Cheatography

SQL Cheat Sheet

by Remidy08 via cheatography.com/159206/cs/33576/

Understanding SQL

Database:	container to store organized data
Database Management System (DBMS):	manipulates the database
Table:	structured list of data of a specific type. There cannot be repeated names for tables in the same database
Schema:	Information about database and table layout and properties
Datatype:	A type of allowed dat in a certain column
Primary Keys:	A column whose values uniquely identify every row in a table. They are not mandatory but most people that create a database use them. These should never be updated or reused

Filtering Data

WHERE	specified right after the table name (before ORDER BY. It is used to filter the data
Operators	used in the WHERE clause. They can be: = (equality); <> or != (nonequality), < (less than), <= (less than or equal), !< (not less than), > (greater than), >= (greater or equal than), !> (not greater than), BETWEEN, IS NULL
AND	used to append conditions to the WHERE clause.

Filtering Data (cont)

OR	instructs the database management system to retrieve rows that match either condition.
IN	used to specify a range of conditions, any of which can be matched. It takes a comma-delimited list of valid values, all enclosed within parentheses.
NOT	can be used before the column to filter on, not just after it. Negates whatever condition comes next to it.

We can do the same with the IN and OR operators, but the IN has the advantage of being easier to read; is easier to use in conjuction with other AND and OR operators; In often executes more quickly; it allows to build subqueries.

Data Manipulation Functions

SUBSTR() (DB2, Oracle, PostgreSQL, and SQLITE) or SUBSTRING() (MariaDB, MySQL and SQL Server)	Extract part of a string
CAST() (DB2, PostgreSQL, SQL Server) or CONVERT() (MariaDB, MySQL, and SQL Server [appears in both])	Data type conversion. Oracle has multiple functions, one for each type
CURRET_DATE (DB2 and PostgreSQL) or CURDATE() (MariaDB and MySQL) or SYSDATE (Oracle) or GETDATE() (SQL Server) or DATE() (SQLite)	Get current date
UPPER()	converts text to uppercase

Data Manipulation Functions (cont)

LEFT()	returns characters from the left of a string
LENGTH() or DATALENGTH() or LEN()	returns the length of a string
LOWER()	converts a string to lower case
RIGHT()	returns characters from the right of a string
SOUNDEX() (PostgreSQL)	returns a string's SOUNDEX value, like the name says, it returns strings with simmilar sounds
DATEPART(yy, column)) (SQL Server) or DATE_PART('year', column) (PostgreSQL))	returns the part of the date that we want to use
EXTRACT(year FROM column)	extracts part of the date with year specifying what part of the date to extract
to_date(date, 'yyyy-mm-dd')	converts strings into dates. It can be used in a BETWEEN statement
YEAR() (DB2, MySQL, and MariaDB)	extracts the year from date.
MONTH() (DB2, MySQL, and MariaDB)	extracts the month from date.
DAY() (DB2, MySQL, and MariaDB)	extracts the day from date.
strftime('%Y', column)	extracts part of a date.
ABS()	returns a number's absolute value

By **Remidy08**
cheatography.com/remidy08/

Not published yet.
Last updated 12th September, 2022.

Data Manipulation Functions (cont)

COS()	returns the trignometric cosine of a specific angle
EXP()	returns the trignometric exponential value of a specific number
PI()	returns the value of pi
SIN()	returns the trignometric sine of a specific angle
SQRT()	returns the trignometric root of a specific number
TAN()	returns the trignometric tangent of a specific angle

SQL functions are not portable, which means they vary between DBMS.
Write comments near functions.

Data sets

UC Irvine Machine Learning Repository

Kaggle datasets

Amazon's AWS datasets

http://dataportals.org/

http://opendatamonitor.eu/

http://quandl.com/

Wikipedia's list of Machine Learning datasets

Quora.com question

Datasets subreddit

Retrieving Data

SELECT	retrieves a specified set of elements (case insensitive). A different number of columns can be called, we just have to write them and separate with ','. If we want all columns we just need to specify '*'

Retrieving Data (cont)

FROM	refers the table we are retrieving the data from
;	used to separate statements
DISTINCT	added just before the column name (it applies to all columns combinations of unique values). It is used when we want a value to appear only once in the output
TOP	Used in Microsoft SQL server to pass how many items, counting from the top, we want to show. Example: SELECT TOP n column FROM table
FETCH FIRST n ROWS ONLT	Used in DB2 to pass how many items, counting from the top, we want to show. It is placed after the table
ROWNUM	Used in Oracle to pass how many items, counting from the top, we want to show. It is placed as if it was a WHERE statement. Example: WHERE ROWNUM <=5
LIMIT	Used in MySQL, MariaDB, PostgreSQL, and SQLite to pass how many items, counting from the top, we want to show. Placed after the table argument with a number next to it.

Retrieving Data (cont)

OFFSET	If we use LIMIT, after we pass it, we can write this argument o specify that we want the next n rows after the previously specified ones. Instead of this, we can use LIMIT m,n, where n referes to the first rows and m to the OFFSET argument
Coments	To create a comment we either use '--', '#', or / *(...)* /, this last one is used for multiline comments.

The first row in a table is row 0 not 1.

Creating Calculated Fields

+ (SQL Server) or || (DB2, Oracle, PostgreSQL, SQLite) or CONCAT() (MySQL, MariaDB)	Used to concatenate/join columns.
RTRIM()	removes white spaces on the right of a column.
LTRIM()	removes white spaces on the left of a column.
TRIM()	removes white spaces on the right and left of a column.

Creating Calculated Fields (cont)

Alias	alternate name for a field value. To do this, we need to place an AS after the calculated field with the pretended name after it. If the alias has more than one word in it, its name should be inclose in quotes (this practice is discoraged)
Curdate()	returns the current date (MySQL and MariaDB)

Calculated fields can include the sum or mutiplication of two columns, such as, column1 * column2.

Grouping Data

GROUP BY	instructs the DBMS to sort the data and group by a certain column. More than when columns can be usesd in this clause. Instead of passing the columns name, we can pass their position
HAVING	filters which groups to include. All the techniques learned with WHERE applies to HAVING as well.

Every expression specified in the select has to be specified in the GROUP BY.
Most SQL implementations do not allow GROUP BY columns with variable length.
NULL can be returned as a group.
The GROUP BY comes before OERDER BY and after WHERE clauses.
Aliases cannot be used.

Working with Subqueries

Query	Any SQL statement, but the term is used to refer to a SELECT statement.
Fully Qualified column names	When we precede the name of a column with the name of the table followed by a '.'. Ex.: table.column
Subquery	This name is normally atributed to a SELECT statement within another SELECT statement. This is most commonly done in a WHERE clause

Subquery SELECT statements can only retrieve a single column.

Joining Tables

SELECT ... FROM column1, column 2	the number of rows retrieved will be the product of the number of rows in each table Cartesian product or cross join).
WHERE	in this case the condition passed into this clause should be the column we want to match in both tables.
INNER JOIN ... ON	used to join tables. We put the columns we want to join, one on each side of the INNER JOIN, with the condition after the ON.

The limit of the maximum number of tables in a join should be accessed in the DBMS documentation.

Sorting Retrieved Data

Clause	usually consists of a keyword and suplied data
ORDER BY	Be sure it is the last clause in the SELECT statement with a column in front of it to mention in which order we should organize the table. It is not mandatory to select the column by which we order the table. Instead of using a column name, we could use its position
DESC or DESCENDING	Added after the column in order by to make the order descending, instead of ascending. The DESC only applies to the column that preceedes it
ASC or ASCEDING	It is the default value of the ORDER BY, does the opposite of the previous one

ORDER BY is case insenstive, so letters like A and a, come in the same order. In some case, if there are foreign characters in the data set, it may be necessary for the data base administrator to change this behavior. By doing this, it is impossible to organize the data like you want, with a simple ORDER BY.

Using Wildcard Filtering

Wildcards	S pecial character used to match parts of a value
LIKE	to use wildcards in search clauses, you must use this operator. To use place it after a column refered in a WHERE cluase with a search pattern in front of it.

Using Wildcard Filtering (cont)

Predicate	expression that evaluates to TRUE, FALSE, or UNKNOWN. Predicates are used in the search condition of WHERE clauses and HAVING clauses, the join conditions of FROM clauses, and other constructs where a Boolean value is required. LIKE is considered a predicate
%	match any number of occurrences (including 0)of any character. Basically, it substitutes any type and number of characters. However, it does not match NULL.
_	it matches a single character. It is not supported by DB2.
[]	used to specify a set of characters, any of which must match a character in the specified position. Sets are not supportes in MySQL, Oracle, DB2, and SQLite
^	negates the meaning of a wildcard. For example, '[^JM]%'.

These types of searches may be case sensitive depending on the DBMS.
Wildcards are rarely positioned in the middle of a search pattern, but there is a situation no included in this case which is looking for email addresses
Some DBMS may add blank spaces to the end of each string in a cell, if this is the case in your DBMS, add % at the end of each search pattern.
Tips:
- Don't overuse wildcards
- Try not to use wildcards at the beggining of the search pattern, it turns it very slow

Summarizing Data

Aggregate functions	functions that operate on a set of rows to calculate and return a single value.
AVG(column)	returns a column's average value. NULL values are ignored by this function.
COUNT(column)	returns the number of row in a column. COUNT(*) to count the number of rows in a table. COUNT(column) count the number of rows which have a value, thus ignoring NULL values.
MAX(column)	returns a comlumn's highest value. It ignores NULL values.
MIN(column)	returns the sum if a column's value. It ignores NULL values.
SUM(column)	returns the sum of a column's values. It ignores NULL values..
TOP (only applies to some DBMSs)	lets you perform calculations on subsets of query results.

Summarizing Data (cont)

TOP PERCENT (only applies to some DBMSs)	lets you perform calculations on subsets of query results.

To calculate multiple averages, we have to use multiple AVG().
In some DBMSs, MAX()/MIN() can be used with multiple columns, in this case, it will return the highest/lowest value of all columns.
We can pass DISTINCT, in between the paretheses, on these functions so we only apply them to distinct values. The DISTINCT can only be used with _COUNT when a column name is specified.