На правах рукописи

УДК 681.3.016

ЩАВЕЛЁВ Леонид Вячеславович

АВТОМАТИЗАЦИЯ ПРОЕКТИРОВАНИЯ СИСТЕМ ОПЕРАТИВНОЙ АНАЛИТИЧЕСКОЙ ОБРАБОТКИ ДАННЫХ (НА ПРИМЕРЕ ИНФОРМАЦИОННО-АНАЛИТИЧЕСКИХ СИСТЕМ В ЭНЕРГЕТИКЕ)

Специальности	05.13.12 -	системы автоматизации проектирования (электротехника, энергетика)
	05.13.14 -	системы обработки информации и управления

Автореферат
диссертации на соискание ученой степени
кандидата технических наук

Иваново 1999

Работа выполнена на кафедре Вычислительной техники и систем автоматизированного проектирования Ивановского государственного энергетического университета.

Научный руководитель -
доктор технических наук, профессор Нуждин В. Н.

Официальные оппоненты -
доктор технических наук, профессор Фомин Б. Ф.;
кандидат технических наук, доцент Голяков С. М.

Ведущая организация -
Московский государственный институт электроники и математики (технический университет), г. Москва

Защита состоится 12 марта 1999 г. В 11 часов в аудитории Б-237 на заседании диссертационного совета Д 063.10.01 при Ивановском государственном энергетическом университете.

С диссертацией можно ознакомиться в библиотеке ИГЭУ.

Отзывы в двух экземплярах, заверенные печатью организации, просим направлять по адресу: 153003, Иваново, ул. Рабфаковская, 34, Совет ИГЭУ.

Автореферат разослан 12 февраля 1999 г.

Ученый секретарь
диссертационного совета,
доктор технических наук, профессор ТАРАРЫКИН С. В.

ОБЩАЯ ХАРАКТЕРИСТИКА РАБОТЫ

Актуальность темы. В постоянно изменяющихся условиях нынешнего динамичного мира сложность управления любой системой, будь то предприятие, ведомство или регион, непрерывно увеличивается, в то время как требования к его качеству значительно возрастают. Для постоянного контролирования множества жизненно важных факторов и нахождения оптимальных решений возникающих проблем необходима информационная поддержка, обеспечиваемая автоматизированными системами поддержки принятия решений (СППР). Потребность в информационной поддержке принятия решений существует и в проектировании, в случаях, когда его объектами становятся крупные народно-хозяйственные системы, структура и функционирование которых описываются большими объемами данных.

Наиболее остро данная проблема стоит в задачах управления и реорганизации больших производственных систем (например, в областях топливной энергетики, электроэнергетики), а также в региональном управлении, то есть там, где стратегические ошибки, возможно, не всегда заметные сразу, с течением времени могут приводить к тяжелым последствиям.

Очевидно, что необходима возможность оперативного доступа лиц, принимающих решения, к достоверным данным о состоянии системы и ситуации вокруг нее. Доступность актуальной информации дает возможность оценить текущее положение, а обзор изменения конкретных характеристик во времени позволяет обнаружить тенденции развития системы и сделать выводы о том, что ожидает ее в будущем. Только обладая всей полнотой сведений о состоянии системы и ее элементов в статике и динамике, можно принимать грамотные решения по применению мер регулирования.

Однако, простое обладание данными не становится решением проблемы. Развитие аппаратных средств быстро сделало ведение крупных банков данных доступным для множества корпоративных пользователей. Но огромные объемы накопленных в них сведений не могут эффективно использоваться без специального инструментария - информационно-аналитических систем (ИАС), позволяющих решить главную проблему заказчика: "слишком много данных, но недостаточно информации".

Очевидно, что разработка каждой конкретной ИАС "с нуля" - слишком расточительный подход. Более эффективным решением может стать использование инструментальных средств, ориентированных на автоматизацию проектирования прикладных ИАС в различных проблемных областях. При этом следует иметь в виду, что в большинстве случаев требуется выполнение аналитической обработки ранее созданных и активно работающих баз данных (БД), структура которых может быть достаточно сложной и не допускающей произвольной модификации существующих таблиц. Исходя из этого и была сформулирована цель диссертационной работы.

Цель диссертационной работы и задачи исследования. Целью работы является разработка подходов и методов организации инструментальной оболочки для автоматизации проектирования и для последующей эксплуатации информационно-аналитических систем (ИАС), выполняющих оперативную аналитическую обработку (многомерный анализ) содержимого реляционных баз данных произвольной структуры и обеспечивающих поддержку принятия решений.

Для достижения поставленной цели в диссертационной работе решаются следующие задачи.

1. Разработка механизмов обеспечения многомерного концептуального представления данных, содержащихся в реляционных базах данных (БД) произвольной структуры.
2. Разработка структуры аналитических метаданных для представления многомерных информационных моделей в реляционных БД, использующихся в качестве содержательного наполнения прикладных ИАС.
3. Разработка инструментария оперативной аналитической обработки данных в составе:
   - системы, предназначенной для создания, редактирования и проверки логической целостности информационных моделей, составляющих аналитические метаданные, и решающей задачу автоматизации проектирования ИАС, ориентированных на конкретную область применения;
   - системы оперативной аналитической обработки данных, содержащихся в реляционных БД произвольной структуры, снабженных аналитическими метаданными с описанием многомерных информационных моделей.
4. Разработка структуры функционального ядра системы оперативной аналитической обработки реляционных данных, открытого для расширения функциональности прикладных ИАС.
5. Отработка технологии построения прикладных ИАС на основе разработанного инструментария.

Методы исследования. При решении поставленных задач в диссертационной работе использованы теория множеств, теория построения информационных систем, теория баз данных, теория объектно-ориентированного проектирования.

Научная новизна работы заключается в разработке метода организации многомерного анализа реляционных данных, накопленных в хранилищах данных произвольной структуры, включающего в себя способ организации многомерных информационных моделей, принцип их описания на уровне аналитических метаданных, порядок их использования при выполнении аналитических запросов к многомерному концептуальному представлению и технологию построения прикладных информационно-аналитических систем на основе созданной инструментальной оболочки.

Получены следующие новые научные результаты.

1. Разработан механизм "виртуальной звезды", определяющий структуру и порядок использования информационных моделей для извлечения многомерной информации из реляционных БД произвольной структуры. В отличие от традиционно применяемых подходов разработанный механизм не требует обязательного приведения баз данных, использующихся для многомерного анализа, к специальному виду, а позволяет анализировать существующие хранилища данных без внесения изменений в их структуру. Предложенная в рамках механизма "виртуальной звезды" возможность разделения уровней обобщения атрибутов многомерной информационной модели на отрезки однородности, а множества ячеек модели - на области однородности получения значений позволяет совмещать в пределах одной модели ячейки со значениями показателей, извлекаемых из различных таблиц исходной БД и имеющих различную природу:

   1) значения полей численного типа таблиц БД;

   2) значения, полученные в результате оперативной агрегации в процессе выполнения запроса;

   3) значения, являющиеся результатами расчета по формулам.
2. Предложен принцип организации аналитических метаданных для хранения многомерных информационных моделей в реляционных БД, являющихся содержательным наполнением прикладных ИАС. Определена структура системных таблиц, вводимых в анализируемую БД для обеспечения конечным пользователям единого многомерного концептуального представления содержащейся в ней разрозненной численной информации.
3. Предложены принципы организации инструментальной системы оперативной аналитической обработки данных в рамках программного комплекса, являющегося инструментальной оболочкой для построения корпоративных ИАС. Согласно им в инструментальной системе оперативной аналитической обработки реляционных данных должно быть произведено явное разделение инструментария конечного пользователя (аналитика), работающего с заранее построенным многомерным концептуальным представлением исследуемого хранилища данных, и инструментария администратора БД, предназначенного для проектирования прикладных ИАС посредством создания многомерных информационных моделей в аналитических метаданных и обеспечивающего возможность организации многомерного анализа данных над хранилищем произвольной структуры. При этом основная функциональность многомерного анализа выносится в отдельное ядро системы оперативной аналитической обработки, к которому в составе проектируемых прикладных ИАС могут подключаться внешние модули интеллектуального анализа данных.

4. Отработана технология построения прикладных ИАС на основе разработанных программных средств.

На защиту выносятся следующие результаты.

1. Механизм "виртуальной звезды", определяющий структуру и порядок использования информационных моделей для извлечения многомерной информации из реляционных БД произвольной структуры.
2. Принцип организации аналитических метаданных для хранения многомерных информационных моделей в реляционных БД, являющихся содержательным наполнением прикладных ИАС.
3. Принципы организации инструментальной системы оперативной аналитической обработки данных в рамках программного комплекса, являющегося инструментальной оболочкой для построения корпоративных ИАС.
4. Технология построения прикладных ИАС на основе разработанных программных средств.

Практическая ценность результатов работы заключается в решении актуальной народнохозяйственной проблемы обеспечения информационной поддержки принятия решений на основе накопленных данных. Это достигнуто за счет разработки инструментальной системы, полностью соответствующей требованиям к продуктам класса OLAP (оперативной аналитической обработки) и обеспечивающей:

1) автоматизацию создания многомерных информационных моделей над реляционными хранилищами данных произвольной структуры и построение на их основе прикладных ИАС для поддержки принятия решений;
2) расширение функциональных возможностей строящихся ИАС с помощью подключения к ядру системы модулей интеллектуального анализа;
3) представление результатов аналитической обработки данных лицам, принимающим решения, в гибко настраиваемом формате табличных, графических и картографических отчетов.

Практическое применение разработанных подходов и средств открывает перспективу использования отечественной технологии многомерного анализа данных в задачах управления и проектирования в сферах промышленности, энергетики, отраслевой и региональной информатизации.

Реализация результатов работы. На основе предложенного способа организации инструментальной оболочки для автоматизации проектирования и для последующей эксплуатации прикладных ИАС в рамках программного комплекса ИнфоВизор разработана инструментальная система оперативной аналитической обработки данных. С ее помощью был построен ряд прикладных ИАС в топливно-энергетической сфере (хранилище данных по учету газа, реестр потребителей газа, БД по учету задолженности потребителей компании Межрегионгаз), электроэнергетике (БД ИПС Уренских электрических сетей) и региональной информатизации (реестры юридических, физических лиц, а также хранилище данных по социально-экономическому

положению Ивановской области). Практическая ценность работы подтверждается актами внедрения.

Апробация работы. Основные результаты работы докладывались и обсуждались: на II Республиканской электронной научной конференции "Современные проблемы информатизации" (Воронеж, 1997); на конференции "Создание и развитие информационной среды вуза: состояние и перспективы" (Ивановская государственная архитектурно-строительная академия, Иваново, 1997); на международной научно-технической конференции "VIII Бернардосовские чтения" (Ивановский государственный энергетический университет, Иваново, 1997); на III Международной электронной научной конференции "Современные проблемы информатизации" (Воронеж, 1998); на научно-практическом семинаре "Новые информационные технологии" (Московский государственный институт электроники и математики, Москва, 1998).

Публикации. По теме диссертации опубликовано 11 печатных работ.

Структура и объем работы. Диссертация состоит из введения, пяти глав и заключения на 150 страницах, а также списка литературы из 105 наименований; содержит 21 рисунок, 8 таблиц и 75 страниц приложений.

СОДЕРЖАНИЕ РАБОТЫ

Во введении определены цели и задачи исследования, обоснована актуальность выбранной темы, сформулированы новые научные результаты, перечислены основные положения, выносимые на защиту.

В первой главе выполнен обзор современных концепций хранения и анализа корпоративных данных. Их разработка в последние несколько лет знаменует собой качественный скачок в развитии информационных систем, вызванный стремлением пользователя получить максимальную отдачу от накапливаемых массивов данных. Рассмотрены концепции: 1) хранилищ данных, 2) оперативной аналитической обработки (многомерного анализа) и 3) интеллектуального анализа данных.

Вообще говоря, информационно-аналитические системы для поддержки принятия решений подразделяются на статические и динамические. Статические СППР, создаваемые в расчете на непосредственное использование лицами, принимающими решения, очень часто оказываются чрезвычайно просты в применении, но жестко ограничены в функциональности. Такие информационные системы руководителя (ИСР) содержат предопределенные множества запросов и, будучи достаточными для повседневного обзора, неспособны ответить на все вопросы к данным, которые могут возникнуть при принятии решений. Поэтому их внешняя простота оборачивается потерей гибкости. Динамические же СППР ориентированы на обработку нерегламентированных, неожиданных запросов к данным. Работа аналитиков с этими системами заключается в интерактивной последовательности

формирования запросов и изучения их результатов, каждый из которых может породить потребность в новой серии запросов.

Поддержка принятия решений на основе накопленных данных с помощью динамических СППР может выполняться в трех базовых сферах.

1. Сфера детализированных данных. Это область действия большинства систем, нацеленных на поиск информации.
2. Сфера агрегированных показателей. Комплексный взгляд на собранную в хранилище данных информацию, ее обобщение и агрегация, гиперкубическое представление и многомерный анализ являются задачами систем оперативной аналитической обработки данных (OLAP).
3. Сфера закономерностей. Интеллектуальная обработка производится методами интеллектуального анализа данных (ИАД, Data Mining), главными задачами которых являются поиск функциональных и логических закономерностей в накопленной информации, построение моделей и правил, которые объясняют найденные аномалии и (с определенной вероятностью) прогнозируют развитие процессов.

Наиболее интересной в этом ряду является концепция оперативной аналитической обработки данных (On-Line Analytical Processing, OLAP), которую предложил в 1993 году идеолог реляционной модели Э. Ф. Кодд. Основные положения концепции OLAP были выражены в 12 требованиях к средствам оперативной аналитической обработки данных, по степени приближения к которым было предложено оценивать продукты данного класса. Главной идеей концепции является обеспечение многомерного концептуального представления информации о предметной области как наиболее естественного с точки зрения профессионала-аналитика. Многомерное концептуальное представление имеет вид множественной перспективы из независимых измерений, вдоль которых могут быть проанализированы определенные совокупности данных. Одновременный анализ по нескольким измерениям данных определяется как многомерный анализ. Каждое измерение включает направления консолидации данных, состоящие из серий последовательных уровней обобщения, где каждый вышестоящий уровень соответствует более общей степени детализации данных по измерению.

Для обеспечения многомерного анализа данных можно или ориентироваться на специальные многомерные СУБД, или оставаться в рамках реляционных технологий. Во втором случае принято собирать заранее агрегированные данные в БД специального вида, построенные по схемам "звезды" или "снежинки". Они предусматривают наличие длинных и "узких" таблиц фактов с численными значениями ячеек гиперкуба и сравнительно небольших и "широких" таблиц измерений.

Но создание для многомерного анализа специальным образом организованных БД не всегда оправдано и возможно, особенно в случаях, когда данные накапливаются в ранее спроектированных БД, структура которых

поддерживается существующим программным обеспечением и не может быть произвольно изменена. В реальных БД информация часто распределена по разным таблицам и, возможно, существует не в явном виде значений полей, а требует оперативного выполнения расчетных операций.

В первой главе также выполнен обзор задач, стадий и методов интеллектуального анализа данных. Намечены пути сближения концепций OLAP и ИАД. Определены проблемы, затрудняющие использование существующих инструментальных средств в области поддержки принятия решений, - в частности, в задачах информатизации отраслевого управления (в энергетической, топливно-энергетической областях) и региональной информатизации, - и поставлены задачи выполняемой работы.

Во второй главе рассмотрен разработанный механизм "виртуальной звезды" для обеспечения многомерного концептуального представления данных, содержащихся в реляционных БД произвольной структуры, в виде агрегированной многомерной информации любой степени обобщения (рис. 1). Показано, что при анализе БД произвольной структуры существует необходимость наличия в многомерной информационной модели не только определений всех возможных информационных измерений (атрибутов) и уровней обобщения данных по каждому из них, но и формального описания правил получения значений, извлекаемых в процессе многомерного анализа, для разных подмножеств ячеек целевого гиперкуба. Для решения этой задачи предложено ввести в информационную модель понятия отрезка однородности и области однородности получения значений. Предложена следующая структура многомерной информационной модели:

$M = (A,L,C,R,N,S)$, где:

- $A = \{a_1, a_2, \ldots, a_{|A|}\}$ - множество атрибутов (элементарных измерений), в базисе которых находятся ячейки численных показателей; каждый атрибут a_i представляет некоторую характеристику, выбранную в качестве одной из осей многомерной модели; $|A|$ фиксированных значений (по одному для атрибута) однозначно задают ячейку в заданном пространстве;
- $L = \{l_1, l_2, \ldots, l_{|L|}\}$ - множество уровней обобщения данных по атрибутам многомерной модели; каждому атрибуту a_i ставится в соответствие непустое подмножество уровней обобщения $L_i \subset L$; при этом с каждым l_j связано множество экземпляров этого уровня обобщения $E_j = \{e_{j1}, e_{j2}, \ldots, e_{j|E_j|}\}$ (точек на координатной оси уровня обобщения);
- $C = \{c_1, c_2, \ldots, c_{|C|}\}$ - множество отрезков однородности получения значений показателей;

- $R = \{r_1, r_2, \ldots, r_{|R|}\}$ - множество областей однородности получения значений показателей;
- $N = \{n_1, n_2, \ldots, n_{|N|}\}$ - множество, определяющее необходимые условия построения запросов на получение значений показателей ячеек областей однородности многомерной информационной модели;
- $S = \{s_1, s_2, \ldots, s_{|S|}\}$ - множество, определяющее достаточные условия построения запросов на получение значений показателей ячеек областей однородности многомерной информационной модели.

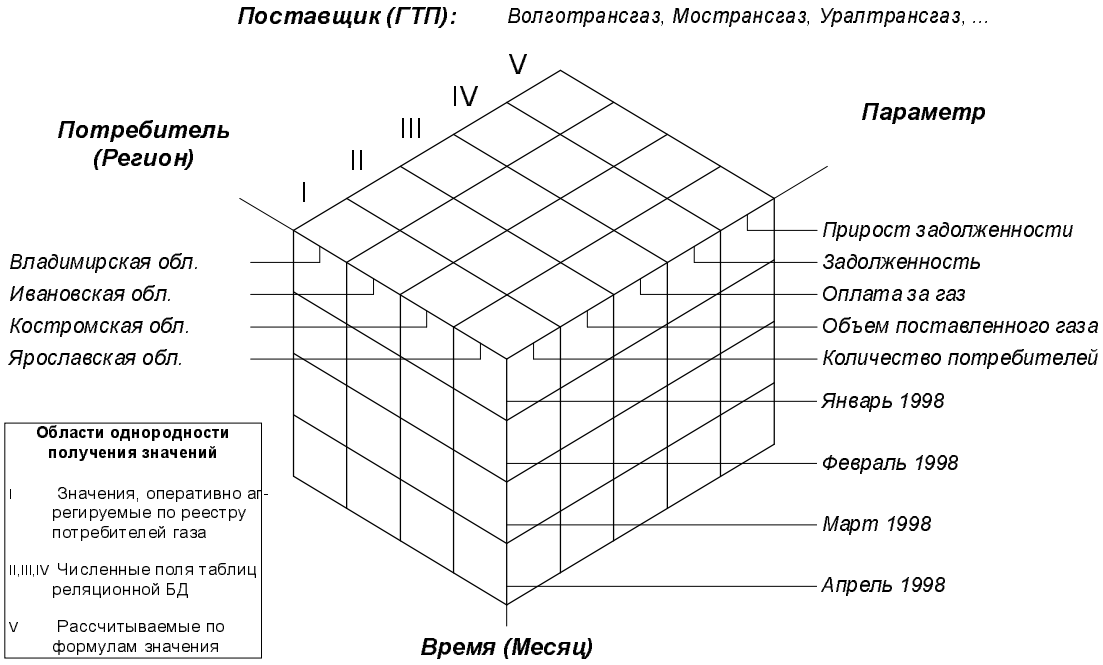

Рис. 1. Пример 4-мерной информационной модели для анализа реляционной БД (хранилища данных по учету газа).

Методы получения значений показателей, ячейки которых имеют в качестве координат по некоторому уровню обобщения l_j разные его экземпляры, могут отличаться, так как необходимые для расчета значений этих показателей данные могут находиться в разных таблицах исходной реляционной БД. Поэтому следует предусмотреть возможность разделения множества экземпляров E_j каждого уровня обобщения l_j (где $1 \le j \le |L|$) на непересекающиеся подмножества E_{jh}, составляющие множество групп однородности $G_j = \{E_{j1}, E_{j2}, \ldots, E_{j|G_j|}\}$ уровня обобщения l_j (здесь $1 \le h \le |G_j|$):

$$\bigcup_{h=1}^{|G_j|} E_{jh} = E_j, \quad \text{где} \quad \forall h_1 \ne h_2 \left(1 \le h_1, h_2 \le |G_j|\right) \; E_{jh_1} \cap E_{jh_2} = \varnothing.$$

На каждом подмножестве экземпляров $E_{jh} \in G_j$ определяется непустое множество отрезков однородности C_{jh}, задающих множество вариантов правил извлечения значений показателей ячеек многомерной информационной модели, имеющих в качестве координаты по оси уровня обобщения l_j любой экземпляр $e_{jt} \in E_{jh}$ $(1 \le t \le |E_j|)$:

$$\bigcup_{h=1}^{|G_j|} C_{jh} = C_j, \quad \text{где} \quad \forall h_1 \ne h_2 \left(1 \le h_1, h_2 \le |G_j|\right) \; C_{jh_1} \cap C_{jh_2} = \varnothing.$$

Таким образом, каждому уровню обобщения l_j соответствует непустое множество отрезков однородности C_j:

$\forall l_j \in L \;\; \exists C_j \ne \varnothing$, такое, что:

$$\bigcup_{j=1}^{|L|} C_j = C \quad \text{и} \quad \forall j_1 \ne j_2 \; (1 \le j_1, j_2 \le |L|) \;\; C_{j_1} \cap C_{j_2} = \varnothing.$$

При этом каждый отрезок однородности $c_k \in C_{jh}$ описывается группой однородности - подмножеством экземпляров $E_{jh} \in G_j$ соответствующего уровня обобщения l_j - и правилом извлечения значений показателей гиперкуба для этого отрезка q_k:

$$\forall c_k \in C_{jh} \;\; c_k = \left(E_{jh}, q_k\right).$$

Множество V комбинаций групп однородности в базисе $A = \left\{a_1, a_2, \ldots, a_{|A|}\right\}$ атрибутов является прямым произведением объединений множеств групп однородности уровней обобщения, связанных с каждым из атрибутов:

$$V = \bigcup_{l_j \in L_1} G_j \times \bigcup_{l_j \in L_2} G_j \times \ldots \times \bigcup_{l_j \in L_{|A|}} G_j.$$

Каждая ячейка полного гиперкуба (расположенная в базисе всех $|A|$ атрибутов), имеющая координаты $\left(e_{j_1 t_1}, e_{j_2 t_2}, \ldots, e_{j_{|A|} t_{|A|}}\right)$, принадлежит области, характеризуемой вектором групп однородности $v_b \in V$:

$$v_b = \left(E_{j_1 h_1}, E_{j_2 h_2}, \ldots, E_{j_{|A|} h_{|A|}}\right),$$

причем каждая i-я координата ячейки является элементом i-й проекции v_b:

$$\forall a_i \in A \;\; e_{j_i t_i} \in \Pi p_i v_b.$$

При рассмотрении проекции полного гиперкуба $\tilde{A} \subseteq A$ $(\tilde{A} \ne \varnothing)$ ячейка $\left(e_{j_{i_1} t_{i_1}}, e_{j_{i_2} t_{i_2}}, \ldots, e_{j_{i_{|\tilde{A}|}} t_{i_{|\tilde{A}|}}}\right)$ принадлежит области, характеризуемой некоторой

проекцией $\tilde{v}_{b'}$, одинаковой для каждого элемента некоторого подмножества векторов групп однородности $\{v_{b_1}, v_{b_2}, \ldots, v_{b_n}\} \subseteq V$, на $\tilde{A}$ атрибутов:

$\tilde{v}_{b'} = \left(E_{j_{i_1} h_{i_1}}, E_{j_{i_2} h_{i_2}}, \ldots, E_{j_{|\tilde{A}|} h_{|\tilde{A}|}}\right)$, такой, что: $\forall a_{i_x} \in \tilde{A} \quad e_{j_{i_x} t_{i_x}} \in Пр_x \tilde{v}_{b'}$.

Поскольку все $|\tilde{A}|$ атрибутов-измерений взаимно перпендикулярны, область каждого вектора $\tilde{v}_{b'}$ содержит $\prod_{a_{i_x} \in \tilde{A}} |Пр_x \tilde{v}_{b'}|$ ячеек. Логично предположить, что для всех этих ячеек правило извлечения значений показателей из реляционной БД должно быть одним и тем же - более того, одно правило может оказаться справедливым для ячеек областей нескольких векторов.

Поскольку на одной группе однородности может быть определено несколько отрезков однородности, каждый вектор $v_b \in V$ может, в общем случае, иметь несколько вариантов реализации через разные вектора комбинаций отрезков однородности $w_d = \left(c_{k_1}, c_{k_2}, \ldots, c_{k_{|A|}}\right)$, которые являются элементами множества W (определяемого аналогично множеству векторов комбинаций групп однородности V):

$$W = \bigcup_{l_j \in L_1} C_j \times \bigcup_{l_j \in L_2} C_j \times \ldots \times \bigcup_{l_j \in L_{|A|}} C_j .$$

При этом вектору каждой комбинации $v_b \in V$ будет соответствовать множество векторов комбинаций отрезков однородности $W_b \subseteq W$:

$W_b = C_{j_1 h_1} \times C_{j_2 h_2} \times \ldots \times C_{j_{|A|} h_{|A|}}$, где

$\forall a_i \in A \quad Пр_1(Пр_i W_b) = \{Пр_i v_b\}$.

Для проекции полного гиперкуба на $\tilde{A} \subseteq A$ атрибутов, аналогично, вектору каждой комбинации $\tilde{v}_{b'}$ соответствует множество векторов комбинаций отрезков однородности $\tilde{W}_{b'}$:

$\tilde{W}_{b'} = C_{j_{i_1} h_{i_1}} \times C_{j_{i_2} h_{i_2}} \times \ldots \times C_{j_{|\tilde{A}|} h_{|\tilde{A}|}}$, где

$\forall a_{i_x} \in \tilde{A} \quad Пр_1(Пр_x \tilde{W}_{b'}) = \{Пр_x \tilde{v}_{b'}\}$.

Основой определения пространственного понятия области однородности получения значений показателей являются понятия группы и отрезка однородности получения значений. В простейшем случае областью однородности является фрагмент гиперкуба, проекциями которого на каждое измерение являются множества экземпляров, принадлежащие одной группе однородности получения значений - вектор комбинации групп однородности (его компоненты соответствуют атрибутам многомерной модели и принимают значения идентификаторов соответствующих групп однородно-

сти для каждого атрибута). Введение понятия области однородности r_u ($1 \le u \le |R|$) позволяет поставить в соответствие множеству векторов комбинаций отрезков однородности один метод расчета значений показателей в ячейках гиперкуба (p_u). Локализация областей однородности в многомерной информационной модели производится с помощью множеств необходимых (N) и достаточных (S) условий построения запросов:

1) $N = \{n_1, n_2, \ldots, n_{|N|}\}$ - множество пар $n_v = (a_i, r_u)$, $a_i \in A$, $r_u \in R$, то есть соответствие между множествами A и R ($N \subseteq A \times R$);

2) $S = \{s_1, s_2, \ldots, s_{|S|}\}$ - множество пар $s_w = (c_k, r_u)$, $c_k \in C$, $r_u \in R$, то есть соответствие между множествами C и R ($S \subseteq C \times R$).

С использованим информации этих множеств область однородности может быть описана как $r_u = (A_u, C_u, p_u)$, где:

- $A_u \subseteq A$ - подмножество атрибутов, выбор которых в запросе обязателен для существования непустых значений в ячейках области однородности:

  $A_u \subseteq \tilde{A}$

  (необходимое условие построения запросов); каждое подмножество атрибутов A_u может быть получено из множества N следующим образом:

  $\forall r_u \in R \quad A_u = \Pi p_1(N \cap A \times \{r_u\})$;

- $C_u \subseteq C$ - подмножество отрезков однородности, такое, что любая ячейка многомерной информационной модели $\left(e_{j_{i_1} t_{i_1}}, e_{j_{i_2} t_{i_2}}, \ldots, e_{j_{i_{|\tilde{A}|}} t_{i_{|\tilde{A}|}}}\right)$ может быть отнесена к области однородности r_u только тогда, когда:

  $\forall a_i \in \tilde{A} \quad \exists!\ c_{k_i} \in C_u$, такое, что $c_{k_i} \in C_{j_i h_i}$ и $e_{j_i t_i} \in E_{j_i h_i}$

  (достаточное условие построения запросов - имеется в виду, что достаточным для построения запроса является выполнение этого условия одновременно с предыдущим); каждое подмножество отрезков однородности C_u может быть получено из множества S следующим образом:

  $\forall r_u \in R \quad C_u = \Pi p_1(S \cap C \times \{r_u\})$;

  таким образом, для ячейки $\left(e_{j_{i_1} t_{i_1}}, e_{j_{i_2} t_{i_2}}, \ldots, e_{j_{i_{|\tilde{A}|}} t_{i_{|\tilde{A}|}}}\right)$, принадлежащей области вектора комбинации отрезков однородности $\tilde{w}_{d'}$, достаточное условие выполняется в той области однородности r_u, в которой множеству C_u принадлежат все компоненты вектора $\tilde{w}_{d'}$;

- p_u - правило получения значений показателей рассматриваемых ячеек, по которому должен быть построен SQL-запрос к реляционной БД при

выполнении обоих рассмотренных условий; оно имеет вид $p_u = (p_what_u, p_from_u, p_where_u, p_groupby_u)$, т. е. определяет дополнения к секциям SQL запросов, формируемых по шаблону:

SELECT <...> FROM <...> WHERE <...> GROUP BY <...>;

аналогично, правила построения запросов, связанные с отрезками однородности $c_k \in C$, имеют вид $q_{k_j} = (q_from_{k_j}, q_where_{k_j}, q_groupby_{k_j})$.

Предложенная организация информационной модели предоставляет требуемую для задачи многомерного концептуального представления данных реляционных БД сложной структуры гибкость, но содержит возможность возникновения двусмысленностей при построении SQL-запросов на получение значений показателей ячеек. Во избежание этого должны выполняться ограничения целостности.

1. Для каждой ячейки многомерной информационной модели в базисе любого подмножества атрибутов $\tilde{A} \subseteq A$, представленных любыми возможными для них уровнями обобщения $\tilde{L} \subseteq L$, может существовать не более чем одна область однородности r_u, для которой в этой ячейке одновременно выполняются необходимое и достаточное условия построения запросов. То есть ячейка может либо не принадлежать ни одной области однородности получения значений (тогда ее значение считается пустым), либо принадлежать только одной области r_u.
2. Для каждой области однородности r_u подмножество отрезков однородности C_u не должно включать более одного отрезка, заданного на любой одной группе однородности E_{jh}:

$$\forall r_u \in R \quad \forall E_{jh} \in \bigcup_{j=1}^{|L|} G_j \quad \left|\left\{c_k \middle| c_k \in C_u \text{ и } \Pi p_1 c_k = E_{jh}\right\}\right| \le 1.$$

Благодаря введению в информационную модель рассмотренных понятий, предложенная концепция виртуальной звезды позволяет обеспечить оперативную аналитическую обработку данных с большим удобством, чем повсеместно используемая стратегия организации БД по принципу схемы звезды или снежинки. Концепция виртуальной звезды позволяет во многом преодолеть ограничения традиционных подходов, благодаря тому, что:

1) не требуется создание БД специальной структуры для целей многомерного анализа, то есть не занимается лишнее дисковое пространство;
2) подход непосредственного получения информации из исходных БД позволяет всегда получать информацию, актуальную на момент запроса, а не на момент последнего обновления таблиц фактов схем типа звезды;
3) любая модификация аналитической модели сводится только к изменению слоя метаданных, не затрагивая структуры исходных БД; более того,

на основе данных одной базы может быть определено несколько информационных моделей (виртуальных звезд).

Естественно, любая построенная по схеме звезды или снежинки БД может быть как частный случай описана через модель виртуальной звезды.

Аналитический запрос к многомерной информационной модели имеет следующий вид.

1. Подмножество атрибутов модели $\tilde{A} \subseteq A$. Оно задает либо полный гиперкуб ($\tilde{A} = A$), либо любую его проекцию ($\tilde{A} \subset A$, $\tilde{A} \neq \varnothing$).
2. Подмножество уровней обобщения $\tilde{L} \subseteq L$, задающее для каждого атрибута $a_i \in \tilde{A}$ соответствующий уровень обобщения данных $l_{j_i} \in L_i$.
3. Непустые подмножества экземпляров $\hat{E}_{j_i} \subseteq E_{j_i}$ для выбранных в запросе уровней обобщения $l_{j_i} \in \tilde{L}$, составляющие множество $\hat{F} = \bigcup_{l_{j_i} \in \tilde{L}} \left\{ \hat{E}_{j_i} \right\}$.

Предложен алгоритм выполнения аналитических запросов к многомерному концептуальному представлению реляционной БД, реализованному в форме многомерной информационной модели. Алгоритм транслирует аналитический запрос к многомерному представлению в серию SQL запросов к исходной БД и объединяет полученные результаты в аналог таблицы фактов схемы звезды, вследствие чего предложенный механизм и был назван механизмом "виртуальной звезды".

В третьей главе предлагается структура аналитических метаданных, в которой могут описываться многомерные информационные модели в рамках предложенного подхода. Аналитические метаданные описывают логическую структуру информационной модели базы данных в виде многомерного гиперкуба или в виде нескольких гиперкубов (виртуальных звезд). Имеющая такие метаданные реляционная БД может использоваться в качестве содержательного наполнения ИАС, обеспечивающей многомерный анализ данных. Системные таблицы аналитических метаданных соответствуют множествам, определяющим многомерную модель: 1) таблица атрибутов; 2) таблица уровней обобщения; 3) таблица отрезков однородности получения значений; 4) таблица областей однородности получения значений; 5) таблица необходимых условий построения запросов; 6) таблица достаточных условий построения запросов.

При этом предельное количество моделей и предельная сложность каждой модели не ограничиваются. Таким образом, проектирование прикладных ИАС при условии разработки основанного на предложенных подходах программного обеспечения должно заключаться во введении в структуру БД подобным образом организованных аналитических метаданных.

В четвертой главе поставлены задачи разработки инструментальной системы оперативной аналитической обработки данных в рамках единого

программного комплекса для проектирования, обслуживания и использования корпоративных ИАС, основанной на предложенном подходе. Дано общее описание программного комплекса ИнфоВизор, созданного в соответствии с этими требованиями. Приведена структура корпоративной ИАС, формируемой на основе данного программного комплекса (рис. 2). Рассмотрены инструментальные системы, входящие в программный комплекс, показаны их подсистемы и модули, перечислены обеспечиваемые функции.

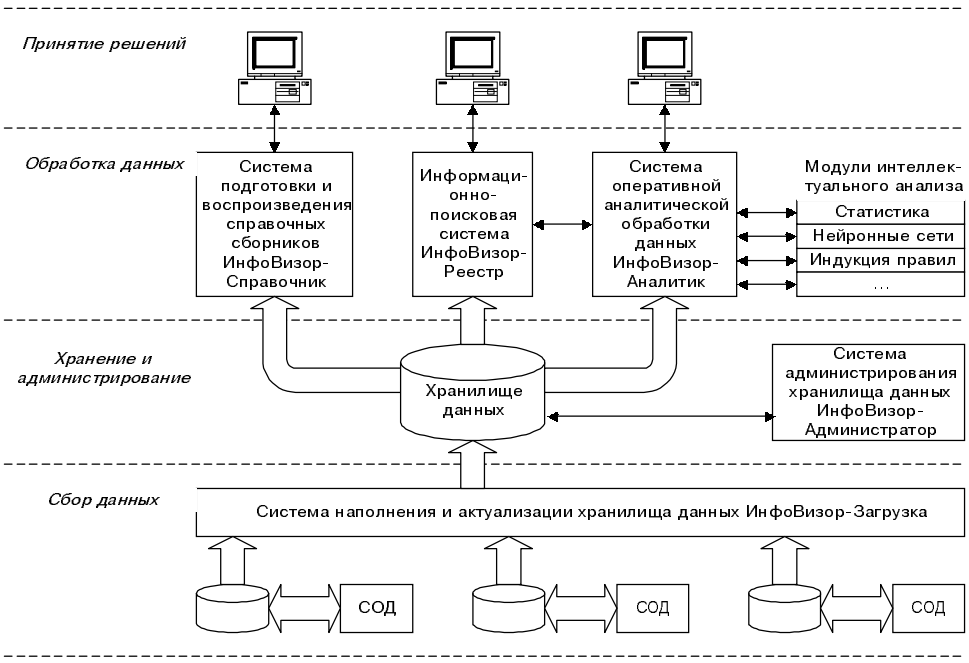

Рис. 2. Общая структура программного комплекса ИнфоВизор.

В пятой главе рассмотрена структура и особенности реализации системы ИнфоВизор-Аналитик, предназначенной для оперативной аналитической обработки реляционных данных, и ИнфоВизор-Администратор, целью которой является автоматизация проектирования прикладных ИАС с помощью создания, редактирования и проверки целостности информационных моделей, составляющих навигационные и аналитические метаданные. Рассмотрена объектная библиотека функционального ядра системы оперативной аналитической обработки данных, предложен механизм динамического подключения к системе расширяющих ее функциональность модулей интеллектуального анализа через функциональное ядро. Рассмотрена технология формирования прикладных ИАС с помощью разработанных программных средств. Приведены примеры ИАС, построенных в топливно-энергетической сфере (компания Межрегионгаз), электроэнергетике (Уренские электрические сети) и региональной информатизации (Администрация Ивановской области).

В заключении подведены итоги работы. Перечислены результаты и выводы работы, определены направления дальнейших исследований.

ОСНОВНЫЕ РЕЗУЛЬТАТЫ И ВЫВОДЫ РАБОТЫ

1. Выполнен обзор современных концепций и методов хранения и анализа корпоративных данных.
2. Предложен механизм "виртуальной звезды" для обеспечения многомерного концептуального представления данных, содержащихся в реляционных БД, в виде агрегированной многомерной информации любой степени обобщения. Разработана структура многомерных информационных моделей, транслирующих аналитические запросы к многомерному концептуальному представлению в серии SQL-запросов к исходным БД.
3. Разработана оригинальная структура аналитических метаданных для представления многомерных информационных моделей в реляционных БД, являющихся содержательным наполнением создаваемых ИАС.
4. Разработана структура инструментальной системы оперативной аналитической обработки данных в рамках программного комплекса, являющегося инструментальной оболочкой для построения корпоративных ИАС.
5. Предложен способ реализации функционального ядра инструментальной системы оперативной аналитической обработки данных в виде независимой объектной библиотеки, допускающий расширение функциональности формируемых прикладных ИАС с помощью подключения к ядру системы оперативной аналитической обработки данных внешних модулей интеллектуального анализа.
6. Разработаны системы ИнфоВизор-Администратор и ИнфоВизор-Аналитик в рамках программного комплекса ИнфоВизор. Система ИнфоВизор-Администратор, реализующая создание, редактирование и проверку логической целостности многомерных информационных моделей, автоматизирует проектирование ИАС, ориентированных на конкретную область применения. Система ИнфоВизор-Аналитик служит основой для создаваемых прикладных ИАС и полностью удовлетворяет требованиям Э. Ф. Кодда к программным средствам класса OLAP (оперативной аналитической обработки данных).
7. Отработана технология построения прикладных ИАС на основе разработанного инструментария. Сформированы прикладные ИАС, предназначенные для многомерного анализа существующих хранилищ данных в топливно-энергетической сфере, электроэнергетике, в сфере региональной власти.

Содержание диссертации отражено в следующих работах.

1. Кокин В. М., Щавелев Л. В. Принципы разработки системы оперативной аналитической обработки информации для поддержки принятия решений // Современные проблемы информатизации: Тез. докл. II Респ. электрон. науч. конф. - Воронеж: Изд-во Воронежского пед. ун-та, 1997. - С. 68-69.

2. Кокин В. М., Щавелев Л. В., Чечулин О. Н. Реализация средств статистического анализа и моделирования в виде библиотеки компонентов для среды визуального программирования // Современные проблемы информатизации: Тез. докл. II Респ. электронной науч. конф. - Воронеж: Изд-во Воронежского пед. ун-та, 1997. - С. 78-79.
3. Щавелев Л. В. Оперативная аналитическая обработка данных для поддержки принятия управленческих решений // Создание и развитие информационной среды вуза: состояние и перспективы. Сб. ст. к конф. / Иван. гос. архитектурно-строит. акад. - Иваново, 1997. - С. 279-283.
4. Концепция корпоративной информационно-поисковой системы / Игнатьева И. Б., Коровкин С. Д., Левенец И. А., Ратманова И. Д., Щавелев Л. В. // VIII Бернардосовские чтения: Тез. докл. междунар. науч.-техн. конф. / Ивановский гос. энергетич. ун-т. - Иваново, 1997. - С. 42.
5. Щавелев Л. В. Разработка средств многомерного анализа оперативных данных для поддержки принятия решений // VIII Бернардосовские чтения: Тез. докл. междунар. науч.-техн. конф. / Ивановский гос. энергетич. ун-т. - Иваново, 1997. - С. 55.
6. Левенец И. А., Щавелев Л. В., Ратманова И. Д. Поли-иерархическая модель представления реляционной структуры хранилищ данных // Новые информационные технологии: материалы науч.-практ. семинара / Моск. гос. ин-т электроники и математики. - М.: 1998. - С. 49-56.
7. Щавелев Л. В., Коровкин С. Д., Левенец И. А. Агрегация и интеллектуальный анализ информации хранилищ данных // Новые информационные технологии: материалы науч.-практ. семинара / Моск. гос. ин-т электроники и математики. - М.: 1998. - С. 108-118.
8. Левенец И. А., Щавелев Л. В. Виды метаданных в корпоративных информационных системах // Современные проблемы информатизации: Тез. докл. III Междунар. электронной науч. конф. - Воронеж: Изд-во Воронежского пед. ун-та, 1998. - С. 109-110.
9. Щавелев Л. В. Современные методы интеллектуального анализа для извлечения знаний из баз данных // Современные проблемы информатизации: Тез. докл. III Междунар. электронной науч. конф. - Воронеж: Изд-во Воронежского пед. ун-та, 1998. - С. 132-133.
10. Щавелев Л. В. Концепция виртуальной звезды для агрегации и многомерного анализа реляционных данных // Современные проблемы информатизации: Тез. докл. III Междунар. электронной науч. конф. - Воронеж: Изд-во Воронежского пед. ун-та, 1998. - С. 143-144.
11. Решение проблемы комплексного оперативного анализа информации хранилищ данных / Коровкин С. Д., Левенец И. А., Ратманова И. Д., Старых В. А., Щавелев Л. В. // СУБД. - 1997. - № 5-6. - С. 47-51.