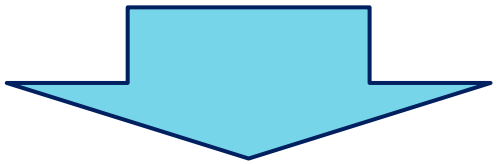
2. Интеграция данных

Проблема:

исходные данные – в источниках с различными моделями и представлением данных, в файлах различных форматов и стандартов кодирования.

При этом в различных источниках могут находиться различные элементы данных, которые должны быть проанализированы совместно.

Для применения методов и алгоритмов анализа данных необходимо, чтобы данные были в едином централизованном источнике и имели унифицированный формат представления.



Важнейший этап подготовки данных к анализу – их ***интеграция*** из множества разнородных источников в один источник, имеющий архитектуру, наиболее соответствующую целям анализа, а также применяемым в процессе анализа алгоритмам.

Задача:

представление совокупности данных из множества независимых источников
(фиксированного или пополняемого)
на основе единой модели данных.

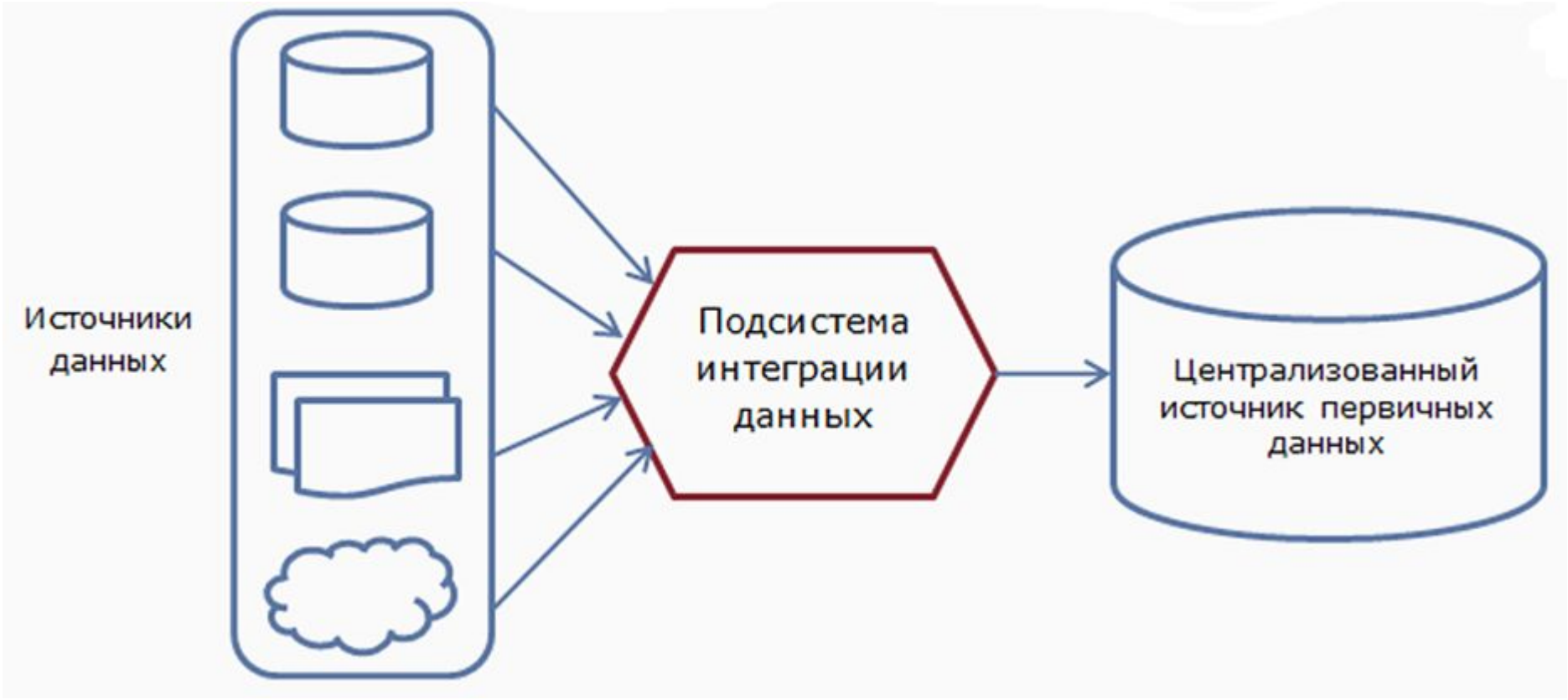
Интеграция данных

Интеграция данных – это процесс объединения данных, находящихся в различных разнородных источниках, в единственном физическом источнике

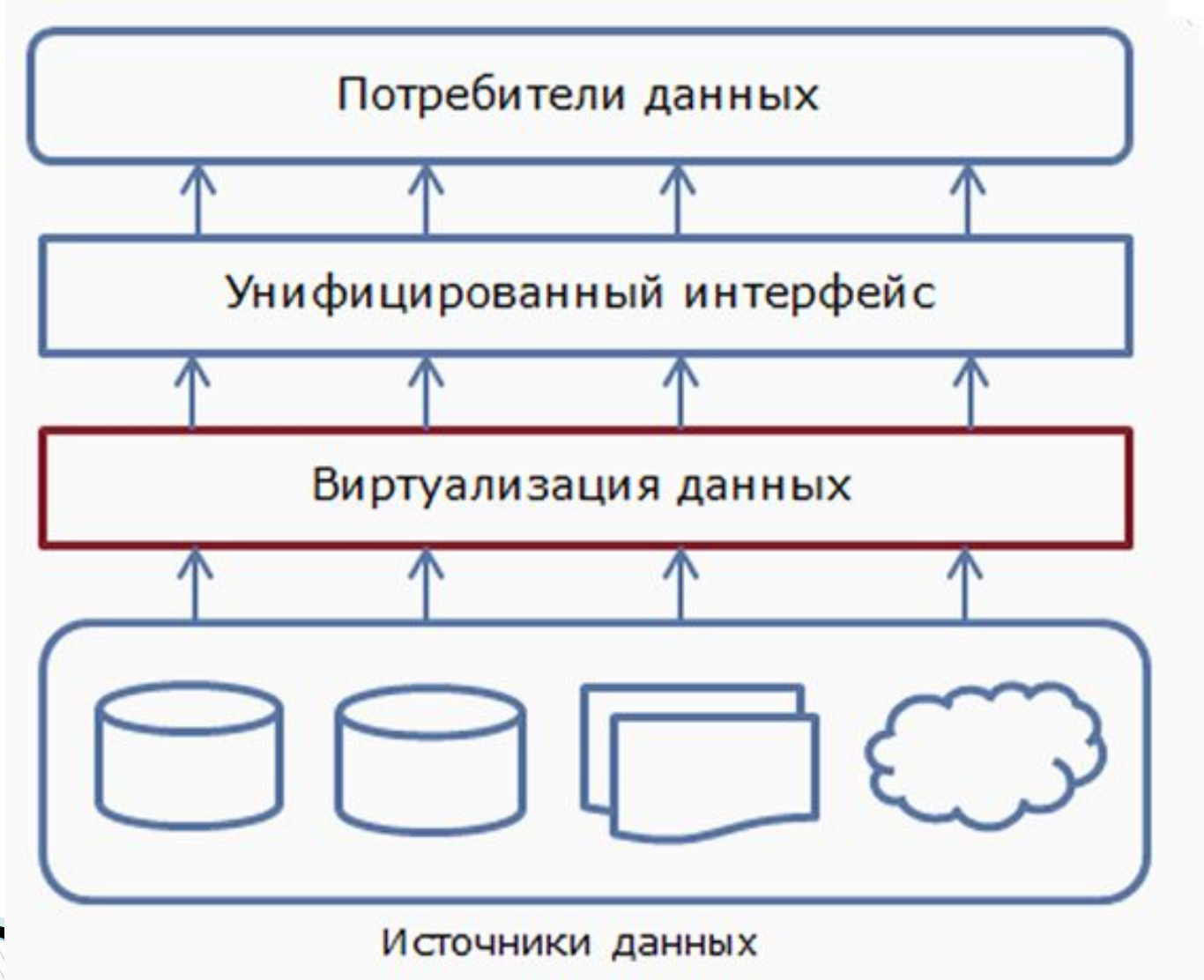
или

обеспечение единого унифицированного интерфейса для некоторой совокупности источников (без физического объединения данных путем копирования информации).

Интеграция данных в едином источнике



Интеграция данных без физического объединения



Источники данных

Под ***источником данных*** в контексте рассматриваемых технологий понимается объект, содержащий структурированные данные.

Замечание.

Если данные в источнике изначально не структурированы, то они должны допускать преобразование в структурированный табличный формат (пример – текстовые файлы, лог-файлы).

В противном случае объект не будет считаться источником данных.

Уровни интеграции данных

- ***Физический уровень***.

  Данные из различных источников преобразуются к единому формату и сохраняются в одном источнике.

- ***Логический уровень***.

  Данные физически размещаются в своих источниках, доступ к ним осуществляется на основе некоторой глобальной схемы, отражающей требуемое совместное представление.

- ***Семантический уровень***.

  Обеспечивается поддержка единого представления данных с учетом их семантических свойств в контексте единой онтологии предметной области.

Семантический подход к интеграции

Учитывает содержательную (контекстную) сторону данных.

Семантическая интеграция основывается на учете природы данных.

Например: можно хранить и анализировать информацию об объеме продаж, выраженном в единицах продукции (штуках, килограммах и т. п.) и в денежном выражении, если добавить дополнительные сведения, связывающие единицы продукции с денежными суммами.

Семантический слой

Такие дополнительные сведения называются ***метаданными*** («данными о данных»).

С помощью метаданных формируется ***семантический слой***, который делает работу с данными более понятной и прозрачной для пользователя.

Способы интеграции данных

- ***Виртуальный***.

  Реализуется с помощью механизма доступа: при выполнении запроса пользователя формируется требуемое представление данных непосредственно из источников.

  Наиболее эффективен, если источники являются динамически обновляемыми.

- ***Материализованный (актуальный)***.

  Формируется полное физическое представление данных, сосуществующее с источниками, на основе которых получено.

  Используется в хранилищах и оперативных складах данных, когда они существуют вместе с источниками.

Быстрые и медленные данные

Данные, накапливаемые в ИС предприятий

Быстрые

Поступают непрерывно; являются сильно детализированными (отражают элементарные события в жизни бизнеса).

Отражают текущие тенденции.

Позволяют принимать оперативные, тактические решения

Медленные

Отражают долгосрочные зависимости и закономерности бизнес-процессов, что позволяет использовать их для решения задач стратегического анализа и поддержки принятия решений

В бизнес-аналитике могут использоваться как быстрые, так и медленные данные.

- При использовании быстрых данных задачи бизнес-аналитики сводятся к построению отчетов, отражающих текущее положение в компании и ее подразделениях, что позволяет вырабатывать тактические решения (подвезти дополнительный товар, создать запас на складе, и т. п.).
- Медленные данные являются историческими и хронологическими подходят для задач описательных и предсказательных моделей *Data Mining* с целью выявления длительных зависимостей и закономерностей, знание которых позволит принимать стратегические решения (смена ассортимента товаров и услуг, изменение ценовой политики и т. п.).

Разные цели и задачи работы с быстрыми и медленными данными

формирование двух классов ИС:

системы оперативного анализа (OLTP-системы);

системы поддержки принятия решений (СППР).

Системы оперативного анализа

OLTP (*On-Line Transaction Processing*) – оперативная (т. е. в режиме реального времени) обработка транзакций.

Под ***транзакцией*** понимается некоторый набор логически связанных операций над базой данных, который рассматривается как единое, завершенное, с точки зрения бизнес-логики, действие над некоторой информацией, связанное с выполнением определенной бизнес-функции (продажа набора товаров по одному чеку, выдача/прием наличных через банковский терминал, оплата услуг телекоммуникационных компаний и т. д.).

Обобщенная схема движения данных в транзакционной системе:

Исключение любой из операций приведет к тому, что связанная с транзакцией бизнес-функция не будет выполнена.

Пример.

В системе бронирования авиабилетов:

- запрос оператора о наличии свободных мест на рейс;
- ответ системы с предоставлением соответствующей информации;
- ввод оператором информации о пассажире, номере заказанного места, оплаченной сумме и т. п.
- передача новой информации в БД и внесение в нее соответствующих изменений;
- передача оператору подтверждения успешного выполнения операции.

Транзакции в СМО выполняются десятки и сотни тысяч раз в день в огромном количестве терминалов, пунктов продаж и т. п. данные в OLTP-системах меняются непрерывно, но небольшими порциями.

Основное требование:

приемлемое время отклика при максимальной загрузке системы.

Особенности OLTP-систем:

- запросы и отчеты являются строго регламентированными (т. к. цель – точное выполнение бизнес-функции); оператор не может создать собственный запрос с целью получения сведений, не предусмотренных регламентом;
- данные быстро теряют актуальность и устаревают (после завершения рейса информация о пассажирах теряет актуальность с точки зрения бизнес-функции бронирования билетов) историчность данных не поддерживается.

СППР

В процессе эксплуатации OLTP-систем – понимание: если транзакционные данные, наколенные в OLTP-системах, не уничтожать после утраты ими бизнес-актуальности, а сохранять, то эти исторические бизнес-данные могут быть использованы для анализа и поиска в них скрытых знаний о бизнес-процессах.

Это отправная точка в появлении и развитии бизнес-аналитики.

Требование к СППР:

возможность получения ответов на нерегламентированные запросы аналитика.

Например: как изменялась динамика визитов покупателей в течение рабочего дня?

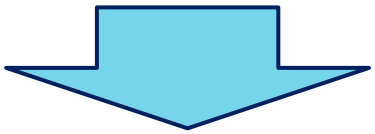
На основе ответов – принятие управленческих решений, направленных на оптимизацию работы гипермаркета (открытие дополнительных касс во время пиковой нагрузки и т. п.).

Такие сложные, нерегламентированные запросы к транзакционным базам данных называются аналитическими (позволяют делать выводы и заключения и использовать их для поддержки принятия решений)

Вопрос:

оптимальна ли OLTP-система, ориентированная на максимально быстрое выполнение простейших запросов, с точки зрения реализации аналитических запросов?

Ответ отрицательный.



В задачах бизнес-аналитики OLTP-системы имеет смысл рассматривать как системы сбора первичных данных, а для хранения данных, подвергаемых анализу, организовать другую систему.


Информационная СППР
Извлечение, преобразование, интеграция
Транзакционная база данных
Хранилище данных (исторические, справочные, метаданные)
Подсистема анализа и моделирования
Управленческие решения
Модели, прогнозы, отчеты

Требования к организации хранения данных в СППР:

- для выполнения нерегламентированных запросов необходима обработка массивов данных из множества разнородных источников;
- для выполнения запросов, связанных с анализом тенденций, прогнозированием протяженных во времени процессов, необходимы исторические данные, накопленные за достаточно длительный период;
- максимальная детализация транзакционных данных (в OLTP-системах) не оптимальна для целей анализа; требуется выполнение агрегации данных на разных уровнях (день – месяц – год, торговая точка – населенный пункт – регион и т. п.) и разного вида (сумма, среднее, минимум, максимум, медиана).

Прежде, чем подвергаться анализу, данные должны пройти подготовку, включающую

- извлечение и объединение данных из множества разнородных источников в централизованную систему хранения;
- преобразование – приведение данных к наиболее удобному для анализа виду (агрегирование, кодирование, квантование и т. п.);
- профайлинг и аудит с последующей очисткой (восстановление полноты и целостности данных – исключение пропусков, дубликатов, противоречий и т. д.).

Современная тенденция – комплексное решение задач подготовки данных в системах бизнес-аналитики.

В частности, аудит, очистка и преобразование часто рассматриваются как часть процесса интеграции данных.

Соответствующий функционал включается в интеграционные системы.

Типы корпоративных данных

- ***Фактографические*** – это данные, отражающие факты, которые описывают процессы, объекты и явления предметной области.

Факт (наблюдение, прецедент, транзакция) представляет собой отдельную запись в БД.

Обычно факт отражает некоторое логически неделимое действие, последовательность которых образует бизнес-процесс в компании.

Примеры: истории продаж, транзакции,
данные биллинга и т. п.

Как правило, фактографические данные являются структурированными, историческими (накапливаются за определенный период) и хронологическими.

Большинство источников, отражающих деятельность компании, используемых для анализа, является фактографическими

- ***Нормативно-справочные данные*** включают различные словари (например, терминологические), справочники (адресов, телефонов и т. п.), классификаторы (ОКПО, ОКАТО), нормативы, кодификаторы и т. п.

  С точки зрения анализа такие данные носят вспомогательный характер (непосредственное построение моделей с их помощью невозможно).

  Могут использоваться для поддержки процесса анализа (обогащения, восстановления и очистки фактографических данных), формирования отчетов.

  Например: неизвестное название города можно восстановить известному по телефонному коду.

- ***Метаданные*** («данные о данных») – разновидность данных, носящих служебный характер.

  Они не отражают течение бизнес-процессов компании, а описывают фактографические и нормативно-справочные данные.

  Содержат информацию о составе данных (например, число страниц документа), содержании (оглавление), статусе, происхождении, форматах представления, условиях доступа и т. д.

Виды метаданных:

- ***технические*** – обеспечивают функционирование баз данных и выполнение запросов к ним;

Примеры: имена таблиц БД и полей в них

- ***бизнес-метаданные*** определяют сущности, хранящиеся в источниках данных, бизнес-термины и определения;

Благодаря им пользователь может оперировать привычными терминами предметной области, которые транслируются в соответствующие запросы

- ***операционные*** – содержат информацию о процессе работы источника данных: происхождение загруженных и преобразованных данных, их статус (активные, архивированные, удаленные), статистику использования и т. д.

Первичные источники данных

Первичными называются источники, данные в которых являются результатом непосредственной регистрации и измерения характеристик бизнес-процессов, объектов или явлений.

Первичные данные могут формироваться как посредством ручного ввода (оператором), так и автоматически (например, с использованием сканера штрих-кода).

Первичные источники, как правило, не используются для аналитической обработки и целей Data Mining.

Они являются источниками данных для ***вторичных*** источников.

Виды первичных источников данных.

- *Транзакционные системы и БД*.
- *Унаследованные системы*.

  Системы, реализованные на основе устаревших технологий, содержащие данные за несколько десятилетий.
- *Облачные источники*.

  Организация хранения данных с использованием ресурсов сторонних организаций.

  Пользователь не видит внутреннюю структуру системы.
- *Файлы, документы*.

  Файлы данных отдельных пользователей.

Вторичные источники данных

Вторичными называются источники, которые получают данные не в процессе их сбора или регистрации, а из первичных источников.

Перенос данных из первичных источников во вторичные (или из одних вторичных источников в другие) реализуется с помощью программно-аппаратного комплекса, называемого ***ETL*** (*Extraction*, *Transformation*, *Loading* – извлечение, преобразование, загрузка).

Виды вторичных источников данных.

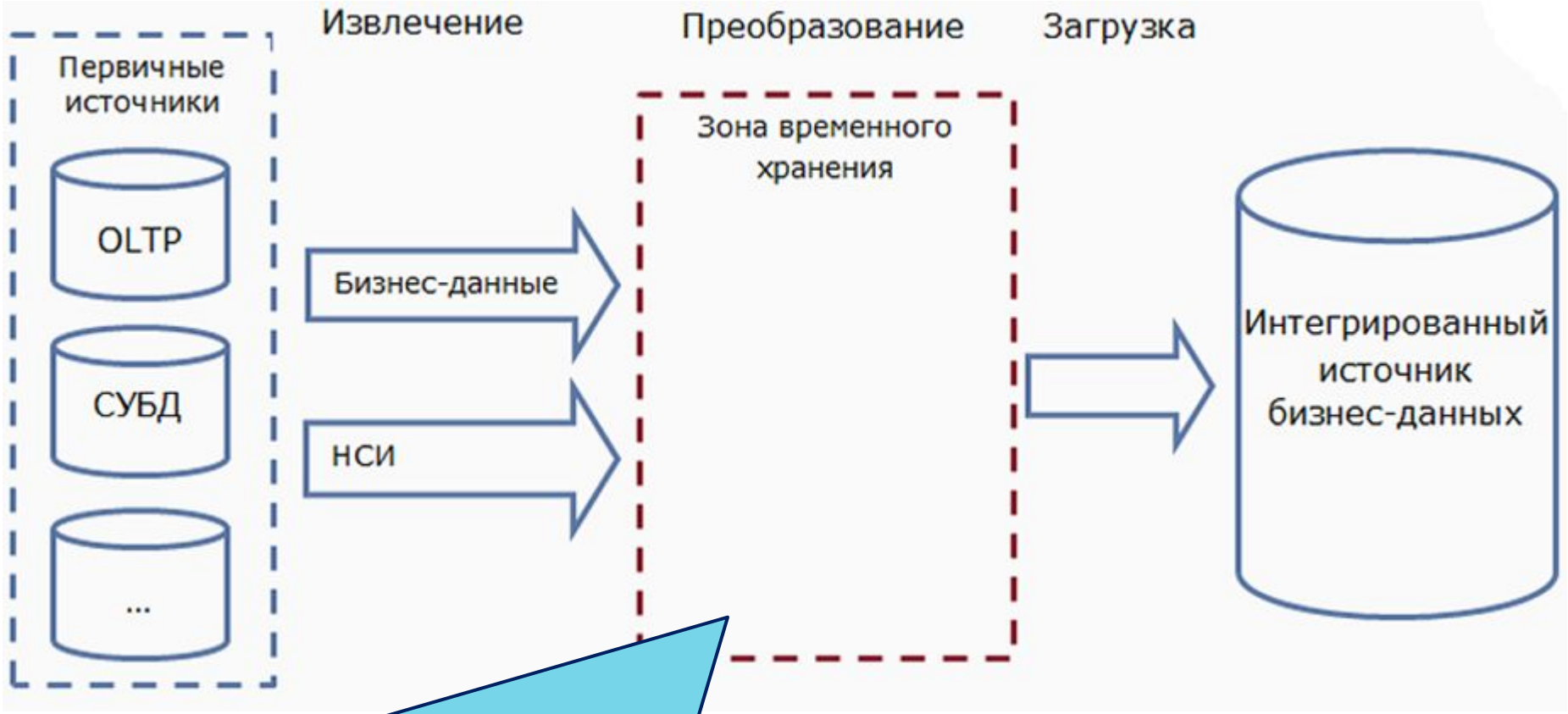
- *Область временного хранения*.
  Используется для промежуточной обработки (очистки, трансформации) и синхронизации данных из разных источников.
- *Оперативный склад данных*.
  БД, в которой хранятся данные реального (или почти реального) времени, используемые для оперативного (тактического) анализа.
- *Хранилище данных*.
- *Витрина данных*.

Процессы ETL

ETL – комплекс программно-аппаратных средств, реализующий

- извлечение данных из различных источников;
- преобразование их к единому формату и модели данных;
- очистку от дубликатов, противоречий и др. факторов, которые могут помешать их анализу;
- загрузку в единый интегрированный источник.

Вариант структуры процессов ETL



Очистка,
приведение к единому представлению (форматам, кодировкам),
преобразование (агрегирование, нормализация, кодирование)
для наибольшего соответствия целям аналитической обработки

«Тонкое место» – преобразование данных в зоне временного хранения:

- вычислительные затраты при увеличении объема извлекаемых данных могут возрастать нелинейно;
- если часть данных значительно опаздывает или временно недоступна, то данные, которые должны по регламенту поступать в интегрированный источник одновременно с этой частью, будут ожидать в зоне временного хранения неопределенно долго (следствием может быть потеря синхронизации хранилища с актуальным состоянием первичных источников).

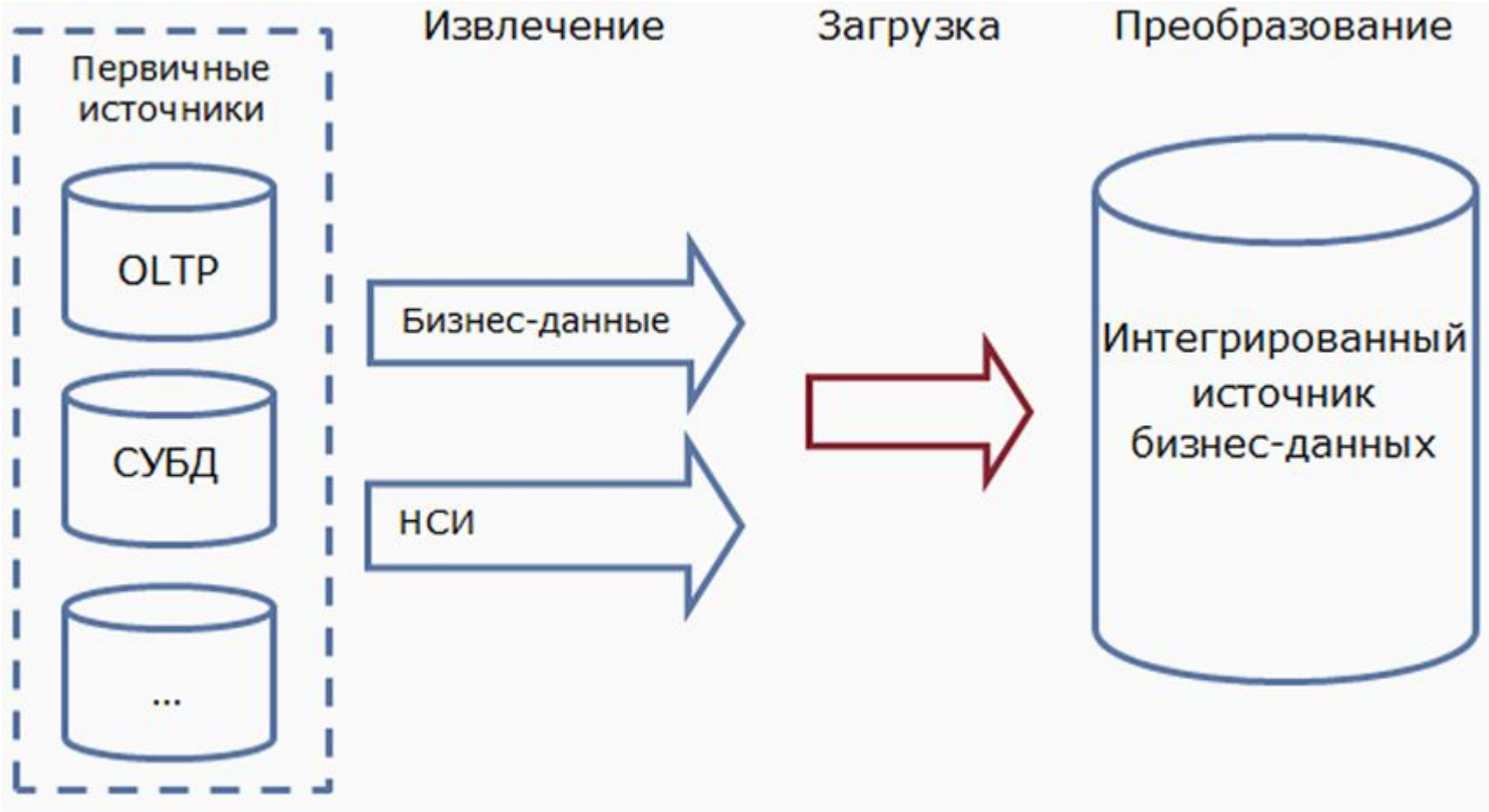
Формула ELT

Решением проблемы может быть перенос большей части *ETL*-операций из зоны временного хранения в интегрированный источник:

данные сначала извлекаются, затем загружаются, и только в интегрированном источнике подвергаются очистке и трансформации (до загрузки может выполняться агрегирование).

Т. е. вместо формулы *ETL*
(извлечение, преобразование, загрузка)
используется формула ***ELT***
(извлечение, загрузка, преобразование).

Вариант структуры процессов ELT



Преимущества такого решения:

- сокращение времени между извлечением данных и загрузкой в ХД;
- данные поступают в ХД, даже если часть их потеряна или запаздывает (это снижает вероятность потери синхронизации с актуальным состоянием источников);
- очистка данных в ХД более эффективна, т. к. очищаются данные, уже подвергшиеся интеграции (в процессе которой могут появляться дубликаты и противоречия).

Наибольшие преимущества ELT перед ETL – в ИС компаний, имеющих значительную территориальную распределенность

Качество данных

Качество данных – ключевой фактор обеспечения корректных и практически полезных результатов.

При этом:

факторы, вызывающие низкое качество данных, не всегда очевидны;

еще менее очевидны проблемы, вызванные низким качеством данных при их анализе.

В каждой компании могут складываться свои требования к качеству данных и причины снижения их качества.

Факторы качества данных

- ***Полнота данных*** – все данные, связанные с некоторым бизнес-процессом, собраны и представлены в полном объеме (отсутствуют неполные столбцы и записи, фрагменты таблиц и т. п.).
- ***Точность*** – представленные значения соответствуют реальным показателям (представлена цена товара, соответствующая действительной), а также не содержат ошибок (пример ошибки – в написании ФИО клиента).

- ***Непротиворечивость*** – связана с пониманием данных.

  Например: таблицы, содержащие данные о различных бизнес-объектах, могут иметь одинаковые имена.

  Или: валюта может быть указана в одном источнике текстовым значение (руб., долл.), в другом – символом (₽, $), в третьем – аббревиатурой (RUR, USD), что может привести к неправильной интерпретации данных.

- ***Актуальность*** данных связана с их своевременным обновлением (может измениться курс валюты, ставка рефинансирования, стоимость материалов и т. д.).

Наиболее эффективный подход к обеспечению качества данных –

организация потока соответствующих работ параллельно основному потоку, связанному с обеспечением функциональности системы.

Уровни очистки данных

- В процессе *ETL* – как элемент подготовки данных к интеграции.
- В консолидированных источниках – как элемент подготовки данных к анализу.
- В бизнес-приложениях – как элемент адаптации к *Data Mining*.

Уровни очистки данных

В контексте алгоритмов и методов моделирования, используемых для решения конкретной задачи

3. Аудит и профайлинг данных (основные понятия)

Аудит качества данных

Процедуры, связанные с обеспечением качества данных, выполняются в рамках специального бизнес-процесса, называемого ***аудитом качества данных***.

Может включать решение различных задач в зависимости от специфики деятельности

Аудит качества данных в контексте *Data Mining* включает

- ***профайлинг данных***;
- оценку влияния выявленных проблем на результаты предполагаемого анализа.

Профайлинг данных

Профайлинг данных – процесс изучения данных с целью понимания их структуры, содержимого и оценки качества.

Аудит данных – это процесс масштаба предприятия (работает с корпоративными данными в целом);

профайлинг реализуется на уровне отдельных источников данных и пользователей в зависимости от конкретных задач анализа.

Профайлинг данных

Обычно включает в себя следующие этапы:

- *структурный анализ* – проверку целостности таблиц данных (может быть нарушена после интегрирования данных из различных источников);
- *анализ контента* – обнаружение пропусков и выбросов данных, дубликатов, противоречий, ошибок, разбор составных значений (ФИО и т. п.);
- *формирование отчетов* – представление результатов в наиболее доступной для интерпретации форме, рекомендации по очистке и предобработке данных.

В *Data Mining* наиболее востребованы следующие задачи:

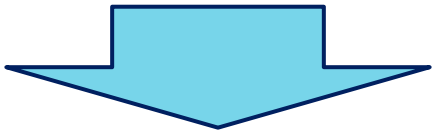
- общая статистика по выборке;
- обнаружение пропусков;
- обнаружение выбросов и экстремальных значений;
- обнаружение дубликатов и противоречий;
- сложные проверки (взаимные проверки между полями и проверки по бизнес-правилам).

Наиболее предпочтительный подход к организации процесса:

- обработка наиболее очевидных проблем перед загрузкой данных в единый интегрированный источник;

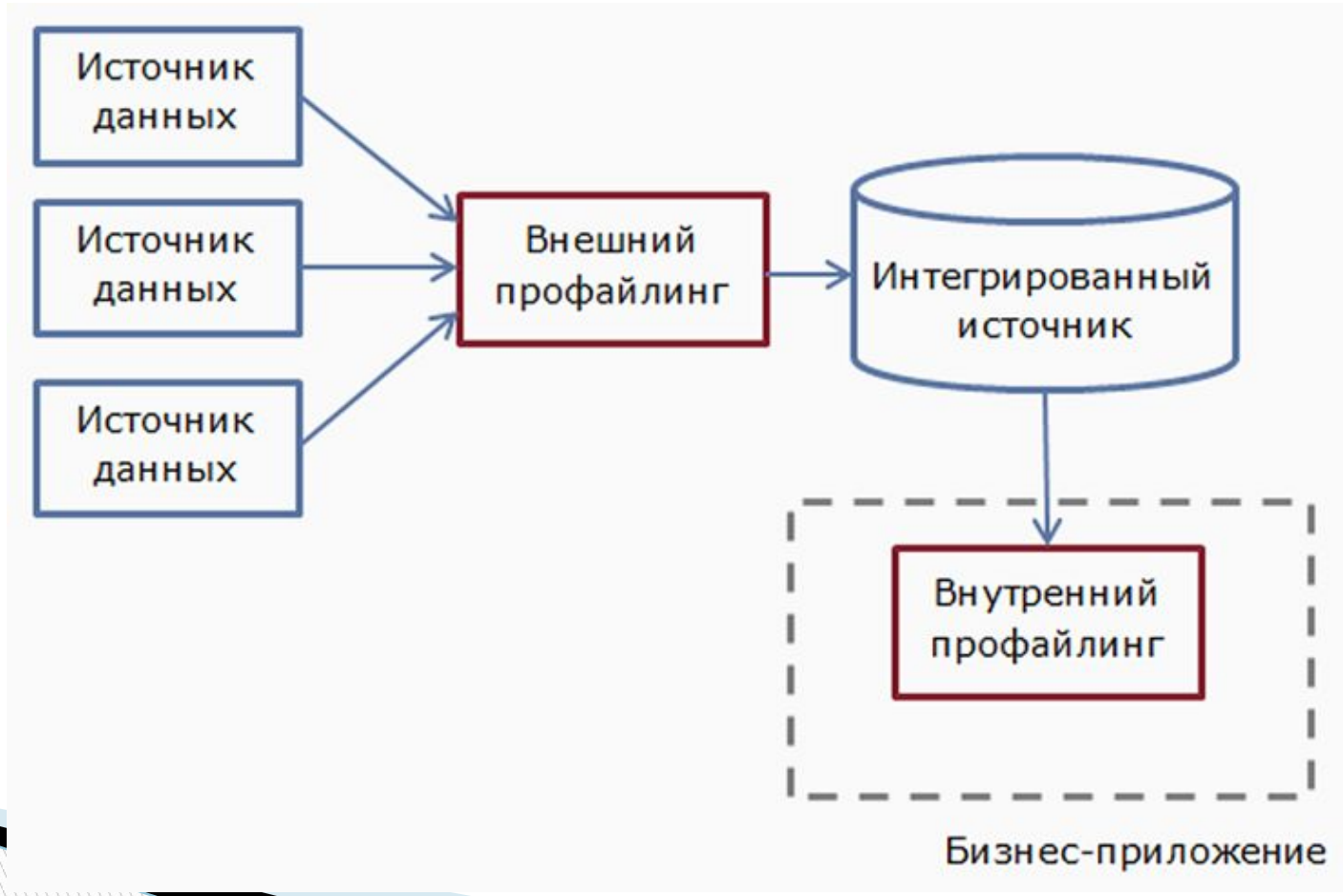
Примеры:
трех уникальных значений в признаке Пол;
несоответствие типа полей источника характеру данных;
некорректный формат даты или отображения дробных чисел и т. п.

- решение остальных проблем в контексте конкретной модели.



В соответствии с этим – два вида профайлинга:

- ***внешний*** (выполняется до загрузки в единый интегрированный источник);
- ***внутренний*** (выполняется в бизнес-приложении).



Отчет по результатам профайлинга данных

Формируется для того, чтобы пользователь мог разработать правильную стратегию предварительной обработки данных.

Такой отчет обычно содержит следующие виды информации:

- характеристики отдельных полей;
- характеристика набора данных в целом;
- резюме относительно возможности использования каждого поля для аналитической обработки.

Отчет по результатам профайлинга данных

Отчет по статистике

Обычно включает следующие характеристики:

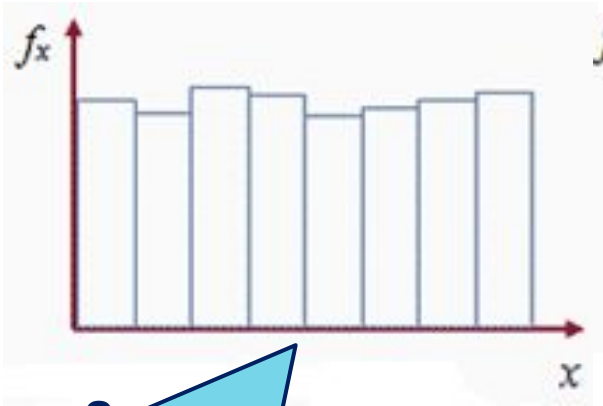
- минимальное и максимальное значения;
- среднее;
- медиана;
- среднеквадратичное отклонение;
- сумма и квадратичная сумма значений поля.

Позволяют обосновать выбор метода предварительной обработки.
Например: значительное различие между средней и медианой свидетельствует о наличии выбросов замена пропусков средним значением – плохое решение

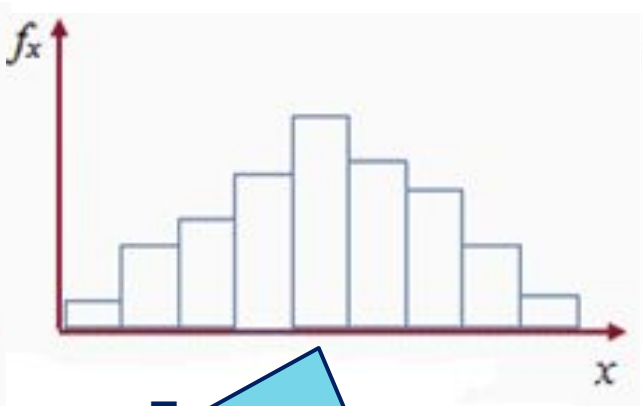
Отчет по статистике

Гистограмма распределения значений поля

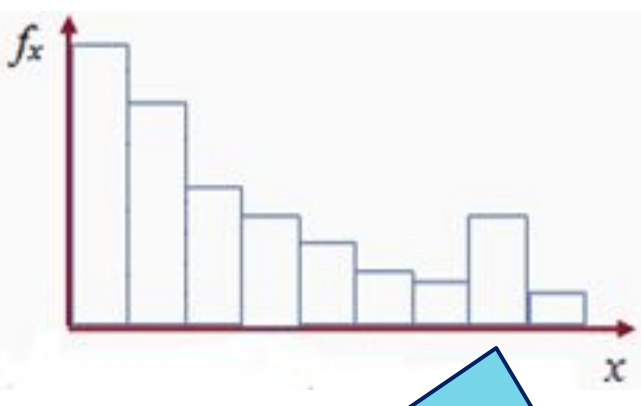
- позволяет определить характер факторов, снижающих качество данных;
- выбрать лучший метод предварительной обработки данных.



Значения равновероятны. Для восстановления пропусков удобно выбрать случайное значение



Есть ярко выраженная мода. Для восстановления пропусков удобно выбрать наиболее вероятное значение



Тяжелый «хвост» распределения. Вероятно наличие выбросов и экстремальных значений

Индикаторы качества данных

Индикаторы качества данных – это показатели, характеризующие качество набора данных в целом.

Примеры.

- Доля записей, не содержащих пропусков, выбросов и экстремальных значений (т. е. пригодных для анализа без предварительной обработки).

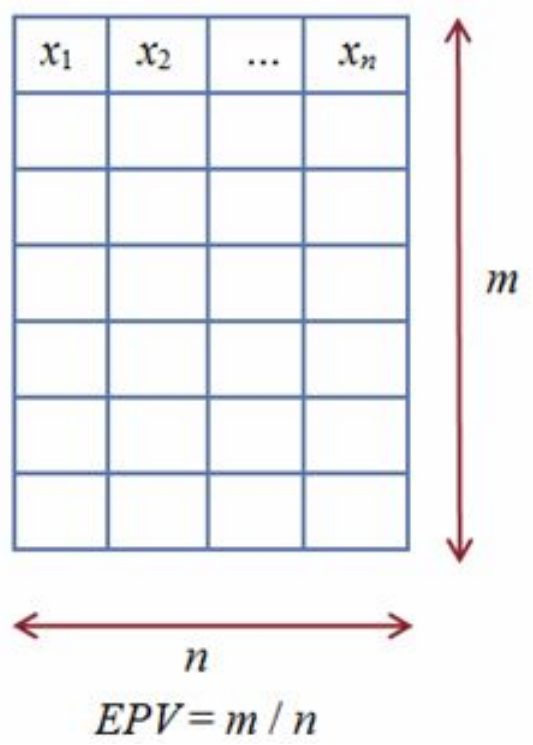
Если доля таких записей менее 50%, то набор данных может быть признан непригодным для анализа (предварительная обработка столь большого числа записей может привести к значительному смещению данных)

Индикаторы качества данных

Примеры.

- Отношение числа записей набора данных к числу столбцов (наблюдений на предиктор, *Evidence per Value* – ***EPV***).

Чем выше *EPV*, тем перспективнее набор данных с точки зрения построения предсказательных моделей. Рекомендации разных специалистов: значение *EPV* должно быть не менее 10-20



Взаимные проверки

Контроль очевидного несоответствия данных в логически связанных полях. Выполняется с помощью некоторых правил.

Примеры правил:

- возраст потенциального заемщика не может быть меньше стажа работы;
- время работы на текущем месте не может быть больше общего стажа работы;
- личный доход клиента не может быть больше семейного дохода.

Взаимные проверки

Красным цветом выделены номера строк с несоответствиями данных в связанных полях.

№	Возраст	Пенсионер	Время работы на текущем месте, мес.	Стаж работы, лет	Личный доход, руб.	Семейный доход, руб.
1	52	0	84	15	9 000	22 000
2	52	0	72	11	19 000	17 000
3	38	0	120	24	20 000	27 000
4	29	0	8	3	5 200	15 000
5	37	0	192	5	10 000	22 000
6	45	0	120	50	14 000	14 000
7	37	1	0	34	10 000	19 000
8	45	0	120	12	14 000	22 500

Использование сведений о бизнес-процессах

Под ***сложными ошибками*** обычно понимают такие, которые невозможно выявить без привлечения знаний об особенностях и логике бизнес-процессов.

Пример.

Цена на один и тот же товар у розничного продавца может меняться (акции оптового поставщика, корпоративные скидки и т. п.).

При этом цена сохраняется для партии товара.

Использование сведений о бизнес-процессах

Пример (продолжение).

Если период D_3-D_1 менее суток, причем $X_1 = X_3$, но $X_1 < X_2$ и $X_2 > X_3$, то это может свидетельствовать об ошибке в записи цены X_2.