4. Хранилища данных

4.1. 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

4.2. 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

4.3. Сравнение систем

1. Характер данных

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

4.3. Сравнение систем
(продолжение)

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

4.3. Сравнение систем
(продолжение)

3. Назначение системы

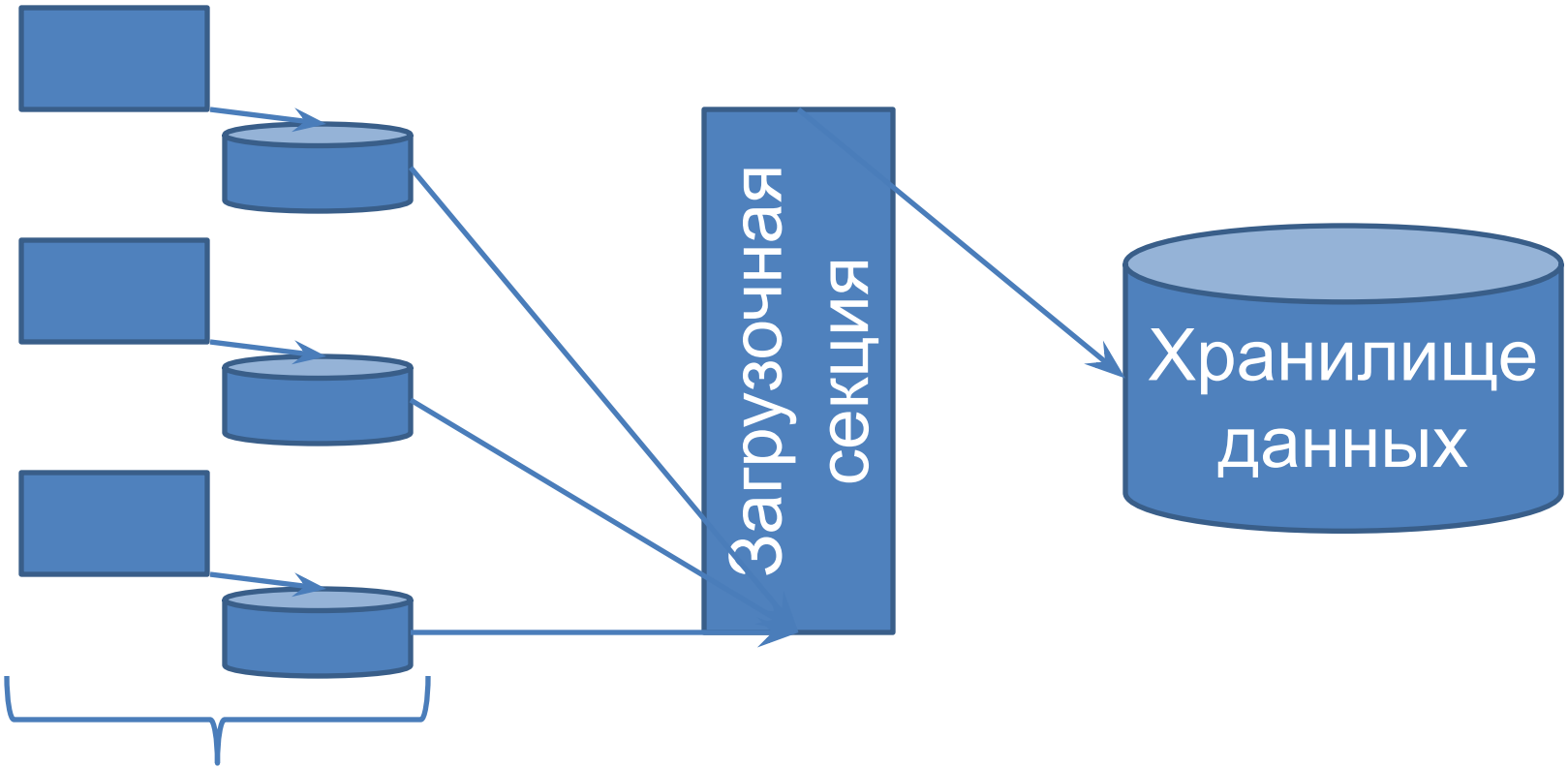
OLTP + базы данных	DSS + хранилища данных
Обработка транзакций	Проведение анализа
Ориентирована на прикладную область	Ориентирована на предметную область
Поддержка принятия повседневных решений	Поддержка принятия стратегических решений

4.3. Сравнение систем
(продолжение)

4. Пользователи

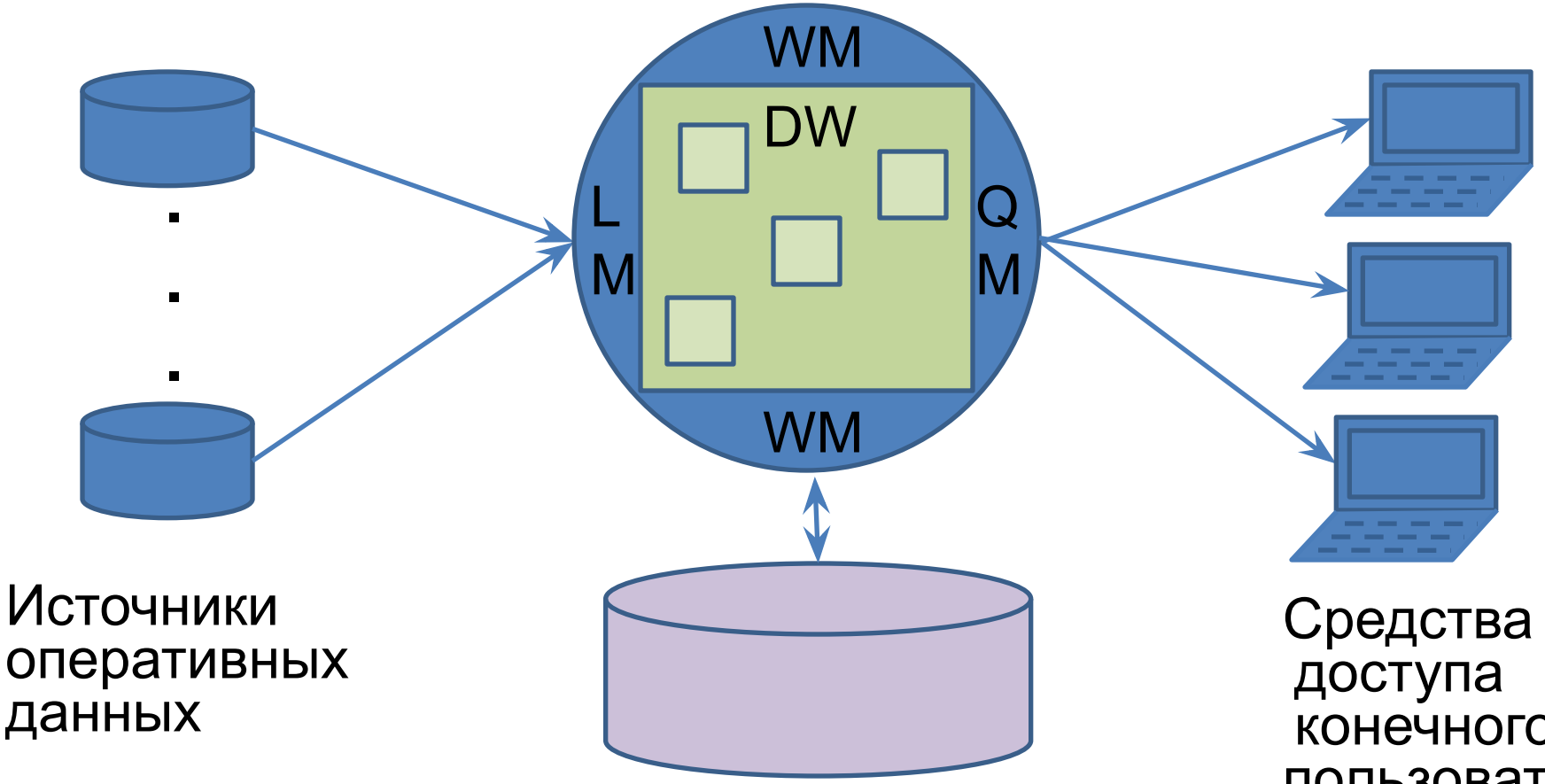
OLTP + баз данных	DSS + хранилища данных
Обслуживает большое количество пользователей исполнительного звена	Обслуживает относительно небольшое количество работников руководящего звена

4.4. Конфигурация хранилища данных

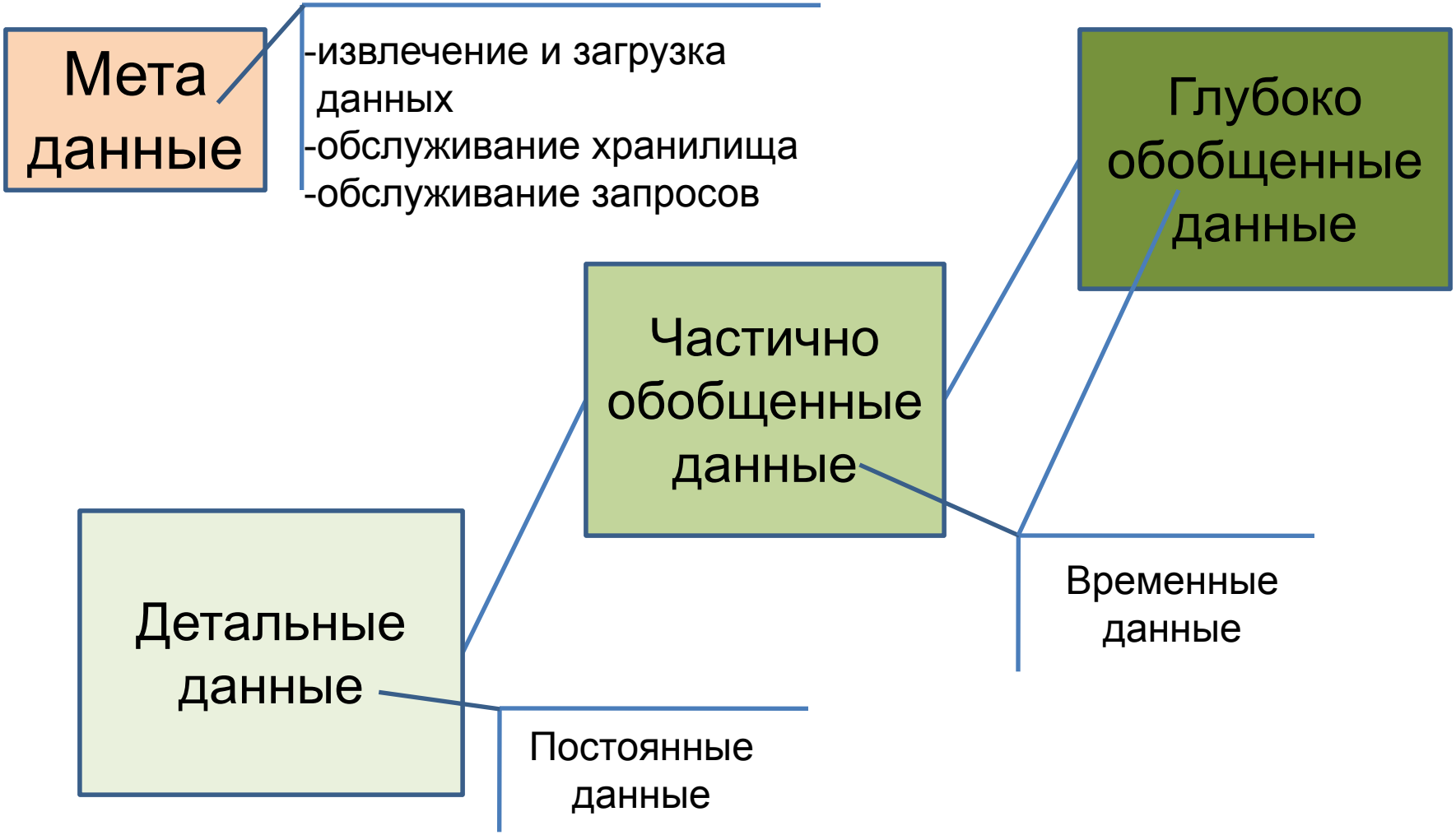


OLTP-системы
источники данных

4.5. Архитектура хранилища данных



4.5. Архитектура хранилища данных
(продолжение)



4.5. Архитектура хранилища данных
(продолжение)

Менеджер загрузки – Load Manager (LM):

внешний (front-end) компонент;

извлечение данных,

загрузка данных в хранилище

– инструменты репликации информации
– генераторы кода
– механизмы динамического преобразования

4.5. Архитектура хранилища данных (продолжение)

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

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

4.5. Архитектура хранилища данных
(продолжение)

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

4.6. 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

4.6. Средства доступа к данным (продолжение)

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

4.6. Средства доступа к данным (продолжение)

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

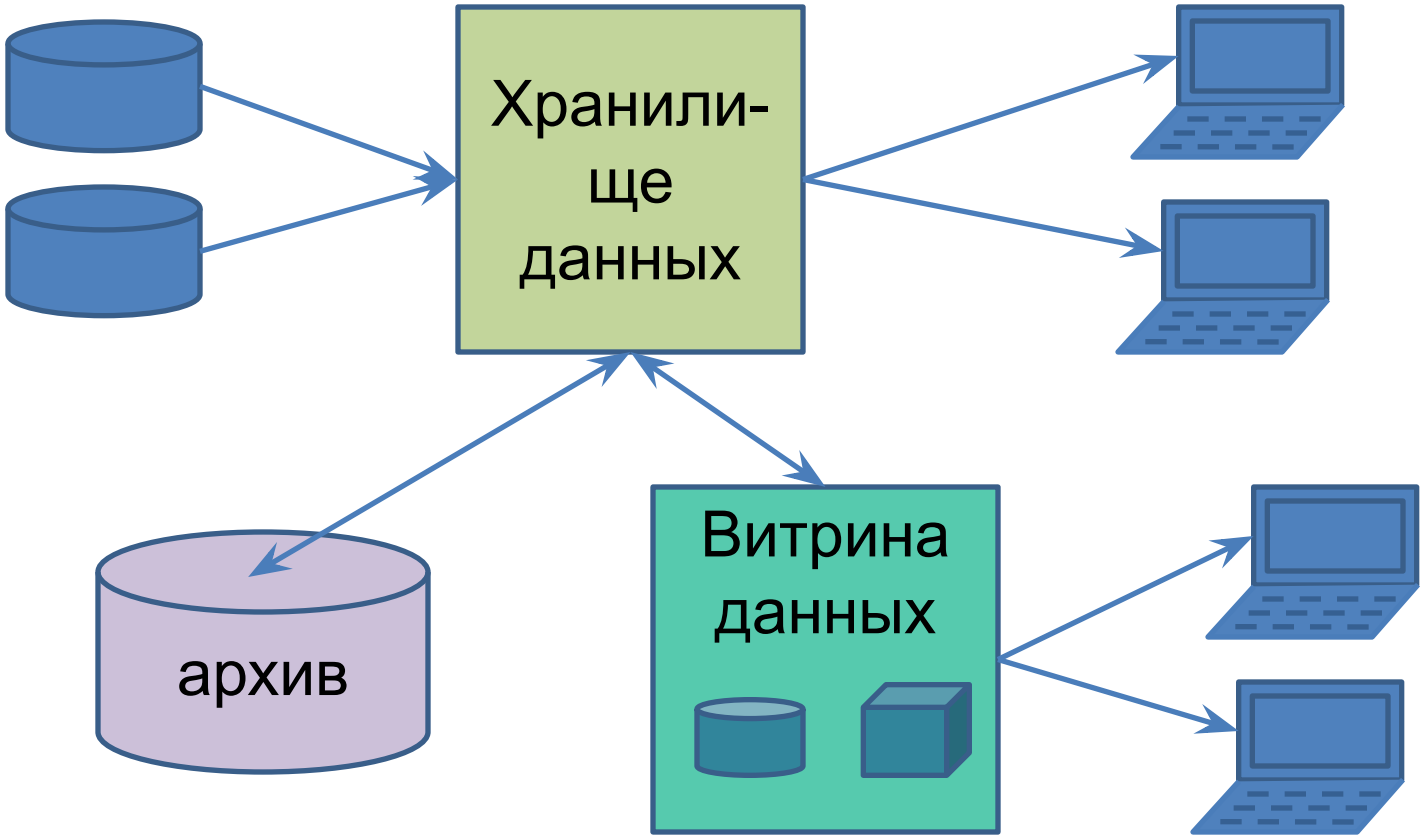
Создание предсказательных моделей
   Intelligent Miner фирмы IBM

4.7. Витрины данных

Data Mart – витрины (магазины) данных

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

4.9. Витрины данных
(продолжение)



4.7. Витрины данных
(продолжение)

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

4.7. Витрины данных
(продолжение)

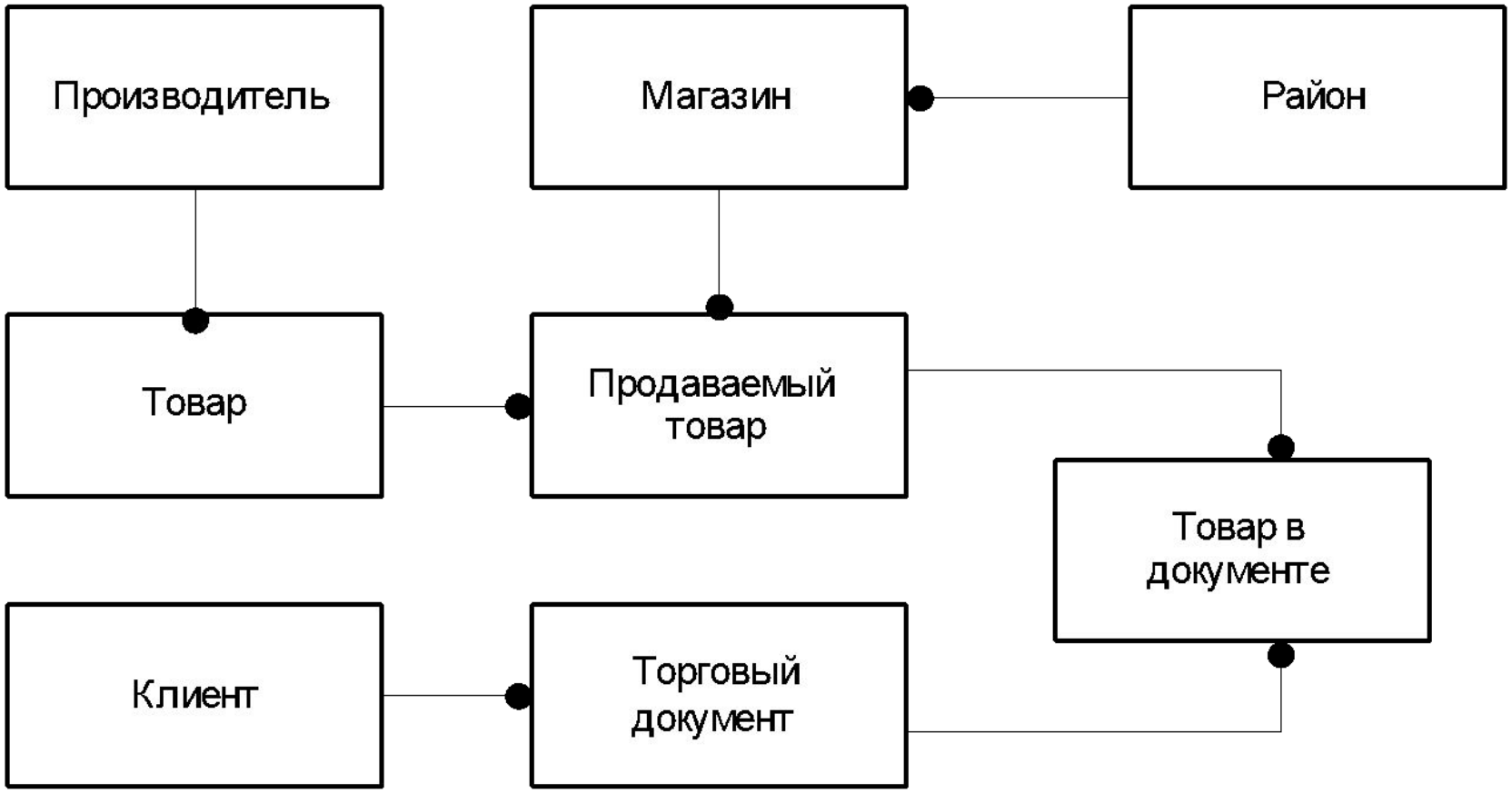
Создание:

- хранилище данных витрины данных
- витрины данных хранилище данных
- хранилище данных + витрины данных

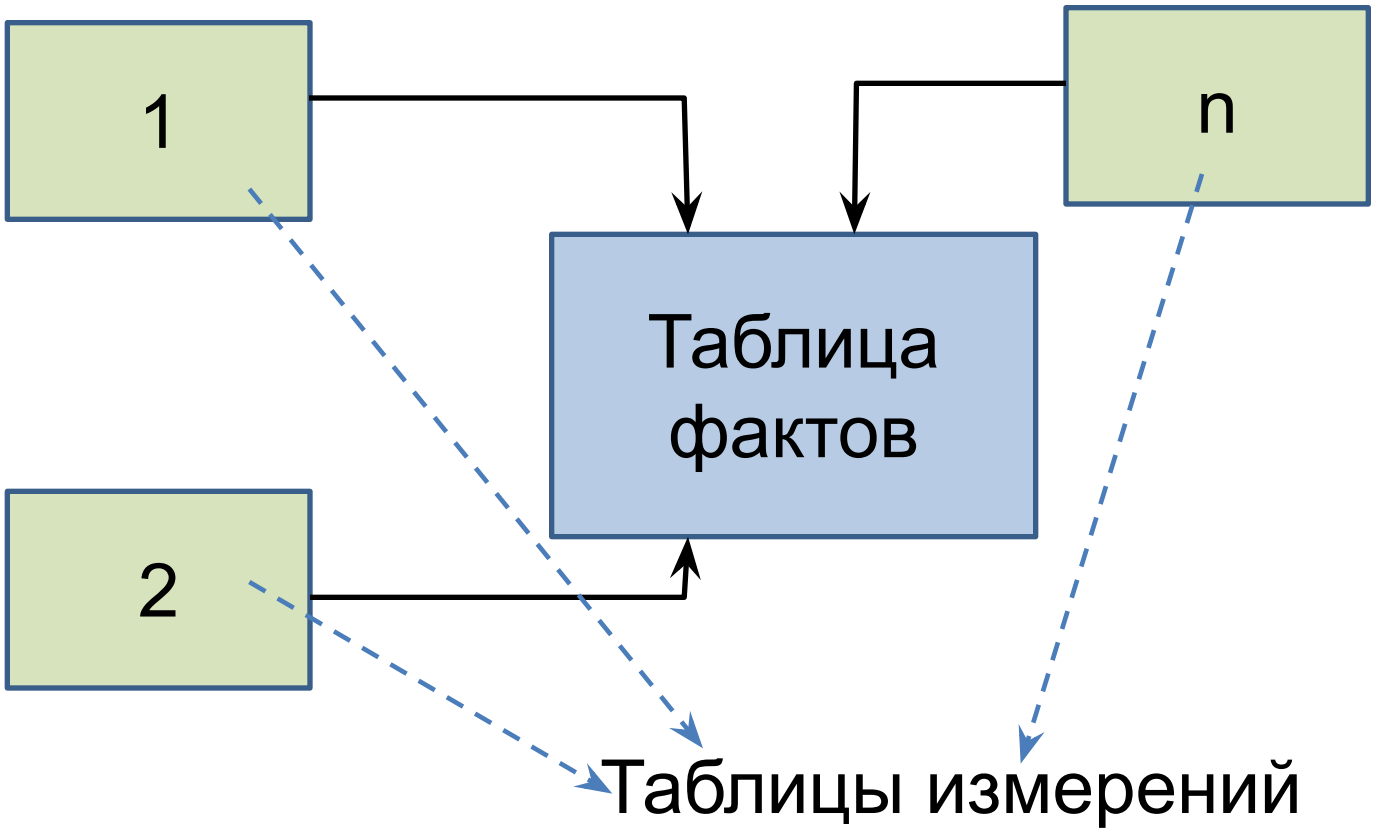
4.8. Проектирование хранилища данных

	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»

4.8. Проектирование хранилища данных (продолжение)

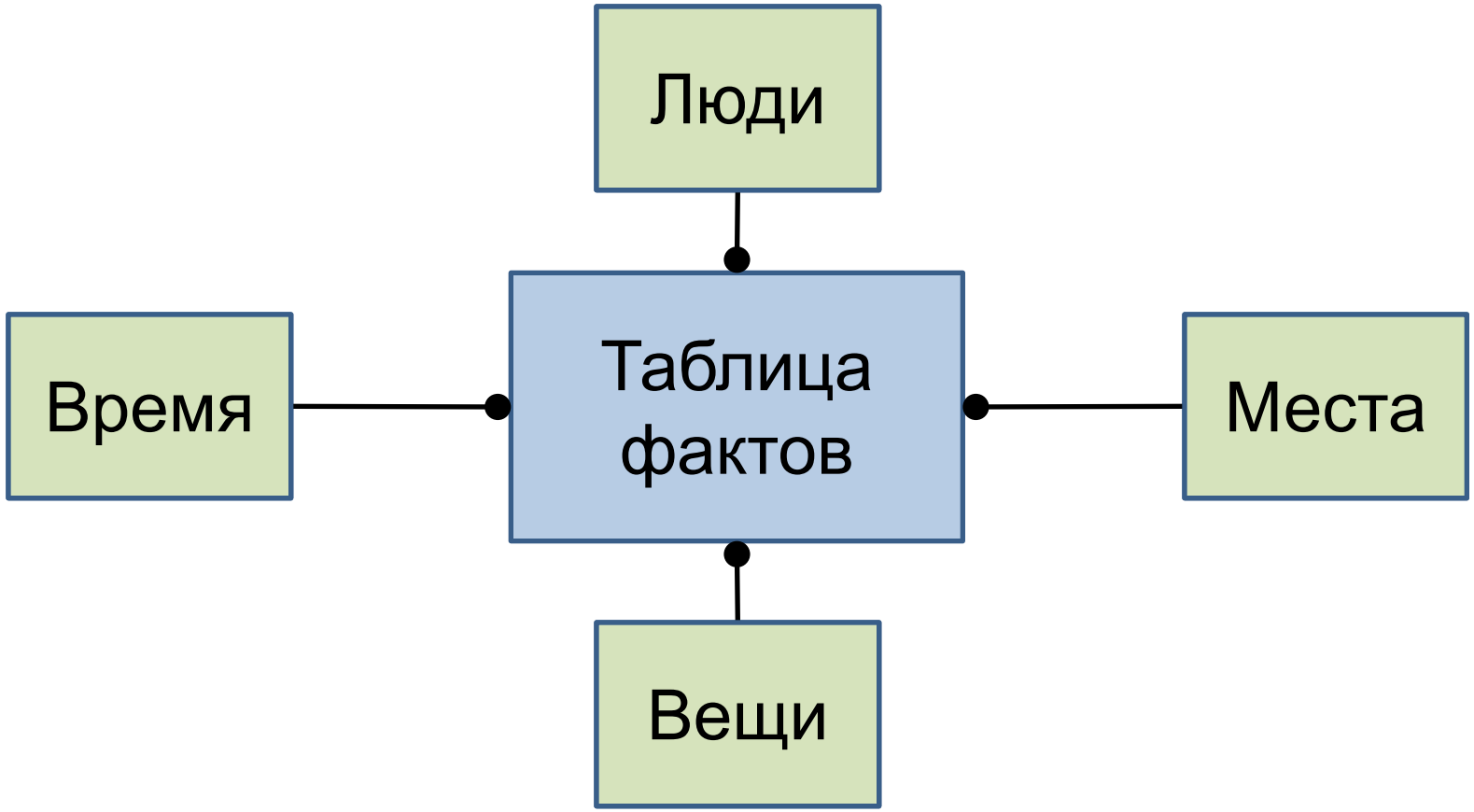


4.9. Схема типа «звезда»

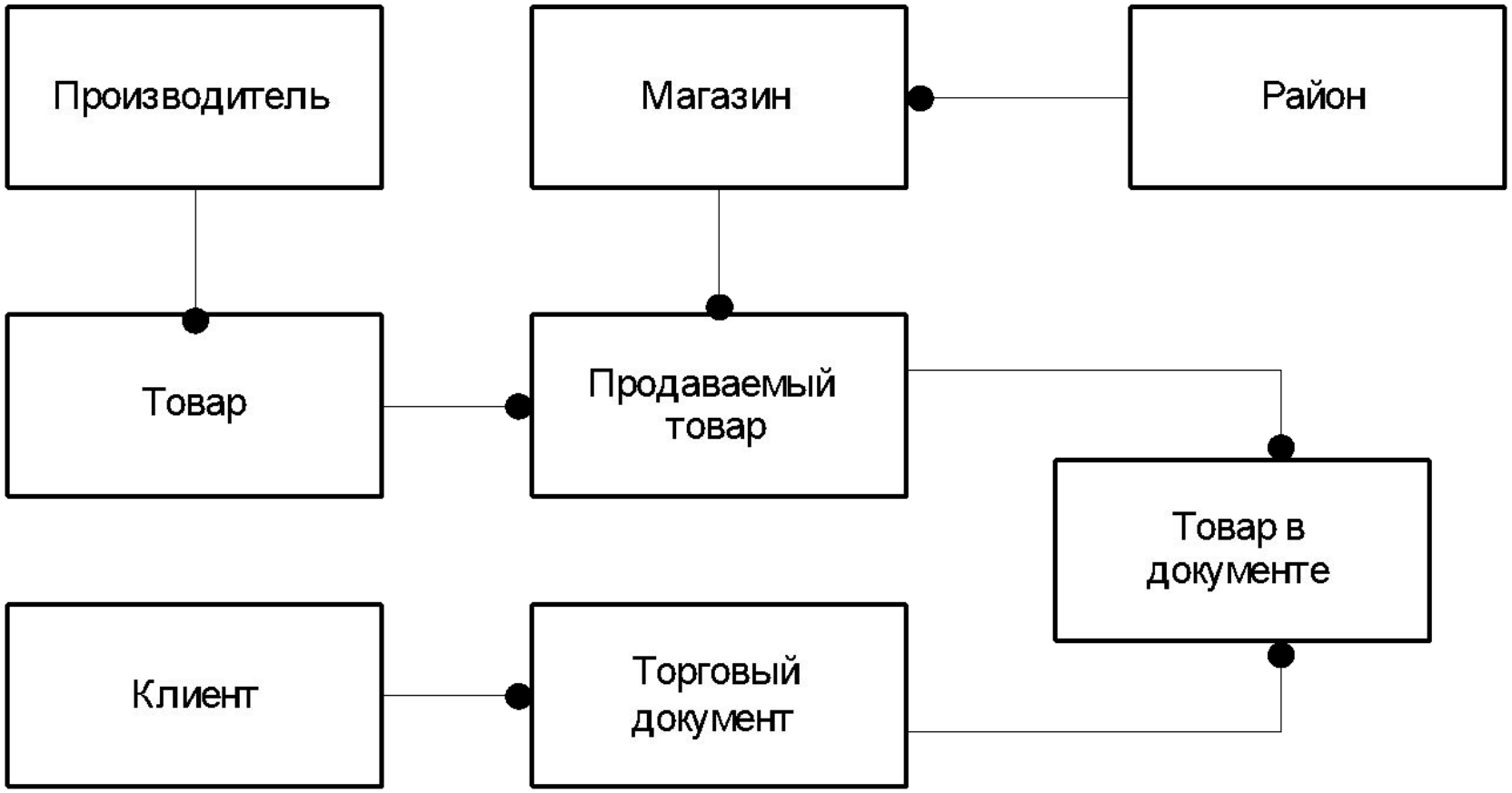


4.9. Схема типа «звезда»
(продолжение)

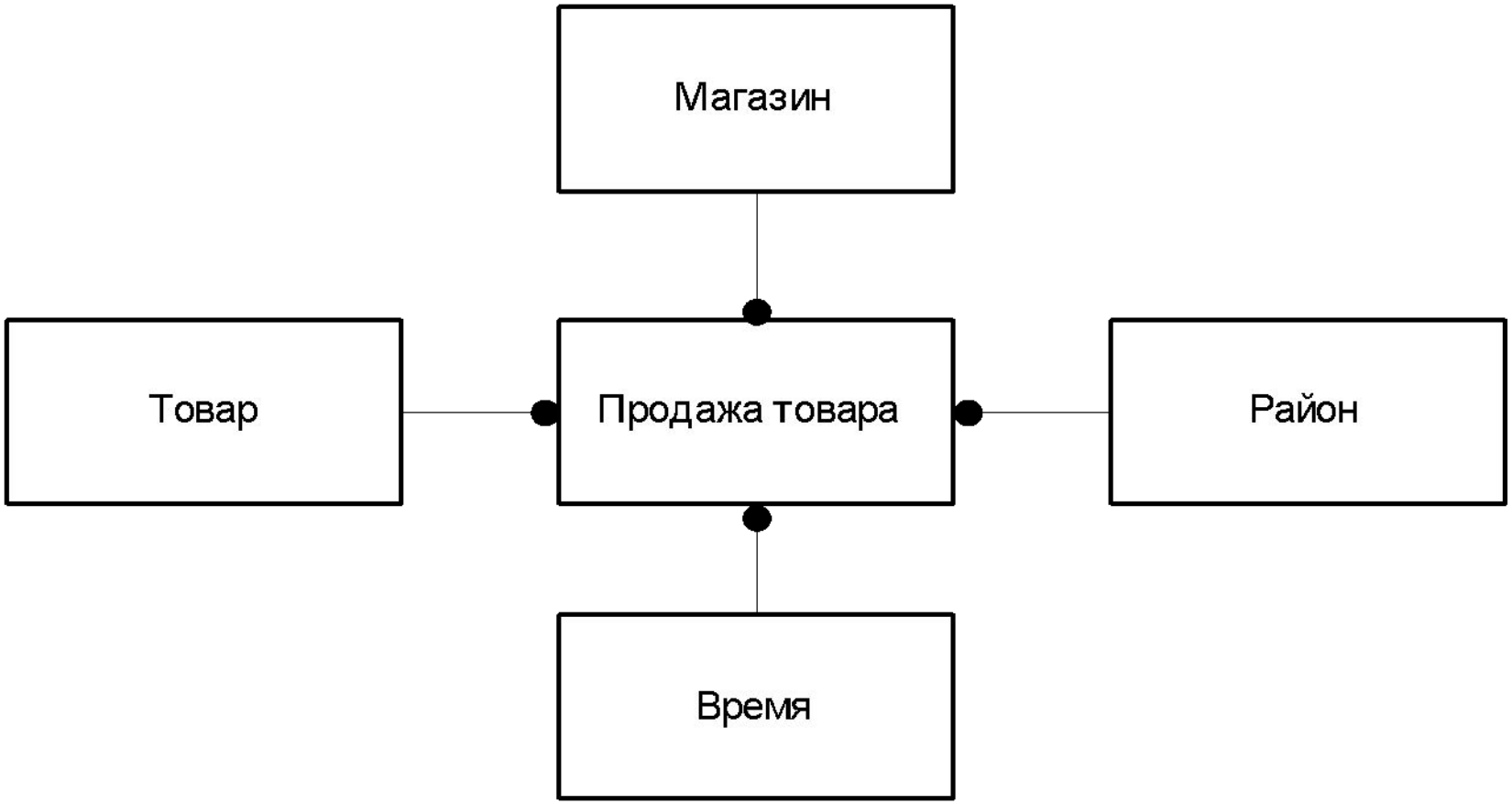
Категории измерений



4.10. Пример проектирования



4.10. Пример проектирования
(продолжение)



4.11. Особенности проектирования

Таблица фактов:

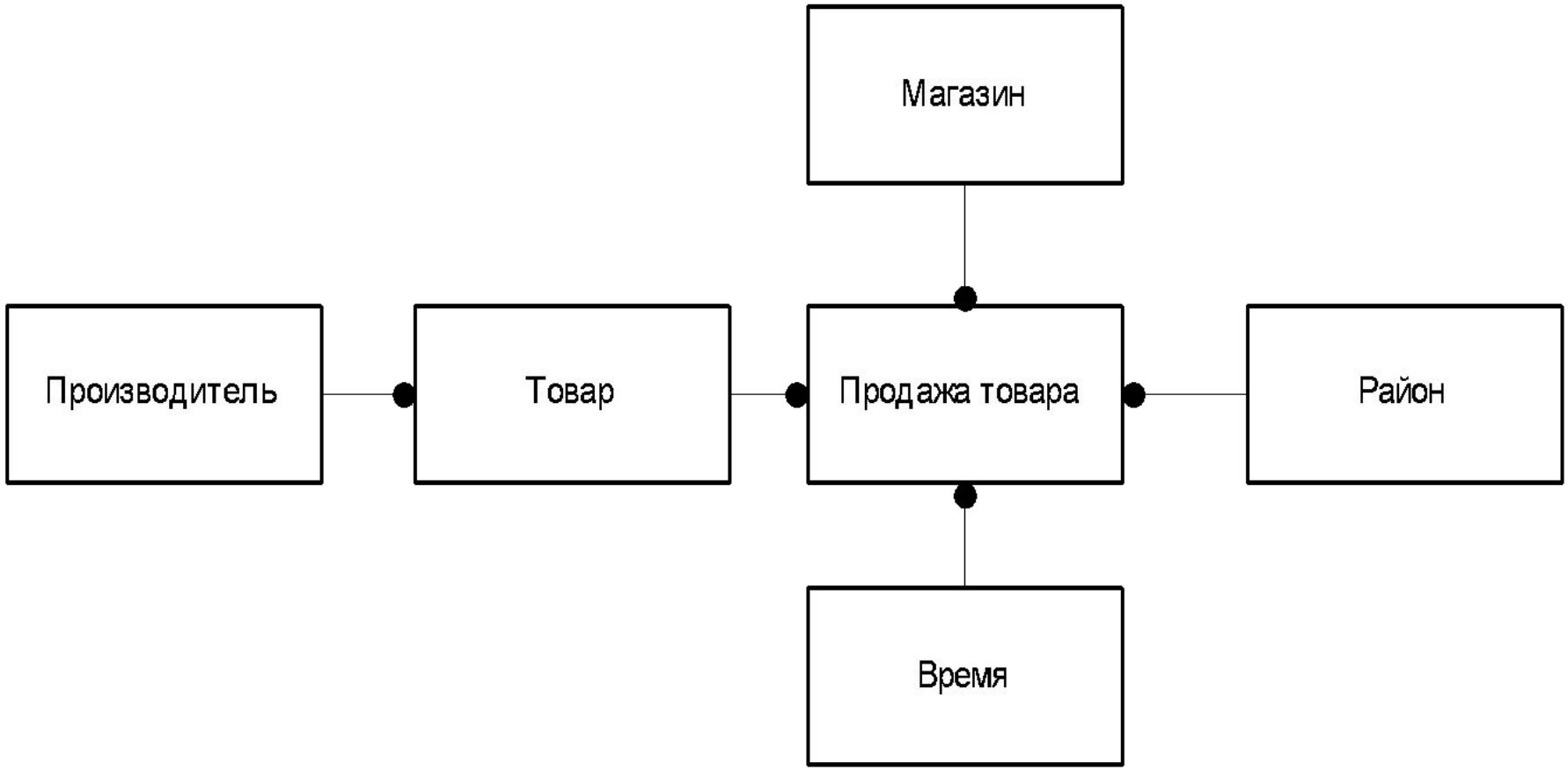
- использование суррогатного ключа
- вычисляемые колонки
  (объем продаж, стоимость в . . .)
- секционирование
  - вертикальное
    (восстановление – через join)
  - горизонтальное
    (восстановление – через union)

4.11. Особенности проектирования (продолжение)

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»

4.11. Особенности проектирования
(продолжение)



4.12. Технология OLAP

Термин OLAP был предложен Коддом в 1993 г.

и определяет архитектуру, которая поддерживает сложные аналитические приложения

Назначение OLAP (Online Analytical Processing) инструментов:

предоставить средства извлечения большого количества записей и вычисления на их основе некоторых итоговых значений

4.13. Правила для OLAP систем

E. Codd, 1993 г.

- Многомерное концептуальное представление данных
- Доступность
- Неизменная производительность подготовки отчетов

4.13. Правила для OLAP систем (продолжение)

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

4.13. Правила для OLAP систем (продолжение)

- Прозрачность
- Динамическое управление разреженностью матриц
- Архитектура клиент-сервер
- Многопользовательская поддержка
- Поддержка интуитивно понятного манипулирования данными

4.14. Критерий FASMI

Fast –

время отклика:

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

4.14. Критерий FASMI
(продолжение)

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

4.14. Критерий FASMI
(продолжение)

Shared –

широкие возможности разграничения доступа к данным и одновременной работы многих пользователей

4.14. Критерий FASMI
(продолжение)

Multidimensional –

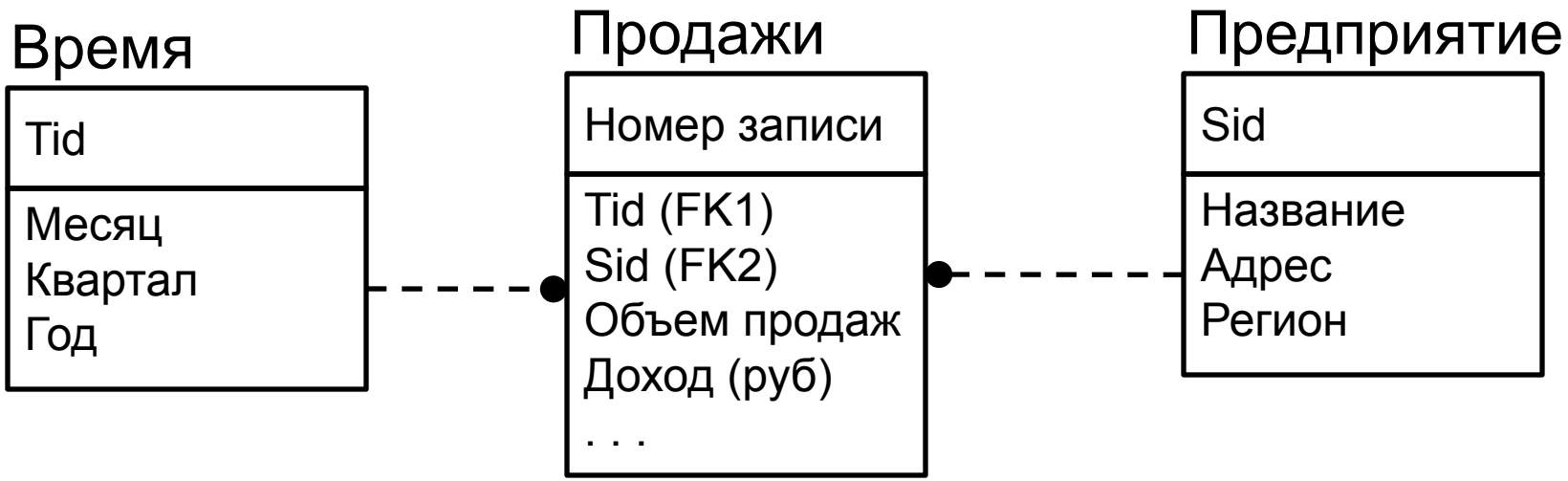
должно быть обеспечено многомерное концептуальное представление данных

Information –

необходимая информация должна быть получена там, где она необходима

4.15. Многомерное представление

Анализ изменения объема продаж и дохода торговых предприятий во времени



4.15. Многомерное представление
(продолжение)

Таблица РБД («плоская»)

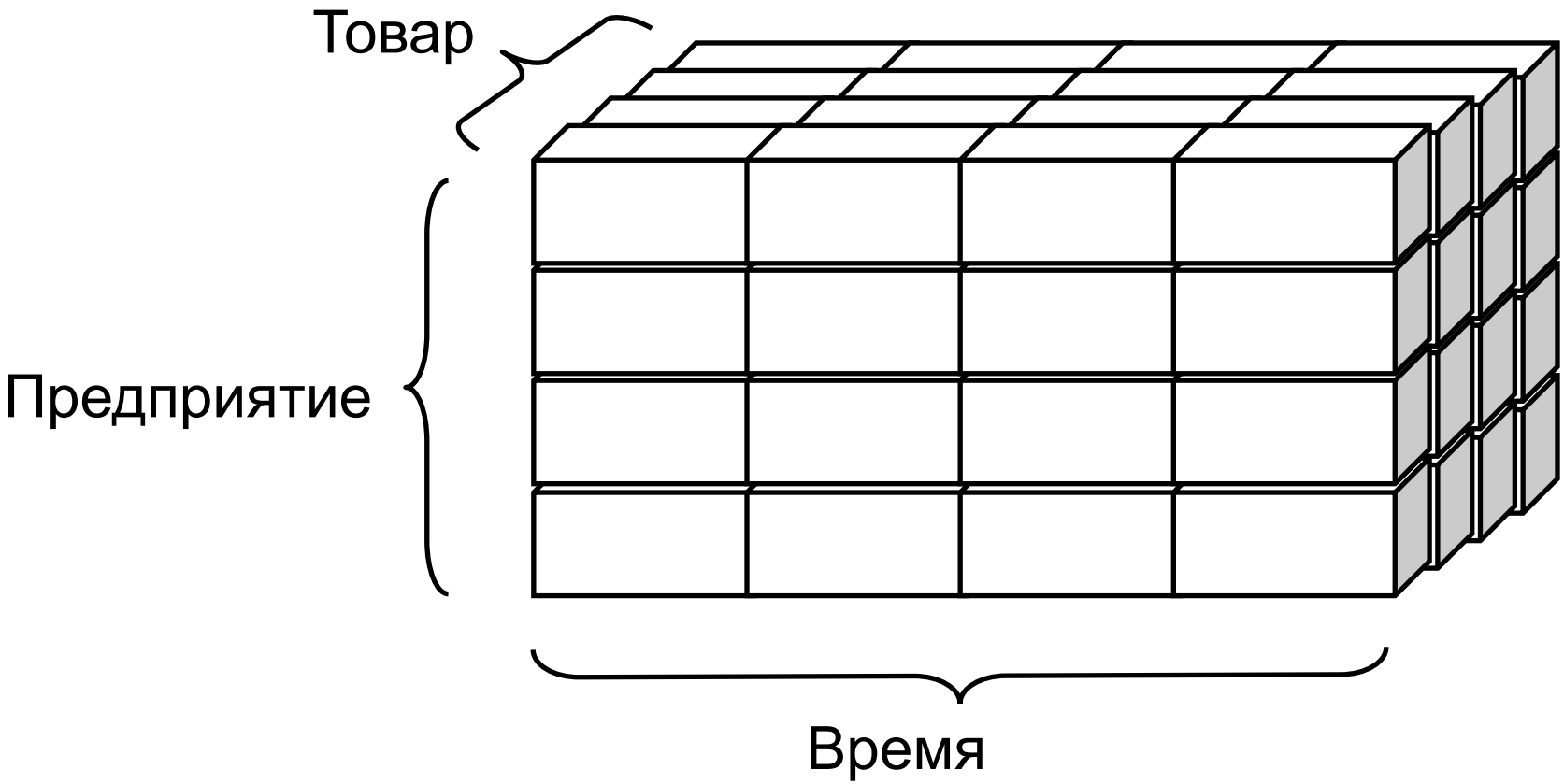
Tid	Sid	Объем продаж	Доход	. . .
1	1	k_{11}	s_{11}	. . .
1	2	k_{12}	s_{12}	. . .
1	3	k_{13}	s_{13}	. . .
1	4	k_{14}	s_{14}	. . .
. . .	. . .	. . .	. . .	. . .
2	1	k_{21}	s_{21}	. . .
. . .	. . .	. . .	. . .	. . .

4.15. Многомерное представление
(продолжение)

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	$k_{31}, s_{31}, \ldots$	$k_{32}, s_{32}, \ldots$	$k_{33}, s_{33}, \ldots$	. . .
. . .	. . .	. . .	. . .	. . .

4.15. Многомерное представление
(продолжение)



4.15. Многомерное представление (продолжение)

Достоинства многомерных структур:

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

4.15. Многомерное представление (продолжение)

Достоинства многомерных структур:

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

4.15. Многомерное представление
(продолжение)

«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.]

4.16. Аналитические операции

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

4.16. Аналитические операции (продолжение)

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

4.16. Аналитические операции (продолжение)

- Разбиение с поворотом (slicing and dicing) – также называется созданием сводной таблицы; позволяет получить представление данных с разных точек зрения

4.17. Категории OLAP инструментов

Berson and Smith, 1997 г.

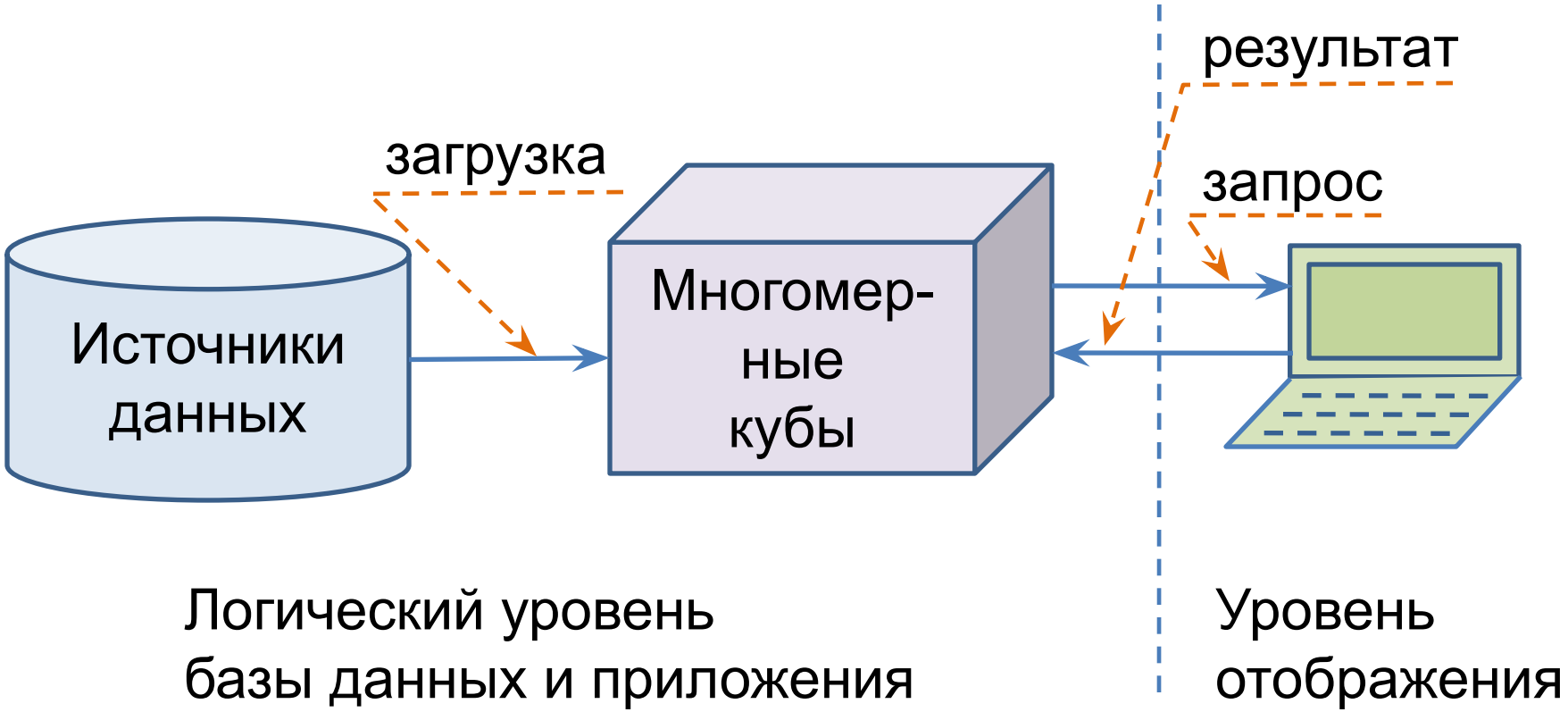
- Многомерные OLAP инструменты – Multidimensional OLAP, MOLAP
- Реляционные OLAP инструменты – Relational OLAP, ROLAP
- Управляемая среда запросов – Managed Query Environment, MQE

4.18. 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

4.18. Многомерный OLAP
(продолжение)



4.18. Многомерный OLAP
(продолжение)

Особенности:

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

4.18. Многомерный OLAP
(продолжение)

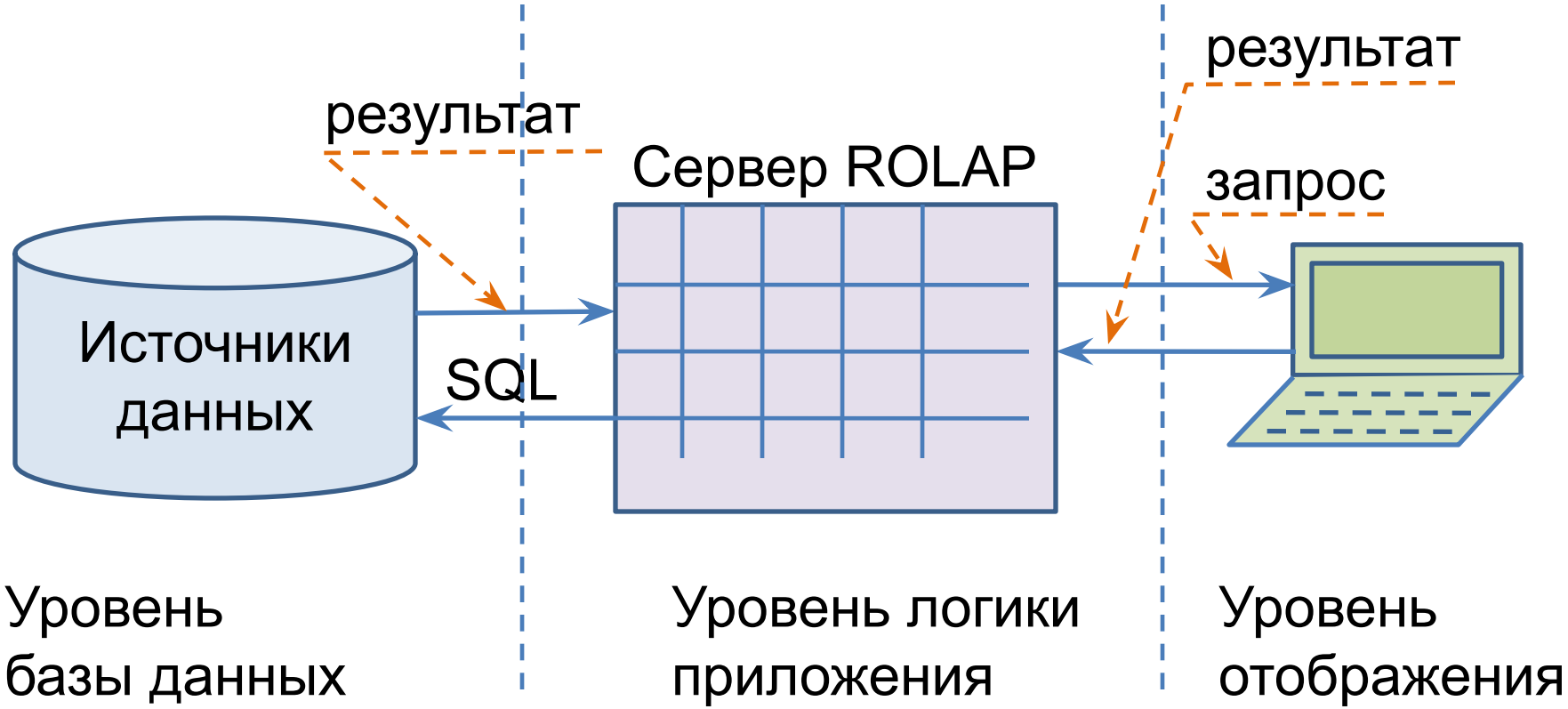
- Просмотр и анализ данных ограничен процессом проектирования структуры данных в соответствии с заранее определенными требованиями
- Необходимы особый набор навыков и знаний, использование специальных инструментов создания и сопровождения базы данных

4.19. 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

4.19. Реляционный OLAP
(продолжение)



4.19. Реляционный OLAP
(продолжение)

Особенности:

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

4.19. Реляционный OLAP
(продолжение)

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

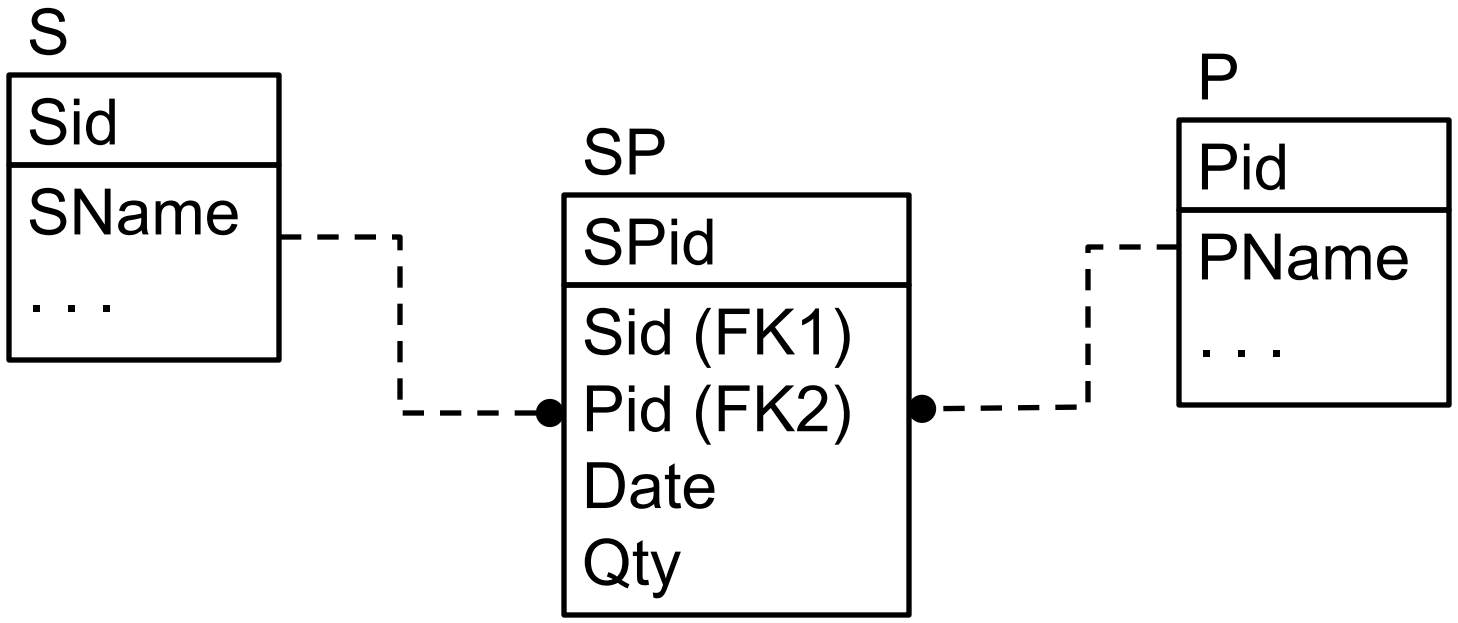
4.20. Дополнительные возможности SQL

Предложение SELECT:

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

4.20. Дополнительные возможности SQL (продолжение)

Пример:



SELECT . . . WITH CUBE | WITH ROLLUP

4.20. Дополнительные возможности SQL (продолжение)

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
```

4.20. Дополнительные возможности SQL (продолжение)

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

4.20. Дополнительные возможности SQL (продолжение)

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
```

4.20. Дополнительные возможности SQL (продолжение)

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

4.20. Дополнительные возможности SQL (продолжение)

	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

4.20. Дополнительные возможности SQL (продолжение)

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
```

4.20. Дополнительные возможности SQL (продолжение)

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

4.20. Дополнительные возможности SQL (продолжение)

	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

5. Платформа EMC Documentum

Области применения ИС

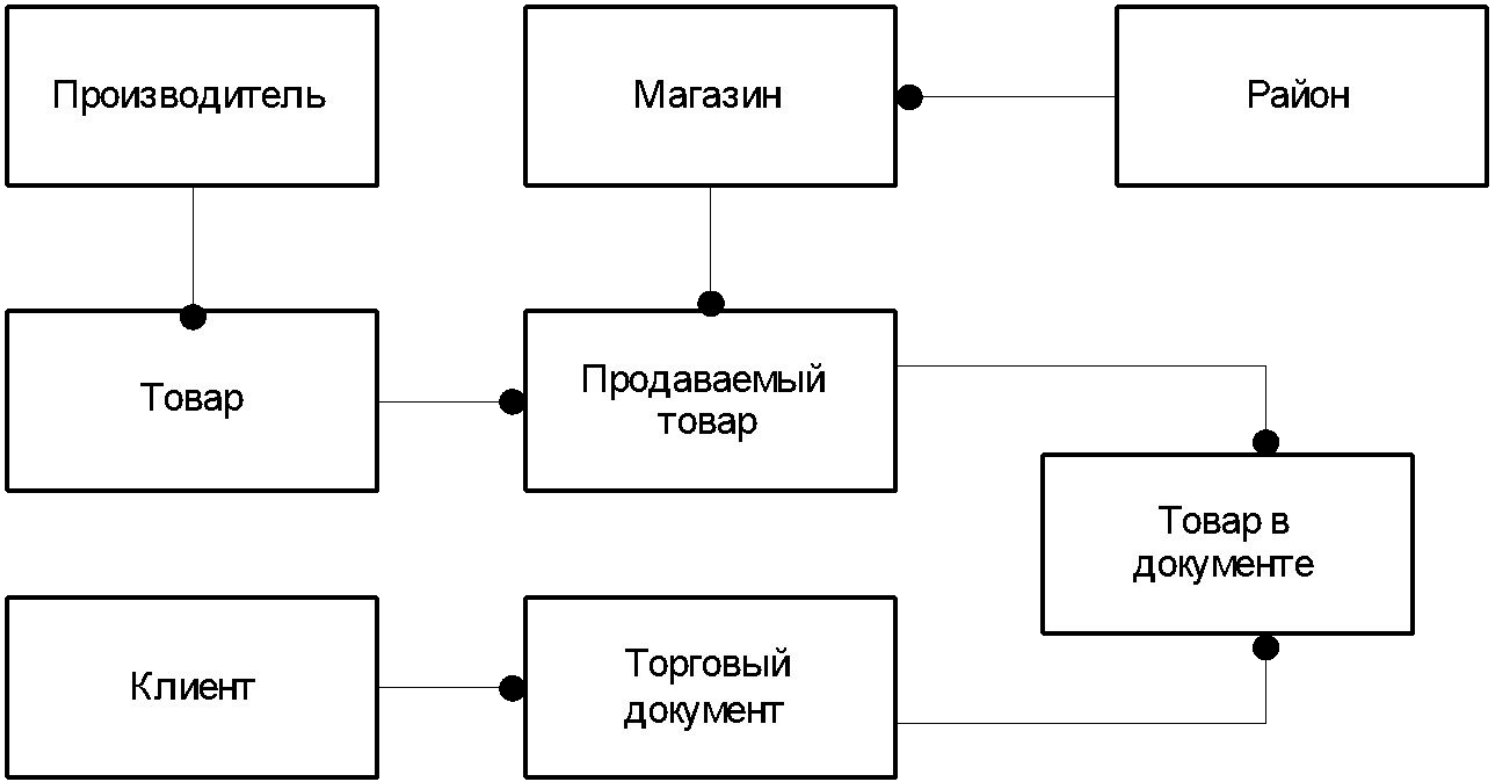
Управление повседневными бизнес процессами (OLTP)

Поддержка принятия стратегических решений (OLAP, Data mining)

Управление информационным содержанием

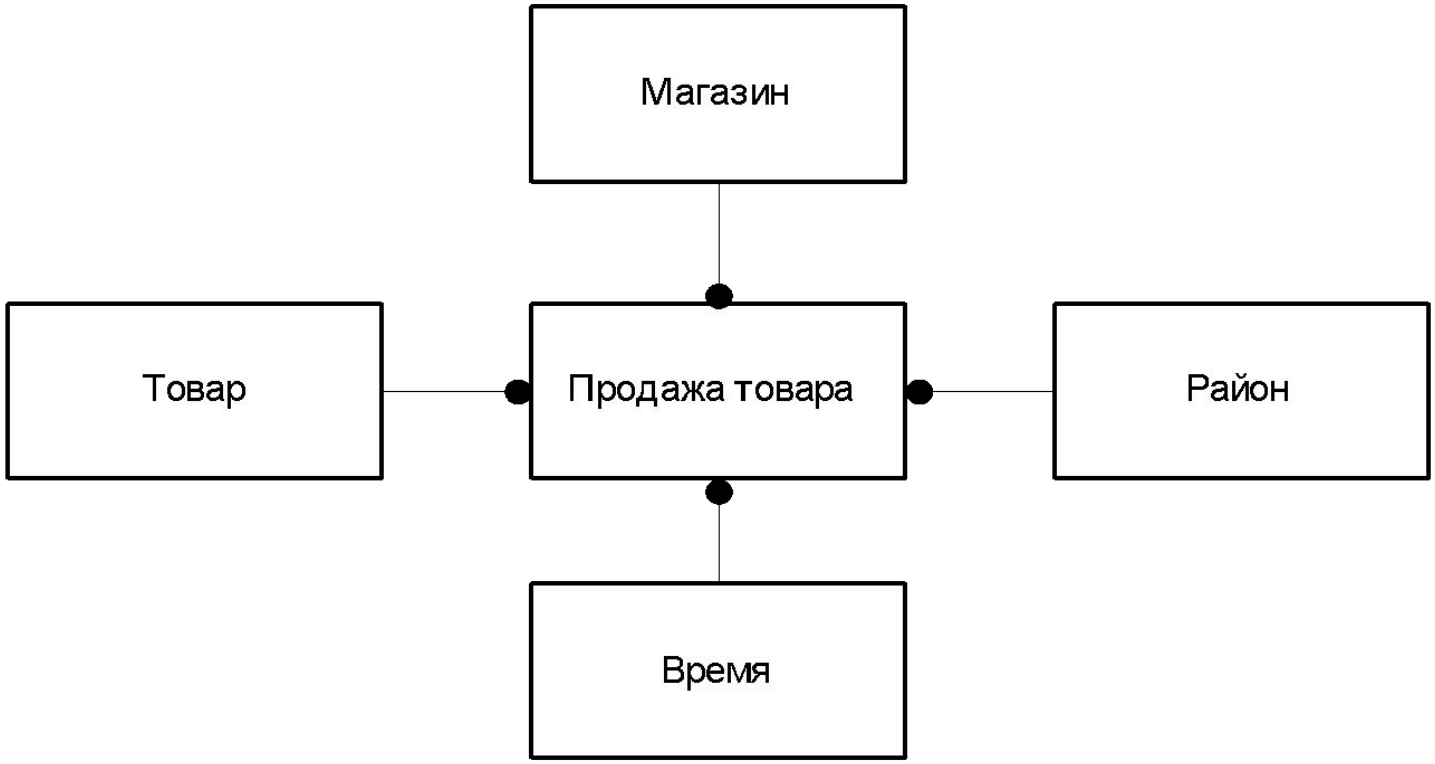
Области применения ИС

- 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

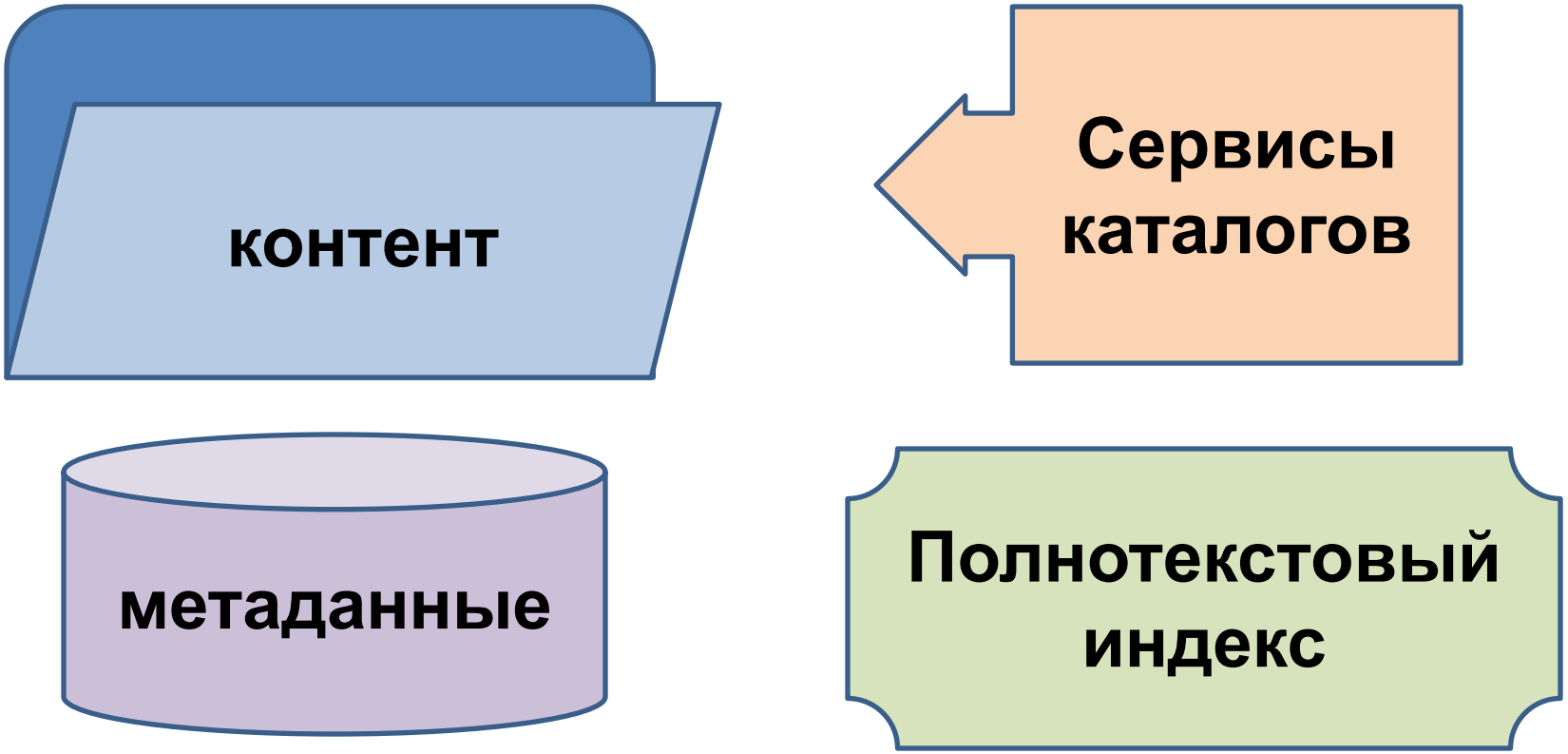
Управление контентом

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

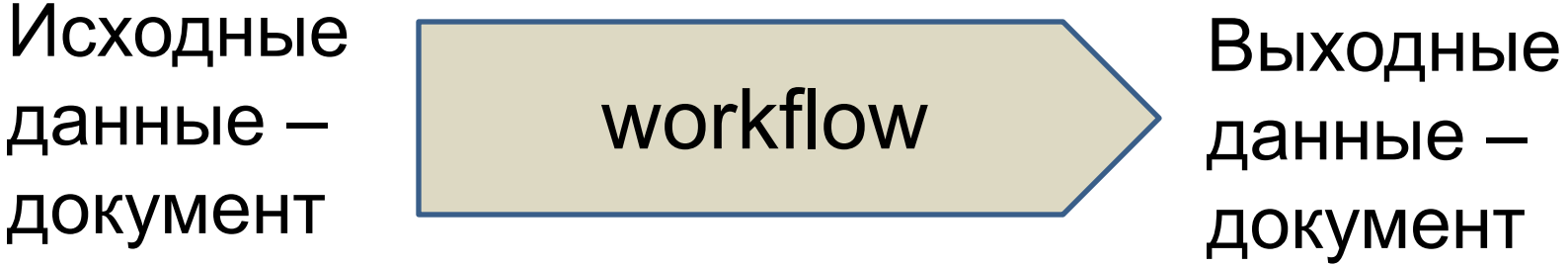
Управление процессами

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

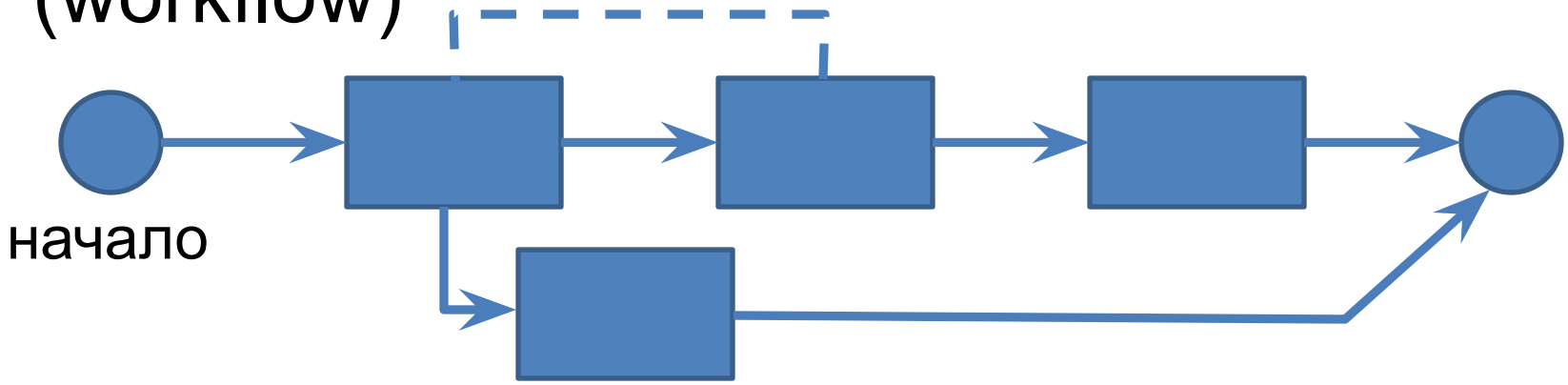


Workflow

Описание процесса

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

Конкретное выполнение работ – процесс (workflow)



Lifecycle

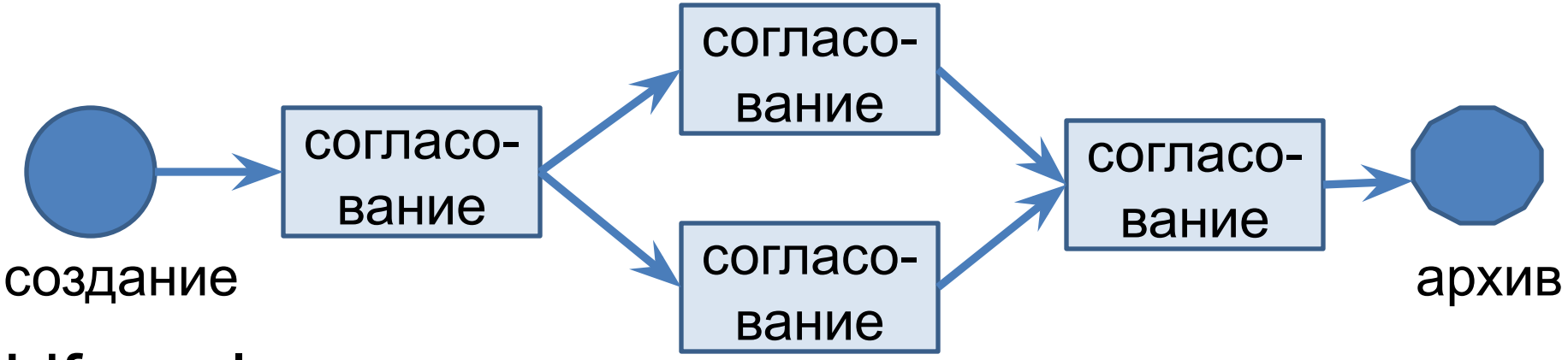
Строго последовательное переключение состояний

Состояния жизненного цикла

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

Пример

Workflow



Lifecycle