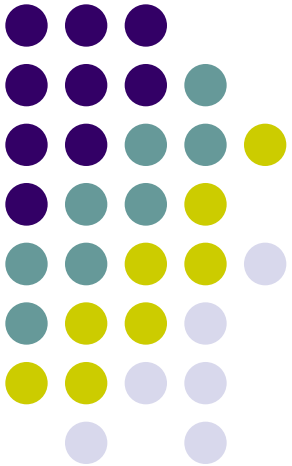
Введение в распределенные методы обработки информации

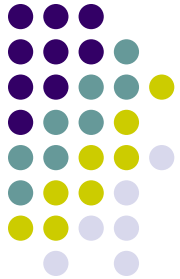
Лекция_№ 4

Технологии интеграции данных

В распределенных системах

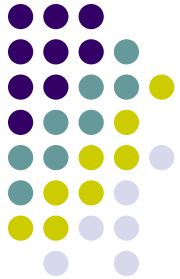


Гомогенные и гетерогенные распределенные БД



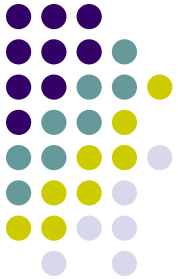
- РБД можно классифицировать на гомогенные и гетерогенные.
- Гомогенной РБД управляет один и тот же тип СУБД.
- Гетерогенной РБД управляют различные типы СУБД, использующие разные модели данных – реляционные, сетевые, иерархические или объектно-ориентированные СУБД.

Гомогенные и гетерогенные распределенные БД



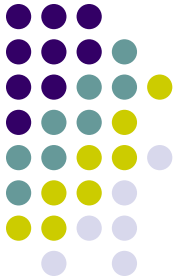
- Гомогенные РБД значительно проще проектировать и сопровождать.
- Кроме того, подобный подход позволяет поэтапно наращивать размеры РБД, последовательно добавляя новые узлы к уже существующей РБД (хорошая масштабируемость).
- Гетерогенные РБД обычно возникают в тех случаях, когда независимые узлы, управляемые своей собственной СУБД, интегрируются во вновь создаваемую РБД

Интеграция данных



- главной проблемой подхода к хранению информации в РБД является разнородность и удаленность источников данных
- целью интеграции является получение единой и цельной картины данных.
- интеграция данных может быть описана с помощью модели, которая включает приложения, продукты, технологии и методы

Модель интеграции данных включает:



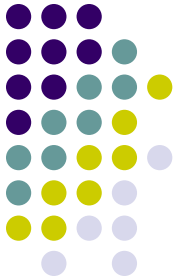
- приложения - это решения, созданные поставщиками в соответствии с требованиями клиентов, которые используют один или несколько продуктов интеграции данных
- продукты - это готовые коммерческие решения, поддерживающие одну или несколько технологий интеграции данных
- технологии реализуют один или несколько методов интеграции данных
- методы - это подходы к интеграции данных, независимые от технологий
- Существует три основных метода интеграции данных: консолидация, федерализация и распространение

Методы интеграции

1
2
3
4
5

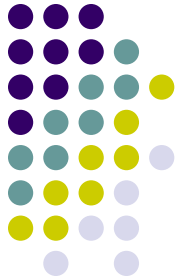
Методы / Этапы	Извлечение	Преобразование	Загрузка в БД
Распространение данных	✓	✗	✓
Федерализация данных	✓	✓	✗
Консолидация данных	✓	✓	✓

Консолидация данных



- Консолидация — комплекс методов и процедур, направленных на извлечение данных из различных источников, обеспечение необходимого уровня их информативности и качества, преобразование в единый формат, в котором они могут быть загружены в хранилище данных или аналитическую систему.

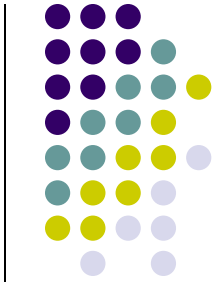
Необходимость консолидации данных



Задачи бизнес-аналитики:

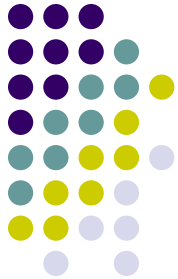
1. Данные на предприятии расположены в различных источниках самых разнообразных форматов и типов:
   - в отдельных файлах офисных документов (Excel, Word, обычных текстовых файлах),
   - в учетных системах («1С:Предприятие», «Парус» и др.),
   - в базах данных (Oracle, Access, dBase и др.).
2. Данные могут быть избыточными или, наоборот, недостаточными.
3. Данные являются «грязными», то есть содержат факторы, мешающие их правильной обработке и анализу (пропуски, аномальные значения, дубликаты и противоречия).

Цели консолидации



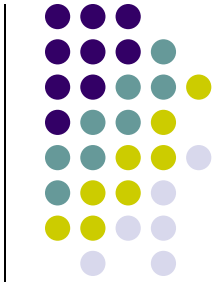
- доведение данных до приемлемого уровня качества и информативности
- организация интегрированного хранения данных в структурах, обеспечивающих их целостность, непротиворечивость, высокую скорость и гибкость выполнения аналитических запросов

Основа консолидации



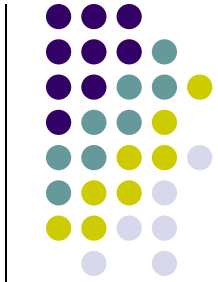
- Консолидация данных является начальным этапом реализации любой аналитической задачи или проекта.
- В основе консолидации лежит процесс сбора и организации хранения данных в виде, оптимальном с точки зрения их обработки на конкретной аналитической платформе или решения конкретной аналитической задачи.
- Сопутствующими задачами консолидации являются оценка качества данных и их обогащение.

Основные критерии оптимальности с точки зрения консолидации данных:



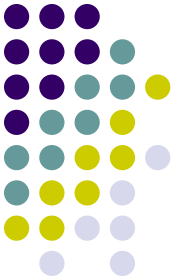
- обеспечение высокой скорости доступа к данным;
- компактность хранения;
- автоматическая поддержка целостности структуры данных;
- контроль непротиворечивости данных.

Источники данных



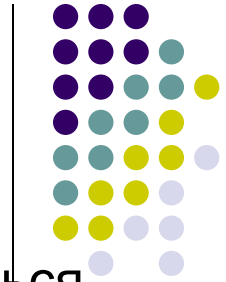
- ***Источник данных*** — объект, содержащий структурированные данные, которые могут оказаться полезными для решения аналитической задачи.
- Объект может считаться источником данных если:
  - используемая аналитическая платформа может осуществлять доступ к данным из этого объекта непосредственно либо после их преобразования в другой формат
  - в противном случае объект не может считаться источником данных.

Основные задачи консолидации данных



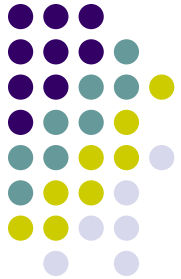
- выбор источников данных, определение типа источников, определение методики организации доступа к источникам
- разработка стратегии консолидации;
- оценка качества данных;
- обогащение;
- очистка;
- перенос в хранилище данных

Выбор источников данных



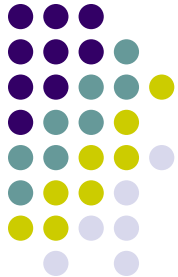
- данные, хранящиеся в отдельных (локальных) файлах
  - ***преимущества***: могут легко создаваться и редактироваться, не требует от персонала специальной подготовки
  - ***недостатки***: не всегда оптимальны с точки зрения скорости доступа к ним, компактности представления данных и поддержки их структурной целостности
- базы данных
  - ***преимущества***: поддерживают целостность данных
  - ***недостатки***: для работы требуют специальной подготовки
- специализированные хранилища данных
  - наиболее предпочтительны для работы с аналитической платформой, поскольку:
  - обеспечивают высокую скорость обмена данными с аналитическими приложениями
  - автоматически поддерживают целостность и непротиворечивость данных
  - главное преимущество ХД — наличие семантического слоя, который дает пользователю возможность оперировать терминами предметной области для формирования аналитических запросов к хранилищу

Разработка стратегии консолидации



- При разработке стратегии консолидации данных необходимо учитывать характер расположения источников данных — локальный, когда они размещены на том же ПК, что и аналитическое приложение, либо удаленный, если источники доступны только через локальную или Глобальную компьютерные сети.
- Характер расположения источников данных может существенно повлиять на качество собранных данных (потеря фрагментов, несогласованность во времени их обновления, противоречивость и т.д.).

Обогащение данных



Обогащение данных— процесс дополнения данных некоторой информацией, позволяющей повысить эффективность решения аналитических задач.

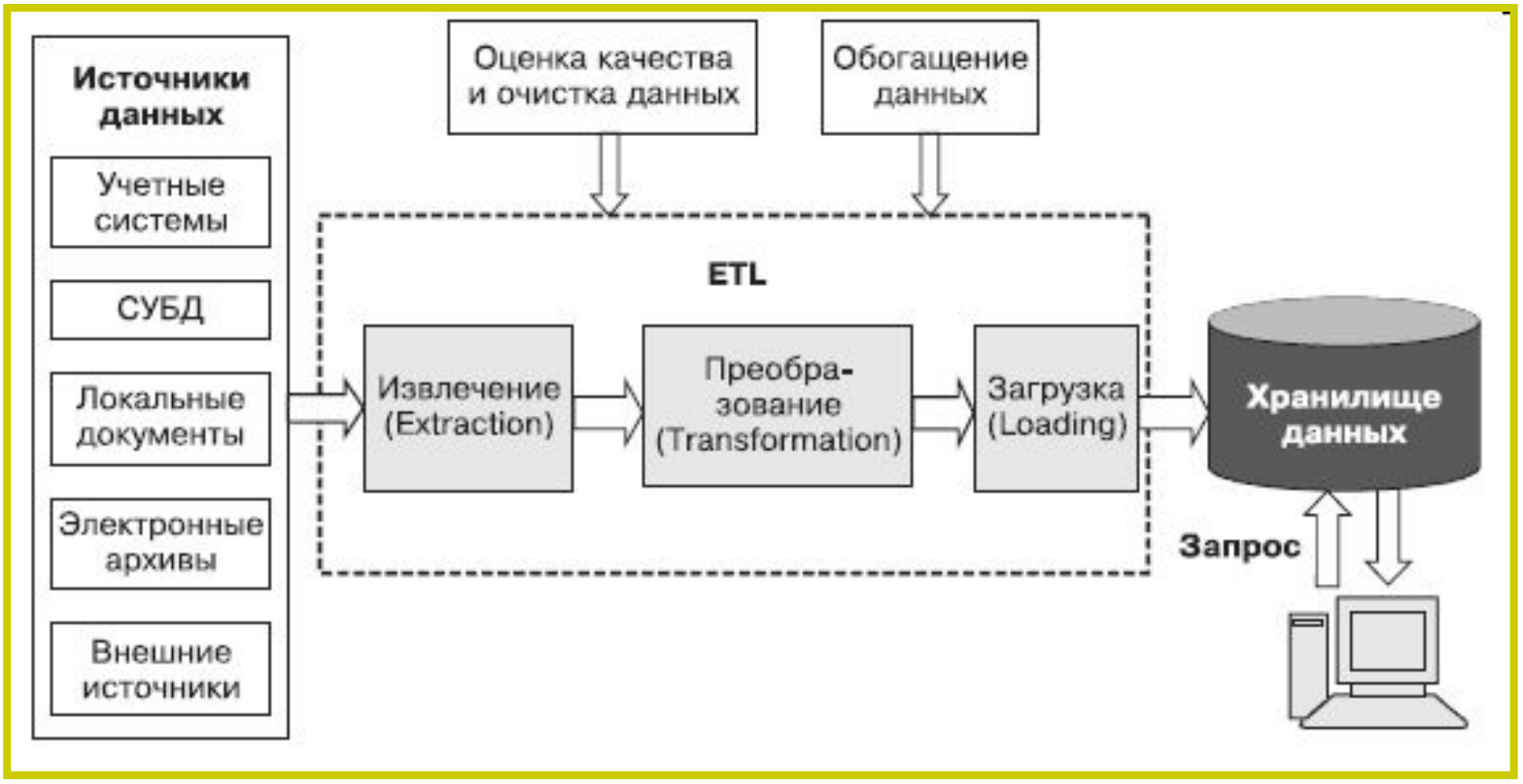
- Обогащение позволяет более эффективно использовать консолидированные данные.
- Обогащение необходимо применять в тех случаях, когда данные содержат недостаточно информации для удовлетворительного решения определенной задачи анализа.
- Обогащение данных позволяет повысить их информационную насыщенность и, как следствие, значимость для решения аналитической задачи.

Очистка данных

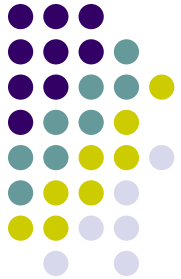
Очистка данных — комплекс методов и процедур, направленных на устранение причин, мешающих корректной обработке: аномалий, пропусков, дубликатов, противоречий, шумов и т.д.

- В большинстве случаев исходные данные являются «грязными», то есть содержат факторы, не позволяющие их корректно анализировать, обнаруживать скрытые структуры и закономерности, устанавливать связи между элементами данных и выполнять другие действия, которые могут потребоваться для получения аналитического решения.
- Поэтому перед тем, как приступить к анализу данных, необходимо оценить их качество и соответствие требованиям, предъявляемым аналитической платформой.
- Если в процессе оценки качества будут выявлены факторы, которые не позволяют корректно применить к данным те или иные аналитические методы, необходимо выполнить соответствующую очистку данных.

Обобщенная схема процесса консолидации



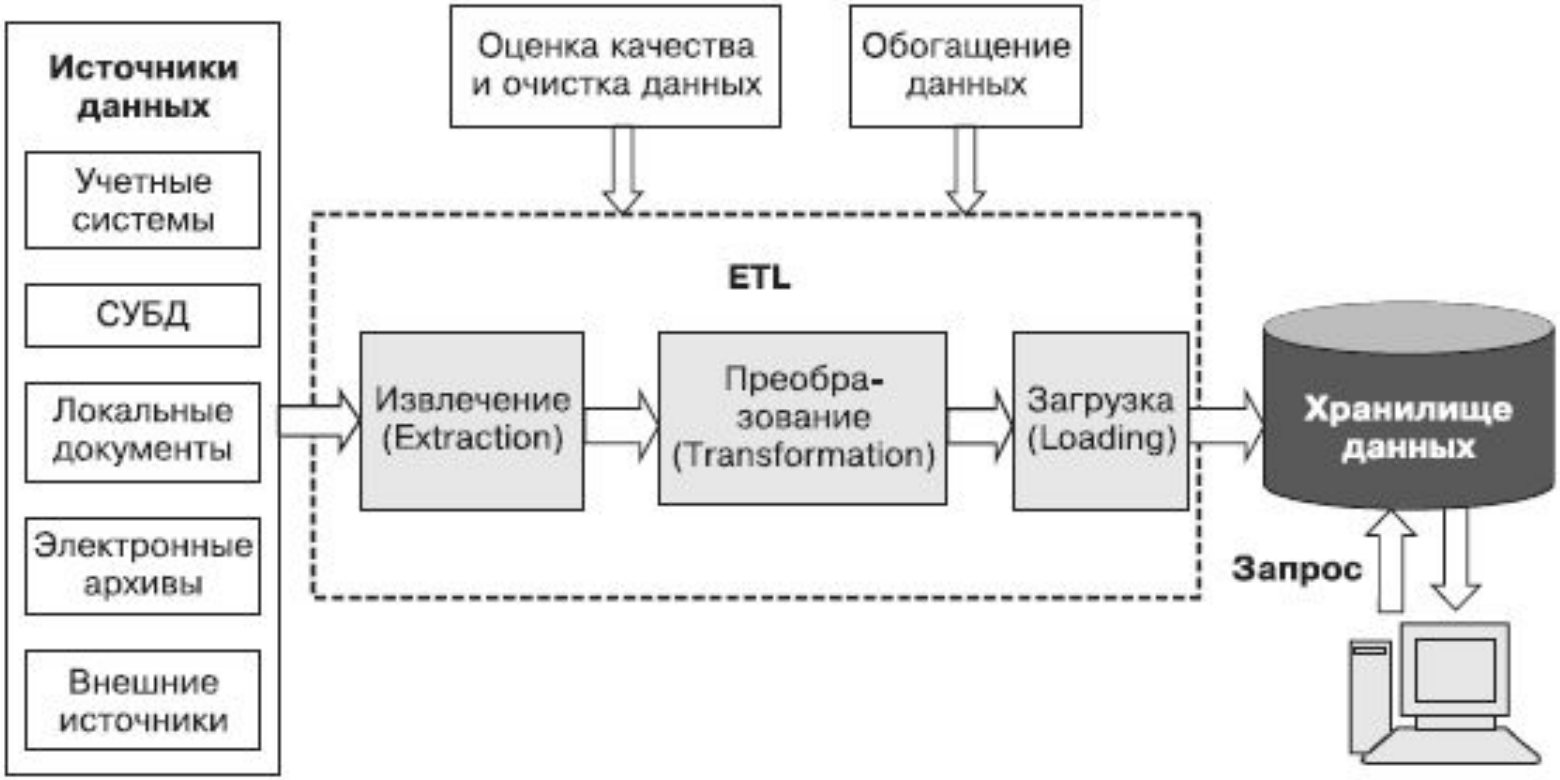
Процесс ETL



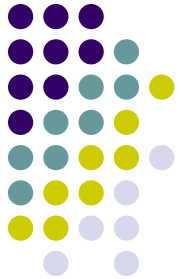
- В основе процедуры консолидации лежит процесс ETL (extraction, transformation, loading).
- Процесс ETL решает задачи:
  - извлечения данных из разнотипных источников,
  - их преобразования к виду, пригодному для хранения в определенной структуре,
  - загрузки данных в соответствующую базу или хранилище
- Если у аналитика возникают сомнения в качестве и информативности исходных данных, то при необходимости он может задействовать процедуры:
  - оценки качества данных,
  - очистки или обогащения данных
- которые также являются составными частями процесса консолидации данных.

Обобщенная структура процесса ETL

Перемещение данных в процессе ETL можно разбить на последовательность процедур, представленных следующей функциональной схемой

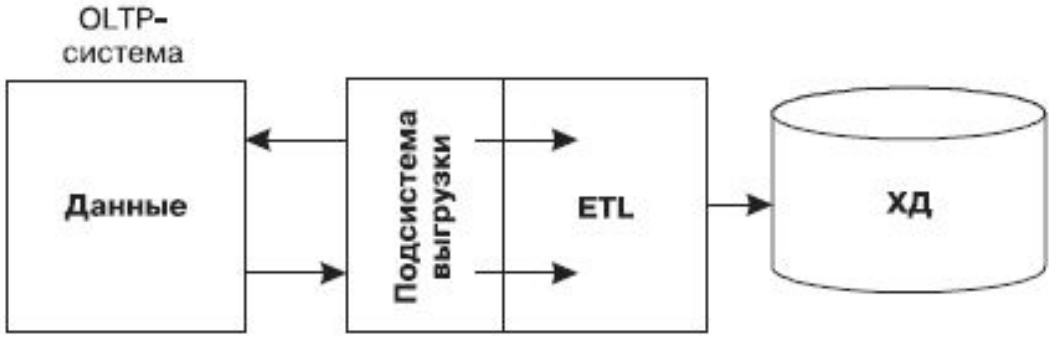


Извлечение данных в ETL



- Начальным этапом процесса ETL является процедура извлечения записей из источника данных и подготовка содержащейся в них информации к процессу преобразования
- Процедуру извлечения можно реализовать двумя основными способами:

1. Извлечение данных с помощью специализированных программных средств

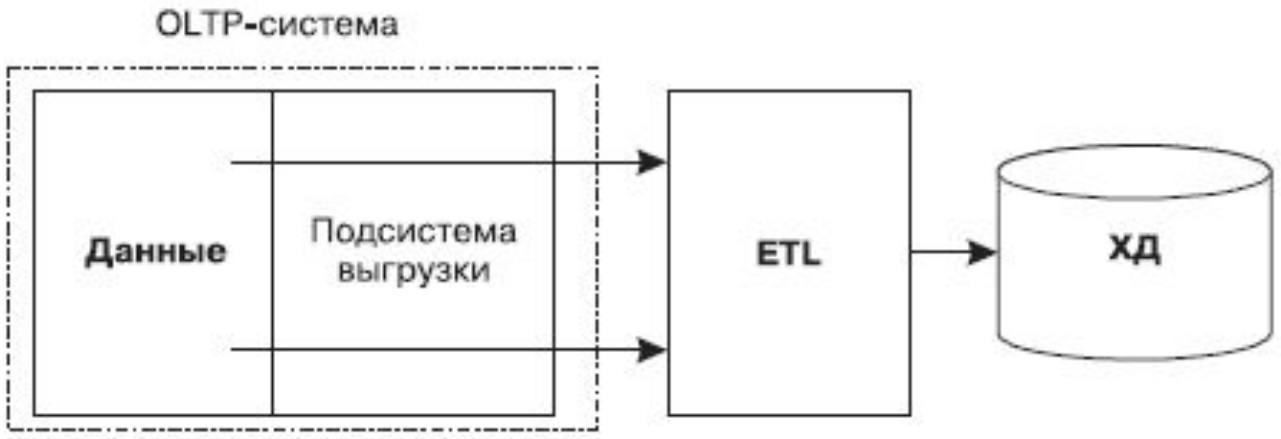


Преимущества:

- позволяет, избежать необходимости оснащать разрабатываемые системы средствами выгрузки,
- позволяет учитывать особенности всего ETL-процесса уже в процессе выгрузки.

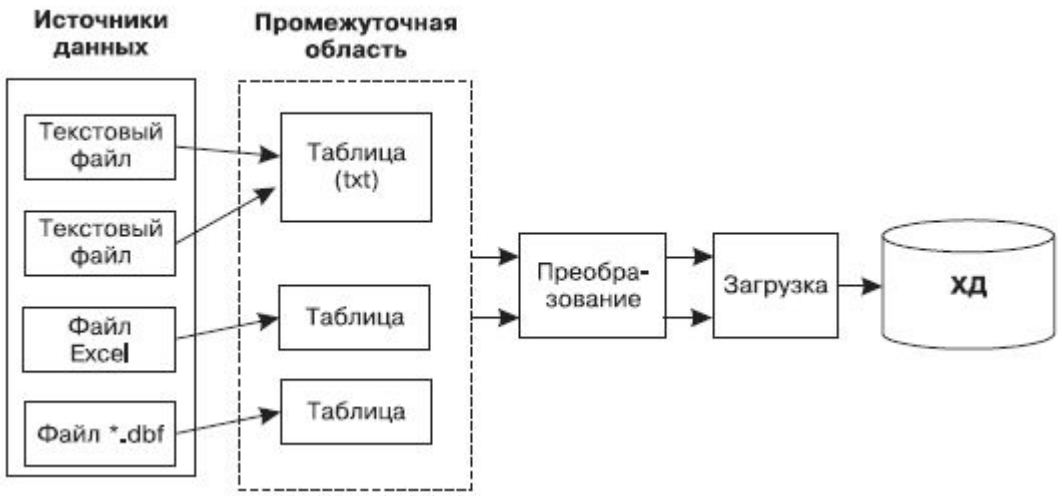
В случае, когда данные извлекаются из локальных источников (отдельных документов, таблиц и т.д.), альтернативы использованию специальных средств нет, поскольку такие виды источников данных не содержат средств выгрузки данных.

2. Извлечение данных средствами той системы, в которой они хранятся



- Поскольку средства «самовыгрузки» разрабатываются с учетом особенностей структуры данных системы, это позволяет адаптировать процедуру извлечения к структуре извлекаемых данных, что в ряде случаев делает процесс более эффективным

Схема организации ETL

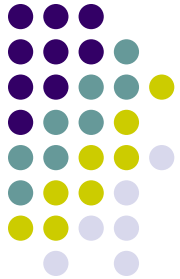


- После извлечения данные помещаются в так называемую промежуточную область, где для каждого источника данных создается своя таблица или отдельный файл (или и то и другое).
- В некоторых случаях, когда требуется выгрузить данные из нескольких источников одного типа, для них создается общая таблица; одно из ее полей указывает на источник, из которого были взяты данные

Процесс преобразования данных в ETL

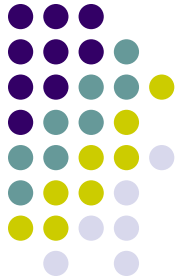
- В процессе преобразования данных в рамках ETL чаще всего выполняются следующие операции:

Преобразование структуры данных



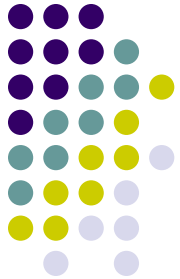
- Во многих случаях данные поступают в хранилище, интегрируясь из множества источников, которые создавались с помощью различных программных средств, методологий, соглашений, стандартов и т.д.
- Данные из таких источников могут отличаться своей структурной организацией: соглашениями о назначении имен полей и таблиц, порядком их описания, форматами, типами и кодировкой данных, например точностью представления числовых данных, используемыми разделителями целой и дробной частей, разделителями групп разрядов и т.д.
- Следовательно, во многих случаях извлеченные данные непригодны для непосредственной загрузки в ХД из-за отличия их структуры от структуры соответствующих целевых таблиц ХД

Агрегирование данных



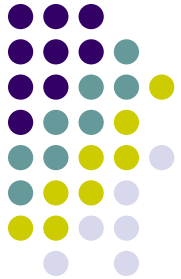
- Как правило, в качестве источников данных для хранилищ выступают системы оперативной обработки данных (OLTP-системы), учетные системы, файлы различных СУБД, локальные файлы отдельных пользователей и т.д. Общим свойством всех этих источников является то, что они содержат данные с максимальной степенью детализации.
- Для достоверного описания предметной области использование данных с максимальным уровнем детализации не всегда целесообразно, поэтому наибольший интерес для анализа представляют данные, обобщенные по некоторому интервалу времени, по группе клиентов, товаров и т.д. Такие обобщенные данные называются агрегированными (иногда агрегатами), а сам процесс их вычисления – агрегированием.
- В результате агрегирования большое количество записей о каждом событии в бизнес-процессе заменяется относительно небольшим количеством записей, содержащих агрегированные значения.

Агрегирование данных



- Фактически при агрегировании производится объединение нескольких записей в одну с вычислением агрегированного значения на основе значений каждой записи.
- При вычислении агрегатов может быть использовано несколько способов.
- Среднее – для данных, расположенных в пределах интервала, в котором они обобщаются, вычисляется среднее значение.
- Затем все записи из данного интервала заменяются одной, содержащей их среднее значение

Пример агрегирования



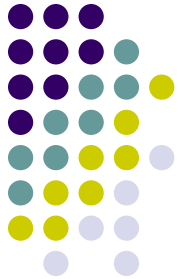
Дата	Цена	Кол-во	Сумма
07.04	150,00	20	3000,00
08.04	135,00	10	1350,00
09.04	220,00	15	3300,00
10.04	173,00	5	865,00
11.04	245,00	24	5880,00
12.04	96,00	12	1152,00
13.04	110,00	320	3520,00

Дата	Среднее кол-во	Средняя сумма	Максимальная сумма	Минимальная сумма	Кол-во сумм	Медиана по сумме
07.04–13.04	16,85	2723,85	5880	865	7	3000

Агрегирование данных

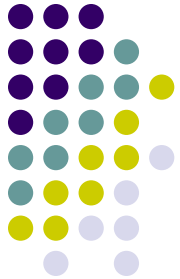
- Из всех возможных вариантов агрегирования следует выбрать наиболее значимые с точки зрения планируемых направлений анализа, а от остальных отказаться.
- Очевидно, можно отказаться от агрегатов, которые имеют малое число подчиненных агрегированных значений (например, агрегирование ежемесячных продаж за квартал), поскольку их легко вычислить в процессе анализа.
- Или, наоборот, можно отказаться от агрегатов с максимальной степенью детализации (например, агрегирование ежедневных продаж).
- Выбор нужных агрегатов всегда определяется особенностями конкретной задачи. При этом следует помнить, что агрегаты, требуемые для анализа, могут быть вычислены и непосредственно при выполнении аналитического запроса к ХД.

Перевод значений



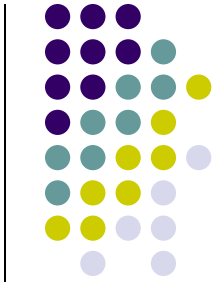
- Часто данные в источниках хранятся с использованием специальных кодировок, которые позволяют сократить избыточность данных и тем самым уменьшить объем памяти, требуемой для их хранения.
- Так, наименования объектов, их свойств и признаков могут храниться в сокращенном виде. В этом случае перед загрузкой данных в хранилище требуется выполнить перевод таких сокращенных значений в более полные и, соответственно, понятные

Создание новых данных



- В процессе загрузки в ХД может понадобиться вычисление некоторых новых данных на основе существующих, что обычно сопровождается созданием новых полей.
- Создание новой информации на основе имеющихся данных тесно связано с таким важным процессом, как обогащение данных, которое может производиться (частично или полностью) на этапе преобразования данных в ETL. Агрегирование также может рассматриваться как создание новых данных.

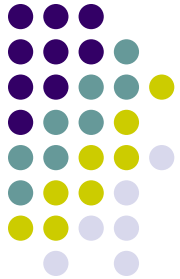
Очистка данных



- Сбор данных в процессе ETL производится из большого числа источников, многие из которых не содержат автоматических средств поддержки целостности, непротиворечивости и корректного представления данных.
- В связи с этим при переносе информации в ХД приходится сталкиваться с потоками «грязных» данных, которые могут стать причиной неправильных результатов анализа и даже сделать невозможным применение некоторых аналитических алгоритмов и методов.
- По этой причине в процессе ETL применяется очистка – процедура корректировки данных, которые в каком-либо смысле не удовлетворяют определенным критериям качества, то есть содержат нарушения структуры данных, противоречия, пропуски, дубликаты, неправильные форматы и т.д.
- Очистка данных – одна из наиболее важных и в то же время наиболее сложных и трудно поддающихся формализации задач ETL-процесса, поскольку набор факторов, снижающих качество данных, весьма разнообразен и может постоянно меняться. Поэтому очистке данных при разработке ETL-процессов уделяют большое внимание.

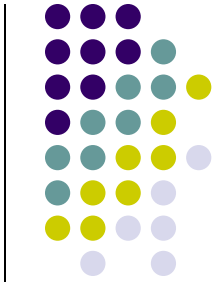
Загрузка данных в хранилище

Постзагрузочные операции



- После завершения загрузки выполняются дополнительные операции над данными, только что загруженными в ХД, перед тем как сделать их доступными для пользователя.
- Такие операции называются постзагрузочными.
- К ним относятся переиндексация, верификация данных и т.д.
- Прежде чем использовать новые данные для анализа, полезно убедиться в их надежности и достоверности.
- Для этих целей можно предусмотреть комплекс верификационных тестов.

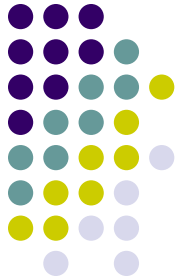
Пример верификационных тестов



- при суммировании продаж по одному измерению результат должен совпадать с соответствующей суммой, полученной по-другому, связанному с ним измерению, то есть сумма продаж по всем товарам за месяц должна соответствовать сумме сделок, заключенных со всеми клиентами за тот же период;
- итоговый показатель за месяц должен соответствовать сумме ежедневных или еженедельных показателей в этом месяце;
- суммарная выручка по всем регионам за текущий месяц должна соответствовать сумме продаж по всем региональным дилерским центрам.

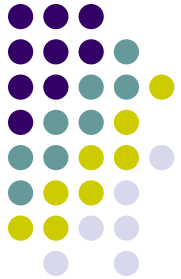
Если тестирование показало, что несоответствия, позволяющие заподозрить потерю или недостоверность данных, отсутствуют, то можно считать загрузку данных в ХД успешной и приступать к анализу новой информации

Пример консолидации данных предприятия



- Процесс сбора, хранения и оперативной обработки данных на типичном предприятии обычно содержит несколько уровней:
  - На верхнем уровне располагаются реляционные SQL-ориентированные СУБД типа SQL Server, Oracle и т.д.
  - На втором — файловые серверы с некоторой системой оперативной обработки или сетевые версии персональных СУБД типа R-Base, FoxPro, Access и т.д.
  - На самом нижнем уровне расположены локальные ПК отдельных пользователей с персональными источниками данных. Чаще всего информация на них собирается в виде файлов офисных приложений — Word, Excel, текстовых файлов и т.д.

Пример консолидации данных предприятия

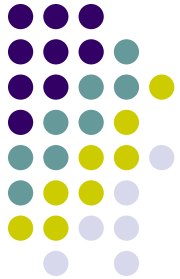


- Из источников данных всех перечисленных уровней информация в соответствии с некоторым регламентом должна перемещаться в ХД. Для этого необходимо:
  - обеспечить выгрузку данных из источников,
  - провести их преобразование к виду, соответствующему структуре ХД,
  - а при необходимости выполнить их обогащение и очистку.

Консолидация данных

- Консолидация данных является сложной многоступенчатой процедурой и важнейшей составляющей аналитического процесса, обеспечивающей высокий уровень аналитических решений.
- Преимуществом консолидации:
  - позволяет осуществлять трансформацию значительных объемов данных (реструктуризацию, согласование, очистку и/или агрегирование) в процессе их передачи от первичных систем к конечным местам хранения.
- Сложности консолидации:
  - поддержка консолидации требует значительных вычислительных ресурсов
  - для поддержки конечного места хранения необходимы существенные ресурсы памяти
  - с учетом постоянно совершенствования аппаратных средств эти сложности не являются неразрешимой проблемой

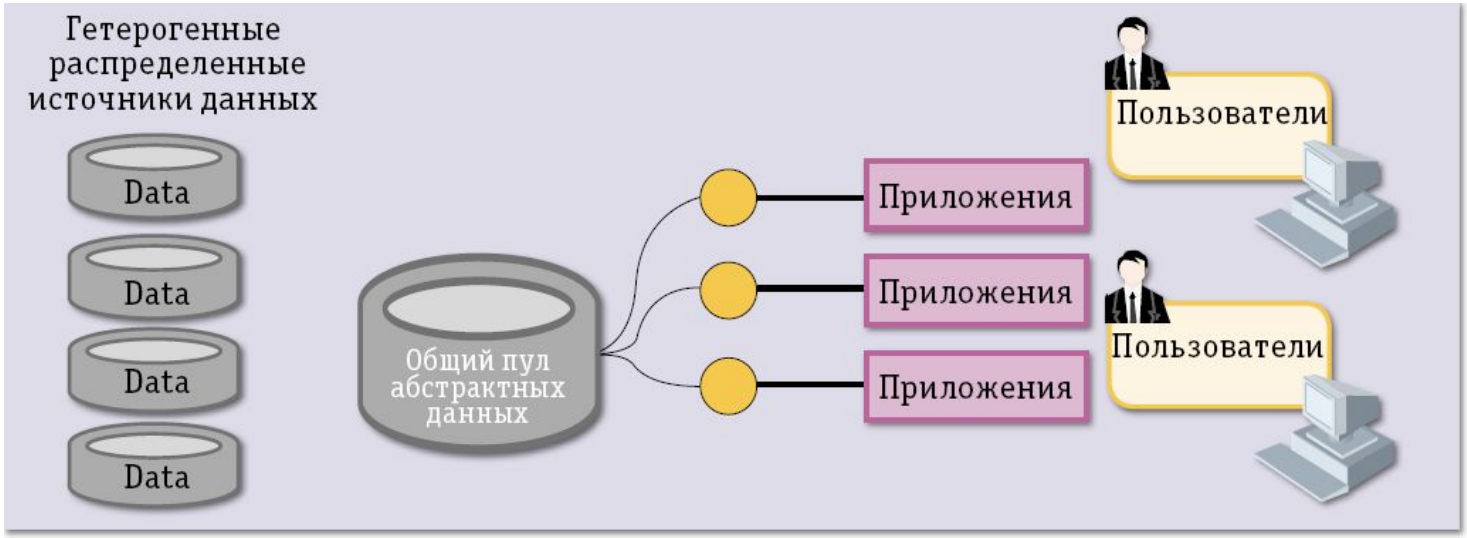
Виртуализации данных



В основе федерализации лежит виртуализация данных

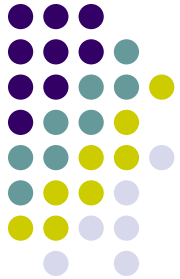
- виртуализация - абстрагировании данных от конкретной формы их хранения
- любая виртуализация подразумевает сбор ресурсов в общий пул и их дальнейшее распределение между потребителями

Общая схема виртуализации данных



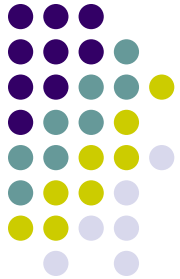
- данные, физически оставаясь на месте, объединяются в один виртуальный пул, а затем поступают в системы бизнес-аналитики, приложения, корпоративные коллажи

Процесс виртуализации:



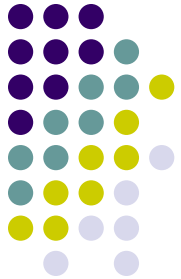
- Виртуализация данных — процесс предоставления данных пользователям посредством интерфейса, скрывающего все технические аспекты хранения данных (способ хранения, местоположение, структура, язык доступа).
- Логически местом для виртуализации данных служит дополнительный промежуточный уровень, изолирующий физическое хранение данных от приложений, которые не должны знать, на каких серверах и в каких базах находятся используемые ими данные.
- При этом могут быть применены самые разные технические приемы

Технические приемы виртуализации



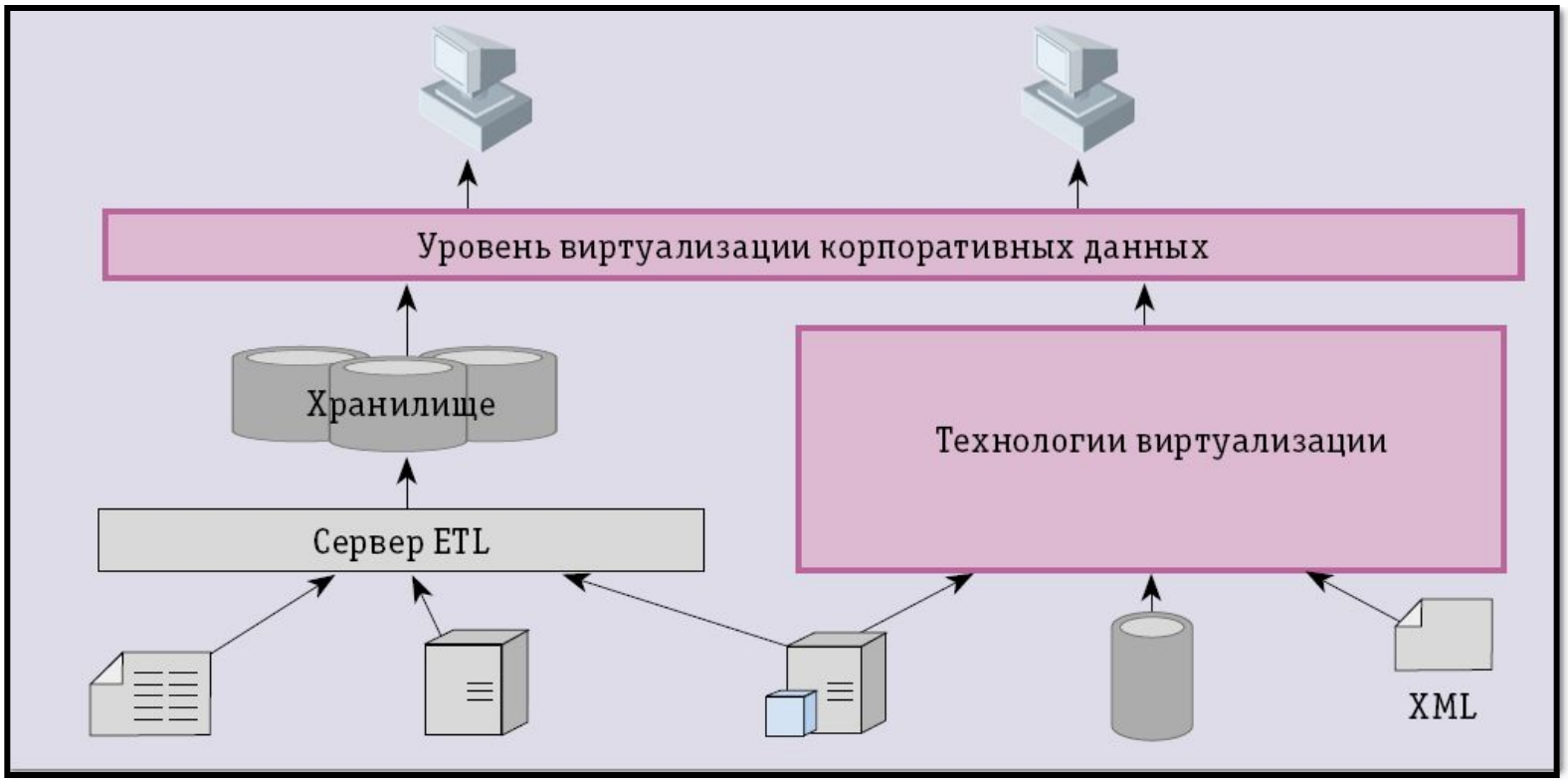
- федерирующий сервер, единообразно представляющий данные из разных источников, с тем чтобы приложения видели данные как одно большое хранилище;
- виртуализация, сосредоточенная в сервисной шине предприятия (Enterprise Service Bus, ESB), выполняющей функции абстрагирования и предоставляющей данные приложениям в форме сервисов;
- облако, содержащее данные (одна из возможных форм виртуализации); где и как хранятся данные, пользователю неизвестно;
- виртуальная база данных в памяти, подпитываемая из физических СУБД;
- собственное решение для конкретной организации.

Федерализация данных

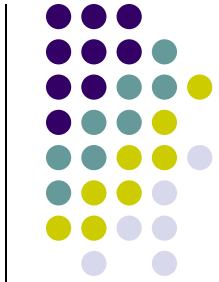


- Федерализация данных — одна из возможных форм организации данных в гетерогенных хранилищах, предусматривающая единообразный доступ к ним.
- Виртуализация не обязательно предполагает федерализацию, но результатом федерации всегда является виртуализация.

Компоненты системы виртуализации

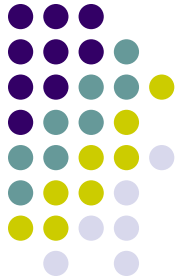


Федерализация данных



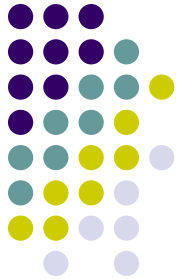
- федерализация данных - это обеспечение единой ***виртуальной*** картины одного или нескольких источников исходных данных.
- федерализация позволяет извлекать данные из различных источников, объединять их и представлять аналитику в режиме реального времени
- при этом физического перемещения данных не происходит: данные остаются у владельцев, доступ к ним всегда осуществляется при необходимости (при выполнении запроса).
- *при федерализации* данных образуется единое виртуальное информационное пространство, данные в котором могут храниться в различных источниках, однако информация о расположении данных недоступна запрашивающей стороне

Федерализация данных



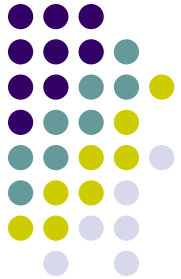
- все необходимые преобразования данных осуществляются при их извлечении из источников
- некоторые федеративные решения могут работать с метаданными, которые отражают семантические связи между элементами данных в источниках
- изучение и профилирование первичных данных, необходимых для федерализации, несильно отличаются от аналогичных процедур, требуемых для консолидации
- интеграция корпоративной информации (Enterprise information integration, сокр. EII) — это пример технологии, которая поддерживает федеративный подход к интеграции данных

Преимущества федерализации данных



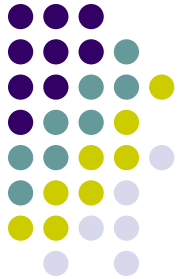
- федеративный подход - обеспечивает доступ к текущим данным и избавляет от необходимости консолидировать первичные данные в новом складе данных
- федерализацию данных возможно использовать в тех случаях, когда стоимость консолидации данных перевешивает бизнес-преимущества, которые она предоставляет (например при подготовке отчетов и оперативной обработки запросов)
- федерализация данных полезна в тех случаях, когда политика безопасности данных и лицензионные ограничения запрещают копирование данных первичных систем
- федерализация могла бы использоваться как кратковременное решение для интеграции данных после приобретения или слияния компаний.

Недостатки федерализации данных



- федерализация данных не очень хорошо подходит для:
  - извлечения и согласования больших массивов данных
  - для тех приложений, где существуют серьезные проблемы с качеством данных в первичных системах.
- федерализация данных оказывает негативное влияние на производительность программы-приложения за счет дополнительных затрат на доступ к многочисленным источникам данных

Распространение данных



- приложения распространения данных осуществляют копирование данных из одного места в другое
- эти приложения обычно работают в оперативном режиме и производят перемещение данных к местам назначения
- обновления в первичной системе могут передаваться в конечную систему синхронно или асинхронно
- большинство технологий синхронного распространения данных поддерживают двусторонний обмен данными между первичными и конечными системами
- примерами технологий, поддерживающих распространение данных, являются интеграция корпоративных приложений (Enterprise application integration, сокр. EAI) и тиражирование корпоративных данных (Enterprise data replication, сокр. EDR)

Распространение данных

Преимущества:

- метод распространения данных может быть использован для перемещения данных в режиме реального времени или близком к нему.
- гарантируется доставка данных и их двустороннее распространение
- Метод распространения данных может использоваться для:
  - уравновешивания рабочей нагрузки,
  - создания резервных копий и
  - восстановления данных, в том числе в случае чрезвычайных ситуаций