L'informatique des entrepôts de données

Daniel Lemire

SEMAINE 9
Les opérations OLAP

9.1. Présentation de la semaine

Nous avons vu la semaine précédente qu'il est possible de définir partiellement le paradigme OLAP avec les schémas en étoile, la table de faits et le cube de données. Cependant, le paradigme prend tout son sens avec les requêtes qu'il permet d'effectuer rapidement : roll-up, drill-down, slice, dice, pivot, iceberg, etc. Un analyste combiners plusieurs de ces requêtes afin d'obtenir l'information qu'il souhaite.

Dans ce cours, nous utiliserons les termes anglais pour désigner les différentes opérations OLAP. Quelques traductions possibles sont proposées dans la Table 1. Dans la pratique, on utilise surtout les termes anglais.

Table 1. Principaux termes désignant les opérations OLAP

terme anglais usuel	traduction possible
roll-up	résumé ou cumul
drill-down	zoom avant
slice	trancher
dice	couper en dés

9.2. Roll-up et drill-down

On effectue un roll-up[1] lorsqu'on souhaite obtenir moins de détails, ou qu'on souhaite obtenir une vue d'ensemble. On peut d'abord faire un roll-up à même une dimension qui dispose d'une hiérarchie. Prenons l'exemple des données suivantes :

France	3
Espagne	1
Suisse	41
Canada	24
É.-U.	14

La dimension Lieu est présentée à la granularité Pays. Si on juge que cela fournit trop de détails inutiles, on peut demander un roll-up vers le continent :

Europe	45
Amérique du Nord	38

Dans cet exemple, nous utilisons la somme de la mesure pour effectuer l'aggrégation, mais on pourrait tout aussi bien utiliser une autre opération comme la moyenne ou le maximum. L'opération contraire, soit le passage des données portant sur les continents vers les données portant sur les pays, est un drill-down.

Il existe aussi une autre forme de roll-up ou de drill-down, qui permet de faire apparaître ou de disparaître une dimension au complet. Prenons l'exemple des pays et de leur mesure. Imaginons que nous souhaiitons décomposer les mesures selon une autre dimension, par exemple la couleur :

[1]On utilise parfois les termes anglais consolidate ou drill-up pour désigner un roll-up.

France	rouge	2
France	bleu	1
France	orange	0
Espagne	rouge	1
Espage	bleu	0
Espage	orange	0
Suisse	rouge	40
Suisse	bleu	1
Suisse	orange	0
Canada	rouge	20
Canada	bleu	4
Canada	orange	0
É.-U.	rouge	10
É.-U.	bleu	4
É.-U.	orange	0

Cette nouvelle table représente un drill-down par rapport à la table originale, parce qu'il y a plus de détails. L'opération inverse, celle qui consiste à faire disparaître l'attribut couleur, est un roll-up.

Il est important d'observer que lorsque l'on fait un drill-down ou un roll-up, les mesures sont modifiées en conséquence. Ainsi, si on fait la somme des mesures, et qu'il ne s'agit que de mesures positives, un roll-up va toujours générer des valeurs de mesures plus importantes. Au contraire, le drill-down réduira la valeur des mesures.

9.3. Dice

Une opération de type dice consiste à générer une nouvelle table, qui comporte autant de dimensions et les mêmes valeurs de mesures, mais où l'on restreint certaines des dimensions à quelques mesures. Prenons l'exemple de la table précédente, qui comprend à la fois des pays et des couleurs. Imaginons que nous voulions éliminer la couleur orange et ne conserver que les pays européens :

France	rouge	2
France	bleu	1
Espagne	rouge	1
Espage	bleu	0
Suisse	rouge	40
Suisse	bleu	1

Le résultat est un dice. Contrairement au drill-down et au roll-up, lorsque l'on effectue un dice, les valeurs ne sont pas modifiées et l'on

conserve les mêmes dimensions et les mêmes valeurs. On se contente d'éliminer certaines des valeurs indésirables.

Si on représente nos données selon deux axes, un pour chaque dimensions, on constate qu'on peut illustrer le dice comme la sélection d'un sous-ensemble des mesures :

	rouge	bleu	orange
France	X	X	
Espagne	X	X	
Suisse	X	X	
Canada			
É.-U.			

Cette illustration explique le terme dice : on a coupé et extrait un sous-cube.

9.4. Slice

L'opération roll-up peut réduire le nombre de dimensions, mais les mesures sont modifiées de manière correspondante. L'opération dice préserve les valeurs des mesures, mais omet un partie des données. L'opération slice, quant à elle, réduit le nombre de dimensions (comme une opération roll-up) mais, tout comme l'opération dice, elle laisse les valeurs de mesures inchangées, se contenant d'omettre une partie des données.

Imaginons que dans l'exemple des pays et des couleurs, nous nous intéressions qu'à la couleur rouge. Dans un tel cas, nous pourrions extraire les enregistrements correspondants dans la table de faits :

France	rouge	2
Espagne	rouge	1
Suisse	rouge	40
Canada	rouge	20
É.-U.	rouge	10

On se rend compte, par contre, que de conserver la dimension couleur ne sert pas à grand chose. Il vaut mieux l'omettre :

France	2
Espagne	1
Suisse	40
Canada	20
É.-U.	10

Si on représente nos données selon deux axes, on voit que l'opération slice correspond à une tranche :

	rouge	bleu	orange
France	X		
Espagne	X		
Suisse	X		
Canada	X		
É.-U.	X		

9.5. Pivot

Le pivot peut désigner l'opération triviale qui consiste à permuter les axes dans une représentation bidimensionnelle des données [**4**]. Par exemple, on pourrait échanger les dimensions couleurs et pays. Cette opération prend son sens dans un chiffrier électronique (par ex. Excel).

Le pivot peut aussi correspondre au remplacement d'un roll-up par un autre. Imaginons que vous ayez une table de faits ayant les dimensions pays, couleur et mois. Après un roll-up sur la dimension mois, vous obtenez une vue (un cuboïde) comportant les dimensions pays et couleur. Vous pourriez alors souhaiter passer de ce roll-up à un roll-up sur la dimension pays ou couleur, générant respectivement une vue comportant les dimensions couleur et mois puis couleur et mois.

On peut expliquer le terme pivot en imaginant un cube en trois dimensions que l'on fait tourner afin d'en consulter les différentes faces.

9.6. Iceberg

Il arrive qu'une table comporte trop d'enregistrements pour qu'on puisse les consulter facilemente, mais sans pour autant qu'un roll-up soit intéressant. Par exemple, dans notre exemple original des pays, la plupart des pays ont une mesure correspondante relativement faible. On pourrait souhaiter ne conserver que les enregistrements correspondant à des mesures significatives. Par exemple, plaçons un seuil de 20 sur la valeur des mesures :

Suisse	41
Canada	24

On pourrait, au contraire, ne s'intéresser qu'au pays ayant une faible mesure. Au lieu de fixer un seuil, on aurait pu se contenter de spécifier que nous ne voulions obtenir que les deux enregistrements ayant les mesures les plus significatives.

On désigne ce type d'opération par le terme iceberg [**2**]. On fait bien sûr référence au fait que, dans un iceberg, seule une petite partie de la glace est visible (hors de l'eau).

9.7. Skyline

Lorsqu'il y plus qu'une mesure, la manière d'appliquer la méthode iceberg n'est pas toujours claire. Prenons l'exemple de restaurants qui ont à la fois un indice de qualité et un indice de prix :

	indice de qualité	indice de prix
Restaurant chez Maurice	5	1
Restaurant chez Pierre	4	1
Macmodal	3	3
La belle patate	1	5
La dame en bleu	1	4

Idéalement, on souhaite sélectionner les restaurants qui ont de bons indices de prix et de qualité. Malheureusement, on voit que le restaurant avec le meilleur indice de qualité a aussi les moins bons prix (le restaurant chez Maurice).

Nous pouvons tout de même éliminer certains candidats indésirables. On dit qu'un enregistrement domine l'autre s'il lui est supérieur ou égal à tous les points de vue, et qu'il lui est supérieur en ce qui a trait à au moins une mesure. On voit que le restaurant chez Pierre est dominé par le restaurant chez Maurice : ils ont les mêmes prix, mais le restaurant chez Maurice a une meilleur qualité. Aussi, le restaurant La belle patate domine le restaurant La dame en bleu.

En excluant tous les enregistrements qui sont dominés par au moins un autre enregistrement, on obtient le skyline [**1**] :

	indice de qualité	indice de prix
Restaurant chez Maurice	5	1
Macmodal	3	3
La belle patate	1	5

Bien que les requêtes de type Skyline ne soient pas systématiquement associées au paradigme OLAP, leur pertinence est indéniable.

9.8. Diamond

Les requêtes de type diamond (ou diamant en français) sont une forme de dice combinée avec des requêtes de type iceberg [**3**]. On cherche à extraire un sous-cube de taille maximal répondant à certaines contraintes. Par exemple, imaginez que vous souhaitiez obtenir une liste de magazins et de types de produits, tel que les ventes des magazins sur les produits sélectionnés excèdent 10 millions et que les ventes des produits dans les magazins sélectionnés excèdent 5 millions (voir Table 2). Il s'agit là d'un exemple typique de requête de type diamond. Bien

Table 2. Ventes (en millions de dollars) avec un diamant 5,10 identifié en gris : on ne conserve que les magazins ayant des ventes de plus de 10 millions $ concernant des types produits rapportant plus de 5 millions $.

	Chicago	Montréal	Miami	Paris	Berlin	Totaux
TV	3.4	0.9	0.1	0.9	2.0	7.3
Caméra	0.1	1.4	3.1	2.3	2.1	9.0
Téléphone	0.2	6.4	2.1	3.5	0.1	12.3
Appareil photo	0.4	2.7	5.3	4.6	3.5	16.5
Console de jeux	3.2	0.3	0.3	2.1	1.5	7.4
Lecteur DVD	0.2	0.5	0.5	2.2	2.3	5.7
Totaux	7.5	12.2	11.4	15.6	11.5	58.2

que simple en apparence, les requêtes diamond peuvent être difficiles à calculer efficacement.

9.9. Questions d'approfondissement

(a) Un ministre du gouvernement canadien souhaite obtenir une nouvelle version du rapport, qui se limite aux statistiques québécoises. Comment nomme-t-on ce type de requête ?

(b) Au lieu d'obtenir des statistiques par employé, le patron souhaite des statistiques par groupe de travail. Comment nomme-t-on ce type de requête ?

(c) Étant donné une table de faits comportant 10 dimensions, combien y a-t-il de pivots possibles sur une vue à deux dimensions ?

(d) Quel type d'opération est plus coûteux : dice, slice, roll-up ou drill-down ?

9.10. Réponses suggérées

(a) Slice.

(b) Roll-up.

(c) Il y a $10 \times 9 = 90$ paires de dimensions distinctes, donc 90 pivots. On doit diviser par deux si on ne souhaite pas distinguer entre (pays, couleur) et (couleur, pays).

(d) Les opérations de type roll-up exigent l'agrégation des mesures, ce qui peut être coûteux. Les opérations dice et slice ne font que de la sélection d'enregistrement, ce qui est peu coûteux lorsque les colonnes sont indexées de manière appropriée.

BIBLIOGRAPHIE

1. S. Börzsönyi, D. Kossmann, and K. Stocker. The skyline operator. In *ICDE '01*, pages 421–430. IEEE Computer Society, 2001.
2. M. Fang, N. Shivakumar, H. Garcia-Molina, R. Motwani, and J. D. Ullman. Computing iceberg queries efficiently. In *VLDB'98*, pages 299–310, 1998.
3. H. Webb, O. Kaser, and D. Lemire. Pruning attribute values from data cubes with diamond dicing. In *International Database Engineering and Applications Symposium (IDEAS'08)*, pages 121–129, 2008.
4. C. M. Wyss and E. L. Robertson. A formal characterization of pivot/unpivot. In *CIKM '05*, pages 602–608, 2005.