Databases for data analytics

https://github.com/evidencebp/databases-course/

Introduction

Goals

- Students will learn to use SQL for data science.
- Students will know how to build a database that will fit their business needs.
- Students will know how to evaluate and protect the data integrity.
- Students will learn basic performance for analytics.

Out of scope

- Database administration (e.g., physical layer, architectures), advanced database programming (e.g., transactions,concurrency), NOSQL databases, and much more.

Guidelines

- Course is “under construction”, let’s use this flexibility
- Learning a language requires a lot of training
  - Acing the course is easier if you come prepared
  - See https://github.com/evidencebp/databases-course/
- Ongoing project to practice and understand
- Understanding why is more important than how
- There are many answers to the same question. You should find the one the fits your needs and explain why.

SQL

- Basic Sql - select, data updates
- Joins
- Grouping and aggregation
- Use for data analytics
- Views

Database design

- Why database are needed in the first place?
- Entity Relations Diagram (ERD)
- Normalization
- Data Definition Language

Data integrity

- Reasons to data integrity problems
- Constraints
- Keys
- Triggers
- Validation

Introduction to performance

- Performance tools
- Indices
- Foreign key removal
- Denormalization
- Temporary storage
- Sampling and aggregations

Why databases are needed?

- For storing data, databases are not needed and csv and other text files are a common solution
- Database are needed for
- A common flexible interface language
- Handling large volumes (e.g., distribution)
- High performance
- Data integrity
- Security
- Multiple users support

Database Management Systems are big software

- The DBMS should store the data AND provide all the requirements
- SQLite, 29K commits
- MySql, 184K commits
- Oracle, 1977 funded, 465.19 billion USD, 159,000 employees company.
- Plenty DBMS exist (E.g. Sql Server, BigQuery, Postgres, Analytics DBs)
- Most DBMS use SQL as interface
- Even those that do not refer to SQL and called NoSQL DBs (e.g., key-value, graph).

Exercise 1

- Install MySql Workbench
- Install IMDB dataset
- Run “select count(*) from imdb_ijs.movies;” and see the you get 388,269
- No need to submit but please do it since it will be required for the next lessons.