АВТОМАТИЧЕСКАЯ РЕПЛИКАЦИЯ ДАННЫХ В КОРПОРАТИВНОЙ ИНФОРМАЦИОННОЙ СРЕДЕ

Шахгельдян К.И.

Введение

Две основные проблемы, с которыми сталкиваются разработчики корпоративной информационной среды (КИС) крупного предприятия на современном этапе, связаны, во-первых, со сложностью объекта автоматизации, во-вторых, с частыми изменениями объекта автоматизации и необходимостью адаптации к этим изменениям в сроки, определяемые временными регламентами функционирования объекта.

Сложность объекта автоматизации и его постоянные изменения приводят к необходимости решать задачи сопровождения и эксплуатации на стадии проектирования и разработки КИС. Задачи эксплуатации и сопровождения требуют в среднем 80% временных и финансовых ресурсов, потраченных на создание и поддержание жизнеспособных информационных систем [1]. Сложность эксплуатации КИС объясняется во многом взаимозависимостью между частями КИС и тем, что изменения в одной части ведут к изменениям в большом числе связанных объектов (данных, сервисов, систем, серверов и т.п.) КИС.

В связи с наличием множества серверов СУБД в КИС приходится решать проблемы интеграции данных. Среди этих проблем, прежде всего, возникает проблема репликации данных. В КИС вуза число репликаций доходит до нескольких сотен. При этом внутри каждой репликации число реплицируемых объектов базы данных в среднем составляет около десятка. Таким образом, в рамках гетерогенной КИС с несколькими филиалами, в которых функционируют собственные серверы баз данных, число реплицируемых объектов достигает нескольких тысяч. В результате возникают две основные проблемы:

1. каким образом отслеживать все требуемые к репликации объекты (изменения в реальном мире происходят постоянно и соответствующие объекты меняются постоянно);

2. каким образом обеспечить синхронизацию реплицируемых объектов (взаимозависимость тысяч объектов отслеживать вручную при условии их постоянных изменений требует слишком больших затрат).

Для организации репликаций данных в основном используются встроенные средства СУБД. Но для КИС вузов характерна гетерогенность, что затрудняет использование встроенных средств репликации. В тех же случаях, когда встроенные средства могут применяться, они не позволяют в полной мере решить проблемы синхронизации объектов, а так же не решают проблем обеспечения качества данных, связанных с нарушением целостности данных.

Стратегия решения проблемы репликации в гетерогенной КИС состоит в обеспечении автоматического выполнения всех процедур, которые могут быть выполнены автоматически.

Управление репликациями, если речь идет о сложной цепочке взаимозависимых объектов, возможно при наличии высокоуровневого формализованного описания. Для формализованного описания в работе [2] предлагается использовать объектно-ориентированный подход. В работах [3, 4] рассматривается описание репликаций между одинаковыми репликами на базе метаописания. Эти решения делают репликации более наглядными (и, соответственно, более управляемыми) для администраторов КИС, но не позволяют их автоматизировать и синхронизировать.

В рамках построения жизнеспособной (а, значит, и адаптивной [1]) КИС ставится вопрос об автоматизации выполнения репликаций. В КИС стоит проблема, как организовать управление репликациями и выполнение репликаций в гетерогенной среде наиболее эффективным образом, с учетом согласования репликаций между собой и между различными объектами КИС.

Онтологический подход

Использование онтологического подхода позволяет автоматизировать организацию репликаций в КИС на базе аксиом и утверждений. В работе [5] автором рассматривались отношения проекции, которые позволяют установить соответствия между понятиями предметной области, ИТ-области и области управления бизнес-процессами и экземплярами понятия – источник данных. Остановимся здесь подробнее на этих и других отношениях, которые позволяют определить алгоритм автоматической репликации данных в гетерогенной КИС.

Наследование как отношение между понятиями X и Y : $P'(X,Y)$ определено в случае, если понятие Y имеет все те же атрибуты, что и понятие X, но на некоторые из которых могут быть наложены более жесткие ограничения, чем на атрибуты понятия X, при этом у Y так же могут быть дополнительные атрибуты. В этом случае понятие Y может наследоваться от X. Ограничение представляют собой понятие условие.

В случае если в производном классе Y добавлены атрибуты по сравнению с понятием X, то определено представление $|Y|_X$, которое позволяет выделять в экземпляре понятия Y часть, которая определяется только семантикой X. Виртуальные атрибуты, т.е. атрибуты, которые являются результатом регулярного выражения, не входят в семантику и два понятия считаются эквивалентными, если их различают только виртуальные атрибуты: $|Y|_X = Y$.

Для отношений наследования определена аксиома

Аксиома 1.

$$\forall X,Y : P'(X,Y) \Rightarrow \overline{Y} \subseteq \overline{X}, \qquad (1)$$

где $\overline{Y}, \overline{X}$ - множества экземпляров понятий соответственно Y, X.

В общем случае отношения проекции определяют связи между понятием X и совокупностью источников данных A : $P(X,A)$. Под совокупностью источников данных здесь понимается набор источников

данных, в которых хранятся экземпляры понятия. Отношения проекции имеет атрибутами тип проекции – на чтение или на чтение/запись и свойства соответствия атрибутов. Свойство соответствия атрибутов отношения проекции представляет собой соответствие между атрибутами понятия X и атрибутами совокупности источников A.

Отношение проекции между одним понятием и разными источниками допускает определения соответствия не со всеми атрибутами понятия. Часть атрибутов в отдельных источниках может не использоваться, и, кроме того, разрешено понятию иметь виртуальные атрибуты, которые не имеют проекции на область ИТ.

Для простоты описания далее совокупность источников данных, с которым у понятия установлены отношения проекции, называется просто источником данных.

При наличии одного источника данных понятие имеет отношения проекции с одним экземпляром понятия источника данных. В КИС вуза источников одних и тех же данных в общем случае больше одного, т.е. могут иметь место многочисленные реплики. В этом случае определяются отношения проекции с несколькими источниками данных.

Определим

- отношение проекции с правом на чтение/запись как $P_1(X,A): P'(P(X,A),P_1(X,A))$ – источник является контейнером экземпляров понятия, при этом разрешено как чтение, так и запись экземпляров, т.е. $P \to type = \mathrm{Re}ad/Write$;
- отношение проекции с правом на чтение $P_2(X,A): P'(P(X,A),P_2(X,A))$ – источник является контейнером экземпляров понятия с правом только на чтение, т.е. $P \to type = \mathrm{Re}adOnly$;.

Все атрибуты отношений, т.е. соответствие между атрибутами понятия и источника данных, относятся к базовому отношению $P(X,A)$.

Аксиома 2.

Если экземпляры какого-то понятия создаются и хранятся в источнике данных, то в иерархии понятий должна существовать проекция на чтения/запись, которая описывает внесение экземпляров в источник. Если все экземпляры понятия X вносятся в источник данных A, то должны быть определены отношения проекции: $X \vee (\forall Y : P'(X,Y)) \vee (\exists Z : P'(Z,X)) \Rightarrow P_1(X,A) \vee P_1(Y,A) \vee P_1(Z,A)$.

Сформулируем утверждения для отношений проекции:

Утверждение 1.

Если определены отношения проекции на чтение/запись, то определены и отношения проекции с тем же источником на чтение: $\forall X, A : P_1(X,A) \Rightarrow P_2(X,A)$.

Здесь и далее доказательства утверждений просты и в силу ограниченности места не приводятся.

Утверждение 2.

В иерархии понятия могут не существовать отношения проекции с правом на чтение/запись, но при этом для всех экземпляров понятий определены отношения на чтения/запись с одним или более источниками данных: $X,Y : P'(X,Y) \Rightarrow \forall A : \neg P(Y,A) \wedge \neg P(X,A) \wedge \overline{X},\overline{Y} \rightarrow \{A_1,\ldots,A_m\}$, где $\{A_1,\ldots,A_m\}$ - источники данных, в которых хранятся экземпляры $\overline{X}$.

Утверждение 3.

Если базовое понятие имеет отношения проекции с некоторым источником, то и производное понятие, для которого верно условие $|Y|_X = Y$, имеет то же отношение проекции с тем же источником данных. $\forall X,Y,\exists A : P'(X,Y) \wedge |Y|_X = Y \wedge P(X,A) \Rightarrow P(Y,A)$, т.е. отношение проекции для случая $|Y|_X = Y$ наследуется.

Утверждение 4.

Если производное понятие по семантике не совпадает с базовым, т.е. $P'(X,Y) \wedge |Y|_X \neq Y$, то для базового понятия отсутствуют отношения проекции чтения/запись: $\forall X,Y : P'(X,Y) \wedge |Y|_X \neq Y \Rightarrow \forall A : \neg P_1(X,A)$.

Утверждение 5.

Если понятие X имеет производное понятие Y : $P'(X,Y)$ и множество экземпляров этих понятий полностью совпадает $\overline{X} = \overline{Y}$, то $\forall A : P(X,A) \Leftrightarrow P(Y,A)$. В более общем случае, если понятие X имеет производные Y,Z : $P'(X,Y), P'(X,Z)$ и экземпляры производных понятий составляют полное множество экземпляров базового понятия: $\overline{X} = \overline{Y} \cup \overline{Z}$, то $\forall A : P(Y,A) \wedge P(Z,A) \Leftrightarrow P(X,A)$.

Утверждение 6.

Если некоторое производное от X понятие Y $P'(X,Y)$ имеет отношение проекции чтение/запись с источником данных $P_1(Y,A)$, то при наличии другого производного от X понятия Z $P'(X,Z)$, экземпляры которого не входят в множество экземпляров понятия Y : $Z \cap Y = \varnothing$, должно быть определено отношения проекции чтения/запись $P_1(Z,B)$, при этом источники A и B могут не совпадать. Если такое понятия Z отсутствует, то это свидетельствует о наличии экземпляров понятий X, которые не имеют средств редактирования.

Атрибуты понятия обычно проецируются на переменные примитивных типов данных и хранятся в основном источнике данных, являясь одним из полей соответствующей таблицы (представления). Основным источником данных считается тот, в котором хранится уникальный идентификатор экземпляра понятия. Свойства понятия могут храниться в основном источнике данных понятия, если отношения между понятиями, которые описывают эти свойства, представляют собой отношения один к одному. Отношения между понятиями типа многие ко многим или один ко многим всегда описываются свойством понятия и имеют отношения проекции с иным источником данных.

Ограничения на отношения могут определять единственность основного источника данных.

Пусть $\left|\overline{S(X,Y)}\right|_{\max}, \left|\overline{S(X,Y)}\right|_{\min}, \left|\overline{S(X,Y)}\right|_{pre}$ - максимальная, минимальная и точная допустимая мощность множества экземпляров свойств $S(X,Y)$ соответственно, а $\left|S(X,Y)_{\mathrm{Re}}\right|$ - действительная мощность множество экземпляров свойств.

Если $\left|\overline{S(X,Y)}\right|_{\max} = 1 \wedge \left|\overline{S(X,Y)}\right|_{pre} = 1$, то в большинстве случаев основной источник данных понятия X совпадает с основным источником данных свойства $S(X,Y)$, т.е. $P(X,A) \wedge P(S(X,Y),A)$. Это означает, что одному экземпляру понятия X соответствует единственный экземпляр понятия Y и это описано хранится в том же источнике, что и X.

Если $\left|\overline{S(X,Y)}\right|_{\max} = 1$, то это означает, что экземпляр понятия X сопоставляется не более чем с одним экземпляром понятия Y, т.е. имеет единственный источник и его атрибут, определяющий отношения, может быть не определен.

Если $\left|S(X,Y)\right|_{\max} \geq 1 \vee \left|S(X,Y)\right|_{pre} \geq 1 \vee \left|S(X,Y)\right|_{\min} \geq 1$, то источник данных, имеющий отношения проекции с $S(X,Y)$, отличается от источника данных, имеющего отношения проекции с X: $P(X,A) \wedge P(S(X,Y),B) \Rightarrow A \neq B$.

Алгоритм автоматический репликации данных

Сформулируем аксиомы, обеспечивающие генерации автоматической репликации.

Аксиома 3.

Если понятие X имеет проекции с источниками A и B с типом доступа чтения и чтения/записи соответственно $P_2(X,A)$, $P_1(X,B)$, то должна существовать репликация всех экземпляров понятия X, хранящихся в источнике B, в источник A $P''(B,A)$ (Рис.1.).

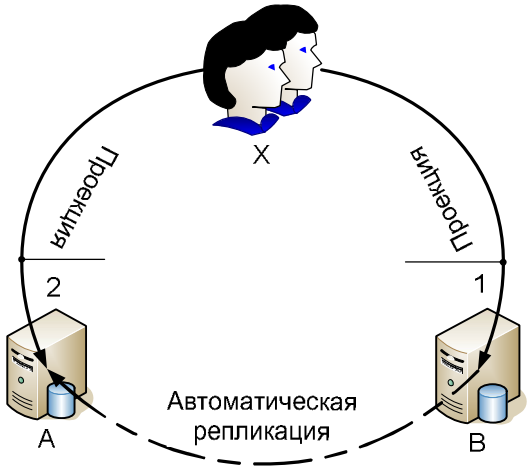

Рис.1. Автоматическая репликация данных на базе отношений проекции

Аксиома 4.

Если понятие X имеет отношения проекции с источником A с правом доступа на чтение $P_2(X, A)$, понятие Y , производное от X (т.е. определены отношения $P'(X, Y)$), имеет отношения проекции с источником B с правом доступа на чтение/запись $P_1(Y, B)$ и $X \neq Y$ ($\overline{Y} \subset \overline{X}$), то должна существовать репликация всех экземпляров из источника B в источник A (Рис.2), т.е. отношения репликации между источниками: $P''(B, A)$. В общем случае понятие Y может содержать дополнительные атрибуты по сравнению с базовым понятием X . В этом случае при автоматической репликации дополнительные атрибуты не реплицируются.

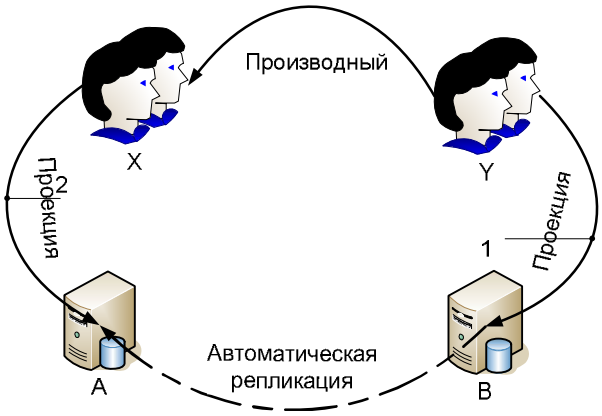

Рис.2. Автоматическая репликация данных на базе отношений наследования и проекции (из производного класса в базовый)

Аксиома 4.

Если понятие X имеет отношения проекции с источником A с правом доступа на чтение/запись $P_1(X, A)$ и понятие Y, производное от понятия X ($P'(X, Y)$), имеет отношения проекции с источником B с правом доступа на чтение $P_2(Y, B)$ и $X \neq Y$, то должна существовать репликация всех объектов класса Y из источника A в источник $B : P''(A, B)$ (Рис.3.).

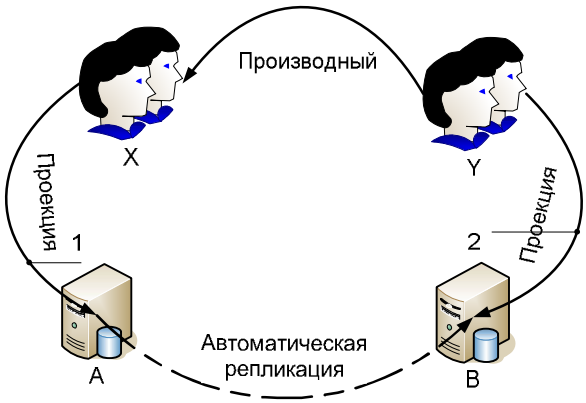

Рис.3. Автоматическая репликация данных на базе отношений проекции и наследования (из базового класса в производный)

Аксиома 5.

Наличие нескольких источников данных, имеющих отношения проекции с правами на чтение с базовым понятием $P_2(X, A), P_2(X, C)$ и наличие отношений проекции между производным классом $P_1(Y, B), P'(X, Y)$, приводит к репликациям вида $P''(B, A), P''(B, C)$ (Рис.4).

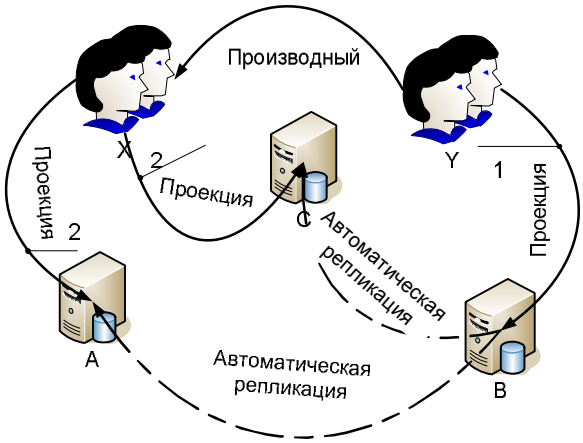

Рис.4. Автоматическая репликация данных в несколько источников на базе отношений

Отношения репликации имеют атрибутом понятие, экземпляры которого должны быть реплицированы.

Алгоритм генерации репликаций состоит в следующем.

1. Определяются отношения проекции с правами чтения/запись для всех понятий в некоторой иерархии. Поиск понятий начинается от базового понятия Object.
2. Для каждого понятия X, имеющего отношения проекции с некоторым источником A_X с правами чтения/запись $P_1(X, A_X)$, определяются все источники, с которым у X есть отношения проекции с правами на чтение $P_2\left(X, B_X^i\right)_{i=\overline{1,N_X}}$. Согласно аксиоме 3 генерируются репликации вида $P''\left(A_X, B_X^i\right)_{i=\overline{1,N_X}}$.
3. Все производные понятия от X (если оно имеет отношения проекции с типом на чтения/запись) должны иметь отношения проекции только с правами на чтение $P'(X, Y_i)_{i=\overline{1,N_Y}}, P_2(Y_i, C_i)_{i=\overline{1,N_Y}}$, отсюда следует, что определена автоматическая репликация $P''(A, C_i)_{i=\overline{1,N_Y}}$.
4. Если первое понятие, которое встретилось в иерархии, имеет отношения проекции только с правами на чтение $P_2(X, A_i)_{i=\overline{1,N_X}}$, то происходит дальнейший поиск по иерархии с целью обнаружения производных

понятий, имеющих отношения проекции с правами на чтения/запись $P'(X,Y_i)_{i=\overline{1,N_Y}}, P_1(Y_i,B_i)_{i=\overline{1,N_Y}}$. Для таких понятий, во-первых, применяется аксиома 3 и определяется репликация в источники с отношениями проекции только на чтение $P_2(Y_i,C_{ij})_{i=\overline{1,N_Y},j=\overline{1,N_{Y_i}}}$ $P''(B_i,C_{ij})_{i=\overline{1,N_Y},j=\overline{1,N_{Y_i}}}$. Во-вторых, для базового понятия X рассматриваются источники, с которыми у X есть отношения проекции с правами только на чтение $P_2(X,A_i)_{i=\overline{1,N_X}}$. В общем случае должны существовать репликации вида $P''(B_j,A_i)_{j=\overline{1,N_Y},i=\overline{1,N_X}}$. Но в более частном случае часть источников B_j, A_i могут совпадать. В случае совпадения источников репликация не генерируется.

В общем случае одно понятие может связно отношениями проекции с несколькими источниками данных (с несколькими таблицами), поэтому при автоматической генерации репликации генерируются репликации для всех источников данных, для которых описаны отношения проекции.

При выполнении автоматической репликации может возникнуть необходимость не в простом копировании значений, а в выполнении некоторого преобразования. Например, понятие X , состоит из двух атрибутов $X=\{x_1,x_2\}$, а при репликации необходимо сохранять понятие с одним атрибутом, равным сумме двух атрибутов X . Для реализации это идеи формируется понятие $Y: P'(X,Y), Y=\{x_1,x_2,y=x_1+x_2\}$. При определении отношений проекции для $P_2(Y,B)$ определяется соответствие только для атрибута y . В результате репликации в источнике B хранятся агрегированные по атрибутам экземпляры понятия X .

Если в производном понятии имеются атрибуты, которые отсутствуют в базовом и определена репликация из источниками базового понятия в источник производного, то атрибуты производного понятия, которые не включены в базовое понятие могут быть определены на основании значений по умолчанию – ограничений для атрибутов понятий в КИС. Значения по умолчанию представляют собой некоторые регулярные выражения. Если значение не определено, то оно устанавливается в null, если это - атрибут

понятия. Для свойств, которые не определены в базовом понятии и не имеют значений по умолчанию в производном понятии, при репликации экземпляры отношений не создаются.

Проблема, которая возникает в задачах репликаций распределенных хранилищ, решаемая другими средствами [6], связана с разными именами исходной и целевой таблицы и полей. Это проблема легко в случае рассматриваемого подхода решается на основании описания отношений проекции одних и тех же атрибутов реплицируемого понятия на источники данных и поля с разными именами.

Еще одной проблемой репликаций является проблема разных типов данных в источнике и приемнике репликации. Решение этой проблемы зависит от того, намеренные или ошибочные эти расхождения. В КИС в модуле контроля корректности описания понятия выполняются проверки на совпадения типов атрибутов источников данных, которые являются проекцией одних и тех же атрибутов понятий в разных источниках. При нахождении различий выполняется уведомление администратора проекта.

В отдельных случаях такие расхождения определены намерено, и для осуществления репликации требуется преобразования типов. Для преобразования могут использоваться либо стандартные средства СУБД.

Обеспечение синхронизации данных

Для реализации синхронных репликаций предлагается использовать автоматически сгенерированные репликации и отношения между понятиями. Как рассматривалось выше, на основании отношений проекции и наследования могут быть сгенерированы репликации. Но порядок их следования должен быть уточнен. Каждая автоматически сгенерированная репликация содержит копирования экземпляров одного понятия.

Рассмотрим пример. Пусть определены отношения проекции $P_1(X, A_X), P_2(X, B_X), P_2(X, C_X)$. Процедура автоматизации репликаций на основании аксиомы 3 делает заключение, что должны существовать

отношения репликации $P''(A_X, B_X), P''(A_X, C_X)$. Пусть существует понятие Y, имеющие свойство, связывающее его с понятием X: $\widetilde{P}(Y, X)$. Для понятия Y определены отношения наследования: $P'(Y, Y_1), P'(Y, Y_2)$ и отношения проекции $P_1(Y_1, A_Y), P_1(Y_2, B_Y), P_2(Y, A_Y), P_2(Y, C_Y)$. Процедура автоматизации репликаций на основе аксиом 4 и 5 делает заключение о необходимости репликаций $P''(B_Y, A_Y), P''(A_Y, C_Y)$.

Таким образом, процедура автоматизации репликаций на основании отношений проекции и наследование сгенерировала репликации

$P''(A_X, B_X), P''(A_X, C_X), P''(B_Y, A_Y), P''(A_Y, C_Y)$.

Для корректного выполнения репликаций необходимо определить порядок их следования и возможность параллельного выполнения.

Порядок следования репликаций с различными источниками/приемникам, определяется наличием отношений между понятиями. Отношения $\widetilde{P}(Y, X)$ определяют свойства понятия Y, поэтому независимым считается понятие X. Таким образом, первым реплицируются экземпляры понятия X: $P''(A_X, B_X), P''(A_X, C_X)$.

Понятие Y связано единственным свойством $\widetilde{P}(Y, X)$ с уже реплицированным понятием, поэтому следующим блоком будет блок репликаций понятия Y. Поскольку источники репликаций различны, то они не могут выполняться параллельно. Последовательность репликаций определяется тем, что один из источников является приемником в другой репликации, поэтому последовательность следующая – первой выполняется репликация $P''(B_Y, A_Y)$, а второй - $P''(A_Y, C_Y)$.

Алгоритм синхронизации репликаций состоит в следующем. Из всех понятий, участвующих в репликациях, выбираются те понятия, которые не имеют никаких отношений ни с одним другим понятием, участвующим в репликациях: $\forall X : \neg P(X, Y)$. Фактически, на первом этапе выбираются независимые понятия. При выполнении репликаций на очередной итерации выбираются те репликации, которые связаны с понятиями либо не

имеющими свойств (т.е. независимыми понятиями), либо их свойства связаны с уже реплицированными понятиями. Внутри блока репликаций одного понятия в случае одного источника для нескольких репликаций, репликации могут выполняться параллельно. В противном случае порядок репликаций внутри блока репликаций одного понятия определяется на базе источников и приемников репликации. В первую очередь выполняются те репликации, приемники в которых не являются источниками в других репликациях внутри блока.

Процесс выполнения автоматических репликаций с синхронизацией, как описано выше выполняется специализированным приложением, входящим в семантический базис КИС. Но при наличии большого числа репликаций узким местом может оказаться сервер приложений, на котором будут выполняться приложение. Для решения проблемы производительности можно использовать другие серверы приложений с распределением на них работы, что реализуется с использованием сервиса баланса нагрузки.

Но здесь возникает проблема синхронизации приложений в распределенной системе. Одним из инструментов синхронизации является событие.

Между репликациями и событиями существуют отношения генерации $P_G(P''(A_X, B_X), \varepsilon_X)$, которые в зависимости от атрибутов определяют начало или завершение репликации.

С помощью установленных событий приложения, реализующие репликации зависимых понятий на различных серверах, узнают о завершенной репликации тех понятий, которые связаны с подготовленным к репликации понятием.

В некоторых случаях репликация $P''(A_Y, C_Y)$ имеет смысл даже, если репликация $P''(B_Y, A_Y)$ не была выполнена. В этом случае на C_Y будут перенесены только те изменения, которые были сделаны за предыдущий период в источнике данных A_Y, без учета изменений в источнике B_Y. Это возможность определяется наличием отношений проекции с правом на

чтение/запись $P_1(Y_1, A_Y)$. Приложение, выполняющее репликацию, определяет невозможность выполнить репликацию $P''(B_Y, A_Y)$, фиксирует неуспех репликации и выполняет репликацию $P''(A_Y, C_Y)$ с уведомлением администратора о неполных данных по экземплярам понятия Y в источнике C_Y.

Проблема уникальных ключей

Одна из проблем, возникающих при репликации данных - это проблема идентификационных ключевых полей. Так как физически базы данных разные, то возникает необходимость синхронизировать между собой первичные ключи таблиц. Существует несколько подходов к решению этой проблемы.

1. Определяется для каждой базы данных диапазон изменения уникальных идентификаторов, при репликации данных в виду разных идентификаторов наложения не произойдет. Но в случае если на поле первичного ключа наложено требование авто увеличения, то диапазон будет нарушен после первой же репликации.
2. Разновидностью первого пункта может быть второй пункт, когда в каждой базе данных идентификаторы вычисляется по некоторой формуле. Например, когда остаток от деления идентификатора на константу определяет, в какой базе генерируется идентификатор. Подход хорошо работает, если это учитывалось с самого начала построения КИС и константа не меняется. Но такое не всегда возможно.
3. Использование дополнительного идентификатора для определения базы данных. Подход требует изменения процедур идентификации уже существующих данных, что не всегда возможно.
4. Изменение кода в базе данных филиала в соответствии с кодами головного вуза. В каждой базе ведется нумерация в соответствии с общими правилами. Но при репликации данных идентификационные поля в базе данных источника репликации заменяются на те, которые

присвоились этим записям в приемнике. Подход опасен тем, что при задержке в репликации возможно использования неизмененных кодов в других задачах.

5. Разновидность 4-го пункта – сохранение в приемнике репликации соответствия идентификационных данных источника и приемника. Поля таблицы соответствия заполняются после выполнения репликации.

Предпочтительными для КИС являются два подхода: первый и последний. Первый подход позволяет определить для каждого источника репликации свой диапазон идентификаторов для тех случаев, где не используется ограничения по автоувеличению для идентификатора записи. Для случаев с автоувеличением идентификаторов используется пятый подход с введением в таблицы дополнительных полей.

Рассмотрим, как решается проблема уникального ключа для автоматической репликации на базе онтологического подхода. Пусть существует базовое понятие X, которое имеет отношения проекции с источником данных A в базе данных головного вуза, при этом доступ определен для чтения $P_2(X, A)$. Понятие Y как производное от понятия X имеет отношение проекции с тем же источником A, но с доступом для чтения/записи $P'(X, Y), P_1(Y, A)$. Производное понятие $Z : P'(X, Z)$ имеет отношения проекции с доступом чтения/записи $P_1(Z, B)$. На основании описания этих понятий и аксиомы 4 формируется репликация, которая обеспечивает копирование из B в A: $P''(B, A)$.

В первую очередь анализируются ограничения наложенные на уникальные идентификаторы в понятиях X, Y, Z. Ограничения на атрибуты производных понятий могут отличаться от ограничений на атрибуты базовых понятий. Если среди ограничений присутствуют ограничения по диапазону, то выполняется проверка на пересечение диапазонов идентификаторов Y, Z. Если диапазоны не пересекаются, то используется первый подход, если диапазоны в ограничениях не определены или диапазоны пересекаются, то

используется пятый подход, который требуется наличия соответствия между экземплярами понятий, имеющих проекцию на разные источники данных.

Первый подход не требует никаких дополнительных описаний, так как при репликации никаких изменений в объектах не происходит. Пятый подход несколько сложнее. При определении отношений проекция для одного и того же понятия (или понятий в одной иерархии) на различные источники данных требуется определение источника для проекции соответствия между экземплярами понятий. Источники, отвечающие за соответствия, определяются либо автоматически (т.е. выполняется их генерация в виде специализированных таблиц) и располагаются они в обеих база данных, либо некоторые, уже существующие таблицы определяются как источники соответствий.

Проблема целостности данных

Проблема возникает в связи с возможным нарушением целостности данных при репликации. Например, в источнике данных A_X ранее существовал экземпляр понятия X, который удалили в связи с тем, что не использовали в настоящем и не собираются использовать в будущем. Но в базе филиала C_Y могли внести информацию, связанную с этим экземпляром (т.е. существуют отношения $\widetilde{P}(Y,X), P(Y,C_Y)$). При попытке реплицировать обновленное понятие $P''(A_X,C_X)$ произойдет ошибка, и репликация не будет выполнена. Ошибка возникает в связи с необходимостью удалить экземпляр понятия X в C_X, при наличии экземпляра Y в C_Y, имеющего отношения $\widetilde{P}(Y,X)$.

Решений этой проблемы может быть несколько.

1. Удалить экземпляры, хранящиеся в C_Y, соответствующие тем, которые затем удалить из C_X.
2. Реплицировать только новые экземпляры X в C_X и не удалять те, которые используются в C_Y.

3. Полностью заменить C_X на A_X и заменить связи с несуществующими экземплярами на некоторые заранее определенные связи (это может быть null или определенный экземпляр из A_X) с уведомлением администратора.
4. Откатить репликации с уведомлением администратор о причине.

Первое решение нельзя признать корректным, так как может быть удалена нужная информация. Второе решение так же нельзя признать корректным в связи с тем, что реплики перестают быть синхронными и это влечет за собой дальнейшие ошибки при репликации оперативных данных.

Наилучшим решением является 3-ье решение, если его можно выполнить автоматически. Для поддержки автоматической замены используется значения по умолчанию для атрибутов понятий.

Пусть описано понятие Y, имеющее отношение с понятием X $\tilde{P}(Y,X)$. При этом один из атрибутов понятия Y $Y \to c$, отвечающий за эту связь, принимает значения из области атрибута $D = \{d_i\}_{i=1}^{N}$ понятия X. Необходимо определить, что атрибут $Y \to c$ в случае выхода из области D принимает определенное значение из D, например, d_1, т.е. $if\ \exists i\ y_i \to c \notin D \Rightarrow y_i \to c_i = d_1$. Для такого определения используется механизм описания ограничений на атрибуты – установки по умолчанию. В установках по умолчанию могут использоваться любые регулярные выражения, включающие константы (в том числе null) и атрибуты этого или другого понятия.

В случае, когда выполнить замену нельзя (отсутствуют или значения по умолчанию, или составляющие регулярного выражения, определяющего значение по умолчанию), то уведомление об ошибке репликации с указанием причины отправляется администратору.

Заключение

В работе рассматривается алгоритм автоматической репликации данных в гетерогенной КИС на основе онтологического описания понятий КИС и отношений их с ИТ-понятиями. Преимуществами алгоритма являются, во-первых, обеспечение с помощью него адаптивности КИС в

вопросе репликации данных (т.е. возможность выполнения репликаций без необходимости их определения администратором КИС), во-вторых, использование онтологического описания, которое генерируется или формируется в процессе создания КИС и используется в различных задачах функционирования и поддержания КИС, не связанных непосредственно с задачей репликации, в-третьих, такой алгоритм позволяет иметь наглядную картину выполнения всех репликаций в КИС.

Литература

[1] Herring C. Viable Software. The intelligent control paradigm for adaptable and adaptive architecture. PhD Thesis.- Australia.-2002.

[2] Хенкин В., Навроцкий С. Опознанные летающие объекты. Открытые системы. №9. 1999.

[3] Berg A., Bode J., Bataille J. Integration of campus wide information systems using a hub and spoke architecture//Proceedings of the 11-th International Conference of European University Information Systems (EUNIS 2005), UK

[4] Kohutkova J. Meta-Transformations in systems integration: the concept and the use// Proceedings of 10-th International Conference of European University Information Systems EUNIS 2004. Slovenia 2004. pp. 253-258.

[5] Шахгельдян К.И. Применение онтологического подхода в корпоративной информационной среде вуза//НТ Ведомости СПбГПУ.-2007.-№4-2 (52).-189-194

[6] Бездушный А.А., Бездушный А.Н., Нестеренко А.К., Серебряков В.А., Сысоев Т.М. Интеграция распределенных данных на основе технологий Semantic Web и рабочих процессов. //Сборник докладов Шестой Всероссийской конференции «Электронные библиотеки: перспективные методы и технологии, электронные коллекции», Санкт-Петербург, 2004