Учебная дисциплина

Корпоративные информационные системы

Лекция 8

Хранилища данных в КИС

Лектор:
Шлаев Дмитрий Валерьевич
кандидат технических наук,
доцент

Учебные вопросы:

1. Понятие хранилища данных
2. Модели данных
3. Ведение НСИ.
4. Состав корпоративного хранилища данных

Введение

В настоящее время совершенствование корпоративного управления становится ключевой стратегической задачей развития и жизнедеятельности любого предприятия. В силу того, что практически все экстенсивные способы совершенствования управления исчерпаны, единственным способом выживания в конкурентной борьбе остаются интенсивные способы улучшения управления. Одним из таких способов является информатизация корпоративного управления за счет внедрения информационных технологий.

1. Понятие хранилища данных

Хранилище данных, как один из важнейших инструментов управления и развития бизнеса является *предметно-ориентированным*, *интегрированным*, *зависимым от времени набором данных*.

Хранилище данных нацелено не только на автоматизацию бизнес-процессов, но и на содержательный анализ информации и предназначено для поддержки принятия решений, а его пользователи - это высший и средний менеджмент организации, аналитики, представители подразделений финансового анализа, маркетинга и других отделов.

Интегрированность данных означает, что, например, данные о клиентах, подразделениях, продуктах и услугах, полученные из различных источников, хранятся согласованно и централизованно. При этом полная информация о клиенте может включать данные, поступившие как из основных транзакционных и информационных (бухгалтерских, торговых либо банковских) систем, так и из фронт-офисного или иного приложения.

Хранилище содержит исторические данные, или зависимый от времени набор данных. Если в оперативных источниках представлены самые последние значения (например, текущее наименование клиента или его физический адрес), то хранилище данных будет содержать в себе всю их предысторию с указанием периода, когда те или иные данные были актуальны.

Назначение продукта

Хранилище данных позволяет собрать в едином, по крайней мере с точки зрения пользователя, месте - супербазе всю информацию, которая может понадобиться управляющему при принятии решения. Источниками данных для информационного хранилища служат в первую очередь данные из разрозненных транзакционных и учетных информационных систем, основанных на различных реляционных СУБД, которые обслуживают повседневную бизнес-деятельность. Источниками необходимой информации могут быть также газеты, радио, телевидение, Интернет и любые другие. При этом предполагается, что данные предварительно должны быть приведены к единым стандартам, очищены от противоречий, структурированы и обобщены с требуемым уровнем детализации.

2. Модели данных

Модель может быть реализована как на реляционной, так и на многомерной СУБД. Центральным компонентом *хранилища* является отраслевая модель данных.

Витрины, построенные на основе хранилища данных или на базе первичных источников, проектируются для удовлетворения потребностей определенной группы пользователей, ориентированных на решение конкретных аналитических задач. Витрины позволяют сравнительно легко обеспечить приемлемую производительность, так как содержат меньший объем данных, заблаговременно их агрегируют и востребованы ограниченным кругом пользователей.

Анализ данных в хранилище реализуется компонентом UniCube, поддерживающим многомерное представление и визуализацию данных с целью их анализа и подготовки отчетов.

Компонент UniCube характеризуется следующими возможностями:

Разделение данных на показатели (переменные) и измерения, определяющие соответственно состояние и пространство бизнеса.

Логическое представление значений показателей в виде многомерных кубов, упорядоченных по равноправным измерениям.

Неограниченное число и количество уровней иерархических связей между значениями измерениями.

Гибкое манипулирование данными. Возможность построение подмножества значений показателя по любому дискриминирующему правилу, определенному на множестве значений его измерений.

Неограниченные возможности агрегирования заданного подмножества значений показателя. Предоставляется возможность вычислять не только сумму значений, но и любой другой определенный пользователем функционал, например, минимум, максимум, среднее, медиану и прочие.

Возможность обработки запросов в "реальном времени" - в темпе процесса аналитического осмысления данных пользователем.

Развитые средства табличного и графического представления данных пользователю.

Крупнейшие компании России внедряют хранилища с середины 90х годов. Предыдущие проекты нельзя назвать неуспешными, так как они решали текущие задачи, в частности, обеспечить руководство компании достоверной непротиворечивой информацией хотя бы по некоторым направлениям деятельности. Однако рост компаний, изменение законодательства и возросшие требования к стратегическому анализу и планированию требуют дальнейшего развития стратегий построения хранилища данных.

Причиной построения хранилищ данных в большинстве случаев являются требования бизнес - пользователей, которые более не в состоянии сводить воедино данные из различных информационных систем.

Источниками данных для будущего хранилища являются транзакционные базы данных (OLTP), унаследованные системы, файловые хранилища, интранет-сайты, разрозненные локальные аналитические приложения. Прежде всего, необходимо определить, где находятся требуемые данные. Поскольку, как правило, эти данные хранятся в различных форматах, необходимо их привести к единому виду, для чего применяются довольно сложные системы извлечения, преобразования и загрузки (Extract, Transformation, Load -ETL) в хранилище данных

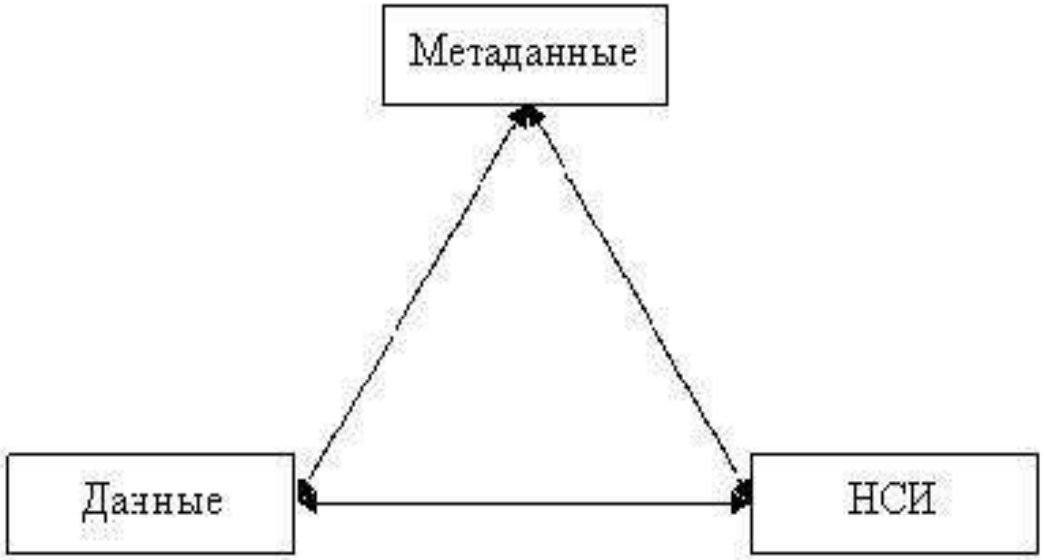
Эта работа не может быть выполнена без сопутствующего анализа метаданных и НСИ. Более того, практика внедрения хранилищ данных показала , что метаданные, созданные и импортированные из различных источников, фактически управляют всем процессом сбора данных.

3. Ведение НСИ (системы централизованного ведения)

Прототипом системы управления метаданными являлись системы словарей-справочников данных, которые были предназначены для логической централизации сведений об информационных ресурсах предприятия и должны были выполнять функции инструмента управления информационными ресурсами предприятия.

Источники данных, в том числе транзакционные системы, содержат метаданные в неявном виде. Например, названия таблиц и имена столбцов в таблицах являются техническими метаданными, а определения сущностей, хранящихся в таблицах, представляют собой бизнес метаданные. Статистика работы приложений, которая может вестись в системах мониторинга, должна быть отнесена к операционным метаданным. Связь между ролями в проекте и правами доступа к базе данных, в том числе правами администрирования, а также данные для аудита и управления изменениями, обычно относятся к проектным метаданным. И, наконец, самая важная часть метаданных - это бизнес метаданные, которые включают в себя бизнес-правила, определения, терминологию, глоссарии, происхождение данных и алгоритмы их обработки.

Структура хранилища данных включает три основных уровня информации: детальные, сводные и архивные данные, а также сопровождающие их метаданные. В настоящее время стало ясно, что этот список должен быть дополнен нормативно-справочной информацией. Связь между данными, НСИ и метаданными можно наглядно представить в виде треугольника.



Как видно из рисунка, все взаимосвязи распадаются на три пары:
данные - метаданные
данные - НСИ
метаданные - НСИ

4. Состав корпоративного хранилища данных

Корпоративное хранилище данных (КХД) преобразует данные, метаданные и НСИ из разнородных источников и предоставляет их пользователям аналитических систем как единую версию правды. Под источниками данных обычно понимают транзакционные базы данных, унаследованные системы, файлы различных форматов, а также иные источники, данные из которых должны быть предоставлены пользователям.

В состав КХД входят:

средства ETL извлечения, преобразования и загрузки данных в центральное хранилище данных;

центральное хранилище данных (ЦХД), предназначенное и оптимизированное для надежного и защищенного хранения данных;

витрины данных, обеспечивающие эффективный доступ пользователей к данным, которые хранятся в структурах, оптимальных для решения конкретных задач пользователей.

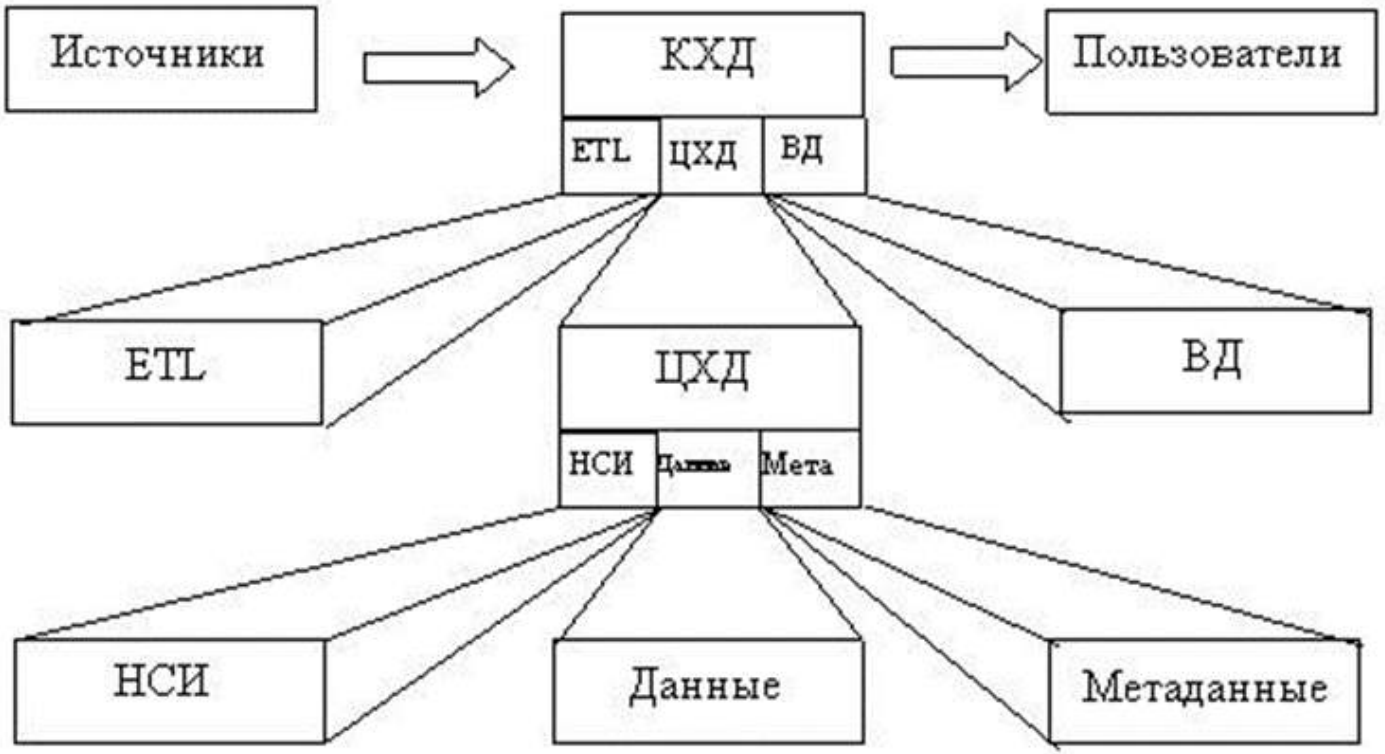
Центральное хранилище включает в себя, прежде всего, три репозитория:

репозиторий нормативно - справочной информации (НСИ);

репозиторий данных;

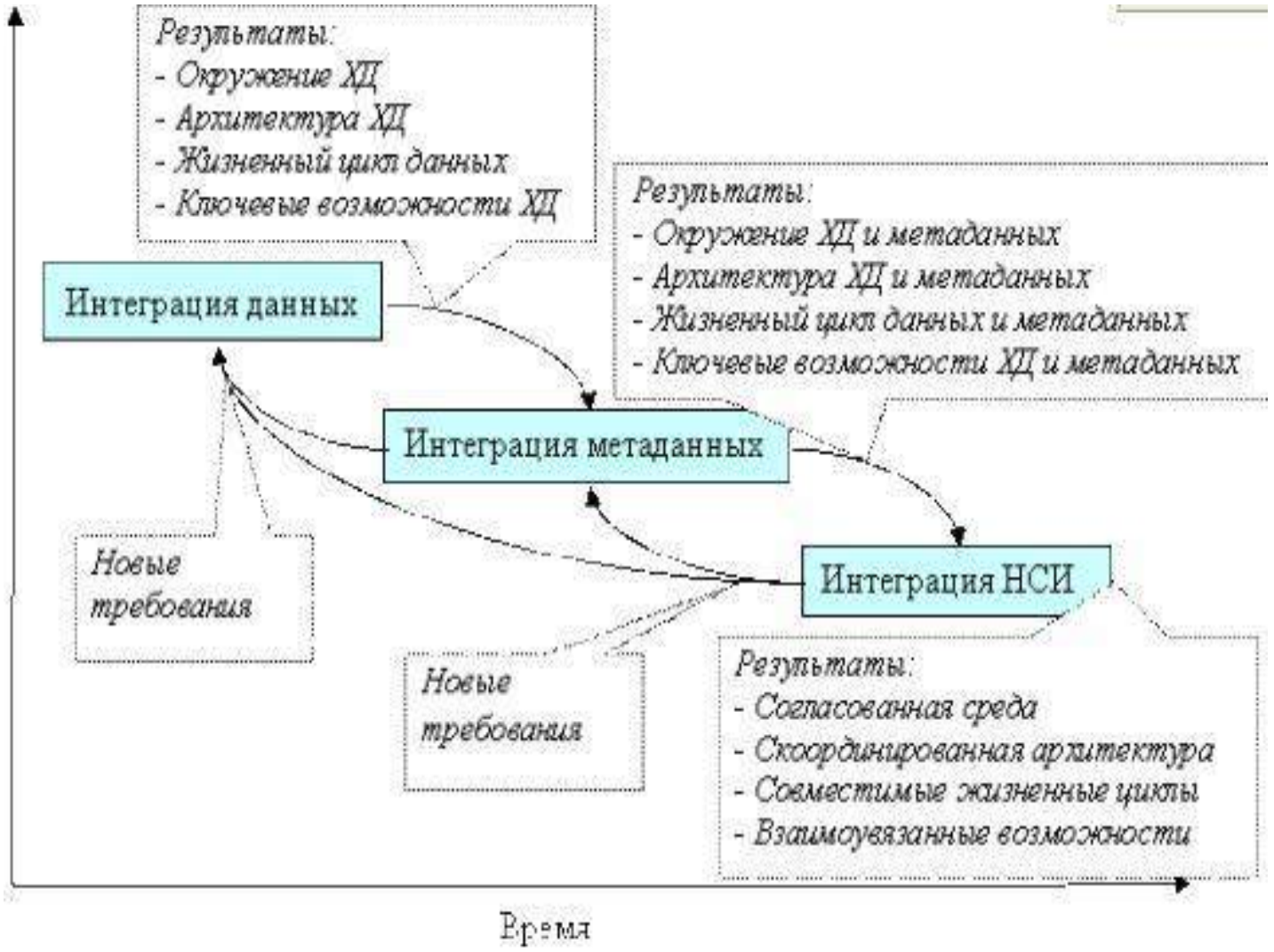
репозиторий метаданных.

В рассмотренную схему не входят оперативный склад данных, зоны промежуточного хранения (staging area), средства доставки данных и доступа к ним, приложения и другие компоненты КХД, несущественные для данного уровня детализации.



Хранилище данных как корпоративная память должно предоставлять целостную непротиворечивую информацию, но обычно это не достигается из-за противоречивой НСИ и недостатка единого понимания смысла данных (т.е. метаданных).

Известным решением является анализ данных и метаданных в рамках проекта интеграции данных, но без создания систем ведения метаданных и НСИ. Внедрение этих систем обычно рассматриваются как отдельные проекты и исполняются после внедрения хранилища данных.


Результаты:
- Окружение ХД
- Архитектура ХД
- Жизненный цикл данных
- Ключевые возможности ХД
Интеграция данных
Результаты:
- Окружение ХД и метаданных
- Архитектура ХД и метаданных
- Жизненный цикл данных и метаданных
- Ключевые возможности ХД и метаданных
Интеграция метаданных
Новые требования
Интеграция НСИ
Новые требования
Результаты:
- Согласованная среда
- Скоординированная архитектура
- Совместимые жизненные циклы
- Взаимоувязанные возможности
Время

Недостатки такого подхода обнаруживаются в процессе эксплуатации, а именно, невысокое качество информации, предоставляемой конечным пользователям хранилища данных, из-за отсутствия согласованного управления метаданными и НСИ, дополнительные расходы на переработку хранилища данных с целью приведения существующих процессов интеграции данных в соответствие с требованиями новых систем управления метаданными и/или НСИ. В результате заказчик получает неэффективную работу трех систем управления данными, метаданными и НСИ, сосуществование модулей с близкой функциональностью, растущую стоимость разработки, высокую стоимость владения и разочарование пользователей из-за расхождения данных, метаданных и НСИ.

Заключение

В настоящее время IBM является единственной компанией, которая предлагает почти полный набор продуктов для осуществления предлагаемой идеи. К ним относятся средства извлечения данных из разнородных источников, средства ведения глоссария метаданных, инструменты проектирования структур данных, средства извлечения и ведения НСИ, современные методологии проектирования среды бизнес - разведки (BI), индустриальные модели данных, а также ПО промежуточного слоя, позволяющее связать компоненты в единую среду информационного обслуживания пользователей.

Идея тройной стратегии, изложенная в данной работе, могла возникнуть в 90-х годах прошлого века. Но ее осуществление было практически невозможно из-за огромных временных, финансовых и трудовых затрат на разработку необходимого инструментария, который стал доступен в последнее время.