

L'approche multidimensionnelle dans les bases de données

Les bases de données relationnelles

L'approche multidimensionnelle permet de traiter simplement des questions qui mettent en jeu de grandes masses de données. Nous considérerons ici qu'une base de données relationnelles est constituée d'un ensemble de tables, chaque table pouvant être décrite sous la forme d'un tableau où les colonnes correspondent à des des variables et les lignes à des enregistrements.

Par exemple, le tableau ci-dessous donne la valeur des variables P1, P2 et B1 de plusieurs secteurs d'activité AE.

AE	P1	P2	B1
B01	800	300	500
B02	500	400	100
B02	600	450	150
B03	200	50	150
B04	700	400	300
B04	900	500	400

La première ligne du tableau donne le nom des différentes variables, les autres lignes correspondent aux enregistrements. Dans une même table, le nombre d'enregistrements est indifférent mais chaque enregistrement doit avoir la même structure, c'est-à-dire comporter les mêmes variables et dans le même ordre. Ainsi, si l'on ajoutait un nouvel enregistrement à cette table, il devrait impérativement comprendre les variables AE, P1, P2 et B1 dans cet ordre.

Nous pouvons considérer que le tableau décrit une économie en l'étudiant sous deux critères différents : les opérations de la comptabilité nationale et les activités économiques. Nous remarquons alors que les deux critères ne sont pas traités de manière symétrique dans la table de notre base de données relationnelle. En effet, on ne peut pas ajouter de nouvelles opérations à la table sans porter atteinte à sa structure alors qu'on peut ajouter autant de nouvelles activités que l'on désire.

Or, il est souvent intéressant de traiter les critères de manière symétrique, surtout si l'on souhaite travailler avec plus de deux critères. Nous allons donc présenter la table de manière différente mais en conservant les mêmes informations et en respectant la symétrie de traitement des critères. Pour cela, nous allons créer une table où chaque valeur du tableau apparaîtra avec les modalités des deux critères dans un même enregistrement, le nombre d'enregistrements étant alors égal au nombre de valeurs du tableau initial. La nouvelle table sera constituée de trois variables, une variable pour chaque critère et une variable pour la valeur. Elle pourra se présenter de la manière suivante :

AE	OP	VAL
B01	P1	800
B02	P1	500
B02	P1	600
B03	P1	200
B04	P1	700
B04	P1	900
B01	P2	300
B02	P2	400
B02	P2	450
B03	P2	50
B04	P2	400
B04	P2	500
B01	B1	500
B02	B1	100
B02	B1	150
B03	B1	150
B04	B1	300
B04	B1	400

On constate que, par rapport à l'ancienne, la nouvelle table est plus volumineuse, mais, en contrepartie, elle présente l'avantage certain d'être plus facilement modifiable puisque l'on peut, sans porter atteinte à sa structure, ajouter autant de nouvelles opérations que l'on désire.

L'autre grand avantage de ce type de tables est qu'il est très facile de travailler avec plus de deux critères. Par exemple, si nous souhaitons introduire également l'année, il est possible de décrire simplement l'ensemble des données avec une table à 4 colonnes comme celle qui est présentée ci-dessous :

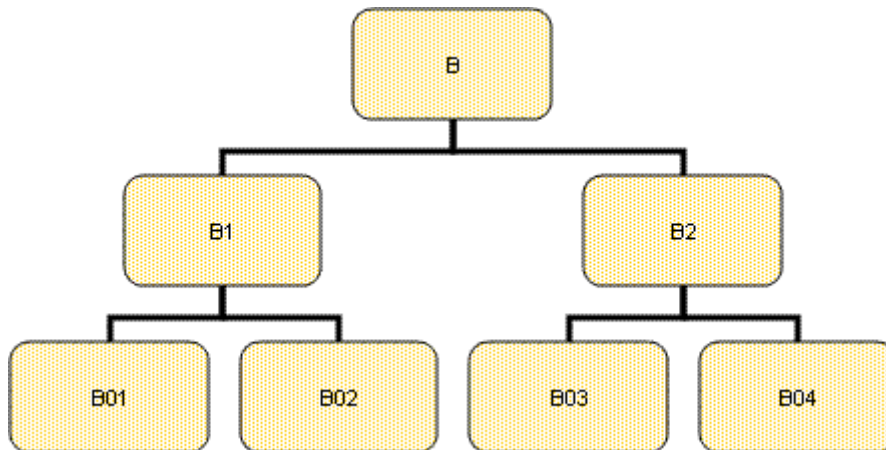
AE	DA	OP	VAL
B01	2004	P1	800
B02	2004	P1	500
B02	2005	P1	600
B03	2004	P1	200
B04	2004	P1	700
B04	2005	P1	900
B01	2004	P2	300
B02	2004	P2	400
B02	2005	P2	450
B03	2004	P2	50
B04	2004	P2	400
B04	2005	P2	500
B01	2004	B1	500
B02	2004	B1	100
B02	2005	B1	150
B03	2004	B1	150
B04	2004	B1	300
B04	2005	B1	400

Cette simplicité de travail avec plusieurs critères est fondamentale pour le comptable national. Déjà, il est facile de voir qu'avec trois critères il aurait été difficile de travailler avec une table du type de la première table présentée car il aurait été nécessaire de multiplier le nombre de colonnes par le nombre d'années. En pratique cela soulève deux problèmes, le premier est qu'il est difficile de travailler avec de nombreuses variables, le second est qu'il faut modifier la structure du tableau chaque fois que l'on veut ajouter une année. Or, le comptable national travaille habituellement avec plus de trois critères. Il utilise, par exemple, l'année, le secteur institutionnel, le secteur d'activité, la branche, le produit, l'opération, le critère ressources/emplois et la source des données. Il est donc pratiquement impossible de travailler avec des tableaux analogues au premier tableau présenté.

Les critères prennent le nom de dimensions et les bases de données où les tables sont structurées de cette manière peuvent être qualifiées de bases de données multidimensionnelles et les tables d'hypercubes.

Les hiérarchies

Les bases de données organisées selon une approche multidimensionnelle sont extrêmement performantes. Elles permettent, en effet, de réaliser très simplement un certain nombre de traitements. Parmi eux, l'un des plus intéressants est l'introduction de hiérarchies. Ainsi, dans notre exemple, nous avons introduit 4 activités différentes correspondant à un niveau de la nomenclature d'activités que nous appellerons, comme les comptes nationaux français le niveau G. Ces secteurs B01, B02, B03, B04 sont regroupés au niveau supérieur, appelé le niveau F, en deux secteurs et B1 (qui regroupe B01 et B02) et B2 (qui regroupe B03 et B04). Au niveau supérieur, appelé niveau E, les deux secteurs B1 et B2 sont regroupés en un seul, le secteur B. Nous avons ici l'exemple d'une hiérarchie sur les secteurs d'activité, l'activité B a deux fils B1 et B2, le secteur B1 a lui-même deux fils B01 et B02, le secteur B2 ayant B03 et B04 pour fils.



On peut souhaiter obtenir les résultats non seulement au niveau G mais également aux niveaux supérieurs. Nous pourrions y parvenir et passer d'un niveau de nomenclature à l'autre en introduisant une table de passage. Il y a deux manières possibles de procéder pour cela, on peut souhaiter disposer de chaque niveau de nomenclature dans deux dimensions séparées ou l'on peut préférer disposer des deux niveaux dans la même dimension.

Dans le premier cas, on fera appel à une table passage du type suivant :

AEG	AEF	AEE
B01	B1	B
B02	B1	B
B03	B2	B
B04	B2	B

Si nous appelons PROD la table de données et HAE la table de passage, la requête suivante nous permet de calculer les données, par exemple, au niveau F :

```

SELECT H.AEF, P.OP, sum(P.VAL) as VAL
FROM PROD P, HAE H
WHERE P.AE=H.AEG
GROUP BY H.AEF, P.OP
;
  
```

Nous obtenons :

AEF	OP	VAL
B1	P1	1900
B2	P1	1800
B1	P2	1150
B2	P2	950
B1	B1	750
B2	B1	850

Nous aurions également pu calculer le niveau E de la même manière en remplaçant dans la requête précédente GROUP BY H.AEF par GROUP BY H.AEE.

Si nous souhaitons faire apparaître dans une même dimension tous les niveaux de la hiérarchie

nous devons utiliser la table suivante :

AEG	AE
B01	B01
B02	B02
B03	B03
B04	B04
B01	B1
B02	B1
B03	B2
B04	B2
B01	B
B02	B
B03	B
B04	B

En utilisant la requête suivante :

```
SELECT H.AE, P.OP, sum(P.VAL) as VAL
FROM PROD P, HAE H
WHERE P.AE=H.AEG
GROUP BY H.AE, P.OP
;
```

Nous obtenons :

AE	OP	VAL
B01	P1	800
B02	P1	1100
B03	P1	200
B04	P1	1600
B01	P2	300
B02	P2	850
B03	P2	50
B04	P2	900
B01	B1	500
B02	B1	250
B03	B1	150
B04	B1	700
B1	P1	1900
B2	P1	1800
B1	P2	1150
B2	P2	950
B1	B1	750
B2	B1	850
B	P1	3700
B	P2	2100
B	B1	1600

Les combinaisons de modalités

Il est souvent utile pour le comptable national de calculer des combinaisons d'un certain nombre de modalités d'une ou de plusieurs dimensions.

Supposons, par exemple, que nous cherchions à vérifier le calcul de B1 (la valeur ajoutée) en faisant la différence entre P1 (la production) et P2 (la consommation intermédiaire). Il suffit pour cela d'introduire une table de passage qui montre comment passer des opérations P1 et P2 à une nouvelle opération B1C. Cette table de passage se présentera alors ainsi :

OPC	OP	VAL
B1C	P1	1
B1C	P2	-1

Si nous appelons PROD la table de données et CALC la table de passage, la requête suivante nous permet de calculer les données :

```
SELECT P.AE, C.OPC, sum(P.VAL*C.VAL) as VAL
FROM PROD P, CALC C
WHERE P.OP=C.OP
GROUP BY P.AE, C.OPC
;
```

Nous obtenons :

AE	OP	VAL
B01	B1C	500
B02	B1C	250
B03	B1C	150
B04	B1C	700

Cette manière de travailler est extrêmement utile pour le contrôle et la maintenance du processus d'élaboration des comptes nationaux. En effet, avec cette approche il est possible de séparer les trois éléments fondamentaux du processus :

- Les données
- Les règles de calcul
- La programmation

Les données et les règles de calcul correspondent, en effet, à deux tables différentes qu'il est possible de modifier et de vérifier séparément, la programmation se réduisant à des requêtes SQL très simples faciles à contrôler.

Ce texte n'engage que son auteur : Francis Malherbe