УДК 004.658.2

Е.А. НОВОХАТСКАЯ
Одесский национальный политехнический университет

РАСЧЕТ КОЭФФИЦИЕНТА МАТЕРИАЛИЗАЦИИ ПРИ ОЦЕНКЕ ЗАПРОСОВ ДЛЯ ОБСЛУЖИВАНИЯ МАТЕРИАЛИЗОВАННЫХ ПРЕДСТАВЛЕНИЙ

В данной работе предложен критерий выбора запросов-кандидатов на материализацию с учетом временных и статистических показателей их выполнения. Полученный числовой коэффициент покрыл такие важные показатели выполнения запросов, как объем затраченных при формировании результата ресурсов, частоту появление запросов в информационной системе, а также частоту обновления базовых таблиц. Последний показатель позволил исключить создание материализованных представлений, обслуживание которых потребовало бы значительного числа вычислительных ресурсов базы данных. Поскольку коэффициент рассчитывается на основании данных журнала транзакций, он может быть использован на этапе поиска запросов-кандидатов на материализацию, когда время выполнения запроса при существующем материализованном представлении еще не известно.

Ключевые слова: коэффициент материализации, материализованное представление, оценка запросов.

К.А. НОВОХАТСЬКА
Одеський національний політехнічний університет

РОЗРАХУНОК КОЕФІЦІЄНТА МАТЕРІАЛІЗАЦІЇ ПРИ ОЦІНЦІ ЗАПИТІВ ДЛЯ ОБСЛУГОВУВАННЯ МАТЕРІАЛІЗОВАНИХ ПРЕДСТАВЛЕНЬ

У даній роботі запропоновано критерій вибору запитів-кандидатів на матеріалізацію з урахуванням часових і статистичних показників їх виконання. Отриманий числовий коефіцієнт покрив такі важливі показники виконання запитів, як обсяг витрачених при формуванні результату ресурсів, частоту появи запитів в інформаційній системі, а також частоту оновлення базових таблиць. Останній показник дозволив виключити створення матеріалізованих представлень, обслуговування яких вимагало б значного числа обчислювальних ресурсів бази даних. Оскільки коефіцієнт розраховується на підставі даних журналу транзакцій, він може бути використаний на етапі пошуку запитів-кандидатів на матеріалізацію, коли час виконання запиту при існуючому матеріалізованому представленні ще не відомо.

Ключові слова: коефіцієнт матеріалізації, матеріалізоване представлення, оцінка запитів.

K.A. NOVOKHATSKA
Odessa National Polytechnic University

CALCULATING THE MATERIALIZATION FACTOR IN QUERY EVALUATION DURING THE MAINTENANCE OF MATERIALIZED VIEWS

In this paper we propose a criterion for selecting the queries to materialize that takes into consideration the timing and statistic indices of query execution. The resulting numerical coefficient covered such important indicators of query execution as the amount of resources consumed during formation of the result, frequency of query occurrence in the information system as well as the refresh rate of base tables. The latter figure would eliminate the creation of materialized views which require a significant number of database computational resources. Since the coefficient is calculated based on the transaction log data, it can be used for finding queries which are candidates to materialize when the query execution time within the existing materialized view is not known yet.

Keywords: coefficient of materialization, materialized view, query evaluation.

Постановка проблемы

Одним из способов повышения производительности работы СУБД является использование материализованных представлений (МП). Они представляют собой предварительно вычисленные и сохраненные на диск результаты выполнения запросов. МП позволяют сократить время выполнения запросов за счет снижения числа обращений к физической памяти и устранения операций сортировки и объединения.

Выбор запросов-кандидатов на материализацию является одной из важных задач в области автоматизации МП [1]. В первую очередь ее нетривиальность обусловлена вычислительной сложностью.

Для поиска запросов-кандидатов на материализацию необходимо проанализировать журнал транзакций за длительный период времени и выбрать запросы, согласно нескольким критериям. Данные критерии связаны с синтаксической схожестью запросов, а также с временными и статистическими показателями выполнения запросов.

Целесообразно создавать МП для наиболее ресурсоемких и часто выполняемых запросов, поступающих в систему [2]. Помимо этого, необходимо учесть, что при каждом обновлении данных в базовых таблицах (БТ) требуется обновить МП, созданных на их основе. Т.е., материализовывать запросы для часто обновляемых данных неэффективно. Поэтому частота обновления БТ является еще одним критерием, который необходимо учитывать при выборе запросов-кандидатов на материализацию.

Таким образом, при сравнении запросов необходимо учесть объем ресурсов, потребляемых при их вычислении, время и частоту их появления в системе, а также частоту обновления БТ, участвующих в запросах.

Для автоматизации задачи поиска запросов-кандидатов на материализацию, необходимо формализовать перечисленные выше критерии и привести их к виду численного коэффициента, который бы позволил эффективно сравнивать запросы между собой.

Анализ последних исследований и публикаций

В работах [3-4] впервые предложена технология создания и обслуживания МП. В том числе сформулированы критерии выбора запросов-кандидатов на материализацию. Для оценки запроса вводится понятие его стоимости, которое определяется как суммарное время выполнения всех одинаковых запросов:

$$S_{Z_k} = \sum_{i=1}^{N_k} T(Z_{ki}),$$

где N_k – количество появлений запроса Z_k;

$T(Z_{ki})$ – время выполнения запроса Z_k в его i-ое появление в системе;

$k=1,K$ – мощность множества входных запросов.

Эффективность МП рассчитывается как разность суммарной стоимости выполнения запросов без применения МП и с его применением:

$$I = \sum_{k=1}^{K} \left(S_{Z_k} - \left(S_{МП_k} + S_{ОБН_k} \right) \right) - S_{ДОП},$$

где $S_{МП\,k}$ – стоимость запроса Z_k при использовании МП;

$S_{ОБН\,k}$ – стоимость обновления МП, созданного на основе запроса Z_k;

$S_{ДОП}$ – дополнительные затраты ресурсов на обслуживание МП.

Предложенное авторами решение на этапе создания МП оценивает запросы только с точки зрения времени их выполнения. Однако, встречаются запросы, которые выполняются в системе единожды, но требуют значительного времени вычисления результата. Такие запросы не должны попадать в кандидаты на материализацию. С другой стороны, запросы, выполняемые относительно быстро, но часто, могут быть материализованы. Таким образом, предложенная авторами стоимость запроса не полностью покрывает сформулированные критерии выбора запросов-кандидатов на материализацию, а именно не учитывает количество затраченных ресурсов, частоту появления запроса и частоту обновления БТ.

В работах [5-6] было продолжено исследование данного вопроса. Была предложена технология оценки эффективности МП, сравнивающая производительность выполнения запросов при использовании МП и без него в разные периоды времени работы системы. Доказано, что одно и то же МП может как положительно, так и негативно влиять на производительность выполнения запросов. На основании данного утверждения были предложены алгоритмы включения и выключения МП, коррелирующие с периодичностью работы информационной системы.

Предложенная технология полностью покрывает вопрос обслуживания МП после его создания, однако она не акцентирует внимание на стадии поиска запросов-кандидатов на материализацию, когда производительность выполнения запросов с использованием МП еще не известна.

Формулировка цели исследования

Целью данной работы является формализация критериев выбора запросов-кандидатов на материализацию и приведение их к виду численного коэффициента, который бы позволил эффективно сравнивать запросы между собой на этапе анализа журнала транзакций СУБД.

Изложение основного материала исследования

Пусть в результате анализа журнала транзакций информационной системы за некоторый продолжительный период времени T было сформировано множество неуникальных запросов разных

типов (*SELECT, INSERT, UPDATE, DELETE*) с указанием времени их выполнения и затраченных ресурсов:

$$Q = \bigcup q < text, \tau, b >,$$

где *text* – текст запроса;

τ – время выполнения запроса;

b – суммарное число блоков данных, прочитанных с диска или буфера и обработанных СУБД для формирования результата.

Сформулируем следующие критерии выбора запроса-кандидата на материализацию:

- время выполнения запроса стремится к максимуму;
- частота выполнения запроса стремится к максимуму;
- число ресурсов, затраченных при выполнении запроса, стремится к максимуму;
- частота обновления базовых таблиц (БТ) стремится к минимуму.

Рассчитаем статистику обновления БТ. Для этого из мультимножества *Q* выделим запросы типа *INSERT, UPDATE* и *DELETE:*

$$Q_\Delta = Q_{INS} \bigcup Q_{UPD} \bigcup Q_{DEL},$$

где $Q_{INS}, Q_{UPD}, Q_{DEL}$ – мультимножества запросов вида *INSERT, UPDATE, DELETE* соответственно.

Для каждого запроса q_Δ из мультимножества Q_Δ проанализируем фразу *FROM* и выделим имена таблиц, участвующих в запросе:

$$Q_\Delta = \bigcup_{i=1}^{N_{Q\Delta}} q_{\Delta i} < text, \tau, b, T, Ctype >,$$

где $N_{Q\Delta} = /Q_\Delta/$ – число запросов q_Δ;

$T = \{T_j\}, j = 1, N_T$ – множество таблиц, участвующих в запросе $q_{\Delta i}$;

N_T – количество таких таблиц;

Ctype = {*I, U, D*} – тип операции (*INSERT, UPDATE* или *DELETE).*

На основании каталога таблиц СУБД составим словарь пользовательских таблиц *V*, каждая запись которого будет описана двойкой вида $\{T_v, f_\Delta\}$, где T_v – имя таблицы, f_Δ – число обновлений данной таблицы.

Опишем алгоритм расчета статистики обновления БТ. Для каждого запроса $q_{\Delta i}$, $i=1, N_{Q\Delta}$ и таблицы T_j, $j = 1, N_T$ найдем соответствующую запись в словаре таблиц *V* и инкрементируем значение f_Δ (рис. 1).

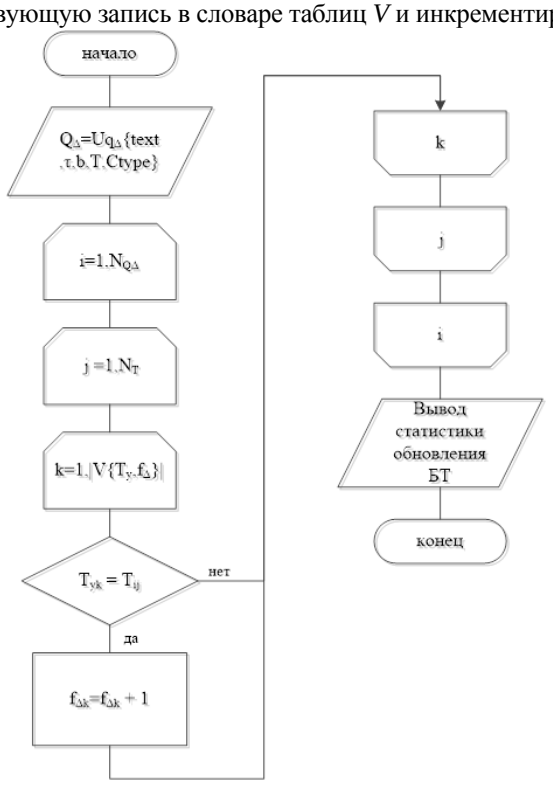

Рис.1. Схема алгоритма расчета статистики обновления БТ

Введем коэффициент *K*, показывающий насколько запрос «пригоден» к материализации. Рассчитаем его с учетом критериев, сформулированных выше.

Шаг 1. Выделение запросов вида *SELECT*

Т.к. МП создаются только для запросов вида *SELECT*, коэффициент *K* имеет смысл рассчитать только для них.

Выделим множество неуникальных запросов типа *SELECT* из мультимножества *Q:*

$$Q_S = \bigcup_{i=1}^{N_{QS}} q_{Si} < text, \tau, b >,$$

где $N_{QS} = /Q_S/$ – число запросов вида *SELECT*.

Шаг 2. Подготовка запросов к анализу.

Для каждого запроса q_{si}, $i=1..N_{QS}$ мультимножества Q_S выполним следующие действия:

- удаление комментариев;
- замена числовых констант на шаблон “@NUMBER”;
- замена строковых констант и дат на шаблон “@LITERAL”;
- замена переменных на шаблон "@BINDING";
- удаление псевдонимов таблиц.

Данный шаг позволяет на первом же этапе идентифицировать одинаковые по синтаксической структуре запросы, отличающиеся лишь значениями констант и переменных, и рассчитать для них общую статистику.

Шаг 3. На основании мультимножества Q_S, содержащего неуникальные запросы, сформируем уникальное множество Q_{US}:

$$Q_{US} = \bigcup_{i=1}^{N_{QUS}} q_{USi} < text, \tau_{US}, c_{UC}, n >,$$

где N_{QUS} – число уникальных запросов q_{US} вида *SELECT*;

n – число вхождений q_{US} запроса в мультимножество Q_S;

τ_{US} – суммарное время выполнения запроса q_{US} за n выполнений;

c_{US} – стоимость выполнения запроса q_{US}. Рассчитывается, как суммарное число блоков данных b_{US}, прочитанных СУБД с диска или буфера при выполнении запроса q_{US} n-раз. Для удобства данное значение будет использовано в процентном соотношении от общего числа потребленных ресурсов:

$$c_{UC} = \frac{\sum_{j=1}^{n} b_{US}}{\sum_{i=1}^{N_{QS}} b_i}.$$

Шаг 4. Для каждого запроса q_{US} проанализируем фразу FROM и выделим имена таблиц, участвующих в запросе:

$$Q_{US} = \bigcup_{i=1}^{N_{QUS}} q_{USi} < text, \tau_{US}, c_{US}, n, T_{US} >,$$

где $T_{US} = \{T_{USj}\}$, $j = 1, N_{TUS}$ – множество таблиц, участвующих в запросе q_{US};

N_{TUS} – количество таких таблиц.

Для каждой таблицы T_{USj} найдем соответствующую ей запись v_{US} в словаре пользовательских таблиц *V:*

$$\{v_{US} < T_v, f_\Delta > | T_v = T_{USj}, v_{US} \in V\}.$$

Для запроса q_{US} выберем максимальное значение f_Δ:

$$F_{\Delta US} = MAX_{j=1}^{N_{TUS}} (f_{\Delta . j}),$$

$$Q_{US} = \bigcup_{i=1}^{N_{QUS}} q_{USi} < text, \tau_{US}, c_{US}, n, T_{US}, F_{\Delta US} >.$$

В дальнейших вычислениях будет использоваться максимальное значение f_Δ, т.к. при обновлении хотя бы одной базовой таблицы МП требуется его инкрементальный пересчет. Таким образом, имеет смысл использовать значение f_Δ, соответствующий наиболее часто обновляемой таблице.

Шаг 5. Рассчитаем коэффициент *K* по следующей формуле:

$$K_{US} = \frac{\tau_{US} c_{US} n}{F_{\Delta US}}. \quad (1)$$

Шаг 6. Приведем запрос q_{US} к результирующему виду.

$$Q_{US} = \bigcup_{i=1}^{N_{QUS}} q_{USi} < text, K_{\Delta US} >.$$

Проанализируем итоговое выражение (1) для коэффициента *K*. В числителе формулы находятся время выполнения запроса τ_{US}, его стоимость c_{US} и число появлений в системе *n*, в знаменателе – частота обновления БТ $F_{\Delta US}$. В общем случае, τ_{US} и c_{US} – прямо пропорциональные величины, поэтому их произведение можно оцениваться как единый показатель.

Рассмотрим различные комбинации значений τ_{US}, c_{US} и *n*, при которых коэффициент *K* будет стремиться к максимуму ($K \rightarrow max$), т.е. соответствующие запросы будут материализованы, при условии, что $F_{\Delta US}$ стремится к минимуму ($F_{\Delta US} \rightarrow min$):

1. $\tau_{US} * c_{US} \rightarrow max$, $n \rightarrow min$.

Данное условие покрывает запросы, которые встречаются редко, но требуют значительного числа ресурсов. К ним относятся, например, различные отчеты, выполняемые в конце отчетного периода (день, неделя, месяц, квартал и т.д.). Использование МП для такой группы запросов эффективно, поскольку оно позволяет существенно снизить нагрузку на СУБД в период сбора статистики и сократить окно обслуживания ИС.

2. $n \rightarrow max$, $\tau_{US} * c_{US} \rightarrow min$.

Запросы, которые выполняются предельно часто, но не потребляют большого числа вычислительных ресурсов, так же попадают в кандидаты на материализацию. Данная группа менее приоритетна с точки зрения создания МП, однако ее материализация позволит в целом повысить производительность работы ИС.

3. $n \rightarrow avg$, $\tau_{US} * c_{US} \rightarrow avg$.

Данное условие покрывает наиболее важную и объемную группу запросов, которые подпадают под критерии материализации. Это запросы со средним значением частоты появления в системе и затрат ресурсов. Именно оптимизация данной группы запросов обычно наиболее существенно влияет на общую производительность ИС.

4. $n \rightarrow max$, $\tau_{US} * c_{US} \rightarrow max$.

При очень больших значениях показателей τ_{US}, c_{US} и *n*, наличие данной группы свидетельствует об ошибке реализации запросов или архитектуры системы. В остальных случаях, запросы этой группы подпадают под предыдущий критерий 3 и являются хорошими кандидатами на материализацию.

Также рассмотрим пограничные значения показателей τ_{US}, c_{US} и *n*, при которых запросы не будет материализованы:

1. $n \rightarrow 0$.

Запрос был выполнен один или несколько раз. Он не встречается с некоторой периодичностью в системе, поэтому не должен быть материализован.

2. $\tau US * cUS \rightarrow 0$

Запрос выполняется предельно быстро и потребляет очень мало ресурсов, поэтому его дальнейшая оптимизация не имеет смысла.

Последнее условие, которое должно быть рассмотрено – $F_{\Delta US} \rightarrow max$. Большие значения $F_{\Delta US}$ снижают ценность запроса с точки зрения материализации, что обусловлено затратами ресурсов на обновление МП. При $F_{\Delta US} \rightarrow max$ запросы материализованы не будут. Это утверждение соответствует одному из сформулированных критериев выбора запросов-кандидатов на материализацию.

Таким образом, были покрыты все сформулированные ранее критерии выбора запросов-кандидатов на материализацию.

Описание эксперимента

Для проведения эксперимента был использован журнал транзакций СУБД, содержащий более 2000 неуникальных запросов. В результате выделения из журнала уникальных запросов было получено 495 записей. Пример распределения коэффициента материализации *K* в зависимости от значений τ, *c* и F_{Δ} представлен на рис. 2.

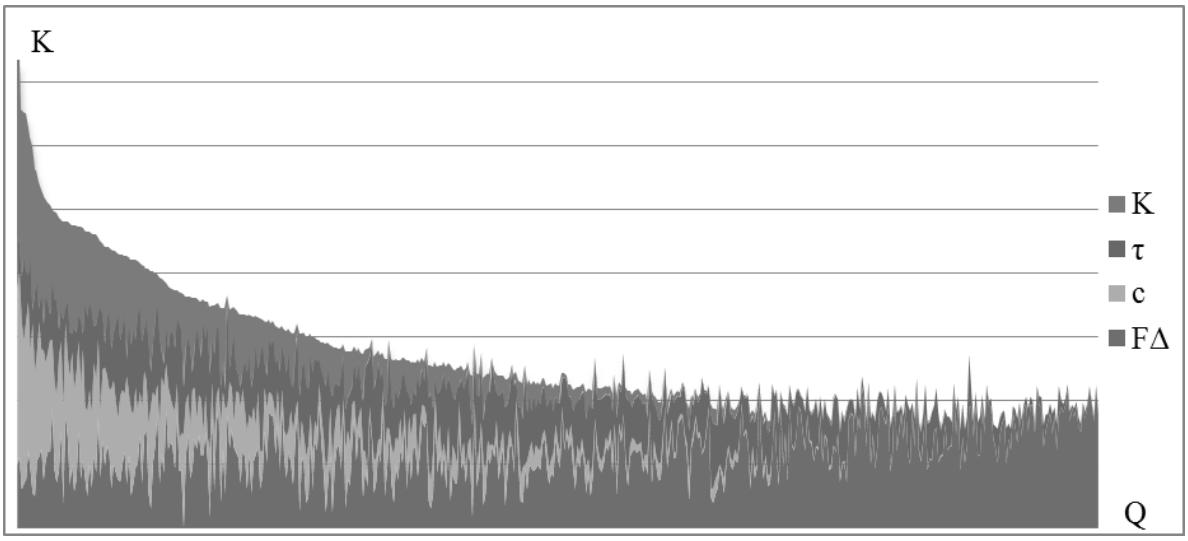

Рис. 2. Пример распределения коэффициента кластеризации для журнала транзакций исследуемой СУБД

Выводы

В данной работе был предложен числовой коэффициент оценки запросов с точки зрения возможности создания МП. Полученный коэффициент покрыл такие важные временные и статистические показатели выполнения запросов, как объем затраченных при формировании результата ресурсов, частоту появление запросов в ИС, а также частоту обновления БТ. Последний показатель позволит отсеять МП, обслуживание которых потребует значительного числа вычислительных ресурсов СУБД. Кроме того, поскольку коэффициент рассчитывается на основании данных журнала транзакций, он может быть использован на этапе поиска запросов-кандидатов на материализацию, когда время выполнения запроса при существующем МП еще не известно.

Список использованной литературы

1. Новохатская Е.А. Методика генерации функций обновления в методе инкрементального обновления материализованных представлений / Е.А. Новохатская, Ю.Н. Возовиков // Вестник СумГУ. Серия «Технические науки». – №3, 2011. – С. 82-96.
2. Новохатская Е.А. Формирование лексем при группировке запросов в методе инкрементального обновления МП / Е.А. Новохатская, А.Б. Кунгурцев // Вестник ЧГТУ. Серия «Технические науки». №1 (71), 2014. – С. 193-199.
3. Кунгурцев А.Б. Анализ возможности применения материализованных представлений в информационных системах / А.Б. Кунгурцев, Куок Винь Нгуен Чан // Праці ОПУ. – 2003. – № 2(20). – С. 102-106.
4. Кунгурцев А.Б. Сравнение запросов в реляционных базах данных для построения материализованных представлений / А. Б. Кунгурцев, Куок Винь Нгуен Чан, Блажко А.А. // Праці УНДІРТ. – Одесса. – 2004. – № (39). – С. 35-38.
5. Кунгурцев А.Б. Поддержка эффективности механизма управления материализованными представлениями / А.Б. Кунгурцев, Ю.Н. Возовиков // Електротехнічні та комп’ютерні системи – 2011. – № 4. – С. 136-140.
6. Кунгурцев А.Б. Технология создания материализованных представлений для реляционных баз данных / А.Б. Кунгурцев, Ю.Н. Возовиков // Праці ОПУ – 2012. – № 2. – С. 170-176.