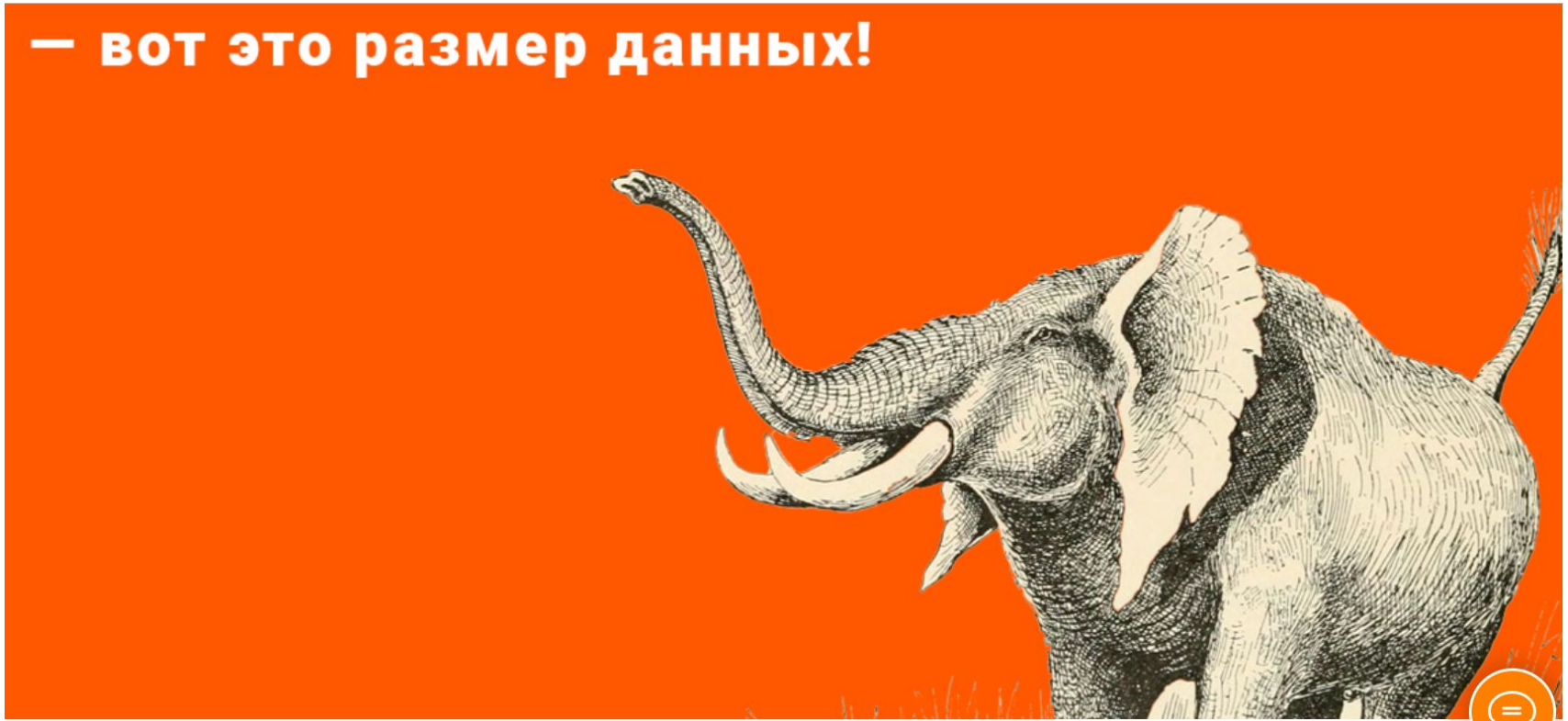
У термина ***Big Data*** есть точная дата рождения — 3 сентября 2008 года был выпущен номер журнала Nature, посвященный влиянию огромных массивов информации на развитие науки.



Предпосылки появления Big Data

1 ***Аналитика данных*** является основным инструментом поиска ***новых знаний*** в массивах данных, необходимых для принятия эффективных управленческих решений. ***Low-code*** становится мейнстримом.

2 ***Развитые средства*** хранения, доставки, интеграции данных позволяют увеличить объем данных, территориальную ***распределенность***, сложность.

3 Конфликт в терминологии. Обилие терминов и их трактовок.

4 Технологии ***Data Mining*** ориентированы на обработку структурированных данных. Но сегодня больший интерес представляют данные из социальных медиа, видео, электронной почты и других распределенных источников.

Источники Big Data

- Торговые сети

Торговые сети регистрируют миллионы клиентских транзакций, пересылают их в хранилища данных, объем которых составляет ***петабайты***.

- Мобильные устройства

Более 5 миллиардов людей по всему миру говорят, обмениваются сообщениями и производят поиск в Интернет с помощью мобильных устройств.

- Автоматические регистраторы

Тысячи автоматических регистраторов по всему миру непрерывно фиксируют погодные условия, и передают метеорологические данные в центры их обработки.

- Социальные сети

Пользователи социальных сетей ежеминутно отправляют десятки

Характеристики категории ***Big Data***:

1) ***Volume*** — ***объем*** данных должен превышать 150 Гб в сутки.

2) ***Velocity*** — ***скорость*** накопления и обработки данных: объем Big Data растет, поэтому для их обработки нужны специальные технологичные инструменты.

3) ***Variety*** — ***разнообразие*** типов данных: они могут быть структурированными, неструктурированными или частично структурированными.

4) ***Variability*** — ***изменчивость***. Потоки Big Data могут иметь свои пики и спады в зависимости от сезона, социальных явлений, изменений в политической ситуации и других факторов.

5) ***Veracity*** — ***достоверность*** и самого массива данных, и результатов аналитики.

6) ***Value*** — ***ценность***.

Пирамида аналитических решений

3) Системы визуального управленческого контроля показателей функционирования компании

2) Специализированные инструменты бизнес-аналитики, ориентированные на бизнес-аналитиков и профильных специалистов

1) Транзакционные учетные системы и хранилища данных — сложные и дорогостоящие технологии, являющиеся фундаментом ИТ-архитектуры любой современной крупной компании

Журнал ПЛАС. Технологии. А.Ю. Медников. Большие Данные и бизнес аналитика (plusworld.ru)

Термин «анализ данных»

Анализ данных – широкое понятие. В общем смысле – это процесс:

- исследования,
- преобразования и
- моделирования данных

с целью извлечения полезной информации и принятия решений.

Для анализа данных применяются различные математические методы.

- Моделирование – универсальный способ, позволяющий обнаружить зависимости, прогнозировать.
- Самое главное: полученные такими образом знания можно тиражировать.

Современное понятие анализа данных

Концепция «модели от данных» требует тщательной подготовки данных – качество данных

Современная бизнес-аналитика делит методы решения задач на две основные группы:

1. извлечение и визуализация данных;
2. построение и использование моделей.

Построение моделей – полученные таким образом знания можно тиражировать.

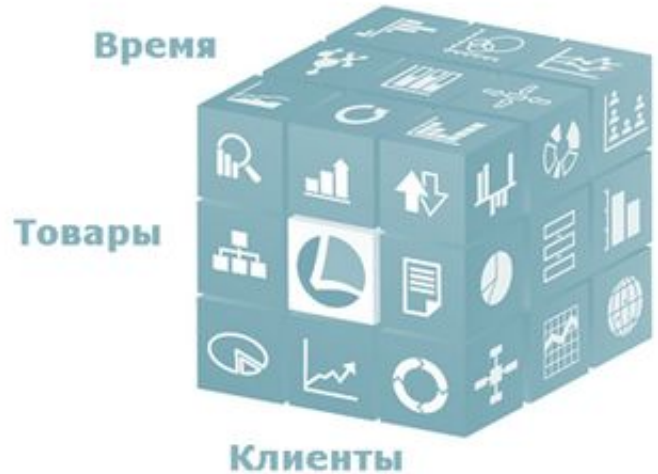
Тиражирование знаний – совокупность инструментальных средств для создания моделей, которые обеспечивают пользователям возможность принятия решений.

Например, в розничной торговле:

- Сколько товара будет продано в следующем периоде?
- Какие клиенты откликаются на акции?
- Какие товары продаются или заказываются вместе?
- Как оптимизировать товарные остатки на складах?

Особенности задач нового 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 эмпирические 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 отчетности* и *интеллектуального анализа данных* (англ.: *Knowledge Discovery in Databases* и *Data Mining*). Сегодня часто все эти технологии рассматривают в контексте термина **бизнес-аналитика**.

Аналитическая пирамида (Analytical stack), предложенная компанией Gartner

BI-платформы, который часто выделяют в отдельную категорию, являются средствами обнаружения знаний (data mining)

OLAP-системы (On-Line Analytical Processing) – системы аналитической обработки данных

Витрины данных (Data marts), как и хранилища, представляют собой структурированные информационные массивы для решения конкретных аналитических задач или обработки запросов определенной группы аналитиков

Хранилища данных (Data warehouse – DW)
Билл Инмон (Bill Inmon), определяет хранилища данных как «предметно-ориентированные, интегрированные, стабильные, поддерживающие хронологию наборы данных, используемые для поддержки принятия управленческих решений»

OLTP (On-Line Transaction Processing) – обработка трансакций в режиме реального времени

Формы представления данных

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

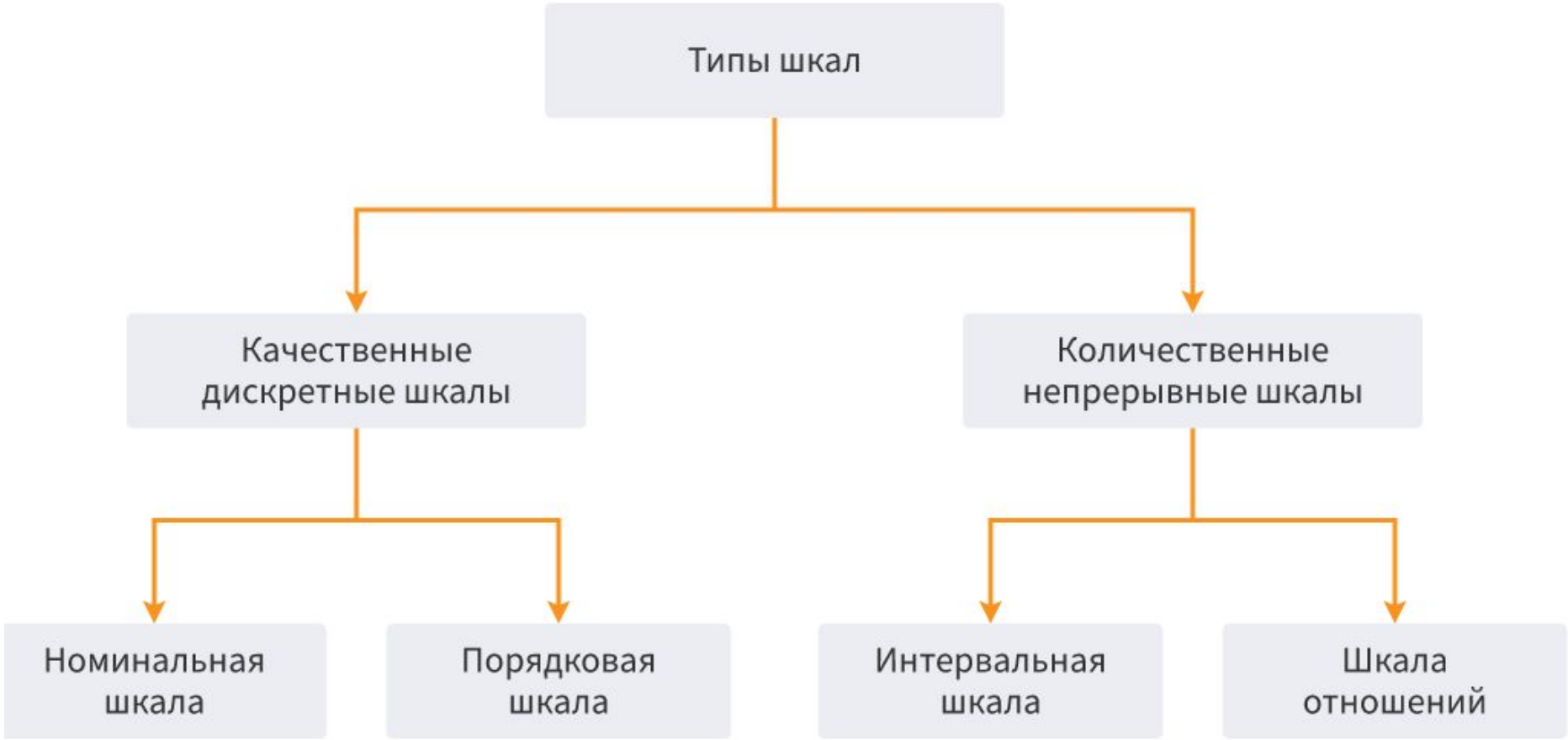
Структурированные данные принято делить на типы:

- Числовой (целый и вещественный);
- Символьный или Строковый;
- Логический (Да/Нет, Ложь/Истина, 1/0);
- Дата/Время.

Неструктурированные данные – данные в произвольной форме:

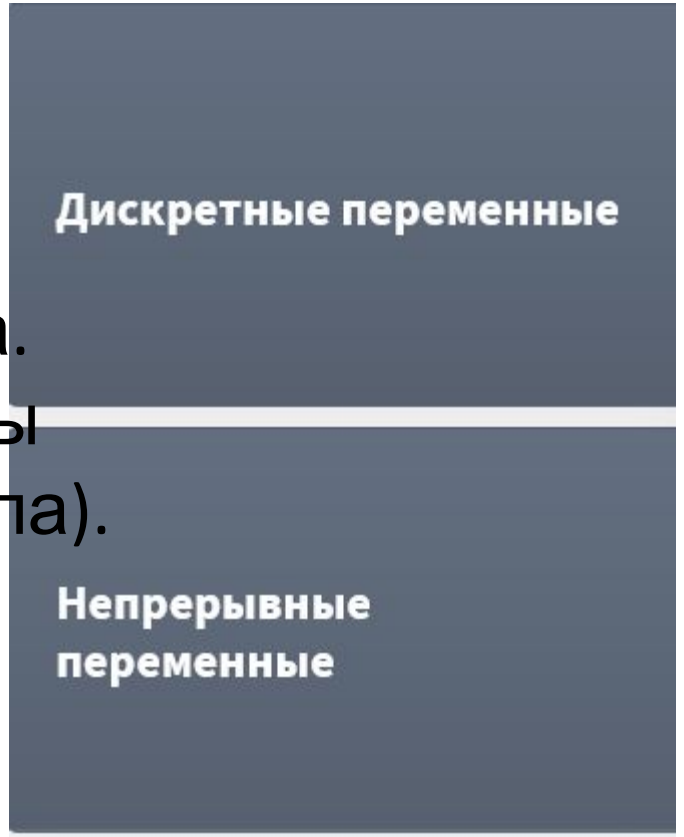
- Видео;
- Речь;
- Аудио;
- Мультимедиа;
- Графика;
- Текст.

Слабоструктурированные данные – правила и форматы определены
в самом общем виде:



Свойства \ Тип шкалы	Номинальная	Порядковая	Интервальная	Отношений
Идентифицируемость	x	x	x	x
Величина (магнитуда)		x	x	x
Равенство интервалов			x	x
Абсолютный ноль				x

По характеру варьирования переменные делятся на:



Дискретные данные являются значениями признака.
С дискретными данными не могут быть произведены никакие арифметические действия (не имеют смысла).
Дискретными данными являются
все данные строкового и логического типа.
Числовые данные являются дискретными
если имеют фиксированное на данный момент значение:
Возраст, Количество студентов в группе, Код товара, Табельный номер и т. д.

Некоторые примеры дискретных данных:

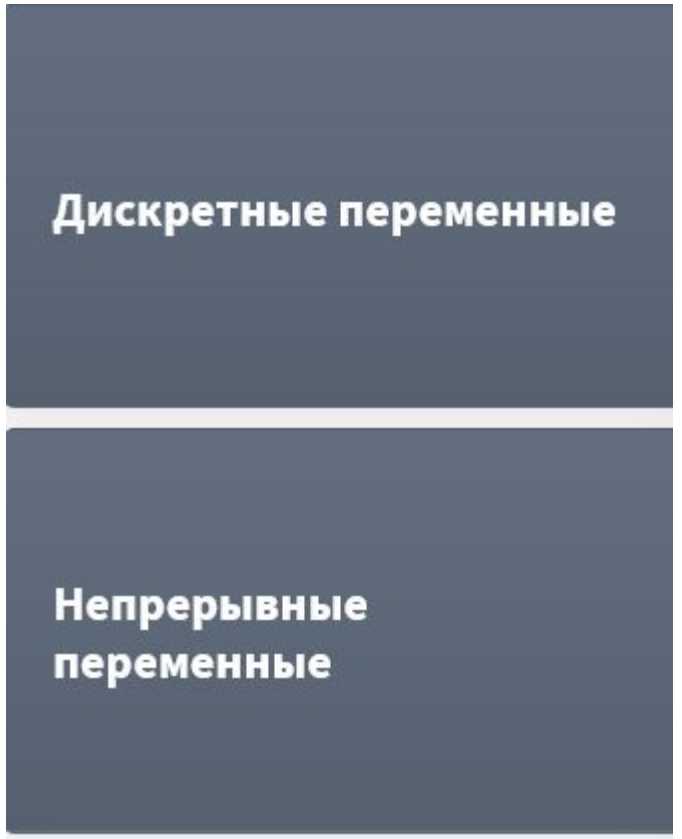
- Количество клиентов, купивших разные товары.
- Количество компьютеров в каждом отделе.
- Количество товаров, которые вы покупаете в продуктовом магазине каждую неделю.

Непрерывные данные — это данные, которые могут принимать
любые значения в некотором интервале.
Над непрерывными данными можно производить арифметические операции и они имеют смысл.

Некоторые примеры непрерывных данных:

- Вес новорожденных малышей.
- Суточная скорость ветра.
- Температура морозильной камеры.

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



Представления наборов 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 рисунках приведены примеры упорядоченных наборов данных – временной ряд (слева, упорядочен по дате) и ряд показаний датчика зонда (справа, упорядочен по глубине скважины).

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	BK	DS
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

Примеры упорядоченных наборов данных

Представления наборов данных

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.

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	2616993

Пример неупорядоченного набора данных

Особо выделяют *транзакционные данные*. Под **транзакцией** подразумеваются 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.

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

Особенности бизнес-данных, накопленных в компаниях

- **Бизнес-данные редко накапливаются специально для решения задач аналитики**
- **Бизнес-данные, как правило, содержат выбросы, ошибки, противоречия и пропуски**
- **С точки зрения анализа объемы хранимых данных очень велики**

1. Абстрагироваться от существующих информационных систем и имеющихся в наличии данных
2. Описать все факторы, потенциально влияющие на анализируемый процесс/объект
3. Экспертно оценить значимость каждого фактора
4. Определить способ представления информации
5. Собрать легко доступные факторы
6. Собрать наиболее значимые факторы
7. Оценить сложность и стоимость сбора средних и наименее важных по значимости факторов

Методы сбора

- Получение из учетных систем: несложная операция, обычно учетные системы имеют развитые методы импорта/экспорта.
- Получение из косвенных источников информации: многие показатели можно оценить по косвенным признакам, например, оценка реального финансового положения жителей региона по объемам покупок товаров для бедных, среднего класса и богатых.
- Использование открытых источников: статистика, отчеты корпораций, маркетинговые исследования, социальные сети и прочее.
- Приобретение данных у специализированных компаний: множество профессионально работающих компаний, стоимость невысокая.
- Проведение собственных мероприятий по сбору данных: дорогостоящий вариант, но всегда существует.
- Ввод данных вручную: данные по экспертным оценкам, трудоемкость высокая.

Информативность данных

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

Среди неинформативных признаков можно выделить четыре типа: (1) **признаки, содержащие только одно значение**; (2) **признаки, содержащие в основном одно значение**; (3) **признаки с уникальными значениями**; (4) **признаки, между которыми имеет место сильная корреляция**, – в этом случае для анализа можно взять только один из них.

Признак
1
1
1
1
1
1
1
1
1
1
1

(1)

Признак
1
1
1
1
1
0
1
1
1
1
1

(2)

№ паспорта
0936-866096
8355-512943
8017-098471
2762-945535
0459-997701
6291-817248
0094-883508
9290-732300
7022-736158
3127-709332
4179-171975

(3)

Пол	Gender
Жен	0
Жен	0
Жен	0
Муж	1
Муж	1
Жен	0
Жен	0
Муж	1
Жен	0
Жен	0
Муж	1

(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.

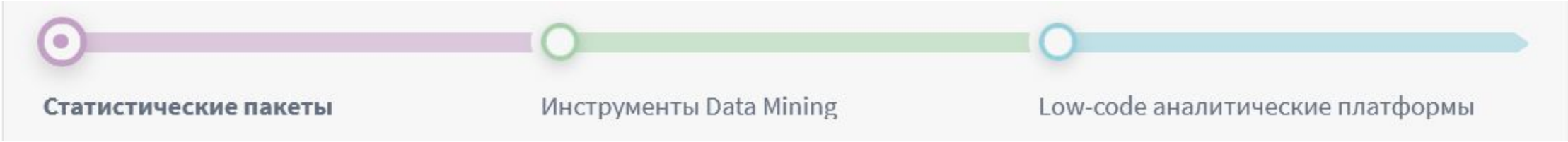
Для *неупорядоченных данных* требования следующие.

- 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 статистически необоснованные шаблоны поведения. Алгоритмы поиска таких шаблонов способны быстро перерабатывать огромные массивы данных. Примерное соотношение между количеством объектов и объемом данных следующее:

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

Инструменты аналитики данных



- Статистические пакеты – хорошая математическая подготовка пользователей; проблемы больших объемов данных; необходимость использования встроенных языков программирования.
- Инструменты Data Mining – возможности современных компьютеров позволяют использовать *хранилища данных*, *Data Mining, Knowledge Discovery in Databases (KDD), Big Data, Deep Learning*.
- Low-code аналитические платформы- специализированные программные системы, автоматизирующие все этапы анализа; аналитические платформы базируются на *low-code* принципах.

Направления развития вычислительной инфраструктуры компании

- Вертикальное масштабирование

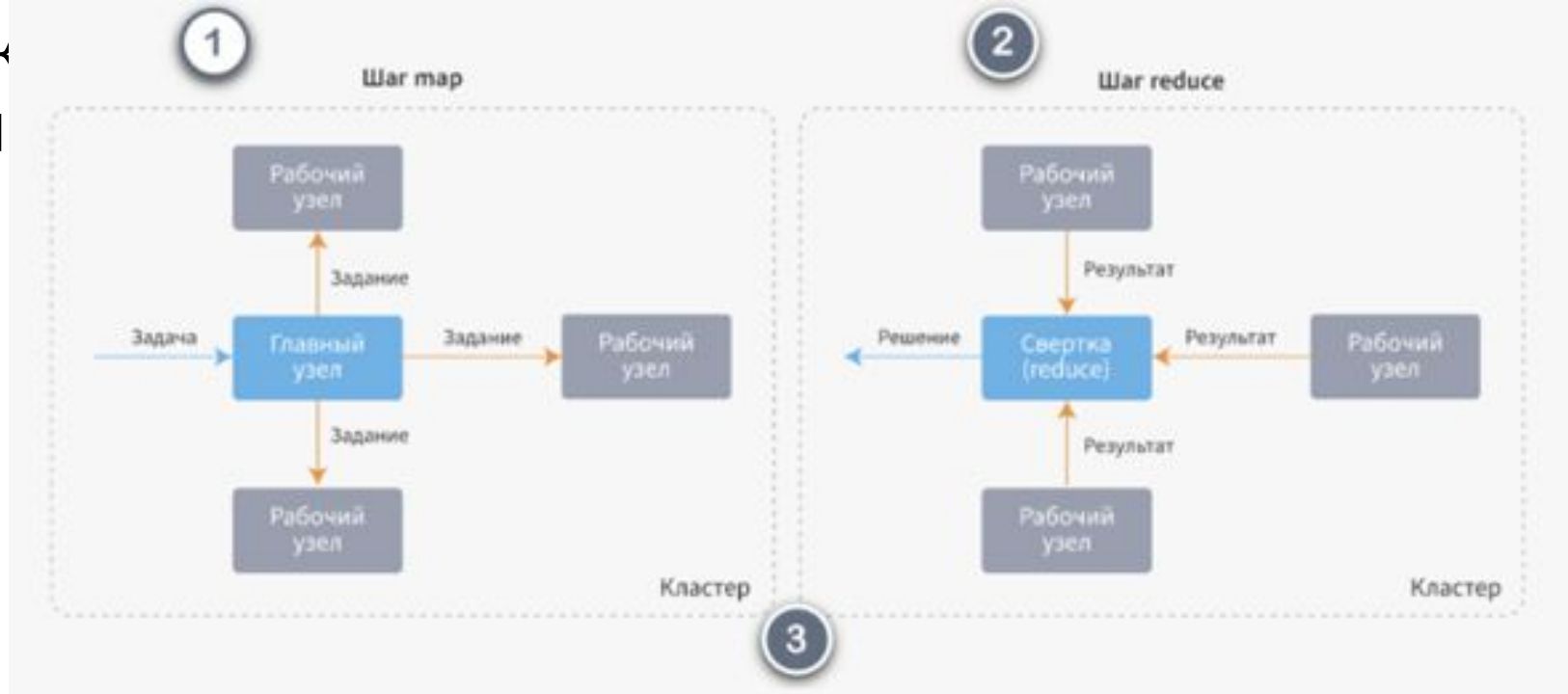
Приобретение более мощного компьютера, то есть добавление ресурсов на единственный вычислительный.

- Горизонтальное масштабирование

Добавление дополнительных недорогих стандартных компьютеров как вычислительных узлов, объединенных в кластер, с распределением работы между ними.

- Большие данные используют технологии ***распределенных вычислений***: вычислительная нагрузка распределяется между некоторым количеством компьютеров-клиентов, которые работают под управлением центрального компьютера.
- Примерами инструментов распределенных вычислений для Больших данных являются ***MapReduce, Hadoop, NoSQL***.

- ***MapReduce*** – модель распределенных вычислений, разработанная компанией Google, которая используется для параллельных вычислений над очень большими (несколько петабайт) массивами данных в распределенных вычислительных сетях.
- Компьютеры в таких сетях делятся на узлы, которые непосредственно производят вычисления, и главные узлы, которые получают задачу, разделяют ее на части и распределяют ее меж
обработки. Данный ш



- После того, как мастер-узел получает от остальных машин сообщение о том, что обработка данных ими закончена (то есть шаг ***map*** завершен), он выдает команду на переход к шагу ***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 C++.

- ***Hadoop*** – проект фонда Apache Software Foundation, свободно распространяемый набор утилит, библиотек и программный каркас для разработки и выполнения распределенных программ, работающих на кластерах из сотен и тысяч узлов.
- ***Hadoop*** используется для реализации поисковых и контекстных механизмов многих высоконагруженных веб-сайтов.
- ***Hadoop*** разработан на основе модели распределенных вычислений MapReduce.
- ***Hadoop*** считается одной из основополагающих технологий 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.

Роль и место Big Data в аналитике данных

- Технологии Knowledge Discovery и Data Mining решают задачи поддержки принятия решений на основе обнаруженных зависимостей и закономерностей в данных, описывающих бизнес-процессы компании.
- 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, но ***Big Data*** оперирует ***петабайтами*** данных неопределенной структуры.
- Роль Big Data с точки зрения предсказательной аналитики заключается в том, чтобы помочь «зачерпнуть» из потока данных образцы, анализ которых поможет описать закономерности всего потока с целью получения знаний о связанных с ним бизнес-процессах.
- Задача Big Data – управление огромными потоками данных из различных распределенных источников, проведение их описательного анализа и формирование наборов данных для построения моделей Data Mining.
- Big Data можно рассматривать ***как технологию подготовки данных*** сверхбольшого, непрерывно возрастающего объема, расположенных в распределенных файловых системах и готовых ***к анализу методами Data Mining.***

Особенности анализа данных уровня Big Data

1) При использовании технологии Big Data в распоряжении исследователя оказывается намного больше данных, причем как структурированных, так и не структурированных.
2) Поэтому для анализа необходимо использовать приложения, «умеющие» работать не только с табличными данными.

3) 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. В случае Big Data такие представления, как правило, отсутствуют.