http://www-adele.imag.fr/users/Didier.Donsez/cours

Conception de Bases Décisionnelles

Didier DONSEZ

Université Joseph Fourier (Grenoble 1)

PolyTech'Grenoble LIG/ADELE

Didier.Donsez@imag.fr

Didier.Donsez@ieee.org

Plan

- Bases de Données Transactionnelles
- La Modélisation Dimensionnelle
- Faits et Dimensions
- Additivité des Attributs
- Mini Dimensions
- Dimensions à évolution lente
- Tables de Faits sans faits
- Estimation de la taille d ’un entrepôt
- Conclusion et Bibliographie

Bases de Données Transactionnelles (Online Transaction Processing)

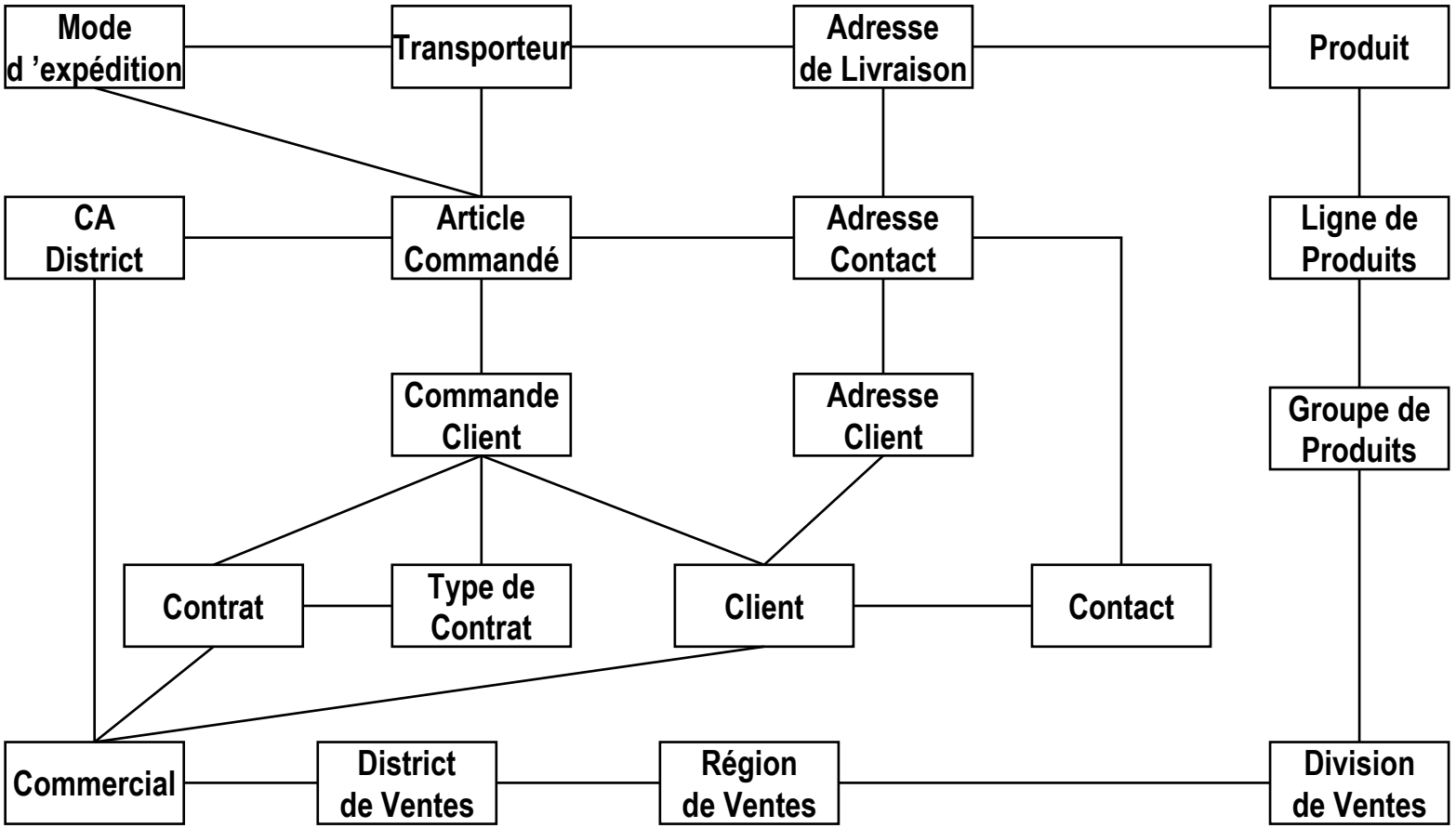
- A quoi sert la normalisation relationnelle
  - Modèle de Dépendances de Données
  - But : Eviter la redondance
  - Inconvénient : Analyse difficile de l ’activité
- Le Temps
  - instantanée de l ’activité
    - BD en changement dite BD « scintillante »
      - besoin de données stables pour des analyses
  - représentation du passé
    - un fardeau pour les systèmes OLTP

Exercice

- Ma table historique
  - Compte(NC, DateOp, Solde)
- Questions :
  - Quel est le solde courant de mon client 525 ?
    - ```
      SELECT    Solde
      FROM      Compte
      WHERE     NC=525
      AND       DateOp=(    SELECT      MAX(DateOp)
                            FROM        Compte
                            WHERE       NC=525
                        )
      ```
  - Quels sont les soldes courants de mes clients ?
      ```

Bases de Données Transactionnelles

- Inconvénient : Analyse de l ’activité par un non-informaticien



Modèle d’après Kimball 97, pp XIX

Objectifs de l ’Entrepôt de Données (ou Base Décisionnelle)

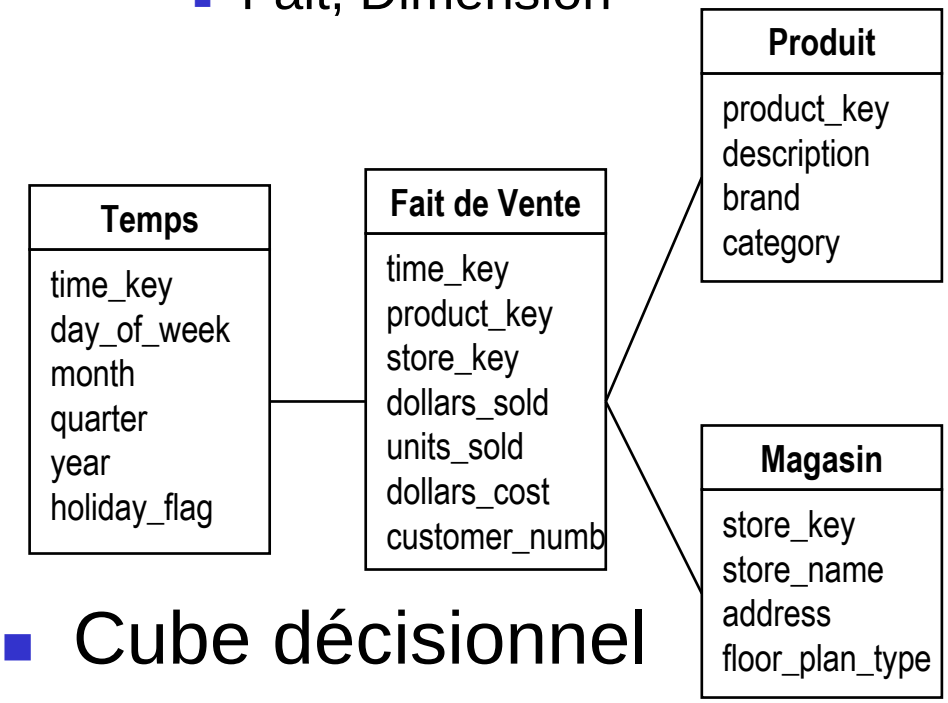
- Accessibilité des informations
  - facile à comprendre donc à utiliser
- Information cohérente
  - idempotence avec le temps
  - incomplétude signalée
- Manipulation des mesures de l ’activité
  - combinaison et séparation (tranches et dès)
- Ensemble de données et de moyens
  - requêtes, analyse, présentation, …
- Publication de données déjà servies

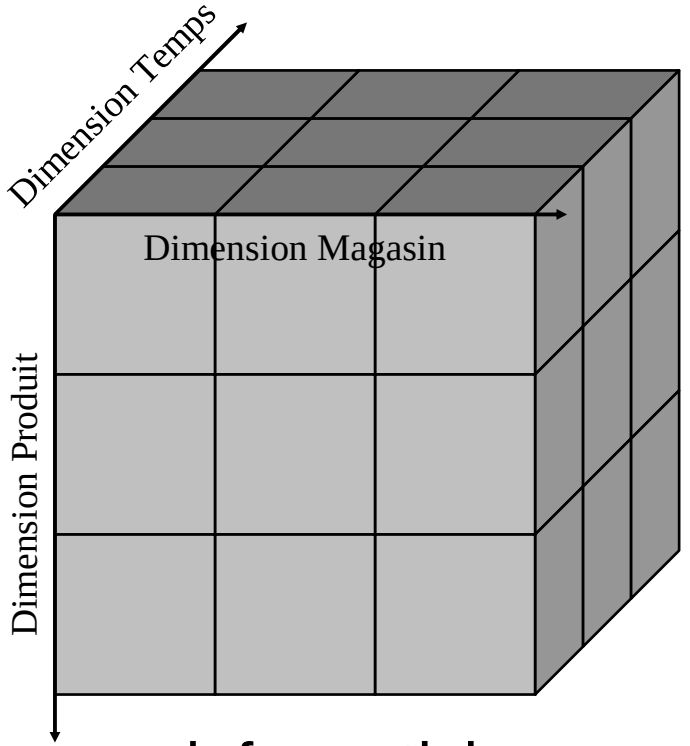
Deux mondes différents

	Information	Le Temps
OLTP	Non redondance	Vue instantanée
DW Entrepôt de Données	Accessibilité	Historique de l ’activité

La Modélisation Dimensionnelle

- modélise l 'activité que l 'on souhaite analyser
  - Modèle en Etoile
    - Fait, Dimension

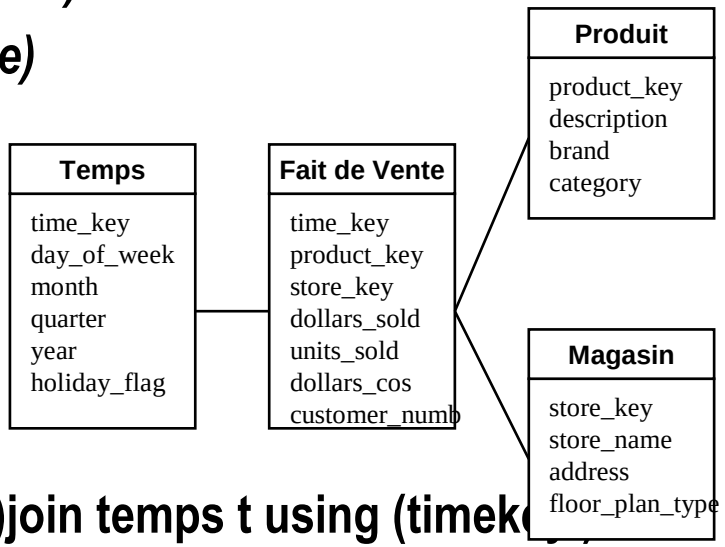




- Cube décisionnel
  - objet simple à manipuler pour des non-informaticiens

Requête Type

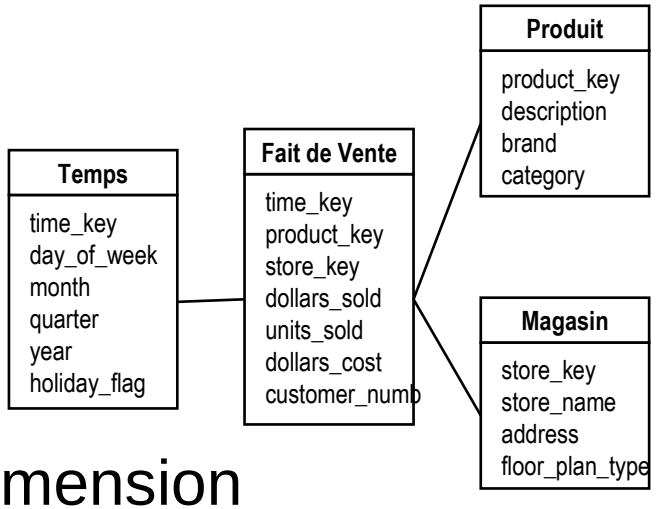
```
select p.brand, sum(fv.dollars_sold), sum(fv.units_sold)
from faitvente fv, produit p, temps t
where fv.productkey = p.productkey  (contrainte de jointure)
and fv.timekey = t.timekey          (contrainte de jointure)
and t.quarter = ' 1 Q 97 '          (contrainte applicative)
group by p.brand
order by p.brand
```

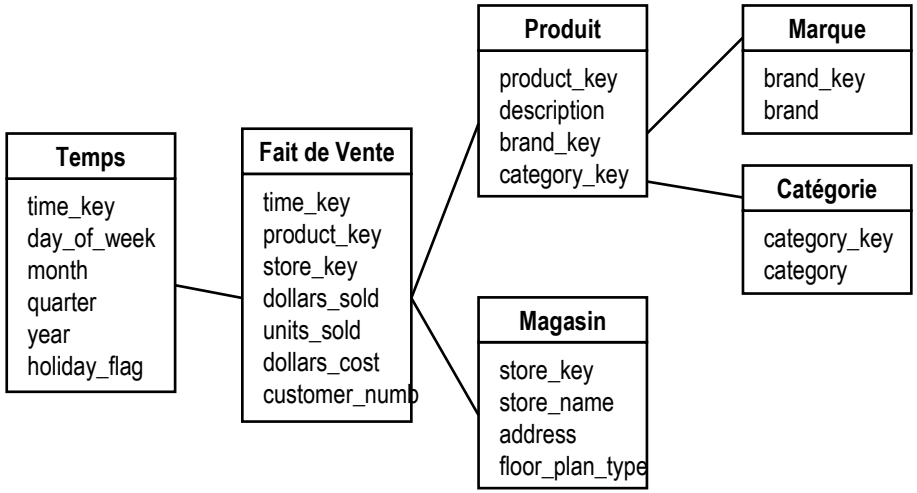



```
select p.brand, sum(fv.dollars), sum(fv.units)
from (faitvente fv join produit p using (productkey))join temps t using (timek
where t.quarter = ' 1 Q 97 '        (contrainte applicative)
group by p.brand
order by p.brand
```

Modèle en étoile
Modèle en flocon (de neige)

- Modèle en étoile
  - Table de Fait
  - Tables de Dimension (1 niveau)
- Modèle en flocon de neige
  - Table de Fait
  - Plusieurs niveaux de Tables de Dimension





Résister à la Normalisation

- Modèle en étoile
  - Taille de dimension plus grosse
- Modèle en flocon
  - Jointures pour reconstruire

- Modèle en étoile >> Modèle en flocon
  - car tables de dimension << table de fait

Processus de Conception

- Choisir le processus à modéliser
- Choisir le grain des faits
  - niveau de détails
    - transactions individuelles
    - récapitulatifs journaliers, mensuels, ...
- Choisir les dimensions
  - typiquement, le temps, le client, le foyer, le produit, le magasin, l ’agence, l ’agent, le contrat, le compte...
- Choisir les mesures de fait
  - de préférence des quantités numériques additives

Tables de Fait

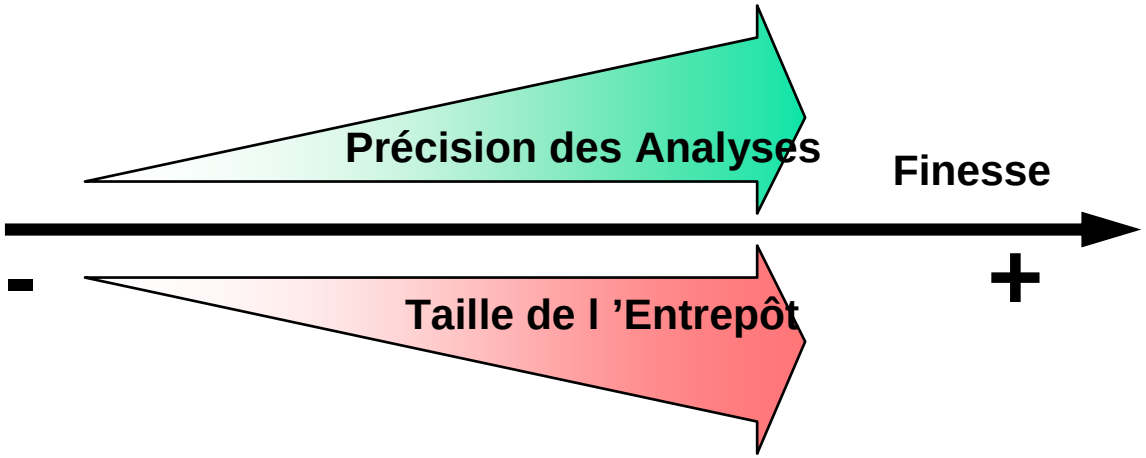
- Fait
  - Grain de mesures de l ’activité
    - chiffre d ’affaire, nombre de vente, gain, nombre de transaction …
      - en général : une valeur numérique
      - comptage des faits sinon
  - Exemple : le Fait de Vente
    - chaque enregistrement de fait représente le total des ventes d ’un produit dans un magasin dans une journée
- Table de Fait
  - relie les tables dites de dimension
- Plusieurs Tables de Fait dans un DW

Tables de Dimension

- Membre d ’une dimension
  - membre spécifique munie de caractéristiques propres
- Description
  - en général textuelle
  - parfois discrète (ensemble limité de valeurs)
    - parfum de glace, couleur d ’habit, …
- Utilisation
  - contrainte applicative
  - entête de ligne (dans des tableaux)
- Remarque importante et Rappel
  - Tables de dimension << Table de fait

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
- Niveau de détail de représentation
  - journée > heure du jour
  - magasin > rayonnage
- Choix de la granularité



Clés dans l ’entrepôt

- Tables de dimension
  - clé primaire
- Tables de fait
  - clé composite ou concaténée
    - clés étrangères des tables de dimension
    - utilisée dans les contraintes de jointure naturelle
- Choix des clés d ’une table de dimension
  - Taille d ’un fait et Coût des comparaisons de jointures
    - valeurs entières anonymes (4 octets)
  - Clés étendues
    - 2 mêmes produits de couleurs différentes = 2 membres
    - Dimension à évolution lente

Additivité des Attributs 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 additionable quelque soit la dimension
      - comptage des faits ou affichage 1 par 1

Additivité des Attributs de Fait

- Exemple
  - quantité vendue, chiffre d ’affaire, coût, nombre de clients, nombre d ’appel ...
- Fait additif
  - quantité vendue, chiffre d ’affaire, coût
- Fait semi additif
  - niveau de stock, de solde (valeurs instantanées)
    - excepté sur la dimension temps
  - nombre de transaction, de client
    - excepté sur la dimension produit
- Fait non additif
  - ex: un attribut ratio
    - ex: marge brute = 1 - Coût/CA

Mesure de Fait Semi-additive

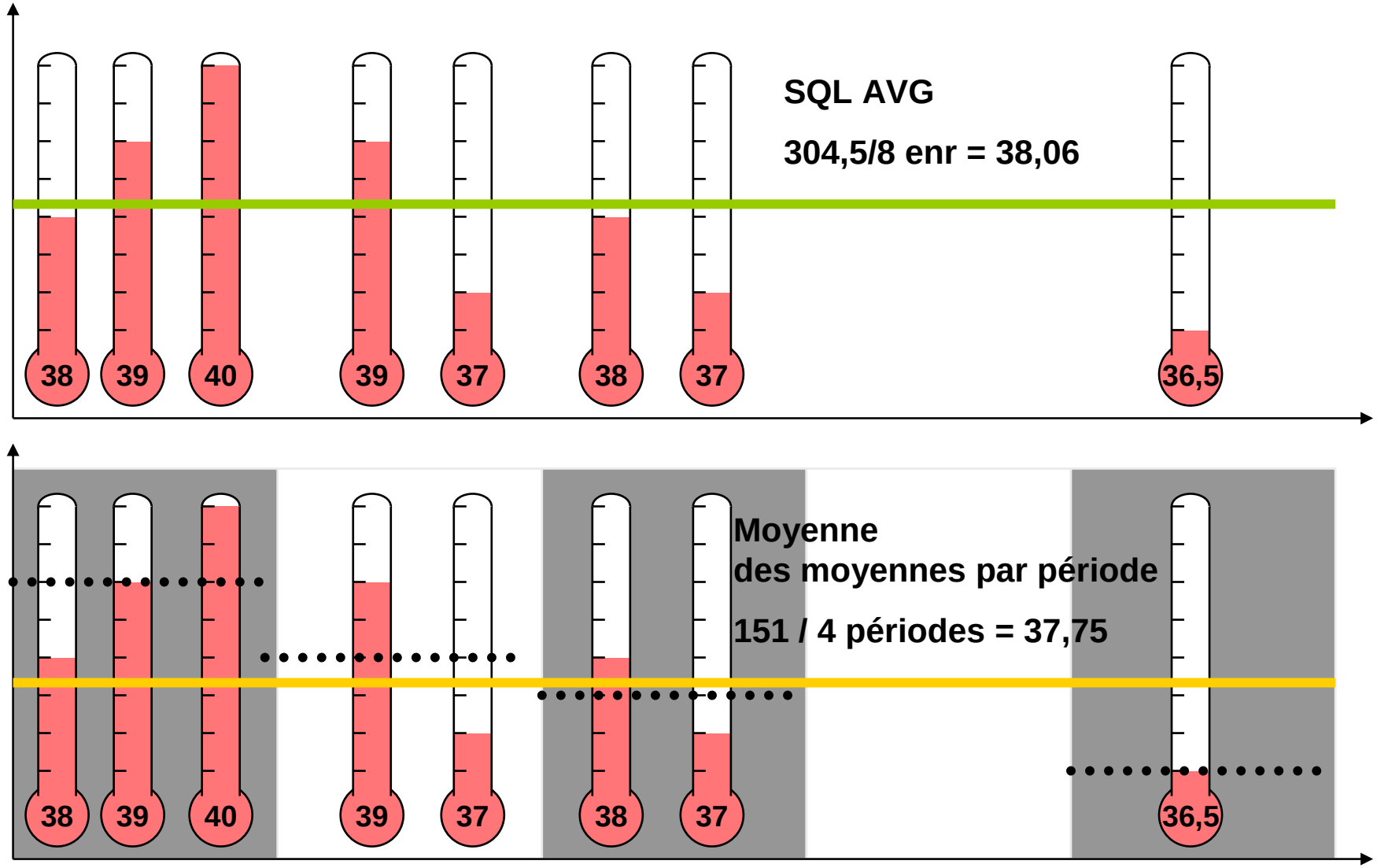
Nombre de Clients, Nombre de Transactions, ...

- Exemple : le nombre de clients et la dimension Produit
  - Soient deux faits (même magasin, même jour)
    - (Papier essuie tout, 20 clients) et (Mouchoir, 30 clients)
  - La somme du nombre de clients sur la dimension Produit n ’a pas de signification
    - car un client peut avoir acheté des mouchoirs et du papier.

- sert uniquement de contrainte applicative
  - nombre de clients ayant acheté des mouchoirs (par mois)

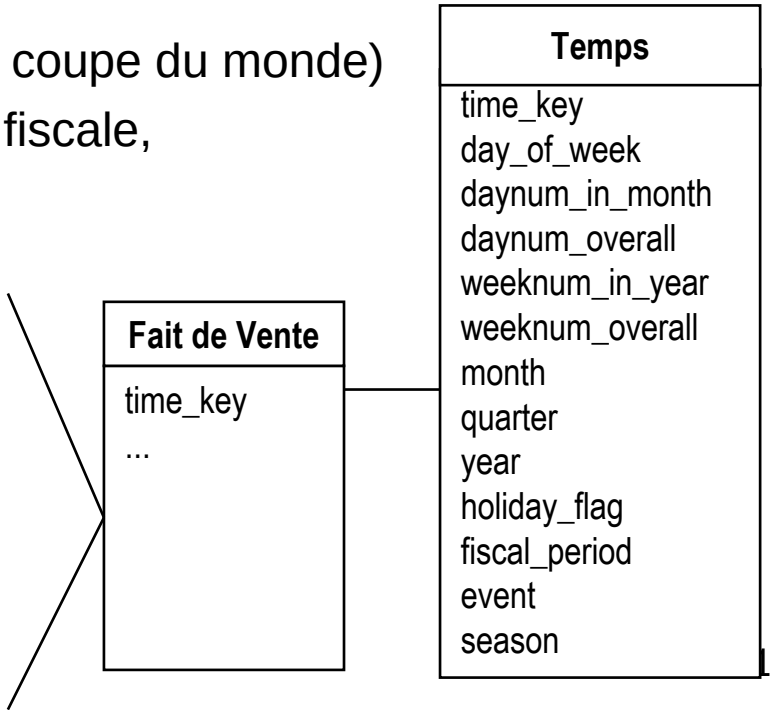
Autre Mesure de Fait Semi-additive

Température patient, Niveau de Stock, ...



Dimension Temps

- Commune à tout entrepôt
- Relié à toute table de fait
- 2 choix d ’implantation
  - Type SQL DATE
  - Calendrier + Table Temps
    - informations supplémentaires
      - événement (match de finale de coupe du monde)
      - jours fériés, vacances, période fiscale,
      - saison haute ou basse, …
- Sémantique du temps
  - Validation
    - occurrence du fait
  - Transaction
    - prise en compte dans l ’entrepôt



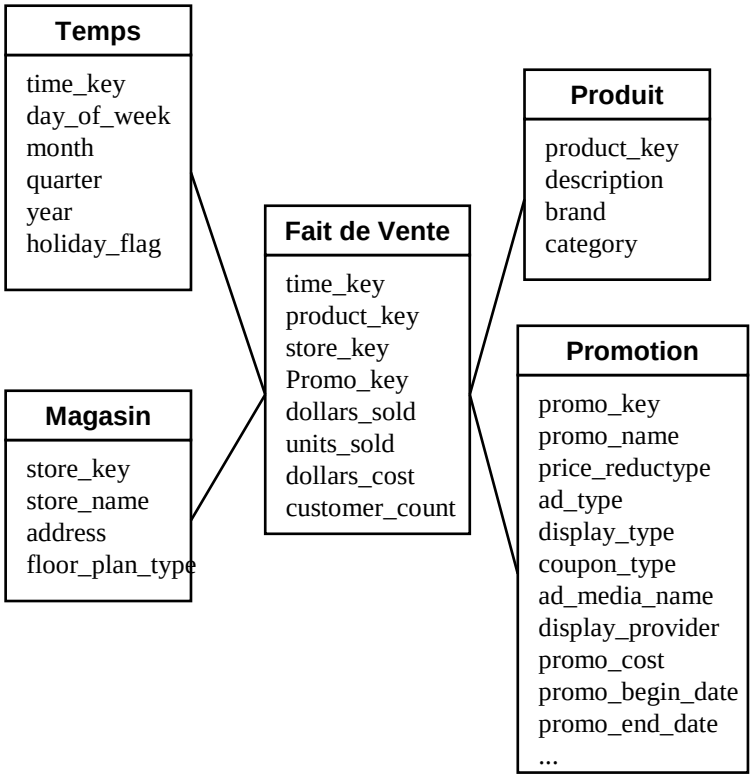
Dimension Temps

- Plusieurs notions de date dans l'entrepôt
  - Date de l'événement
  - Date de transaction
  - Date de chargement
  - Date de requête

- Cf SGBDs Temporels
  - Temps de référence pour les requêtes
    - Quel était le nombre de clients
      quand il était Noël ?
      - Les chargements effectués après Noël ne sont pas pris en compte
  - Voir Chris Date, « Introduction aux Bases de Données », 7ème édition, Chapitre 22

Dimension Causale

- dimension qui provoque le fait
  - ex: la dimension Promotion
    est supposée avoir provoqué le Fait de Vente

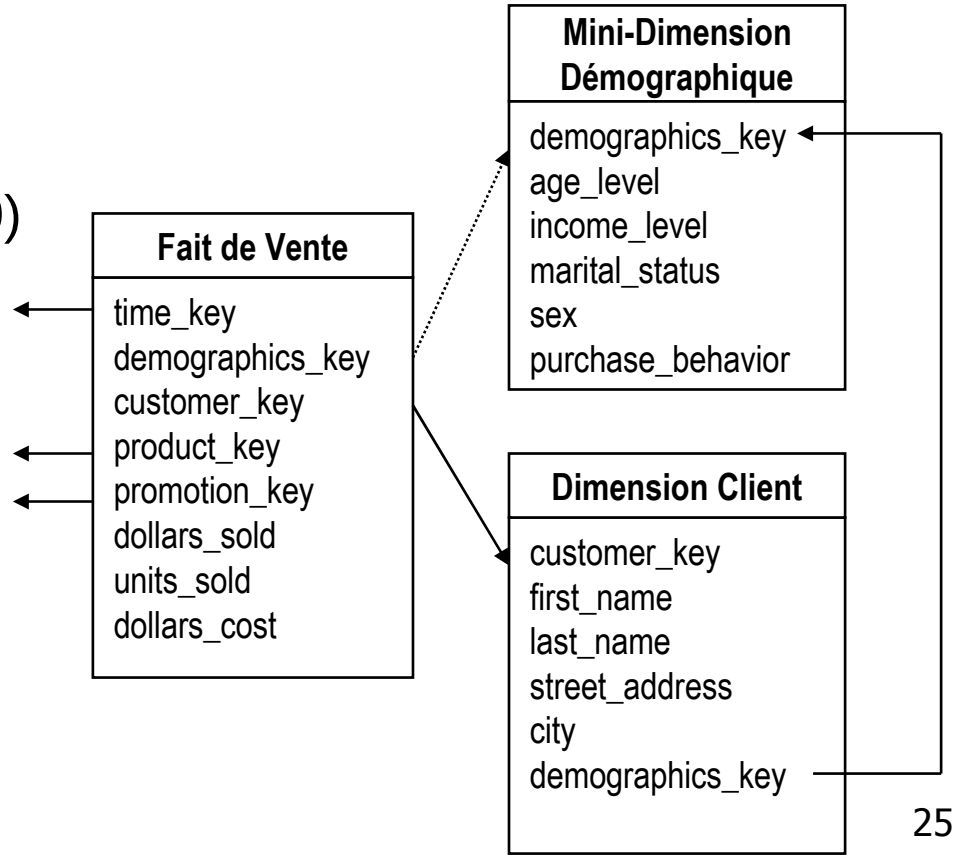


Grandes Dimensions

- Nombreux membres
  réduire la taille des tables
  - dimension Produits (300.000)
  - dimension Clients (10.000.000)
- Solutions
  - ☹ L ’appel du Flocon de Neige
    - tables de dimension secondaires (déportées) associée à une table de dimension
    - Faible gain de place et Navigation compromise
  - ☺ Mini Dimensions
    - Mini dimensions démographiques pour les clients

Mini Dimensions Démographiques

- Dimension client
  - nombreux enregistrements, nombreux attributs
- Solutions
  - ☹Flocons
  - ☺Mini-Dimension
    - Combinaisons (<100000) d 'intervalles de valeurs démographiques



Dimensions à évolution lente (i)

- Changement de description des membres dans les dimensions
  - un client peut changer d ’adresse, se marier, ...
  - un produit peut changer de noms, de formulations
    - « Tree ’s » en « M&M », « Raider » en « Twix », « Yaourt à la vanille en Yaourt » en « saveur Vanille », « bio » en « Activa »
- Choix entre 3 solutions
  - écrasement de l ’ancienne valeur
  - versionnement
  - valeur d ’origine / valeur courante
- Remarque
  - quand la transition n ’est pas immédiate : il reste pendant un certain temps des anciens produits en rayon
  - Solution : 2 membres différents

Dimensions à évolution lente (ii)

- 3 solutions
  - Ecrasement de l 'ancienne valeur
    - renoncer à suivre les situations passées
    - mais correction d 'informations erronées
  - Versionnement
    - clé étendue d 'un numéro de version
    - partitionnement automatique de l 'historique
  - Valeur courante / Valeur d 'origine
  - et Valeur courante / Valeur antérieure
    - l 'ancienne valeur n 'est utile que pendant un certain temps pour étudier les effets d 'une transition
    - exemple: renouvellement d 'une force de vente
- Mini dimension à évolution lente

Dimensions à évolution lente
Ecrasement de l ’ancienne valeur

- renoncer à suivre les situations passées
- mais correction d ’informations erronées

Temps

#T:JJ:MM:AA:Event
1201:14:02:99:St Valentin

Produit

#P:Descr
66:Bague
77:Fleur

Fait de Vente

#T:#C:#P:Prix
200:100:77:100
201:100:77:100
202:100:77:100
202:100:66:10000
568:100:77:20
1115:100:77:100
1116:100:77:100
1117:100:66:50000
1200:100:77:100

Client

#C: #V:Nom:SitMarital
100:Didier:~~Divorcé~~Marié

Dimensions à évolution lente
Versionnement

- clé étendue d ’un numéro de version
- partitionnement automatique de l ’historique

Temps
#T:JJ:MM:AA:Event
1201:14:02:99:St Valentin

Produit
#P:Descr
66:Bague
77:Fleur

Fait de Vente
#T:#C:#P:Prix
200:100:1:77:100
201:100:1:77:100
202:100:1:77:100
202:100:1:66:10000
568:100:2:77:20
1115:100:3:77:100
1116:100:3:77:100
1117:100:3:66:50000
1200:100:4:77:100

Client
#C: #V:Nom:SitMarital:DateEffet
100:1:Didier:Célibataire:10
100:2:Didier:Marié:203
100:3:Didier:Divorcé:567
100:4:Didier:Marié:1118

Dimensions à évolution lente
Valeur d 'origine / valeur courante

- l'ancienne valeur n 'est utile que pendant un certain temps pour étudier les effets d 'une transition

Temps
#T:JJ:MM:AA:Event
1201:14:02:99:St Valentin

Produit
#P:Descr
66:Bague
77:Fleur

Fait de Vente
#T:#C:#P:Prix
200:100:77:100
201:100:77:100
202:100:77:100
202:100:66:10000
568:100:77:20
1115:100:77:100
1116:100:77:100
1117:100:66:50000
1200:100:77:100

Client
#C: #V:Nom:SMcour:SMorig:DateEffet
100:Didier:Marié:Célibataire:1118
102:Paul:Célibataire:NULL:NULL

Dimensions à évolution lente
Valeur antérieure / valeur courante

- La valeur antérieure n ’est utile que pendant un certain temps pour étudier les effets d ’une transition

Temps
#T:JJ:MM:AA:Event
1201:14:02:99:St Valentin

Produit
#P:Descr
66:Bague
77:Fleur

Fait de Vente
#T:#C:#P:Prix
200:100:77:100
201:100:77:100
202:100:77:100
202:100:66:10000
568:100:77:20
1115:100:77:100
1116:100:77:100
1117:100:66:50000
1200:100:77:100

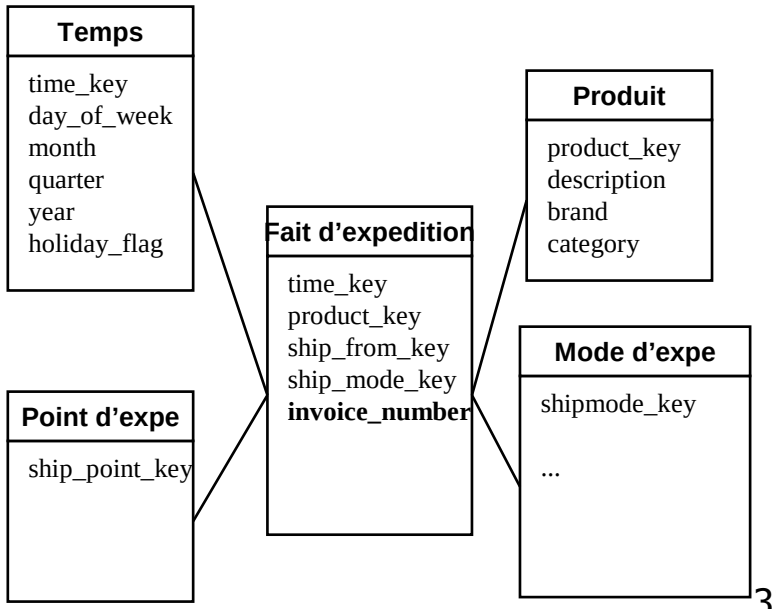
Client
#C: #V:Nom:SMcour:SMant:DateEffet
100:Didier:Marié:Divorcé:1118
102:Paul:Célibataire:NULL:NULL

Dimension Client Douteuse

- Dimension Client dans laquelle la même personne peut apparaître de nombreuses fois
  - orthographes légèrement différentes
  - attributs différents

Dimension Dégénérée

- Dimension sans attribut
  - Pas de table
  - Mais la clé de dimension est dans la table de fait
- Exemple
  - numéro de facture (invoice number),
  - numéro de ticket
  - …



Bases hippocratiques

- garantir la sécurité des données personnelles
  - Cf serment d'Hippocrate des médecins
- Règles à respecter
  - Spécification des objectifs
  - Consentement
  - Collection limitée
  - Limitation d'usage
  - Limitation de divulgation
  - Limitation de conservation
  - Exactitude
  - Sûreté
  - Ouverture
  - Conformité

- Agrawal, R., Kiernan, J., Srikant, R., Xu, Y., Hippocratic Databases, International Conference on Very Large Data Bases (VLDB), Hong Kong, China, 2002.

Remerciement à Medhi Benzine

Anonymisation et dégradation des données

- Contexte
  - Données sensibles (dossier patient, …)
  - Mais Traitement statistique (épidémiologie, …)
- Problème des « quasi-identifiants »
  - Aux Etats-Unis, 87 % des individus sont identifiés par le groupe d'attributs: <date de naissance, sexe, code postal>
- k-anonymat
  - dégrader les attributs constituant le *quasi-identifiant* de manière à rendre ces attributs pour un individu identiques à ceux de k-1 autres individus
    - Sweeney, L., k-anonymity : a model for protecting privacy. International Journal on Uncertainty, Fuzziness and Knowledge-based Systems, 10(5) : Pages 557-570, 2002

Remerciement à Medhi Benzine

Exemple de k-Anonymat

	Non Sensitive			Sensitive
	Zip code	Age	Nationality	Condition
1	13053	28	Russian	Heart Disease
2	13068	29	American	HeartDisease
3	13068	21	Japanese	Viral Infection
4	13053	23	American	Viral Infection
5	14853	50	Indian	Cancer
6	14853	55	Russian	Heart Disease
7	14850	47	American	Viral Infection
8	14850	49	American	Cancer
9	13053	31	American	Cancer
10	13053	37	Indian	Cancer
11	13068	36	Japanese	Cancer
12	13068	35	American	Cancer

	Non Sensitive			Sensitive
	Zip code	Age	Nationality	Condition
1	130**	< 30	*	Heart Disease
2	130**	< 30	*	HeartDisease
3	130**	< 30	*	Viral Infection
4	130**	< 30	*	Viral Infection
5	1485*	≥ 40	*	Cancer
6	1485*	≥ 40	*	Heart Disease
7	1485*	≥ 40	*	Viral Infection
8	1485*	≥ 40	*	Cancer
9	130**	3*	*	Cancer
10	130**	3*	*	Cancer
11	130**	3*	*	Cancer
12	130**	3*	*	Cancer

	Non Sensitive			Sensitive
	Zip code	Age	Nationality	Condition
1	1305*	≤ 40	*	Heart Disease
4	1305*	≤ 40	*	Viral Infection
9	1305*	≤ 40	*	Cancer
10	1305*	≤ 40	*	Cancer
5	1485*	> 40	*	Cancer
6	1485*	> 40	*	Heart Disease
7	1485*	> 40	*	Viral Infection
8	1485*	> 40	*	Cancer
2	1306*	≤ 40	*	HeartDisease
3	1306*	≤ 40	*	Viral Infection
11	1306*	≤ 40	*	Cancer
12	1306*	≤ 40	*	Cancer

Remerciement à Medhi Benzine

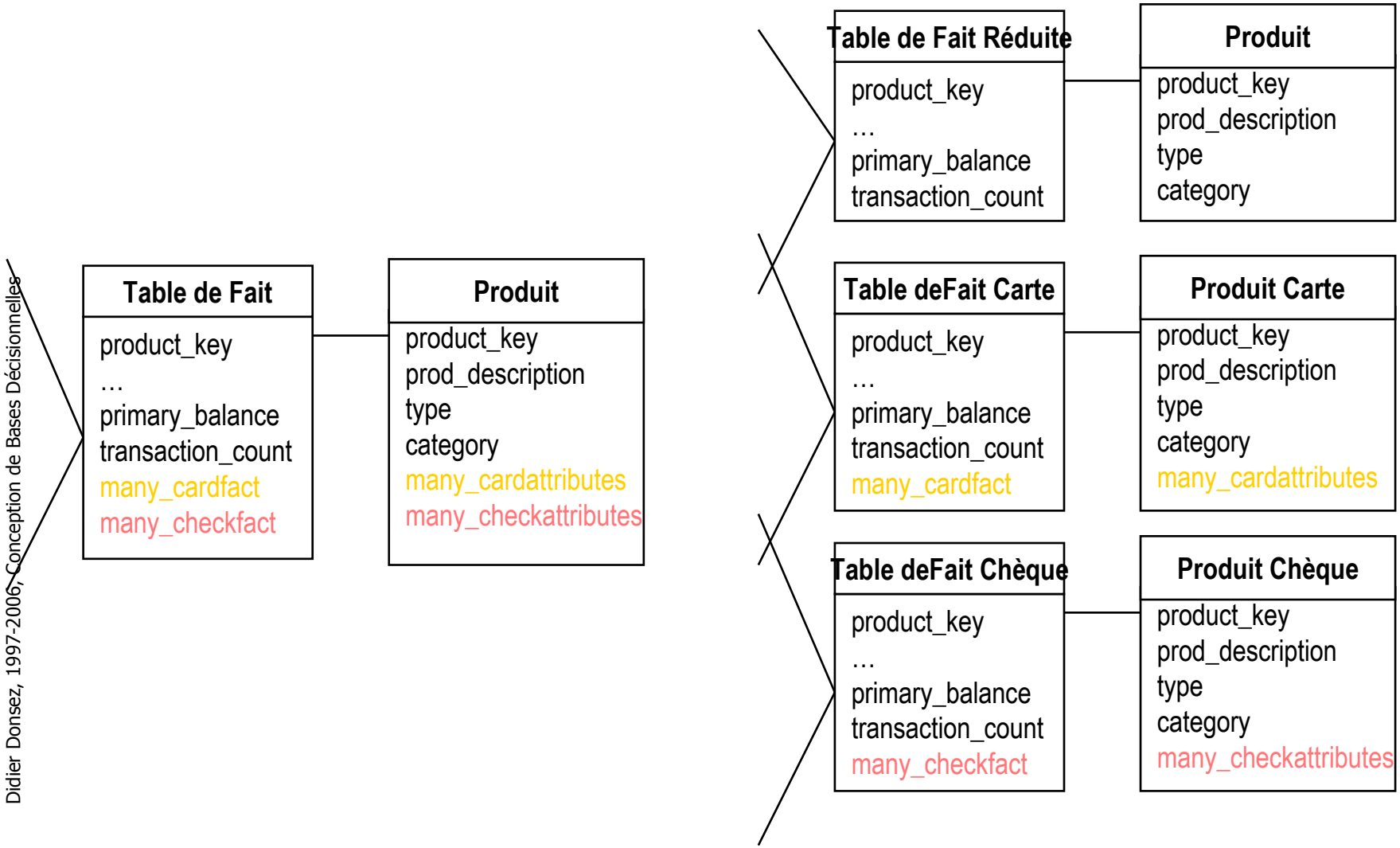
Tables de Suivi d ’Evénements

- souvent sans mesure
  - cours, enseignant, étudiant
  - hôpital, médecin, patient, diagnostic
  - parties d ’un accident
- Comptage des faits

Tables de Faits Réduites et Particularisées

- Application avec des produits hétérogènes
  - compte chèque, compte d 'épargne, ...
  - police et sinistre automobile, habitation, ...
- Conception (économie de place)
  - tables de fait réduite
    - tous les enregistrements de fait reduit aux mesures communes
    - table dimension réduite aux attributs communs
  - 1 table particularisé de fait par produit hétérogène
    - seuls les enregistrements concernant le produit
    - 1 table de dimension par produit
      - attributs particuliers
- Remarque : pas de OODW pour l'instant !

Tables de Faits Réduites et Particularisées



Codage des Clés et des Mesures

- Mesure de fait
  - valeurs entières (4 octets)
  - parfois plus
    - ex: PNB des USA au cent près
- Clés
  - valeurs entières anonymes (4 octets)
    - réduit la taille de l ’enregistrement de fait
    - réduit le coût CPU des comparaison de jointure
  - la correspondance clé opérationnelle et clé entrepôt est faite à l ’extraction

Estimation de la taille de l 'entrepôt

- Dimensionner l 'entrepôt
  - Choix des granularités
  - Choix d 'une machine/SGBD cible (benchmark)
- Exemple : Supermarché
  - Dimensions
    - Temps : 4 ans * 365 jours = 1460 jours
    - Magasin : 300
    - Produit : 200000 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 de champs de clé = 4
    - Nb de champs de fait = 4
    - Table des Faits = 8,76.10^9 * 8 champs * 4 octets = **280 Go**

Estimation de la taille de l 'entrepôt

- Exemple : Ligne d 'article en Grande Distribution
  - Temps : 3 ans * 365 jours = 1095 jours
  - CA annuel = 80 000 000 000 $
  - Montant moyen d 'un article = 5 $
  - Nb de champs de clé = 4
  - Nb de champs de fait = 4
  - Nombre de Faits = 3*(80.10^9/5) = 48.10^9
  - Table de Faits= 48.10^9 * 8 champs * 4 octets = **1,59 To**
- Exemple : Suivi d 'appels téléphoniques
  - Temps : 3 ans * 365 jours = 1095 jours
  - Nombre d 'appel par jour = 100 000 000
  - Nb de champs de clé = 5
  - Nb de champs de fait = 3
  - Table des Faits = 1095.10^8 * 8 champs * 4 octets = **3,49 To**
- Exemple : Suivi d 'achats par carte de crédit
  - Temps : 3 ans * 12 mois = 36 mois
  - Nombre de compte carte = 50 000 000
  - Nombre moyen d 'achat par mois par carte = 50
  - Nb de champs de clé = 5
  - Nb de champs de fait = 3
  - Table des Faits = 54.10^9 * 8 champs * 4 octets = **1,73 To**

Conclusion

- Résister à la normalisation
- ...

Bibliographie - Livre

- Ralph Kimball, Entrepôts de Données, Ed. Intl Thomson Pub., 1997 et 2000, ISBN 2-84180-021-0
  - la bible du concepteur ☺☺☺
  - contient un outil (StarTracker) et des bases d ’exemples.
  - Son nouvel ouvrage sort en 04/2002 (www.rkimball.com)
- Rob Mattison, Data Warehousing -Strategies, Technologies and Technics, IEEE Computer Society 1996, ISBN 0-07-041034-8, 55$
  - la méthodologie d ’organisation ☺
- Jean Michel Franco, Le Data Warehouse / Le Data Mining, Eyrolles, 1997
  - un survol en français 😐

Bibliographie - Livre

- Ralph Kimball, Laura Reeves , "Concevoir et déployer un data warehouse Guide de conduite de projet ", Ed Eyrolles

- Ralph Kimball, "The Data Warehouse Toolkit: The Complete Guide to Dimensional Modeling, 2nd Edition"

- Ralph Kimball, "The Data Webhouse Toolkit: Building the Web-Enabled Data Warehouse"