1) What is Load Balancing?

A: If one server gets too many requests, it can crash or slow down. Load balancing distributes traffic evenly across multiple servers.

Q: How does it work?

A: A load balancer sits in front of your servers. It checks which server has the least load and sends the user's request there.

Benefits:

✅ Prevents server crashes
✅ Handles more users
✅ Improves uptime

Analogy: Like a security guard sending customers to the shortest cash counter line.

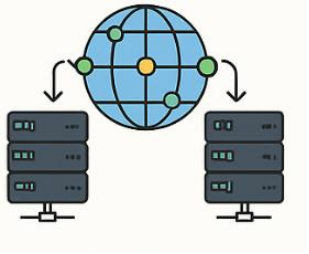
2) What is a CDN (Content Delivery Network)?

Q: Why does a website load slowly in some countries?

A: Because the server might be far away. Data has to travel a long distance.

Q: How does a CDN help?

A: A CDN has many servers worldwide. It stores copies of your website's static files (images, videos, scripts) close to the user.



Benefits:

✅ Loads websites faster globally
✅ Reduces your main server's workload
✅ Improves user experience everywhere

Analogy: Like having multiple warehouses near customers instead of shipping everything from one city.

3) What is Caching?

Q: Why do servers repeat the same work for every user?

A: Without caching, the server fetches or computes data every time someone asks — even if it's the same data.

Q: How does caching solve this?

A: Cache stores frequently requested data temporarily. The server can send the cached copy instead of recalculating.

Benefits:

✅ Faster responses
✅ Fewer database calls
✅ Great for static data like homepages or profiles

Analogy: Like remembering the answer to a frequently asked question instead of looking it up every time.

4) What is a Database Index?

Q: Why do some database queries take a long time?

A: Because the database may have to scan every row to find what you need.

Q: How does an index make queries faster?

A: An index is like a table of contents for your data. It lets the database jump directly to the matching rows instead of checking every single one.

Benefits:

✅ Speeds up data retrieval
✅ Reduces server load
✅ Improves user experience for searches

Analogy: Like using a book's index to find a topic instead of reading every page.

5) What is Horizontal Scaling?

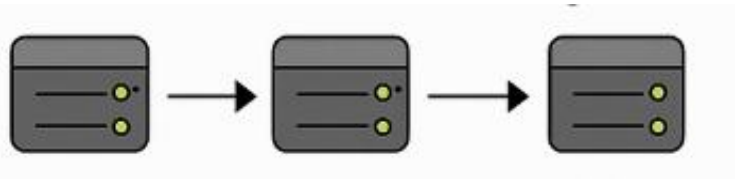
Q: What if one server can't handle all the traffic?

A: The server can get overloaded, slowing down the system or causing crashes.

Q: How does horizontal scaling solve this?

A: Instead of making one server more powerful, you add more servers to share the load.

Benefits:



✅ Handles more traffic
✅ Provides better reliability
✅ Easier to grow your system

Analogy: Like adding more checkout counters in a supermarket during rush hour.

6) What is a Message Queue?

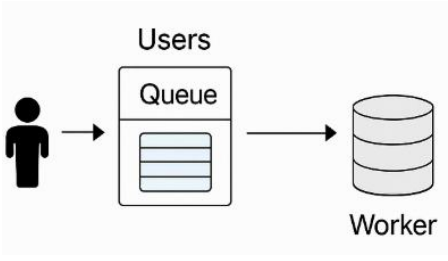
Q: Why do some apps process tasks slowly when many users act at once?

A: Because doing everything immediately can overwhelm the system.

Q: How does a message queue help?

A: A message queue temporarily stores tasks and lets workers process them one by one or in batches, smoothing out spikes in traffic.

Benefits:



✅ Prevents system overload
✅ Handles tasks asynchronously
✅ Improves reliability during high traffic

Analogy: Like customers taking a token and waiting their turn instead of crowding the service desk.

7) What is Replication?

Q: How do databases avoid downtime or data loss if one server fails?

A: Replication copies data from one database server (master/primary) to one or more others (slaves/replicas), so if one fails, another can take over.

✅ Benefits:

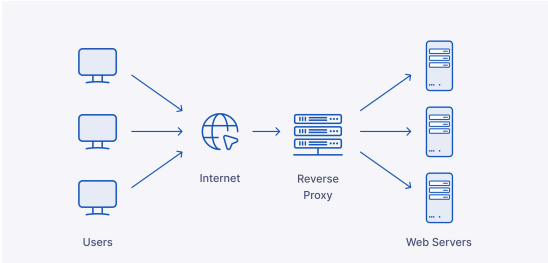
- High availability
- Data redundancy
- Better read performance (reads can be distributed)

Analogy: Like having duplicate keys for a locker, so you're never locked out.

8) What is a Reverse Proxy?

Q: How do websites hide their internal servers from users and add security?

A: A reverse proxy sits in front of web servers, forwarding client requests while handling things like caching, SSL termination, and load balancing.



✅ Benefits:

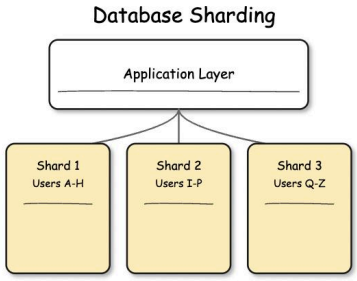
- Security (hides internal structure)
- Caching for faster responses
- Simplifies scaling

Analogy: Like a receptionist who filters and forwards calls to the right employee.

9) What is Sharding?

Q: How do you handle very large databases that can't fit on a single machine?

A: Sharding splits a database into smaller pieces (shards), each holding a subset of data. These shards can be distributed across multiple servers.



✅ Benefits:

- Handles massive datasets
- Allows horizontal scaling
- Reduces bottlenecks

Analogy: Like splitting a phonebook into different books for different cities.

10) What is Sticky Session?

Q: In load balancing, how do you make sure a user keeps talking to the same server?

A: Sticky sessions (session affinity) make sure requests from the same client always go to the same backend server, often using cookies or IP hash.

✅ Benefits:

- Keeps user session state consistent
- Simplifies session management

Analogy: Like always visiting the same cashier at a shop who remembers your order.