Phase 1

1️⃣ Capacity (Agility, Elasticity, Scalability)

- **Capacity** = How much your system can handle.
- **Agility** = How quickly you can adjust resources.
- **Elasticity** = Ability to **grow or shrink automatically** based on demand.
- **Scalability** = Ability to **handle more load** without breaking.

Example:
Imagine a website:

- On weekdays, 50 users visit → one small server is enough.
- On weekends, 5000 users visit → you add more servers automatically (elasticity).

2️⃣ Availability

- **Availability** = Your service stays **up and running**, even if something goes wrong.
- Think of it like a **backup plan** for failure.

Example:

- You have a website hosted in one data center. If that data center goes down, your site goes offline → low availability.
- If you have two data centers and traffic switches to the second when the first fails → high availability.

3️⃣ Blast Radius

- **Blast radius** = How much damage happens if something fails.
- Smaller blast radius → less impact when failure happens.

Example:

- One giant server crashes → your entire app goes down → huge blast radius.
- Many small servers → if one crashes, only part of your app is affected → small blast radius.

4️⃣ Disaster Recovery

- **Disaster recovery** = How quickly you can **restore your service after a big failure**.

Example:

- Your data center floods → you switch to another region and your website is back online.

5️⃣ Vertical Scaling vs Horizontal Scaling

Vertical Scaling (Scale Up)

- Add **more power to one server** (CPU, memory).
- **Problem:**
    - Often requires **downtime** to upgrade.

    - Not all apps can handle huge servers.

- **Example:**
    - Upgrade a small server from 2 CPUs → 16 CPUs → your server must restart.
    - If it crashes during upgrade → downtime.

Horizontal Scaling (Scale Out)

- Add **more servers** instead of making one bigger.
- **Better for cloud** → handles failure and traffic smoothly.
- **Example:**
    - Your website has 2 small servers → traffic increases → add 3 more servers automatically.
    - Traffic decreases → remove 1 server.
    - You **never go below 2 servers** to ensure availability.

Rule of thumb in cloud: Horizontal scaling is safer, more flexible, and keeps your service always available.

✅ Simple Analogy

- **Vertical Scaling** = One big pizza → hard to eat, and if it burns, you lose all.
- **Horizontal Scaling** = Many small pizzas → easy to share, if one burns, others are still fine.
-