Université de Sherbrooke

Bases de données dimensionnelles

Interrogation

UdeS:BDD_20

Christina KHNAISSER *(christina.khnaisser@usherbrooke.ca)*

CoFELI/Scriptorum/BDD_30-Interrogation, version 1.0.1.b, en date du 2025-11-20

— en vigueur —

Plan

Introduction

Le présent document a pour but de présenter les différentes catégories de requêtes pour les bases de données dimensionnelles.

Évolution du document

La première version du document a été établie sur la base des travaux publiés par Adamson et plus particulièrement [Adamson2010a], chapitres 4, 9, 14 et 15.

1. Mise en contexte

Une base de données dimensionnelles s'intéresse à l'évaluation des processus d'un domaine.

Les besoins analytiques permettent d'identifier des requêtes pour extraire les données nécessaires pour l'évaluation.

2. Catégorie de requêtes

- *Browser*
- *Drill down*
- *Drill up*
- *Drill across*
- *Pivot*

2.1. Browser

Browser est une catégorie de requête qui consiste à explorer les données d'une dimension.

SQL Query

```
SELECT DISTINCT
   product.category
FROM
   product
ORDER By
   product.category
```

Query Results

```
CATEGORY
========
.
.
.
Fasteners
Folders
Packaging
Pens
Measurement
Notebooks
Storage
.
.
.
```

SQL Query

```
SELECT DISTINCT
   product.category,
   product.product
FROM
   product
WHERE
   product.category = "Packaging"
ORDER BY
   product.product
```

Query Results

```
CATEGORY     PRODUCT
========     =======
Packaging    Box - Large
Packaging    Box - Medium
Packaging    Box - Small
Packaging    Clasp Letter
Packaging    Envelope #10
Packaging    Envelope Bubble
```

Figure 1-7 Browse queries and their results

Figure 1. Exploration de la relation Produit [Adamson2010a]

2.2. Drill up and down

- *Drill down* est une catégorie de requête qui consiste à **ajouter** une ou plusieurs dimensions de la même hiérarchie à une requête pour raffiner le résultat.
- *Drill up* est une catégorie de requête qui consiste à **retirer** une ou plusieurs dimensions de la même hiérarchie à une requête pour synthétiser le résultat.

Exercices

- Calculer le taux de réussite des personnes étudiantes (drill-up)
  - par programme d'études
  - par département
  - par facultés
- Calculer le taux d'occupation des lits d'hôpital (drill-down)
  - par hôpital
  - par service
  - par département

Exemple 1. Calcul de la vente totale [Adamson2010a]

Hiérarchie de Produit

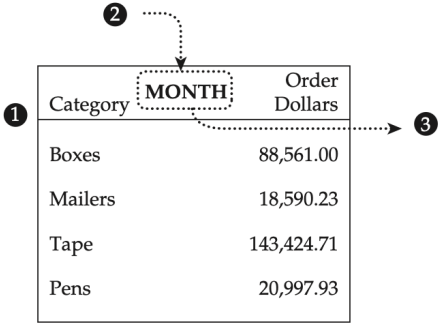
- Calculer la vente totale par produit
- Calculer la vente totale par catégorie

Hiérarchie de Moment

- Calculer la vente totale par mois
- Calculer la vente totale par année

Plusieurs dimensions

- par catégorie de produit
- par mois par catégorie de produit



Category	Month	Order Dollars
Boxes	January	23,520.20
	February	33,120.03
	March	31,920.77
Mailers	January	8,544.00
	February	7,763.01
	March	2,283.22
Tape	January	49,920.01
	February	29,088.37
	March	64,416.33
Pens	January	14,997.84
	February	2,880.01
	March	3,120.08

Figure 7-1 Adding dimensional detail

2.3. Drill across

Drill across est une catégorie de requête qui consiste à fusionner de requêtes provenant de plusieurs étoiles (plusieurs relations factuelles).

L'objectif de ces requêtes est de comparer deux ou plusieurs processus.

Pour pouvoir fusionner des faits, les attributs des dimensions communes les mêmes.

Exemple

- Évaluer la qualité des produits par produit et par mois (ventes et retours)
  - nombre de ventes
  - nombre de retours
  - montant total des ventes
  - montant total des retours

2.3.1. Multiplicité des tables de faits

ORDER_FACTS

day_key	customer_key	product_key	quantity_key
123	777	111	100
123	777	222	200
123	777	333	50

SHIPMENT_FACTS

day_key	customer_key	product_key	quantity_shipped
456	777	111	100
456	777	222	75
789	777	222	125

```
SELECT
   product.product,
   SUM( order_facts.quantity ordered ),
   SUM( shipment_facts.quantity_shipped )
FROM
   product,
   day,
   order_facts,
   shipment_facts
WHERE
   order_facts.product_key = product.product_key AND
   order_facts.day_key = day.day_key AND
   shipment_facts.product_key = product.product_key AND
   shipment_facts.day_key = day.day_key AND
   ...additional qualifications on date...
GROUP BY
   product.product
```

The order for product 222 is double counted

product	sum(quantity_ordered)	sum(quantity_shipped)
Product 111	100	100
Product 222	400	200

The order for product 333 does not appear

Figure 4-10 Joining two fact tables leads to trouble

Figure 2. Problèmes de jointure de deux relations factuelles [Adamson2010a]

Les données doivent être recueillies en trois étapes qui consistent à passer d'une étoile à une autre :

1. calculer le résultat d'un processus à la fois selon la granularité choisie
2. joindre les résultats par pairs en utilisant une jointure complète (*full outer join*),
3. ajouter des attributs calculés aux besoins.

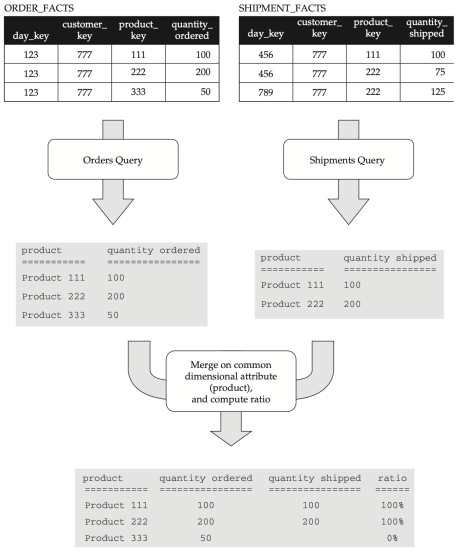


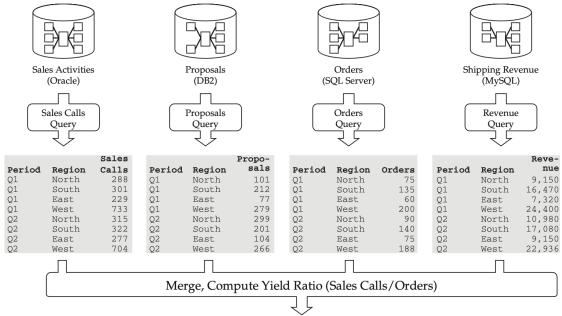
Figure 4-11 Drilling across orders and shipments

Figure 3. Requête Drill across avec 2 relations factuelles [Adamson2010a]

Notez les jointures décomposées en produits cartésiens (FROM) et restrictions (WHERE). Plusieurs membres de la communauté de pratique ont une telle phobie (injustifiée) des jointures) qu'ils les simulent ainsi (pensant sans doute les éviter), complexifiant en fait le travail du SGBD (qui doit recomposer les jointures) afin de pouvoir les optimiser adéquatement. De même pour le réviseur humain qui doit les vérifier. On remarque par ailleurs que c'est une source fréquente d'erreur (restriction erronée ou absente).

En conclusion, l'écriture explicite des jointures est adéquate, en particulier du point de vue de la validité, de l'efficacité et de l'efficience.

2.3.2. Multiplicité des sources



Sales Report
2008

Period	*Region*	*Sales Calls*	*Proposals*	*Orders*	*Revenue*	*Yield*
Q1	North	288	101	75	9,150	26%
	South	301	212	135	16,470	45%
	East	229	77	60	7,320	26%
	West	733	279	200	24,400	27%
Q2	North	315	299	90	10,980	29%
	South	322	201	140	17,080	43%
	East	277	104	75	9,150	27%
	West	704	266	188	22,936	27%

Figure 4-13 Drilling across four fact tables

Figure 4. Requête Drill across avec 4 relations factuelles [Adamson2010a]

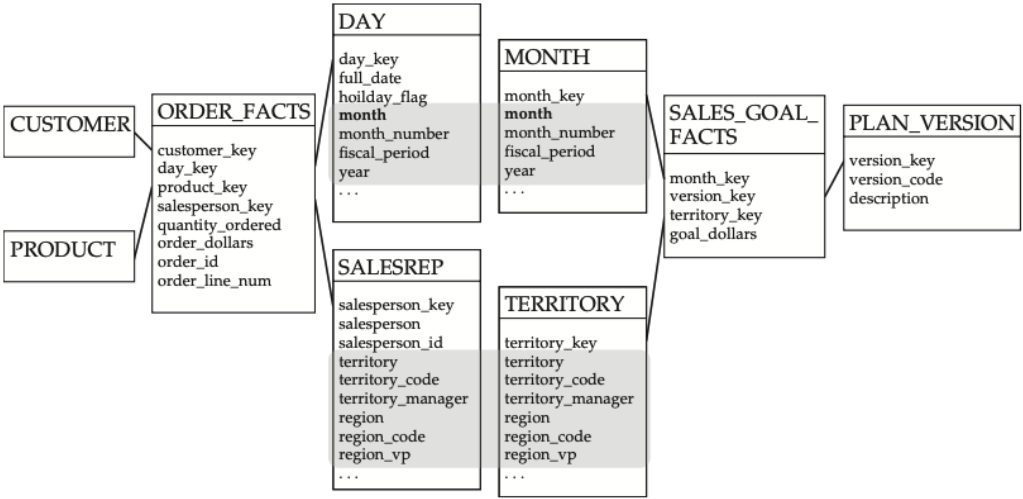


Figure 5-3 These stars do not share common dimension tables but do share common dimension attributes

Figure 5. Dimensions conformes [Adamson2010a]

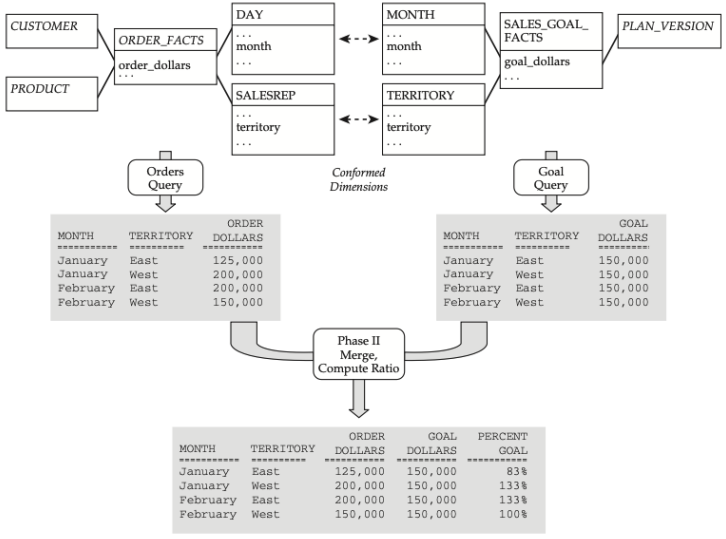


Figure 5-4 Drilling across order_facts and sales_goal_facts

Figure 6. Fusion des dimensions conformes [Adamson2010a]

On constate aisément qu'il est préférable de traiter la multiplicité des sources au moment de l'alimentation.

On s'assure ainsi que le travail est fait, bien fait et fait une seule fois, simplifiant ainsi l'expression de toutes les requêtes nécessitant lesdites sources.

Voici quelques (mauvaises) raisons pour lesquelles ce traitement n'est pas réalisé au bon moment (alimentation) ou au bon endroit (*staging region*) :

- Qui va payer pour ce travail préparatoire ? Que ceux qui en ont besoin le fassent !
  - Pour quelle raison fait-on un entrepôt de données déjà ?
- La conciliation des sources dont j'ai besoin est différente de celle réalisée lors de l'alimentation

- Ce ne peut être si on respecte la sémantique et les prédicats des attributs.

2.4. Pivot

- *Pivot* est une catégorie de requête qui consiste *reformater* le résultat d'une requête en transformant les lignes en colonnes ou vice-versa. Le résultat de ces requêtes permet de faciliter la présentation des données selon un format de rapport ou pour les algorithmes d'intelligence artificielle.

```
ACCOUNT   TRANSACTION_TYPE  SUM(AMOUNT)
========  ================  ===========
01-2211   Credits             20,301.00
01-2211   Debits            - 17,691.30
07-4499   Credits              1,221.23
07-4499   Debits           -   2,220.01
  .
  .
  .
```

```
ACCOUNT       DEBITS     CREDITS
========  ==========  ==========
01-2211    17,691.30   20,301.00
07-4499     2,220.01    1,221.23
  .
  .
  .
```

Figure 7. Requête pivoting [Adamson2010a]

3. Schémas dérivés

La base de données analytique sert à répondre à diverses questions pour évaluer l'état d'une organisation afin de permettre aux gestionnaires de prendre des décisions éclairées.

L'objectif principal qui a amené à la création du modèle dimensionnel est la facilité d'écrire des requêtes et l'optimisation des performances.

- Restructurer les données
- Synthétiser les données

Les schémas dérivés offrent une structure adaptée à un besoin d'analyse spécifique.

3.1. Fusion des faits (*merged fact*)

La comparaison d'évènements provenant de plusieurs relations factuelles. (élimine plusieurs étapes d'une requête drill-across).

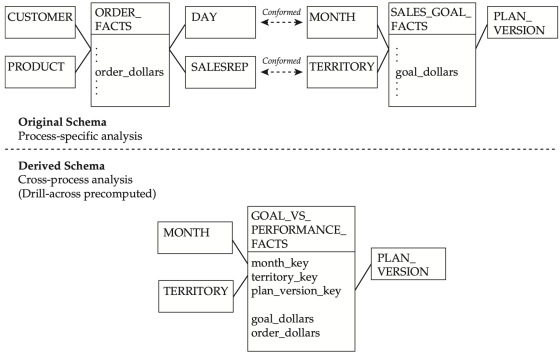


Figure 14-1 A merged fact table compares plans to actuals

Figure 8. Fusion des faits [Adamson2010a]

3.2. Pivotement des faits (*Pivoted fact*)

La transposition des lignes en colonnes pour certains types de rapports.

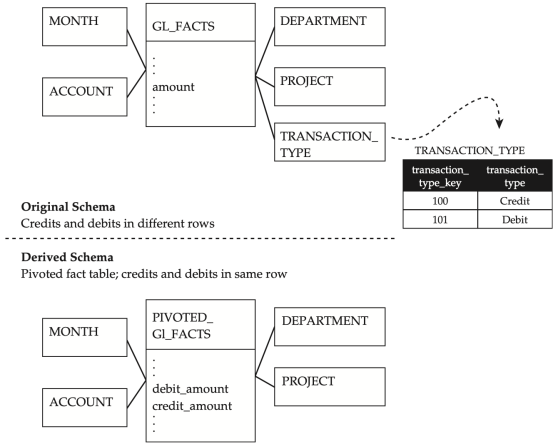


Figure 14-2 A pivoted fact table

Figure 9. Pivoter des faits [Adamson2010a]

3.3. Partitionnement des faits (*Sliced fact*) :

La sélection d'un sous-ensemble des faits selon une ou plusieurs dimensions.

Exemple, partitionné par magasin, région, par département.

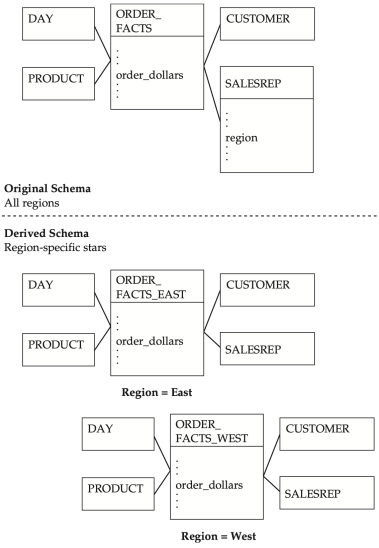


Figure 14-3 Sliced fact tables

Figure 10. Partitionner les faits [Adamson2010a]

3.4. Aggrégation des faits

Synthétiser des mesures selon une ou plusieurs dimensions.

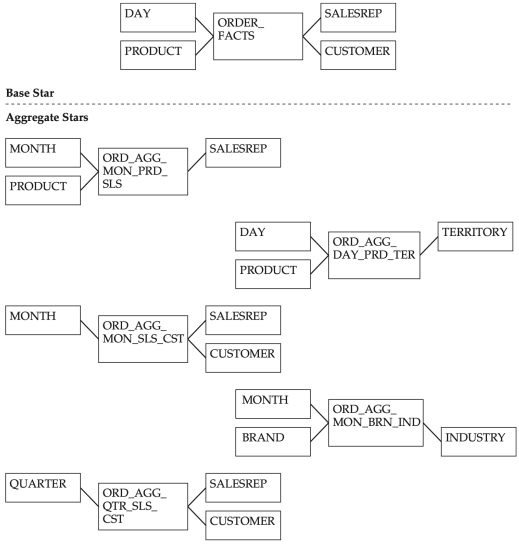


Figure 15-2 A base star and several aggregates

Figure 11. Aggréger les faits [Adamson2010a]

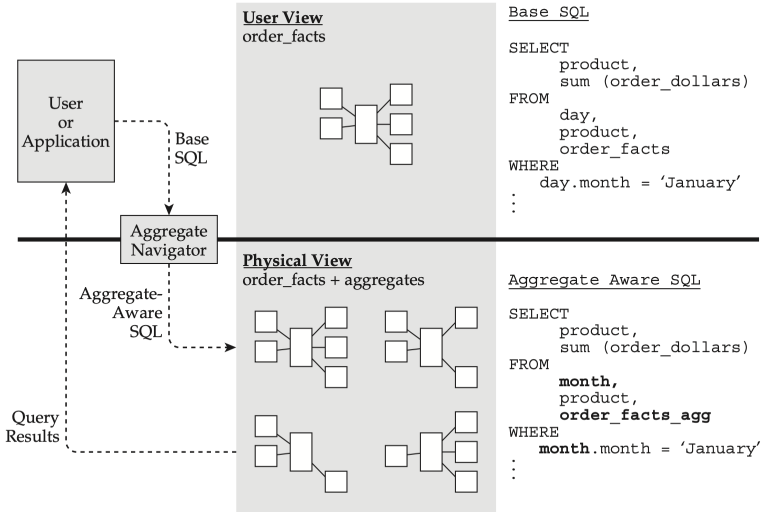


Figure 15-4 Aggregate navigator

Figure 12. SGBD - Nagivation des étoiles aggrégées [Adamson2010a]

3.5. Et les vues ?

Pourquoi n'utilise-t-on pas des vues pour représenter les tables des schémas dérivés ?

La raison invoquée est l'inefficience (supposée) des vues lorsqu'elles sont utilisées à plusieurs reprises (sans changements aux données des tables sur lesquelles elles sont construites).

Cet argument ne prend pas en compte la possibilité de « matérialiser » les vues, de les « rematérialiser » au besoin et même d'automatiser la « rematérialisation » au besoin.

Il est vrai que ces possibilités

- sont apparues tardivement (au cours des années 1990) et qu'elles ne sont pas connues de tous ;

- ne sont pas disponibles dans certains jouets utilisés en lieu et place de SGBD.

:-)

4. Règles de pratique

- Décrire rigoureusement le besoin d'analyse pour pouvoir choisir la catégorie de requêtes qui convient le mieux.
- N'essayez jamais de joindre deux relations factuelles directement ou par l'intermédiaire d'une dimension commune. Cela peut produire des résultats inexacts.
- Si vous utilisez des logiciels de construction de requêtes, la définition des catégories de requêtes peu variée. Lisez attentivement la documentation avant de construire votre requête.

Références

[Adamson2010a]
Christopher ADAMSON ;
The complete reference star schema ;
McGraw-Hill, New York (NY, US), 2010 ;
ISBN 978-0-07-174432-4.

Produit le 2025-11-21 12:12:15 UTC

Université de Sherbrooke