

УДК 551.501 + 556.043

**Ю.В. Доронина, канд. техн. наук, доцент,
М.Р. Валентюк**

Севастопольский национальный технический университет

Студгородок, г. Севастополь, Украина, 99053

E-mail: root@sevgtu.sebastopol.ua

ОПТИМИЗАЦИЯ ПРОЦЕССА АРХИВИРОВАНИЯ ДАННЫХ В СКЛАДСКОМ ХОЗЯЙСТВЕ АВТОМАТИЗИРОВАННОГО ПРОИЗВОДСТВА

Рассматриваются аспекты оптимизации процесса архивирования данных. Исследуются возможности разработки современного автоматизированного архива, в том числе для гидрометеорологической отрасли.

Современное развитие всех отраслей народного хозяйства связано с распространением компьютерной техники и сохранением данных различных видов. Накопление информации на технических носителях в складском хозяйстве приводит к различного рода сложностям, связанным как с физическим разрастанием архива, что затрудняет поиск информации и возможность ее сохранения, так и с разнородностью форматов данных, предназначенных для обработки и анализа.

Во многих организациях наибольшее распространение получило развитие автоматизации документооборота. Для обеспечения сохранности данных используется принцип создания специальных структур хранения – архивов данных. Создание архива данных подразумевает решение двух важных задач: непосредственного архивирования и создания каталога архива. Таким образом, *целью данной статьи* является исследование возможности оптимизации процесса складирования данных для автоматизированного производства.

Важнейшей частью современного архива на технических носителях, помимо самих данных, является некоторая программная оболочка, автоматизирующая процессы контроля хранимости данных. Эти функции можно возложить на так называемую автоматизированную архивную систему (ААС), представляющую собой совокупность взаимосвязанных технических и программных средств, скоординированных в заданной области. В целом, ААС – это комплекс программно-технических средств, включая информационное обеспечение, предназначенное для формирования, пополнения и поддержания в корректном состоянии архив на технических носителях (АТН).

Не следует понимать ААС только как базу данных (БД), поскольку ААС имеет более широкое понятие, в которое входит и четкая организация работ по приему, использованию, хранению информации, что обеспечивается теоретически обоснованной структурой ААС, а также ведение документации, включая синхронизацию и обеспечения целостности хранящейся информации.

Функции ААС выполняет автоматизированное рабочее место (АРМ) инженера АТН в комплексе с методическим обеспечением ААС. В больших автоматизированных архивных комплексах (ААК) к АРМ инженера АТН добавляется АРМ пользователя, которое представляет собой среду для оптимального выполнения запросов (транзакций) по архивной информации складского характера. Наиболее естественным решением для ААК является также наличие локальной сети между несколькими клиентскими автоматизированными рабочими местами.

Создание АТН представляет собой сложную, многоэтапную задачу, что обусловлено следующими факторами:

- 1) частой необходимостью бессрочного хранения данных;
- 2) централизованным хранением данных, при котором возникает отчужденность архивных структур от проектировщиков;
- 3) различиями форматов архивных данных;
- 4) большим объемом архивов.

Исходя из главных задач архива, необходимо предусмотреть следующее:

- экономичное, компактное и надежное накопление и долговременное безопасное сохранение документов на разных носителях информации;
- контроль над сохранностью документов, которые временно сохраняются в организациях и подразделениях;
- оперативное обеспечение потребностей организаций и учреждений различных отраслей экономики и форм собственности продукцией архива документов (работах, информационных услугах) для организации эффективного производства и научных исследований, в том числе и на договорных началах.

Дополнительным и важным ограничением в современных условиях должен выступать минимальный уровень финансовых затрат. Первый этап автоматизации представлен созданием непосредственно проекта и концептуальных основ ААС. На втором этапе возникает задача разработки

так называемого «форматного шлюза», обеспечивающего соответствие или преобразование иноформатных данных для хранения на сервере, с которого, в свою очередь имеется возможность выхода в другие системы, т.е. необходимо параллельно с разработкой новой современной архивной системы иметь так называемую «буферную зону», в которой будет в течение некоторого времени осуществляться перевод уже имеющихся данных в новый общесовместимый формат. В функции «форматного шлюза» входит и декодирование данных из локальных форматов.

На следующем этапе продолжается устранение многоформатности данных как последствия наличия разрозненных локальных систем обработки и хранения, и создаются условия для параллельной работы в «буферной зоне» и в современной ААС с постепенным отмиранием первой.

В практике архивации на персональных компьютерах принято структурировать размещение файлов данных на носителях в виде так называемого дерева каталогов, представляющего собой систематизированное объединение файлов в группы, затем эти группы в более крупные группы и так далее до главных корневых каталогов. Подобное структурирование архива существенно облегчает работу с ним на уровне пользователя (удобство и быстрота поиска нужного файла) и на уровне ЭВМ.

Следует отметить, что идеи оптимизации процесса архивирования данных складского хозяйства возможно распространить и на другие отрасли народного хозяйства, где также особо важным звеном производства являются данные. Так, например, складирование данных для гидрометеорологии является важным этапом технологического процесса контроля данных. Организация гидрометеорологической (ГМ) информации должна обладать следующим важным свойством: возможностью оперативного доступа (выборки данных) (ОД) при условии высокого качества контроля данных (качества данных) (КД). Учитывая при этом постоянную тенденцию информации к накоплению, возникает задача построения такой организации ГМ данных на базе современных компьютерных технологий и технических средств, которая обладала бы этим свойством [1, 2].

Очевидное на первый взгляд решение этой задачи – в непосредственной организации БД – является далеко не тем оптимальным шагом, который бы дал возможность совместить ОД и КД. Можно добавить еще множество необходимых условий для формирования такой структуры ГМ данных: это и удобная среда пользователя (пользовательский интерфейс) для ввода ГМ данных, и учет стандартов проверок введенных данных (не менее трех), совместимые форматы ГМ данных для обмена.



Рисунок 1 – Виды представления данных для архивирования

Поставленная задача, таким образом, сводится к организации оперативного доступа при всех указанных условиях. Решение этой задачи на местах неприемлемо. Оно свелось бы к накоплению локальных разрозненных БД, которые будут иметь, по всей вероятности, несовместимые структуры. Имеется ряд факторов, которые делают проблему структурирования архива гидрометеорологических данных, в общем случае, очень сложной.

Поэтому, как уже указывалось выше, должен применяться единый подход на строго научной основе. Эта основа описана далее как теория организации структур для хранения информации.

Поставленная задача усложнена еще и тем, что обычно на предприятиях архив в той или иной форме уже существует и входящий поток данных в него среднем постоянен. Создание и внедрение в этих условиях принципиально нового подхода к организации архива данных на технических носителях практически невозможно. Поэтому, должен быть применен метод постепенного (поэтапного) перевода существующего архива в современный АТН. При этом необходимо постоянное согласование с существующими стандартами, форматами данных, применяемыми технологиями. Дополнительным и важным ограничением в современных условиях должен выступать минимальный уровень финансовых затрат.

Одна из наиболее важных проблем в деле архивации данных состоит в эффективном распределении архива на технических носителях. Разработка соответствующих методик и критериев также входит в круг задач, решаемых в данной работе. Предлагается ввести некоторые понятия. Принимается допущение, что в архивах на предприятиях автоматизированных производств в основном данные хранятся в виде электронных таблиц (это характерно и для параллельно рассматриваемой гидрометеорологической отрасли).

Пусть S-формы – это структуры данных, предназначенные для эффективного обмена и использования (в виде БД), преобразованные из библиотеки электронных таблиц (БЭТ) по определенному принципу. Иными словами, S-форма – это блок поименованных данных, образованный по смысловому признаку. Например: данные по отказам оборудования, организованные в директории по годам, далее в более крупные директории по типам таблиц и т.д. S-формы определяются, в основном, совокупностью классифицированных запросов. Таким образом, ставится задача классификации запросов и далее – классификации S-форм.

Минимальная S-форма строится при создании для каждого файла из БЭТ соответствующего файла формата .dbf (связь: один к одному):

$$S_{\min} = \begin{cases} 1FILE(.xls) \\ 1FILE(.dbf) \end{cases} = S_0,$$

где .xls – файл формата таблиц Excel (рабочий формат); .dbf – файл формата dBase (формат локальных таблиц); .gdb – файл базы данных InterBase – сервера (тип системы клиент – сервер).

Следует отметить, что форматы выбраны произвольно и в общем случае могут быть соответствующими новыми форматами разрабатываемой структуры.

Максимальная S-форма строится при создании для всех файлов из БЭТ одного файла формата .gdb

$$S_{\max} = \begin{cases} NFILE(.xls) \\ 1FILE(.dbf) \end{cases}.$$

Для $S_{\max}$ характерно наличие единой программы обработки запросов, единого интерфейса.

Промежуточные S-структуры, разработанные для гидрометеорологического архива определяются следующим образом:

$$\begin{cases} S_1^1 = 1Год(1Станция(1FILE(.dbf))), \\ S_1^2 = iЛет(1Станция(1FILE(.dbf))), \\ S_1^3 = (блок).Лет(1Станция(1FILE(.dbf))); \end{cases}$$

$$\begin{cases} S_2^1 = 1Год(TГМ_j(1Станция(1FILE(.dbf)))), \\ S_2^2 = iЛет(TГМ_j(1Станция(1FILE(.dbf)))), \\ S_2^3 = (блок).Лет(TГМ_j(1Станция(1FILE(.dbf)))); \end{cases}$$

$$\begin{cases} S_3^1 = NСтанций(1Год(TГМ_j(1FILE(.dbf)))), \\ S_3^2 = (список)Станций(1Год(TГМ_j(1FILE(.dbf)))); \end{cases}$$

$$\begin{cases} S_4^1 = NСтанций(iЛет(TГМ_j(1FILE(.dbf)))), \\ S_4^2 = (список)Станций(iЛет(TГМ_j(1FILE(.dbf)))), \\ S_4^3 = NСтанций((блок).Лет(TГМ_j(1FILE(.dbf)))), \\ S_4^4 = (список)Станций((блок).Лет(TГМ_j(1FILE(.dbf)))); \end{cases}$$

$$\begin{cases} S_5^1 = NСтанций(iЛет(1FILE(.dbf))), \\ S_5^2 = (список)Станций(iЛет(1FILE(.dbf))), \\ S_5^3 = NСтанций((блок).Лет(1FILE(.dbf))), \\ S_5^4 = (список)Станций((блок).Лет(1FILE(.dbf))). \end{cases}$$

При формировании запросов Z-серий учитываются количественные показатели параметров структур. Например, при трех основных типах ГМ таблиц (ТГМ 1, ТГМ 2, ТГМ 7) возможно использование их как в виде вложенных каталогов, так и корневых. Таким образом, последовательность компонентов в S-формах может нарушаться в зависимости от построенного запроса.

Нумерация запросов соответствует нумерации S-форм. Например, S-форме типа S_3^1 соответствует серия запроса Z_3^1 :

$$\begin{aligned} Z_3^1 &= TГМ_j(iГод(NСтанций)) \\ Z_3^1 &= \max T_{\text{годо}}(1994-1996гг.(SEV, HM, YAL)) \end{aligned}$$

Для начального формирования S-форм необходимо составить список параметров структур.

На рисунке 2 изображена структура архива на технических носителях морских прибрежных наблюдений Гидрометеорологической обсерватории Севастополя, построенная на базе S-форм.

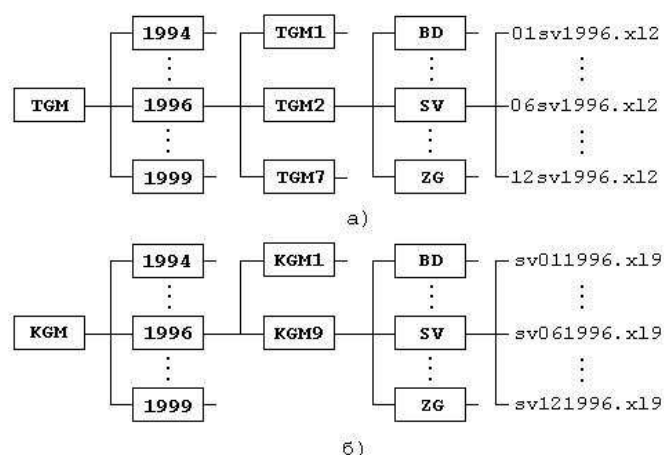


Рисунок 2 – Структура архива на технических носителях

В таблицах 1 и 2 приведены сводные данные по архивируемым таблицам с целью выбора типа физического носителя.

Таблица 1 – Статистические характеристики

Таблица	Средний объем	Сред. отклон.	Результат
TGM2	40,3 КБ	1,6 КБ	40,3 ± 4,8 КБ
TGM7	35,9 КБ	0,4 КБ	35,9 ± 1,2 КБ
TGM1	124,0 КБ	4,0 КБ	124 ± 12 КБ

Таблица 2 – Параметры архивирования

Таблица	Количество станций	Объем таблиц за год	Кол-во лет на Zip*	Кол-во лет на CDD**
TGM2	24	-	-	-
TGM7	15	6,3 МБ	15,9	95,2
TGM1	31	45,0 МБ	2,2	13,3

*Zip-дискета 100 Mb

**CD-Disk 600 Mb

С целью оптимизации архивирования данных был проведен анализ возможности сжатия этих данных. Степень сжатия зависит от нескольких причин, а именно:

- большое значение имеет тип сжимаемых данных. Лучше всего сжимаются графические, текстовые файлы; для них степень сжатия может быть от пяти до сорока процентов; хуже сжимаются файлы исполняемых программ, загрузочных модулей, файлы мультимедиа;

- большое значение имеет метод сжатия;

- немаловажно и то, какой архиватор используется; при выборе типа архиватора обычно руководствуются следующими соображениями: чтобы степень сжатия была как можно выше, а времени на упаковку и распаковку файлов уходило как можно меньше.

Для изучения эффективности сжатия одного файла, содержащего ГМ данные, было взято четыре архиватора – WinRar 3.50, 7-Zip 4-42, PowerArchiver 2006 v.9.61 и WinUNA 2.0.

Был взят файл типа TGM-1 размером 158 Кб. Файл представляет собой электронные таблицы Excel.

Результаты исследования представлены в таблице 3.

Таблица 3 – Сравнительная характеристика эффективности сжатия

Архиватор	Расширение архива	Размер архива	Сжатие, %
WinUHA	*.uha	20 Kb	87,34
WinRAR	*.rar	25 Kb	84,18
7-Zip	*.7z	18 Kb	88,61
PowerArchiver	*.bh	34 Kb	78,48
PowerArchiver	*.cab	24 Kb	84,81
PowerArchiver	*.lzh	34 Kb	78,48
PowerArchiver	*.gzip	34 Kb	78,48
PowerArchiver	*.zip	34 Kb	78,48

Сравнение размеров показало, что эффективнее сжимает архиватор 7-Zip, в результате работы которого появляются архивы с расширением *.7z. При этом файл сжимается на 88,6 %.

Единственным недостатком исследуемых архиваторов является то, что они не учитывают специфику архивируемого файла, т.е. его структуру. При рассмотрении файла можно заметить, что файлы в пределах типа имеют идентичные «шапки» таблиц. Различием являются только данные, которые вписываются в таблицы. Следовательно, для уменьшения объема хранимых данных целесообразно разбивать файл на две части – «шапки» таблиц и сами данные. Одинаковые части файлов («шапки таблиц») хранить в отдельных файлах, так называемых шаблонах, а файлы с данными преобразовать таким образом, чтобы убрать избыточную и известную информацию при этом не потеряв связей с таблицами [3].

Поэтому для увеличения эффективности сжатия файлов был разработан программный модуль, учитывающий описанные особенности.

Усредненные значения полученных файлов-архивов для каждого типа файлов, содержащих гидрометеорологические данные, представлены в таблице 4. Таким образом, файл типа ТГМ-1 сжимается на 98,7 %.

Таблица 4 – Усредненные значения файлов-архивов

Тип файла	ТГМ-1	ТГМ-2	ТГМ-7	КГМ-1	КГМ-9а	КГМ-9т	КГМ-9э
Размер архива, Кб	2	2	1	2,8	1	1	1

Таким образом, полученные возможности алгоритмизации деятельности в строгой системе рабочей документации ведут к оперативности работы, простоте в нахождении любого документа, воспроизведении подобных документов на основе уже имеющихся, избавляет от рутинных работ, эффективны при научной обработке и статистическом анализе данных и просто при хранении громадных объемов данных. Эти исследования, как было показано, могут быть применимы не только для складского хозяйства автоматизированного производства, но и для непроизводственной, гидрометеорологической отрасли. Были предложены новые структуры распределения архива на технических носителях. Все это также служит для оптимального распределения рабочего времени, что является основой научной организации труда (НОТ).

Библиографический список

1. Доронина Ю.В. Автоматизация цикла гидрометеорологических данных / Ю.В. Доронина // Сб. научн. тр. УкрНИГМИ. — 1999. — Вып. 247. — С. 250 – 254.
2. Глинских А.Н. Архитектуры систем хранения данных NAS и SAN. Современное состояние и перспективы развития / А.Н. Глинских, А.А. Николаев // Электронный номер газеты “Компьютер-Информ”. — 2000. — № 9. — 48 с.
3. Иванова А.Ф. Аналитическая обработка данных / А.Ф. Иванова // Электронный номер журнала “ВУТЕ/Россия”. — 2006. — № 2. — 23 с.

Поступила в редакцию 26.01.2007 г.