УДК 004.651

# МАТЕМАТИЧЕСКОЕ МОДЕЛИРОВАНИЕ ВЫБОРА ФИЗИЧЕСКОЙ ОРГАНИЗАЦИИ БАЗ ДАННЫХ

**А.А. Карпук**

*Национальный банк Республики Беларусь. Расчетный центр, Anatoly_Karpuk@bisc.by*

Базы данных (БД) и системы управления базами данных (СУБД) используются для решения экономических задач в автоматизированных информационных системах (АИС) предприятий, банков и финансовых компаний. СУБД лежат в основе систем мониторинга и прогнозирования развития отраслей и экономики страны в целом. Объем БД и время выполнения приложений, взаимодействующих с БД средствами СУБД, существенно зависят от решений, принятых при проектировании логической и физической организации БД. На ранних этапах развития теории и практики БД вопросы проектирования логической и физической организации БД рассматривались в многочисленных публикациях. В настоящее время в книгах по СУБД описываются, в основном, только методы проектирования логической организации БД. Это объясняется тем, что современные СУБД Oracle, Microsoft SQL Server, IBM DB2, MySQL и др. при создании БД по умолчанию создают физические структуры БД, которые в большинстве случаев удовлетворяют требованиям пользователей по объемам БД и по времени выполнения приложений. Однако для АИС реального времени, а также для информационно–аналитических систем задача проектирования оптимальной физической организации БД по–прежнему остается актуальной.

Классическая концепция БД в соответствие со стандартом ANSI/SPARC предусматривает три уровня описания данных: внешний, концептуальный и внутренний. На каждом уровне используется соответствующая модель данных. Описание БД в контексте конкретной модели данных называется схемой БД. Внешние схемы БД описывают данные в том виде, в котором они доступны пользовательским приложениям. Концептуальная схема описывает БД в виде, едином для всех приложений и не зависящем от используемого в СУБД представления данных в среде хранения и путей доступа к ним. Внутренняя схема БД описывает представление данных в среде хранения и пути доступа к ним. Внешние и концептуальная схема относятся к логической организации данных, а внутренняя схема – к физической организации данных.

Для сложных предметных областей практически невозможно сразу построить концептуальную и внешние схемы БД, поэтому в современных методологиях и CASE–средствах проектирования БД концептуальный уровень описания данных разбит на два: информационно–логический (инфологический) и даталогический. На инфологическом уровне БД описывается в виде, не зависящем от используемой СУБД, а на даталогическом уровне БД описывается на языке описания данных конкретной СУБД. Автором разработана методология описания предметной области АИС [1], в которой на заключительном этапе строится обобщенная инфологическая модель предметной области (так называемая каноническая модель), для построения которой используется расширенная реляционная модель данных. Множество допустимых структурных компонентов этой модели данных имеет вид $K = \{D, A, Q, EQ\}$, где $D = \{D_s\}, s = \overline{1, d}$ – множество доменов, $A = \{A_i\}, i = \overline{1, a}$ – множество атрибутов, $Q = \{Q_j\},\ j = \overline{1, N},\ Q_j \subset A$ – множество отношений в четвертой нормальной форме, $EQ = \{Q_j \rightarrow Q_i \mid Q_i, Q_j \in Q\}$ – множество функциональных связей (ФС) различных типов между отношениями.

Задача выбора физической организации БД состоит в нахождении такого отображения канонической схемы предметной области на физические структуры данных, поддерживаемые выбранной СУБД, при котором достигает экстремума функция, выражающая критерий эффективности, и выполняется система ограни-

чений, учитывающих максимальные допустимые объемные и временные характеристики, а также отражающих связи между параметрами структуры хранения. В работе рассматривается задача отображения канонической схемы предметной области на память прямого доступа, в результате решения которой можно получить обобщенную физическую структуру БД, не зависящую от типа СУБД, а затем отобразить ее на физические структуры конкретной СУБД и построить внутреннюю схему БД. Предполагается, что СУБД функционирует в операционной среде, поддерживающей последовательный и прямой способы организации наборов данных (файлов) и базисный метод доступа с блокированием хранимых записей.

При построении математической модели для выбора физической организации БД предполагаются известными следующие объемные и временные характеристики канонической модели предметной области: длины значений каждого атрибута; среднее и максимальное число кортежей каждого отношения; среднее и максимальное число кортежей отношения, функционально определяющих один кортеж зависимого отношения для всех ФС между отношениями; среднее и максимальное число значений вторичных (поисковых) ключей отношений; количество приложений (функциональных задач), учитываемых при проектировании БД; среднее число выполнения операций включения (INSERT), удаления (DELETE), обновления (UPDATE) кортежей каждого отношения, поиска (SELECT) кортежей каждого отношения по каждому основному и вторичному ключу при выполнении каждого приложения. В качестве критерия проектирования физической структуры БД обычно выбирают объем БД $V$ или среднее время выполнения приложений $T_l, l = \overline{1, Z}$, где $Z$ – количество приложений. Наиболее общим является интегральный критерий в виде функции затрат

$$K = \varphi_0 V + \sum_{l=1}^{Z} \varphi_l T_l,$$

где $\varphi_l, l = \overline{0, Z}$ – известные весовые коэффициенты.

Искомыми параметрами физической организации БД, используемыми при построении математической модели, являются: распределение отношений канонической схемы предметной области по типам логических записей линейной или иерархической структуры; способ организации внутренней структуры каждого типа записи; способ организации каждого логического файла; способы индексирования; распределение логических файлов по областям хранения (физическим файлам), размеры страниц областей. При построении модели для нахождения этих параметров вводится ряд переменных, через которые выражаются величины $V$ и $T_l$, используемые в целевой функции, а также в ограничениях вида $V \le V^{\max}$, $T_l \le T_l^{\max}$.

В работах [2, 3] были получены расчетные соотношения для вычисления объема БД $V$ и среднего времени выполнения приложений $T_l, l = \overline{1, Z}$, при известных объемных и временных характеристиках канонической модели предметной области в зависимости от искомых параметров физической организации БД. Анализ полученной математической модели для выбора физической организации БД показывает, что задача поиска параметров физической организации данных, обеспечивающих глобальный минимум функции затрат при выполнении ограничений на объем БД и среднее время выполнения приложений, является NP – трудной. Число искомых переменных задачи достаточно велико, между переменными существуют различные взаимосвязи, целевая функция и функции системы ограничений являются нелинейными. В силу этого задачу разработки внутренней схемы БД предлагается рассматривать в виде совокупности задач субоптимизации по локальным критериям оптимальности с использованием формально – эвристических методов.

С учетом того, что приложения разрабатываются и включаются в систему не одновременно, можно предложить следующий подход к разработке внутренней схемы БД. Для первого включаемого в систему приложения или группы приложений строится концептуальная схема БД и выбираются параметры внутренней схемы БД, обеспечивающие минимальный объем БД. Производится оценка среднего времени выполнения каждого включаемого в систему приложения при выбранных параметрах внутренней схемы БД. Если ограничение на среднее время выполнения для рассматриваемого приложения не выполняется, то производится модификация внутренней схемы БД, причем допускается только такая модификация, которая не увеличивает среднее время выполнения каждой операции над каждым отношением канонической модели предметной области. Среди всех модификаций, удовлетворяющих ограничению на среднее время выполнения рассматриваемого приложения, выбирается модификация, доставляющая минимальное приращение целевой функции. Полученная внутренняя схема используется в качестве начальной при включении в систему очередного приложения. В силу выбора допустимых модификаций внутренней схемы БД при таком подходе среднее время выполнения ранее включенных в систему приложений никогда не увеличивается.

При построении начальной внутренней схемы БД и ее модификации решаются следующие задачи субоптимизации по локальным критериям оптимальности: распределение отношений канонической схемы предметной области по типам логических записей; выбор способа организации внутренней структуры каждого типа записи; выбор способа организации каждого логического файла; выбор способа индексирования каж-

дого логического файла; распределение логических файлов по областям хранения; выбор размеров страниц областей.

Построим математическую модель задачи распределение логических файлов по областям хранения. Пусть в результате решения задачи распределения отношений канонической схемы предметной области по типам логических записей, математическая модель которой построена в работе [4], получено $f$ типов логических записей. Каждому типу логических записей соответствует логический файл $F_p, p=\overline{1, f}$, через $L_p$ обозначим максимальный объем файла $F_p$. Для каждой пары файлов $F_p$ и $F_q$, $p=\overline{1, f}, q=\overline{1, f}$, из канонической модели предметной области определим $C_{pq}$ – суммарное количество связей между отношениями, вошедшими в состав логических записей файлов $F_p$ и $F_q$. Требуется распределить логические файлы по $b$ областям таким образом, чтобы суммарное количество связей между файлами, находящимися в одной области, было максимальным. Максимальный объем каждой области ограничен величиной $L^{\max}$. Математическая модель задачи распределение логических файлов по областям хранения имеет вид:

$$\sum_{s=1}^{b}\sum_{p=1}^{f}\sum_{q=1}^{f} C_{pq}\, x_{ps}\, x_{qs} \to \max;$$

$$\sum_{s=1}^{b} x_{ps} = 1 \text{ для всех } p=\overline{1, f};$$

$$\sum_{p=1}^{f} L_p x_{ps} \le L^{\max} \text{ для всех } s=\overline{1, b};$$

$$x_{ps} \in \{0,1\} \text{ для всех } p=\overline{1, f},\ s=\overline{1, b}.$$

***Список использованных источников***


1. Карпук, А.А. Информационное моделирование предметной области автоматизированных информационных систем / А.А. Карпук // Технологии информатизации и управления: Сб. научн. ст. Вып. 2 / Редкол.: А.М. Кадан (отв. ред.) [и др.]. – Минск: БГУ, 2011. – С. 24–30.

2. Карпук, А.А. Методы оценки объемов памяти при проектировании физической структуры базы данных / А.А. Карпук // Вопросы радиоэлектроники. Сер. ОВР. – 1990. – Вып. 11. – С. 3–12.

3. Карпук, А.А. Методы оценки производительности СУБД в АСУ специального назначения / А.А. Карпук // Вопросы специальной радиоэлектроники. Сер. СОИУ. – 1992. – Вып. 1. – С. 12–17.

4. Карпук, А.А. Задача выбора типов объектов при проектировании базы данных / А.А. Карпук // Вопросы специальной радиоэлектроники. Сер. СОИУ. – 1987. – Вып. 10. – С. 79–85.