Introduction to Data Engineering

1) What is Data Engineering:

- the modern data Ecosystem:
  - the Core concept is that data is a valuable asset
  - The ecosystem's purpose is to turn raw data into "actionable insights" for decision making.
  - this is achieved through a collaboration of several specialized roles.

- Key players in data Ecosystem:
  - data Engineer: build the highway for data the Architect
  - dat analyst: drives on the highway to rapport on traffic the Storyteller
  - data Scientist: use the highway to predict where new cities will be built ~~Predictive~~
  - Business Analyst: decides why we need the highway and where it should go for economic growth. translator

- data Engineer tasks:
  - collecting source data
  - processing data
  - storing data
  - making data Available to users securely.

2) the Data Ecosystem and languages for Data professionals:

- Datatypes & Sources:
  - structured: rigid Schema (DBtable, spread)
  - semi-structured: Flexib " with tags (Json, xml, emails)
  - unstructured: not defin schem (img, video, soci media posts)
  - *Source: rational/non rational DB APIs, web scraping, Datastreams (IoT, GPS)

- Data repositories (storage):
  - OLTP (transactional): For day to day operation (e.g banking transaction optimized for fast writes.
  - OLAP (Analytical): For complex analysis and reporting. optimized for fast reads on large Datasets.
    - * Data Wirehouse: central repository for structured cleaned data from multiple Sources.
    - * Data Mart: subsection of a data wirehouse
    - * Data lake: stores vast amounts of a raw data in its nativ format (struct-nonstruct-semi structured)

- Data Process & Pipeline's:

  entire process of moving data from source to destin

  - ETL: data transformed befor loaded (like Data wirehouses)
  - ELT: data loaded first (data lake) big data (clouds)

- Essential Languages: SQL, python R, shell scripting.

- comparison:

reposito types	Description & purpose	Key characteristics	exemple
DB	collecting data for input storage search, modificat	- relational RDBMS: table format, SQL, ACID compilant - non relational (No SQL) schemasless, scales easily	RDBMS: OLTP (mySQL PostgreSQL) NoSQL: Large, diverse fast changig data
DWirehouse or Data (mart)	structured	- optimized for OLAP - Analysis ready	Business Analy (Business intell & reporting) google big Query, Snowflace
DLake	massive amount of raw data (stru/unstur/semi)	- stores everything. - highly scalable.	AWS S3 Azure Data-Lake.
Big data	handle high Volum, Velocity, Variety.	- handles high Volum Velocity, Variety	Hadoop HDFS

- No SQL DB types:
  - Key-Value store (Redis, Dynamo DB).
  - Documented based (like JSON) (MongoDB, CouchDB)
  - column-based (Cassandra HBase.
  - Graph-based uses Nodes (data and edges (Neo4j, Cosmos DB).

RDMS (SQL) when have struct data and need ACID compliance.

NOSQL (not only) you have high Valume, Variety, or need high scalability and Flexibility (Not ACID)

- Data Gathering & Import methods:
  - SQL: extract data From relational databases (Select, Filter, Sort, Group..)
  - APIs: how applic programm access data From an endpoint (like web service or data marketplace).
  - Web Scraping: process Autom to download data from websites (text, contact info, images, .. from unstructured web pages.
  - Data streams: Handling continous real time data flows from sources (IoT, sensor, social media....).

- Data exchange: using formel platforms to securely share data.
- where import data?!
  - * structured: from relation DB
  - * unst/semi: Goes into NoSQL or Datalakes

- Data Wrangling (Data cleani) process to make raw Data usable it involves Exploration, transformation, Validation.

  Key tasks: → pandas-numpy
  - join: combin column from diff tab
  - Union: " rows from diff tables
  - Normalization: removing unused dat
  - Denormalization: combine data from multiple tables.
  - cleaning tasks: handle missing values, remove duplicate data, remove irrele data, Fix syntax error, standardiz data (make tex lower case) convert datatypes.

- Governance & Complince
  - ensure that Data secure privat trustworthy
  - adher laws & regulation