Data Curation Techniques

Siddharth R

Data Modeling and Design

- The process for converting data into a usable form is known as data modeling and design.
- Data modeling has become more challenging because of the variety of new data sources and use cases.
- A data swamp is a poorly managed data where stored data lacks organization, metadata, and governance, making it difficult to retrieve, understand, or use - Example: Write Once, Read Never (WORN)

Data Integration & Interoperability

Challenges with Heterogeneous Data:

- Format Incompatibility - JSON , XML, CSV, SQL, NoSQL
- Semantic Inconsistency
- Structural Differences
- Scalability

Example Scenario: Smart City - Integration challenges in IoT data

Principles of good data architecture

- Agility is the foundation for good data architecture
- Good data architecture is flexible and easily maintainable.
- Bad data architecture is tightly coupled, rigid, overly centralized
- The AWS Well-Architected Framework consists of six pillars:
  - Operational excellence
  - Security
  - Reliability
  - Performance efficiency
  - Cost optimization
  - Sustainability

Plan for Failure

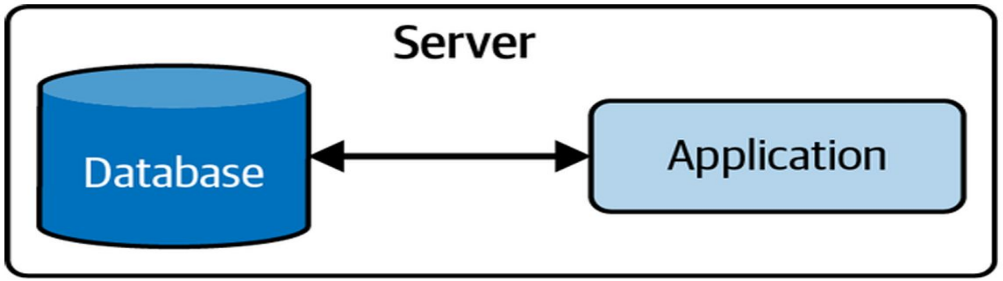
Key terms evaluation failure scenario

1. Availability
2. Recovery time objective
3. Recovery point objective

Tightly coupled vs loosely coupled

- Every part of a domain and service is vitally dependent upon every other domain and service. This pattern is known as tightly coupled.
- Services that do not have strict dependence on each other, in a pattern known as loose coupling.
- Architecture tiers:
  - Single tier
  - Multi tier
  - Monolithic
  - Microservices

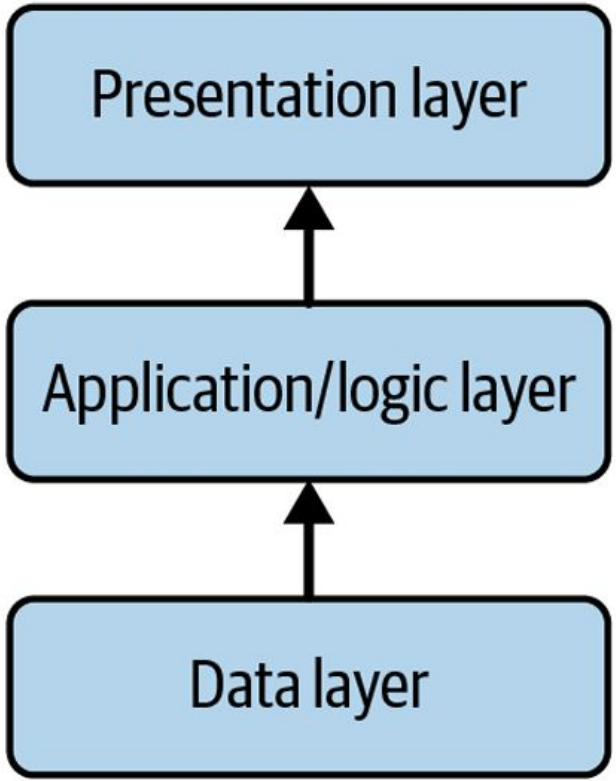
Single Tier



- Database and application are tightly coupled, residing on a single server
- Entire architecture fails if any one fails
- Good for prototyping
- Not advised for production
- Resource availability issue

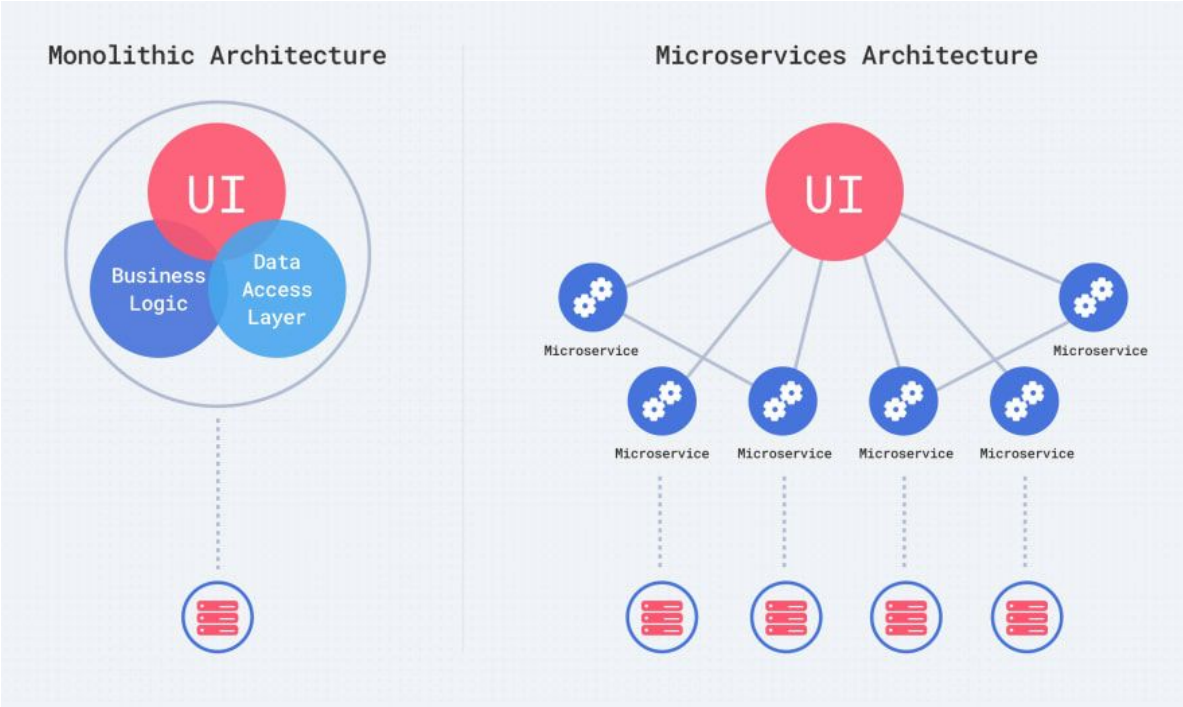
Multitier

- Also known as n-tier
- Composed of separate layers: data, application, business logic, presentation, etc.
- Resources are spread across various tiers.



Monoliths & Microservices

- A monolithic application is one large, self-contained unit.
- All parts of the app—UI, business logic, database access, etc.—are packaged together and run as a single service.
- In Microservices, each service is independent and talks via APIs.
- Can be scaled, updated, and deployed independently.



Data Storage System

- Database
- Data warehouse
- Data lake
- Data mart
- Cloud storage

Data Storage Abstraction

- Hot
- Warm
- Cold

Thank You !!!