**Базы данных и SQL**

База данных (БД) – структурированное хранилище информации, организованные таким образом, чтобы обеспечить хранение, управление и доступ к информации. Чаще всего эти данные взаимосвязаны. Сама по себе БД – это просто хранилище данных, остальные свойства базы данных обеспечиваются за счет системы управления базой данных.

Система управления базой данных (СУБД) – программные средства для управления данными. Отвечает за поддержку языка БД (обеспечивает интерфейс для взаимодействия с данными), за механизм хранения и извлечения данных, за оптимизацию процессов извлечения данных.

**Виды БД:**

*Без использования СУБД*

|  |  |
| --- | --- |
| Файловые | Данные хранятся в виде текста в файле. |
| Иерархические | Имеет древовидную структуру (узел может иметь только одного родителя) Пример: файловая система. |
| Сетевые | Имеет структуру графа (узел может иметь более одного родителя) |

*С использованием СУБД*

|  |  |
| --- | --- |
| Реляционные | Структура в виде таблиц, связанных между собой |
| NoSQL | Общее название для нереляционных БД. |

**Типы СУБД:**

NoSQL (not only SQL) – Нереляционные СУБД, в которых данные могут быть организованы в виде иных структур, отличных от таблиц, ориентированных под свои задачи.

Используется язык отличный от SQL.

Отсутствие жесткой схемы данных (Приятное следствие отсутствия схемы — эффективность работы с разреженными (sparse) данными. Если в одном документе есть поле date\_published, а во втором — нет, значит никакого пустого поля date\_published для второго создано не будет)

Отказ от атомарности и согласованности данных в пользу горизонтального масштабирования.

1) Ключ-значение – база данных организована в виде словаря. Некая таблица, имеющая уникальный ключ и соответствующее ему значение. Подходит для кэширования. [Redis]

Высокая скорость, легкое масштабирование. Не подходит для сложных структур (сущности с разными типами данных) и сложных запросов.

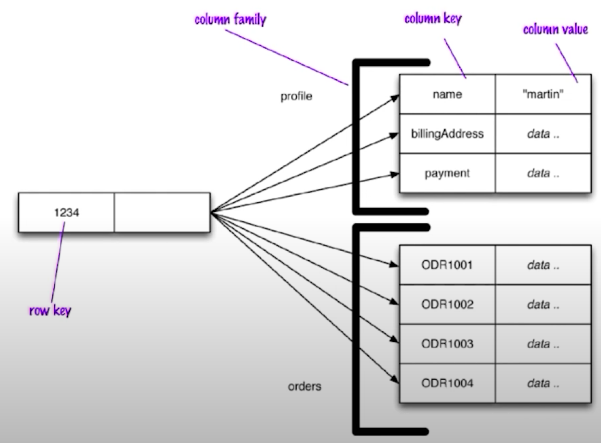
2) Колоночные – Данные организованы в виде таблиц так же, как и в реляционной СУБД, но не построчно (для чтения колонки придется вычитать всю строку до нужной колонки), а «поколоночно», т.е. колонка — это отдельная структура, в которой содержатся данные для всех строк по определенному ключу.

Подходит для аналитических запросов на больших объемах данных (эффективна при запросах для более 100 миллионов строк) – OLAP. [Apache druid]

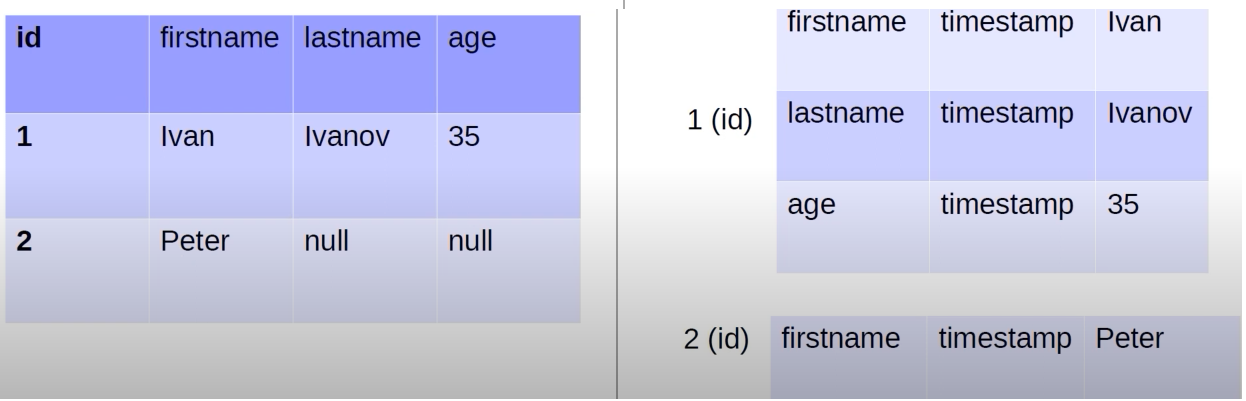
3) Ширококолоночные (Wide Column Stores) – Каждому ключу соответствует набор column family – логическая группировка колонок, каждая из которых состоит из множества сущностей ключ (названия колонки) – значение.

В отличие от реляционных не имеют жесткой схемы (в случае добавления или отсутствия значений не нужно хранить null) (см. рис.2), лучшее сжатие, горизонтально масштабируема [HBase, Cassandra]

Общая структура:



Реляционная СУБД Колоночная СУБД



4) Документо-ориентированная – Данные организованы в виде коллекций (аналог таблицы), которые в свою очередь содержат документы (аналог строки в таблице). Документ представлен парами ключ-значение, имеет формат JSON или BSON

Гибкий – не требует схем данных, данные хранятся в одной сущности – не требует соединений, как в реляционной, горизонтально масштабируемый

5) Графовые – имеют структуру графа, необходимы в ситуации, когда нужно отразить большое кол-во связей между сущностями. [Neo4j, AllegroGraph]

6) Реляционные – Данные организованы в виде отношений (таблиц) которые имеют жесткую схему организации, систему операций над которыми определяет реляционная алгебра, операции выражаются языком SQL. Согласованность данных в приоритете, заточена под транзакции, которые поддерживают свойства ACID. Построчное хранение данных – быстрая вставка, но накладные расходы при чтении данных.

**CAP теорема –** распределенная система не может в полной мере обеспечивать свойства consistency (согласованность), availability (доступность), partition tolerance (устойчивость к разделению)

Consistency – Гарантирует, что запрос к системе выдаст последнюю версию изменений, даже если изменение было произведено на другом узле – актуальность данных.

Availability - Каждый запрос завершается успешным ответом.

Partition tolerance – Система продолжает функционировать при наличии разделения/разрыва сети между узлами

CA системы – такая система подразумевает, что если выходит из строя один из компонентов системы, то она перестает работать

CP – система не гарантирует, что ответ от нее будет получен в разумные сроки (пока запрашиваемые данные не будут синхронизованы на всех узлах, клиент будет ждать ответ на запрос)

AP – система не гарантирует консистентность данных, такие системы не подразумевают того, что данные будут всегда противоречивы, они говорят о том, что данные не будут констстентны в каждый момент времени

Данная «теорема» является условной, не является теоремой в математическом понимании и содержит много противоречий [1](https://habr.com/ru/articles/231703/), [2](https://habr.com/ru/articles/322276/), [3](https://habr.com/ru/articles/328792/), [4](https://habr.com/ru/articles/258145/)

Для описания свойств SQL и NoSQL систем используются ACID и BASE

**ACID –** свойство транзакций, направленные на обеспечение надежности и целостности данных.

*Atomicity* (Атомарность) – Транзакция не может быть зафиксирована частично, то есть не может иметь промежуточных состояний, она либо выполнена полностью, либо не выполнена вовсе.

*Consistency* (Согласованность) - свойство, гарантирующее, что каждая успешная транзакция зафиксирует только допустимые результаты. То есть, это гарантия того, что при успешной транзакции будут выполнены все правила, ограничения, которые предъявляет система к конкретным данным, иначе транзакция не будет выполнена и данные в системе вернутся к прежнему состоянию.

*Isolation* (Изолированность) – Параллельные транзакции (транзакции, работающие с одними и теми же данными) не должны влиять на результаты выполнения друг друга.

Durability (Надежность) – Если клиент получил подтверждение от системы подтверждение завершения транзакции, то эти изменения не будут отменены из-за сбоя.

***Уровни изоляции для ACID -***

**BASE** – альтернативный подход к построению распределенных систем или свойства распределенных систем, в которых Consistency и Availability выполнены не в полной мере.

Basically Available (Базовая доступность) – на любой запрос будет дан ответ, но он может отражать неконсистентные данные.

Soft state (Неустойчивое состояние) – данные в системе могут меняться, даже если изменений в системе нет

Eventual consistency (согласованность в конечном итоге) – при отсутствии изменений в системе, в конечном итоге, все данные будут согласованы.

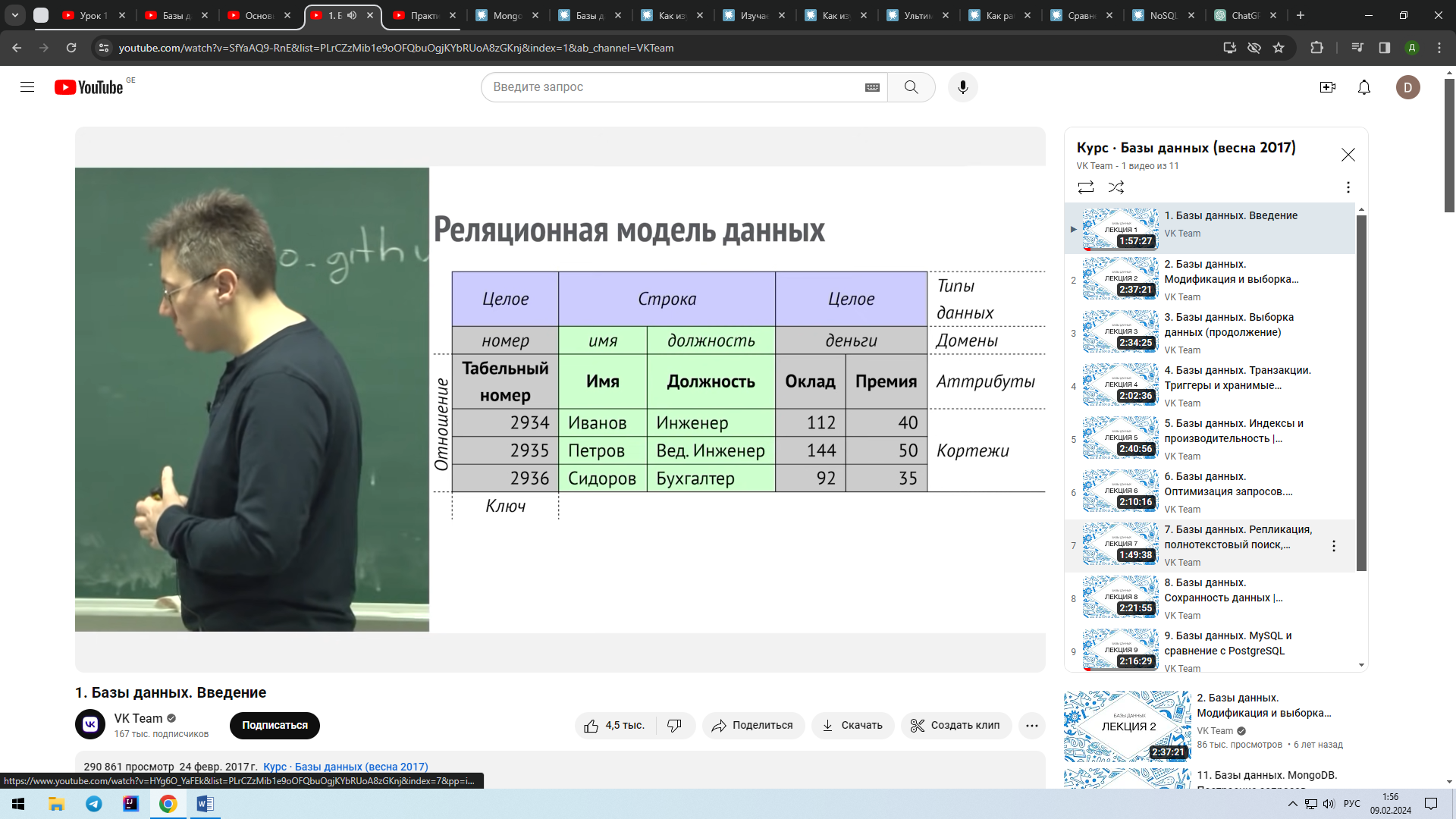
**OLAP/OLTP –** подходы для обработки данных в БД

*OLAP* **(**Online Analytical Processing**) –** Предназначен для аналитической обработки данных: анализ и агрегация больших объемов данных, заточен под чтение, а не под изменение данных.

*OLTP*(Online Transaction Processing) – Предназначен для обработки транзакций в реальном времени.

**Подробнее про реляционные СУБД**

*Реляционная модель данных:*



*Атрибут (столбец) –* Имеет Тип данных, Домен, Имя

*Домен –* некое условие для типа данных (ограничение)

*Ключ* – атрибут, который однозначно идентифицирует каждую запись в таблице (должен обладать свойством уникальности и иметь значение отличное от Null).

В качестве первичного ключа может служить либо естественный ключ (т.е. уже существующий атрибут). Минусы – проблемы, в случае его изменения, так как нужно будет поменять его во всех таблицах, может занимать больше памяти. Либо суррогатный – к отношению добавляется новый атрибут никак его не характеризующий, например, порядковый номер или id, в качестве значения используется число с автоинкрементом.

*Строка/кортеж* – одна запись в таблице, т.е. набор значений каждого атрибута для определенного ключа

*Отношение -* таблица

*Сущность* – Объект предметной области, который может быть описан

Результирующий набор – результат запроса на языке SQL

**Реляционная алгебра и SQL –** любой запрос на языке SQL представляет из себя набор из перечисленных операций из реляционной алгебры.

|  |  |  |  |
| --- | --- | --- | --- |
| Выборка | Извлекает только те строки, которые удовлетворяют заданному предикату (условию). | SELECT \* FROM “Люди” WHERE “Возраст” >=34 |  |
| Проекция | Получить только указанные атрибуты | SELECT “Возраст”, ”Вес” FROM “Люди” |  |
| Объединение | Возвращает отношение, содержащее все уникальные кортежи, которые встречаются в отношениях. | SELECT \* FROM “Люди” UNION SELECT \* FROM “Персонажи” |  |
| Пересечение | Возвращает только совпадающие кортежи | SELECT \* FROM “Люди” NATURAL JOIN “Персонажи” |  |
| Разность | Возвращает отношение, где первое отношение берется как основа и повторяющиеся значения из второго отношения не включаются в результирующее | SELECT \* FROM “Люди” NATURAL LEFT JOIN “Персонажи” WHERE “Персонажи” IS NULL |  |
| Произведение | Каждая запись из первого отношения комбинируется с каждой записью из отношения (декартово произведение) | SELECT \* FROM “Люди”, “Работники” |  |
| Деление | Из первого отношения берутся записи, которые включают все комбинации из второго отношения |  |  |
| Соединение | Записи из отношений соединяются по ключу | **SELECT \* FROM “asd” JOIN “DSA” ON** |  |

Объединение пример: При операции объединения двух отношений с разными атрибутами на выходе получаем отношение, содержащее все атрибуты из обоих отношений, если атрибуты отношений совпадают, то в конечном отношении будут содержаться только уникальные строки.

|  |  |
| --- | --- |
|  |  |

**Категории запросов**

DDL (Data Definition Language) – CREATE, ALTER, DROP – команды для определения и изменения структуры БД

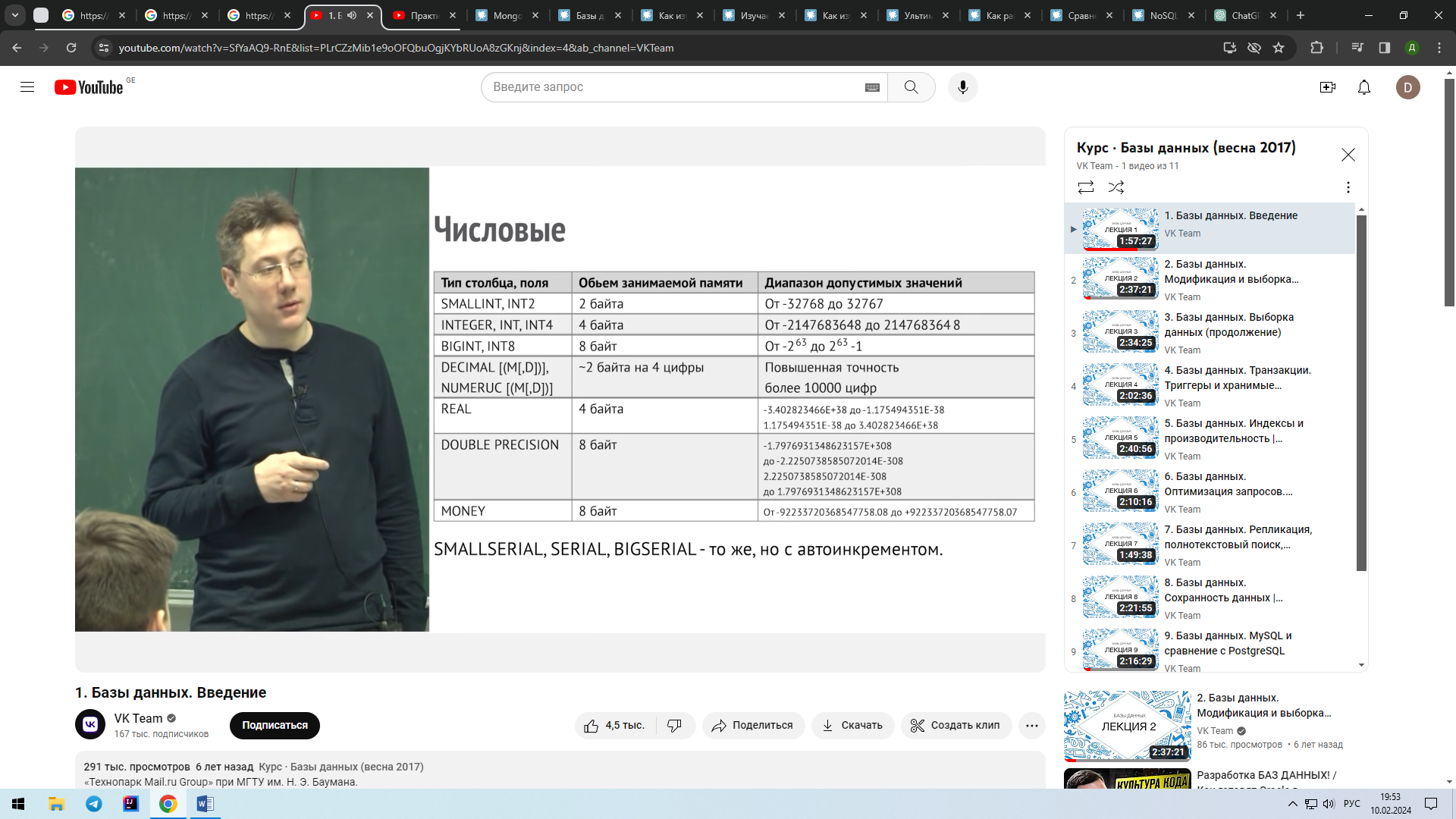
DML (Data Manipulation Language) – SELECT, INSERT, UPDATE, DELETE, MERGE – команды для манипулирования данными внутри таблицы

DCL (Data Control Language) – GRANT, REVOKE, DENY - команды для управления правами доступа и безопасностью данных.

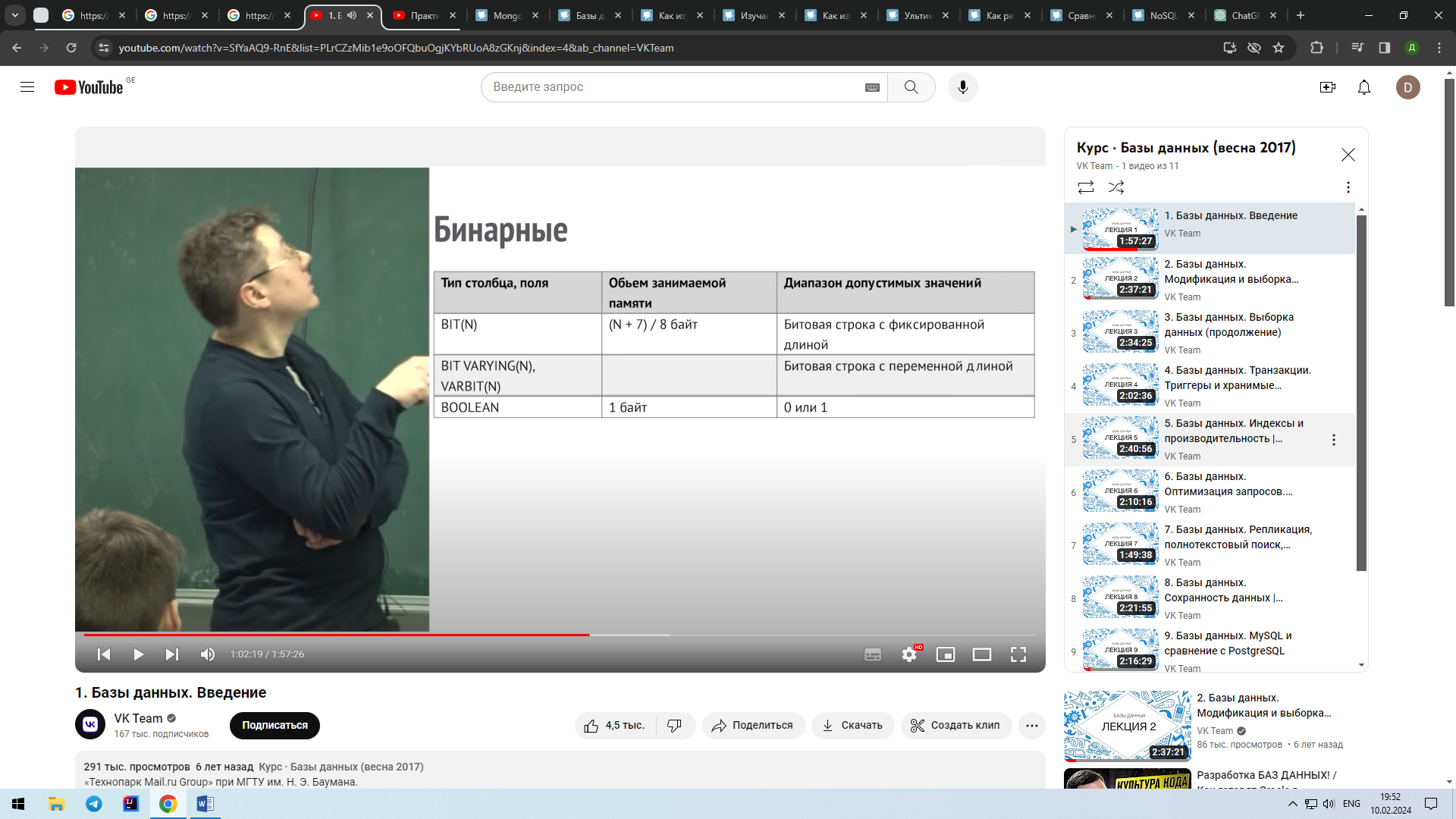
TCL (Transaction Control Language) - COMMIT, ROLLBACK, SEVEPOINT - Команды для управления транзакциями.

**Типы данных в PostgreSQL**

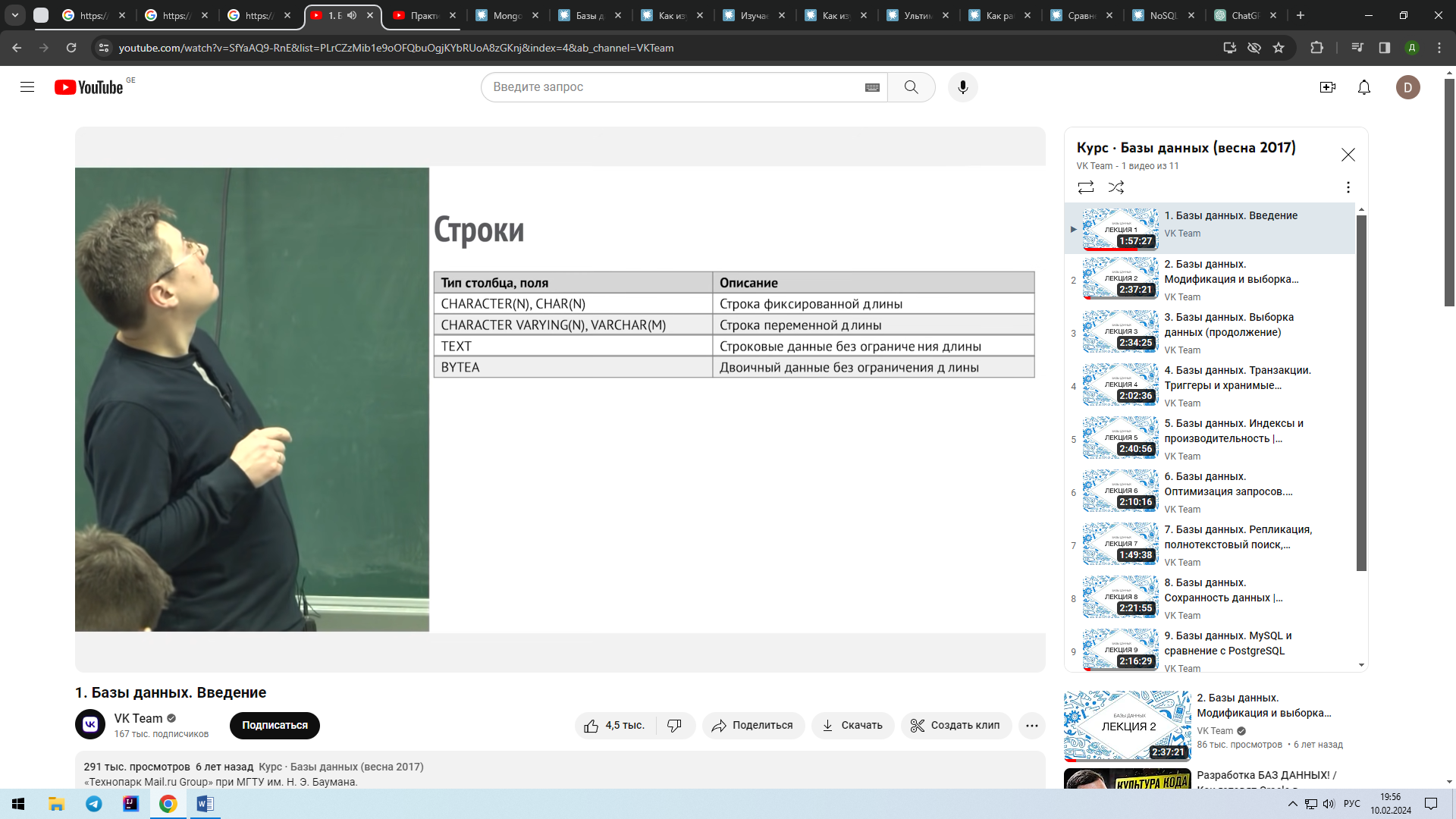
*Числовые*

*decimal/numeric* –используется для обозначения денежных значений (можно использовать тип MONEY, но он менее точен) m –максимальное общее кол-во цифр, d – максимальное кол-во цифр после запятой. То есть целая часть может содержать не больше m-d цифр. Если d больше допустимого, то оно будет округлено до допустимого.

Serial – беззнаковые, автоинкрементируемые

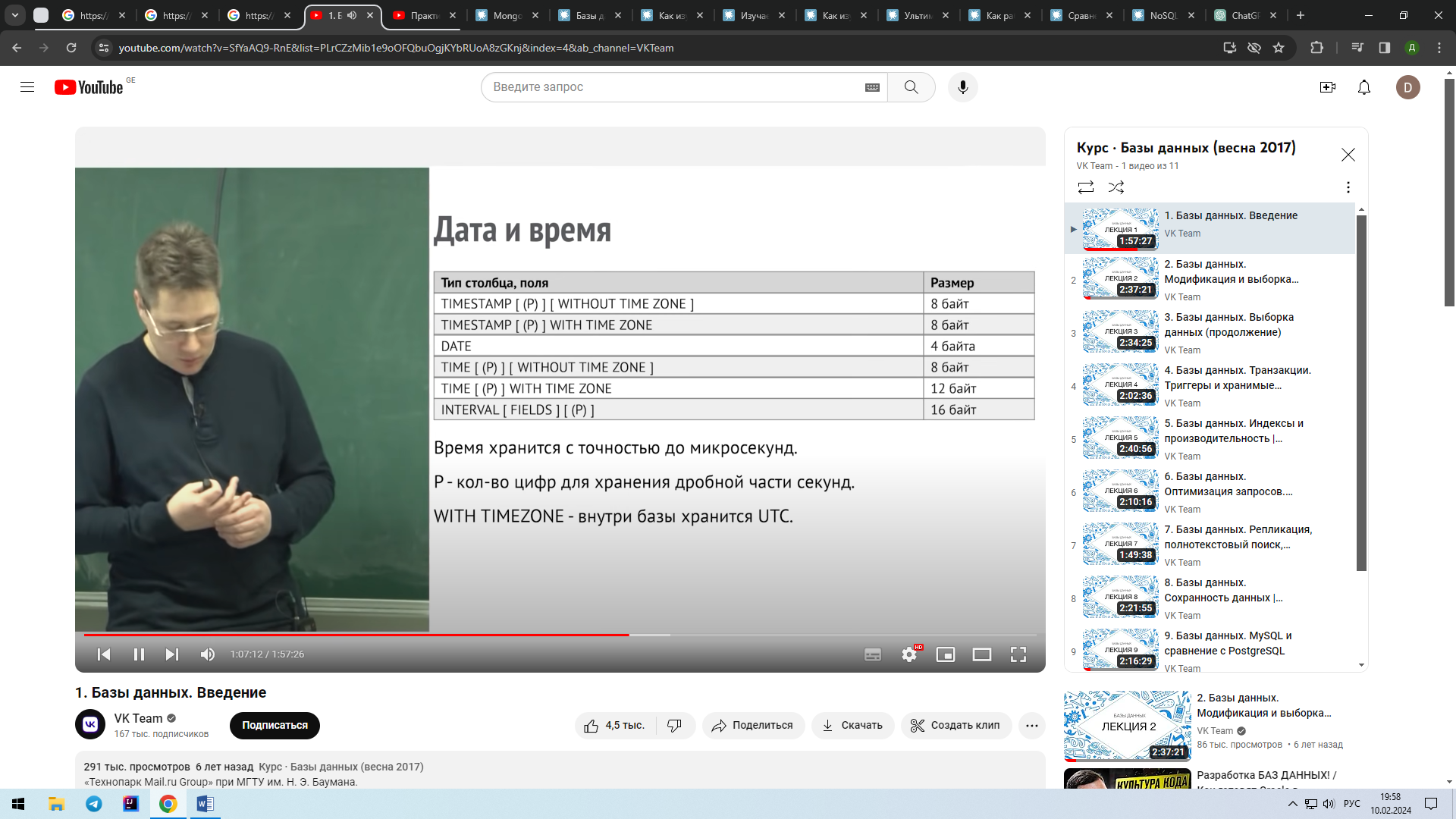


Bit(n) – фиксированная длина, если меньше – дописываются нули.

Varbit(n) – фиксированная длина, если меньше – нули не дописываются  


Сhar(n) – фикс длина, если меньше – дописываются пробелы

Varchar(n) – фикс длина, если меньше – пробелы не дописываются – более оптимальный вариант в большинстве случаев



Date хранит число, месяц год и время (часы, минуты, секунды)

Timestamp – Date + миллесекунды

Null – значение без типа или отсутствие значения

Поддерживает Arrays, JSON, XML, геометрические типы, custom-типы.

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

**Select** – команда для извлечения информации из бд.

Общий вид команды:

Select <Имя столбца> from <Имя таблицы> where<условия> order by;

*Distinct* – убирает дубликаты из результирующего набора строк.

SELECT DISTINCT column1 FROM table\_name; - выведет уникальные значения для column1

SELECT DISTINCT column1, column2 FROM table\_name; - Выведет уникальные комбинации значений для column1 и column2

SELECT column1 DISTINCT column2 FROM table\_name; - Такой синтаксис недопустим по причине возможного несоответствия кол-ва значений двух атрибутов (у первого выведутся все значения, включая неуникальные, у второго неуникальные отбрасываются, что ведет к несоответствию кол-ва строк и неопределенностью в заполнении строк данными)

*As* – алиас, используется для задания псевдонима столбца или таблицы в результирующем запросе (относится к получаемому выводу, сама таблица не меняется)

SELECT \* FROM students AS name\_alias; - для переименования таблицы.

SELECT id AS student\_id, name AS student\_name FROM students; - для переименования колонок

Оператор as можно опускать.

*Expressions* – результаты вывода для каждого аргумента можно изменять перед выводом посредством операций и выражений (изменяется вывод, но база не меняется).

Арифметические операции, операции сравнения, текстовые операции (конкатенация), логические (and, or, not), функции (sum, avg, count и тд) и др.

1) Арифметические операции - Подразумевают операции сложения, вычитания, умножения, деления, процентный остаток. Могут быть произведены как между аргументом и числовыми значениями, так и между аргументами.

2) Текстовые операции – Основная текстовая операция – конкатенация, с помощью оператора ||, помимо нее есть функции для работы со строками типа concat, upper, lower, length, substring, trim, left, right.

Операции сравнения, логические операции и null операции: Эти типы выражений используются в контексте оператора Where, с помощью которого задаются условия результирующей выборки значений.

3) Операции сравнения – стандартные (>, <, = и тд), null операции (is null, is not null) так же можно отнести к операциям сравнения, но так как любое сравнение с null дает false (даже с другим null) их вынесли отдельно. При сравнении текстовых значений используется лексикографический порядок (как в словаре).

4) Логические операции используются для комбинирования нескольких операций сравнения – and, or, not. and – вводит несколько условий для соответствия, or – вводит условия, одно из которых должно быть выполнено, not – исключает случаи, подходящие под условия (where not <условие>).

Оператор between так же задает промежуток значений. Является аналогом записи

Where аргумент between table\_1 >= значение and <= значение;

Оператор in(список значений) является аналогом оператора =, выводит значения, соответствующие одному из значений в списке, может быть заменен на where <колонка> = <значение> or <колонка> = <значение>

Like – оператор для поиска среди текстовых типов по заданному шаблону. Для поиска используются маски:

% - любое кол-во любых символов

\_ - один любой символ

[A-B] – диапазон допустимых символов, если [^A-B] или [!A-B] – отрицание этого диапазона

| - или

Escape – вводит символ для экранирования, например LIKE 'ST$\_%' ESCAPE '$'; - здесь \_ означает не оператор для like, а символ, который нужно найти

5) Функции

|  |  |
| --- | --- |
| *single-row* - кол-во аргументов соответствует кол-ву результатов | |
| Lower() |  |
| Upper() |  |
| Initcap() | первая буква каждого слова будет иметь верхний регистр, остальные нижний (разделителями выступают пробелы и спец символы). |
| Concat() | аналог ||, |
| Length() |  |
| Lpad(a, b, c)  Rpad(a, b, c) | используется для дополнения строки слева/справа символами до определенной длины (a – модифицируемая строка, b – длина результирующей строки, c – добавляемые символы) |
| Trim() | Удаляет заданный символ с начала - leading, конца – trailing или двух концов – both слова.  Trim (a b from c) – a – модификатор удаления, b – символ, который нужно удалить, c – откуда удалить. Если не задать модификатор, по умолчанию будет both, если не задать символ, то по умолчанию будет пробел. |
| Position() | Возвращает индекс первого совпадения с подстрокой для символа.  Position(‘substring’ in ‘string’ [from start\_position]) - substring – искомая строка, string – источник поиска, [from start\_position] – начиная с какого символа искать |
| Substring() | Возвращает подстроку из заданного диапазона.  Substr(a, b, c) – a – источник поиска, b – индекс начала поиска (если аргумент отрицательный - будет вестись отчет с конца), c – кол-во символов (необязательный аргумент)  Так же можно использовать синтаксис Substr(a from b for c) |
| Replace() | Заменяет искомые символы на заданные.  Replace(a, b, c) – a –источник для поиска, b – что заменяем c – на что заменяем.  Для удаления нужных символов c задается как ‘’; |
| Round() | Функция для округления значений.  Round(a, b) – a – число для округления, b – точность (кол-во цифр после запятой) |
| Trunk() | Отсекает значения, не входящие в диапазон точности (без округления)  Trunk(a, b) – число для «округления», b – точность (если отрицательное, то отсечется целая часть, но сохранятся разряды)  Пример: select trunc(123.45, -2); вывод: 100 |
| *multiple-row* - результат всегда один, независимо от кол-ва аргументов | |
| Count() | Null в учет не берется |
| Sum() |  |
| Avg() |  |
| Group by() |  |
| Max()/Min() |  |
| Nested group функции |  |
| Having |  |

Order by <аргумент> <asc / desc> <nulls first / last>– задает сортировку вывода. Asc, desc

Аргумент может быть именем аргумента, выражением, алиасом, номером аргумента в select

Несколько аргументов для сортировки – определяет более точный порядок сортировки. Если по первому аргументу значения повторяются, то они по сути никак не будут отсортированы и для таких случаев можно использовать другой аргумент, по которому они так же будут отсортированы

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

SELECT column\_name, data\_type, character\_maximum\_length

FROM information\_schema.columns

WHERE table\_name = 'students';

Запрос к системной таблице, которая хранит информацию обо всех таблицах

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_