

УДК 004.9:342.56

## ПРОЕКТИРОВАНИЕ ХРАНИЛИЩА ДАННЫХ – ИНФОРМАЦИОННОЙ ОСНОВЫ ИНТЕЛЛЕКТУАЛЬНОГО АНАЛИЗА В СФЕРЕ СУДЕБНОГО ПРОИЗВОДСТВА

Бальзамов А.А.

ГОУВПО «Мордовский государственный университет им. Н. П. Огарева», г. Саранск  
8-927-172-55-96, [anblz@yandex.ru](mailto:anblz@yandex.ru)

**Аннотация.** Рассмотрены вопросы проектирования хранилища данных для целей интеллектуального анализа данных и поддержки принятия решений в сфере судебного производства. Приведен обзор имеющихся программных продуктов в данной сфере.

**Ключевые слова:** судебное производство, система поддержки принятия решения, интеллектуальный анализ данных, оперативная база данных, хранилище данных, нормализованные хранилища, размерностные хранилища, многомерные базы данных.

### Постановка задачи

При построении модели и выборе математических методов системы поддержки принятия решения (СППР) в судебном производстве уже рассматривались вопросы целесообразности автоматизации данного направления<sup>1</sup>. Сложность принятия максимально объективного решения, большая доля субъективности, высокая ответственность лица, принимающего решение, также необходимо отметить как причины использования механизмов СППР при принятии судебных решений. Среди целей были указаны:

1. *Обеспечение более высокой степени объективности и прозрачности процесса принятия решения в судебном производстве за счет формирования системы поддержки принятия решений.*
2. *Разработка механизмов оценки качества принятых решений.*

При дальнейшем изучении предметной области стало очевидно, что в базе данных (БД) для СППР наряду с методом рассуждения по прецедентам<sup>2</sup> (case-based reasoning, CBR), необходимо использовать *методы интеллектуального анализа данных (ИАД, в зарубежных источниках «Data Mining» - «DM»)*. Таким образом, наряду с методом получения решения путем поиска подобных проблемных ситуаций в памяти, хранящей прошлый опыт решения задач, и адаптации найденных решений к новым условиям (CBR), получим механизмы анализа и выявления скрытых закономерностей и взаимосвязей между переменными в массивах данных (DM). Использование методов ИАД даст возможность решать еще одну группу задач и достичь определенной цели:

3. *Аналитическая обработка данных судебного производства для целей улучшения качества принятия решения.*

Решение данной задачи позволит лицу, принимающему решение (ЛПР), основываясь только на знании предметной области и не зная структуры и состава данных, получать необходимую ему информацию. Смысл применения методов ИАД в данном случае кроется в характере получаемой информации:

1. *Не очевидный характер получаемой информации (процесс обнаружения в "сырых" данных ранее неизвестных нетривиальных практически полезных и доступных интерпретации знаний, необходимых для принятия решений<sup>3</sup>).*

2. Широкие возможности по представлению данных (интерпретация, визуализация).

### **Обзор готовых решений ИАД в сфере судебного производства**

В сфере автоматизации процессов судебного производства мировых судей<sup>4</sup> (объект автоматизации – мировые судьи), включая делопроизводство, документооборот, информационно-правовое обеспечение, статистика, банк судебных решений и прочие, существует ряд решений охватывающих одну или несколько типизированных подсистем:

1. Судебное делопроизводство, документооборот, статистика, банк судебных решений: ПК «Астрея» (Агора-Софт), ПК «Мировой судья» (Крок), АРМ «Мировой Судья» (г. Волгоград).

2. Информационно-правовое обеспечение, банк судебных решений: ИПО «Гарант» (Гарант), СПС «КонсультантПлюс» (Консультант), ИПС «Кодекс» (Кодекс).

Решения из обеих групп позволяют в той или иной степени вести обобщения материалов судебной практики. Типичный круг задач:

- поиск по реквизитам материала;
- полнотекстовый поиск;
- сортировка;
- группировка.

В первой группе программ, как правило, информация о судебных решениях хранится в структурированном и формализованном виде (ряд сущностей в базе данных), во второй – полнотекстовые документы на естественном языке, проиндексированные по ограниченному ряду признаков (тематика, название, вид, дата и т.д.). Возможности статистических подсистем первой группы решений ограничена набором фиксированных форм и, в некоторых случаях, возможностью применения нефиксированных OLTP-запросов к БД.

Таким образом, имеющиеся решения не обладают полноценным инструментарием для интеллектуального анализа данных, другими словами, не могут производить новые знания на основе имеющихся в системе. К слову, исторически методы интеллектуального анализа данных первым проникли в область бизнеса. Затраты на их использование с лихвой покрываются получаемой выгодой. В сфере «Электронного государства» современные методы ИАД применяются крайне ограниченно. Хотя их использование иногда просто необходимо там, где процесс принятия решения должностным лицом очень сложен, субъективен и ответственен.

Таким образом, построение системы автоматизации судебного производства с использованием механизмов ИАД и СППР представляется целесообразным и эффективным.

### **Хранилище данных – информационная основа для ИАД**

Стоит отметить, что общая концептуальная архитектура решения, описываемого в данной статье, представляет собой совокупность подсистем<sup>4</sup> (судебное делопроизводство, документооборот, «помощник судьи», информационно-правовая подсистема, статистика, банк судебных решений), которые базируются на единой реляционной базе данных и совокупности внешних источников данных, представленных, как правило, в виде XML или документов на естественном языке (ЕЯ). Очевидно, что для целей ИАД и СППР в основе целесообразно использовать данную БД и необходимые внешние источники. Но, для приведения OLTP-данных к виду, удобному для разностороннего анализа, необходимо промежуточное преобразование данных и создание отдельного *хранилища данных (ХД)* (англ. *Data Warehouse*), нацеленного на содержательный анализ данных, а не на автоматизацию бизнес-процессов.

Целесообразность создания выделенного ХД заключается в следующем:

1. Использование неподготовленных OLTP-данных для интеллектуального анализа приведет к большим расходам ресурсов сервера БД, вследствие появления очень громоздких, ветвящихся и вложенных запросов.

2. Возможность переориентации данных на новые комплексные единицы хранения информации (введение объекта «прецедент»), адекватные задачам принятия решения и анализа.

3. Более гибкая работа с метаданными.

4. Возможность сохранения истории различных данных, изменяющихся во времени.

Таким образом, хранилище данных сформирует промежуточное звено между средствами ИАД, СППР и оперативной БД программного комплекса (рис. 1).

### Выбор модели хранилища данных

Известно, что существуют два архитектурных направления построения ХД - нормализованные хранилища данных и размерностные хранилища<sup>5</sup>.

Проектировщики ХД исторически поделены на два лагеря в соответствии с приверженностью к одной из двух основных схем проектирования. Основным научным фундаментом размерностных схем (Dimensional Modeling) стали работы Ральфа Кимбалла (Ralph Kimball), сторонники нормализованных хранилищ (E-R Modeling) придерживаются основных принципов, сформированных Биллом Инмоном (Bill Inmon)<sup>6</sup>.

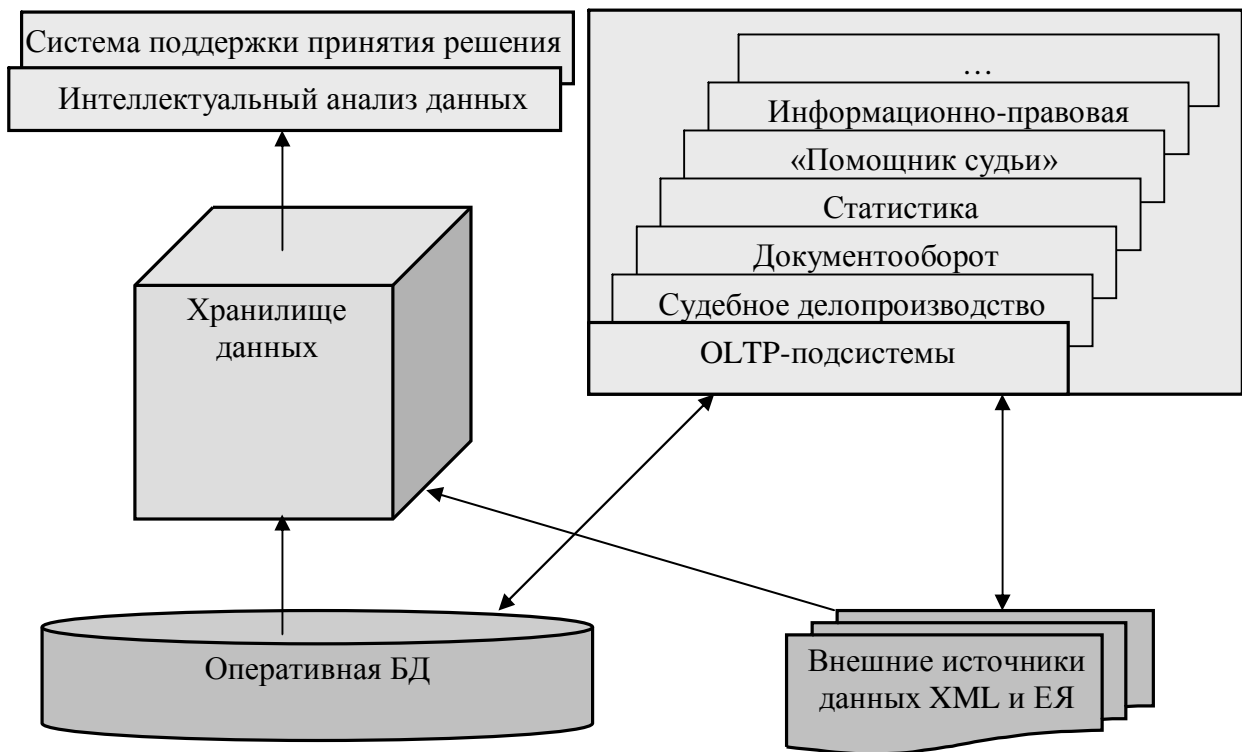


Рисунок 1. Архитектура решения

В нормализованных хранилищах данные находятся в предметно ориентированных таблицах третьей нормальной формы - витрины данных. Нормализованные хранилища характеризуются как простые в создании и управлении, недостатки нормализованных хранилищ - высокая избыточность информации, проблемы с интеграцией данных из нескольких таблиц одновременно. Тем не менее, нормализованные хранилища имеют ряд неоспоримых достоинств:

1. Наиболее правильное (идеализированное) описание предметной области.
2. Поддержка правил данных (в основном, один-ко-многим).
3. В ряде случаев существенно меньший объем ХД.
4. Относительно простая процедура загрузки данных.

Размерностные хранилища используют схему "звезда" или "снежинка". При этом в центре звезды находятся данные (таблица фактов) а размерности образуют лучи звезды.

Различные таблицы фактов совместно используют таблицы размерностей, что значительно облегчает операции объединения данных из нескольких предметных таблиц фактов. Размерности часто создаются в третьей нормальной форме (медленно изменяющиеся размерности), для протоколирования изменения в размерностях. Основными достоинствами размерностных хранилищ являются:

1. Простота и понятность для разработчиков и пользователей.
2. Благодаря более эффективному хранению данных и формализованным размерностям облегчается и ускоряется доступ к данным, особенно при сложном анализе.

Основной недостаток - более сложные процедуры подготовки и загрузки данных, а также изменения размерностей данных.

При выборе типа хранилища необходимо учесть ряд особенностей задачи *аналитической обработки данных судебного производства*:

1. Существование центральных информационных объектов таких, как «судебный прецедент», «судимость», «сторона по делу» и т.д. Наличие возможности определения новых объектов.
2. Наличие негромоздкого интерпретатора запросов ЕЯ. Простота создания новых отчетов для пользователя.
3. Наличие одной основной центральной базы OLTP-данных и набора внешних источников.

Исходя из обозначенных особенностей адекватным задаче видится использование размерностного хранилища, использующего схему «звезда» (или «снежинка» при необходимости переопределения размерностей).

### **Проектирование размерностного ХД**

Стоит отметить, что проектирование размерностных ХД имеет в своей основе принципы проектирования многомерных БД. Исходя из концепции построения такого рода хранилищ необходимо выделить *центральную таблицу фактов* (или несколько таблиц) и *таблицы размерностей*, которые идентифицируют принадлежность фактов определенным категориям, указанным в таблицах размерностей. Также возможно наличие таблиц, которые идентифицируют не факты, а размерности более низкого уровня (хотя в классической теории многомерных БД такое не допускается). В методологии Ральфа Кимбала указана необходимость проектирования и документирования *шины хранилища данных* (англ. data warehouse bus) в виде матрицы. Ее строки представляют бизнес-процессы организации, а столбцы - согласованные измерения. В ячейках матрицы - факты того, имеет ли отношение данное измерение к данному бизнес-процессу. Наряду с проектированием ХД необходимо разработать механизмы для *извлечения данных* из систем-источников, преобразования данных и их загрузки в целевую базу данных (англ. ETL - Extract, transform, and load ).

Для формирования *центральной таблицы фактов* выберем базовую степень детализации таблицы фактов. Как правило, любая таблица фактов по Кимбалу базируется либо на уровне транзакции (transaction grain), либо периодического снимка (periodic snapshot grain), либо аккумулирующего снимка (accumulating snapshot grain)<sup>7</sup>. При решении нашей задачи будем использовать аккумулирующий снимок судебного прецедента, который будет включать в себя различные факты судебного производства: общая информация о деле; ссылки на стороны по делу, на информацию о составе преступления (правонарушения), на события по делу, на процессуальные меры (пресечения, обеспечения производства), на рассмотрение и решение по делу; на результат дела в отношении обвиняемого (правонарушителя) и т.д. Вопреки классической теории построения размерностных ХД на концептуальном уровне выделим две основные таблицы фактов:

1. Информация о судебном деле.
2. Информация о стороне по делу.

При формировании *таблиц размерностей* кроме основных следует учитывать и служебные (измерение «аудит» для контроля доступа к данным ХД). По отношению к

обработке изменений записей медленно меняющихся измерений в проектировании многомерных моделей имеются два основных способа:

1. Новые значения атрибутов измерений записываются поверх старых. Таким образом, при этом типе обработки история изменений теряется (SCD1). Это простые измерения: «характеристики стороны по делу», «признаки состава преступления», «характеристики субъекта» и т.д.

2. При изменении атрибута измерения в таблицу измерения вставляется новая запись с новым значением суррогатного ключа (SCD2). Данная обработка изменений характерна для таких измерений, как «процессуальные меры», «состояние дела», «решение по делу» и т.д.

Стоит отметить, что все фиксированные формы судебной отчетности (или отдельные показатели) центрируются вокруг двух объектов: «Дело», «Сторона по делу». Совокупность фактов по делу и фактов о стороне по делу назовем «прецедент». Исходя из этого на концептуальном уровне сгруппируем размерности вокруг двух таблиц фактов (рис. 2).

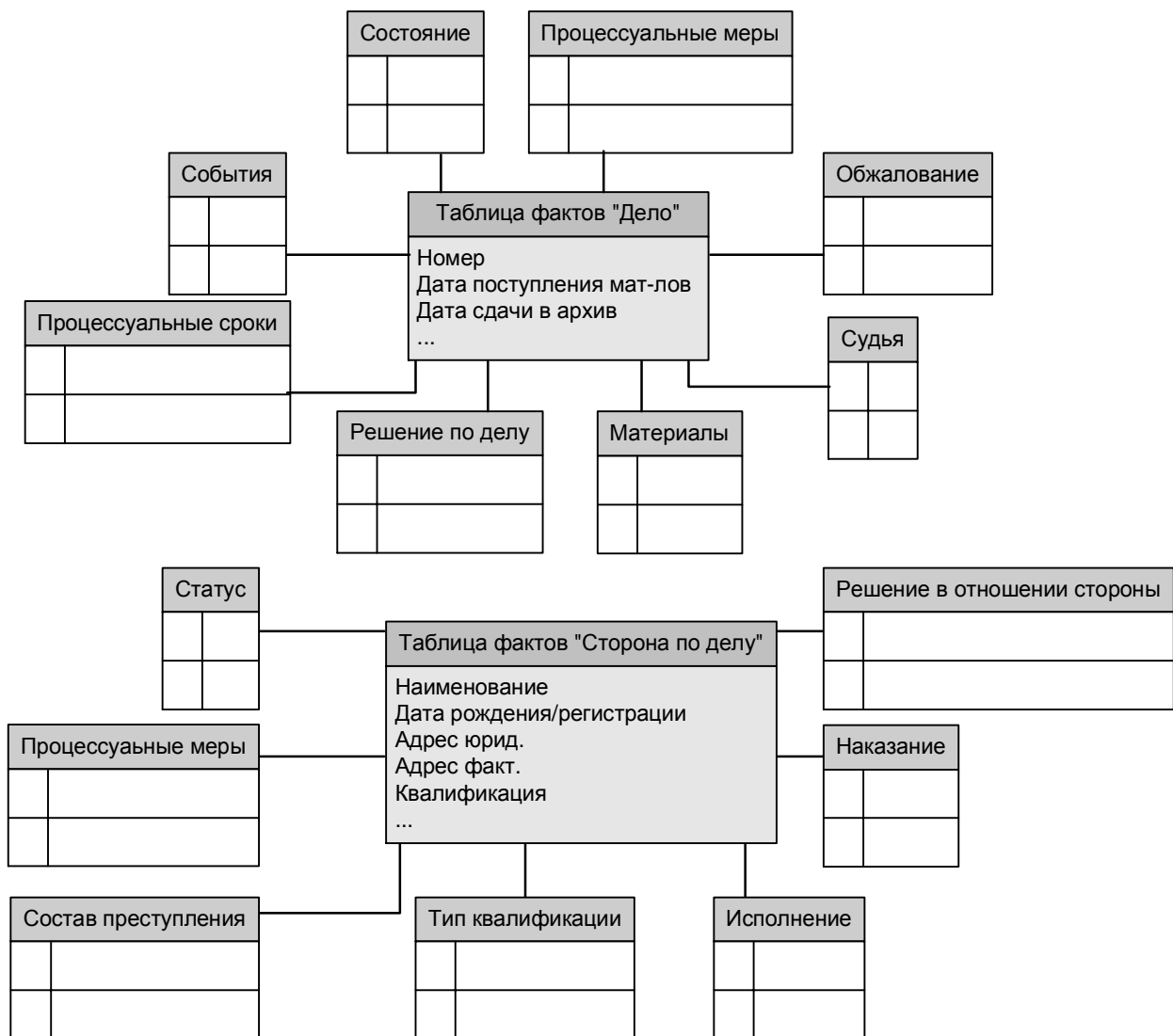


Рисунок 2. Фрагмент концептуальной модели ХД

Этап проектирования ХД очень важен, так как ошибки данного этапа обходятся дороже всего. Хранилище, некорректно отображающее предметную область, сделает применения методов ИАД затруднительным и в ряде случаев даст неверные результаты. Поэтому важно максимально продуктивно использовать имеющиеся методологии, так как

они содержат обобщенные рекомендации и опыт проектирования ХД для различных областей.

### Литература

- <sup>1</sup> Бальзамов А.А. Построение модели и выбор методов СППР в судебном производстве. - Электронное издание "Электроника и информационные технологии", 2008, выпуск № 2 (4). зарегистрировано 11.01.2009 под номером 0420800067/0020 - <http://fetmag.mrsu.ru> – 8с.
- <sup>2</sup> Бальзамов А.А. Бальзамов А.Ю. Применимость СВР-метода для СППР судебного производства. - XXXVI Огаревские чтения: Мат-лы научн. конф. в 3 ч. ч3. Технические науки. – Саранск: Изд-во Мордов. ун-та, 2009. с.142–143
- <sup>3</sup> Data Mining – добыча данных. – [http://www.basegroup.ru/library/methodology/data\\_mining/](http://www.basegroup.ru/library/methodology/data_mining/)
- <sup>4</sup> Бальзамов А.А., Бальзамов А.Ю. Разработка автоматизированной системы судебного делопроизводства. - Наука и инновации в Республике Мордовия. Мат-лы VII респ. научн.-практ. конф. (8–13 февр. 2008 г.). – Саранск, 2008. с.564 – 568
- <sup>5</sup> Хранилище данных. Библиографическая статья. – <http://ru.wikipedia.org>
- <sup>6</sup> Dimensional Modeling and E-R Modeling In The Data Warehouse By Joseph M. Firestone, Ph.D. White Paper No. Eight. June 22, 1998
- <sup>7</sup> Ralph Kimball, Ph.D. Fundamental Grains. Intelligent Enterprise Magazine. March 30, 1999 Volume 2 Number 5 – [http://intelligent-enterprise.informationweek.com/db\\_area/archives/1999/993003/warehouse.jhtml](http://intelligent-enterprise.informationweek.com/db_area/archives/1999/993003/warehouse.jhtml)