Решение проблемы комплексного оперативного анализа информации хранилищ данных

С. Д. Коровкин, И. А. Левенец, И. Д. Ратманова,
В. А. Старых, Л. В. Щавелёв

(СУБД. - 1997. - № 5-6)

Управление любой сложной социально-экономической системой - от предприятия до региона или отрасли - весьма затруднительно без обратной связи, которая заключается в отслеживании и анализе данных, отражающих состояние этой системы и ситуацию вокруг нее. Постоянная доступность актуальной информации дает возможность оценить текущее положение дел, а обзор изменения конкретных характеристик во времени позволяет обнаружить тенденции развития системы и сделать выводы о том, что ожидает ее в будущем. Таким образом, обладая всей полнотой сведений о состоянии системы и ее элементов в статике и динамике, управляющий персонал может принимать грамотные решения по применению мер регулирования. Такое управление основано на знании и потому более эффективно, чем принятие важных решений вслепую.

На современном этапе проблемы ведения баз данных (БД) оперативной информации на разных уровнях управления перестали быть непреодолимыми. Более того, разработана концепция Хранилищ Данных - архитектура построения корпоративных информационных систем, получившая развитие вследствие желания конечных пользователей иметь непосредственный единообразный доступ к необходимым им данным, источники происхождения которых организационно и территориально распределены, а анализ которых может способствовать принятию решений [10]. Билл Инмон, автор концепции Хранилищ Данных, определил их как "предметно ориентированные, интегрированные, неизменчивые, поддерживающие хронологию наборы данных, организованные с целью поддержки управления", призванные выступать в роли "единого и единственного источника истины", обеспечивающего менеджеров и аналитиков достоверной информацией, необходимой для оперативного анализа и принятия решений [8]. Ричард Хакаторн, другой основоположник этой концепции, писал, что цель Хранилищ Данных - обеспечить для организации "единый образ существующей реальности" [7].

В этом контексте наиболее актуальной проблемой является обеспечение интегрированного взгляда на сложный объект управления в целом, комплексного анализа собранных о нем сведений и извлечения из огромного объема детализированных данных некоторой полезной информации - знаний о закономерностях его развития.

Поддержка принятия управленческих решений на основе накопленной информации может осуществляться в трех основных областях [9].

- **Область детализированных данных**. Это сфера действия большинства транзакционных систем (OLTP), нацеленных на *поиск информации*. В большинстве случаев реляционные СУБД отлично справляются с возникающими здесь задачами.
- **Область 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) [5, 4]. Здесь можно или ориентироваться на специальные многомерные СУБД, или (что, как правило, предпочтительнее) оставаться в рамках реляционных технологий. Во втором случае заранее агрегированные данные могут собираться в БД звездообразного вида, либо агрегация информации может производиться на лету в процессе сканирования детализированных таблиц реляционной БД.

- **Область закономерностей**. *Интеллектуальная обработка* производится методами интеллектуального анализа данных (ИАД, Data Mining) [3], главными задачами которых являются поиск функциональных и логических закономерностей в накопленной информации, построение моделей и правил, которые объясняют найденные аномалии и/или (с определенной вероятностью) прогнозируют развитие рассматриваемых процессов.

Программный комплекс ИнфоВизор является инструментальной оболочкой для создания информационно-аналитических систем (ИАС), покрывающих все перечисленные типы обработки данных. Основой представления информации о структуре используемых совместно с инструментарием ИнфоВизора реляционных БД являются метаданные - *навигационные* (предназначенные для задач поиска и модификации данных), *аналитические* (созданные для реализации многомерного анализа накопленных сведений) и *операционные* (обеспечивающие наполнение хранилища данных из внешних информационных источников в автоматическом режиме по заранее утвержденным сценариям). Благодаря наличию таких метаданных практически любая непротиворечиво спроектированная БД может быть адаптирована к использованию в качестве содержательного наполнения ИАС на основе этого инструментария. Таким образом достигается инвариантность программного обеспечения по отношению к структуре представления исходной информации.

Являясь большим программным комплексом, ИнфоВизор состоит из ряда программных продуктов, которые могут использоваться автономно или совместно, взаимодействуя и дополняя друг друга.

ИнфоВизор-Загрузка. Для наполнения централизованного хранилища данных и постоянного поддержания его в актуальном состоянии почти всегда требуется регулярно обновлять его содержание свежей информацией из информационных источников - баз данных уровня подразделения или проблемно-ориентированных баз. Данный продукт позволяет создавать и автоматически выполнять сценарии обновления информации хранилища. Администратор базы данных формирует регламент загрузки новой информации из источников (*операционные метаданные*), после чего запланированные действия автоматически выполняются в нужные моменты времени. Таким образом, предлагаемый инструмент дает возможность решить задачу интеграции и согласования разрозненных данных.

ИнфоВизор-Реестр. Этот продукт предназначен для обслуживания и использования корпоративных информационно-поисковых систем в рамках хранилища данных. По мнению авторов, наиболее адекватно информационные сущности описываются с помощью иерархии объектов. Не случайно именно иерархическое представление принято в моделировании для анализа структуры сложных систем. В свое время такой подход был успешно реализован в отечественной СУБД ИНЕС [2], а сегодня используется в СУБД Datacom/DB

компании Applied Data Research, Inc. (ADR) и Adabas компании Software AG [1], а также в некоторых постреляционных СУБД. В данных системах логическое представление информации в виде иерархической модели соответствует физическому методу хранения данных. Однако, по многим причинам, практически всегда кажется более предпочтительным вести Хранилище Данных с помощью серверов реляционных СУБД: во-первых, в настоящее время ряд ведущих производителей этих серверов (Informix, Oracle, Sybase и др.) значительно вырвались вперед по сравнению со своими конкурентами по уровню производительности машин баз данных; во-вторых, Хранилища Данных практически всегда создаются в качестве надстройки над множеством существующих систем обработки данных (СОД) [4], которые производят обслуживание проблемно-ориентированных (ведомственных) реляционных баз данных с помощью унаследованных систем (legacy systems); наконец, язык SQL для реляционной модели является единственным языком манипулирования данными, имеющим общепринятый мировой стандарт. Поэтому было принято решение, оставаясь на реляционной платформе, воспользоваться преимуществами иерархического и сетевого подхода в процессах семантического моделирования структуры хранилища данных и навигации по находящимся там информационным объектам.

Решение данной проблемы может быть получено путем введения слоя *навигационных метаданных* между физическим хранилищем и программными приложениями его обработки, который обеспечивал бы оперативную трансляцию реляционных данных в термины более естественной семантической модели. Вид этой семантической модели наиболее адекватно отражается метафорой дерева смысловых атрибутов, то есть аналогом традиционной иерархической модели хранения данных. Можно сказать, что на новом витке спирали исторического развития мы в некоторой степени возвращаемся к незаслуженно забытым идеям иерархического и сетевого подхода к построению информационных систем.

Поскольку практически любая реляционная база данных может быть описана слоем *навигационных метаданных* и представлена как иерархия связанных друг с другом атрибутов информационных объектов, подсистемы ввода данных, корректировки и поиска информации не должны изменяться при модификации структуры базы данных или при изменении потребностей пользователей в подробностях получаемой информации.

ИнфоВизор-Аналитик. Если предыдущие инструменты имели дело с детализированными данными, то ИнфоВизор-Аналитик в первую очередь предназначен для комплексного многомерного анализа численной агрегированной информации, то есть является системой класса ROLAP - оперативной аналитической обработки реляционных данных.

Авторами разработан оригинальный механизм извлечения агрегированной многомерной информации любой степени обобщения из реляционных БД произвольной структуры. Идея системы ИнфоВизор-Аналитик заключается во введении в Хранилище Данных *аналитических метаданных*, которые содержат набор информационных моделей (*"виртуальных звезд"*), позволяющих извлекать из него обобщенную численную информацию в многомерном базисе измерений. За основу аналитических метаданных взято понятие многомерного гиперкуба. Осями гиперкуба (*атрибутами* информационной модели) могут быть любые способы консолидации данных

[5], представляемые справочниками экземпляров некоторых объектов (множеством точек оси). Одной из осей гиперкуба может являться время, что позволяет анализировать динамические процессы. Совокупность выбранных экземпляров от каждой оси однозначно характеризует ячейку гиперкуба в пространстве объявленных атрибутов. Каждая ячейка гиперкуба может содержать численное значение - *показатель*, - смысл которого определяется координатами ячейки - набором экземпляров, по одному от каждого атрибута.

Получение значений показателей производится по результатам SQL-запросов к реляционной БД, которые автоматически генерируются ядром системы, исходя из диалогового запроса пользователя-аналитика к информационной модели. При этом результирующими значениями показателей информационной модели могут быть:

1) значения полей численного типа БД (предварительно агрегированные значения);
2) функции агрегирования SQL (avg, count, max, min, sum, а также другие функции агрегирования, поддерживаемые SQL-сервером [6]);
3) формулы (SQL-выражения, имеющие на выходе одно численное значение).

Атрибут - это один из признаков показателя, который может быть представлен в виде оси информационной модели. В литературе его часто называют *измерением*. Но в предлагаемой нами концепции понятие измерения более сложно: на этапе обработки ранее извлеченной информации и формирования отчетов измерения могут быть не только получены непосредственно из атрибутов, но и являться группировками нескольких независимых осей - атрибутов. Это делается для снижения общего количества измерений гиперкуба ради удобства конечного пользователя (довольно трудно манипулировать абстрактными пространствами с высокой мерностью).

Для обеспечения возможности рассматривать показатели с разными степенями агрегации реализация каждого атрибута, в общем случае, может быть выполнена с помощью нескольких конкретных справочников экземпляров - различных *уровней обобщения атрибута*. Так, атрибуту "Местоположение" можно поставить в соответствие несколько реализаций - уровень детализированного справочника СОАТО, уровень городов и районов, уровень субъектов Российской Федерации и так далее. Атрибут "Время" может иметь уровни обобщения с точностью до года, квартала, месяца и так далее.

Результаты серии SQL-запросов, сформированной по каждому конкретному оперативному заказу аналитика (диалоговому запросу), имеют совершенно идентичный формат (набор полей идентификаторов экземпляров справочников выбранных уровней обобщения выбранных атрибутов плюс целевое поле показателя), поэтому могут быть составлены в одну временную таблицу (например, оператором UNION), которая в совокупности со справочниками уровней обобщения атрибутов может рассматриваться как самостоятельная база данных, имеющая *схему звезды* (Star Scheme) [4, 5]. Тот факт, что центральная таблица этой БД является временной, то есть оперативно создается и заполняется по требованию аналитика в ответ на его нерегламентированный запрос, является основанием наименования описанного метода - "Концепция виртуальной звезды". Таким образом, предлагаемая концепция является целостным законченным оригинальным способом реализации гибких информационно-аналитических систем класса OLAP. Система ИнфоВизор-Аналитик основывается именно на этой концепции.

Новизна предлагаемого подхода "виртуальной звезды" по сравнению с существующими механизмами, реализованными в продуктах OLAP, заключается в следующем.

1. Введение понятий группируемого справочника уровня обобщения атрибута и области однородности извлечения значений позволяет в рамках одной информационной модели обращаться к значениям трех типов: 1) предварительно агрегированным в полях численного типа, 2) полученным в результате оперативной агрегации в процессе выполнения запроса и 3) являющимся результатами расчета заданных формул. Таким образом, построенный пользователем диалоговый запрос может возвратить значения, принадлежащие ко всем этим типам, в рамках одного гиперкуба. При этом в процессе построения информационной модели можно достичь допустимого компромисса между быстротой доступа к заранее агрегированным значениям и актуальностью значений, агрегируемых оперативно. Даже для разных экземпляров одного справочника одной оси могут быть заданы различные методы доступа, исходя, например, из частоты обращения к соответствующим значениям.
2. Реализация гибкого механизма группировки атрибутов - элементарных измерений - в процессе рассмотрения сформированного гиперкуба в укрупненные измерения позволяет существенно понизить его мерность и облегчить восприятие конечным пользователем содержащейся в нем информации.

Разработанная авторами система ИнфоВизор-Аналитик реализует клиентскую часть описанной OLAP-системы и позволяет производить оперативный многомерный анализ численных данных реляционных БД произвольной структуры. Этот продукт доступен в варианте двухзвенной архитектуры "клиент-сервер" и в варианте для архитектуры *Intranet*. Установка intranet-версии продукта на web-сервере - может одновременно обслуживать большое количество удаленных пользователей, которым необходим лишь сетевой навигатор на рабочем месте.

Ядро ROLAP-системы обеспечивает извлечение агрегированных показателей в многомерном базисе измерений. В таком виде численная информация удобна для просмотра и построения отчетов, но в том, что касается собственно аналитической обработки, продукты ROLAP сами по себе могут выполнять только простейшие действия. Для сложных способов анализа информации инструментарий ROLAP может рассматриваться исключительно как средство подготовки исходного материала. Поэтому ядро системы ИнфоВизор-Аналитик реализовано таким образом, что к нему могут обращаться внешние модули интеллектуального и проблемно-ориентированного анализа. Дополнение системы такими модулями в каждом конкретном случае ее применения позволяет перейти в область закономерностей и решать задачи из сферы Data Mining.

ИнфоВизор-Администратор. Этот продукт является основным инструментом администратора хранилища данных. С его помощью могут создаваться и редактироваться *информационные модели*, составляющие навигационные и аналитические метаданные, а также сценарии загрузки информации из внешних источников - операционные метаданные.

Список литературы

1. Л. Л. Винокуров, Д. В. Леонтьев, А. Ф. Гершельман. СУБД ADABAS - основа универсального сервера баз данных // СУБД. - 1997. - №2. - С. 36-40.
2. Н. Е. Емельянов. Введение в СУБД ИНЕС. - М.: Наука, 1988.
3. Н. Кречетов, П. Иванов. Продукты для интеллектуального анализа данных // *ComputerWeek-Москва*. - 1997. - N 14-15. - С. 32-39.
4. А. А. Сахаров. Концепции построения и реализации информационных систем, ориентированных на анализ данных // *СУБД*. - 1996.- N 4. - С. 55-70.
5. E. F. Codd, S. B. Codd, C. T. Salley. Providing OLAP (On-Line Analytical Processing) to User-Analysts: An IT Mandate. - E. F. Codd & Associates, 1993.
6. J. Gray, S. Chaudhuri, A. Bosworth, A. Layman, D. Reichart, M. Venkatrao, F. Pellow, H. Pirahesh. Data Cube: A Relational Aggregation Operator Generalizing Group-By, Cross-Tab, and Sub-Totals // *Data Mining and Knowledge Discovery*. - 1997. - N 1. - P. 29-53.
7. D. Hackathorn. Reinventing Enterprise Systems Via Data Warehousing. - Washington, DC: The Data Warehousing Institute Annual Conference, 1995.
8. W. H. Inmon. Building The Data Warehouse (Second Edition). - NY, NY: John Wiley. - 1993.
9. K. Parsaye. New Realms of Analysis: Surveying Decision Support // *Database Programming & Design*. - 1996. - N 4. - P. 26-33.
10. N. Raden. Данные, Данные и только данные // ComputerWeek-Москва. - 1996. - №8. - С. 28.