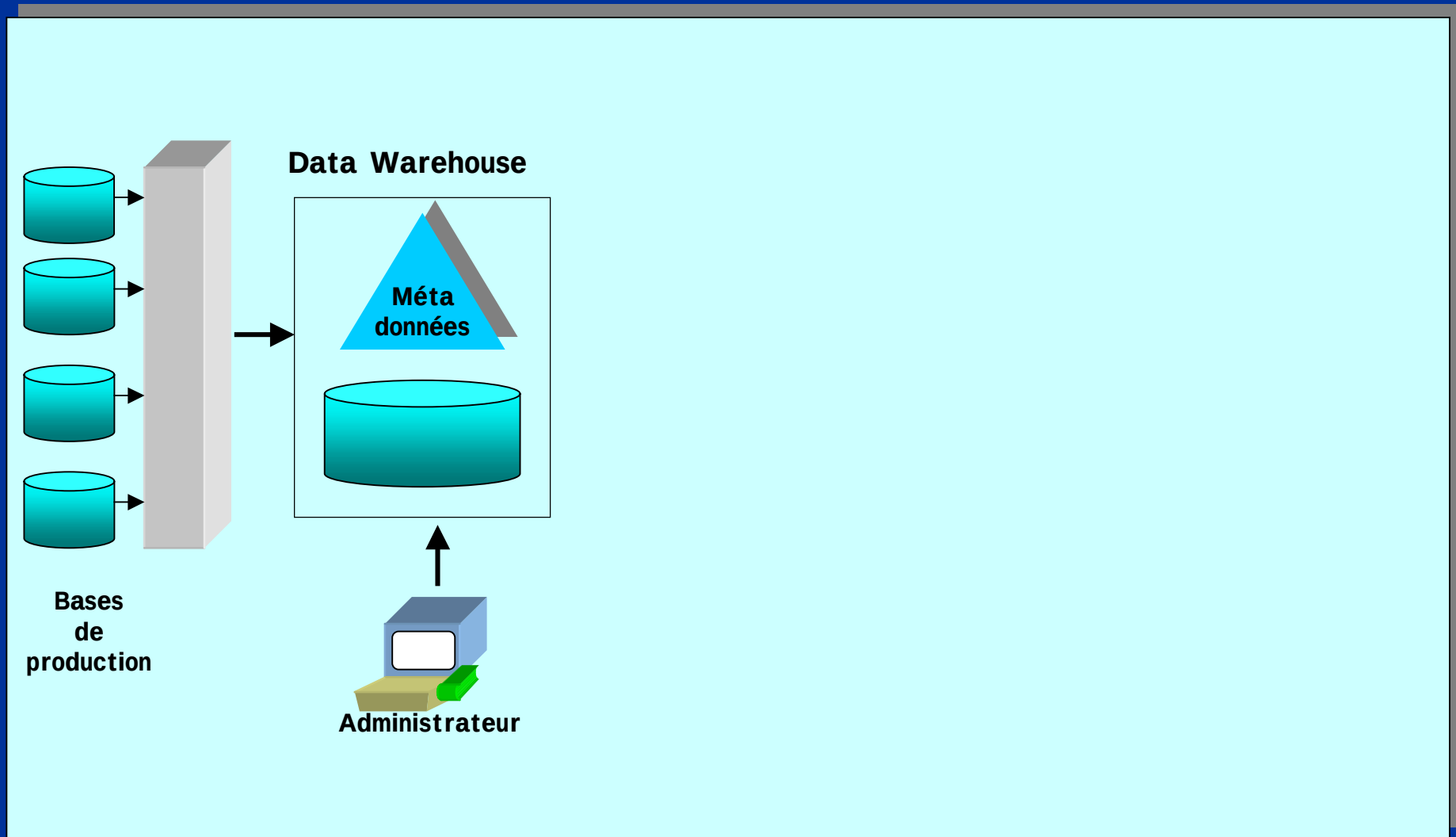
- **Transformation (suite)**
- Les données doivent alors être filtrées afin d'éliminer les données aberrantes : données sans valeurs avec des valeurs manquantes (exemple de la valeur 99).
- Souvent dans les bases de production, certaines données sont sémantiquement fausses.
- Pour avoir une alimentation de qualité, il faut avoir une bonne connaissance des données à entreposer et des règles qui les régissent. Savoir corriger les données pour les doter d'un vrai sens sémantique.
- On peut dédoubler des données pour gagner au niveau de la cohérence.

- **Transformation (suite)**
 - L'ensemble des données sources, après nettoyage ou transformation d'après des règles précises ou par application de programmes (*pour un contrôle de vraisemblance par des méthodes statistiques*), seront restructurées et converties dans un **format cible**.
 - Il faut synchroniser les données pour que les valeurs agrégées obtenues soient cohérentes. Avant de passer à la phase de chargement.

- **Chargement**

- ☑ C'est l'opération qui consiste à charger les données nettoyées et préparées dans le DW.
- ☑ C'est une opération qui risque d'être assez longue. Il faut mettre en place des stratégies pour assurer de bonnes conditions à sa réalisation et définir la politique de rafraîchissement.
- ☑ C'est une phase plutôt mécanique et la moins complexe.

Architecture Décisionnelle



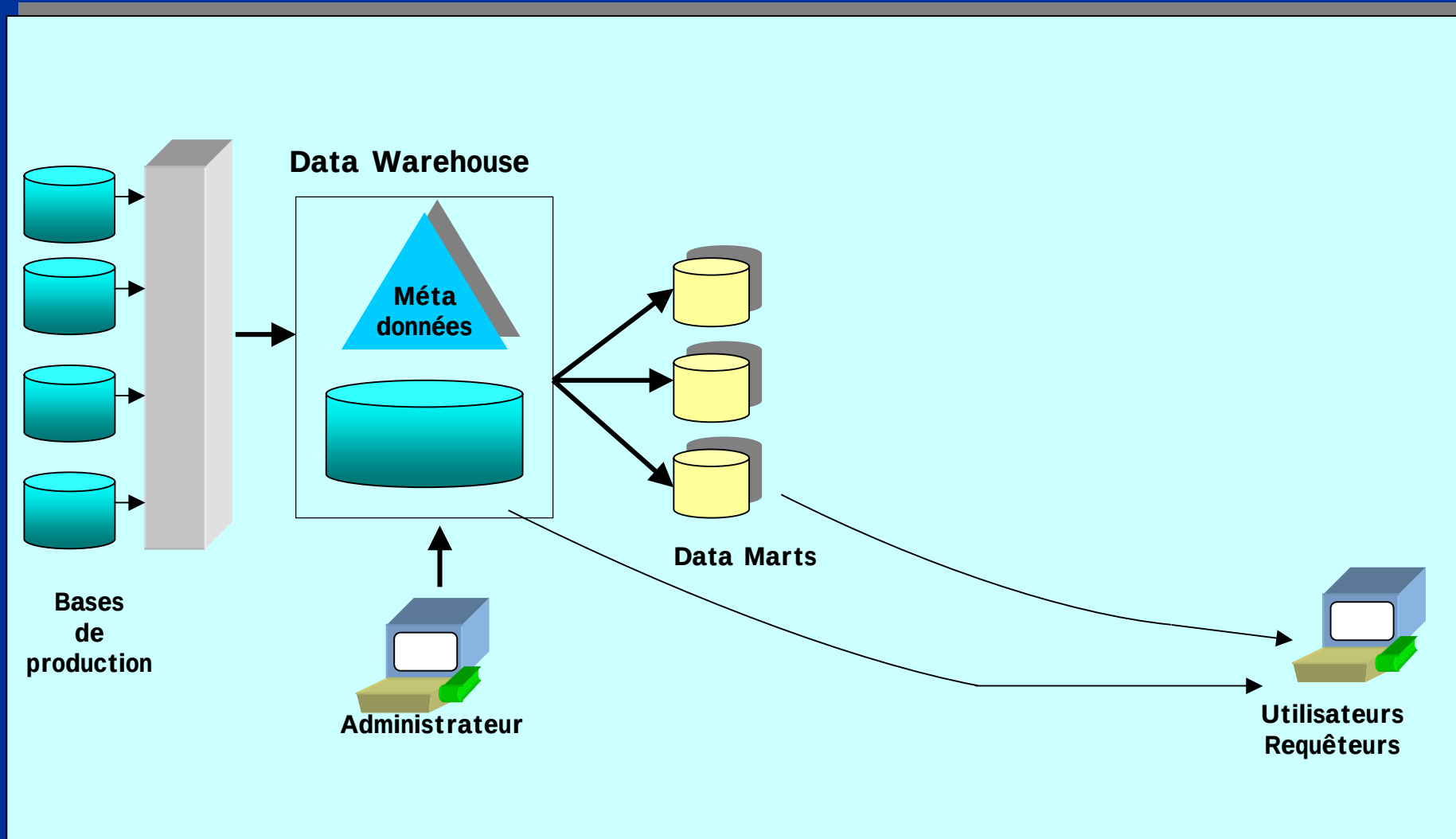
Administration du DW

- ✍ La fonction d'administration porte sur un **aspect fonctionnel** (qualité et la pérennité des données) mais aussi sur un **aspect technique** (maintenance, optimisation, sécurisation,...)
- ✍ Elle concerne l'ensemble des tâches du processus d'entreposage de la sélection des données de production à la mise à disposition pour construire les espaces d'analyse.
- ✍ L'administrateur du DW doit maîtriser la gestion des données (*données, provenance des données, méta-données*).
- ✍ Les données agrégées sont aussi une production (information) de l'entreprise comme les données de production (ERP), doivent être entreposées.
- ✍ Ainsi le développement du DW témoignera, aussi bien de la production de base que de l'activité informationnelle (pilotage de l'entreprise).
- ✍ Les requêtes portent plus souvent sur les agrégats que sur les données de base. (80% - 20%)

- **Administration**

- ✘ Le DW est un aspect physique du SI de l'entreprise. Il doit être par conséquent évolutif. Les données doivent donc changer. On doit procéder à d'autres alimentations et donc gérer l'actualisation des données.
- ✘ Il existe des outils qui prennent en charge les tâches de rafraîchissement des données.
- ✘ Ils procèdent par replication pour propager les maj effectuées dans les BD sources, dans le DW.
- ✘ Le mécanisme de réplication et une opération de copie de données d'une BD vers une ou plusieurs BD.
- ✘ Les répliquations sont alors asymétriques synchrones ou asynchrones ou alors symétriques synchrones ou asynchrones.
- ✘ Le rafraîchissement des données peut se faire également par des processus de transformation qui exploitent les méta-données.

Architecture Décisionnelle



La restitution

Les data marts

Simple “**magasins de données**” (Data marts), on y stockera des données portant sur une des activités de l’entreprise.

Ceux sont en quelque sorte des vues métier.

Exemple Data mart Comptabilité, Data mart RH,.....

Ces mini DW peuvent alors être considérés comme des espaces d’analyse, du fait que les données sont bien moins nombreuses et surtout qu’elles sont thématiques.

Ils peuvent également servir de bases de construction à des cubes de données.

Les cubes données et les technologies OLAP

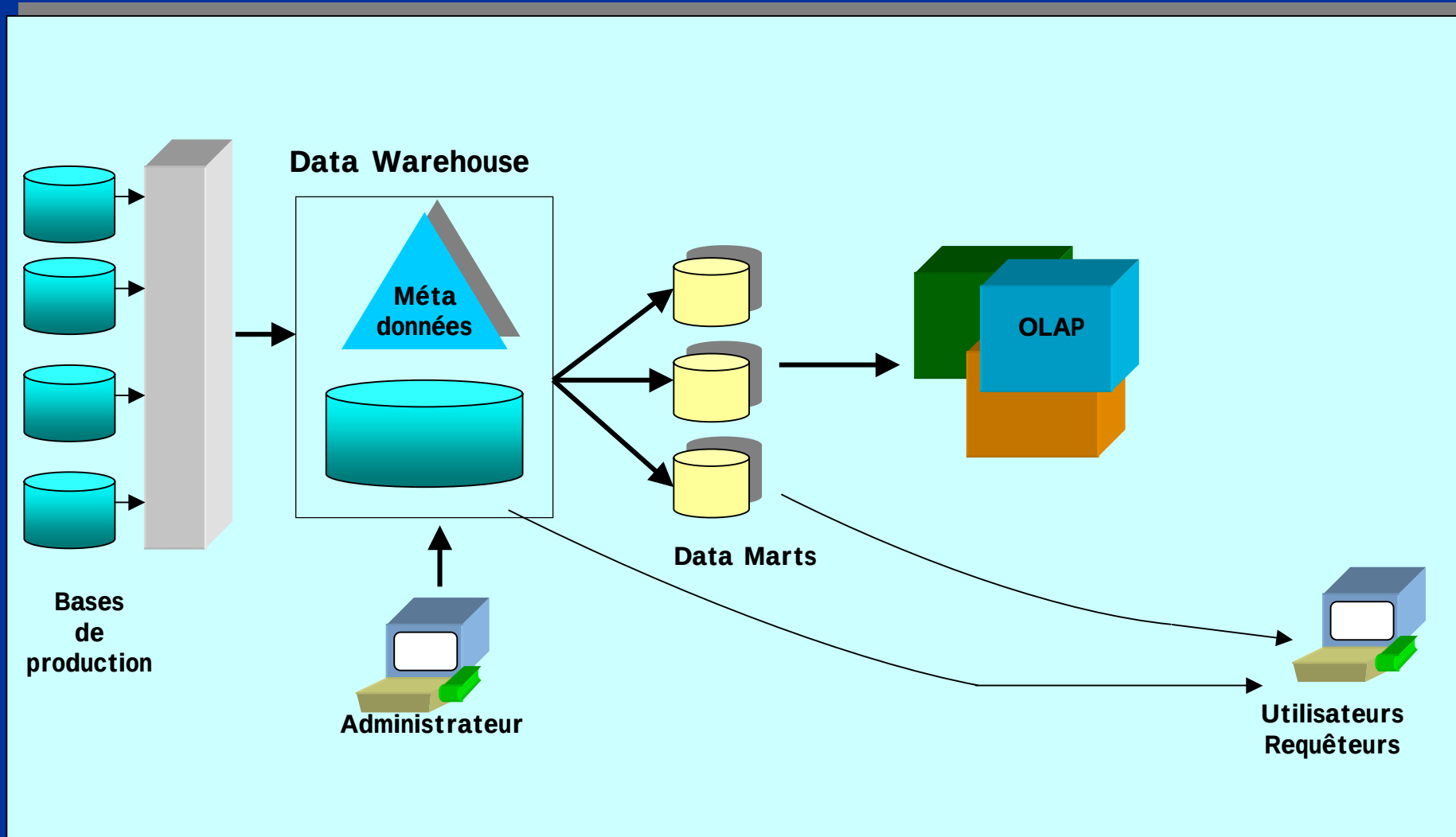
OLTP

- Les applications conçues pour des opérations quotidiennes dans les BDs.
- Ces transactions nécessitent des données détaillées et actualisés.
- Les BD vont de quelques milliers de Mo à des Go.
- .

OLAP

- Les données sont historisées, résumées, consolidées.
- Les DW contiennent des données sur une longue période de temps.
- Les DW vont de centaines de Go à des To (Téra octets).

Architecture Décisionnelle



Les 12 règles d'OLAP :

- 1°) une vue multidimensionnelle des données.
- 2°) La transparence vis à vis de l'utilisateur qui doit accéder à la BD par l'intermédiaire d'outils simples (tableur, par ex).
- 3°) La BD doit disposer d'un modèle et d'outils permettant d'accéder à de multiples sources, d'effectuer les conversions et extractions nécessaire pour alimenter la Base OLAP.
- 4°) Le modèle de données, le nombre de dimensions ou le nombre de niveaux d'agrégation doivent pouvoir changer, sans remettre en cause le fonctionnement de la base.
- 5°) Architecture Client/ Serveur.
- 6°) Toutes les dimensions définies dans le modèle de données doivent être accessibles pour chacune des données.
- 7°) Gestion des matrices creuses. Les parties vides du cube multidimensionnel doivent être stockées de manière à ne pas détériorer les temps d'accès.
- 8°) Accessibilité simultanément par plusieurs utilisateurs.
- 9°) Toutes les données stockées ou calculées dans le cube doivent être accessibles et les règles de gestion doivent toujours s'y appliquer. Toutes les tranches de cube doivent être visualisées.
- 10°) Navigation aisée dans les données pour les utilisateurs, de manière intuitive.
- 11°) Outil de présentation des données.
- 12°) Nombre illimité de dimensions et de niveaux d'agrégation.

Représentation des données

Les données sont perçues à travers plusieurs dimensions. Elles sont qualifiées de **multidimensionnelles**, indépendamment de leur support (tables relationnelles ou tableaux multidimensionnels)

Produit	Region	Ventes
Clous	Est	50
Clous	Ouest	60
Clous	Centre	100
Vis	Est	40
Vis	Ouest	70
Vis	Centre	80
Boulons	Est	90
Boulons	Ouest	120
Boulons	Centre	140
Nettoyeurs	Est	20
Nettoyeurs	Ouest	10
Nettoyeurs	Centre	30

Représentation des données dans une table relationnelle

	Est	Ouest	Centre
Clous	50	60	100
Vis	40	70	80
Boulons	90	120	140
Nettoyeurs	20	10	30

Représentation des données dans un tableau multidimensionnel

les requêtes seront de type : “ **Quelle est le total des ventes dans l’Est?**”.

On peut calculer divers totaux.

Tables relationnelles : on peut traiter quelques centaines d’enregistrements par seconde.

Tableau multidimensionnel : on peut rajouter en lignes et colonnes plus de 10000 valeurs par seconde.

Pour accélérer les temps de réponses, il est préférable de pré-calculer tous les totaux et sous totaux possibles.

Produit	Region	Ventes
Clous	Est	50
Clous	Ouest	60
Clous	Centre	100
Clous	Total	210
Vis	Est	40
Vis	Ouest	70
Vis	Centre	80
Vis	Total	190
Boulons	Est	90
Boulons	Ouest	120
Boulons	Centre	140
Boulons	Total	350
Nettoyeurs	Est	20
Nettoyeurs	Ouest	10
Nettoyeurs	Centre	30
Nettoyeurs	Total	60
Total	Est	200
Total	Ouest	260
Total	Centre	350
Total	Total	810

Pour le calcul de ces totaux : 28 accès en lecture et 8 accès en écriture.

Un SGBDR lit 200 enregistrements par seconde et en écrit environ 20 par seconde.

	Est	Ouest	Centre	Total
Clous	50	60	100	210
Vis	40	70	80	190
Boulons	90	120	140	350
Nettoyeurs	20	10	30	60
Total	200	260	350	810

OLAP consolide entre 20 et 30000 cellules par seconde.

La valeur ALL remplace une colonne d'agrégats.

Magasin	Date	Rayon	CA Ventes
Mag1	1/2/96	010	3500
Mag1	6/2/96	010	2500
Mag1	10/2/96	010	2900
Mag1	ALL	010	8900
Mag1	...	...	...

S'il y a N attributs concourant à la construction du cube : il y aura **$2^N - 1$ agrégations.**

Soit $C_1, C_2, \dots, C_N$ les cardinalités des N attributs : le cube aura **$\prod(C_i + 1)$ enregistrements.**

Dans la tables VENTES si on a $2 \times 3 \times 3 = 18$ tuples :
dans le cube on aura $3 \times 4 \times 4 = 48$ tuples.

L'approche relationnelle.

(MicroStrategy MS ; Informix's Metacube MC , Information Advantage IA)

L'ensemble des données est stocké dans une BDR. Les données sont sous forme d'enregistrements.

VENTES (Magasin, Rayon, Date, CA Ventes, Nb Client)

```
Select Magasin, Date , Sum(CA Ventes)
From VENTES
Group By Magasin, Date
```

Nouveaux opérateurs d'agrégation : cube , rollup.

L'approche multidimensionnelle

Arbor Software : hyperion (Codd & co...), Express Oracle , LightShip (de Pilot)

Il s'agit de stocker les données dans des tableaux multidimensionnels.
Ces tableaux peuvent être éparses.

On y stocke dans les **cellules** les mesures (valeurs à observer), les données représentant les dimensions sont les **coordonnées** de ces valeurs.

Un tableau = { **cellules** }

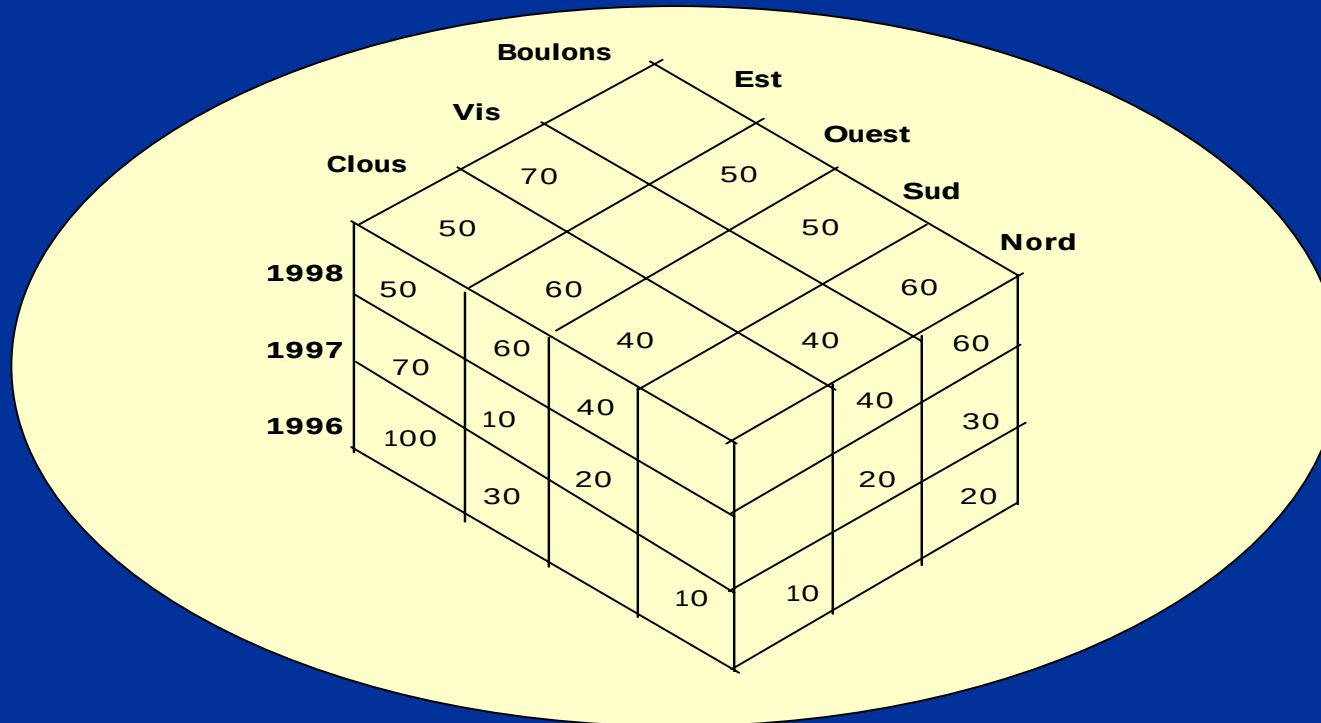
Les cellules contenant les données originelles : **entrées**

Les totaux calculés sont les **sorties**

Les valeurs des dimensions sont les **membres**

Les données dans les cellules représentent une **variable** dimensionnée,
exemple: CA par Produits et Régions.

Représentation des données dans un cube



On peut *agréger* les **produits** par *catégories* ou *détailler* les **régions** par *villes*.

BD éparses

- Plus on a de dimension plus on a de cellules.
Une partie des produits est vendue $\Rightarrow$ des cellules sans valeur : données éparses.

Exemple :

On dispose de 100 000 données (eq. Tuples)

4 dimensions ayant une cardinalité de 30 modalités chacune:

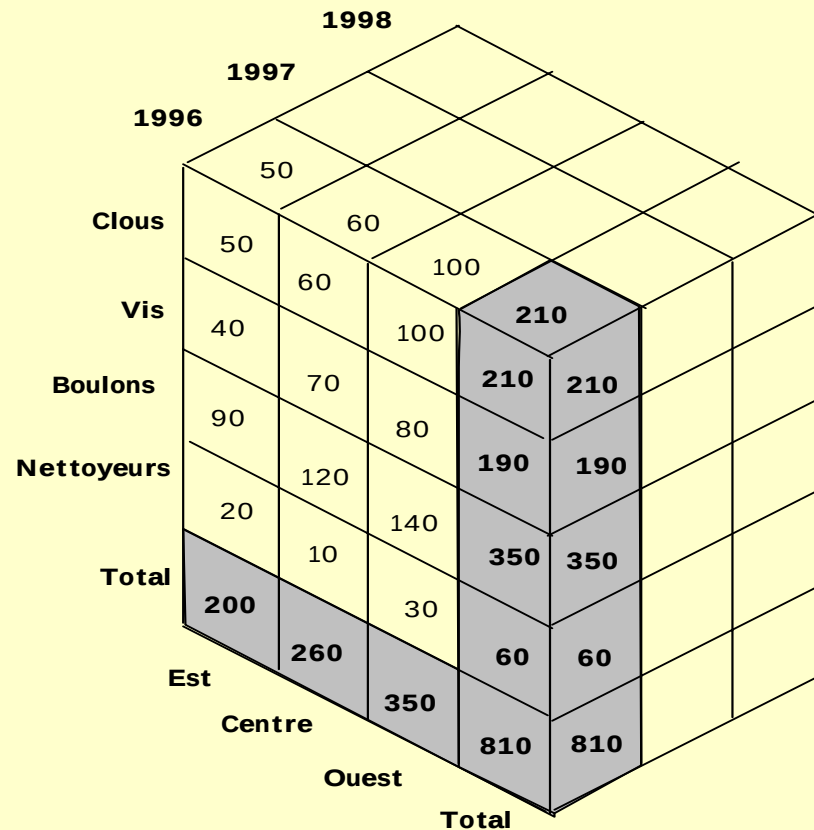
$$30 * 30 * 30 * 30 = 810\,000 \text{ cellules}$$

(dont 710 000 vides : 12,3% seulement sont pleines)

- Une BD est considérée comme éparses si elle a moins de **40%** de ses cellules “ peuplées ”.
- Techniques de compression des données

Les différents opérations d'OLAP

Roll up ; **Drill down** : Agrégation de données ; forage de données :

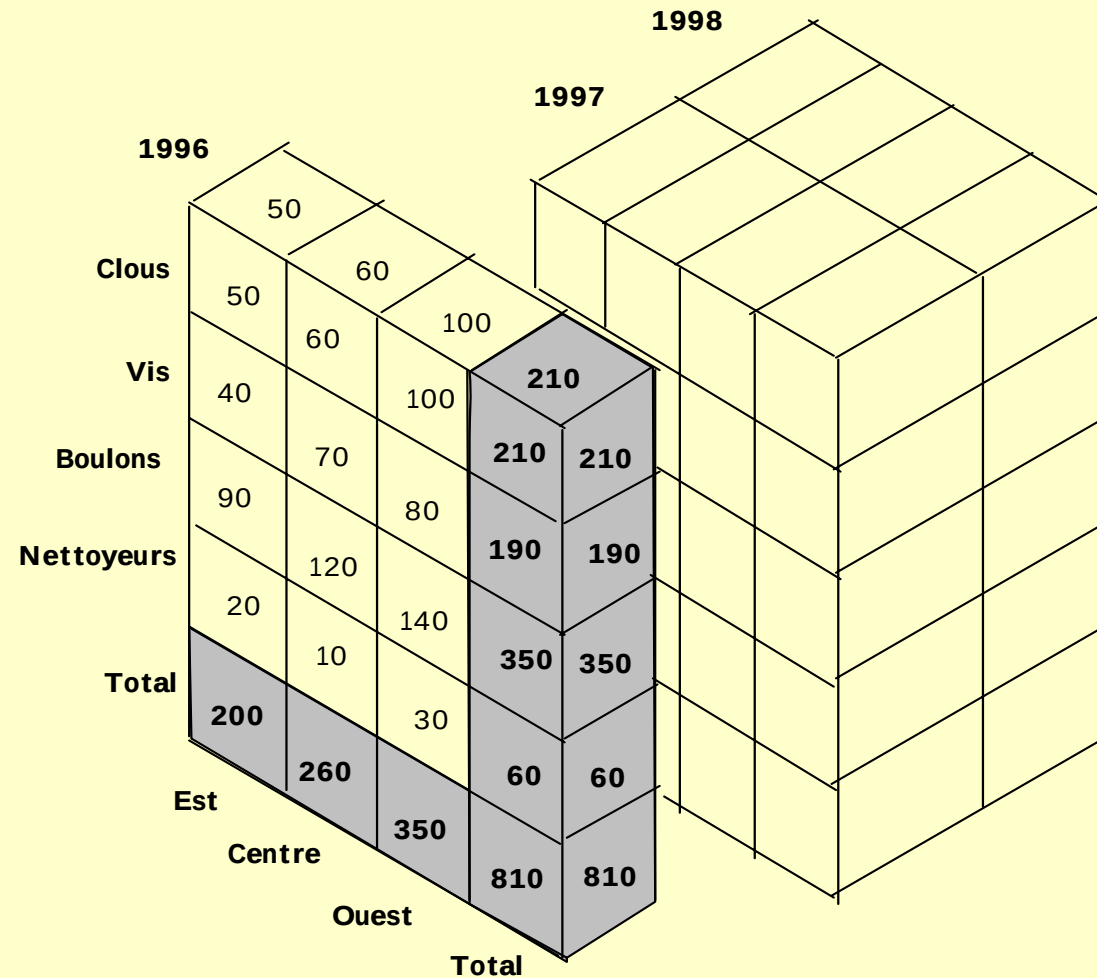


L'opération du Drill peut se décliner en plusieurs autres opérations :

- **Drill accross** : Drill latéral, comparaison sur des mesures dans plusieurs tables de faits ;
- **Drill through** : voir l'information à travers plusieurs dimensions
- **Reach through** : voir l'information en profondeur, jusqu'aux données de base.

Slice & Dice

Sélection & projection des données



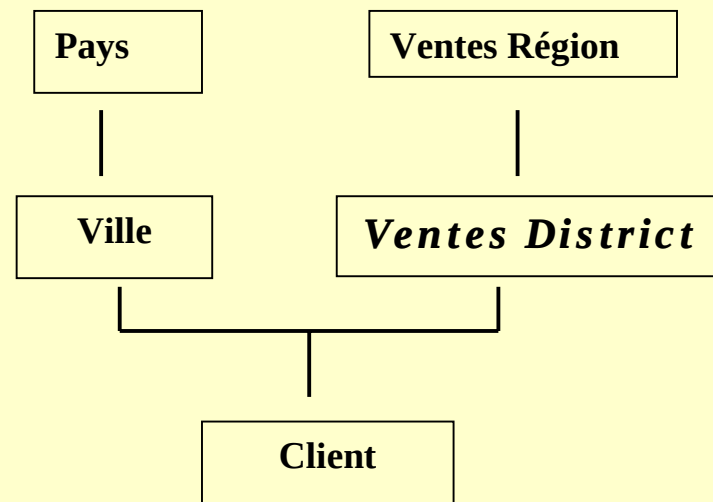
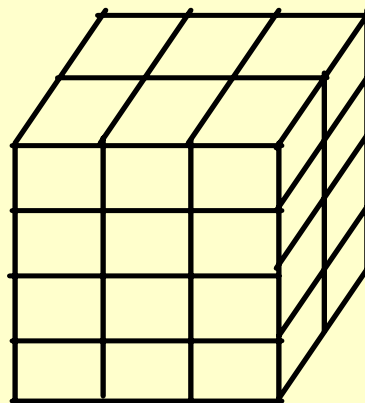
Dimension avec des classes : hiérarchie

Exemple : *Villes* → *Régions* → *Pays* → *Continent*.
Les dimensions peuvent être organisées en hiérarchie

Hiérarchies multiples dans une dimension :

Région
District
Client

Pays
Ville
Client



Nesting

Emboîtement des dimensions pour une représentation en 2-D

		1996	1997	1998
Clous	Est	50	60	
	Ouest	60		60
	Centre	100	110	120
Vis	Est	40	50	
	Ouest	70		80
	Centre	80	90	
Boulons	Est	20	40	50
	Ouest	10		
	Centre	30	40	50

Pivoting

Transposition des dimensions

L'approche hybride (HOLAP)

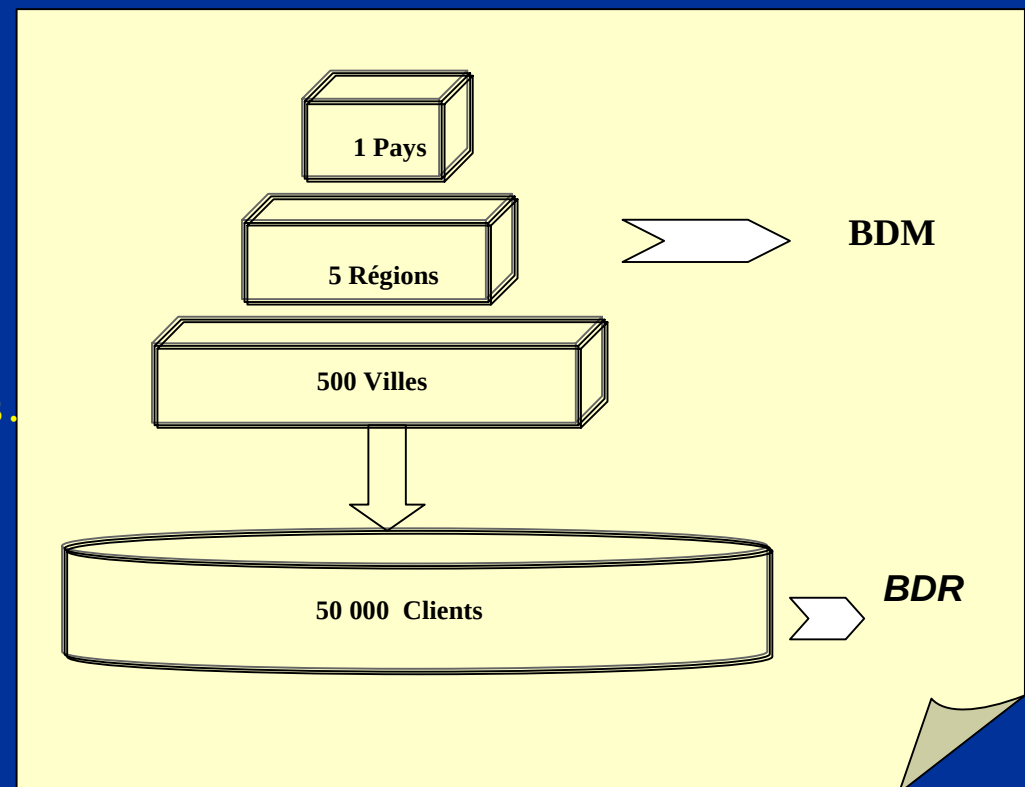
Approche relationnelle : 30% du temps est consacré aux I/O.

Approche multidimensionnelle : 20%. (70% calculs et 10% décompression)

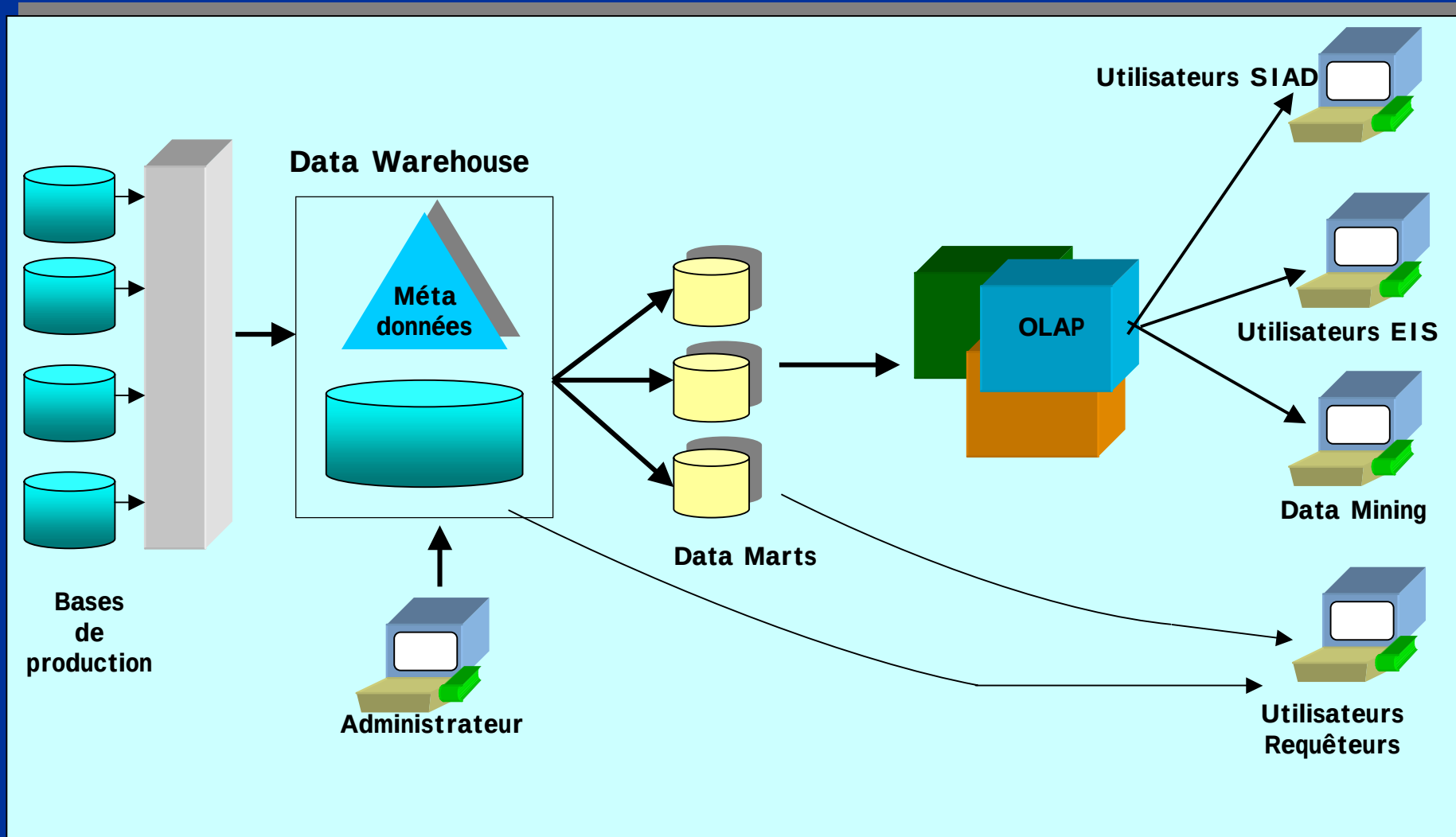
La 3^o voie préconisée consiste à utiliser les **tables** comme **structure** permanente de **stockage** des données et les **tableaux** comme **structure** alors des requêtes.

La démarche consisterait en 3 étapes :

1. Charger les données d'une table vers un tableau.
2. Calculer le cube de ce tableau selon les méthodes initialement présentées.
3. Stocker les résultats (données agrégées) dans un table.



Architecture Décisionnelle



Tous les outils pouvant synthétiser, **explorer**, **confirmer**, **expliquer**, **prédire** les données sont des outils de restitution

Les différentes techniques de restitution :

- ☞ La représentation cartographique ou par color coding

- ☞ Les requêteurs

 - Les outils graphiques d'interrogations

 - Les outils EIS / SIAD

- ☞ Les Outils statistiques et le **Data Mining**

 - d'analyse exploratoire** (techniques statistiques, visualisation, recherche d'associations, typologies, réseaux de neurones)

 - d'analyse confirmatoire** (techniques de régressions, de segmentations, réseaux de neurones)

 - d'analyse de modèles prédictifs** (techniques de régressions, de segmentations, réseaux de neurones).

Merci
pour votre attention