Оглавление

Тема 1. Работа с данными

Что такое данные

Данные – это воспринимаемые человеком факты, события, сообщения, измеряемые характеристики, регистрирующие сигналы. Специфика данных в том, что они, с одной стороны, существуют независимо от наблюдателя, а с другой – становятся собственно «данными» лишь тогда, когда существует целенаправленно собирающий их субъект.

В итоге данные должны быть тем основанием, на котором возводятся все заключения, выводы и решения. Они вторичны по отношению к цели исследования и предметной области, но первичны по отношению к методам их обработки и анализа, извлекающим из данных только ту информацию, которая потенциально доступна в рамках отобранного материала.

Данные получаются в результате измерений. Под измерением понимается присвоение символов образцам в соответствии с некоторыми правилами. Эти символы могут быть буквенными или числовыми. Числовые символы могут представлять категории или быть числовыми.

Шкалы

Шкала наименований используется для измерения значений качественных признаков. Значением такого признака является наименование класса эквивалентности, к которому принадлежит рассматриваемый объект. Примерами значений качественных признаков являются названия государств, цвета, марки автомобилей и т.п.

Процедура измерения в номинальной шкале состоит в классификации (группировке) объектов таким образом, что объекты одного класса однородны (одинаковы) по анализируемому признаку или свойству, тогда как объекты из разных классов различаются по анализируемому признаку или свойству.

Например, группу людей можно разбить на две группы по полу или по семейному положению, на несколько групп по цвету глаз или волос, по месту проживания, по образованию, по профессии и т. д.

При построении шкал наименований должны быть выполнены следующие требования:

- каждый член некоторого множества объектов должен быть отнесен лишь к одному классу объектов (или к собирательному классу **прочие объекты);**
- ни один из объектов не может быть отнесен одновременно к двум или большему числу классов.

Приписываемые объектам символы, которые могут быть цифрами, буквами, словами, специальными символами, являются не более чем метками соответствующих классов. Отличительной особенностью номинальной шкалы выступает принципиальная невозможность упорядочить классы по измеряемому признаку – к ним невозможно применять суждения **больше-меньше, хуже-лучше** (например, пол, национальность, профессия, сорт вина, предпочитаемая музыка и. Т.п.). Единственное отношение, устанавливаемое на шкале наименований, это отношение тождества – обьекты, принадлежащие к одному классу, считаются **тождественными,** к разным – **различными.**

Если идентифицированные классы обозначить цифрами (1-м, 2-м 3-м и т.д.) что удобно при компьютерной обработке результатов измерения, то такие цифры не являются числами, не обладают свойствами чисел и к ним неприменимы арифметические действия. С величинами, измеряемыми в шкале наименований, можно выполнять только одну операцию – проверку их совпадения или несовпадения. По результатам такой проверки можно дополнительно вычислять частоты заполнения (вероятности) для различных классов, которые могут использоваться для применения рае личных методов статистического анализа

Частным случаем шкалы наименований является шкала, с помощью которой фиксируется наличие или отсутствие у объекта определенного свойства, его соответствие – несоответствие определенному требованию. Признак, который измеряется по дихотомической шкале наименований, называется альтернативным Он может принимать всего два значения: 0 – объект не обладает измеряемым качеством; 1 – обладает. Наиболее распространенные примеры таких шкал это: пол (мужчина – женщина), успешностъ выполнения задания (справился – не справился), соответствие норме (норма – патология), проголосовал за – проголосовал против и т.п.

Порядковая шкала (ранговая) более совершенна, чем номинальная, поскольку она строится на отношении тождества и порядка и позволяет устанавливать предпочтения между различными объектами. Она применяется

для упорядочения объектов по одному или нескольким признакам. Наибольшее распространение порядковые шкалы получили при измерении и сравнении качественных свойств, которые нельзя оценить непосредственно каким-либо числом.

Однако при этом, как правило, качественным суждениям человека приписывают количественные оценки, которые называются баллами. Баллы – это обычно натуральные числа, которые показывают ранг тех или иных объектов и следуют в порядке убывания или возрастания их предпочтительности. Например, используя порядковую шкалу, руководитель может оценить исполнительскую дисциплину или квалификацию своих сотрудников, выставляя им следующие баллы: 2 – низкая, 3 – средняя, 4 – высокая, 5 – очень высокая. Числа в этой шкале определяют только порядок следования объектов по их предпочтительности, но не позволяют утверждать, в какой степени один объект предпочтительнее, чем другой.

В частности, оценки показателей в порядковой шкале могут иметь всего два значения: 0 (не выполнил) и 1 (выполнил). Такие показатели часто встречаются в управленческой практике. Они применяются для оценивания таких заданий или работ, для которых обязательно выполнение всех требований. Если хоть что-то не выполнено, то и все задание считается невыполненным.

Например, проект считается выполненным, если соблюдены все требования заказчика. Если хоть одно требование не выполнено, и подписи заказчика на акте приемки-сдачи работ не стоит, значит; и проект в целом еще не завершен. Или, скажем; нельзя заключить договор на строго определенных условиях на 99%. Можно либо заключить, либо не заключить. Снова получаем только два возможных значения показателя.

Для порядковой шкалы допустимыми считаются любые преобразования показателей, которые не нарушают порядок следования объектов. Показатели, измеряемые в порядковой шкале; несут уже гораздо больше информации и позволяют судить об отношениях предпочтения между объектами типа **лучше-хуже, больше-меньше** и другие.

Однако в этой шкале также отсутствуют понятия масштаба и начала отсчета. Поэтому значения показателей, имеющих порядковую шкалу не позволяют ответить на вопрос: на сколько или во сколько раз один объект предпочтительнее, чем любой другой?

В отличие от двух предыдущих шкал, в **шкале интервалов** существует единица измерения, либо реальная (физическая), либо условная, при помощи которой можно установить количественные различия между объектами в отношении измеряемого свойства. Равные разности чисел в этой шкале будут означать равные различия в количествах измеряемого свойства у разных объектов, или у одного и того же объекта в разные моменты времени. Однако, то, что одно число оказывается в несколько раз больше другого, не обязательно говорит о таких же отношениях в количествах из меряемых свойств. В шкале интервалов может быть задействована вся числовая ось, но при этом ноль не указывает на отсутствие измеряемого свойства, т.к. нулевая точка часто является произвольной (например, как в шкале температуры по Цельсию), либо вообще отсутствует, как в некоторых шкалах психологических тестов.

Примерами шкалы интервалов являются календарное время, температурные шкалы Цельсия и Фаренгейта. Шкала оценок с заданным количеством баллов часто рассматривается как интервальная в предположении, что минимальное и максимальное положения на шкале соответствуют некоторым крайним оценкам или позициям и интервалы между баллами шкалы имеют одинаковую длину. К шкалам интервалов относится абсолютное большинство измерительных шкал, применяемых в науке, технике и быту: рост и вес, возраст, расстояние, сила тока, время (длительность промежутка между двумя событиями).

Шкала отношений (пропорциональные шкалы) классифицирует объекты пропорционально степени выраженности измеряемого свойства. Есть **абсолютный нуль (0),** указывающий на полное отсутствие измеряемого свойства.

В шкале отношений также существует единиц а измерения, при помощи которой объекты можно упорядочить в отношении измеряемого свойства и установить количественные различия между ними. Особенностью шкалы отношений является то, что к числам этой шкалы применимы все математические операции, следовательно, возможен ответ на вопрос, во сколько раз одно значение больше или меньше другого. В этой шкале обязательно, по крайней мере; теоретически, присутствует нуль, который говорит об абсолютном отсутствии измеряемого свойства Большинство ныне существующих физических шкал (длины, массы, времени, температуры по Кельвину и т. д.) являются примерами шкал отношений.

Показатели, измеряемые в шкале отношений, наиболее распространены в теории и практике управления. Например, к ним относятся прибыль, объем продаж, объем производства, доля фирмы на рынке, уровень риска, издержки, показатели рентабельности, затраты времени и другие показатели, для которых существует естественное начало отсчета **(нулевая точка).**

Основное различие между шкалами интервалов и отношений со стоит в том что шкала отношений имеет **абсолютный нуль,** не зависящий от произвола наблюдателя и соответствующий полному отсутствию измеряемого признака, а на шкале интервалов нуль устанавливается произвольно или в соответствии с некоторыми условными договоренностями.

Важно, что в ряду шкал – наименований, порядка, интервалов, отношений – увеличивается мощность шкал:

- качественные измерения с меняются количественными;
- возрастают возможности оценки свойств объектов, различий и отношений их свойств;
- увеличиваются возможности применения арифметических операций статистических мер и критериев;
- расширяются пределы инвариантности измерений.

Более мощные шкалы обладают всеми возможностями шкал менее мощных, что связывает все шкалы в единую систему измерений.

Метрические шкалы – это шкалы, у которых есть единицы измерения (например, метр, м/с). К ним относится шкала отношений. **Неметрические шкалы –** это шкалы, у которых нет единицы измерений. К ним относятся шкала наименований, порядковая и интервальная шкала.

Выбор данных для заданной шкалы

Определение того, в какой шкале измерен о явление (представлен признак) – ключевой момент анализа данных: любой последующий шаг, выбор любого метода зависит именно от этого. Поэтому первым шагом любого статистического анализа является определение типа исследуемых данных и отнесение их к той или иной шкале измерения – *номинальной, порядковой, интервальной* или *шкале отношений.*

Определение того, в какой шкале измерен о явление (представлен призн ак) – ключевой момент анализа данных: любой последующий шаг, выбор любого метода зависит именно от этого. Поэтому первым шагом любого

статистического анализа является определение типа исследуемых данных и отнесение их к той или иной шкале измерения – *номинальной, порядковой, интервальной* или *шкале отношений.*

Числовые (скалярные) **переменные** определяют числовые величины, измеряемые на некоторой *интервальной шкале* (относительной шкале) или *шкале отношений* (абсолютной шкале).

Скалярные величины, измеренные на интервальной шкале; могут сравниваться, упорядочиваться, складываться, вычитаться. С ними можно выполнять все обычные операции над числами, такие как вычисление среднего и оценку изменчивости. Примеры таких величин – *время, высота местности, температура по Цельсию* – это величины, которые по физической природе либо не имеют абсолютного нуля, либо допускают свободу выбор а в установлении начала отсчета. Для числовых величин, измеренных на относительной или абсолютной шкале, помимо операций сравнения и упорядочивания применимы любые арифметические действия.

Примеры величин, измеренных в шкапе отношений – *вес, длина, электрическое сопротивление, температура по Кельвину, деньги,* в абсолютной шкале – *количество предметов.* Эти шкалы имеют абсолютную нулевую точку, которая характеризует полное отсутствие измеряемого свойства, а абсолютная шкала – еще и абсолютную безразмерную единицу.

Таблица 1 Пример переменных на интервальной шкале и шкале отношений

Переменная	Шкала
Температура	Интервальная шкала
Экзаменационная оценка	Интервальная шкала
Грегорианский календарь	Интервальная шкала
Вес	Шкала отношений
Возраст в годах	Шкала отношений
Зарплата в рублях	Шкала отношений

Шкалы могут приводиться одна к другой: количественная – к ординальной или номинальной, ординальная шкала – к номинальной. Приведение одной шкалы к другой называют *понижением шкалы,* оно ведет к потере некоторой части информации об изучаемых признаках. Обратные операции считаются некорректными. Понижение шкал применяется при анализе переменных, измеренных в разных шкалах.

Тип данных определяет, какими статистическими методами эти данные могут обрабатываться и анализироваться. В следующей таблице приводятся статистические характеристики, которые имеют смысл в различных шкалах.

Правильный выбор шкалы дня измерения показателей имеет большое значение и зависит от наличия необходимой информации и цели, которая преследуется при проведении оценивания.

Так, использование метрических шкал требует более полной информации по сравнению с номинальными или порядковыми шкалами, а получение этой информации связано с дополнительными затратами ресурсов и времени. Поэтому при выборе типа шкалы всегда необходимо учитывать особенность решаемой задачи: как эта информация будет использована в дальнейшем.

Если задача состоит в ранжировании чего-либо или кого-либо (например, сотрудников) по некоторому признаку (например, по их квалификации), то нет необходимости измерять количественные характеристики, а достаточно измерить лишь качественные и ограничиться порядковой шкалой. По мере получения дополнительной информации можно переходить к более совершенным шкалам.

Тип данных определяет множество значений, набор операций которые можно применять к таким значениям, и возможно, способ реализации хранения значений и выполнения операций.

Дискретные и непрерывные данные

По характеру варьирования количественные признаки делятся на *дискретные* и *непрерывные.*

Дискретные данные являются значениями признака, общее число которых конечно или бесконечно, но может быть подсчитано при помощи натуральных чисел от одного до бесконечности. С дискретными данными не могут быть произведены никакие арифметические действия; либо они не имеют смысла

Дискретными данными являются все данные строкового и логического типа. Дискретными могут быть и числовые данные. Например, показатель **Код товара,** принимающий значение целого типа, является дискретным, так как операции сложения, вычитания, умножения над этим показателем не имеют смысла.

Непрерывные данные – это данные, значения которых могут принимать какое угодно значение в некотором интервале. Над непрерывными данными

можно производить арифметические операции: *сложение, вычитание, умножение* и *деление,* и они имеют смысл.

Примерами непрерывных данных являются: *возраст, рост, любые стоимостные показатели, количественные оценки* (ВНП страны, количество товара, объем отгрузки, вес отгрузки, прибыль и т.д.).

Качественные и количественные данные

Существует два типа переменных, значения которых образуют наборы данных: *качественные (категорийные)* и количественные *(числовые).*

Значения переменных, которые регистрируются с помощью чисел, имеющих содержательный смысл, называют **количественными** данными. Существуют две разновидности числовых данных: *дискретные* и *непрерывные.*

Например, многие результаты деятельности организации можно оценить количественно в объективных единицах измерения (рублях, часах, процентах и так далее). Для результатов, допускающих количественное измерение, используют количественные показатели. Значения таких показателей выражаются в виде некоторого действительного числа, имеющего определенный физический или экономический смысл.

К ним относятся все финансовые показатели *(выручка, чистая прибыль, постоянные и переменные издержки, показатели рентабельности, оборачиваемости, ликвидности* и другие), а также часть рыночных показателей *(объем продаж, доля рынка, размер/рост клиентской базы* и так далее) и показателей, характеризующих эффективность бизнес-процесс о в, деятельность по обучению, по развитию предприятия (например: *производительность труда; производственный цикл; время выполнения заказа; текучесть персонала; количество сотрудников, прошедших обучение* и другие).

Однако большинство характеристик и результатов работы организации, подразделений и сотрудников строгому количественному измерению не поддаются. Для их оценивания используют качественные показатели.

Данные; регистрирующие определенное качество, которым обладает объект или явление; называются **качественными.**

Значения качественных (атрибутивных) признаков часто выражаются в словесной форме. Примерами качественных данных являются *форма*

собственности предприятия, должность, которую занимает сотрудник на предприятии, тип акции и т п.

Даже если значения категорийных величин – числа (например, полу человека приписать соответственно числа 0 и 1), то обрабатывать эти числа как количественные данные нельзя.

Часто качественные показатели измеряют с помощью экспертных оценок; то есть субъективно, путем наблюдения за процессом и результатами работы. К ним, например, относятся такие показатели, как относительная конкурентная позиция предприятия, индекс удовлетворенности клиентов, индекс удовлетворенности персонала, качество и своевременность представления документов, соблюдение стандартов и регламентов, выполнение поручений руководителя *и многие другие.*

Качественные данные бывают двух типов: *порядковые,* для которых существует имеющий содержательный смысл порядок, и номинальные, для которых нет содержательно интерпретируемого порядка.

Порядковые данные можно ранжировать и использовать это ранжирование при проведении статистического анализа. Примером порядковых данных являются ответы на вопросы анкеты, содержащей следующие варианты ответов: **да; больше да, чем нет; больше нет, чем да; нет.** Хоть и можно выразить эти ответы числами (например, 4, 3, 2, 1), но предложенная шкала оценок носит субъективный характер. Нельзя считать, что разница между ответами 4 и 3 такая же; как и между ответами 2 и 1. Также нельзя считать, что ответ 3 в три раза лучше ответа 1.

Для измерения качественных и количественных показателей используются разные шкалы.

Количественными называют показатели, значения которых измеряются в любой метрической шкале. *Качественными* называют показатели, значения которых измеряются в номинальной или порядковой шкале.

В силу того, что символы, присваиваемые объектам в соответствии с порядковыми и номинальными шкалами, не обладают числовыми свойствами, даже если записываются с помощью цифр, эти два типа шкал получили общее название качественных, в отличие от количественных шкал интервалов и отношений.

Символы, приписываемые объектам в соответствии с количественными измерительными шкалами *(интервальная шкала* и *шкала отношений),* могут быть только числами.

Шкалы интервалов и отношений имеют общее свойство, отличающее их от качественных шкал: они предполагают не только определенный порядок между объектами или их классами, но и наличие некоторой единицы измерения, позволяющей определять, насколько значение признака у одного объекта больше или меньше; чему другого. Другими словами, на обеих количественных шкалах, помимо отношений **тождества** и **порядка,** определено отношение **разности,** к ним можно применять арифметические действия **сложения** и **вычитания.**

Тема 2. Анализ данных. Основные понятия и определения

Понятие Анализ данных

Сферы применения информационных технологий в современном обществе чрезвычайно велики. Под информационными технологиями мы будем понимать процессы получения, преобразования и потребления информации.

Основу информационных технологий составляют информационные процессы создания, регистрации, обработки, накопления, хранения, поиска и передачи данных и информации. Эти процессы обычно называют технологиями. Рассмотрим эти процессы подробнее.

Обработка данных – процесс выполнения последовательности операций над данными с целью преобразования их в информацию. Обработка данных может происходить в интерактивном и фоновом режимах.

Технологией обработки данных называют взаимосвязанные операции, выполняемые в строго определенной последовательности с момента передачи данных на обработку до получения заданных результатов.

К обработке данных можно отнести и технологии, построенные на анализе данных, которые стали актуальны в последние годы.

Анализ данных – широкое понятие. Сегодня существуют десятки его определений. В самом общем смысле *анализ данных – это процесс исследования, фильтрации, преобразования и моделирования данных с целью извлечения полезной информации и принятия решений.* В процессе анализа данных исследователь производит совокупность действий с целью формирования определенных представлений о характер е явления, описываемого этими данными. Как правило, для анализа данных используются раз личные математические методы.

Наша задача – проследить эволюцию анализа данных (в современном пони мании этого термина) от самых истоков до бизнес-аналитики.

По одной из классификаций, анализ данных вырос из задач прикладной математики. Кроме анализа данных в ней выделяют еще две задачи: *вычислительная математика и идентификация моделей.* Так как исторически они возникли первыми, их еще называют классическими подходами прикладной математики.

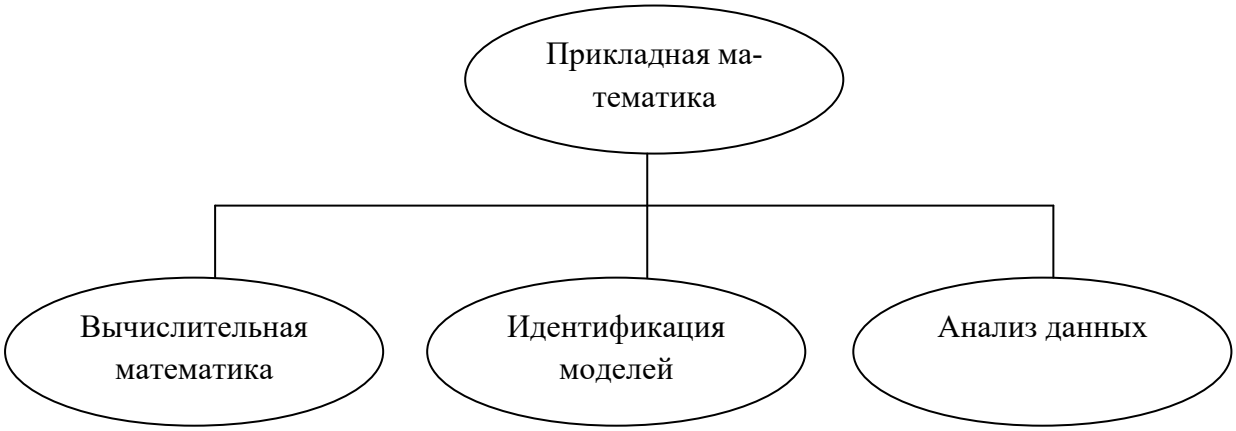


Рисунок 1. Классические подходы прикладной математики

Вычислительная математика решает задачу вычисления одних характеристик изучаемого объекта или явления по известным значениям других его характеристик. При этом модель объекта считается известной, а зависимости между характеристиками описываются аналитическим выражением в виде уравнения или системы уравнений или неравенств. Проблемы, возникающие при решении таких задач, связаны, в основном, с большими объемами вычислений и с защитой от погрешностей, накапливающихся в компьютере из-за округления

Задача **идентификации модели** формулируется по-другому: известен набор переменных; влияющих на целевую характеристику известен также общий вид зависимости между характеристиками, но коэффициенты, показатели степени и другие параметры модели неизвестны, и, чтобы их определить, используются протоколы наблюдений, отражающие значения одних характеристик при разных значениях других.

Делается серия предположений о значениях неизвестных параметров модели и эти предположения проверяются на протоколах. В результате выбираются такие значения параметров, при которых модель с заданной точностью позволяет по одним (входным) характеристикам определять другие (выходные или целевые) характеристики. К таким задачам принято относить дедуктивные процедуры математический статистики: корреляционный и регрессионный анализы, факторный анализ, численные методы оптимизации и т.п.

Задачи **идентификации моделей** требуют от исследователя ответственности за правильный выбор параметров модели (сама модель

считается известной). Наличие этого рискованного шагав процессе решения задачи также лишает результат ореола строгой математической чистоты.

Кроме того, на любом этапе развития прикладной математики возникают реальные задачи, для решения которых **нет готовых математических моделей** и времени для ожидания их появления нет. Поэтому в середине XX века у исследователей приходит пони мание ограниченности вычислительной математики, и ведутся поиски новых парадигм анализа данных.

Анализ данных Тьюки

Для полноты картины добавим, что в 1960-1970-х годах компьютеры были еще очень слабы в области отображения информации, особенно графической. И в эту формализованную и перегруженную теорией среду в 1962 году практически «ворвался» Джон *Тъюки (John Тикеу) с* концепцией **разведочного агализа данных** (англ.: *Exploratory Data Analisis, EDA).* Тьюки был убежден, что можно многое узнать из данных, просто визуализируя их; что нужно больше внимания уделятъ использованию данных для выдвижения гипотез.

Этот первичный этап анализа он назвал *разведочным,* а важнейшим его элементом определил широкое использование визуального представления много мерных данных. Для этого данные представляются в виде графиков, схем условных рисунков, таблиц, особенностью которых является **наглядность** – возможность увидеть признаки каких-либо закономерностей. Так; ящичная диаграмма (англ.: *box plot)* и диаграмма рассеяния (англ.:scatterplot) получили широкое распространение в статистике именно после работ Тьюки. К разведочному анализу данных относятся также методы, связанные с линейным проецированием, упрощением описания с помощью факторного анализа и многомерного шкалирования.

Разведочный анализ – это в большей степени подход, чем теория, синтез детерминированных, стохастических и эвристических подходов к анализу выборочных наблюдений. Всего в своей концепции Тьюки выделил три этапа анализа данных:

1. разведочный;
2. подтверждающий и ли конфирматорный;
3. итоговый.

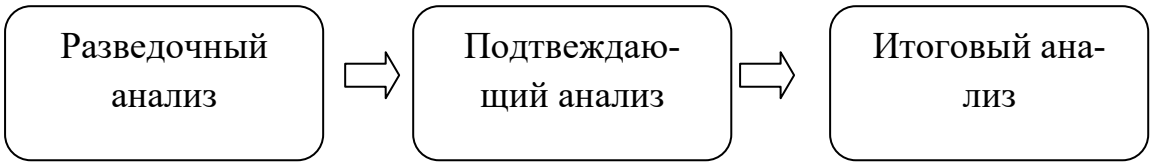


Рисунок 2. Три этапа анализа данных Тьюки

Можно сказать, что, концепция разведочного анализа Тьюки есть родитель таких современных технологий, как KDD, Data Mining, Big Data, которые послужили толчком к развитию разнообразных методов визуализации многомерных данных.

Если на первом этапе цель исследователя – выявить внутренние вероятностные и геометрические закономерности в данных для формирования и верификации тех или иных рабочих гипотез о связях между переменны ми, когда отсутствуют априорные представления о природе этих связей, то на следующего подтверждающем анализе, ставится задача проверки соответствия с формулированных гипотез полученным эмпирическим данным, вычисляются итоговые статистические оценки моделей и определяются их погреши ости. На третьем этапе анализа данных проводится экспертный анализ результатов и их обобщение. В случае необходимости, на всех этапах исследования возможны итерационные уточнения и обобщения.

Можно сказать, что концепция разведочного анализа Тьюки есть родитель таких современных технологий бизнес-аналитики, как KDD, Data Mining, Big Data, также послужила толчком к развитию разнообразных методов визуализации многомерных данных.

Что касается самого процесса анализа данный то по Тьюки он выглядит по-другому: вместо последовательности действий *Модель-Анализ-Данные* идет *Данные-Анализ (разведочный)-Модель-Анализ (подтверждающий)*. Таким образом, отправной точкой служат данные; характеризующие исследуемый объект или явление. Модель «следует» за данными, а не наоборот; как в классическом подходе. Двунаправленные стрелки показывают, что анализ данных носит циклический характер: на первых итерациях выдвинутые гипотезы могут потребовать дополнительных экспериментальных данных или наблюдений, уточнений. Это существенно облегчит подбор способов более глубокой обработки данных на этапах подтверждающего анализа.

Если результаты разведочного анализа говорят в пользу некоторой модели, то ее правильность можно проверить, применив к новым (то есть вновь

измеренным) данным явления или объекта. В вычислительной математике и математической статистике такие действия невозможны; именно по этой причине процедуры разделения выборок на обучающие и тестовые множества и валидации на них моделей пришли в *бизнес-аналитику* из концепции анализа данных, заложенной Тьюки.

Так рождается анализ данных в современном понимании, где решается задача анализа явлений, для которых еще нет математических моделей. Есть только наборы экспериментальных данных **входы-выходы** и даже только **входы,** представленные в виде массивов или таблиц.

Конструирование моделей и определение параметров этих моделей является основным предметом внимания современного анализа данных Исследователи отвечают за привнесение эвристических гипотез о возможных формах (моделях) зависимостей, параметр ах предполагаемых законов распределений и так далее. Наряду с *дедуктивным* аппаратом при решении этих задач используются *индуктивные* методы, реализованные в алгоритмах *машинного обучения*.

Однако концепция «моделей от данных» требует тщательного подхода к качеству самих исходных данных, поскольку ошибочные, зашумленные данные могут привести к моделям и выводам, не имеющим никакого отношения к действительности. Поэтому в анализе данных важную роль играют *интеграция, подготовка* и *очистка данных.*

В нашем курсе, да и вообще в бизнес-аналитике, говоря об анализе данных, мы будем предполагать использование именно подхода к анализу заложенного Тьюки. Современная динамичная среда, задачи, стоящие перед бизнесом огромные объемы накопленных данных не оставляют шанса на значительный успех классических подходов к анализу.

Современное понятие анализа данных

Концепция анализа данных, предложенная Тьюки, стала только отправной точкой в развитии анализа данных Появление персональных компьютеров и доступность специализированного программного обеспечения для визуализации и статистического анализа, безусловно, открыло массу новых возможностей широкому кругу исследователей, но этого было недостаточно.

К концу 80-х годов объем накопленной информации увеличивался каждые 20 месяцев. Объемы корпоративных баз и хранилищ данных росли еще

быстрее. Такой «информационный взрыв» потребовал очередного пересмотра подходов к анализу данных. Среди основных факторов, которые привели к этому помимо *значительных объемов данных,* также отметим следующие.

- Требование принятия решений в *режиме реального времени* или близком к реальному.
- *Разнородный характер данных,* измеренных в различных шкалах и необходимость их обработки разнотипными средствами и сведение результата в некоторой единой инструментальной среде.
- Осознание *потребности* в создании средств автоматического выделения и анализа скрытых зависимостей.
- *Сложный характер объекта управления* (взаимодействие большого множеств в разнородных процессов и подсистем), ограничивающий использование как традиционных моделей и методов, так и интеллектуальных экспертных систем продукционного типа.

Таким образом, перед исследователями встают новые задачи, характеризующиеся следующими особенностями.

- Объект исследования характеризуется большими объемами данных, требуется анализ в ограниченное время.
- Гарантий того, что данные хорошего качества, нет, требуется проводить их аудит, очистку и обогащение;
- Формальная модель объекта отсутствует (нет полного и непротиворечивого аналитического описания).
- Необходимо уметь выделять параметры, определяющие поведение объекта исследования в тех или иных ситуациях;
- Необходимо уметь обобщать имеющуюся информацию, выделяя неявно представленные зависимости (то есть те эмпирически е правила, которые позволяют предсказывать поведение модели объекта в новых обстоятельствах).

Ответом на эти вызовы стало появление в 90-х годах технологий хранилищ и витрин данных (англ.: Data Warehousing), аналитической от четности и интеллектуального анализа данных (англ.: Knowledge Discovery in Databases и Data Mining). Сегодня часто все эти технологии рассматривают в контексте термина бизнес-аналитика.

Понятие Бизнес-аналитика

Бизнес-аналитика – это перевод на русский язык англоязычных терминов *Business Intelligence (BI)* и *Business Analytics.* Строго говоря, это немного разные по смыслу понятия, а название «бизнес-аналитика» уже прочно закрепилось в качестве перевода для *Business Intelligence.* Для простоты будем считать, что это синонимы, хотя, если вникнуть в детали, то окажется, что *Business Intelligence* – это один из базовых сегментов *Business Analytics.*

Итак; термин бизнес-аналитика, *Business Intelligence*, впервые появился в 1958 году в статье исследователя из IBM Ханса Питера Луна. Он определил этот тер мин как: «Возможность понимания связей между представленными фактами». Широкого распространения в те годы термин не получил. В конце 80-хгг Говард Дреснер (позже – аналитик компании Gartner) определил *Business Intelligence* как общий термин, описывающей «концепции и методы для улучшения принятия бизнес-решений с использованием систем на основе бизнес-данных».

Можно признать удачным и емким определение авторитетной консалтинговой компании IDC: *бизнес-аналитика – эта инструменты и приложения для поиска, анализа, моделирования и доставки информации, необходимой для принятия решений*. Современная бизнес- аналитика – мультидисциплинарная область, находящаяся на стыке информационных технологий, баз данных, алгоритмов интеллектуальной обработки информации, математической статистики и методов визуализации и ее корни идут из концепции анализа данных по Тьюки.

В проектах по бизнес-аналитике присутствуют четыре важные составляющие: *эксперт, гипотеза, аналитики руководитель проекта.* Дадим им определения.

Эксперт – специалист предметной области, профессионал, который за годы обучения и практической деятельности научился эффективно решать задачи, относящиеся к конкретной предметной области.

Эксперт – ключевая фигура в процессе анализа. По-настоящему эффективные аналитические решения можно получить не на основе одних лишь компьютерных программ, а в результате сочетания лучшего из того, что может человек и компьютер. Эксперт выдвигает гипотезы (предположения) и

для проверки их достоверности либо просматривает некие выборки различными способами, либо строит те или иные модели.

Гипотезой в анализе данных часто выступает предположение о влиянии какого-либо фактора или группы факторов на результат. К примеру, при построении прогноза продаж допускается предположение; что на величину будущих продаж существенно влияют продажи за предыдущие периоды и остатки на складе. При оценке кредитоспособности потенциального заемщика выдвигается гипотеза, что на кредитоспособность влияют социально-экономические характеристики клиента: возраст; образование, семейное положение и т п.

В крупных проектах по бизнес-аналитике участвуют, как правило, несколько экспертов, аналитиков, а кроме того, руководитель проекта.

Аналитик – специалист в области анализа и моделирования Аналитик на достаточном уровне владеет какими-либо инструментальными и программными средствами анализа данных, например, методами *Data Mining.*

Аналитик играет роль «мостика» между экспертами, то есть является связующим звеном между специалистами разных уровней и областей. Он собирает у экспертов различные гипотезы, выдвигает требования к данным, проверяет гипотезы и вместе с экспертами анализирует полученные результаты. Аналитик должен обладать системными знаниями, так как помимо задач анализа на его плечи часто ложатся технические вопросы, связанные с базами данных интеграцией с источниками данных, тестированием и производительностью.

Поэтому в дальнейшем главным лицом в анализе данных мы будем считать аналитика, предполагая, что он тесно сотрудничает с экспертами предметных областей.

В обязанности руководителя проектов входят функции координации действий всех участников проекта, решение спорных вопросов, планирование и контроль сроков проекта.

Итак; следуя концепции анализа данных Тьюки, современная бизнес-аналитика делит методы решения задач на две основные группы:

- извлечение и визуализация данных;
- построение и использование моделей.

На рисунке 3 приведена общая схема такого анализа.

Таким образом, группе задач *Извлечение и визуализация данных,* по сути, соответствует этап разведочного анализа данных, а групп е *Построение и использование моделей* – этап подтверждающего анализа.

Конечно, эти соответствия условны, ведь тот же подтверждающий анализ Тьюки опирался на методы математической статистики, а в бизнес-аналитике – это не только статистические методы, но и различные описательные и предсказательные модели *Data Mining,* алгоритмы машинного обучения.

Рисунок 3. Общая схема анализа данных в современном понимании

Чтобы получить новые знания об исследуемом объекте или явлении, не обязательно строить сложные модели. Часто достаточно «посмотреть» на данные в нужном виде, чтобы сделать определенные выводы или выдвинуть предположение о характере зависимостей в системе; получить ответ на интересующий вопрос. Это помогает сделать визуализация.

В случае визуализации аналитик некоторым образом формулирует запрос к информационной системе, извлекает нужную информацию из различных источников и просматривает полученные результаты. На их основе он делает выводы, которые и являются результатом анализа. Существует множество способов визуализации данных:

- *OLAP* (кросс-таблицы и кросс-диаграммы);
- таблицы;
- диаграммы, гистограммы;
- карты, проекции, срезы и т. п.

Несомненными достоинствами визуализации являются относительная простота создания и введения в эксплуатацию подобных систем и возможность их применения практически в любой сфере деятельности. Кроме того, в этом случае по максимуму используются знания эксперта в предметной области и его способность принимать во внимание многие трудно формализуемые факторы, влияющие на бизнес.

Недостатками визуализации являются не способность людей обнаружить достаточно сложные и нетривиальные зависимости, а также невозможность отделить знания от эксперта и тиражировать знания.

Этапы моделирования

Построение моделей – универсальный способ изучения окружающего мира, позволяющий обнаруживать зависимости, прогнозировать, разбивать на группы и решать множество других важных задач. Но самое главное: полученные таким образом знания можно тиражировать.

Тиражирование знаний – совокупность инструментальных средств для создания моделей, которые обеспечивают конечным пользователям возможность использовать результаты моделирования для принятия решений, без необходимости понимания методик, при помощи которых эти результаты получены.

Процесс построения моделей в бизнес-аналитике состоит из нескольких шагов (см. рисунок 4).

Формулирование цели моделирования. При построении модели следует отталкиваться от задачи, которую можно рассматривать как получение ответа на интересующий заказчика вопрос. Процесс построения моделей в бизнес-аналитике показан на рис. 4.

Рисунок 4. Процесс построения моделей в бизнес-аналитике

Например, в розничной торговле к таким вопросам относятся следующие:

- Какова структура продаж за определенный период?
- Какие клиенты приносят наибольшую прибыль?
- Какие товары продаются или заказываются вместе?
- Как оптимизировать товарные остатки на складах?
- В этом случае можно говорить о создании модели прогнозирования продаж, модели выявления ассоциаций и т д.
- *Подготовка и сбор данных* Современная бизнес-аналитика полагается на использование данных, подготовить и систематизировать которые отдельная задача.

Если качество модели неудовлетворительное, то процесс построения модели повторяется, как это было показано на рис. 4.

На практике рассмотренные подходы к анализу комбинируются. Например, визуализация данных наводит аналитика на некоторые идеи, которые он пробует проверить при помощи различных мод елей, а к полученным результатам снова применяются методы визуализации. Механизмы визуализации и построения моделей дополняют друг друга.

Тема 3: Методология CRISP-DM

Стандарт CRISP-DM

Рассмотрев процесс анализа данных, нельзя не упомянуть межотраслевой стандарт CRISP-DM (англ: Cross Industry Standard Process for Data Mining). По сути это популярная методология ведения проектов в бизнес-аналитике, особенно если в них используются модели Data Mining.

CRISP-DM был разработан в конце 1996 года четырьмя «ветеранами» из молодых компаний на рынке бизнес-аналитики: ISL (поглощена SPSS Inc), NCR Corporation, Daimler-Benz и OHRA.

На рисунке 6 приведена модель жизненного цикла проекта по анализу данных, который состоит из шести этапов. При этом последовательность этапов не является строгой. Стрелки указывают наиболее важные и частые зависимости между этапами.

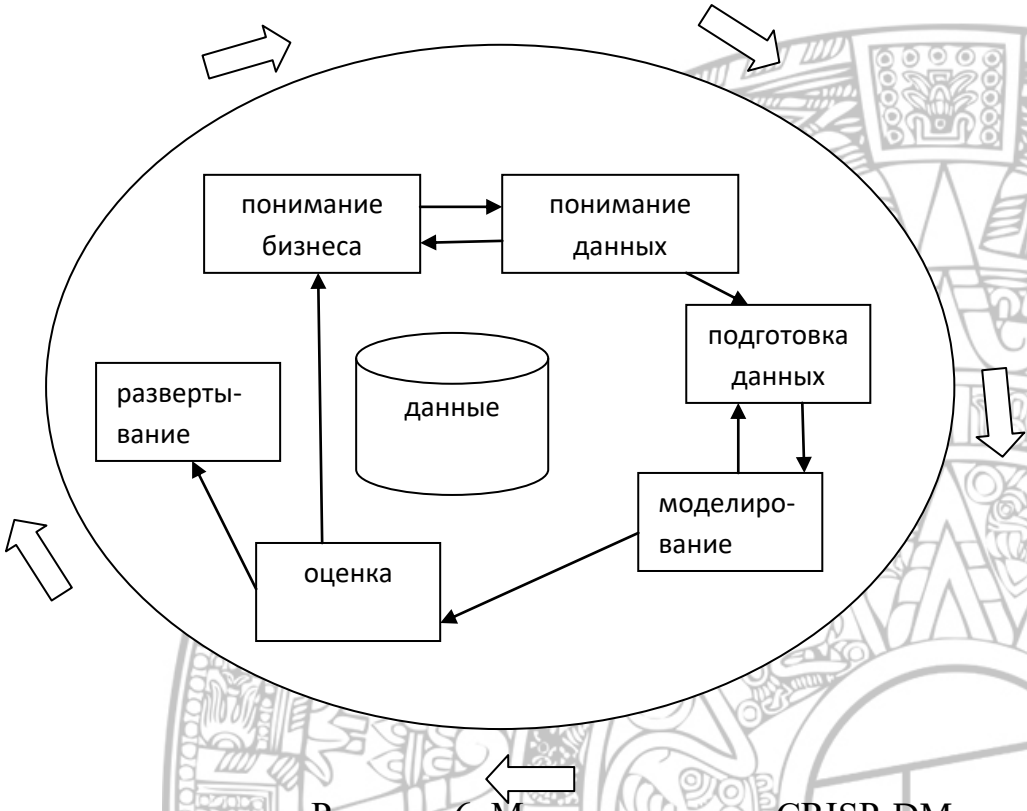


Рисунок 6. Модель процесса CRISP-DM

Внешний круг на рис. 6. указывает на цикличность процесса анализа данных, который продолжается и после развертывания проекта. Необходимо

постоянно совершенствовать свои модели для того, чтобы они давали лучшие результаты и не устаревали.

Понимание бизнеса. Первый этап посвящен определению целей проекта и требований к результату с точки зрения бизнеса. Далее необходимо сформулировать их на языке анализа данных и классов задач Data Mining, а также разработать предварительный план проекта.

Понимание данных начинается с первоначального сбора данных визуализации и разведочного анализа, выявления проблем с качеством данных. Цель этапа – понять структуру данных, обнаружить интересные подмножества для формирования и последующей проверки гипотез.

Этап подготовки данных ставит целью получить итоговый набор данных, которые будут использоваться при моделировании, из исходных первичных источников.

Процедуры подготовки данных могут выполняться много раз. Они включают в себя отбор таблиц, записей и атрибутов, а также конвертацию и очистку данных для моделирования.

Моделирование. На этом этапе идет выбор методов и алгоритмов моделирования, строятся модели, а их параметры настраиваются на оптимальные значения. Как правило, для решения любой задачи анализа данных существует несколько подходов. Некоторые подходы накладывают особые требования на представление данных. Таким образом, часто бывает нужен возврат на шаг назад к фазе подготовки данных.

Оценка. На этом этапе проекта уже построена модель и получены количественные оценки её качества. Перед тем, как внедрять эту модель, необходимо убедиться, что основная бизнес-цель проекта достигнута. Возможно, придется какие-то вопросы рассмотреть более детально. В конце этапа принимается решение по использованию результатов анализа данных на практике.

Развертывание. В зависимости от требований этот этап может быть простым, например, составление финального отчета, или сложным, например, встраивание модели в бизнес-процесс. Обычно развертывание – это забота клиента. Но важно дать понять клиенту, что ему нужно сделать для того, чтобы начать использовать полученные модели.

Можно назвать следующие преимущества методологии *CRISP-DM.*

- Модель процесса CRISP-DM универсальна и подходит для ведения проектов по бизнес-аналитике в любых отраслях.
- Нет привязки к конкретным программным продуктам или инструментам.
- Методология близка по духу к технологии извлечения знаний из баз данных –*Knowledge Discovery in Databases, KDD*

Формы представления данных

Данные, описывающие реальные объекты, процессы и явления, могут быть представлены в различных формах, измерены в различных шкалах и иметь определенный тип и вид.

По степени структурированности выделяют следующие формы представления данных:

- неструктурированные;
- структурированные;
- слабоструктурированные**.**

К *неструктурированным* относятся данные, произвольные по форме, включающие тексты и графику, мультимедиа (видео, речь, аудио). Эта форма представления данных широко используется, например, в Интернете, а сами данные предоставляются пользователю в виде отклика поисковыми системами.

Структурированные данные отражают отдельные факты предметной области. Структурированными называются данные, определенным образом упорядоченные и организованные с целью обеспечения возможности применения к ним некоторых действий (например, визуального или компьютерного анализа). Это основная форма представления сведений в базах данных.

Организация того или иного вида хранения данных (структурированных или неструктурированных) связана с обеспечением доступа к ним. Под доступом понимается возможность выделения элемента данных (или множества элементов) среди других элементов по каким-либо признакам с целью выполнения некоторых действий над элементом.

Одной из самых распространенных моделей хранения структурированных данных является таблица. В ней все данные упорядочиваются в двумерную структуру, состоящую из столбцов (поля, колонки, переменные, атрибуты, признаки) и строк (записи, прецеденты, примеры).

В ячейках такой таблицы содержатся элементы данных: символы, числа, логические значения.

Неструктурированные данные непригодны для обработки напрямую методами анализа и подвергаются специальным приемам структуризации. Например, в анализе текстов при структурировании из исходного текста может быть сформирована таблица с частотами встречаемости слов, и уже такой набор данных будет обрабатываться методами, пригодными для структурированных данных.

Слабоструктурированные данные – это данные, для которых определены некоторые правила и форматы, но в самом общем виде. Например, строка с адресом, строка в прайс-листе, Ф ИО и т. п. В отличие от неструктурированных, такие данные с меньшими усилиями преобразуются к структурированной форме, однако без процедуры преобразования они тоже непригодны для анализа. На рис. 7. приведен пример стандартизации строки с адресом.

390045 г. Рязань, ул. Ленина, д. 45 корп

Поле	**Значение**
Индекс	390045
Город	Рязань
Улица	Ленина
Дом	45
Корпус	1

Рисунок 7. Пример стандартизации строки с адресом

Подавляющее большинство методов анализа данных работает только с хорошо структурированными данными, представленными в табличном виде, поэтому дальнейшее изложение во всем курсе ведется применительно к структурированным данным. Сбор информации в структурированном виде осуществляется на этапе подготовки данных к анализу и обсуждается в соответствующей теме.

Типы данных

Поля структурированных данных принято делить на четыре типа:

- числовой, который бывает *целым (количество товара, код товара* и т п.) и *вещественным (цена, скидка* и т п.);
- символьный или строковый *(фамилия, наименование, адрес, пол, образованием* т п.);
- логический *(Да/Нет, Ложь/Истина, 0/1):*
- дата/время.

С типами дата/время и логический (ЛОЖЬ, ИСТИНА) все просто. Остановимся подробнее на символьных и числовых типах.

Номинальные переменные

Данные, представляющие собой значения категорийных (символьных) переменных, измеряются по номинальной, либо по порядковой шкале. Номинальные переменные могут принимать значения, измеренные на шкале наименований, состоящей из наименований категорий, которые никак естественным образом не упорядочиваются. Никаких соотношений, кроме равенства или неравенства, между такими значениями нет. Эти данные могут упорядочиваться, с ними не могут быть произведены никакие арифметические действия.

Таблица 2 Примеры номинальных переменных

Переменная	Категории
Наличие машины	Да Нет
Кредитная история	Положительная Отрицательная Нет данных
Провайдер	Мегафон МТС Билайн

Как правило, по умолчанию алгоритмы анализа данных считают, что категорийные (символьные) данные будут обрабатываться как номинальные.

Ординальные переменные

Переменные, измеренные на шкале порядка (их еще называют ординальными), имеют упорядоченные категории. К таким переменным можно отнести различные балльные или экспертные оценки с очевидным упорядочением значений. Балльная оценка успеваемости (неудовлетворительно, удовлетворительно, хорошо, отлично) являются типичным примером порядковой величины.

Измерения в ординальной шкале содержат информацию только в порядке следования величин, но не позволяют количественно выразить, насколько или во сколько раз одно значение больше или меньше другого. Для таких данных применимы только операции сравнения и ранжирования: равно, не равно, больше, меньше; арифметические действия не могут быть произведены.

Таблица 3 Примеры ординальных переменных

Переменная	Упорядоченные категории
Образование	Среднее Среднее специальное Высшее Ученая степень
Оценка продукции	Плохо Удовлетворительно Хорошо Очень хорошо
Дисконтная карта	Бронзовая Серебряная Золотая

Не все алгоритмы анализа данных умеют работать специальным образом с ординальными переменными, учитывая порядок следования категорий.

Числовые, дискретные и непрерывные переменные

Числовые переменные определяют числовые величины, измеряемые на некоторой интервальной шкале (относительной шкале) или шкале отношений (абсолютной шкале). Такие величины, измеренные на интервальной шкале, могут сравниваться, упорядочиваться, складываться, вычитаться. С ними можно выполнять все обычные операции над числами, как вычисление среднего и оценку изменчивости.

Тип данных определяет, какими методами эти данные могут обрабатываться и анализироваться.

По характеру варьирования переменные делятся на дискретные и непрерывные.

Определение. Дискретные данные являются значениями признака, общее число которые конечно или бесконечно, но может быть подсчитано при помощи натуральных чисел от одного до бесконечности. С дискретными данными не могут быть произведены никакие арифметические действия, либо они не имеют смысла.

Дискретными данными являются данные строкового и логического типа. Дискретными могут быть и числовые данные. Например, показатель Код товара, принимающий значение целого типа, является дискретным, так как операции сложения, вычитания, умножения над этим показателем не имеют смысла.

Непрерывные данные – это данные, которые могут принимать любое значение в некотором интервале. Над непрерывными данными можно производить арифметические операции: сложение, вычитание, умножение и деление, и они имеют смысл.

Возможные сочетания характеров и типов данных приведены в табл. 4.

Таблица 4 Соответствие между типами и видами данных

Тип данных	**Вид данных**	
	Непрерывный	**Дискретный**
Числовой	+	+
Строковый		+
Логический		+
Дата/время	+	+

Примерами непрерывных данных являются: возраст, рост, любые стоимостные показатели, количественные оценки (ВВП страны, количество товара, объем отгрузки, все отгрузки, прибыль и так далее).

Тема 4. Представления наборов данных

Упорядоченные и неупорядоченные наборы данных

По отношению к задаче анализа наборы данных могут быть упорядоченными и неупорядоченными.

В упорядоченном наборе данных каждому столбцу соответствует один признак, а в каждую строку заносятся упорядоченные по какому-либо признаку события с интервалом периода между строками. Часто таким признаком выступает время. На рисунке приведены примеры упорядоченных наборов данных – временной ряд (упорядочен по дате) и ряд показателей датчика зонда (упорядочен по глубине скважины).

Таблица 5 Примеры упорядоченных наборов данных

Дата	Количество	Сумма
01.02.2017	4	283,31
01.02.2017	1	72,48
01.02.2017	1	173,32
02.02.2017	6	294,84
02.02.2017	2	405,76
02.02.2017	12	303,13
02.02.2017	1	210,50
03.02.2017	6	512,16
03.02.2017	3	156,96

Глубина	ВК	05
887,9	8,85	0,218
888,1	9,627	0,216
888,3	14,584	0,217
888,5	21,647	0,215
888,7	17,172	0,216
888,9	6,118	0,215
889,1	2,886	0,217
889,3	2,506	0,219

В неупорядоченном наборе каждому столбцу соответствует признак, а в каждую строку заносится пример (ситуация, прецедент), соответственно, упорядоченность строк не требуется. Пример такого набора данных приведен на рисунке.

В табл. 6 показан пример неупорядоченного набора данных.

Таблица 6 Пример неупорядоченного набора данных

Номер	Банк	Город	Филиалы	Собственные активы
2	Внешторгбанк	Москва	32	23236327
3	Газпромбанк	Москва	27	9255041
4	Альфа-Банк	Москва	17	12446938
5	ОАО «ПСБ»	Санкт-Петербург	44	1275859
6	Банк Москвы	Москва	34	3335734
7	АКБ «ДИБ»	Москва	0	261 6993

Особо выделяют транзакционные данные. Под транзакционными подразумевается несколько объектов или действий, являющихся логически связанной единицей.

Этот способ представления используется алгоритмами анализа покупок (чеков) в супермаркетах. Но в общем случае речь может идти о любых связанных объектах или действиях. В табл. 7 показан пример транзакции покупки товара в магазине.

Таблица 7 Пример транзакции

Одна транзакция →

Код транзакции	Товар
10200	Йогурт «Чудо» 0,4
10200	Батон «Рязанский»
10201	Вода «Боржоми» 0,5
10201	Сахарный песок

Подготовка данных к анализу

Бизнес-аналитика базируется на различных алгоритмах извлечения закономерностей из данных, результатом работы которых являются модели и знания. Таких алгоритмов довольно много, но они не способны гарантировать качественное решение. Никакой, даже весьма изощренный, метод сам по себе не даст хорошего результата, так как критически важным является качество исходных данных. Чаще всего именно оно становится причиной неудачи бизнес-аналитических проектов.

Несмотря на то, что существуют специальные методы аудита и повышения качества данных, понимание и соблюдение принципов сбора и подготовки

данных значительно облегчит построение моделей и позволит получить хорошие результаты. К тому же, среди аналитиков и руководителей проекта распространено мнение, подкрепленное практикой, что до 80% процесса анализа данных – это время, потрачено на их подготовку.

Особенности бизнес-данных

Бизнес-данные редко накапливаются специально дня решения задач анализа. Предприятия и организации собирают данные для коммерческих целей: ведения учета, проведения финансового анализа, составления отчетности, принятия решений и т. п. Этим бизнес-данные отличаются от экспериментальных данных, которые собираются для исследовательских целей. Основными потребителями бизнес-данных обычно являются лица, принимающие решения в компаниях.

Бизнес-данные, как правило, содержат ошибки, выбросы, противоречия и пропуски. Это следствие того, что компании не собирают данные с целью анализа. В них появляются ошибки различной природы, что снижает качеств о данных.

С точки зрения анализа объемы *хранимых данных очень велики*. Современные базы данных содержат гигабайты и терабайты информации. Для ресурсоемких алгоритмов анализа данных таблицу объемом 100 тыс. записей можно считать большой, поэтому при построении моделей это нужно учитывать, например, использовать масштабируемые алгоритмы, способные работать на больших наборах данных.

Отмеченные особенности бизнес-данных влияют как на сам процесс анализа, так и на подготовку и систематизацию данных.

Формализация данных – принципы

При сборе данных следует придерживаться следующих принципов.

Абстрагироваться от существующих информационных систем и имеющихся в наличии данных. Большие объемы накопленных данных совершенно не говорят о том, что их достаточно для анализа в конкретной компании. Необходимо отталкиваться от задачи и подбирать данные для ее решения, а не брать имеющуюся информацию.

К примеру, при построении моделей прогноза продаж опрос экспертов показал, что на спрос очень влияет цветовая характеристика товара. Анализ

имеющихся данных продемонстрировал, что информация о цвете товарной позиции отсутствует в учетной системе. Значит, нужно каким-то образом добавить эти данные; иначе не стоит рассчитывать на хороший результат использования моделей.

Описать все факторы, потенциально влияющие на анализируемый процесс/объект. Основным инструментом здесь становится опрос экспертов и людей, непосредственно владеющих проблемной ситуацией. Необходимо максимально использовать знания экспертов о предметной области и, полагаясь на здравый смысл, постараться собрать и систематизировать максимум возможных предположений и гипотез.

Экспертно оценить значимость каждого фактора. Эта оценка не является окончательной, она будет отправной точкой. В процессе анализа вполне может выясниться, что фактор, который эксперты посчитали очень важным, таковым не является, и наоборот, незначимый, с их точки зрения, фактор может оказывать значительное влияние на результат.

Определить способ представления информации – число, дата, да/нет, категория (т. е. *тип данных).* Определить способ представления, то есть формализовать, некоторые данные просто. Например, объем продаж в рублях – это определенное число. Но довольно часто бывает непонятно, как представить фактор. Чаще всего такие проблемы возникают с качественными характеристиками.

Например, на объемы продаж влияет качество товара. Качество – сложное понятие; но если этот показатель действительно важен, то нужно придумать способ его измерения. Скажем, качество можно определять по количеству брака на тысячу единиц продукции либо оценивать экспертно, разбив на несколько категорий – *отлично/хорошо/удовлетворительно/плохо.*

Собрать легкодоступные факторы. Они содержатся в первую очередь в источниках структурированной информации – учетных системах, базах данных и т.д.

Обязательно собрать наиболее значимые, с точки зрения экспертов, факторы. Вполне возможно, что без них не удастся построить качественную модель.

Оценить сложность и стоимость сбора средних и наименее важных по значимости факторов. Некоторые данные легкодоступны, их можно извлечь из существующих информационных систем. Но есть информация, которую

непросто собрать, например сведения о конкурентах, поэтому необходимо оценить, во что обойдется сбор данных. Сбор данных не является самоцелью. Если информацию получить легко, то, естественно, ее нужно собрать. Если сложно, то необходимо соизмерить затраты на ее сбор и систематизацию с ожидаемыми результатами.

Информативность данных

Одной из распространенных ошибок при сборе данных из структурированных источников является стремление взять для анализа как можно больше признаков, описываемых объекты. Между тем предварительная оценка данных, которая проводится при помощи разведочного анализа данных, существенно помогает в определении информативности признаков.

Среди информативных признаков можно выделить четыре типа (1) признаки, содержащие только одно значение; (2) признаки, между которыми имеет место сильная корреляция, – в этом случае для анализа можно взять только один из них.

Таблица 8 Примеры неинформативных данных

Признак	Признак	№ паспорта	Пол	Gender
1	1	0936-866096	Жен	0
1	1	8355-512928	Жен	0
1	0	8017-098418	Муж	1
1	1	0094-732300	Муж	1
(1)	(2)	(3)	(4)	

Требования к данным

Существуют определенные начальные требования к минимальным объемам данных для моделирования. В зависимости от представления данных и решаемой задачи эти требования различны.

Для *временных рядов,* которые относятся к упорядоченным данным, требования следующие.

- Если для бизнес-процесса (например, *продажи)* характерна сезонность/цикличность, то необходимо иметь данные хотя бы за один полный сезон/цикл с возможностью варьирования интервалов *(понедельное, помесячное* и так далее).
- Максимальный горизонт прогнозирования зависит от объема данных: данные за 1,5 года – прогноз возможен максимум на 1 месяц; данные за 2-3 года – на 2 месяца.
- Для *неупорядоченных данных* требования следующее:
- Количество примеров (прецедентов) должно быть значительно больше количества факторов (столбцов).
- Желательно, чтобы данные покрывали как можно больше ситуаций реального процесса.

Транзакционные данные. Анализ транзакций целесообразно производить на большом объеме данных, иначе могут быть выявлены статические шаблоны поведения. Алгоритмы поиска таких шаблонов способны быстро перерабатывать огромные массивы данных. Примерное соотношение между количеств ом объектов и объемом данных следующее:

- 300-500 объектов – от 10 тыс. транзакций; – 500-1000 объектов – более 300 тыс. транзакций;
- 500-1000 объектов – более 300 тыс. транзакций.

Тема 5. Сбор данных

Понятие Сбор данных

Качество принятых решений, выработанных в ходе анализа бизнес-данных, зависит не только от эффективности используемых методов и алгоритмов, но и от того, насколько правильно подобраны и подготовлены сами исходные данные.

Начальным этапом реализации любого BI-проекта или просто решения задачи анализа является **сбор данных**. Это отдельная задача, которая решается внедрением различных автоматизированных систем обработки информации. Поэтому мы предполагаем, что бизнес-процессы первичной регистрации и сбора фактографических данных каким-то образом в компании уже запущены. В зависимости от масштабов бизнеса компании и степени его автоматизации различается объём и характер регистрируемой информации.

Так, кроме самих продаж сеть розничных магазинов может собирать данные с сенсоров систем подсчета посетителей, статистику и порядок переходов на страницах в интернет-магазине и т. д. А технических комплексах системы сбора, а телеметрической информации представляют собой сложный программно-аппаратный комплекс, способный в непрерывном режиме регистрировать сотни и тысячи параметров.

Методы сбора данных

Есть несколько методов сбора необходимых для анализа данных.

Получение из учетных систем. Обычно в учётных системах есть различные механизмы построения отчетов и эксперта даннах, поэтому извлечение нужной информации из них чаще всего относительно несложная операция.

Получение данных из косвенных источников информации. О многих показателях можно судить по косвенным признакам, и этим нужно воспользоваться. Например, можно оценить реальное финансовое положение жителей определенного региона следующим образом. В большинстве случаев имеются несколько товаров, предназначенных для выполнения одной и той же функции, но отличающихся по цене: товары для бедных, средних и богатых. Если получить отчет о продажах товаров в интересующем регионе и проанализировать пропорции, в которых продаются товары для бедных,

средних и богатых, то можно предположить, что чем больше доля дорогих изделий из одной товарной группы, тем более состоятельны в среднем жители данного региона.

Использование открытых источников. Большое количество данных присутствует в таких открытых источниках, как статистические сборники, отчеты корпорации, опубликованные результаты маркетинговых исследований, социальные сети и прочее.

Приобретение данных у специализированных компаний. На рынке работает множество компаний, которые профессионально занимаются сбором данных и предоставлением результатов клиентам для последующего анализа. Собираемая информация обычно предоставляется в виде различных таблиц и сводок, которые с успехом можно применить при анализе. Стоимость получения подобной информации чаще всего относительно невысока.

Проведение собственных мероприятий по сбору данных. Этот вариант сбора данных может быть достаточно дорогостоящим, но в любом случае он существует.

Ввод данных вручную. Данные вводятся по различного рода экспертным оценкам сотрудниками организации. Такой метод является наиболее трудоемким.

Методы сбора информации существенно отличаются по стоимость и необходимому времени, поэтому следует соизмерять затраты с результатом. Возможно, от сбора некоторых данных придется отказаться, но факторы, которые эксперты оценили как наиболее значимые, нужно собрать и обязательно, несмотря на стоимость этих работ, либо вообще не проводить анализ.

Данные должны быть собраны в таблицы базы данных, в текстовые файлы с разделителями, в файлы MS Excel, то есть должны быть представлены в структурированном виде. Кроме того, необходимо унифицировать представление данных: один и тот же объект должен везде описываться одинаково.

Подготовка данных

Когда первичный сбор данных налажен, возникает следующая проблема. В процессе реализации проектов по бизнес-аналитике оказывается, что исходные данные содержатся в источниках с различными моделями и представлением

данных, файлах разно образных приложений, форматов и стандартов кодирования. Между тем, для эффективного применения различных методов и алгоритмов анализа данных, особенно, использующих машинное обучение, необходимо, чтобы данные были объединены в едином централизованном источнике и имели унифицированный формат представления. Кроме этого, в различных источниках могут располагаться различные элементы данных (например, признаки или атрибуты исследуемых бизнес-процессов), которые должны быть проанализированы совместно.

Поэтому одним из важнейших этапов подготовки данных к анализу является их ***интеграция*** из множества разнородных источников в один, имеющий архитектуру, которая наилучшим образом соответствует целям анализа, а также применяемым методами алгоритмам. Иными словами, задача заключается в представлении совокупности данных из множества независимых источников на основе единой модели данных. При этом множество источников может быть *фиксированным* или *пополняемым,* а сами источники данных могут быть *неизменчивыми* или с *обновляемым контентом.*

В настоящее время разработало большое количество различных подходов, методов и рекомендаций для решения задачи интеграции данных с целью их дальнейшей аналитической обработки приложениями традиционного инструментария BI и Data Mining. Зачастую это приводит к определенной путанице понятий и определений, затрудненной оценке достоинств и недостатков различных подходов и методов, а также границ применения тех или иных архитектурных решений. Поэтому вначале мы рассмотрим основные понятия и классификацию подходов и методов, применяемых в системах интеграции данных (СИД), а также краткую историю проблемы.

Интеграция данных

Задача ***объединения данных*** из разнородных источников известна с 60-х годов прошлого века. Её появление связано с развитием баз данных, а точнее, необходимостью помещать в них данные из различных источников. Важной проблемой является также актуальность данных, поддержание которой требует многократного выполнения процедуры дозагрузки новых данных и синхронизации нового единого источника с текущим со стоянием первичных источников.

В середине 70-х годов XX в. активизировались разработки *распределенных систем баз данных* и сформировались более четкие представления о многоуровневой архитектуре системы управления базой данных и о моделях данных как методе моделирования реальности.

В конце 1990-х – начале 2000-х значительная часть разработок в области интеграции данных касается проблем *семантической интеграции.* Она связана не со структурированием интеграционной архитектуры, а с разрешением семантических конфликтов между разнородными источниками данных. Например, если две компании хотят объединить свои базы данных, некоторые бизнес-понятия и определения в их схемах могут иметь разные значения. Скажем, в одной базе данных товары представлены их наименованиями, а в другой – кодами из справочника. Общая стратегия для решения таких проблем предполагает использование *онтологий,* которые явно определяют условия схемы и таким образом способствуют разрешению семантических конфликтов.

Примерно с 2010 года начали активно проводиться исследования, которые позволяли отказаться от «изоляционизма» в данных, когда каждая модель данных, рассматриваемая при их интеграции, соответствовала модели данных в определенном источнике. Более совершенным подходом к моделированию данных стала разработка моделей, основанных на *унификации структуры метаданных* (о метаданных мы будем говорить позже).

Интеграция данных (ИД) – это процесс объединения данных, находящихся в различных разнородных источниках, в единственном физическом источнике или обеспечение единого унифицированного интерфейса для некоторой совокупности источников (при этом физического объединения данных путем копирования информации не происходит).

Роль интеграции особенно возрастает при увеличении объёмов данных, разнообразии их структур, типов и видов, что чрезвычайно актуально для таких современных технологий, как Data Mining, Data Science и Big Data.

Ранее неоднократно упоминалось понятие ***источник данных.*** В контексте рассматриваемых технологий это ***объект, содержащий структурированные данные.*** Необходимо, чтобы аналитическая платформа могла получать доступ к данным из этого источника непосредственно, либо после их преобразования в совместный формат. В противном случае, очевидно, что объект не может считаться источником данных.

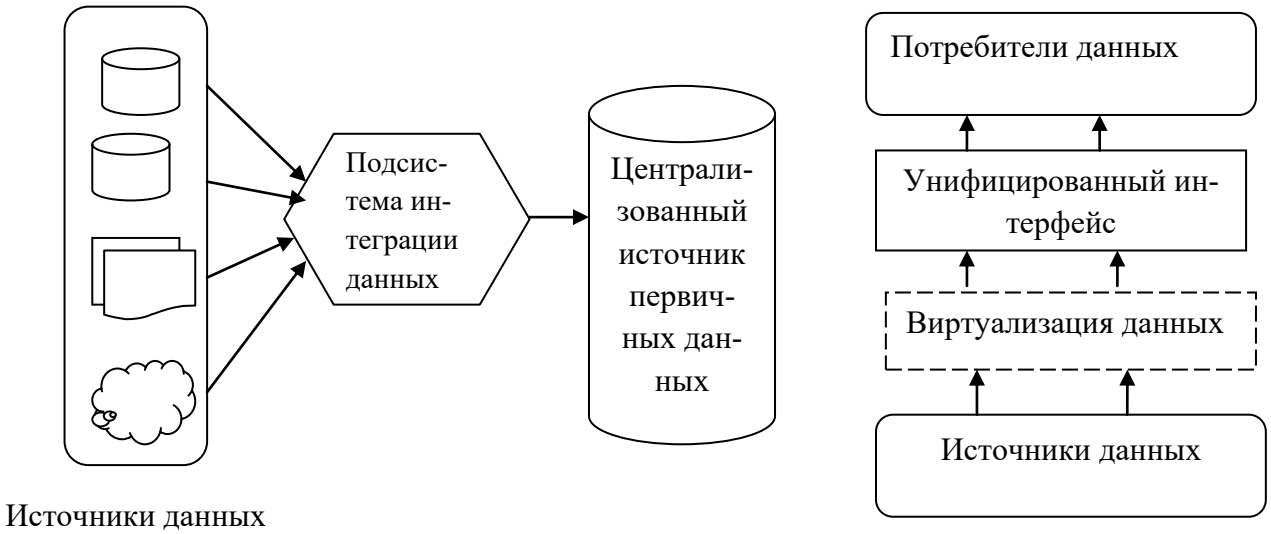


Рисунок 8. Интеграция данных

Источник данных

К последнему случаю относятся, в частности, ***веб-логи*** (специальные файлы, в которые заносятся все действия пользователя на сервере) и текстовые документы. Изначально это неструктурированные данные. Однако к веб-логу можно применить парсер, а из текстового документа извлечь частоты слов, превратив полезную информацию для анализа в структурированный табличный вид.

Обычно руководителям проектов по бизнес-аналитике «с нуля» приходится сталкиваться со следующей ситуацией. Во-первых, данные на предприятии расположены в различных источниках самых разнообразных форматов и типов – в отдельных файлах офисных документов (Excel, Word, обычных текстовых файлах), в учетных и прикладных системах (1С:Предприятие, CRM и другие), в базах данных (Oracle, MS SQL и другие).

Во-вторых, данные могут быть избыточными или, наоборот, недостаточными. А в-третьих, данные являются «грязными», то есть содержат факторы, мешающие их правильной обработке и анализу (пропуски, выбросы, неоднородность форматов, дубликаты).

В настоящее время можно выделить **два основных подхода** к построению систем интеграции данных.

Синтаксический – является более «старым» и основан на внешнем сходстве объединяемый данный. Например, критерием для объединения может служить одинаковый тип или модель данный нескольких источников, их связь с

одним предприятием или бизнес-процессом. Но то, что источники относятся к одному и тому же бизнес-процессу, вовсе не означает, что они совместимы и могут анализироваться вместе.

Действительно, при анализе продаж данные по номерам автомобилей, на которых производился вывоз товара, и фамилий водителей вряд ли будут способствовать улучшению качества прогнозирования продаж. С другой стороны, совместный анализ данных из различных источников, внешне никак не связанных, может дать неожиданные и значимые результаты.

Идея синтаксического подхода сформировалась исторически, когда разработчики первых баз данных нашли удачное, как тогда казалось, решение, отделив данные от контекста и получив возможность работы с ними в чистом виде (так называемое «инженерное» мировоззрение, сложившееся с 1360-х годов: данные отдельно – контекст отдельно).

В итоге родилась концепция: данные – в компьютере, а контекст – в человеке. В рамках нее данные рассматривались как примитивное сырье, наборы битов и байтов, а их содержательная сторона никого не интересовала. Более того, отсутствовало даже внятное определение, что такое данные, и были аморфные высказывания, вроде «данные – это представление фактов и идей в формализованном виде, пригодном для передачи и обработки в некотором информационном процессе» и прочее.

Если данных немного и они несложные, то использование такой концепции не вызывало проблем, но в условиях взрывного роста объемов и многообразия данных она становилась несостоятельной, от чего в наибольшей степени страдали аналитики, которым требовалось видеть не отдельные факты развития анализируемых процессов и явлений, а картину в целом.

Семантический подход к интеграции данных учитывает не только внешнюю, но и содержательную, контекстную сторону данных. Семантическая интеграция основывается на знании и учете природы данных. Например, мы сможем вместе хранить и анализировать информацию о продажах, выраженную в единицах объема проданного товара (штуках, килограммах, метрах и так далее) и денежном выражении, если добавим к ним дополнительные сведения, которые свяжут единицы с денежными суммами.

Такие дополнительные сведения называются **метаданными** – «данными о данных». Метаданные создавали дополнительные сложности, но открывали новые возможности. С помощью метаданных формировался так называемый *се-*

мантический слой, который делал работу с данными более понятной и прозрачной для пользователя.

Основная проблема интеграции без учета семантики данных заключалась в том, что одни и те же объекты данных могли интерпретироваться по-разному. Чтобы избежать этого, возникла необходимость явно выразить семантику и включать ее в данные: данные должны содержать в себе описания собственной семантики.

Семантическая интеграция обеспечивает объединение только тех данных, которые соответствуют или наиболее близки одними тем же сущностям в окружающем мире.

Первые попытки создания систем с семантической интеграцией были предприняты в 1930-х г. Тогда впервые стали использовать понятие *онтологий* в приложении к компьютерным данным.

Рисунок 8. Основные подходы к построению систем интеграции данных

Интеграция данных является одной из наиболее проблематичных областей информационных технологий. Методы, используемые для ее реализации, а также их сложность и особенности зависят от множества факторов: требуемого уровня интеграции, свойств отдельных источников, их количества, разнообразия и динамики.

Всего выделяют **три уровня** и **два способа** интеграции данных.

Различают следующие уровни интеграции:

- **Физический уровень.** Интеграция на физическом уровне является наиболее простой задачей. Данные из различных источников преобразуются к единому формату и сохраняются в одном источнике.
- **Логический уровень.** Данные по-прежнему физически размещаются.

в своих источниках, доступ к ним реализуется на основе некоторой глобальной схемы, отражающей их требуемое совместное представление.

- **Семантический уровень.** Обеспечивает поддержку единого представления данных с учетом их семантических свойств в контексте единой онтологии предметной области.

В настоящее время наибольший интерес представляют два последних уровня, которые позволяют обеспечить лучшие интеграционные решения на больших объемах данных, имеющих сложную структуру.

Также выделяют два способа интеграции:

- **Виртуальный** – реализуется с помощью механизма доступа, который при выполнении запроса пользователя формирует требуемое представление данных непосредственно из источников. Наиболее эффективен, если источники данных являются динамически обновляемыми.
- **Материализованный** (актуальный) – формируется полное физическое представление данных, сосуществующее с источниками, на основе которых оно было получено. Очевидно, что для динамически изменяющихся источников данный подход не удобен, поскольку каждый раз при изменении источника нужно переформировывать физическое представление. Данный подход, в частности, используется в хранилищах и оперативных складах данных, когда они существуют вместе с источниками, данными из которых они были заполнены.

Еще одним аспектом, оказывающим значительное влияние на решение задачи интеграции данных, является *неоднородность источников.* При этом само понятие неоднородности различается в зависимости от используемого уровня интеграции. Для физического уровня наиболее характерна неоднородность форматов данных (требуется их преобразование к единому формату, поэтому, чем выше разнообразие форматов, тем сложнее задача). Для логического уров-

ня следует говорить о разнообразности моделей и схем данных, поэтому при интеграции требуется построить некоторую глобальную модель, соответствующую требуемому представлению. В этом случае задача усложняется не с ростом числа источников и разнообразия их форматов, а с ростом разнообразия моделей данных в них.

При разработке системы интеграции данных обычно требуется решить типичный набор задач, к которым относятся следующие:

- Разработка архитектуры СИД.
- Разработка интегрирующей модели данных, являющейся основой единого пользовательского интерфейса СИД.
- Разработка методов представления моделей данных и построение отображений, поддерживаемых отдельными источниками данных.
- Интеграция метаданных, используемых в системе источников данных.
- Преодоление неоднородности источников данных.
- Разработка механизмов семантической интеграции источников данных.

Тема 6. Интеграция данных и бизнес-аналитика

Быстрые и медленные данные

Данные, накапливаемые в информационных системах компаний, делят на два вида: **быстрые** и **медленные.** *Быстрые* данные поступают непрерывно, сплошным потоком и являются сильно детализированными, поскольку отражают элементарные события в жизни бизнеса. Быстрые данные отражают текущие тенденции в бизнесе, и позволяют принимать оперативные, тактические решения.

Медленными называют данные, которые не являются медленно меняющимися или перемещающимися, а отражают долгосрочные зависимости и закономерности бизнес-процессов, что позволяет использовать их для решения задач стратегического анализа и поддержки принятия решений.

Бизнес-аналитика может использовать как быстрые, так и медленные данные. При использовании быстрых данных, задачи бизнес-аналитики сводятся к построению отчетов, отражающих текущее положение в компании и в ее подразделениях, что позволяет вырабатывать тактические решения (подвезти дополнительный товар, создать оперативный запас на складе).

Медленные данные, являющиеся историческими и хронологическими, больше подходят для задач описательных и предсказательных моделей Data Mining с целью выявления длительных зависимостей и закономерностей, знание которых позволяет принимать стратегические решения (смена ассортимента товаров и услуг, изменение ценовой политики и так далее).

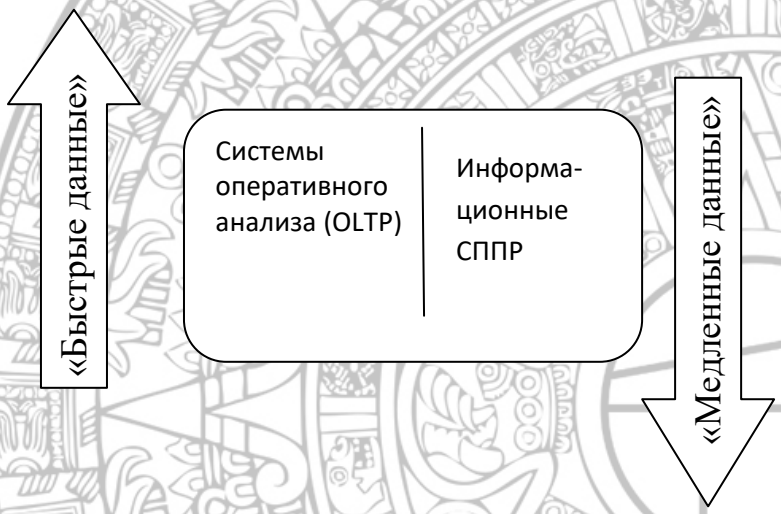


Рисунок 9. Виды данных в информационных системах компаний

Разные цели и задачи работы с быстрыми и медленными данными привели к формированию двух классов систем – оперативного анализа и информационных систем поддержки принятия решений (СППР).

При этом в СППР возникает задача использования данных из множества разнородных источников, расположенных как внутри, так и снаружи предприятия. Для этого данные консолидируются и интегрируются в единый источник, откуда извлекаются с помощью аналитических запросов для анализа. Рассмотрим два класса этих систем более подробно.

Системы оперативного анализа

С середины 80-х годов прошлого века начался период бурного развития и внедрения информационных систем для организации сбора и хранения различного рода бизнес-информации. Как правило, они представляли собой корпоративные системы, предназначенные для оперативной обработки данных, и обслуживали бухгалтерию, архивы, телефонные сети, регистрацию документов, банковские операции, и так далее.

С появлением персональных компьютеров такие системы стали доступными для малого и среднего бизнеса. Системы оперативной обработки информации получили название OLTP (англ.: *On-Line Transaction Processing* – оперативная, то есть в режиме реального времени, обработка транзакций).

Под *транзакцией* в данном случае понимают 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.

Примерами транзакций могут быть: выдача или прием наличных через банковский терминал, продажа набора товаров в супермаркете по одному чеку, бронирование авиа или железнодорожного билета, оплата услуг телекоммуникационных компаний и другие действия при массовом обслуживании клиентов.

Логическое единство операций в транзакции подразумевает, что исключение любой из них делает всю транзакцию бессмысленной. Например, в системе продажи и бронирования авиабилетов из многочисленных пунктов продаж непрерывно стекается информация об уже проданных билетах, которую вводят со своих рабочих мест операторы. В той же базе данных формируется информация о свободных местах. С точки зрения данной задачи транзакция включает в себя набор следующих действий:

- запрос оператора о наличии свободных мест на тот или иной рейс;
- отклик от системы с предоставлением соответствующей информации;
- ввод оператором информации о клиенте, номере заказанного места и оплаченной сумме (возможно, будет присутствовать еще какая-либо служебная вспомогательная информация);
- передача новой информации в базу данных и внесение в нее соответствующих изменений;
- передала оператору подтверждения о том, что операция выполнена успешно.

Исключение любой из перечисленных операций приводит к тому, что связанная с транзакцией бизнес-функция не будет выполнена. Обобщенная схема движения данных в транзакционной системе представлена на рисунке 10.

Транзакции в системах массового обслуживания выполняются десятки и сотни тысяч раз в день в огромном количестве банковских терминалов, пунктов продаж билетов вокзалов и аэропортов, контрольно-кассовых пунктах магазинов крупных торговых сетей.

Рисунок 10. Обобщенная схема движения данных в транзакционной системе

Как следствие, данные в транзакционных OLTP-системах меняются непрерывно, но небольшими порциями. Основным требованием при этом является минимальное время отклика даже при максимальной загрузке системы.

Очевидно, что запросы и отчеты в OLTP системах являются строго *регламентированными* с целью точного выполнения заданной бизнес- функции: оператор системы не может создать собственный запрос с целью получения каких-либо дополнительных сведений, кроме предусмотренных регламентом.

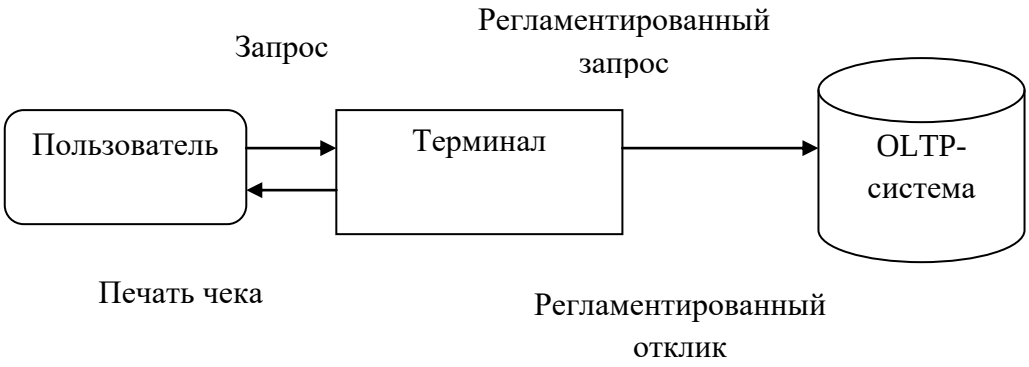


Рисунок 11. Пример транзакции OLTP-системы

Кроме этого, данные в OLTP-системах очень быстро теряют актуальность и устаревают. Действительно, как только рейс завершен, информация о пассажирах и занятых ими местах теряет смысл с точки зрения текущей деятельности компании и, спустя некоторое время, удаляется. Следовательно, историчность данных в OLTP-системах не поддерживается.

Системы поддержки принятия решений

В некоторых случаях транзакционные данные, накапливаемые в базах данных OLTP-систем, не уничтожались после утраты бизнес-актуальности, а сохранялись с целью последующего использования для формирования отчетности для государственных регулирующих органов и внутреннего применения. Как следствие, в компаниях стали накапливаться исторические бизнес-данные. За длительные промежутки времени их могло накопиться достаточно много для того, чтобы организация и поддержка их хранения обходились компаниям достаточно дорого.

Поэтому у руководства компаний закономерно возник вопрос: а нельзя ли использовать эти данные для поиска в них скрытых знаний о бизнес-процессах, которые помогут принимать лучшие управленческие решения. Эту ситуацию можно назвать отправной в появлении и развитии бизнес-аналитики. На рис. 12 показана схема информационной системы поддержки принятия решений.

Например, обладая исторической информацией о количестве чеков и времени их формирования в гипермаркете, можно оптимизировать работу магазина, открыв дополнительные кассы в часы пиковой нагрузки. Однако для принятия такого решения недостаточно использовать регламентированные запросы.

Аналитику могут понадобиться более сложные; *нерегламентированные запросы,* предполагающие некоторую обработку данных. Например, как изме-

нялась динамика визитов клиентов в течение рабочего дня? Или какой была динамика остатков на зарплатных картах клиентов банка?

Такие сложные; нерегламентированные запросы к транзакционным базам данных называются аналитическими, поскольку позволяют делать определенные выводы и заключения и использовать их для поддержки принятия решений.

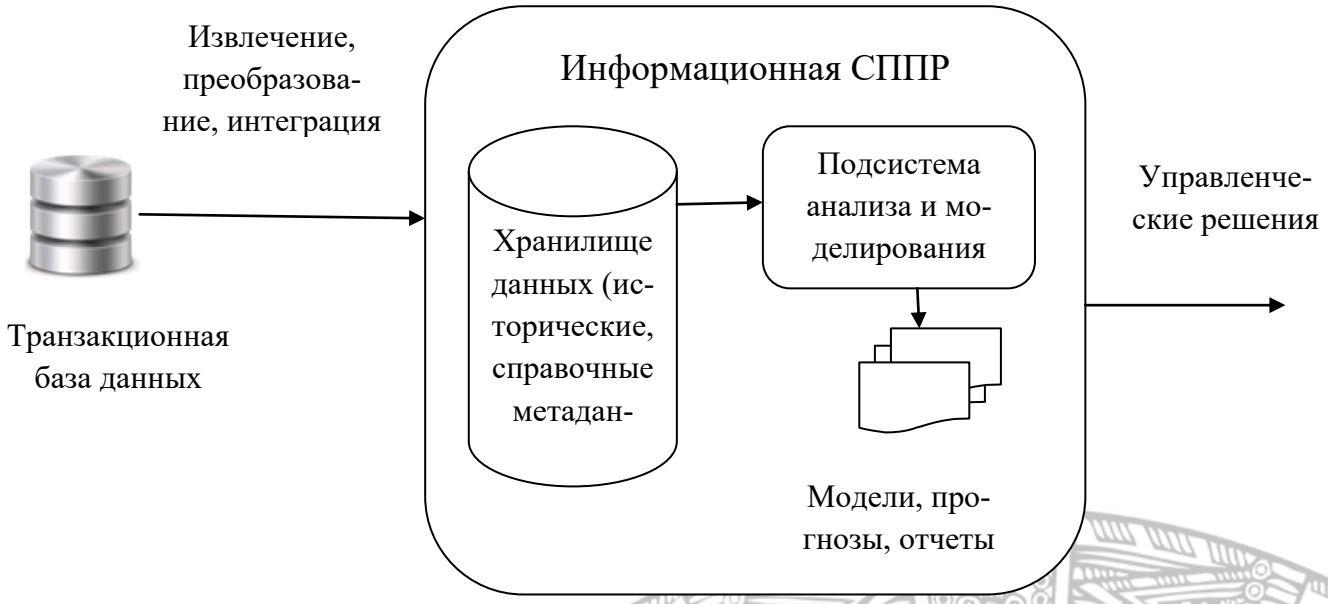


Рисунок 12. Система поддержки принятия решений

При этом возникает вопрос: а оптимальна ли OLTP- система, ориентированная на максимально быстрое выполнение простейших запросов, с точки зрения возможности реализации аналитических запросов, где главное не скорость, а точность, полнота и непротиворечивость отклика? Именно поэтому для задач бизнес-аналитики потребовались другие системы для организации хранения данных, а сами OLTP-системы стали рассматриваться как системы сбора первичных данных.

Осознание бизнесом перспектив, открываемых анализом данных, привело к развитию нового класса систем – информационных систем поддержки принятая решений (СППР), в которых время отклика, хотя и играло важную роль, но не было критичным фактором.

В процессе разработки информационных СППР и методологии их применения обнаружилось, что для эффективной работы данные должны быть организованы несколько иным способов, чем тот, который применяется в OLTP-системах. В частности:

- 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, что не обеспечивается обычными OLTP-системами;
- транзакционные данные в OLTP-системах являются максимально детализированными, что не оптимально с точки зрения анализа. При аналитической обработке предпочтение отдается данным с некоторым уровнем их обобщения. Когда, например, единицей анализируемого показателя является не каждая отдельная покупка, а сумма (среднее, медиана, максимум, минимум) покупок, сделанных за определенный временной период (день, неделю, месяц).

Таким образом, прежде чем подвергаться обработке в СППР, бизнес-данные должны пройти определенную подготовку, включающую:

- **интеграцию** – извлечение и объединение данных из множества разнородных источников в централизованную систему хранения;
- **преобразование** – приведение данных к наиболее удобному для анализа виду (агрегирование, кодирование и так далее).

Кроме интеграции и преобразований наборов данный, почти всегда требуется предварительный их *профайлинг и аудит* с последующей *очисткой* – восстановлением нарушения полноты и целостности данных, исключением из них пропусков, дубликатов, противоречий и других факторов, мешающих их корректному анализу.

Разница между OLTP-системами и информационными СППР

Рассмотрим разницу между OLTP-системами и информационными СППР, сравним их по списку свойств:

Современной тенденцией стало комплексное решение перечисленных задач подготовки данных в системах бизнес-аналитики. В частности, аудит, очистку и преобразование часто рассматривают как часть процесса интеграции, поскольку проблемы в данных могут возникать именно при его неудачном выполнении и должны выявляться и устраняться до загрузки в централизованную систему хранения.

Таблица 9. Разница между OLTP-системами и информационными СППР

	Цели использования данных	Уровень обощения (детализация данных)	Требования к качеству данных	Формат хранения данных	Время хранения
OLTP	Формарование отчетности, простые алгоритмы обработки	Максимально детализированы	«Сырые» данные с ошибками, прорусками и т.д.	Могут храниться в любых форматах	В пределах отчетного периода (как правило, 1-2 года)
СППР	Аналитическая обработка с целью поиска скрытых закономерностей. Построения прогноза и т.д.	Различные уровни детализации (обобщения)	Данные, прошедшие профалинг, аудит, очистку	Хранятся и обрабатываются в едином формате	Годы, десятилетия

	Изменение данных	Приодичность обновления	Доступ к данным	Характер выполнения запросов	Время выполнения запросов
OLTP	Данные могут добавляться, изменяться, удаляться	Часто, но в небольших объемах	Обеспечивается доступ ко всем текущим (оперативным данным)	Стандартные (регулярные), настроенные заранее	Несколько секунд
СППР	Допускается только добавление новых данных; ранее добавленные данные изменяться не должны	Редко, но в больших объемах (в соответствии с регламентом)	Обеспечивается доступ к историческим данным с соблюдение их хронологии	Нерегламентировнные, формируемые аналитиком «на лету»	До нескольких минут

Именно поэтому, говоря об интеграции данных в системах бизнес-аналитики, подразумевают не только само объединение данных из разнородных источников, но и выполнение их очистки и преобразования. Поэтому соответствующие функции включаются в интеграционные системы.

Тема 7. Источники данных

Виды источников данных

В процессе аналитической обработки данных происходит их поэтапное преобразование в соответствии с целями и задачами анализа. Сначала данные извлекаются из различных источников, затем преобразуются и объединяются, подвергаются аудиту, профайлингу и базовой очистке. Таким образом, формируется последовательность наборов данных, каждый из которых является результатом обработки предыдущего. Следовательно, можно указать *первичный, вторичный* и так далее набор данных.

Соответственно и источники данных можно разделить по видам на **первичные** и **вторичные.**

Но прежде чем говорить о видах источников, стоит остановиться на несколько другом аспекте. Как уже отмечалось, одной из проблем, с которыми сталкиваются компании при реализации аналитических проектов, является отсутствие комплексного подхода при разработке систем интеграции, учитывающего разнородный характер данных. В информационных системах компании и внешнем окружении накапливаются не только основные данные, описывающие бизнес-процессы, но и сведения служебного и технического характера. В этой связи по характеру интегрируемых данных их можно разделить на несколько типов:

- фактографические;
- нормативно-справочные;
- метаданные

Фактографические – это данные, отражающие факты, которые описывают процессы, объекты и явления предметной области (часто их называют историческими). Здесь факт (наблюдение, прецедент, транзакция) представляет собой отдельную запись в базе данных. Обычно факт отражает некоторое логически неделимое действие, последовательность которых образует бизнес-процесс в компании. Примерами фактографических данных являются истории продаж, курсов валют, транзакции, данные биллинга и т. д.

Собственно, большинство источников, отражающих деятельность компании и используемых для ее анализа, являются фактографическими. Как правило, фактографические данные являются структурированными, историческими

(накапливаются за определенный период) и хронологическими (порядок следования записей соответствует хронологии фактов).

Нормативно-справочные (англ.: *reference data)* включают различного рода словари (например, терминологические), справочники (адресов, телефонов), классификаторы (ОКПО, ОКАТО), нормативы, кодификаторы, рубрикаторы и так далее.

С точки зрения анализа нормативно-справочные данные носят вспомогательный характер. Непосредственно строить предсказательные модели с их помощью нельзя. Однако их использование может служить для поддержки аналитического процесса, обогащения, восстановления и очистки фактографических данных, формирования отчетов. Например, если город клиента неизвестен, то его можно восстановить по телефонному коду.

Нормативно-справочные данные бывают **внешними** и **внутренними.**

Внешними называются такие нормативно-справочные данные, которые содержат информацию, не относящуюся к бизнес-процессам событиям и явлениям, происходящим внутри предприятия, реализующего аналитический проект. Обычно внешние источники содержат нормативные документы отраслевого и ведомственного характера, отчеты государственных органов статистики и финансового регулирования, курсы валют и ценных бумаг на биржах, информацию из бюро кредитных историй и так далее.

Внутренними называются нормативно-справочные данные, содержащие информацию, циркулирующую внутри компании. Типичными внутренними источниками, используемыми в анализе данных, являются классификаторы товаров и изделий, результаты маркетинговых исследований и так далее.

Следует отметить, что физическая локализация данных не определяет их как внешние или внутренние. Например, отчет о продажах компании, размещенный в общем доступе на сайте в сети Интернет, не становится внешним источником, так как по-прежнему содержит внутреннюю информацию компании.

Для нормативно-справочных данных часто применяют синоним – *нормативно-справочная информация* и соответствующее сокращение – НСИ. Существуют специализированные программные решения для ведения НСИ – так называемые MDM-системы.

Метаданные или «данные о данных» – разновидность данный, носящий служебный характер. Они не отражают течение бизнес-процессов компании, а описывают фактографические и нормативно-справочные данные. Как правило,

метаданные не представляются пользователю в явном виде, однако играют важную роль в процессе анализа данных и их подготовки (в частности, интеграции).

Метаданные содержат информацию о составе данных (например, число страниц документа), содержании (оглавление), статусе, происхождении, локализации, качестве, форматами формах представления, условиях доступа, приобретения и использования, авторских, имущественных и смежных с ними правах на данные и др. В состав метаданных могут входить: каталоги, справочники, реестры, (позже мы также обсудим метаданные хранилищ данных).

Можно выделить **три** основных вида метаданных:

- **технические –** обеспечивают функционирование баз данных, а также выполнение запросов к ним. Примерами технических метаданных являются имена таблиц базы данных и полей в них;
- **бизнес-метаданные –** определяют сущности, хранящиеся в источниках данных, бизнес-термины и определения. Благодаря бизнес-метаданным, пользователь может оперировать при работе с данными привычными терминами предметной области, которые транслируются в соответствующие запросы к источникам;
- **операционные метаданные** – содержат информацию о процессе работы источника данных: происхождение загруженных и преобразованных данных, их статус (активные, архивированные или удаленные), статистику использования, сообщения об ошибках и т. д.

Первичные источники данных

Первичными называются источники, данные в которых являются результатом непосредственной регистрации и измерения характеристик бизнес-процессов, объектов и явлений. Первичные данные могут формироваться как посредством ручного ввода (например, оператор вручную переносит данные из анкеты клиента), или автоматически (например, с использованием сканера штрих-кода в кассовых пунктах супермаркетов).

Первичные данные могут собираться в соответствии с точным пониманием целей и задач анализа (собираются только те данные, которые гарантированно будут анализироваться). Но иногда собираются все доступные данные, что в перспективе делает возможным расширение спектра решаемых задач анализа (хотя это может привести к дополнительным временным, финансовым и ре-

сурсным затратам). Методология сбора первичных данных может быть заранее известной или может производиться «как есть».

Первичные источники, как правило, не используются для аналитической обработки данных и целей Data Mining, поскольку данные в них являются «сырыми» (англ.: *Raw Data,* необработанные данные, полученные в ходе регистрации бизнес-процесса). Поэтому они являются источниками для *вторичных* источников данных.

Можно выделить четыре способа сбора первичных данных.

Опрос – систематический сбор информации от респондентов посредством личных контактов с ними, по телефону, почте или через Интернет. Опрос может дать сведения об отношениях респондента к тем или иным явлениям социально-экономической жизни, мнение о товарах и услугах компании, готовности воспользоваться новыми товарами и услугами на рынке и так далее. Для фиксации ответов используются вопросники или анкета.

Если опрос производится с помощью электронный средств, то соответствующие наборы данный могут формироваться автоматически, если с помощью «бумажного» анкетирования, то результаты затем должны переноситься операторами в электронную форму.

С точки зрения разработки стратегии подготовки данных знание способа имеет значение. Вероятность ошибок и несоответствий при «бумажном» методе возрастает, поскольку респондент может неразборчиво указать сведения в анкете, оператор может допустить опечатку при ручном вводе и так далее. При полностью компьютеризованной процедуре опроса вероятность проблем в данных меньше.

Наблюдение – метод, с помощью которого изучают и фиксируют реальное поведение бизнес-процессов, как текущее, так и ретроспективное. Наблюдение может быть открытым или скрытым.

Эксперимент – имеет место тогда, когда один или несколько параметров бизнес-процесса изменяется, а один или несколько других – контролируется. Например, руководство супермаркета может изменять число работающих кассовых пунктов, регистрируя при этом среди ее количество ожидающих очереди покупателей. Недостатком эксперимента является некоторая искусственность условий его проведения, а также невозможность учета и контроля всех изменяющихся параметров.

Имитация – метод, основанный на применении компьютерных моделей. Вначале строится модель контролируемых и неконтролируемых факторов. Для имитации не требуется сотрудничество со стороны клиентов, и она позволяет учитывать множество взаимосвязанных показателей. Однако имитация сложная сильно зависит от положенных в основу модели предположений.

Виды источников первичных данных

Рассмотрим основные виды систем для сбора и хранения первичных данных. Первый из них – это *OLTP-системы и базы данных.* На первых мы подробно уже останавливались ранее, заметим только, что практически все современные OLTP-системы основаны на технологиях баз данных. Кроме того, первичными источниками могут выступать и базы данных, не регистрирующие транзакционную информацию. В основном к ним относятся разнообразные источники внешней и внутренней нормативно-справочной информации, структурированные массивы маркетинговых данных и тому подобное.

Напомним, следует различать понятия *базы данных* и *системы управления базами данных* (СУБД). **База данных –** это собственно файл, который содержит данные в определенном формате, обычно, в виде таблиц. Но базы данных масштаба предприятия содержат не одну таблицу, а десятки, сотни и даже тысячи. Очевидно, что «ручное» манипулирование отдельными таблицами в этом случае неэффективно. Поэтому для автоматизации процесса управления данными в базах применяют специальные программные средства – **СУБД.** С помощью СУБД осуществляется организация хранения, поиска и доступа к данным.

Второй вид первичного источника – так называемая *унаследованная информационная система.* Процесс разработки и внедрения информационных ресурсов для хранения и обработки данных, а также накопления самих данных в некоторых компаниях может длиться годами и даже десятилетиями. Очевидно, что за это время информационные системы и технологии, на которых они построены, могут устаревать морально и технически. Например, до сих пор можно столкнуться с ситуацией, когда организация использует базы данных на основе «плоских» таблиц в формате DBF, с которыми работали такие приложения, как dBase, FoxBase, FoxPro и другие. В современных условиях использование таких систем неэффективно.

Причин такого «зависания» в прошлом может быть множество. Например, консерватизм, когда люди, стоявшие у истоков системы, считают ее оптимальным решением в конкретном случае. Отсутствие средств, непонимание руководством роли современных информационных технологий в борьбе за конкурентные преимущества также могут быть в списке причин.

Сложно найти корпорацию возрастом больше 25 лет, в которой не использовались бы информационные подсистемы, созданные на основе ранних аппаратно-программных платформ компании IBM. Базы данных таких подсистем содержат громадные объемы потенциально ценной информации, и компания просто не может обойтись без их использования. С другой стороны, такие системы очень трудно сопровождать и поддерживать.

Такие системы называют унаследованными.

Для того чтобы эффективно использовать данные, накопившиеся в унаследованных системах, компании стремятся перевести их на новые технологии хранения и обработки данных. Но часто возникает проблема, что работоспособность унаследованной системы может быть настолько важна для компании, что эту систему нельзя вывести из использования даже на короткое время.

Типичным подходом к модернизации унаследованных систем является поэтапная замена элементов старой системы на элементы новой.

В процессе модернизации вокруг старой системы создается оболочка новой системы – оборудования, программного обеспечения, интерфейса. Затем поэтапно производится миграция данных из старых структур в новые, до тех пор, пока не произойдет их полная замена. В некоторых случаях старую систему оставляют функционировать параллельно с модернизированной.

Современной тенденцией является экспоненциальный рост объемов данный. В этих условиях ресурсы информационных систем компаний могут оказаться недостаточными для хранения всей информации, представляющей потенциальный интерес для анализа. Решением проблемы может быть использование ресурсов, предоставляемых сторонней организацией.

Но поскольку клиентов у такой компании может быть очень много и территориально они распределены по всему миру, логично организовывать хранение на большом количестве компьютеров, распределенных в сети Интернет, то есть компания, реализующая такие сервисы, «сдает» всем желающим «площади» для хранения данных.

Основной особенностью такого рода систем является то, что пользователь не видит ее внутреннюю структуру, которая как бы скрыта от него в «облаке». Именно поэтому называют **облачными.** Использование облачных источников данных имеет свои преимущества и недостатки. К **преимуществам** относятся:

- отсутствие необходимости приобретать и содержать собственную аппаратную и программную инфраструктуру;
- оплата пользователем только того объема хранения, который реально занимают его данные, а не сервера целиком;
- процедуры по резервированию и сохранению целостности данных производятся провайдером облачного сервиса, освобождая клиента от этих задач;

Но вместе с тем, можно выделить и **недостатки** облачного подхода:

- возникают проблемы безопасности при пересылке данных;
- из-за необходимости пересылки данных и большой нагрузке на облачный сервис, как правило, скорость доступа к данным ниже, чем при использовании локальных источников;
- из-за технических проблем, провайдера облачного сервиса или нарушения работы каналов связи данные могут оказаться недоступными.

Кроме корпоративных информационных систем, в которых задачи сбора, хранения и обработки данных решаются централизованно, с помощью специализированного программного обеспечения, баз данных, эксплуатируемых квалифицированным персоналом, большие объемы данных могут накапливаться в отдельных файлах и документах работников и подразделений организации. Основной причиной использования таких файлов является простота ввода и редактирования информации.

Файлы данных отдельных пользователей могут иметь значительный интерес с точки зрения анализа и обнаружения скрытых закономерностей в бизнес-процессах и явлениях.

Более того, существует категория аналитиков, которые специально занимаются исследованием файлов отдельных пользователей, будучи уверенными, что именно там могут быть обнаружены наиболее интересные и полезные знания. Эта уверенность основана на двух фактах:

- никто лучше работников на местах не знает всех особенностей бизнес-процессов, поэтому они могут собирать и представлять данные именно

таким образом, чтобы они наилучшим образом отражали наиболее существенные аспекты бизнеса;

- данные, собираемые в централизованных базах данных, зачастую проходят через «призму инженерного мировоззрения», и их представление приводится к оптимальному виду с точки зрения IT-специалиста, но не с точки зрения бизнес-аналитика;

Именно поэтому «копание» в локальных файлах отдельных пользователей помогает обнаружить самые неожиданные аспекты бизнеса.

Главной проблемой работы с локальными файлами пользователей является разнообразие приложений и форматов, на основе которых они были созданы. Работник, которого никто не регламентирует в плане выбора приложения, воспользуется тем, что есть под рукой или тем, что более привычно.

В простейшем случае, это могут быть текстовые файлы, в которых пользователь выровнял колонки таблиц с помощью табуляции или пробелов (текстовые файлы с разделителями). Достаточно популярны для сбора данных форматы текстовых редакторов (MS Word, OpenOffice Writer, Pages и другие), табличные процессоры (MS Excel, OpenOffice Calc, Kcells и т. д.). Нередко данные размещаются в HTML и XML-документах. Разнообразие приложений и форматов усложняет задачу интеграции пользовательских данных в корпоративные системы.

Еще одной важной проблемой локальных файлов пользователей является *низкое качество данных.* В офисных приложениях отсутствуют средства контроля структуры и целостности данных (ничто не мешает пользователю добавлять неполные записи, размещать в одном столбце значения различных типов), единые требования к именам столбцов, поддержка уникальности и непротиворечивости записей. А первичный ввод данных вручную ведет к появлению большого числа ошибок и опечаток.

Поэтому, не смотря на то, что пользовательские файлы и документы могут представлять значительный интерес для анализа, процесс их использования при интеграции данных может оказаться трудоемким.

Вторичные источники данных

Вторичными являются источники, которые получают данные не в процессе их сбора и регистрации, а из первичных источников. Вторичные источники

являются источниками данных для аналитических систем и бизнес-приложений, а также других вторичных источников.

Перенос данных из первичных источников во вторичные (а также из одних вторичных источников в другие), реализуется с помощью программно-аппаратного комплекса, называемого ETL (англ.: *Extract, Transform, Load* – извлечение, преобразование, загрузка). Иногда так же называют и процесс переноса данных из источника к получателю. ETL извлекает данные из различных источников, выполняет их преобразование (очистку, преобразование к формату, совместимому с получателем данных), а также загрузку данных. Если данные извлекаются из множества источников, имеющих различные форматы, то в ETL может производиться их интегрирование.

В корпоративных информационных системах используются несколько типов вторичных источников. Рассмотрим их подробнее.

Область временного хранения (англ.: *Staging Area)* – используются для промежуточной обработки (очистки, трансформации) и синхронизации данных, поступающих из различных источников. Синхронизация необходима потому, что в некоторых случаях данные, особенно из удаленных источников, могут запаздывать. Поэтому «опережающие» данные будут ожидать интегрирования с «опаздывающими» в областях временного хранения.

Оперативный склад данных (англ.: *Operational Data Stone)* – база данных, в которой хранятся оперативные данные – данные реального (или почти реального) времени, используемые для оперативного (тактического) анализа данных с целью поддержки принятия решений. Синоним – *транспортная база данных.*

Хранилище данных (англ.: *Data Warehouse,* ХД) – предметно-ориентированный, интегрированный, неизменяемый, хронологический источник данных, специально разработанный для подготовки отчетов и анализа с целью поддержки принятия решений в организации. Большинство ХД обеспечивают высокую скорость обмена данными с аналитическими приложениями, автоматически поддерживают целостность и непротиворечивость данных. Главное преимущество ХД перед остальными типами источников данных – наличие *семантического слоя,* который дает пользователю возможность оперировать терминами предметной области для формирования нерегламентированных запросов к хранилищу

Витрина данных (англ.: *Data Mart)* – массив тематической информации, ориентированный на пользователей одной рабочей группы или подразделения компании. Может заполняться как на основе хранилища данных, так и непосредственно из первичных источников.

Говоря о хранилище данных, часто выделяют:

- **корпоративное ХД,** содержащее все данные организации. В этом случае хранилище является центром информационной инфраструктуры организации, и на него замыкаются основные потоки данных, циркулирующие внутри нее. Альтернативой может быть использование нескольких хранилищ, созданных для отдельных подразделений или бизнес-процессов;
- **центральное ХД** – также содержит все данные организации и замыкает потоки данных, но при этом является доступным всем пользователям внутри компании. Альтернативой является организация доступа к ХД только отдельных категорий пользователей;

Вообще, четкой границы между понятиями корпоративного и центрального хранилищ данных не существует. Более того, некоторые авторы не делают различия между ними. Иногда в понятие корпоративного ХД включают обслуживающую его информационную инфраструктуру, а центральное хранилище полагают ее частью.

При загрузке данных в ХД средствами ETL принимают все возможные меры, чтобы данные были очищены *от пропусков, выбросов, дубликатов* и *противоречий.* Поэтому ХД является наиболее «чистым» источником данных в корпоративной информационной системе и является наиболее предпочтительным с точки зрения поддержки процесса анализа данных и формирования отчетности.

Отметим, что информационная инфраструктура компании не обязательно содержит все перечисленные виды вторичных источников. В следующих Темах мы рассмотрим свойства вторичных источников, их роль и место в корпоративной информационной системе компании более подробно.

Тема 8-9. Интеграция данных

Методы и интеграция данных

Исторически сначала появился метод интеграции **точка-точка.**

В процессе интеграции источников связь между ними может осуществляться непосредственно, или с использованием некоторой системы – посредника (медиатора или информационного хаба), которая обеспечивает согласование частный схем данный источников с глобальной схемой хранилища, выполнение всех необходимых преобразований данных, а также единый интерфейс пользователя.

Если же интеграция данных производится без использования посредника, то говорят, что имеет место интеграция типа «точка-точка» (англ. *point-to-point).* В этом случае каждый источник и получатель данных должны «знать» о существовании друг друга.

Таким образом, при использовании метода «точка-точка» поток данных перемещается непосредственно от источника к получателю, минуя какие-либо промежуточные системы. Такая схема является наиболее простой, но число потоков данных может экспоненциально возрастать при подключении новых источников. Кроме этого, затруднена модернизация таких систем.

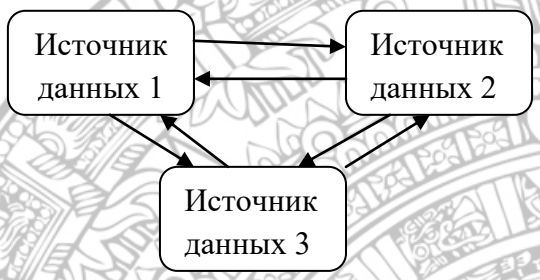


Рисунок 13. Метода «точка-точка

Консолидация имеет место в том случае, когда интеграция производится на физическом уровне с помощью материализованного представления. При этом данные извлекаются (англ.: *extract)* из источников, преобразуются к единому формату (англ.: *transform) и* загружаются (англ.: *load)* в консолидированный источник данный (корпоративное хранилище или оперативный склад данных).

Таким образом, в процессе консолидации происходит физическое перемещение (копирование) данных. Создается полное материализованное представ-

ление интегрированных данных, отчужденное от исходных источников и сосуществующее с ними.

Кроме этого, следует отметить «однонаправленный» характер консолидации – данные из распределенных источников перемещаются в ХД, но их обратное распространение не предусмотрено.

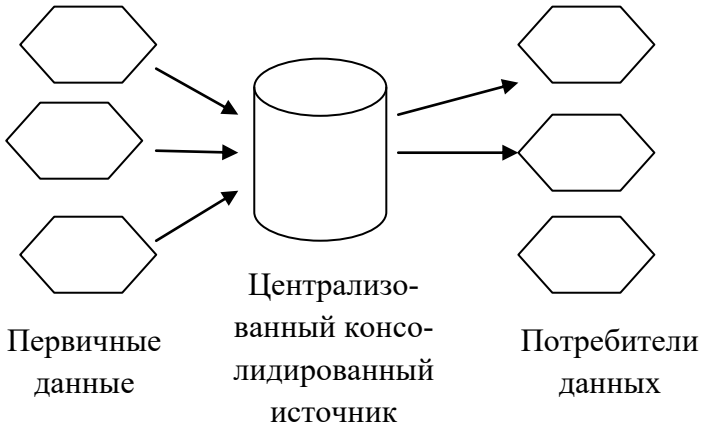


Рисунок 14. Консолидация данных

Недостатки подхода очевидны:

- совместное существование источников и консолидированного ХД удваивает требования к ресурсам дисковой памяти;
- состояния консолидированного ХД практически невозможно синхронизировать с текущим состоянием источников, поскольку данные в ХД всегда будут появляться с задержкой. Таким образом, в ХД могут быть периоды неактуальности, когда анализ данных может дать искаженные, относительно текущего положения дел, результаты;
- интеграция новых источников данных проблематична, поскольку для нее потребуется изменять весь процесс ETL и структуру метаданных ХД.

Преимуществами консолидации являются:

- физическое наличие ХД повышает устойчивость системы интеграции данных к сбоям и нарушениям в работе оборудования. В частности, если источники перестанут быть доступными из-за проблем с сетью, данные в консолидированном ХД будут по-прежнему доступны (хотя и без возможности актуализации);
- при использовании ХД больше возможностей для поддержания *целостности, непротиворечивости* и *качества данных.*

Кроме метода «точка-точка» существует более современный подход к интеграции информационных систем – сервисный подход (англ.: SOA – Service Oriented Architecture). В этом случае система интеграции строится на основе сервисов данных, которые образуют; своего рода, семантический слой (уровень абстракции), который разделяет бизнес-логику и механизмы передачи и преобразования данных. На рис. 14 показана схема сервисного подход к интеграции информационных систем.

Под сервисами данных понимают определенный вид технологий для доступа к произвольным источникам данных. Они образуют еще один системный уровень абстракции, скрывающий от бизнес-приложений физические характеристики источников данных и механизмы доступа к ним.

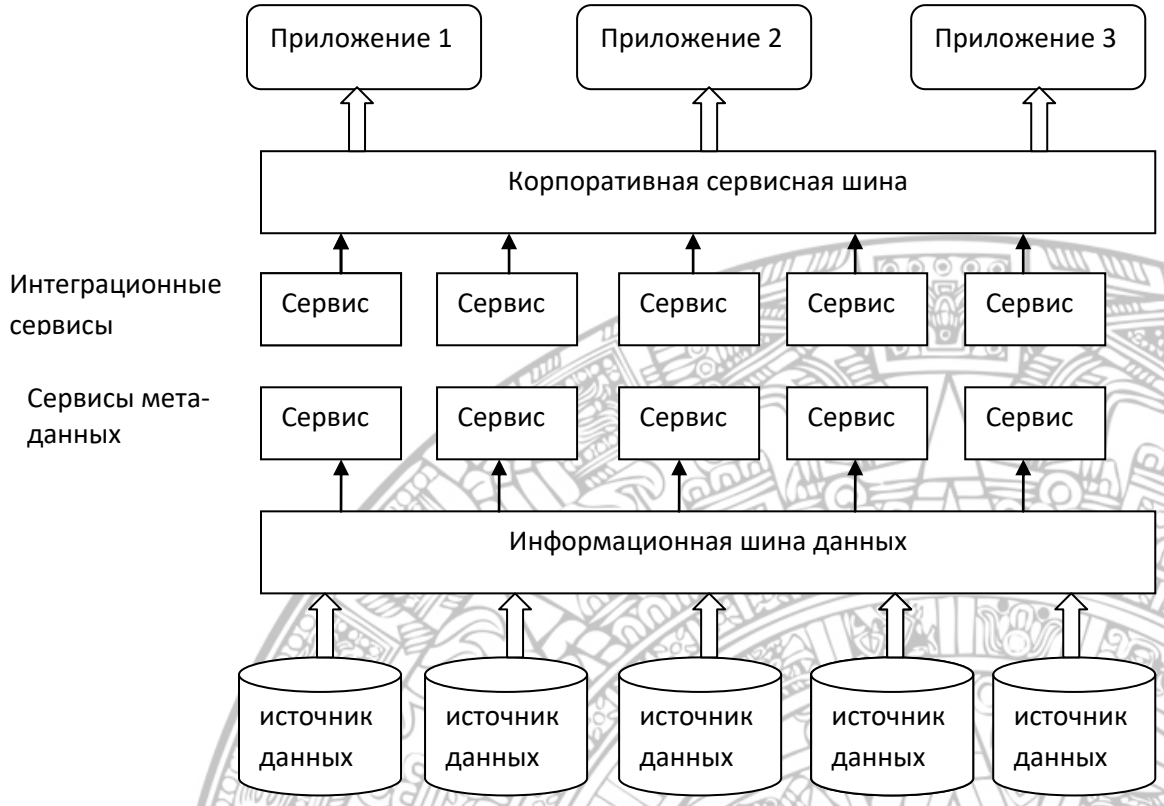


Рисунок 14. Сервисный подход к интеграции информационных систем

Преимуществами сервисного подхода являются:

- один и тот же сервис может использоваться для различных бизнес-процессов;
- оперативность изменения – изменения в одном сервисе распространяются на все бизнес-процессы, в которых этот сервис использовался;

- масштабируемость – благодаря возможности распределения вычислительной нагрузки сервисно-ориентированные системы более устойчивы к пиковым нагрузкам.

Недостатки:

- высокая сложность интеграции с внешними приложениями, не предоставляющими сервисов для доступа к данным;
- для эффективного информационного взаимодействия все ресурсы, входящие в состав схемы, должны быть хорошо структурированными;
- некоторые реализации SOA накладывают существенные ограничения на объемы передаваемых данных.

Таким образом, если интеграционные архитектуры других типов ориентированы на приложения, то системы, основанные на SOA, ориентированы на бизнес-процессы.

Компоненты корпоративной информационной фабрики

Современная высококонкурентная бизнес-среда требует от компаний непрерывного поиска путей повышения их эффективности. При этом только улучшение технологий и снижение себестоимости товаров и услуг не обеспечивает достаточного уровня поддержки конкурентоспособности. Выпуск большего объема продукции за меньшую стоимость, что еще в недавнем прошлом считалось основой конкурентоспособности, в настоящее время не гарантирует успеха.

Чтобы успешно конкурировать, предприятия должны сосредотачиваться не на улучшении отдельных решений, а формировать комплекс возможностей, которые обеспечат лучшую систему управления и поддержки принятия управленческих решений в целом. Однако если рассматривать работу IT-отделов компаний, можно обнаружить, что они уже активно заняты: созданием и внедрением систем, которые обещают повысить конкурентоспособность: хранилищ, и оперативных складов данных, витрин данных, многомерных и реляционных СУБД, приложений BI и Data Mining.

Часто такие решения носят «точечный» характер с целью получить какие-то частные сиюминутные выгоды за минимальное время. Между тем, для достижения стратегических целей, направленных на долгосрочную перспективу элементы информационной инфраструктуры компании должны объединяться

на основе некоторой модели, или, как принято говорить, информационной экосистемы.

Информационная экосистема – это система, состоящая из нескольких компонентов, каждый из которых обслуживает некоторое сообщество людей с целью формирования единой и сбалансированной информационной среды. Со временем баланс и отношения между компонентами экосистемы меняются под воздействием внешних факторов. Поэтому возможность быстрой адаптации и восстановления баланса является признаком эффективной информационной экосистемы.

Таким образом, информационная экосистема позволяет компаниям дополнять традиционные оперативные системы возможностями бизнес-аналитики и управления бизнесом. Кроме того, она обеспечивает комплексную модель для осмысления и использования разнообразных информационных технологий, которые трансформируют нашу информационную парадигму Физическим воплощением информационной экосистемы является корпоративная информационная фабрика.

Термин корпоративная информационная фабрика был впервые введен Б. Инмоном в начале 1980-х (англ. Corporate Information Factory – CIF)

Корпоративная информационная фабрика имеет с одной стороны, достаточно общую структуру, и в то же время является уникальной для каждой компании. Структуру CIF удобно представлять в виде нескольких уровней, на каждом из которых расположены наборы систем, реализующих похожие функции. Обычно выделяют **шесть** уровней.

Уровень источников данных. Включает разнообразные источники первичных данных, таких, как OLTP и унаследованные системы, офисные документы, базы данных, файловые архивы, любые файлы, содержащие структурированные данные.

Уровень извлечения, преобразования и загрузки данных. Программно-аппаратный комплекс, реализующий извлечение данных из различных источников, преобразование к единому формату и загрузку в интегрированное хранилище.

Уровень хранения данных. Обеспечивает надежное, защищенное от несанкционированного доступа, хранение данных.

Уровень распределения данных. Выполняет предоставление данных из хранилища различным потребителям. При этом решается задача выборки дан-

ных, преобразования их структуры в форму наиболее удобную для потребителя (реструктуризация), а также физической передачи данных потребителю.

Уровень предоставления данных. Содержит источники данных для конечных пользователей, которым предоставляются данные. В большинстве случаев это различные витрины данных, но они могут и отсутствовать (тогда пользователи напрямую обращаются к источникам уровня хранения).

Уровень бизнес-приложений. Содержит приложения, реализующие различные виды анализа данных, формирующие отчетность и решающие задачи автоматизации управления бизнес-процессами компании.

Следует отметить, что в зависимости от специфики компании, некоторые уровни могут содержать дополнительные компоненты, а некоторые из представленных, напротив, могут отсутствовать.

Не все они являются необходимыми для конкретной системы: их состав определяется целями и задачами, решаемыми с ее помощью. Например, если компания не использует оперативный анализ данных с целью поддержки тактических решений, то включение в систему оперативного склада не имеет смысла.

Далее мы рассмотрим наиболее типичные составляющие корпоративной информационной фабрики, обеспечивающие решение всего спектра задач бизнес-аналитики и начнем с уровня хранения и предоставления данных. Поскольку центральное хранилище данных и витрины данных будут изучаться отдельно, сейчас обсудим подробнее остальные компоненты присутствующие на данном уровне.

Репозиторий нормативно-справочной информации (НСИ)

Одним из видов первичных данных, играющим важную роль в процессе анализа, является нормативно-справочная информация или НСИ. Традиционно сложилось мнение, что центральное хранилище данных должно содержать только бизнес-данные и метаданные. Поэтому некоторые компании начинают разработку и планирование CIF без выделения средств и ресурсов наведение НСИ. В то же время, попытка внесения изменений и доработок уже на этапе внедрения системы, неизбежно сказывается на сроках, бюджете и качестве проекта.

При интеграции нескольких источников несогласованность НСИ может вызвать серьезные проблемы. Например, в справочнике телефонов одной сис-

темы номера представлены через тире, а в другой – сплошные; в одном названия компаний представлены с полным указанием организационно-правовой формы, а в другой с сокращенным. Кроме этого НСИ, хотя и редко, но подвержены изменениям, что вызывает проблему синхронизации НСИ различных систем.

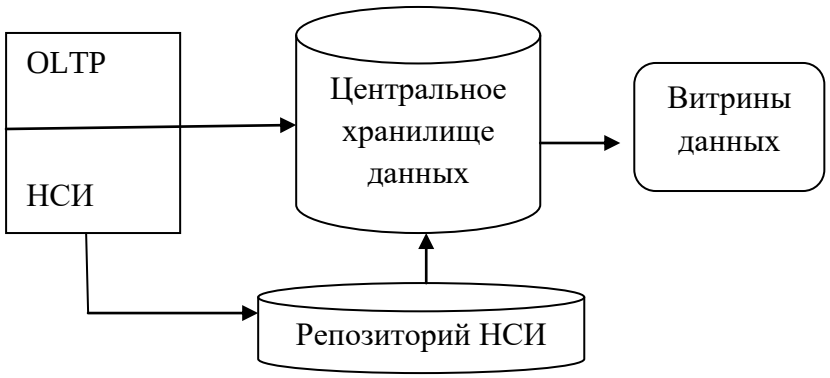


Рисунок 15. Централизованное хранилище данных

Построение информационных систем большинства организаций производилось поэтапно, подразделения компаний переходили на использование компьютеров в разное время, используя при этом разнообразные аппаратные и программные средства. Это приводило к тому, что НСИ этих подразделений также формировались с использованием разнообразных средств, на основе различных стандартов, форматов и кодировок.

Сама по себе, такая ситуация не вызывает проблем но только пока информация используется внутри подразделения и не выходит на корпоративный уровень. И как только это происходит, несогласованность форматов НСИ может вызвать значительные трудности при интегрировании данных в CIF.

Именно поэтому целесообразно создать централизованный репозиторий НСИ, а не хранить их распределено.

Это даст следующие преимущества:

- хранение НСИ в рамках единой модели, согласованной с остальными компонентами системы;
- ведение НСИ на основе корпоративных и отраслевых стандартов, классификации и кодирования;
- обеспечение единого регламента и технологической среды для доступа к НСИ, а также ведения экспертами классификаторов и справочников;

- поддержание необходимого уровня безопасности НСИ и их синхронизации, исключение дублированной, ошибочной и противоречивой информации;
- возможность внедрения классификаторов и справочников НСИ в действующие управленческие, аналитические и другие системы, что позволяет сократить расходы на ведение НСИ;
- оперативность использования НСИ для формирования отчетов.

Мастер-данные

Итак, в корпоративной информационной фабрике присутствует несколько видов данных транзакционные, бизнес-данные, нормативно-справочные и метаданные. Бизнес-данные отражают лечение бизнес-процессов в компании (например, историю продаж, закутою) и используются для поддержки принятия решений, а нормативно-справочные – содержат набор допустимых значений, используемых в таблица бизнес-данных.

Таким образом, непосредственно в процессе анализа оказываются задействованы бизнес-данные и нормативно-справочные. В зарубежной литературе их часто объединяют под одним термином – основные данные или мастер-данные (англ. Master Data).

Следует отметить, что четкой границы, отделяющей мастер-данные от остальных видов данных в CIF вообще говоря, не существует Действительно, некоторые авторы отождествляют мастер-данные с нормативно-справочными. Другие предлагают относить к мастер-данным любые нетранзакционные данные.

Принято считать, что мастер-данные описывают различные бизнес-объекты, и это описание должно быть согласованным, единообразным по всему предприятию (идентичные бизнес-сущности должны интерпретироваться одинаково независимо оттого, в каком бизнес-процессе они задействованы).

Между тем, транзакционные данные описывают элементарные факты: деятельности организации, а для того, чтобы перейти к описанию бизнес-объектов, они должны быть преобразованы – агрегированы, интегрированы и так далее, т. е. приведены к бизнес-данным.

Репозиторий метаданных содержит бизнес- правила и определения, терминологию, информацию о происхождении данных и алгоритмах их обработки, описанные на языке бизнеса, описания таблиц и полей (атрибутов) и другую

служебную информацию. Подробное изучение данной темы выходит за рамки курса бизнес-аналитики.

Не являясь транзакционными по своей природе, мастер-данные часто поддерживают их обработку и преобразование. В некоторых случаях к мастер-данным относят и метаданные. Мастер-данные играют большую роль в информационной фабрике, поскольку от них существенно зависит корректность отчетности и результатов анализа. Поэтому в информационных технологиях в настоящее время возникло и активно развивается направление, связанное с управлением основными данными (англ.:Master Data Management – MDM).

Мастер-данные могут храниться как в центральном репозитории, где они интегрируются из различных источников, так и распределено, при этом доступ к ним производится по ссылкам. Кроме НСИ, чаще выделяют следующие виды мастер-данных:

Мастер-данные предприятия (англ. Enterprise Master Data) – отдельный источник основных бизнес-данных, используемых во всех системах, приложениях и бизнес процессах внутри предприятия (сек отделах, подразделениях, дочерних компаниях и странах).

Рыночные мастер-данные (англ. Market Master Data) – отдельный источник основных бизнес-данных внутри определенного сегмента рынка. Являясь, по сути, внешними данными, рыночные мастер-данные должны быть совместимыми со специфическими для данной предметной области системами.

Материальные мастер-данные (англ: Material Master Data) – данные о запасных частях, сырье и продуктах, используемых в системах планирования материальных ресурсов предприятия.

Возможны и другие виды мастер-данных, отражающих специфики их использования в определенных предметных областях.

Оперативный склад данных

Как уже обсуждалось, все бизнес-данные можно разделить на оперативные и исторические, или быстрые и медленные. Анализ больших интервалов исторических данных позволяет выявлять глобальные зависимости и тенденции, знание которых дает возможность разрабатывать стратегические решения. Между тем, бизнес-процессы постоянно подвергаются воздействию различных внешних и внутренних факторов, которые заставляют их флуктуировать отно-

сительно главной тенденции. При стратегическом анализе такие флуктуации рассматриваются как мешающие факторы и обычно подавляются.

Однако с точки зрения оперативного анализа, когда требуются некоторые тактические решения, направленные на ближнесрочную перспективу представляет интерес исследование локальных закономерностей. Действительно, если в глобальном масштабе продажи растут, что позволяет говорить о перспективах расширения бизнеса то локальные падения продаж, вызванные некоторыми неизвестными факторами, действующими на коротких временных интервалах, могут принести серьезные потери.

Тем не мене, если удастся выявить такие краткосрочные закономерности и разработать соответствующие тактические решения, можно минимизировать потери. Но для этого требуется анализировать не все исторические данные, а только текущие, или взятые за некоторый относительно короткий ретроспективный интервал, то есть оперативные.

Поэтому при интеграции данных целесообразно перед их поступлением в централизованный источник данных создать дополнительную структуру, в которой сохраняются и анализируются оперативные данные с целью выработки тактических бизнес-решений. Такие источники входят в архитектуру практически любой информационной фабрики CIF Это оперативные склады данных (англ.: Operational Data Store, OSD, ОСД), идея которых была предложена Е. Инмоном в 1999 году.

Таким образом, оперативный склад данный – это элемент архитектуры корпоративной информационной фабрики, содержащий объектно-ориентированную, интегрированную и готовую к использованию информацию реального (или почти реального) масштаба времени, не являющийся первичным источником.

Преимуществами использования ОСД являются:

- обеспечение быстрого доступа к важным оперативным данным;
- при использовании ОСД компания всегда имеет представление о текущих параметрах бизнес-процессов и успешности текущей деятельности;
- повышается эффективность формирования оперативных отчетов и снижается нагрузка по запросам к первичным информационным источникам;
- при наличии центрального ХД в него ускоряются процессы загрузки, поскольку часть данных уже находится в ОСД;
- повышаются возможности администрирования: можно

разграничить права аналитиков, занимающихся тактическим анализом оперативных данных из ОСД и аналитиков, проводящих стратегический анализ исторических данных из ЦХД.

Существуют и другие преимущества использования ОСД, связанные с различными аспектами его роли и места в корпоративной информационной фабрике.

Данные оперативного склада регулярно обновляются. Каждый раз, когда данные изменяются в оперативных системах и внешних источниках, соответствующие им данные из оперативного склада также должны быть изменены.

Зоны временного хранения

Зоной временного хранения (англ: *Staging Area*) называют буферную базу данных, необходимую для выполнения некоторых внутренних служебных технологических операций над данными, перемещаемыми из источников в ЦХД. Например, в процессе ETL не только извлекаются данные из источников, но и выполняются различные преобразования над ними: приведение к единому формату, кодирование, проверка значений на непротиворечивость и так далее.

Эти действия выполняются не в источниках, и не в ЦХД, поэтому процесс ETL должен использовать какие-то промежуточные места для хранения данных. Это и есть зоны временного хранения.

В более широком понимании, зонами временного хранения являются любые области хранения данных, предназначенные для выполнения операций внешними пользователями или системами в соответствии с бизнес-требованиями обработки данных. Выделение зон временного хранения в отдельный компонент информационной фабрики необходимо, так как для этих зон требуется создание дополнительных средств администрирования, мониторинга обеспечения безопасности и аудита.

Кроме того, зоны временного хранения необходимы для синхронизации данных из различных источников в процессе их интегрирования.

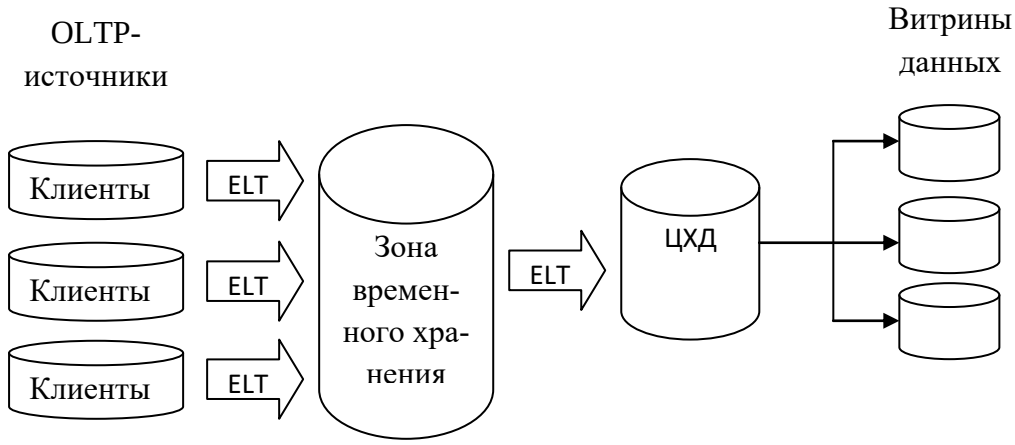


Рисунок 16. Выделение зон временного хранения в CIF

Например, данные, связанные со сбытом продукции, извлекаются ежедневно, а итоги финансовой деятельности – еженедельно. Следовательно, данные, пришедшие раньше, будут ожидать интегрирования с «опаздывающими» данными в зонах временного хранения.

Тема 10. Процессы информативной корпоративной фабрики

Процессы ELT и ETL

Важнейшим процессом, реализуемым в корпоративной информационной фабрике, является извлечение данных из различных источников, их преобразование к единому формату и модели данных, очистка от дубликатов, противоречий и других факторов, которые могут помещать их анализу, а также загрузка в единый интегрированный источник. Для реализации данного процесса в CIF включается специальный комплекс аппаратно-программных средств, называемых ETL, (англ.:*Extract ,Transform, Load* – извлечение, преобразование, загрузка).

Поскольку число систем и объем генерируемых ими данных в CIF может быть очень велик, важнейшими требованиями к аппаратно-программному комплексу ETL являются высокая пропускная способность и вычислительная производительность. Кроме того, желательно, чтобы ETL-система была универсальной, то есть могла извлекать и переносить данные как можно большего числа типов и форматов.

Вариант структуры процессов ETL в контуре корпоративной информационной фабрики представлен на рисунке 16 (здесь и далее мы полагаем, что в рамках проекта решаются только задания, связанные с интеграцией бизнес-данных и НСИ, ведение метаданных компании явно не выделяется).

Данные извлекаются из одного или нескольких источников и помещаются в зону временного хранения, где образуют промежуточные таблицы. Там данные проходят очистку и преобразуются к единому представлению (форматам, кодировкам), а также выполняется их преобразование (агрегирование, нормализация, кодирование), чтобы сделать их наиболее подходящими для последующих задач аналитической обработки.

Для решения данной проблемы было предложено большинство ETL-операций перенести из зоны временного хранения в интегрированный источник. То есть данные сначала извлекаются, затем загружаются, и только в интегрированном источнике подвергаются очистке и трансформации (до загрузки может проводиться агрегирование).

На рис. 16 приведен вариант структуры процессов ETL в контуре CIF.

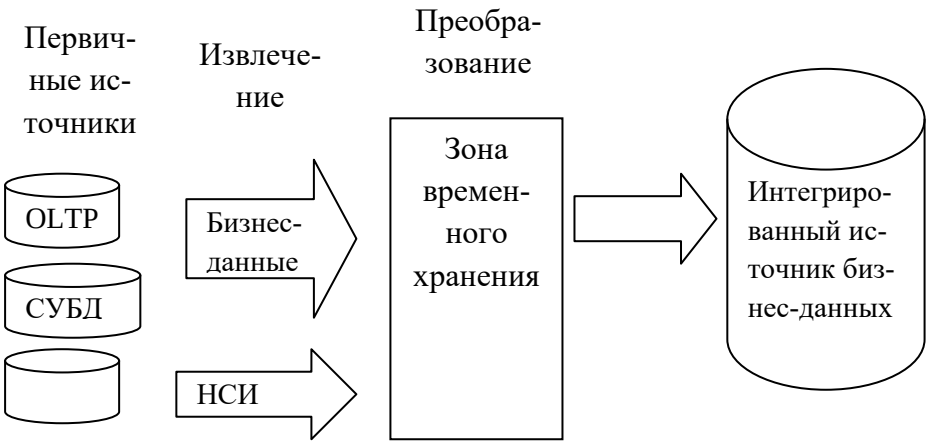


Рисунок 16. Вариант структуры процессов ETL в контуре CIF

Таким образом, производится переход от формулы «извлечение, преобразование, загрузка» к «извлечение, загрузка, преобразование» (англ.: *Extract, Load, Transform – ELT*).

Перенос места обработки данных в интегрированный источник – чаще всего это хранилище – дает следующие преимущества:

- сокращается время между извлечением данных и их загрузкой в хранилище;
- данные поступают в хранилище, даже если часть их потеряна или значительно запаздывает что снижает вероятность потери синхронизации с актуальным состоянием источников;
- очистка данных в хранилище более эффективна, чем в области временного хранения, поскольку очищаются данные, уже подвергшиеся интеграции, в процессе которой могут появляться дубликаты и противоречия.

Наибольшие преимущества ELT перед ETL, проявляются в информационных системах компаний, имеющих значительную территориальную распределенность.

Качество данных – Data Quality

В большинстве случаев данные, извлекаемые из различных источников и интегрируемые в центральном хранилище информационной фабрики, не отвечают определенным требованиям по качеству. Основными причинами этого являются:

- изначально низкое качество первичных источников данных –

далеко не всегда их владельцы заботятся о поддержании качества данных. Особенно проблемными в этом смысле являются источники, в которых данные собираются с помощью ручного ввода, который порождает ошибки, опечатки, пропуски и т.д.;

- отсутствие понимания причин снижения качества данных, стратегии и регламента повышения качества;
- неудачный выбор программных средств, недостаточное внимание к качеству данных в процессе разработки архитектуры CIF;
- попытка перенести мероприятия, направленные на повышение качества данных, на завершающие этапы проекта.

Практика бизнес-аналитики показывает, что качество данных является ключевым факторам обеспечения корректных и практически полезных результатов. Из-за низкого качества данных результаты анализа могут оказаться искаженными, и на их основе приняты неверные управленческие решения, что может вызвать тяжелые последствия для бизнеса.

Например, из-за ошибок и неполноты данных о заемщиках, модель оценки вероятности дефолта, используемая в банке, может формировать ошибочные решения по выдаче кредитов с точки зрения рисков.

Факторы, вызывающие низкое качество данных, далеко не всегда очевидны. Еще менее очевидны проблемы, которые могут быть вызваны низким качеством данных при их анализе. Как следствие, заказчики и финансовые руководители проектов не видят причины для дополнительных затрат, направленных на борьбу с низким качеством данных. Многоаспектность понятия качество данных, большое количество факторов, от которых оно зависит – зачастую не позволяют ясно и чет сформулировать критерии качества.

В каждой компании могут складываться свои, особенные требования качеству данных, а также индивидуальные особенности причин снижения качества. Поэтому задана разработчика архитектуры CIF заключается в реализации процессов, средств и методов обеспечения качества данных Качество данных должно обеспечиваться на всех этапах разработки информационной системы: от постановки задачи до введения ее в эксплуатацию

Качество и метаданные

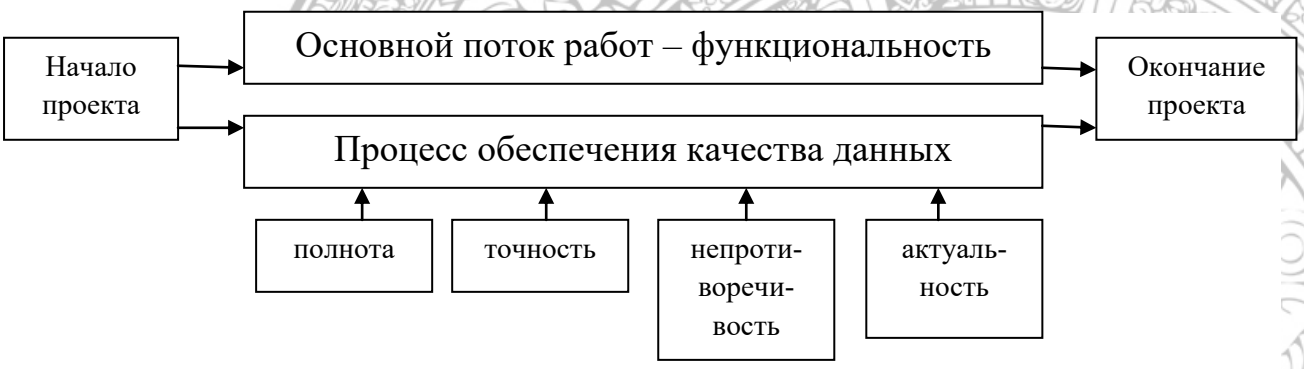
При решении задачи обеспечения качества данных в CIF ключевую роль играют метаданные.

Действительно, при постановке задачи (отправном этапе любого проекта) в первую очередь формулируются бизнес-правила определения, терминология, разрабатываются бизнес-словари и глоссарии, алгоритмы анализа. То есть фактически создаются бизнес-метаданные. И если на этом начальном этапе проблеме качества данных будет уделено достаточное внимание, то это обеспечит хороший задел для достижения высокого уровня качества системы в целом.

В процессе разработки CIF формируются технические метаданные: наименование таблиц и имен полей в них, установка связей и отношений между таблицами, разработка алгоритмов обработки данных в соответствии с бизнес-правилами.

На этапе ввода системы в эксплуатацию создаются эксплуатационные метаданные – журналы учета активности пользователей и использования вычислительных ресурсов, статистика работы приложений и так далее. Основная цель технических метаданных – снижать вероятность отклонения работы системы от установленного регламента, что также способствует повышению качества данных. Включение этапа управления качеством CIF жизненный цикл метаданных позволяет решить проблемы с качеством данных до того, как они повлияют на проект.

Обеспечение качества данных



Наиболее эффективным подходом к обеспечению качества данных является организация потока соответствующих работ параллельно основному потоку, связанному с обеспечением функциональности системы. На практике наиболее критическими факторами качества данных являются полнота,

точность, непротиворечивость и актуальность.

Полнота данных означает, что все данные, связанные с некоторым бизнес-процессом, собраны и представлены в полном объеме – в них отсутствуют неполные столбцы и записи, фрагменты таблиц и так далее. Точность указывает, что представленные значения данных соответствуют реальным показателям (например, представлена цена товара, соответствующая реальной), а так же не содержат ошибок (неправильно указано ФИО клиента).

Непротиворечивость связана с пониманием данных. Например, таблицы, содержащие данные о различных бизнес-объектах, могут иметь одинаковые имена. Или в одном случае валюта может быть указана текстовым значением (руб., долл.), а в другом – символом ($) или аббревиатурой (RUR, USD), что может привести к непониманию и неправильной интерпретации данных.

Актуальность данных связана с их своевременным обновлением. Например, может изменяться курс валюты; ставка рефинансирования, стоимость используемых материалов и так далее. Если данные вовремя не обновлять, то использование устаревших значений приведет к формированию некорректных отчетов, смещенных результатов анализа и другим «вредным» для управления бизнесом явлениям.

Для разработки эффективной стратегии борьбы за качество данных очень важно понимать причины его снижения, а также связь факторов качества с работой системы. Например:

- появление неполных данных может быть связано с неправильной организацией их сбора;
- увеличение нагрузки на операторов ручного вода данных (например, вследствие сокращения персонала), скорее всего, приведет к возрастанию числа ошибок,
- следствием несогласованности моделей данных из различных источников может стать появление противоречий (в одном источнике ФИО клиента указывается полностью, а в другом- с инициалами);
- актуальность данных может быть нарушена из-за непродуманного регламента загрузки, или из-за низкой пропускной способности программно-аппаратного комплекса ETL.

Уровни отчистки данных

Если проследить путь перемещения данных в корпоративной

информационной фабрике от первичных источников до аналитических и бизнес-приложений, можно увидеть, что данные несколько раз подвергаются процедуре очистки. Обычно она реализуется на трёх уровнях:

- 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.

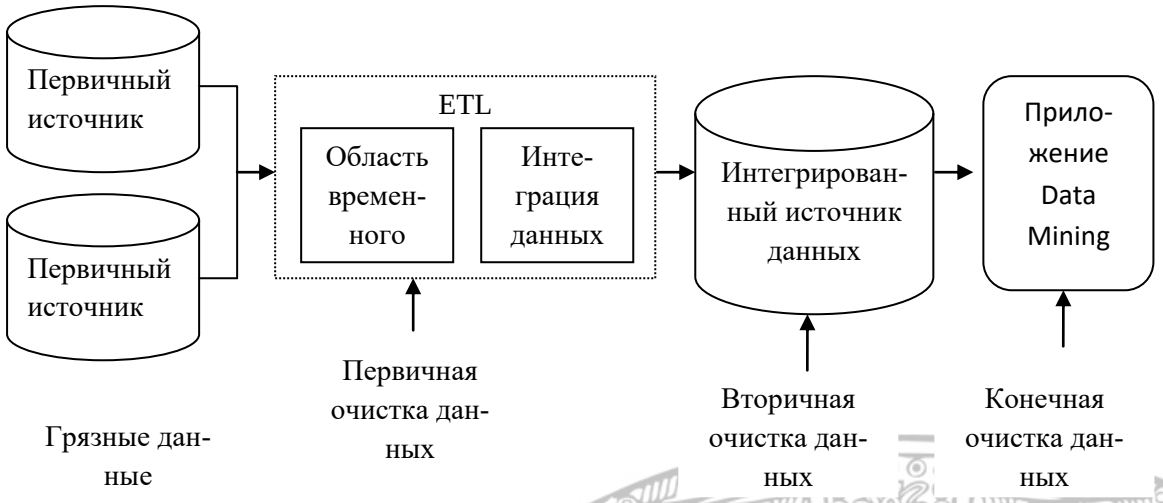


Рисунок 17. Путь перемещения данных в CIF

Очистка данных в ETL

Необходимость очистки данных в процессе ETL обусловлена тем, что в первичных источниках они являются самыми «грязными» и содержат все факторы, снижающие качество: пропуски, дубликаты и противоречия, выбросы и экстремальные значения, нарушения целостности и структуры данных. При этом набор этих факторов может значительно варьироваться зависимости от типа источника.

Например, нарушения структуры и целостности данных менее характерны для баз данных, поскольку в них присутствуют средства их автоматического контроля и поддержки. Напротив, в файлах офисных приложений (текстовых и табличных процессорах, текстовых файлах с разделителями и такс далее) отсутствуют средства поддержки полноты и целостности данных, поэтому в них можно ожидать весь спектр возможных проблем с качеством.

Это связано с преимущественно ручной формой ввода данных, возможной низкой квалификацией пользователей, а также тем, что изначально сбор данных производился без нацеливания на их интегрирование в какую-либо

информационную систему и анализ. Повсеместно могут встречаться несогласованность форматов и представлений данных.

Таким образом, методы и алгоритмы очистки данных, применяемые в процессе ETL, должны соответствовать видам источников данных и их происхождению. При этом целью является то, чтобы данные поступали в интегрированный источник максимально очищенными и готовыми к дальнейшей аналитической обработке или распределению потребителям.

Для очистки данных в ETL должна быть предусмотрена зона временного хранения, где данные помещаются в промежуточные таблицы, к которым применяются различные алгоритмы аудита и очистки данных. Сам процесс очистки в ETL должен производиться до выполнения интеграции данных, а особое внимание следует уделять проблемам, которые могут помешать интеграции.

Очистка данных в консолидированных источниках

Решить все проблемы с данными на этапе их очистки в ET. до желаемого состояния удается далеко не всегда. Поэтому типичной является ситуация, когда в консолидированный источник все еще попадают «грязные» данные. Причин этого может быть несколько.

- Очистка в ETL производится автоматически, без участия пользователя, в соответствии с ранее созданными настройками и заданными алгоритмами, которые выбираются исходя из априорных сведений о происхождении данных, степени их «загрязнения» и типичных проблемах. Поэтому процесс очистки в ETL не является адаптивным появление новых проблем, алгоритм обработки которых изначально не был предусмотрен, приведет к тому, что эти проблемы не будут решены: Например, если задан только алгоритм преобразования номеров телефонов, написанных через тире, в сплошные, то появление номера, написанного через пробелы, станет «неожиданностью» для системы, и ошибка не будет обработана.
- Не все проблемы в данных могут быть обнаружены и распознаны на этапе ETL. Например, интерпретация некоторого значения как экстремального обычно зависит от контекста решаемой задачи анализа, который на этапе ETL еще не известен.
- Поскольку программно-аппаратный комплекс ETL должен обеспечивать высокую пропускную способность и загрузку данных в

соответствии с жестким регламентом время на поиск и обнаружение проблем на этапе ETL является ограниченным, что может не позволить обнаружить и обработать все проблемы.

- Сам процесс интеграции может порождать проблемы в данных, особенно в случае плохой согласованности данных из различных источников. Наиболее типичными проблемами, возникающими в процессе интеграции, являются дубликаты и противоречия.

Очистка данных в бизнес-приложениях

Низкое качество является наиболее критичным для анализа данных в бизнес-приложениях классов BI, Data Mining, Big Data. Это связано с тем, что на основе «сырых» данных невозможно построить корректную модель или составить прогноз с приемлемой точностью. Закономерности и зависимости, извлеченные из таких данных, будут искаженными, и, следовательно, непригодными для поддержки принятия решений и управления бизнесом.

До поступления данных в аналитические приложения они проходят двойную очистку в ETL и консолидированном источнике, что позволяет ожидать, что их качество будет достаточным для проведения различных видов анализа. Тем не менее, почти любое аналитическое приложение содержит свой набор средств для профайлинга и очистки данных.

Необходимость третьего уровня очистки данных обусловлена тем, что она производится непосредственно в контексте алгоритмов и методов моделирования, используемых для решения конкретной задачи анализа. Действительно, алгоритм обработки тех или иных проблем в данных и выбор его параметров (например, метод восстановления пропусков, способ и степень подавления выбросов) зависят от конкретной задачи и модели Data Mining.

Например, выбросы могут ухудшить результат обучения нейронной сети, поэтому обучающие данные должны быть сглажены. Однако степень этого сглаживания должна быть такова, чтобы не подавить и изменения данных, отражающие искомые закономерности и зависимости. Поэтому для выбора оптимальной процедуры очистки для конкретной аналитической задачи могут потребоваться многократные эксперименты.

Таблица 10 Процессы ETL и SRD

ETL	SRD
Извлекает данные из различных внешних систем и источников	Извлекает данные из интегрированного источника (чаще всего – ЦХД)
На входе «сырые» данные, которые требуется преобразовать к единому формату и модели представления	На входе очищенные и хорошо структурированные данные, которые требуется преобразовать в формат, используемый приложением-потребителем
Загружает данные в интегрированный источник данных	Доставляет данные различным системам-потребителям, чаще всего витринам данных

Данные, загруженные посредством процесса ETL/ELT в центральное хранилище, затем будут использоваться витринами данных. Поэтому важным процессом, реализуемым в CIF, является распределение данных из ЦХД различным потребителям. Такой процесс должен производить выборку реструктуризацию и доставку данных потребителями в литературе его упоминают как SRD (англ.: Sample, Restructure, Deliver).

Тема 11. Базовые архитектуры корпоративной информационной фабрики

Архитектура корпоративной информационной фабрики

Ранее мы рассмотрели компоненты, которые входят в состав корпоративной информационной фабрики, а также протекающие в ней процессы. Очевидно, что набор компонентов, из которых будет состоять конкретная CIF а также способ организации их взаимодействия (будем называть это архитектурой CIF), зависит от особенностей бизнеса компании.

Выбор архитектуры зависит от множества факторов – предметной области, степени территориальной распределенности, необходимости оперативного принятия решений, ресурсов, которые компания может использовать для создания CIF и так далее.

Например, если компания не имеет достаточно средств для разработки и внедрения CIF, то она отдаст предпочтение более дешевым архитектурам, не содержащим централизованного ХД, как наиболее затратного компонента. Если бизнес является территориально распределенным, то для синхронизации поступления данных из удаленных источников особое внимание потребуется уделять областям временного хранения.

Таким образом, архитектура CIF разрабатывается для конкретного бизнеса таким образом, чтобы обеспечить максимально эффективное управление им. В настоящее время разработано большое количество разнообразных архитектур CIF. Тем не менее, из них, можно выбрать несколько архитектур, являющихся базовыми, на основе которых могут быть разработаны конфигурации, учитывающие специфику конкретного бизнеса.

Все архитектуры можно разделить на две основных группы: использующие ЦХД, и в которых оно отсутствует. К первой группе относятся централизованное ХД с ETL; централизованное ХД с оперативным складом данных, централизованное ХД с витринами данных.

Архитектуры, не использующие ЦХД: независимые витрины данных; только оперативный склад данных; системы управления основными данными (MDM-системы).

Извлечение данных из различных источников, преобразование их к единому формату и загрузка в ЦХД осуществляется с помощью процесса ETL. При

такой архитектуре подсистема переноса данных и подсистема ранения оказываются полностью разделенными. Однако является ли такая архитектура оптимальной, до сих пор ведутся дискуссии.

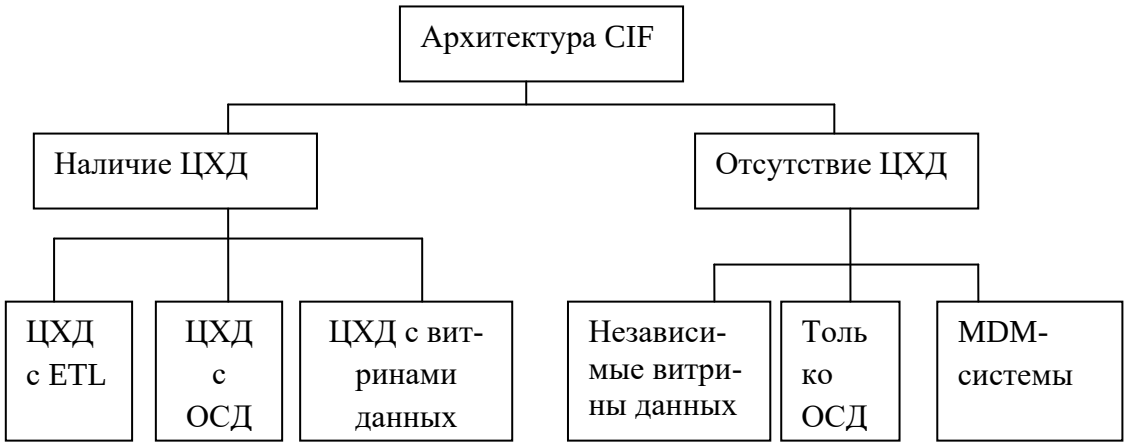


Рисунок 18. Архитектура CIF

Централизованное ХД с ETL

Чтобы корректно работать с источниками данных, процессам ETL. должны быть доступны сведения о них, такие, как структуры и форматы данных, используемые стандарты представления и кодирования, регламент работы транзакционных систем и так далее. Эти сведения содержатся в метаданных. Попытка игнорировать метаданные в процессе ETL неизбежно приводит к снижению качества данных в хранилище.

Как следствие, пользователи могут потерять доверие к ЦХД, и будут пытаться получить доступ к источникам непосредственно, минуя хранилище. Как правило, это приводит к возрастанию необоснованных временных затрат пользователей. Например, аналитик, вместо того, чтобы решать свои задачи, львиную долю времени будет тратить на получение доступа к данным.

Еще одной проблемой является использование в процессе ETL нормативно ‘справочной информации при возрастании числа источников и увеличении потока данных пропускной способности ETL оказывается недостаточно.

Решением данных проблем является организация работы ETL в эффективном взаимодействии с метаданными и НСИ. Загрузка данных из ETL может производиться как в репозитории ХД, так и в оперативный склад или зону временного хранения. С учетом возрастания объемов данных и повышенных требований к пропускной способности в ETL должна широко использоваться технология параллельной обработки.

Кроме этого, подсистема ETL должна удовлетворять еще ряду противоречивых требований. В частности, обладать масштабируемостью (вычислительные затраты должны возрастать линейно при возрастании объема данных), возможностью завершить загрузку, даже если один или несколько источников данных оказались временно недоступны, содержать средства управления метаданными и НСИ. Попытка удовлетворить этим требованиям приводит к критическому ухудшению пропускной способности ETL-процесса, что заставляет разработчиков искать альтернативные архитектуры обеспечения взаимодействия источников и хранилища.

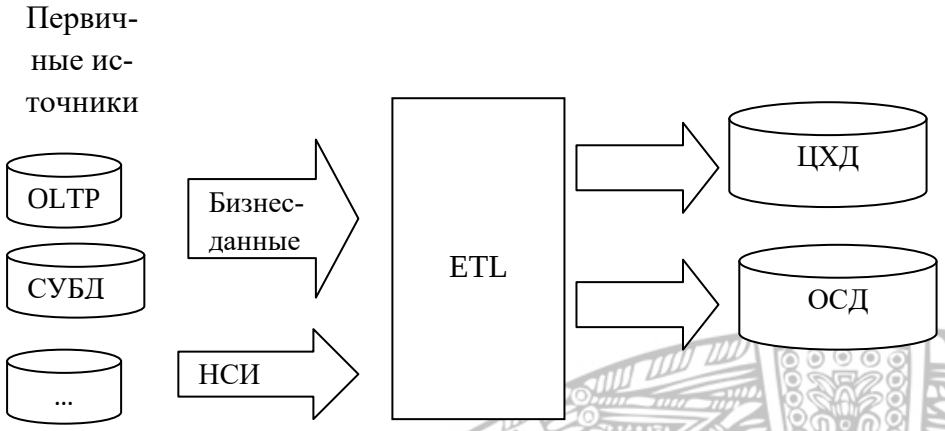


Рисунок 18. Централизованное ХД с ETL

Централизованное ХД с ОСД

Как было отмечено ранее, важным компонентом корпоративной информационной фабрики является оперативный склад данных, ориентированный на решение задач тактического анализа с целью поддержки оперативных управленческих решений. ОСД содержит данные реального времени или с незначительной ретроспективой, отражающей текущее положение в бизнесе. Применение ОСД позволяет «развязать» работу аналитиков, открывает дополнительные возможности в плане администрирования, а также позволяет избежать возможных повреждений основных данных в ЦХД.

Можно выделить три возможных подхода к использованию ОСД в общей структуре CIF – **последовательный, параллельный и независимый**.

Последовательный подход предполагает, что данные из различных источников сначала загружаются в ОСД, а оттуда – в ЦХД. При этом бизнес- приложения, реализующие инструменты тактического анализа, напрямую используют данные из ОСД, минуя витрины данных. В то же время, бизнес-приложения,

реализующие инструменты стратегического анализа, могут использовать данные из ЦХД как посредством витрин данных, так и напрямую.

Главное преимущество прямого доступа бизнес-приложений к ОСД обусловлено тем, что использование витрин данных может значительно повысить время, требуемое для получения данных, необходимых для тактического анализа. Действительно, тактический анализ в компании может проводиться фактически непрерывно, по мере поступления новой информации в реальном времени, поэтому и обращение к ОСД может быть очень частым. В этих условиях процесс обновления витрин данных может оказаться весьма длительным, что неприемлемо для тактического анализа.

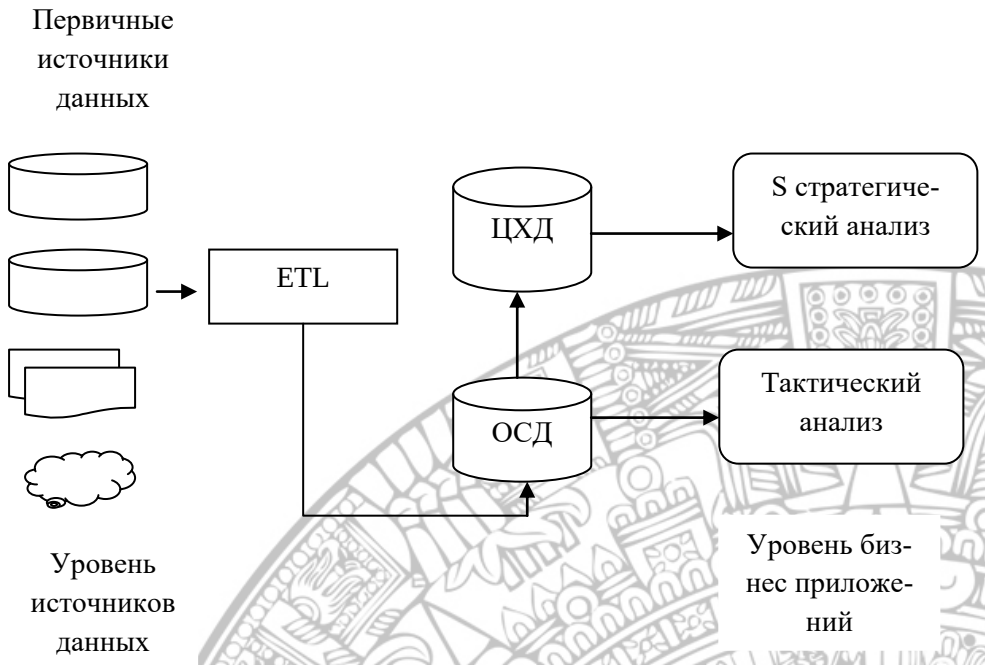


Рисунок 19. Архитектура последовательного соединения ОСД и ЦХД

Важным преимуществом последовательного подхода является возможность использования ОСД в качестве зоны временного хранения, в которой могут выполняться операции по очистке и преобразованию данных.

Действительно, в процессе тактического анализа оперативных данных могут обнаруживаться противоречия, дубликаты, пропуски, выбросы и другие факторы, снижающие качество данных, не устраненные в процессе ETL. Это позволяет применить к ним различные алгоритмы и методы очистки, что обеспечит поступление» ЦХД более качественных данных.

В то же время, задачи стратегического анализа, ориентированные на использование ретроспективных данных, охватывающих достаточно длительные

периоды времени (ежемесячно, ежеквартально), не требуют частого обращения к ЦХД. Поэтому затраты времени на формирование витрин данных не критичны.

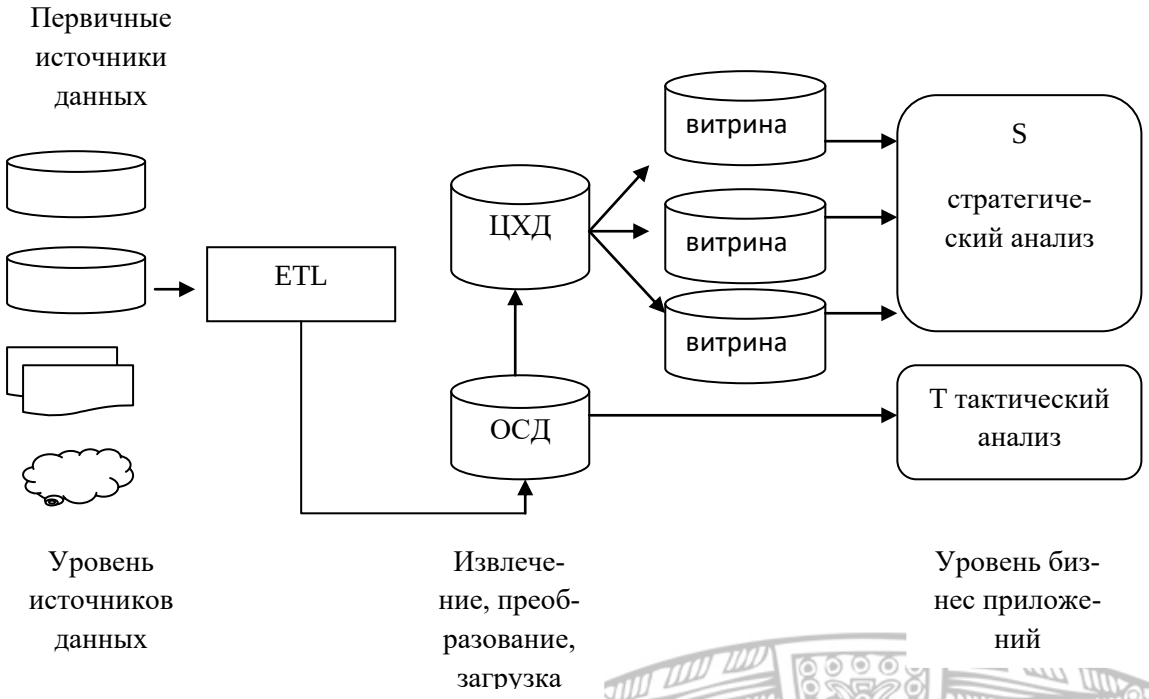


Рисунок 20. Архитектура параллельного соединения ОСД и ЦХД

Таким образом, архитектура параллельного соединения ОСД и ЦХД, при которой данные из ОСД используются бизнес-приложениями напрямую, а из ЦХД – через витрины данных, представляет практический интерес.

При использовании независимого подхода данные с выхода ETL одновременно поступают и в ОСД и в ЦХД. Преимущества этого подхода:

- данные с выхода ETL быстрее оказываются в ЦХД, что **улучшает** его синхронизацию с источниками данных,
- если в процессе тактического анализа в данные, расположенные в ОСД, вносятся нежелательные изменения, то они не попадут в ЦХД.

Недостатками подхода являются:

- отсутствует дополнительный этап контроля и повышения качества данных, реализуемый в ОСД при параллельном соединении;
- регламент прохождения данных через ОСД должен соответствовать регламенту загрузки в ЦХД, что снижает время, доступное для работы с данными при их тактическом анализе.

Независимый подход, таким образом, наиболее целесообразен в тех случаях, когда к защите данных ЦХД предъявляются повышенные требования.

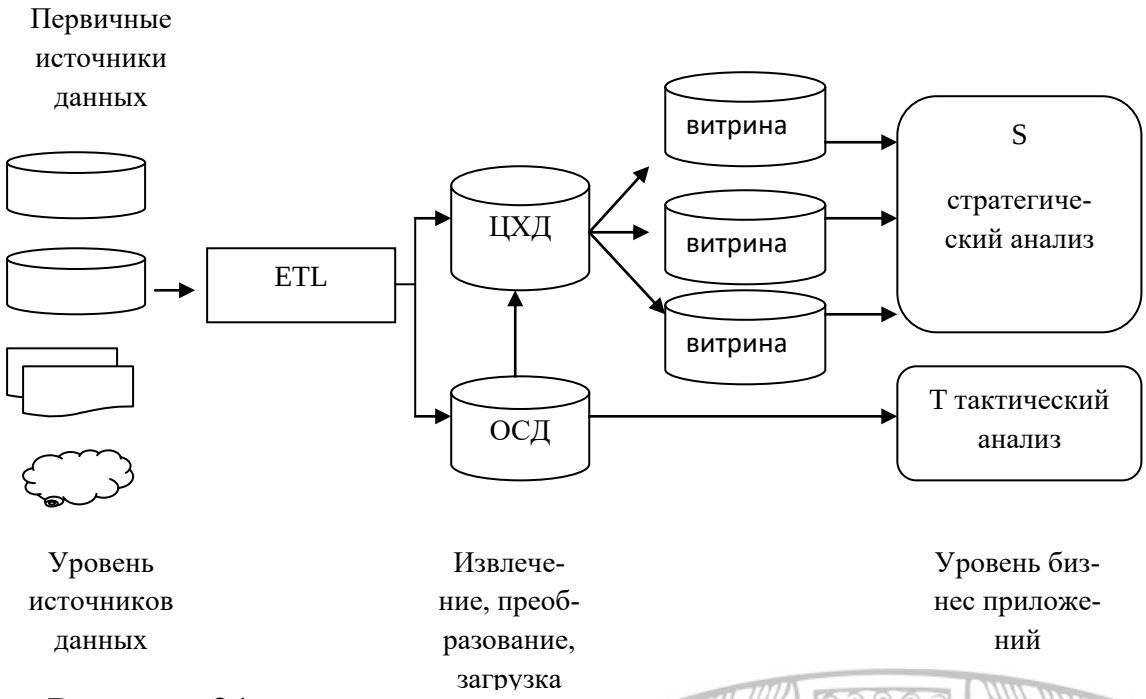


Рисунок 21. Архитектура независимого соединения ОСД и ЦХД

ЦХД является основным местом хранения исторических данных, используемых для целей бизнес-аналитики. Однако оно не решает всех задач, поскольку:

- по мере того, как ЦХД развивается, обеспечить доступ к данным становится все труднее, усложняется администрирование и обеспечение информационной безопасности;
- ЦХД является инструментом масштаба предприятия, поэтому не может содержать данные в виде, удобном для работы всеми подразделениями, по всем направлениям бизнеса;
- хранилище используется одновременно многими пользователями, которые конкурируют между собой за доступ к данным внутри ЦХД, вызывая конфликты и нагружая хранилище;
- коллективный доступ большого числа пользователей может привести к искажению информации в ЦХД, ее частичной или даже полной потере;
- недостаточная пропускная способность и ненадежность телекоммуникационных линий в случае территориально распределенной компании.

Все эти и многие другие проблемы, связанные с эксплуатацией ЦХД, позволяет снять использование витрин данных.

Возможно два способа соединения витрин данных с ЦХД – **независимое и параллельное**. При параллельном соединении каждая витрина данных имеет свое подключение к ЦХД. Данный подход имеет свои преимущества и недостатки.

Преимуществом является высокая надежность – проблемы доступа, возникающие у какой-либо витрины, не влияют на работу других витрин. С другой стороны, такой способ доступа не оптимален с точки зрения нагрузки на ЦХД: если одни и те же данные будут требоваться нескольким витринам, то число обращений к хранилищу будет равно числу этих витрин.

Независимое соединение предполагает наличие процедур SRD, когда данные из хранилища извлекаются одной распределяющей системой и предоставляются витринам независимо.

Недостатком подхода является снижение надежности: проблемы в программно-аппаратном комплексе SRD приводят к нарушению работы всех витрин данных. Преимуществом является оптимизация нагрузки на хранилище: если несколько витрин запрашивают одну и ту же информацию из ЦХД, система SRD обращается к хранилищу единственный раз, а затем распределяет данные по витринам.

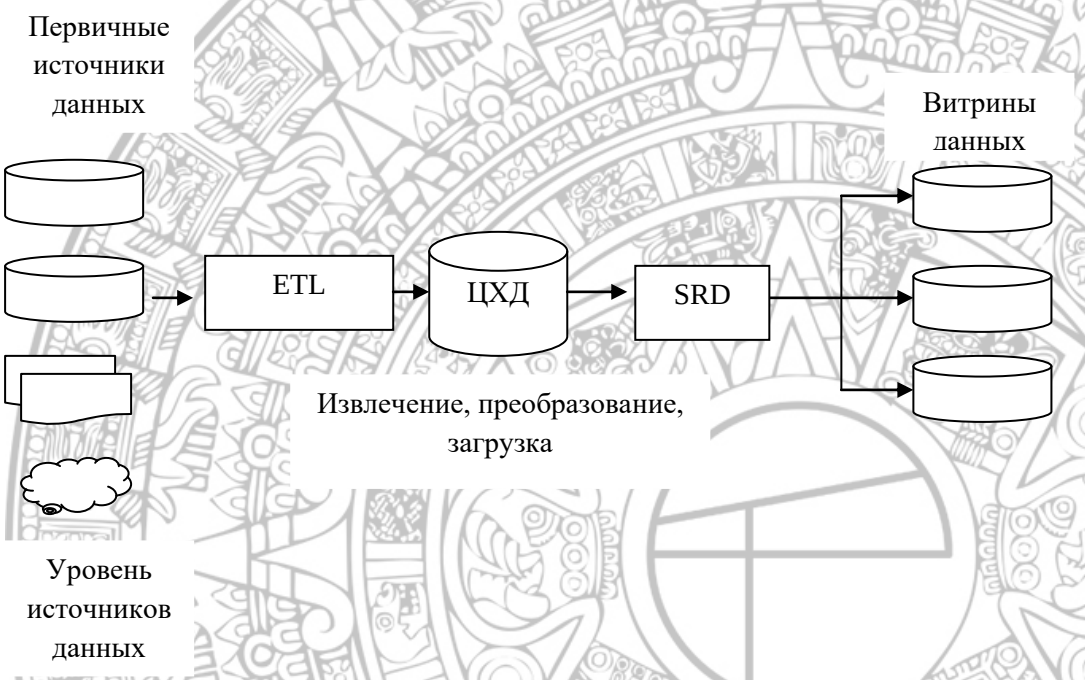


Рисунок 21. Независимое соединение витрин данных с ЦХД

Таким образом, использование SRD фактически снижает число пользователей ЦХД до одного, что является оптимальным вариантом с точки зрения балансировки нагрузки на хранилище, атаке в плане администрирования и защиты информации. Эти несвойственные хранилищу функции выводятся в систему SRD и независимые витрины данных.

Независимые витрины данных

Рассмотрим варианты архитектур без централизованного ХД. Первый – с независимыми витринами данных. Независимые витрины данных характеризуются несколькими признаками. Во-первых, данные в каждую витрину поступают непосредственно из первичных источников, без использования ЦХД. Во-вторых, эти витрины данных построены независимо друг от друга в плане используемого программного обеспечения и аппаратных средств. Данная архитектура имеет следующие недостатки:

- по мере того, как число независимых витрин данных растёт, бесконтрольно возрастает и количество избыточных данных. Причиной избыточности является возможное дублирование данных из хранилища загружаемых в каждую витрину;
- поскольку хранилище данных отсутствует процедуры очистки, интеграции и трансформации данных приходится выполнять для каждой витрины. Это приводит к увеличению вычислительной нагрузки на систему в целом, а также требует, чтобы в каждом подразделении, использующем витрину данных, имелся собственный специалист по ETL, что ведет к дополнительным затратам;
- каждое подразделение компании строит свою витрину данных независимо от других подразделений, используя различные стандарты, форматы и модели данных, не говоря уже о метаданных. Следствием является плохая согласованность результатов анализа даже одних и тех же данных в масштабе всего предприятия;
- поскольку независимые витрины данных читают данные из различных источников, то, если, например, пяти витринам потребуется информация о клиентах, то она будет считана пять раз, что приводит к ухудшению масштабируемости системы;
- поскольку независимые витрины данных ведутся отдельными подразделениями изолированно друг от друга, ни одна из них не дает общего

взгляда на работу компании, что мешает выработке согласованных управленческих решений.

Несмотря на указанные недостатки архитектуры, многие компании все же используют ее. Это обусловлено следующими причинами:

- низкая стоимость и простота реализации;
- возможность построения системы для отдельно взятого подразделения, если построение корпоративной системы для компании слишком дорого.

В большинстве случаев построение системы интеграции данных с независимыми витринами – начальный этап разработки архитектуры корпоративной информационной фабрики или вынужденная мера по причине экономии средств. По мере развития компании, использующие независимые витрины, отходят от данной архитектуры и переходят к использованию полноценных ЦХД.

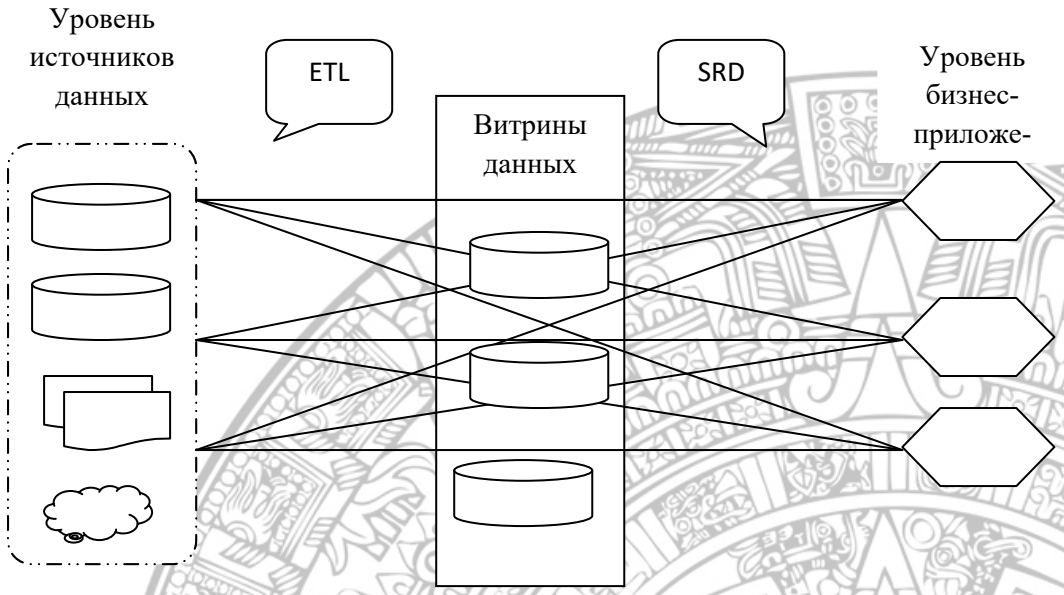


Рисунок 22. Архитектура без централизованного ХД

Только оперативный склад данных

Прямая загрузка из ETL в витрины может вызвать ряд проблем Основной из них является отсутствие возможности восстановления данных, если они были искажены или утеряны в процессе передачи: Это связано с тем, что выполнение процедур ETL не предполагает долговременного хранения данных. Другой проблемой является снижение качества данных, поскольку этап его контроля и улучшения внутри ЦХД не производится.

Снизить остроту проблем, связанных с прямой загрузкой витрин данных из первичных источников, позволяет использование ОСД. Действительно, ОСД содержит оперативные данные – данные реального времени или с небольшой ретроспективой. Изначально, ОСД играет роль буфера между транзакционными детализированными данными и ЦХД. Так же, как ЦХД, ОСД является предметно-ориентированным, но не содержит историю данных и подвержено изменчивости.

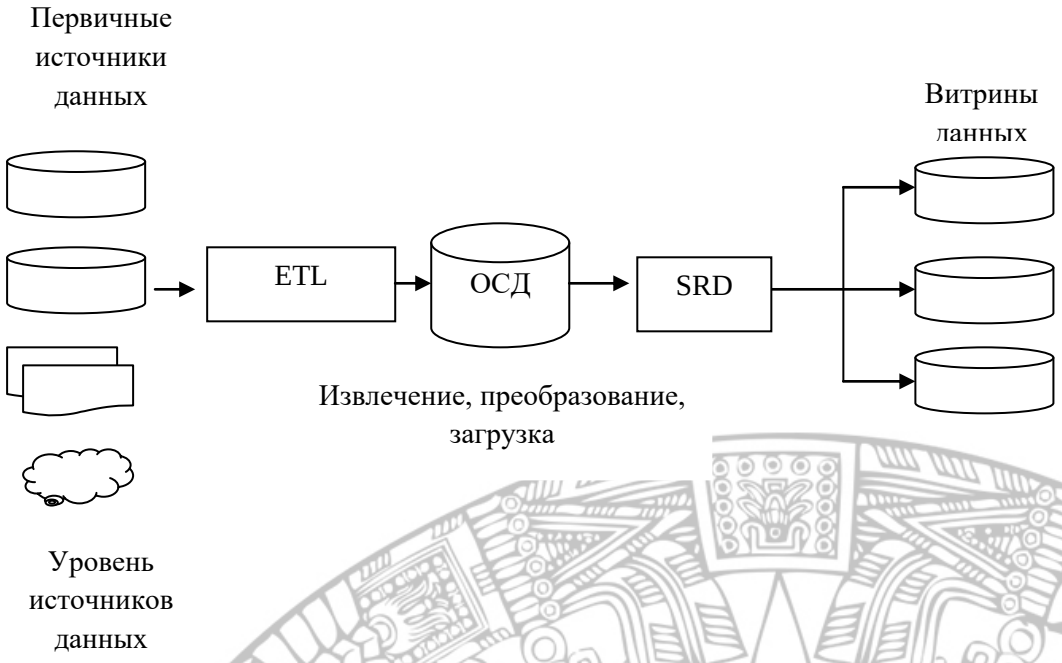


Рисунок 23. Архитектура с оперативным складом данных

Если передача данных от источников к витринам данных производится без использования ЦХД, то введение ОСД позволяет реализовать в них очистку данных.

Тема 12. Системы управления мастер-данными

Понятие MDM-системы

Кроме транзакционных данных, отражающих элементарные факты бизнес-процессов, в корпоративной информационной фабрике должна содержаться информация о клиентах (ФИО, номер паспорта, адрес, телефон), товарах (наименования, цены), услугах, сотрудниках, материалах и других объектах, являющихся субъектами деятельности компании. Ранее было отмечено, что такие данные часто называют основными или мастер-данными.

Пусть была выполнена транзакция, в ходе которой клиент приобрел у компании стройматериалы для строительства. Пример приведен в табл.12.

Таблица 11 Пример транзакции

Поле	Значение
Дата	21/03/2014
Клиент	ИП Борисов К.И.
Адрес	г. Москва, ул. Новая, 12
Телефон	(495)78952789
Товар	Кирпич
Цена за единицу, руб.	11,00
Количество, шт.	10000
Сумма, руб.	110000,00

Если клиент уже пользовался услугами компании, то информация о нем уже содержится в соответствующем справочнике. Информация о покупаемом товаре (наименование, цена, другие свойства) также занесена в справочник. Таким образом, в транзакции появляются только несколько новых значений данных – дата, количество, сумма, которые оператор заносит вручную. Справочные значения оператор выбирает из списков, что снижает вероятность ошибки.

«Поведение» мастер-данных существенно отличается от транзакционных. Если дата, количество проданного товара и сумма различны для каждой транзакции, то наименования и цены меняются достаточно редко. Тем не менее, они также могут содержать факторы, мешающие корректному анализу данных и формированию отчетности.

Для них типичны противоречия (разные ФИО для одних паспортных данных, некорректные форматы имен людей, дать адресов и телефонов, аномальные значения цен и других характеристик объектов, и так далее), дубликаты

(одного и того же клиента внесли в справочник два раза). Поэтому мастер-данные должны непрерывно контролироваться на предмет выявления и исправления обнаруженных проблем.

Для реализации этих задач в информационных системах компаний организуется совокупность процессов и инструментов, называемых управление мастер-данными (англ.:Master Data Management – MDM). Так же именуют и комплекс программно-аппаратных средств управления мастер -данными в компании. MDM обеспечивает процессы формирования мастер-данных (сбора из внешних источников, повышения качества за счет очистки и обогащения и так далее) и их распределение для дальнейшего использования в других компонентах корпоративной информационной фабрики.

Основной задачей MDM является формирование и поддержка достоверной, непротиворечивой и актуальной информации, а также обмен данными с другими компонентами корпоративной информационной фабрики, которые обычно имеют свои модели данных.

Роль и место MDM-системы в структуре CIF

Комплекс мероприятий, реализующих разработку и создание MDM-системы в рамках существующей или только создающейся CIF, называют MDM проектом.

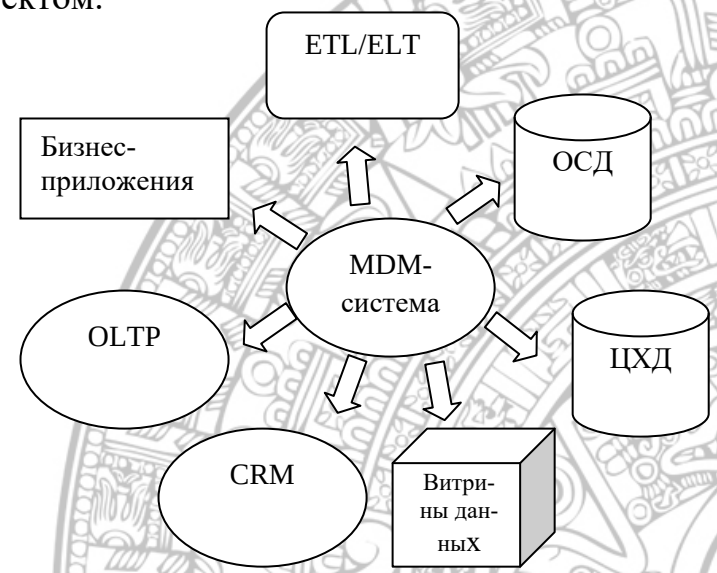


Рисунок 24. Архитектура MDM-проекта

Архитектура MDM-проекта строится по принципу «звезды», в центре которой расположена MDM система, а к ней подключены остальные компоненты и процессы информационной фабрики.

Происхождение мастер-данных и НСИ

Рассмотрим процесс формирования данных на примере работы: небольшой оптовой компании, процесс оформления заказов в которой реализуется по схеме, представленной на рисунке.

Все связи в схеме имеют отношение один-ко-многим: один менеджер осуществляет взаимодействие с множеством клиентов, с каждым клиентом связано несколько заказов, в каждом заказе присутствуют несколько товаров и так далее.

Исходя из представленной схемы, можно отметить наличие в ней трех видов данных:

- транзакционные данные, относящиеся непосредственно к бизнес-операциям (поставка товара по заказу определенного клиента, реализуемая определенным менеджером),
- мастер-данные компании, описывающие ключевые объекты деятельности информацию о клиентах, номенклатуре товаров и услуг;
- ссылочные данные – данные, используемые всеми информационными системами компании: справочники единиц измерений, словари используемых сокращений, тематические классификаторы и так далее.

Следует отметить, что из-за отсутствия четких критериев, позволяющих разделить мастер-данные и ссылочные данные их объединяют общим понятием НСИ. Очевидно, что в MDM-системе присутствуют также и метаданные.

Архитектуры MDM-систем

В настоящее время разработано несколько архитектур MDM-систем. Мы рассмотрим три основных: консолидированные, транзакционные и реестровые.

Консолидированные MDM-системы. Важнейшей задачей MDM-системы является автоматизация управления данными о клиентах (англ. Customer Data Integration, CDI). Это особенно актуально для территориально распределенных компаний, имеющих обширную сеть филиалов и офисов продаж, в которых может использоваться различное оборудование, программное обеспечение и методы учета. Для таких компаний важным является формирование единого клиентского справочника, который будет содержать полные, достоверные и непротиворечивые данные о каждом клиенте компании или так называемую «зо-

лотую запись» (англ.: Golden Record, «единая версия правды»). Золотая запись используется всеми структурами компании и позволяет сформировать единый взгляд на клиента, использование MDM-системы – осуществить переход к клиенто-ориентированной модели бизнеса.

При формировании золотой записи к данным извлекаемым из первичных учетных систем, применяется профайлинг (профилирование), в процессе которого выявляются ошибки, несогласованность форматов и представлений, наличие дубликатов и противоречий. Затем данные подвергаются очистке и интеграции, обогащению и стандартизации.

Консолидированные MDM -системы

«Замусоренность» клиентской базы приводит к тому, что компания даже не имеет понятия, сколько у нее на самом деле клиентов, какие из них являются прибыльными или убыточными, лояльными или склонными к уходу. Любой общий анализ и отчетность по такой базе не дадут достоверной картины.

Только объединение информации о клиенте из множества источников позволяет получить полное представление о нем на уровне, достаточном для эффективной организации взаимодействия с ними. Поэтому CDI системы содержат полный набор средств для профилирования и анализа качества данных о клиентах, а также для его повышения.

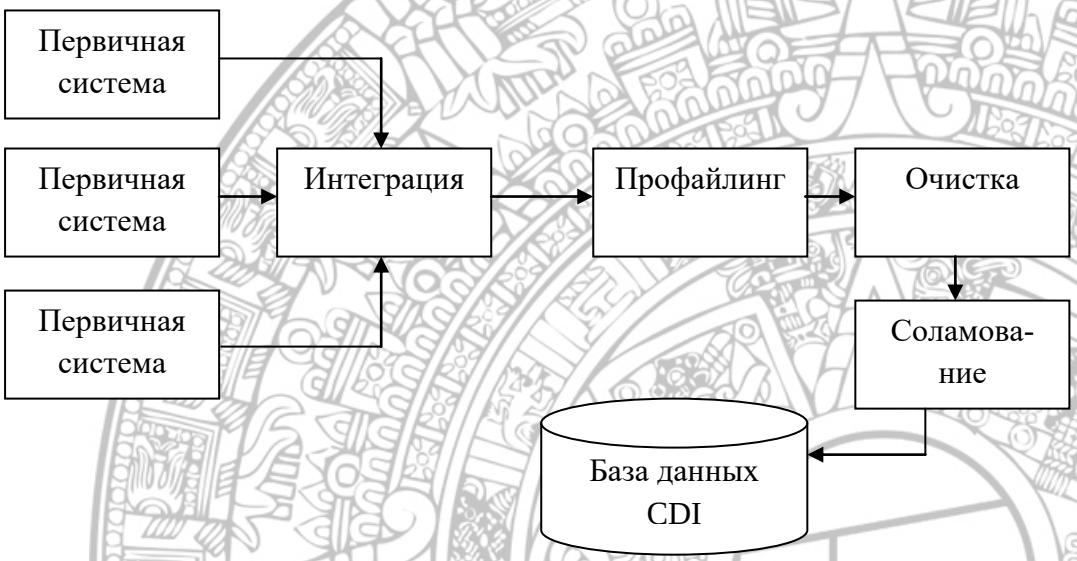


Рисунок 25. Консолидированные MDM-системы

В процессе профайлинга выявляются общие проблемы источников данных, например, число (или процент) клиентов, для которых отсутствуют адреса или номера телефонов. Согласование (стандартизация) – приведение к единому

заданному формату, например, телефонных номеров к сплошному (без тире и пробелов) формату, В процессе очистки производится исправление ошибок и несоответствий в данных.

Например, если статистический анализ показал, что для 200 клиентов с именем Сергей указан пол Мужской, а для пяти – Женский, то может быть автоматически обнаружена ошибка ввода данных. Исправить ее просто – заменить значения, вероятность появления которых совместно с другими мала, на альтернативные.

Транзакционные MDM-системы

Транзакционные MDM-системы: Данный вид MDM-систем ориентирован на создание единой платформы для ведения НСИ, которая должна обслуживать всех участников внутренних бизнес процессов компании, и обычно содержит информацию, описывающую материально-технические ресурсы предприятия, ее продукты, товары и услуги. Такую модель управления мастер-данными часто называют Product Information Management, PIM.

В отличие от модели CDI, которая предполагает консолидацию НСИ в едином репозитории, модель PIM использует специальный узел – MDM Transaction Hub, через который бизнес-приложения получают оперативный доступ к транзакционным и мастер-данным. При этом мастер данные частично хранятся в источниках транзакционных данных, а частично – в репозитории, входящем в состав хаба.

При этом если между репозиторием мастер-данных хаба и мастер-данными источников связь двунаправленная, то бизнес-приложения имеют доступ к мастер-данным и транзакционным данным в режиме только чтения. В репозитории мастер-данных транзакционного хаба содержатся: метаданные, описывающие модель мастер-данных, собственно мастер-данные; история изменения мастер-данных, ссылочные данные единицы измерения, коды стран-изготовителей, места хранения и так далее.

Преимуществом модели PIM относительно CDI является то, что в ней проще поддерживается актуальность мастер-данных, в то время как в базе данных CDI изменения мастер-данных отслеживаются со значительной задержкой.

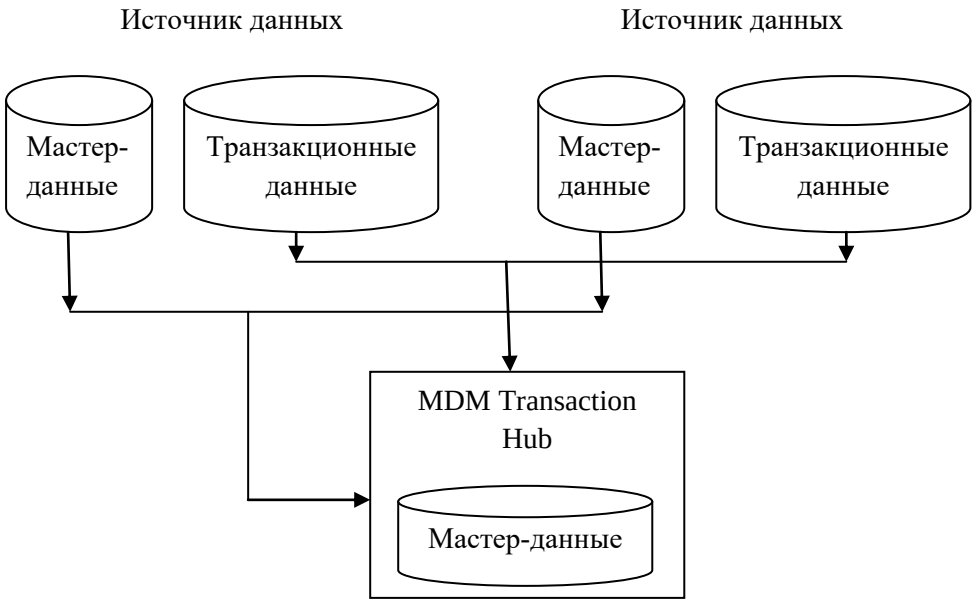


Рисунок 26. Транзакционные MDM-системы

Реестровые MDM-системы

В основе идеи реестровой архитектуры MDM системы лежит идея получения «виртуальной золотой записи» на основе перекрестных запросов к источникам данных, которые производят поиск связанных записей, имеющих отношение к определенной сущности в мастер-данных (клиенту, товару, изделию и так далее).

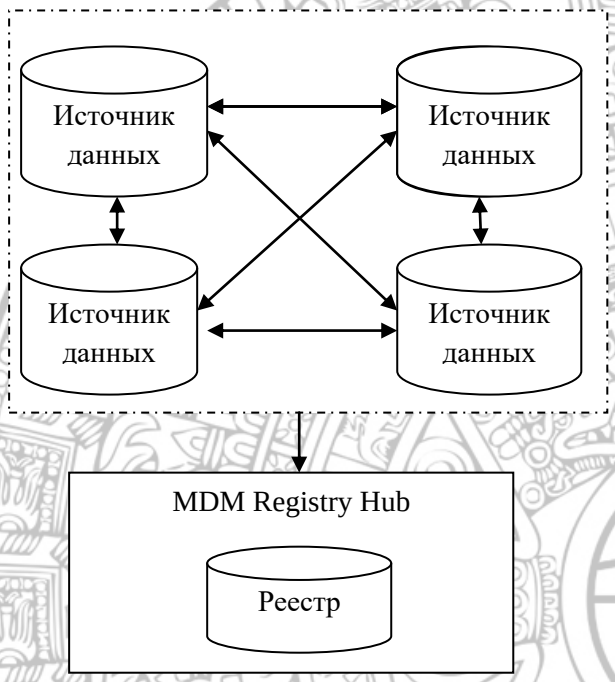


Рисунок 27. Реестровые MDM-системы

Такие записи поступают в центральный MP хаб, где из них формируется реестр, обеспечивающий «золотой взгляд» на данную бизнес-сущность с точки зрения различных информационных систем компании.

При этом, хотя «золотая запись» формируется в некотором центральном узле MDM-системы (хабе), связь между источниками данных осуществляется без посредников, то есть по принципу точка-точка.

Преимуществами использования реестровой архитектуры (см. рис. 27) является оперативность и низкая стоимость, а также минимум операций по согласованию данных. Недостатком является сравнительно высокая нагрузка на систему, поскольку обращения производятся одновременно к большому количеству источников данных.

Свойства данных в MDM-системе

При выборе архитектуры MDM системы и разработке стратегии ее внедрения в корпоративную информационную фабрику очень важно понимать особенности и свойства данных, находящихся в ней, в контексте конкретного проекта внедрения MDM. Набор видов данных и их свойств для каждого конкретного случая может различаться, но можно выделить наиболее типичные – ссылочные (НСИ), мастер-данные и транзакционные данные. Сравним их основные свойства по критериям.

Таблица 12. Основные свойства данных в MDM системе

Критерий	Ссылочные данные (НСИ)	Мастер данные	Транзакционные данные
Содержание и назначение	Справочники, нормативы, сокращения кодификаторы, стандарты, словари, бизнес, правила в виде диапазонов, пороговых значений констант и т.д. Связующее звено меду мастер и транзакционными данными.	Основные объекты компании, прошедшие процедуры очистки и согласования. Определяются типом архитектуры MDM-системы	Содержат собственно транзакции Часто представлены комбинацией мастер-данных, НСИ и. дополнительных атрибутов
Скорость изменения модели данных	Практически не подвержена изменениям на протяжении всего периода использования	Меняется редко и в основном при адаптация к изменениям бизнес-процессов и расширению моделей в транзакционных системах	Меняется часто в зависимости от новых бизнес-требований к системе

Окончание табл. 13

Скорость изменения данных	Очень низкая	Низкая, данные меняются при изменении мастер-данных в одной из информационных систем (или в MDM-системе) и формирования «золотой записи»	Высокая. Меняются в каждой транзакции
Чувствительность к качеству данных	Очень высокая. Недопустимы противоречия, дубликаты, несоответствия форматов	Высокая	Высокая. Качество данных определяется средствами их очистки в транзакционной системы
Время существования	На протяжении всего времени существования бизнеса предприятия	Формируются и поддерживаются на протяжении всего времени существования бизнеса предприятия	Определяется временем существования транзакционной системы
Количество данных	Небольшое, отражает сущности, количество которых заведомо конечно	Большое, но ограничено	Не ограничено
Связь с другими видами данных	Не ссылаются ни на какие другие виды данных и не содержат их	Содержат или ссылаются на НСИ	Могут содержать или ссылаться на любые другие виды данных
Способы распространения	Централизовано и однонаправленно – списками, файлами, выгрузками, репликацией между БД	Двунаправленные каналы, работающие в реумах «онлайн» и «офлайн»	Определяется спецификой транзакционной системы, любые способы интеграции
Места появления данных внутри CIF	Транзакционные системы. MDM-система. Глоссарий предприятия	MDM-система. Транзакционные системы	Транзакционные системы

Преобразование данных

Преобразование данных как часть ETL-операций

Преобразование данных в том или ином виде выполняется во многих компонентах корпоративной информационной фабрики: в процессе переноса и загрузки данных в интегрированный источник или области временного хранения (ETL), непосредственно при подготовке данных к анализу в бизнес-приложении (SRD). Такая распределенность процесса преобразования обусловлена тем, что на каждом этапе он преследует различные цели.

Операции преобразования могут производиться для обеспечения технической и логической совместимости данных, их подготовки к извлечению, переносу в хранилище данных и так далее. Например, адреса часто вводят одной строкой. В то же время для анализа могут представлять интерес отдельные компоненты адреса, которые имеют как текстовый формат (*улица, город*), так и числовой (*номер дома, офиса*). С помощью трансформации можно распределить соответствующие элементы по отдельным полям и преобразовать их в нужный формат.

Основными целями преобразования данных на этапе процесса ETL являются:

- приведение их в соответствие с моделью данных, используемой в хранилище;
- осуществление корректной консолидации и собственно загрузка в хранилище.

Возникает вопрос: если преобразованию данных так много внимания уделяется на этапе интеграции данных, то зачем включать средства преобразования в аналитическое приложение (что, несомненно, усложняет и делает более дорогим его разработку)? Ответ прост. **Не все данные поступают в бизнес-приложение из систем, где они прошли предварительную подготовку**. Но главная причина заключается в том, что трансформация данных в этих системах в большей степени носит *технический характер* и слабо связана с возможными методами, алгоритмами и целями анализа.

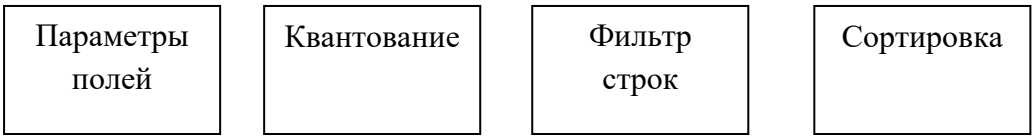
Основные методы преобразования данных

Параметры полей. Позволяют изменять имена, типы, метки и назначения полей исходной выборки данных. Например, если поле, содержащее числовую информацию, в источнике данных по какой-либо причине имеет строковый тип, значения этого поля не могут обрабатываться как числа. Чтобы работа с числовыми данными этого поля стала возможной, их следует преобразовать к числовому типу.

Квантование. Позволяет разбить диапазон возможных значений числового признака на заданное количество интервалов и присвоить номера интервалов или иные метки попавшим в них значениям.

Фильтр строк. Оставляет только те записи, которые удовлетворяют заданным условиям.

Сортировка. Позволяет изменить порядок следования записей исходной выборки данных в соответствии с алгоритмом, определенным пользователем В некоторых случаях сортировка дает возможность упростить визуальный анализ выборки, оперативно определить наибольшие и наименьшие значения признаков и так далее.



Обогащение данных. Включает в себя несколько операций, позволяющих дополнить выборку недостающей информацией из других выборок, если исходная содержит недостаточно данных для анализа. Операция **слияния**, в частности, позволяет объединить две таблицы по одноименным полям, **дополнение данных** – использовать одноименные поля для дополнения одной таблицы полями из других, которые отсутствуют в первой. При **объединении** к записям исходной выборки добавляются все записи другой, а в случае операции **соединения** добавляются все выбранные поля.

Табличная подстановка значений и кодирование. Позволяет производить замену значений в исходной выборке данных на основе, так называемой таблицы подстановки. Таблица подстановки содержит пары «исходное значение-новое значение». Каждое значение выборки данных проверяется на соответствие исходному значению таблицы подстановки, и если такое соответствие найдено, то значение выборки изменяется на соответствующее новое значение из таблицы подстановки. Это очень удобный способ для автоматической корректировки значений.

Группировка. Очень часто информация, интересующая аналитика, в таблице оказывается «разбавлена» посторонними данными, разобщена, разбросана по отдельным полям и записям. Используя группировку, можно обобщить нужную информацию, объединить ее в минимально необходимое количество полей и значений.

Вычисляемые значения. Иногда для анализа требуется информация, которая отсутствует в явном виде в исходных данных, но может быть получена на основе вычислений над имеющимися значениями, например, если известны цена и количество товара, то сумма может быть рассчитана как их произведение. Для этих целей в аналитическое приложение включается своего рода калькулятор, который позволяет выполнять над данными исходной выборки различные

вычисления. Поскольку анализируемые данные могут быть различных типов (строковый, числовой, дата/время, логический), то механизм расчетов должен поддерживать работу не только с числовыми данными, но и с данными других типов, например, выделять подстроку, выполнять логические операции и так далее.

Преобразование упорядоченных данных. Позволяет оптимизировать представление таких данных с целью обеспечения дальнейшего анализа, например, решения задачи прогнозирования временного ряда или группировки по временному периоду

Транспонирование. Служит для вращения набора данных, когда строки необходимо сделать столбцами и наоборот.

Тема 13. Технология OLAP и ее особенности

Бизнес-аналитика располагает различными методиками, а также средствами автоматизации, которые служат для поддержки принятия решений:

- подсистемы информационно-поискового анализа;
- подсистемы оперативной аналитической обработки данных (OLAP);
- подсистемы так называемой «добычи данных», или интеллектуального анализа (методы и алгоритмы «Data Mining»).

Понятие OLAP

Традиционным подходом к организации баз данных и соответствующих обслуживающих приложений является OLTP-подход.

OLTP или Online Transaction Processing – это обработка транзакций в реальном времени. Структура такой базы данных сильно нормализована и оптимизирована для выполнения коротких идущих большим потоком транзакций, при этом клиенту требуется от системы минимальное время отклика. Обрабатываемый и сохраняемый OLTP-системой в течение дня объем данных может достигать нескольких гигабайт. Примерами применения OLTP-подхода могут служить системы учета биржевых, банковских операций, системы бухгалтерского и складского учёта и т.д.

Благодаря нормализации в таких системах значительно снижается избыточность данных и вычислительные потребности на операции обновления, что делает OLTP-системы идеальными для хранения данных. Однако сложность структуры таблиц и большие объемы накопленных данных приводят к снижению скорости выполнения сложных запросов на извлечение данных (например, посчитать прибыль организации по кварталам за последние пять лет), снижению производительности системы в целом.

В результате эти системы оказываются непригодными для решения задач, диктуемых бизнес-аналитиками.

Поиск решения данной проблемы привел к формированию совершенно нового подхода, получившего название *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. Цель таких систем – проверка гипотез пользователя-аналитика.

Централизация и удобное структурирование – это далеко не все, что нужно аналитику. Ему ведь еще требуется инструмент для просмотра, визуализации информации. Традиционные отчеты, даже построенные на основе единого хранилища, лишены одного – гибкости. Их нельзя «покрутить», «развернуть» или «свернуть», чтобы получить желаемое представление данных. Конечно, можно вызвать программиста (если он захочет придти), и он (если не занят) сделает новый отчет достаточно быстро – скажем, в течение часа.

Получается, что аналитик может проверить за день не более двух идей. А ему (если он хороший аналитик) таких идей может приходить в голову по нескольку в час. И чем больше **«срезов» и «разрезов»** данных аналитик видит, тем больше у него идей, которые, в свою очередь, для проверки требуют все новых и новых «срезов». Вот бы ему такой инструмент, который позволил бы разворачивать и сворачивать данные просто и удобно. В качестве такого инструмента и выступает OLAP.

Хотя OLAP и не представляет собой необходимый атрибут хранилища данных, он все чаще и чаще применяется для анализа накопленных в этом хранилище сведений.

Компоненты, входящие в типичное хранилище, представлены на рис. 28.

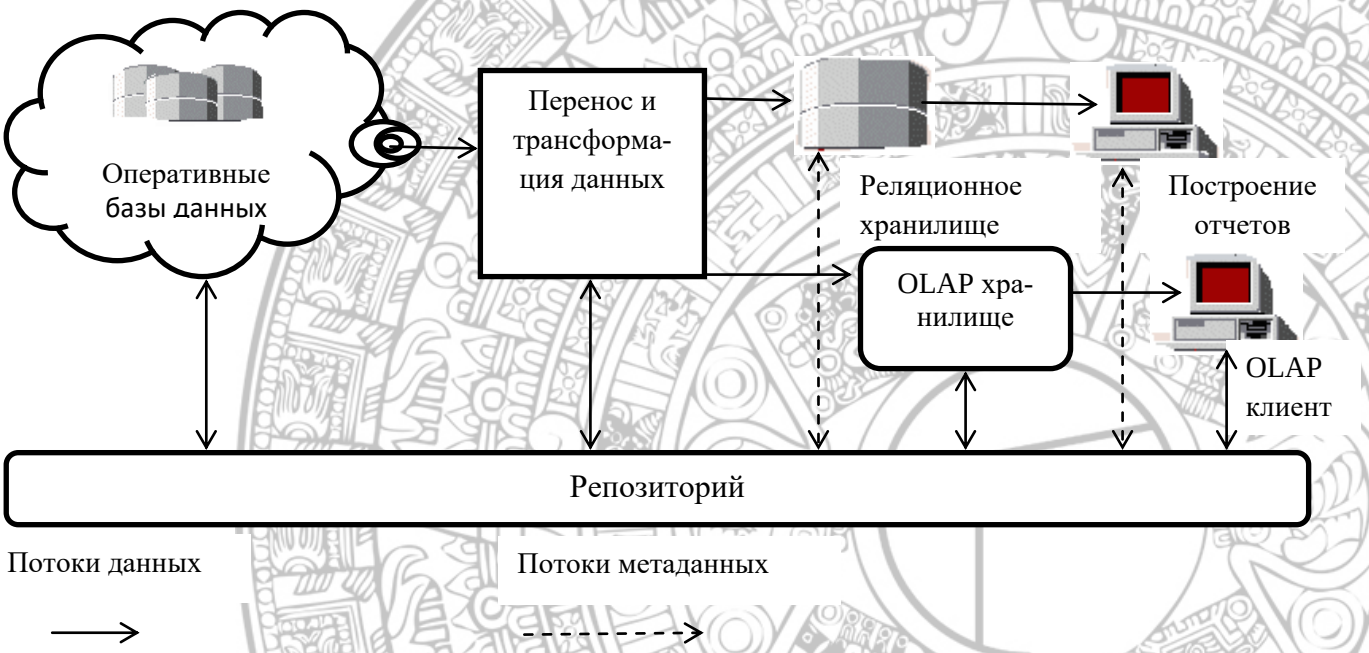


Рисунок 28. Структура хранилища данных

Оперативные данные собираются из различных источников, очищаются, интегрируются и складываются в реляционное хранилище. При этом они уже доступны для анализа при помощи различных средств построения отчетов. Затем данные (полностью или частично) подготавливаются для OLAP-анализа. Они могут быть загружены в специальную базу данных OLAP или оставлены в реляционном хранилище.

Важнейшим его элементом являются метаданные, т. е. информация о структуре, размещении и трансформации данных. Благодаря им обеспечивается эффективное взаимодействие различных компонентов хранилища.

Таким образом, можно определить **OLAP как совокупность средств многомерного анализа данных, накопленных в хранилище.** Теоретически средства OLAP можно применять и непосредственно к оперативным данным или их точным копиям (чтобы не мешать оперативным пользователям). Но мы тем самым рискуем наступить на уже описанные выше грабли, т. е. начать анализировать оперативные данные, которые напрямую для анализа непригодны.

Законы OLAP

Основоположником термина OLAP является Эдгар Кодд, известный как классик теории реляционных баз данных. В 1993 году он опубликовали статью «Обеспечение OLAP (оперативной аналитической обработки) для пользователей-аналитиков», в которой были изложены 12 законов, заложившие основу концепции аналитической обработки данных в реальном времени. Позднее, в 1995 году эти правила были дополнены еще шестью. Ниже будет рассмотрен полный список выдвинутых Коддом требований к OLAP-системам, позволяющий глубже понять их идеологические основы.

1. *Многомерное концептуальное представление данных*. Эта особенность – основа технологии OLAP. Вместо привычной модели данных реляционных источников, основанной на плоской системе координат, пользователь получает в свое распоряжение интуитивно понятную многомерную модель, в которой данные организуются в виде многомерных кубов (гиперкубов). Осями многомерной системы координат служат основные атрибуты анализируемого бизнес-процесса (товар, регион, тип покупателя, время и т.д.). На пересечениях осей (измерений) многомерной системы координат находятся данные, количест-

венно характеризующие процесс – меры (объемы, остатки на складе, издержки и т. п.).

2. *Интуитивное манипулирование данными*.

3. *Доступность*: OLAP как посредник между гетерогенными источниками данных и представлением для конечного пользователя.

4. *Пакетное извлечение против интерпретации*. Требуется, чтобы продукт в равной степени эффективно обеспечивал доступ и к собственному хранилищу данных, и к внешним данным.

5. *Модели анализа OLAP*. Требуется, чтобы OLAP-системы поддерживали формирование настраиваемых отчетов, формирование разрезов и группировок данных, проверку гипотез (ответы на вопрос «что, если…?») и модели поиска целей.

6. *Архитектура «клиент-сервер»*. Требуется также, чтобы серверный компонент был бы достаточно интеллектуальным для того, чтобы различные клиенты могли подключаться с минимумом усилий и программирования.

7. *Прозрачность*. Это требование означает, что пользователь получает все необходимые данные из OLAP-машины, не подозревая, откуда они берутся.

8. *Многопользовательская поддержка*. Инструменты OLAP должны обеспечивать одновременный доступ (чтение и запись), интеграцию и конфиденциальность.

9. *Обработка ненормализованных данных*. Данное требование указывает на необходимость интеграции между OLAP-машиной и ненормализованными источниками данных. Модификации данных, выполненные в среде OLAP, не должны приводить к изменениям данных, хранимых в исходных внешних системах.

10. *Сохранение результатов OLAP: хранение их отдельно от исходных данных*.

11. *Исключение отсутствующих значений*. Отсутствующие значения должны отличаться от нулевых значений.

12. *Обработка отсутствующих значений*. Все отсутствующие значения будут игнорироваться OLAP-анализатором без учета их источника.

13. *Гибкость формирования отчетов*. Измерения должны быть размещены в отчете так, как это нужно пользователю.

14. *Стандартная производительность отчетов*. Требуется, чтобы производительность формирования отчетов существенно не падала с ростом количества измерений и размеров базы данных.

15. *Автоматическая настройка физического уровня*. Требуется, чтобы OLAP-системы автоматически настраивали свою физическую схему в зависимости от типа модели, объемов данных и разреженности базы данных.

16. *Универсальность измерений*. Все измерения должны быть равноправны, каждое измерение должно быть эквивалентно и в структуре, и в операционных возможностях.

17. *Неограниченное число измерений и уровней агрегации*. Кодд предлагает, что в случае принятия некоторого максимума, он должен обеспечивать хотя бы 15 измерений, а предпочтительнее – 20.

18. *Неограниченные операции между размерностями*. Все виды операций должны быть дозволены для любых измерений.

Альтернативным набором критериев определения OLAP является широко известный сформулированный Найджелом Пендсом и Ричардом Критом в 1995 г. *тест FASMI*, или *Fast Analysis of Shared Multidimensional Information* – Быстрый Анализ Разделяемой Многомерной Информации:

- FAST (Быстрый) – означает, что система должна обеспечивать выдачу большинства ответов пользователям в сжатые сроки; при этом самые простые запросы обрабатываются в течение одной секунды и лишь немногие – более 20;
- ANALYSIS (Анализ) – означает, что система может справляться с любым логическим и статистическим анализом, характерным для данного приложения, и обеспечивает его сохранение в виде, доступном для конечного пользователя;
- SHARED (Разделяемый) – означает, что система осуществляет все требования защиты конфиденциальности (возможно до уровня ячейки) и, если множественный доступ для записи необходим, обеспечивает блокировку модификаций на соответствующем уровне;

- MULTIDIMENSIONAL (Многомерный) – означает, что система должна обеспечить многомерное концептуальное представление данных, включая полную поддержку для иерархий и множественных иерархий; многомерность – ключевой критерий;
- INFORMATION (Информация) – требуемая информация должна быть получена там, где она необходима.

В табл. 14 показаны основные характеристики технологий OLAP и OLTP, чтобы окончательно развести эти два понятия.

Таблица 13 Сравнительная характеристика технологий OLAP и OLTP

Признак сравнения	OLAP	OLTP
Объем хранимой информации	Большой объем информации	Большой объем информации
Хранение данных	Синхронизированная информация из различных баз данных с использованием общих классификаторов	Зачастую различные базы данных для отдельных подразделений
Отношение к нормализации	Ненормализованная схема, существуют дубликаты данных	Нормализованная схема, дубликаты данных отсутствуют
Частота изменения данных	Производится редко через пакетную загрузку	Интенсивное изменение данных
Специфика режима работы с данными	Система выполняет сложные ранее не регламентированные запросы над большими объемами данных с широким применением агрегатных функций, группировок; анализ временных зависимостей	Система работает в транзакционном режиме; транзакции малы по объему обрабатываемой информации; наборы процедур, запросов определены заранее
Пользователи	Малое количество пользователей (менеджеры, аналитики)	Большое количество пользователей-операторов

Как отмечалось ранее OLTP-системы, приспособленные для хранения данных, оказались непригодными для задач аналитиков. OLAP-системы же опти-

мизированы для выполнения операций чтения над большими объемами данных. Высокая скорость выполнения сложных аналитических запросов OLAP-системами связана с особенностями построения используемых ими многомерных структур (многомерные базы данных, или OLAP-кубы):

- OLAP-системы строятся на базе денормализованных источников – хранилищ данных; в итоге в базе данных OLAP могут содержаться избыточные данные, но в то же время положительным моментом в упрощении структуры связей таблиц является повышение скорости выполнения запросов;
- многомерный куб содержит в себе не только сами данные, но и их агрегаты (обобщенные показатели) по различным измерениям; то есть в базе данных OLAP хранятся заранее посчитанные системой показатели, которые потенциально могут потребоваться бизнес-аналитику.
- Многомерность в OLAP-системах может быть представлена на трех уровнях:
- многомерное представление данных – средства на стороне клиента, обеспечивающие многомерную визуализацию и манипулирование данными; слой многомерного представления абстрагирован от физической структуры данных и воспринимает данные как многомерные;
- многомерная обработка – средство (язык) формирования многомерных запросов (традиционный реляционный язык SQL здесь оказывается непригодным) и процессор, умеющий обработать и выполнить такой запрос;
- многомерное хранение – средства физической организации данных, обеспечивающие эффективное выполнение многомерных запросов.

Первые два уровня в обязательном порядке присутствуют во всех OLAP-средствах. Третий уровень, хотя и является широко распространенным, не обязателен, так как данные для многомерного представления могут извлекаться и из обычных реляционных структур; процессор многомерных запросов в этом случае транслирует многомерные запросы в SQL-запросы, которые выполняются реляционной СУБД.

Виды OLAP-серверов

В соответствии с требованием прозрачности OLAP-систем способ реализации многомерной модели скрыт от пользователя. Однако способ реализации важен, поскольку от него зависят производительность решения и требуемые ре-

сурсы. Существует три основных способа реализации многомерной модели: MOLAP, ROLAP, HOLAP.

MOLAP (Multidimensional OLAP, или многомерный OLAP) – исходные и агрегатные данные хранятся в многомерной базе данных. Хранение данных в многомерных структурах позволяет манипулировать данными как многомерным упорядоченным массивом, благодаря чему скорость вычисления агрегатных значений одинакова для любого из измерений.

Физически данные хранятся в «плоских» файлах, при этом куб представляется в виде одной плоской таблицы, в которую построчно вписываются все комбинации элементов всех измерений с соответствующими им значениями мер.

В силу своих особенностей использование MOLAP является эффективным при следующих условиях:

- объем исходных данных для анализа не слишком велик (не более нескольких гигабайт), т. е. уровень агрегации данных достаточно высок;
- набор информационных измерений стабилен (MOLAP чувствителен к изменению многомерной модели);
- наименьшее время отклика системы на нерегламентированные запросы является критичным параметром (MOLAP обеспечивает высокую скорость поиска и выборки);
- требуется использование сложных встроенных функций для выполнения вычислений над ячейками куба, возможность написания пользовательских функций (MOLAP легко справляются с задачами включения в информационную модель разнообразных встроенных функций).

ROLAP (Relational OLAP, или реляционный OLAP) – исходные данные остаются в той же реляционной базе данных, где они изначально и находились. Агрегатные же данные помещают в специально созданные для их хранения служебные таблицы в той же базе данных.

Достоинства ROLAP:

- в большинстве случаев корпоративные хранилища данных реализуются средствами реляционных СУБД, и инструменты ROLAP позволяют производить анализ непосредственно над ними;
- в случае переменной размерности задачи, когда изменения в структуру измерений приходится вносить достаточно часто, ROLAP-системы с

динамическим представлением размерности являются оптимальным решением, т. к. в них такие модификации не требуют физической реорганизации БД;

- реляционные СУБД обеспечивают значительно более высокий уровень защиты данных и хорошие возможности разграничения прав доступа.

Недостатки ROLAP:

- меньшая производительность в сравнении с MOLAP. Для обеспечения производительности, сравнимой с MOLAP, реляционные системы требуют тщательной проработки схемы базы данных и настройки индексов.

HOLAP (Hybrid OLAP, или гибридный OLAP) – исходные данные остаются в той же реляционной базе данных, где они изначально находились, а агрегатные данные хранятся в многомерной базе данных. Серверы HOLAP применяют подход ROLAP для разреженных областей многомерного пространства и подход MOLAP – для плотных областей. Серверы HOLAP разделяют запрос на несколько подзапросов, направляют их к соответствующим фрагментам данных, комбинируют результаты, а затем предоставляют результат пользователю.

Понятие OLAP-куба

Главный постулат OLAP – многомерность в представлении данных. В терминологии OLAP для описания многомерного дискретного пространства данных используется понятие куба, или гиперкуба.

Куб представляет собой многомерную структуру данных, из которой пользователь-аналитик может запрашивать информацию. Кубы создаются из **фактов и измерений**.

Факты – это данные об объектах и событиях в компании, которые будут подлежать анализу. Факты одного типа образуют меры (measures). Мера есть тип значения в ячейке куба.

Измерения – это элементы данных, по которым производится анализ фактов. КолТема таких элементов формирует атрибут измерения (например, дни недели могут образовать атрибут измерения «время»). В задачах бизнес-анализа коммерческих предприятий в качестве измерений часто выступают такие категории, как «время», «продажи», «товары», «клиенты», «сотрудники», «географическое местоположение».

Измерения чаще всего являются иерархическими структурами, представляющими собой логические категории, по которым пользователь может анализировать фактические данные. Каждая иерархия может иметь один или не-

сколько уровней. Так иерархия измерения «географическое местоположение» может включать уровни: «страна – область – город». Измерение может иметь несколько иерархий (при этом каждая иерархия одного измерения должна иметь один и тот же ключевой атрибут таблицы измерений).

Куб может содержать фактические данные из одной или нескольких таблиц фактов и чаще всего содержит несколько измерений. Любой конкретный куб обычно имеет конкретный направленный предмет анализа.

На рис. 29 показан пример куба, предназначенного для анализа продаж продуктов нефтепереработки некоторой компанией по регионам. Данный куб имеет три измерения (время, товар и регион) и одну меру (объем продаж, выраженный в денежном эквиваленте). Значения мер хранятся в соответствующих ячейках (cell) куба.

Каждая ячейка уникально идентифицируется набором членов каждого из измерений, называемого кортежем. Например, ячейка, расположенная в нижнем левом углу куба (содержит значение $98399), задается кортежем [Июль 2005, Дальний Восток, Дизель]. Здесь значение $98399 показывают объем продаж (в денежном выражении) дизеля на Дальнем Востоке за июль 2005 года.

Стоит обратить также внимание на то, что некоторые ячейки не содержат никаких значений: эти ячейки пусты, потому что в таблице фактов не содержится данных для них.

Конечной целью создания подобных кубов является минимизация времени обработки запросов, извлекающих требуемую информацию из фактических данных. Для реализации этой задачи кубы обычно содержат предварительно вычисленные итоговые данные, называемые агрегациями (aggregations). Т.е. куб охватывает пространство данных большее, чем фактическое – в нем существуют логические, вычисляемые точки.

Вычислять значения точек в логическом пространстве на основе фактических значений позволяют функции агрегирования. Наиболее простыми функциями агрегирования являются SUM, MAX, MIN, COUNT. Так на рис. 29, например, используя функцию MAX, для приведенного в примере куба можно выявить, когда произошел пик продаж дизеля на Дальнем Востоке и т.д.

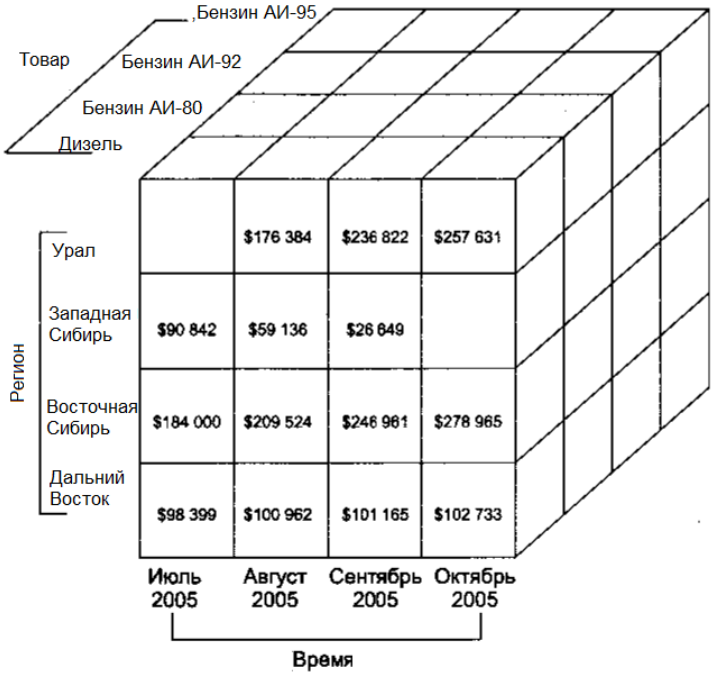


Рисунок 19. Куб с информацией о продажах нефтепродуктов в различных регионах

Решением этой проблемы является внедрение специального атрибута, объединяющего все элементы измерения. Этот атрибут (создается автоматически) содержит всего один элемент – All («Все»). Для простых функций агрегирования, например, суммы, элемент All эквивалентен сумме значений всех элементов фактического пространства данного измерения.

Важной концепцией многомерной модели данных является подпространство, или подкуб (sub cube). Подкуб представляет собой часть полного пространства куба в виде некоторой многомерной фигуры внутри куба. Так как многомерное пространство куба дискретно и ограничено, подкуб также дискретен и ограничен.

Операции над OLAP-кубами

Над OLAP-кубом могут выполняться следующие операции:

1. срез;
2. вращение;
3. консолидация;
4. детализация.

На рис. 30 показан срез, он является частным случаем подкуба. Это процедура формирования подмножество многомерного массива данных, соответствующее единственному значению одного или нескольких элементов измерений, не входящих в это подмножество.

Например, чтобы узнать, как продвигались продажи нефтепродуктов во времени только в определенном регионе, а именно на Урале, то необходимо зафиксировать измерение «Товары» на элементе «Урал» и извлечь из куба соответствующее подмножество (подкуб).

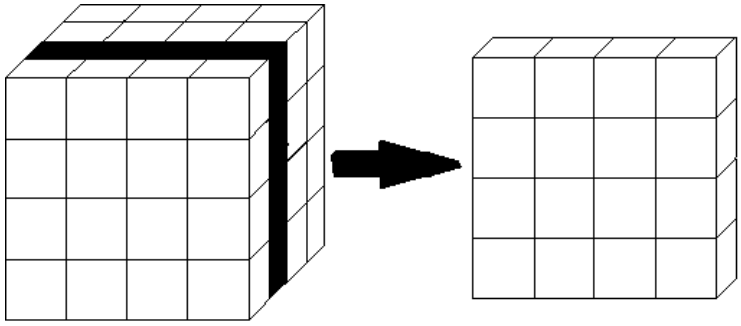


Рисунок 20. Срез OLAP-куба

На рис. 31показано вращение – операция изменения расположения измерений, представленных в отчете или на отображаемой странице. Например, операция вращения может заключаться в перестановке местами строк и столбцов таблицы. Кроме того, вращением куба данных является перемещение вне табличных измерений на место измерений, представленных на отображаемой странице, и наоборот.

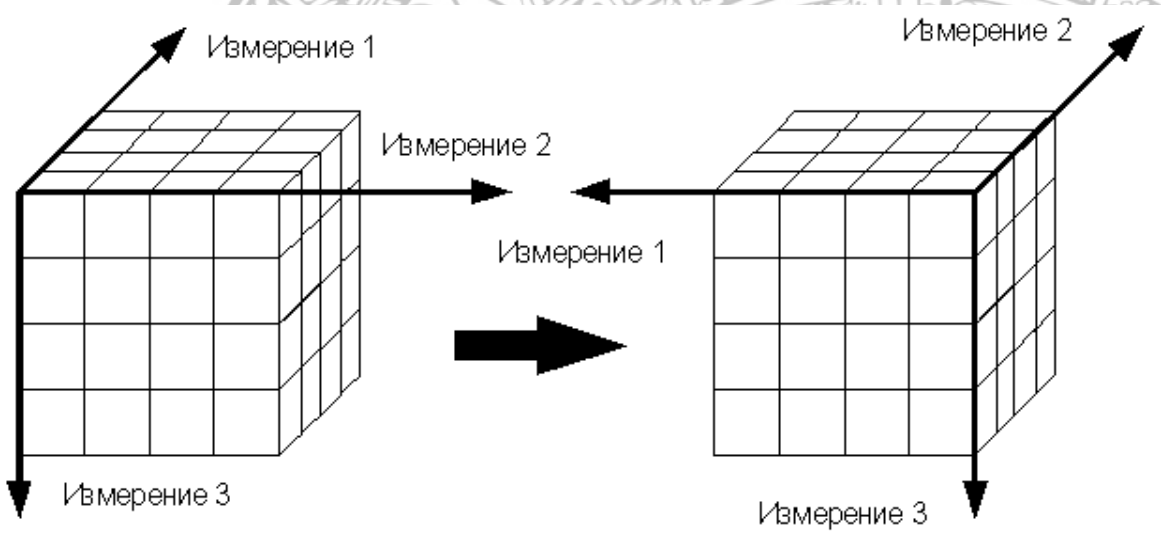


Рисунок 31. Вращение OLAP-куба

На рис. 32 консолидация – операция перехода от детального представления данных к агрегированному. Например, переход к просмотру данных о продажах не по месяцам, а по годам.

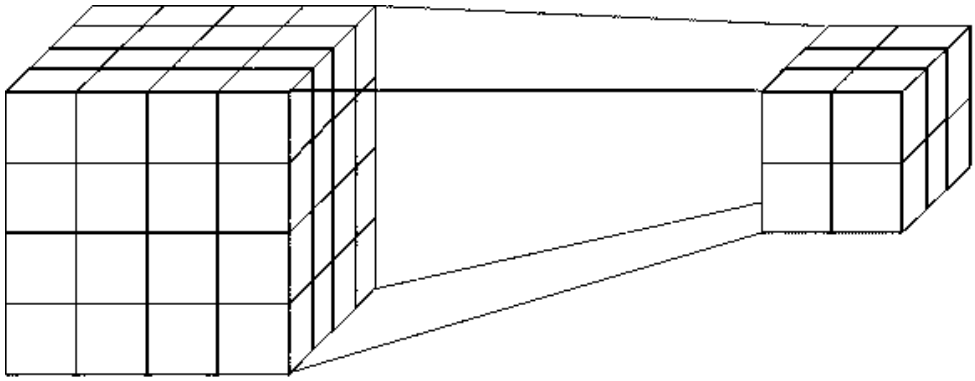


Рисунок 32. Консолидация данных OLAP-куба

На рис. 32выполнена детализация – обратная консолидации операция, которая определяет переход от агрегированного представления данных к детальному.

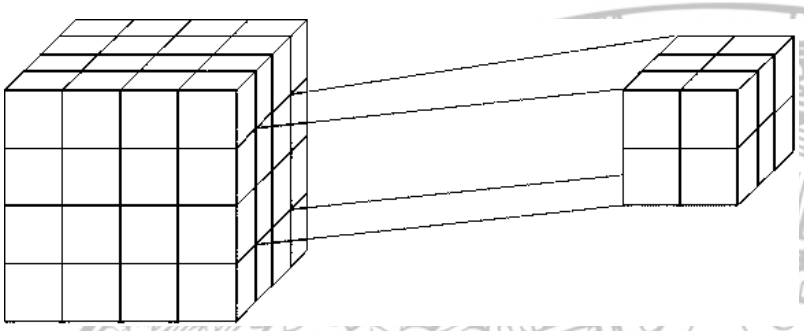


Рисунок 33. Детализация данных OLAP-куба

Направление детализации (и консолидации) может быть задано как по иерархии отдельных измерений, так и согласно прочим отношениям, установленным в рамках измерений или между измерениями.

Использование технологии OLAP

Существует несколько категорий продуктов, обеспечивающих ту или иную часть функциональности OLAP. В первую очередь, их можно разбить на OLAP-серверы и OLAP-клиенты. OLAP-серверы обеспечивают создание и наполнение кубов, а также выполнение многомерных запросов и передачу многомерных данных клиенту, реализуя при этом какой-то из интерфейсов обмена, который может быть стандартным, либо принятым у одного разработчика

OLAP-решений. OLAP-клиенты предоставляют возможность работы с многомерными данными, их визуализации и пользовательской обработки. Они подразделяются на группы в зависимости от функциональной нагруженности.

Самым простым OLAP-клиентом является такой, который не может работать без OLAP-сервера. Такой клиент образует интерактивную оболочку для доступа к данным OLAP-сервера (примером является Analysis Manager из набора MS SQL Server 2000 Analysis Services). Более сложные клиенты могут, как работать с OLAP-серверами, так и создавать клиентские кубы из реляционных баз и сохранять их для локальной работы. Наиболее популярным из OLAP-клиентов этого вида является Microsoft Excel (другим примером может быть Cube Analyser из пакета Deductor, Loginom).

Тема 14. Аналитические платформы. Инструменты бизнес-аналитики

Даже эффективные технологии извлечения закономерностей из данных, такие как Knowledge Discovery и Data Mining, не представляют особой ценности без инструментальной поддержки в виде соответствующего программного обеспечения. Рынок программных средств продолжает формироваться по сей день. На рис. 34 представлена следующая классификация программного обеспечения бизнес-аналитики.

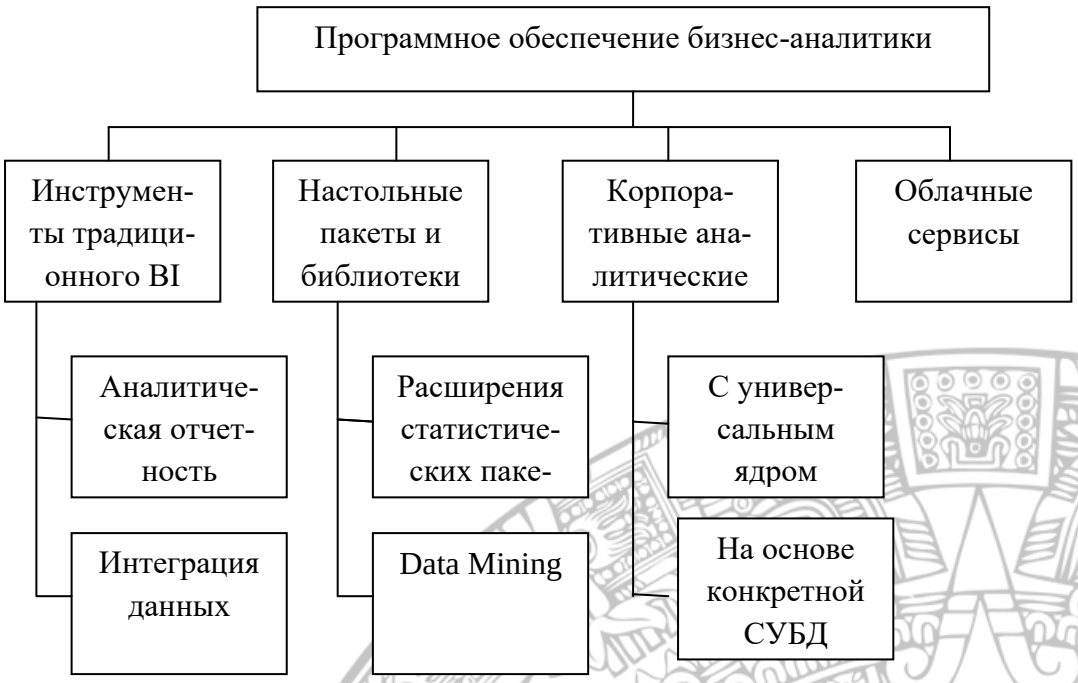


Рисунок 34. Вариант классификации программных решений в области бизнес-аналитики

Первая категория программных средств – инструменты традиционной бизнес-аналитики – является базовой для компаний, и включает в себя средства формирования и выполнения нерегламентированных запросов и отчетности, а также выполнение простого анализа. Сюда входят нерегламентируемые запросы, отчеты, средства многомерного анализа, инструментальные панели, диаграммы и графики, а также несложные вычисления «на лету».

Задача построения полноценной системы аналитической отчетности, как правило, не обходится без решения вопросов интеграции данных из разнородных источников, поэтому к традиционным инструментам также относят средства извлечения данных из внешних источников, их преобразования и очистки,

чтобы они соответствовали нуждам бизнес-модели компании. В качестве консолидированного источника данных обычно выступают: хранилище данных, витрины данных, оперативный склад данных, база данных с мастер-данными.

Для настольного применения востребованы пакеты и библиотеки с алгоритмами Data Mining. Они существуют двух видов: как получившие развитие *статистические и математические* пакеты *прикладных программ,* в которые добавились новые алгоритмы, и *приложения нового образца,* изначально создававшиеся как инструменты *Data Mining,* с визуальным проектированием логики обработки данных.

Такие пакеты ориентированы в основном на профессиональных пользователей. Их отличительные особенности:

- слабая интеграция с промышленными источниками данных;
- конвейерная (поточная) обработка новых данных затруднительна или реализуется встроенными языками программирования и требует высокой квалификации;
- из-за использования пакетов на локальных рабочих станциях обработка больших объемов данных затруднена. В статистических пакетах, кроме того, бедные средства очистки и предобработки данных

Плюсом статистических пакетов является их широкая распространенность. Настольные *Data Mining* пакеты могут быть ориентированы на решение всех классов задач *Data Mining* или какого-либо одного, например кластеризации или классификации. Вместе с тем эти пакеты предоставляют богатые возможности в плане алгоритмов, что достаточно для решения исследовательских задач. Существует немало свободно распространяемых настольных пакетов *Data Mining* с открытыми исходными кодами.

Создание эффективных прикладных решений промышленного уровня с помощью таких пакетов затруднено.

Аналитические платформы

Аналитические платформы изначально ориентированы на комплексный анализ данных и предназначены для создания и эксплуатации прикладных корпоративных решений в различных областях.

Аналитическая платформа – специализированное программное решение (или набор решений), которое содержит в себе все инструменты для извлечения закономерностей из «сырых» данных: средства интеграции данных и

управления метаданными, извлечение и преобразование данных, алгоритмы средства визуализации и распространения результатов среди пользователей, а также возможности «конвейерной» обработки новых данных, коллективную работу над моделями, включая управление их версиями и разграничение прав доступа.

Аналитическая платформа обычно состоит из следующих компонентов.

- Аналитический сервер.
- Клиентское приложение – рабочее место аналитика.
- Клиентское приложение – рабочее место пользователя.
- Подсистема управления метаданными и реализация многомерной модели, данных (хранилище или витрина данных).
- Репозитарий моделей.
- Интеграционный сервер.

Рассмотрим их подробнее.

Ядро платформы – аналитический сервер. Через него связаны и функционируют все остальные компоненты платформы. Кроме того, сервер призван «прогонять» новые данные через существующие модели.

Модели, описывающие выявленные закономерности, правила и прогнозы, находятся в специальном источнике данных – репозитории моделей. Отметим две важные функциональные особенности репозитория:

- поддержка разграничения прав доступа к моделям и коллективной работы над ними;
- поддержка версионности моделей.

В аналитической платформе, как правило, присутствуют гибкие и развитые средства интеграции данных и работы в едином пространстве метаданных, для чего используется реальный или виртуальный многомерный источник: хранилище, витрина данных.

В аналитической платформе, как правило, присутствуют гибкие и развитые средства интеграции данных и работы в едином пространстве метаданных, для чего используется реальный или виртуальный многомерный источник: хранилище, витрина данных. На рис. 35 показана взаимосвязь компонентов аналитической платформы.

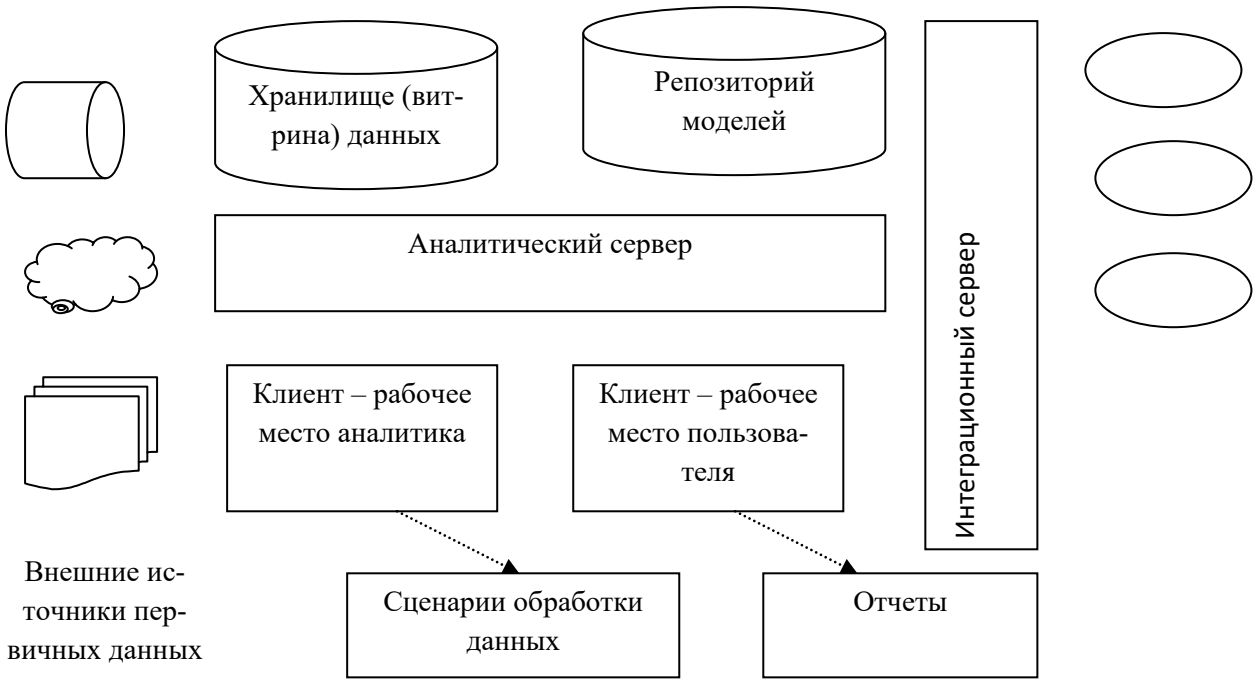


Рисунок 35. Взаимосвязь компонентов аналитической платформы

Логику обработки данных, построение и использование моделей, панели отчетов и визуализаторов реализует аналитик через специальное клиентское приложение. Через клиентские приложение работает и конечный пользователь – разница между ними в том, что пользователю, как правило, доступны отчеты, которые он может просматривать, но не изменять.

Пользовательское приложение-клиент обычно обращается к аналитическому серверу посредством веб-интерфейса; приложение для аналитика может быть как с веб-интерфейсом, так и полноценным приложением операционной системы.

Еще один важный компонент аналитической платформы – интеграционный сервер. Он предоставляет специальный механизм обмена данными со сторонними приложениями, реализует так называемую сервис-ориентированную архитектуру. Обмен данными осуществляется посредством XML-запросов. На основе этой технологии строятся большинство сервисов, предоставляющих бизнес-аналитику как услугу (например, продажа прогнозов финансовых котировок на основе моделей Data Mining).

На рис. 36 показаны компоненты аналитической платформы.

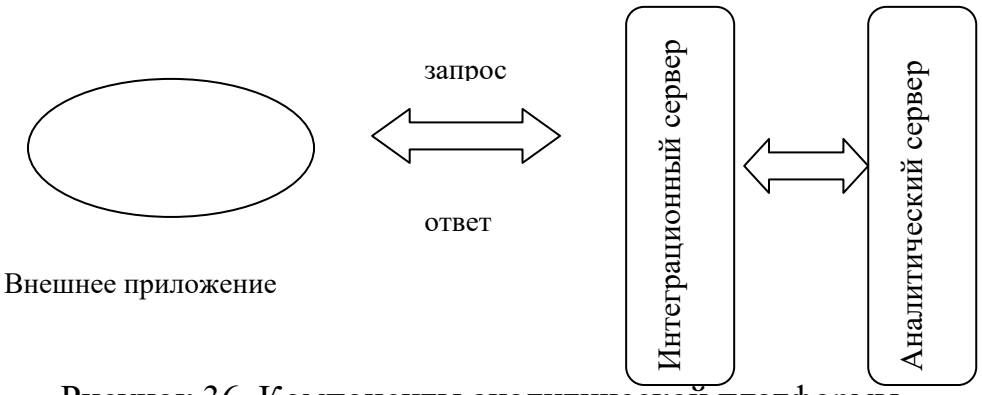


Рисунок 36. Компоненты аналитической платформы

Реализация компонентов аналитической платформы может быть «привязана» к определенной *системе управления базами данных* (СУБД). Инструменты бизнес-аналитики как бы встраиваются в СУБД. Отличительные особенности такого подхода:

- высокая производительность;
- алгоритмы анализа данных по максимуму используют преимущества СУБД;
- жесткая привязка всех технологий анализа к одной СУБД;
- сложность в создании прикладных решений, поскольку работа с СУБД ориентирована на программистов и администраторов баз данных.

Основным препятствием на пути все более широкого применения методов и программных средств анализа данных является сложность инструментов. Поэтому важно освободить аналитика от необходимости программировать код или макросы. Своеобразным ответом на это требование стало появление **языков визуального моделирования.** Сегодня их наличие является стандартом де-факто в полноценной аналитической платформе.

Язык моделирования позволяет аналитику в визуальной среде строить последовательности шагов по обработке данных от получения «сырых» данных до конечного результата. Шаги представляют собой набор атомарных по отношению к данным операций, каждую из которых можно представить отдельным узлом Примеры таких операций: *выборка данных, фильтрация, сортировка, добавление нового столбца, построение модели* и т. п. Набор узлов образует графическую диаграмму. Такой способ представления очень близок к рассуждениям и действиям аналитика, которые он так или иначе проделывает.

Общие особенности языков моделирования в аналитических платформах следующие.

- Базовым узлом, с которого начинается диаграмма, является узел импорта, поскольку в аналитических платформах обычно отсутствуют средства для ручного ввода данных; предполагается, что данные уже имеются в каких-либо источниках.
- Графическое изображение, соответствующее какому-либо узлу несет в себе большой семантический смысл. Оно помогает аналитику различать узлы по функциям и определять их активность (часто еще не выполненный узел обозначается иконкой серого цвета, а выполненный-цветной).
- Диаграмма описывает формализованную последовательность действий над данными, и эти действия можно повторить на совершенно других данные предварительно настроив соответствие входов.

Существуют две формы представления диаграмм в виде дерева и в виде графа. Каждая из форм представления имеет как достоинства. Так и недостатки. У деревьев более жесткая структура по сравнению с графами, поэтому, к примеру, отображение двух узлов, сливающихся в один, затруднено. Вместе с тем дерево более компактно (в графе обязательно присутствие узлов, которые занимают место на диаграмме), что очень важно при большом количестве узлов, и позволяет выполнять множество интуитивно понятных операций, связанных с манипулированием ветвями (копирование, удаление, перенос и так далее). На рынке аналитических платформ наиболее распространена форма диаграмм в виде графов.

Облачные сервисы

Сегодня все больше и больше направлений бизнеса работают на **облачных технологиях –** это когда появляется возможность получить вычислительные мощности и программное обеспечение «как услугу», а это значит; что заказчику не нужно заботиться ни о работоспособности инфраструктуры, ни о программном обеспечении – эти обязанности теперь лежат на плечах поставщика облачных услуг (IaaS – IT- инфраструктура как услуга; SaaS – программное обеспечение как услуга).

Использование решений бизнес-аналитики в виде облачного сервиса подходит не всем компаниям, оно малоприменимо для организаций, которые

работают с секретными данными. Тем не менее, с каждым годом все больше компаний выбирают SaaS – модель как более экономичную и достаточно надежную. Многие вендоры-производители программного обеспечения переводят свои инструменты в облачные технологии. Таким образом, аналитики строят модели Data Mining «в облаке», после чего экспортируют их и встраивают в свои бизнес-процессы.

Отметим, что к облачной аналитике относится и понятие модель как услуга: использование результатов моделирования (модели разработаны поставщиком облачной услуги) путем получения прогнозов, оценок; вероятностей и т.п. для заданных входных воздействий. Яркий пример – покупка скорингового балла заемщика в бюро кредитных историй (балл рассчитывается моделью или каскадом моделей, построенными на данных, накопленных в бюро кредитных историй).

Тема 15. Большие данные. Наука о данных

Технологии *Knowledge и Data Mining* развиваются и совершенствуются уже более двух десятилетий, и за это время стали основным инструментом поиска новых знаний в массивах данных необходимых для принятия эффективных управленческих решений. На основе Data Mining с успехом строятся системы углубленной бизнес-аналитики, ставшие привычным элементом информационного ландшафта компаний.

Параллельно с методологией анализа развивались и средства доставки, хранения интеграции данных что было обусловлено непрерывным возрастанием их объема, территориальной распределенности, сложности, требованиям по более адекватному описанию окружающего мира и событий в нем. Постепенно становилось очевидным, что потребности общества в области обработки «сырых» данных с целью их преобразования в полезные знания уже не могли удовлетворяться развитием существующих подходов и методов, а требовали переосмысления с учетом новых реалий.

Также назрел конфликт в терминологии. Обилие терминов и их трактовок; так или иначе связанных с анализом данных, способно запутать даже опытного исследователя.

В данной лекции мы обсудим термины, характеризующие последние тенденции в анализе данных – *Большие данные (Big Data) и Наука о данных (Data Science).*

Предпосылки появления Big Data

Современные тренды в развитии бизнес-аналитики накладывают рад ограничений на использование приложений *Data Mining.* Основным из этих ограничений является то, что технологии *Data Mining* ориентированы, прежде всего, на обработку структурированных данных. Между тем, все больший интерес представляют собой данные, поступающие в режиме реального времени из социальных медиа, видео и фото регистраторов, электронной почты и других распределенных источников, расположенных во внешнем окружении. Основным свойством таких источников является наличие н врастающего высокоскоростного потока данных с неопределенной структурой. При этом

объем данных, фактически, ничем не ограничен, и может достигать терабайт и даже петабайт.

За последние несколько лет человечество произвело информации больше, чем за всю историю своего существования. И рост продолжается экспоненциально. Так; согласно данным исследовательской компании IDC**,** к 2020-му году объем данных в компаниях вырастет в 50 раз по сравнению с текущим состоянием.

Это во многом было обусловлено экспоненциальным ростом количества вычислительных средств, приложений и пользователей, участвующих в формировании глобальных потоков данных. Современная эпоха-эпоха мобильной связи, мобильного Интернета, социальных сетей, блогов привела к появлению миллиардов пользователей и миллионов приложений. Динамика развития инфраструктуры данных представлена в табл. 15.

Таблица 14 Динамика развития инфраструктуры данных

Эпоха	Технологии	Пользователи	Приложения
До середины 1980-х	Терминалы Мейнфреймы	Миллионы	Тысячи
Середина 1 980-х-2010	Клиент-серв ер ЛВС Интернет Персональные компьютеры	Сотни миллио-нов	Десятки тысяч
2010-по настоящее время	Мобильные устройства Со-циальные сети Мобильные приложения Аналитика Больших данных	Миллиарды	Миллионы

Термин Big Data

Необходимость обработки качественно новых объемов структурированных и неструктурированных данных показала, что традиционные подходы к их хранению и анализу стали неэффективными, а, следовательно, необходимы новые технологии. Аналитики рассуждают следующим образом, Мы не знаем, нужна ли нам информация, а если нужна, то какая, до тех пор, пока не проанализируем ее.

Стоимость хранения информации настолько снизилась, что появилась возможность собирать всё больше данных и анализировать их, руководствуясь принципом, Мы не знаем, чего мы не знаем. Например, может быть

обнаружено, что площадь того или иного цвета на обложке журнала влияет на вероятность его продаж в определенном периоде.

Итак; возникла проблема построения новой вычислительной инфраструктуры, которая была бы эффективной и не очень дорогой. Ключом к построению такой инфраструктуры и стал комплекс технологий, известный в настоящее время как Большие данные – Big Data.

Попытка разобраться, что собой представляют Большие данные, вооружившись только привычными понятиями и терминами анализа данных, вряд ли увенчается успехам. Некоторые авторы рассматривают Big Data как Data Mining с кардинально увеличенными возможностями в плане объемов хранимых и обрабатываемых данных, а также скорости доступа к ним. Другие авторы рассматривают Data Mining как небольшую составляющую Big Data. А третьи вообще не упоминают Data Mining в контексте Big Data. И все же попробуем разобраться что, с точки зрения специалиста по бизнес-аналитике, представляет собой термин Big Data, в чём его принципиальное отличие от Data Mining, какие новые перспективы и возможности открывают Большие данные?

Приведем примеры источников, порождающих Большие данные.

- Крупные розничные торговые сети регистрируют ежечасно миллионы клиентских транзакций, которые пересылаются в хранилища данных, объем которых может составлять несколько петабайт.
- Более 5 миллиардов людей по всему миру говорят; обмениваются сообщениями и производят поиск в Интернет с помощью мобильных устройств.
- Тысячи автоматических регистраторов по всему миру непрерывно фиксируют погодные условия, и передают метеорологические данные в центры их обработки.
- Пользователи социальных сетей ежеминутно отправляют десятки миллионов сообщений.

Термин Big Data шире, чем просто очень много данных: в разном контексте под ним могут подразумеваться и данные большого объема, и технология их обработки, а также проекты, рынок и даже компании, активно использующие эти технологии.

Характеристики технологии Big Data

Термин Big Data, Большие данные**;** в научный и корпоративный обиход вошел в 2008 году. С точки зрения бизнес-аналитики, Big Data можно определить как технологию в области аппаратного и программного обеспечения, которая интернирует, организует, управляет и анализирует данные, характеризующиеся четырьмя характеристиками: объемам, разнообразием, изменчивостью и скоростью.

Поскольку в англоязычном варианте эти характеристики обознаются, соответственно Volume, Variety, Variability и Velocity, то их часто называют четыре V. Таким образом, если до сих пор характеристикой данных, определяющей организацию их обработки, был объем, то Big Data предполагает использование трех дополнительных параметров:

- разнообразие – отражает тот факт, что в отличие ль технологий Data Mining, ориентированных на анализ данных, структурированных в виде таблиц, технологии Big Data должны позволять обрабатывать неструктурированные данные самых различных форматов, в том числе, текст, аудио и видео;
- изменчивость – возможность проводить обработку данных, которые могут непрерывно изменяться, что отличается от концепции хранения данных, используемых в бизнес-аналитике с базовым принципом не изменчивости данных;
- скорость – указывает на то, что требуется анализировать данные, которые не являются заранее консолидированными и в некотором неизменчивом источнике, а представлены непрерывным потоком, поступающим по телекоммуникационным каналам.

У компаний, которым необходимо хранить и анализировать непрерывно возрастающие объемы данных, есть возможность выбора двух направлений развития вычислительной инфраструктуры.

Первый – приобрести более мощный компьютер с большим количеством процессоров, объемом оперативной памяти, дискового пространства и так далее. Это *называется масштабированием по вертикали,* то есть добавление ресурсов на единственный вычислительный узел.

На рис. 37 показаны характеристики технологии Big Data.

Рисунок 37 Характеристики технологии Big Data

Второй вариант – *горизонтальное масштабирование,* которое базируется на добавлении дополнительных вычислительных узлов, то есть предоставляет возможность добавлять в систему дополнительные компьютеры и распределять работу между ними. Горизонтальное масштабирование позволяет построить высоконадежное решение с обеспечением должной степени резервирования на базе недорогих стандартных компьютеров. Сотни и даже тысячи маломощных компьютеров, объединенных в *кластер,* могут обеспечивать вычислительную мощность суперкомпьютеров.

Для решения задач аналитической обработки массивов данных, которые по своей локализации, размерам и структуре соответствуют *Большим данным,* используются технологии распределенных вычислений: вычислительная нагрузка распределяется между некоторым количеством (чем больше, тем лучше) компьютеров-клиентов, которые работают под управлением некоторого управляющего центрального компьютера. Последний распределяет «задания» между клиентскими машинами, получает результаты обработки и формирует из них общий результат. Современные реализации распределенных вычислительных систем позволяют эффективно обрабатывать терабайты, петабайты и даже экзабайты данных.

Инструменты распределения вычислений для Big Data

Для того чтобы построить инфраструктуру для больших данных понадобятся две подсистемы: большое число компьютеров (узлов) относительно небольшой мощности, объединенных в вычислительную сеть (кластер), а также программное обеспечение; способное распределять вычислительные ресурсы сети при решении сложных задач.

Кратко рассмотрим программные инструменты, которые в литературе наиболее часто упоминают как основу создания информационной инфраструктуры для Big Data – MapReduce, Hadoop и NoSQL.

MapReduce – модель распределённых вычислений, разработанная компанией **Google,** используемая для параллельных вычислений над очень большими (несколько петабайт) массивами данных в распределенных вычислительных сетях. Компьютеры в таких сетях делятся на узлы, которые непосредственно производят вычисление и главные узлы, которые получают задачу, разделяют ее на части и распределяют ее между рабочими узлами для предварительной обработки. Данный шаг называется **map.**

После того, как мастер-узел получает от остальных машин сообщение о том, что обработка данных и ми закончена (то есть шаг **map** завершён), он выдает команду на переход **reduce** (свертка), в процессе которого формируется результат, возвращаемый на мастер узел для формирования итогового решения

При этом MapReduce это не какая-то конкретная программа, а метод организации распределенных вычислений, который может быть реализован с помощью программы, написанной на каком-то, наиболее удобном в конкретном случае языке (например, в реализации MapReduce в Google используется язык C++).

Hadoop – проект фонда **Apache Software Foundation,** свободно распространяемый набор утилит, библиотеки программный каркас для разработки и выполнения распределённых программ, работающих на кластерах из сотен и тысяч узлов. Используется для реализации поисковых и контекстных механизмов многих высокой загруженных веб-сайтов. Разработан на основе модели распределенных вычислений MapReduce. Считается одной из основополагающих технологий Big Data.

На рис. 37 показаны инструменты распределенных вычислений для Big Data.

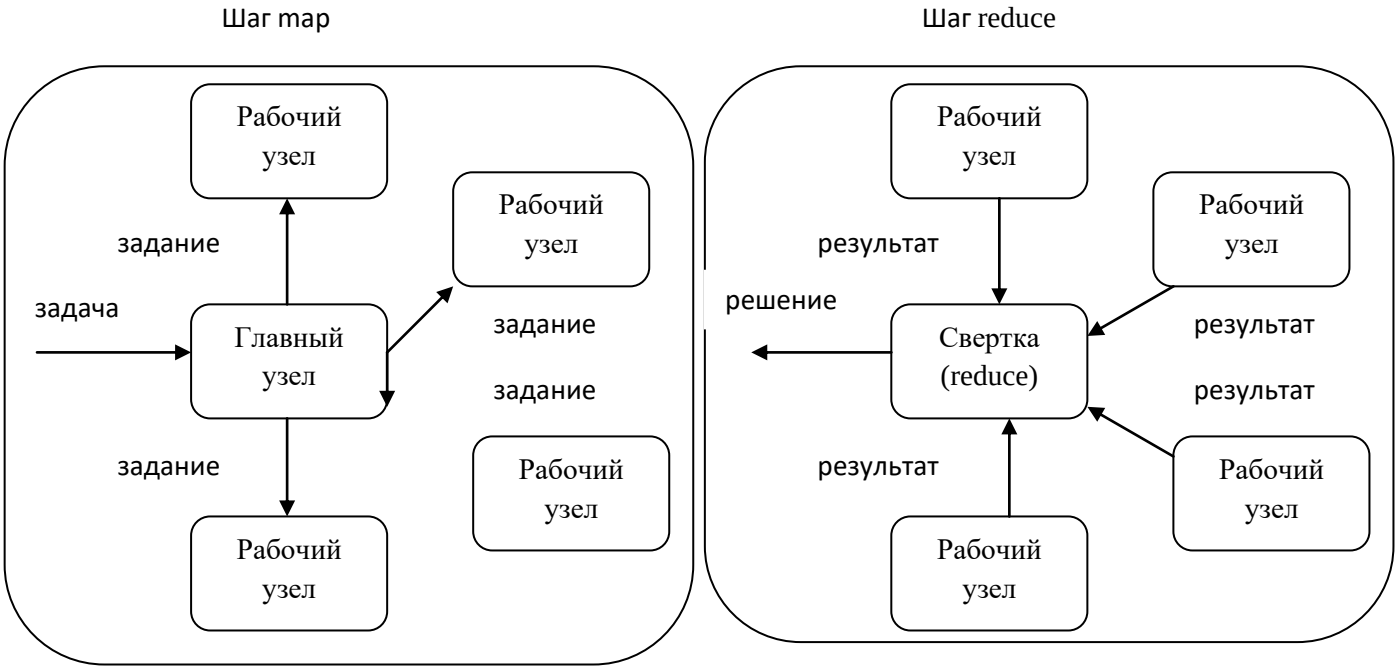


Рисунок 37. Инструменты распределенных вычислений для Big Data

NoSQL – группа подходов, которые для хранения и обработки данных используют параллельные распределенные системы интернет-приложений (например, поисковые системы), но при этом отказываются от традиционных реляционных систем управления базами данных с доступом к данным с помощью языка SQL. Отсюда и термин – NoSQL.

Роль и место Big Data в анализе данных

Технологии Knowledge Discovery и Data Mining решают задачи поддержки принятия решений на основе обнаруженных зависимостей и закономерностей в данные описывающих бизнес-процессы компании. При этом предполагается, что чем больше данных будет задействовано, тем лучше будут полученные решения. Именно поэтому появление Больших данных очень быстро привело к появлению Большой аналитики или аналитики Больших данных.

Для создания моделей Data Mining необходимы структурированные данные, и далеко не всегда огромное число обучающих примеров, которое способно предоставить Big Data, способствует улучшению качества модели. Тогда возникает естественный вопрос: как соотносятся Big Data, оперирующие петабайтами данных неопределенной структуры и относительно небольшие наборы выборок для построения предсказательных моделей, которые должны быть жестко структурированы. Роль Big Data с точки зрения предсказательной

аналитики заключается в том, чтобы помочь «зачерпнуть» из стремительного и бурного потока данных образцы, анализ которых поможет описать закономерности всего потока с целью получения знаний о связанных с ним бизнес-процессах.

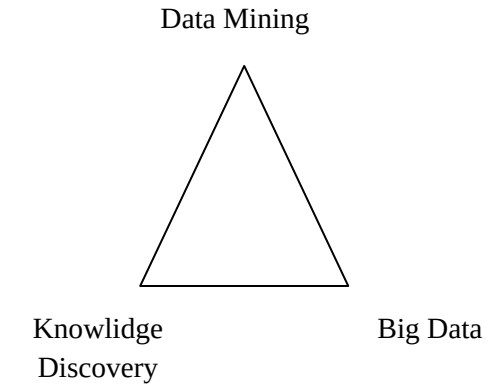


Рисунок 38. Место Big Data в анализе данных

Таким образом, задача Big Data – управление огромными потоками данных из различных распределенных источников (Интернета, мобильных приложений, биржевой информации, аудио и видеорегистраторов, данных и так далее), проведение их описательного анализа, и формирование наборов данных для построения моделей Data Mining.

Таким образом, Big Data правильнее рассматривать как технологию подготовки данных сверхбольшого, непрерывно возрастающего объема, расположенных в распределенных файловых системах и готовых к анализу методами Data Mining и бизнес-аналитики.

В этой связи возникает вопрос: существуют ли какие-либо особенности анализа данных уровня Big Data относительно обычных данных, то есть можно ли использовать к ним термин Большая аналитика?

Во-первых, при использовании технологий Big Data в распоряжении исследователя о называется намного больше данных, причем как структурированные, так и неструктурированных. Поэтому для анализа необходимо использовать приложения, «умеющие» работать не только с табличными данными.

Во-вторых, при работе с данными уровня бизнес-аналитики, исследователь в большинстве случаев имеет представление о характере, природе и происхождении используемых данных, что очень важно при интерпретации

результатов их исследования. В случае *Больших данных* такие представления, как правило, отсутствуют.

Data Science – краткая история понятия

Завершим наше обсуждение еще одним термином – наука о данных или Data Science. Несмотря на то, что впервые этот термин *наука о данных* прозвучал почти 50 лет назад, массово это понятие вошло в лексикон специалистов в области информационных технологий сравнительно недавно. Да и само понятие за это время значительно эволюционировало.

Бурное развитие информационных технологий привело к тому, что данные становились предметом все более пристального внимания исследователей в различных областях как потенциальный источник ценных знаний, использование которых в бизнесе обещает получение конкурентных преимуществ. Данные стали называть «новой нефтью». Параллельно развивалась и обогащалась *наука о данных.*

Впервые термин наука о данных был введен профессором Копенгагенского университета Питером Науром в 1966 году. В основе концепции П. Наура лежит представление о данные как о сырье; из которого могут быть сделаны те или иные полезные продукты для использования в других областях знаний и наук.

Следовательно, можно описать «жизненный цикл» данных с момента их появления и до момента практического внедрение разработанных на их основе, продуктов. В этом контексте, наука о данных представляет собой дисциплину, которая изучает этот жизненный цикл. Хотя П. Наура считают «пионером» в области науки о данных, многие исследователи рассматривают ее истоки в разведочном анализе Тьюки.

В 2001 году У. С. Кливленд опубликовал статью, в которой предложил рассматривать науку о данных, как самостоятельную дисциплину «в контексте информатики и интеллектуального анализа данных».

В частности он отметил, что «имеет место ограниченность знаний специалистов в области информационных технологий относительно подходов и методов организации поиска знаний, с другой стороны, статистики недостаточно хорошо знают информационные технологии.

Таким образом, по его мнению, «наука о данных должна связывать статистику и достижения в области компьютерной обработки данных».

В 2002 году при Международном совете по науке начал издаваться журнал Наука о данных**,** который рассматривал проблемы описания информационных систем и их приложений. В нем Data Science определили как «дисциплину, объединяющую в себе различные направления статистики, Data Mining, машинное обучение и применение баз данных для решения сложных задач, связанных с обработкой данных».

В 2005 году Национальный научный фонд США определил исследователей в области науки о данных как специалистов в области «информационных и компьютерных технологий, баз данных и программного обеспечения, а также программистов, экспертов по различным дисциплинам, библиотекарей, архивистов и других работников, которые участвуют в процессе сбора и обработки цифровых данных».

Значимыми вехами в развитии и становлении науки о данных стали работы под руководством Г. Пятецкого-Шапиро, в которых были заложены основы методик Knowledge Discovery и Data Mining. И, наконец, последним кирпичиком в фундамент науки о данных в современном ее понимании, являлось появление технологии Big Data.

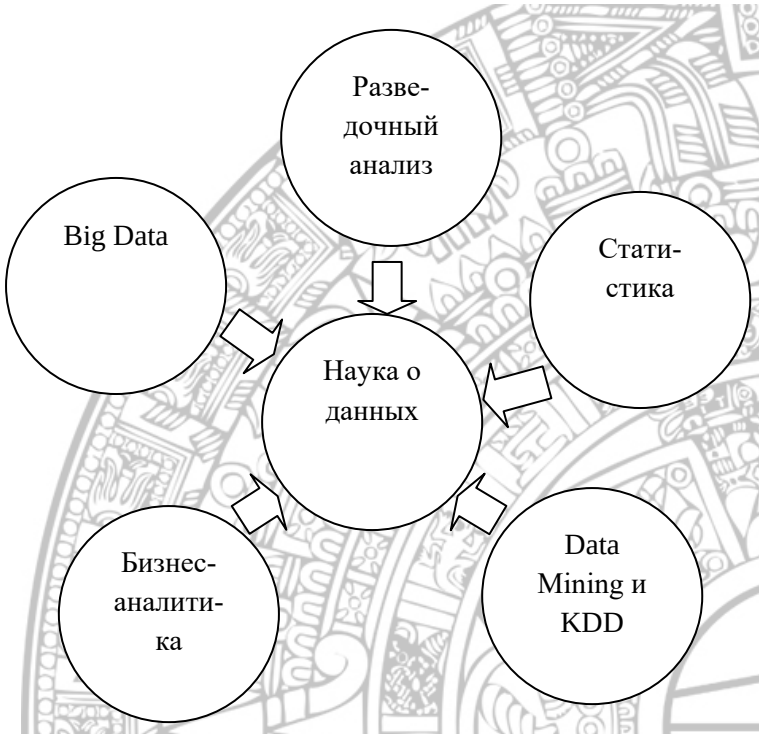


Рисунок 38. Общеупотребительное понятие науки о данных

Таким образом, в широком смысле, наука о данных решает практические задачи компьютерной обработки данных с целью получения полезных знаний. Зародившись как академическая дисциплина, наука о данных последовательно включала в себя как традиционные статистические подходы и разведочный анализ, так и технологии Knowledge Discovery, Data Mining и бизнес-аналитику. Важным направлением развития науки о данных в последние годы является обработка больших наборов данных Big Data.

Следует отметить, что общепринятым понятием науки о данных, как дисциплины, интегрирующей все направления использования данных для обнаружения в них полезных знаний (вместо набора различных терминов, употребляющихся ранее), стало примерно в 2010 году. Поэтому часто науку о данных считают наукой, которой еще нет, которая находится на стадии становления формирования.

Специалист по данным и бизнес-аналитике

Появился даже специальный термин для обозначения профессии специалиста по данным data scientist. Поэтому очень интересен вопрос: чем отличается бизнес-аналитик от data scientist.

Data scientist – понятие, которое подразумевает специалиста широкого профиля, занимающегося обработкой разных данных, различными способами и для различных целей, и включает в себя бизнес-аналитику, как частичный случай. То есть, скорее бизнес-аналитик является в какой-то степени специалистом по данным, но не наоборот.

Бизнес-аналитик – это специалист, который может про анализировать данные некоторой компании с целью принятия управленческих решений и который хорошо разбирается в прикладных задачах бизнес-аналитики. Например, ответить на вопрос, как снизить отток клиентов? Как увеличить выдачи кредитов при сохранении качества текущего кредитного портфеля?

Отличительным навыком бизнес-аналитик от *специалиста по данным* является наличие у первого развитых способностей анализировать бизнес-ситуацию качественно, интуитивно, с учетом трудно формализуемой информации, которую нельзя задать формулами. В тоже время *специалист по данным* владеет широким спектром навыков – умеет программировать, хорошо разбирается в нюансах алгоритмов машинного обучения.

Мы свели в таблицу основные характеристики бизнес-аналитика и data scientist`a, для лучшего понимая отличительных особенностей между ними. Хотя и те, и другие, опираются в целом на один и тот же фундамент в виде визуализации, статистики, машинного оборудования, Data Mining, аналитических пакетов и платформ.

Таблица 15 основные характеристики бизнес-аналитика и data scientist`a

Знания и профессиональные навыки	Специалист по данным	Бизнес-аналитик
Алгоритмы	Статистика и машинное обучение, методы оптимизации – требуется глубокое знание, позволяющее самостоятельно реализовывать как базовые алгоритмы, так и их модификации.	Требуется знание того, как работают основные алгоритмы Data Minig и как настройки алгоритмов влияют на их работу
Программирование и базы данных	Требуется знание основ алгоритмизации, популярных скриптовых языко в и платформ программирования, принципов параллельного программирования, реляционных баз данных и SQL,концепции MapReduce и Hadoop	Желательно знание SQL
Владение инструментами анализа	Статистические пакеты, библиотеки Data Mining, аналитические платформы, инструменты для визуализации, в том числе умение комбинировать решение из нескольких программных модулей	Использует преимущественно аналитические платформы с визуальными языками моделирования без необходимости программирования
Знание предметной бизнес- области	Требуется минимально, т.к. работа ведется с данными, не связанные прямо с экономическими показателями (например, профили соц. сетей, веб-логи)	Требуется
Изобретательность	Требуется	Требуется
Оценка экономического и финансового эффекта от внедрения моделей	Навык не обязателен	НАВЫК ОБЯЗАТЕЛЕН

Является ли Data Science наукой?

В настоящее время в среде ученых и специалистов ведутся активные дебаты – правильно ли называть Data Science наукой. Мнения на этот счет разнятся. Некоторые предлагают считать науку о данных частью статистики. Другие: считают, что *наука о данных* – просто красивое выражение, не имеющее реального содержания, но используемое вместо *бизнес-аналитики* в университетских курсах, как более «академичное». Третьи полностью идентифицируют науку о данных с *Big Data.* Специалисты в области статистики склонны больше видеть в Data Science науку, а специалисты в области компьютерных технологий-данные.

Слово **наука** подразумевает знания, полученные путем систематических исследований. В первом приближении, это совокупность систематических действий, с помощью которых из данных извлекаются знания в форме проверяемых выводов, заключений и предсказаний. Можно сказать, что статистика решает те же самые задачи. В чем же отличие Data Science от статистики?

Во-первых, статистические методы ориентированы на работу с качественными, структурированными данными. Если данные являются «сырыми», содержат пропуски, дубликаты, выбросы, то результаты статистического анализа могут оказаться смещенными. Поэтому необходима определенная предобработка и очистка данных.

Во-вторых, все больше возрастает объем и роль неструктурированных и слабо структурированных данных (видео, аудио, изображения, речь, текст, тэги и так далее), непосредственное применение статистических методов к которым невозможно. Еще одной проблемой является множественность и разнообразие источников данных, что требует их интеграции. Таким образом, применения в современных условиях для поиска знаний в данных только статистических методов недостаточно.

Что касается сравнения науки о данных с *Big Data,* то можно сказать, что для того, что бы получить ценные и полезные знания на основе небольшого объема данных необязательно привлекать *Большие данные.* Хорошему специалисту может оказаться достаточным для решения задачи нескольких сотен наблюдений, а другому и терабайт не поможет.

Таким образом, под общим названием Data Science в настоящее время существует множество различных, слабо систематизированных подходов,

методов и технологий для анализа данных различного объема с целью поиска знаний. Однако наукой это можно называть весьма условно, поскольку для науки должны быть разработаны строгие предметы и методы исследований, что применительно к Data Science, вообще говоря, в настоящее время отсутствует. Поэтому Data Science правильнее считать междисциплинарным направлением информационных технологий, включающих все аспекты работы с данными с целью извлечения из них полезных знаний.

Тема 16. Методы визуализации. OLAP-анализ

В данной лекции мы познакомимся с основными классами визуализаторов и изучим базовые, входящие в состав любого инструментального BI-средства.

Одной из важнейших составляющих бизнес-аналитики является визуализация – представление данных в виде, который обеспечивает наиболее эффективную работу пользователя. Способ визуализации должен максимально полно отражать поведение данных, содержащуюся в них информацию, тенденции, закономерности и так далее.

При этом выбор способа визуализации зависит от характера исследуемых данных и от задачи анализа, а также от предпочтений пользователя.

Многие связывают визуализацию только с интерпретацией, оценкой качества и достоверности результатов анализа. Однако это в корне неверно. Визуализацию необходимо применять на всех этапах анализа без исключения. На практике в процессе анализа данных пользователь непрерывно работает с различными визуализаторами.

Сегодня существует несколько сотен способов визуализации табличных данных. Эта область знаний интенсивно развивается.

Цели и задачи визуализации

Итак, визуализация используется на разных этапах аналитического процесса для достижения следующих целей и решения следующих задач:

- визуализация источников данных;
- визуализация загруженной выборки;
- визуализация результатов предобработки;
- визуализация промежуточных результатов;
- визуализация результатов анализа.

Рассмотрим каждую из задач подробнее.

Визуализация Источников данных

В источнике данных перед их загрузкой в аналитический контур корпоративной информационной фабрики аналитику требуется визуально оценить:

- характер, тип и поведение данных,
- динамический диапазон значений;

- степень гладкости;
- наличие факторов, снижающих качество данных, таких как шумы, аномальные и пропущенные значения.

Визуальный анализ источника данных позволяет:

- увидеть, соответствуют ли данные ожидаемым,
- оценить степень пригодности данных к анализу;
- выдвинуть гипотезы о закономерностях процессов, описываемых данными;
- определить, какие виды очистки и предобработки необходимо применить к данным

Кроме того, визуализация источников данных позволяет определить метод загрузки данных в аналитический контур и параметры, которые при этом должны быть использованы. Например, для корректной загрузки данных из текстового файла с разделителями необходимо правильно определить символ-разделитель, используемый формат даты и времени, расположение заголовков столбцов и так далее.

Для визуализации источников данных можно использовать приложения, в которых они были созданы (текстовые редакторы, СУБД, электронные таблицы и так далее). Кроме того, аналитические платформы содержат собственные средства предварительного просмотра источников данных.

Методы визуализации

В настоящее время в бизнес-аналитике и Data Sience используется несколько десятков основных методов визуализации. Выбор метода определяется особенностями и характером данных, спецификой решаемой задачи и, наконец, предпочтениями пользователя. Рассмотрим методы визуализации, приняв за основу следующую классификацию.

Табличные и графические. Как правило, таблицы применяются в том случае, когда пользователю необходимо работать с отдельными значениями данных, вносить изменения, контролировать форматы данных, пропуски, противоречия и так далее. Графические методы позволяют лучше увидеть общий характер данных – закономерности, тенденции, периодические изменения. Кроме того, графические методы более эффективно сопоставляют данные: достаточно

построить графики двух исследуемых процессов на одной системе координат чтобы оценить степень их сходства и различия.

При изучении различных видов визуализации удобнее рассматривать их не по отдельности, а в контексте задач, для которых они наиболее часто применяются. Можно выделить следующие группы методов визуализации:

- общего назначения – применяются для решения типовых задач анализа данных визуальной оценки качества и характера данных, распределения значений признаков, статистических характеристик и т. д. В них можно выделить дез подвида – простые и сложные. К последним, в частности, относится OLAP-анализ – комплекс методов для визуализации многомерных данных;
- оценка качества моделей – позволяет оценивать различные характеристики моделей, такие как точность, эффективность, достоверность результатов, интерпретируемость, устойчивость и так далее;
- интерпретация результатов анализа – служаг для представления конечных результатов анализа в виде, наиболее удобном с точки зрения их интерпретации пользователем

Подсистемы визуализации данных содержатся не только в специализированных аналитических платформах, но и практически во всех программных средствах, которые связаны с обработкой данных, – от офисных приложений до систем компьютерной математики. Однако в аналитических платформах визуализации данных уделяется особое внимание, поскольку она является одной из составляющих аналитического процесса, без которой невозможно эффективно решить поставленные задачи.

Наилучших результатов можно добиться, если считать визуализацию не отдельной подсистемой, а такой же частью аналитического процесса, как, например, подготовка данных, аудит и профайлинг моделирование. Даже если для построения качественной модели данных недостаточно, визуализация позволяет выдвигать гипотезы, делать выводы на основе экспертных оценок, разрабатывать способы повышения информативности данных

Визуализаторы общего назначения

Можно выделить набор средств визуализации, которые очень часто используются в бизнес-аналитике. Такие средства визуализации называются визуализаторами общего назначения. Начнем с простых визуализаторов. К ним относятся:

- таблицы;
- график;
- диаграммы;
- гистограммы.

Сложные визуализаторы общего назначения

Их сложность, в первую очередь, связана с пониманием этих визуализаторов: диаграммы и гистограммы привычны большинству пользователей, чего не скажешь о визуализаторах, получивших широкое распространение в последнее десятилетие. Обычно это многомерные визуализаторы, позволяющие представить информацию из нескольких измерений за счет использования различных цветов, форм, размеров и расположения объектов анализа:

- OLAP-анализ;
- географические и тепловые карты;
- дерево-карта или плоское дерево;
- диаграмма связей;
- облако данных;
- площадная диаграмма;
- график рассеивания.

OLAP-анализ

Большинство реальных бизнес-процессов являются сложными, поскольку в них участвует много объектов, которые находятся в самых разнообразных отношениях и с каждым из которых может быть связано несколько числовых характеристик. Поэтому при визуализации данных часто встает вопрос: как представить сложные данные в таком виде, чтобы человек мог их осмыслить и интерпретировать.

Например, если исследуемым процессом являются продажи, то приходится иметь дело с наименованиями товаров и групп товаров, с городами, в которых они продавались, с датами продаж, с информацией о поставщиках, покупателях, местах хранения и так далее. На первый взгляд, данная проблема стоит не так уж остро. Если необходимо отслеживать несколько объектов (продажу нескольких товаров), то достаточно построить соответствующее количество линий на трафике для каждого товара или таблицу.

Но товары могут различаться по цене и объемам продаж в сотни и даже тысячи раз, что вызовет проблему при построении нескольких графиков в одной системе координат. Кроме того, данные могут иметь различную степень детализации, что также затрудняет их табличное или графическое представление. И наконец, ситуацию заводит в тупик огромное количество записей, которые накапливаются в корпоративных базах и хранилищах данных.

Все эти трудности, возникающие при обработке больших и сложных массивов данных, создали предпосылки для появления метода, визуализации многомерных табличных данных – OLAP-анализа.

В основе OLAP лежит многомерное представление данных (принцип многомерного хранения данных подробно рассматривается в разделе Основы хранилищ и витрин данных), которые могут быть разделены на количественные и качественные. Качественные данные представляют собой значения, выраженные в категориальной форме. Обычно это наименования товаров, групп товаров, организаций, названия городов, ФИО сотрудников и так далее. С каждым объектом связаны признаки, количественно описывающие его. Для товара это может быть пена, количество или сумма; для города, в котором расположено торговое представительство, – расстояние до него и количество жителей; для сотрудника – заработная плата и стаж работы.

С каждым качественным значением анализируемого бизнес процесса связаны один или несколько количественных показателей. В рамках многомерной модели данные, качественно описывающие исследуемый бизнес- процесс, называются измерениями. Измерениями могут быть Товар, Город, Клиент, Организация др. К ним относят и дату Данные, количественно описывающие процесс или объект называются фактами (пли показателями), Примеры фактов: Количество, Сумма, Возраст, Доход, Торговая наценка и др. Другими словами, с каждым измерением связаны один или несколько фактов.

Измерения несут смысловую нагрузку; а факты – количественную. Чтобы достоверно отделить измерения от фактов, достаточно сопоставить значение с вопросом Измерения позволяют ответить на вопросы Что? (товар), Кто? (клиент), Когда? (дата) и Где? (город). Факты отвечают на единственный вопрос – Сколько?

Чтобы построить OLAP-куб, пользователь должен указать системе следующие параметры:

- какие измерения и факты включать в куб;
- методы агрегации значений фактов.

Визуализация OLAP-куба производится с помощью специального вида таблиц, которые строятся на основе срезов OLAP-куба, содержащих необходимую пользователю информацию. Срезы, в свою очередь, являются результатом выполнения соответствующего запроса к источнику данных.

Как правило, в процессе построения срезов пользователь с помощью мыши и клавиатуры манипулирует заголовками измерений, добиваясь наиболее информативного представления данных в кубе. В зависимости от положения заголовков измерений в таблице автоматически формируется запрос к базе или хранилищу данных. Запрос извлекает данные из базы или хранилища, после чего OLAP-ядро системы визуализирует их.

Таким образом, OLAP-куб можно использовать не только как метод визуализации, но и как средство оперативного формирования отчетов и представления информации в нужном разрезе (так называемая аналитическая отчетность).

Наибольший интерес OLAP-куб представляет с точки зрения визуального анализа данных, поиска особенностей и закономерностей в данных. При этом существуют два подхода:

- Аналитик может задаться несколькими вопросами: Какой товар самый популярный, Где продажи были наилучшими и с чем это связано и т. д. Затем с помощью манипуляций заголовками измерений выбирается такое представление куба, которое позволяет ответить на поставленные вопросы.
- Можно подойти к проблеме и с другой стороны – последовательно перебирать возможные варианты представления данных, которые обеспечивает куб, и с их помощью сопоставлять данные, выявлять закономерности и связи между элементами данных, выдвигать гипотезы и так далее.

Умелое использование OLAP-анализа и работа с кубом порой позволяют получить результаты даже в тех случаях, когда методы Data Mining оказываются малоэффективными (например, из-за недостатка или низкого качества данных).

Манипуляции с измерениями

Рассмотрим еще три операции с измерениями.

Изменение порядка следования измерений. При работе с несколькими измерениями в кубе закладывается возможность выбрать порядок их отображения.

Отбор значений измерений. Часто необходимость в отображении всех возможных значений измерения (например, дат, товаров и т. д.) отсутствует. Поэтому, чтобы не загромождать куб ненужными данными, лишние значения измерений могут быть временно скрыты, а если понадобится — отображены снова. При этом выбирать отображаемые значения измерений можно или непосредственного из списка или с помощью фильтрации по какому-либо условию.

Транспонирование. Если при работе с кубом выяснится, что значения измерений, которые отображаются в столбцах, удобнее отображать в строках или наоборот, то соответствующее преобразование можно легко произвести с помощью операции транспонирования.

При работе с OLAP-кубами широко применяется еще одна операция, называемая **детализацией** (англ: drill down – проникновение, более детальное исследование). Ее необходимость вызвана тем, что в большинстве случаев значения в кубе являются агрегированными, например, в пределах некоторой даты или интервала дат. В то же время пользователя могут интересовать и атомарные (то есть детализированные) значения, на основе которых были получены агрегированные.

Операция детализации заключается в отображении набора записей выборки данных, в результате агрегирования которых было получено соответствующее значение OLAP-куба.

Географические карты

Карты позволяют наглядно представить данные, связанные с географическим расположением исследуемых объектов и процессов. Это могут быть демографические данные (например, распределение показателей смертности или рождаемости по различным региона), а также данные, отражающие миграцию населения и рабочей силы, обеспечение регионов энергоносителями, динамику распространения эпидемий и т. д.

В бизнес-аналитике это информация, связанная с уровнем потребления в регионах, характером спроса и предложения по различным видам товаров, сведения о продажах, осуществляемых региональными дилерами, о логистических и транспортных потоках и т.д.

Тепловые карты

Впрочем, карты не обязательно должны быть связаны с географией. В бизнес-аналитике почти всегда приходится иметь дело с объектами, которые описываются двумя признаками и более. То есть выборки, образованные такими объектами, являются многомерными. В этом случае может возникнуть проблема с визуализацией результатов, поскольку представление многомерных объектов на плоских визуализаторах (графиках, диаграммах) не всегда удобно и корректно отображает результаты.

Поэтому актуальны визуализаторы, позволяющие адекватно представлять многомерные данные. В частности, распространение получили двумерные тепловые карты (англ. heat maps), где каждому значению признака соответствует один из оттенков в заранее выбранной цветовой гамме.

Карта-дерево

К семейству карт можно отнести и метод визуализации, получивший название дерево-карта или плоское дерево (англ.: treemap). Он является очень эффективным при изображении численных атрибутов элементов (размер, стоимость, значение), организованных в большие иерархии. Базовая идея метода состоит в том, чтобы изобразить дерево, каждая вершина которого имеет численный атрибут в виде прямоугольника (или некоторой другой геометрической фигуры) таким образом, чтобы площади изображений вершин дерева были пропорциональны их значениям атрибута.

Диаграмма связей

Нередко требуется исследовать характер и степень взаимной зависимости между различными объектами. Для анализа можно использовать визуализацию связей, когда объекты представляются в виде некоторых значков, а связи между ними – в виде линий, соединяющих соответствующие значки. При этом сила связи, то есть степень взаимной зависимости объектов, может показываться различными способами. Чаще всего для этого используют:

- толщину линии: чем сильнее связь между объектами, тем толще соединяющая их линия;
- цвет линии, при этом выбираются оттенки определенного спектра.

В принципе, допускается одновременно использовать и толщину, и цвет, хотя такие диаграммы более сложны для восприятия. Кроме того, можно по-разному располагать объекты на плоскости.

Облако данных

Облако данных (англ: data Cloud) – это данные, в которых используется другой цвет и/или размер шрифта для обозначения числовых данных. Родителем этого визуализатора можно считать облако тегов, в котором сравниваются ключевые слова или фразы значения), содержащиеся внутри фрагмента текста (набора данных), задавая каждому из них свой размер шрифта.

Площадная диаграмма

Площадная диаграмма (англ: Bubble Chart) – это смесь графика и диаграммы, когда по двум осям расставлен набор точек, соответствующий значениям. При этом сами точки не соединены и имеют различную величину или цвет, которые задаются дополнительными показателями.

График рассеивания

График рассеивания (англ: scatterplot) показывает распределение ограниченного набора точек, соответствующих значениям по осям.

Список использованной литературы

1. Бариленко В.И. Основы бизнес-анализа: учебное пособие / под ред. В. И. Бариленко. – М.: КНОРУС, 2013. – 234 с.
2. Бергер А.Б. MS SQL Server 2005 Analysis Services. OLAP и многомерный анализ данных / А.Б. Бергер. – СПб.: BHV, 2017. – 928 с.
3. Бизунок В.К. Горчинская О.Ю., Ладыженский Г.М. Системы поддержки принятия решений для банков . http:// www . olap . ru /
4. Барсегян А.А. Методы и модели анализа данных: OLAP и Data Mining / А.А. Барсегян М.С. Куприянов, В.В. Степаненко, и др.. – М.: СПб: БХВ, 2013. – 336 с.
5. Годин А. М. Статистика: учебник / А. М. Годин. – Москва: Дашков и К, 2016. – 451 с.
6. Зинченко А. П. Статистика: учебник / А. П. Зинченко. – Москва: Колосс, 2016. – 566 с.
7. Крянев А.В. Метрический анализ и обработка данных / А.В. Крянев, Г.В. Лукин, Д.К. Удумян. – М.: Физматлит, 2017. – 308 с.
8. Лесковец Ю. Анализ больших наборов данных / Ю. Лесковец, А. Раджараман. – М.: ДМК, 2016. – 498 с.
9. Миркин Б.Г. Введение в анализ данных: Учебник и практикум / Б.Г. Миркин. – Люберцы: Юрайт, 2016. – 174 с.
10. Мхитарян В.С. Анализ данных: учебник для академического бакалавриата / под ред. В.С. Мхитаряна. – М.: Изд. Юрайт, 2017 – 490 с.
11. Ниворожкина Л.И. Статистические методы анализа данных: Учебник / Л.И. Ниворожкина, С.В. Арженовский, А.А. Рудяга. – М.: Риор, 2018. – 320 с.
12. Паклин Н.Б. Бизнес-аналитика: от данных к знаниям: учебное пособие [для вузов] / Н. Б. Паклин, В. И. Орешков. – 2-е изд., – Санкт-Петербург [и др.]: Питер, 2013. – 701 с.
13. Тюрин Ю.Н. Анализ данных на компьютере / Ю.Н. Тюрин, А.А. Макаров. – М.: МЦНМО, 2016. – 368 с.
14. Тьюки Дж. Анализ результатов наблюдений. Разведочный анализ. – М.: Мир, 1981. – 696 с.
15. Федоров А., Елманова Н. Введение в OLAP . Компьютер Пресс № 4 – 12, 2001

16. Хрусталёв Е.М. Агрегация данных в OLAP-кубах. http :// www . olap . ru /

17. Чубукова И.А. Data Mining: Учебное пособие / И.А. Чубукова. 2-е изд., М.: Университет Информационных технологий; БИНОМ. Лаборатория знаний, 2008. 382 с.

18. Щавелёв Л.В. Оперативная аналитическая обработка данных: концепции и технологии. http :// www . olap . ru /

19. Курс: Анализ данных [Электронный ресурс]:, 2013. – Электрон. текстовые дан. on-line. – Загл. с титул. экрана. – URL: ttp://do.ssau.ru/moodle/course/view.php?id=442 (Дата обращения 23.12.2015)

20. Coursera – бесплатные онлайн-курсы от ведущих университетов мира | Coursera [Электронный ресурс]: [б. и.], 2019. – Электрон. текстовые дан. on-line. – Загл. с титул. экрана. – URL: https://www.coursera.org (Дата обращения 23.09.2019).

21. Справка о подробных системных требованиях для работы с платформой Loginom. – URL: https://loginom.ru/system-requirements (Дата обращения 23.09.2019).

22. Официальный сайт компании. – URL: https://loginom.ru (Дата обращения 23.09.2019).

Приложение. Аналитическая платформа Loginom в примерах и задачах

Инструменты для бизнес-аналитики

До появления аналитических платформ анализ данных осуществлялся в основном в статистических пакетах. Их использование требовало высокой квалификации пользователя. Большинство алгоритмов, реализованных в статистических пакетах, не позволяло эффективно обрабатывать большие объемы информации. Для автоматизации рутинных операций приходилось использовать встроенные языки программирования.

В конце 80-х гг. произошел стремительный рост объема информации, накаливаемый на машинных носителях, и возросли потребности бизнеса по применению анализа данных. Ответом этому стало появление новых парадигм в анализе: хранилища данных, машинное обучение, Data Mining, Knowledge Discovery in Databases, Big Data, Deep Learning. Это позволило популяризировать анализ данных, вывести его на промышленную основу и решить огромное число бизнес-задач с большим экономическим эффектом.

Результатом развития анализа данных стали специализированные программные системы – аналитические платформы, которые полностью автоматизируют все этапы от получения интеграции данных до эксплуатации моделей и интерпретации результатов.

Сегодня – Loginom является ярким представителем как настольной, так и корпоративной системой анализа данных последнего поколения.

Loginom – это аналитическая платформа, основа для создания для законченных прикладных решений в области анализа данных. Реализованные в Loginom технологии позволяют на базе единой архитектуры пройти все этапы построения аналитической системы: от консолидации данных до построения и визуализации полученных результатов.

Loginom является новым поколением платформы, ранее называвшейся Deductor. Реализованная в Loginom архитектура позволяет добиться максимальной гибкости при создании законченного решения. Благодаря данной архитектуре можно собрать в одном аналитическом приложении все необходимые инструменты анализа и реализовать автоматическое выполнение подготовленных сценариев.

Платформа включает средства, позволяющие максимально сократить сроки разработки, быстро создавать и выводить на рынок новые прикладные решения, а также адаптировать их в соответствии с изменяющимися требованиями компаний. Создание законченного решения занимает минимум времени: достаточно получить данные, определить сценарий обработки и задать место для экспорта полученных результатов, при этом не требуется знание языков программирования.

В Loginom особое внимание уделено решению проблем повторного использования сценариев обработки данных. Для этого используется уникальная комбинация из структурного и объектно-ориентированного подходов к моделированию.

В процессе развертывания и использования аналитической платформы с ней взаимодействуют различные категории пользователей: аналитик, лицо, принимающее решение, администратор,

Задача 1. Используя данные роста 5 апельсиновых деревьев, посаженных 31.12.1968, выполнить следующие задания:

- Произвести отбор данных динамики каждого дерева в отдельный набор данных.
- Произвести выборки записей из исходной таблицы по возрасту: 664, 1004, 1372, 1582. В выборках 1 и 3 произвести сортировку по возрастанию величины обхвата дерева, а в выборках 2 и 4 – по убыванию обхвата дерева соответственно.
- Найти средние значения обхвата 5 деревьев по возрастам.
- Используя возможности визуализации отобразить на совместном графике динамику размеров деревьев. По графику определить самое большое дерево по итогам наблюдений.

Источники:

1. Draper, N. R. and Smith, H. (1998), Applied Regression Analysis (3rd ed), Wiley (exercise 24.N).

2. Pinheiro, J. C. and Bates, D. M. (2000) Mixed-effects Models in S and S-PLUS, Springer.

Таблица 16 Данные роста апельсиновых деревьев, посаженных 31.12.1968

№	Номер дерева	Возраст, дни	Обхват, мм	№	Номер дерева	Возраст, дни	Обхват, мм
1.	1	118	30	19.	3	1231	115
2.	1	484	58	20.	3	1372	139
3.	1	664	87	21.	3	1582	140
4.	1	1004	115	22.	4	118	32
5.	1	1231	120	23.	4	484	62
6.	1	1372	142	24.	4	664	112
7.	1	1582	145	25.	4	1004	167
8.	2	118	33	26.	4	1231	179
9.	2	484	69	27.	4	1372	209
10.	2	664	111	28.	4	1582	214
№	Номер дерева	Возраст, дни	Обхват, мм	№	Номер дерева	Возраст, дни	Обхват, мм
11.	2	1004	156	29.	5	118	30
12.	2	1231	172	30.	5	484	49
13.	2	1372	203	31.	5	664	81
14.	2	1582	203	32.	5	1004	125
15.	3	118	30	33.	5	1231	142
16.	3	484	51	34.	5	1372	174
17.	3	664	75	35.	5	1582	177
18.	3	1004	108				

Задача 2. Провинции Швейцарии

Стандартизованный показатель рождаемости и социально-экономические показатели для каждой из 47 франкоязычных провинций Швейцарии примерно в 1888 году.

СКР – Суммарный коэффициент рождаемости – является наиболее точным показателем уровня рождаемости, данный коэффициент характеризует среднее число рождений у одной женщины в гипотетическом поколении за всю её жизнь при сохранении существующих уровней рождаемости в каждом возрасте независимо от смертности и от изменений возрастного состава.

Сельское хозяйство – % мужчин, занятых в сельском хозяйстве.

Экзамен – % призывников, получивших высшую оценку на экзамене по армии

Образование – % образования за пределами начальной школы для призывников.

Католик – % католиков.

Младенческая Смертность – живорожденных, которые живут менее 1 года.

Задание:

1. Провести визуальный анализ выборки. Построить гистограммы распределения характеристик, попарные графики зависимостей. Определить, исходя из графиков, наиболее очевидные зависимости и несостоятельные взаимоопределения.

2. Провести корреляционный анализ переменных, определить наиболее и наименее коррелирующие характеристики, для каждого их четырех возможных критериев. На основе данных корреляционного анализа построить кубы значений критериев автокорреляции.

3. Провести факторный анализ переменных, используя ограничение на количество факторов равное 2. Затем используя диаграмму, расположить значения полученных факторов на графике разброса. Оценить получившиеся группы.

4. Откватовать исходные данные плиточным способом по одному из предложенных полей. Сгруппировать данные по выходным квантам, рассчитав средние от остальных характеристик.

Данные доступны по ссылке: https://opr.princeton.edu/archive/pefp/switz.aspx

Источники:

1. Project «16P5», pages 549-551 in

2. Mosteller, F. and Tukey, J. W. (1977) Data Analysis and Regression: A Second Course in Statistics. Addison-Wesley, Reading Mass.

3. indicating their source as Data used by permission of Franice van de Walle. Office of Population Research, Princeton University, 1976. Unpublished data assembled under NICHD contract number No 1-HD-O-2077.

4. Becker, R. A., Chambers, J. M. and Wilks, A. R. (1988) The New S Language. Wadsworth & Brooks/Cole.

Задача 3. Исследование болезни сердца пациентов

Данные о пациентах с болезнью сердца

Данные имеют достаточно большой вид для печати, поэтому приводиться в виде таблицы не будут.

Информация о наборе данных:

База данных содержит 76 атрибутов, но все опубликованные эксперименты относятся к подмножеству 14 из них. В частности, база данных Кливленда принадлежит исследователям ML на сегодняшний день. Поле «target» относится к наличию болезни сердца у пациента. Это целочисленное значение от 0 (нет присутствия) до 4-х. Эксперименты с базой данных Кливленда были сосредоточены на простой попытке отличить присутствие (значения 1,2,3,4) от отсутствия (значение 0).

Имена и номера социального страхования пациентов были недавно удалены из базы данных, заменены фиктивными значениями.

Задание:

1. При помощи инструмента «Визуализация» просмотреть статистику по данным наблюдений в выборке. Сделать начальные выводы об информативности переменных наблюдений.

2. Провести кластеризацию на основе инструментов: «Кластеризация», «EM кластеризация», «Самоорганизующиеся сети». Сгруппировать данные по кластерам. Оценить средние характеристики количественных переменных. Сделать качественные выводы о механизме разбиения на группы.

3. Провести классификацию на основе инструментов: «Логистическая регрессия» и «Классификация (нейросети)». Сравнить процент ошибок на

обучающей и тестовой выборках для двух моделей. В качестве маркера или выходной переменной использовать переменную target.

Используются только 14 атрибутов:

1. # 3 (age) возраст в годах
2. # 4 (sex) (1 = мужчина; 0 = женщина)
3. # 9 (cp) тип боли в груди
4. # 10 (trestbps) артериальное давление в покое (в мм рт. ст. при поступлении в больницу)
5. # 12 (chol) уровень холестерина в сыворотке в mg/dl
6. # 16 (fbs) (уровень сахара в крови натощак> 120 mg/dl) (1 = верно, 0 = неверно)
7. # 19 (restecg) результаты электрокардиографии в покое
8. # 32 (thalach) достигнутая максимальная частота сердечных сокращений
9. # 38 (exang) стенокардия, вызванная физической нагрузкой (1 = да, 0 = нет)
10. # 40 (oldpeak) Депрессия ST, вызванная физическими упражнениями относительно отдыха
11. # 41 (slope) наклон пика упражнений сегмента ST
12. # 44 (ca) количество крупных сосудов (0-3), окрашенных по цвету
13. # 51 (thal) 3 = нормально; 6 = исправленный дефект; 7 = обратимый дефект
14. # 58 (num) (прогнозируемый атрибут)

Полная документация по тегу:

3 age: возраст в годах

4 sex: пол (1 = мужчина; 0 = женщина)

9 cp: тип боли в груди

- значение 1: типичная стенокардия
- значение 2: атипичная стенокардия
- значение 3: неангинальная боль
- значение 4: бессимптомное

10 trestbps: артериальное давление в покое (в мм рт.ст. при поступлении в больницу)

12 chol: сыворотка холесторальная в mg/dl

16 fbs: (уровень сахара в крови натощак> 120 mg/dl) (1 = правда, 0 = ложь)

19 restecg: результаты электрокардиографии в состоянии покоя

- значение 0: нормальное ;
- значение 1: с аномалией волны ST-T (инверсия зубца T и / или повышение или депрессия ST> 0,05 мВ).
- значение 2: показывает вероятную или определённую гипертрофию левого желудочка по критериям Эстеса ;

32 thalach: максимальная достигнутая частота сердечных сокращений

38 exang: стенокардия, вызванная физической нагрузкой (1 = да; 0 = нет)

40 oldpeak: депрессия ST, вызванная физической нагрузкой относительно отдыха

41 slope: наклон пикового упражнения ST сегмента

- Значение 1: восходящий
- Значение 2: плоский
- Значение 3: наклон вниз

44 ca: количество крупных сосудов (0-3), окрашенных при флюороскопии

51 thal:

3 = нормальное;

6 = исправленный дефект;

7 = обратимый дефект.

58 num: ангиографический статус заболевания

- значение 0: <50% сужение диаметра
- значение 1:> 50% сужение диаметра

Данные доступны по ссылке:

https://www.kaggle.com/ronitf/heart-disease-uci

Задача 4. Прогнозирование поведения клиентов банка

Имея данные о клиенте банка, можем ли мы построить классификатор, который может определить, уйдут они или нет?

Поля:

RowNumber

CustomerId

Surname

CreditScore

Geography – Страна

Gender – Пол

Age – Возраст

Tenure – Собственность

Balance – Баланс счета

NumOfProducts – сколько счетов, банковских счетов, связанных продуктов у клиента.

HasCrCard – Имеет ли клиент кредитную карту.

IsActiveMember – Является ли клиент активным пользователем? (Субъективно, но для концепции).

EstimatedSalary – Расчетная заработная плата клиента.

Exited – Клиенты все-таки покинули банк?

Задание:

1. Просмотреть статистику выборки по каждой переменной.
2. Оценить распределение целевой переменной.
3. Провести классификацию.
4. Провести кластеризацию.
5. Оценить результат работы самоорганизующихся сетей.

Данные доступны по ссылке:

https://www.kaggle.com/barelydedicated/bank-customer-churn-modeling

Задача 5. Неподготовленные данные, мировая динамика.

Задача с общей постановкой на фильтрацию, модификацию, визуализацию и анализ данных.

Данные представлены тремя наборами из наблюдений по каждой стране данных по каждому из годов с 1960-2016 по трем критериям:

Популяция.

Средний коэффициент рождаемости (CKP).

Ожидаемая продолжительность жизни.

Данные разбиты на 3 набора и состоят из колонок, то есть данные находятся в слабоструктурированном виде, следовательно, необходимо провести некую предобработку этих данных.

Первоочередные задачи:

1. Структурировать данные, привести их к виду, приемлемому для анализа (разобранный пример)
2. Визуализация данных, построение срезов для коэффициентов и динамики численности.

3. (Оптимально) провести кластерный анализ, разбить страны на отдельные группы, приоритетно по численности, затем по СКР и качеству жизни. На основе данных кластеров, визуализировать данные временных рядов для каждого из кластеров в отдельности, и провести качественный анализ внутри групп.

Данные доступны по ссылке:

https://www.kaggle.com/gemartin/world-bank-data-1960-to-2016