ВИКОРИСТАННЯ ГРАНИЧНИХ СУМ ДЛЯ РЕАЛІЗАЦІЇ СКЛАДНИХ ЗАПИТІВ У БАГАТОВИМІРНОМУ ПРОСТОРІ ДАНИХ

УДК 504.05:004.047

СКАРГА-БАНДУРОВА Інна Сергіївна

к.т.н., доцент, кафедра КІ, Технологічний інститут Східноукраїнського національного університету ім. В. Даля

Наукові інтереси: теорія прийняття рішень, інформаційні технології в промисловій безпеці та екології, медична інформатика.

ВСТУП

Можливість подання інформації в абстрактній і незалежної від реалізації формі має вирішальне значення в життєвому циклі інформаційної системи, не тільки на етапі її проектування, але й у фазі експлуатації. Процеси обробки та аналізу даних істотно залежать від використовуваної моделі представлення даних, оскільки вибір моделі багато в чому визначає набір застосовних операцій для обробки даних і швидкість проведення аналізу даних.

Це особливо вірно в контексті створення сховищ даних (Data Warehouse), оперативної аналітичної обробки даних (On-Line Analytical Processing) і добування інформації з даних (Knowledge Discovery in Databases) при реалізації природоохоронних заходів, де через рівень складності розробка додатків та управління ними представляють собою тривалий ітераційний процес, спрямований на постійне вдосконалення існуючої структури і усунення помилок.

До інформаційного забезпечення природоохоронних інституцій відносяться збір, обробка, аналіз, синтез даних, побудова моделей, створення баз даних для користувачів. Важливим питанням у поданні запитів та аналізі потоків екологічних даних є те, що вони за своєю природою є багаторівневими і тому вимагають, великомасштабного моделювання [1]. Великі потоки даних породжувані набором джерел даних природно вимагають для своєї обробки відповідних засобів розширеного аналізу, моделей інтелектуального аналізу даних, що виходять за рамки традиційних рішень відмінних від СУБД з SQL- інтерфейсами. В даний час є ряд робіт [2-10], в яких відображений сучасний рівень використання інформаційних засобів в аналізі та обробці даних. У [11] багатовимірний простір даних визначається як варіація реляційної моделі, яка використовує багатовимірні моделі для організації даних і характеризує відносини між ними. При обробці багатовимірних даних у переважній кількості випадків вирішується завдання їх агрегування. Зокрема, агрегації даних присвячені роботи [12-14], функції агрегації для текстових сховищ даних представлені в [7], в роботах [15,16] розглядаються технології формування багатовимірних запитів для документальних сховищ даних. Операції агрегатного типу описані в патентах [17,18], моделювання джерел багатовимірних даних в [19,20].

Основними вимогами до подання та обробки даних є [21]:

- використання багатовимірного подання даних з підтримкою ієрархій і множинних ієрархій;
- підтримка статистичного, оперативного та інтелектуального аналізу даних, а також аналізу, що визначається бізнес-процесами організації, незалежно від програми, візуалізації результатів в доступному для кінцевого користувача вигляді;
- забезпечення однаково високої швидкості виконання всіх запитів до системи.

ПРОБЛЕМНА СИТУАЦІЯ

Стосовно останньої вимоги, доцільність розробки і використання методів оптимізації сумарних запитів до баз даних не викликає сумнівів, як для звичайних, так і

для багатовимірних БД, однак підходи можуть відрізнятися. Так, для реляційної БД у [22] наводиться приклад запиту виду:

$$R(A_1, A_2, A_3, A_4) = \\ = \sigma_{A_2=11 \wedge A_3=55}(R_1(A_1, A_2) \times R_2(A_3, A_4)). \quad (1)$$

Якщо припустити, що відношення $R_1(A_1,A_2)$ містить 10 000 записів, причому 15 записів з $A_2 = 11$, а відношення $R_2(A_3,A_4)$, представлене файлом, що містить 20 000 записів, в якому 50 записів з $A_3 = 55$ і спробувати відразу ж виконувати запит, то для виконання декартова добутку необхідно здійснити 200000000 звернень до записів.

Запит (1), перетворений до виду (2) дозволяє істотно скоротити обсяг необхідної обчислювальної роботи:

$$R(A_1, A_2, A_3, A_4) = \\ = (\sigma_{A_2=11} R_1(A_1, A_2)) \times (\sigma_{A_3=55} R_2(A_3, A_4)), \quad (2)$$

де, для обчислення $\sigma_{A_2=11} R_1(A_1, A_2)$ потрібно виконати 10 000 звернень до записів, а для обчислення $\sigma_{A_3=55} R_2(A_3, A_4)$ 20 000 звернень. І для остаточного формування відповіді на запит необхідно виконати 750 звернень до записів файлів проміжних результатів. У підсумку буде потрібно виконати 30 750 звернень, що істотно менше, ніж у попередньому випадку.

Що стосується багатовимірних структур, то швидкість виконання запитів є ще більш критичною по відношенню до типів запитів і обсягів вибірки. За даними розробника, у системі обробки інформації про навколишнє середовище і здоров'я населення EHIPS [23], при простому запиті реальний час читання в БД може досягати 10 годин і більше (табл.1).

Таблиця 1

Результати читання БД для реальної вибірки згідно даних [23]

Блок і координати	Кількість інтервалів для реальної вибірки	Об'єм, МБ	Час читання БД
Блок викиду	3×1000×200 = 600 000	2,5	Орієнтовно: 5-10 хвилин
Забруднювач	200		
Джерела	1000 (на 1 підприємство)		При простому SQL-запиті: >1 години
Час	1		
Територія	3 підприємства		
Блок захворювань	10×150×379×110 = 62 535 000 Дані: 10×150×366×7 = 3 843 400	250	Орієнтовно: 0,5-1 година
Діагноз	10		При простому SQL-запиті: >10 годин
Статевозріла група	150 (по містах)		
Час	379		
Територія	110 (7 поліклінік)		

В рамках означеної проблеми в роботі вирішується завдання знаходження сумарних значень у виділених інтервалах багатовимірного куба.

Постановка завдання. Куб даних являє фактичні дані, на яких фокусується аналіз і пов'язує виміри з координатами, визначеними на множині рівнів вимірювань. Вимірювання у загальному випадку представляє бізнес-перспективу, при якій аналіз даних повинен бути виконаний і організований у вигляді ієрархії рівнів, що відповідають різним варіантам угруповання її елементів.

Проблема знаходження сумарних значень у виділених інтервалах у d-вимірному кубі даних формулюється у вигляді задачі знаходження граничних сум масиву A:

$$Sum([l_1; h_1], [l_1; h_2], ..., [l_d; h_d]) = \\ = \sum_{i_1=l_1+1}^{h_1} \sum_{i_2=l_2+1}^{h_2} ... \sum_{i_d+1}^{h_d} A(i_1, ..., i_d)$$

.

Для зручності запису, діапазон усіх цілих чисел ij позначений як $[l_j;h_j]$, де $l_j < ih \le h_j$, тобто діапазон не містить нижню межу l_i і включає в себе верхню межу h_i.

Область $([l_1;h_1],[l_1;h_2],...,[l_d;h_d])$ позначає d - мірний простір, обмежений $l_j < ij \le h_j$ у вимірі j для всіх $j \in D$.

Розмір області відповідає числу цілих точок, визначених у ній. Так розмір області $([l_1;h_1],[l_1;h_2],...,[l_d;h_d])$ дорівнює $\prod_{j=1}^{d}(h_j - l_j)$.

Приклад формулювання завдання для задачі інвентаризації відходів в інформаційно-аналітичній системі управління природоохоронною діяльністю:

Знайти значення сумарних викидів від постійних джерел із середнім значенням, заданим на інтервалі від l_1 до h_1 протягом декількох років, з року l_2 по рік h_2 по всім виробництвам.

ОСНОВНА ЧАСТИНА

Традиційний підхід до вирішення даної задачі передбачає послідовне виконання ряду операцій, що містять реляційні запити join, group by, select <fields list> from і/або низку операцій на багатовимірних кубах (прості: level_climbing, packing, projection, і складні, такі як navigation, slicing та ін.), однак, аналогічно представленому вище прикладу, ряд авторів [24-26] відзначає суттєвий час затримки при реалізації запиту до багатовимірної бази даних.

Враховуючи значний обсяг аналізованих даних, у роботі пропонується використовувати підхід на основі граничних (префіксних) сум.

Згідно [27], префіксна сума - це часткова сума значень у попередніх осередках вихідного масиву.

Загальна схема знаходження сумарних значень у виділених інтервалах d-вимірного куба включає виконання наступних основних етапів (рис.1):

1. Подання куба даних у вигляді d-вимірного масиву A.

2. Трансформація d-вимірного масиву A в масив префіксних сум P.

3. Вибір інтервалів для розрахунку.

4. Розрахунок граничних сум для зазначеної області.

Крок 1. Вибір підмножини вимірювань в кубі даних.

Крок 2. Трансформація d-вимірного масиву A в масив префіксних сум P.

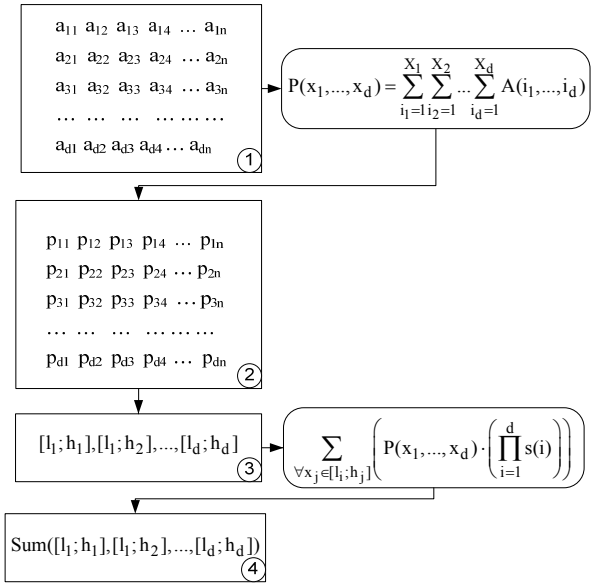

Рисунок 1 – Схема рішення задачі знаходження граничних сум у в d-вимірному кубі

Для розрахунку набору префіксних сум уздовж обраних вимірювань, заснованого на агрегатних значеннях куба використовуємо алгоритм обчислення префіксної суми, описаний в [18,27].

$$P(x_1,...,x_d) = \sum_{i_1=1}^{X_1}\sum_{i_2=1}^{X_2}\dots\sum_{i_d=1}^{X_d} A(i_1,...,i_d), \qquad (3)$$

У результаті розрахунку (3) масив A перетворюється в масив префіксних сум P (рис.2).

A

a11	a12	a13	a14	…	a1n
a21	a22	a23	a24	…	a2n
a31	a32	a33	a34	…	a3n
…	…	…	…	…	…
am1	am2	am3	am4	…	amn

P

p_{11}	p_{12}	p_{13}	p_{14}	…	p_{1n}
p_{21}	p_{22}	p_{23}	p_{24}	…	p_{2n}
p_{31}	p_{32}	p_{33}	p_{34}	…	p_{3n}
…	…	…	…	…	…
p_{m1}	p_{m2}	p_{m3}	p_{m4}	…	p_{mn}

Рисунок 2 – Перетворення d-мірного масиву A в масив префіксних сум P

Крок 3. Генерація діапазону сум на основі розрахованих префіксних сум.

Для розрахунку діапазону використовується ф. (4).

$$Sum([l_1;h_1],[l_1;h_2],...,[l_d;h_d]) = \\ = \sum_{\forall x_j \in [l_i;h_j]} \left(P(x_1,...,x_d) \cdot \left(\prod_{i=1}^{d} s(i) \right) \right), \qquad (4)$$

.

$$s(i) = \begin{cases} 1 & при \quad x_i = h_i, \\ -1 & при \quad x_i = l_i. \end{cases}$$

де

Крок 4. Підрахунок набору префіксних сум уздовж обраного виміру.

При d = 2, гранична сума $Sum([l_1;h_1],[l_2;h_2])$ може бути обчислена наступним чином:

$$Sum([l_1;h_1],[l_2;h_2]) = \\ = P(h_1,h_2) - P(h_1,l_2) - P(l_1,h_2) + P(l_1,l_2). \qquad (5)$$

Наприклад, для отримання сумарних викидів від постійних джерел із середнім значенням, заданим на інтервалі від l_1 до h_1 протягом декількох років, з року l_2 по рік h_2 по масиву префіксних сум P', отриманому з A' достатньо використати ф. (5):

A'

	1	2	3	4	5	6
1	20	30	10	20	30	40
2	15	20	40	30	50	10
3	20	10	10	40	30	15

P'

	1	2	3	4	5	6
1	20	50	60	80	110	150
2	35	85	135	185	265	315
3	55	115	175	265	375	440

$$Sum([l_1;h_1],[l_1;h_2]) = \\ = P(h_1,h_2) - P(h_1,l_{2-1}) - P(l_{1-1},h_2) + P(l_{1-1},l_{2-1}) = \\ = p_{34} - p_{14} - p_{31} + p_{11}.$$

$$Sum([2;4],[2;3]) = 265 - 80 - 55 + 20 = 150.$$

A'

		l1			h1		
		1	2	3	4	5	6
	1	20	30	10	20	30	40
l2	2	15	20	40	30	50	10
h2	3	20	10	10	40	30	15

P'

		l1-1	l1		h1		
		1	2	3	4	5	6
l2-1	1	20	50	60	80	110	150
l2	2	35	85	135	185	265	315
h2	3	55	115	175	265	375	440

Висновки. Розглянута концепція орієнтована на зберігання і обробку даних при реалізації складних запитів до багатовимірних баз даних. Отримані результати дають підставу вважати, що завдяки збереженню масиву префіксних сум P, який відповідає розміру куба даних, всі запити будь-якого діапазону для даного куба будуть мати постійний час відповіді, незалежно від розміру куба, обмеженого безпосередньо запитом. Відповідь на запит значення діапазону може вимагати доступ до деяких осередків куба даних для додаткової інформації, однак середній час і складність значно знижені.

ЛІТЕРАТУРА:

1. Han, J. Stream Cube: An Architecture for Multi-Dimensional Analysis of Data Streams / J. Han, Y. Chen, G. Dong, J. Pei, B. W. Wah, J. Wang, Y. D. Cai / Distributed and Parallel Databases. – 2005. - 18(2). - pp. 173–197.
2. Barsegjan A.A. Metody i modeli analiza dannykh: OLAP i Data Mining / A. A. Barsegjan, M. S. Kuprijanov, V. V. Stepanenko, I. I. Kholod. – SPb.: BKhV-Peterburg, 2004. – 336 s.
3. Kurenkov, N.I. Osobennosti analiza mnogomernykh dannykh [Ehlektronnyjj resurs] / N. I. Kurenkov – Rezhim dostupa www. URL: http://prr.ru/Papers/analyze.doc – 15.04.2014 g.
4. Shakhovs'ka, N. B. Skhovishha ta prostori danikh / N. B. Shakhovs'ka, V. V. Pasichnik. - L'viv: L'vivs'ka politekhnika, 2009. – 244 s. – ISBN 978-966-553-796-0.
5. Data handling in science and technology. Scientific data ranking methods: theory and applications / M. Pavan, R. Todeschini (Eds.). – Elsevier, 2008. – 215 p.
6. Ravat, F. A Conceptual Model for Multidimensional Analysis of Documents / F. Ravat, O. Teste, R. Tournier, G. Zurlfluh / Proc. of 26th Int. Conf. on Conceptual Modeling, Auckland, New Zealand, November 5-9, 2007. - 2007. - p. 550-565.

7. Ravat, F. OLAP Aggregation Function for Textual Data Warehouse [Ehlektronnyjj resurs] / F. Ravat, O. Teste, R. Tournier. – Rezhim dostupa: www. URL: ftp://irit.fr/IRIT/SIG/ICEIS_Agg.pdf – 15.04.2014 g.
8. Samet, H. Foundations of multidimensional and metric data structures / H. Samet. - The Morgan Kaufmann Series, 2006. – 993 p.
9. Samtani, S. Recent advances and research problems in data warehousing / S. Samtani, M. Mohania, V. Kumar, Y. Kambayashi // Proc. of the Workshops on Data Warehousing and Data Mining: Advances in Database Technologies. – 1998. – pp. 81-92.
10. West, M. Developing high quality data models / M. West. – Morgan Kaufmann, 2011. - 408 p.
11. O'Brien, J. Management Information Systems / J. O'Brien, G. Marakas. - McGraw-Hill/Irwin, 2010. – 712 p.
12. Khrustalev, E.M. Agregacija dannykh v OLAP-kubakh [Ehlektronnyjj resurs] / E. M. Khrustalev. – Rezhim dostupa: www. URL: http://www.interface.ru/misc/mut.htm - 6.02.2014 g.
13. Gray, J. Data Cube: A Relational Aggregation Operator Generalizing Group-By, Cross-Tab, and Sub-Totals / J. Gray, S. Chaudhuri, A. Bosworth, A. Layman, D. Reichart, M. Venkatrao, F. Pellow, H. Pirahesh, / Data Mining and Knowledge Discovery. – 1997. - 1(1). – 29–54.
14. Lenz, H.-J. Summarizability in OLAP and Statistical Data Bases / H.-J. Lenz, A. Shoshani // Proc. of 9th Int. Conf. on Scientific and Statistical Database Management. – 1997. – pp.132-143.
15. Nassis, V.Conceptual Design of XML Document Warehouses / V. Nassis, R. Rajugan, T. S. Dillon, J. W. Rahayu, // 6th Int. Conf. on Data Warehousing and Knowledge Discovery (DaWaK), LNCS 3181, Springer. – 2004. – pp.1–14.
16. Tseng, F.S.C. The concept of document warehousing for multi-dimensional modeling of textual-based business intelligence. / F.S.C Tseng, A.Y.H. Chou //J. of Decision Support Systems (DSS). – 2006. – vol. 42(2). – p. 727–744.
17. 17. Efficient multidimensional data aggregation operator implementation: US Patent 5,822,751 A / Gray J., Reichart D. C. ; assignee: Microsoft Corporation. – Appl. No. 768,105; filed: 16.12.1996 ; date of patent: 13.10.1998.
18. Method and system for performing range max/min queries on a data cube: US Patent 5,926,820 A / Agrawal R, Ho Ch.-T., Megiddo N. ; assignee: International Business Machines Corporations. – Appl. No. 08/808,046; filed: 27.02.1997 ; date of patent: 20.07.1999.
19. Modeling multidimensional data sources : US Patent 2007/0208721 A1 / Zaman K. A., Song S., Suen E.Sh.-L. – Appl. No. 10/726,338; filed: 01.12.2003 ; date of patent: 06.09.2007.
20. Populating data cubes using calculated relations: US Patent 6,970,874 B2 / Egilsson A. S., Gudbjartsson H. ; assignee: Decode genetics ehf. – Appl. No. US 10/216,670; filed: 08.08.2002 ; date of patent: 29.10.2005.
21. Belov, V.N. Modeli mnogomernogo predstavlenija i obrabotki dannykh na osnove algebry kortezhejj v informacionno-analiticheskojj sisteme : avtoref. dis. kand. tekhn. nauk: 05.13.17, 05.13.01 / V. N. Belov. – Penza., 2012. – 20 s.
22. Sedov, N.N. Vvedenie v komp'juternuju matematiku : ucheb. posobie. / N. N. Sedov. - Rossijjsk. gosud. otkr. tekhn. un-t putejj soobshh. M.: 2002. – 89 c.
23. Sistema obrabotki informacii ob okruzhajushhejj srede i zdorov'e naselenija EHIPS / [Elektronnijj resurs] – Rezhim dostupu: www. URL: http://www.iki.rssi.ru/ehips/ - 12.03.2014 g.
24. Kudrjavcev, Ju. A. Algoritmy ehffektivnojj obrabotki MOLAP-kubov: avtoref. dis. … kand. fiz.-mat. nauk : 05.13.11 / Ju. A. Kudrjavcev. – M., 2009. - 17 s.
25. Gray, J. Data Cube: A Relational Aggregation Operator Generalizing Group-By, Cross-Tab, and Sub-Totals / J. Gray, S. Chaudhuri, A. Bosworth, A. Layman, D. Reichart, M. Venkatrao, F. Pellow, H. Pirahesh, / Data Mining and Knowledge Discovery. – 1997. - 1(1). – 29–54.
26. Method and system for performing range max/min queries on a data cube: US Patent 5,926,820 A / Agrawal R, Ho Ch.-T., Megiddo N. ; assignee: International Business Machines Corporations. – Appl. No. 08/808,046; filed: 27.02.1997 ; date of patent: 20.07.1999.
27. Prefix sum pass to linearize a-buffer storage: US Patent 2008/0316214 / Peeper C. ; assignee: Microsoft Corporation. – Appl. No. 11/766,091; filed: 20.06.2007 ; date of patent: 25.12.2008.