EDM Laboratory

GWE #3

Database, Data Warehouse, Data Lakes

Brief Explanation

GWE #3

EDM Laboratory

Database vs Data Warehouse vs Data Lake

Data Warehouse

- integrated
- for reporting purpose

VS

Data Lake

- Schema-on-read
- for ML purpose

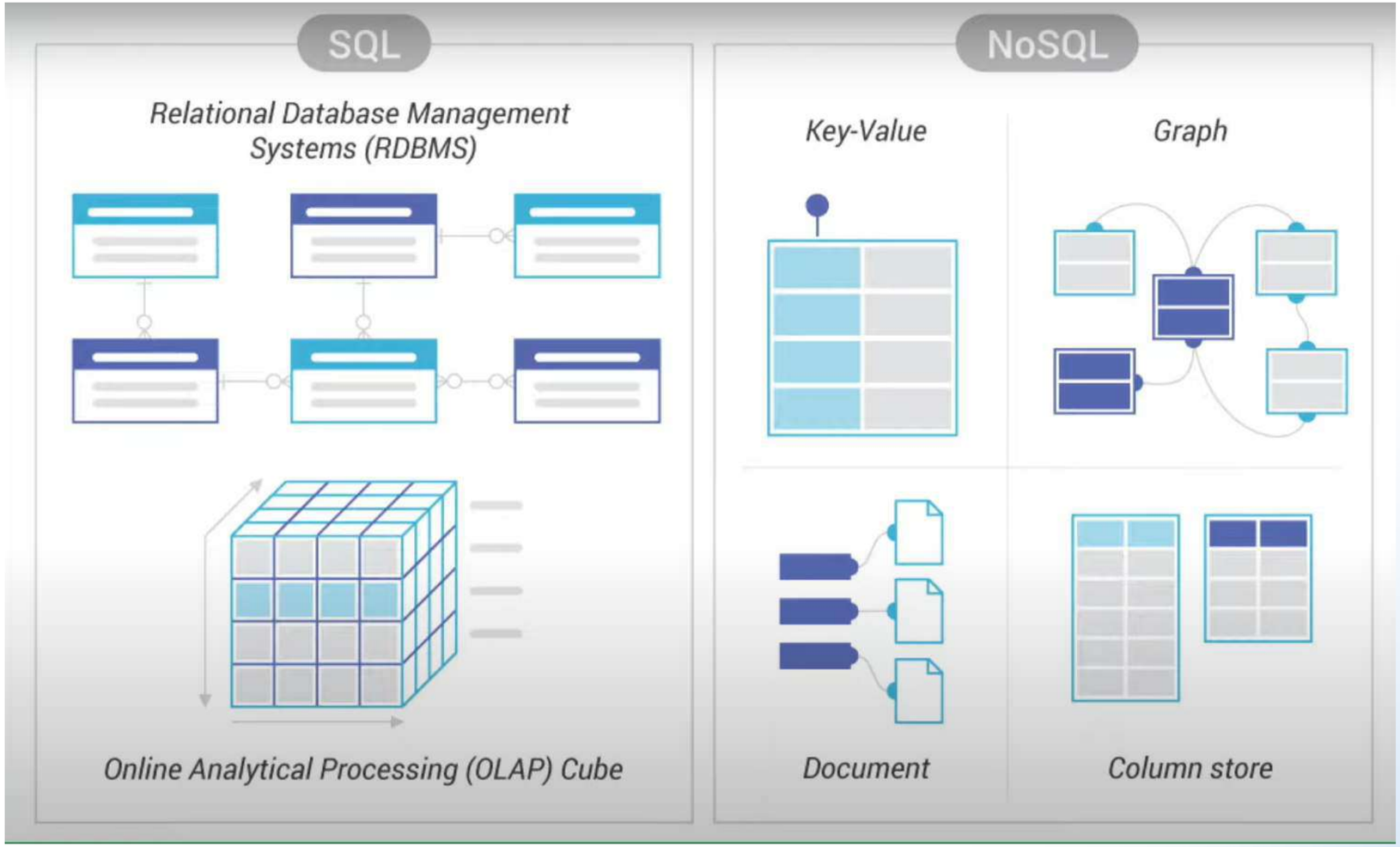
VS

Database

- CRUD
- Operational purpose (software)

EDM Laboratory

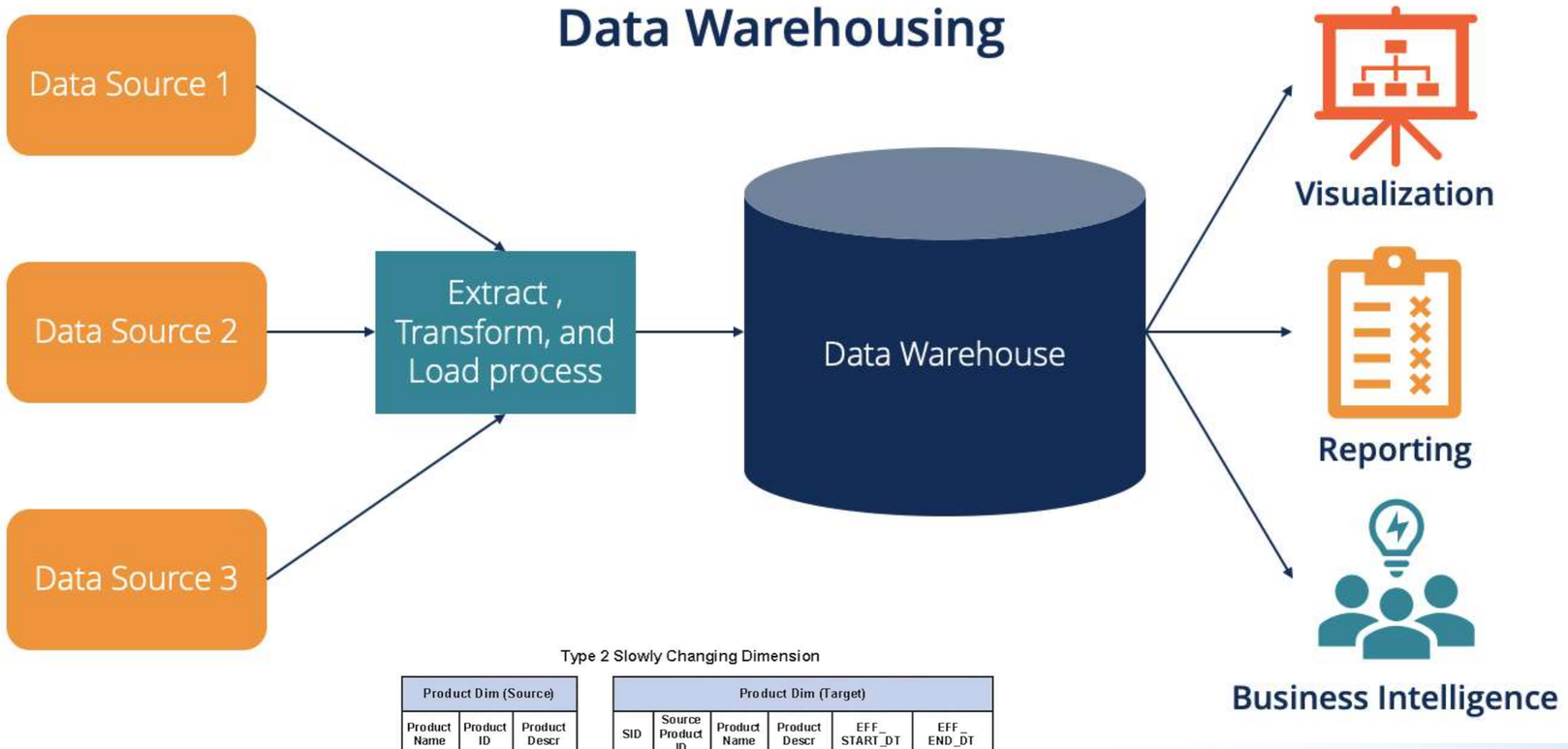
Database



GWE #3

EDM Laboratory

Data Warehouse



Type 2 Slowly Changing Dimension

Product Dim (Source)		
Product Name	Product ID	Product Descr
12 inch box	012	12 inch glued box
10 inch box	010	10 inch ~~glued~~ box 10 inch **pasted** box

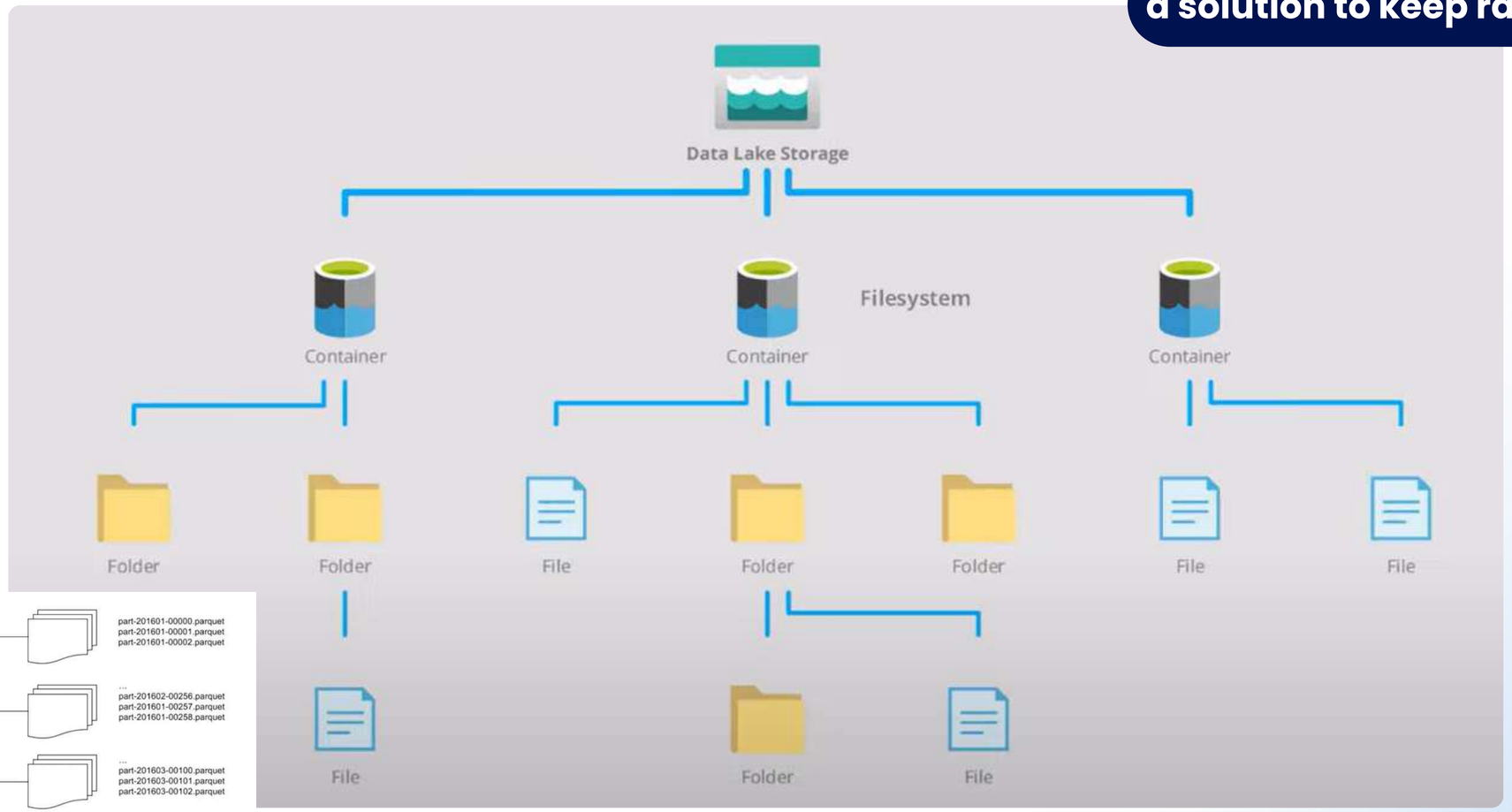
Product Dim (Target)					
SID	Source Product ID	Product Name	Product Descr	EFF_ START_DT	EFF_ END_DT
0001	012	12 inch box	12 inch glued box	Jan-01-1753	Dec-31-9999
0002	010	10 inch box	10 inch glued box	Jan-01-1753	May-12-06
0003	**010**	**10 inch box**	**10 inch pasted box**	**May-12-06**	**Dec-31-9999**

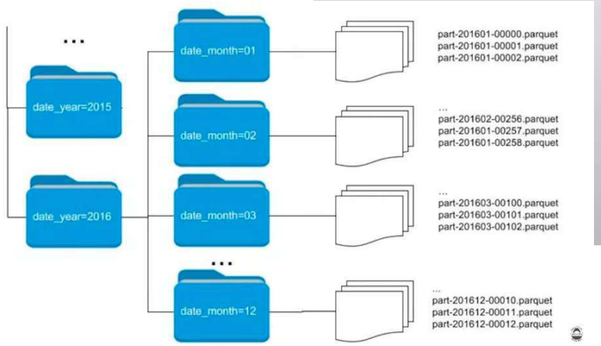
GWE #3

Data Lakes

EDM Laboratory

a solution to keep raw data





GWE #3

EDM Laboratory

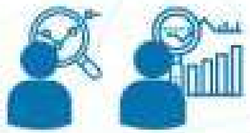
Kenapa Data Lakes

Data Lakes

Data

unstructured

Users



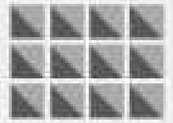
Data Scientists,
Data Analysts

Use cases

Stream Processing,
Machine Learning,
Real time analysis

Data Warehouse

Data



Structured

Users

Business Analysts

Use cases

Batch Processing,
BI, Reporting

Raw

Data Lakes contain unstructured, semi structured and structured data with minimal processing. It can be used to contain unconventional data such as log and sensor data

Large

Data Lakes contain vast amounts of data in the order of petabytes. Since the data can be in any form or size, large amounts of unstructured data can be stored indefinitely and can be transformed when in use only

Undefined

Data in data lakes can be used for a wide variety of applications, such as Machine Learning, Streaming analytics, and AI

Refined

Data Warehouses contain highly structured data that is cleaned, pre-processed and refined. This data is stored for very specific use cases such as BI.

Smaller

Data Warehouses contain less data in the order of terabytes. In order to maintain data cleanliness and health of the warehouse, Data must be processed before ingestion and periodic purging of data is necessary

Relational

Data Warehouses contain historic and relational data, such as transaction systems, operations etc

GWE #3

Kenapa Data Lakes

EDM Laboratory

- Kepemilikan dapat dibagi secara terbatas
- Simple Data Management with more Context
- Fondasi bagi AI dan ML
- More Security & Governance

EDM Laboratory

Bahas sedikit tentang DBMS

SQL

S = Structured

Q = Query

L = Language

NoSQL

N = Not

O = Only

S = Structured

Q = Query

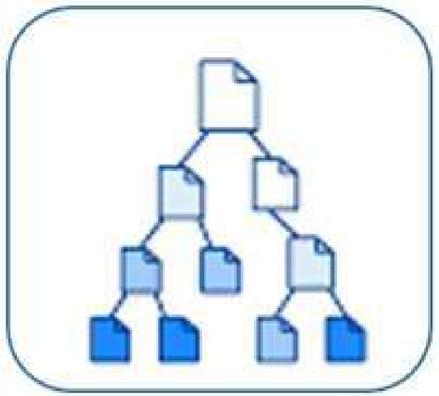
L = Language

EDM Laboratory

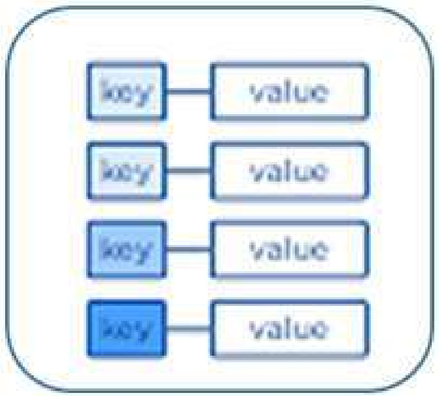
Bahas sedikit tentang DBMS

NoSQL

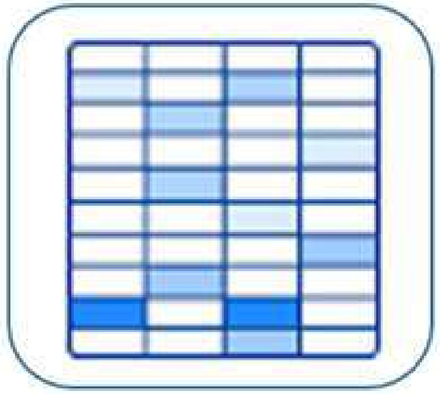
-> sistem manajemen database yang tidak memiliki relasi.



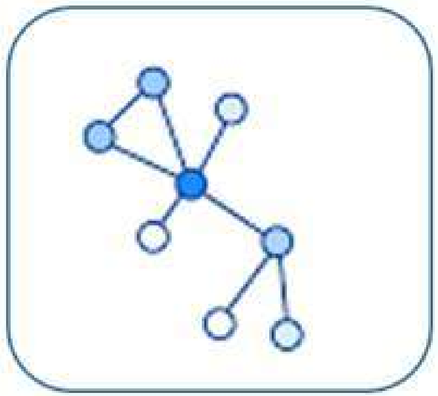
Document Store



Key-Value Store



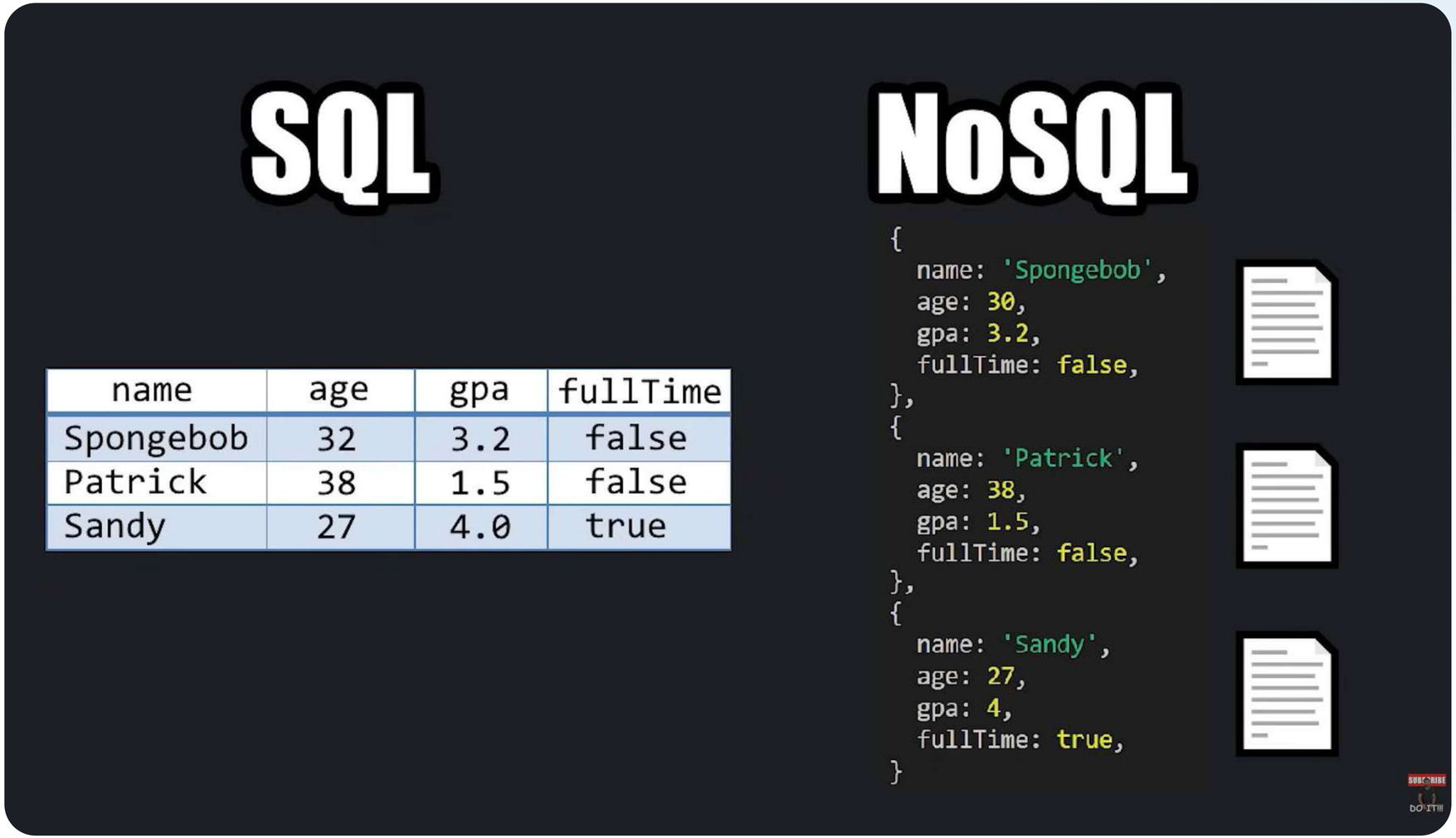
Wide-Column Store



Graph Store

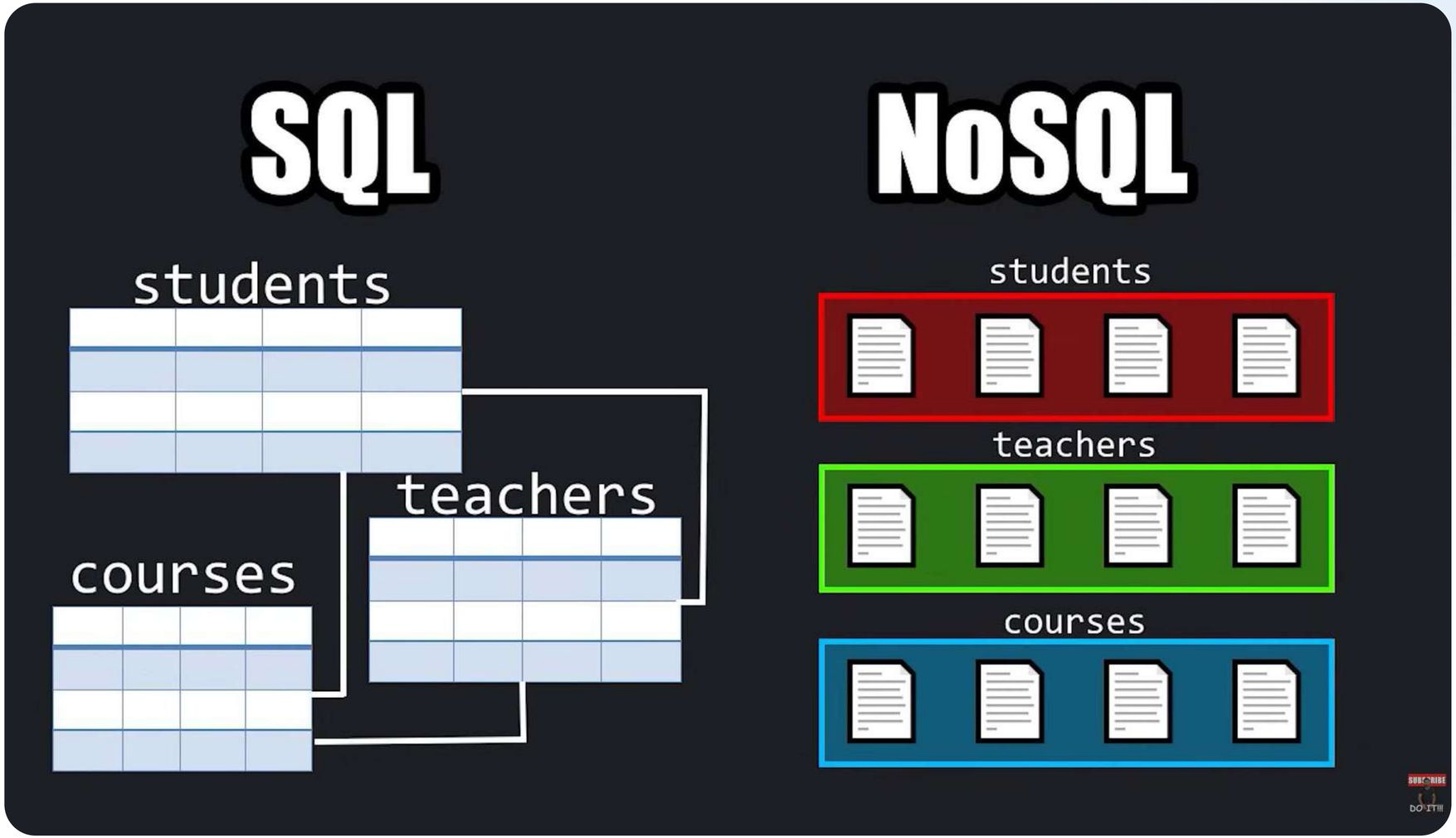
EDM Laboratory

Bahas sedikit tentang DBMS



EDM Laboratory

Bahas sedikit tentang DBMS



each row = each document

EDM Laboratory

Pros & Cons

SQL

+ Easy querying on relationships
+ More organized & structured (minim error)
+ Atomic

- Prepared structure
- Hard to Scale (horizontal scaling)

NoSQL

+ Flexible
+ Enabled Sharding

- Ease loss of consistency

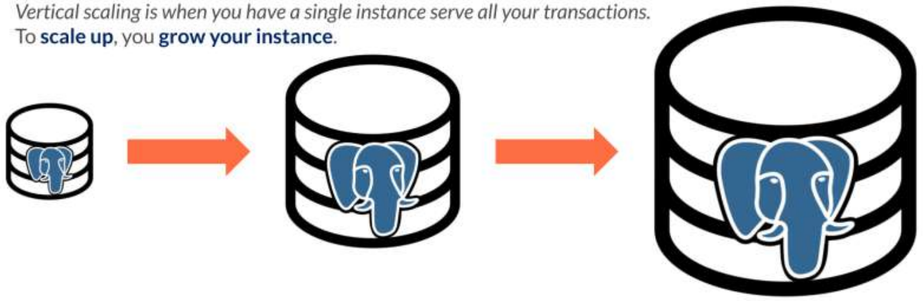
each row = each document

EDM Laboratory

Pros & Cons

Vertical Scaling

Vertical scaling is when you have a single instance serve all your transactions.
To **scale up**, you **grow your instance**.



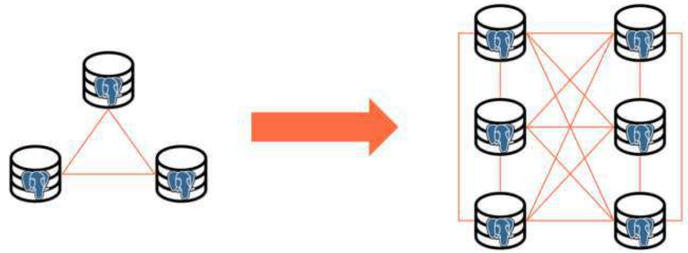
2 vCPU
2GB memory
50GB storage

32 vCPU
128GB memory
2TB storage

96 vCPU
768GB memory
64TB storage

Horizontal Scaling

Horizontal scaling is when you have a multiple instances serving your transactions.
To **scale out**, you **add instances** (called nodes).



3 nodes
2 vCPU each
2GB memory each
50GB storage each

6 nodes
2 vCPU each
2GB memory each
50GB storage each

Pros & Cons


```
{
  title : 'movie',
  actors : [
    {
      _id: 'actor_id1',
      name: 'actor1'
    },
    {
      _id: 'actor_id'2,
      name: 'actor2'
    }
  ],
  plot: '...',
  reviews: [...],
  ...
}

{
  name : 'actor1',
  movies : [
    {
     _id: 'movie_id1',
     name: "movie1'
    },
    {
     _id: 'movie_id2',
     name: "movie2'
    }
 ],
 biography: '...',
 pictures: [...],
 ...
```

EDM Laboratory

So which one?