УДК 621.3.037.3

АКТУАЛИЗАЦИЯ СВЕДЕНИЙ О ДАННЫХ ИНФОРМАЦИОННОЙ СИСТЕМЫ СРЕДСТВАМИ АКТИВНОГО СЛОВАРЯ–СПРАВОЧНИКА ДАННЫХ

А.В. БУДИХИН, Д.А. БУРОВ, А.В. НИКОЛАЕВ, А.В. ОСТРОУХ

Статья представлена доктором технических наук, профессором Шапкиным В.С.

В работе рассматриваются модели информационных систем, в которых процесс актуализации метаданных приложений, баз данных и других компонентов выполняется системой словаря–справочника данных на этапе сопровождения. Исследовано два различных типа систем, получены их временные и вероятностные оценки.

Введение

Любая современная информационная система (ИС) включает в себя в качестве одного из важнейших компонентов базу данных (БД). При этом эксплуатация базы данных предполагает создание резервных копий файлов с информацией. Содержащиеся в базе сведения считаются актуальными на определенные моменты времени. В случае повреждения файлов с данными или необходимости очередной актуализации сведений может быть осуществлено восстановление или приращение информации соответственно. При этом информация, загруженная в базу данных, становится доступной пользователям, является актуальной, а сама база есть единственно верный источник данных.

Метаданные информационной системы включают в себя и метаданные БД, изменение которых производится средствами СУБД. Однако при этом очень часто сведения о других компонентах системы не подвергаются вообще никакой синхронизации либо актуализируются неточно и с опозданием [1]. Для эффективного решения проблемы синхронизации метаданных словаря–справочника данных (ССД) информационной системы с метаданными рабочих станций (локальными банками данных, удаленными серверами, клиентскими рабочими местами и т.п.) необходимо определить устойчивую и эффективную модель. Такая модель основывается на расширенной архитектуре системы словаря–справочника данных, так как она полностью обеспечивает обмен метаданными между всеми компонентами информационной системы.

Теоретический анализ

В ряде исследований, направленных на повышение эффективности использования информационной системы [2, 3], используются, в том числе, различные подходы к построению системы ССД. В зависимости от требований, предъявляемых к ИС, архитектура ССД может упрощаться или усложняться. В рассматриваемых работах основное внимание уделяется взаимодействию с ССД пользователей, получающих метаданные в виде информации о назначении тех или иных элементов системы, разработчиков, администраторов баз данных и других специалистов. Другими словами, подробно рассматриваются различные варианты архитектуры, позволяющие повысить эффективность взаимодействия человека со сведениями о системе [4].

Однако синхронизация метаданных также влияет на эффективность работы пользователей. Ситуация, когда часть пользователей обращается к информации, которой просто не существует для других участников системы, не способствует правильной и эффективной эксплуатации приложений. Между тем, подобное возникает в том случае, когда вопросам актуализации ССД уделяется мало времени либо не уделяется вообще нисколько. Практика показывает, что во

многих организациях, даже в тех, в которых данные, обрабатываемые системой, влияют на жизнь и здоровье человека, процесс обновления и актуализации метаданных ИС может быть отодвинут на задний план. Это, конечно, не означает, что несвоевременное обновление документации, пользовательских приложений и словарей БД непременно приведет к критическим последствиям, но может в значительной степени ухудшить производительность на местах и эффективность системы в целом.

В общем, процесс актуализации метаданных в ССД выглядит следующим образом:

1. Какой–либо участник системы инициирует изменения.

2. Администратор данных вносит изменения в ССД.

3. Разработчики приложений и администраторы базы данных вносят соответствующие изменения на своих участках и фиксируют их в ССД.

4. Подготавливается документация на изменения.

В случае использования пассивного словаря–справочника данных пользователи могут обращаться к ССД с запросами и получать актуальную информацию [5]. Приложения, которые были изменены с учетом внесенных обновлений, также становятся актуальными. На места (в филиалы, территориально удаленные отделения) рассылаются обновления приложений, инструкции по обновлению локальных баз данных. После выполнения актуализации в каждом подразделении можно говорить о завершении синхронизации. При этом в тот период, когда метаданные уже обновлены в ССД (центральном Репозитарии организации), на местах могут продолжать использовать приложения предыдущей версии, которые не учитывают внесенных изменений. Период актуализации ССД в данном случае может затянуться на неопределенный срок. Чем больше рабочих станций в системе, тем выше неопределенность ее состояния. Основная проблема в данном случае заключается в концепции самого словаря–справочника данных, а именно, в его пассивной роли.

Другой вариант – использовать активные ССД, способные не только передавать запрашиваемую информацию, но и непосредственно влиять на нее. В этом случае приложения, базы данных, документация и прочие информационные ресурсы предприятия могут быть обновлены практически одновременно и полностью автоматически. Пользователи узнают об изменениях, произошедших в системе не из сопроводительного письма с указаниями по установке, а непосредственно во время работы, когда интерфейс приложения изменится (например, появится новое поле).

Существуют два принципиально разных способа активной актуализации метаданных:

1. ССД самостоятельно обновляет все рабочие станции;

2. Каждая рабочая станция запрашивает ССД о состоянии метаданных и обновляет их в случае необходимости.

В идеальном случае, при первом подходе все рабочие станции всегда доступны (линии связи исправны, системы находятся в рабочем состоянии), при втором, – ССД в состоянии удовлетворить запросы сразу всех рабочих станций. И в том, и в другом случае предполагается, что система ССД доступна и находится в рабочем состоянии.

Для проведения исследований необходимо приблизить модель к реальной системе. Понятно, что чем больше рабочих станций может обслужить ССД второй модели (пропускная способность), тем выше ее эффективность, так же как и выше эффективность в случае минимального количества недоступных рабочих станций (первый случай). В проводимом исследовании выбрана ситуация, при которой максимально допустимое количество недоступных рабочих станций и пропускная способность ССД установлена в половину от имеющегося количества рабочих станций. Данное значение можно считать оптимальным для исследования системы.

В рассматриваемой модели за рамки выводится процесс обновления ССД администраторами данных. То есть предполагается, что данные в базе метаданных на момент обращения к ней рабочих станций, а также в случае обращения к рабочим станциям ССД актуальны и готовы к рассылке.

С учетом всего вышесказанного рассмотрим моделируемые системы. Систему, в которой ССД обращается к рабочим станциям, обозначим S_1, а систему, в которой рабочие станции обращаются к ССД – S_2.

Пусть имеется n рабочих станций W в системе S_1. Пусть некоторая рабочая станция W_i доступна и активна в момент времени t_k. Тогда ССД – C – отправив запрос и убедившись в доступности станции, пошлет блок информации с метаданными Q_l и получит подтверждение Z_l о фиксации изменений. Выполнение описанных действий означает актуализацию сведений ССД на рабочей станции W_i. Примем за условную временную единицу T_i время, реально необходимое для выполнения данной операции на станции i. Будем считать, что на каждой станции время, требуемое для фиксации изменений в системе, не зависит от удаленности рабочей станции от ССД. Поэтому условную временную единицу обозначим T.

Пусть теперь некоторая W_k недоступна. В этом случае предполагается, что сбой носит временный, устранимый характер. Он может быть вызван ошибкой протокола передачи данных, либо загруженностью рабочей станции, либо техническими мероприятиями, проводимыми на рабочей станции. Учитывая производительность современных аппаратных комплексов и предполагаемые объемы метаданных, T может принимать значение, равное как нескольким секундам, так и нескольким минутам. Принимая во внимание встречающиеся на практике временные сроки, требующиеся для устранения описанных и подобных сбоев, время, требующееся для восстановления работоспособности станции, будем выбирать из интервала $\{5T…20T\}$. Значение, меньшее $5T$, представляется незначительным ввиду близости его к T, а значение, большее $20T$, нецелесообразным, так как является тормозящим работу системы и возникает в отдельных случаях критических сбоев и внештатных ситуаций, которые выходят за рамки модели. После устранения сбоя необходимо выполнить все перечисленные действия по актуализации, то есть затратить время T.

В случае актуализации рабочих станций в системе S_2 рассуждения те же, за исключением того, что недоступной может быть сама система ССД. Так как она ограничена в ресурсах, то есть может обслуживать одновременно половину имеющихся рабочих станций. В этом случае Q_l – запрос, направляемый от ССД к рабочей станции, становится в очередь и ожидает, когда ССД сможет его обработать. Время на обработку l–го запроса также выбирается равным T. При этом также возможен вариант описанного для системы S_1 сбоя. Так как интервал в этом случае выбирается такой же, как и в системе S_1, а выбор любого значения из интервала является равновероятным, можно говорить об адекватности данной модели. На практике возникновение любого из описанных сбоев также является случайным событием, а вероятность возникновения того или иного события практически одинакова.

Требуется определить зависимость времени обработки запросов по актуализации системы от количества рабочих станций, а также влияние количества рабочих станций на вероятность возникновения сбоя в системе.

Методика и проведение исследования

Для проведения эксперимента воспользуемся имитационным моделированием. Для моделирования случайных величин будем использовать генератор псевдослучайных чисел ЭВМ с равномерным законом распределения. Обозначим полученную случайную величину отказа рабочих станций – X с математическим ожиданием M_X и дисперсией D_X. Понятно, что оба этих параметра неизвестны. Аналогично Y – время выполнения актуализации, а M_Y и D_Y – математическое ожидание и дисперсия случайной величины Y. Для получения более точных результатов выполним ряд экспериментов N. При исследовании модели N положим равным одному миллиону итераций. В результате будет получена совокупность N численных результатов. Среднее арифметическое полученных значений нельзя рассматривать в качестве математического ожидания величины X или Y. Для получения несмещенной оценки проведем еще серию из N экспе-

риментов. Полученный числовой ряд случайных величин содержит средние значения для каждой серии из N экспериментов. В качестве математического ожидания будем рассматривать среднее этих значений. В качестве оценки дисперсии будем использовать исправленную статистическую дисперсию.

В качестве параметров при моделировании могут выступать фиксированные значения пропускной способности для S_2 (максимально допустимое значение отказов рабочих станций для S_1), которые мы установили в половину имеющихся в системе рабочих станций, то есть пятьдесят процентов всех рабочих станций.

В качестве варьируемых параметров возьмем количество рабочих станций в системе.

Моделирование будет производиться с помощью программы, имитирующей действия, происходящие в системе, написанной на языке C++.

Описание результатов и выводы

Как было сказано выше, для уменьшения среднего квадратического отклонения полученных результатов исследование проводилось при значительном количестве итераций $N = 1000{,}000$.

Полученные результаты приведены в сводной таблице.

Таблица

Результаты экспериментов

		S_1				S_2			
№ п.п.	Кол–во рабочих станций	P сбоя, %	E сбоя	T mod	E mod	P сбоя, %	E сбоя	T mod	E mod
1	10	19,96	0,05	35,87	0,09	26,91	0,07	48,41	0,13
2	15	23,30	0,08	59,36	0,26	22,65	0,09	62,72	0,23
3	30	23,31	0,09	116,57	0,90	24,39	0,05	131,50	0,58
4	45	23,41	0,06	174,91	1,67	24,78	0,06	198,38	1,38
5	90	24,39	0,07	357,23	4,02	22,13	0,08	357,69	4,22
6	135	24,69	0,06	539,34	4,31	20,40	0,08	498,02	4,43
7	270	24,66	0,08	1072,82	6,90	18,69	0,11	909,89	4,48
8	405	24,63	0,09	1605,30	7,24	17,68	0,07	1296,22	7,12
9	810	24,57	0,11	3186,33	20,22	15,49	0,08	2303,45	8,36
10	1215	24,20	0,10	4668,13	33,20	14,40	0,05	3243,89	10,62
11	2430	24,23	0,09	9285,88	76,94	13,01	0,07	5936,13	58,28

В таблице вероятность возникновения сбоя в системе при выполнении актуализации обозначена как *P сбоя*. В системе S_1 эта величина означает процент рабочих станций, который был недоступен для ССД во время обновления, но обновление было выполнено после некоторого времени ожидания. Время ожидания включается в общее время моделирования (время выполнения) *T mod*. Кроме него сюда входит и время, требующееся на выполнение операции актуализации. *E сбоя* и *E mod* – соответствующие ошибки полученных результатов. В системе S_2 величина *P сбоя* означает вероятность возникновения ситуации, при которой к ССД отправляется запросов больше, чем может быть обработано.

Графики зависимости времени обработки запросов по актуализации системы от количества рабочих станций представлены на рис. 1:

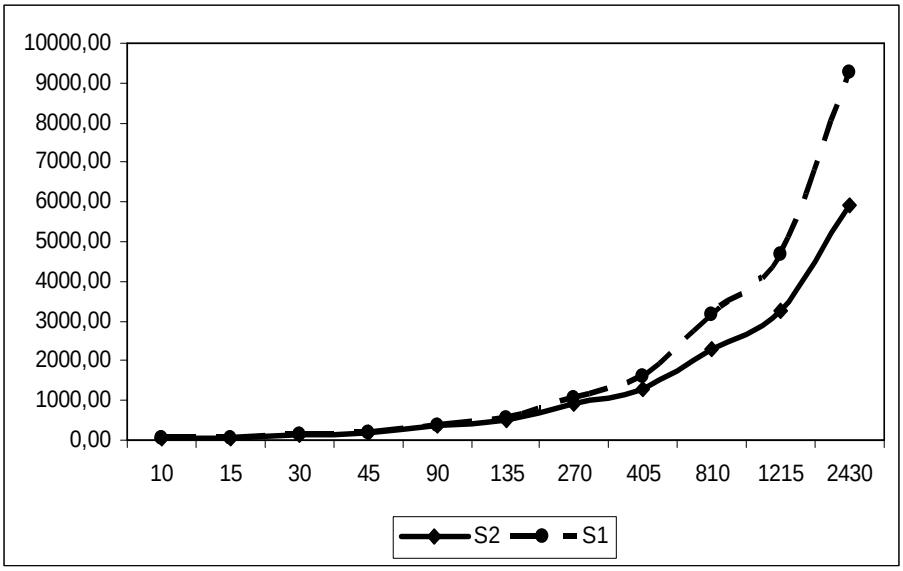

Рис. 1. Зависимость времени обработки запросов от количества рабочих станций в системе

На рис. 2 показаны зависимости вероятности сбоя от количества рабочих станций в системе.

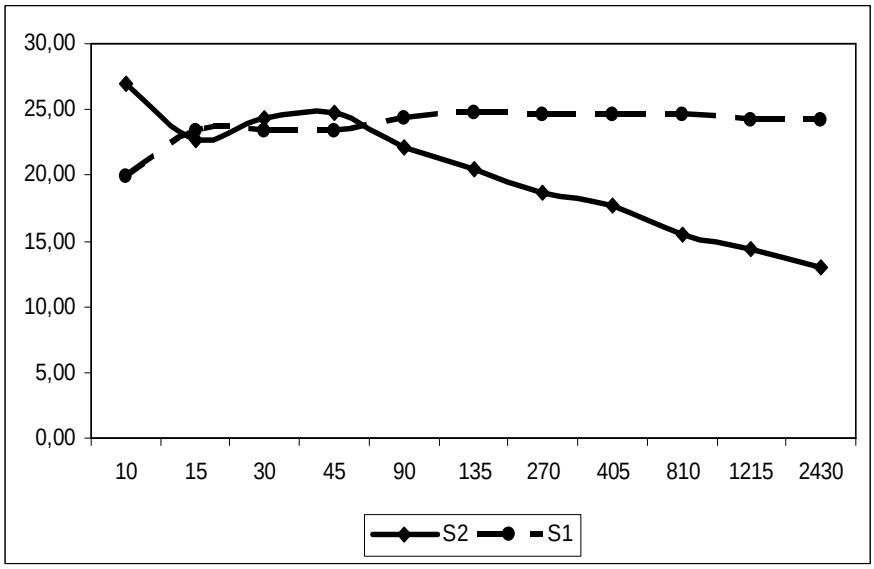

Рис. 2. Влияние количества рабочих станций на вероятность возникновения сбоя в системе

Приведенные в таблице результаты позволяют сделать следующие выводы:

1. В системе, в которой словарь–справочник данных инициирует обновления и обращается к каждой рабочей станции (S_1), вероятность возникновения сбоя при незначительном увеличении количества рабочих станций возрастает. Дальнейшее увеличение количества рабочих станций не приводит к существенному изменению вероятности, значение которой стабилизируется и стремится к половине максимально возможного (определенного) числа сбоев.

2. В системе S_1 с увеличением количества рабочих станций время выполнения нелинейно возрастает. В рассматриваемой модели системы при количестве рабочих станций, большем 150, динамика времени выполнения резко меняется.

3. В системе, в которой рабочие станции обращаются к ССД (S_2), при уменьшении числа рабочих станций вероятность сбоя повышается, вследствие того, что растет вероятность одновременного обращения рабочих станций к ССД. Между тем это значение не может увеличиваться до своих теоретических максимумов, так как аппаратные возможности ССД должны обеспечивать одновременное подключение значительного для системы количества рабочих станций. При этом с ростом числа рабочих станций пропускная способность ССД не должна понижаться меньше половины числа возможных подключений, так как это может привести к значительным простоям и сбоям. В результате проведенного исследования было установлено, что с ростом количества рабочих станций вероятность сбоя уменьшается (рис. 2). Это объясняется тем, что падает вероятность одновременного подключения такого значительно числа станций к ССД.

4. В системе S_2 с увеличением числа рабочих станций время выполнения возрастает меньше, чем при аналогичных значениях количества рабочих станций для первой системы.

5. При количестве рабочих станций $W < 45$ эффективнее система S_1. Эффективность систем S_1 и S_2 можно считать одинаковыми в диапазоне $50 < W < 100$. При этом вероятность возникновения сбоев меньше в системе S_2. При дальнейшем увеличении W эффективнее оказывается система S_2.

Заключение

Хотя в рассмотренных моделях система, в которой рабочие станции обращаются к ССД, с увеличением их количества, оказалась более эффективной чем первая, полученные временные оценки могут считаться чересчур высокими для выполнения операции актуализации и сопровождения информационной системы. Хорошие результаты показаны при значениях $50 < W < 100$. Поскольку на практике ограничить количество рабочих станций не представляется возможным, хорошим решением данной проблемы может являться применение параллельных систем ССД.

Рассмотренная задача может быть исследована более глубоко за счет усложнения моделей с целью получения более подробной оценки системы.

ЛИТЕРАТУРА

1. Кудинов А. Хранилище данных как основа корпоративной интеграции // PCWeek/RE. 2006. № 20.

2. Горохов Д., Чернов В. Методы организации хранения данных в СУБД //Открытые системы. 2003. №3.

3. Локшин М.В., Кравец О.Я. Разработка методов распараллеливания запросов в гетерогенных РСУБД // Системы управления и информационные технологии. 2005. №4(21). С. 71 – 76.

4. Тэнненбаум Э. Решения для метаданных // Addison Wesley. 2002

5. Marco D. Meta Data Repositories: Where We've Been and Where We're Going // DM Review. 2002. № 2.

METADATA ACTUALIZATION OF INFORMATION SYSTEM BASED ON ACTIVE DATA DICTIONARY SYSTEM FACILITIES

Budikhin A.V., Burov D.A., Nikolaev A.V., Ostroukh A.V.

This article presents Information System models in which Data Dictionary System initiate program, database and other system components metadata actualization at an exploitation phase. Two different models of Information Systems are analyzed in the article. Probabilistic and time estimates are received.

Сведения об авторах

Будихин Анатолий Владимирович, 1955 г.р., окончил МВТУ им.Баумана (1978), кандидат технических наук, доцент кафедры автоматизированных систем управления МАДИ (ГТУ), автор более 70 научных работ, область научных интересов – автоматизация производственных процессов (промышленность), проектирование баз данных, словарей–справочников данных и информационных систем.

Буров Дмитрий Анатольевич, 1967 г.р., окончил МАДИ (1989), старший преподаватель кафедры корпоративных информационных систем Российского нового университета, автор 9 научных работ, область научных интересов – консалтинг, проектирование баз данных и корпоративных информационных систем, разработка прикладных интерфейсов информационных систем.

Николаев Александр Владимирович, 1982 г.р., окончил МАДИ (ГТУ) (2003), аспирант кафедры автоматизированных систем управления МАДИ (ГТУ), автор 3 научных работ, область научных интересов – автоматизация процессов и производств на промышленных предприятиях, проектирование баз данных и информационных систем, разработка прикладных интерфейсов информационных систем.

Остроух Андрей Владимирович, 1975 г.р., окончил МАДИ (1996), кандидат технических наук, доцент кафедры автоматизированных систем управления МАДИ (ГТУ), докторант, автор более 50 научных работ, область научных интересов – автоматизация управления строительными предприятиями, проектирование баз данных и информационных систем, разработка цифровых образовательных ресурсов.