

# **Использование СУБД для хранения и поиска информации о физико-химических свойствах веществ**

Г.В. Белов (gbelov@imail.ru)

Термоцентр им. В.П. Глушко, ИТЭС ОИВТ РАН, 127412, Москва, Ижорская 13/19

Проведен анализ возможностей использования СУБД различных классов для хранения и поиска данных о термодинамических свойствах индивидуальных веществ. Результаты анализа могут быть полезны при проектировании и создании информационных систем для моделирования физико-химических процессов.

The analysis of possibility of usage of various RDBMS for thermodynamic data storage and data search is accomplished. The results of analysis may be useful for the development of information systems intended for modeling of complex physico-chemical processes.

Для хранения информации о термодинамических и термохимических свойствах индивидуальных веществ, как правило, используются файлы специального вида [1, 2, 3, 4], которые создаются разработчиками программного обеспечения в соответствии с их представлениями об особенностях этих данных и требованиями к их защищенности от несанкционированного доступа (подход на основе плоских файлов). В том случае, если список параметров, сведения о которых хранятся на диске, невелик, такой подход вполне оправдан, поскольку он обеспечивает эффективный поиск и выборку информации, а также естественным образом защищает ее от тех, кто не знает структуры используемой файловой системы. Однако в общем случае условие компактности данных не выполняется, а если, кроме того, информационно-справочная система интенсивно развивается, такой подход становится не эффективным, поскольку трудности, обусловленные необходимостью изменения структуры файлов и самое главное – необходимостью модификации программного обеспечения, становятся чрезмерно большими, см. также [5]. Более подробно недостатки использования файловых систем для создания баз данных рассмотрены в работе [6]. В этой ситуации представляется целесообразным использование какой-либо стандартной СУБД в сочетании со средством программирования типа Delphi, C++ Builder или Visual Basic.

Покажем на нескольких примерах, насколько эффективным является поиск информации в базе данных о термодинамических свойствах веществ, при реализации ее с использованием различных СУБД. В данном случае производительность СУБД определяется главным образом характеристиками ее машины баз данных, то есть набором служебных процедур и функций, при помощи которых осуществляется доступ к информации реляционных таблиц определенного формата (Paradox, Access, InterBase). Так доступ к таблицам Paradox может осуществляться через BDE (Borland Database Engine), а доступ к таблицам Access - через MS Jet. Реализация пользовательского интерфейса в настоящей работе выполнена с использованием Delphi 5. Для того, чтобы обеспечить доступ к таблицам в формате Access использованы ADO (Microsoft ActiveX Data Objects).

Для оценки производительности различных СУБД была использована реальная база данных о термодинамических свойствах индивидуальных веществ, структура которой приводится в приложении. База данных, содержащая сведения 2527 веществах, которые образованы 79 химическими элементами, состоит из пяти таблиц. Данные были нормализованы в соответствии с требованиями, предъявляемыми к реляционным базам данных, [6,7].

В таблице DATA1 хранится только первичный индекс и химическая формула, которая является главным идентификатором вещества. В таблице DATA2 приведены общие сведения, относящиеся к молекуле данного вещества (индекс CAS, название, молекулярная масса, параметры потенциала Леннард-Джонса, дипольный момент и т.д.). Таблицы DATA1 и DATA2 связаны соотношением 1:1. В принципе, данные из этих таблиц можно объединить, однако при этом время поиска возрастает за счет увеличения размера записи. Таблица DATA3 хранит информацию о веществе в данном фазовом состоянии (твердом, жидком или газообразном) термодинамические свойства вещества при атмосферном давлении и комнатной температуре, энтальпию образования и т.д. Таблица DATA1 связана с таблицей DATA3 соотношением 1:∞. Таблица DATA4, связанная с таблицей DATA3 отношением ∞:1, содержит значения коэффициентов полинома, аппроксимирующего зависимость приведенной энергии Гиббса от температуры, а также границы температурных интервалов. Наконец, в таблице DATA5 хранятся сведения о химических элементах, из которых состоит молекула вещества, и числе атомов каждого из элементов в молекуле.

Основная масса запросов к базе данных по термодинамическим свойствам веществ имеет вид:

- «ВЫБРАТЬ ЧТО-ЛИБО ИЗ БД, ГДЕ  
а) ЗАДАН СПИСОК ХИМИЧЕСКИХ ЭЛЕМЕНТОВ  
б) ЗАДАН СПИСОК ФОРМУЛ ВЕЩЕСТВ»

Иногда возможны дополнительные ограничения типа «ТОЛЬКО ГАЗЫ» или «ТОЛЬКО ГАЗООБРАЗНЫЕ ИОНЫ».

Процедура выборки данных по заданному списку формул веществ как правило не вызывает проблем. Поэтому для оценки скорости выполнения запроса важно, прежде всего, определить время, затрачиваемое на выборку формул из базы данных по заданному списку химических элементов.

Выборку формул можно осуществить тремя способами:

1. при помощи SQL запроса,
2. путем линейного поиска в базе данных, т.е. последовательного просмотра всех записей соответствующих таблиц и выбора тех, которые удовлетворяют заданным критериям,
3. с использованием фильтра.

Было проведено исследование быстродействия (в указанном выше смысле) четырех реляционных СУБД:

- 1) Paradox 7 + BDE (Borland),
- 2) Access + MSJet OLE DB 4.0 (Microsoft),
- 3) InterBase 6.0 (Borland),
- 4) DBISAM 2.04 (Elevatesoftware).

Испытания проводились на компьютере Celeron 400, RAM = 128 Mbytes.

Для выборки формул использовался SQL запрос следующего вида

```
SELECT * FROM DATA1
WHERE NOT EXISTS
  (SELECT * FROM DATA5
   WHERE DATA5.D1_NO = DATA1.D1_NO
   AND NOT (EL_SYMB IN (ListOfElements)))
```

где ListOfElements – список символов химических элементов через запятую. Результаты испытаний приведены в табл. 1.

Процедура линейного поиска выполнялась двумя способами. В первом случае (прямой поиск) таблицы DATA1 и DATA5 были связаны отношением «главная-подчиненная» по полю D1\_NO. Таблица DATA1 последовательно просматривалась сверху вниз, при этом проверялся список элементов в подчиненной таблице DATA5. Критерием выбора текущего значения поля Formula из DATA1 являлось условие вхождения значений поля EL\_SYMB всех записей таблицы DATA5 в список элементов ListOfElements.

Во втором случае (альтернативный поиск) таблицы DATA1 и DATA5 были независимы. Осуществлялся последовательный просмотр таблицы DATA5, записи в которой были упорядочены по полю D1\_NO. При этом итерационно формировался текущий список элементов для всех полей, имеющих одинаковое значение поля D1\_NO. После смены номера D1\_NO, который означал переход к новому веществу, текущий список элементов сравнивался со списком ListOfElements. В том случае, если все элементы текущего списка входили в ListOfElements, выполнялся поиск записи в DATA1 по предыдущему значению поля D1\_NO и производился выбор значения поля Formula.

В последнем случае, когда выборка формул производилась при помощи фильтра, в процедуре фильтрации использовался анализатор формул веществ. Формула вещества попадала в число отобранных, если список химических элементов, полученных в результате ее анализа, либо совпадал со списком ListOfElements, либо целиком в нем содержался. С точки зрения программной реализации этот случай является наиболее трудоемким, поскольку алгоритм процедуры анализа химических формул веществ, который зависит от правил их записи, может быть довольно сложным. Кроме того, указанный подход является наименее гибким из трех способов поиска.

Как видно из таблицы 1 поиск при помощи SQL-запроса наиболее эффективен в случаях использования СУБД InterBase и Access. Время альтернативного поиска во всех случаях меньше времени прямого поиска (причем в некоторых случаях существенно меньше). В свою очередь применение фильтра в ряде случаев помогает еще больше ускорить процедуру поиска.

Следует отметить, что рассмотренные СУБД относятся к разным классам. Так Paradox и Access по сложившейся терминологии принято называть настольными СУБД, InterBase - типичная клиент - серверная СУБД, а DBISAM можно отнести к разряду «сверхлегких» СУБД, поскольку ее машина баз данных чрезвычайно компактна - около 300 КБайт, причем никакие дополнительные драйверы не используются.

Приведенные данные о быстродействии носят скорее качественный, чем количественный характер, поскольку был рассмотрен только один тип запросов. Однако, некоторые выводы все-таки можно сделать. Как видно из таблицы 1, DBISAM 2.04 является наиболее медленной из четырех рассмотренных СУБД, а наиболее универсальной СУБД является InterBase, поскольку ее машина баз данных обеспечивает эффективное выполнение SQL запросов и быструю навигацию по базе данных. Следует подчеркнуть, что в большинстве испытаний время выборки слабо зависело от числа элементов в списке, т.е. запросы типа «ВЫБРАТЬ ВСЕ ФОРМУЛЫ» и «ВЫБРАТЬ ФОРМУЛЫ ВСЕХ ВЕЩЕСТВ, КОТОРЫЕ ОБРАЗОВАНЫ ListOfElements» выполнялись практически за один промежуток времени.

Время выполнения аналогичных запросов с использованием БД на основе плоских файлов (все данные размещены в обычных файлах специального вида) составляет примерно 0.01 – 0.1с (естественно, в этом случае возможен только линейный поиск). Таким образом, выигрыш от использования специализированных файловых структур для хранения данных очень невелик, а иногда и вовсе отсутствует, поэтому использование реляционных СУБД для хранения больших массивов информации о физико-химических свойствах веществ можно считать вполне оправданным.

Таблица 1. Примерное время выборки данных с использованием различных СУБД, с.

| СУБД           | Время выполнения SQL-запроса. | Время прямо-го линейного поиска. | Время альтер-нативного поиска. | Время поиска с использованием фильтра. |
|----------------|-------------------------------|----------------------------------|--------------------------------|----------------------------------------|
| Paradox + BDE  | 50                            | 0.5                              | 0.2                            | 0.2                                    |
| Access + MSJet | 1                             | 100                              | 3.5                            | 0.5                                    |
| OLE DB 4.0     |                               |                                  |                                |                                        |
| InterBase 6.0  | 0.7                           | 3.5                              | 0.9                            | 0.2                                    |
| DBISAM 2.04    | 3.5 <sup>1)</sup>             | 3                                | 1                              | 0.5                                    |

<sup>1)</sup> Реализация языка SQL в СУБД DBISAM 2.04 не имеет оператора EXISTS, поэтому для выборки был использован запрос вида

```
select distinct FORMULA, D1_NO from DATA1, DATA5
where (DATA1.D1_NO = DATA5.D1_NO)
and DATA5.D1_NO not in (select D5.D1_NO from DATA5 D5
where (D5.EL_SYMB not in (ListOfElements)))
```

## Приложение. Структура базы данных

### **DATA1. Список формул индивидуальных веществ.**

| Поле      | Тип       | Назначение     |
|-----------|-----------|----------------|
| D1_NO     | Autoinc   | Первичный ключ |
| Formula   | Text (20) | Формула        |
| IonCharge | Integer   | Заряд          |

### **DATA2. Сведения о молекуле вещества.**

| Поле   | Тип       | Назначение                                         |
|--------|-----------|----------------------------------------------------|
| D1_NO  | Longint   | Внешний ключ                                       |
| CASidx | Longint   | CAS индекс                                         |
| Name   | Text (60) | Название вещества                                  |
| ...    |           | прочая информация, относящаяся к молекуле вещества |

### **DATA3. Сведения о веществе в данном состоянии.**

| Поле   | Тип      | Назначение                                    |
|--------|----------|-----------------------------------------------|
| D3_NO  | Autoinc  | Первичный ключ                                |
| D1_NO  | Longint  | Внешний ключ                                  |
| Pstate | Text (2) | Фазовое состояние                             |
| DH0    | Single   | $\Delta_f H^\circ(0)$                         |
| DH298  | Single   | $\Delta_f H^\circ(298.15)$                    |
| Cp298  | Single   | $C_p^\circ(298)$                              |
| S298   | Single   | $S^\circ(298)$                                |
| H298H0 | Single   | $H^\circ(298.15) - H^\circ(0)$                |
| ...    |          | прочие сведения о веществе в данном состоянии |

**DATA4.** Коэффициенты полинома, аппроксимирующего зависимость приведенной энергии Гиббса от температуры.

| Поле  | Тип      | Назначение                                |
|-------|----------|-------------------------------------------|
| D4_NO | Autoinc  | Первичный ключ                            |
| D3_NO | Longint  | Внешний ключ                              |
| Tmin  | Single   |                                           |
| Tmax  | Single   |                                           |
| F1    | Double   |                                           |
| F2    | Double   |                                           |
| F3    | Double   |                                           |
| F4    | Double   |                                           |
| F5    | Double   |                                           |
| F6    | Double   |                                           |
| F7    | Double   |                                           |
| P     | Text (2) | Дополнительный признак фазового состояния |

**DATA5.** Служебная таблица.

| Поле    | Тип      | Назначение                               |
|---------|----------|------------------------------------------|
| D5_NO   | Autoinc  | Первичный ключ                           |
| D1_NO   | Longint  | Внешний ключ                             |
| EL_SYMB | Text (2) | Символ химического элемента              |
| EL_QTY  | Single   | Число атомов EL_SYMB в молекуле вещества |

## Литература

1. Gurvich, L.V., Iorish, V.S. et al. IVTANTHERMO - A Thermodynamic Database and Software System for the Personal Computer. User's Guide. CRC Press, Inc., Boca Raton, 1993.
2. Белов Г.В., Иориш В.С., Юнгман В.С. Моделирование равновесных состояний термодинамических систем с использованием ИВТАНТЕРМО для Windows. Теплофизика высоких температур.-2000.-т.38, №2.-С.191-196.
3. Применение ЭВМ для термодинамических расчетов металлургических процессов / Г.Б. Синярев, Н.А. Ватолин, Б.Г. Трусов, Г.К. Моисеев. – М.: Наука, 1982. – 263с.
4. Bale C.W., Eriksson G. Metallurgical Thermochemical Databases – a Review. Canadian Metallurgical Quarterly, Vol. 29, No. 2, pp.105-132.- 1990.
5. Математическое моделирование высокотемпературных процессов в энергосиловых установках / В.Е. Алемасов, А.Ф. Дрегаллин, В.Г. Крюков, В.И. Наумов.-М.: Наука, 1989.-256с.
6. Коннолли Т., Бегг К., Страчан А. Базы данных: проектирование, реализация и сопровождение. Теория и практика. – М.: Издательский дом «Вильямс», 2000. – 1120с.
7. Дейт К.Дж. Введение в системы баз данных. - К.:Диалектика, 1998.-784с.