Introduction to Data Engineering

Duties of a Data Engineer

1. Build ETL/ELT pipelines to move data from sources to warehouses/lakes.
2. Ensure data accuracy, consistency, and security.
3. Work with analysts, scientists, and business teams to deliver usable data.

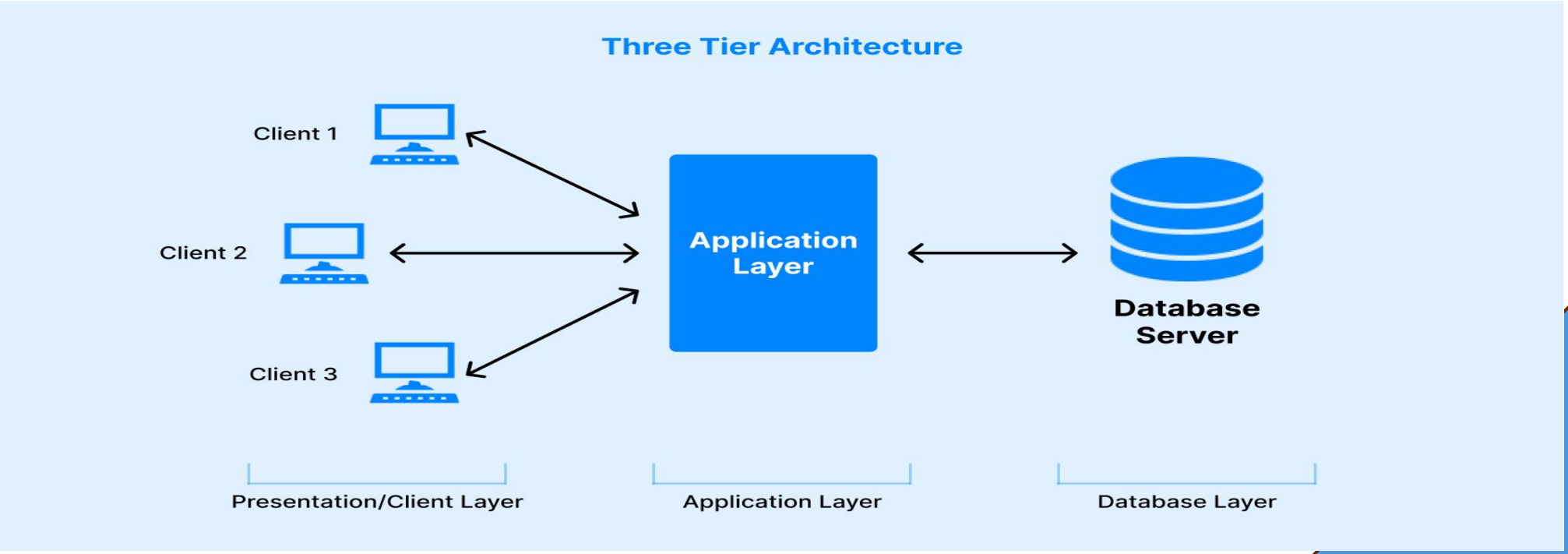
What's Expected of a Senior Data Engineer

A senior data engineer goes beyond building pipelines; they focus on performance, scalability, and cost efficiency.

1. Cost Optimization
2. Performance & Optimization
3. Architecture & Leadership

Database

A database is basically a container that holds several types of objects and data in an organized fashion. Generally, one database is used for a particular application or purpose, though this is not a hard and fast rule.



Common Types of Storage

1. **Relational Databases (RDBMS)**
    - Structure: Tables with rows and columns (like spreadsheets).
    - Examples: SQL Server, MySQL, Oracle, PostgreSQL.
2. **NoSQL Databases**
    - Structure: Flexible; not limited to tables. Includes:
        - Document stores (JSON-like documents) — e.g., MongoDB,Couchbase
        - Key-Value stores — e.g., Redis, DynamoDB
        - Column-family stores — e.g., Cassandra, HBase
        - Graph databases — e.g., Neo4j, Amazon Neptune
3. **Object Storage**
        - Raw unstructured data (multimedia, data lakes, backups).
        - Best for**:** Unstructured data like images, videos, backups, logs.
        - Examples**:** Amazon S3, Azure Blob Storage, Google Cloud Storage.

Data Structure

Structured Data

- Organized in fixed schema (rows, columns, fields)
- Stored in relational databases, easy to query with SQL
- **Formats:** CSV, XLS/XLSX, DB tables
- **Examples:** SQL tables, spreadsheets

Semi-Structured Data

- No strict schema, but uses tags/keys for organization
- Flexible, partially organized
- **Formats:** JSON, XML, YAML, BSON, Avro, Parquet, Logs
- **Examples:** Web data (JSON/XML), emails, NoSQL docs

Unstructured Data

- No predefined schema or structure
- Harder to search and analyze without preprocessing
- **Formats:** TXT, PDF, DOC/DOCX, MP3, MP4, JPEG/PNG, WAV
- **Examples:** Multimedia (images, audio, video), documents, social media posts

Microsoft SQL Server

SQL Server is Microsoft's relational database management system (RDBMS). An RDBMS stores data in tables according to the relational model.

A SQL Server database must comprise at least two files. One is the data file with the default extension .mdf, and the other is the log file with the default extension .ldf.

Layers of the SQL server Database

- **Client Layer**: SSMS (or any app that connects)
- **Database Engine Layer**: MSSQLSERVER service
- **Storage Layer**: MDF/NDF/LDF files on disk

Creating a Database

1. Create a folder (e.g. BDPN_SQL_SERVER)
2. Find the SQL Server service account name
3. Grant the SQL Server service account name **Full Control** on the folder
4. Verify SQL Server can see the folder
5. Run the CREATE DATABASE statement using that folder as the location

https://github.com/Stephen-Data-Engineer-Public/BDPN/blob/main/SQL_SERVER/Database.md

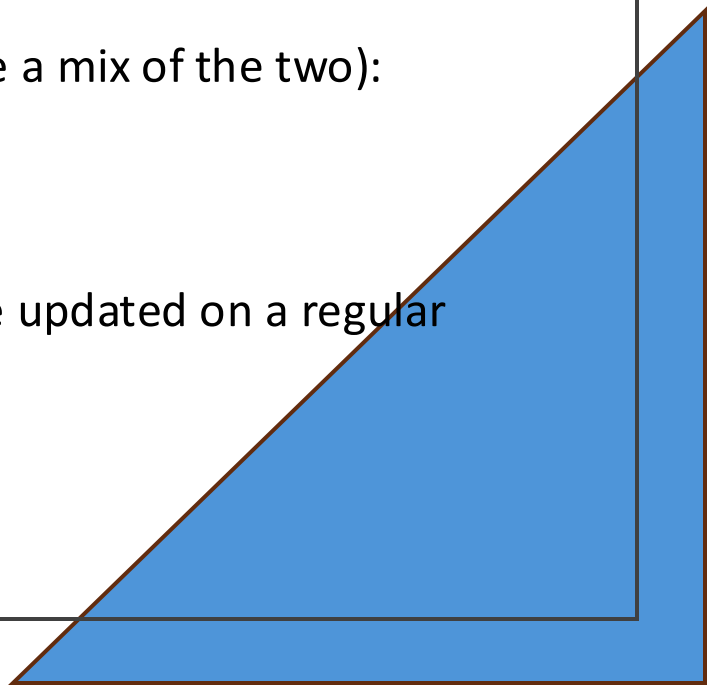
There are two types of workloads (although in real life, a lot of databases are a mix of the two):

OLTP workloads have two main characteristics:

- A lot of small queries are executed.
- A lot of new rows are added to the database and existing rows need to be updated on a regular

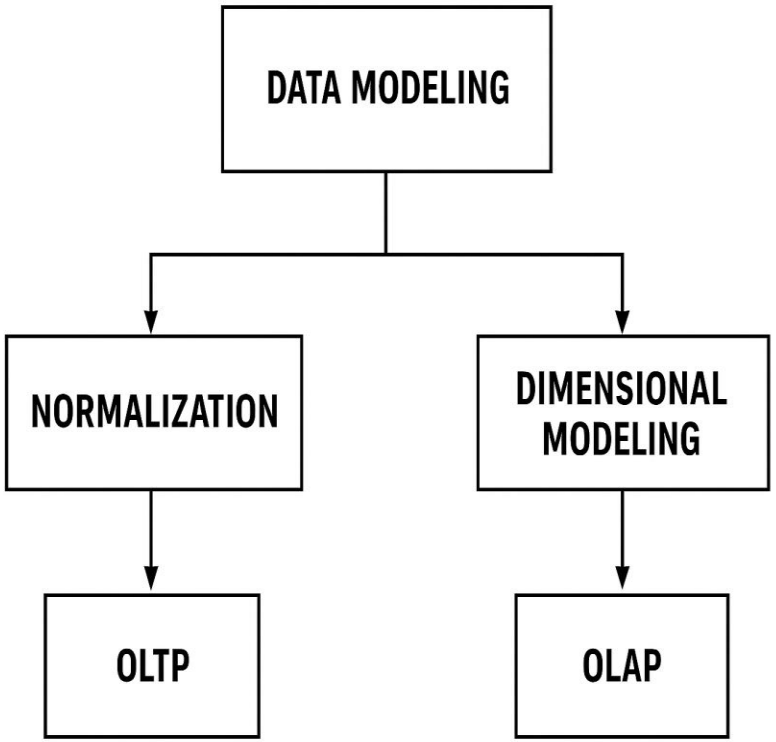
OLAP workloads also have two main characteristics:

- The workload is (almost) read-only.
- Most queries use large datasets.



Data Modelling

Data modelling : The overall process of designing how data is stored, organized, and related in a database.



Normalization

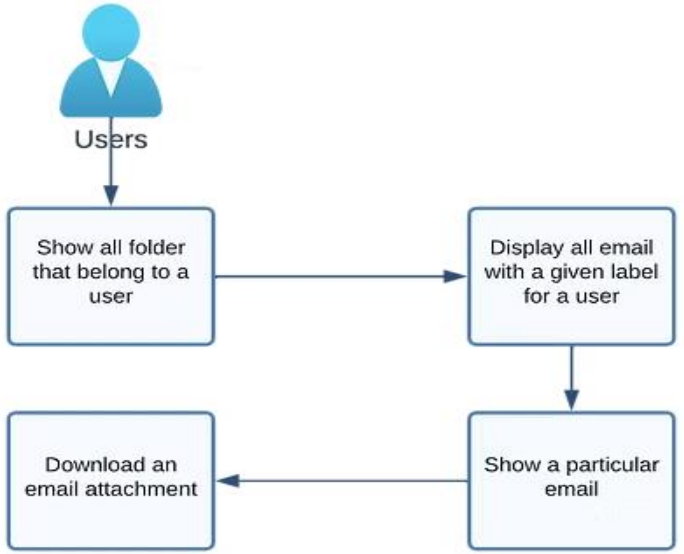
- Definition: A technique used in Online Transaction Processing (**OLTP**) database design to reduce redundancy and improve data integrity.
- Purpose: Organizes data into tables with clear relationships, usually applying rules called normal forms.

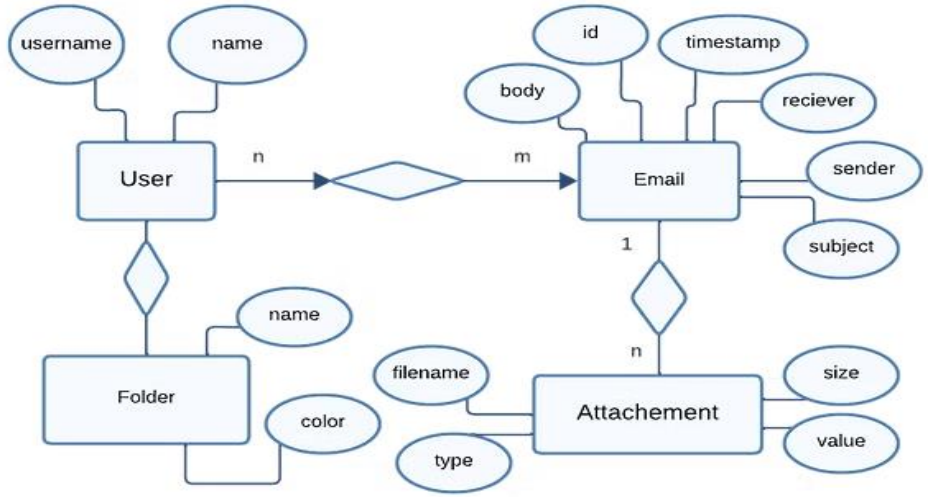
Dimensional Modelling

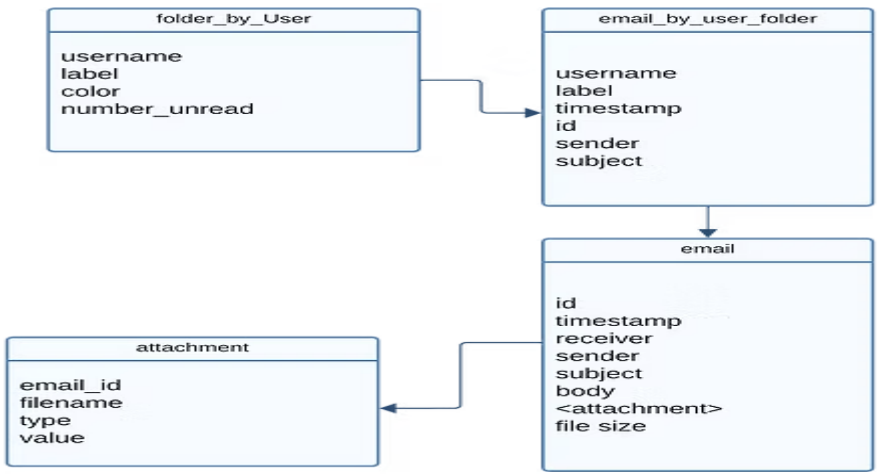
- Definition: A technique used in Online Analytical Processing (**OLAP**) database design, especially data warehouses, to optimize for querying and reporting.
- Purpose: Uses facts and dimensions in star or snowflake schemas to make complex queries faster and easier to understand.

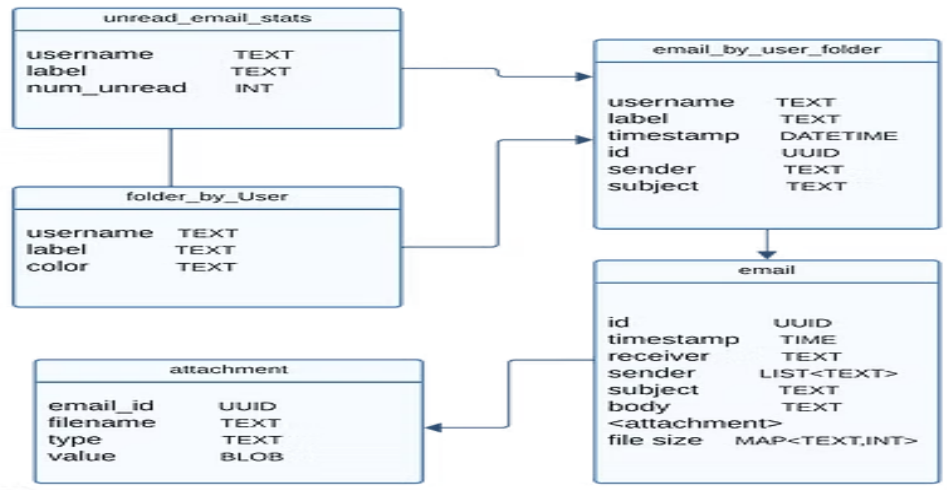
Process of Data Modelling

- **Conceptual** modelling is a process of identifying and visually mapping the moving pieces, or entities, of a business operation.
- **Logical** modelling is the bridge between the business's conceptual operating model and the physical structure of the database.
- **Logical** modelling is the bridge between the business's conceptual operating model and the physical structure of the database.(e.g., SQL Server, Oracle).


Users
Show all folder that belong to a user
Display all email with a given label for a user
Download an email attachment
Show a particular email


username
name
id
timestamp
body
reciever
User
n
m
Email
sender
subject
1
name
n
filename
size
Folder
color
Attachement
type
value


folder_by_User
username
label
color
number_unread
email_by_user_folder
username
label
timestamp
id
sender
subject
email
id
timestamp
receiver
sender
subject
body
<attachment>
file size
attachment
email_id
filename
type
value


unread_email_stats
username TEXT
label TEXT
num_unread INT
folder_by_User
username TEXT
label TEXT
color TEXT
email_by_user_folder
username TEXT
label TEXT
timestamp DATETIME
id UUID
sender TEXT
subject TEXT
email
id UUID
timestamp TIME
receiver TEXT
sender LIST<TEXT>
subject TEXT
body TEXT
<attachment>
file size MAP<TEXT,INT>
attachment
email_id UUID
filename TEXT
type TEXT
value BLOB

Creating a Table

`CREATE TABLE` defines table name + columns:

- Specify column name, data type, and `NULL/NOT NULL`.

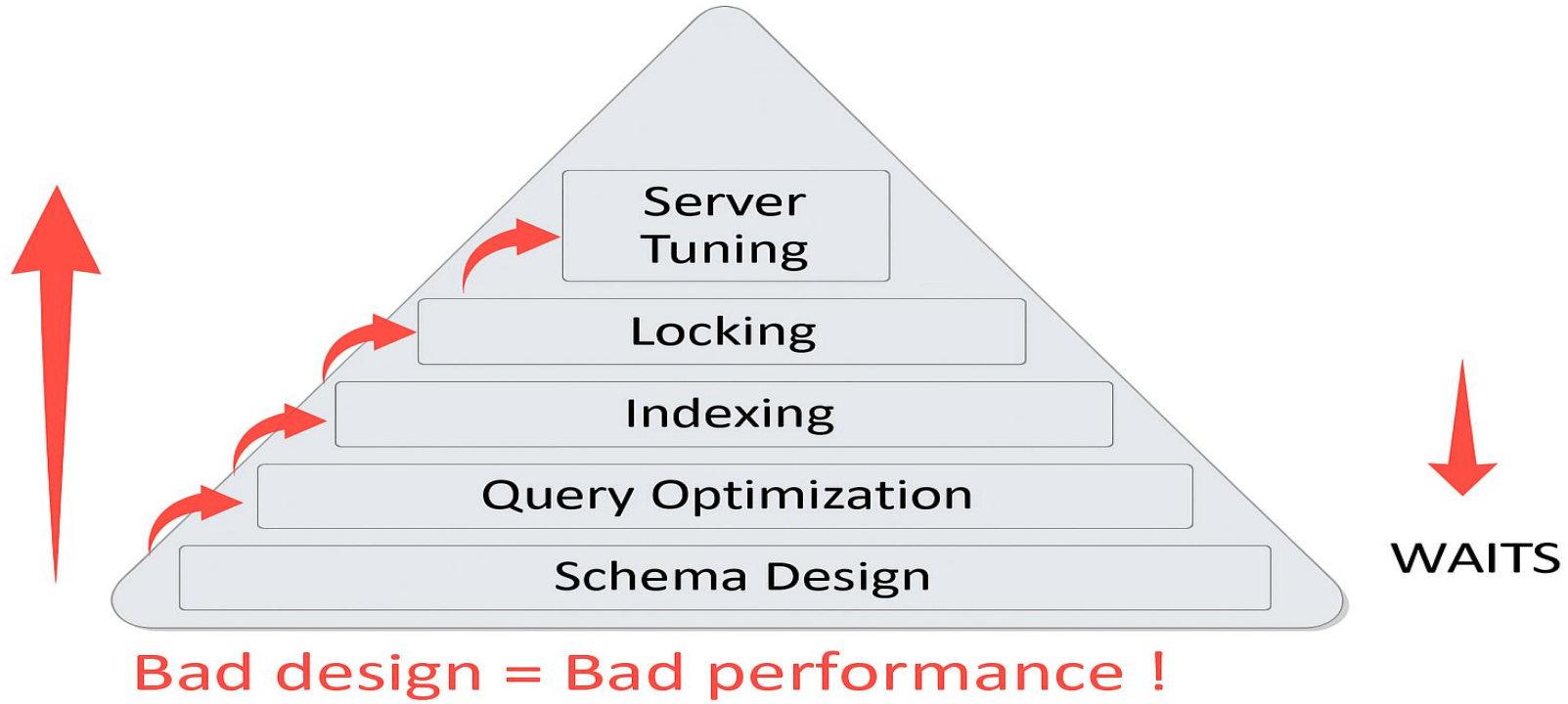
Data Integrity Constraints

- Primary Key: Ensures row uniqueness, no NULLs.
- Unique: Enforces uniqueness, allows one NULL
- Foreign Key: Enforces referential integrity.
  - Example: `FOREIGN KEY(empid) REFERENCES Employees(empid)`
- Check: Ensures condition is met.
  - Example: `CHECK(salary > 0.00)`
- Default: Supplies a value when none provided.
  - Example: `DEFAULT(SYSDATETIME())`

https://github.com/Stephen-Data-Engineer-Public/BDPN/blob/main/SQL_SERVER/Tables.md

Why is Data Modelling Important

Data modelling ensures that data is structured, consistent, and optimized, enabling faster queries, reliable analytics, and long-term scalability.



BDPN
Data careers for everyone

Project

Scenario:
A retail company wants to organize its data for reporting and operations. You are tasked with designing and building the database.

- A working RetailDB database.
- Populated tables.
- A catalog document explaining the schema and design.

https://github.com/Stephen-Data-Engineer-Public/BDPN/tree/main/Dataset/Data%20Modeling