Introduction to SQL Databases

Introduction

While the role of a data scientist does involve fun stuff like building and improving machine learning models, there are many other necessary skills that must be mastered.

Data Cleaning/Preparation is a major task which nearly always must be done before getting to the work of analysis or modeling.

Data Wrangling/Munging or the process of transforming data from the raw form to a usable form or just in moving it from one place to another is another major task. Quite often, the data you will be working with will be stored in a relational database, and in order to access it, you will need to use Structured Query Language (SQL)

What is a Database?

- A systematic collection of data
- Supports **storage** and **manipulation** of data
- Access to a database is usually done through a Database Management System (DBMS) (eg. SQL Server, MySQL, Oracle)

Relational Databases / SQL Database / Transactional Database

- Most commonly used structure in enterprise scenarios
- Data is organized into one or more tables of columns and rows, with a unique key identifying each row.
- Tables are linked together through the use of keys, which ensure **referential integrity**
- Relational databases almost exclusively use SQL (Structured Query Language) to perform transactions.

Structured Query Language (SQL)

SQL is a programming language designed to manage data held in a relational database management system.

SQL comes in different flavors with slight variations on syntax and features depending on the type of RDBM you are working with:

- PostgreSQL
- Transact SQL (T-SQL) for Microsoft SQL Server
- SQLite
- MySQL
- PL-SQL for Oracle

Keywords

The **Structured** part of SQL refers to the format and keywords that make up the **Query.**

- SELECT
- FROM
- AS
- LIMIT
- DISTINCT
- COUNT
- WHERE
- AND
- OR
- BETWEEN
- IN
- (NOT) NULL
- LIKE
- AVG(), SUM(), MAX(), MIN(), etc.
- ORDER BY
- GROUP BY
- HAVING

Concepts

The **Keywords** allow you to perform operations such as:

- Selecting all columns from a table
- Selecting single columns from a table
- Aggregating data
- Finding unique values
- Slicing data (with multiple criteria)
- Selecting/avoiding null values
- Math
- Aliasing
- Organizing output results

Format of a Query