Проектирование баз данных

Часть 5. Хранилища данных и OLAP

Использование MS SQL Server Analysis Services 2008 для построения хранилищ данных

Автор: В.В. Полубояров (http://www.intuit.ru/department/database/mssqlsas2008)

5.1.

Хранилище данных и OLAP.

Назначение.

Основные характеристики

Бизнес-анализ

(BI, Business Intelligence)

– это категория приложений и технологий для сбора, хранения, анализа и публикации данных, позволяющая корпоративным пользователям принимать лучшие решения.

В русскоязычной терминологии подобные системы называются также системами поддержки принятия решений (СППР).

Сбор и хранение информации, а также решение задач информационно-поискового запроса эффективно реализуются средствами систем управления базами данных (СУБД) с помощью OLTP (Online Transaction Processing)-подсистем.

Непосредственно OLTP-системы не подходят для полноценного анализа информации.

Почему?

В силу противоречивости требований, предъявляемых к OLTP-системам и СППР.

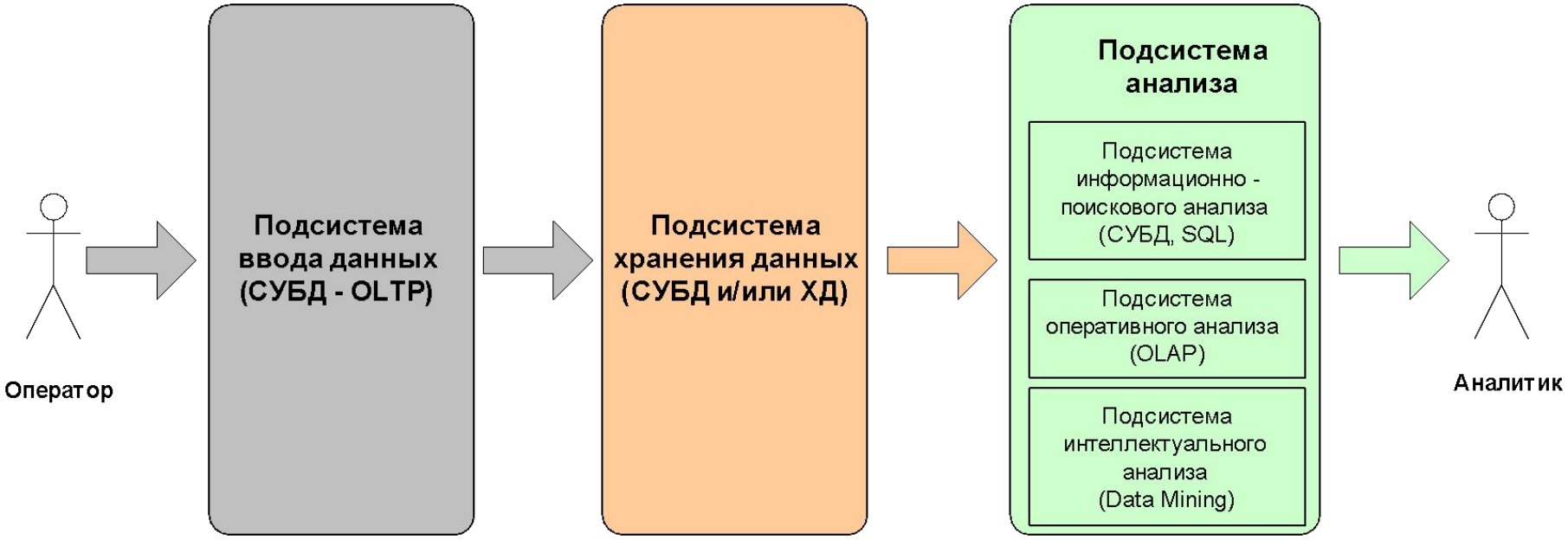
Для предоставления необходимой для принятия решений информации обычно приходится собирать данные из нескольких транзакционных баз данных различной структуры и содержания.

Основная проблема при этом состоит в несогласованности и противоречивости этих баз-источников, отсутствии единого логического взгляда на корпоративные данные.

В основе концепции ХД лежит идея разделения данных, используемых для оперативной обработки и для решения задач анализа, что позволяет оптимизировать структуры хранения.

ХД позволяет интегрировать ранее разъединенные детализированные данные, содержащиеся в исторических архивах, накапливаемых в традиционных OLTP-системах, поступающих из внешних источников, в единую базу данных, осуществляя их предварительное согласование и, возможно, агрегацию.

Архитектура СППР



Подсистема анализа может быть построена на основе:

- подсистемы информационно-поискового анализа на базе реляционных СУБД и статических запросов с использованием языка SQL;

- подсистемы оперативного анализа. Для реализации таких подсистем применяется технология оперативной аналитической обработки данных OLAP, использующая концепцию многомерного представления данных;

- подсистемы интеллектуального анализа, реализующие методы и алгоритмы Data Mining.

Понятие хранилищ данных

ХД – предметно-ориентированный, интегрированный, редко меняющийся, поддерживающий хронологию набор данных, организованный для целей поддержки принятия решений.

Предметная ориентация означает, что ХД интегрируют информацию, отражающую различные точки зрения на предметную область.

Интеграция предполагает, что данные, хранящиеся в ХД, приводятся к единому формату. Поддержка хронологии означает, что все данные в ХД соответствуют последовательным интервалам времени.

Структура СППР с физическим ХД

- При загрузке данных из OLTP-системы в ХД происходит дублирование данных.
- В ходе этой загрузки данные фильтруются, поскольку не все из них имеют значение для проведения процедур анализа.
- В ХД хранится обобщенная информация, которая в OLTP-системе отсутствует.

Виртуальные хранилища данных

В системе виртуальных ХД данные из OLTP-системы не копируются в единое хранилище. Они извлекаются, преобразуются и интегрируются непосредственно при выполнении аналитических запросов в режиме реального времени. Фактически такие запросы напрямую передаются к OLTP-системе.

Достоинства виртуального ХД:

- минимизация объема хранимых данных;
- работа с текущими, актуальными данными.

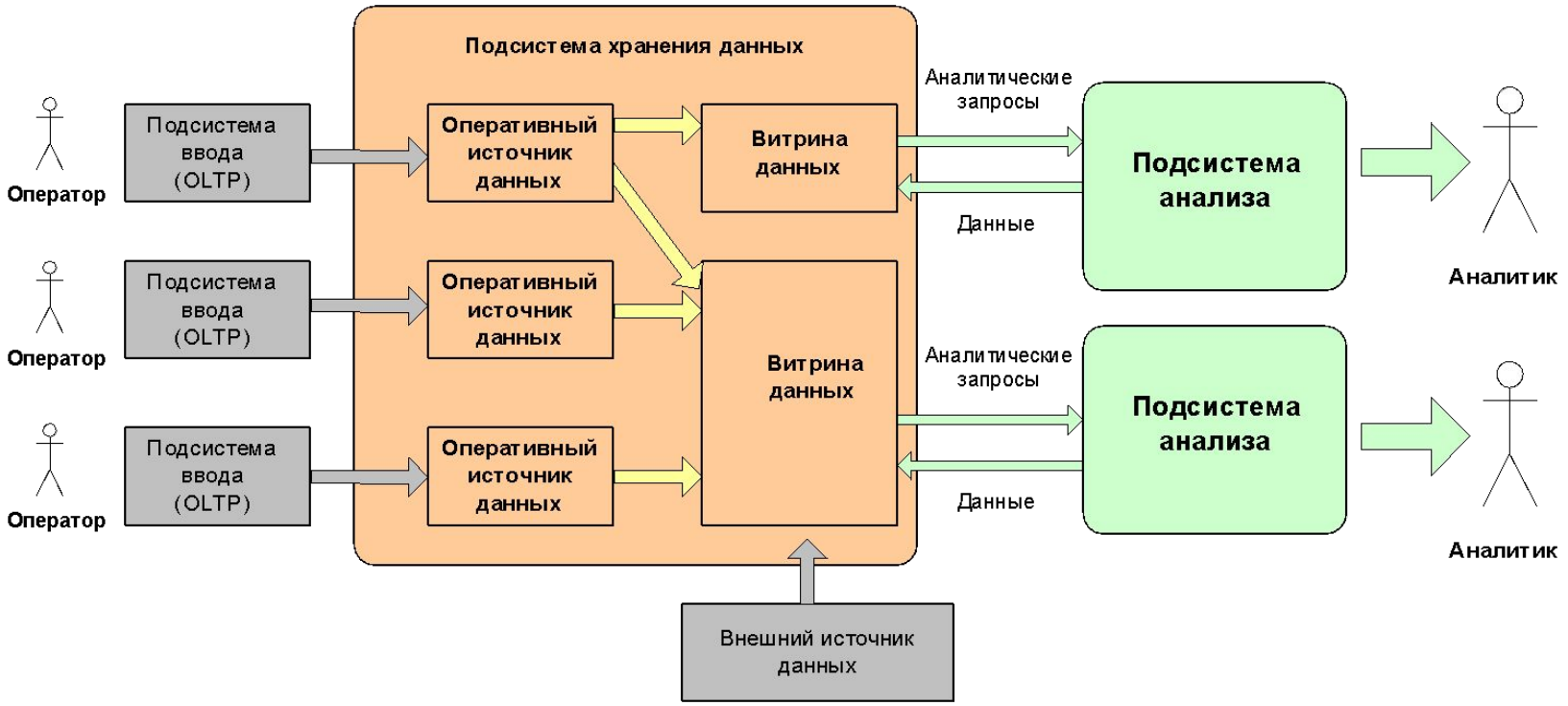
Недостатки виртуального ХД:

- более высокое, по сравнению с физическим ХД время обработки запросов;
- необходимость постоянной доступности всех OLTP-источников;
- снижение быстродействия OLTP-систем;
- OLTP-системы не ориентированы на хранение данных за длительный период времени, по мере необходимости данные выгружаются в архивные, поэтому не всегда имеется физическая возможность получения полного набора данных в ХД.

Проблемы построения хранилищ данных

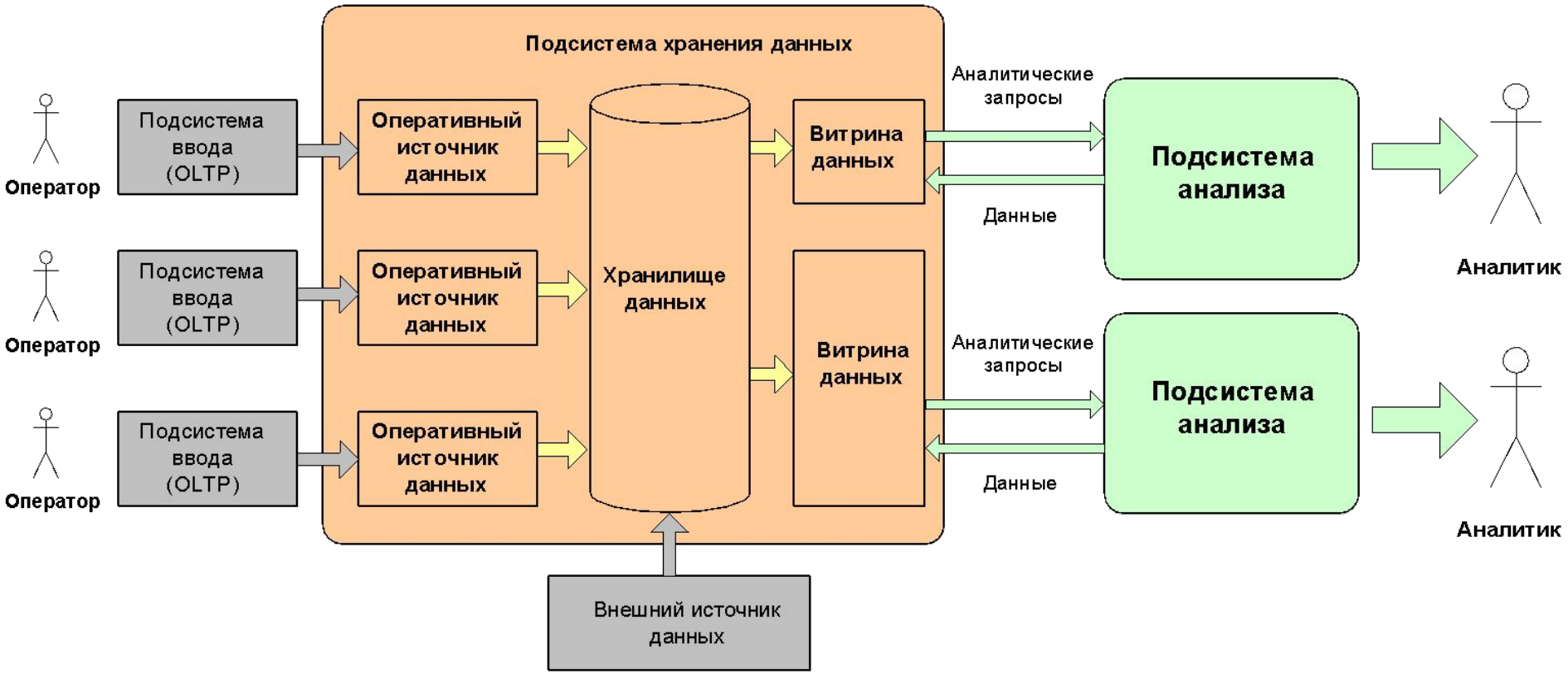
1. Интеграция разнородных данных.
2. Эффективное хранение и обработка больших объемов данных.
3. Организация многоуровневых справочников метаданных.
4. Обеспечение информационной безопасности ХД.

Структура СППР с самостоятельными витринами данных (ВД)



- ВД содержит данные, ориентированные на конкретного пользователя, существенно меньше по объему, и для ее реализации требуется меньше затрат.
- ВД могут строиться как самостоятельно, так и вместе с ХД.
- ВД внедряются гораздо быстрее и быстрее виден эффект от их использования.

Структура СППР
с хранилищами данных и витринами данных



5.2.

Понятие и модель данных OLAP

Понятие OLAP

OLAP (Online Analytical Processing) – технология оперативной аналитической обработки данных, использующая методы и средства для сбора, хранения и анализа многомерных данных в целях поддержки процессов принятия решений.

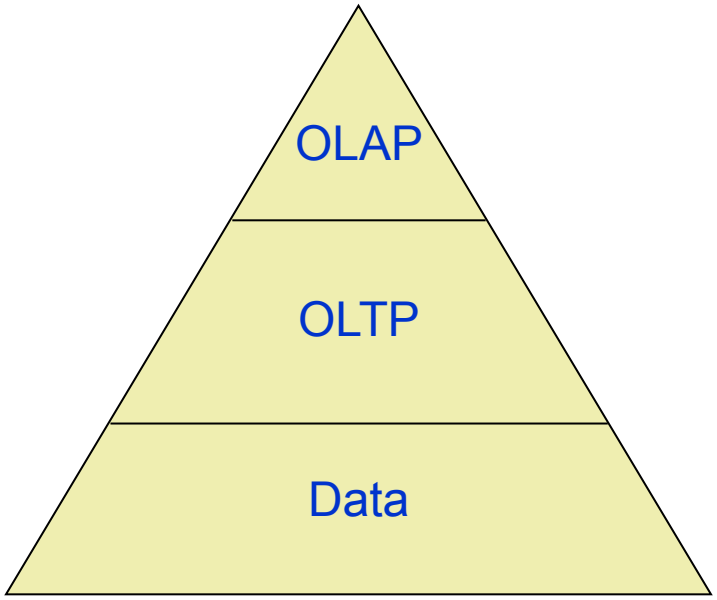
Основное назначение OLAP-систем – поддержка аналитической деятельности, произвольных запросов пользователей - аналитиков. Цель OLAP-анализа – проверка возникающих гипотез.

OLTP – On-Line Transaction Processing,

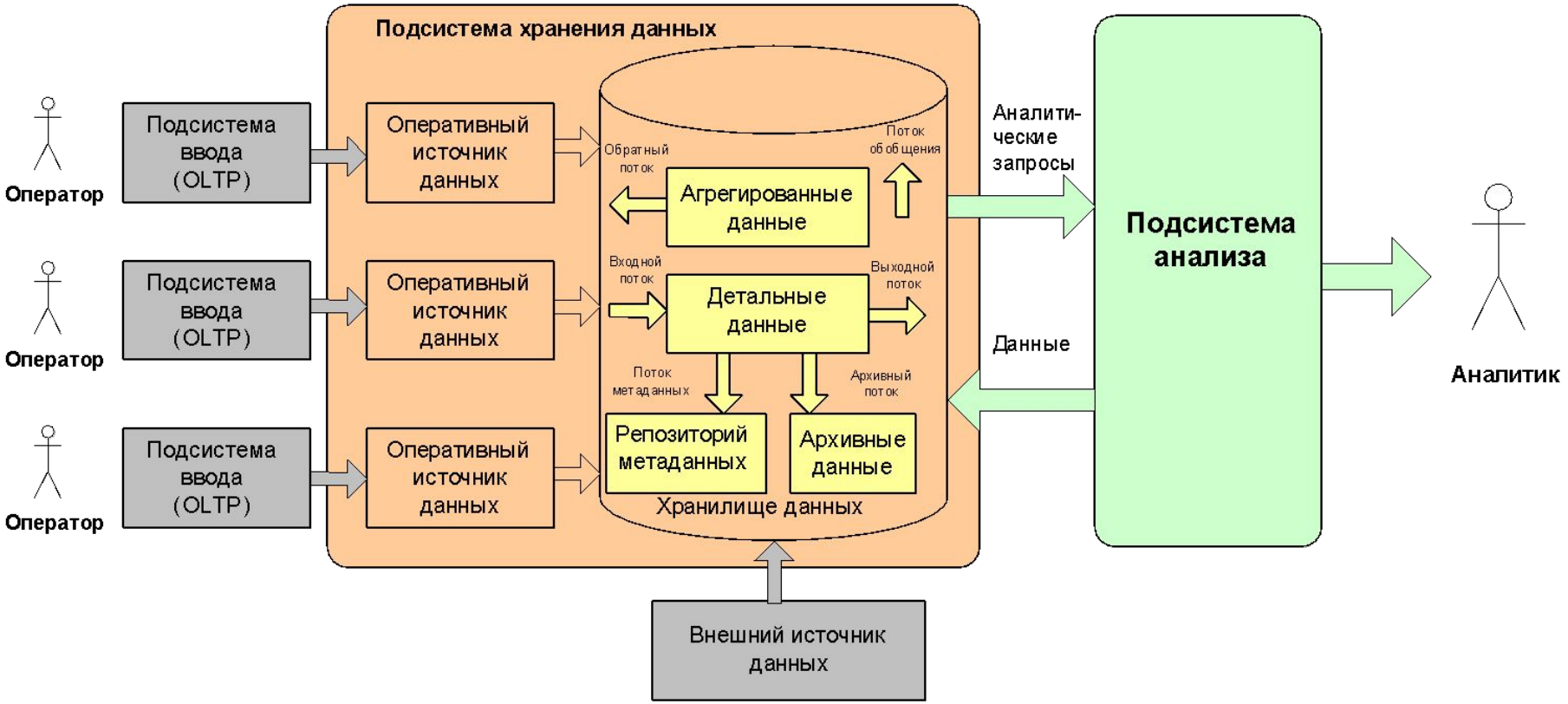
оперативная транзакционная обработка данных

OLAP – On-Line Analytical Processing

оперативная аналитическая обработка данных



Категории данных в хранилищах данных



1. Детальные данные
2. Агрегированные (обобщенные) данные
3. Метаданные

Информационные потоки в хранилищах данных

- входной поток - образуется данными, копируемыми из OLTP-систем в ХД; данные при этом часто очищаются и обогащаются путем добавления новых атрибутов;
- поток обобщения - образуется агрегированием детальных данных и их сохранением в ХД;
- архивный поток - образуется перемещением детальных данных, количество обращений к которым снизилось;
- поток метаданных - образуется потоком информации о данных в репозиторий данных;
- выходной поток - образуется данными, извлекаемыми пользователями;
- обратный поток - образуется очищенными данными, записываемыми обратно в OLTP-системы.

OLAP и OLTP.
Характеристики и основные отличия

Характеристики OLTP системы

- Большой объем информации
- Часто различные БД для разных подразделений
- Нормализованная схема, отсутствие дублирования информации
- Интенсивное изменение данных
- Транзакционный режим работы
- Транзакции затрагивают небольшой объем данных
- Обработка текущих данных – мгновенный снимок
- Много клиентов
- Малое время отклика – несколько секунд

OLAP и OLTP.
Характеристики и основные отличия

Характеристики OLAP системы

- Большой объем информации
- Синхронизированная информация из различных БД с использованием общих классификаторов
- Ненормализованная схема БД с дубликатами
- Данные меняются редко, Изменение происходит через пакетную загрузку
- Выполняются сложные нерегламентированные запросы над большим объемом данных с широким применением группировок и агрегатных функций.
- Анализ временных зависимостей
- Небольшое количество работающих пользователей – аналитики и менеджеры
- Большее время отклика (но все равно приемлемое) – несколько минут

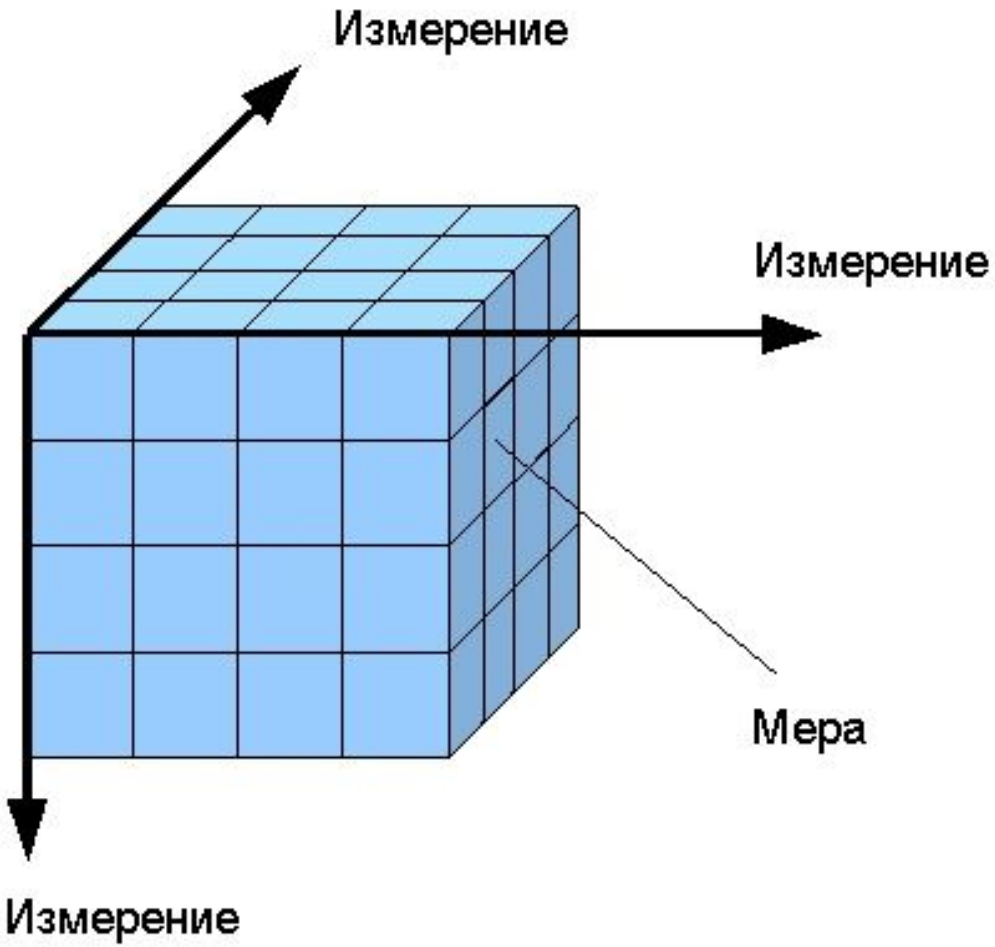
5.3.
Правила Кодда
для OLAP систем

1. Концептуальное многомерное представление
2. Прозрачность.
3. Доступность.
4. Постоянная производительность при разработке отчетов.
5. Клиент-серверная архитектура.
6. Общая многомерность.
7. Динамическое управление разреженными матрицами.
8. Многопользовательская поддержка.
9. Неограниченные перекрестные операции.
10. Интуитивная манипуляция данными.
11. Гибкие возможности получения отчетов.
12. Неограниченная размерность и число уровней агрегации.

5.4.

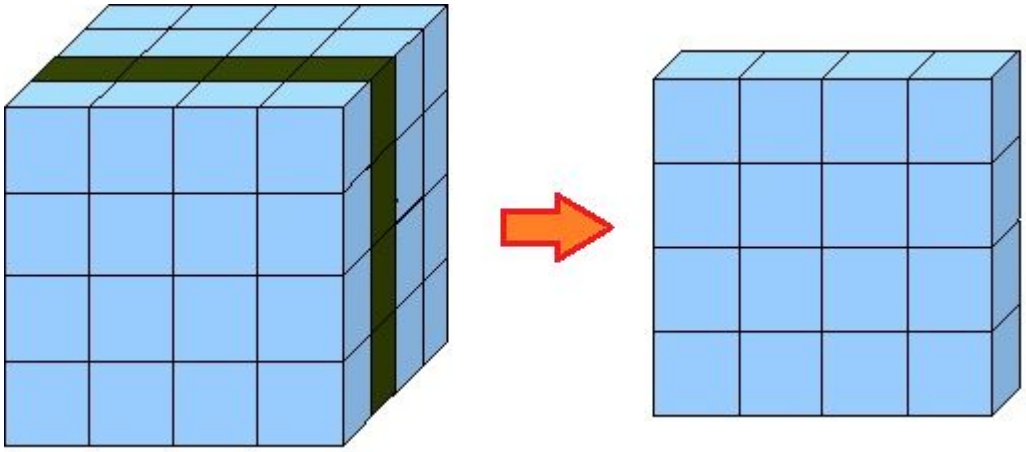
Структура OLAP-куба

Гиперкуб

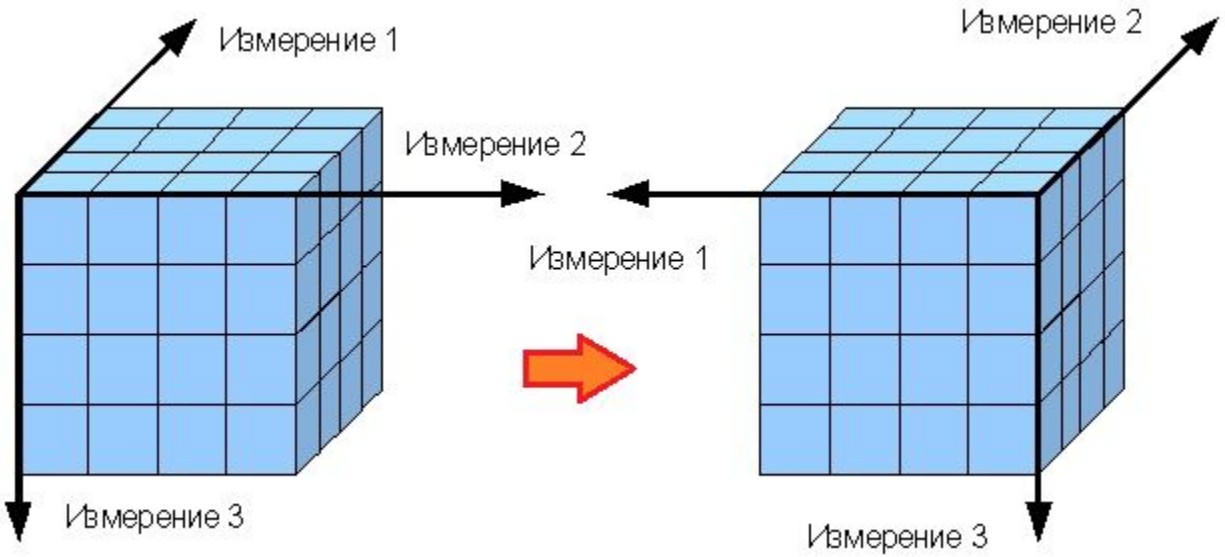


Операции, выполняемые над гиперкубом

1. Срез

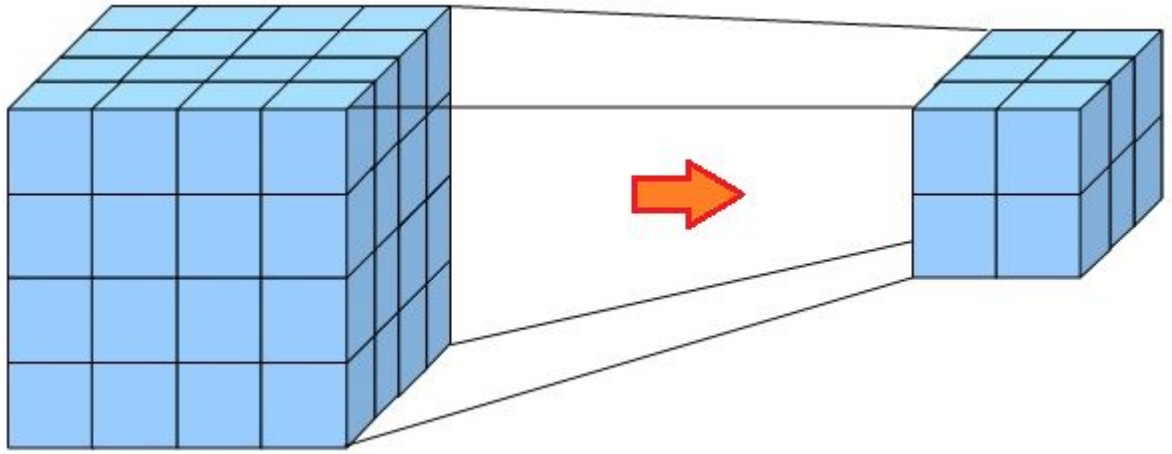


2. Вращение

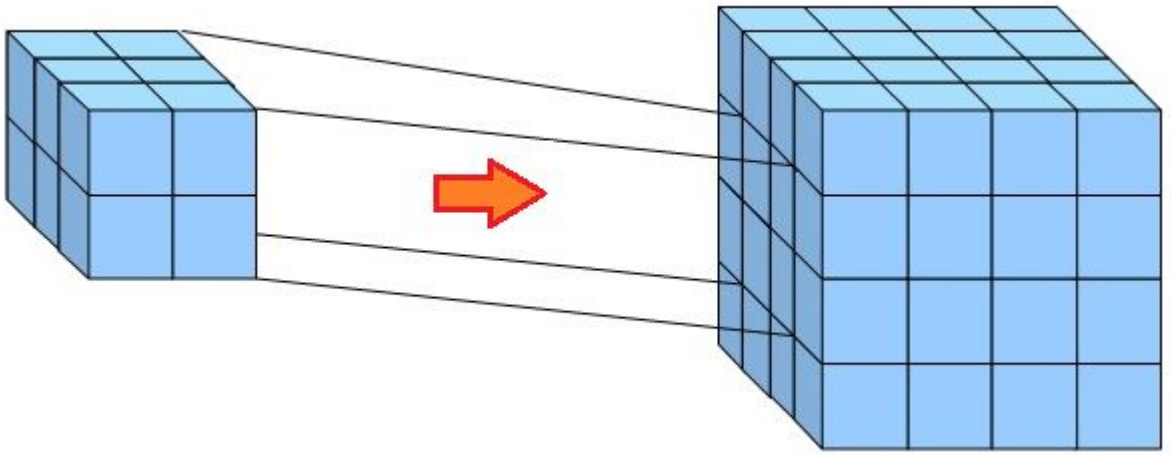


Операции, выполняемые над гиперкубом

3. Консолидация



4. Детализация



Фрагмент хранилища данных для OLAP

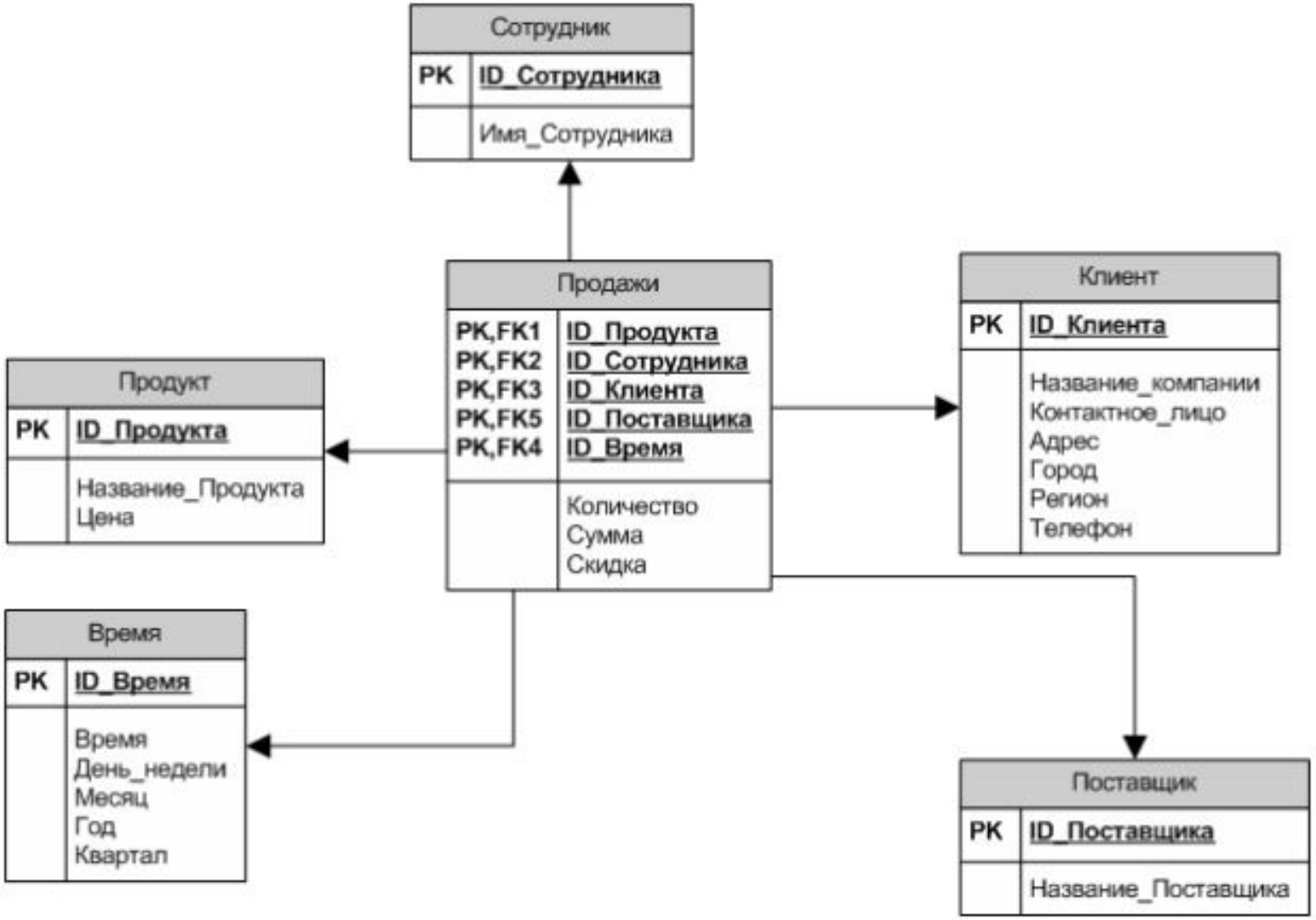


Таблица фактов

Основные типы таблиц фактов

1. Факты, связанные с транзакциями (Transaction facts).

2. Факты, связанные с "моментальными снимками" (Snapshot facts).

3. Факты, связанные с элементами документа (Line-item facts).

4. Факты, связанные с событиями или состоянием объекта (Event or state facts).

Таблица измерений

Таблицы измерений содержат неизменяемые либо редко изменяемые данные.

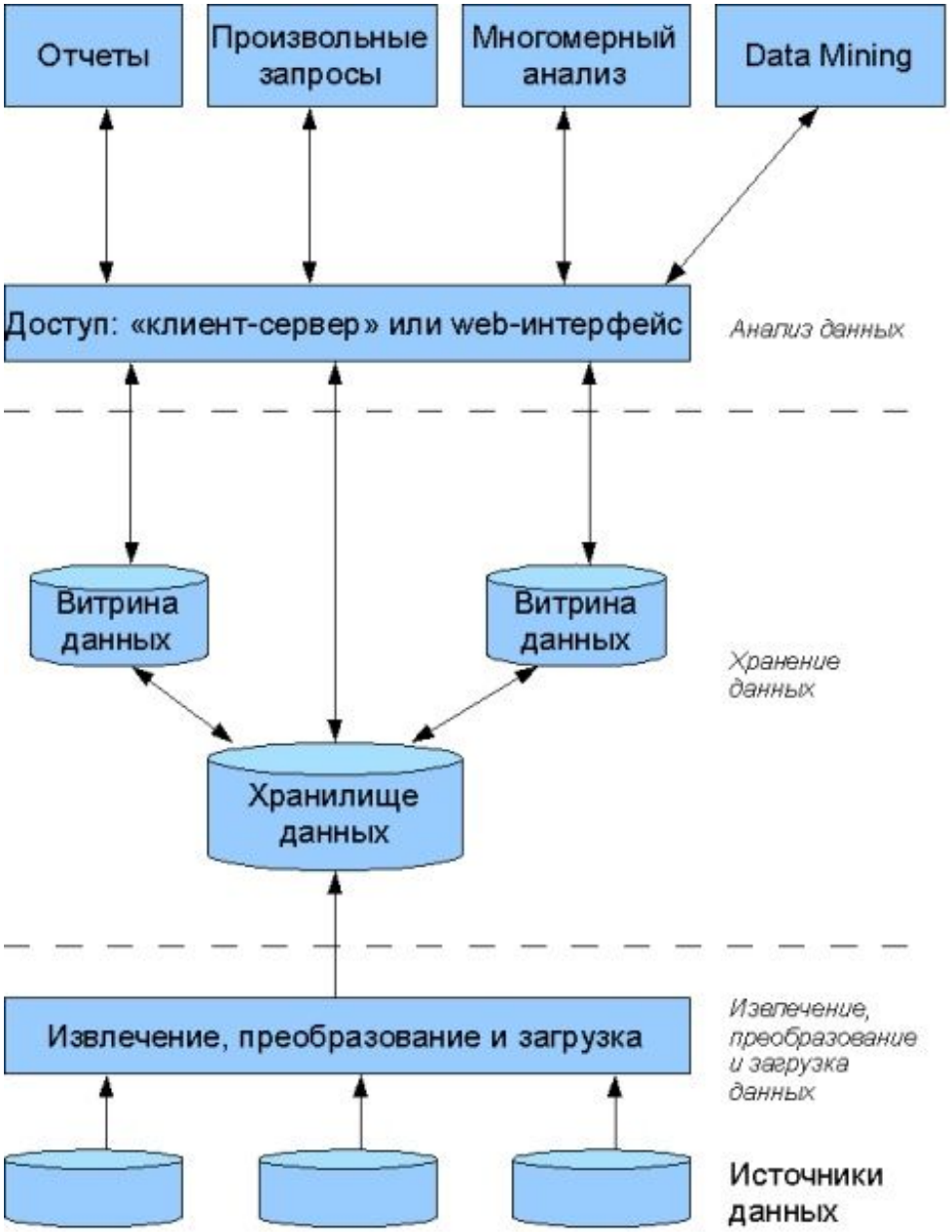
Таблицы измерений также содержат как минимум одно описательное поле (обычно с именем члена измерения) и, как правило, целочисленное ключевое поле (обычно это суррогатный ключ) для однозначной идентификации члена измерения.

Если будущее измерение, основанное на данной таблице измерений, содержит иерархию, то таблица измерений также может содержать поля, указывающие на "родителя" данного члена в этой иерархии.

Каждая таблица измерений должна находиться в отношении "один ко многим" с таблицей фактов.

Скорость роста таблиц измерений должна быть незначительной по сравнению со скоростью роста таблицы фактов

Архитектура OLAP-систем



5.5.

Реализация OLAP

Типы OLAP - серверов

- MOLAP (Multidimensional OLAP)
- ROLAP (Relational OLAP)
- HOLAP (Hybrid OLAP)

MOLAP - сервер

Детальные и агрегированные данные хранятся в многомерной базе данных.

Хранение данных в многомерных структурах позволяет манипулировать данными как многомерным массивом, благодаря чему скорость вычисления агрегатных значений одинакова для любого из измерений.

Однако в этом случае многомерная база данных оказывается избыточной, так как многомерные данные полностью содержат детальные реляционные данные.

MOLAP - сервер

Преимущества

- Высокая производительность.
- Структура и интерфейсы наилучшим образом соответствуют структуре аналитических запросов.
- Многомерные СУБД легко справляются с задачами включения в информационную модель разнообразных встроенных функций.

MOLAP - сервер

Недостатки

- MOLAP могут работать только со своими собственными многомерными БД и основываются на патентованных технологиях для многомерных СУБД, поэтому являются наиболее дорогими.

- По сравнению с реляционными, очень неэффективно используют внешнюю память, обладают худшими по сравнению с реляционными БД механизмами транзакций.

- Отсутствуют единые стандарты на интерфейс, языки описания и манипулирования данными.

- Не поддерживают репликацию данных, часто используемую в качестве механизма загрузки.

ROLAP - сервер

ROLAP-системы позволяют представлять данные, хранимые в классической реляционной базе, в многомерной форме или в плоских локальных таблицах на файл-сервере, обеспечивая преобразование информации в многомерную модель через промежуточный слой метаданных.

Агрегаты хранятся в той же БД в специально созданных служебных таблицах. В этом случае гиперкуб эмулируется СУБД на логическом уровне.

ROLAP - сервер

Преимущества

- Работа с очень большими БД
- Развитые средства администрирования.
- Инструменты ROLAP позволяют производить анализ непосредственно над хранилищем данных.
- В случае переменной размерности задачи ROLAP не требуют физической реорганизации БД, как в случае MOLAP.
- Системы ROLAP могут функционировать на гораздо менее мощных клиентских станциях, чем системы MOLAP.
- Более высокий уровень защиты данных и хорошие возможности разграничения прав доступа.

ROLAP - сервер

Недостатки

- Ограниченные возможности с точки зрения расчета значений функционального типа.

- Меньшая производительность, чем у MOLAP. Для обеспечения сравнимой с MOLAP производительности реляционные системы требуют тщательной проработки схемы БД и специальной настройки индексов. Но в результате этих операций производительность хорошо настроенных реляционных систем при использовании схемы "звезда" сравнима с производительностью систем на основе многомерных БД.

HOLAP - сервер

Детальные данные остаются в той же реляционной базе данных, где они изначально находились, а агрегатные данные хранятся в многомерной базе данных

Схемы реализации OLAP в реляционных системах

- Схема «Звезда»
- Схема «Снежинка»

Схема «Звезда»

Каждое измерение содержится в одной таблице.

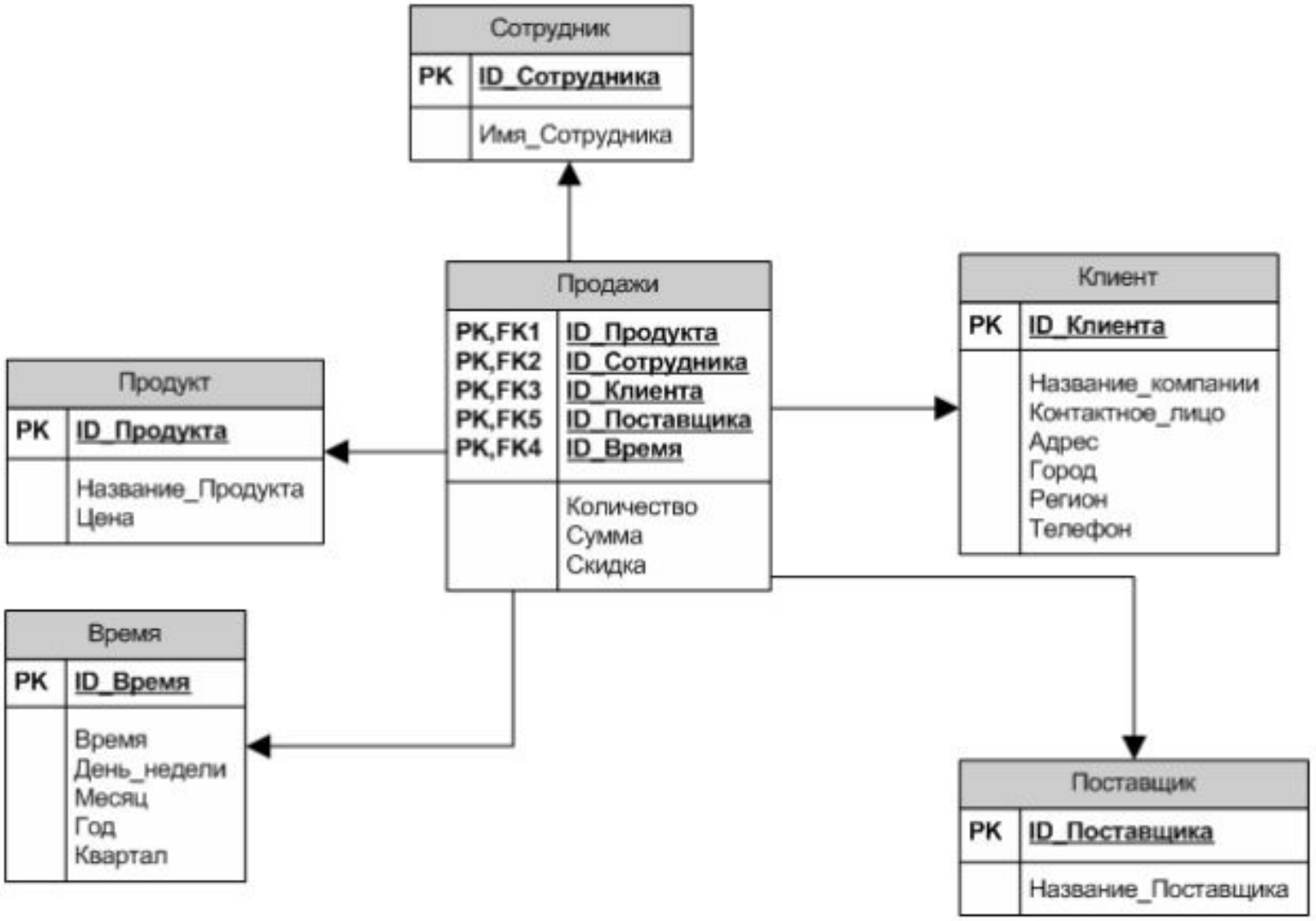


Схема «Звезда»

Особенности:

- Одна таблица фактов (fact table), которая сильно денормализована является центральной в схеме, может состоять из миллионов строк и содержит суммируемые или фактические данные, с помощью которых можно ответить на различные вопросы.

- Несколько денормализованных таблиц измерений (dimensional table) имеют меньшее количество строк, чем таблицы фактов, и содержат описательную информацию. Эти таблицы позволяют пользователю быстро переходить от таблицы фактов к дополнительной информации.

- Таблица фактов и таблицы размерности связаны идентифицирующими связями, при этом первичные ключи таблицы размерности мигрируют в таблицу фактов в качестве внешних ключей. Первичный ключ таблицы факта целиком состоит из первичных ключей всех таблиц размерности.

- Агрегированные данные хранятся совместно с исходными.

Схема «Звезда»

Преимущества

Благодаря денормализации таблиц измерений упрощается восприятие структуры данных пользователем и формулировка запросов, уменьшается количество операций соединения таблиц при обработке запросов. Некоторые промышленные СУБД и инструменты класса OLAP / Reporting умеют использовать преимущества схемы "звезда" для сокращения времени выполнения запросов.

Схема «Звезда»

Недостатки

Денормализация таблиц измерений вносит избыточность данных, возрастает требуемый для их хранения объем памяти.

Если агрегаты хранятся совместно с исходными данными, то в измерениях необходимо использовать дополнительный параметр - уровень иерархии.

Схема «Снежинка»

Существует измерение, которое содержится в нескольких таблицах

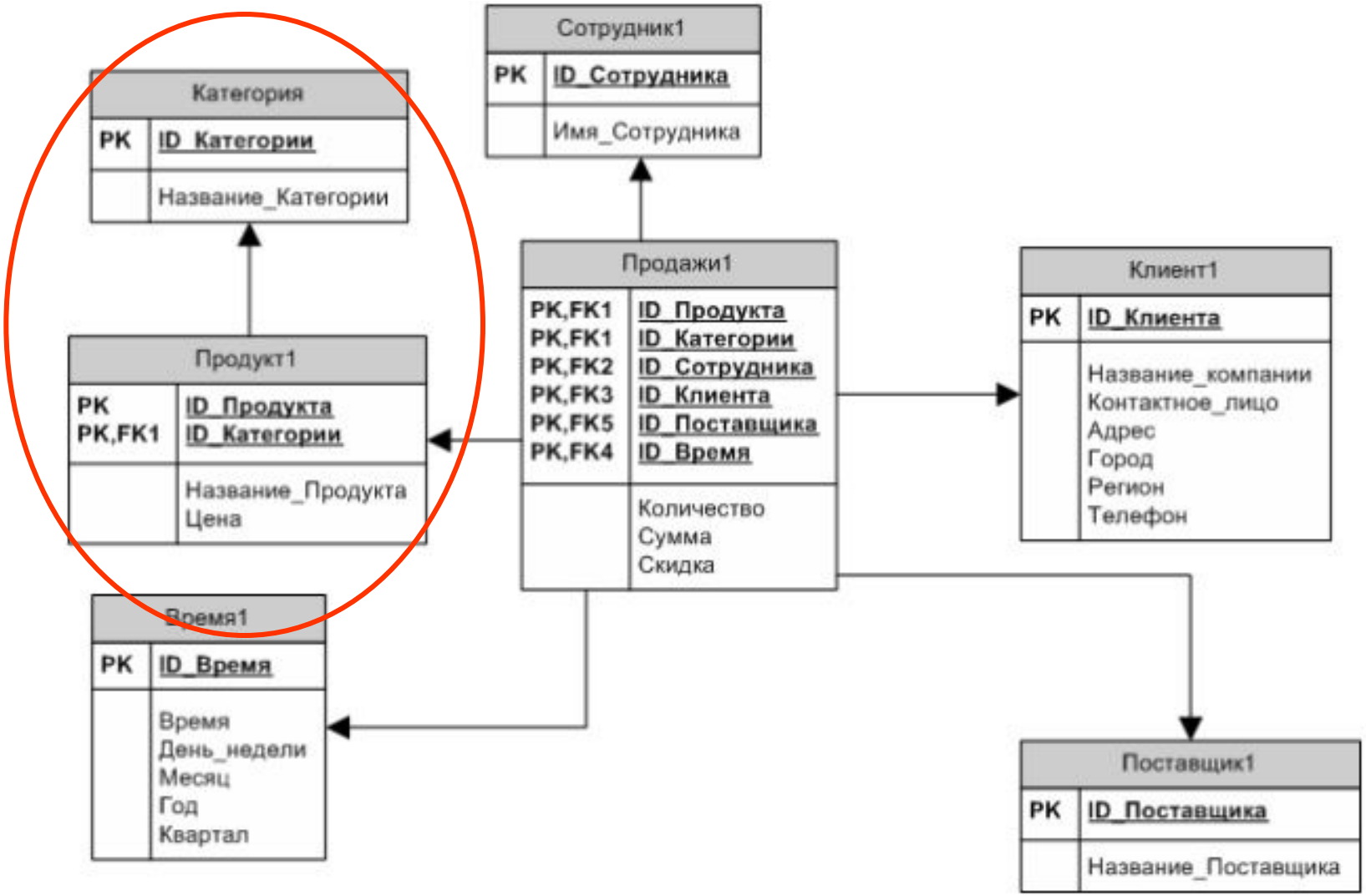


Схема «Снежинка»

Особенности:

- Одна таблица фактов (fact table), которая сильно денормализована является центральной в схеме, может состоять из миллионов строк и содержать суммируемые или фактические данные, с помощью которых можно ответить на различные вопросы.

- Несколько таблиц измерений (dimensional table), которые нормализованы в отличие от схемы "звезда". Имеют меньшее количество строк, чем таблицы фактов, и содержат описательную информацию. Эти таблицы позволяют пользователю быстро переходить от таблицы фактов к дополнительной информации. Первичные ключи в них состоят из единственного атрибута (соответствуют единственному элементу измерения).

- Таблица фактов и таблицы размерности связаны идентифицирующими связями, при этом первичные ключи таблицы размерности мигрируют в таблицу фактов в качестве внешних ключей. Первичный ключ таблицы факта целиком состоит из первичных ключей всех таблиц размерности.

- В схеме "снежинка" агрегированные данные могут храниться отдельно от исходных

Схема «Снежинка»

Преимущества

Нормализация таблиц измерений

в отличие от схемы "звезда" позволяет

минимизировать избыточность данных и более

эффективно выполнять запросы,

связанные со структурой значений измерений.

Схема «Снежинка»

Недостатки

За нормализацию таблиц измерений иногда приходится платить временем выполнения запросов.