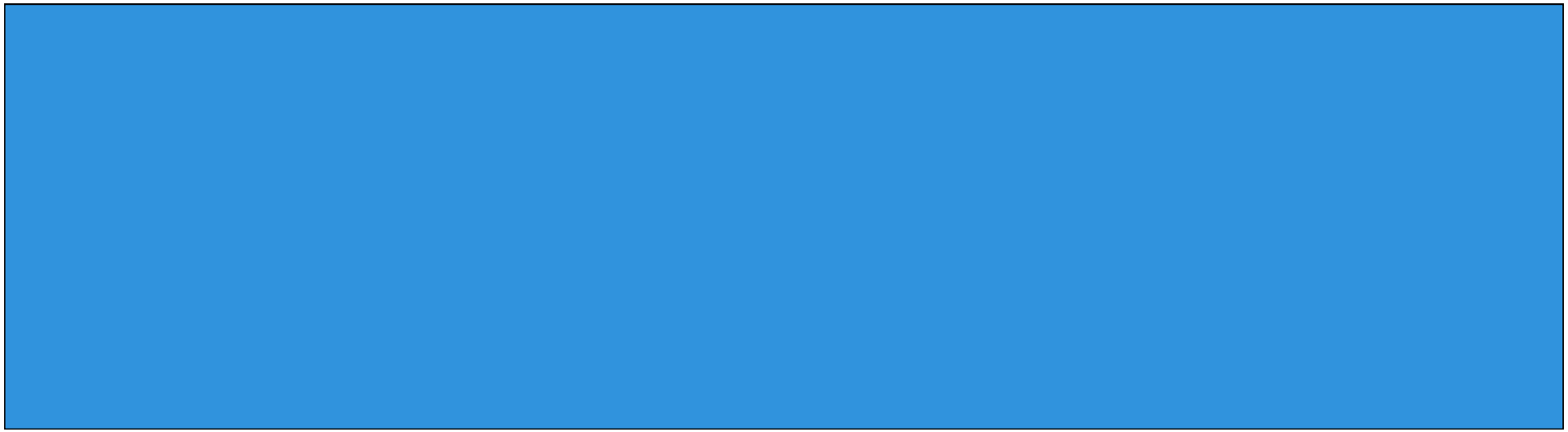
Entrepôts de données



Opérationnel vs Décisionnel

- SI **opérationnel** (SIO) : support à la réalisation des activités d'un ensemble de processus métier.
- Par exemple,
  - pour un processus de vente, enregistrement des commandes des clients, et expédition de leurs articles
  - Pour une gestion RH, enregistrement des informations sur les salariés, les contrats, les salaires, génération des fiches de paie, …
- Bases de données relationnelles, normalisées
- Applications « OLTP » (online transaction processing)

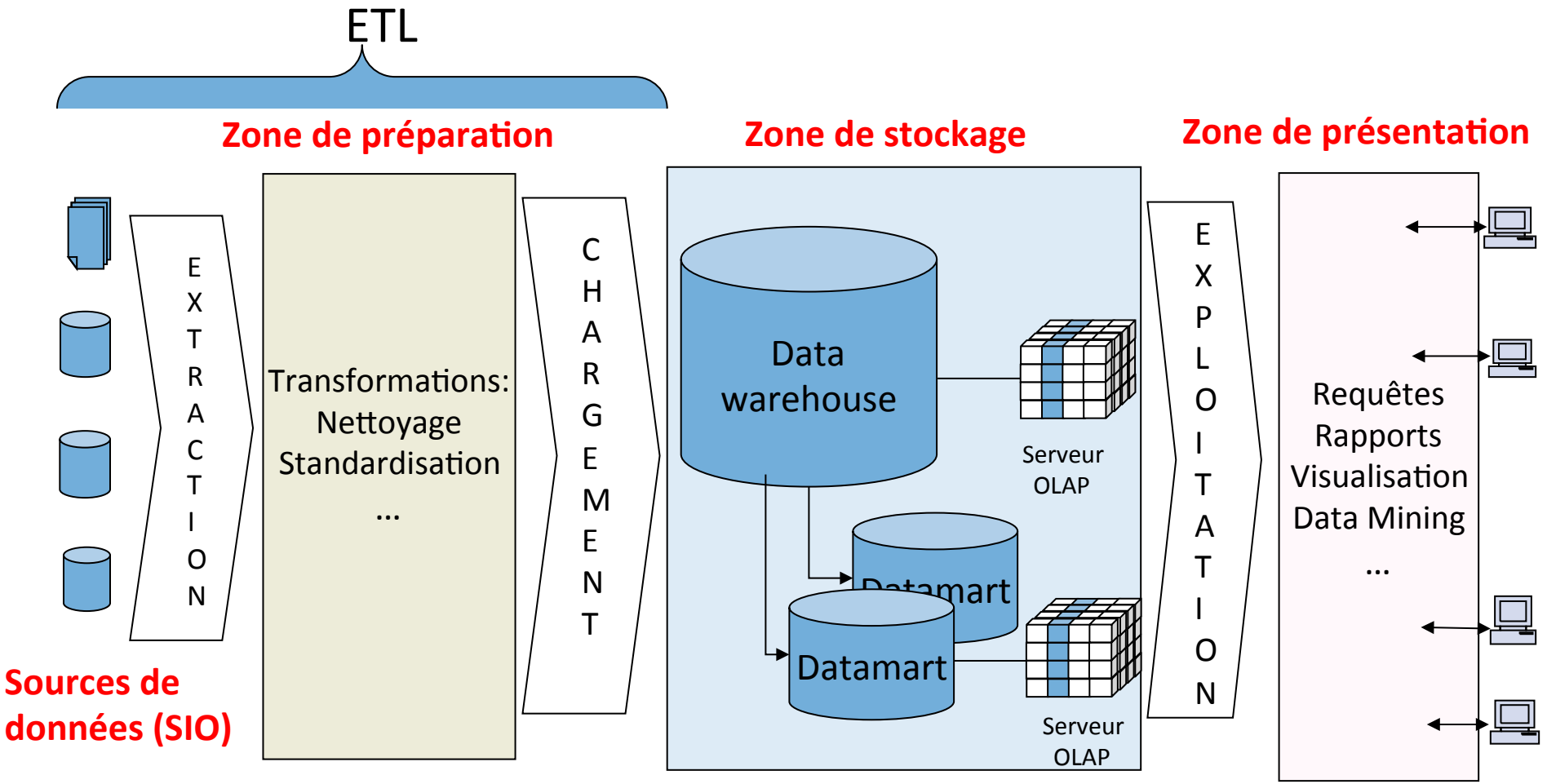
Opérationnel vs Décisionnel

- SI **décisionnel** (SID) : support à l'évaluation de la performance des processus, prise de décision.
- Par exemple,
  - Pour un processus de vente, quelle est l'évolution du chiffre d'affaire pour chaque catégorie de produit, quelle est la répartition des clients par secteur géographique, …
  - Pour un processus RH, quelle est la répartition du taux de CDD par département, …
- BDD multidimensionnelles, dénormalisées
- Applications OLAP (online analytics processing)

Opérationnel vs Décisionnel

Données opérationnelles	Données décisionnelles
Orientées application, détaillées, précises au moment de l'accès.	Orientée activité (thème, sujet), condensées, représente des données historiques.
Mise à jour interactive possible de la part des utilisateurs (insert, delete, update)	Pas de mise à jour interactive, données utilisées pour de l'analyse (reporting), donc accès en lecture principalement.
Accédées de façon unitaire par une personne à la fois (transactions unitaires).	Utilisées par l'ensemble des analystes, gérées par sous-ensemble (en lecture seule, pas de verrou sur les données).
Haute disponibilité en continu. Données critiques qui ne peuvent pas être perdues (sauvegardes/ réplication, …). L'accès au données doit être rapide.	Exigence différente, haute disponibilité ponctuelle. Les données récentes peuvent être rechargées à partir du SIO en cas de perte des données. L'accès aux données peut être lent (car requêtes complexes et volumétrie élevée).
Uniques (pas de redondance en théorie, schéma relationnel normalisé 3NF)	Peuvent être redondantes (pas de normalisation 3NF), modèle multi-dimensionnel.
Données dédiées au métier de l'entreprise, utilisées potentiellement par tous les salariés.	Données dédiées aux prises de décision, utilisées par les analystes et les décideurs.
Petite quantité de données utilisées par un traitement (en général), requêtes « simples »	Grande quantité de données utilisée par les traitements, requêtes complexes (agrégations, calculs)
Réalisation des opérations au jour le jour	Cycles de vie différents, données collectées sur le long terme.
Volume de données assez faible	Volume de données très grand, car données historisées et redondance.

Architecture générale du SID



Architecture générale

Sur la figure précédente, on remarque une distinction entre :

- **Data warehouse** (entrepôt de données), collection centralisée de toutes les données des processus métiers de l'entreprise
- **Datamart** (magasin de données), données relatives à un besoin particulier, un processus métier. Par exemple données RH.

Le data warehouse et/ou les datamarts forme(nt) le cœur du SID.

Entrepôt de Données (Data Warehouse)

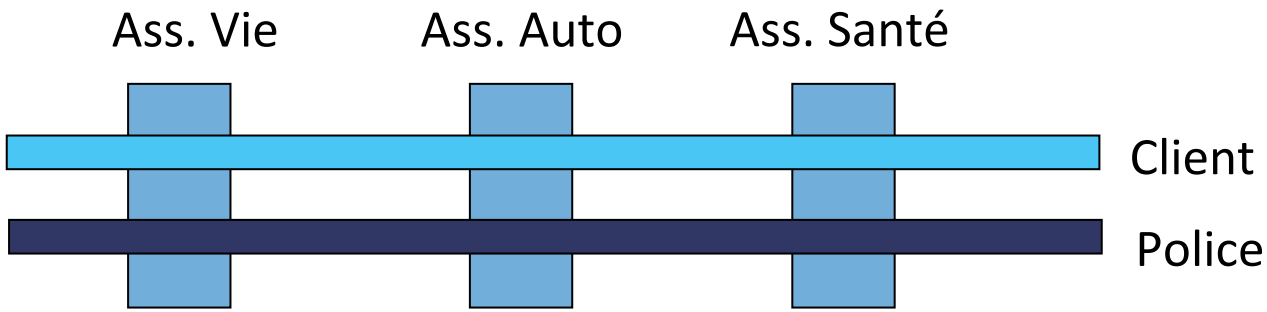
Définition de Bill Inmon (1996) **:**

«Le Data Warehouse est une collection de données orientées sujet, intégrées, non volatiles et historisées, organisées pour le support d'un processus d'aide à la décision.»

Data Warehouse

Les 4 caractéristiques des data warehouse

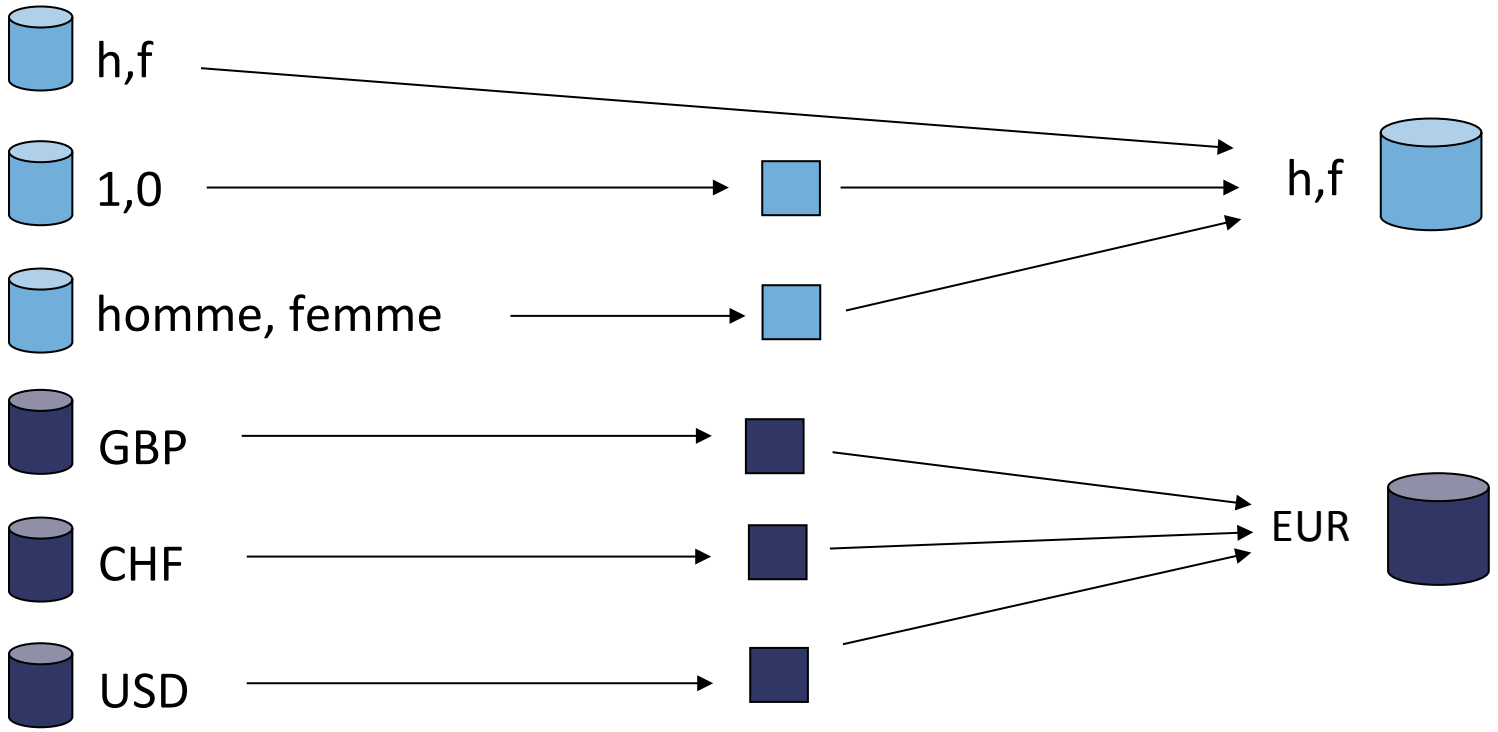
1. Données orientées sujet:
   - Regroupe les informations par thèmes ou par différents métiers (clients, produits, risques, …)
   - Pas la même organisation que le SIO : le sujet est transversal aux structures fonctionnelles de l'entreprise.



Les 4 caractéristiques des data warehouse

2. Données intégrées:

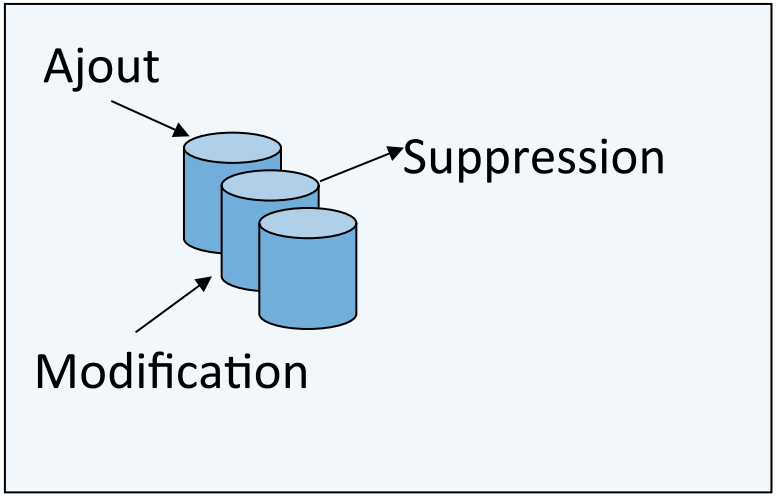
normalisation des données, définition d'un référentiel unique



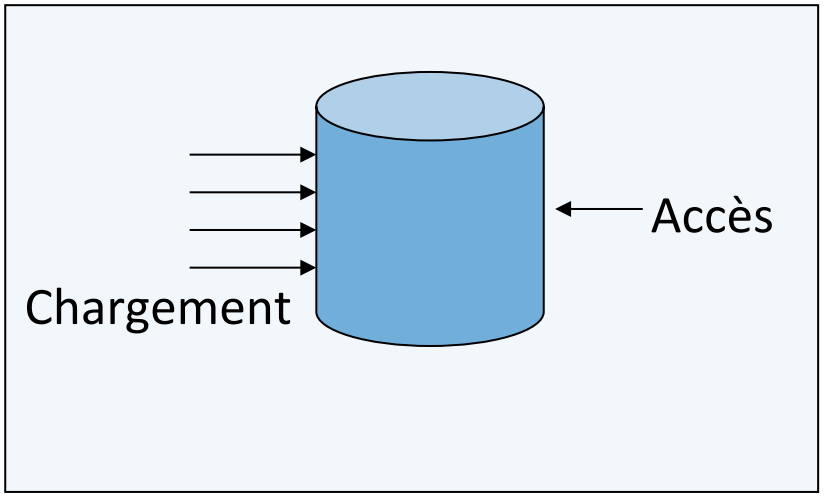
Les 4 caractéristiques des data warehouse

3. Données non volatiles

- Pas de changement au fil du temps (read-only)
- Traçabilité des informations et des décisions prises.
- Copie des données de production, uniquement des inserts



Bases de production (SIO)



Entrepôts de données (SID)

Les 4 caractéristiques des data warehouse

4. Données historisées/datées

- Les données persistent dans le temps
- Mise en place d'un référentiel temps (horodatage)

Base de production

Image de la base en Mai 2005

Répertoire

Nom	Ville
Dupont	Paris
Durand	Lyon

Image de la base en Juillet 2006

Répertoire

Nom	Ville
Dupont	Marseille
Durand	Lyon

Entrepôt de données

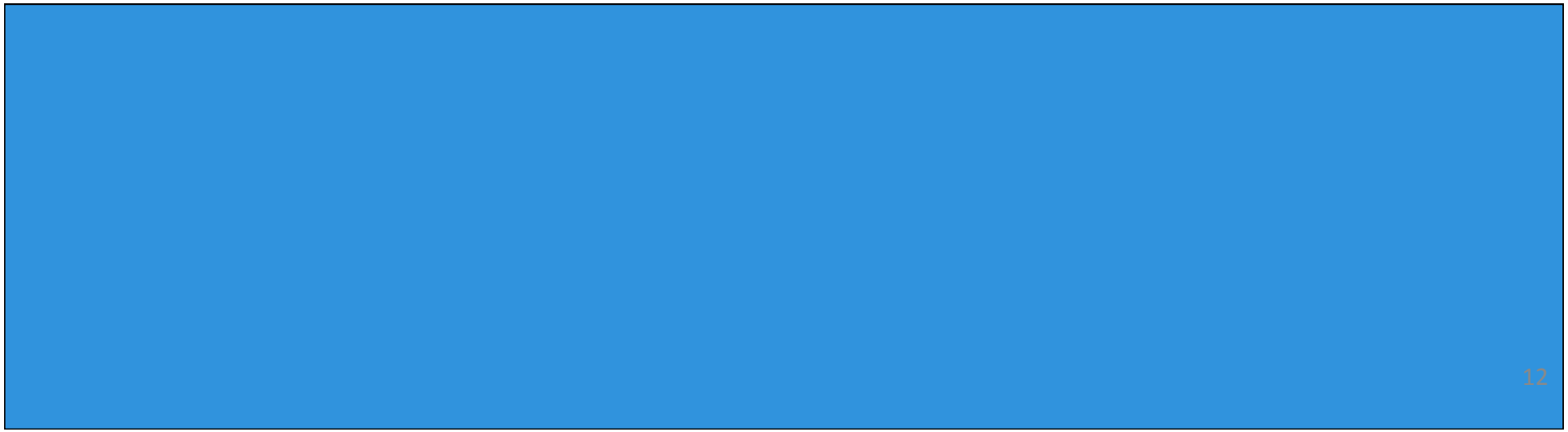
Calendrier

Code	Année	Mois
1	2005	Mai
2	2006	Juillet

Répertoire

Code	Année	Mois
1	Dupont	Paris
1	Durand	Lyon
2	Dupont	Marseille

Modélisation d'un SID



Exemple : Grande distribution

- plusieurs magasins
- quelques milliers de produits
  - unités de stock (***SKU*** ou *stock keeping units*)
  - code barre (***UPC*** ou *universal product code*)
    - pour les produits livrés par les fournisseurs (2/3 des produits)
- points de vente (***POS*** ou *point of sale*), magasins
  - avec scanning des codes à barres
- influence sur les ventes: ajustement des prix et promotion
  - réduction de prix temporaires (***TPR*** ou *temporary price reduction*)
  - présentation des rayons (shelf display)
  - présentation des têtes de gondoles (end aisle display)

Objectifs de la modélisation

1. Rendre les données facilement accessibles
   Manipulation aisée, outils conviviaux.
   Rapidité.
2. Présentation cohérente
   Données nettoyées, vérifiées, crédibles.
   Codage et représentation documentés.
   Assurer la qualité des données.
3. Architecture évolutive
   Compatible avec les nouvelles requêtes.
   Résistance aux changements.
   Compatibilité ascendante.
   Modifications documentées.

L'entrepôt de données doit servir à la prise de décision Il doit être accepté par les utilisateurs (conduite du changement).

Les utilisateurs peuvent se passer de l'entrepôt de données, pas du SI opérationnel.

Travail du concepteur

- Ecouter les utilisateurs : besoins, décisions . . .
- Cibler les « meilleurs » utilisateurs.
- Sélectionner les données pertinentes, modéliser l'entrepôt
- Concevoir des outils de visualisation/ interrogation simples.
- Contrôler la validité des données présentées.
- Enrichir constamment la base.

Recueil des besoins

OBJECTIF PRINCIPAL

Qu'attendez-vous principalement du Data Warehouse ?

DECISIONS

Quelles décisions avez-vous à prendre ? (**Quoi ?**)

Quels sont les critères qui influencent la prise de décision ? (**Comment ?**)

Dans quel(s) but(s) les décisions sont-elles prises ? (**Pourquoi ?**)

DIFFICULTES ACTUELLES

Quelles sont les difficultés actuellement rencontrées dans la prise de décision, difficultés en rapport avec les données

- précision des données (détails, actualisation, vérification)
- synthèse des données (regroupements)
- évolution (temps)
- autres...

ACTUALISATION DES INFORMATIONS

Quels sont les besoins concernant la fréquence de mise à jour des informations proposées par le Data Warehouse ?

PRESENTATION DES INFORMATIONS

- Quelles sont vos préférences dans la présentation des informations
- tableaux, graphiques, ?
- Type de graphiques : barres-graphes, "camemberts", nuages de points ... ?
- Existe-t-il une présentation actuelle ou habituelle à conserver ?

Modélisation

Pour une base de données opérationnelle (SIO), on utilise le modèle entité association, avec une normalisation 3NF (cf cours de licence), mais

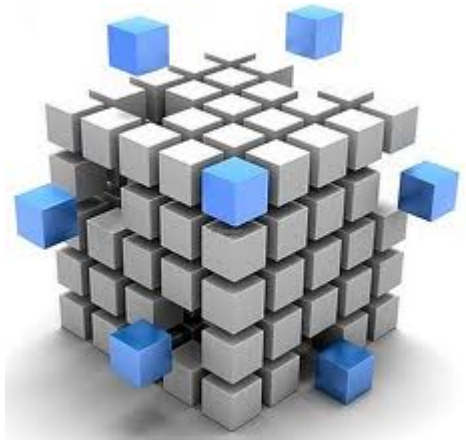
1. La normalisation 3NF entraine la création de **nombreuse tables**. L'exécution d'une requête décisionnelle requiert de **nombreuses jointures suivies d'agrégats.**
2. La normalisation 3NF se justifie par la difficulté de gérer la **redondance** dans un système OLTP, lors des **mises à jour** de la table. Cette justification disparaît dans un système en lecture seule.

Modèle multidimensionnel

- Face aux 2 points précédents, un nouveau modèle adapté aux SID a été défini : **le modèle multidimensionnel**.
- Dans ce modèle, les activités réalisées par un processus métier sont décrites en terme de **faits**, mesurés par des **indicateurs**, ainsi qu'en terme de **dimensions**.
- Un fait représente 1 événement (*la vente d'un produit à un client dans un magasin à une certaine date*)
- Un indicateur **mesure** un fait (*montant de la vente*)
- Les dimensions décrivent le contexte (qui ? *Un client*, où ? *Dans un magasin*, quoi ? *Un produit*, quand ? *A une date donnée*, …)

vision OLAP

- Une mesure = une valeur selon n dimensions, d'où la notion de (hyper)**cube** de données.
- Un fait est un cube atomique à l'intérieur d'un plus grand cube.
- Requête OLAP : manipulation de ce cube.



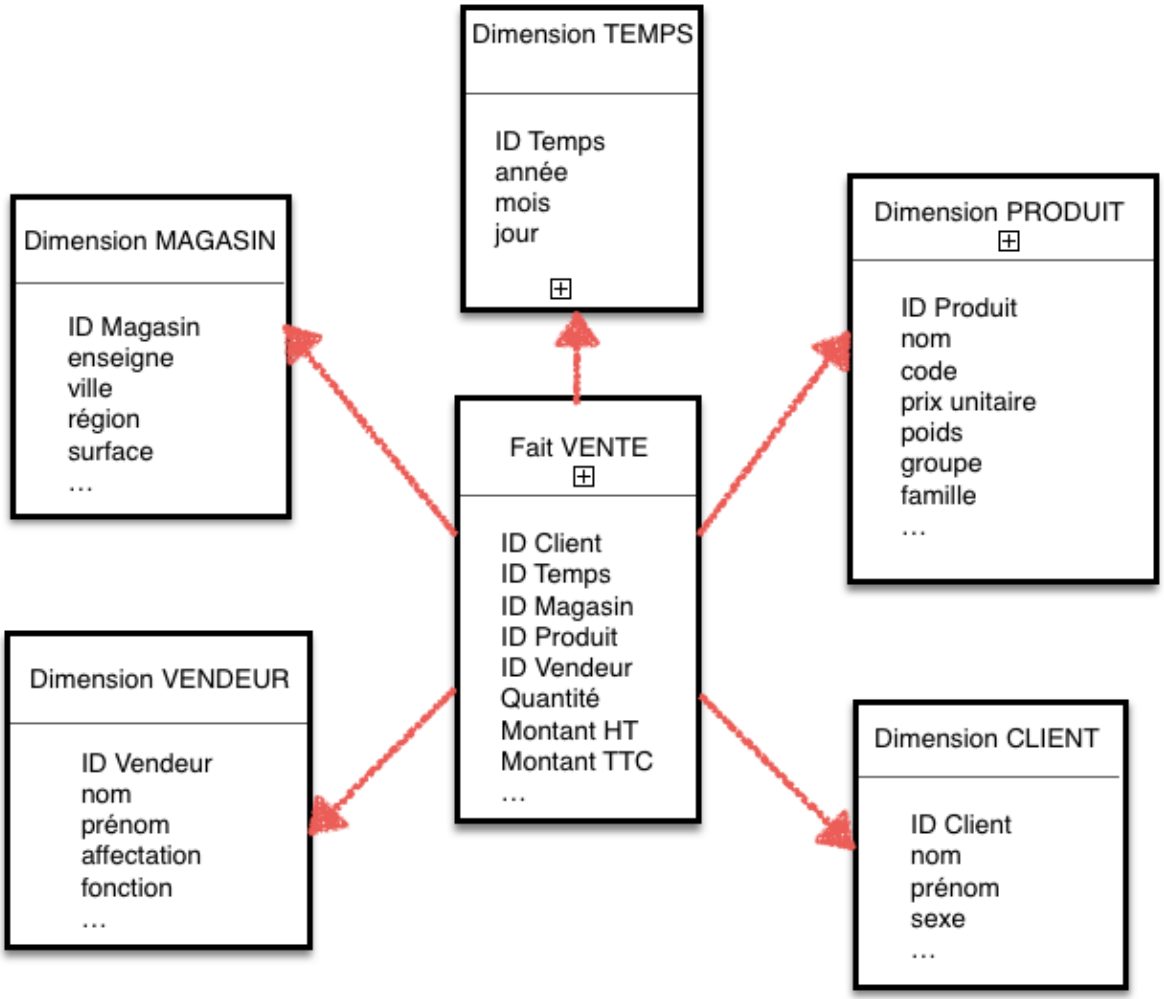
Multidimensionnel : les Faits

- Les faits sont rangés dans une table, qui est la plus **volumineuse** de la base.
- Une ligne = un fait observé à un instant T ou bien le fait qu'il s'est passé un événement lié à une activité métier (vente, contrat de travail, …).
- La table de fait contient à la fois la définition du **contexte** (clés étrangères sur les tables dimensions) et les **indicateurs**, mesures du fait.
- Il faut choisir le **grain** nécessaire **:** i.e. le niveau de détail des faits

Modèle en étoile

- une entité centrale : la table de fait
  - objets de l'analyse, traduit une activité, un événement lié à un processus métier
  - taille très importante
  - beaucoup de champs, en particulier des mesures – les indicateurs.
- des entités périphériques : les tables de dimensions
  - critères de l'analyse, le contexte du fait.
  - taille peu importante (en comparaison aux faits)
  - peu de champs

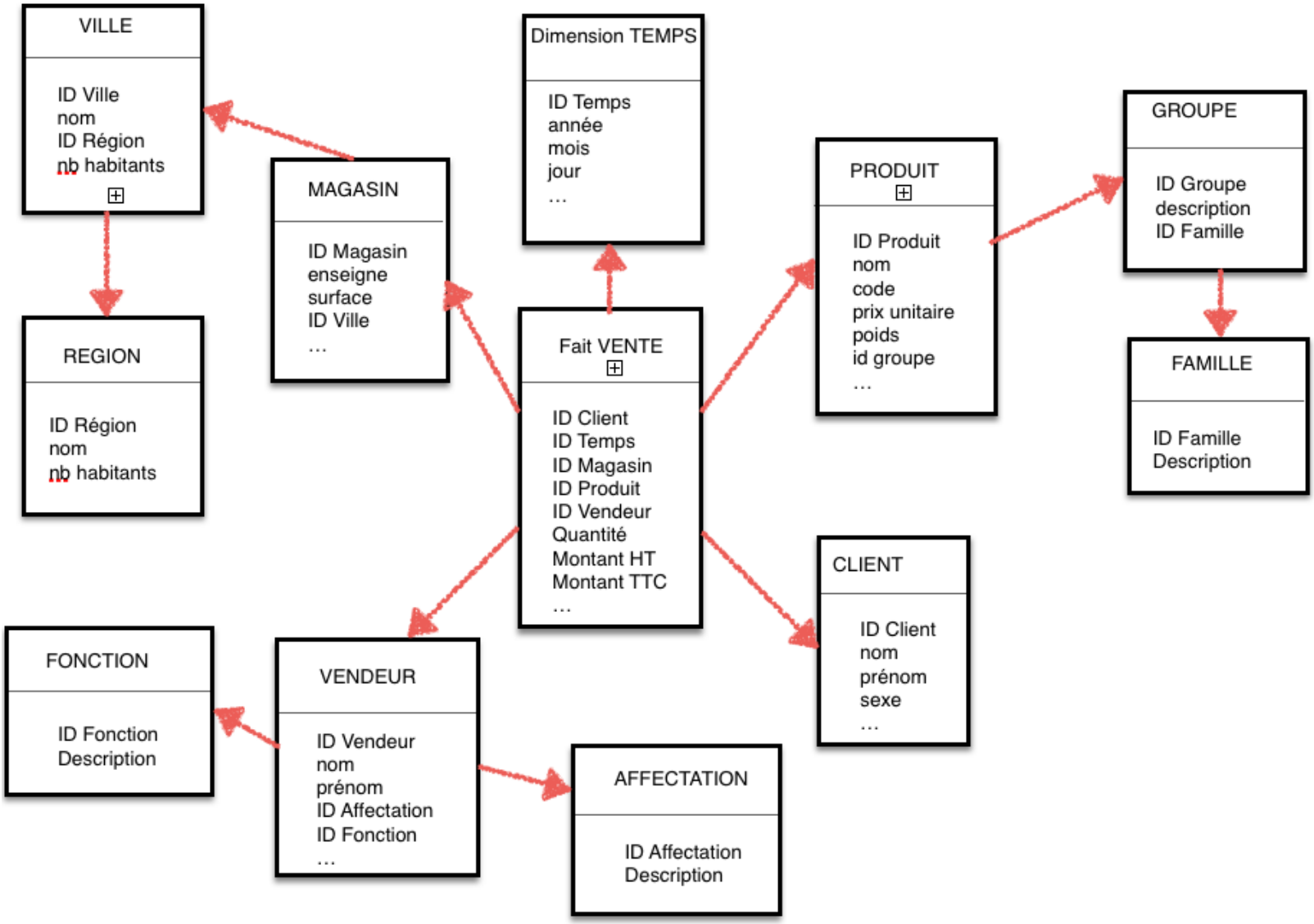
Modèle en étoile



Modèle en flocon

- On peut augmenter la lisibilité du modèle en regroupant certaines dimensions, ou en « dépliant » une dimension qui contient plusieurs niveaux de granularité.
- On définit ainsi des **hiérarchies**, qui peuvent être par exemple géographiques ou organisationnelles.
- A faire 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.

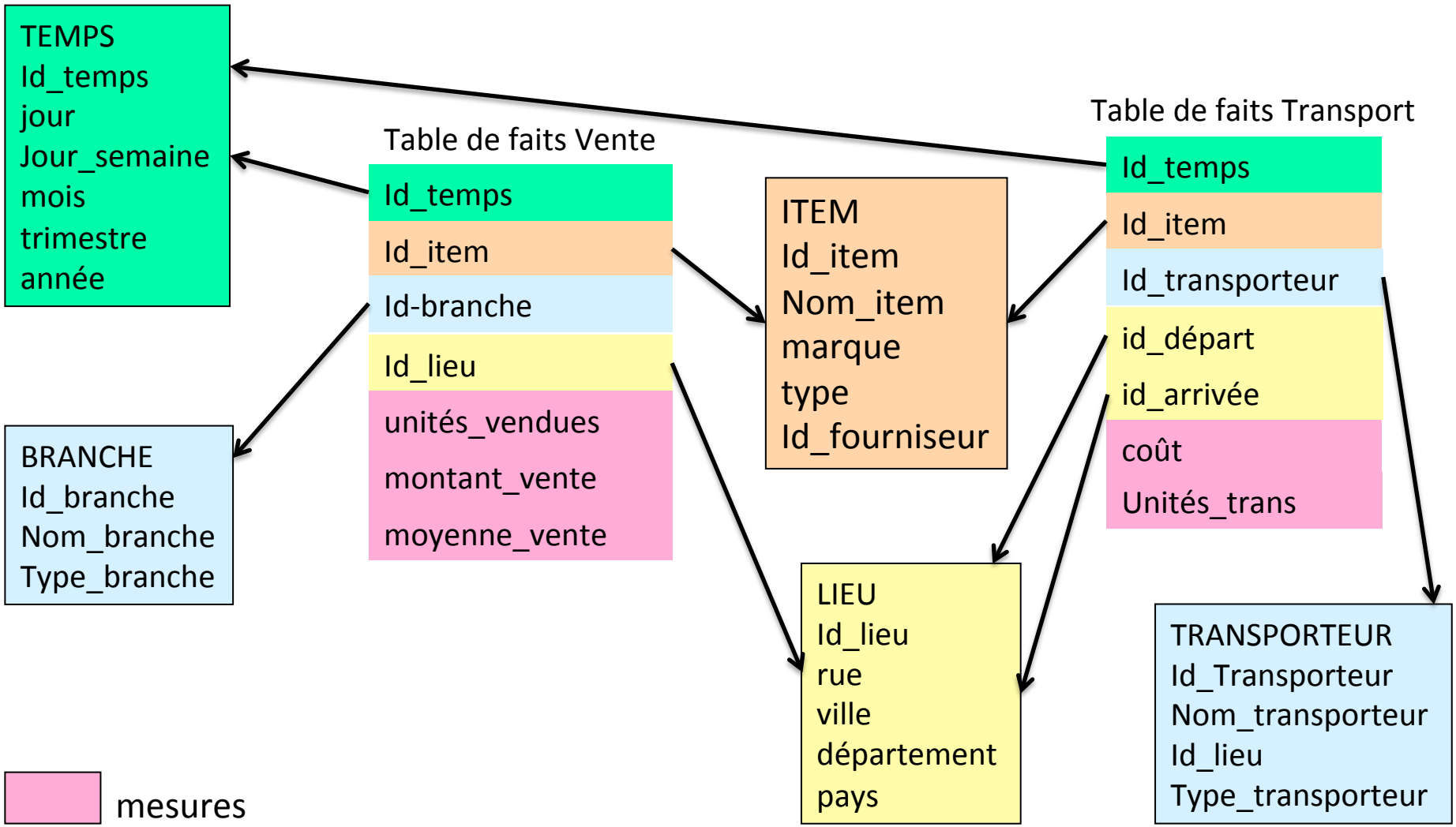
Modèle en flocon



Modèle en constellation

Lorsque le SID contient plusieurs tables de faits, qui se partagent des dimensions communes, on parle de **modèle en constellation**.

Modèle en constellation



Types de faits

- 3 principaux types :
  - **Transaction** : par exemple, la relation Achat montrée précédemment.
  - **Instantané** : photographie à un instant T. Exemples : gestion des stocks sous la forme d'un inventaire (et non le suivi des mouvements de stocks), relevés météo, ….
  - ***Instantané accumulé*** : cycle de vie d'un processus métier (par exemple, recrutement pour un service RH)

Il est ensuite possible de construire des relations factuelles plus complexes, qui font la synthèse de plusieurs relations factuelles.

Granularité / Finesse des Faits

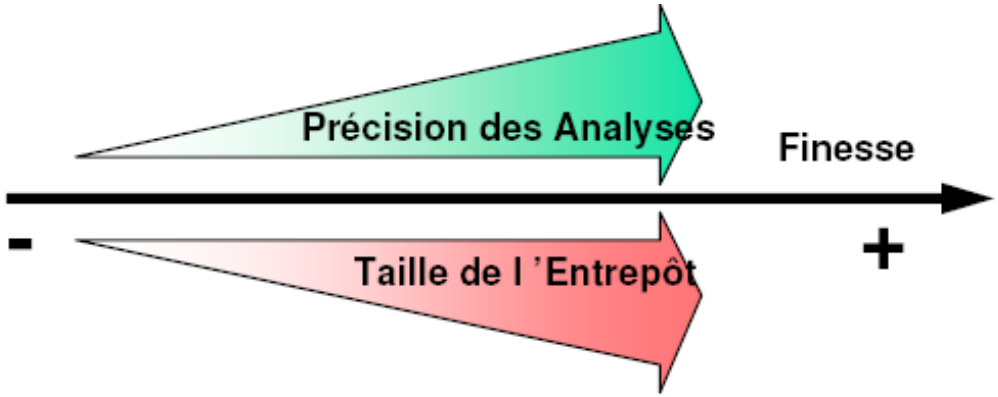
- Tables éparses

  hypothèse d ’un monde fermé :

  s ’il y a pas de fait (vente = 0$), on ne le représente pas

- Choix de la granularité

  les informations exprimées dans le grain le plus bas possible car les requêtes font souvent des coupes dans l ’entrepôt selon des critères précis



Grain - exemples

- TEMPS : mouvement journalier des articles vendus
  - ou + détaillé : pour chaque transaction
    - chaque passage au point de vente (un ticket client)
    - grand volume
    - si identification du client (carte de fidélité, …), analyse plus fine du comportement
  - Ou – détaillé : enregistrement hebdomadaire (ou mensuel)
    - mais ignorance de nombreux phénomènes
      - mesure des actions quotidiennes
      - analyse par jour (différence entre lundi ou week-end, …)
- PRODUIT : au niveau du SKU (unité de stock)
  - ou par marque, unité d ’emballage, …

Additivité des mesures de Fait

- Plusieurs millions de faits à résumer
  - compter les faits
  - additionner les mesures
- Propriété d ’additivité
  - Fait additif : additionnable suivant toutes les dimensions
  - Fait semi additif : additionnable seulement suivant certaines dimensions
  - Fait non additif : non additionnable quelque soit la dimension

Additivité des mesures de Fait

Exemples :

- Fait additif

  quantité vendue, montant vente HT, montant vente TTC

- Fait semi additif

  niveau de stock : additif sauf sur la dimension **temps** car **valeur instantanée** (cf types de relations factuelles).

- Fait non additif

  Ratio, marge brute = 1 - Coût/CA

Dimensions

- Les dimensions sont utilisées pour analyser les faits. On parle aussi **d'axe d'analyse**.
- Une dimension comporte un niveau (comme la dimension client), ou bien plusieurs niveaux formant une ou plusieurs hiérarchies (comme la dimension temps).

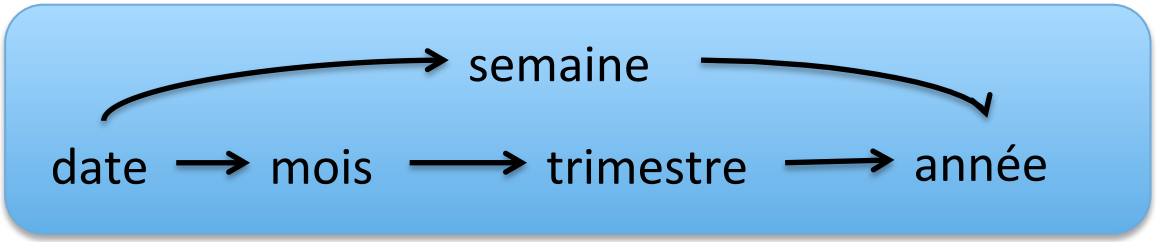
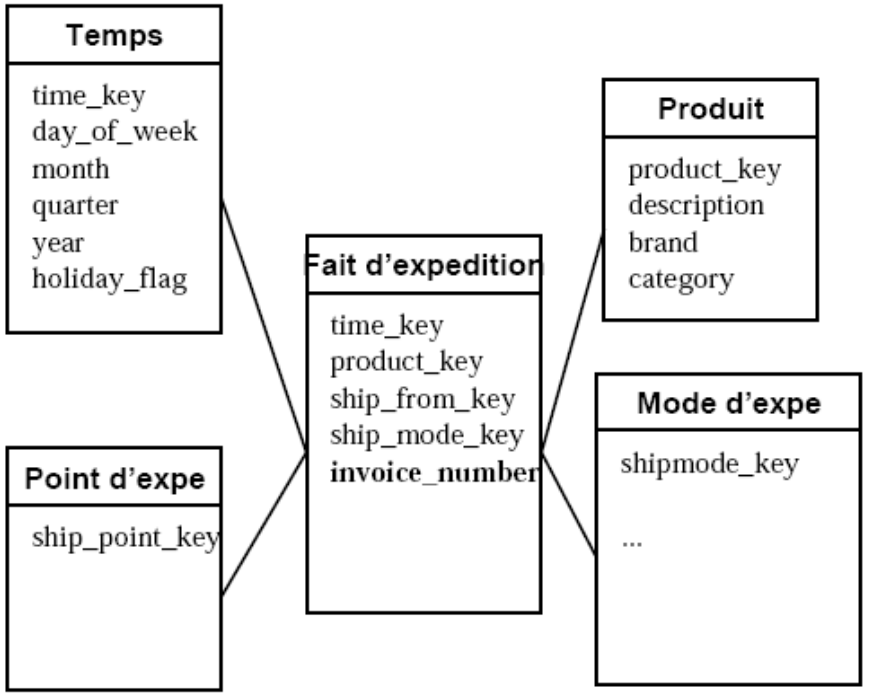


Table de dimension

- Une dimension est en général traduite en une table.
- Pour des raisons de volumétrie, on peut décomposer une dimension hiérarchisée en plusieurs tables en utilisant une normalisation 3NF. On obtient ainsi un modèle en Flocon (cf dimension Produit avec Groupe et Famille).

Dimension Dégénérée

- Dimension sans attribut
  - Pas de table ➨ la clé de dimension est dans la table de fait
- Exemples :
  - numéro de facture (invoice number),
  - numéro de ticket



Dimension Temps

- Il existe toujours une dimension Temps, liée à l'historisation.
- 2 choix d 'implantation
  - Type SQL DATE (dimension dégénérée dans la table des faits, mais souvent insuffisant)
  - **Calendrier + Table dimension Temps**
- Sémantique du temps
  - Validation : date d'occurrence du fait
  - Intégration : date de prise en compte dans l 'entrepôt

Dimension Temps

- L'utilisation d'une table dimension Temps permet d'ajouter des informations supplémentaires : événement (match de finale de coupe du monde) ; jours fériés, vacances, période fiscale ; saison haute ou basse, …
- De plus, une table dimension Temps permet de stocker des dates auxquelles il n'y a pas de fait associé.

Dimension Temps

- Le type SQL Date seul (dimension dégénérée) est donc bien souvent insuffisant mais une date peut servir de clé de la table Temps
- chaque jour peut être caractérisé par
  - le jour de la semaine (lundi, …)
  - le numéro du jour du mois (1, 2, …)
  - s 'il est le dernier jour du mois (O/N)
  - le numéro du jour (calendrier julien à partir d 'une date donnée)
  - le numéro de semaine dans l 'année (1, 2, … 52)
  - le numéro du mois (1, 2, …12)
  - le mois (janvier, février, …)
  - le mois d'une année (avril 2014, …)
  - l'année (2014, 2015, …)
  - le trimestre (1er, 2ème, …)
  - la période fiscale
  - s 'il est férié ou non
  - la saison (printemps, été, …)
  - un événement (final de foot, …)
  - ...

Dimension *TEMPS*

id temps
jourSemaine
noJourMois
dernJour
noJour
noSemaine
noMois
mois
moisAnnée
année
trimestre
périodeFisc
férié
Saison
événement
…

Liens entre les tables

- Table de dimension
  - clé primaire simple
- Table de fait
  - clé primaire constituée (d'une partie) des clés étrangères vers les tables de dimension, ou des dimensions dégénérées.
  - Jointure entre la table de faits et les tables de dimensions

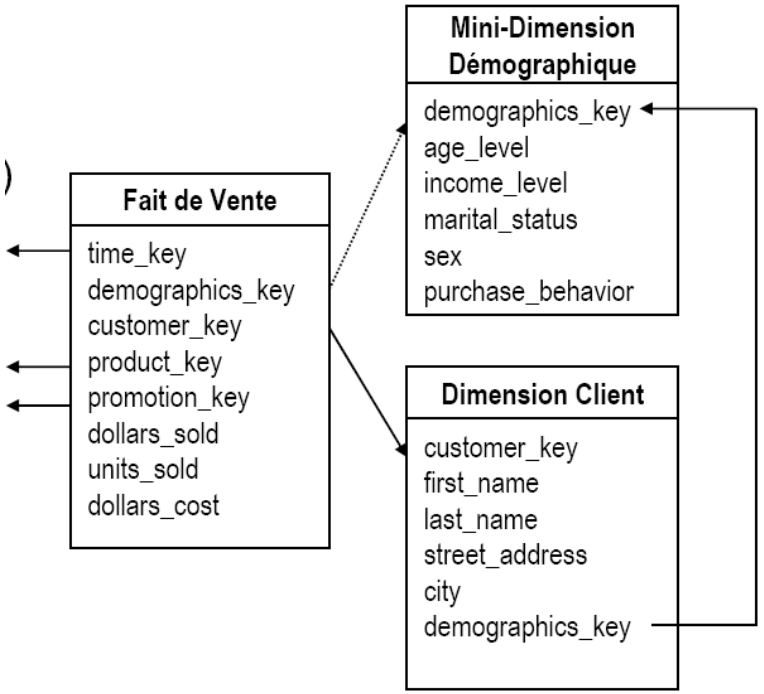
Grandes Dimensions

- Nombreux membres
- réduire la taille des tables
  - dimension Produits (300.000)
  - dimension Clients (10.000.000)
- Solutions
  - Utilisation du Flocon de Neige
    - tables de dimension secondaires (déportées) associée à une table de dimension, parente dans une hiérarchie
    - inconvénient : parfois faible gain de place et navigation compromise
  - Mini Dimensions

Mini-dimension

- Mini-Dimension : on sépare les attributs à **évolution lente** de ceux à **évolution rapide (mini-dimension)**
- Exemple : table Client

  Combinaisons (<100000) d'intervalles de valeurs démographiques



Evolution d'une dimension

- Changement de description des membres dans les dimensions
  - un client peut changer d 'adresse, se marier, ...
  - un produit peut changer de noms, de formulations («Raider» en «Twix», «Yaourt à la vanille» en «Yaourt saveur Vanille», …)
- Choix entre 3 solutions
  - écrasement de l 'ancienne valeur : en cas de correction d'une erreur, ou de renommage
  - gestion de versions : quand ça a une influence sur les statistiques (changement situation maritale, produit qui change de catégorie, …)
  - valeur d 'origine / valeur courante : moins utilisé, quand on a besoin des 2 versions simultanément

Solution 1 : Écrasement de l'ancienne valeur

Exemple : Un produit change de groupe

Clé produit	Description	Groupe
12345	Intelli-Kids	~~Logiciel~~ Jeux Educatif

- Avantage: Facile à mettre en œuvre
- Inconvénients:
  - Perte de la trace des valeurs antérieures des attributs
  - Perte de la cause de l'évolution dans les faits mesurés, et problème de cohérence : avant la modification, le produit comptait dans les chiffres du groupe « Logiciel ».

Solution 2 : gestion de versions

- Avantages:
  - Permet de suivre l'évolution des attributs
  - Permet de segmenter la table de faits en fonction de l'historique
- Inconvénient:
  - Accroit le volume de la table
  - Faire le lien entre les différentes versions du produit, pour des statistiques sur le long terme (si les clés sont différentes).
- Implémentation :
  1. Utiliser une clé de substitution, ou
  2. Intégrer une date ou un numéro de version dans la clé (clé composite)

num produit	version	Description	Groupe
12345	1	Intelli-Kids	Logiciel
12345	2	Intelli-Kids	Jeux Educatif

Solution 3 : 2 attributs (avant/après)

- Avantage:

  Avoir deux visions simultanées des données :

  - Voir les données récentes avec l'ancien attribut
  - Voir les données anciennes avec le nouvel attribut

- Inconvénient:

  Inadapté pour suivre plusieurs valeurs d'attributs intermédiaires

Clé produit	Description	Ancien groupe	Nouveau groupe
12345	Intelli-Kids	Logiciel	Jeux Educatif

Estimation de la taille de l ’entrepôt

- Dimensionner l ’entrepôt : connaître comment sont codés les attributs des tables, connaître le nombre de lignes. *Rappel : La table de fait est largement plus volumineuse que les dimensions.*
- Choix des granularités : influence sur la volumétrie. Trouver un compromis entre la précision des analyses et la taille.
- L’estimation de la taille et de son évolution dans le temps permettent le choix d ’une machine/SGBD cible (benchmark)

Codage des Clés et des Mesures

- Mesure de fait
  - valeurs entières (souvent 4 octets, parfois plus)
  - nombres flottants (4 ou 8 octets selon simple/double précision)
- Clés
  - valeurs entières (4 octets), artificielles
    - réduit la taille de l'enregistrement de fait
    - réduit le coût CPU des comparaisons de jointure
  - quelquefois, nécessité de faire à l'extraction une correspondance entre clé opérationnelle et clé entrepôt.

Exemple 1

- Dimensions
  - Temps : 4 ans * 365 jours = 1460 jours
  - Magasin : 300
  - Produit : 200 000 références GENCOD (10% vendus chaque jour)
  - Promotion : un article est dans une seule condition de promotion par jour et par magasin
- Fait
  - 1460 * 300 * 20000 * 1 = 8,76 milliards d ’enregistrements
  - Nb d’attributs de clé = 4
  - Nb d’attributs mesures (sur 4 octets) = 4
- Table des Faits

  Volume = 8,76.10^9 * 8 attributs * 4 octets = 280 Go

Exemples 2 et 3

- Exemple : Ligne d ’articles en Grande Distribution
  - Temps : 3 ans * 365 jours = 1095 jours
  - CA annuel = 80 000 000 000 $
  - Montant moyen d ’un article = 5 $
  - Nb d’attributs de clé = 4
  - Nb d’attributs mesures = 4
  - Nombre de Faits = 3*(80.10^9/5) = 48.10^9
  - Table de Faits= 48.10^9 * 8 attributs * 4 octets = 1,5 To
- Exemple : Suivi d ’appels téléphoniques
  - Temps : 3 ans * 365 jours = 1095 jours
  - Nombre d ’appels par jour = 100 000 000
  - Nb d’attributs de clé = 5
  - Nb d’attributs mesures = 3
  - Table des Faits = 1095*10^8 * 8 attributs * 4 octets = 3,5 To