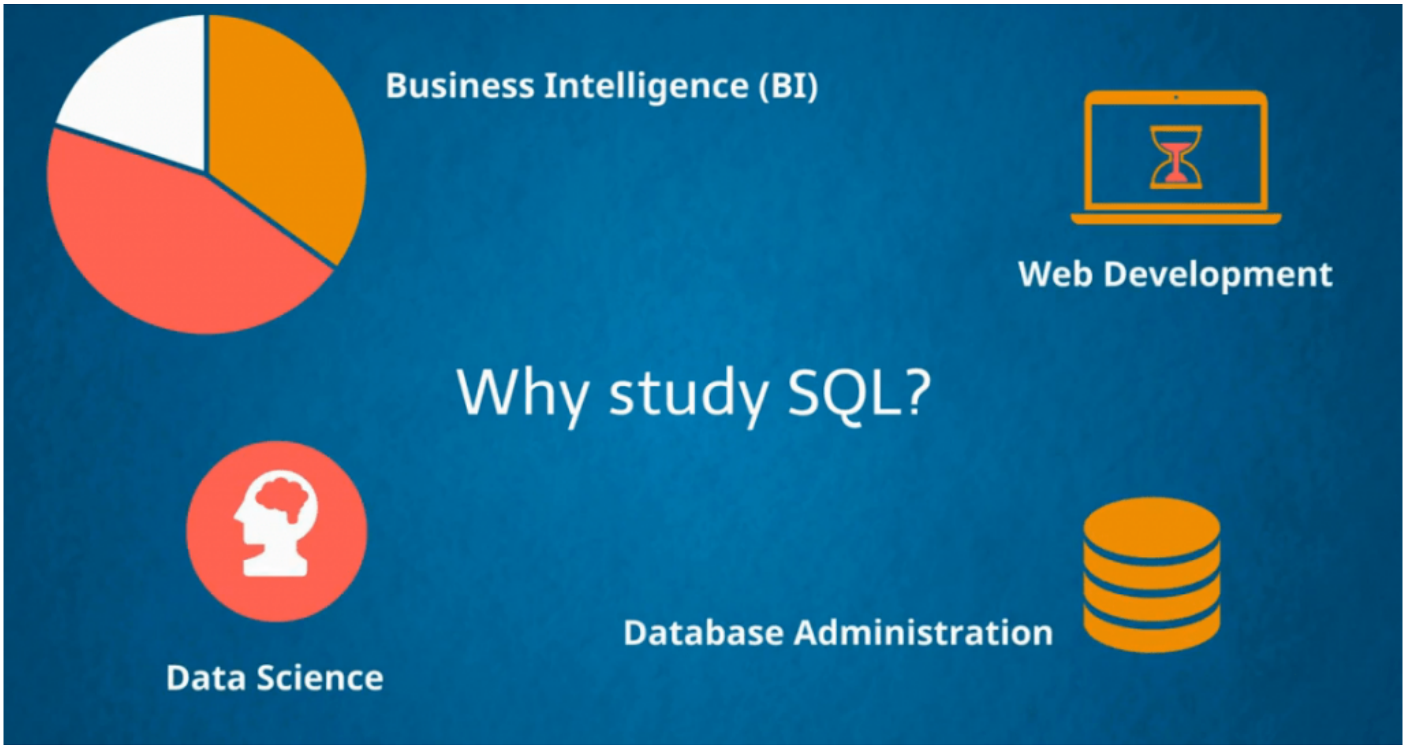
Road to Data Science in 50 Days - Day 6

Databases for Data Science



In the past 5 days, we have covered most of the statistics prerequisite for Data Science. However, if you are completely new to the terms that were introduced in last 5 days, i would suggest to take a couple of weeks to understand and practise some problems for all the terms, and then catch up with this github repo.

From today onwards, we will learn the about the Databases and the importance of Database in Data Science.

Database/SQL is one of the most requested skills in Data Science but it is often neglected by many data science practitioner as most of them want to directly jump to the magical part of Data Science i.e. Modelling. Learning and mastering database / Cloud is a skill in itself and can bag you different roles to make your career in Data Science.

Introduction

Much of the world’s data resides in databases. SQL (or Structured Query Language) is a powerful language which is used for communicating with and extracting data from databases. A working knowledge of databases and SQL is a must if you want to become a data scientist

As an example, let’s consider car manufacturers. Each car manufacturer might have a database composed of many different tables (eg. one for each of the different car models fabricated). In each of these tables, will then be stored different metrics about each car model sales in different countries.

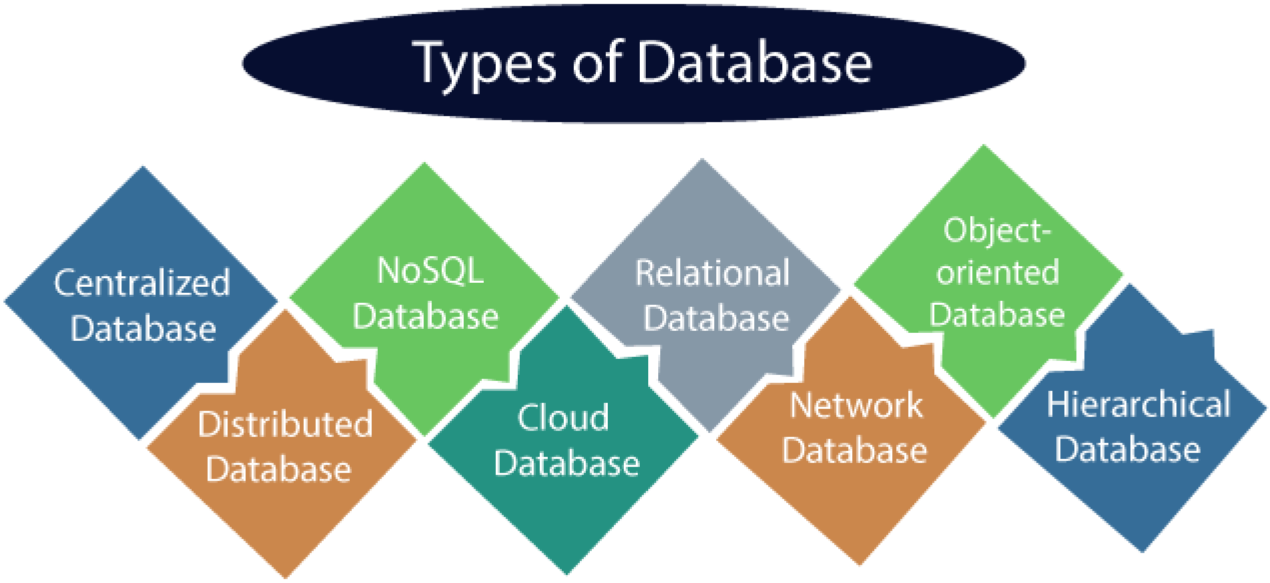
Together with Python and R, SQL is now considered to be one of the most requested skills in Data Science (Figure 1). Some of the reason why SQL is so requested nowadays are:

- About 2.5 quintillion bytes of data is generated every day. In order to store such large amounts of data, it is strictly necessary to make use of databases.
- Companies are now giving more and more importance to the value of data. Data can, for example, be used to: analyse and solve business problems, make predictions on market trends and understand customer needs.

One of the main advantages of using SQL, is that when performing operations with data, this is accessed directly (without any need to copy it beforehand). This can considerably speed up workflow executions.

But SQL is not the only database that is primarily used in Database management and Data Science.

Types of Database:



Depending upon the usage requirements, there are following types of databases available in the market −

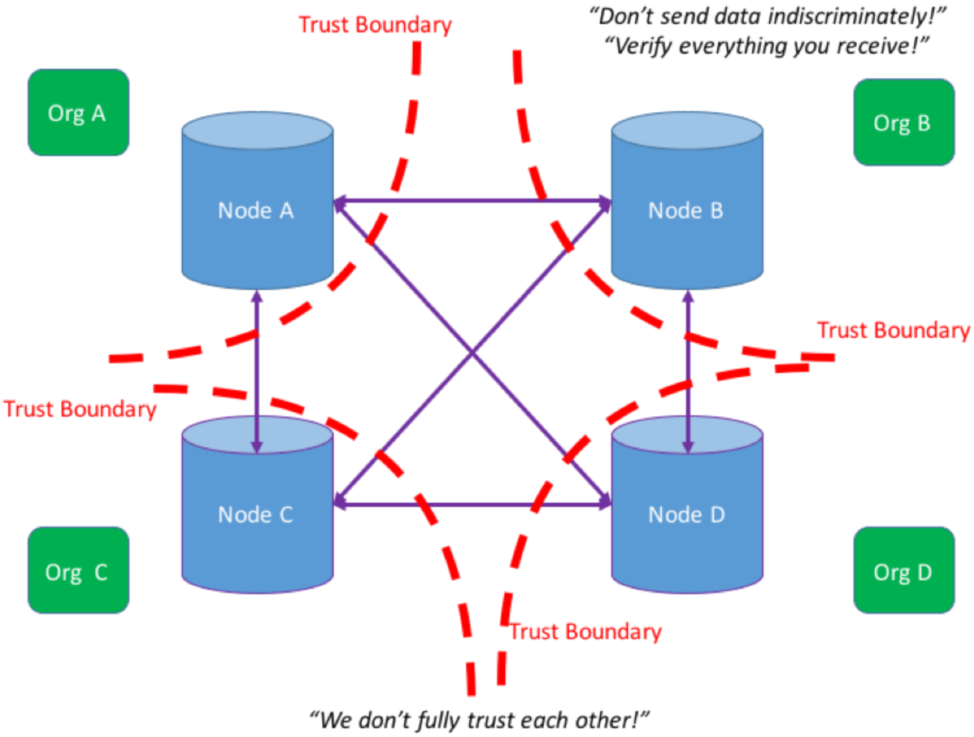
- Centralised database.
- Distributed database.
- Personal database.
- End-user database.
- Commercial database.
- NoSQL database.
- Operational database.
- Relational database.
- Cloud database.
- Object-oriented database.
- Graph database.

1. Centralised Database

The information(data) is stored at a centralized location and the users from different locations can access this data. This type of database contains application procedures that help the users to access the data even from a remote location.

Various kinds of authentication procedures are applied for the verification and validation of end users, likewise, a registration number is provided by the application procedures which keeps a track and record of data usage. The local area office handles this thing.

2. Distributed Database


Trust Boundary
"Don't send data indiscriminately!"
"Verify everything you receive!"
Org A
Org B
Node A
Node B
Trust Boundary
Trust Boundary
Node C
Node D
Org C
Org D
Trust Boundary
"We don't fully trust each other!"

Just opposite of the centralized database concept, the distributed database has contributions from the common database as well as the information captured by local computers also. The data is not at one place and is distributed at various sites of an organization. These sites are connected to each other with the help of communication links which helps them to access the distributed data easily.

You can imagine a distributed database as a one in which various portions of a database are stored in multiple different locations(physical) along with the application procedures which are replicated and distributed among various points in a network.

There are two kinds of distributed database, viz. homogenous and heterogeneous. The databases which have same underlying hardware and run over same operating systems and application procedures are known as homogeneous DDB, for eg. All physical locations in a DDB. Whereas, the operating systems, underlying hardware as well as application procedures can be different at various sites of a DDB which is known as heterogeneous DDB.

3. Personal Database

Data is collected and stored on personal computers which is small and easily manageable. The data is generally used by the same department of an organization and is accessed by a small group of people.

4. End User Database

The end user is usually not concerned about the transaction or operations done at various levels and is only aware of the product which may be a software or an application. Therefore, this is a shared database which is specifically designed for the end user, just like different levels' managers. Summary of whole information is collected in this database.

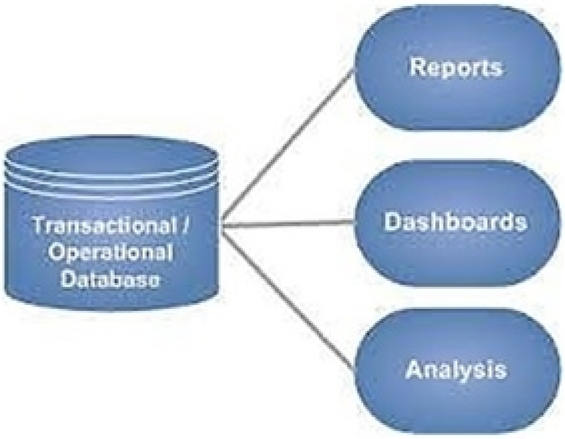
5. Commercial Database

These are the paid versions of the huge databases designed uniquely for the users who want to access the information for help. These databases are subject specific, and one cannot afford to maintain such a huge information. Access to such databases is provided through commercial links.

6. NoSQL Database

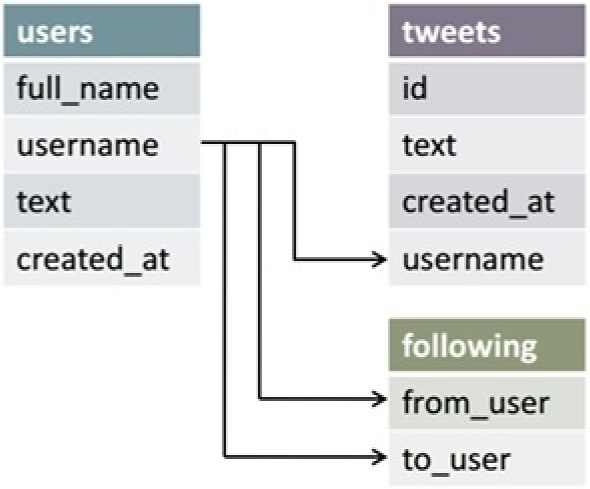
These are used for large sets of distributed data. There are some big data performance issues which are effectively handled by relational databases, such kind of issues are easily managed by NoSQL databases. There are very efficient in analyzing large size unstructured data that may be stored at multiple virtual servers of the cloud.

7.Operational Database



Information related to operations of an enterprise is stored inside this database. Functional lines like marketing, employee relations, customer service etc. require such kind of databases.

8. Relational Databases



These databases are categorized by a set of tables where data gets fit into a pre-defined category. The table consists of rows and columns where the column has an entry for data for a specific category and rows contains instance for that data defined according to the category. The Structured Query Language (SQL) is the standard user and application program interface for a relational database.

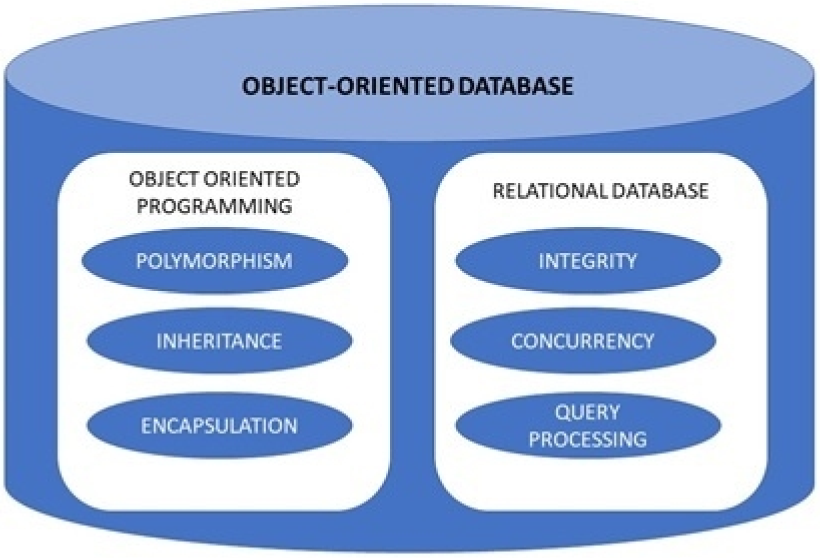
There are various simple operations that can be applied over the table which makes these databases easier to extend, join two databases with a common relation and modify all existing applications.

9. Cloud Databases

Now a day, data has been specifically getting stored over clouds also known as a virtual environment, either in a hybrid cloud, public or private cloud. A cloud database is a database that has been optimized or built for such a virtualized environment. There are various benefits of a cloud database, some of which are the ability to pay for storage capacity and bandwidth on a per-user basis, and they provide scalability on demand, along with high availability.

A cloud database also gives enterprises the opportunity to support business applications in a software-as-a-service deployment.

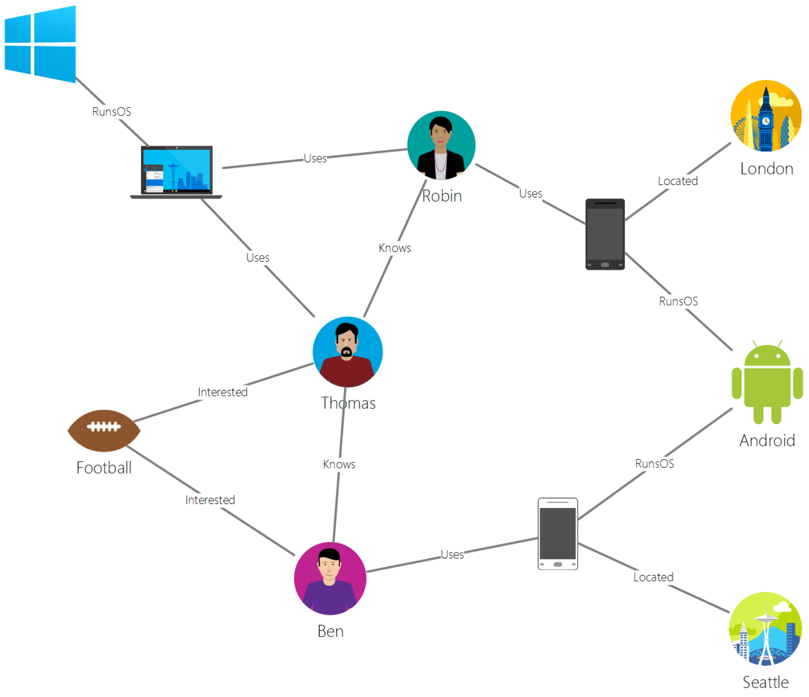
10. Object-Oriented Databases



An object-oriented database is a collection of object-oriented programming and relational database. There are various items which are created using object-oriented programming languages like C++, Java which can be stored in relational databases, but object-oriented databases are well-suited for those items.

An object-oriented database is organized around objects rather than actions, and data rather than logic. For example, a multimedia record in a relational database can be a definable data object, as opposed to an alphanumeric value.

11. Graph Databases



The graph is a collection of nodes and edges where each node is used to represent an entity and each edge

Importance of SQL in Data Science

Data Science is the study and analysis of data. In order to analyze the data, we need to extract it from the database. This is where SQL comes into the picture. Relational Database Management is an important part of Data Science. While many modern industries have geared their product management with NoSQL, SQL remains the ideal choice for many CRM, business intelligence tools and in office operations.

Many database platforms are modelled after SQL. This is because it has become a standard for many database systems. As a matter of fact, modern big data systems like Hadoop, Spark make use of SQL for maintaining relational database systems and processing structured data. While Hadoop provides features for batch SQL, Impala and Apache Drill provide interactive query capabilities.

On the other hand, Apache Spark uses the powerful in-memory SQL system to accelerate the processing of queries.

Furthermore, in order to become a data scientist, knowledge of SQL is a must. Many interview questions of Data Science start with SQL queries. Therefore, SQL is essential for Data Science. Therefore, from the above description, we conclude that:

- A Data Scientist needs SQL in order to handle structured data. This structured data is stored in relational databases. Therefore, in order to query these databases, a data scientist must have a sound knowledge of SQL.
- As a matter of fact, Big Data Platforms like Hadoop provides an extension for querying SQL commands for manipulating data through HiveQL.
- In order to experiment with data through the creation of test environments, data scientists make use of SQL as their standard tool.
- In order to carry out data analytics with the data that is stored in relational databases like Oracle, Microsoft SQL, MySQL, we need SQL.
- SQL is also essential for carrying out data wrangling and preparation. Therefore, when dealing with various Big Data tools, you will make use of SQL.

What SQL Skills are required for Data Science?

The aspiring Data Scientists must have the following necessary SQL skills:

1. Knowledge of Relational Database Model

A Relational Database Model System (RDBMS) is the primary and foremost necessary concept for an aspiring Data Scientist. In order to store structured data, you must know RDBMS in-depth. You can then access, retrieve and manipulate the data through SQL. An RDBMS is a standard for every data platform. Even the advanced big data platforms consist of an RDBMS section for processing structured information.

2. Knowledge of the SQL commands

A Data Scientist must know these following SQL commands –

- Data Query Language
- Data Manipulation Language

- Data Definition Language
- Data Control Language

3. Null Value

Null is used to represent a missing value. A field that contains Null value is blank in a table. However, a Null value is different than a zero value or a field that contains blank spaces.

4. Indexes

With the help of special lookup tables, a database search engine can locate values in a row easily. With SQL indexing, we can quickly load the data into the database.

5. Joins

Table joins are the most important concepts of relational databases that a data scientist must know. There are two types of joins – Inner Join and Outer Join. They are then further divided into Inner, Left, Right, Full etc.

6. Primary & Foreign Key

A primary key represents unique values in a database. With the help of a primary key, we are able to distinguish each line and record from the database. A Foreign Key, on the other hand, is used to connect two tables together.

7. SubQuery

A subquery is the nested query that is embedded in another query. There are four important subqueries in SQL – SELECT, INSERT, UPDATE and DELETE. It will return the information to the primary query.

8. Creating Tables

Data Science makes use of organized relational tables, and therefore, it is necessary to know how to create tables in SQL.

SQL

SQL is followed by a unique set of rules and guidelines called Syntax.

All the SQL statements start with any of the keywords like SELECT, INSERT, UPDATE, DELETE, ALTER, DROP, CREATE, USE, SHOW and all the statements end with a semicolon (;).

The most important point to be noted here is that SQL is case insensitive, which means SELECT and select have same meaning in SQL statements. Whereas, MySQL makes difference in table names. So, if you are working with MySQL, then you need to give table names as they exist in the database.

Various Syntax in SQL

SQL SELECT Statement

```
SELECT column1, column2....columnN
FROM   table_name;
```

SQL DISTINCT Clause

```
SELECT DISTINCT column1, column2....columnN
FROM   table_name;
```

SQL WHERE Clause

```
SELECT column1, column2....columnN
FROM   table_name
WHERE  CONDITION;
```

SQL AND/OR Clause

```
SELECT column1, column2....columnN
FROM   table_name
WHERE  CONDITION-1 {AND|OR} CONDITION-2;
```

SQL IN Clause

```
SELECT column1, column2....columnN
FROM   table_name
WHERE  column_name IN (val-1, val-2,...val-N);
```

SQL BETWEEN Clause

```
SELECT column1, column2....columnN
FROM   table_name
WHERE  column_name BETWEEN val-1 AND val-2;
```

SQL LIKE Clause

```
SELECT column1, column2....columnN
FROM   table_name
WHERE  column_name LIKE { PATTERN };
```

QL ORDER BY Clause

```
SELECT column1, column2....columnN
FROM   table_name
WHERE  CONDITION
ORDER BY column_name {ASC|DESC};
```

SQL GROUP BY Clause

```
SELECT SUM(column_name)
FROM   table_name
WHERE  CONDITION
GROUP BY column_name;
```

SQL COUNT Clause

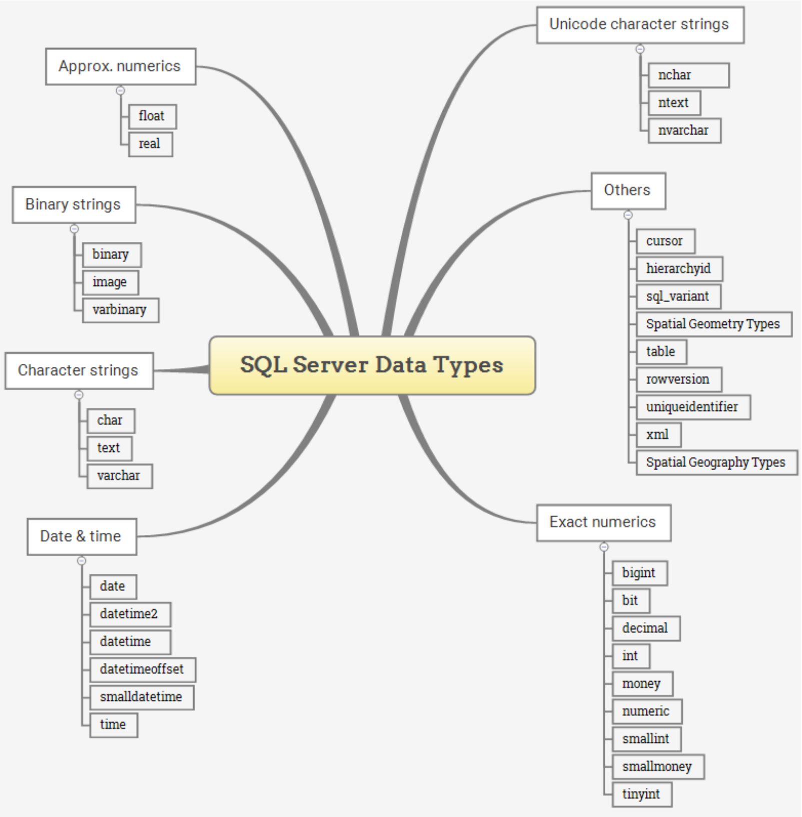
```
SELECT COUNT(column_name)
FROM   table_name
WHERE  CONDITION;
```

SQL HAVING Clause

```
SELECT SUM(column_name)
FROM   table_name
WHERE  CONDITION
GROUP BY column_name
HAVING (arithematic function condition);
```

SQL - DataTypes

SQL Data Type is an attribute that specifies the type of data of any object. Each column, variable and expression has a related data type in SQL. You can use these data types while creating your tables. You can choose a data type for a table column based on your requirement.



SQL Server offers six categories of data types for your use which are listed below −

Exact Numeric Data Types

- int
- bigint
- smallint
- tinyint
- bit
- decimal
- numeric
- money
- smallmoney

Approximate Numeric Data Types

- float
- real

Date and Time Data Types

- Datetinme
- smalldatetime
- date
- time

Character Strings Data Types

- char
- varchar
- varchar(max)
- text

Unicode Character Strings Data Types

- nchar
- nvarchar
- nvarchar(max)
- ntext

Binary Data Types

- binary
- varbinary
- varbinary(max)
- image

Misc Data Types

- sql_variant
- timestamp
- uniqueidentifier
- xml
- cursor
- table

To learn about these data types in details, refer: https://www.tutorialspoint.com/sql/sql-data-types.htm (https://www.tutorialspoint.com/sql/sql-data-types.htm)

SQL - Operators

What is an Operator in SQL?

An operator is a reserved word or a character used primarily in an SQL statement's WHERE clause to perform operation(s), such as comparisons and arithmetic operations. These Operators are used to specify conditions in an SQL statement and to serve as conjunctions for multiple conditions in a statement.

- Arithmetic operators
- Comparison operators
- Logical operators
- Operators used to negate conditions

SQL has following operators:

SQL Arithmetic Operators

- **+** (Addition) - Adds values on either side of the operator.
- **-** (Subtraction) - Subtracts right hand operand from left hand operand.
- ***** (Multiplication) - Multiplies values on either side of the operator.
- **/** (Division) - Divides left hand operand by right hand operand.
- **%** (Modulus) - Divides left hand operand by right hand operand and returns remainder.

SQL Comparison Operators

- **=** Checks if the values of two operands are equal or not, if yes then condition becomes true.
- **!=** - Checks if the values of two operands are equal or not, if values are not equal then condition becomes true.
- **<>** - Checks if the values of two operands are equal or not, if values are not equal then condition becomes true.
- **>** - Checks if the value of left operand is greater than the value of right operand, if yes then condition becomes true.
- **<** - Checks if the value of left operand is less than the value of right operand, if yes then condition becomes true.
- **>=** - Checks if the value of left operand is greater than or equal to the value of right operand, if yes then condition becomes true.
- **<=** - Checks if the value of left operand is less than or equal to the value of right operand, if yes then condition becomes true.
- **!<** - Checks if the value of left operand is not less than the value of right operand, if yes then condition becomes true.
- **!>** - Checks if the value of left operand is not greater than the value of right operand, if yes then condition becomes true.

SQL Logical Operators

- **ALL** - The ALL operator is used to compare a value to all values in another value set
- **ANY** - The ANY operator is used to compare a value to any applicable value in the list as per the condition.

- **BETWEEN** - The BETWEEN operator is used to search for values that are within a set of values, given the minimum value and the maximum value.
- **EXISTS** - The EXISTS operator is used to search for the presence of a row in a specified table that meets a certain criterion.
- **IN** - The IN operator is used to compare a value to a list of literal values that have been specified.
- **LIKE** - The LIKE operator is used to compare a value to similar values using wildcard operators.
- **NOT** - The NOT operator reverses the meaning of the logical operator with which it is used. Eg: NOT EXISTS, NOT BETWEEN, NOT IN, etc. This is a negate operator.
- **OR** - The OR operator is used to combine multiple conditions in an SQL statement's WHERE clause.
- **IS NULL** - The NULL operator is used to compare a value with a NULL value.
- **UNIQUE** - The UNIQUE operator searches every row of a specified table for uniqueness (no duplicates).

For the next couple of days we will perform create a database and perform some queries and understand the applications of SQL in Data Science

References and Links to Follow:

Top Databases used in Data Science - https://analyticsindiamag.com/top-databases-used-in-machine-learning-projects/ (https://analyticsindiamag.com/top-databases-used-in-machine-learning-projects/)

Types of Databases - https://www.tutorialspoint.com/Types-of-databases (https://www.tutorialspoint.com/Types-of-databases)

SQL Tutorial - https://www.tutorialspoint.com/sql/sql_tutorial.pdf (https://www.tutorialspoint.com/sql/sql_tutorial.pdf)

Video Tutorials:

SQL Basics for Beginers by edureka: https://www.youtube.com/watch?v=zbMHLJ0dY4w (https://www.youtube.com/watch?v=zbMHLJ0dY4w)

SQL Tutorials Full Database Course: https://www.youtube.com/watch?v=HXV3zeQKqGY (https://www.youtube.com/watch?v=HXV3zeQKqGY)