WHAT IS SCALABILITY IN SYSTEM DESIGN?

educative

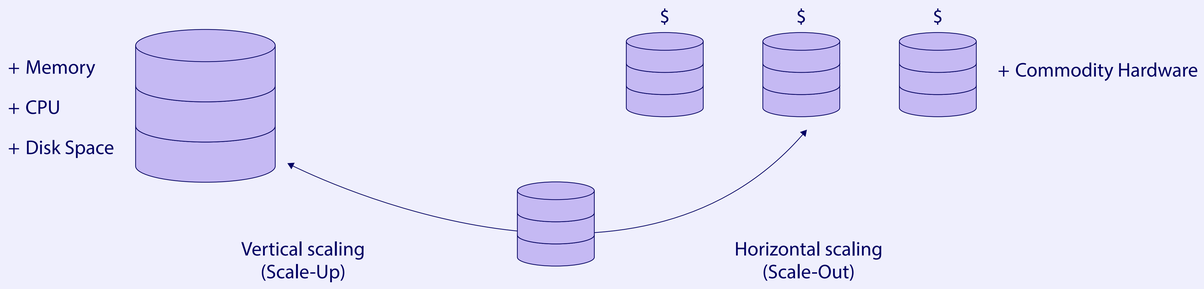
What is Scalability in System Design?

The ability of a system to handle an increasing amount of workload without compromising performance

Types of Scalability

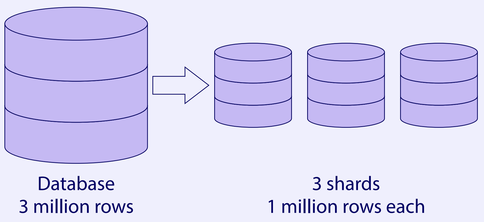
- Vertical scaling refers to scaling by providing additional capabilities to an existing device
- Horizontal scaling refers to scaling by increasing the number of machines in the network

What It Looks Like

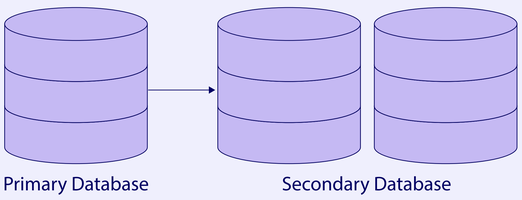


Scalability Techniques

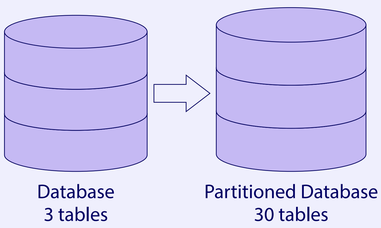
Sharding:
Distributes data across multiple servers in smaller units



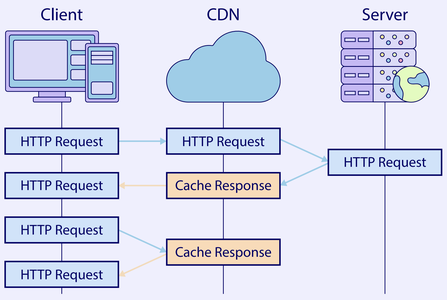
Replication:
Creates and maintains copies of data across multiple servers



Partitioning:
Divides a database into smaller, organised segments



Caching:
Improves response times through efficient retrieval of frequently accessed data



How to Achieve It

- Modular Design
- Load Balancing
- Caching
- CDNs
- Elasticity
- Asynchronous Processing

What to Avoid

- Monolithic Architectures
- Stateful Components
- No Load Testing

Key Principles to Consider

- CAP Theorem
- Microservices
- Event-Driven Architecture
- ACID Transaction