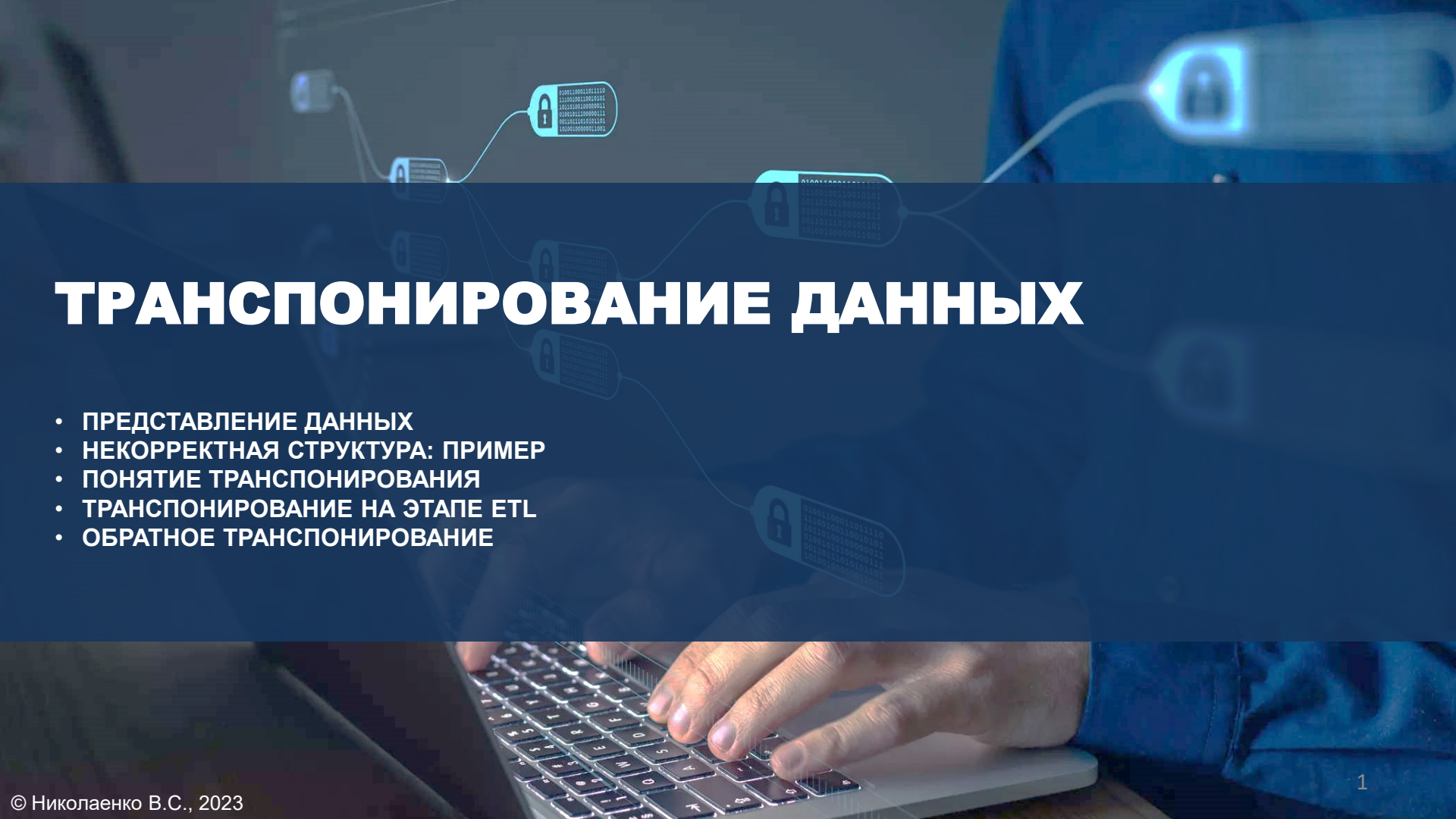


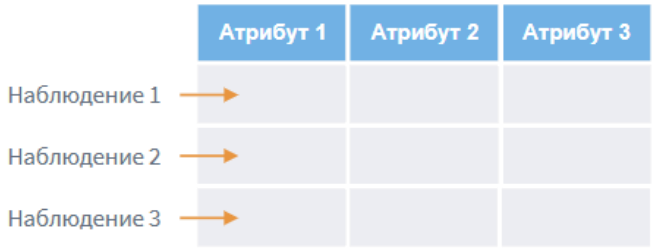
ТРАНСПОНИРОВАНИЕ ДАННЫХ

- ПРЕДСТАВЛЕНИЕ ДАННЫХ
- НЕКОРРЕКТНАЯ СТРУКТУРА: ПРИМЕР
- ПОНЯТИЕ ТРАНСПОНИРОВАНИЯ
- ТРАНСПОНИРОВАНИЕ НА ЭТАПЕ ETL
- ОБРАТНОЕ ТРАНСПОНИРОВАНИЕ

ПРЕДСТАВЛЕНИЕ ДАННЫХ

Большинство алгоритмов аналитики данных могут применяться только к структурированным данным, то есть данным, представленным в виде таблиц, в которых каждый столбец представляет собой некоторый атрибут или признак, а каждая запись – наблюдение, описывающее состояние анализируемого объекта или процесса.

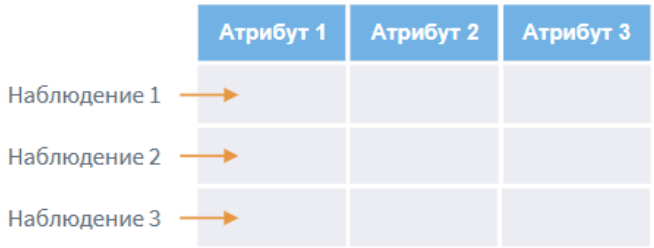
Однако даже если данные являются структурированными, это еще не гарантирует, что структура таблицы соответствует требованиям той или иной задачи анализа.



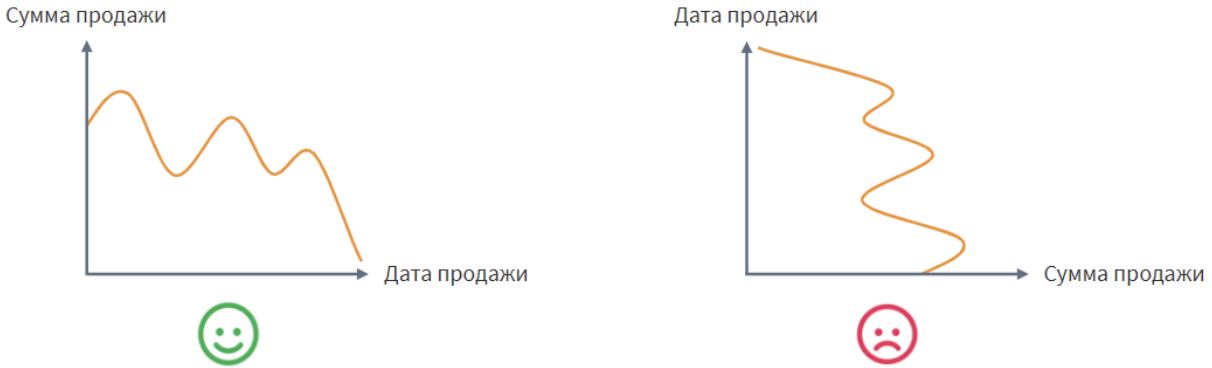
ПРЕДСТАВЛЕНИЕ ДАННЫХ

Действительно, для одного и того же набора данных можно построить множество табличных представлений. При этом некоторые представления способны привести к частичной и даже полной утрате смысла данных.

В лучшем случае это может привести к невозможности построить на основе данных модель или применить к ним готовую модель. В худшем можно получить некорректные результаты анализа, сделать по ним неправильные выводы и заключения.



ПРЕДСТАВЛЕНИЕ ДАННЫХ



Иллюстрацией к такой ситуации может служить построение графика. Для получения корректного графического представления данных значения независимой переменной (например, **Дата продажи**) откладываются по горизонтальной оси, а зависимой (например, **Сумма продажи**) – по вертикальной. Если сделать наоборот, то разобраться в таком графике будет очень непросто.

ПРЕДСТАВЛЕНИЕ ДАННЫХ

Ситуации, когда структура таблицы, в которой содержатся анализируемые данные, не соответствует решаемой задаче, чаще всего возникают тогда, когда данные поступают с рабочих мест отдельных пользователей, которые строят таблицу, как умеют, или как им удобно.

ПРЕДСТАВЛЕНИЕ ДАННЫХ

Дата	01.03.17	02.03.17	03.03.17	04.03.17	05.03.17	06.03.17
Товар						
Количество						
Сумма						
Остаток						

Приведем пример. Пусть имеется плоская таблица, по которой требуется построить модель для прогнозирования временного ряда, которую пользователь заполнил в виде, представленном на слайде.

Использование такой таблицы в качестве источника данных для алгоритма прогнозирования крайне неудобно, поскольку даты, товары и количественные показатели продаж должны образовывать столбцы (поля), а наблюдения – строки (записи).

ПРЕДСТАВЛЕНИЕ ДАННЫХ

Дата	01.03.17	02.03.17	03.03.17	04.03.17	05.03.17	06.03.17
Товар						
Количество						
Сумма						
Остаток						

В данном же случае наблюдение – это информация о товаре, его количестве, сумме и остатке на определенную дату. При этом наблюдения образуют столбцы, что противоречит самой идее структуризации.

ПРЕДСТАВЛЕНИЕ ДАННЫХ

Дата	01.03.17	02.03.17	03.03.17	04.03.17	05.03.17	06.03.17
Товар						
Количество						
Сумма						
Остаток						

Кроме того, поля источника данных должны быть типизированы, то есть содержать данные только одного типа. В представленной таблице поля содержат значения различных типов: строковые – для наименований товаров, и числовые – для суммы, количества и остатков, то есть тип данных каждого столбца – переменный.

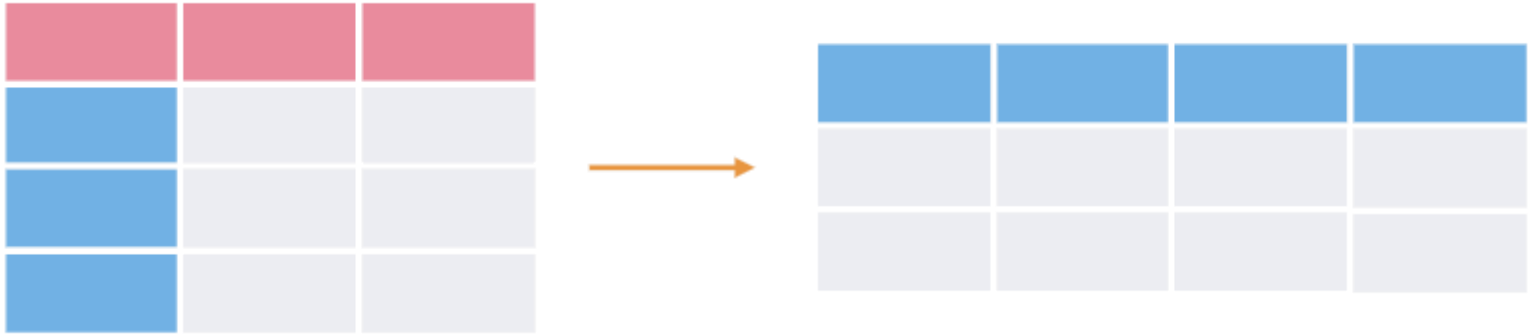
Такая ситуация в большинстве случаев приведет к ошибке несоответствия типов данных при попытке применить к данным из таблицы какую-либо обработку.

НЕКОРРЕКТНАЯ СТРУКТУРА: ПРИМЕР

Дата	Товар	Количество	Сумма	Остаток
01.03.2017				
02.03.2017				
03.03.2017				
04.03.2017				
05.03.2017				
06.03.2017				

Ситуацию можно легко исправить, если применить к таблице операцию, которая преобразует строки таблицы в столбцы, а столбцы – в строки. В результате мы получим те же данные, но другой структуры.

ПОНЯТИЕ ТРАНСПОНИРОВАНИЯ



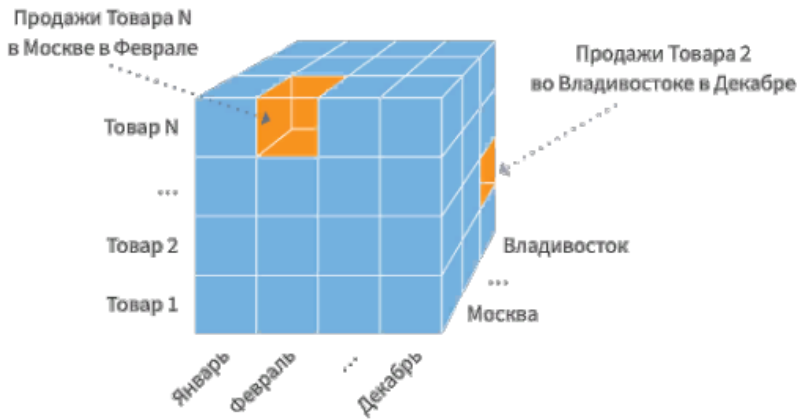
Транспонирование – это термин из теории матриц, который обозначает операцию, преобразующую столбцы матрицы в строки, а строки – в столбцы.

При работе с таблицами, содержащими анализируемые данные, этот термин имеет более широкий смысл. Такие таблицы могут иметь сложную структуру, десятки полей измерений и фактов. В результате возникает противоречие между «плоской» структурой таблицы и «многомерным» характером содержащихся в ней данных.

ПОНЯТИЕ ТРАНСПОНИРОВАНИЯ

Традиционно в бизнес-аналитике для выбора наиболее удобного представления многомерных данных используются OLAP-кубы, но OLAP-куб – это всего лишь средство визуализации многомерных данных.

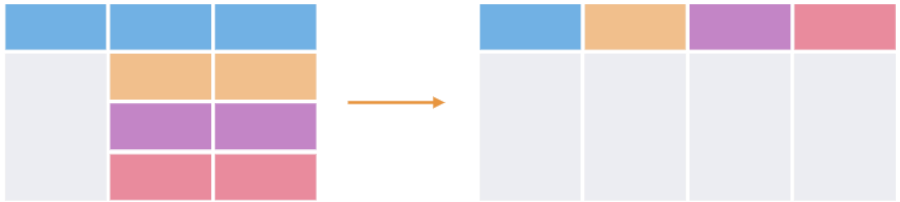
Манипулируя OLAP-кубом для выбора оптимального представления, мы не изменяем саму структуру источника данных. Можно сказать, что с помощью OLAP -куба пользователь управляет представлением данных, а с помощью транспонирования – структурой источника данных.



ТРАНСПОНИРОВАНИЕ НА ЭТАПЕ ETL

Транспонирование на этапе ETL используется для оптимизации структуры источника данных сточки зрения определенной задачи. С помощью транспонирования можно не только менять местами строки и столбцы таблицы, но и проводить более сложные манипуляции с ее структурой.

В качестве источников данных часто выступают файлы баз данных самых различных форматов, поэтому, получая источник данных, аналитик заранее не знает, в каком виде эти данные будут загружены в информационную систему компании.



ТРАНСПОНИРОВАНИЕ НА ЭТАПЕ ETL

Таблица 1

ФИО клиента	Атрибут	Значение атрибута
Иванов И.И.	Возраст	36
Иванов И.И.	Доход	18 000
Иванов И.И.	Стаж работы	15
Сидоров П.В.	Возраст	45
Сидоров П.В.	Доход	24 000
Сидоров П.В.	Стаж работы	21

Таблица 2

ФИО клиента	Возраст	Доход	Стаж работы
Иванов И.И.	36	18 000	15
Сидоров П.В.	45	24 000	21

Например, возможна ситуация, когда данные из первичного источника представлены структурой, как это демонстрирует таблица 1. Каждое наблюдение (клиент) оказывается размещенным в нескольких строках, а атрибуты не образуют поля.

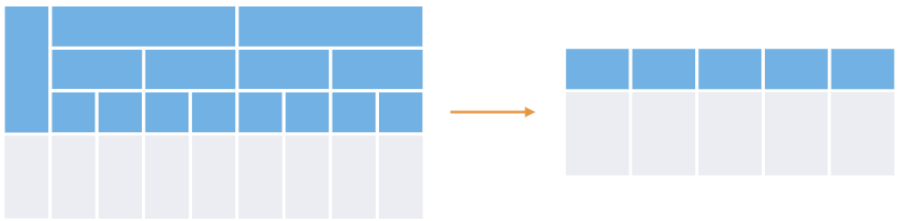
С помощью транспонирования можно «развернуть» таблицу 2, как показано на слайде.

ОБРАТНОЕ ТРАНСПОНИРОВАНИЕ

Еще одной проблемой, с которой часто приходится сталкиваться аналитикам при работе с источниками данных, являются нарушения в их структуре.

Если в витринах, хранилищах и базах данных регулярность структуры данных поддерживается автоматически, то в файлах отдельных пользователей структура таблиц не является жестко заданной.

Особенно характерна данная ситуация для файлов электронных таблиц, где в случае сложной, иерархической структуры данных появляются заголовки, общие для нескольких полей.



ОБРАТНОЕ ТРАНСПОНИРОВАНИЕ

Дата	Группа 1						Группа 2			
	Товар A		Товар B		Товар C		Товар D		Товар E	
	Кол-во	Сумма	Кол-во	Сумма	Кол-во	Сумма	Кол-во	Сумма	Кол-во	Сумма
01.03.17										
02.03.17										
03.03.17										

Типичный пример представлен на слайде, где в таблице присутствуют три измерения: **Дата**, **Товар** и **Группа товара**. При этом измерение **Товар** является иерархически подчиненным измерению **Группа товара**. Часто встречаются еще более сложные структуры, например, товары могут группироваться по городам, где они продавались, фирмам-поставщикам и клиентам и так далее.

Операция **обратного транспонирования** позволяет избавиться от структурных нарушений в таблице.

ОБРАТНОЕ ТРАНСПОНИРОВАНИЕ

В результате обратного транспонирования таблицы может быть получена структура, показанная на слайде.

В каждой записи новой таблицы содержится наблюдение по каждой дате, группе товара и отдельному товару, структура данных является полностью регулярной, а столбцы – типизированы. Такая таблица вполне отвечает требованиям, предъявляемым источникам данных, хотя и выглядит более громоздкой и избыточной.

Для выполнения операции транспонирования пользователь должен определить, какие именно измерения и факты исходной таблицы должны войти в транспонированную таблицу, а также, какие измерения должны отображаться в столбцах, а какие – в строках.

При этом для выполнения операции транспонирования должно быть выбрано хотя бы одно поле фактов, поскольку именно факты являются «связующим звеном» измерений.

Иными словами, какие бы изменения в таблице не происходили в результате ее транспонирования, факты должны быть жестко связаны со своими измерениями.

Дата	Группа товара	Товар	Кол-во	Сумма
01.03.17	Группа 1	Товар A		
01.03.17	Группа 1	Товар B		
01.03.17	Группа 1	Товар C		
01.03.17	Группа 2	Товар D		
01.03.17	Группа 2	Товар E		
02.03.17	Группа 1	Товар A		
02.03.17	Группа 1	Товар B		
02.03.17	Группа 1	Товар C		
02.03.17	Группа 2	Товар D		
02.03.17	Группа 2	Товар E		
03.03.17	Группа 1	Товар A		
03.03.17	Группа 1	Товар B		
03.03.17	Группа 1	Товар C		