# Hyrje në Inxhinierimin e të dhënave

Dr. Enida Sheme

# Përmbajtja e lëndës “Përpunimi i të dhënave me Python”

- Gjithsej 5 kredite = 57 ore
- Leksione 2 kredite
- Seminar 1 kredit, pjesëmarrja 75%
- Laborator 1 kredit, pjesëmarrja 100%
- Projekt 1 kredit, fiton / nuk fiton dhe vlerësim me deri në 20 pikë.

# Fjalët kyce mbi përmbajtjen e lëndës

- Data engineering (Inxhinierimi i të dhënave)
- Modelimi i të dhënave
- Big data
- Extract Transfer Load
- Gjuha e programimit Python dhe libraritë për përpunimin e të dhënave
- Modelet e Bazës së të Dhënave relacionale dhe jo-relacionale
- Data Warehouse (DWH)
- Teknologji Cloud për të ruajtur dhe përpunuar të dhënat

# Njohuritë paraprake

- Programim i orientuar nga objekti
- Programim web
- Algoritmikë
- Bazë të dhënash
- Sisteme të shpërndara

# Vlerësimi

- 20 pikë Projekt kursi
- 80 pikë Provimi përfundimtar = Teori dhe ushtrime nga materialet e leksioneve, seminareve dhe laboratoreve.

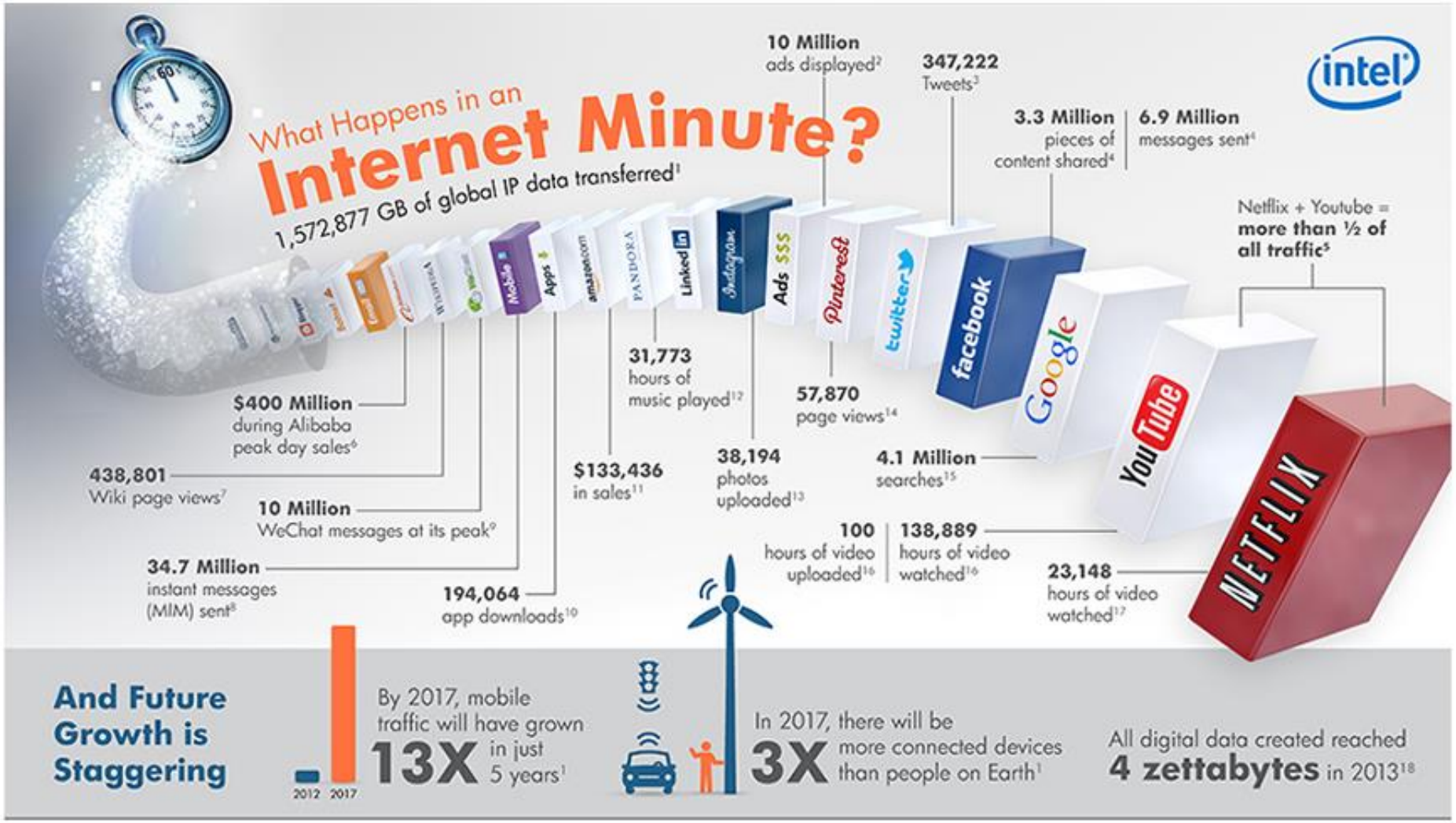
# Literatura për lëndën

- Leksione në formë slidesh nga Titullari i lëndës.
- Ralph Kimball, Margy Ross. “The data warehouse toolkit”
- Martin Kleppmann. “Designing Data-intensive applications”
- Paul Crickard. “Data Engineering with Python”
- Tom White. “Hadoop: the definitive guide”
- Luiz André Barroso, Urs Hölzle. “The Datacenter as a Computer: An Introduction to the Design of Warehouse-Scale Machines”

# Data: the new oil

- Matematikani britanik Clive Humby deklaroi në 2006 se “data is the new oil”
- Domethenia realiste është se të dhënat/data, ashtu si lënda djegëse, nuk është e dobishme në gjendjen e papërpunuar.
- Të dhënat duhen filtruar, përpunuar dhe të kthehen në dicka të përdorshme/ dobishme;
- Vlera e të dhënave qëndron në potencialin që ato mbartin për të përftuar “insight” dhe ndërtuar strategji.

# Statistika mbi gjenerimin dhe përdorimin e të dhënave



# Ekosistemi i të dhënave

- Ekosistemi i te dhenave i referohet gjuhes se programimit, paketave, algoritmeve, sherbimeve Cloud, dhe infrastrukturen e pergjithshme qe perdoret nga nje kompani per te grumbulluar, ruajtur, analizuar dhe perpunuar te dhenat.
- Data project lifecycle (fazat / perberesit e ekosistemit te te dhenave)

# Përbërësit e ekosistemit të të dhënave

1. Sensing

- Sensing refers to the process of identifying data sources for your project. It involves evaluating the quality of data so you can better understand whether it’s valuable. This evaluation includes asking such questions as:
- Is the data accurate?
- Is the data recent and up to date?
- Is the data complete?
- Is the data valid? Can it be trusted?
- Data can be sourced from internal sources, such as databases, spreadsheets, CRMs, and other software. It can also be sourced from external sources, such as websites or third-party data aggregators.

# Përbërësit e ekosistemit të të dhënave

2. Collection

- Once a potential data source has been identified, data must be collected.
- Data collection can be completed through manual or automated processes. Generally it isn't feasible to manually perform large-scale data collection. That's why data scientists use programming languages to write software designed to automate the data collection process.
- KEYS:
- Various programming languages: These include R, Python, SQL, and JavaScript
- Code packages and libraries: Existing code that's been written and tested and allows data scientists to generate programs more quickly and efficiently
- APIs: Software programs designed to interact with other applications and extract data

# Përbërësit e ekosistemit të të dhënave

3. Wrangling

- Data wrangling is a set of processes designed to transform raw data into a more usable format. Depending on the quality of the data in question, it may involve merging multiple datasets, identifying and filling gaps in data, deleting unnecessary or incorrect data, and “cleaning” and structuring data for future analysis.
- KEYS
- Algorithms, programming languages, data wrangling tools

# Përbërësit e ekosistemit të të dhënave

4. Analysis

- After raw data has been inspected and transformed into a readily usable state, it can be analyzed. Depending on the specific challenge your data project seeks to address, this analysis can be diagnostic, descriptive, predictive, or prescriptive. While each of these forms of analysis is unique, they rely on the same processes and tools.
- KEYS:
- Algorithms: A series of steps or rules to be followed to solve a problem (in this case, the analysis of various data points)
- Statistical models: Mathematical models used to investigate and interpret data
- Data visualization tools: These include Tableau, Microsoft BI, and Google Charts, which can generate graphical representations of data. Data visualization software may also have other functionality you can leverage.

# Përbërësit e ekosistemit të të dhënave

5. Storage

- Throughout all of the data life cycle stages, data must be stored in a way that’s both secure and accessible. The exact medium used for storage is dictated by your organization’s data governance procedures.
- Keys:
- Cloud-based storage solutions: These allow an organization to store data off-site and access it remotely
- On-site servers: These give organizations a greater sense of control over how data is stored and used
- Other storage media: These include hard drives, USB devices, CD-ROMs, and floppy disks

# Rëndësia e “data engineering” në ekosistemin e të dhënave

- Cdo perberes ndervepron me njeri tjetrin.
- Njohja e te dhenave e pergatit kompanine per sfidat e mundshme dhe zgjidhje sa me eficente per to.
- Shembull, diskutim.

# Koncepti “Big data”

- Koncept relativ

Diskutim:

- Extremely Large dataset
- That may be analyzed computationally
- To reveal patterns, trends, associations…

# 3 V-të e “Big data”

**Volume**

- Data larger than a single machine (CPU, RAM, disk)
- infrastructures and techniques that scale by using more machines

**Velocity**

- Endless stream of new events
- New data arrives continuously
- Data stream technologies

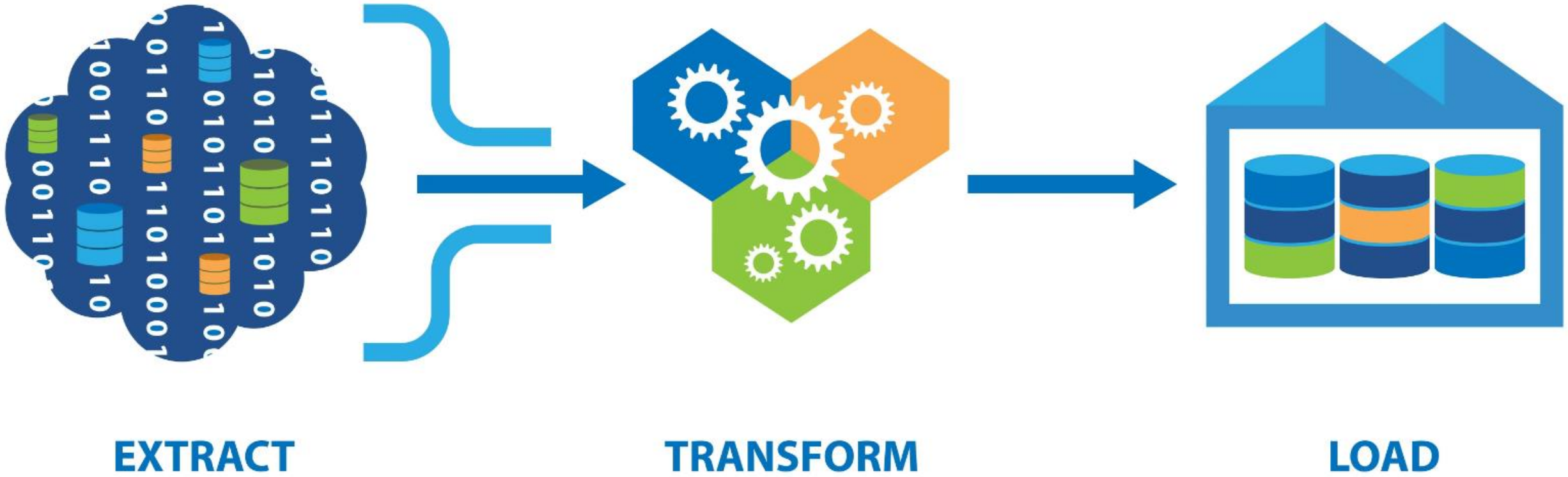
**Variety**

- dirty, incomplete, inconclusive data (e.g. text in tweets)
- Semantic and technical complications

# Dallimet midis:

- **Data science**: analizimi i te dhenave per te gjetur modele (patterns) dhe tendenca, me qellim parashikimin apo pritshmerite per rezultate te ardhme.
- **Data analysis**: analizimi i te dhenave per te kryer nje permbledhje te nje gjendje te shkuar apo aktuale ne format grafik, qe mund te interpretohet nga njeriu.
- **Data engineering**: pergatitja e nje zgjidhje per analizimin e te dhenave nga data scientist.

# ETL = Extract Transform Load



# Extract

- Shembull: Dua te di sa shitje ka patur nga cdo shites ne nje kompani shitjesh, ne nje periudhe 1 mujore, dhe sasia totale e te ardhurave te sjella nga shitesi me i mire.

Te dhenat

- Mijera/miliona rekorde
- Te grumbulluara nga baze te dhenash apo faqe internet

- Nderlidhja me bazen e te dhenave: SQL query, SELECT

# Transform

- Merge & combine nga disa datasete, si psh sistemi CRM, sistemi i anketimeve per feedback nga klientet, sistemi i shitjeve, etj.
- Fshirja e duplikatave
- Pastrimi i te dhenave nga te dhenat jo te sakta, te korruptuara, jo te plota (qe nuk mbartin informacion), etj

# Load

- Pas transformimit te te dhenave, rezultati ruhet ne format raporti, apo ngarkohet ne nje baze te dhenash ose eksportohet ne Cloud storage.

- Nderfaqja me bazen e te dhenave: SQL query INSERT

- Per forma me te nderlikuara ETL, cdo kompani e perpunimit te te dhenave ka mjetet e veta si: Oracle Data Integrator, IBM DataStage, Azure data factory, AWS, Google, Microsoft Power BI, etj

# Shembull praktik ETL

| Emer Shites | Sasia e produkteve # | Vlera totale EU |
|---|---|---|
| A. B | 132 | 2100 |
| C. D | 225 | 3450 |
| E. F | 339 | 4500 |
| J. K | 145 | 3550 |
| M. N | 199 | 4230 |
| ..... | ... | ... |

# Qartësia/Saktësia e të dhënave (how clean Is my data?)

- Rëndësia e të dhënave të pastra

- Eficence me e larte
- Saktesi me e larte
- Besueshmeri me e larte

# Python

- Pse Python është zgjedhja më popullore për detyra mbi “data engineering”?

Libraritë kryesore në Python:

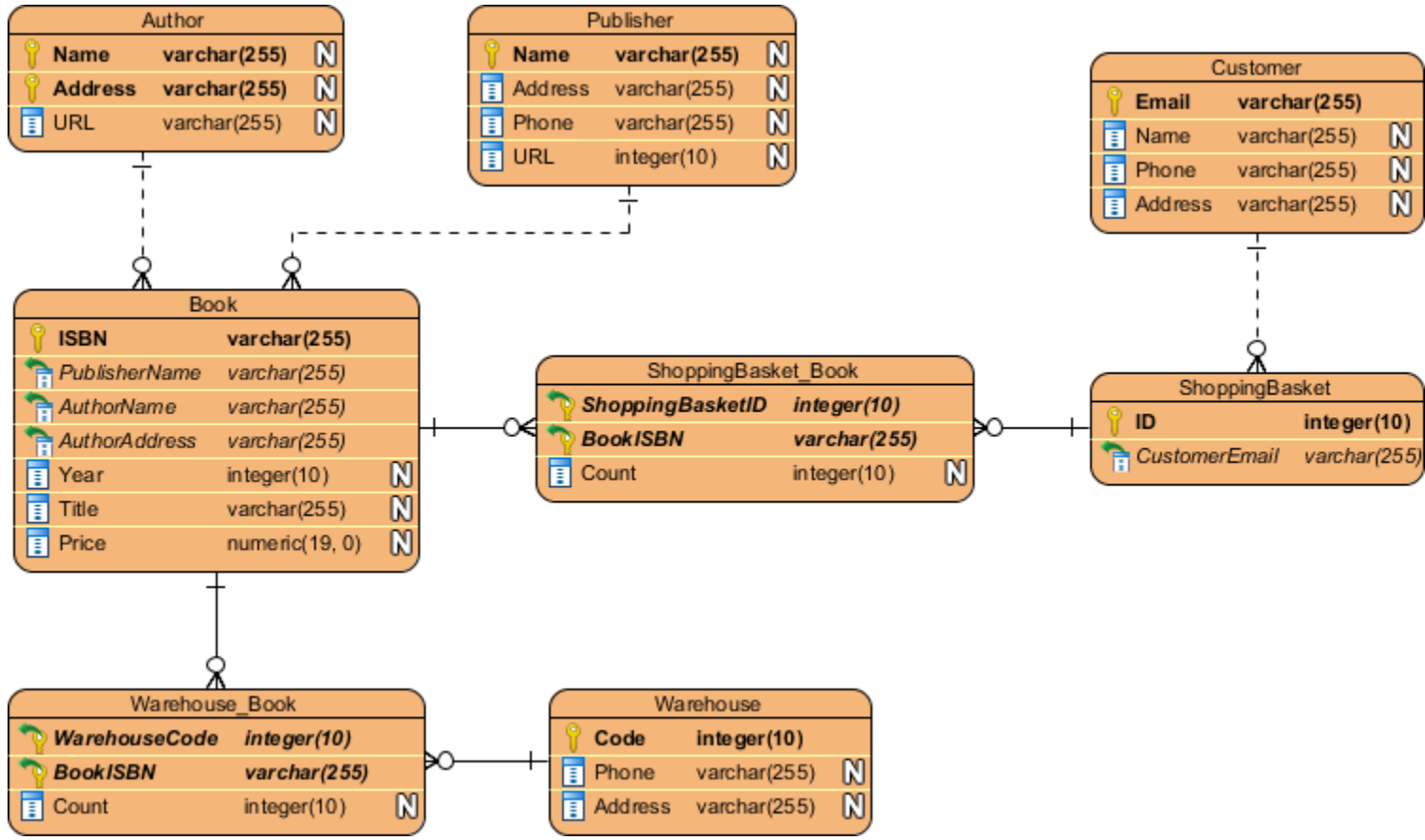
- NumPy
- Pandas
- Scikit

# Modelimi i të dhënave

- Pse është i rëndësishëm modelimi i të dhënave në inxhinierimin e të dhënave?

# Entity-Relationship Diagram

- Krijimi i nje modeli për të dhënat
- Identifikimi i entiteteve dhe atributeve
- Përcaktimi i marrëdhënieve
- Zbatimi i rregullave të Normalizimit



# Mjete për modelimin e të dhënave

ERD drawing software

- Draw.io
- Lucidchart

Database design tool

- MYSQL Workbench