**База данных (БД)** – это совокупность данных, хранящихся и упорядоченных в соответствии с определенной **структурой.**

Какой протокол используется для передачи данных между распределенными приложениями на линух — серверах? FTP/SSH (ответы разнятся, точный не знаю)

Распределённая система - это совокупность компьютеров, которые обмениваются данными и синхронизуются в общей сети, образуя единый компьютер для конечного пользователя

Структура таблицы в обработке и хранении данных - это перечень полей и их типов данных, которыми представлена таблицы данных?

Модель данных определяет, как и каким образом данные будут располагаться в бд и как к ним будет предоставляться доступ?

В формате данных JSON данные хранятся в виде таблицы, с разлинованными строками и столбцами в виде отдельных объектов документа

Среди шкал выберите те, в которых меры среднего может послужить среднее арифметическое: Интервальная шкала

Для чего необходимо ведение журналов транзакций транзакционными системами РСУБД?

Поддержка целостности данных, Журналирование запросов, Восстановление реляционной БД

Транзакция - группа операторов определения манипуляции данных переводящих базу данных из одного согласованного состояния в другое согласованное состояние

Выберите пример правильно составленного запроса с использованием агрегирующей функции SUM:

SELECT SUM(price) FROM Orders

Упорядочивание записи в таблице по строковой переменной осуществляется по правилу:

Посимвольное лексикографическое правило в соответствии с положением символа в кодовой таблице символов

Фильтрация данных - операция выборке строк (объектов) или наблюдений из таблицы данных в соответствии с логическим правилом сравнения значений выбранного атрибута с определённым значением.

Обогащение данных - это процесс дополнения сырых данных той информацией, которая в исходном виде в них отсутствует, но необходима для качественного анализа

Витрина данных — это часть хранилища данных, секционированная для отделов или направлений бизнеса (например, продажи, маркетинг или финансы)

Выберите верные стадии ETL процесса, расположите их в верном порядке следования. Запишите число без пробелов, состоящее из цифр последовательности.

341 (Extract Transform Load) извлечение преобразование загрузка

Сопоставьте компоненты конвейера обработки данных с их функциями.

Сортировка - Упорядочивание записей в таблице данных в соответствии с порядком значений выбранного поля или полей... ,

Фильтрация - Выборка данных на основе логических выражений сравнений над значениями в полях таблицы данных,

Слияние - Объединение записей по совпадающей структуре полей нескольких таблиц , Группировка - Вычисление агрегационных функций над подвыборками данных, сформированных на основе совпадения значений в выбранном поле групп

Аналитические базы данных - специализированные ***колоночные*** РСУБД, оптимизированные для быстрой выборки данных из витрин

Расположите стадии проведения аналитики данных в правильном порядке

25143678

Процесс загрузки данных в долговременное хранилище данных заключается в:

2. переносе данных из внешних источников в структуры хранилища данных

3. загрузке данных из сети Интернет файловое хранилище сервера

В чем состоят основные отличия ETL от ELT процессов?

3. Реализация стадии предобработки данных перед загрузкой в долговременное хранилище данных для структуризации информации

4. Работа, в основном, со структурированными данными для решения задач анализа

Какие задачи решает аналитика данных?

3. Получение новых знаний о предметной области исходя из данных, отражающих реализацию конкретного бизнес процесса на основе методов обработки данных

Какие задачи решает агрегация (группировка) данных?

Подсчёт агрегатированных функции (статистик) по выборкам показателей по группам совпадающих дискретных признаков

Какие факторы ограничивают кратное повышение производительности в решениях на базе распределённых архитектур?

Соединение исполнителей в сети и помехи;

Модель данных, предполагающая наличие ориентированной связи между объектами, от родительского объекта к дочерним называется \_\_\_

иерархическая

Возможно ли использование одновременно двух агрегирующих функций: SELECT MIN(price), MAX(price) FROM Orders

да, в результате мы получим мин и макс стоимости

Основными функциями СУБД являются:

Управление данными во внешней памяти, Управление буферами оперативной памяти, Поддержка языков базы данных, Журнализация и резервное копирование базы данных

Выберите верные утверждения, относящиеся к понятию неструктурированных данных:

не имеет предопределенной организации и имеет множество форм, большое количество различных форматов данных, хранят в озёрах данных файловых бд

Какое ключевое слово SQL запроса позволяет упорядочить строки в соответствии со значениями в выбранных полях?

ORDER BY

**Структурированными** называются данные, отражающие отдельные факты предметной области и упорядоченные определенным образом с целью обеспечения возможности применения к ним различных методов обработки

**Большие данные** – разнообразные данные, которые поступают с постоянно растущей скоростью и объем которых постоянно растет

Частное облако – инфраструктура и вычислительные ресурсы принадлежат и эксплуатируются одной организацией

**Интеллектуальный анализ данных** (data mining) исследует закономерности, которые нельзя выявить при обычном просмотре данных, и использует методы математического и статистического анализа.

**Серверные ОС –** предназначены для управления программным обеспечением, которое в свою очередь обслуживает всех пользователей сети, как внутренней, так и внешней

Серверные ОС не предоставляют специализированный графический инструментарий управления системой и управляются напрямую с использованием командной строки

Серверные ОС являются более предпочтительными с точки зрения экономии ресурсов и гибкости использования управления системой на основе команд

Наибольшее распространение получили Linux системы с UNIX-подобными утилитами командной строки

Что не является синонимом к «строке» - признак

**Дискретные данные** являются значениями признака, общее число которых конечной или бесконечно, но может быть подсчитано при помощи натруральных чисел.

**Непрерывные данные** – это данные, которые могут принимать любые значения в некотором интервале.

СУБД работает в соответствии со **структурой**, на которой строится база данных

**Схема БД** – список таблиц, их атрибутов, типов данных, ограничений, ключей и связей между таблицами, необходимый для корректной организации хранения данных в памяти вычислительного устройства и доступа к данным извне, как на запись, так и на чтение.

Доступ к данным в реляционных СУБД классически осуществляется с помощью языка **DML**, подъязыка SQL.

**Аналитические базы данных** — специализированные колоночные РСУБД, оптимизированные для быстрой выборки данных из витрин.

Основаны на системах массивных параллельных вычислений (MPP) и базах данных, поддерживающих такой режим работы.

Предназначены такие БД для обработки данных на стадии работы непосредственно с витринами больших данных.

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

Профайлинг данных включает в себя следующие этапы:

Подведение общих описательных статистик по выборке.

Обнаружение пропусков.

Обнаружение выбросов и экстремальных значений.

Обнаружение дубликатов и противоречий.

Сложные проверки.

**Чистые данные** представляют собой табличный набор наблюдений в котором каждой строке данных соответствует полный перечень атрибутов c адекватными значениями.

**Грязные данные** - это неверные, недостаточные, не несущие никакой пользы.

Среди шкал выберите те, на которых невозможно ввести отношение порядка  
Номинальная

**Сортировка табличных данных** – преобразование, упорядочивающее набор объектов (строк) или наблюдений в связи с правилом упорядочивания по выбранным атрибутам.

Какая функция из перечисленных НЕ является агрегационной? Корень числа

**Агрегирование данных —** это сбор информации из баз данных с целью подготовки комбинированных наборов данных для обработки данных.

Свойства распределенной системы: совместное использование ресурсов; Совместная обработка; масштабируемость; обнаружение отказов; прозрачность.

Верные стадии очистки данных: обработка пропусков, удаление дубликатов и противоречий, обработка выбросов (нетипичных значений), верификация целостности данных

Способы обработки ситуаций с выбросами и экстремальными значениями:  
удаление, замена, интерпретация выбросов и использование спец. моделей

Какие виды обогащения данных существуют: локальное, внешнее

... это система аналитической обработки данных. Она предназначена для подготовки отчетов, построения прогностических сценариев и выполнения статистических расчетов. (OLAP)

... это выполнение различных преобразований данных с целью их подготовки к анализу или моделированию

(Трансформация данных)