1. **Назовите источники появления БД**

Большие данные (Big Data) могут появляться из различных источников, включая:

1. Социальные сети: Данные из социальных сетей, таких как Facebook, Twitter, Instagram и LinkedIn, содержат огромное количество информации о поведении пользователей, их предпочтениях и взаимодействиях.

2. Интернет вещей (Internet of Things, IoT): Устройства IoT, такие как датчики, мобильные устройства, умные дома и промышленное оборудование, генерируют большое количество данных, которые могут быть использованы для анализа и принятия решений.

3. Онлайн-транзакции: Данные от онлайн-транзакций, такие как покупки в интернет-магазинах, финансовые операции и т.д., могут быть использованы для анализа покупательского поведения, выявления тенденций и прогнозирования спроса.

4. Медицинские данные: Данные из медицинских учреждений, электронных медицинских записей, изображений медицинского оборудования и геномных данных образуют огромный объем информации, который может быть использован для улучшения диагностики и лечения.

5. Научные исследования: Большие данные также могут возникать из научных исследований, включая геномику, астрофизику, климатические модели и другие области, где данные собираются с помощью крупных экспериментов и наблюдений.

1. **В каких областях используются большие данные**

Большие данные (Big Data) широко используются во многих различных областях из-за их способности предоставлять ценную информацию и понимание на основе объема, скорости и разнообразия данных. Некоторые из областей, где большие данные находят применение, включают:

1. Бизнес и маркетинг: Большие данные помогают компаниям анализировать рынок, понимать потребности клиентов, улучшать стратегии продаж и маркетинга.

2. Здравоохранение: В медицине большие данные используются для улучшения диагностики, прогнозирования заболеваний, разработки персонализированных методов лечения и управления медицинскими данными.

3. Финансы: Большие данные помогают финансовым учреждениям в анализе рисков, прогнозировании трендов на рынке, борьбе с мошенничеством и улучшении инвестиционных стратегий.

4. Транспорт и логистика: Данные помогают оптимизировать маршруты, улучшить логистику, предсказывать трафик и снижать издержки.

5. Наука и исследования: В различных областях науки, таких как астрономия, геномика, климатология, большие данные используются для моделирования, анализа экспериментов и прогнозирования.

6. Государственное управление: Большие данные помогают правительствам в принятии решений на основе данных, улучшении общественных сервисов, мониторинге общественного мнения и борьбе с преступностью.

Это лишь некоторые области применения больших данных, и их использование продолжает расширяться по мере развития технологий анализа данных.

1. **Основные вызовы (4V, 8V):**

4V:

* Volume (Объем);
* Velocity (Скорость передачи, создания и обработки данных );
* Variety (Разнообразие данных);
* Veracity (Достоверность).

8V + 1:

* Volume (Объем);
* Velocity (Скорость передачи, создания и обработки данных );
* Variety (Разнообразие данных);
* Veracity (Достоверность);
* Visualization (Визуализация);
* Viscosity (Вязкость данных);
* Value (Значимость);
* Variability (Изменчивость);
* Virtualization (Виртуализация).

Variability – Изменчивость данных в зависимости от контекста.

Visualization – Необходимо представить данные в формате понятному человеку.

Virtualization – Изоляция вычислительных процессов и ресурсов друг от друга, крч говоря к обработке тема относится.

Viscosity – Сложность сбора данных.

1. **Определение термина «Большие данные», источники получения.**

Большие Данные – это некоторый набор подходов, практик, применяемых для анализа, извлечения ценностей из некоторого набора данных, при этом этот набор данных не поддаётся адекватному анализу с помощью традиционных методов обработки данных ввиду своего размера и/или потока данных.

См 1. вопрос

1. **Перечислите и охарактеризуйте логические слои для работы с большими данными**

* Получение данных – на этом слое осуществляется непосредственно получение, сбор данных из различных источников различными способами, будь то веб-скрэппинг или данные с различных датчиков.
* Хранение данных – на этом слое осуществляется непосредственно хранение данных в различных хранилищах как NoSQL, так и SQL, централизованно ил распределённо, всё зависит от конкретной имплементации.
* Обработка данных – на этом слое производится предварительна обработка данных ( отсекается различный мусор, происходит их нормализация и т.д.) Производится с помощью различных инструментов таких как Apache Airflow, Hadoop, Spark.
* Анализ данных – непосредственно слой, на котором подготовленные данные подвергаются статистическому анализу, выявляются различные закономерности, извлекаются полезная информация. Происходит дата-майнинг и т.д.
* Слой представления – слой отвечающий за визуализацию найденных полезных данных, полученных на предыдущем этапе. Отчёты, графики, 3д-модели – всё это реализуется здесь.
* Слой безопасности – можно выделить отдельный слой, отвечающий за безопасность собираемых и обрабатываемых данных, шифрование данных при передаче, ограничение доступа к хранилищам различными способами и т.д.

1. **Данные, информация, знания – в чём отличия?**

Данные -> Информация -> Знания

* Данные – набор фактов;
* Информация – сведения, уменьшающие неопределённость;
* Знания – сведения, позволяющие действовать с предсказуемым результатом.

1. **Области деятельности, где эффективно используются Большие Данные**

Методы бигдата, используются в большом количестве областей (Наука, Гос. Управление, здравоохранение, Ритейл, Финтех и т.д). Если приводит примеры из лекций, то следует вспомнить:

* АнтиФрод система Visa, которая позволяет определить мошеннические переводы
* МинТруд ФРГ выявил с помощью БД людей, несправедливо получающих пособия по безработице.
* Больница Торонто отслеживает и прогнозирует состояние детей-младенцев в реальном времени на основе собираемых данных.\
* Антиплагиат-система – анализирует огромные массивы текстовых данных и распознаёт плагиат.

1. **Сравнение различных продуктов для аналитики больших данных**

Изображение выглядит как текст, снимок экрана, Параллельный, Шрифт

Автоматически созданное описание

1. **Сравнительная характеристика R VS Python**

**В общем:**

* R-это язык программирования, созданный статистиками и майнерами данных для статистического анализа и графики, поддерживаемый фондом R foundation for statistical computing. R также обеспечивает высококачественную графику, а также имеет некоторые популярные библиотеки, которые помогают в аналитических частях, таких как R Markdown и Shiny.
* Python-это простой, легкий, полноценный и объектно-ориентированный язык программирования высокого уровня, который используется для веб-разработки или разработки ПО для анализа данных, а также в приложениях на основе графического интерфейса, таких как игры, графический дизайн, веб-приложения.

**Скорость:**

Для сравнительно небольших данных – Python быстрее, для больших – R – перехватывает инициативу.

**Тенденции:**

Переход с R на Python более вероятен чем наоборот.

**Глубокое обучение:**

В данном случае Python выигрывает у R, т.к в Python есть библиотеки tensorflow и keano, пусть они и доступны в R.

**Вакансии:**

В целом намечается тенденция на увеличение вакансий на Python.

Но нельзя однозначно выбирать что-то одно, возможно лучшим решением будет сочетание сильных сторон обоих языков.

**Пример кейса:**

*Анализ новостного потока*. Парсинг на R (rvest) + NLP на Python + параметризованный отчет на R (RMarkdown Parameterized Reports).

1. **Охарактеризовать конструкции языка R.**

* Программы на R состоят из *выражений,*
* *Выражения* состоят из *объектов* и *функций,*
* *Объекты* (векторы, списки, функции),
* Имена переменных называются *символами.*
* *Среда –* это набор символов.
* *Функции* – это объекты, которые принимают на вход объекты (*аргументы*) и выдают объекты.

Также R содержит цикл for, while, repeat, классы реализованы в привязке к существующим объектам.

При присваивании переменной объекта, он действительно присваивается в переменной в рамках текущей среды.

При присвоении объект копируется, а не передаётся ссылка на него.

Пример синтаксиса:

X <- ifelse(2 >5,5,2)

1. **(11,12,13) Перечислить типы языка R, структуры данных, примеры.**

Изображение выглядит как текст, снимок экрана, Шрифт, число

Автоматически созданное описание

|  |  |  |
| --- | --- | --- |
| **Vector** | **Является упорядоченным набором однотипных объектов** | **1:500, c(1,2,3,4,5), c(1,2,c(2,5),1,2)** |
| **List** | **Отличается от вектора тем, что может быть поименован и хранить разные типы данных, в том числе даже функции** | **list(**  **first=c(1,2,5),**  **1:5,**  **5+10**  **)** |
| **Array** | **Трехразмерная структура, хранит однотипные данные** | array(0, dim = c(2, 2, 3)) |
| **Matrix** | **Двухразмерная структура, хранит однотипные данные** | matrix(0, dim = c(2, 2)) |
| **DataFrame** | **По своеё сти представляет экселевскую таблицу, есть заголовок, колонки и строки, может хранить разные типы данных** | **n = c(2, 3, 5)**  **s = c("aa", "bb", "cc")**  **b = c(TRUE, FALSE, TRUE)**  **df = data.frame(n, s, b)** |
| **Factor** | **По своей сути является энумом** | x <-c("female", "male", "male", "female")  gender <-factor(x) |



1. **Принцип массивных (векторных и матричных вычислений R)**

Принцип векторизованных вычислений применим не только к векторам как таковым, но и к более сложным объектам R - [матрицам](http://r-analytics.blogspot.de/2011/07/r_10.html), [спискам](http://r-analytics.blogspot.de/2011/07/r_13.html) и [таблицам данных](http://r-analytics.blogspot.de/2011/07/r_22.html) (для R разницы между последними двумя типами объектов не существует - фактически таблица данных является списком из нескольких компонентов-векторов одинакового размера). В базовой комплектации R имеется целое семейство функций, предназначенных для организации векторизованных вычислений над такими объектами.

1. **Основные понятия мат. Статистики и дескриптивный анализ**

* Генеральная совокупность – всё множество объектов, к которому формируется некоторая гипотеза.
* Выборка – подмножество генеральной совокупности.
* Репрезентативность выборки – параметр, характеризующий возможность обобщить выводы сделанные в рамках выборки на генеральную совокупность.
* Параметры – характеристики генеральной совокупности
* Статистики – характеристики выборки
* Дескриптивный анализ - статистические методы обработки данных, их систематизация, наглядно представленная в виде таблиц и графиков, а также количественное описание данных с помощью статистических показателей. / Вид исследования направленный на получение факт. Данных о распространяемости изученных явлений различ. Его направлений.

1. **Генеральная совокупность и выборка**

* Генеральная совокупность – всё множество объектов, к которому формируется некоторая гипотеза.
* Выборка – подмножество генеральной совокупности.
* Репрезентативность выборки – параметр, характеризующий возможность обобщить выводы сделанные в рамках выборки на генеральную совокупность.
* Параметры – характеристики генеральной совокупности
* Статистики – характеристики выборки

Чем выше изменчивость некоторого параметра, тем больше нужно брать выборку.

Для сравнения двух выборок их размер должен быть примерно одинаков и общая численность не менее 50 человек.

*Способы отбора выборки:*

* Простой рандомизированный отбор
* Отбор по свойствам генеральной совокупности, изучаются те параметры, которые влияют на изменчивость изучаемого свойства. Затем определяется процентное соотношение групп различающихся по этим свойствам в генеральной совокупности, а затем в выборке поддерживается то же соотношение по подгруппам.

1. **Шкалы Измерений**

Изображение выглядит как текст, снимок экрана, Шрифт, диаграмма

Автоматически созданное описание

Мощность в данном контексте означает насколько много данных может предоставить шкала исследователю.

Номинальная – используется только для качественной классификации, например пол, цвет, раса и т.д.

Порядковая - эта шкала уже позволяет ранжировать результаты, к примеру (понравился, средне, не понравился как оценка фильму).

Интервальные – эта шкала позволяет сравнивать результаты относительно друг-друга в числовом виде, т.е получается определить насколько значение, например температура 40 не только выше 20, но и то что увеличение с 0 до 40 вдвое больше чем с 20 до 40. И при этом уровни значений у этой шкалы – одинаковы.

Относительные – эта шкала, расширяет предыдущую, устанавливая некоторую точку отсчёта, а именно 0. Как пример – масса тела, объем

Изображение выглядит как текст, снимок экрана, Шрифт, число

Автоматически созданное описание

1. **Меры центральной тенденции, их сравнительный анализ**

Мера центральной тенденции – это число, которое характеризует выборку по уровню выраженности некоторого признака.

*Меры*:

**Мода** – самое часто встречающееся значение в выборке. На графике распределения частот ей соответствует *максимум* (или же *вершина*). Если вершин несколько, то выборка – *мультимодальная*, иначе – *унимодальная.*

Изображение выглядит как линия, диаграмма, снимок экрана, График

Автоматически созданное описание

**Медиана** – центральное значение в отсортированной по порядку выборке.

Изображение выглядит как текст, снимок экрана, Шрифт, линия

Автоматически созданное описание

**Среднее –** Сумма всех значений в выборке, поделённое на количество элементов в ней

Применимость в различных шкалах:

Изображение выглядит как текст, снимок экрана, линия, число

Автоматически созданное описание

1. **Нормальное распределение, уровень статистической достоверности**

Распределение – функция от X вычисляющая вероятность наличия X в некоторой выборке.

Нормальное распределение Гаусса – распределение вероятностей, которое задаётся функцией:

Изображение выглядит как Шрифт, текст, белый, линия

Автоматически созданное описание, где

***μ -*** среднее значение и задаёт положение кривой на оси, задаёт её вершину.

σ - стандартное отклонение, задаёт ширину «колокола» графика

Изображение выглядит как линия, диаграмма, График, скат

Автоматически созданное описание

Изображение выглядит как текст, снимок экрана, линия, диаграмма

Автоматически созданное описание

**1. Статистическая значимость**: Это концепция, которая позволяет оценить, насколько результаты исследования или эксперимента статистически значимы, то есть насколько вероятно получить такие же или более экстремальные результаты случайно. Если результат статистически значим, это означает, что он вероятно не является результатом случайности.

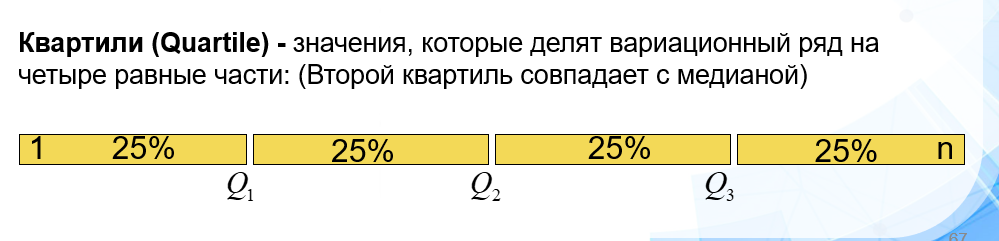
**2. p-уровень значимости**: p-уровень значимости (или p-value) представляет собой вероятность получить наблюдаемые данные (или более экстремальные) при условии, что нулевая гипотеза верна. Чем меньше значение p, тем сильнее доказательства против нулевой гипотезы. Обычно используются стандартные уровни значимости, такие как 0.05 или 0.01: если p < 0.05, то результат считается статистически значимым на уровне 0.05.

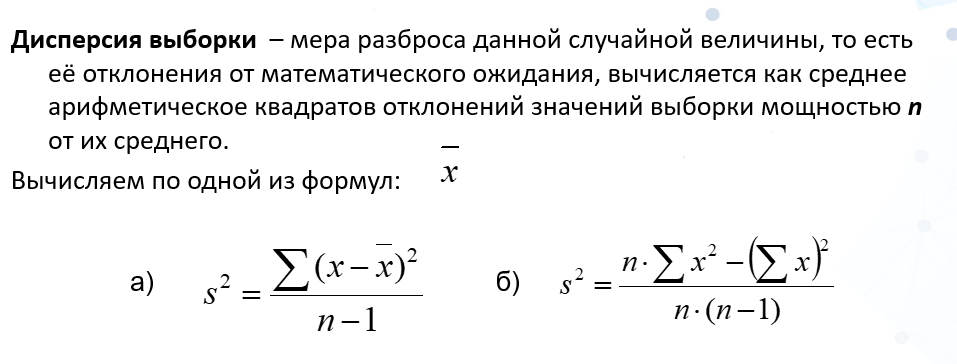
* P ≤ 0,05 – обычный уровень, т. е. получен статистически значимый результат.
* P ≤ 0,01 – высокий уровень, т. е. выявлена выраженная закономерность.
* P ≤ 0,001 – очень высокий уровень.

Итак, статистическая значимость оценивает общую значимость результатов исследования, а p-уровень значимости предоставляет конкретную вероятность получить такие данные при условии верности нулевой гипотезы. Они являются важными инструментами для принятия статистически обоснованных выводов на основе данных и проверки гипотез.

1. **Свойства описательных статистик (Дескриптивный анализ). Меры изменчивости.**

**Размах** – Разница между максимальным и минимальным значением в выборке.





**σ - Стандартное отклонение – является корнем дисперсии выборки**

