УДК 371.315.7 (575.2) (04)

ПРИМЕНЕНИЕ НОВЫХ ИНФОРМАЦИОННЫХ ТЕХНОЛОГИЙ ПРИ ОБРАБОТКЕ СТАТИСТИЧЕСКИХ ДАННЫХ

В.Ф. Бабак – докт. техн. наук, проф.
Т.Г. Турчанова – ст. преподаватель
И.В. Хмелева – ст. преподаватель

The methods and means of large statistical data processing on an example of Kyrgyz farming are analyzed. Recommendations on application of modern methods in statistic analysis are given.

До начала 90-х годов обработкой данных занималась математическая статистика, которая до определенного времени была основным инструментом при анализе данных, размещенных только в локальных базах данных (БД). В настоящее время в связи с обработкой большого потока информации она уже не справляется. Появились потребности в решении масштабных задач и совершенствовании технологии записи и хранения данных. Обрабатываемые данные характеризовались следующим:

- ◊ неограниченным объемом;
- ◊ разнородностью представляемых структур;
- ◊ многообразием требуемых результатов анализа;
- ◊ распределенным хранением информации.

Современные базы данных состоят из двух компонентов: файлов, содержащих данные, и их описаний. Они размещены в системах управления базами данных (СУБД). Это программный продукт, используемый приложениями для доступа к данным и выполняющий следующие функции: определение данных; обеспечение их сохранности и соблюдение правил, установленных пользователем; манипулирование данными; выборка данных; безопасность.

Традиционно, приложения, работающие с базой данных, делятся на локальные и функционирующие в архитектуре клиент-сервер [1].

Одновременно с усложнением решаемых задач усложнялись и совершенствовались программы для работы с базами данных. Появились деления на одно- и многопользовательские СУБД. Возникла дополнительная классификация клиентских приложений на "тонкие" и "толстые", появились разнообразные способы связи между клиентом и сервером, алгоритмы обслуживания очередей клиентов и способы управления транзакциями. Согласно новой классификации все приложения для работы с базами данных делятся на группы в зависимости от числа уровней обработки данных [2].

Те программы, которые раньше назывались локальными (независимо от способа связи с СУБД), чаще всего сейчас входят в число одноуровневых приложений, так как обработка данных в них ведется в единственном месте. Клиент-серверные приложения стали подразделяться на двухуровневые (классический клиент-сервер) и трехуровневые (клиент-сервер с программным обеспечением промежуточного слоя).

Анализ имеющихся средств обработки данных. До сегодняшнего дня в республике в основном применялись одноуровневые программные комплексы. Базы данных в этих системах решали только задачи хранения и представления данных, из-за чего даже самые незначительные изменения входных данных тре-

бовали адаптации программного кода и перекомпиляции программ. Вопросы сохранения целостности данных были полностью возложены на СУБД с индексно-последовательной организацией файлов, что часто приводило к возникновению проблем при некорректном изменении данных извне. Для работы с такими системами нужны были опытные операторы, к тому же такие системы не позволяли повысить производительность.

При статистическом анализе фактографических данных можно выделить две основные фазы: сбор данных и получение выходных аналитических результатов.

К фазе сбора данных предъявляются следующие требования [3]:

- ◊ обслуживание большого числа пользователей, интенсивно добавляющих и модифицирующих данные;
- ◊ наличие гибкой структуры, которая требует минимальных усилий для внесения новых, модификации или удаления старых типов объектов;
- ◊ обеспечение целостности данных и высокой степени дополнительных контролей.

Вторая фаза обследования требует только быстрый и простой доступ к статическим данным для построения отчётов.

Различные требования к фазам ввода и вывода делают нецелесообразным их размещение в одной БД. В то же время особенностью реализации БД для выборочных обследований является то, что в процессе ввода система получает только выборочную совокупность, хотя для анализа требуются стопроцентные (сгенерированные на генеральную совокупность) данные. В настоящее время в республике применяются два метода получения аналитических данных для выборочных обследований: приведение выборки к генеральной совокупности в статической форме и использование взвешенных коэффициентов без приведения к статической форме.

Применение новых средств обработки данных. При обработке данных по последней переписи населения ГВЦ НСК была сделана попытка использовать серверную часть программного обеспечения для автоматизированной системы учёта и статистической обработки данных, т.е. попытка перейти от технологии настольных баз данных на клиент-серверную архитектуру. Применение денормализованной структуры БД потребовало написания слишком сложного программного обеспечения для клиентских приложений, так как вся бизнес-логика находилась на стороне клиента.

После проведения фазы ввода данных структура таблиц БД была адаптирована для получения отчётов и статистических таблиц. Однако при получении выходных данных по переписи населения возможности SQL применялись не достаточно эффективно, из-за чего не были использованы все преимущества денормализованной структуры.

Клиентскому приложению посылались результаты выполнения запросов и по сети передавались только те данные, которые в действительности были нужны клиенту. При этом значительно снизилась нагрузка на сеть. Кроме того, SQL-сервер оптимизировал полученный запрос таким образом, чтобы он был выполнен за минимально возможное время. Все это повысило быстродействие системы и снизило время ожидания результата запроса [4].

При выполнении запросов сервером существенно повышается степень безопасности данных, поскольку правила целостности данных задаются на стороне сервера и являются едиными для всех приложений, использующих эту базу данных. Мощный аппарат транзакций, поддерживаемый SQL-серверами, блокирует одновременное изменение одних и тех же данных различными пользователями и предоставляет возможность откатов к первоначальным значениям при внесении в базу данных изменений, закончившихся аварийно.

Данные в серверной базе данных обычно физически хранятся на диске в виде одного большого файла, что в сочетании с назначаемыми каждому пользователю паролями и привилегиями существенно повышает защиту данных от намеренной порчи и хищений.

Следующим шагом в усовершенствовании работы с БД было использование технологии клиент-сервер при создании автоматизированной системы учёта и статистической обработки данных переписи населения, занятого в сельскохозяйственном секторе экономики республики. При этом возник целый ряд информационных проблем, связанных с обеспечением надежности хранения и обработки данных.

В настоящей работе предложены следующие решения этих проблем.

Во-первых, повысить степень нормализации данных, вплоть до 5НФ. В сильно нормализованной БД основная часть вопросов, касающихся целостности данных, решается уже на уровне физической структуры БД. Сильно нормализованная БД малочувствительна к внесению новых типов объектов. Кроме того, перед написанием хранимых процедур, обеспечивающих ввод, контроль и корректировку данных, в БД корректировки должны быть определены все типы ошибок по нарушению целостности данных и логических контролей. Во-вторых, было предложено использовать трехуровневую клиент-серверную архитектуру, что позволило выделить бизнес-логику обработки данных в отдельный уровень.

В процессе проектирования программного комплекса были разработаны:

- ◊ алгоритм управления информационными потоками между функционально различными модулями программного комплекса;
- ◊ алгоритм инициализации форм ввода разделов;
- ◊ алгоритм логического контроля над введенными данными;
- ◊ алгоритм учета первоначальных ошибок введенных данных;
- ◊ алгоритм формирования выходных данных (статистических таблиц).

Для обеспечения статистической обработки данных была предложена идея использования хранилищ данных со схемой "звезда" [5] и построена концептуальная модель, обеспечивающая объединение различных классов анализируемых данных в одну однородную структуру. В результате хранилище представляло собой всего один объект, который способен содержать микро-данные статистического обследования в виде единиц учёта со всеми их атрибутами. Схема работы системы баз данных изображена на рисунке.

Основное назначение разработанной БД – корректировка и хранение полной исторической картины о возникновении и корректировке логических ошибок в различных статистических обследованиях.

Кроме того, для обследований, в которых процессы ввода и корректировки данных разделены, база данных корректировки выполняет функции хранилища некорректной информации. Например, если на этапе ввода формуляр переписи не пройдёт логический контроль, его ответы будут помещены в журнал неоткорректированных данных центральной БД корректировки, а не в базу ввода формуляров. Такой подход позволяет сохранять в первичных базах данных только "чистую" (проверенную) информацию [6].

Центральное хранилище содержит микроданные регистра и формуляров в форме, удобной для получения выходных данных. Опытные пользователи, занимающиеся построением отчетов и статистических таблиц, имеют доступ чтения микроданных центрального хранилища.

При проектировании БД центрального хранилища использовался подход моделирования звёздных схем. Следуя этому подходу, сложная структура микроданных регистра и формуляров путём денормализации была трансформирована в несколько "широких" таблиц фактических данных, связанных со справочниками. Громоздкие структуры метаданных, такие как части адресов, имена и организационно-правовые формы, также были приведены к более простым формам.

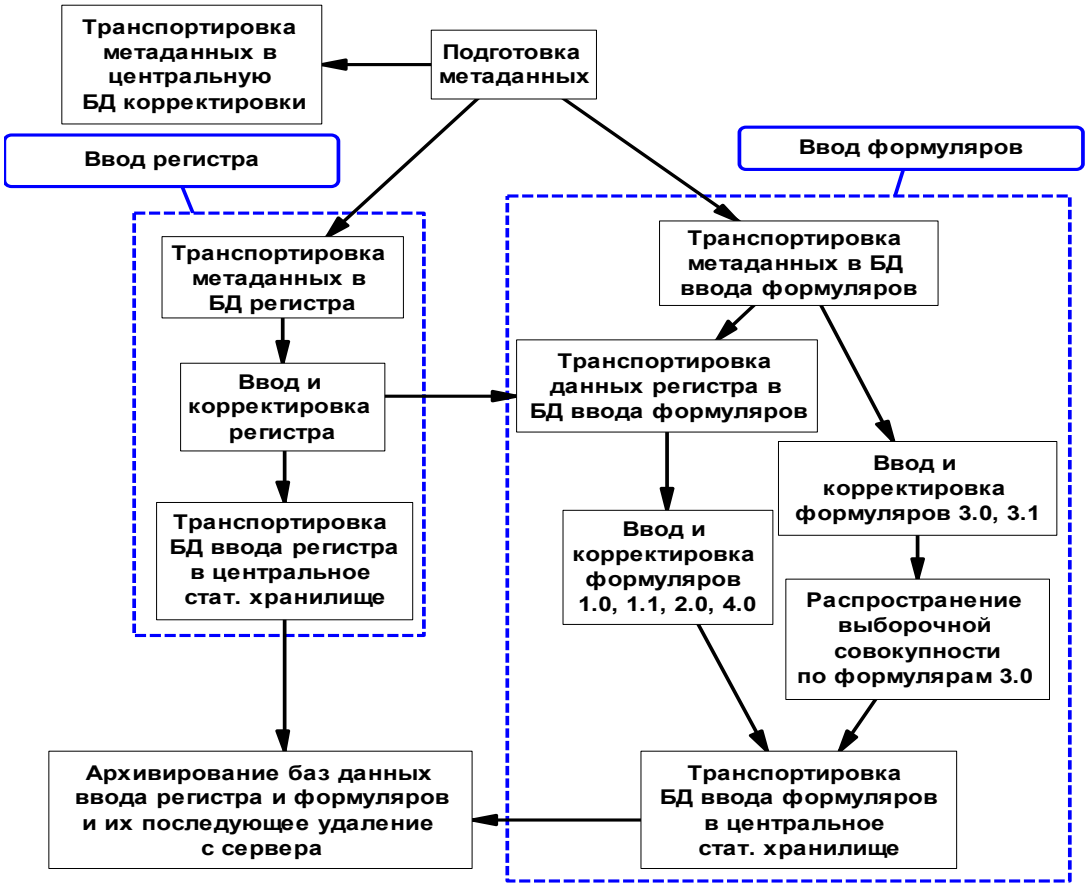
Было сформировано три основных типа таблиц.

1. Таблицы по категориям (с фиксированным боковиком в разрезе территорий).
2. Таблицы по территориям (с плавающим боковиком в разрезе территорий).
3. Группировочные таблицы (с фиксированным боковиком в разрезе территорий и категорий).

Каждая таблица имеет отдельную шапку, которая находится в формате Excel, и отдельную постановку (алгоритм формирования данных для таблицы из данных БД). Получение данных для таблиц реализовано средствами серверной части БД (хранимые процедуры, функции), а преобразование этих данных в формат Excel – средствами клиентской части.

Существует два подхода при проектировании клиентской части:

- ◊ разрабатываются приложения для формирования каждой таблицы, причем значительная часть SQL-кода находится внутри приложения;



Концептуальная схема обработки статистической информации по переписи населения.

◊ разрабатывается одно клиентское приложение для формирования статистических таблиц.

При первом подходе процесс получения статистических таблиц становится очень трудоемким и не дает разработчику возможности сконцентрироваться на какой-либо одной части (клиентской или серверной при внесении изменений в постановку). К тому же усложняется процесс тестирования ПО, а также делается невозможным процесс формирования таблиц многопоточным.

При втором подходе происходит четкое разделение клиентской и серверной частей и становится возможным создание многопоточной системы формирования таблиц. Отпадает необходимость перекомпиляции приложения, за исключением появления нового типа таблиц. Значительно усложняется каждая из частей, клиентская и серверная, но за счет разделения этих частей упрощается процесс тестирования.

Физическая модель статистического хранилища представляет собой один объект, который располагается в одной таблице, полученной из БД путём её денормализации (1НФ-2НФ). По окончании фазы ввода БД сохраняется на долговременном носителе и удаляется с сервера, поэтому статистическое хранилище должно содержать неагрегированные микроданные. Но так как для большинства конечных пользователей требуются агрегированные данные или статистические таблицы, предложено формировать индексированные виды, специальные функции или таблицы с агрегированными данными статистического хранилища. Это позволит квалифицированным пользователям БД получать широкий спектр статистических данных.

В работе предложен новый метод получения 100%-ных данных путём клонирования выборки. Здесь под клонированием данных формуляра подразумевается создание полных копий всех его атрибутов. Данный алгоритм был предложен специалистами Швейцарского вычислительного центра и использует метод случайной выборки с заменой [7].

Основное преимущество клонирования в том, что после такого распространения создаётся иллюзия 100%-ной выборки.

В результате проведенного анализа можно отметить тенденцию к переходу от старых методов обработки данных с использованием одноуровневых БД и языков Clipper, FoxPro, Excel к новым методам обработки и хранения данных (использование архитектуры клиент-сервер и максимальное использование возможностей SQL). В качестве дальнейшего развития современных подходов к обработке данных предполагается перейти на трехуровневую архитектуру БД и применение методов и средств интеллектуального анализа данных, позволяющих получать так называемые скрытые данные из БД.

Литература

1. Классификация приложений для работы с базами данных http://studiodelphin.com/referat/
2. Многоуровневые системы клиент-сервер http://www.infonet.spnet.ru/osp/nets/1997/06/source/72.htm.
3. *Тиори Т., Фрай Дж.* Проектирование структур баз данных. – М.: Мир, 1985.
4. *Уинкуп С.* Microsoft SL Server в подлиннике. – СПб.: BVH, 1998.
5. *Heming C., Halle B.* Handbook of Relational database design. – Addison-Wesley, 1989.
6. *Cocharan W.G.* Sampling Techniques // John Wiley & Sons, 1977.
7. MSDN Library.