NANYANG TECHNOLOGICAL UNIVERSITY SINGAPORE

NTU Academy for Professional and Continuing Education

(SCTP) Advanced Professional Certificate

Data Science and AI

Data Science Archetypes

Step 1 : Take the quiz

Step 2 : Share results

https://www.menti.com/aleew5mqi6tz

8460 6080

NANYANG TECHNOLOGICAL UNIVERSITY SINGAPORE

2.1 Introduction to Big Data and Data Engineering

Module Overview

2.1 Introduction to Big Data and Data Engineering

2.2 Data Architecture

2.3 Data Encoding and Data Flow

2.4 Data Extraction and Web Scraping

2.5 Data Warehouse

2.6 Data Pipelines and Orchestration

2.7 Data Orchestration and Testing

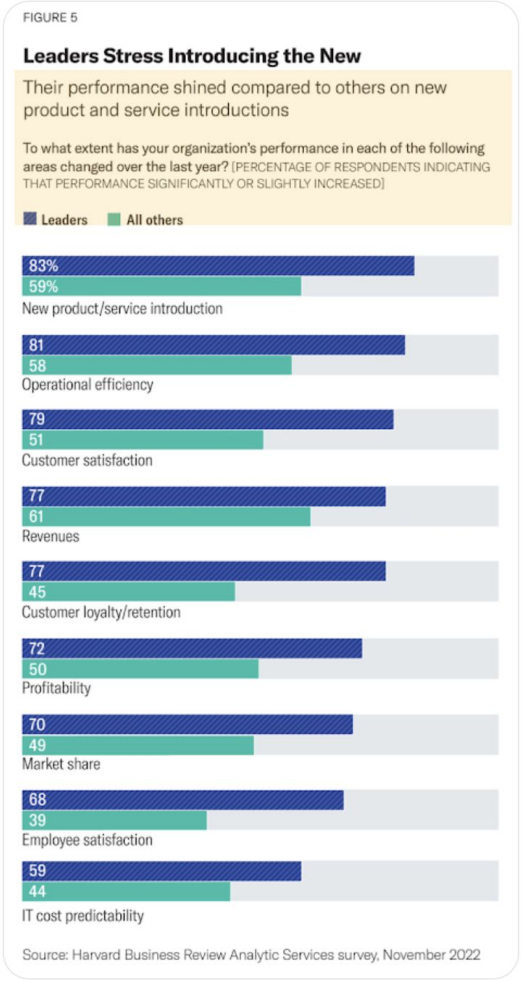
2.8 Out of Core/Memory Processing

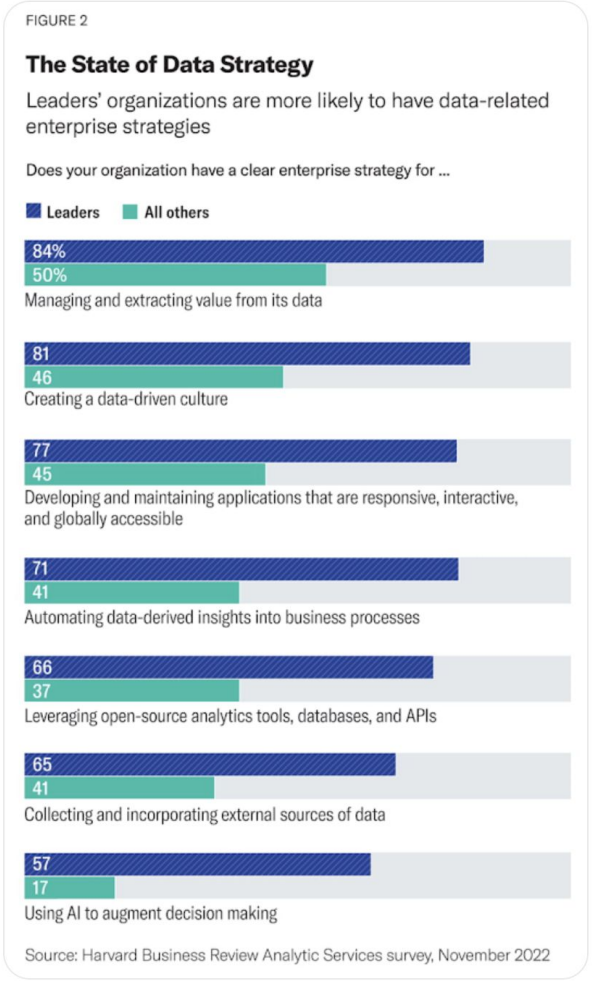
2.9 Big Data Ecosystem and Batch Processing

2.10 Event Streaming and Stream Processing

Big on data: Study shows why data-driven companies are more profitable than their peers

A Harvard Business Review survey of more than 360 executives reveals how data leaders use AI and analytics to power decision making and thrive in crisis.





Agenda

- Introduction to Big Data, 5Vs
- Big Data Tools and Technologies
- Structured vs Unstructured Data
- NoSQL Database Types
- OLTP vs OLAP
- ACID vs BASE
- Introduction to Data Engineering
- Data Engineering Lifecycle

Big Data

- Volume and Variety
  - Massive amounts of structured and unstructured data
- Velocity and Veracity
  - Fast moving data streams requiring reliable processing

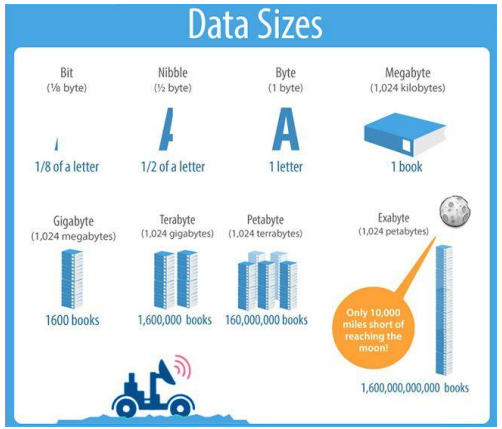
Data Engineering

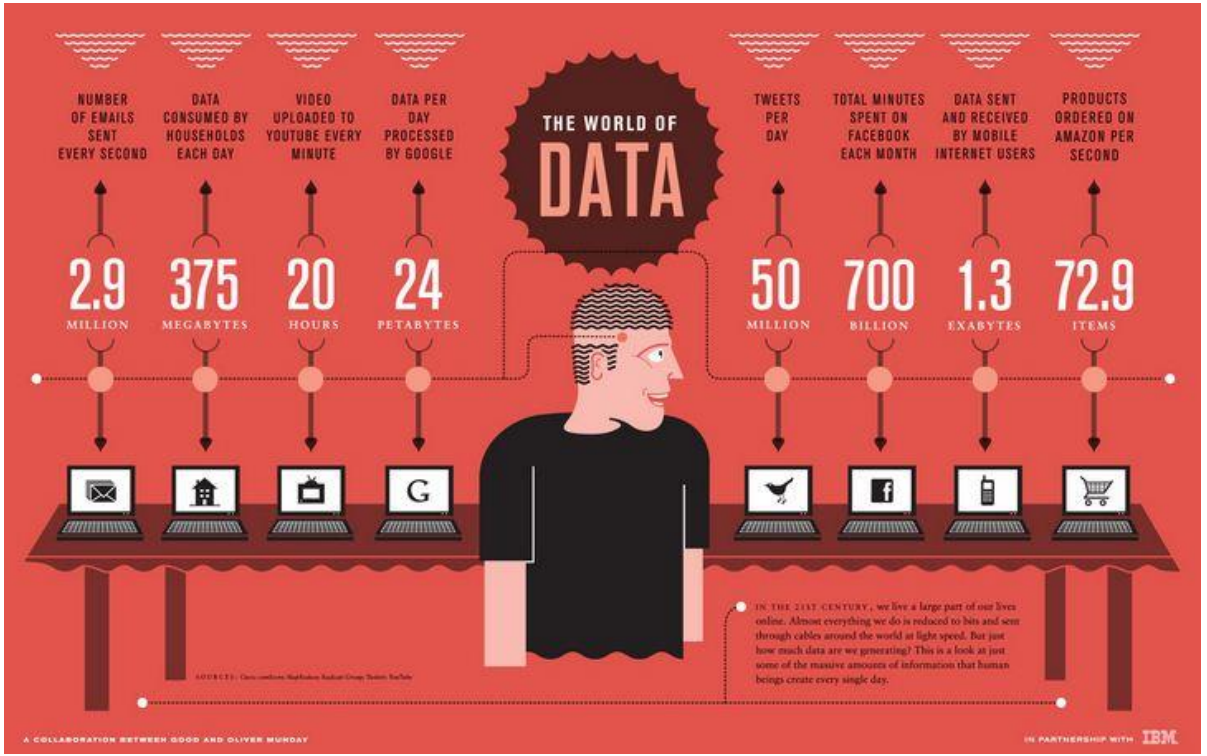
- Lifecycle management
  - Generation
  - Storage
  - Ingestion
  - Transform
  - Serve
- Data engineering provides the framework to extract value from big data

Big Data describes large volume of data:

- *structured* and *unstructured*

Large amounts of data that cannot be processed using traditional methods.





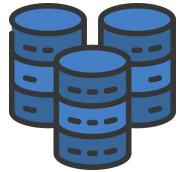
What are the Vs of Big Data?

https://www.menti.com/aleew5mqi6tz

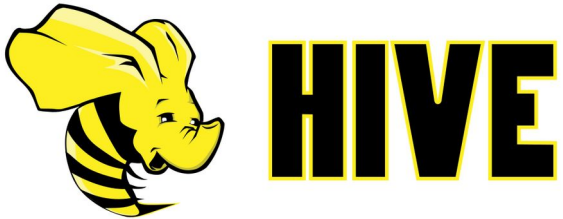
Or use QR code

5 Vs of Big Data

- **Volume**: The total amount of data produced by both machines and people.
- **Velocity**: How fast data is created and processed.
- **Variety**: The range of data types, encompassing both structured and unstructured data.
- **Veracity**: The reliability of data, including its biases, noise, and abnormalities, affecting insight accuracy.
- **Value**: The usefulness of data, defined by the insights it enables, the decisions it guides, and its impact on business outcomes.

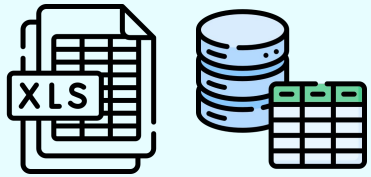


Big Data Tools and Technologies



Structured and Unstructured Data

Structured Data



Organized in tables, easy to search/analyze. *Examples*: *Relational databases (DuckDB), spreadsheets (Excel).*

Unstructured Data

Lacks predefined organization, hard to search/analyze. *Examples: images, videos, social media posts.* Makes up ~80% of generated data.

Relational databases need structured schema

NoSQL databases are better for storing large amount of unstructured data, offering scalability and flexibility.

Database Recap

Relational Database

- Also known as SQL database
- Based on relational model and has predefined schema
- Suited for structured data
- Uses SQL (Structured Query Language) for querying

NoSQL Database

- “Not-only” SQL, or non-relational database
- Flexible schema
- Suited for semi-structured / unstructured data
- SQL optional

Databases: OLTP and OLAP

OLTP (Online Transaction Processing)

For real-time data entry and retrieval.

Optimised for **transactional** processing (banking transactions, e-commerce orders, online reservations, etc.)

Usually **normalized**.

A transaction is a single logical unit of work that accesses and possibly **modifies** the contents of a database.

Both **read** and **write** operations.

Examples: MySQL, Oracle, SQL Server

OLAP (Online Analytical Processing)

For complex data analysis.

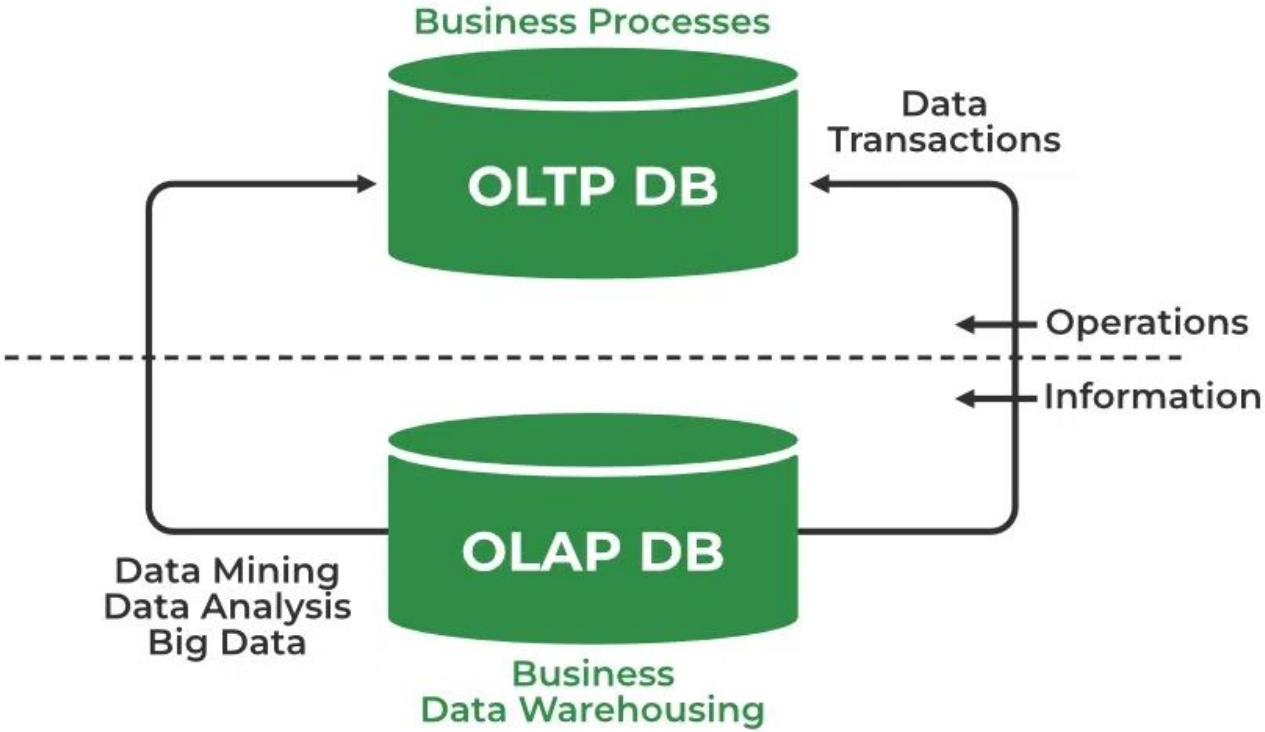
Optimized for **analytical** processing (business reporting, process management, budgeting, forecasting)

Usually **denormalized**.

An analytical **query** performs selection, filtering and aggregation on data

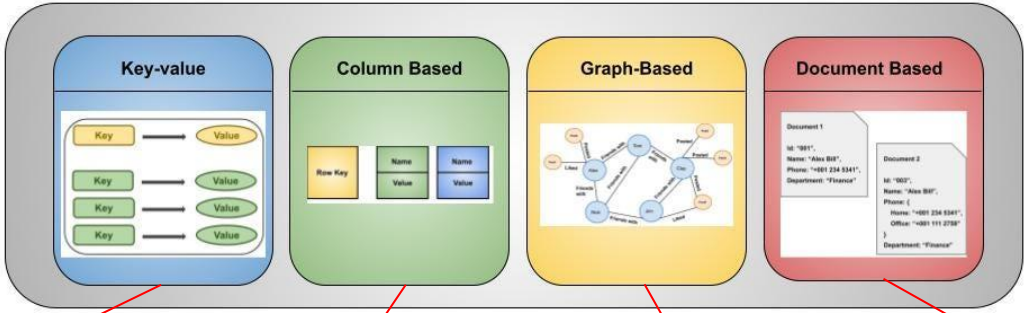
Usually **read** operations, rarely write.

Examples: data warehouses, data marts, etc.



Source: [Difference Between OLAP and OLTP in Databases | GeeksforGeeks]

Types of NoSQL databases



Stores data in key-value pairs. (*Redis, DynamoDB*)

Stores data in columns instead of rows (*Cassandra, HBase*)

Stores data in nodes and edges. (*Neo4j, OrientDB*)

Stores data in documents (*MongoDB, CouchDB*)

🍋 ACID

- **A**tomicity, **C**onsistency, **I**solation, **D**urability
- Guarantees all transactions are valid, even in event of failures
- High data integrity and reliability
- **Can lead to performance bottlenecks**
- **Challenging to scale**

🧼 BASE

- **B**asically **A**vailable, **S**oft state, **E**ventual consistency
- Prioritize availability and scalability over immediate consistency
- Lower latency and higher throughput
- Relaxes some ACID properties
- **Temporary data inconsistency**
- **Possible stale data**

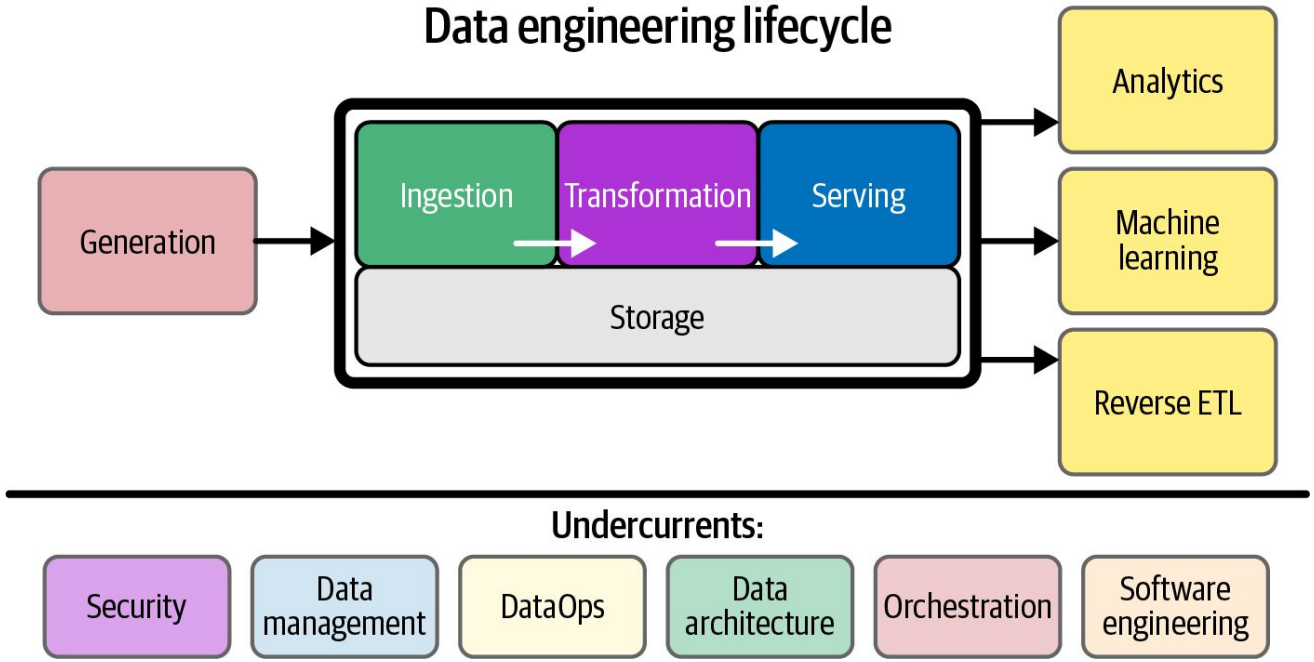
ACID vs BASE

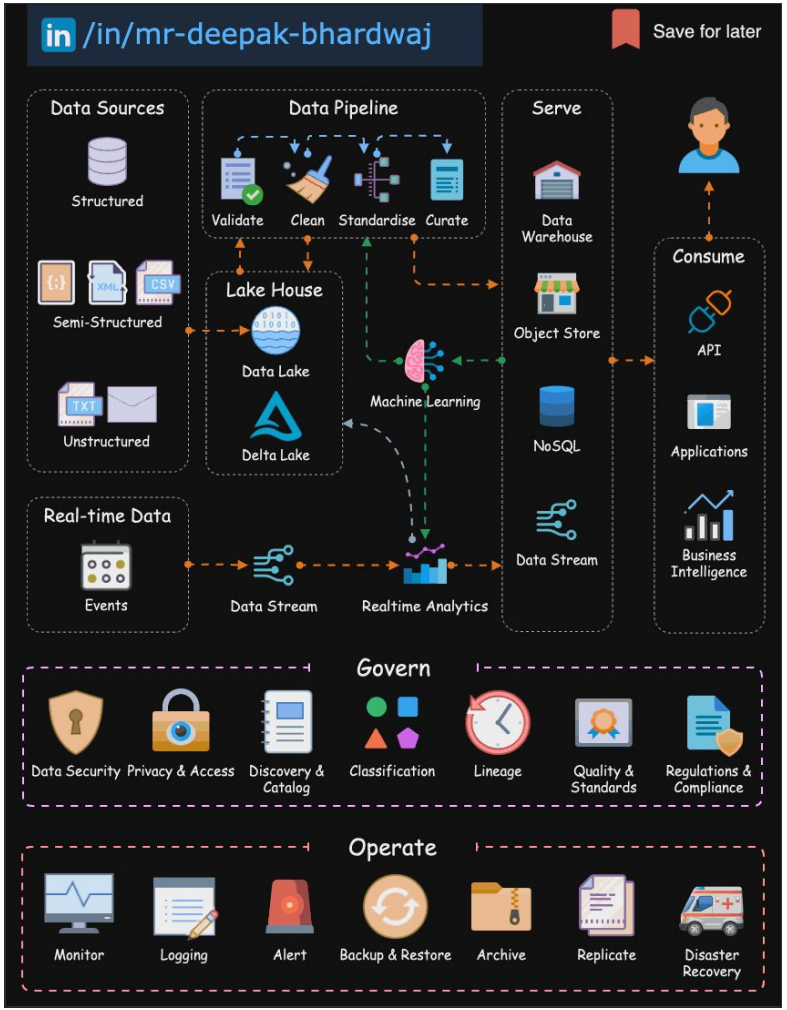
https://www.menti.com/aleew5mqi6tz

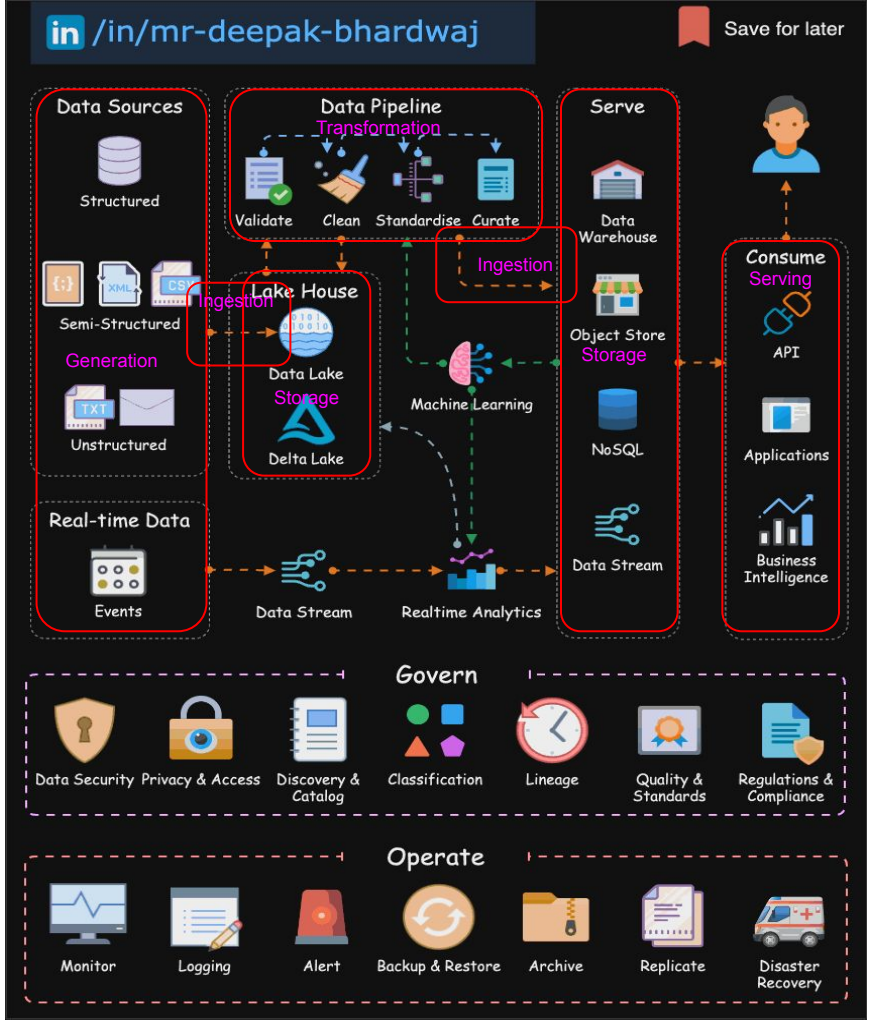
Or use QR code

Data Engineering Lifecycle

- Generation
- Storage
- Ingestion
- Transformation
- Serving







End of Lesson - Exit Ticket

Survey Link

https://www.menti.com