МАТЕМАТИЧЕСКОЕ МОДЕЛИРОВАНИЕ, КОМПЬЮТЕРНЫЕ ТЕХНОЛОГИИ В ТЕОРИИ УПРАВЛЕНИЯ СЛОЖНЫХ ПРОЦЕССОВ

МЕТОДИКА ПРОЕКТИРОВАНИЯ РАСПРЕДЕЛЕННЫХ ИНФОРМАЦИОННЫХ ХРАНИЛИЩ

А.Ю. Иванов, доктор технических наук, профессор;
В.С. Горшков. Санкт-Петербургский университет ГПС МЧС России

Рассмотрена последовательность и содержательное наполнение этапов проектирования распределенных информационных хранилищ автоматизированных систем МЧС России. Представлена общая схема проектирования. Определены цели, указаны исходные данные, виды работ и результаты реализации каждого этапа.

Ключевые слова: распределенное информационное хранилище, проектирование, логическая структура, логический фрагмент, физический фрагмент

DESIGN METHOD OF DISTRIBUTED DATA WAREHOUSES

A.Y. Ivanov, V.S. Gorshkov.
Saint-Petersburg university of State fire service of EMERCOM of Russia

The article discusses sequence and contents of distributed data warehouses of automated systems of EMERCOM of Russia. A common design scheme is given. Aims are defined, initial data, work kinds and results of each stage realization are indicated.

Key words: distributed data warehouse, design, logical structure, logical fragment, phsycal fragment

В настоящее время в практику МЧС России активно внедряются автоматизированные системы. Их информационную основу составляют базы данных. Потребность в решении должностными лицами органов управления задач аналитического характера определила необходимость перехода от традиционных баз данных к более сложным образованиям – информационным хранилищам (ИХ). Иерархическое построение и территориальное рассредоточение системы управления МЧС России обусловили следующую особенность: информационные хранилища должны быть распределенными [1].

Одной из основных задач, стоящих перед разработчиками распределенных информационных хранилищ (РИХ), является определение порядка и содержания работ, связанных с их проектированием.

В соответствии с традиционной схемой разработки баз данных, архитектурой РИХ и уровнями представления данных целесообразно рассмотреть совокупность работ, определяющую общую схему построения хранилища (рис.), которую можно разделить на следующие этапы: концептуальное моделирование; логическое проектирование; фрагментирование логической структуры, распределение фрагментов и физическая реализация [2,3].

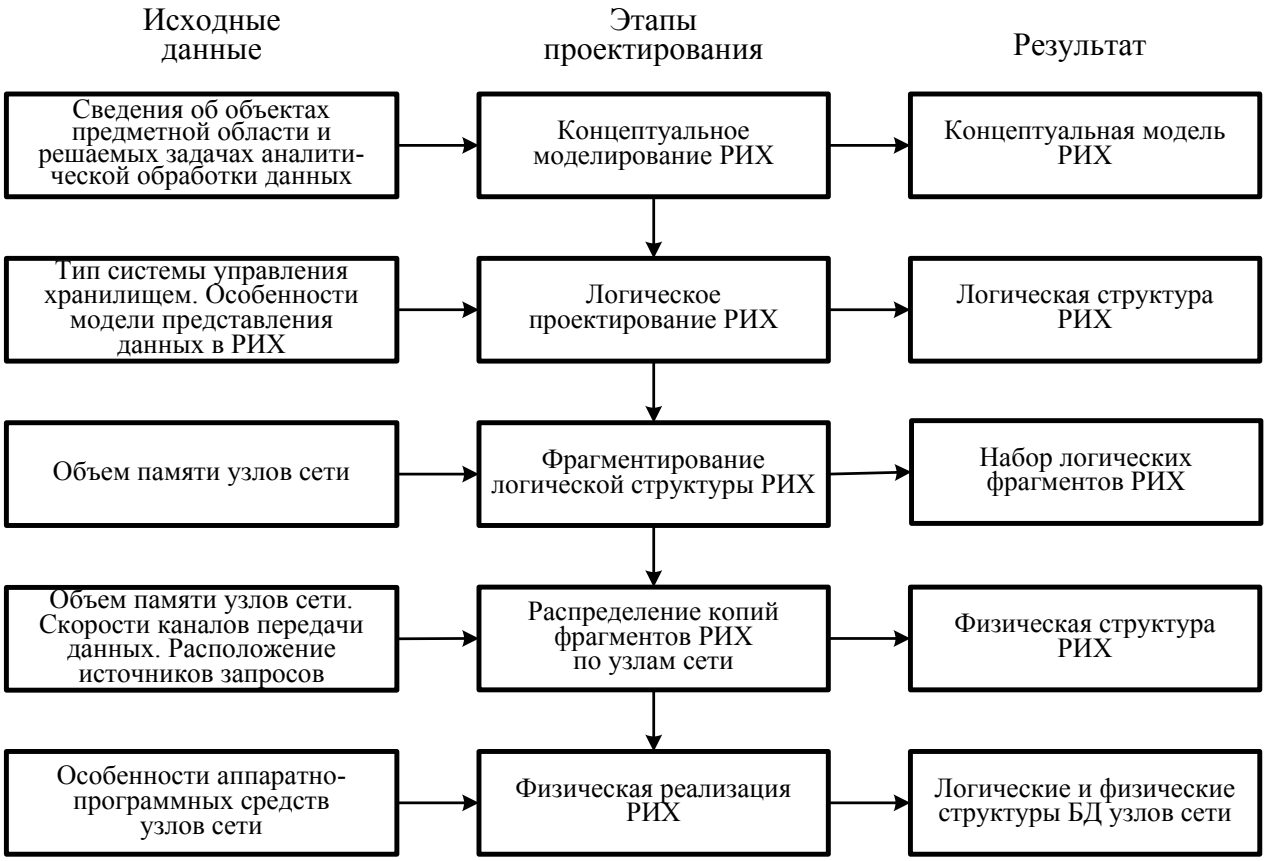

Рис. Общая схема проектирования распределенных информационных хранилищ

В процессе проектирования происходит преобразование концептуальной модели предметной области в логическую структуру РИХ, которая в свою очередь трансформируется в физическую структуру. Технология построения хранилища допускает итеративный характер выполнения этапов.

Этап *концептуального моделирования* предназначен для изучения как предметной области, так и задач аналитической обработки данных и интеллектуального анализа данных, а также для формирования концептуальной модели хранилища.

Исходными данными для этого этапа выступают:

– предполагаемые запросы пользователей к данным, находящимся в хранилище, их семантика и внешние характеристики, например, частота реализации;

– требования к процессу исполнения запросов, такие как временные параметры, показатели достоверности и т.п.;

– характеристики функциональных задач, связанных с аналитической обработкой данных, а именно: исходные данные, частота решения и др.

На этапе концептуального проектирования выполняются следующие работы [4]:

– определение детализированных данных (атрибутов-факторов), подлежащих хранению и многомерному анализу;

– разработка структуры информационных измерений ИХ, включающей формирование набора измерений, определение иерархий данных, их уровней и элементов и определение способов агрегирования данных;

–определение источников детализированных данных;

– определение прав доступа к детализированным и агрегированным данным.

Определение детализированных данных является одним из основных видов работ. Детализированные данные (атрибуты-факторы) – это информационные элементы, находящиеся в ячейках многомерного гиперкуба. Они определяются исходя из потребностей пользовате-

лей хранилища. При определении атрибутов устанавливается формат их представления, а также область допустимых значений (домен).

Формирование набора информационных измерений хранилища обусловлено необходимостью установления размерности гиперкуба и семантики его измерений. Количество измерений в концептуальной модели не ограничивается и выделяется исходя из потребностей пользователя.

Выявление источников детализированных данных требуется для предварительной обработки данных, загружаемых в хранилище. В ходе этого вида работ устанавливается:

– в какой операционной базе данных или в каком внешнем источнике хранятся атрибуты, и каков их формат;

– как преобразовать данные в различных источниках к единому ранее установленному формату;

– каким способом вычислять предварительно определяемые агрегаты;

– какова периодичность вычисления агрегатов и загрузки хранилища детализированными и агрегированными данными.

В дальнейшем для реализации предварительного вычисления агрегатов, преобразования форматов и периодической загрузки хранилища следует воспользоваться возможностями, имеющимися в предполагаемой к использованию системами управления базами данных (СУБД), или сформировать для этой цели дополнительные программные компоненты.

Определение прав доступа к детализированным и агрегированным данным необходимо для обеспечения требований по безопасности данных. Важность данного вида работ, в ходе которого формируется концептуальная схема доступа, обусловлена тем, что в многомерном хранилище в одной зоне безопасности могут находиться данные, относящиеся к различным уровням конфиденциальности. Следовательно, для удовлетворения требований по безопасности информации в хранилище следует либо использовать средства повышенных классов безопасности, либо проводить декомпозицию единого гиперкуба на множество мелких.

Таким образом, результатом этого этапа выступают:

– множество атрибутов-факторов, форматы их представления и исходные домены;

– множество информационных измерений и их состав;

– источники и процедуры получения детализированных данных;

– процедуры и периодичность формирования агрегированных данных;

– схема разграничения доступа к детализированным и агрегированным данным.

Все перечисленные позиции в совокупности составляют общий замысел построения или *концептуальную модель* проектируемого распределенного информационного хранилища.

Этап *логического проектирования* связан с созданием логической структуры информационного хранилища.

Исходными данными для логического проектирования являются результаты предыдущего этапа.

Последовательность работ на этапе включает в себя:

– выбор программно-инструментального средства создания информационного хранилища, к которым относятся традиционные СУБД или специализированные программные средства, получившие название «системы управления информационными хранилищами (СУИХ);

– разработку исходной логической структуры БД ИХ, соответствующей поддерживаемой СУБД или СУИХ модели представления данных;

– оптимизацию логической структуры ИХ.

В информационных хранилищах детализированные и агрегированные данные могут храниться либо в виде реляционных, либо многомерных структур. Поэтому в настоящее время применяются три способа хранения данных;

– специализированные многомерные системы управления информационными хранилищами, поддерживающие многомерную модель данных и технологию многомерной опера-

тивной аналитической обработки данных (МОАОД от англ. *MOLAP – Multidimensional OLAP*). В соответствии с таким подходом исходные и агрегированные данные хранятся в многомерной базе данных. Хранение данных в многомерных структурах позволяет манипулировать данными как многомерным массивом, благодаря чему скорость вычисления агрегированных значений одинакова для любого из измерений. Однако в этом случае многомерная база данных оказывается избыточной, так как многомерный куб полностью содержит исходные реляционные данные;

– традиционные реляционные СУБД, поддерживающие только реляционную модель данных и технологию реляционной оперативной аналитической обработки данных (РОАОД, от англ. *ROLAP – Relational OLAP*). Технология предполагает, что исходные данные остаются в той же реляционной БД, где они изначально находились. Агрегированные данные помещают в специально созданные для их хранения служебные таблицы той же БД.

– гибридные СУИХ, являющиеся по сути реляционными, но обладающие возможностями создания и поддержки многомерных объектов в соответствии с технологией гибридной оперативной аналитической обработки данных (ГОАОД, от англ. *HOLAP Hybrid OLAP*). Особенностью технологии является то, что исходные данные остаются в той же реляционной БД, где они изначально находились, а агрегированные данные хранятся в многомерной БД.

Некоторые средства (СУИХ, СУБД) поддерживают хранение данных только в реляционных структурах, другие – только в многомерных. Большинство из них сочетает все три способа хранения данных. Выбор способа хранения зависит от объема и структуры исходных данных, требований к скорости выполнения запросов и частоты обновления гиперкубов. Однако следует заметить, что многомерные и гибридные СУИХ в настоящее время находятся в стадии экспериментальных исследований и пока не вышли на уровень широкого распространения.

Целью построения исходной логической структуры ИХ является удовлетворение требований по полноте отображения данных, осуществляемого в ходе преобразования концептуальной модели в логическую структуру ИХ.

Практикой выработан ряд эмпирических правил, которых следует придерживаться на этапе логического проектирования [4]:

– для каждого гиперкуба концептуальной модели данных первоначально формируется своя подсхема типа «звезда», в которой имеется одна таблица факторов и столько таблиц измерений, сколько измерений имеется у гиперкуба;

– в случае наличия иерархии элементов вдоль некоторого измерения соответствующая таблица измерений декомпозируется на несколько таблиц, число которых соответствует числу уровней иерархии;

– для декомпозированных таблиц измерений, соответствующих высшим уровням иерархии, формируются дополнительные таблицы факторов, содержащие агрегированные по этим уровням данные.

Возможности оптимизации логической структуры заключаются в том, что, как известно, исходный вариант логической структуры в случае реляционной или гибридной технологии не является единственно возможным. В качестве допустимых операций его преобразования выступают следующие:

– объединение таблиц, соответствующих уровням иерархии информационных измерений;

– объединение или декомпозиция таблиц факторов, содержащих детализированные и (или) агрегированные данные;

– декомпозиция единого гиперкуба на более мелкие.

Эти и другие операции, применяемые к исходной логической структуре и ее последующим модификациям, позволяют сформировать множество альтернативных вариантов логической структуры, каждый из которых характеризуется своей оперативностью обработки аналитических запросов различных видов.

Ограничением обычно выступает требование недопустимости совместного нахождения

в одной таблице атрибутов разных уровней конфиденциальности, которое вытекает из концептуальной схемы доступа.

В качестве критерия оптимизации целесообразно выбрать среднее время обработки аналитических запросов. Ограничением обычно выступает требование недопустимости совместного нахождения в одной таблице атрибутов разных уровней конфиденциальности, которое вытекает из концептуальной схемы доступа.

В настоящее время разработанных и апробированных точных методов оптимизации логической структуры ИХ не существует.

Результатом этапа логического проектирования выступает глобальная логическая структура распределенного информационного хранилища.

Этап *фрагментирования* логической структуры нацелен на получение набора логических фрагментов.

Исходными данными для этого этапа выступают:

– глобальная логическая структура РИХ;

– характеристики узлов сети, в которых предполагается размещение частей хранилища, в частности, объем внешней памяти.

Рассматриваемый этап связан с делением глобальной логической структуры на логические фрагменты (ЛФ).

Глобальная структура разделяется на множество логических фрагментов. Естественно, что при этом должно выполняться требование о сохранении информации, то есть логические фрагменты в совокупности должны содержать все сведения, имеющиеся в хранилище. Дополнительно на процесс формирования разделов накладываются ограничения по их допустимому размеру, времени реакции на запрос и надежности обращения. Вследствие этого в ЛФ рекомендуется объединять часто используемые информационные единицы, чтобы улучшались характеристики времени ответа на запрос. Допустимый размер каждого ЛФ, как неделимой совокупности данных, определяется фиксированным объемом памяти в каждом узле сети. И в общем случае ограничения на класс допустимых расчленений накладывает емкость внешних запоминающих устройств в узлах.

При формировании логических фрагментов используются операции проекции, селекции, вертикального и горизонтального срезов.

Формальные методы, позволяющие оптимизировать фрагментирование глобальной логической структуры, в настоящее время отсутствуют в силу следующих причин:

– трудность формирования критерия оптимизации;

– корреляция процесса фрагментирования с последующим процессом распределения копий логических фрагментов.

Результатами являются совокупность логических фрагментов и размер каждого из них.

Этап *распределения* копий логических фрагментов необходим для формирования содержательного наполнения узлов сети необходимыми данными.

Исходными данными для этого этапа являются:

– логические фрагменты и их параметры;

– требования к процессу обработки аналитических запросов (временные характеристики, частота и др.);

– перечень и характеристики узлов сети, в которых предполагается размещать информационные компоненты хранилища;

– скорости каналов передачи данных;

– расположение должностных лиц, в интересах которых будут исполняться запросы.

Решение задачи распределения копий ЛФ, называемых физическими фрагментами, в существенной степени определяется принятой стратегией.

При принятии к реализации стратегии дублирования решение находится довольно просто. Более того, при этом отпадает необходимость выполнения предыдущего этапа. При анализе характеристик узлов сети требуется установить целесообразность размещения в каждом из них хранилища в полном объеме. Ответ на поставленный вопрос в большинстве случаев

предопределен и зависит от структуры сети, объема памяти в ее узлах и здравого смысла.

Значительно сложнее представляется задача размещения при использовании стратегии разделения и особенно сложной при смешанной стратегии. В случае реализации стратегии разделения необходимо: во-первых, расчленить глобальную структуру на логические фрагменты и, во-вторых, разместить каждый фрагмент в конкретном узле с учетом ограничений на размещение. Задача является итеративной и возможно, что расчленение глобальной логической структуры потребуется проводить неоднократно. Если же используется смешанная стратегия, решение становится более сложным: каждый логический фрагмент может быть размещен в любом числе узлов. Количество перестановок фрагментов растет очень быстро, и это является одной из причин того, что ограничиваются нахождением не оптимального, а рационального размещения.

Неформальным критерием при решении задачи распределения физических фрагментов может выступать правило «80/20», которое предполагает, что 80 % запросов к хранилищу должно исполняться локально, а 20 % – удаленно. Существенный отход от этого критерия в сторону снижения первой величины и повышения второй приводит к следующему. Во-первых, ухудшаются показатели оперативности реализации запросов. Во-вторых, возрастает нагрузка на коммуникационную среду, особенно в случае высокой интенсивности обновления идентичных физических фрагментов.

Ограничениям являются объемно-временные характеристики узлов сети и пропускная способность каналов коммуникационной среды.

Решение задачи оптимального размещения фрагментов по узлам сети методами линейного программирования возможно при довольно жестких допущениях о характере потока запросов, предопределенном числе и неизменности этих запросов, заранее известном числе хранимых фрагментов. Количество переменных и ограничений прогрессирует с увеличением числа узлов сети и поэтому получение решения возможно лишь для задач малой размерности. Решение также усложняется при предъявлении менее жестких требований к характеру запросов пользователей. Подходы к решению используют в своей основе метод динамического программирования. Если же типовые запросы первоначально неизвестны, то используются статистические методы для определения этих запросов, а результаты их определения служат входными данными для решения задачи распределения фрагментов.

Результатом этапа выступают наборы физических фрагментов РИХ, в последующем размещаемые в назначенных узлах сети. Совокупность этих наборов образует физическую структуру распределенного информационного хранилища.

Этап *физической реализации*. Целью данного этапа является преобразование фрагментированной структуры хранилища в структуру данных локального уровня.

На этом этапе выполняются следующие работы:

– в случае использования технологий МОАОД и ГОАОД полученные при фрагментации многомерные данные преобразуются к реляционному виду, пригодному для обработки локальными СУБД;

– на каждом узле формируются так называемые ведомости рассылки, в которые включаются адреса идентичных физических фрагментов для последующей актуализации дублированных данных.

Другие работы, проводимые на узлах сети, соответствуют традиционному процессу формирования локальных баз данных.

В целом процесс разработки РИХ отличается высокой трудоемкостью и значительными материальными затратами и носит итеративный характер. Задача выбора наилучшего соответствия между характеристиками распределенного информационного хранилища и методами распределения данных в сети требует всестороннего анализа, так как принятые проектные решения оказывают непосредственное влияние на реализацию и последующее функционирование информационно-аналитической системы в целом.

Литература

1. Иванов А.Ю., Горшков В.С. Концепция распределенных информационных хранилищ // Проблемы управления рисками в техносфере. 2010. № 1[13]. С.67–74.

2. Голенищев Э.П., Клиненко И.В. Информационное обеспечение систем управления. Сер. Учебники и учеб. пособ. Ростов-н/Д.: Феникс, 2003. 285 с.

3. Иванов А.Ю., Саенко И.Б. Основы построения и проектирования реляционных баз данных. СПб.: ВАС, 1998. 80 с.

4. Саенко И.Б. Теоретические основы многомерно-реляционного представления данных и их применение для построения баз данных АСУ связью. СПб.: ВУС, 2001. 176 с.