УДК 004.652

ПРОЕКТИРОВАНИЕ ХРАНИЛИЩА ДАННЫХ СОЦИАЛЬНОЙ ИНФОРМАЦИИ

Киреев В.А.

Научный руководитель – ст. преподаватель Барков В.И.

Сибирский федеральный университет

В современных условиях применение информационных технологий в муниципальном управлении является неотъемлемой частью управленческого процесса. Данная работа посвящена проектированию системы учета населения. Основное внимание уделено проектированию хранилища данных, содержащего социальную информацию. Под социальной информацией понимается совокупность данных о гражданине, требуемых для социального взаимодействия и осуществления процесса управления.

Можно выделить следующие задачи, которые должна решать система учета населения:

1. Уточнение персональных данных граждан по заданному адресу, уточнение адреса проживания граждан по полным или неполным идентификационным данным.
2. Уточнение предыдущих мест проживания гражданина, его миграцию, уточнение состава граждан, ранее проживавших по указанному адресу, т.е. ретроспективный поиск.
3. Получение количественных оценок жителей заданной территории в заданный период времени, в том числе дифференцированных количественных оценок граждан различного возрастного диапазона, пола, гражданства и т.п.
4. Получение аналитических оценок изменений количества и состава населения за заданный период по заданным территориям.
5. Получение проблемно-ориентированных списков граждан по заданным критериям (списки избирателей, списки потенциальных пенсионеров на ближайшие годы, списки допризывников и т.п.).

Данные задачи невозможно решить только в рамках существующих в Красноярске программных систем. Поэтому была поставлена задача исследовать модели данных, обеспечивающие решение этих задач.

В подсистемах ввода данных, называемых OLTP (Online transaction processing), выполняется операционная (транзакционная) обработка данных. Для реализации этих подсистем используют обычные системы управления базами данных (СУБД). Для реализации подсистемы хранения используют современные СУБД и концепцию хранилищ данных. Подсистема анализа состоит из трех частей: подсистемы информационно-поискового анализа на базе реляционных СУБД и статических запросов SQL (Structured Query Language); подсистемы оперативного анализа на основе технологии оперативной аналитической обработки данных OLAP (Online analytical processing), использующая концепцию многомерного представления данных; подсистемы интеллектуального анализа, реализующей методы и алгоритмы Data Mining («добыча данных»).

Практика использования OLTP-систем показала неэффективность их применения для полноценного анализа информации. Такие системы успешно решают задачи сбора, хранения и поиска информации, но они не удовлетворяют требованиям к анализу

данных. Для объединения в рамках одной системы OLTP-подсистем и подсистем анализа используется концепция хранилищ данных.

В настоящее время преобладают два основных подхода к архитектуре построения хранилищ данных. Это – корпоративная информационная фабрика – CIF (Corporate Information Factory), основоположником которой является Билл Инмон; и хранилище данных с архитектурой шины (Data Warehouse Bus, сокр. BUS), идеологом которой является Ральф Кимболл. Для достижения общей цели Инмон и Кимболл используют разные пути решения проблем сбора информации, управления информацией и аналитики для поддержки принятия решений. Инмон рекомендует начинать с создания крупных централизованных хранилищ данных, сопровождающихся несколькими более мелкими базами данных, служащими аналитическим нуждам (позднее ставшие известными, как «витрины данных»). Кимболл напротив рекомендует начать с создания нескольких витрин данных, которые служат аналитическим потребностям отделов, а затем интегрировать эти витрины данных для согласованности через «информационную шину» (Information Bus). Кроме того, различия в подходах Инмона и Кимболла выражаются в структуре данных. Инмон придерживается реляционной модели в третьей нормальной форме в отличие от Кимболла, который выступает за создание многомерной модели в виде схем «звезда» и «снежинка».

Для решения поставленной задачи – работе с социальной информацией – более подходит «восходящий» подход Ральфа Кимболла, поскольку данные о гражданах собираются различными ведомствами. Так же реляционная модель с третьей нормальной формой затрудняет работу с системой ввиду большого количества данных, которое приводит к долгой работе запросов к СУБД с такими нормализованными данными.

Важным условием работы с социальной информацией является возможность расширять данные за счет информации, получаемой из других источников. В качестве добавляемых сущностей используется концепция измерений многомерных баз данных. Многомерное моделирование предусматривает использование измерений для предоставления максимально возможного контекста для фактов. В отличие от реляционных баз данных, контролируемая избыточность в многомерных базах данных, в общем, считается оправданной, если она увеличивает информационную ценность. Многомерные базы данных рассматривают данные как кубы, которые являются обобщением электронных таблиц на любое число измерений. Факты представляют субъект, который необходимо проанализировать. Факты однозначно определяются комбинацией значений измерений; факт существует только тогда, когда ячейка для конкретной комбинации значений не пуста. Таблица фактов является основной таблицей хранилища данных. Как правило, она содержит сведения об объектах или событиях, совокупность которых будет в дальнейшем анализироваться. Обычно говорят о четырех наиболее часто встречающихся типах фактов – транзакционные факты; факты, связанные с «моментальными снимками»; факты, связанные с элементами документа; факты о событиях или состоянии объекта. Таблица фактов, как правило, содержит уникальный составной ключ, объединяющий первичные ключи таблиц измерений. Таблицы измерений содержат неизменяемые либо редко изменяемые данные. Каждая таблица измерений должна находиться в отношении «один ко многим» с таблицей фактов.

В настоящее время применяются три способа хранения данных: MOLAP (детальные и агрегированные данные хранятся в многомерной базе данных); ROLAP (все данные хранятся в реляционной базе данных); HOLAP (детальные данные хранятся

в реляционной базе данных, а агрегатные – в многомерной). При проектировании хранилища данных социальной информации было решено выбрать структуру ROLAP.

Поскольку личные данные человека в течение его жизни постоянно подвергаются изменениям, требуется, чтобы хранилище данных поддерживало возможность сохранять информацию о событиях и интервалах времени, соответствующих определенным событиям в жизни человека, таких как смена паспорта, фамилии, места жительства и т.п. При переходе к темпоральной базе данных для каждого факта можно указать тот промежуток времени, когда этот факт являлся истинным в моделируемом мире, представленном в базе данных. Подобное представление времени, когда с данными связывается промежуток времени их актуальности, называется *модельным*, или *действительным* (*valid*) временем. Другим типом времени является транзакционное, т.е. время добавления записи или ее удаления из базы данных.

В многомерном моделировании темпоральное расширение обеспечивается медленно меняющимися измерениями (Slowly changing dimensions, SCD), т.е. измерениями, не ключевые атрибуты которых имеют тенденцию со временем изменяться. Всего существует 6 основных типов SCD, которые определяют как в модели отражена история изменений. Тип 0 заключается в том, что данные после первого попадания в таблицу далее не изменяются. Тип 1 – это обычная перезапись старых данных новыми. Тип 2 заключается в создании для каждой версии отдельной записи в таблице с добавлением поля – ключевого атрибута данной версии, например, номер версии, дата изменения или дата начала и конца интервала существования версии. При использовании Типа 3 в самой записи содержатся дополнительные поля для предыдущих значений атрибута, а при получении новых данных, старые данные перезаписываются текущими значениями. Тип 4 заключается в том, что история изменений содержится в отдельной таблице. Гибридный тип или тип 6 (1+2+3)является комбинацией вышеназванных методов и заключается во внесении дополнительной избыточности: берется за основу тип 2, добавляется суррогатный атрибут для альтернативного обзора версий (тип 3), и перезаписывается одна или все предыдущие версии (тип 1).

На Рис. 1 показана схема хранилища данных. Для оперативного и ретроспективного поиска данных используется таблица фактов «История изменений». Таблица фактов «История изменений» по сути является бесфактовой, фактом является само наличие записи в таблице. Однако можно считать фактами модельное и транзакционное время. Здесь используется так называемая схема «снежинка», поскольку некоторые измерения таблицы фактов содержатся в связанных таблицах. Таблица фактов «Количественная оценка населения» позволяет получить количество мужчин, женщин, избирателей и пенсионеров по адресу, гражданству, дате и возрасту. Здесь используется схема «звезда» - таблица фактов является центральной в схеме и соединяет все таблицы измерений. Преимуществами схемы «звезда» благодаря денормализации таблиц являются упрощение восприятия структуры данных пользователем, упрощение формулировки запросов, уменьшение количества операций соединения таблиц при обработке запросов. Недостатком является внесение избыточности данных, что приводит к возрастанию требуемого для их хранения объем памяти. Но на практике оказывается, что при использовании схемы «снежинка» выигрыш по памяти не так велик за счет большого объема таблицы фактов.

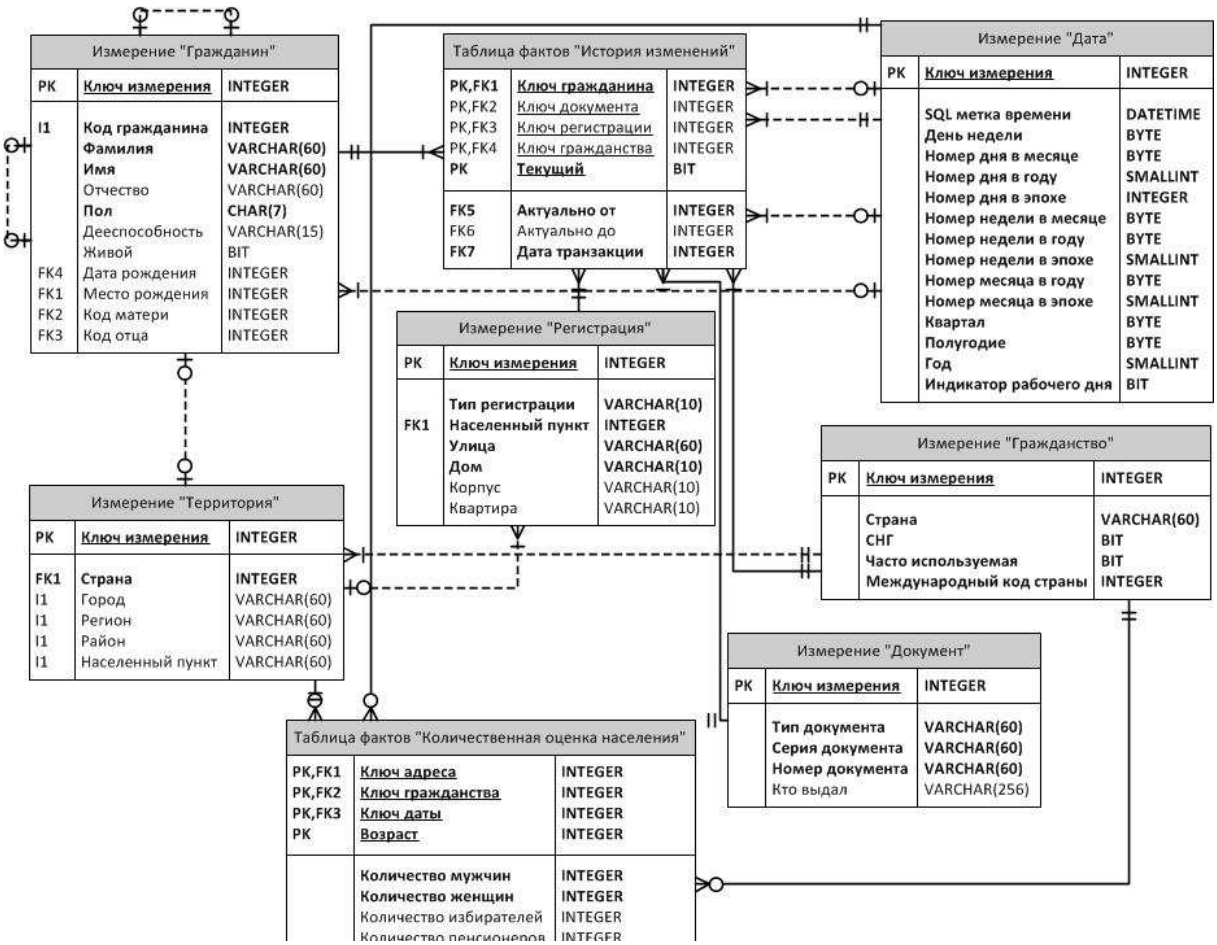

Рис. 1: Схема хранилища данных

В хранилище представлены измерения, моделирующие основную информацию о гражданах, предоставляемую в ФМС:

1. Измерение «Гражданин» - личные данные;
2. Измерение «Документ» - информация о выданных документах;
3. Измерение «Гражданство» - информация о гражданстве;
4. Измерение «Регистрация» - информация об адресе и типе регистрации;
5. Измерение «Адрес» - данные о территории места рождения или регистрации;
6. Измерение «Дата» - данные о дате для темпоральных операций.

Такая структура позволяет обеспечить гибкость при получении нужной информации для последующего анализа, а так же расширяемость хранилища данных за счет добавления социально-значимой информации из других источников, которая бы представлялась в виде новых измерений. Построение хранилища данных является итерационным процессом. Данная структура предоставляет витрину данных для территориальной избирательной комиссии и представляет собой первую итерацию данного процесса. Предполагается, что дальнейшее развитие этой структуры должно привести к созданию единого регистра данных о населении территориальной единицы для обеспечения более успешного управления.