Редактор *В. П. Мирошкина*

<https://studopedia.org/2-15554.html>

*Корректор В. П. Мирошкина*

Подготовка оригинал-макета*В. П. Мирошкина*

Макет обложки ?????

Подписано в печать 22.03.12 г.

Формат 60х84 1/16. Гарнитура «Таймс»

Усл. печ. л. 15,4. Уч.-изд. л. 15,0.

Тираж 50 экз. Заказ ………

Отпечатано с готового оригинал-макета

в типографии «Папирус»

300041, г. Тула, ул. Ленина, 12

**РОССИЙСКОЙ ФЕДЕРАЦИИ**

**Федеральное государственное**

**Автономное образовательное учреждение**

**Высшего профессионального образования**

**“Южный федеральный университет”**

**ИССЛЕДОВАНИЕ МЕТОДОВ ОРГАНИЗАЦИИ**

**И РАБОТЫ С ХРАНИЛИЩАМИ ДАННЫХ**

**Методические указания**

**К выполнению лабораторных работ**

**По дисциплине**

**ПРОЕКТИРОВАНИЕ ИНФОРМАЦИОННЫХ СИСТЕМ**

**Часть 2**

Для студентов специальности 080801

**Прикладная информатика (в экономике)**

и направления 230700

**Прикладная информатика**



Таганрог 2012

УДК ХХХХХХХ

Составитель Г.И. Гончаренко

Исследование методов организации и работы с хранилищами данных: Методические указания к выполнению лабораторных работ по дисциплине «Проектирование информационных систем». Ч. 2. – Таганрог: Изд-во ЮФУ, 2012. – 32 с.

В методических указаниях рассмотрены основные моменты, связанные с проектированием и реализацией хранилищ данных (ХД) в информационных системах. Рассмотрены различные варианты архитектуры ХД, современные средства моделирования и проектирования ХД и витрин данных, вопросы организации аналитических исследований в рамках систем поддержки принятия решений с использованием ХД. Приведено содержательное описание порядка выполнения лабораторных работ. Требования к заданиям и варианты приведены в соответствии с тематикой выполнения первой части работ. Предназначено для студентов специальности 080801 «Прикладная информатика (в экономике)»и направления 230700 «Прикладная информатика».

Ил.: 12. Библиограф.: 6 назв.

Рецензент Кавчук А.А., канд. техн. наук, доцент кафедры САиТ ЮФУ

Один из крупнейших специалистов в области хранилищ данных Билл Инмон, охарактеризовал Хранилища Данных (ХД) как: "предметно-ориентированные, интегрированные, неизменчивые, поддерживающие хронологию наборы данных, организованные для целей поддержки управления", обеспечивающие менеджеров и аналитиков достоверной информацией, необходимой для анализа и принятия решений [1].

 По Инмону данные организованы по предметным областям. Все данные о некотором предмете (бизнес-объекте) собираются (обычно из множества различных источников), очищаются, согласовываются, дополняются, агрегируются и представляются в единой, удобной для их использования в бизнес-анализе форме. На рис. 1.1 показан пример того, как данные могут быть организованы по предметной области Продажи.

Рис. 1.1. Пример организации данных по предметным областям

Подобная организация данных носит название схемы «Звезда». Авторство такого варианта структуры данных принадлежит Р. Кодду, понимавшему недостатки предложенной им реляционной модели по отношению к многомерным данным и данным для аналитических исследований. Схема «Звезда» обычно содержит одну большую таблицу, называемую таблицей факта (*fact table)*, помещенную в центр, и окружающие ее меньшие таблицы, называемые таблицами измерений (*dimensional table)*, соединенными c таблицей факта в виде звезды радиальными связями. В этих связях таблицы измерений являются родительскими, таблица факта - дочерней.

Данные при этом должны быть интегрированы или согласованны в соответствии с теорией моделирования и проектирования ХД, авторство которой, наряду с вышеупомянутым Инмоном, делит еще один известный ученый Ральф Кимболл.

При этом данные, поступающие в хранилище из различных оперативных приложений и других источников, необходимо привести к некоторому общему формату. Возможно, потребуется произвести «чистку» и «проверку» оперативной информации, полученной в режиме реального времени. Данные могут быть избыточными, храниться одновременно в нескольких местах, быть синхронизированными или несинхронизированными между собой и иметь противоречивое представление. Например, в отделениях международного банка в разных странах сведения о счетах могут храниться в различных базах данных и отражать, естественно, состояние дел в национальной валюте. Для проведения общего анализа эти данные необходимо привести к одной и той же валюте и представить единый обменный курс на определенный момент времени. Должны быть устранены такие аномалии, как повторение одного и того же имени у разных данных, использование разных имен для одних и тех же данных и т.д.

|  |  |
| --- | --- |
| https://avatars.mds.yandex.net/get-direct/329956/qttXQopYdpKZ_SPQjeQbCw/x180**[Создай сайт](https://an.yandex.ru/count/7YkxMneYAW450Au1CPI6jre00000EFIz6K02I09Wl0Xe172-qO3X1801pjVQbGM80QYHuyaWa07MjB6v39W1seMQZH2W0RIlghOCg064xvsB4BW1thQ4eml00GBO0SRJtGZW0Vp9hGRe0Khu0Th2thu1Y083e0A6wja6kGBU1-V8TD_StV02ee-kkWlu0eA0W82Gi2Q00_sFgDKCY0FrjvNm3gW31h031EW4uWlu1AwE5uW5heuNa0NCoHUW1S7t1gW5aF05i0MGy0Mu1PV_1S05pSaNo0MO_mNG1RhI0k05HF050PW6WD2iymwW1fAX0QW6ag41oGPzoOetRY3wAj46VLt0hcuW-YhP1W000F1C0000gGSp7A0mu0545B07W82GDC07mhRqnW_G1mBW1uOAW0W8q0YwYe21m910u0Y3zQm3W0e1mGe00000003mFzWA0k0AW8bw-0h0_1M82mQg2n2zSE8KCnm000wcHk4wQ0K0m0k0emN82u3Kam7P2pCSe33W0KGKw0kkZXVm2mQ83Dh2thu1w0m2yWq0-Wq0WWu0a0x0X3sW3i24FG00?test-tag=168779698935905&stat-id=3&" \t "_blank)**[самостоятельно!](https://an.yandex.ru/count/7YkxMneYAW450Au1CPI6jre00000EFIz6K02I09Wl0Xe172-qO3X1801pjVQbGM80QYHuyaWa07MjB6v39W1seMQZH2W0RIlghOCg064xvsB4BW1thQ4eml00GBO0SRJtGZW0Vp9hGRe0Khu0Th2thu1Y083e0A6wja6kGBU1-V8TD_StV02ee-kkWlu0eA0W82Gi2Q00_sFgDKCY0FrjvNm3gW31h031EW4uWlu1AwE5uW5heuNa0NCoHUW1S7t1gW5aF05i0MGy0Mu1PV_1S05pSaNo0MO_mNG1RhI0k05HF050PW6WD2iymwW1fAX0QW6ag41oGPzoOetRY3wAj46VLt0hcuW-YhP1W000F1C0000gGSp7A0mu0545B07W82GDC07mhRqnW_G1mBW1uOAW0W8q0YwYe21m910u0Y3zQm3W0e1mGe00000003mFzWA0k0AW8bw-0h0_1M82mQg2n2zSE8KCnm000wcHk4wQ0K0m0k0emN82u3Kam7P2pCSe33W0KGKw0kkZXVm2mQ83Dh2thu1w0m2yWq0-Wq0WWu0a0x0X3sW3i24FG00?test-tag=168779698935905&stat-id=3&" \t "_blank)**Сайт** своими руками. Дарим 1000 рублей на счёт! [umi.ru](https://an.yandex.ru/count/7YkxMneYAW450Au1CPI6jre00000EFIz6K02I09Wl0Xe172-qO3X1801pjVQbGM80QYHuyaWa07MjB6v39W1seMQZH2W0RIlghOCg064xvsB4BW1thQ4eml00GBO0SRJtGZW0Vp9hGRe0Khu0Th2thu1Y083e0A6wja6kGBU1-V8TD_StV02ee-kkWlu0eA0W82Gi2Q00_sFgDKCY0FrjvNm3gW31h031EW4uWlu1AwE5uW5heuNa0NCoHUW1S7t1gW5aF05i0MGy0Mu1PV_1S05pSaNo0MO_mNG1RhI0k05HF050PW6WD2iymwW1fAX0QW6ag41oGPzoOetRY3wAj46VLt0hcuW-YhP1W000F1C0000gGSp7A0mu0545B07W82GDC07mhRqnW_G1mBW1uOAW0W8q0YwYe21m910u0Y3zQm3W0e1mGe00000003mFzWA0k0AW8bw-0h0_1M82mQg2n2zSE8KCnm000wcHk4wQ0K0m0k0emN82u3Kam7P2pCSe33W0KGKw0kkZXVm2mQ83Dh2thu1w0m2yWq0-Wq0WWu0a0x0X3sW3i24FG00?test-tag=168779698935905&stat-id=3&" \t "_blank)[Адрес и телефон](https://an.yandex.ru/count/7YkxMqCjkTC50Au1CPI6jre00000EFIz6K02I09Wl0Xe172-qO3X1801pjVQbGM80QYHuyaWa07MjB6v39W1seMQZH2W0RIlghOCg064xvsB4BW1thQ4eml00GxO0SRJtGZW0Vp9hGRe0Khu0Th2thu1Y083e0A6wja6kGBU1-V8TD_StV02ee-kkWlu0eA0W82Gi2Q00_sFgDKCY0FrjvNm3gW31h031EW4uWlu1AwE5uW5heuNa0NCoHUW1S7t1gW5aF05i0MGy0Mu1PV_1S05pSaNo0MO_mNG1RhI0k05HF050PW6WD2iymwW1fAX0QW6ag41oGPzoOetRY3wAj46VLt0hcuW-YhP1W000F1C0000gGSp7A0mu0545B07W82GDC07mhRqnW_G1mBW1uOAW0W8q0YwYe21m910u0Y3zQm3W0e1mGe00000003mFzWA0k0AW8bw-0h0_1M82mQg2n2zSE8KCnm000wcHk4wQ0K0m0k0emN82u3Kam7P2pCSe33W0KGKw0kkZXVm2mQ83Dh2thu1w0m2yWq0-Wq0WWu0a0x0X3sW3i24FG00?test-tag=168779698935905&stat-id=3&)Скрыть рекламу:Не интересуюсь этой темойТовар куплен или услуга найденаНарушает закон или спамМешает просмотру контента

|  |
| --- |
| Спасибо, объявление скрыто. |

 |

[Яндекс.Директ](https://direct.yandex.ru/?partner)

Так, на рис. 1.2 показано, как можно привести значение данных к единому формату на примере четырех баз данных, в которых обозначение пола имело разное представление.

Необходимо не забывать, что понятие согласованности, лежит в основе достоверности анализа данных. Пока не проведена работа по взаимному согласованию значений данных из различных источников, сложно говорить об их достоверности.



*Рис. 1.2. Пример приведения данных к единому формату*

*Информация в аналитических информационных системах (АИС) стабильна. Данные загружаются в хранилище и затем предоставляются системе поддержки принятия решений (СППР). Важно отметить существенную разницу между оперативными данными и данными, пригодными для анализа. Оперативные, или транзакционные данные, составляющие основу так называемых OLTP – систем, – это детальная информация о сделанных заказах, выписанных счетах, денежных переводах и т.п. Они предназначены для приложений, которые выполняют повседневные задачи. Поэтому оперативные данные быстро и часто изменяются, существуют в реальном времени (каждое последующее значение заменяет предыдущее), тесно привязаны к конкретному приложению и, как правило, имеют сложный для понимания конечным пользователем формат. В отличие от оперативных данных информация в хранилище меняется тогда, когда этого захочет пользователь. Она загружается в хранилище через определенные промежутки времени и является непротиворечивой благодаря проведенным преобразованиям оперативных данных .*

*Информация, представленная в ХД, отражает историю изменения данных. Хранилище данных представляет собой последовательность моментальных снимков корпоративной информации через определенные, заранее заданные промежутки времени. Хранилище может пополняться ежедневно, еженедельно или ежемесячно - это определяется в процессе его организации. Главное, что аналитик получает не только абсолютное значение величины, но и возможность проследить историю ее изменения за определенный период времени. Маркетинг-менеджер, например, сможет посмотреть, как шла реализация товара в течение трех лет, анализируя данные о продажах, которые накапливались в хранилище к концу каждой рабочей недели.*

*С позиций накопления производственного опыта в интересах его последующего использования в интересах поддержки принятия решений в АИС его хранение только в транзакционных системах имеет следующие очевидные недостатки:*

*- распределенное хранение в нескольких, как правило, разнородных БД существенно затрудняет использование этих данных из-за сложности их одновременной выборки из нескольких источников, возможных пересечений перечней хранимых данных и (или) нарушений их согласованности (целостности);*

*- запросы на выборку ретроспективных данных снижают реактивность транзакционных систем;*

*- глубина ретроспективного анализа определяется алгоритмами работы транзакционных систем.*

*В силу указанных недостатков, а также принципиальных различий между системами OLTP и АИС (см. рис. 1.3 и табл. 1) данные транзакционных систем практически бесполезны для непосредственного применения в АИС. Таким образом, складывается парадоксальная ситуация – "отсутствие информации при ее наличии или даже избытке".*

|  |
| --- |
|   |
|   | http://konspekta.net/studopediaorg/baza2/578255622666.files/image008.gif |

*Рис. 1.3.Отличие хранилища данных от обычной базы данных*

*Таблица 1*

# Архитектура хранилища данных

[⇐ Предыдущая](https://studopedia.org/2-15555.html)[1](https://studopedia.org/2-15554.html)[2](https://studopedia.org/2-15555.html)**3**[4](https://studopedia.org/2-15557.html)[5](https://studopedia.org/2-15558.html)[6](https://studopedia.org/2-15559.html)[7](https://studopedia.org/2-15560.html)[8](https://studopedia.org/2-15561.html)[Следующая ⇒](https://studopedia.org/2-15557.html)

К проектированию хранилищ данных обычно предъявляются следующие требования:

· структура данных хранилища должна быть понятна пользователям;

· должны быть выделены статические данные, которые регулярно модифицируются: ежедневно, еженедельно, ежеквартально;

· должны быть упрощены требования к запросам с целью исключения запросов, которые могли бы требовать множественных утверждений SQL в традиционных реляционных СУБД;

· должна быть обеспечена поддержка сложных запросов SQL, которые требуют последовательной обработки тысяч или миллионов записей [1].

ХД это огромный проект, включающий в себя согласование между многими частями организации. Соответствующая архитектура может помочь созданию проекта. На рис. 2.1 приводится распространенный вариант двухуровневой архитектуры построения хранилища данных.

В рамках этой архитектуры данные из операционной СУБД проходят через процесс трансформации (ETL – процесс) и затем загружаются в хранилище данных через сервер хранилища данных. Доступ к хранилищу данных для извлечения данных для принятия решений осуществляется через сервер.

Трансформационный процесс достаточно трудоемкий. Для него необходимо иметь четкое понятие о предметной области хранилища данных. Для лучшего понимания предметной области необходимо начать построение хранилища данных с концептуальной модели.

|  |
| --- |
|   |
|   | http://konspekta.net/studopediaorg/baza2/578255622666.files/image009.gif |

Рис. 2.1. Двухуровневая архитектура построения хранилищ данных

**Преимущества подобной архитектуры заключаются в следующем:**

* данные хранятся в единственном экземпляре;
* минимальные затраты на хранение данных;
* отсутствуют проблемы, связанные с синхронизацией нескольких копий данных;
* данные консолидируются на уровне предприятия, что позволяет иметь единую картину бизнеса.

**Недостатки двухуровневой архитектуры:**

* данные не структурируются для поддержки потребностей отдельных пользователей или групп пользователей;
* возможны проблемы с производительностью системы;
* возможны трудности с разграничением прав пользователей на доступ к данным.

**В случае выбора архитектуры СППР на основе трехуровневого хранилища данных, последнее представляет собой единый централизованный источник информации для всего предприятия. Витрины данных отражают подмножества данных из хранилища, организованные для решения задач отдельных подразделений компании. Конечные пользователи имеют возможность доступа к детальным данным хранилища, в случае если данных в витрине недостаточно, а также для получения более полной картины состояния бизнеса (рис. 2.2).**

|  |  |
| --- | --- |
| https://avatars.mds.yandex.net/get-direct/238679/NZgmDo-bz75yvCJ3dsYPfw/wy150[**Joomla** 3! Освоить легко и быстро](https://an.yandex.ru/count/NA2O0HBCySa50BG1CQw6jre00000EFIz6K02I09Wl0Xe1720mEo02e01o-hc0eW1h96rjY-G0Ro8fCqhc07ayRgJAQ01ieYapIke0S3nkfCfk06AdDQN6C010jW1kBpb4k01ZC-F0-W1dW7u0G680WEW0kwnfGAv0hl8tZhqE9Nqy0B7tjN31lW2We20W93G8u03zu3EsWk80_VeefaEc0E7mGMe0mQm0mIu1Fy1w0IQ2_W4vemPY0NcZ1cG1RsL6g05wyC7g0Nei0Qm1UYm1hW5oDC6m0MzbHh81SZJ1j05oEy2u0Lxy0K1c0Rk_veAe0QIeG6e1fAX0Sa6VScADsuW-YhH1drTmAvk8FegsGO0003mJ0000907SAa7Cno0WHe8H1Im1u20a3J01yAszCOFq0S2u0U62e082D08keg0WO0Gu0Zts981W0e1mGe00000003mFzWA0k0AW8bw-0h0_1M82mQg2n300zoRCnm009GwrU8wQ0K0m0k0emN82u3Kam7P2pCSW84Q24GKw0lcZ1dm2mQ83EItuRu1w0m2yWq0-Wq0WWu0a0x0X3sW3i24FG00?test-tag=156685071030369&stat-id=3&)Хотите научиться работать с **Joomla** 3? Начните с этого курса. 6+[alex-kurteev.ru](https://an.yandex.ru/count/NA2O0HBCySa50BG1CQw6jre00000EFIz6K02I09Wl0Xe1720mEo02e01o-hc0eW1h96rjY-G0Ro8fCqhc07ayRgJAQ01ieYapIke0S3nkfCfk06AdDQN6C010jW1kBpb4k01ZC-F0-W1dW7u0G680WEW0kwnfGAv0hl8tZhqE9Nqy0B7tjN31lW2We20W93G8u03zu3EsWk80_VeefaEc0E7mGMe0mQm0mIu1Fy1w0IQ2_W4vemPY0NcZ1cG1RsL6g05wyC7g0Nei0Qm1UYm1hW5oDC6m0MzbHh81SZJ1j05oEy2u0Lxy0K1c0Rk_veAe0QIeG6e1fAX0Sa6VScADsuW-YhH1drTmAvk8FegsGO0003mJ0000907SAa7Cno0WHe8H1Im1u20a3J01yAszCOFq0S2u0U62e082D08keg0WO0Gu0Zts981W0e1mGe00000003mFzWA0k0AW8bw-0h0_1M82mQg2n300zoRCnm009GwrU8wQ0K0m0k0emN82u3Kam7P2pCSW84Q24GKw0lcZ1dm2mQ83EItuRu1w0m2yWq0-Wq0WWu0a0x0X3sW3i24FG00?test-tag=156685071030369&stat-id=3&)Скрыть рекламу:Не интересуюсь этой темойТовар куплен или услуга найденаНарушает закон или спамМешает просмотру контента

|  |
| --- |
| Спасибо, объявление скрыто. |

 |

[Яндекс.Директ](https://direct.yandex.ru/?partner)

   

**

*Рис. 2.2. Трехуровневая архитектура построения хранилищ данных*

***Eе преимущества заключаются в том, что:***

* *создание и наполнение витрин данных упрощено, поскольку наполнение происходит из единого стандартизованного надежного источника очищенных нормализованных данных;*
* *имеется (построена) общая структура данных предприятия;*
* *витрины данных синхронизированы и совместимы с общей структурой данных предприятия;*
* *существует возможность сравнительно легкого расширения хранилища и добавления новых витрин данных;*
* *гарантированная производительность.*

***Недостатки трехуровневой архитектуры:***

* *избыточность данных, ведущая к росту объема хранилища;*
* *требуется согласованность с принятой архитектурой многих областей с потенциально различными требованиями (например, скорость внедрения иногда конкурирует с требованиями следовать архитектурному подходу).*

*Тем не менее, несмотря на указанные недостатки, последняя конструкция наилучшим образом отвечает возможности решения самых разнообразных аналитических и управленческих задач в любой сфере деятельности.*

*Очевидно, что для принятия качественного решения необходим достаточно полный и глубокий анализ данных, который просто невозможно провести без специального программного обеспечения. Конечно, можно было бы воспользоваться и более простыми программными продуктами такими, как электронные таблицы. Но это годилось бы только для единичных решений, имеющих место один раз в большой промежуток времени.*

*Предприятие, как правило, ставит перед собой задачу консолидации – накопления – данных различных отделов, для того, чтобы в будущем избежать проволочек с ее получением, а также для ее своевременного обновления. Кроме того, принятие решения в условиях современности предполагает автоматизацию всего процесса анализа данных, начиная от сбора данных и заканчивая формированием необходимой аналитической отчетности. И, наконец, самым важным в формировании механизма принятия решений является возможность поставить его на поток, т. е. обеспечить неоднократное его использование.*

*Все эти функции сочетают в себе аналитические платформы. Сегодня на рынке программного обеспечения предлагается большое количество таких программных решений. В настоящих методических указаниях мы остановимся на одном из них, а именно платформе Deductor Studio Academic 5.2. Поскольку предоставление полного лицензионного пакета осуществляется на коммерческой основе, работа с данной платформой будет осуществляться в рамках учебной версии.*

*Кроме возможности работать с единым источником информации, руководители и аналитики должны иметь удобные средства визуализации данных, агрегирования, поиска тенденций, прогнозирования. Несмотря на многообразие аналитической деятельности, можно выделить типовые технологии анализа данных, каждой из которых соответствует определенный набор инструментальных средств. Вместе с хранилищем данных эти средства обеспечивают полное решение для автоматизации аналитической деятельности и создания корпоративной информационно-аналитической системы.*

|  |
| --- |
|  |

# Краткое описание порядка работы с DEDUCTOR Studio

[⇐ Предыдущая](https://studopedia.org/2-15556.html)[1](https://studopedia.org/2-15554.html)[2](https://studopedia.org/2-15555.html)[3](https://studopedia.org/2-15556.html)**4**[5](https://studopedia.org/2-15558.html)[6](https://studopedia.org/2-15559.html)[7](https://studopedia.org/2-15560.html)[8](https://studopedia.org/2-15561.html)[Следующая ⇒](https://studopedia.org/2-15558.html)

Архитектуру платформы Deductor можно изобразить в виде следующей схемы, представленной на рис. 2.3.

|  |  |
| --- | --- |
|

|  |
| --- |
| Deductor Studio |

 |

|  |  |
| --- | --- |
|

|  |
| --- |
| Deductor Warehouse |

 |

Рис. 2.3. Архитектура системы Deductor

Deductor состоит из двух компонентов – аналитического приложения Deductor Studio и многомерного хранилища данных Deductor Warehouse.

Кроме того, выделяют еще несколько модулей – Deductor Viewer, Deductor Server и Deductor Client. Они представляют собой соответственно рабочее место конченого пользователя, сервер для удаленной аналитической обработки и библиотеку для работы с Deductor Server.

**Deductor Warehouse («хранилище»)** – это многомерное хранилище данных, аккумулирующее всю необходимую для анализа предметной области информацию. Физически оно представляет собой реляционную базу данных, которая содержит таблицы для хранения информации и таблицы связей, обеспечивающих целостное хранение сведений.

Методология хранилища такова, что информация хранится в виде процессов, каждый из которых имеет определенный набор измерений и фактов. Измерения могут быть простыми списками, например, дата, либо содержать дополнительные столбцы – атрибуты. Такая схема хранения получила название «Снежинка» или «Звезда».

Пример структуры типа «Звезда» приведен на риc. 1.1. Подобная схема реализуется посредством перевода реляционных таблиц в многомерное представление с помощью специального семантического слоя. Это механизм, при помощи которого пользователь может заменить исходные бизнес-термины (например, «продукт», «клиент», «продажа») в операции с базой данных, и наоборот. Это очень удобно, поскольку от пользователя не требуется знание структуры хранения данных или языка запросов. Выбор информации из хранилища осуществляется с помощью Мастера Импорта. С его помощью пользователь сам выбирает нужные ему данные. Система сама формирует специфичный для каждой СУБД (будь то Oracle, MS SQL или Firebird).

Загрузка данных в Deductor Warehouse осуществляется с помощью интегрированных модулей – Deductor Studio и/или Deductor Server. Схему доступа к хранилищу данных можно изобразить как на рис. 2.4.

|  |  |
| --- | --- |
|

|  |
| --- |
| Deductor Studio |

 |



|  |  |
| --- | --- |
|

|  |
| --- |
| Deductor Viewer |

 |



|  |
| --- |
|   |
|   | http://konspekta.net/studopediaorg/baza2/578255622666.files/image017.gif |

|  |  |
| --- | --- |
| https://avatars.mds.yandex.net/get-direct/238679/NZgmDo-bz75yvCJ3dsYPfw/wy150[**Joomla** 3! Освоить легко и быстро](https://an.yandex.ru/count/8XKwZ6dl_b050BG1CTU6jre00000EFIz6K02I09Wl0Xe1720mEo02e01o-hc0eW1h96rjY-G0Ro8fCqhc07ayRgJAQ01ieYapIke0S3nkfCfk06AdDQN6C010jW1kBpb4k01ZC-F0-W1dW7u0G680WEW0kwnfGAv0cZQmwrEMhvJy0BoYgJ32FW2We20W93G8803zu3EsWk80_VeefaEc0E7mGMe0mQm0mIu1Fy1w0Je2VW4neyQY0N6ZngG1RsL6g05YEi7g0M7qWQm1OVI1hW5oDC6m0MzbHh81SZJ1j05m_u2u0Lxy0K1c0RayFq9e0QIeG6e1fAX0Sa6VScADsuW-YhH1drTmAvk8FegsGO0003mJ0000907SAa7CnnWkVCHH1Im1u20a3J01yAszCOFq0S2u0U62e082D08keg0WO0Gu0Zts981W0e1mGe00000003mFzWA0k0AW8bw-0h0_1M82mQg2n3K98nmCnm00828I-KwQ0K0m0k0emN82u3Kam7P2pCSOBdp4KGKw0l6Znhm2mQ83EItuRu1w0m2yWq0-Wq0WWu0a0x0X3sW3i24FG00?test-tag=143490931497057&stat-id=3&)Хотите научиться работать с **Joomla** 3? Начните с этого курса. 6+[alex-kurteev.ru](https://an.yandex.ru/count/8XKwZ6dl_b050BG1CTU6jre00000EFIz6K02I09Wl0Xe1720mEo02e01o-hc0eW1h96rjY-G0Ro8fCqhc07ayRgJAQ01ieYapIke0S3nkfCfk06AdDQN6C010jW1kBpb4k01ZC-F0-W1dW7u0G680WEW0kwnfGAv0cZQmwrEMhvJy0BoYgJ32FW2We20W93G8803zu3EsWk80_VeefaEc0E7mGMe0mQm0mIu1Fy1w0Je2VW4neyQY0N6ZngG1RsL6g05YEi7g0M7qWQm1OVI1hW5oDC6m0MzbHh81SZJ1j05m_u2u0Lxy0K1c0RayFq9e0QIeG6e1fAX0Sa6VScADsuW-YhH1drTmAvk8FegsGO0003mJ0000907SAa7CnnWkVCHH1Im1u20a3J01yAszCOFq0S2u0U62e082D08keg0WO0Gu0Zts981W0e1mGe00000003mFzWA0k0AW8bw-0h0_1M82mQg2n3K98nmCnm00828I-KwQ0K0m0k0emN82u3Kam7P2pCSOBdp4KGKw0l6Znhm2mQ83EItuRu1w0m2yWq0-Wq0WWu0a0x0X3sW3i24FG00?test-tag=143490931497057&stat-id=3&" \t "_blank)Скрыть рекламу:Не интересуюсь этой темойТовар куплен или услуга найденаНарушает закон или спамМешает просмотру контента

|  |
| --- |
| Спасибо, объявление скрыто. |

 |

[Яндекс.Директ](https://direct.yandex.ru/?partner)

**Рисунок 3. Схема доступа к хранилищу данных в Deductor**

Рис. 2.4.Схема доступа к хранилищу данных в Deductor

Еще однo достоинство Warehouse - позволять проводить процедуру загрузки данных в любое время. Стоит отметить, что использование хранилища не является обязательным при анализе, поскольку все необходимые действия можно провести с любыми табличными данными. Однако практика показывает, что применение Deductor Warehouse позволяет значительно ускорить создание законченного решения, обеспечить более высокую производительность и облегчить работу с информацией конечному пользователю.

**Deductor Studio** является аналитическим ядром платформы Deductor. Оно содержит полный набор механизмов импорта, визуализации и экспорта данных для быстрого и эффективного анализа информации.

Deductor Studio в общем понимании является рабочим местом аналитика. В этом приложении осуществляется формализация знаний эксперта. Этот модуль включает в себя различные инструменты обработки данных, включая импорт/экспорт данных, их очистку и предобработку. Этапы анализа данных представлены на следующей схеме (рис. 2.5):

|  |  |
| --- | --- |
|

|  |
| --- |
| Предобработанные данные |

 |

|  |  |
| --- | --- |
|

|  |
| --- |
| Трансформированные данные |

 |

|  |  |
| --- | --- |
|

|  |
| --- |
| Хранилище данных |

 |

|  |  |
| --- | --- |
|

|  |
| --- |
| Выборка |

 |

|  |  |
| --- | --- |
|

|  |
| --- |
| Предобработка |

 |

|  |  |
| --- | --- |
|

|  |
| --- |
| Трансформация |

 |

|  |  |
| --- | --- |
|

|  |
| --- |
| Data Mining |

 |

|  |  |
| --- | --- |
|

|  |
| --- |
| Интерпретация |

 |

|  |  |
| --- | --- |
|

|  |
| --- |
| Знания |

 |

|  |  |
| --- | --- |
|

|  |
| --- |
| Шаблоны |

 |

|  |  |
| --- | --- |
|

|  |
| --- |
| Исходные данные |

 |



Рис. 2.5. Процесс извлечения знаний из данных в Deductor Studio

На начальном этапе в программу загружаются или импортируются данные из какого-либо произвольного источника. Кроме Deductor Warehouse, такими источниками являются файлы MS Excel и MS Access, Dbase, CSV-файлы, ADO-источники (позволяют получить данные из любого ODBC-источника, будь то Oracle, MS SQL, Sybase и прочих), а также из системы 1С: Предприятие.

Обычно в программу загружаются не все данные, а определенная выборка, необходимая для дальнейшего анализа. После получения выборки по ней можно составить определенную статистику и представить ее в наиболее предпочтительном для пользователя виде, например, в виде диаграмм или гистограмм.

Следующим шагом является преобразование данных (если этого требуют методы анализа). Примерами такой трансформации могут быть преобразование нечисловых данных в числовые, либо дискретизация непрерывных данных.

После этого к данным применяются методы более глубокого анализа. На этом этапе выявляются скрытые зависимости и закономерности в данных, на основании которых строятся различные модели – шаблоны. Одним из таких методов является DataMining.

Известно, что DataMining обозначает процесс извлечения знаний из сырых, необработанных данных. В Deductor Studio этот метод представлен следующим набором алгоритмов: нейронные сети, линейная регрессия, прогнозирование, автокорреляция, деревья решений, самоорганизующиеся карты (карты Кохонена), ассоциативные правила.

Результатом обработки также является набор данных, который, в свою очередь, может быть обработан. Импортированный набор данных, а также данные полученные на каждом этапе обработки, могут быть экспортированы для последующего использования в других учетных системах.

В Deductor Studio поддерживаются следующие форматы: MS Excel, MS Word, HTML, XML, Dbase, буфер обмена Windows, текстовый файл с разделителями.

Результаты каждого действия можно отобразить с помощью следующих способов: OLAP-кубы, плоская таблица, диаграмма, гистограмма, статистика, анализ по принципу «что, если», граф нейросети, дерево (иерархическая система правил).

Способ возможных отображений зависит от выбранного метода обработки данных. Например, нейросеть содержит визуализатор «Граф нейросети», специфичный только для нее. Но в большинстве случаев многие способы отображения, например диаграммы, пригодны для любых методов обработки.

Таким образом, архитектура Deductor Studio может быть представлена в следующем виде (рис.2.6).



Рис. 2.6. Архитектура Deductor Studio

В условиях цикла лабораторных работ на кафедре прикладной информатики доступна учебная версия Deductor Academic. К сожалению, многие возможности Deductor Warehouse и Deductor Studio в этой версии недоступны. Так, для загрузки аналитических данных требуется их преобразование в текстовый формат с разделителями.

Подробная инструкция работы с учебной версией Deductor размещена на кампусе в разделе «Мои материалы» лектора дисциплины «Проектирование информационных систем».

|  |
| --- |
|  |

# Проектирование хранилища данных

[⇐ Предыдущая](https://studopedia.org/2-15557.html)[1](https://studopedia.org/2-15554.html)[2](https://studopedia.org/2-15555.html)[3](https://studopedia.org/2-15556.html)[4](https://studopedia.org/2-15557.html)**5**[6](https://studopedia.org/2-15559.html)[7](https://studopedia.org/2-15560.html)[8](https://studopedia.org/2-15561.html)[Следующая ⇒](https://studopedia.org/2-15559.html)

В настоящее время существуют фактические стандарты построения корпоративных информационно-аналитических систем, основанных на концепции хранилища.

Эти стандарты опираются на современные исследования и общемировую практику создания хранилищ данных и аналитических систем.

В общем виде архитектура корпоративной информационно-аналитической системы описывается схемой с тремя выделенными слоями (рис. 2.7):

– извлечение, преобразование и загрузка данных;

– хранение данных;

– анализ данных (рабочие места пользователей).

На рис. 2.8 показаны основные этапы проектирования информационно – аналитической системы.



Рис. 2.7. Архитектура информационно-аналитической системы

|  |  |  |  |
| --- | --- | --- | --- |
|   |   |   |   |
|   |   | http://konspekta.net/studopediaorg/baza2/578255622666.files/image022.gif |  |
|   | http://konspekta.net/studopediaorg/baza2/578255622666.files/image023.gif |  |  |
|   |  |  |  |

*Рис. 2.8. Этапы проектирования базы данных*

*Концептуальная модель заключается в решении самых насущных задач. Практически это означает, что на первой стадии важно создать ясную картину будущего содержания хранилища данных в самых грубых чертах, затем выбрать наиболее актуальную задачу, и руководствуясь концепцией архитектуры хранилища данных работать над ней, не забывая об общей картине. Такой подход позволяет получить ощутимые результаты в очень короткие сроки и при этом сохранить уверенность в том, что дальнейшее расширение создаваемого хранилища не приведет к перестройке всей его структуры.*

*С самого начала необходимо оценить условия для реализации проекта. Затем необходимо получить четкое “видение” основных проблем организации в целом, а также определить основные источники данных, чтобы, получив общую, концептуальную модель деятельности организации, использовать ее как базис для стратеги построения хранилища.*

*На стадии сбора требований осуществляется определение круга пользователей, связанных с решаемой на данном этапе проблемой, и их опрос. Параллельно, осуществляется анализ имеющихся источников информации и способов доступа к ним. В результате этой стадии конкретизируется:*

*· структура данных в источниках*

*· основные предметные области, связанные с задачей*

*· требования пользователей*

*На этой же стадии вводится четко определенный и согласованный со всеми участниками проекта критерий завершенности проекта или его очередной итерации.*

|  |  |
| --- | --- |
| https://avatars.mds.yandex.net/get-direct/329956/qttXQopYdpKZ_SPQjeQbCw/x180**[Создай сайт](https://an.yandex.ru/count/LgqkrX42-Mm50Au1CV26jre00000EFIz6K02I09Wl0Xe172-qO3X1801pjVQbGM80QYHuyaWa07MjB6v39W1seMQZH2W0RIlghOCg064xvsB4BW1thQ4eml00GBO0SRJtGZW0Vp9hGRe0Khu0Th2thu1Y083e0A6wja6kG8txgW99C2SIF02jhFaZmRu0eA0W82Gm2Q00_sFgDKCY0FrjvNm3gW31h031EW4X0hu1FN_5eW5zVyMa0NCoHUW1PRp1gW5kkm5i0Mwx0Mu1PV_1S05pSaNo0MO_mNG1URG0k05HF050PW6WD2iymwW1fAX0QW6ag41oGPzoOetRY3wAj46VLt0hcuW-YhP1W000F1C0000gGSp7032vnT45B07W82GDC07mhRqnW_G1mBW1uOAW0W8q0YwYe21m910u0Y3zQm3W0e1mGe00000003mFzWA0k0AW8bw-0h0_1M82mQg2n2SFTFXCnm0096AoEOwQ0K0m0k0emN82u3Kam7P2pCS0CBd5qGKw0lr_nRm2mQ83Dh2thu1w0m2yWq0-Wq0WWu0a0x0X3sW3i24FG00?test-tag=169329454749793&stat-id=3&" \t "_blank)**[самостоятельно!](https://an.yandex.ru/count/LgqkrX42-Mm50Au1CV26jre00000EFIz6K02I09Wl0Xe172-qO3X1801pjVQbGM80QYHuyaWa07MjB6v39W1seMQZH2W0RIlghOCg064xvsB4BW1thQ4eml00GBO0SRJtGZW0Vp9hGRe0Khu0Th2thu1Y083e0A6wja6kG8txgW99C2SIF02jhFaZmRu0eA0W82Gm2Q00_sFgDKCY0FrjvNm3gW31h031EW4X0hu1FN_5eW5zVyMa0NCoHUW1PRp1gW5kkm5i0Mwx0Mu1PV_1S05pSaNo0MO_mNG1URG0k05HF050PW6WD2iymwW1fAX0QW6ag41oGPzoOetRY3wAj46VLt0hcuW-YhP1W000F1C0000gGSp7032vnT45B07W82GDC07mhRqnW_G1mBW1uOAW0W8q0YwYe21m910u0Y3zQm3W0e1mGe00000003mFzWA0k0AW8bw-0h0_1M82mQg2n2SFTFXCnm0096AoEOwQ0K0m0k0emN82u3Kam7P2pCS0CBd5qGKw0lr_nRm2mQ83Dh2thu1w0m2yWq0-Wq0WWu0a0x0X3sW3i24FG00?test-tag=169329454749793&stat-id=3&" \t "_blank)**Сайт** своими руками. Дарим 1000 рублей на счёт! [umi.ru](https://an.yandex.ru/count/LgqkrX42-Mm50Au1CV26jre00000EFIz6K02I09Wl0Xe172-qO3X1801pjVQbGM80QYHuyaWa07MjB6v39W1seMQZH2W0RIlghOCg064xvsB4BW1thQ4eml00GBO0SRJtGZW0Vp9hGRe0Khu0Th2thu1Y083e0A6wja6kG8txgW99C2SIF02jhFaZmRu0eA0W82Gm2Q00_sFgDKCY0FrjvNm3gW31h031EW4X0hu1FN_5eW5zVyMa0NCoHUW1PRp1gW5kkm5i0Mwx0Mu1PV_1S05pSaNo0MO_mNG1URG0k05HF050PW6WD2iymwW1fAX0QW6ag41oGPzoOetRY3wAj46VLt0hcuW-YhP1W000F1C0000gGSp7032vnT45B07W82GDC07mhRqnW_G1mBW1uOAW0W8q0YwYe21m910u0Y3zQm3W0e1mGe00000003mFzWA0k0AW8bw-0h0_1M82mQg2n2SFTFXCnm0096AoEOwQ0K0m0k0emN82u3Kam7P2pCS0CBd5qGKw0lr_nRm2mQ83Dh2thu1w0m2yWq0-Wq0WWu0a0x0X3sW3i24FG00?test-tag=169329454749793&stat-id=3&" \t "_blank)[Адрес и телефон](https://an.yandex.ru/count/LgqkraWDQhu50Au1CV26jre00000EFIz6K02I09Wl0Xe172-qO3X1801pjVQbGM80QYHuyaWa07MjB6v39W1seMQZH2W0RIlghOCg064xvsB4BW1thQ4eml00GxO0SRJtGZW0Vp9hGRe0Khu0Th2thu1Y083e0A6wja6kG8txgW99C2SIF02jhFaZmRu0eA0W82Gm2Q00_sFgDKCY0FrjvNm3gW31h031EW4X0hu1FN_5eW5zVyMa0NCoHUW1PRp1gW5kkm5i0Mwx0Mu1PV_1S05pSaNo0MO_mNG1URG0k05HF050PW6WD2iymwW1fAX0QW6ag41oGPzoOetRY3wAj46VLt0hcuW-YhP1W000F1C0000gGSp7032vnT45B07W82GDC07mhRqnW_G1mBW1uOAW0W8q0YwYe21m910u0Y3zQm3W0e1mGe00000003mFzWA0k0AW8bw-0h0_1M82mQg2n2SFTFXCnm0096AoEOwQ0K0m0k0emN82u3Kam7P2pCS0CBd5qGKw0lr_nRm2mQ83Dh2thu1w0m2yWq0-Wq0WWu0a0x0X3sW3i24FG00?test-tag=169329454749793&stat-id=3&)Скрыть рекламу:Не интересуюсь этой темойТовар куплен или услуга найденаНарушает закон или спамМешает просмотру контента

|  |
| --- |
| Спасибо, объявление скрыто. |

 |

[*Яндекс.Директ*](https://direct.yandex.ru/?partner)

*Логическая модель данных является визуальным представлением структур данных, их атрибутов и бизнес-правил. Логическая модель представляет данные таким образом, чтобы они легко воспринимались бизнес-пользователями. Проектирование логической модели должно быть свободно от требований платформы и языка реализации или способа дальнейшего использования данных.*

*Разработчик модели использует требования к данным и результаты анализа для формирования логической модели данных. Разработчик приводит логическую модель к третьей нормальной форме и проверяет ее на соответствие корпоративной модели данных.*

*После сравнения логической и корпоративной моделей данных и внесения всех необходимых изменений важно повторно рассмотреть модель с точки зрения точности и полноты. Для достижения хороших результатов полезно проводить экспертный анализ, а также совместный анализ с бизнес партнерами и командой разработчиков.*

*Логическая модель использует сущности, атрибуты и отношения для представления данных и бизнес-правил. Сущности представляют собой объекты, о которых корпорация заинтересована хранить данные. Атрибуты - это данные, которые корпорация заинтересована хранить. Отношения описывают взаимосвязи между сущностями в терминах бизнес-правил.*

*Сущности представляют собой объекты, данные о которых корпорация заинтересована сохранять. Сущностями могут быть вещественные объекты, такие как персона или книга, но они могут представлять и абстрактные концепции, такие как центр затрат или производственная единица. Описывая сущность необходимо использовать фактографические подробности, которые уникально ее идентифицируют. Каждый экземпляр сущности должен быть отдельным и отличным от всех других экземпляров этой сущности. Например, модель данных для хранения информации о клиентах должна обеспечивать способ, позволяющий отличить одного клиента от другого.*

*Атрибуты представляют данные об объектах, которые необходимо иметь корпорации. Атрибуты представляются именами существительными, которые описывают характеристики сущностей.*

*Отношения представляют взаимосвязи между объектами, о которых корпорация заинтересована хранить данные.Отношения выражаются глаголами или глагольными фразами, которые описывают взаимосвязь.*

*В основу концептуальной и логической моделей проектируемой ИАС положена модель «как должно быть» («to be»), разработанная студентами в рамках первого цикла лабораторных работ в соответствии с полученным вариантом задания.*

*Физическое проектирование транслирует сущности, атрибуты и отношения между ними в таблице, колонки и физические зависимости, которые описываются посредством SQL и проецируется на выбранную технологию для хранения и управления базами данных. При построении физической модели для хранилища данных, необходимо учитывать огромное количество запросов в данной базе, по сравнению с операционной базой данных. Поэтому получение данных из хранилища требует совершенно иного подхода. Одними из популярных средств физического проектирования хранилищ данных является CA ERwin Process Modeler 7.3 (ранее BPwin) и AllFusion ERwin Data Modeler 7.3 (ERwin).*

*CA ERwin Process Modeler можно использовать для графического представления бизнес-процессов. Графически представленная схема выполнения работ, обмена информацией, документооборота визуализирует модель бизнес-процесса. Графическое изложение этой информации позволяет перевести задачи управления организацией из области сложного ремесла в сферу инженерных технологий.*

***CA ERwin Process Modeler (BPwin)****помогает четко документировать важные аспекты любых бизнес-процессов: действия, которые необходимо предпринять, способы их осуществления и контроля, требующиеся для этого ресурсы, а также визуализировать получаемые от этих действий результаты. CA ERwin Process Modeler повышает бизнес-эффективность ИТ-решений, позволяя аналитикам и проектировщикам моделей соотносить корпоративные инициативы и задачи с бизнес-требованиями и процессами информационной архитектуры и проектирования приложений. Таким образом, формируется целостная картина деятельности предприятия: от потоков работ в небольших подразделениях до сложных организационных функций.*

***AllFusion ERwin Data Modeler (ранее: ERwin)****- CASE-средство для проектирования и документирования баз данных, которое позволяет создавать, документировать и сопровождать базы данных, хранилища и витрины данных. Модели данных помогают визуализировать структуру данных, обеспечивая эффективный процесс организации, управления и администрирования таких аспектов деятельности предприятия, как уровень сложности данных, технологий баз данных и среды развертывания.
AllFusion ERwin Data Modeler (ERwin) предназначен для всех компаний, разрабатывающих и использующих базы данных, для администраторов баз данных, системных аналитиков, проектировщиков баз данных, разработчиков, руководителей проектов. AllFusion ERwin Data Modeler позволяет управлять данными в процессе корпоративных изменений, а также в условиях стремительно изменяющихся технологий.*

|  |  |
| --- | --- |
| https://avatars.mds.yandex.net/get-direct/238679/NZgmDo-bz75yvCJ3dsYPfw/wy150[**Joomla** 3! Освоить легко и быстро](https://an.yandex.ru/count/FWc6Jbunezu50BG1CV66jre00000EFIz6K02I09Wl0Xe1720mEo02e01o-hc0eW1h96rjY-G0Ro8fCqhc07ayRgJAQ01ieYapIke0S3nkfCfk06AdDQN6C010jW1kBpb4k01ZC-F0-W1dW7u0G680WIW0kwnfGAv0ZVkg0aam9n8y0Blee3u1VW2ue20W9309e03zu3EsWk80_VeefaEc0E7mGMe0mQm0mIu1Fy1w0IL2lW4mPCHY0N1an6G1RsL6g05qv85g0MNhWIm1PUk1BW5oDC6m0MzbHh81SZJ1j05b_i1u0Lxy0K1c0RayFq9e0QIeG6e1fAX0Sa6VScADsuW-YhH1drTmAvk8FegsGO0003mJ0000907SAa7CnmGl_yNH1Im1u20a3J01yAszCOFq0S2u0U62e082D08keg0WO0Gu0Zts981W0e1mGe00000003mFzWA0k0AW8bw-0h0_1M82mYg2n0mgEpyCnm00A29pkOwQ0K0m0k0emN82u3Kam7P2pCS4B__5qGKw0l1an7m2mQ83EItuRu1w0m2yWq0-Wq0WWu0a0x0X3sW3i24FG00?test-tag=169329454749793&stat-id=4&)Хотите научиться работать с **Joomla** 3? Начните с этого курса. 6+[alex-kurteev.ru](https://an.yandex.ru/count/FWc6Jbunezu50BG1CV66jre00000EFIz6K02I09Wl0Xe1720mEo02e01o-hc0eW1h96rjY-G0Ro8fCqhc07ayRgJAQ01ieYapIke0S3nkfCfk06AdDQN6C010jW1kBpb4k01ZC-F0-W1dW7u0G680WIW0kwnfGAv0ZVkg0aam9n8y0Blee3u1VW2ue20W9309e03zu3EsWk80_VeefaEc0E7mGMe0mQm0mIu1Fy1w0IL2lW4mPCHY0N1an6G1RsL6g05qv85g0MNhWIm1PUk1BW5oDC6m0MzbHh81SZJ1j05b_i1u0Lxy0K1c0RayFq9e0QIeG6e1fAX0Sa6VScADsuW-YhH1drTmAvk8FegsGO0003mJ0000907SAa7CnmGl_yNH1Im1u20a3J01yAszCOFq0S2u0U62e082D08keg0WO0Gu0Zts981W0e1mGe00000003mFzWA0k0AW8bw-0h0_1M82mYg2n0mgEpyCnm00A29pkOwQ0K0m0k0emN82u3Kam7P2pCS4B__5qGKw0l1an7m2mQ83EItuRu1w0m2yWq0-Wq0WWu0a0x0X3sW3i24FG00?test-tag=169329454749793&stat-id=4&" \t "_blank)Скрыть рекламу:Не интересуюсь этой темойТовар куплен или услуга найденаНарушает закон или спамМешает просмотру контента

|  |
| --- |
| Спасибо, объявление скрыто. |

 |

[*Яндекс.Директ*](https://direct.yandex.ru/?partner)

*AllFusion ERwin Data Modeler (ERwin) позволяет наглядно отображать сложные структуры данных. Удобная в использовании графическая среда AllFusion ERwin Data Modeler упрощает разработку базы данных и автоматизирует множество трудоемких задач, уменьшая сроки создания высококачественных и высокопроизводительных транзакционных баз данных и хранилищ данных. Данное решение улучшает коммуникацию в вашей организации, обеспечивая совместную работу администраторов и разработчиков баз данных, многократное использование модели, а также наглядное представление комплексных активов данных в удобном для понимания и обслуживания формате.*

*Программные средства CA ERwin Process Modeler и AllFusion ERwin Data Modeler, являясь самыми современными средствами проектирования информационных аналитических систем, имеют ряд недостатков. Во-первых, они имеют только англоязычный интерфейс и, во-вторых, не имеют доступного по стоимости учебного варианта (в отличие от того же Дедуктора).*

*Поэтому при проектировании хранилища данных в настоящем цикле лабораторных работ рекомендовано пользоваться одной из доступных реляционных СУБД (например, Access) с последующей трансформацией аналитической информации в текстовую форму.*

|  |
| --- |
|  |

# К моделированию схем измерений

[⇐ Предыдущая](https://studopedia.org/2-15558.html)[1](https://studopedia.org/2-15554.html)[2](https://studopedia.org/2-15555.html)[3](https://studopedia.org/2-15556.html)[4](https://studopedia.org/2-15557.html)[5](https://studopedia.org/2-15558.html)**6**[7](https://studopedia.org/2-15560.html)[8](https://studopedia.org/2-15561.html)[Следующая ⇒](https://studopedia.org/2-15560.html)

Моделирование схемы измерений обеспечивает представление данных и операций, специально направленных на процесс поддержки принятия решений в хранилищах данных.

Как известно, реляционная модель была впервые предложена Эдгаром Коддом, сотрудником исследовательской лаборатории корпорации IBM, в 1970 году. В настоящее время реляционная модель является доминирующей для приложений коммерческой обработки данных. Но тот же Э. Кодд позже отмечал недостатки реляционной модели в отношении систем поддержки принятия решений, в особенности для многомерных представлений данных.

В качестве примера рассмотрим компанию, которая продает колу в разных регионах страны. В частности, компания продает 4 разновидности колы (кола, диетическая кола, лимонная кола и апельсиновая кола) в 5 разных регионах страны (Северо-востоке, Среднем западе, Западе, Юго-западе и Юго-востоке). Каждый регион состоит из нескольких штатов, а каждый штат из нескольких городов. Для того чтобы хранить данные по продажам за день по каждому продукту и региону в реляционной базе данных, необходима таблица, состоящая из 3 полей (Продукт, Регион, Продажи) и 20 записей.

Такое представление данных может быть комплексным. Допустим, что компания решила добавить пятый продукт. Для того чтобы отслеживать продажи нового продукта по регионам, необходимо добавить 5 новых рядов в таблицу (каждый ряд для каждого региона). Во-вторых, данные в табл. 2 представляют данные по продажам за один день (допустим 1 Января). Для того чтобы получить данные за все 365 дней 2003 года, необходимо добавить четвертую колонку к данным по продаже и умножить 20 строк на 365, получив в итоге 7300 строк.

Таблица 2

|  |  |  |
| --- | --- | --- |
| ПРОДУКТ | РЕГИОН | ПРОДАЖИ |
| Кола | Северо-восток |  |
| Кола | Средний запад |  |
| Кола | Запад |  |
| Кола | Юго-запад |  |
| Кола | Юго-восток |  |
| Диетическая кола | Северо-восток |  |
| Диетическая кола | Средний запад |  |
| Диетическая кола | Запад |  |
| Диетическая кола | Юго-запад |  |
| Диетическая кола | Юго-восток |  |
| Лимонная кола | Северо-восток |  |
| Лимонная кола | Средний запад |  |
| Лимонная кола | Запад |  |
| Лимонная кола | Юго-запад |  |
| Лимонная кола | Юго-восток |  |
| Апельсиновая кола | Северо-восток |  |
| Апельсиновая кола | Средний запад |  |
| Апельсиновая кола | Запад |  |
| Апельсиновая кола | Юго-запад |  |
| Апельсиновая кола | Юго-восток |  |

В упомянутом выше примере размер реляционной базы данных очень сильно увеличивается, что значительно затрудняет понимание и управление такой базой данных.

|  |  |
| --- | --- |
| https://avatars.mds.yandex.net/get-direct/329956/qttXQopYdpKZ_SPQjeQbCw/x180[**Создай сайт**самостоятельно!](https://an.yandex.ru/count/SS0G8LrVj1G50Au1CGA7jre00000EFIz6K02I09Wl0Xe172-qO3X1801pjVQbGM80QYHuyaWa07MjB6v39W1seMQZH2W0RIlghOCg064xvsB4BW1thQ4eml00GBO0SRJtGZW0Vp9hGRe0Khu0Th2thu1Y083e0A6wja6kGArOZvdSsoC4l02x-IPdWRu0eA0W82Gm2Q00_sFgDKCY0FrjvNm3gW31h031EW4s0du1Cgv5eW5ohaMa0NCoHUW1OdU1gW5oTe5i0N9sWMu1PV_1S05pSaNo0MO_mNG1U380k05HF050PW6WD2iymwW1fAX0QW6ag41oGPzoOetRY3wAj46VLt0hcuW-YhP1W000F1C0000gGSp78I961n45B07W82GDC07mhRqnW_G1mBW1uOAW0W8q0YwYe21m910u0Y3zQm3W0e1mGe00000003mFzWA0k0AW8bw-0h0_1M82mQg2n1NwjUKCnm008MyrESwQ0K0m0k0emN82u3Kam7P2pCSX8aO74GKw0lAkHRm2mQ83Dh2thu1w0m2yWq0-Wq0WWu0a0x0X3sW3i24FG00?test-tag=169329454749793&stat-id=3&)**Сайт** своими руками. Дарим 1000 рублей на счёт! [umi.ru](https://an.yandex.ru/count/SS0G8LrVj1G50Au1CGA7jre00000EFIz6K02I09Wl0Xe172-qO3X1801pjVQbGM80QYHuyaWa07MjB6v39W1seMQZH2W0RIlghOCg064xvsB4BW1thQ4eml00GBO0SRJtGZW0Vp9hGRe0Khu0Th2thu1Y083e0A6wja6kGArOZvdSsoC4l02x-IPdWRu0eA0W82Gm2Q00_sFgDKCY0FrjvNm3gW31h031EW4s0du1Cgv5eW5ohaMa0NCoHUW1OdU1gW5oTe5i0N9sWMu1PV_1S05pSaNo0MO_mNG1U380k05HF050PW6WD2iymwW1fAX0QW6ag41oGPzoOetRY3wAj46VLt0hcuW-YhP1W000F1C0000gGSp78I961n45B07W82GDC07mhRqnW_G1mBW1uOAW0W8q0YwYe21m910u0Y3zQm3W0e1mGe00000003mFzWA0k0AW8bw-0h0_1M82mQg2n1NwjUKCnm008MyrESwQ0K0m0k0emN82u3Kam7P2pCSX8aO74GKw0lAkHRm2mQ83Dh2thu1w0m2yWq0-Wq0WWu0a0x0X3sW3i24FG00?test-tag=169329454749793&stat-id=3&" \t "_blank)[Адрес и телефон](https://an.yandex.ru/count/SS0G8GHG9yO50Au1CGA7jre00000EFIz6K02I09Wl0Xe172-qO3X1801pjVQbGM80QYHuyaWa07MjB6v39W1seMQZH2W0RIlghOCg064xvsB4BW1thQ4eml00GxO0SRJtGZW0Vp9hGRe0Khu0Th2thu1Y083e0A6wja6kGArOZvdSsoC4l02x-IPdWRu0eA0W82Gm2Q00_sFgDKCY0FrjvNm3gW31h031EW4s0du1Cgv5eW5ohaMa0NCoHUW1OdU1gW5oTe5i0N9sWMu1PV_1S05pSaNo0MO_mNG1U380k05HF050PW6WD2iymwW1fAX0QW6ag41oGPzoOetRY3wAj46VLt0hcuW-YhP1W000F1C0000gGSp78I961n45B07W82GDC07mhRqnW_G1mBW1uOAW0W8q0YwYe21m910u0Y3zQm3W0e1mGe00000003mFzWA0k0AW8bw-0h0_1M82mQg2n1NwjUKCnm008MyrESwQ0K0m0k0emN82u3Kam7P2pCSX8aO74GKw0lAkHRm2mQ83Dh2thu1w0m2yWq0-Wq0WWu0a0x0X3sW3i24FG00?test-tag=169329454749793&stat-id=3&)Скрыть рекламу:Не интересуюсь этой темойТовар куплен или услуга найденаНарушает закон или спамМешает просмотру контента

|  |
| --- |
| Спасибо, объявление скрыто. |

 |

[Яндекс.Директ](https://direct.yandex.ru/?partner)

Исследование данных в табл. 3 показывает, что в ней существует два основных измерения, а именно ПРОДУКТ и РЕГИОН. Данная таблица может быть значительно упрощена за счет четкого описания этих измерений в многомерной таблице.

Таблица 3

|  |
| --- |
| ПРОДУКТ |
| Регион | Кола | Диетическая кола | Лимонная кола | Апельсиновая кола |
| Северо-восток |  |  |  |  |
| Средний запад |  |  |  |  |
| Запад |  |  |  |  |
| Юго-запад |  |  |  |  |
| Юго-восток |  |  |  |  |

Многомерное представление данных более понятно для понимания и расширения. Например, для добавления дополнительного пятого продукта требуется просто добавить колонку к табл. 3. Добавление третьего измерения “ВРЕМЯ” представлено в виде трехмерного куба, показанного на рис. 2.9.

Также можно представить трехмерную модель как книгу, состоящую из 365 страниц, где каждая страница содержит данные по продажам продукта в регионах за определенный день года. К тому же такая модель более компактна, и требует меньше дискового пространства, так как общие поля не дублируются, как в табл. 2.

Совершенно очевидно, что многомерное представление данных также обеспечивает наиболее эффективный процесс обработки аналитических запросов, по сравнению с реляционным представлением.



Рис. 2.9.Представление трехмерного куба

В заключение необходимо отметить, что многомерная база данных имеет два преимущества над традиционными базами данных. Во-первых, она легка для понимания и представления. Во-вторых, многомерная структура предлагает значительные улучшения представления процессов выполнения запросов, особенно когда число и размер измерений увеличивается, с чем сталкиваются разработчики хранилищ данных.

# ПОРЯДОК ВЫПОЛНЕНИЯ ЦИКЛА ЛАБОРАТОРНЫХ РАБОТ ПО ПРОЕКТИРОВАНИЮ ИАС

[⇐ Предыдущая](https://studopedia.org/2-15559.html)[1](https://studopedia.org/2-15554.html)[2](https://studopedia.org/2-15555.html)[3](https://studopedia.org/2-15556.html)[4](https://studopedia.org/2-15557.html)[5](https://studopedia.org/2-15558.html)[6](https://studopedia.org/2-15559.html)**7**[8Следующая ⇒](https://studopedia.org/2-15561.html)

При выполнении настоящего цикла лабораторных работ студенты опираются на результаты лабораторных работ по CASE-методологии проектирования информационных систем, которые они выполняли в предыдущем семестре.

Предметная область проектирования остается неизменной. Каждый из студентов получает в дополнение к ранее выданному варианту дополнительное задание на обработку ряда аналитических запросов к созданному для этой цели инварианту хранилища данных.

В рамках лабораторной работы №1 студенты выполняют модернизацию имеющейся модели информационной системы в зависимости от перечня и содержания аналитических запросов. Для этой цели используется моделирующий пакет «Ramus Educational». В данной модели должны найти отражение как обработка традиционных OLTP-запросов, так и аналитическая обработка данных.

При выполнении лабораторной работы №2 в рамках виртуального «Хранилища» моделируется информационная структура типа «Звезда» или «Снежинка» (в зависимости от содержания аналитических запросов). Для целей моделирования рекомендуется использовать либо реляционную СУБД типа Access, либо табличный процессор типа Excel (в зависимости от необходимости дополнительной обработки данных). Студенты, имеющие возможность применения платформы ERwin, могут использовать для целей моделирования Хранилища программные средства, входящие в ее состав.

Лабораторная работа №3 настоящего цикла посвящена проведению аналитических исследований в рамках полученного задания. Для этих целей используется учебно-аналитический пакет Deductor Academic (см. раздел 2.3 настоящих методических указаний). Если для целей моделирования структуры Хранилища использовался не ERwin, то предварительно структуры данных Хранилища переводятся в текстовый формат с разделителями (см. инструкцию по работе с Дедуктором).

При выполнении лабораторной работы №4 студенты осуществляют демонстрацию разработанной ИАС, защищают результаты проектирования.

|  |
| --- |
|  |

# Библиографический список. 1. Кузнецов С.А. (Центр Информационных Технологий), Артемьев В

[⇐ Предыдущая](https://studopedia.org/2-15560.html)[1](https://studopedia.org/2-15554.html)[2](https://studopedia.org/2-15555.html)[3](https://studopedia.org/2-15556.html)[4](https://studopedia.org/2-15557.html)[5](https://studopedia.org/2-15558.html)[6](https://studopedia.org/2-15559.html)[7](https://studopedia.org/2-15560.html)**8**

1. Кузнецов С.А. (Центр Информационных Технологий), Артемьев В. (ГЦИ ЦБ РФ).Обзор возможностей применения ведущих СУБД для построения хранилищ данных (Data Warehouse). http://citforum.ru/database/kbd98/glava15.shtml

2. Хэйлер Э. Новый подход к построению корпоративного Хранилища данных: разрешение сложностей при подготовке отчетности на всех уровнях организации.

http://lib.zabspu.ru/citforum/consulting/BI/corp\_bank/index.htm

3. R-Style Computers. Первые сетевые системы хранения данных (NAS) от российского производителя R-Style Computers. Построение больших хранилищ данных с низкой стоимостью хранения. http://r-style.donpac.ru/new/new153.htm

4. Маклаков С.В. Хранилища данных и их проектирование с помощью CA ERwin.- Москва: КомпьютерПресс, - 2001 № 1.

5. Сенько А.М., Якшин М.М. Система виртуальной интеграции разнородных баз данных // Технологии информационного общества -Интернет и современное общество: Труды VII Всероссийской объединенной конференции. – СПб.: Изд-во СПбГУ, 2004.- С. 142-143.

6. Горин С.В., Тандоев А.Ю. (Фирма Алконс Софт). Применение CASE-средства ERwin 2.5 для информационного моделирования в системах обработки данных.

http://citforum.ru/database/kbd96/ 65.shtml

**Гончаренко Георгий Иванович**

**ИССЛЕДОВАНИЕ МЕТОДОВ ОРГАНИЗАЦИИ**

### Обзор возможностей применения ведущих СУБД для построения хранилищ данных (DataWarehouse)

*Сергей Кузнецов, Центр Информационных Технологий,
Валерий Артемьев, ГЦИ ЦБ РФ*

**Основные понятия
Сравнение оперативных и аналитических ИС с точки зрения обеспечения данными**

1. Основным источником информации, поступающей в оперативную БД является деятельность корпорации. Для проведения анализа данных требуется привлечение внешних источников информации (например, статистических отчетов). *Хранилище данных должно включать как внутренние корпоративные данные, так и внешние данные*.
2. Для оперативной обработки требуются свежие данные за несколько последних месяцев, для проведения достоверных анализа и прогнозирования в хранилище данных нужно иметь информацию о деятельности корпорации и состоянии рынка на протяжении нескольких лет. *Объем аналитических БД как минимум на порядок больше объема оперативных.*
3. Во многих крупных корпорациях одновременно существуют несколько оперативных ИС с собственными БД (по историческим причинам). Оперативные БД могут содержать семантически эквивалентную информацию, представленную в разных форматах, с разным указанием времени ее поступления, иногда даже противоречивую. *Хранилище данных должно содержать единообразно представленную и согласованную информацию, максимально соответствующую содержанию оперативных БД. Необходима компонента для извлечения и "очистки" информации из разных источников.*
4. Оперативные ИС создаются в расчете на решение конкретных задач. Информация из БД выбирается часто и небольшими порциями. Обычно набор запросов к оперативной БД известен уже при проектировании. *Набор запросов к аналитической базе данных предсказать невозможно. Хранилища данных существуют, чтобы отвечать на нерегламентированные (ad hoc) запросы аналитиков*. Можно рассчитывать только на то, что запросы будут поступать не слишком часто и затрагивать большие объемы информации. *Размеры аналитической БД стимулируют использование запросов с агрегатами (сумма, минимальное, максимальное, среднее значение и т.д.).*
5. Оперативные БД по своей природе являются сильно изменчивыми, что учитывается в используемых СУБД (нормализованная структура БД, строки хранятся неупорядоченно, B-деревья для индексации, транзакционность). *При малой изменчивости аналитических БД (только при загрузке данных) оказываются разумными упорядоченность массивов, более быстрые методы индексации при массовой выборке, хранение заранее агрегированных данных.*
6. Для оперативных ИС обычно хватает защиты информации на уровне таблиц. *Информация аналитических БД настолько критична для корпорации, что требуются большая грануляция защиты(индивидуальные права доступа к определенным строкам и/или столбцам таблицы).*

**Концепция хранилища данных**

Хранилище данных - предметно-ориентированный, интегрированный, неизменчивый, поддерживающий хронологию набор данных, организованный для целей поддержки управления.

Подход построения хранилища данных для интеграции неоднородных источников данных принципиально отличается от подхода динамической интеграции разнородных БД. Реально строится новое крупномасштабное хранилище, управление данными в котором происходит по другим правилам, чем в исходных оперативных БД.

В основе концепции хранилища данных лежат две основные идеи:

(1) *Интеграция разъединенных детализированных данных* (детализированных в том смысле, что они описывают некоторые конкретные факты, свойства, события и т.д.) *в едином хранилище*. В процессе интеграции должно выполняться согласование рассогласованных детализированных данных и, возможно, их агрегация. Данные могут поступать из исторических архивов корпорации, оперативных баз данных, внешних источников.

(2) *Разделение наборов данных и приложений, используемых для оперативной обработки и применяемых для решения задач анализа.*

**Общая архитектура аналитических ИС**



**Потоки данных в информационном хранилище**



**Свойства информационных хранилищ**

Уильям Инмон, считающийся основателем нового направления развития технологии БД, дал классическое определение информационного хранилища в 1990 г. Он охарактеризовал его как специальным образом администрируемую базу данных, содержимое которой имеет следующие свойства:

* Предметная ориентация
* Интегрированность данных
* Инвариантность во времени
* Неразрушаемость - cтабильность информации
* Минимизация избыточности информации

*Предметная ориентация*

В отличие от БД в традиционных OLTP-системах, где данные подобраны в соответствии с конкретными приложениями, информация в DW ориентирована на задачи поддержки принятия решений.. Для системы поддержки принятия решений требуются "исторические" данные - факты продаж за определенные интервалы времени. Хорошо спроектированные структуры данных DW отражают развитие всех направлений бизнеса компании во времени.

Поскольку в DW-технологии объекты данных выходят на первый план, то особые требования предъявляются к структурам БД, используемым для создания информационных хранилищ.. Принципиально отличаются и структуры баз данных для OLTP- и DW-систем. Во втором случае в них помещается только та информация, которая может быть полезной для работы систем поддержки принятия решений (DSS).

*Интегрированность данных*

Данные в информационное хранилище поступают из различных источников, где они могут иметь разные имена, атрибуты, единицы измерения и способы кодировки. После загрузки в DW данные очищаются от индивидуальных признаков, т. е. как бы приводятся к общему знаменателю. С этого момента они представляются пользователю в виде единого информационного пространства.

Если в четырех разных приложениях пол клиента кодировался четырьмя различными способами, то в информационном хранилище будет использована единая для всех данных схема кодировки (например, f,m).

*Инвариантность во времени*

В OLTP-системах истинность данных гарантирована только в момент чтения, поскольку уже в следующее мгновение они могут измениться в результате очередной транзакции. Важным отличием DW от OLTP-систем является то, что данные в них сохраняют свою истинность в любой момент процесса чтения.

В OLTP-системах информация часто модифицируется как результат выполнения каких-либо транзакций. Временная инвариантность данных в DW достигается за счет введения полей с атрибутом "время" (день, неделя, месяц) в ключи таблиц. В результате записи в таблицах DW никогда не изменяются, представляя собой снимки данных, сделанные в определенные отрезки времени. В DW содержатся как бы моментальные снимки данных. Каждый элемент в своем ключе явно или косвенно хранит временной параметр, например день, месяц или год.

*Неразрушаемость - cтабильность информации*

В OLTP-системах записи могут регулярно добавляться, удаляться и редактироваться. В DW-системах, как следует из требования временной инвариантности, однажды загруженные данные теоретически никогда не меняются. По отношению к ним возможны только две операции: начальная загрузка и чтение (доступ). Это и определяет специфику проектирования структуры базы данных для DW. Если при создании OLTP-систем разработчики должны учитывать такие моменты, как откаты транзакций после сбоя сервера, борьба с взаимными блокировками процессов (deadlocks), сохранение целостности данных, то для DW данные проблемы не столь актуальны - перед разработчиками стоят другие задачи, связанные, например, с обеспечением высокой скорости доступа к данным.

*Минимизация избыточности информации*

Поскольку информация в DW загружается из OLTP-систем, возникает вопрос, не ведет ли это к чрезмерной избыточности данных? Нет, утверждает Билл Инмон. На самом деле избыточность минимальна (около 1%!), что объясняется следующими причинами:

* при загрузке информации из OLTP-cистем в DW данные фильтруются. Многие из них вообще не попадают в DW, поскольку лишены смысла с точки зрения использования в системах поддержки принятия решений;
* информация в OLTP-системах носит, как правило, оперативный характер, и данные, потеряв актуальность, удаляются. В DW, напротив, хранится историческая информация, и с этой точки зрения перекрытие содержимого DW данными OLTP- систем оказывается весьма незначительным;
* в DW хранится некая итоговая информация, которая в базах данных OLTP-систем вообще отсутствует;
* во время загрузки в DW записи сортируются, очищаются от ненужной информации и приводят к единому формату. После такой обработки это уже совсем другие данные.

**Основные компоненты информационного хранилища**

*ПО промежуточного слоя*

Обеспечивает сетевой доступ и доступ к базам данных. Сюда относятся сетевые и коммуникационные протоколы, драйверы, системы обмена сообщениями и пр.

*Транзакционные БД и внешние источники информации*

Базы данных OLTP-систем исторически предназначались для эффективной обработки структур данных в относительно небольшом числе четко определенных транзакций. Из-за ограниченной целевой направленности "учетных" систем применяемые в них структуры данных плохо подходят для систем поддержки принятия решений. Кроме того, возраст многих установленных OLTP-систем достигает 10 - 15 лет.

*Уровень доступа к данным*

Относящееся сюда ПО обеспечивает общение конечных пользователей с информационным хранилищем и загрузку требуемых данных из транзакционных систем. В настоящее время универсальным языком общения служит язык структурированных запросов (SQL).

*Загрузка и предварительная обработка*

Этот уровень включает в себя набор средств для загрузки данных из OLTP-систем и внешних источников. Выполняется, как правило, в сочетании с дополнительной обработкой: проверкой данных на чистоту, консолидацией, форматированием, фильтрацией и пр.

*Информационное хранилище*

Представляет собой ядро всей системы - один или несколько серверов БД.

*Метаданные*

Метаданные (репозиторий, "данные о данных"). Играют роль справочника, содержащего сведения об источниках первичных данных, алгоритмах обработки, которым исходные данные были подвергнуты, и т. д.

*Уровень информационного доступа*

Обеспечивает непосредственное общение пользователя с данным DW посредством стандартных систем манипулирования, анализа и предоставления данных типа MS Excel, MS Access, Lotus 1-2-3 и др.

*Уровень управления (администрирования)*

Отслеживает выполнение процедур, необходимых для обновления информационного хранилища или поддержания его состояния. Здесь программируются процедуры подкачки данных, перестройки индексов, выполнения итоговых (суммирующих) расчетов, репликации данных, построения отчетов, формирования сообщений пользователям, контроля целостности и др.

**Проблемы интеграции данных**

Остановимся на некоторых проблемах реализации хранилища данных:

* Неоднородность программной среды
* Распределенный характер организации
* Повышенные требования к безопасности данных
* Необходимость наличия многоуровневых справочников метаданных
* Потребность в эффективном хранении и обработке очень больших объемов информации

**Неоднородность программной среды**

Хранилище данных практически никогда не создается на пустом месте. Почти всегда конечное решение будет разнородным, т.е. в нем будут использоваться автономно разработанные программные средства. Прежде всего это касается формирования интегрированного согласованного набора данных, которые могут поступать из разнородных баз данных, электронных архивов, публичных и коммерческих электронных каталогов, справочников, статистических сборников. При построении хранилища данных приходится решать задачу построения единой, согласованно функционирующей информационной системы на основе неоднородных программных средств и решений. При выборе средств реализации хранилища данных приходится учитывать множество факторов, включающих уровень совместимости различных программных компонентов, легкость их освоения и использования, эффективность функционирования и т.д.

**Распределенный характер организации**

В концепции хранилища данных предопределено то, что операционная аналитическая обработка может выполняться в любом узле сети независимо от места расположения основного хранилища. Хотя при аналитической обработке данные только читаются, и потребность в синхронизации отсутствует, для достижения эффективности необходимо поддерживать репликацию данных в разных узлах сети. (На самом деле, все не так просто. Одним из требований к хранилищам данных является то, чтобы свежая информация поступала в хранилище как можно быстрее. Т.е. потенциально любая модификация оперативной БД может инициировать добавление данных к хранилищу данных, а тогда потребуется обновить и все реплики, для чего синхронизация все-таки нужна.)

**Повышение требований к безопасности данных**

Собранная вместе согласованная информация об истории развития корпорации, ее успехах и неудачах, о взаимоотношениях с поставщиками и заказчиками, об истории и состоянии рынка дает возможность анализа прошлой и текущей деятельности корпорации и построения прогнозов для будущего. Эта информация настолько ценна для корпорации, что нельзя допустить возможности ее утечки (на самом деле, если хранилище данных одной корпорации попадет в руки аналитиков другой корпорации, то все аналитические прогнозы первой корпорации сразу станут неверными). В системах, основанных на хранилищах данных, оказывается недостаточной защита данных в стиле языка SQL, которую обеспечивают обычные коммерческие СУБД (этот уровень защиты соответствует классу C2 в соответствии с классификацией Оранжевой Книги Министерства обороны США). Для обеспечения должного уровня защиты доступ к данным должен контролироваться не только на уровне таблиц и их столбцов, но и на уровне отдельных строк (это уже соответствует классу B1 Оранжевой Книги). Приходится также решать вопросы аутентификации пользователей, защиты данных при их перемещении в хранилище данных из оперативных баз данных и внешних источников, защиты данных при их передаче по сети.

**Необходимость наличия многоуровневых справочников метаданных**

Если роль метаданных (обычно содержащихся в таблицах-каталогах) в оперативных информационных системах достаточно ограничена, то для OLAP-систем наличие развитых метаданных и средств их предоставления конечным пользователям является одним из основных условий успешной реализации. Например, прежде, чем менеджер корпорации задаст системе свой вопрос, он должен понять, какая информация имеется, насколько она актуальна, можно ли ей доверять, сколько времени может занять формирование ответа и т.д. Для пользователя OLAP-системы требуются метаданные, по крайней мере, следующих типов:

(1) Описания структур данных, их взаимосвязей.

(2) Информация о хранимых на хранилище данных и поддерживаемых им агрегатах данных.

(3) Информация об источниках данных и о степени их достоверности. Одна и та же информация могла попасть в хранилище данных из разных источников. Пользователь должен иметь возможность узнать, какой источник был выбран основным, и каким образом производились согласование и очистка данных.

(4) Информация о периодичности обновлений данных. Желательно знать не только то, какому моменту времени соответствуют интересующие его данные, но и когда они в следующий раз будут обновлены.

(5) Информация о владельцах данных. Пользователю OLAP-системы может оказаться полезной информация о наличии в системе данных, к которым он не имеет доступа, о владельцах этих данных и о действиях, которые он должен предпринять, чтобы получить доступ к данным.

(6) Статистические оценки времени выполнения запросов. До выполнения запроса полезно иметь хотя бы приблизительную оценку времени, которое потребуется для получения ответа, и объема этого ответа.

**Потребность в эффективном хранении и обработке очень больших объемов
информации**

Уже сейчас известны примеры хранилищ данных, содержащих терабайты информации. По данным консалтинговой компании Meta Group, около половины корпораций, использующих или планирующих использовать хранилища данных, предполагает довести их объем до сотен гигабайт. Проблемой таких больших хранилищ является то, что накладные расходы на внешнюю память возрастают нелинейно при возрастании объема хранилища. Исследования, проведенные на основе тестового набора TPC-D, показали, что для баз данных объемом в 100 гигабайт потребуется внешняя память объемом в 4.87 раза большая, чем нужно собственно для полезных данных. При дальнейшем росте баз данных этот коэффициент увеличивается.

**Реализация хранилищ и витрин данных**

Варианты реализации хранилищ данных

* Виртуальное хранилище данных
* Витрины данных
* Глобальное хранилище данных
* Многоуровневая архитектура хранилища данных

**Виртуальное хранилище данных**

В его основе - репозиторий метаданных, которые описывают источники информации (БД транзакционных систем, внешние файлы и др.), SQL-запросы для их считывания и процедуры обработки и предоставления информации. Непосредственный доступ к последним обеспечивает ПО промежуточного слоя. В этом случае избыточность данных нулевая. Конечные пользователи фактически работают с транзакционными системами напрямую со всеми вытекающими отсюда плюсами (доступ к "живым" данным в реальном времени) и минусами (интенсивный сетевой трафик, снижение производительности OLTP-систем и реальная угроза их работоспособности вследствие неудачных действий пользователей-аналитиков).

**Витрина данных**

Витрина данных (Data Mart) по своему исходному определению - это набор тематически связанных баз данных, которые содержат информацию, относящуюся к отдельным аспектам деятельности корпорации. По сути дела, витрина данных - это облегченный вариант хранилища данных, содержащий только тематически объединенные данные. Целевая база данных максимально приближена к конечному пользователю и может содержать тематически ориентированные агрегатные данные. Витрина данных, естественно, существенно меньше по объему, чем корпоративное хранилище данных, и для его реализации не требуется особо мощная вычислительная техника.

**Глобальное хранилище данных**

В последнее время все более популярной становится идея совместить концепции хранилища и витрины данных в одной реализации и использовать хранилище данных в качестве единственного источника интегрированных данных для всех витрин данных. Тогда естественной становится такая трехуровневая архитектура системы:

На **первом уровне** реализуется корпоративное хранилище данных на основе одной из развитых современных реляционных СУБД. Это хранилище интегрированных в основном детализированных данных. Реляционные СУБД обеспечивают эффективное хранение и управление данными очень большого объема, но не слишком хорошо соответствуют потребностям OLAP-систем, в частности, в связи с требованием многомерного представления данных.

На**втором уровне** поддерживаются витрины данных на основе многомерной системы управления базами данных (примером такой системы является Oracle Express Server). Такие СУБД почти идеально подходят для целей разработки OLAP-систем, но пока не позволяют хранить сверхбольшие объемы данных (предельный размер многомерной базы данных составляет 10-40 Гбайт). В данном случае это и не требуется, поскольку речь идет о витринах данных. Заметим, что витрина данных не обязательно должна быть полностью сформирована. Она может содержать ссылки на хранилище данных и добирать оттуда информацию по мере поступления запросов. Конечно, это несколько увеличивает время отклика, но зато снимает проблему ограниченного объема многомерной базы данных.

Наконец, на **третьем уровне** находятся клиентские рабочие места конечных пользователей, на которых устанавливаются средства оперативного анализа данных.

**Подходы и имеющиеся решения**

**Компания IBM**

Решение компании IBM называется *A Data Warehouse Plus*. Целью компании является обеспечение интегрированного набора программных продуктов и сервисов, основанных на единой архитектуре. Основой хранилищ данных является семейство СУБД DB2. Преимуществом IBM является то, что данные, которые нужно извлечь из оперативной базы данных и поместить в хранилище данных, находятся в системах IBM. Поэтому естественная тесная интеграция программных продуктов.

Предлагаются три решения для хранилищ данных:

(1) *Изолированная витрина данных.* Предназначен для решения отдельных задач вне связи с общим хранилищем корпорации.

(2) *Зависимая витрина данных.* Аналогичен изолированной витрине данных, но источники данных находятся под централизованным контролем.

(3) *Глобальное хранилище данных.* Корпоративное хранилище данных, которое полностью централизовано контролируется и управляется. Глобальное хранилище данных может храниться централизовано или состоять из нескольких распределенных в сети рынков данных.

**Oracle**

Решение компании Oracle в области хранилищ данных основывается на двух факторах: широкий ассортимент продуктов самой компании и деятельность партнеров в рамках программы Warehouse Technology Initiative. Возможности Oracle в области хранилищ данных базируются на следующих составляющих:

* наличие реляционной СУБД Oracle 7, которая постоянно совершенствуется для лучшего удовлетворения потребностей хранилищ данных;
* существование набора готовых приложений, обеспечивающих возможности разработки хранилища данных;
* высокий технологический потенциал компании в области анализа данных;
* доступность ряда продуктов, производимых другими компаниями.

**Hewlett Packard**

Работы, связанные с хранилищами данных, выполняются в рамках программы OpenWarehouse. Выполнение этой программы должно обеспечить возможность построения хранилищ данных на основе мощных компьютеров HP, аппаратуры других производителей и программных компонентов. Основой подхода HP являются Unix-платформы и программный продукт Intelligent Warehouse, который предназначен для управления хранилищами данных. Основа построения хранилищ данных, предлагаемая HP, оставляет свободу выбора реляционной СУБД, средств реинжиниринга и т.д.

**NCR**

Решение компании направлено на решение проблем корпораций, у которых одинаково сильны потребности и в системах поддержки принятия решений, и в системах оперативной аналитической обработки данных. Предлагаемая архитектура называется *Enterprise Information Factory* и основывается на опыте использования системы управления базами данных Teradata и связанных с ней методах параллельной обработки.

**Informix Software**

Стратегия компании в отношение хранилищ данных направлена на расширение рынка для ее продукта *On-Line Dinamic Parallel Server*. Предлагаемая архитектура хранилища данных базируется на четырех технологиях: реляционные базы данных, программном обеспечении для управления хранилищем данных, средствах доступа к данным и платформе открытых систем. Три последние компонента разрабатываются партнерами компании. После выхода Универсального Сервера, основанного на объектно-реляционном подходе, можно ожидать, что и он будет использоваться для построения хранилищ данных.

**SAS Institute**

Компания считает себя поставщиком полного решения для организации хранилища данных. Подход основан на следующем:

* обеспечение доступа к данным с возможностью их извлечения из самых разнообразных хранилищ данных (и реляционных, и нереляционных);
* преобразование данных и манипулирование ими с использованием 4GL;
* наличие сервера многомерных баз данных;
* большой набор методов и средств для аналитической обработки и статистического анализа.

**Sybase**

Стратегия компании в области хранилищ данных основывается на разработанной ей архитектуре *Warehouse WORKS*. В основе подхода находится реляционная СУБД Sybase System 11, средство для подключения и доступа к базам данных OmniCONNECT и средство разработки приложений PowerBuilder. Компания продолжает совершенствовать свою СУБД для лучшего удовлетворения потребностей хранилищ данных (например, введена побитная индексация).

**Software AG**

Деятельность компании в области хранилищ данных происходит в рамках программы *Open Data Warehouse Initiative*. Программа базируется на основных продуктах компании ADABAS и Natural 4GL, собственных и приобретенных средствах извлечения и анализа данных, средстве управления хранилищем данных SourcePoint. SourcePoint позволяет автоматизировать процесс извлечения и пересылки данных, а также их загрузки в хранилище данных.

### Заключение

Совместное применение новых информационных технологий:

* Технология хранилищ данных и оперативной аналитической обработки
* Технологии Web и Java

позволит создать информационную инфраструктуру корпорации и упростить доступ к данным для оперативного анализа.

*Центр Информационных Технологий
Москва, Ленинские горы, МГУ им. М.В. Ломоносова, 2-ой учебный корпус
Тел.: (095) 932-9212, 932-9213, 939-0783
Факс: 939-3670, 939-0805
E-mail:**info@citmgu.ru* *URL:*[*http://www.citmgu.ru*](http://www.citmgu.ru/)