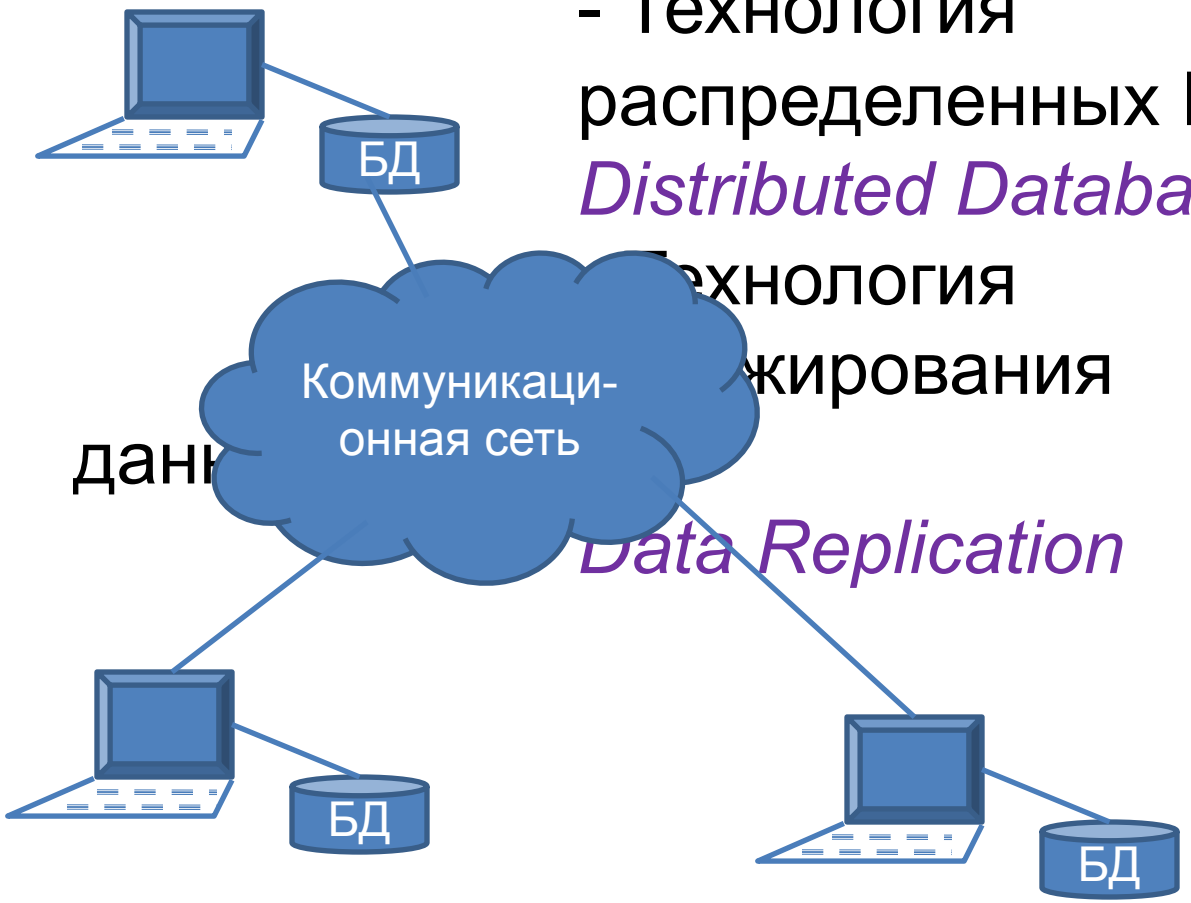
Распределенные системы

Общая характеристика



- Технология распределенных БД *Distributed Database*

ехнология
жирования
дан

Data Replication

Общая характеристика

Системы распределенных баз данных – набор узлов, связанных вместе коммуникационной сетью:

- каждый узел обладает своими собственными системами баз данных;
- узлы работают согласованно, предоставляя доступ к данным на любом узле сети.

Общая характеристика

Распределенная БД – тип *виртуального* объекта

На каждом узле:

– собственные базы данных

– собственные локальные пользователи

– собственные СУБД и средства управления транзакциями

Общая характеристика

Два вида систем распределенных БД:

– *однородные*

– *неоднородные*

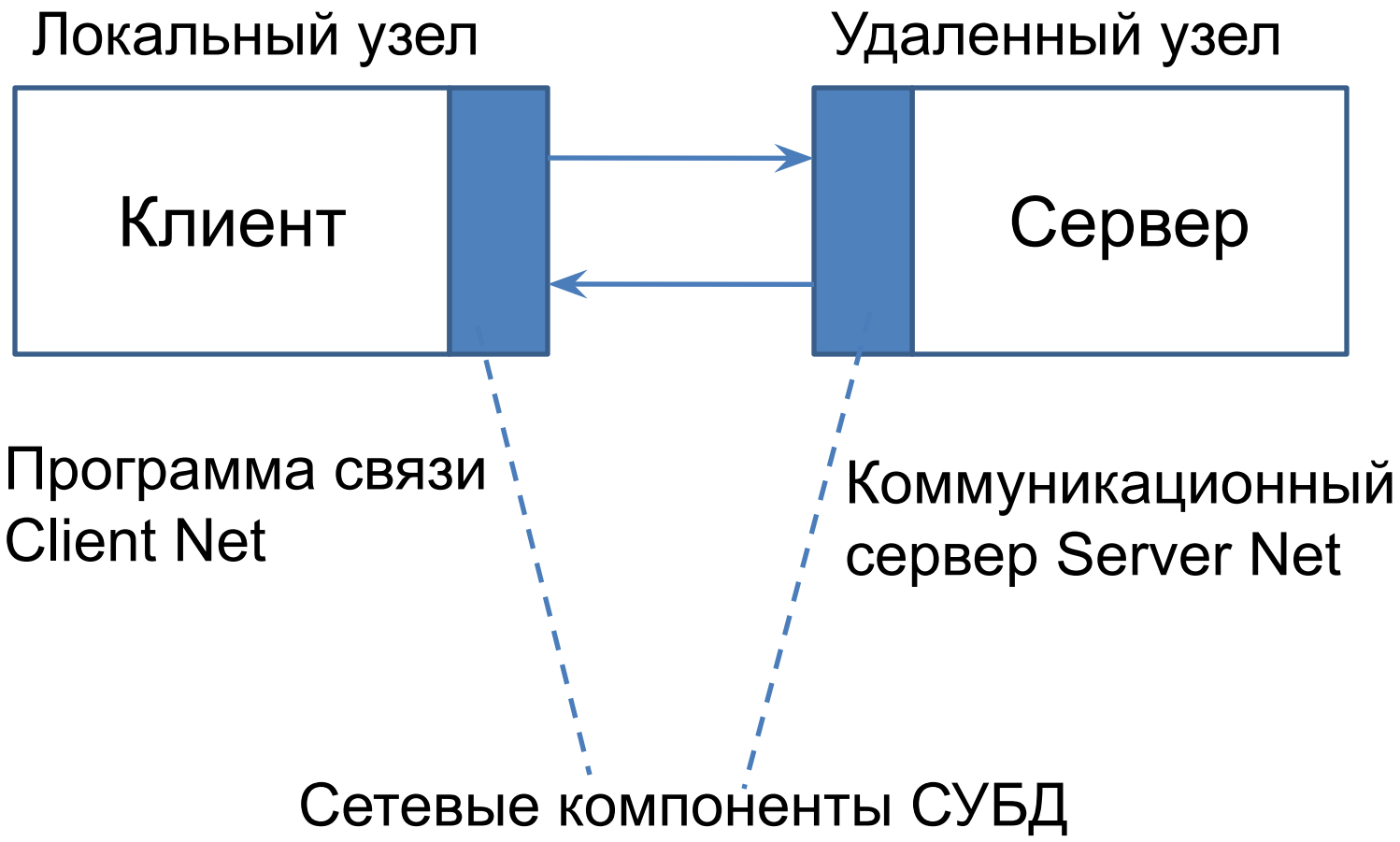
Фундаментальный принцип системы распределенных БД:

Для пользователя система распределенных БД должна выглядеть точно так же, как нераспределенная система

Общая характеристика

- Проблемы сетевого взаимодействия
- Проблемы доступа к данным
- Проблемы распределенных СУБД
- Проблемы тиражирования (репликации) данных
- Проблемы неоднородных систем

Проблемы сетевого взаимодействия



Проблемы сетевого взаимодействия

Требования:

– прозрачность сети
– независимость от аппаратного обеспечения
– автоматическая трансляция кодов
– независимость от СУБД

Проблемы сетевого взаимодействия

Прозрачность сети:

– независимость от использования сетевого аппаратного обеспечения

– независимость от протоколов сетевого обмена

Коммуникационный сервер должен поддерживать как можно более широкий диапазон сетевых протоколов

Проблемы сетевого взаимодействия

Независимость от аппаратного обеспечения:

– необходимость согласования форматов представления данных

Задача коммуникационного сервера – на уровне обмена данными обеспечить согласование их форматов между удаленными и локальными узлами

Проблемы сетевого взаимодействия

Автоматическая трансляция кодов:

– необходимость преобразования кодов символов в соответствии с используемыми таблицами кодов (ASCII, EBCDIC)

Коммуникационный сервер должен решать проблему трансляции кодов для каждой взаимодействующей пары

Проблемы сетевого взаимодействия

Независимость от СУБД:

Все экземпляры СУБД, функционирующие на различных узлах сети, должны поддерживать один и тот же интерфейс

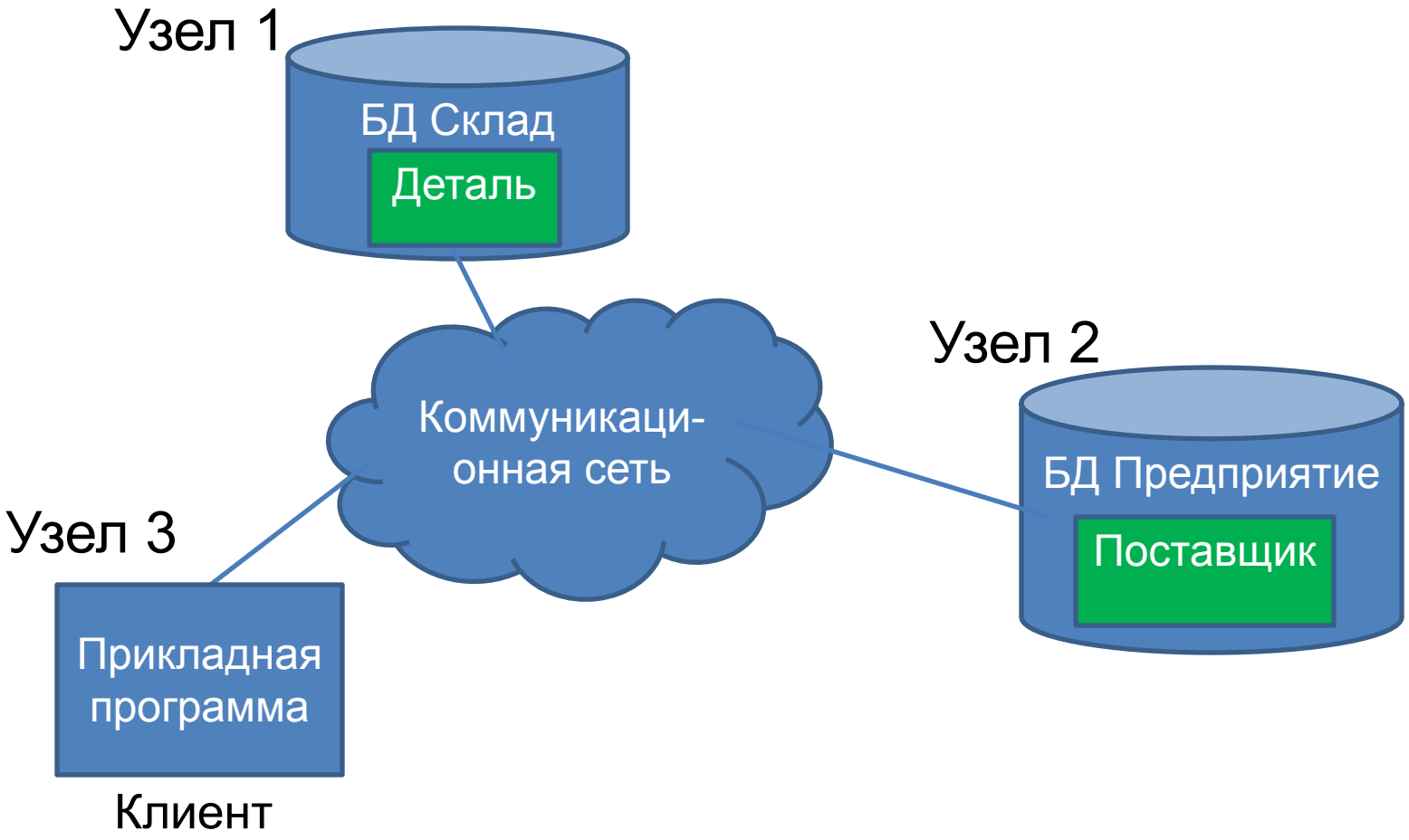
Проблемы доступа к данным

Требования:

– прозрачность (независимость от) расположения
– прозрачность (независимость от) фрагментации

Проблемы доступа к данным

Прозрачность расположения



Проблемы доступа к данным

Имя объекта в MS SQL Server:
сервер.база_данных.пользователь.объект

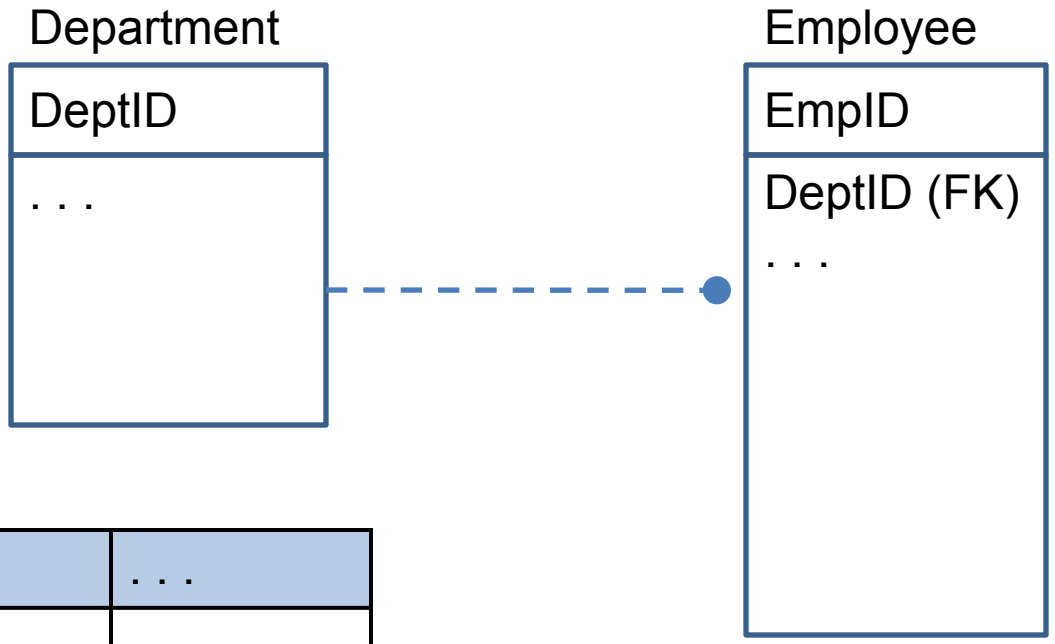
Прозрачный (для пользователя) доступ к удаленным данным предполагает использование в прикладных программах такого интерфейса с сервером БД, который позволяет переносить данные в сети с одного узла на другой, не требуя при этом модификации текста прикладной программы

Проблемы доступа к данным

Прозрачность фрагментации:

В системе поддерживается фрагментация данных, если некое хранимое отношение в целях физического хранения можно разделить на части, или фрагменты, хранимые на разных узлах сети

Проблемы доступа к данным



DeptID	Region	. . .
D1	Москва	. . .
D2	Хабаровск	. . .
. . .	. . .	. . .

Проблемы доступа к данным

Предполагается, что:

– все фрагменты данного отношения независимы

– фрагменты не должны допускать потерю информации

Проблемы распределенных СУБД

Задачи:

– управление именами в распределенной среде
– обработка распределенных запросов
– управление распределенными транзакциями

Проблемы распределенных СУБД

Управление именами в распределенной среде

Организация системного каталога:

1. *Централизованный каталог*
2. *Полностью тиражируемый (реплицированный) каталог*
3. *Секционированный (локальный) каталог*
4. *1 + 3*

. . .

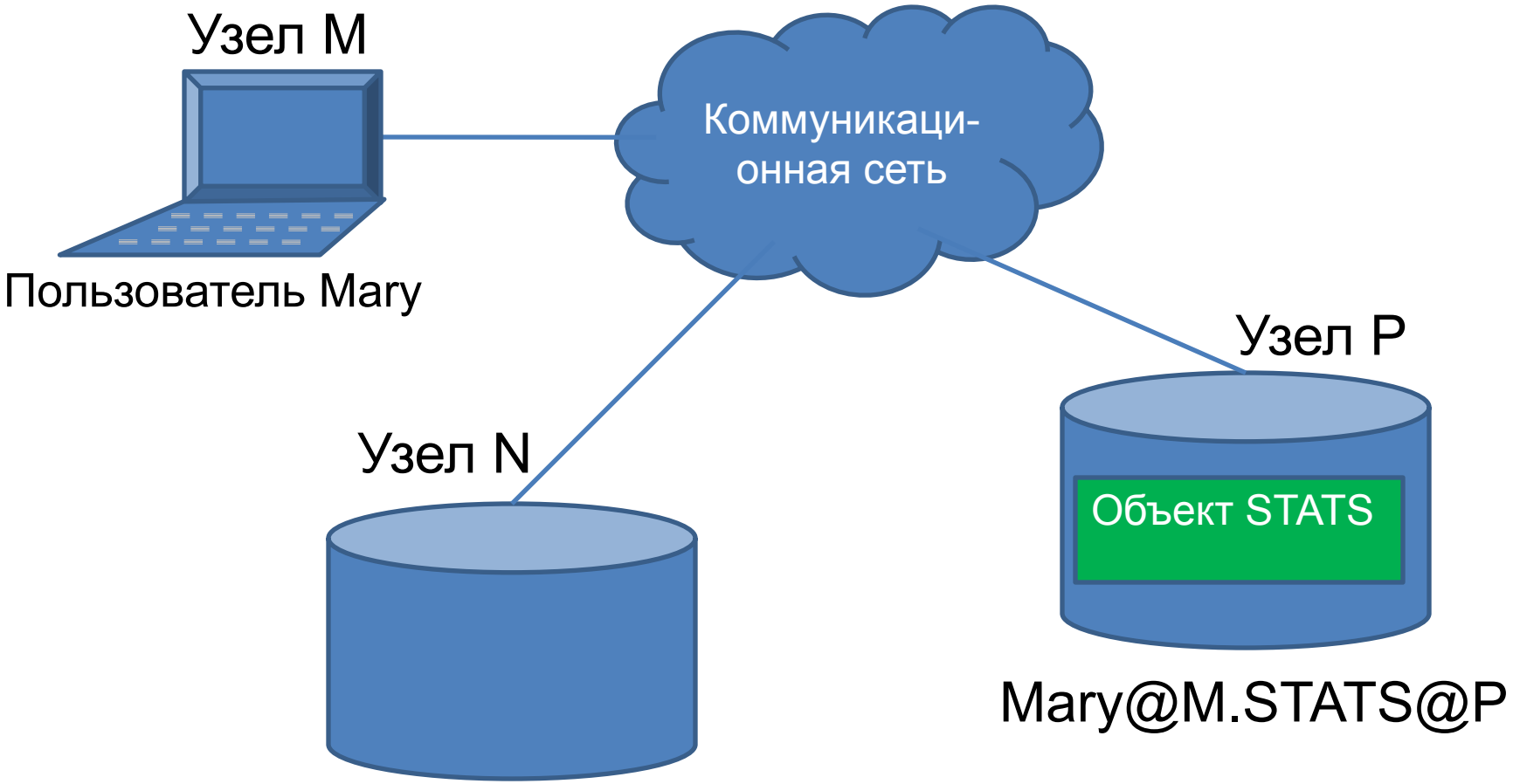
Проблемы распределенных СУБД

Управление именами в System R*

Системное имя объекта:

кто_создал@где_создал.
имя_объекта@где_размещен

Проблемы распределенных СУБД



Проблемы распределенных СУБД

Синоним системного имени:

CREATE SYNONYM *имя_синонима*

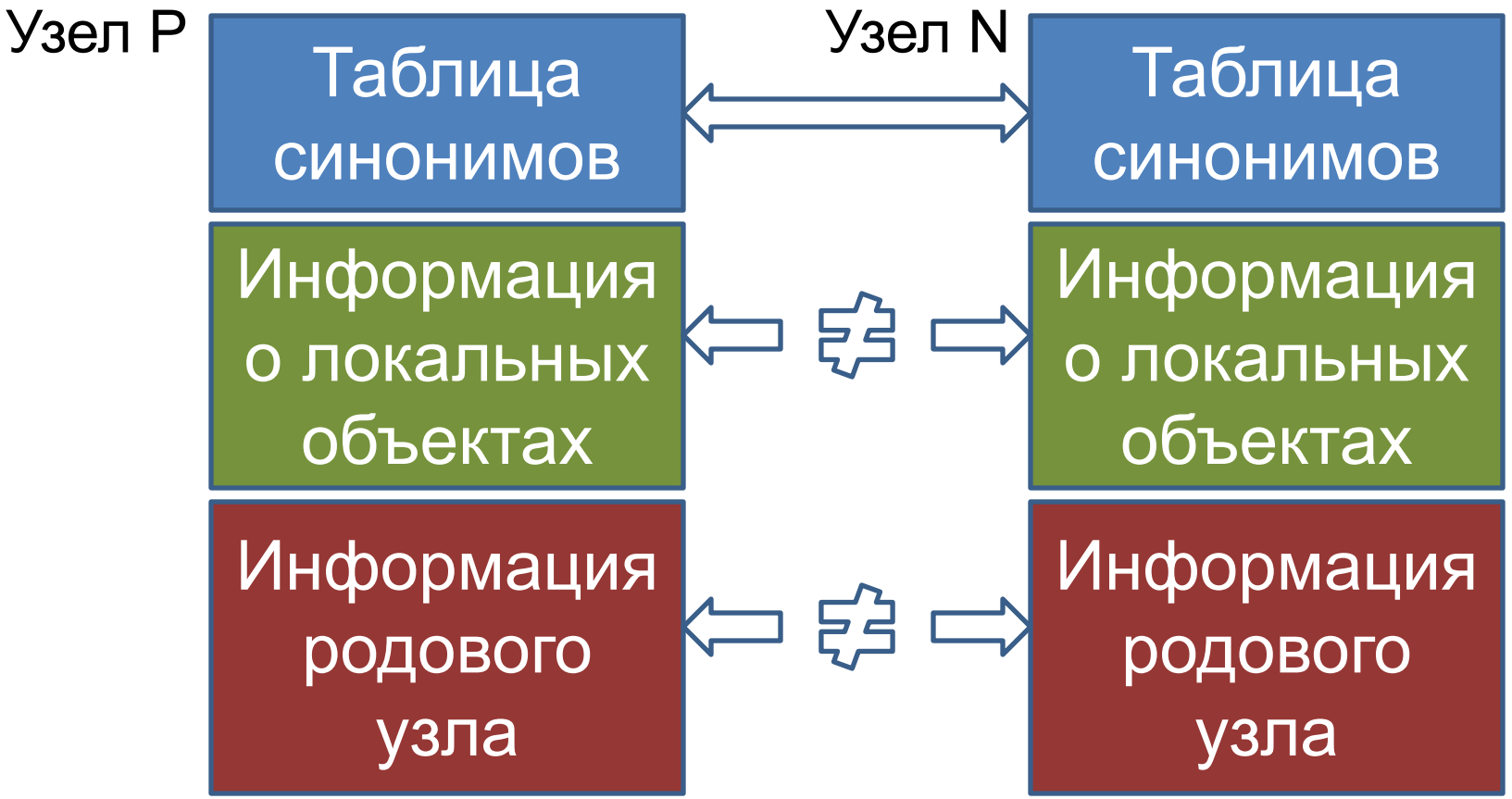
FOR *имя_объекта*

CREATE SYNONYM MSTATS FOR

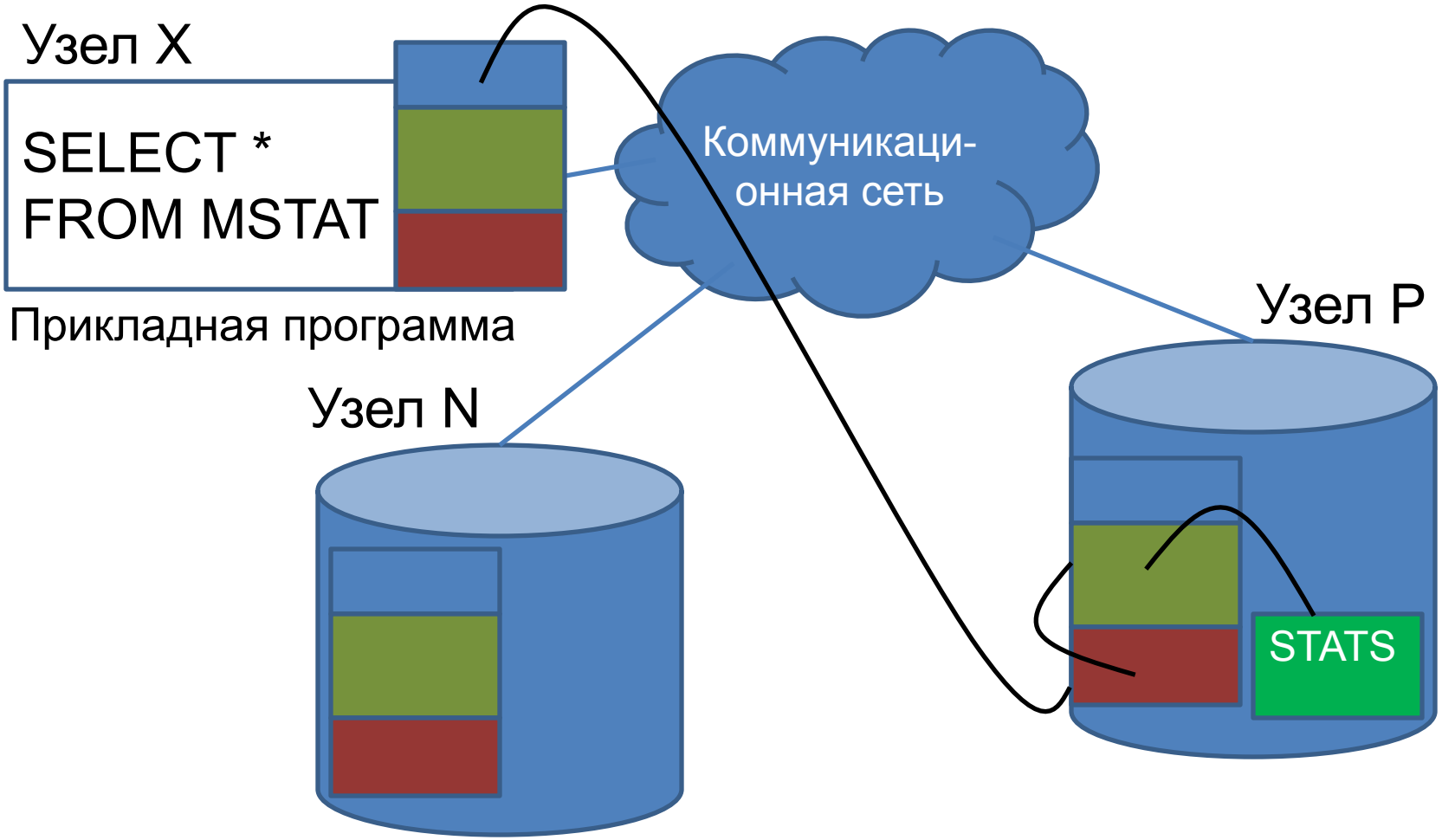
Mary@M.STATS@N

Проблемы распределенных СУБД

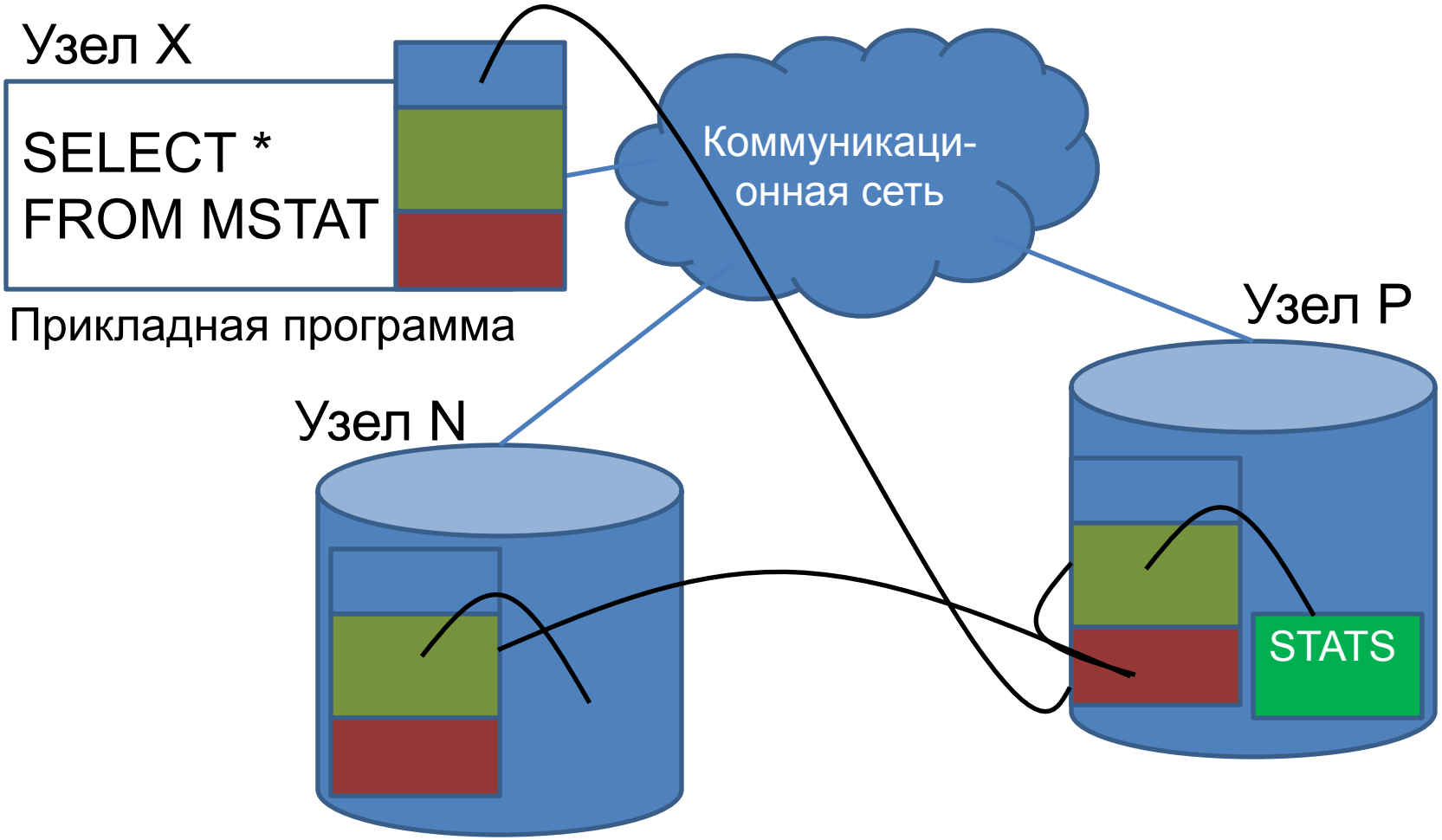
Структура распределенного каталога



Проблемы распределенных СУБД

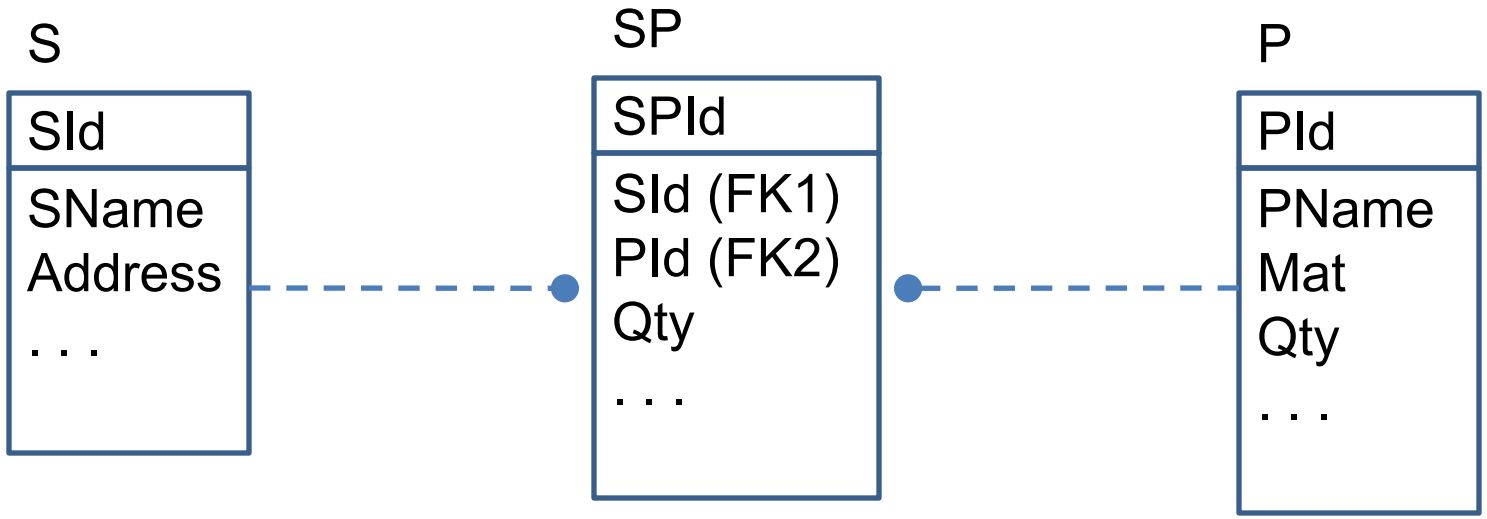


Проблемы распределенных СУБД



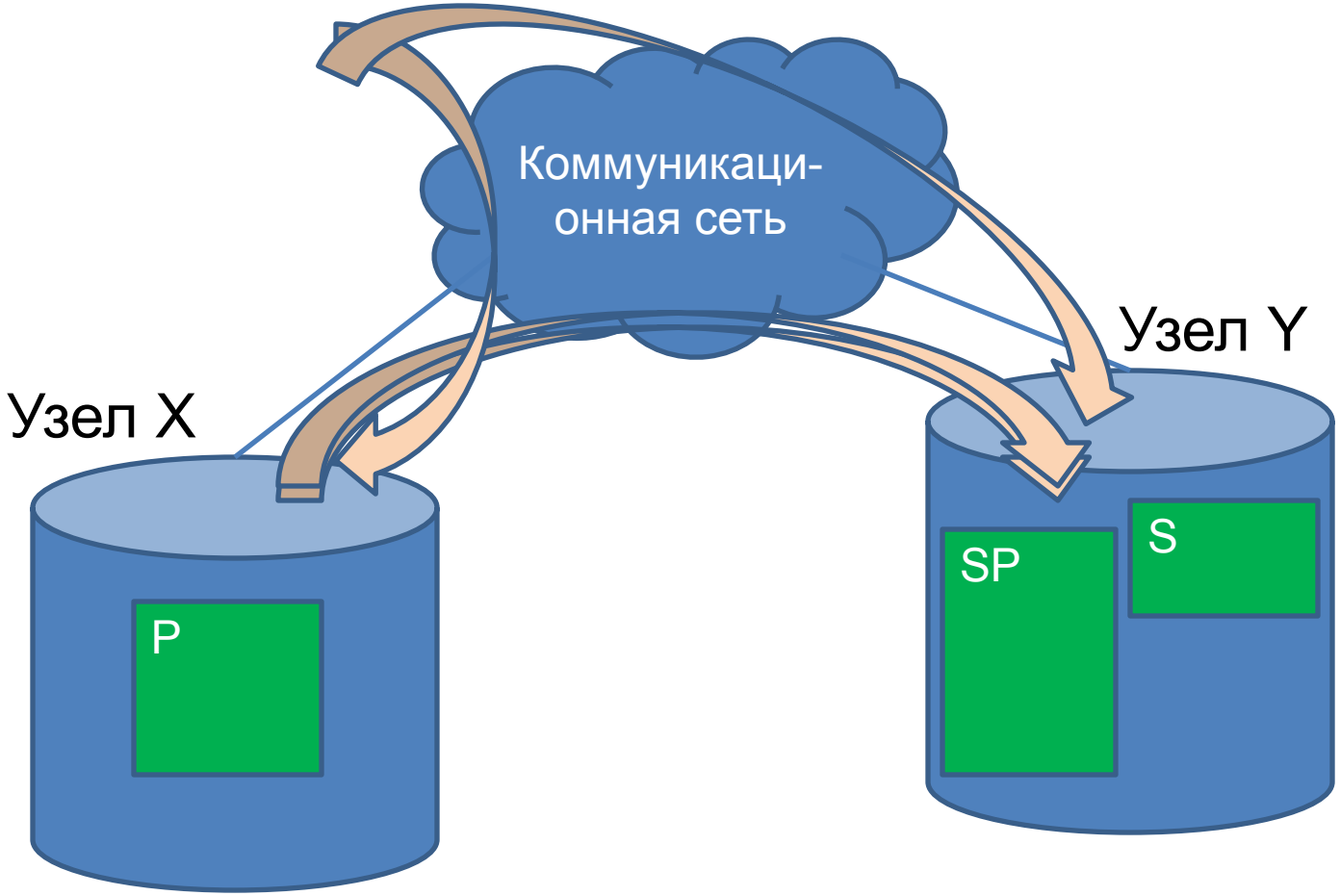
Проблемы распределенных СУБД

Обработка распределенных запросов




```
SELECT S.SName FROM S, P, SP WHERE
   (S.Address = 'NNN' AND S.SId = SP.SId AND
   SP.PId = P.PId AND P.Mat = 'MMM')
```

Проблемы распределенных СУБД



Проблемы распределенных СУБД

Обработка распределенных запросов (query processing) – преобразование декларативного определения запроса в операции манипулирования данными низкого уровня

Проблемы распределенных СУБД

Централизованные СУБД	Распределенные СУБД
• декомпозиция запроса	• декомпозиция запроса • локализация данных • глобальная оптимизация запроса
• оптимизация запроса	• оптимизация запроса

Проблемы распределенных СУБД

Декомпозиция запроса – трансляция с языка SQL в выражение реляционной алгебры

Оптимизация запроса – выбор «наилучшей» стратегии выполнения запроса из множества альтернатив (минимальная сумма затрат, необходимых для выполнения запроса)

Проблемы распределенных СУБД

Локализация данных – преобразование выражения реляционной алгебры с учетом физического размещения данных

Глобальная оптимизация – поиск наилучшей стратегии выполнения запроса с учетом коммуникационных операций пересылки данных

Проблемы распределенных СУБД

Управление распределенными транзакциями

Выполнение транзакции, инициированной в некотором узле сети N, влечет инициирование транзакции и в других узлах:

$T = \{ T_N, T_X, T_Y, \ldots \}$

Проблемы распределенных СУБД

В распределенных БД транзакция, выполнение которой заключается в обновлении данных на нескольких узлах сети, называется *глобальной*, или *распределенной* транзакцией.

Глобальная транзакция состоит из нескольких *агентов*, или *локальных* транзакций.

Проблемы распределенных СУБД

Для глобальной транзакции – свойства АСИД.

Проблемы:

– управление параллелизмом

– управление восстановлением

Проблемы распределенных СУБД

Управление параллелизмом

Также основано на механизме блокировок

Свойство сериализуемости транзакций:

Ни одна блокировка от имени какой-либо транзакции не должна устанавливаться после снятия хотя бы одной ранее установленной блокировки

Проблемы распределенных СУБД

Свойство глобальной сериализуемости:

Выполнение множества распределенных транзакций является сериализуемым тогда и только тогда, когда:

- *выполнение этого множества транзакций является сериализуемым на каждом узле,*
- *порядок сериализации этих транзакций на всех узлах один и тот же*

Проблемы распределенных СУБД

Методы блокирования:

Централизованное блокирование – centralized locking

Распределенное (децентрализованное) блокирование – distributed (decentralized) locking

Проблемы распределенных СУБД

Централизованное блокирование

Единая таблица блокировок для всей распределенной БД, управляемая единым менеджером блокировок

Проблемы:

– производительность
– надежность

Проблемы распределенных СУБД

Распределенное (децентрализованное) блокирование

Управление блокировками распределено между всеми узлами системы; взаимная координация менеджеров блокировок

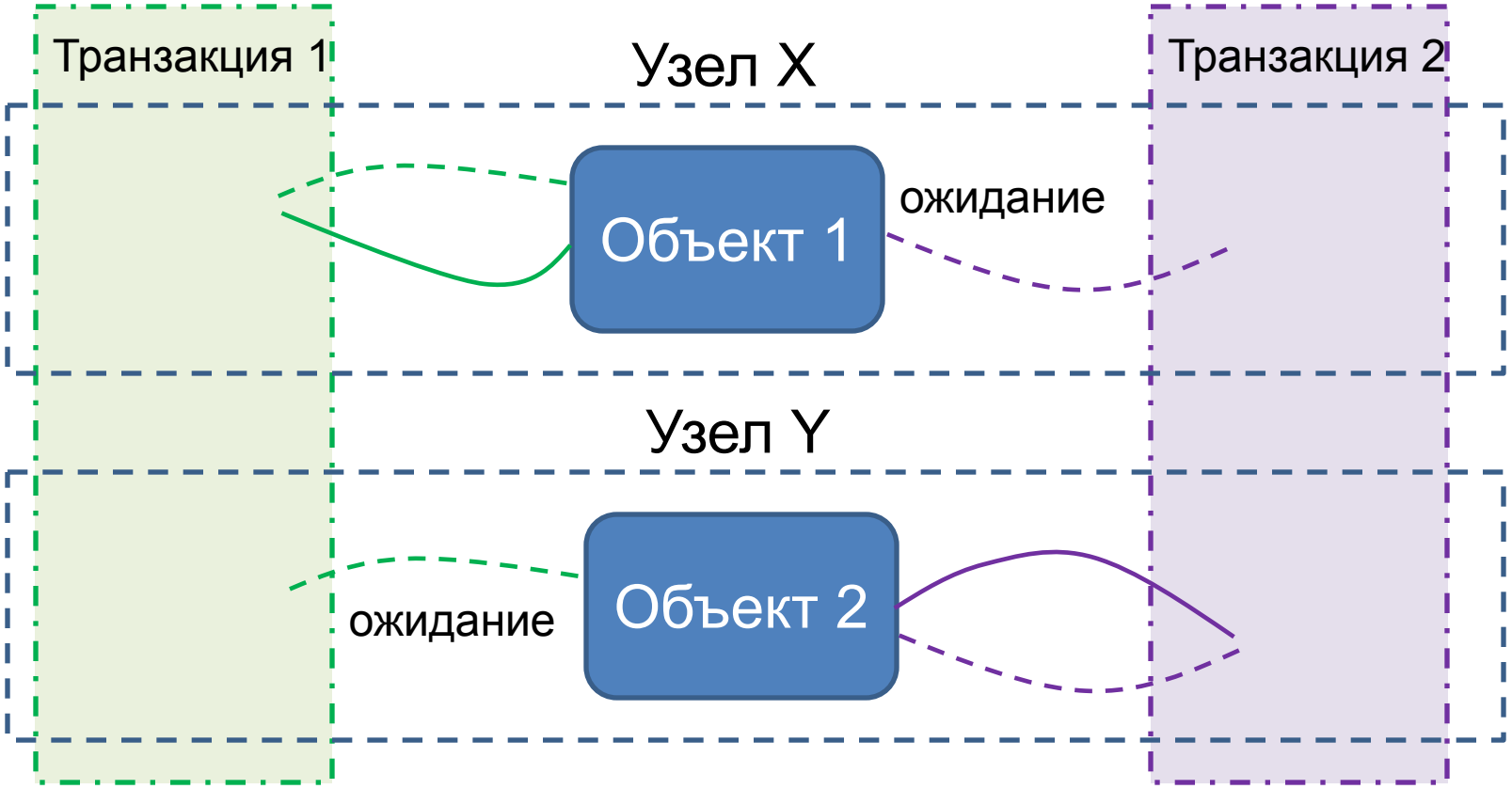
Проблемы:

– более сложные алгоритмы

– выше коммуникационные затраты

Проблемы распределенных СУБД

Проблема тупиков (deadlock)



Проблемы распределенных СУБД

Управление восстановлением

Типы сбоев:

– программный (сбой транзакции)
– мягкий (сбой системы, узла; потеря данных в оперативной памяти)
– жесткий (сбой носителей; потеря данных во внешней памяти)
– коммуникационные сбои

Проблемы распределенных СУБД

Коммуникационные сбои:

– ошибки в сообщениях (сетевой протокол)
– нарушение упорядоченности сообщений (сетевой протокол)
– потерянные (не доставленные) сообщения (СУБД)
– повреждение линий связи (СУБД)

Проблемы распределенных СУБД

Свойства транзакции:

Атомарность – протокол 2PC

Согласованность

Изолированность

Долговременность – протокол распределенного восстановления

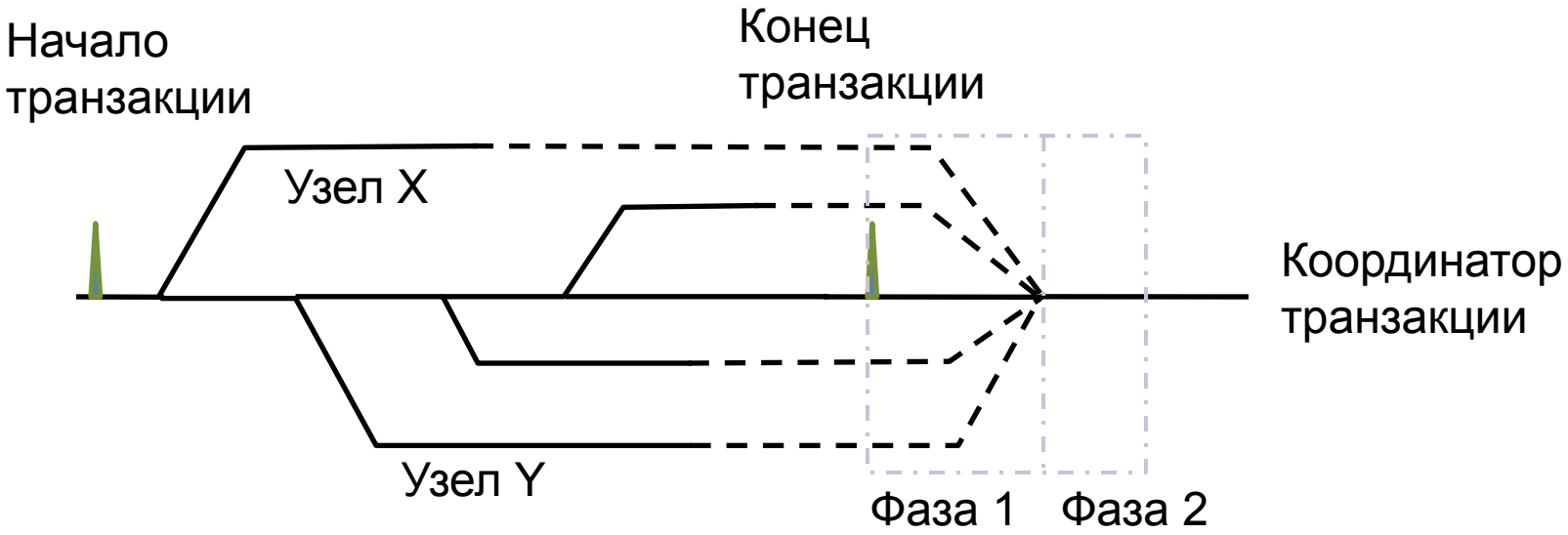
Проблемы распределенных СУБД

Атомарность

$T = \{ T_N, T_X, T_Y, \ldots \}$

- commit – должны быть зафиксированы изменения для всех локальных транзакций
- rollback – должны быть аннулированы изменения для всех локальных транзакций
- Журнал распределенной транзакции

Проблемы распределенных СУБД



Протокол двухфазной фиксации – 2PC (two-phase commit)

Проблемы распределенных СУБД

- Фаза 1 – Подготовиться к фиксации

  Локальные журналы транзакций, ответ
- Фаза 2 – Принятие решения

  Глобальный журнал транзакции, фиксация решения, информирование участников

Проблемы распределенных СУБД

Важно:

- Каждый участник глобальной транзакции должен делать то, что ему предписано координатором во время фазы 2
- Именно появление записи решения в журнале координатора отмечает переход с фазы 1 на фазу 2.

Проблемы распределенных СУБД

Особенности:

- Функция координатора выполняется узлом, на котором инициирована распределенная транзакция.
- Координатор должен обмениваться данными с каждым узлом-участником.
- Локальные узлы – участники процесса двухфазной фиксации должны выполнять любые действия, предписанные координатором, и теряют локальную автономность.

Проблемы распределенных СУБД

Проблемы:

1. Односторонний выбор участником аварийного завершения

Узел X проголосовал за откат транзакции – не ожидает ответа от координатора

Проблемы распределенных СУБД

Проблемы:

2. Блокирующий характер протокола 2PC

Узел X проголосовал за фиксацию транзакции; ожидает ответа от координатора – сбой на линии связи

Проблемы распределенных СУБД

Протокол восстановления

- Ищется запись в журнале координатора: есть – можно восстановить, нет – откат
- Потеря связи с локальным узлом:
  - во время фазы 1 – откат транзакции
  - во время фазы 2 – попытки завершить транзакцию, пока связь не будет восстановлена

Проблемы распределенных СУБД

- Сбой координатора:
  - до начала процедуры фиксации: начать процесс фиксации после восстановления
  - координатор в состоянии готовности (фаза 1): перезапустить процедуру фиксации после восстановления
  - после принятия решения (фаза 2): никаких действий

Технология тиражирования данных

Концепции:

- Отказ от распределения данных
- Все данные дублируются на каждом узле сети (где они обрабатываются)
- Транзакции в системе выполняются и завершаются локально

Технология тиражирования данных

В системе поддерживается репликация данных (Data Replication), если заданное хранимое отношение или заданный фрагмент могут быть представлены несколькими разными копиями, или репликами, хранимыми на разных узлах сети

Технология тиражирования данных

Независимость от репликации: пользователи, по крайней мере, с логической точки зрения, должны работать таком режиме, как будто данные не реплицированы вовсе.

Технология тиражирования данных

Преимущества:

- данные всегда расположены там, где они обрабатываются;
- большая доступность: пока остается доступной хотя бы одна реплика;
- большая надежность хранения данных: всегда можно восстановить целостное состояние БД, если существует хотя бы одна ее реплика на каком-либо узле сети.

Технология тиражирования данных

Главный недостаток:

- нарушение тождественности всех копий

Требование:

при обновлении некоторого реплицированного объекта все копии этого объекта также должны обновляться (проблема тиражирования обновлений).

Технология тиражирования данных

- Тиражирование данных – это асинхронный перенос изменений объектов исходной БД в принимающие БД, принадлежащие различным узлам распределенной системы

В составе СУБД – сервер тиражирования данных (репликатор)

Технология тиражирования данных

Главная проблема – нарушение целостности данных:

- При последовательном обращении к разным копиям – когда передавать информацию об изменениях?
- При параллельном обращении – нужно ли (и как) запрещать доступ к данным со стороны других пользователей?

Технология тиражирования данных

Задача 1 – стратегия обновления копий:

- синхронное обновление
- асинхронное обновление

Задача 2 – стратегия доступа к данным:

- все копии доступны для обновления
- только некоторые копии доступны для обновления

Технология тиражирования данных

Стратегии доступа к данным

Только некоторые копии доступны для обновления – концепция первичной копии:

- изменения выполняются только на узле, выделенном для первичной копии
- на остальных узлах – только чтение данных
- за тиражирование изменений отвечает узел первичной копии

Технология тиражирования данных

Технология тиражирования данных

Концепция первичной копии:
одновременный доступ – за счет блокировок первичной копии

Используется:

- в системах поддержки принятия решений
- в системах поддержки мобильных пользователей

Технология тиражирования данных

Все копии доступны для изменения – проблемы с точки зрения синхронизации обновлений:

- обеспечение целостности данных
- обеспечение доступности данных

Оптимистические и пессимистические протоколы управления транзакциями

Технология тиражирования данных

Пессимистические протоколы – предпочтение обеспечению целостности: на отдельных узлах сети не допускается выполнение любых транзакций, для которых не существует гарантии не нарушения целостности базы данных

Алгоритмы управления параллелизмом:

- метод 2PC
- метод распределенных блокировок

Технология тиражирования данных

Оптимистические протоколы – предпочтение обеспечению доступности данных: допускается независимое обновление данных в каждой копии, даже если после объединения всех изменений вероятен переход базы данных в несогласованное состояние

Технология тиражирования данных

Стратегии обновления копий

Тиражирование обновлений:

- синхронное
- асинхронное

Технология тиражирования данных

Синхронное обновление: обновление всех копий – часть самой транзакции; используется протокол 2PC, но по сети передаются только изменения данных

Недостатки:

- транзакция не может быть завершена, если один из узлов недоступен
- дополнительная нагрузка на сеть

Технология тиражирования данных

Асинхронное обновление: обновление целевых баз данных после выполнения обновлений исходной базы данных.

Задержка – от нескольких секунд до нескольких часов и даже дней

Гарантируется, что в какой-то момент времени данные во всех копиях будут синхронизированы

Технология тиражирования данных

Кто инициирует распространение обновлений:

- узел, на котором выполнены изменения
- узел, которому нужны обновленные данные

Технология тиражирования данных

Репликации в MS SQL Server

Терминология:

- Издатель – Publisher: сервер, который предоставляет информацию из своих баз данных другим серверам
- Подписчик – Subscriber: сервер, копирующий информацию от издателя

Технология тиражирования данных

- Дистрибьютор – Distributor: промежуточный север, принимающий данные от издателя и распространяющий их подписчикам

Технология тиражирования данных

- Публикация – набор статей для обновления, принадлежащих одной базе данных
- Статья – минимальный набор данных, рассматриваемый системой репликации как одно целое (обычно – таблица базы данных)

Технология тиражирования данных

Функции издателя:

- создание публикации
- отслеживание изменений, вносимых в данные
- подготовка публикации к тиражировнию

Технология тиражирования данных

Типы репликации

1. Только издатель может изменять публикацию
   - репликация моментальных снимков
2. Все могут изменять публикацию
   - подписчики незамедлительного обновления
   - репликация сведением
   - отложенные обновления

Технология тиражирования данных

Репликация моментальных снимков – Snapshot Replication

Для тиражирования данных используются моментальные снимки – полная копия публикации, сохраняемая в специальном файле

Технология тиражирования данных

Подписчики незамедлительного обновления – Immediate Updating Subscriber

Подписчик, изменяя свою копию данных, одновременно должен выполнить изменение данных на издателе

- Нет конфликтов изменения данных
- Постоянное соединение между подписчиком и издателем

Технология тиражирования данных

Репликация сведением – Merge Replication

Самый сложный тип репликации

- Не требуется постоянное соединение подписчика с издателем
- Подписчики работают автономно, накапливая изменения данных
- На издателе объединяются все изменения данных

Технология тиражирования данных

На издателе могут быть обнаружены конфликты изменений

Специальные алгоритмы разрешения конфликтов, в основе которых – «шкала приоритетов»

Технология тиражирования данных

Отложенное обновление – Queue Updating

Обновления, выполненные на подписчике, применяются на издателе с некоторой задержкой

- Постоянное соединение с издателем отсутствует
- Соединение периодически устанавливается

Технология тиражирования данных

Подписчик записывает информацию о выполненных изменениях в очередь; если информация об изменениях не может быть записана в очередь – изменения не фиксируются

Запомненные в очереди данные переносятся на издатель

Также возможны конфликты изменений

Технология тиражирования данных

Методы обновления информации на подписчиках

1. Принудительная репликация – Push Subscription

- Инициатор – издатель
- Требуется постоянное соединение
- Интервалы обновления подписчиков устанавливаются на дистрибьюторе

Технология тиражирования данных

2. Репликация по запросу – Pull Subscription

- Инициатор – подписчик
- Для каждого подписчика на дистрибьюторе – свой набор данных, отражающий изменения
- Не требуется постоянное соединение

Хранилища данных

Основные понятия

Системы оперативной обработки транзакций – Online Transaction Processing (OLTP)

Системы поддержки принятия решений – Decision Support System (DSS)

Усовершенствованная технология баз данных:

– специальные средства управления процессом хранения информации
– мощные инструменты анализа накопленных данных

Определение

Bill Inmon, 1993 г.

Хранилище данных (Data Warehouse) – это предметно-ориентированный, интегрированный, привязанный ко времени и неизменяемый набор данных, предназначенный для поддержки принятия решений

Сравнение систем

1. Характер данных

OLTP + баз данных	DSS + хранилища данных
Текущие данные	Исторические данные
Подробные сведения	Обобщенные данные
Динамические данные	Статические данные

Сравнение систем

2. Обработка данных

OLTP + базы данных	DSS + хранилища данных
Повторяющийся способ обработки	Нерегламентированный, неструктурированный, эвристический способ
Высокая интенсивность обработки транзакций	Средняя и низкая интенсивность обработки транзакций
Предсказуемый способ использования	Непредсказуемый способ использования

Сравнение систем

3. Назначение системы

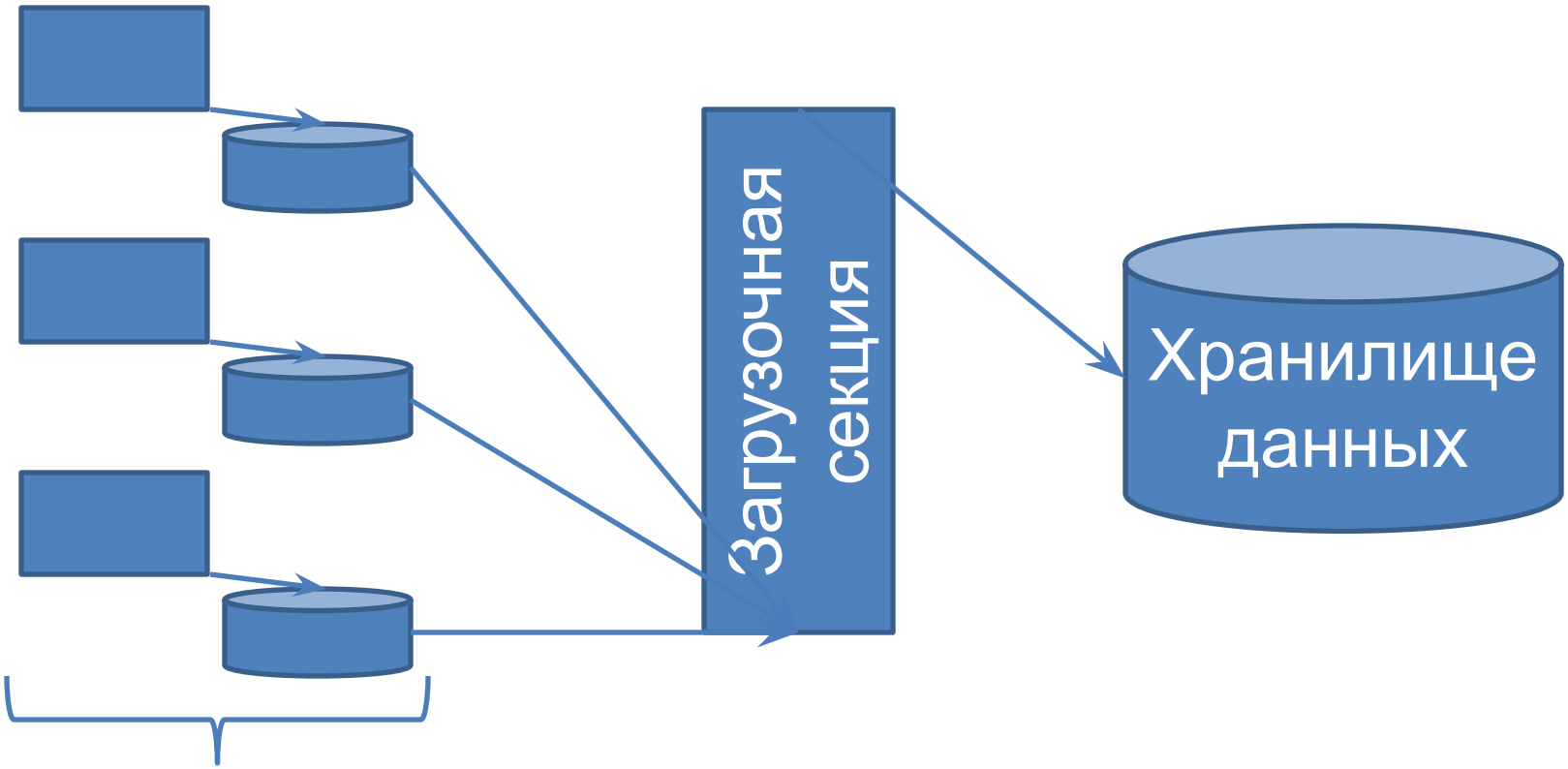
OLTP + базы данных	DSS + хранилища данных
Обработка транзакций	Проведение анализа
Ориентирована на прикладную область	Ориентирована на предметную область
Поддержка принятия повседневных решений	Поддержка принятия стратегических решений

Сравнение систем

4. Пользователи

OLTP + баз данных	DSS + хранилища данных
Обслуживает большое количество пользователей исполнительного звена	Обслуживает относительно небольшое количество работников руководящего звена

Конфигурация хранилища данных



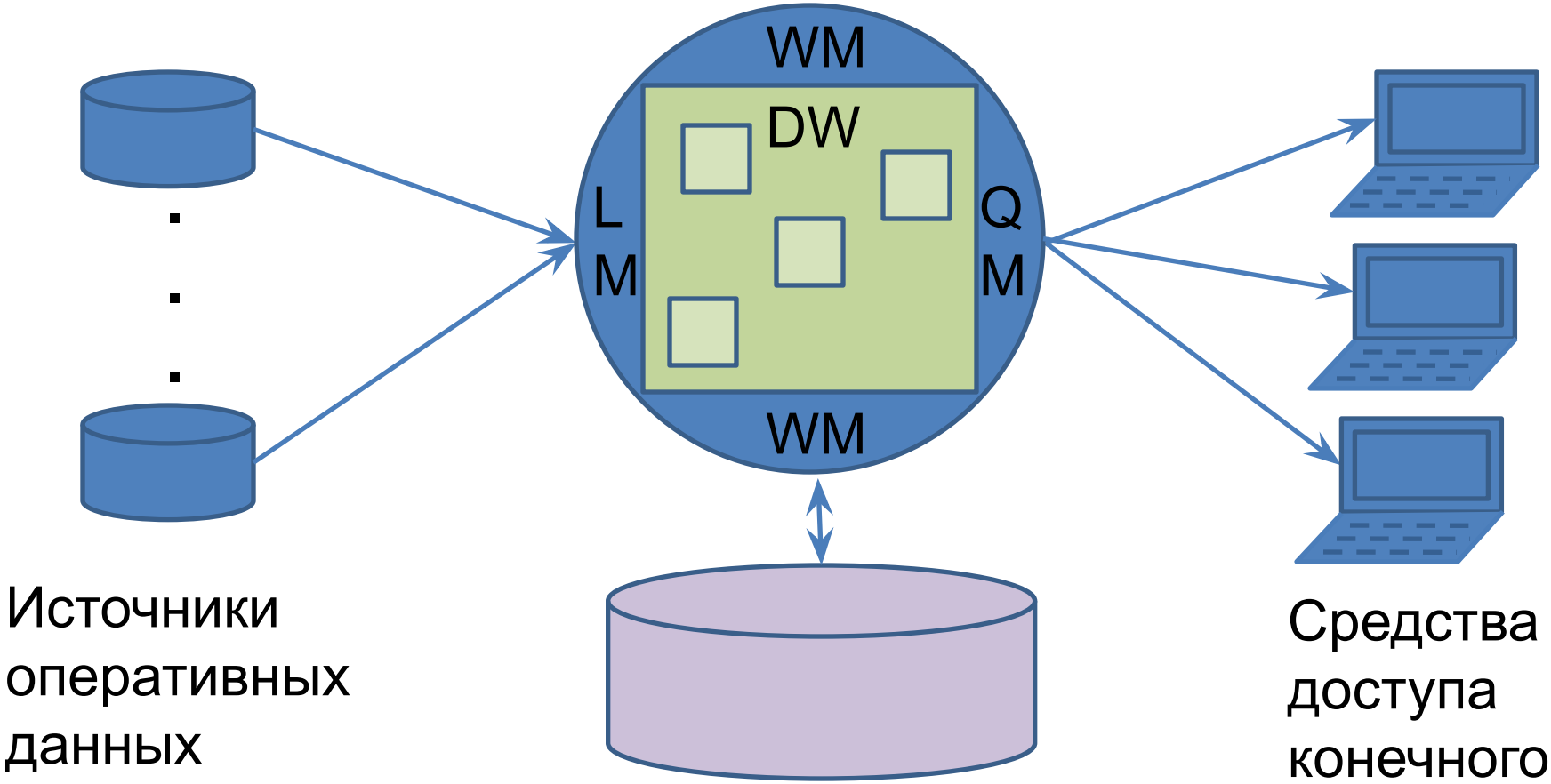
OLTP-системы
источники данных

Загрузочная секция

Назначение:

- устранение несогласованности, фрагментарности, дубликатов и пропусков – очистка данных (data scrubbing)
- обеспечение совместимости данных с другими источниками – расслоение (slicing) и расщепление (dicing) данных

Архитектура хранилища данных



Источники оперативных данных

Архив и резервные копии

Средства доступа конечного пользователя

Архитектура хранилища данных

Менеджер загрузки – Load Manager (LM): внешний (front-end) компонент; извлечение данных, загрузка данных в хранилище

- инструменты репликации информации
- генераторы кода
- механизмы динамического преобразования

Архитектура хранилища данных

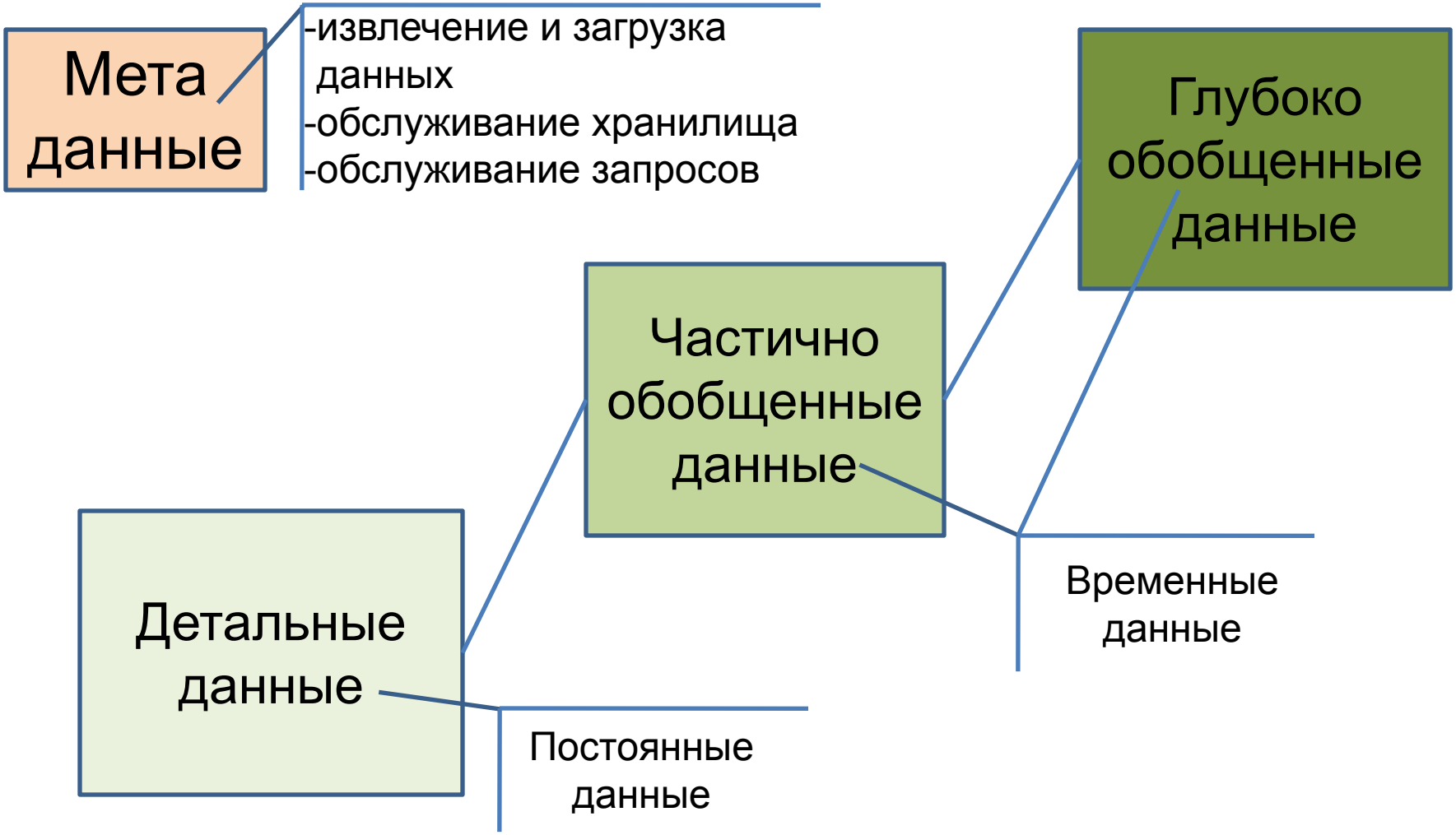
Менеджер хранилища – Warehouse Manager (WM): управление информацией, помещенной в хранилище данных

- анализ непротиворечивости данных
- создание необходимых индексов
- денормализация
- обобщение
- резервное копирование

Архитектура хранилища данных

Менеджер запросов – Query Manager (QM): внутренний (back-end) компонент; управление запросами пользователей. Создается на базе предоставляемых СУБД инструментов доступа к данным и инструментов мониторинга хранилища

Структура хранилища данных



Средства доступа к данным

1. Инструменты информационной системы руководителя – Executive Information System (EIS; сейчас – Everybody Information System); предоставление поддержки управляющему персоналу всех уровней.

Предопределенный набор сценариев обработки данных и составления отчетов

Express Analyzer фирмы Oracle

Средства доступа к данным

2. Инструменты оперативной аналитической обработки – Online Analytical Processing (OLAP); оценка эффективности деятельности предприятия, предсказание объемов продаж и планирование товарных запасов.

Построение и выполнение нерегламентированных запросов

Express Server фирмы Oracle

Средства доступа к данным

3. Инструменты разработки данных – Data mining; открытие новых осмысленных корреляций, распределений и тенденций, создание предсказательных, а не ретроспективных моделей.

Создание предсказательных моделей

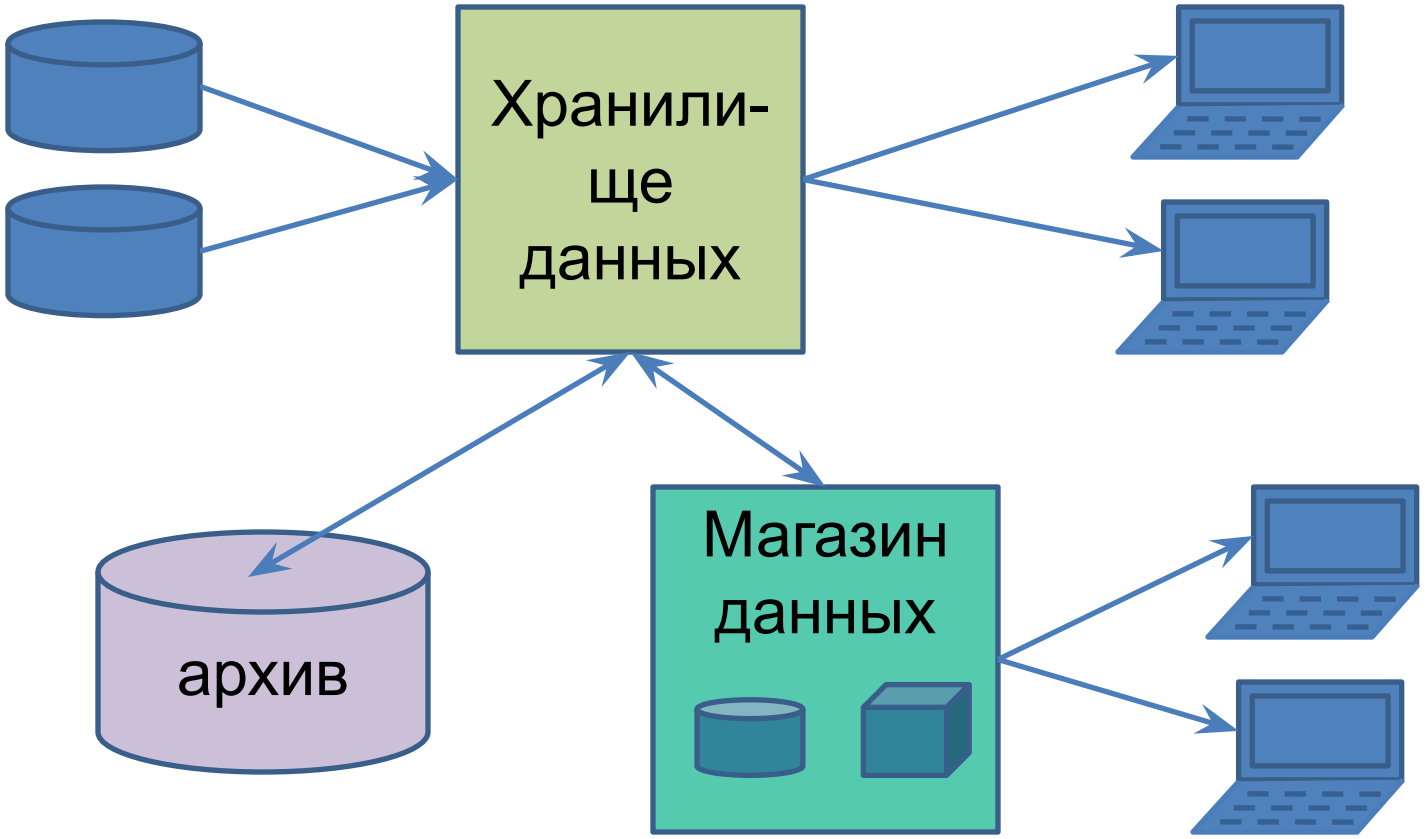
Intelligent Miner фирмы IBM

Витрины данных

Data Mart – (магазины данных) – подмножество хранилища данных, которое поддерживает требования отдельного подразделения или деловой сферы организации

- доступ к данным, которые приходится анализировать чаще других
- предоставление данных в форме, соответствующей коллективному представлению подразделения
- сокращение времени ответа на вопрос

Витрины данных



Витрины данных

Отличие от хранилища данных:

- отвечает требованиям только одного из подразделений организации или некоторой ее деловой сферы
- обычно не содержит детальных оперативных сведений
- структура информации более понятна и проста в управлении

Проектирование хранилища данных

	Базы данных	Хранилища данных
Исходные данные к информационному моделированию	Бизнес логика	Цель исследований
Критерий информационного моделирования	Достоверность и согласованность данных	Время выполнения запросов
Загрузка данных	Ручная, в соответствии с бизнес логикой	Автоматическая загрузка по расписанию из оперативных источников
Информационная модель	Диаграмма сущность – связь	Схема типа «звезда»

Схема типа «звезда»

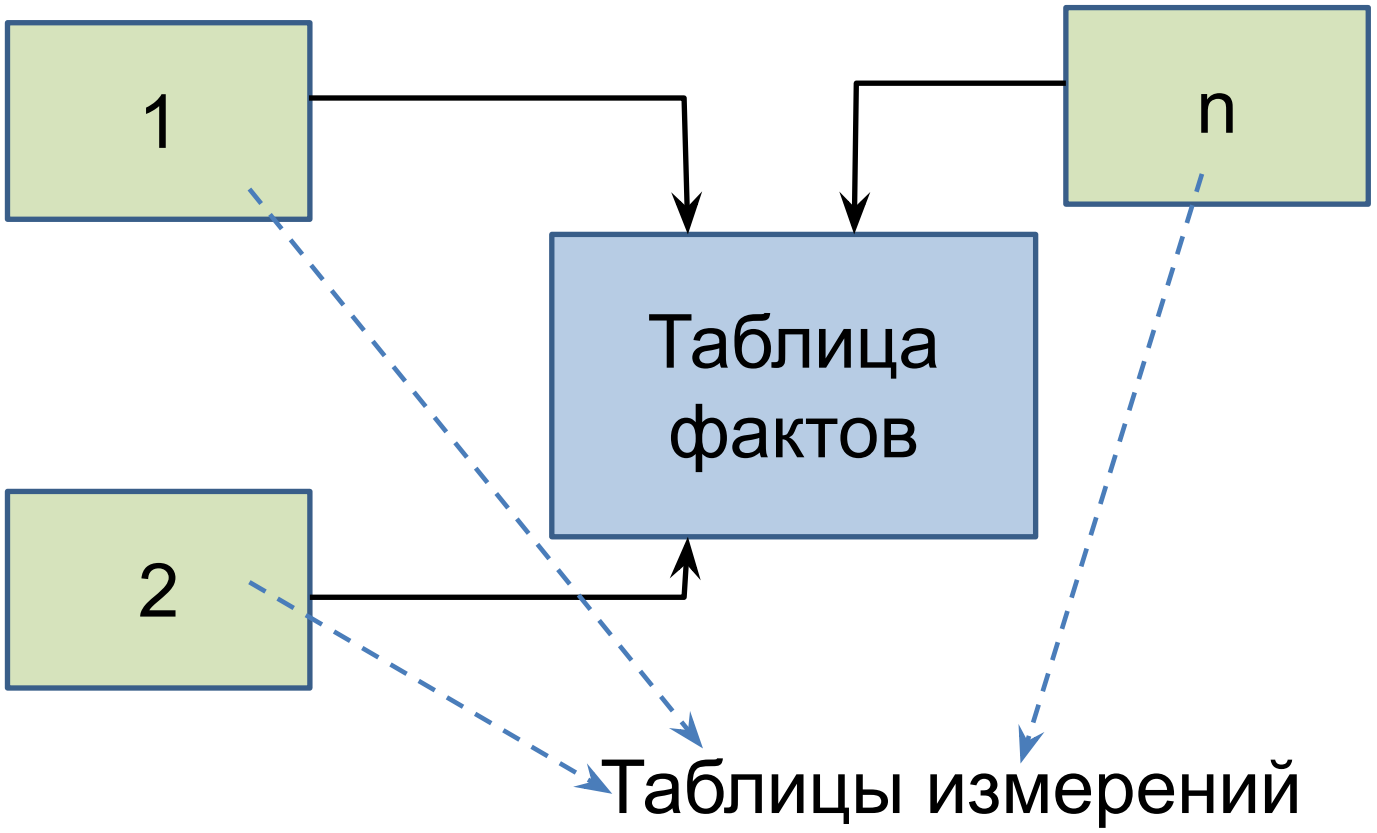


Схема типа «звезда»

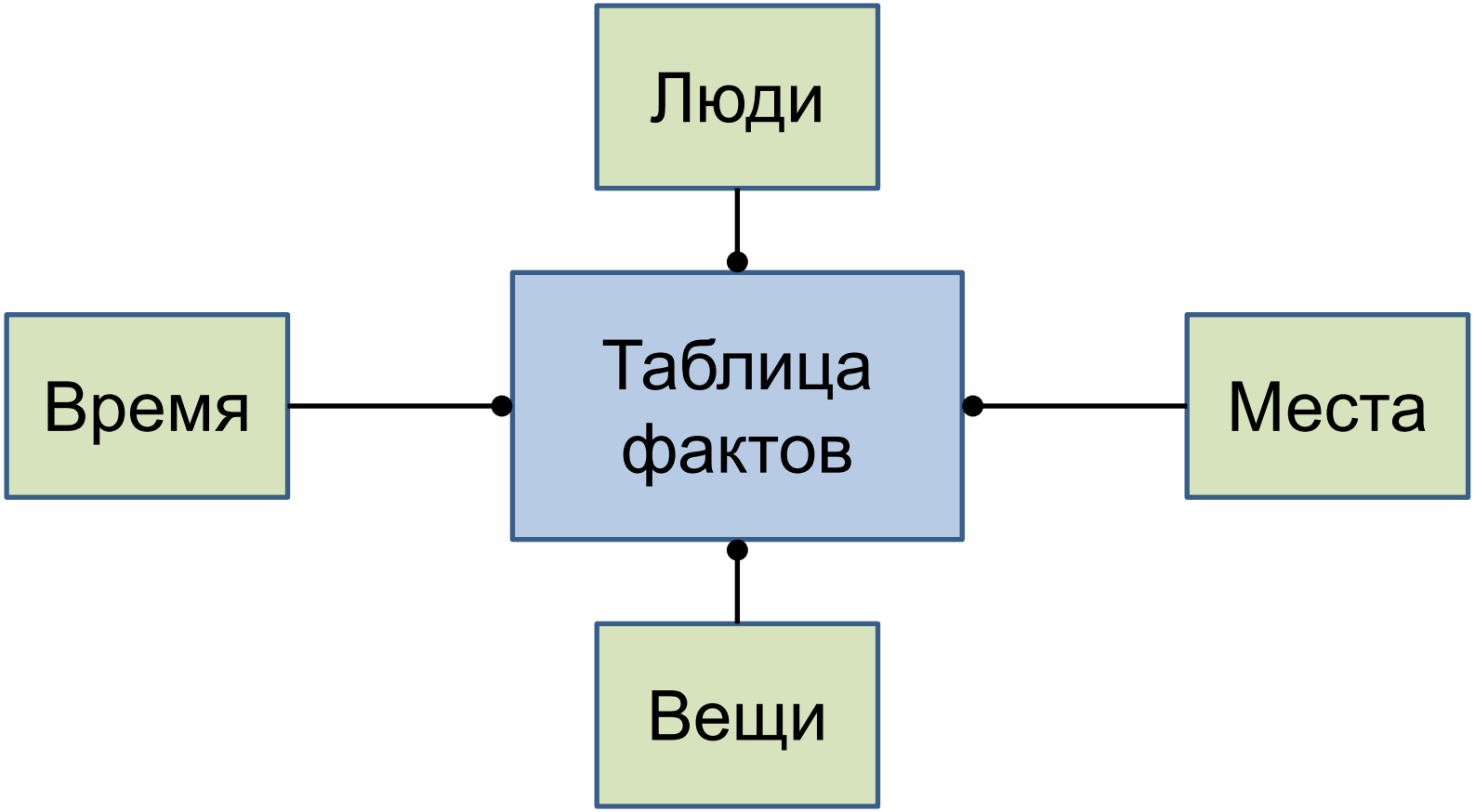
Таблица фактов (fact table) – количественные значения; деловые факты, определяющие фактическую сущность; детальные данные, представляющие собой основные виды бизнес деятельности организации и факторы, влияющие на данный бизнес или его сектор

Схема типа «звезда»

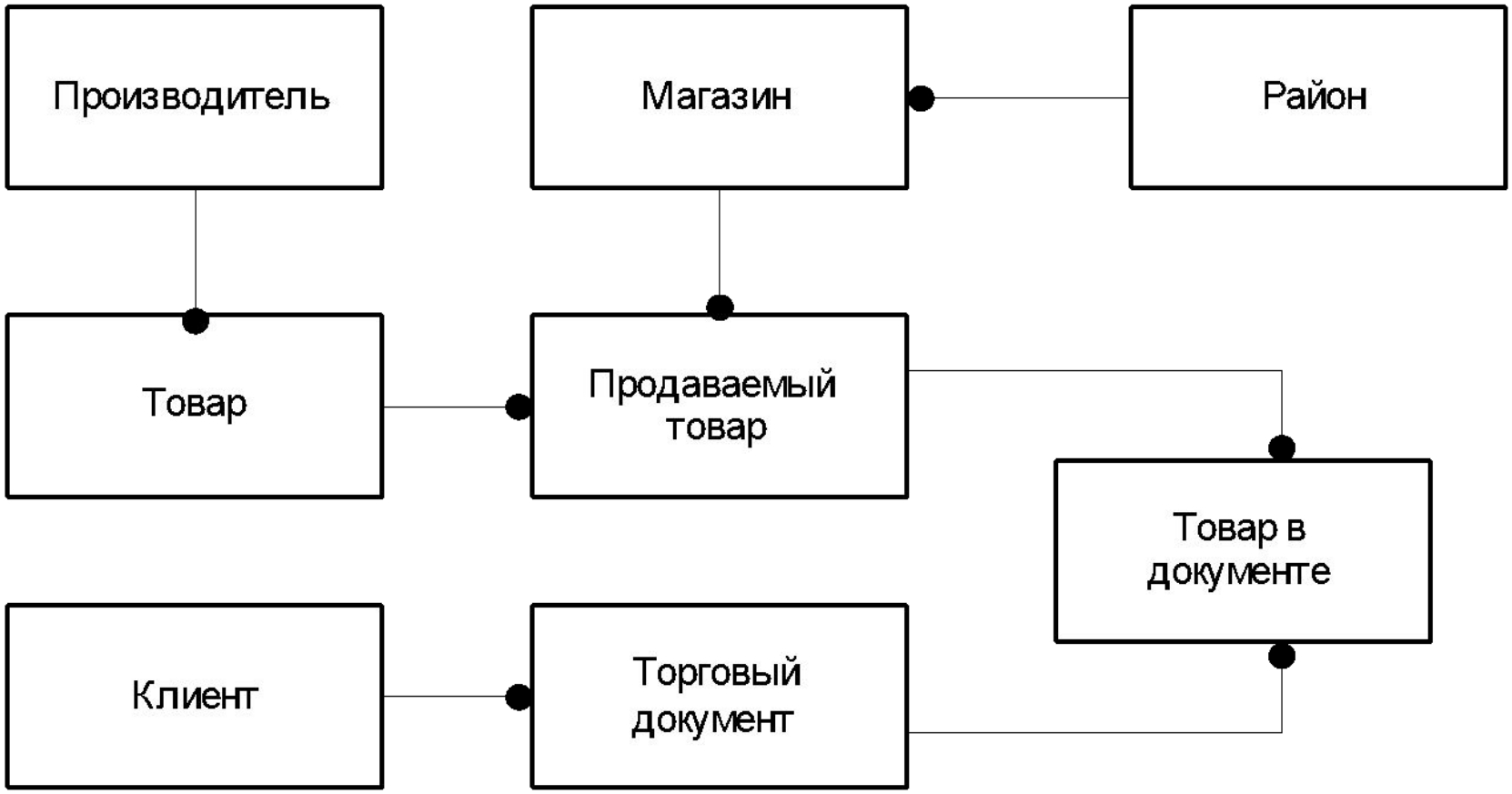
Таблицы измерений (dimension tables) – *дескриптивные* (описательные) значения; справочные данные, или данные деловых измерений; элементы, которые могут оказывать определенное влияние или порождать различные тенденции в развитии фактов

Схема типа «звезда»

Категории измерений



Пример проектирования



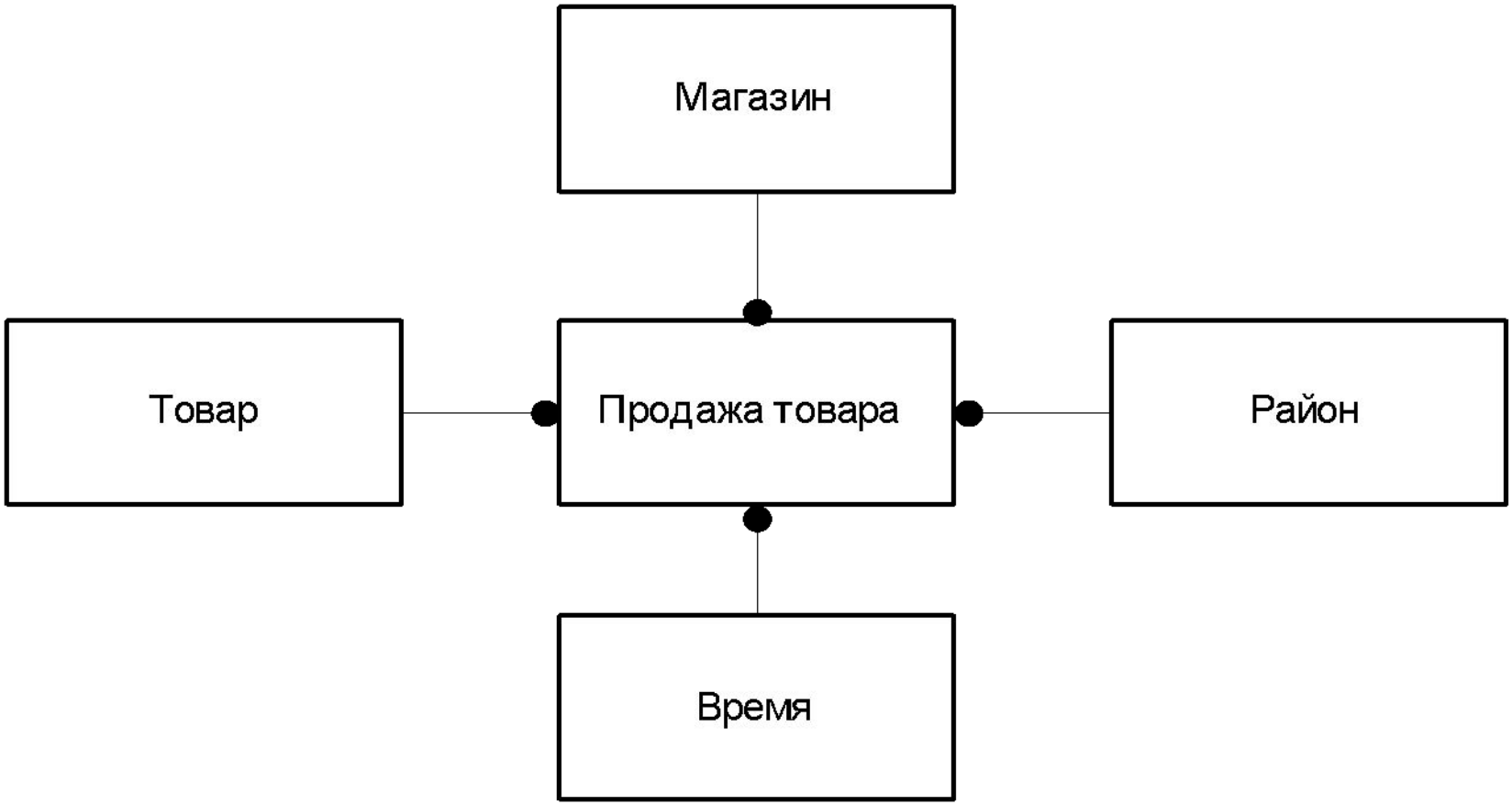
Области применения ИС

Управление повседневными бизнес процессами (OLTP)

Поддержка принятия стратегических решений (OLAP, Data mining)

Управление информационным содержанием

Пример проектирования



Особенности проектирования

Таблица фактов:

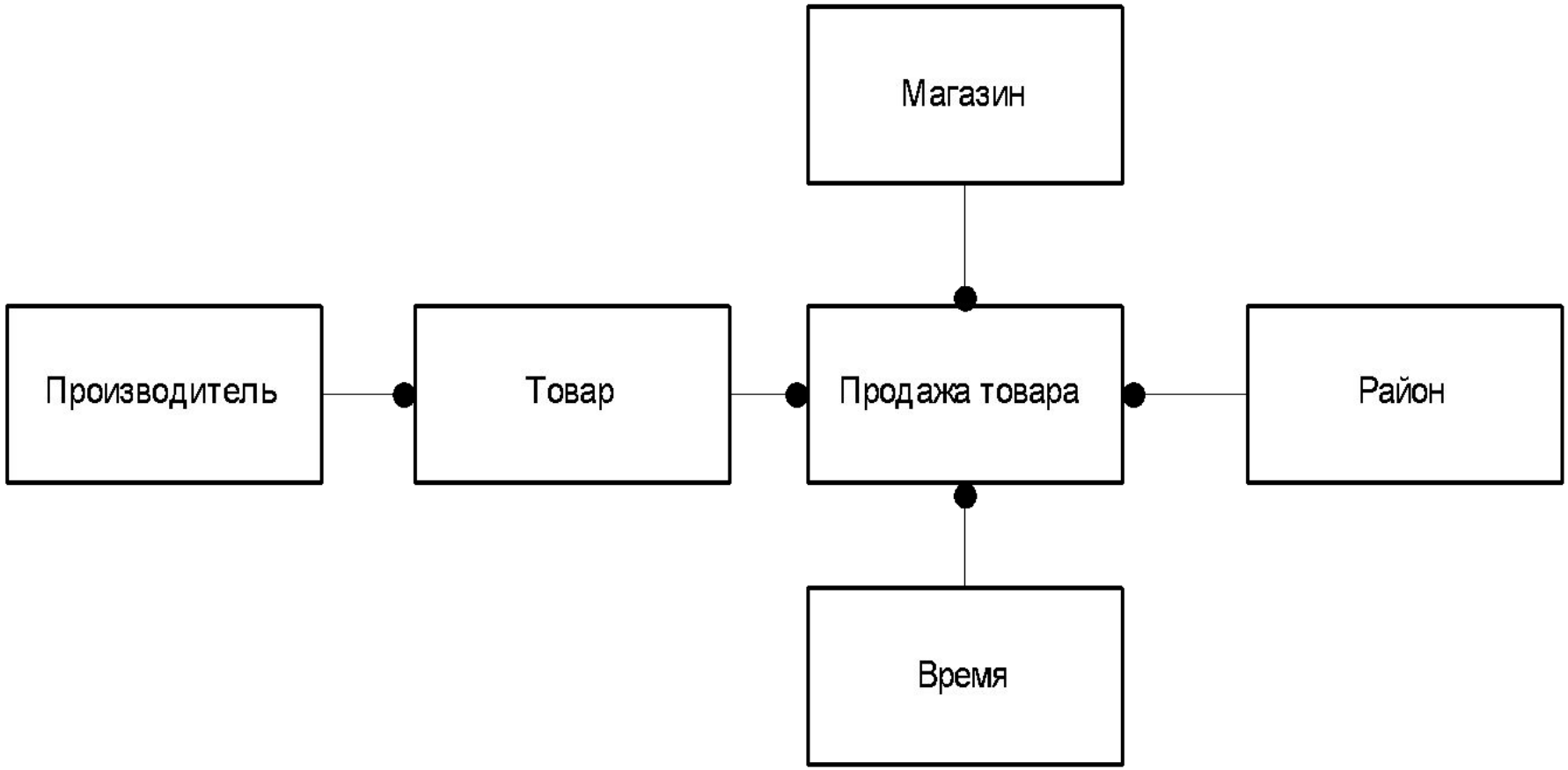
- использование суррогатного ключа
- вычисляемые колонки (объем продаж, стоимость в . . .)
- секционирование
  - вертикальное (восстановление – через join)
  - горизонтальное (восстановление – через union)

Особенности проектирования

Таблицы измерений:

- существующие таблицы OLTP базы данных (*Товар*, *Магазин*)
- новые измерения (из других таблиц базы данных – *Район* или из элементов таблиц базы данных – *Время*)
- денормализация таблицы измерений
- развертывание измерений – схема типа «снежинка»

Особенности проектирования



Технология OLAP

Назначение OLAP (Online Analytical Processing) инструментов: предоставить средства извлечения большого количества записей и вычисления на их основе некоторых итоговых значений.

Термин OLAP был предложен Коддом в 1993 г. и определяет архитектуру, которая поддерживает сложные аналитические приложения.

Технология OLAP

Критерий FASMI:

Fast – время отклика:

- среднее ~ 5 сек;
- для простых запросов - ~ 1 сек;
- для самых сложных - ~ 20 сек;
- более 30 сек – недопустимо

Технология OLAP

Analysis – система должна справляться с любым логическим и статистическим анализом, характерным для данного приложения; пользователь может определять новые вычисления как часть анализа и формировать нужные отчеты без необходимости программирования

Технология OLAP

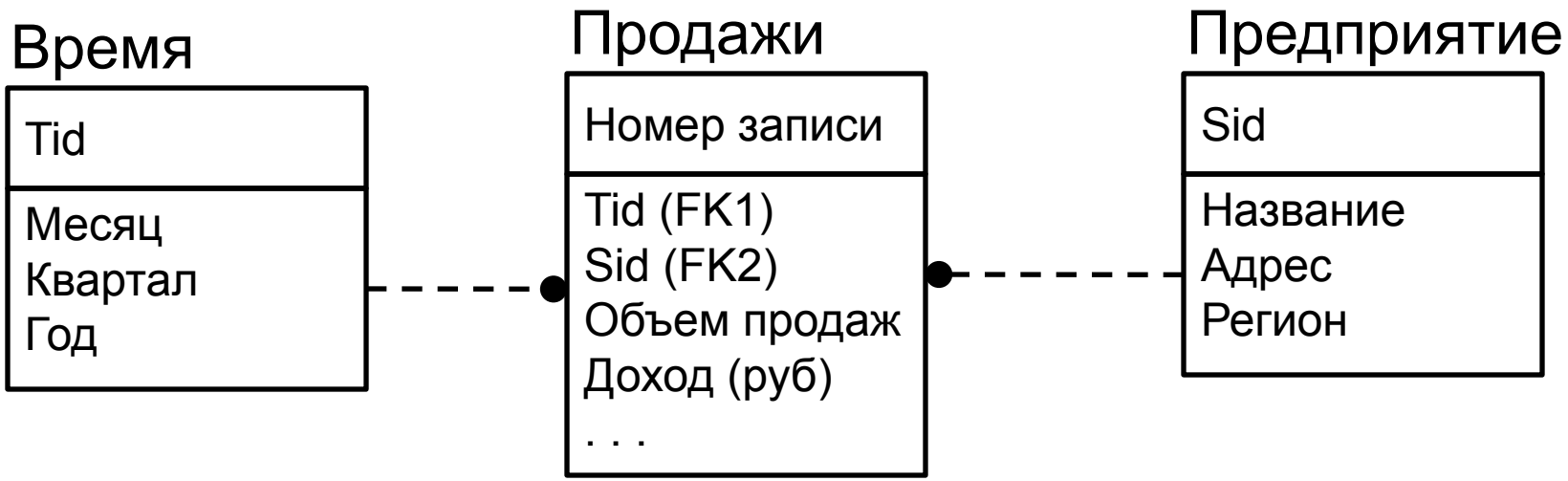
Shared – широкие возможности разграничения доступа к данным и одновременной работы многих пользователей

Multidimensional – должно быть обеспечено многомерное концептуальное представление данных

Information – необходимая информация должна быть получена там, где она необходима

Многомерное представление

Анализ изменения объема продаж и дохода торговых предприятий во времени



Многомерное представление

Таблица РБД («плоская»)

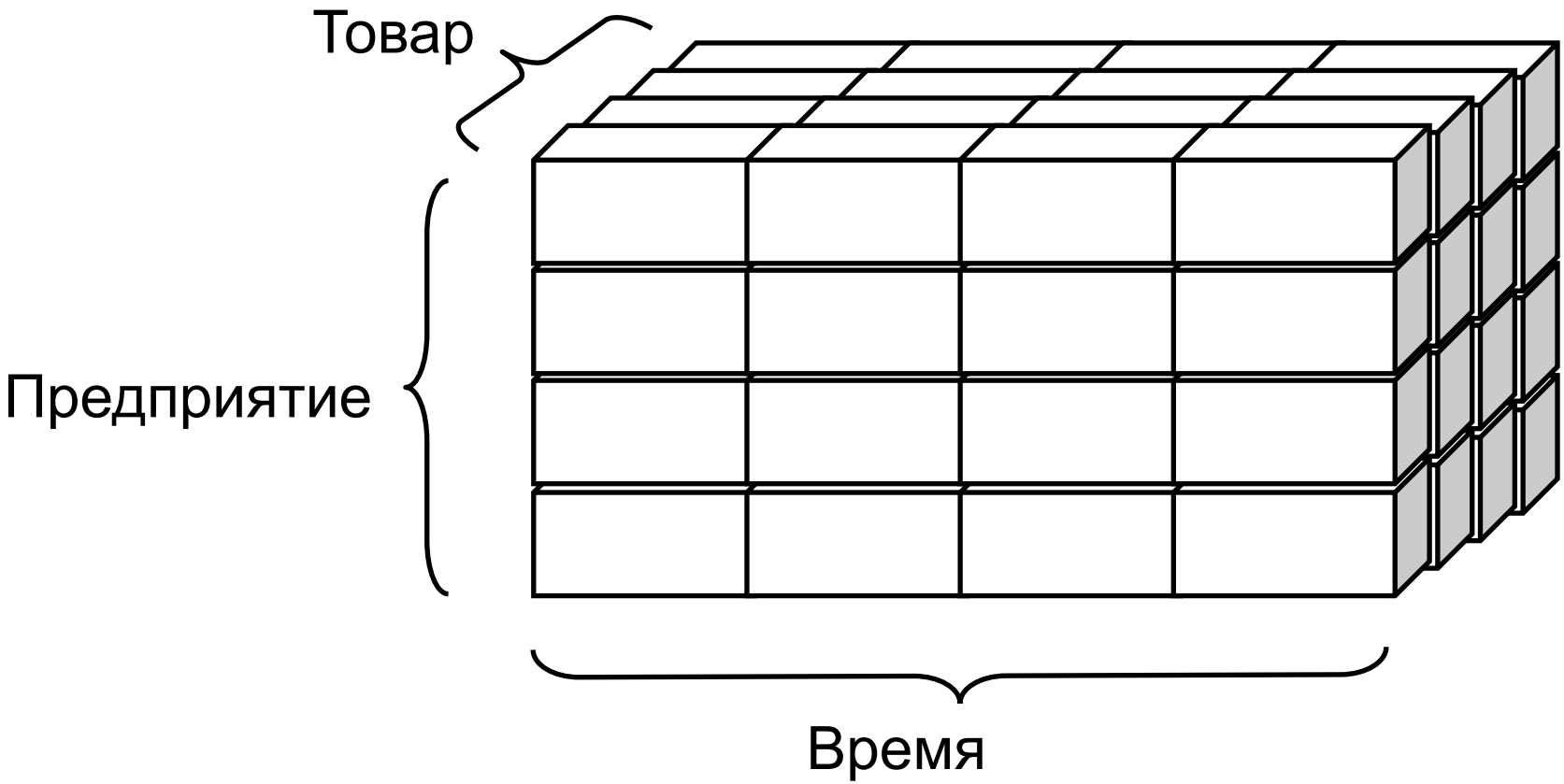
Tid	Sid	Объем продаж	Доход	. . .
1	1	k_{11}	s_{11}	. . .
1	2	k_{12}	s_{12}	. . .
1	3	k_{13}	s_{13}	. . .
1	4	k_{14}	s_{14}	. . .
. . .	. . .	. . .	. . .	. . .
2	1	k_{21}	s_{21}	. . .
. . .	. . .	. . .	. . .	. . .

Многомерное представление

Двухмерное представление

Tid / Sid	1	2	3	. . .
1	$k_{11}, s_{11}, \ldots$	$k_{12}, s_{12}, \ldots$	$k_{13}, s_{13}, \ldots$	. . .
2	$k_{21}, s_{21}, \ldots$	$k_{22}, s_{22}, \ldots$	$k_{23}, s_{23}, \ldots$	. . .
3	$k_{31}, s_{31}, \ldots$	$k_{32}, s_{32}, \ldots$	$k_{33}, s_{33}, \ldots$	. . .
. . .	. . .	. . .	. . .	. . .

Многомерное представление



Многомерное представление

Достоинства многомерных структур:

- очень компактны
- обеспечивают простые средства просмотра и манипулирования элементами данных, обладающих многими взаимосвязями

Многомерное представление

Достоинства многомерных структур:

- легко расширяются при включении новой размерности
- допускают выполнение операций матричной арифметики, позволяющих легко вычислять средние и общие значения

Многомерное представление

«Типичная реляционная СУБД способна сканировать всего несколько сотен строк в секунду, тогда как типичная многомерная СУБД способна выполнять обобщающие операции со скоростью до 10000 строк в секунду и даже выше.»

[Коннолли Т. и др.]

Аналитические операции

- Консолидация – обобщающие операции, такие как простое суммирование значений (свертка), или расчет с использованием сложных выражений, включающих другие связанные данные

Аналитические операции

- Нисходящий анализ (drill-down) – операция, обратная консолидации; включает возможность отображения подробных сведений для рассматриваемых консолидированных данных;

Аналитические операции

- Разбиение с поворотом (slicing and dicing) – также называется созданием сводной таблицы; позволяет получить представление данных с разных точек зрения. Например, одно представление – сведения о доходах от продаж товаров указанного типа по каждому району, другое представление – данные о доходах магазинов в каждом районе

Аналитические операции

Предварительное обобщение, использование иерархической структуры размерностей и управление заполнением пространства кубов позволяют значительно сократить размер базы данных и исключить потребность многократного вычисления одних и тех же значений

Правила для OLAP систем

E. Codd, 1993 г.

- Многомерное концептуальное представление данных
- Доступность (доступ к требуемым для анализа данным)
- Неизменная производительность подготовки отчетов (количество измерений, степень обобщения данных)

Правила для OLAP систем

- Неограниченные перекрестные операции между размерностями
- Неограниченное число измерений и уровней обобщения
- Гибкость средств формирования отчетов

Категории OLAP инструментов

Berson and Smith, 1997 г.

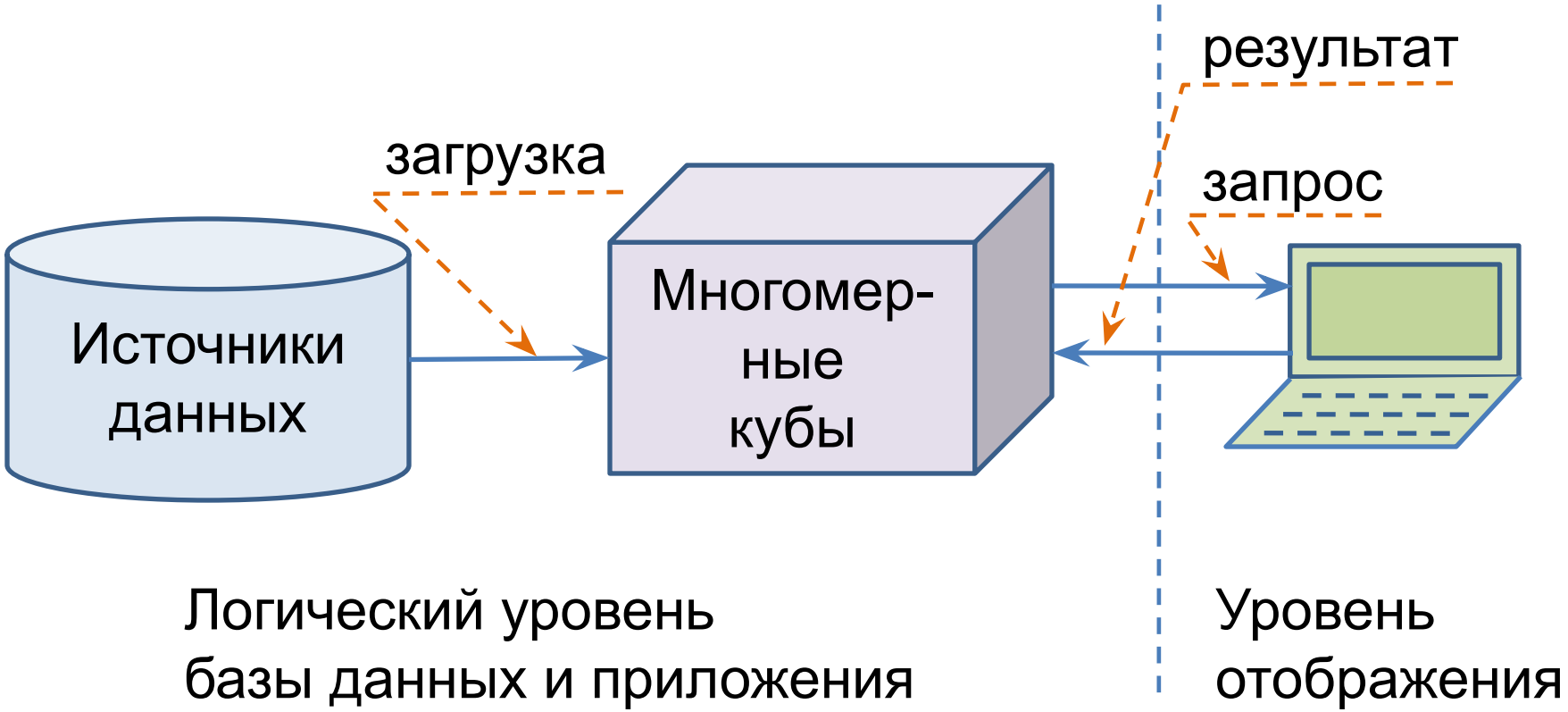
- Многомерные OLAP инструменты – Multidimensional OLAP, MOLAP
- Реляционные OLAP инструменты – Relational OLAP, ROLAP
- Управляемая среда запросов – Managed Query Environment, MQE

Многомерный OLAP

Специализированные структуры данных и многомерные СУБД

- Данные обобщаются и хранятся в соответствии с их предполагаемым использованием
- Высокая производительность
- Тесное взаимодействие с уровнем приложения и уровнем отображения

Многомерный OLAP



Многомерный OLAP

Особенности:

- Используемые структуры данных обладают ограниченной способностью поддержки нескольких предметных областей и осуществления доступа к подробным сведениям

Многомерный OLAP

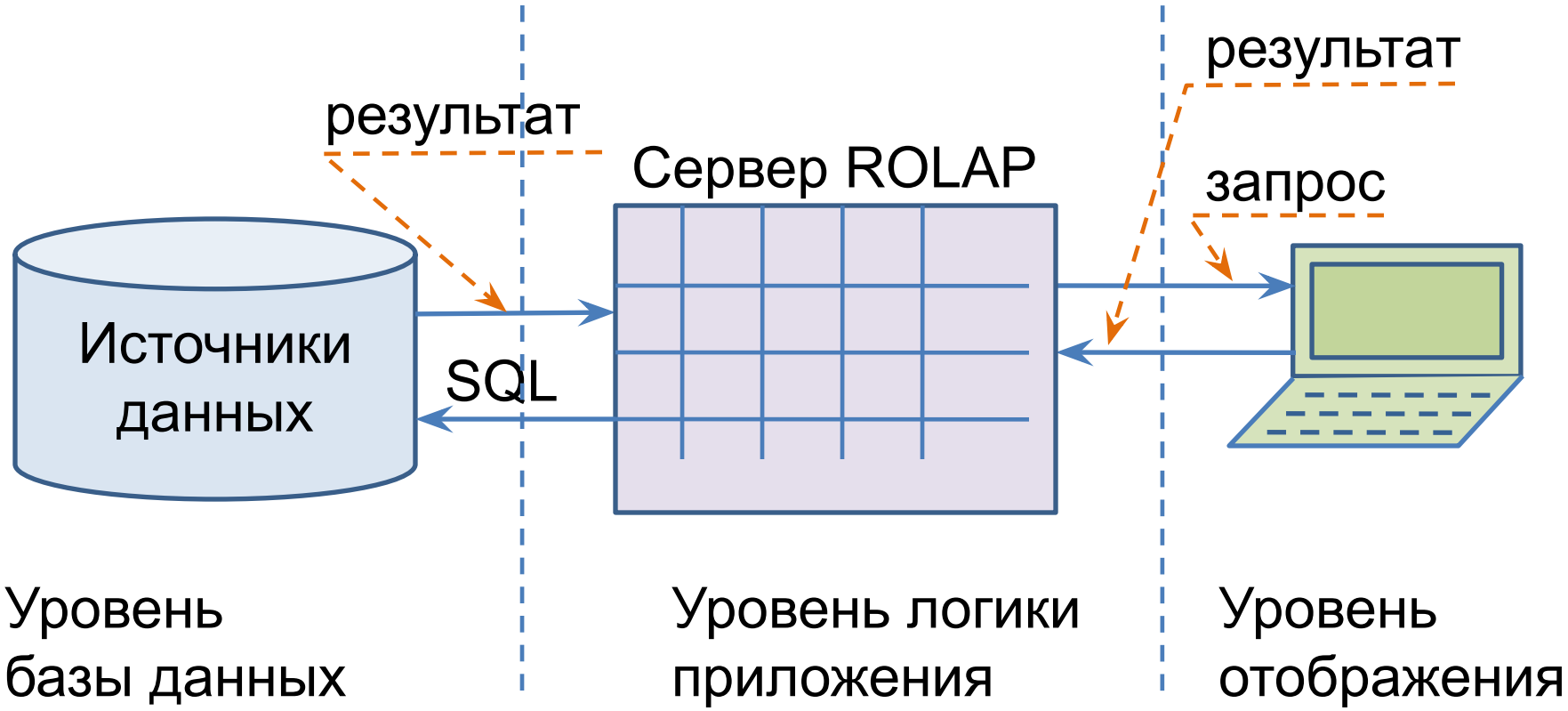
- Просмотр и анализ данных ограничен процессом проектирования структуры данных в соответствии с заранее определенными требованиями
- Необходимы особый набор навыков и знаний, использование специальных инструментов создания и сопровождения базы данных

Реляционный OLAP

Взаимодействие с СУБД – уровень метаданных

- Нет необходимости создания статичной многомерной структуры данных
- Дополнительные средства поддержки функций многомерного анализа
- Создание сильно денормализованной базы данных

Реляционный OLAP



Реляционный OLAP

Особенности:

- Необходима разработка промежуточного ПО для многомерных приложений (преобразование отношений РБД в многомерную структуру)

Реляционный OLAP

- Требуется разработка инструментов, предназначенных для создания устойчивых многомерных структур со вспомогательными компонентами администрирования этих структур

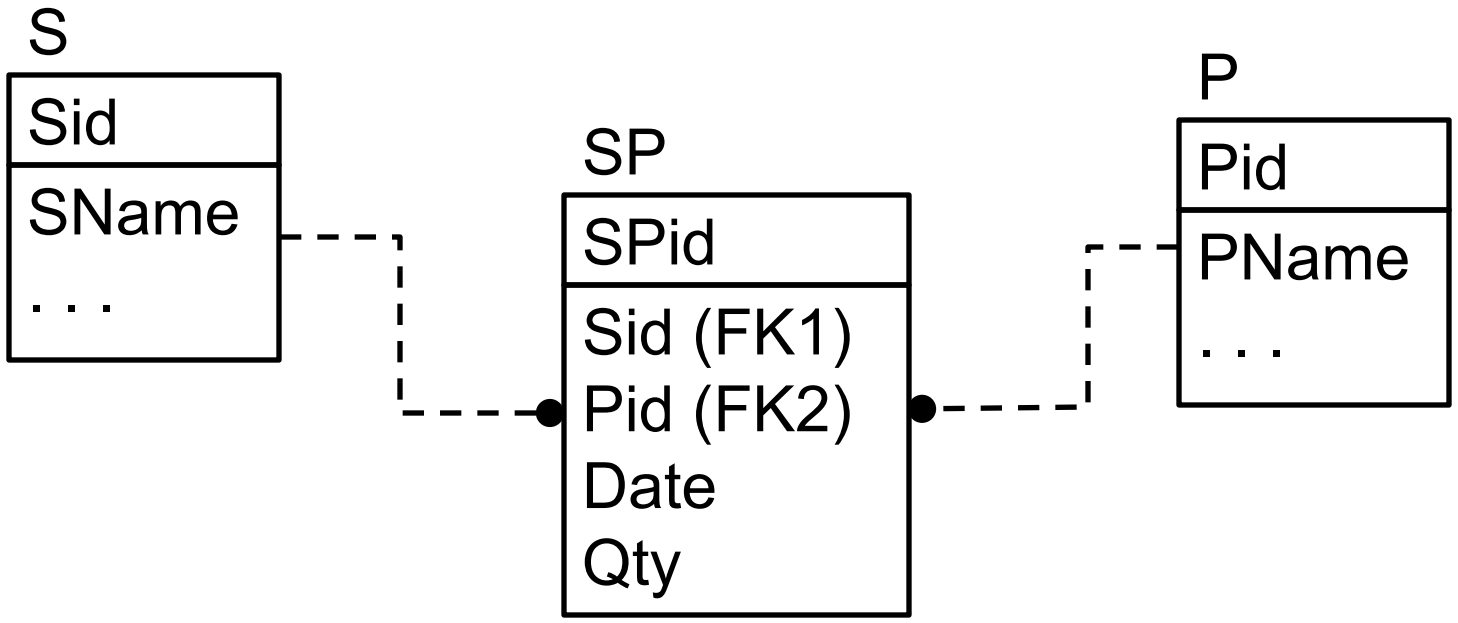
Дополнительные возможности SQL

Предложение SELECT:

SELECT . . . FROM . . .
GROUP BY . . .
WITH ROLLUP | WITH CUBE

Дополнительные возможности SQL

Пример:



SELECT . . . WITH CUBE | WITH ROLLUP

Дополнительные возможности SQL

Пример:

```
SELECT   SName, PName, sum(qty) as
   sum
FROM   S join SP on S.Sid = SP.Sid
join P on SP.Pid = P.Pid
GROUP BY   SName, PName
```

Дополнительные возможности SQL

SName	PName	sum
АО ИМИ	болт	200
АО МММ	болт	400
АО ИМИ	винт	100
АО ИПИ	винт	200
АО ИВТ	гайка	400
АО ИМИ	гайка	100
АО МММ	гайка	400
АО ИМИ	шайба	300

Дополнительные возможности SQL

Пример:

```
SELECT   SName, PName, sum(qty) as
   sum
FROM   S join SP on S.Sid = SP.Sid
join P on SP.Pid = P.Pid
GROUP BY   SName, Pname
WITH   ROLLUP
```

Дополнительные возможности SQL

SName	PName	sum
АО ИВТ	гайка	400
АО ИВТ	NULL	400
АО ИМИ	болт	200
АО ИМИ	винт	100
АО ИМИ	гайка	100
АО ИМИ	шайба	300
АО ИМИ	NULL	700
. . .	. . .	. . .
NULL	NULL	2100

Дополнительные возможности SQL

	болт	винт	гайка	шайба	итог
АО ИВТ			400		400
АО ИМИ	200	100	100	300	700
АО ИПИ		200			200
АО МММ	400		400		800
					21000

Дополнительные возможности SQL

Пример:

```
SELECT   SName, PName, sum(qty) as
   sum
FROM   S join SP on S.Sid = SP.Sid
join P on SP.Pid = P.Pid
GROUP BY   SName, Pname
WITH    CUBE
```

Дополнительные возможности SQL

SName	PName	sum
АО ИВТ	гайка	400
АО ИВТ	NULL	400
АО ИМИ	болт	200
АО ИМИ	. . .	. . .
АО ИМИ	NULL	700
. . .	. . .	. . .
NULL	болт	600
. . .	. . .	. . .
NULL	NULL	2100

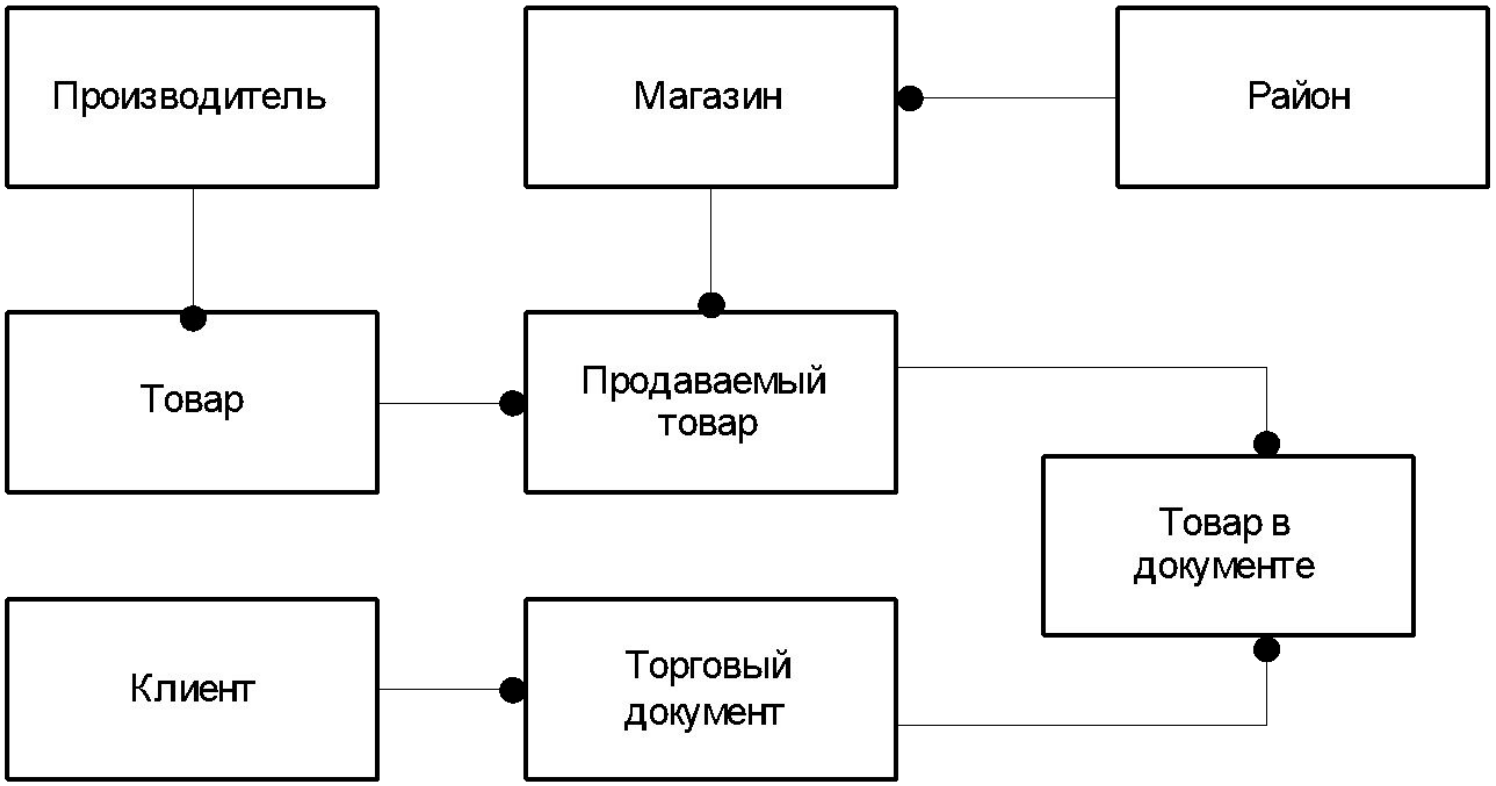
Дополнительные возможности SQL

	болт	винт	гайка	шайба	итог
АО ИВТ			400		400
АО ИМИ	200	100	100	300	700
АО ИПИ		200			200
АО МММ	400		400		800
	600	300	900	300	21000

Платформа EMC Documentum

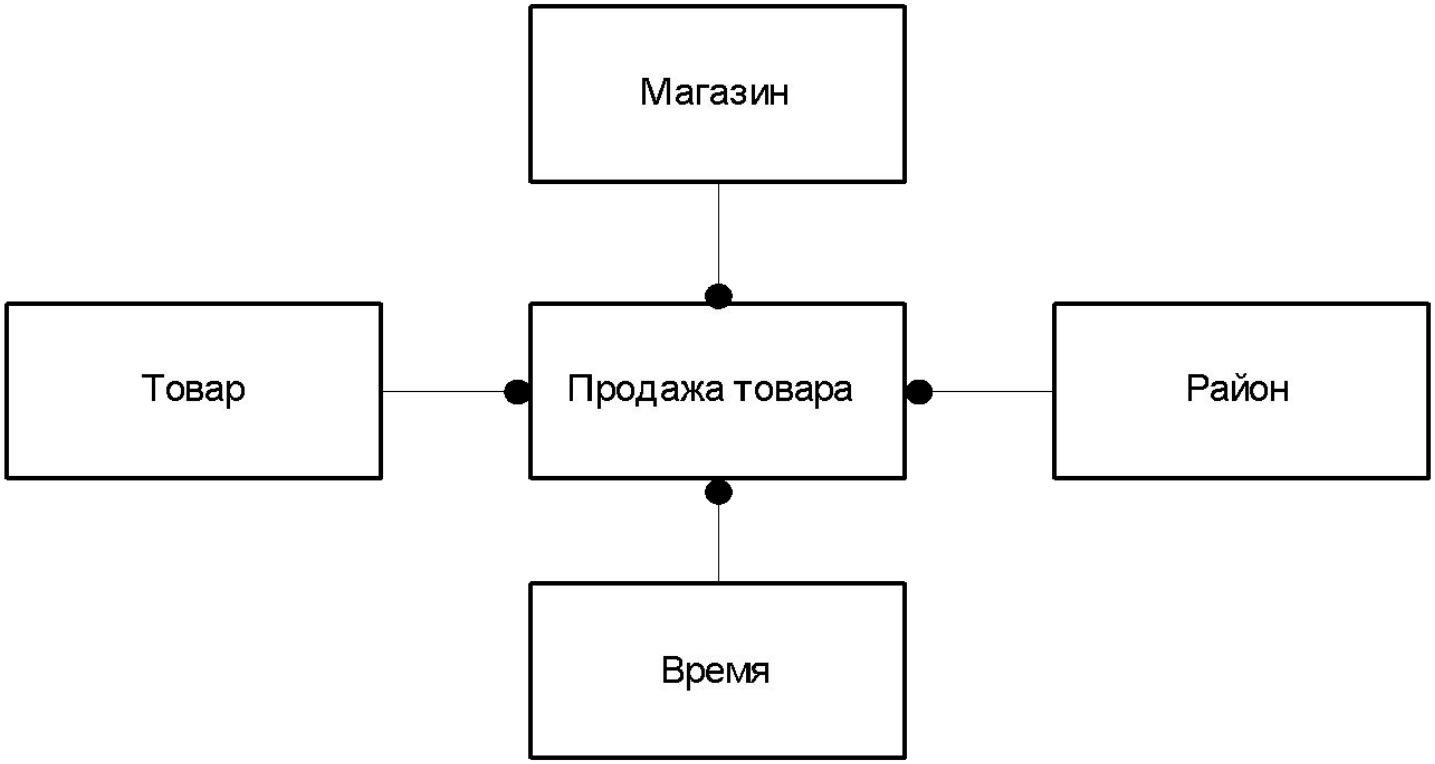
Области применения ИС

- Управление повседневными бизнес процессами (OLTP)



Области применения ИС

- Поддержка принятия стратегических решений (OLAP, Data mining)



Области применения ИС

- Enterprise Content Management (ECM) – стратегии, методы и инструментальные средства, используемые для ввода/сбора, управления, хранения, архивирования и доставки информационного содержания (контента) и документов, относящихся к ключевым процессам организации

Информационное содержание

Информационное содержание (контент) – информационные объекты, хранящиеся в различных форматах, которые можно извлекать, повторно использовать публиковать

(Коммерческие документы, сообщения электронной почты, образы документов, мультимедийные файлы, …)

Управление контентом

- Создание и сохранение документов
- Обработка документов – поиск, управление версиями, . . .
- Получение доступа к содержимому – управление доступом, аудит, . . .
- Управление бизнес процессами – автоматизация, жизненный цикл контента, . . .

Управление контентом

Системы управления контентом (CMS, Content Management System) – управление неструктурированными данными

Элемент контента

Метаданные

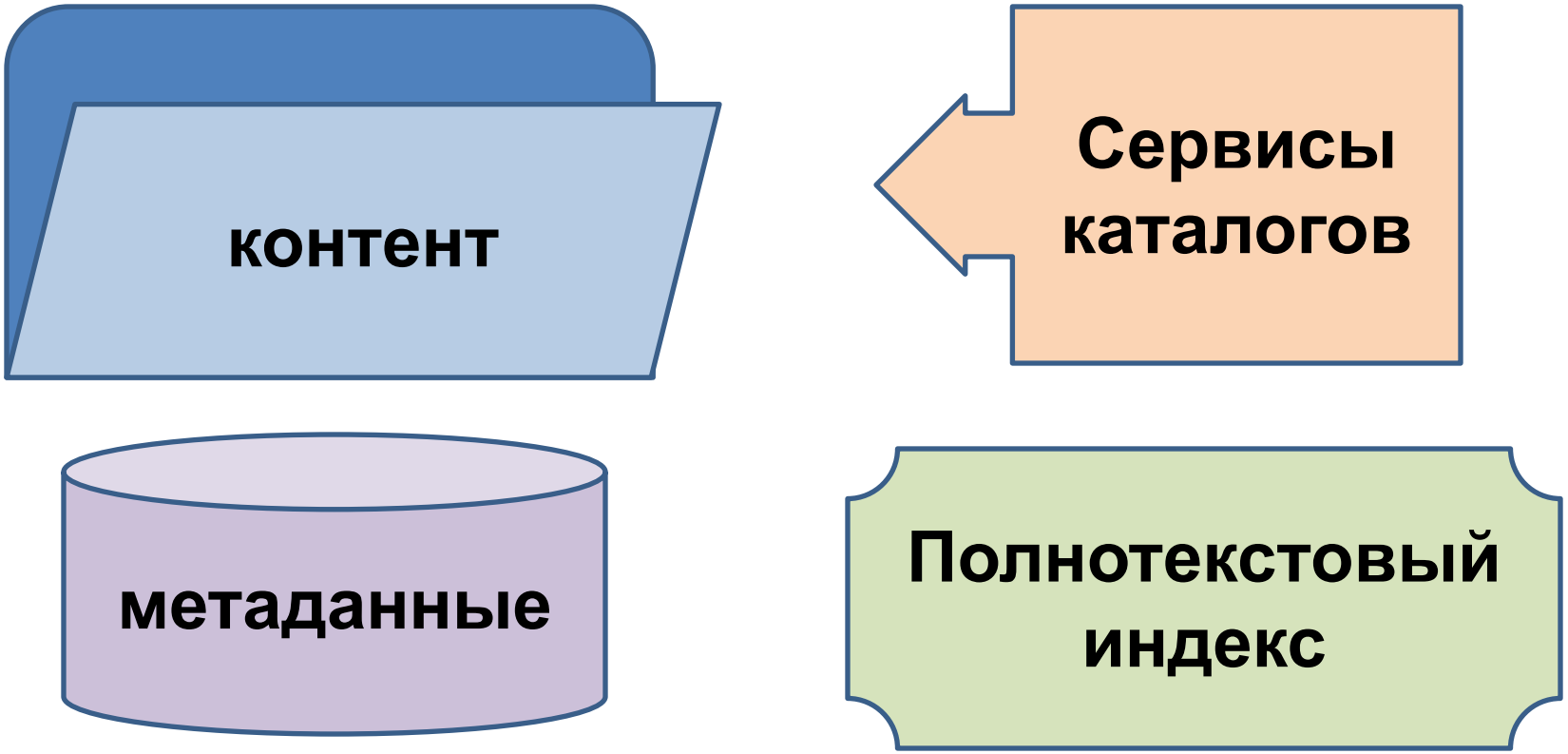
Управление контентом

Репозиторий – управляемый блок хранения контента и метаданных

Инфраструктура репозитория

- Компоненты репозитория
- Сервисы репозитория
- Сервисы безопасности

Компоненты репозитория



Сервисы репозитория

- Объектная модель данным
- Управление связями объектов
- Словарь данных
- Сервисы хранения
- Поиск / запросы
- Жизненный цикл
- Распределенные / федеративные сервисы

Сервисы безопасности

- Управление доступом
- Управление правами
- Разрешения
- Аудит
- Шифрование

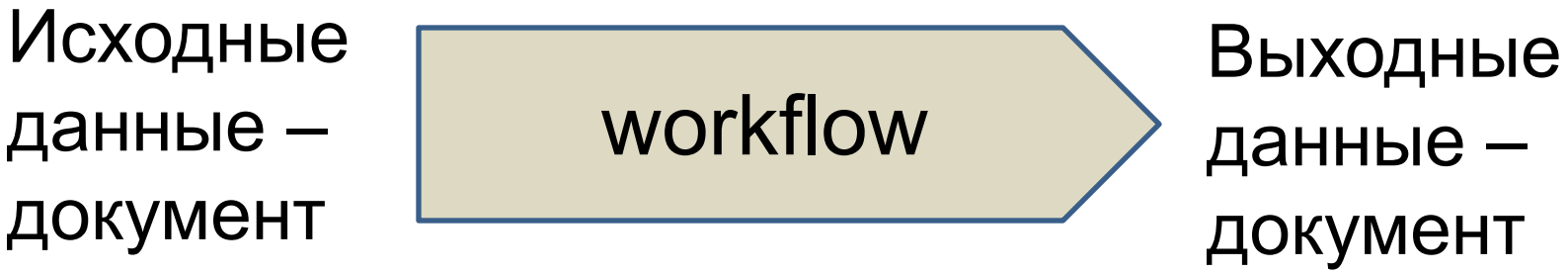
Управление процессами

Workflow – представляет бизнес процессы и приложения, ориентированные на события. Может быть определен для документов, папок и виртуальных документов

Lifecycle – последовательность состояний, в которых в которых может находиться отдельный документ

Workflow

Бизнес процесс – набор связанных действий, которые создают некоторый результат, преобразуя исходные данные в более значимые выходные данные

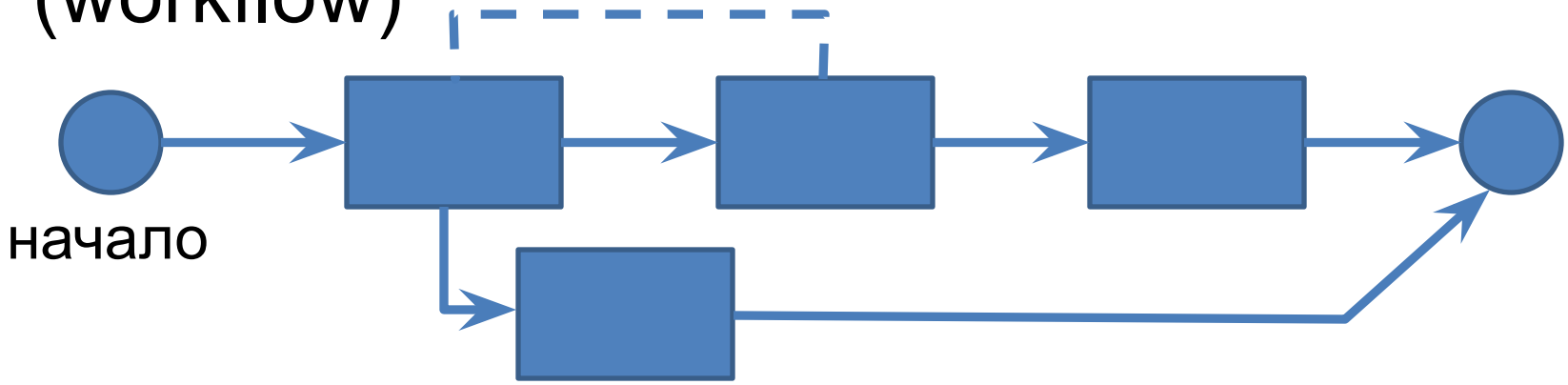


Workflow

Описание процесса

- Задача (activity)
- Исполнитель (performer)
- Поток информации (flow)

Конкретное выполнение работ – процесс (workflow)



Lifecycle

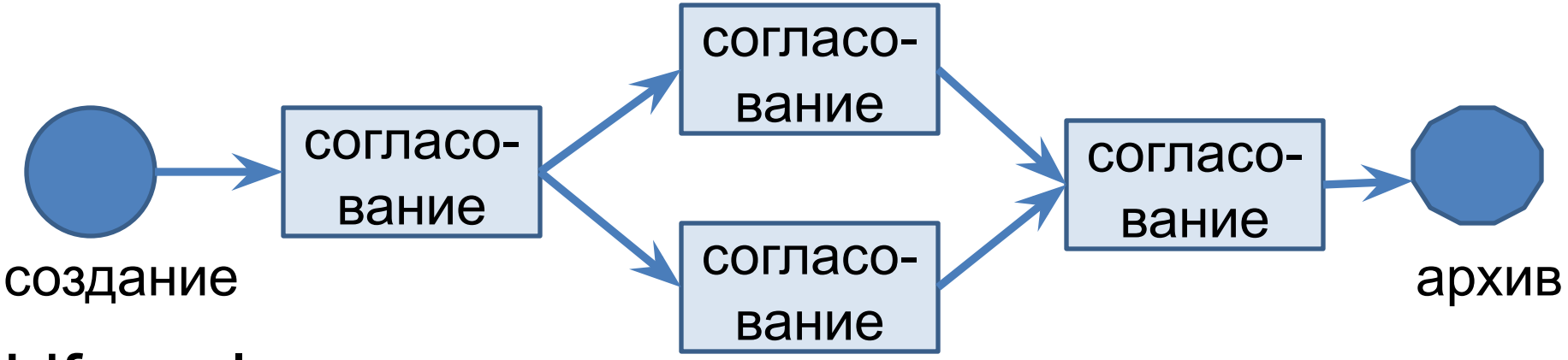
Строго последовательное переключение состояний

Состояния жизненного цикла

- Стартовое – создание документа, ввод содержимого
- Промежуточные состояния – различные стадии документа
- Конечное состояние – передача документа в архив

Пример

Workflow



Lifecycle