УДК 004.62; 004.63; 004.65; 004.72; 004.77

ОБЕСПЕЧЕНИЕ ЭКОНОМИЧНОСТИ, ХАРАКТЕРИЗУЕМОЙ ЗАТРАТАМИ ВЫЧИСЛИТЕЛЬНЫХ РЕСУРСОВ

Гаврилова М. В., Зарипова Е. Ф., Ощепков Д. Е.

Закрытое акционерное общество «Международный Деловой Консалтинг», Москва, Россия (107140, г. Москва, ул. Верхняя Красносельская, д.11А, стр.3), e-mail: a.zhiganov@delcon.ru

Одной из наиболее важных задач при осуществлении двунаправленного обмена данными является обеспечение экономичности процесса синхронизации – задача оптимизации процесса обновления и создания записей в целевых таблицах баз данных. Для оптимизации этих процессов необходимо выполнить анализ взаимодействия баз данных при поиске актуальных данных. Наиболее оптимальным вариантом является создание реляционных правил на стороне модуля двухстороннего обмена, который будет осуществлять автономное формирование базы HASH записей для каждого элемента синхронизируемых БД. Это обеспечит экономию вычислительных ресурсов, так как не будет требоваться передача всех данных на транспортном уровне, а передаваться будет только HASH сумма. И только при идентификации появления новой информации будет передано содержимое данных, записанное в синхронизируемой записи. Аналогичная экономия вычислительных ресурсов возникает и при записи в целевые базы данных актуальных данных, поскольку обновление неактуализируемого контента происходить не будет.

Ключевые слова: двунаправленный обмен данными, базы данных, вычислительные ресурсы.

ECONOMICAL PROVISION, CHARACTERIZED BY THE COMPUTATIONAL BURDEN

Gavrilova M. V., Zaripova E. F., Oschepkov D. E.

"International Business Consulting" Joint-Stock Company, Moscow, Russia (107140, Moscow, V.Krasnoselskaya street, 11A-3), e-mail: a.zhiganov@delcon.ru

One of the most important tasks in the implementation of a bi-directional data exchange is to ensure the efficiency of the synchronization process - the problem of optimizing the process of creating and updating records in the target database tables. To optimize these processes is necessary to analyze the interaction databases when searching for relevant data. The best option is to create a relational rules on bilateral exchange of the module, which will carry out independent building a base of HASH records for each element of the synchronized database. This will provide savings in computational resources, as it will not require the transfer of data at the transport level, and will only be transmitted HASH programming. Only the identification of new information will be transferred content data recorded in the synchronized record. Similar saving computing resources arises when writing to the target database of actual data because the update is not continuously updated content will not occur.

Keywords: bi-directional data exchange, databases, computing resources.

Одной из наиболее важных задач при осуществлении двунаправленного обмена данными является обеспечение экономичности процесса синхронизации – задача оптимизации процесса обновления и создания записей в целевых таблицах БД, поскольку именно эти операции вызывают наибольшие ресурсные затраты. Для оптимизации этих процессов необходимо выполнить анализ взаимодействия баз данных при поиске актуальных данных.

Довольно часто для реализации практических задач требуется обеспечить возможность модифицирования узкоспециализированных агрегационных интернет-систем, без затрат на разработку однотипных функций для каждой из них и обеспечить более эффективное управление данными, что приведет к повышению производительности труда.

Следовательно, добавление полей синхронизации и временных меток в базы данных пользователей недопустимы.

В таком случае требуется выполнить анализ хранилища данных и его перенос, выполняя подходы и методы двунаправленного обмена данными.

Сравнение или сопоставление структур исходных и целевых хранилищ данных выполняется в рамках анализа хранилищ данных. Под задачей определения семантических соответствий между элементами схем принимается сопоставление схем. На сопоставление схем накладываются ограничения на варианты решения, так как она считается AI-полной задачей. В [3] показывается разделение подходов к сопоставлению схем. Синонимом для сопоставления схем в русском языке является «сравнение схем».

Точность результата и трудоемкость считаются наиболее важными характеристиками методов и подходов к сопоставлению схем. Большое внимание уделяют трудоемкостям предлагаемых алгоритмов, так как AI-полнота сопоставления схем обуславливает необходимость корректировки результатов сопоставления человеком, а большие размеры схем увеличивают время выполнения сопоставления. Например, при использовании сопоставления графов следует учитывать, что, как правило, задача сопоставления графов не отнесена ни к классу P, ни к классу NP-полных задач, а задача изоморфизма подграфу (один из типов сопоставления графов) является NP-полной. Именно поэтому, если возникает необходимость сопоставления схем, то выбор метода для его выполнения нужно производить, руководствуясь допустимыми соотношениями точности и трудоемкости.

Использование разных моделей данных может сильно осложнить процесс двунаправленного обмена данными между хранилищами данных разных типов из-за сопоставления структур. В [2] говорится о двунаправленном обмене реляционных баз данных в объектно-ориентированные или XML базы данных.

Вся сложность состоит в том, что правил сравнения элементов различных моделей данных в готовом и однозначном виде нет.

Есть два подхода к сопоставлению элементов разных моделей данных:

а) определение правил сравнения для каждой возможной пары моделей данных;

б) использование промежуточного концептуального представления, в котором будут представлены исходные и целевые структуры данных.

Так, в качестве промежуточного представления предлагаются:

а) граф;

б) семантические модели;

в) XML файлы.

У такого подхода есть и свои недостатки, например, возможная потеря семантики при переходе к промежуточному представлению, а также увеличение трудоемкости. Процесс двунаправленного обмена данными между хранилищами данных одного типа не сталкивается с вышеописанными проблемами, но при этом ограничивается область её применения. При выполнении сравнения структур данных существуют сложности при двунаправленном обмене данными между хранилищами данных одного типа. Они связаны с различиями конкретных реализаций хранилищ данных. В реляционных СУБД могут использоваться разные стандарты языка SQL, а, например, в одном хранилище данных длина данных типа «строка» ограничена, в другом – нет. Различные представления объектов моделируемого мира и связей между ними также осложняют сравнение структур данных исходных и целевых хранилищ данных. Говоря терминами ER-модели Чена, которые описываются в [1], одно и то же явление моделируемого мира может быть представлено как атрибут, как множество сущностей и как множество связей. Следовательно, появляется проблема сопоставления объектов и их связей в силу возможности использования различных их представлений.

К этой проблеме также относится использование эквивалентных конструкций моделей данных и разных точек зрения при представлении объектов моделируемого мира, таких как:

а) использование разных имен (синонимов, гиперонимов);

б) ограничение целостности;

в) ограничение типов данных.

С решением некоторых вопросов связано и осуществление переноса. Одним из этих вопросов является и определение порядка переноса данных. Такое решение находится в зависимости от наличия связей между объектами моделируемого мира, представленными в хранилище данных. Сложностей не возникнет, если в хранилище данных представлен лишь один тип объектов моделируемого мира без связей. Также не возникнет проблем, если в хранилище будет представлено несколько типов объектов, но без связей. В противном случае, если будет представлено несколько типов объектов связанных между собой, могут возникнуть проблемы с осуществлением переноса. В [4] при рассмотрении процесса двунаправленного обмена в реляционных базах данных, был предложен подход, который представляет собой комбинацию семантики и одного или нескольких атомарных значений. Выделяются два типа фактов – независимые и зависимые. Процесс двунаправленного обмена фактов состоит из двух подпроцессов: переноса независимых фактов и переноса зависимых фактов. Схожие идеи заложены и в другие методы и подходы, например, в методы и системы для двунаправленного обмена данными, описанные в статьях [5]. Однако предложенные

методы не решают проблему возникновения циклических связей между объектами разного или одного типов.

Определение значений уникальных идентификаторов объектов моделируемой области также является важным для переноса данных. Если перенос производится в непустое хранилище данных, то возможны коллизии уникальных идентификаторов, а при переносе в хранилище данных другого типа или другой реализации или в хранилище данных со структурой, отличающейся от структуры исходного хранилища, могут возникнуть проблемы, такие как невозможность преобразования одного типа данных уникальных идентификаторов в другой и т.д. Также усложняет перенос данных то, что вариант выхода из проблемной ситуации в виде изменения значений уникальных идентификаторов влечет за собой необходимость изменения значений этих уникальных идентификаторов в связях между объектами моделируемого мира.

При двунаправленном обмене данными необходимо выполнять их преобразование. Сложность допустимых преобразований в конкретных методах и подходах определяет их применимость для выполнения определенного проекта по обмену данными. В [4] предполагается, что в исходных данных нет ненужной информации, и данные корректны, и не рассматривают не выражающие семантики ключи и сложные трансформации данных. Генерация сценариев для предопределенных шаблонов преобразования предусмотрены в [5] для реляционных баз данных.

Исходя из вышеизложенного, оптимальным вариантом является создание реляционных правил на стороне модуля двухстороннего обмена, который будет осуществлять автономное формирование базы HASH записей для каждого элемента синхронизируемых БД. Это обеспечит экономию вычислительных ресурсов, так как не будет требоваться передача всех данных на транспортном уровне, а передаваться будет только HASH сумма. И только при идентификации появления новой информации будет передано содержимое данных, записанное в синхронизируемой записи (рисунок 1).

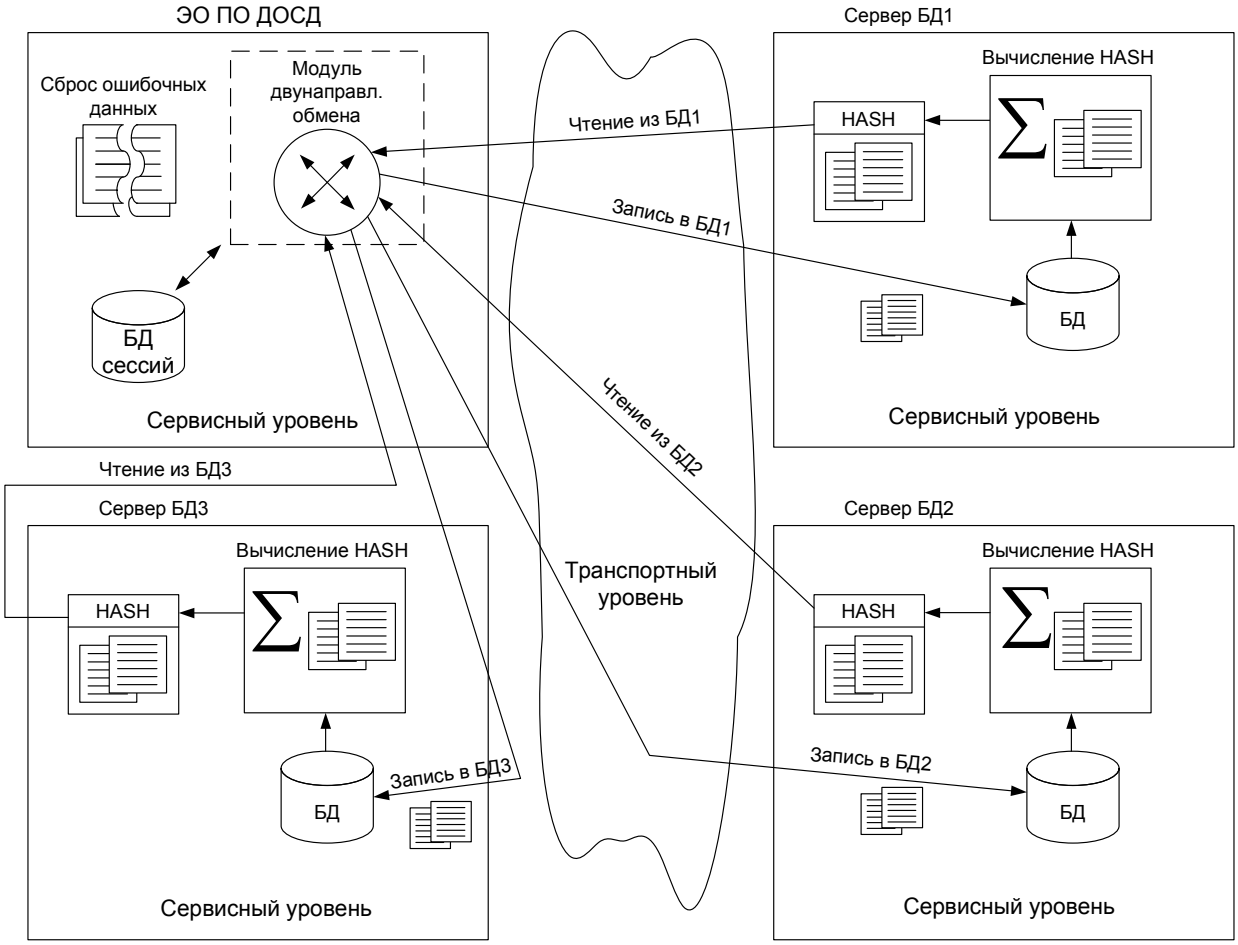

Рисунок 1. Двунаправленный обмен данными с использованием HASH

Аналогичная экономия вычислительных ресурсов возникает и при записи в целевые базы данных актуальных данных, поскольку обновление не актуализируемого контента происходить не будет.

В свою очередь, наличие автономной базы HASH связок в совокупности с временным штампом обеспечат обслуживание более 2-х нитей двунаправленного обмена без потери качества обслуживания. При этом количество нитей будет определяться техническими характеристиками вычислительного комплекса.

Таким образом, будет обеспечиваться экономичность, характеризуемая экономией затрат вычислительных ресурсов.

Описанные решения реализованы в экспериментальном образце программного обеспечения, реализующего двунаправленный обмен структурированными данными, между несвязными интернет-ресурсами. Данная работа выполняется при финансовой поддержке Минобрнауки в рамках государственного контракта № 14.514.11.4009.

Список литературы

1. Чен П. Модель «сущность-связь» – шаг к единому представлению о данных // СУБД. – 1995. – № 3. – С. 137–158.

2. Alam M. Migration from relational database into object oriented database / M. Alam, S. Wasan // Journal of Computer Science. – 2006. – Vol. 2, Issue 10. – P. 781–784.

3. Anan M. Managing uncertainty in schema matcher ensembles / M. Anan, A. Gal // Scalable Uncertainty Management First International Conference (SUM 2007). Washington, DC, USA, October 10–12, 2007. Proceedings. – Springer Berlin / Heidelberg, 2007. – P. 60–73.

4. Hudicka J. DataMIG – a data migration management tool suite [Electronic resource] // Dulcian, Inc. – Electronic data. – URL: http://www.dulcian.com/Articles/DataMIG.htm (reference date: 05.09.2010).

5. Maatuk A. Relational database migration: a perspective / A. Maatuk., A. Ali, N. Rossiter // Database and Experts Applications. 19th International Conference (DEXA 2008). Turin, Italy, September 1-5, 2008. Proceedings. – Springer Berlin / Heidelberg, 2008. – P. 676–683.

Рецензенты:

Мисюрин Сергей Юрьевич, д.ф.-м.н., заведующий лабораторией, Институт машиноведения им. А. А. Благонравова РАН, г. Москва.

Сахвадзе Георгий Жорович, д.т.н., в.н.с., Институт машиноведения им. А. А. Благонравова РАН, г. Москва.