Data cleaning

(1)

(a) There are 2 reason for using SQL to cleaning data?

(i) Data primarily stay in data base, so in 1st level if we do something it only possible with SQL.

(2nd) SQL is really fast with huge data.

These are the 2 reason we have know SQL for.

(b) in data wranging there are 3 process?

(1st) gathering data

(2nd) Assesing data

(3rd) Cleaning data

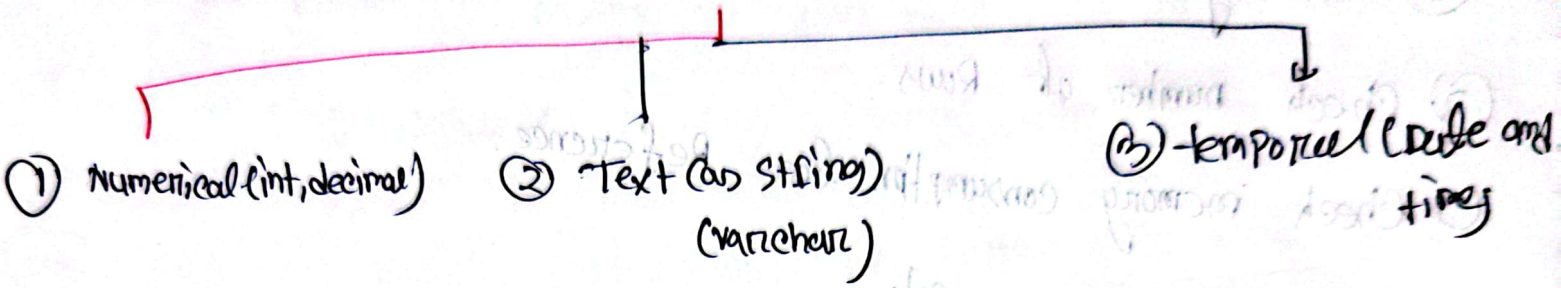
(a) Dat Assesing have 2 way to do?

(1) manually by scaning data with your eyes.

(2) through Programming

String Data type

Sql data types we can create 3 primary data types



String data type:

① CHAR: The data type is used to store fixed-length string.

④ if i make a char[10] (10 space will consume)

② if i store [Hello] in the char, although it seems, it will consume 5 space, since we specify [10] space, so it will consume 10 space by adding space to [Hello -----] like this.

④ length 255

③ we can't store more than 10 char, since we specify only up to 10 characters are allowed.

④ where we can use CHAR: when we want to store fixed number of data eg like { Phone numbers }

② Varchar :

① in varchare we specify a upper limit like (255)

if we store 1 word it will occupy 1 space

" " " 10 " " " " 10 "

it is actually dynamic

② varchar generally use for those column, we don't the length of the data. like: Name, email,

③ TEXT :

① This data type is used to store larger amounts of variable-length string data than Varchar.

② it can store up to 65,535 character.

③ used in text data, such as blog post or comments.

④ MEDIUMTEXT :-

① This data type is used to store even larger amount of text data than text.

② it can store up to 16,777,215 character.

③ generally used to in {long-form articles on legal documents} {privacy-polices, terms and conditions}

⑤ LONGTEXT :-

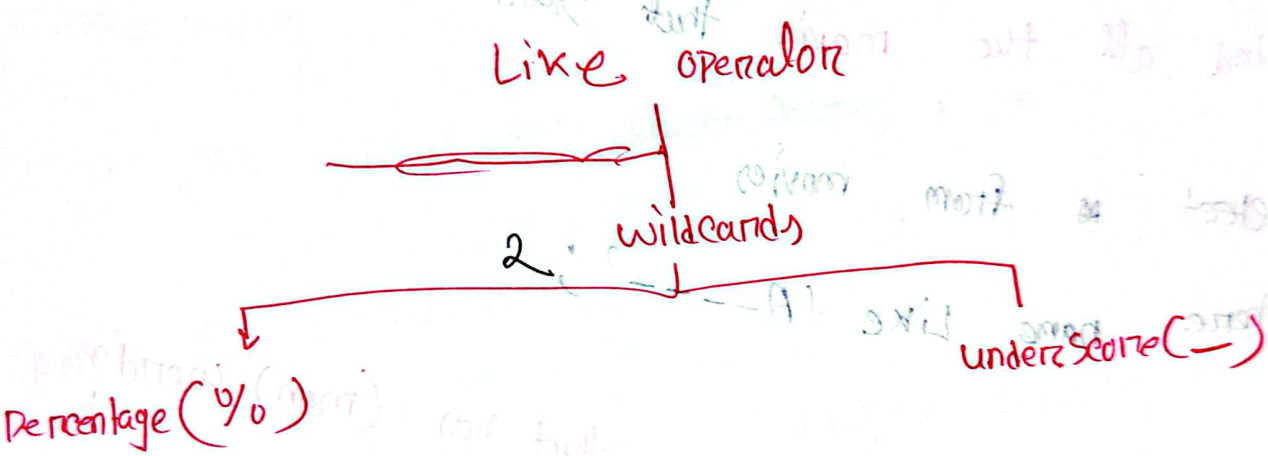
① This data type is used to store the largest amount of text data.

② it can store up to 4,294,967,295 char.

③ used in entire books or large collection of data.

wildcards

⊙ The Like operator in mysql is used to match a string value against a pattern using wildcard characters. it is commonly used in select, where and join clause to filter on join rows based on a pattern match



% this Represent zero, one or more character,

while (_) " " a single character

(40) find all the movie name, that has only 5 character?

```
Select * from movies
where name Like '_____';
```

{ one underscore Represent on word/char }

(41) find all the movie that start with (A and has 5 character)

```
Select * from movies
where name Like 'A____';
```

(42) find all the movie name that has (man) word?

```
select * from movies.
where name like '%man%';
```

(43) find all the movie name that start with (man) word?

```
Select * from movies.
where name like 'man%'
```

④ find all the movie that end with man word;

```
select * from movies
where name like ('%man');
```

string function

→ upper / lower operation

```
select name, upper(name), lower(name) from movies;
```

⇒ Concate and concat_ws (with separation)

⑦ I want to write movie name and direction name?

```
select concat(name, ' => ', director) from movies;
```

⑧ also adding star name

```
select concat(name, ' => ', director, ' => ', star) from
movies.
```

(#) in SQL there are 2 types of type conversion are there

↳ Implicet (Internally understand (10) this number need to convert into string inorder to concate)

↳ Explicet (using cast function)

Concat_ws (with seperator)

① Select concat_ws('=>', name, directon, star) from movies;

Substring → last 5 character

(#) i need to extract 5 character from movie name.?

start → end

Select name, substr(name, 1, 5) from movies;

(-0) if we don't Provide last end value, it will alltomatically select last index of string.

Select name, substr(name, 4) from movies;

⑩ we can even put negative value.

⑪ so if i want first character from last five character

```
select name, substr(name, -5, 1) from movies;
```

⑫ Replace string operation

① Select Replace("hello world", "world", 'india')
my string — the substring I want to replace — Replace world.

⑬ change the movie name that has a substring man change it with women.

```
select Replace(name, 'man', 'women') from movies.
```

⑭ Reverse function

⑮ find those movie that has a palindrome word;

```
select name from movies
where name = Reverse(name)
```

⑭ Char_length vs length

both are not same, the main difference between. CHAR_length and Length is that CHAR_length return the length of string. wher as Length Return the the lengths of a string in bytes.

① both are same return character length, but there is a subtle difference.

Café ⇒ if apply length then i will get 5

⇒ but " we apply char_length " " 4

② insert (str, position, length, newstr.

str: the original string to insert into.

position: the position at which to insert the new substring. The first position is 1.

length: the number of character to replace

newstr: the new substring to insert.

Select insert ("Hello world", 7, 0, "india") ⇒ Hello indiaworld

select insert ("Hello world", 7, 5, india) ⇒ Hello india

Left Right same as substring

The shining

Select name, Left(name, 3), Right(name, 3) from movies;

- Left ⇒ The. (first word three →)
- Right ⇒ ing (last word three ←)

Repeat operation

Select Repeat(name, 3) from movies;

Trim [ltrim and rtrim]

Trim → Remove space from first and last.

Select ('Trim(" nitish ");

⇒ nitish.

(#) we can also remove any character ?

Select Trim(Both '.' From "....nitish....");

⇒ nitish.

⑩ if want ~~all~~ to character .....shanto

select Trim (Leading "." from "....shanto")

=> shanto

⑪ if I want ~~from~~ character shanto ----

select Trim (Trailing "." from "shanto----")

=> shanto

⑫ LTRIM and RTrim only Remove space.

Left Right

LTRIM(" ---- shanto ----"); => shanto ----

RTRIM(" ---- shanto ----"); => ---- shanto.

Substring_index (Python split function) super important

① Select substring_index("www.campusx.in", ".", 1)

⇒ www

1 ↓ first occurance cuts the string result from

② Select substring_index("www.campusx.in", ".", 2)

⇒ ww.campusx.

③ substring_index("www.campusx.in", ".", -1) =

⇒ in.

④ Strcmp (String compaire)

the strcmp() function return an integer that indicates the relationship between the two string.

- if str1 is less than str2, the function return a negative integer.
- if str1 " greater " " " " " positive int.
- if same str1 = str2 then 0

select strcmp("mumbai", "Delhi") => 1

m আসলে আসে D থেকে।

(ii) locate (Hello, world)

=> locate actual told where the index of world start.

select locate("w", 'hello world') => 7.

(iii) LPAD and RPad

doing padding. suppose this is phone number 1811146290 (10) i want add +880, (4)

select LPAD('1811146290', 14, "+880")

select RPAD('1811146290', 14, "+880")