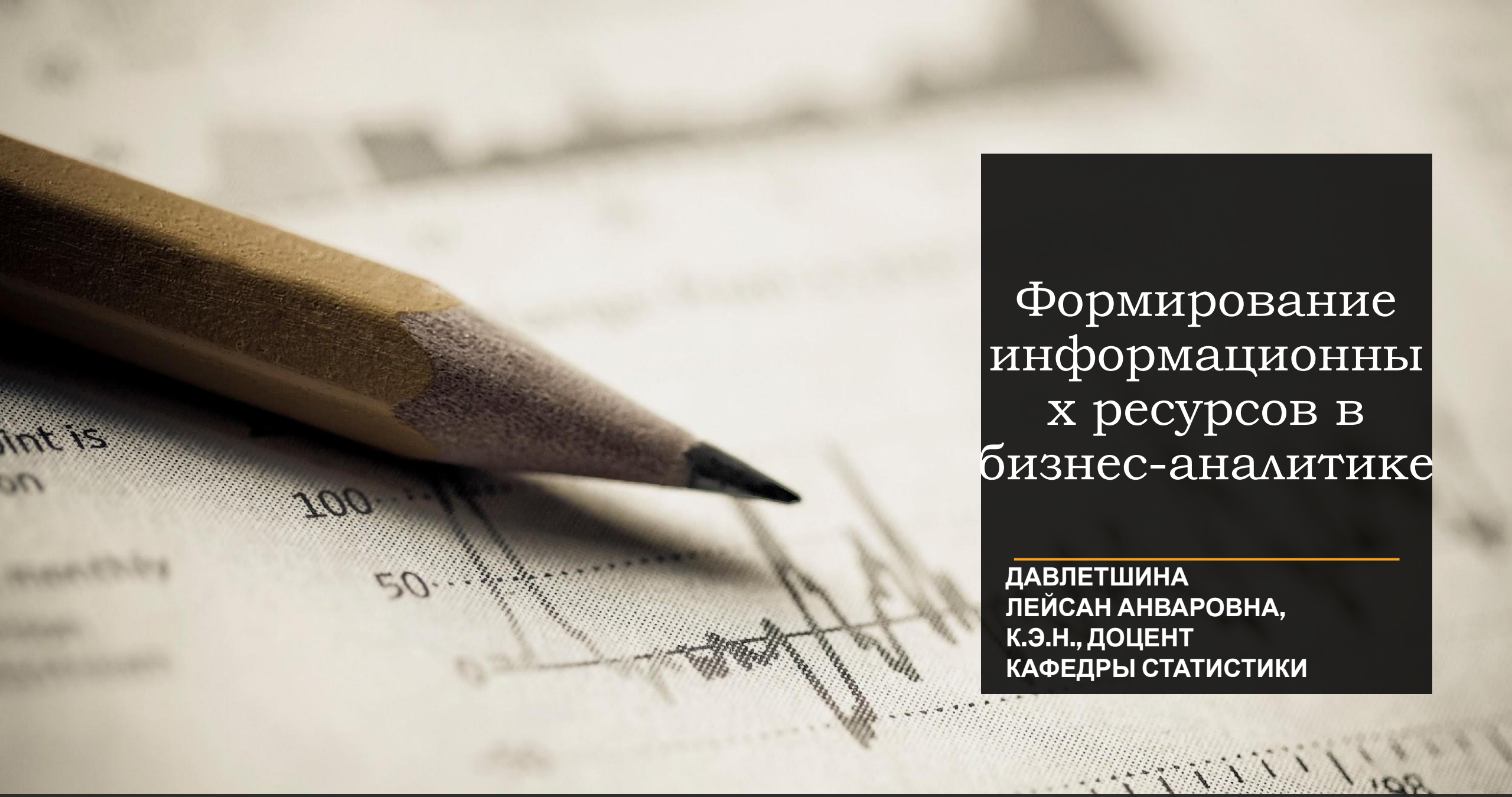


Формирование информационных ресурсов в бизнес-аналитике

ДАВЛЕТШИНА ЛЕЙСАН АНВАРОВНА,
К.Э.Н., ДОЦЕНТ
КАФЕДРЫ СТАТИСТИКИ

Структура курса

1	Изучение основных баз данных. Обоснование выбора базы данных на основе принятия решения
2	Информационная культура и архитектура. Интеллектуальная среда бизнес-анализа
3	Технологические изменения. Состояние российского рынка информации. Определение достоверных источников информации
4	Формирование информационных ресурсов для анализа из достоверных источников
5	Особенности формирования информационных потребностей субъектов управления
6	Определение необходимых источников информации для формирования выборочной совокупности данных из вторичных источников
7	Основные правила при ведении статистических регистров. Консолидация информации для формирования достоверной статистики из различных ведомств.
8	Оперативный мониторинг и анализ бизнес-ситуаций
9	Современные проблемы в развитии прикладных информационных систем
10	Способы ведения контроля по учету статистической информации

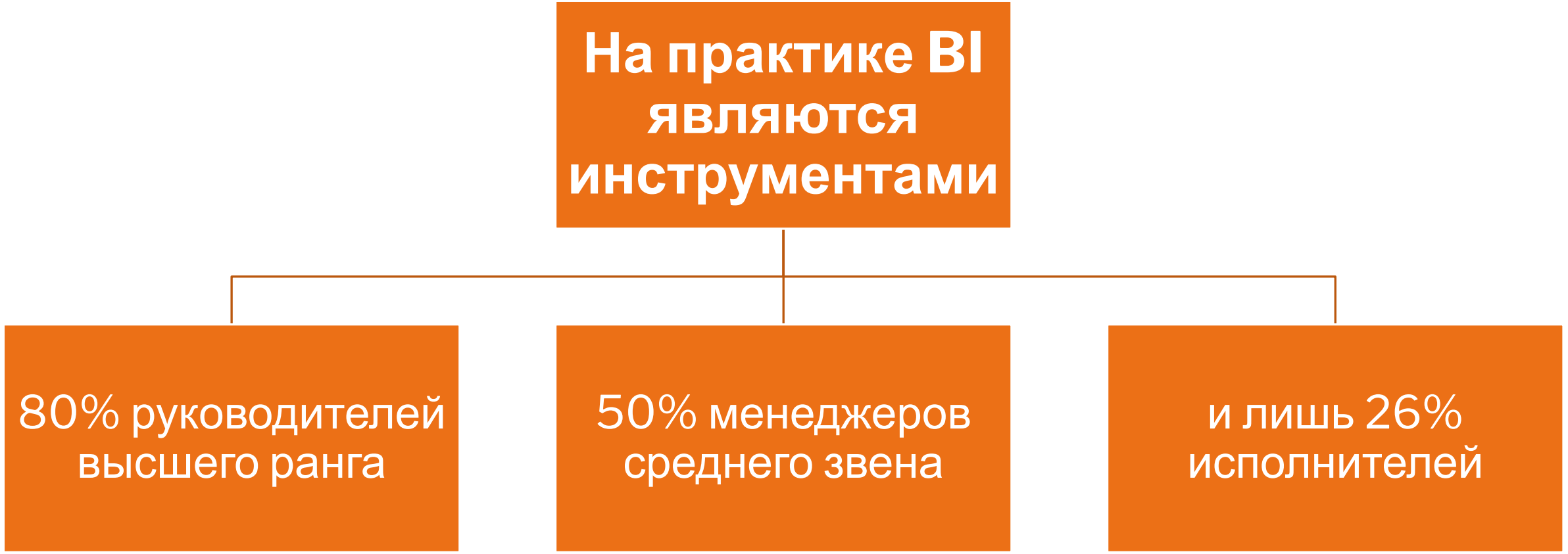
Тема 1. Изучение основных баз данных. Обоснование выбора базы данных на основе принятия решения

Тема 1. Изучение основных баз данных. Обоснование выбора базы данных на основе принятия решения

1. Данные, информация, знания
2. От транзакционных систем к системам аналитическим
3. Модели данных хранилищ данных
4. Сценарий функционирования хранилища данных

1.1. Данные, информация, знания

Средства бизнес-анализа (Business Intelligence, BI)



Факторы сдерживания внедрения средств бизнес-анализа:

- сотрудники не знают их возможностей (57%)
- высокая стоимость таких решений (54%)
- недостаточная квалификация пользователей (34%)

Основные ожидания руководства компаний в отношении средств BI

- всем сотрудникам необходимо овладеть навыками принятия решений с их помощью
- должны совершенствоватьс я технологии и повышаться уровень доступности
- упрощение использования, обеспечении точности и недвусмысленности предоставляемой информации

Информация – это отражение реального (материального, предметного) мира, выражаемого в виде сигналов и знаков. Понятие «**информация**» многозначно, и поэтому пока строго не определено, но можно сказать, что информация - *это содержание сообщения, сигнала памяти, а также сведения, имеющиеся в них.*

Информацию, как продукт производства и применения, отличает, прежде всего, *предметное содержание*. Информация подразделяется по виду обслуживаемой ею человеческой деятельности на:

- научную
- техническую
- производственную
- социальную
- правовую
- управленческую
- экономическую
- финансовую и т.д.

Информация при управлении организацией позволяет

- определять стратегические, тактические и оперативные цели и задачи организации
- осуществлять контроль за текущим состоянием организации, ее подразделений и процессов в них
- принимать обоснованные и своевременные решения, координировать действия подразделений в достижении целей

Особенности информации:

· информация сама по себе не материальна, а материальны носители информации

· информацию можно получить, запомнить, записать, забыть, передать, стереть, но ее нельзя получить там где ее нет.

· если человек получил некоторую информацию, то скольким бы лицам он ее не сообщил, у него количество информации от этого не уменьшится, а у всех получивших от него эту информацию также не превысит первоначальной порции.

Данные - это величины, их отношения, словосочетания, факты, преобразование и обработка которых позволяет извлечь новую информацию, т.е. знание о том или ином предмете, процессе, явлении.

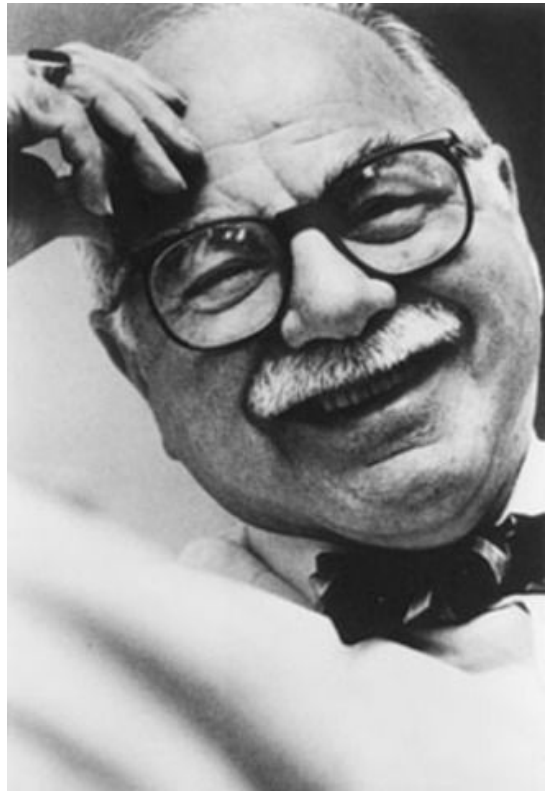
Данными обычно называют сведения :

- плановые,
- учетные,
- нормативные.

Все данные считаются **полезными**, даже *некачественные*. Просто полезность *некачественных* данных *отрицательна*.

Знания – есть некоторая более высокая степень организации данных.

По определению Даниэла Белла* «знание – это совокупность организованных высказываний о фактах или идеях, представляющих обоснованное суждение или экспериментальный результат и передается другим посредством некоторого средства коммуникации в систематизированной форме».



Даниел Белл — американский социолог и публицист, создатель теории постиндустриального (информационного) общества, профессор Гарвардского университета.

Если **данные** – это *отдельные факты*, характеризующие объекты, процессы и явления *в предметной области*, а также их свойства, то **знания** – это *выявленные закономерности* предметной области.

Знания имеют свои определенные свойства:	знания могут быть представлены в форме данных, например в виде текста на некотором формальном носителе, в виде сети, задающей связи разного рода
	знания обладают способностью управлять информационными процессами (вычислениями)
	знания всегда используются для решения задач
	знания делятся на отдельные фрагменты – описание объектов, процессов, ситуаций, явлений.

Что позволяет делать информация при управлении организацией?

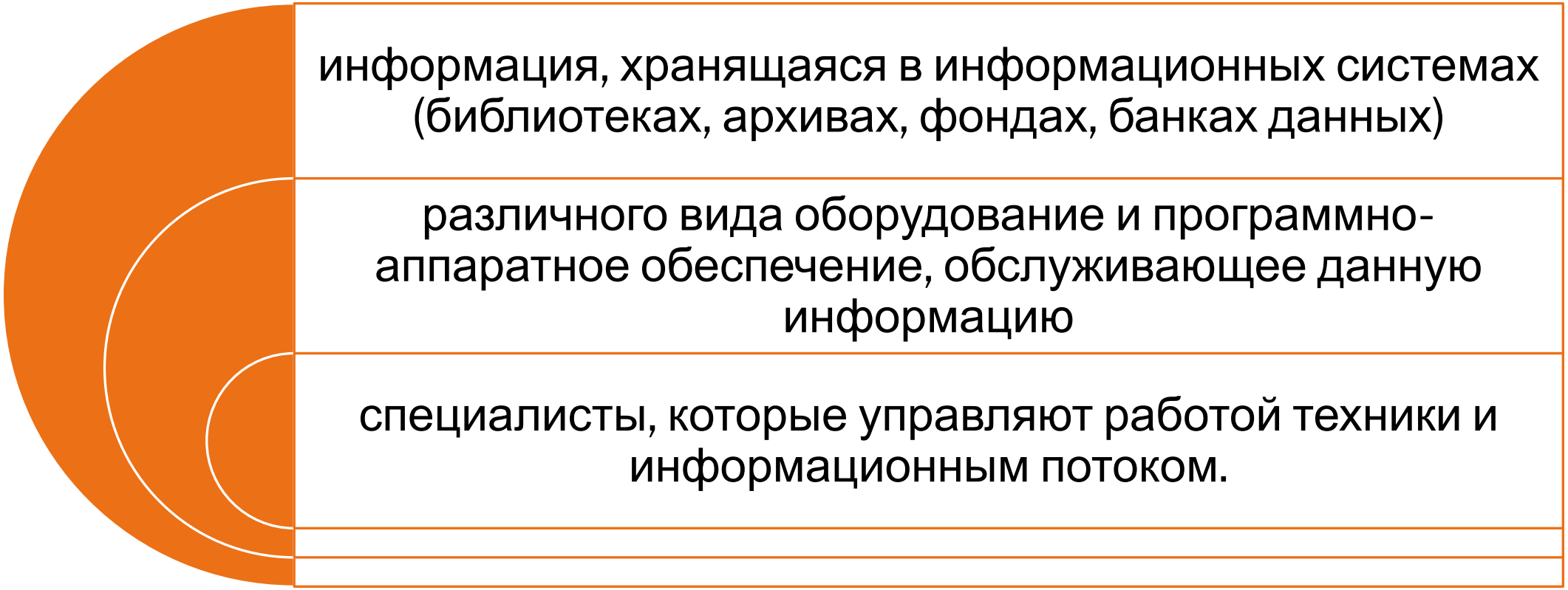
1. определять стратегические, тактические и оперативные цели и задачи организации

2. осуществлять контроль за текущим состоянием организации, ее подразделений и процессов в

3. принимать обоснованные и своевременные решения

4. координировать действия подразделений в достижении целей

Информационные ресурсы – это совокупность всех информационных продуктов произведенным обществом.



Информационный продукт - это информационные ресурсы всех видов, программные продукты, базы и банки данных и другая информация, представленные в форме товара.

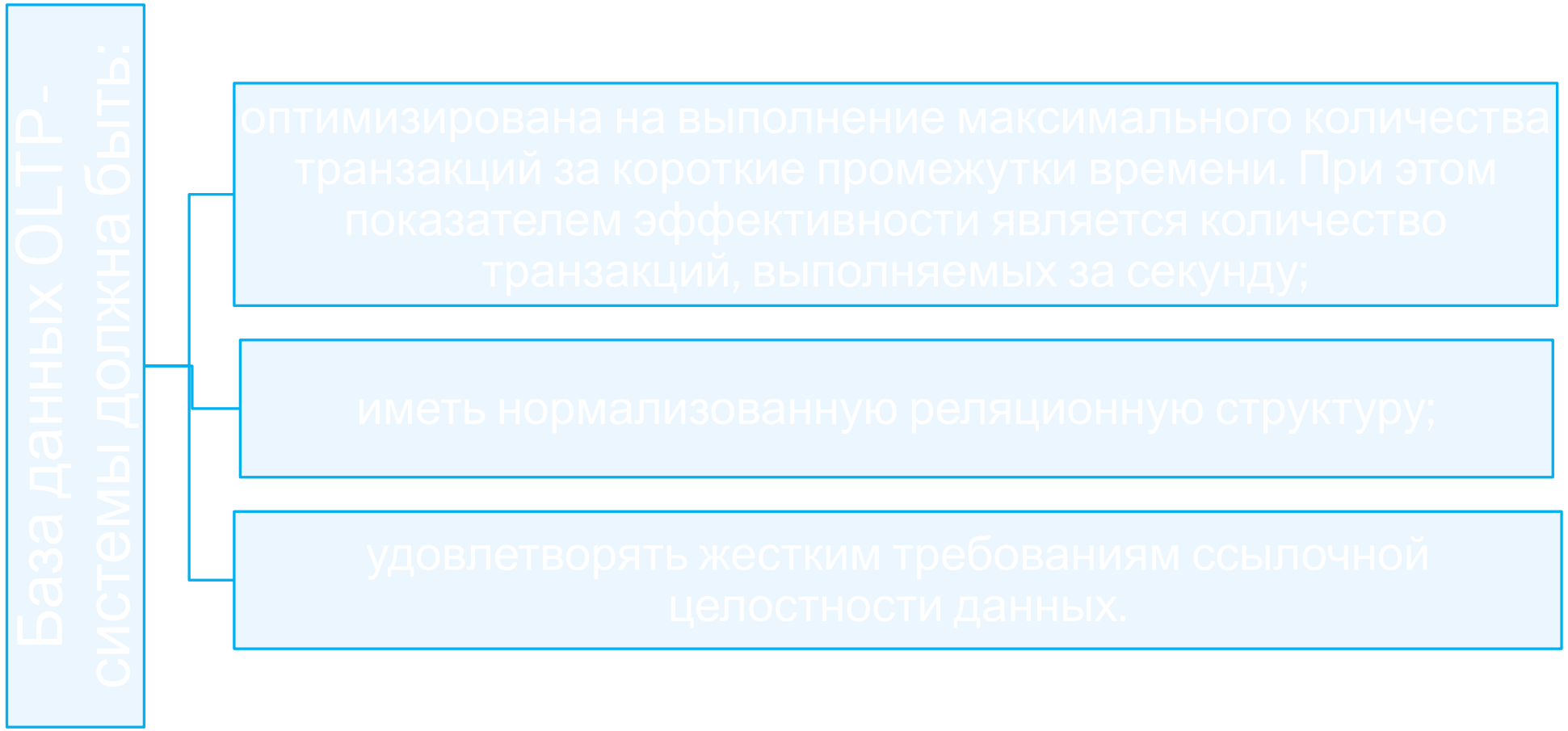
ИП в форме различного рода информации является источником человеческих знаний выраженных в:

- программных средствах
- базах данных
- служб экспертного обеспечения

1.2. От транзакционных систем к системам аналитическим

Для автоматизации операционных задач (учет платежей в бюджет, учет расходов бюджета, учет клиентов, учет договоров, учет заказов, учет взаиморасчетов, учет запасов и пр.), которые решаются сотрудниками «нижнего» звена финансовых учреждений, производственных, консалтинговых компаний и других организаций, традиционно используются учетные системы, называемые также **OLTP-системами** или **транзакционными** системами

OLTP-система осуществляет *учет и хранение первичной информации* о работе организации, обрабатывая огромное количество транзакций, производя «горы» данных, связанных с операционной деятельностью.



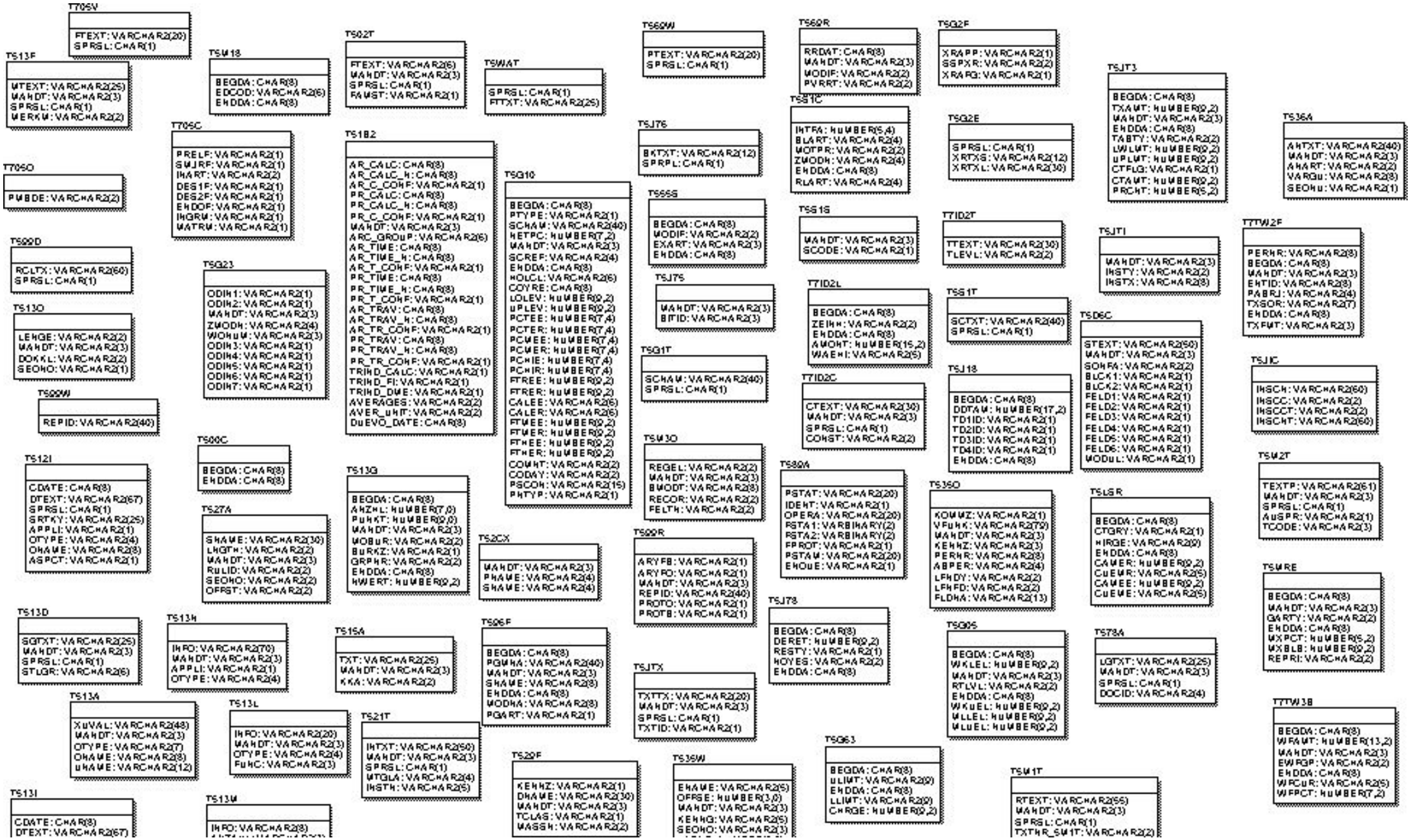


Рисунок 1 – Фрагмент структуры базы данных OLTP-системы

Получение агрегированной информации из базы данных OLTP-системы часто требует выполнения операций **соединения** по многим таблицам, содержащим *большое число записей*.

Такие запросы могут выполняться часами, **перегружая OLTP-систему**, нарушая тем самым нормальную работу подразделений организации.

Кроме того, для **планирования и оптимизации ресурсов** руководителям нужно знать, как на загрузку предприятия *влияют сезонные и годовые тренды*.

Например, можно *сравнивать продажи* в течение первого квартала этого года с продажами в течение первого квартала предшествующих лет или попытаться *оценить влияние* новой компании маркетинга, проходящей в течение определенных периодов, рассматривая продажи в течение тех же самых периодов.

Однако OLTP- системы **не предназначены** *для хранения, анализа информации за длительный период времени*.
Данные в большинстве OLTP-систем архивируются сразу после того, как они становятся неактивными

Таким образом характеристики OLTP-системы **не позволяют** использовать их *для оперативного анализа* непосредственно лицами, принимающими решения.

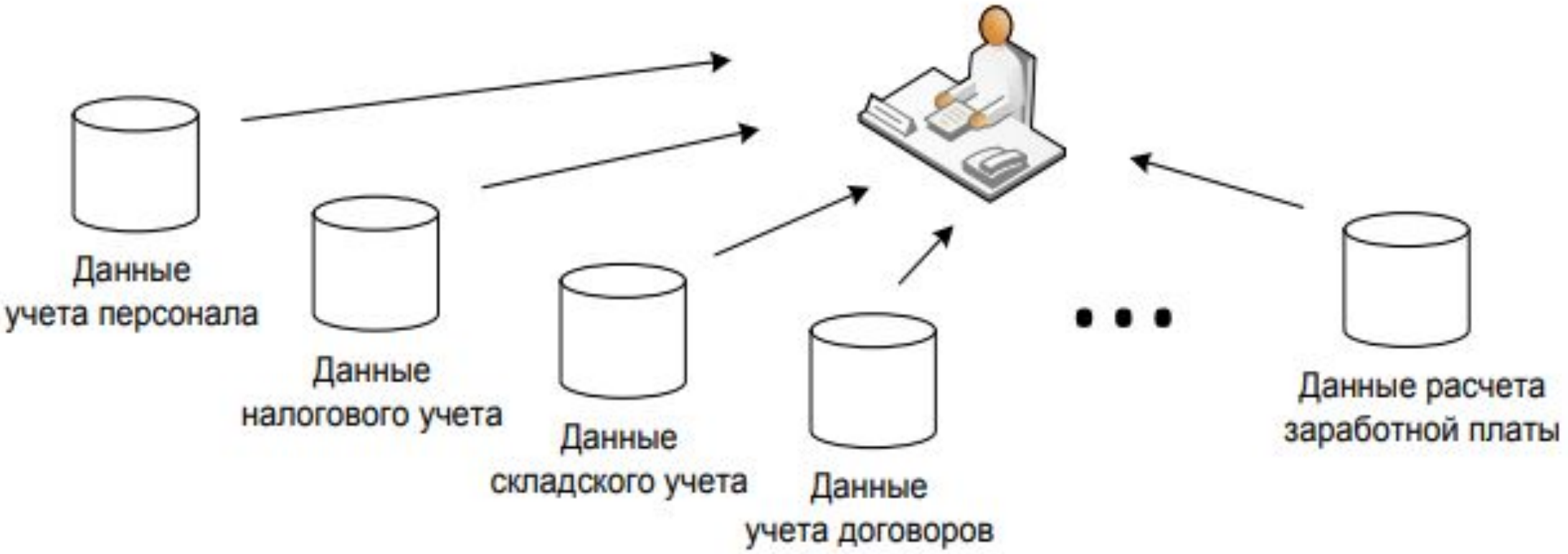


Рисунок 2 – Пример многообразия OLTP-систем в организации

Основная проблема многообразия источников состоит в *несогласованности* и *противоречивости* содержащихся в них данных, в *отсутствии единого логического взгляда* на корпоративные данные.

Успешная деятельность организации невозможна без *принятия обоснованных управленческих решений*.

Такие **решения** могут быть построены на основе *всестороннего анализа* результатов выполнения бизнес-процессов *в самой организации*, а также *многочисленных внешних факторов*.

Время принятия решений в современных условиях *сокращается*.

Роль информационных технологий, поддерживающих процессы бизнес-анализа и принятия управленческих решений, *возрастает*.

В результате наблюдается активное развитие особого класса информационных систем – **информационно-аналитических систем**, ориентированных на *оперативную аналитическую обработку данных*, извлекаемых из множества источников данных как внутри, так и вне организации, предназначенных для помощи управляющему персоналу организации в принятии обоснованных своевременных решений.

Рисунок 3 – Базовые технологии информационно-аналитической системы

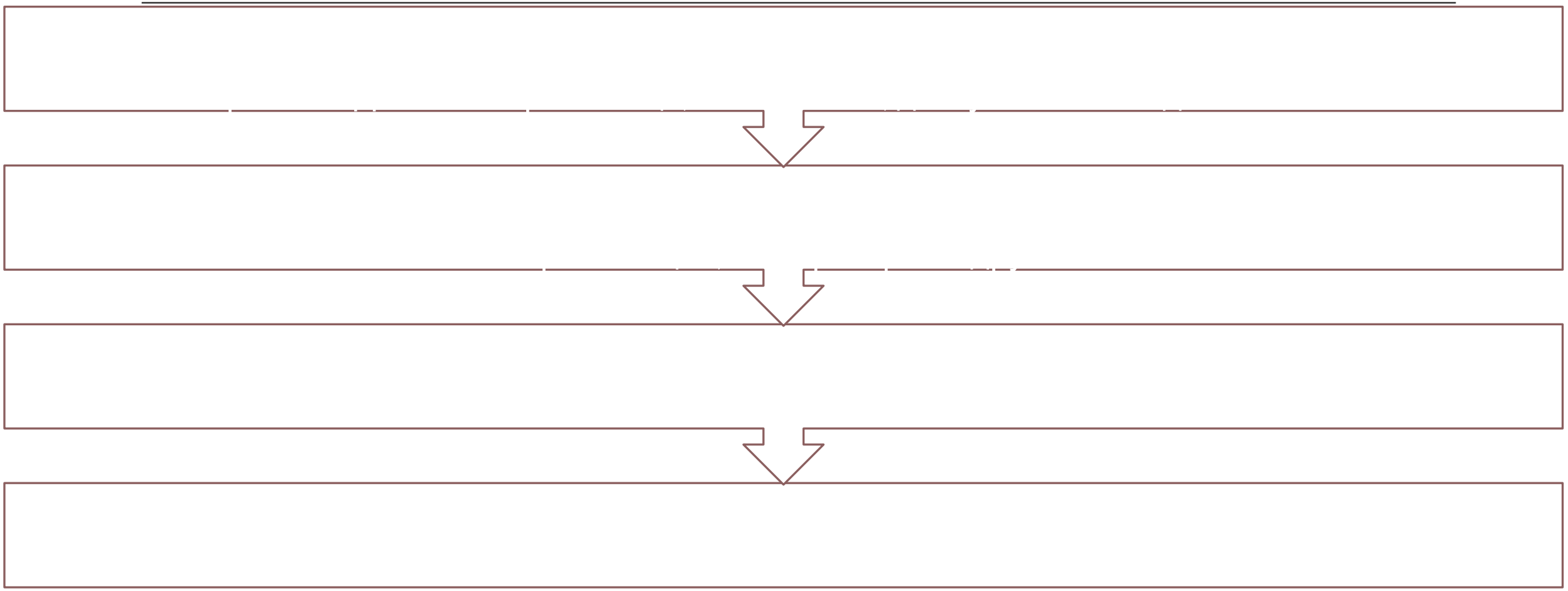
Информационно-аналитическая система базируется на нескольких информационных технологиях.

Как правило информационно-аналитическая система сочетает:

- технологию хранилищ данных,
- технологию оперативного анализа данных,
- технологию интеллектуального анализа данных,
- современные технологии визуализации

Рисунок 4 – Укрупненная структура аналитической системы

Упрощенно структуру информационно-аналитической системы можно представить следующим образом:



1.3. Модели данных хранилищ данных

Модель данных хранилища может быть представлена на трех уровнях:

Концептуальная модель

- является отражением предметной области, в рамках которой планируется построение хранилища данных, а также спектра аналитических задач, для решения которых разрабатывается хранилище.

Логическая модель

- расширяет концептуальную, уточняя состав таблиц и взаимосвязи между ними, добавляя колонки, а также определения для таблиц и колонок.

Физическая модель

- описывает реализацию объектов логической модели на уровне объектов конкретной базы данных.

Концептуальная модель данных хранилища

В основе **концептуальной модели хранилища** лежит многомерная модель данных. Основными понятиями многомерной модели данных являются: **куб** (гиперкуб) данных, **факты, показатели** (меры), **измерения, иерархии, агрегаты, срез**.

Поскольку многомерная модель данных предназначена для предоставления *BI-информации*, все показатели, измерения, иерархии должны иметь названия, доступные для понимания лицам, принимающим решения в организации.

Кроме того, модель предусматривает возможность сопровождать названия объектов развернутыми описаниями.

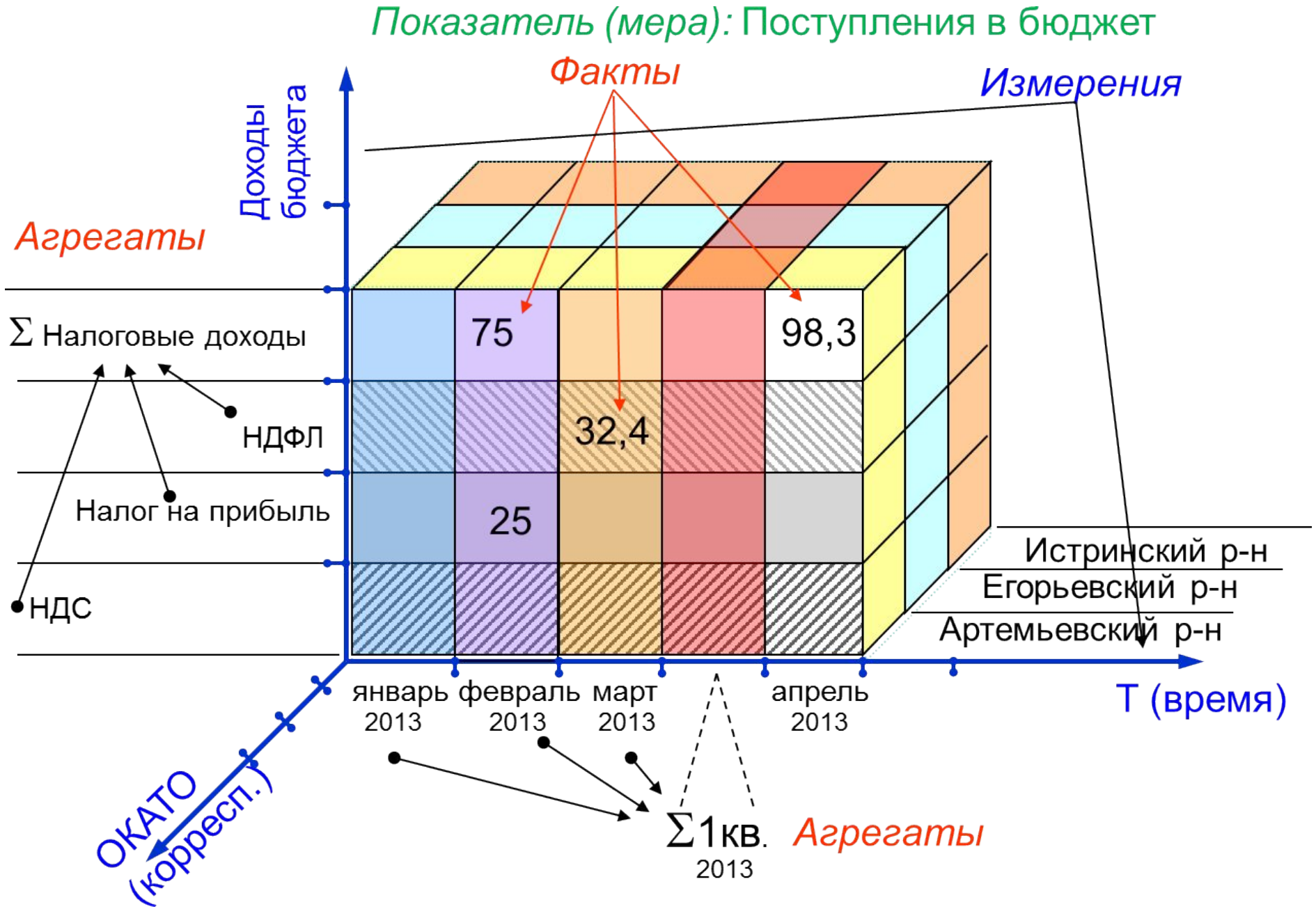


Рисунок 5 – Пример куба данных для анализа поступлений в бюджет

Каковы были поступления в бюджет по налогу на прибыль от корреспондентов Артемьевского района за февраль 2013 г.?

Каковы были поступления в бюджет по налогу на прибыль и НДС от корреспондентов Артемьевского и Егорьевского районов за январь и февраль 2013 г.?

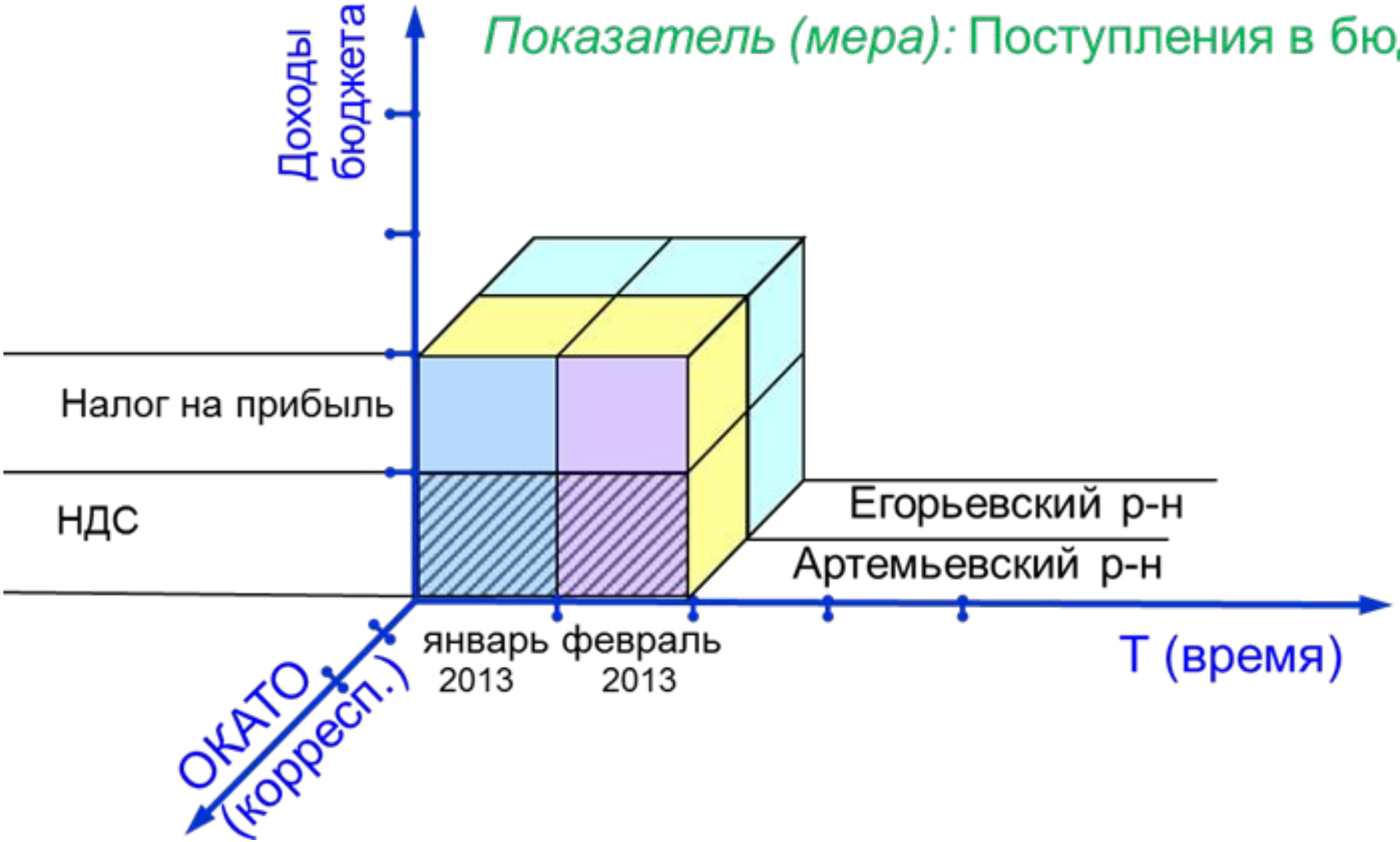
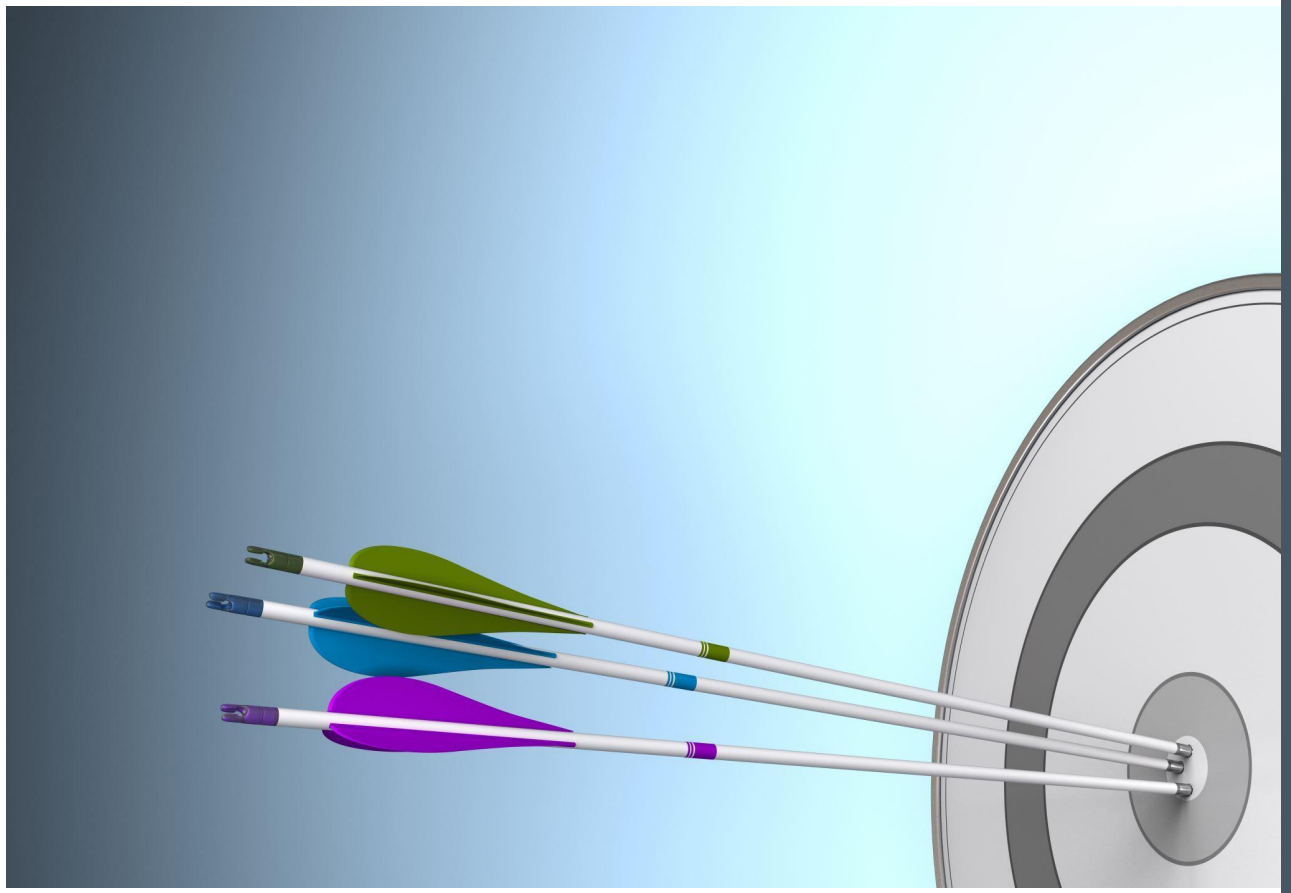


Рисунок 6 – Пример среза куба данных



Многомерная модель позволяет делать срезы куба данных и поворачивать их нужной гранью *любым удобным* аналитику образом.

Таким образом, используя **концептуальную модель хранилища**, аналитик и/или лицо, принимающее решения, могут *оперативно самостоятельно* сформировать запрос в соответствии с текущей аналитической задачей.

Логическая модель данных хранилища

Концептуальная модель хранилища (*гиперкуб*) на логическом уровне представляется с помощью **схемы «Звезда»** или ее варианта - **схемы «Снежинка».**

Схема «Звезда» имеет одну **таблицу фактов** и несколько **таблиц измерений**. Таблицы измерений являются **денормализованными**.

Схема «Снежинка» имеет одну таблицу фактов и несколько нормализованных таблиц измерений.

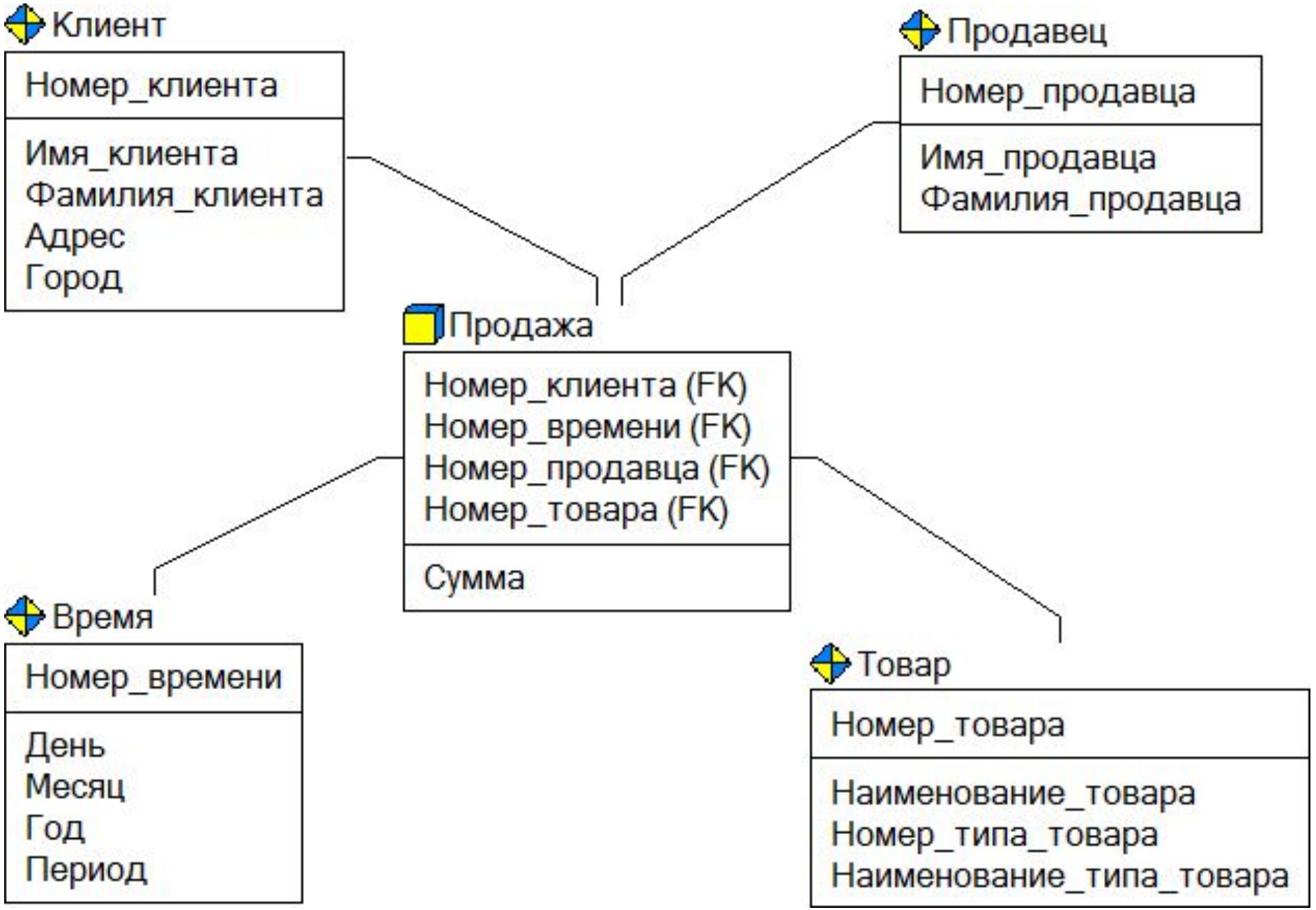


Рисунок 7 – Пример схемы «Звезда» для анализа продаж

Невозможно создать *универсальную денормализованную структуру* данных, обеспечивающую высокую производительность при выполнении любого аналитического запроса. Поэтому схема «Звезда» строится так, чтобы *обеспечить наивысшую производительность* при выполнении **одного** самого важного **запроса**, либо для группы похожих запросов.

Так, хранилище, разработанное на основе схемы на рис. 7, позволит оперативной ответить на вопросы, связанные продажами в компании.

Таблица фактов является *центральной таблицей* в схеме «Звезда». Таблицу фактов окружают меньшие таблицы, называемые **таблицами размерности** или **измерений**. Таблицы измерений содержат *неизменяемые либо редко изменяемые данные* типа справочника.

В рассмотренном примере на рис. 7 предполагается, что записи в таблицах «Клиент», «Продавец», «Время», «Товар» не изменяются или изменяются редко.

Таблицы измерений соединены с *таблицей фактов* в виде **звезды** радиальными связями.

В этих связях *таблицы размерности* являются **родительскими**, *таблица факта* – **дочерней**.

Таблица фактов и таблицы размерности связаны **идентифицирующими связями**, при этом первичные ключи таблицы размерности мигрируют в таблицу фактов в качестве внешних ключей, как показано на рис. 7.

В ~~краткосрочных пилотных проектах для~~ сокращения времени раз работки хранилища иногда используют **схему «Снежинка»** с ее *нормализованными таблицами измерений*, так как она ближе по структуре к источникам – нормализованным базам данных OLTP-систем.

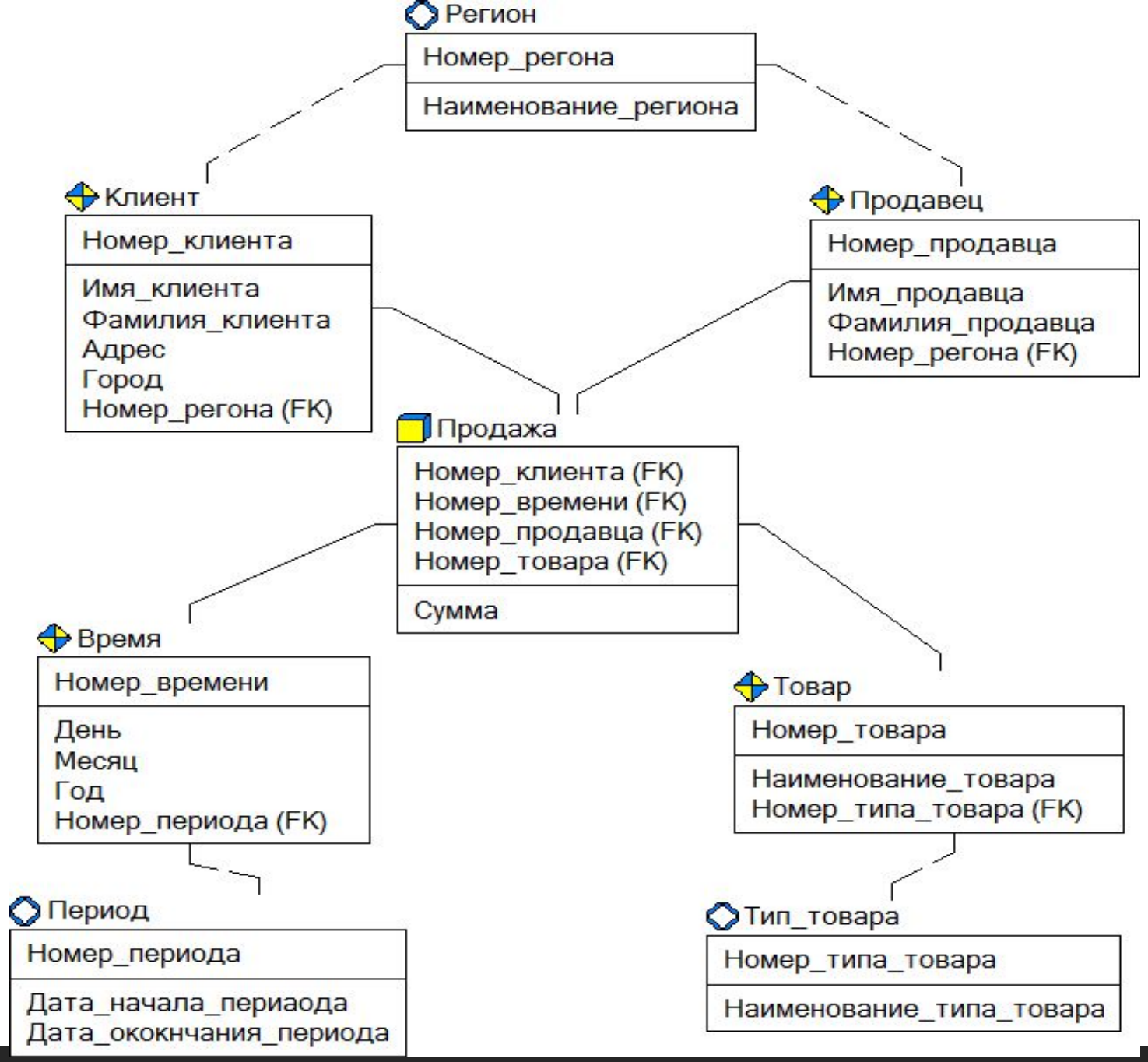


Рисунок 8 – Пример схемы «Снежинка» для анализа продаж

Дополнительные таблицы измерений в схеме «Снежинка», обычно соответствуют верхним уровням **иерархии измерения** и находятся в соотношении «один ко многим» с главной таблицей измерений, соответствующей нижнему уровню иерархии.

Эти дополнительные таблицы иногда называют **консольными**. На схеме, представленной на рис. 8, таблицы «Регион», «Период», «Тип товара» являются консольными.

Консольные таблицы могут быть связаны только с *таблицами измерений*, причем консольная таблица в этой связи родительская, а таблица измерений – дочерняя.

Связь может быть **идентифицирующей** или **неидентифицирующей**.

Консольная таблица не может быть связана с таблицей факта.

Она используется **для нормализации данных** в таблицах измерений.

Например, в результате анализа данных по видам запросов к хранилищу было обнаружено, что количество запросов о продажах с детализацией по наименованию товара в десять раз меньше, чем количество запросов о продажах по типам товаров.

В таких случаях изменяют схему «Звезда»: таблицу измерения разбивают на две отдельные таблицы, связав их неидентифицирующей связью, как показано на рис. 9.

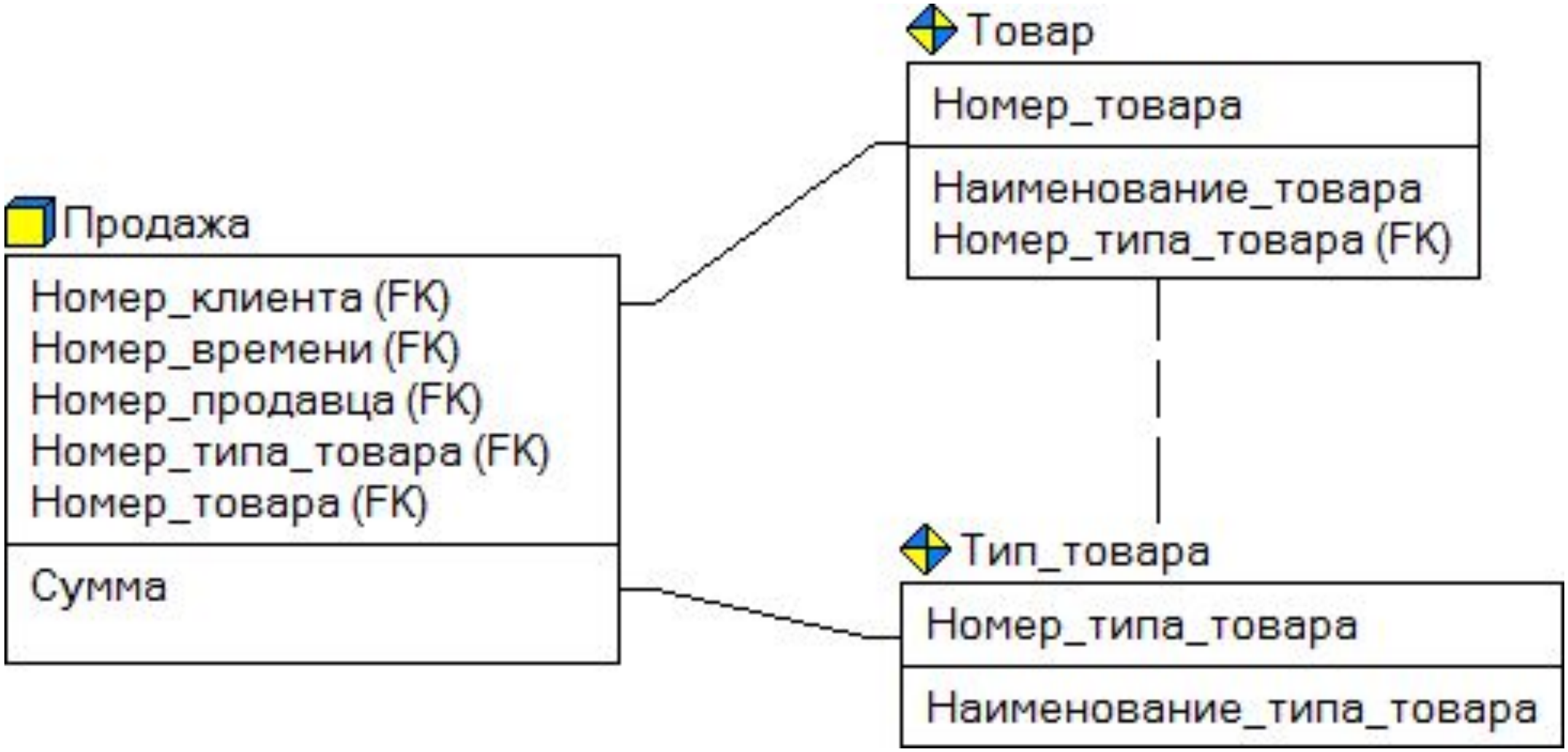
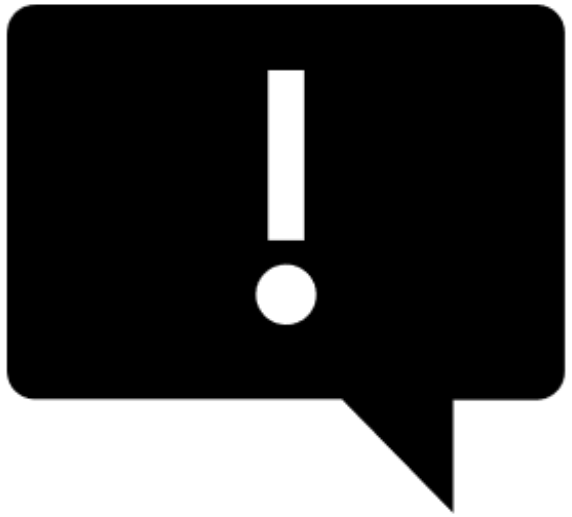


Рисунок 9 – Таблица измерения разбита на две связанные таблицы



Хранилище данных предприятия содержит *данные из разных сфер деятельности* и состоит из набора связных схем «Звезда» или «Снежинка».

Согласование разных схем производится через *общие таблицы измерений.*

Физическая модель данных хранилища

Физическая модель может быть представлена с помощью схем «*Звезда*» или «*Снежинка*», *дополненными описанием* реализации объектов логической модели на уровне объектов конкретной базы данных.

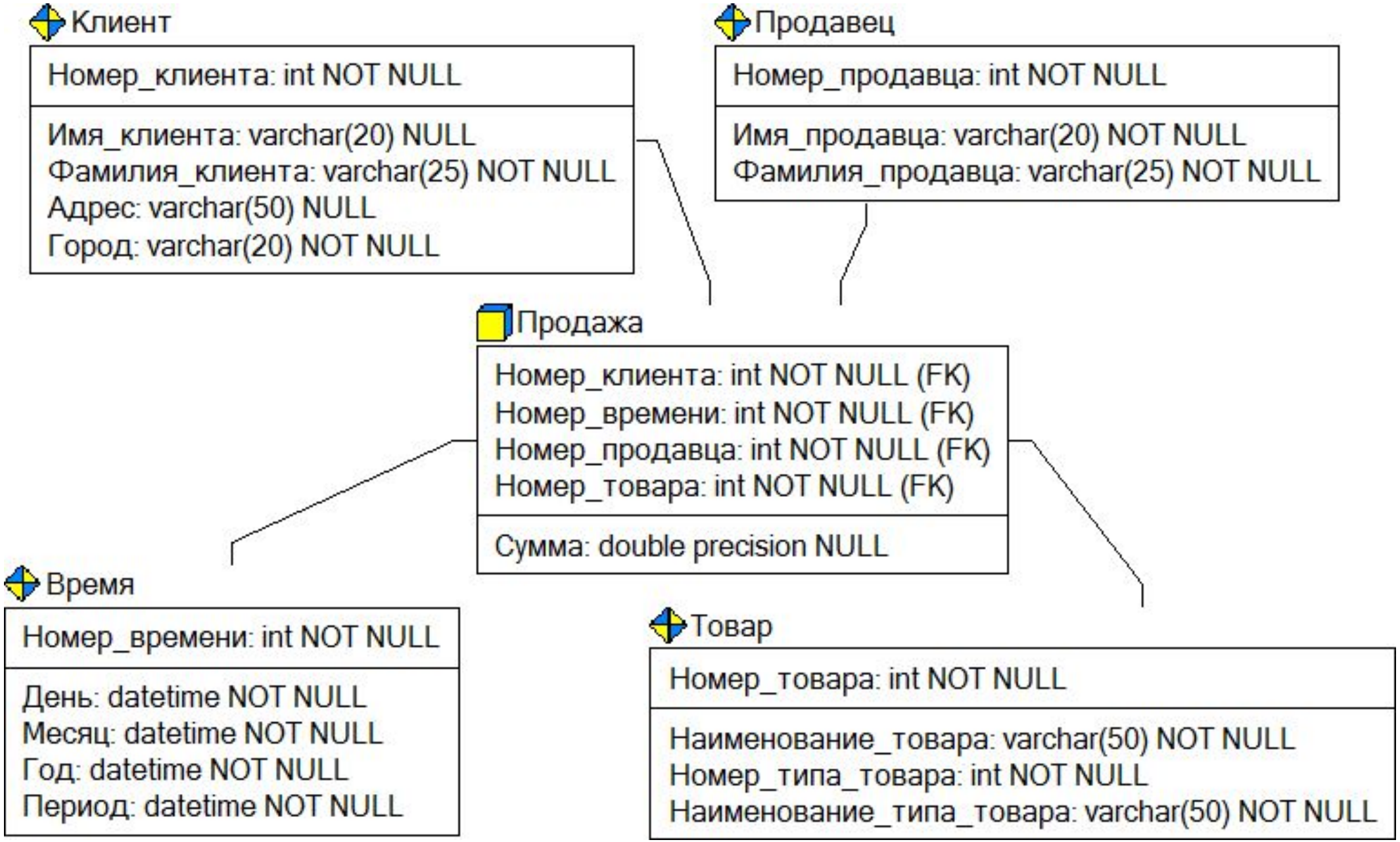
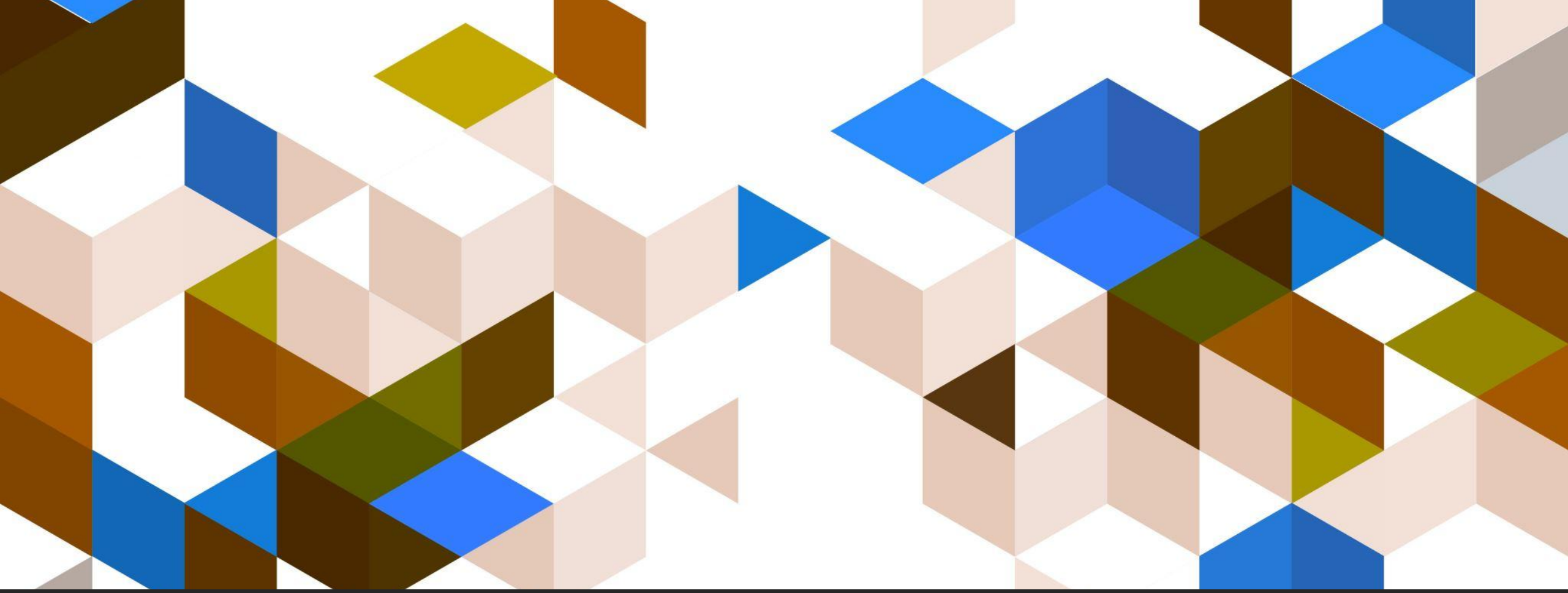


Рисунок 10 – Пример физической модели хранилища

На рис. 10 представлена **физическая модель хранилища**, соответствующая логической модели на рис. 7, отображающая *ряд дополнительных характеристик объектов базы данных* SQL Server 2012, а именно: тип данных колонок, а также возможность отсутствия значений колонок (опция Null/Not Null).



1.4. Сценарий функционирования хранилища данных

сценарий функционирования хранилища данных

выглядит следующим образом

Рисунок 11 – Сценарий функционирования хранилища данных

Извлечение, преобразование и загрузка данных

Данные поступают из различных внутренних транзакционных систем, от подчиненных структур, от внешних организаций в соответствии с установленным регламентом, формами и макетами отчетности.

Вся эта информация пр*оверяется, согласуется, преобразуется и помещается* в хранилище и витрины данных.

После этого пользователи с помощью специализированных инструментальных средств получают необходимую им информацию для построения различных *табличных и графических представлений, прогнозирования, моделирования и выполнения других аналитических задач.*

В *ходе наполнения хранилища* информация подвергается **многоступенчатой обработке**.

На первом этапе производится буферизация сведений, поступающих из разных источников информации, что позволяет стабилизировать процесс загрузки и выполнять сверку присланных данных. *Первую буферную зону называют шлюзом*.

На следующем этапе данные преобразуются и перемещаются из одной буферной зоны в другую. Здесь устраняется структурная неоднородность данных, проверяется их целостность, служебным полям хранилища присваиваются соответствующие значения, решаются задачи интеграции данных. На этом этапе производится также очистка информации. Сведения, не прошедшие процедуры преобразования, очистки или проверки, попадают в специальную зону, где их в полуавтоматическом режиме можно исправить, дополнить, после чего заново загрузить в хранилище.

Рисунок 12 – Загрузка данных в хранилища данных

Процесс **настройки хранилища данных** подразумевает выполнение следующих шагов: