

# Introduction

## Big Data Systems

Dr. Rubi Boim

# Agenda for today

- 5 V's of Big Data
- Cloud computing
- Highly available / highly Scalable
- Managed vs Unmanaged services

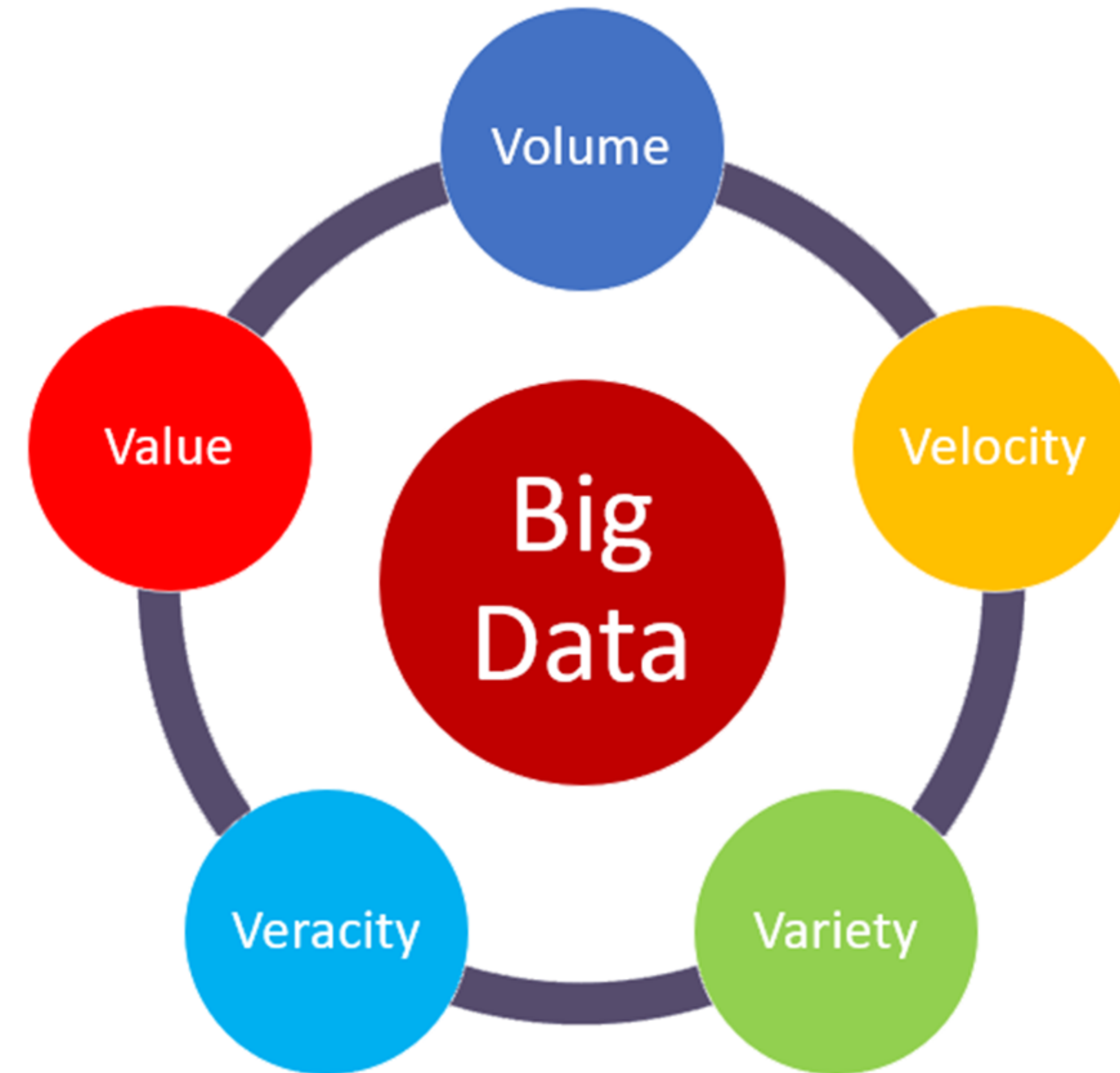
# When data is Big Data?



Discussion

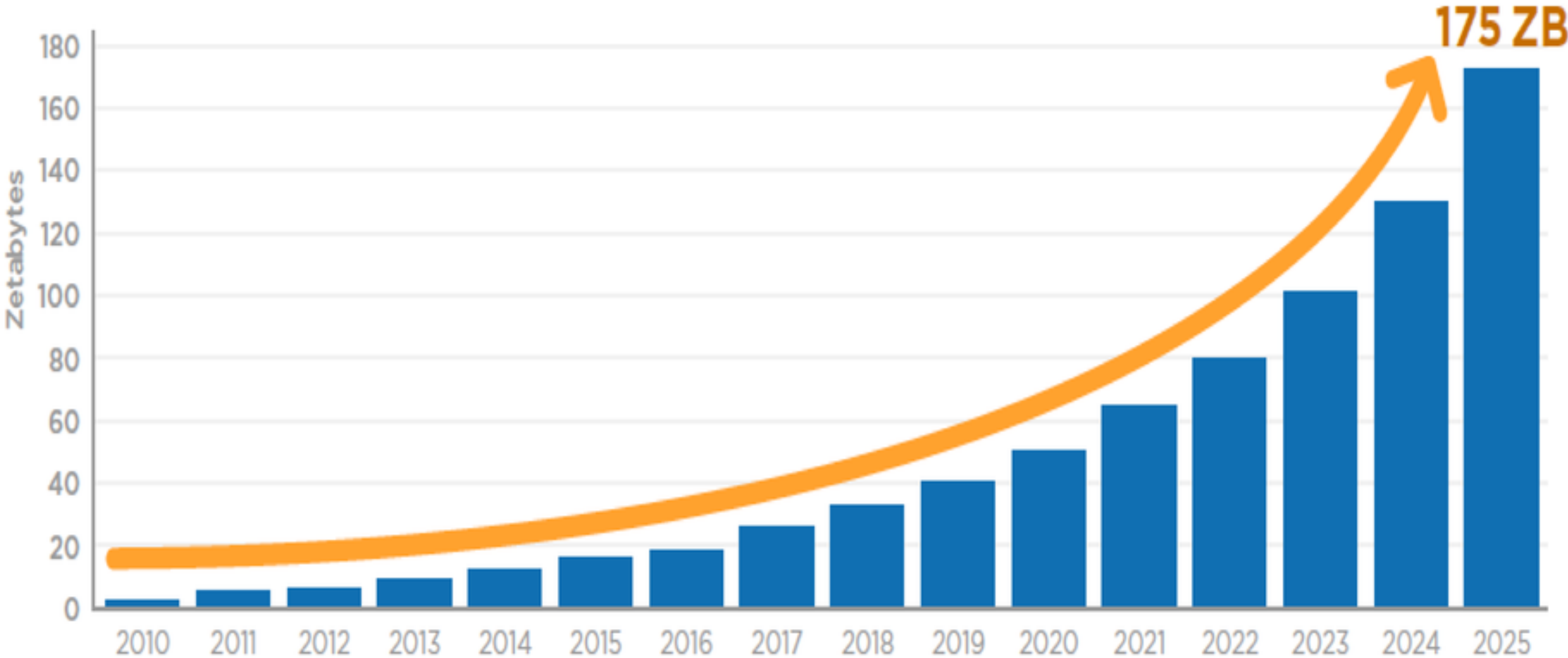
# 5 V's of Big Data

- Volume
- Velocity
- Variety
- Veracity
- Value



# Volume

- Data is rapidly increasing  
(due to cloud computing, mobile and more)



Source: Data Age 2025, sponsored by Seagate with data from IDC Global DataSphere, Nov 2018

Value	Metric
1000	kB kilobyte
1000 <sup>2</sup>	MB megabyte
1000 <sup>3</sup>	GB gigabyte
1000 <sup>4</sup>	TB terabyte
1000 <sup>5</sup>	PB petabyte
1000 <sup>6</sup>	EB exabyte
1000 <sup>7</sup>	ZB zettabyte
1000 <sup>8</sup>	YB yottabyte

# Volume

- **Data is rapidly increasing**  
(due to cloud computing, mobile and more)

**As of 2020, WhatsApp users send over 100 billion messages each day**

# Velocity

The speed at which data is generated

- Frequency of data generation (write)  
everything is measured
- Frequency of data processing (read)  
real time experience

# Variety

- **Structured data**  
info, transactions...
- **Semi structured data**  
logs, sensor data...
- **Unstructured data**  
images, video, audio...



# Veracity

The truthfulness or reliability of the data

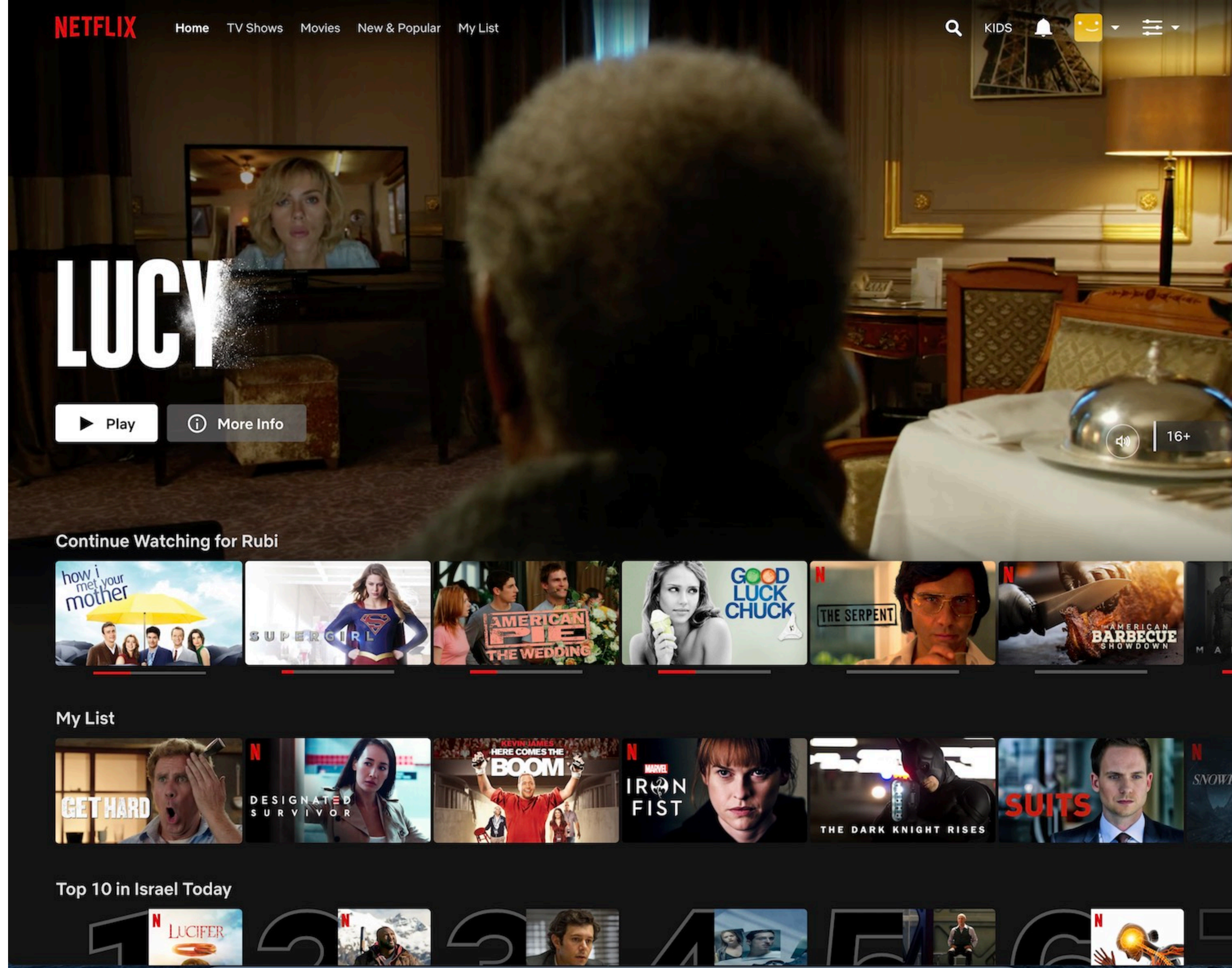
- data quality of captured data can vary greatly
  - bias
  - abnormalities
  - inconsistencies
  - duplication

# Value

The final result.

- which questions were answered
- hidden insights (machine learning)
- collecting data without use is, well, useless

- Volume
- Velocity
- Variety
- Veracity
- Value





# Cloud computing



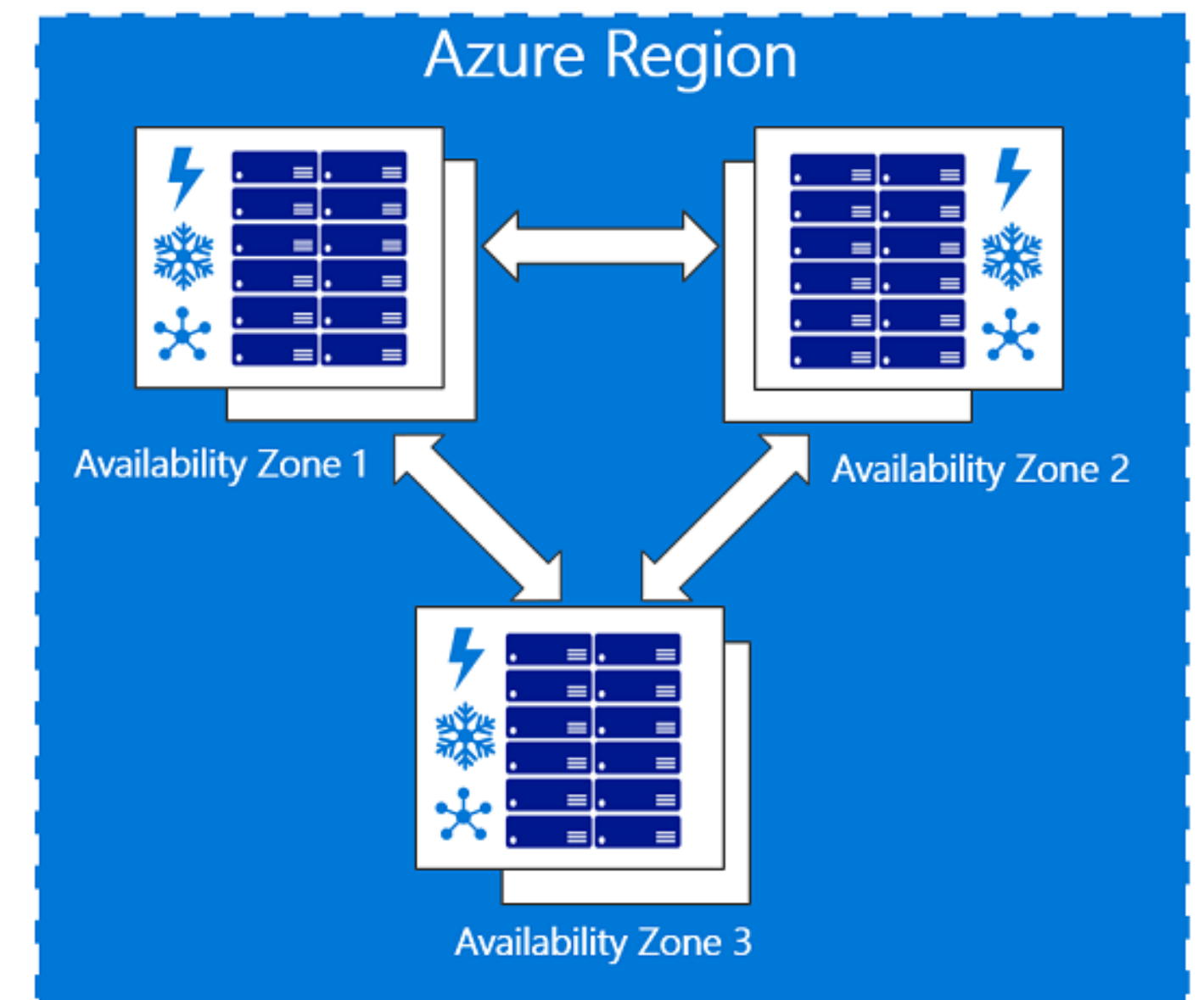
# Cloud computing

Not so relevant for the “regular Database course”, but for Big Data it is crucial

# Region / AZ / EL

- **Region**  
Cluster of data centers in a physical location
- **Availability Zone**  
a discrete data center with redundant power, networking, and connectivity in a Region
- **Edge Location**  
access to the network with limited services (usually CDN)

- (Names may vary between cloud providers)



## AWS Global Infrastructure Map

The AWS Cloud spans 105 Availability Zones within 33 geographic regions around the world, with announced plans for 12 more Availability Zones and 4 more AWS Regions in Germany, Malaysia, New Zealand, and Thailand.

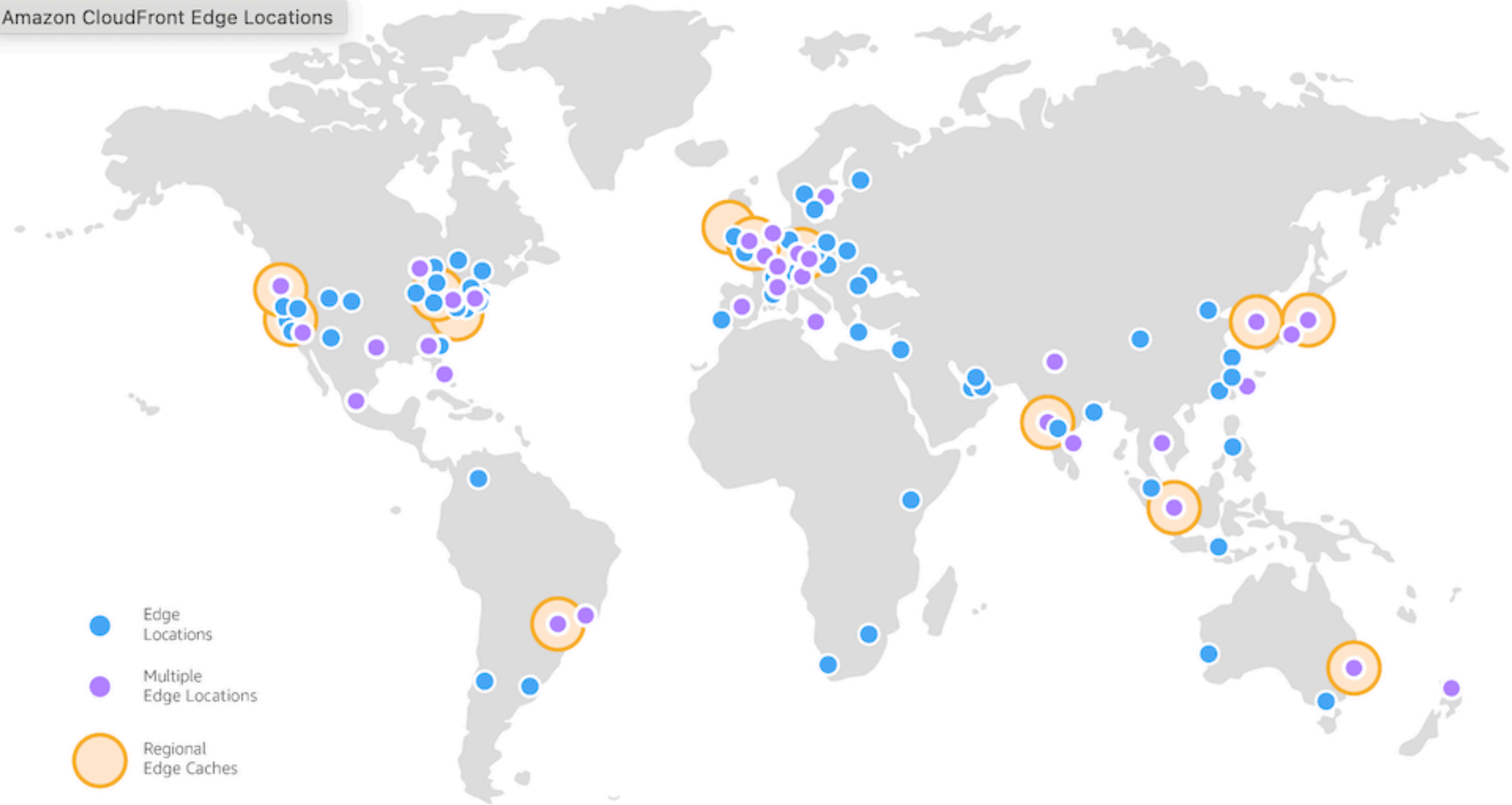


# AWS regions (January 2024)

List view

Regions Coming soon

Amazon CloudFront Edge Locations

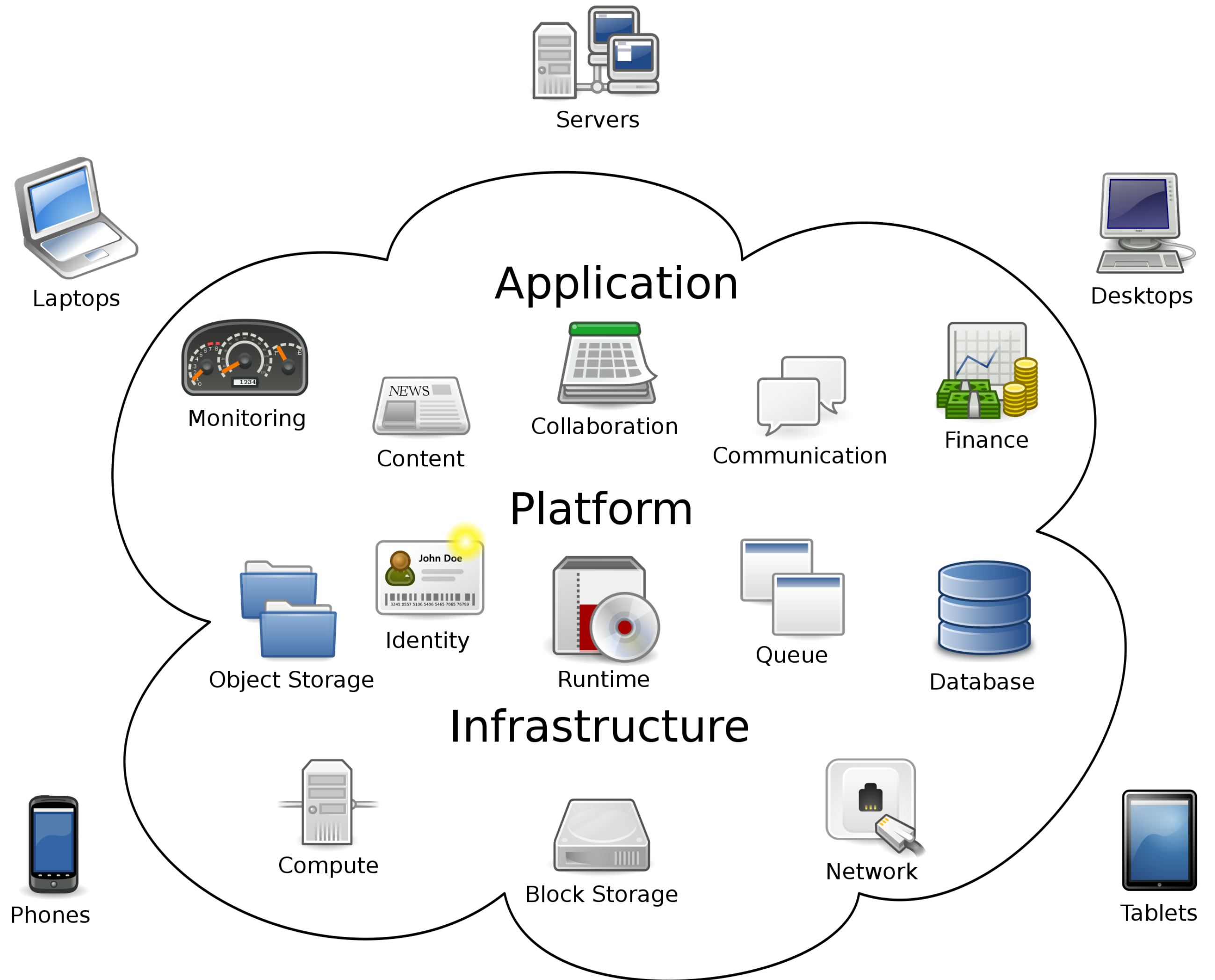


AWS edge locations (January 2024)



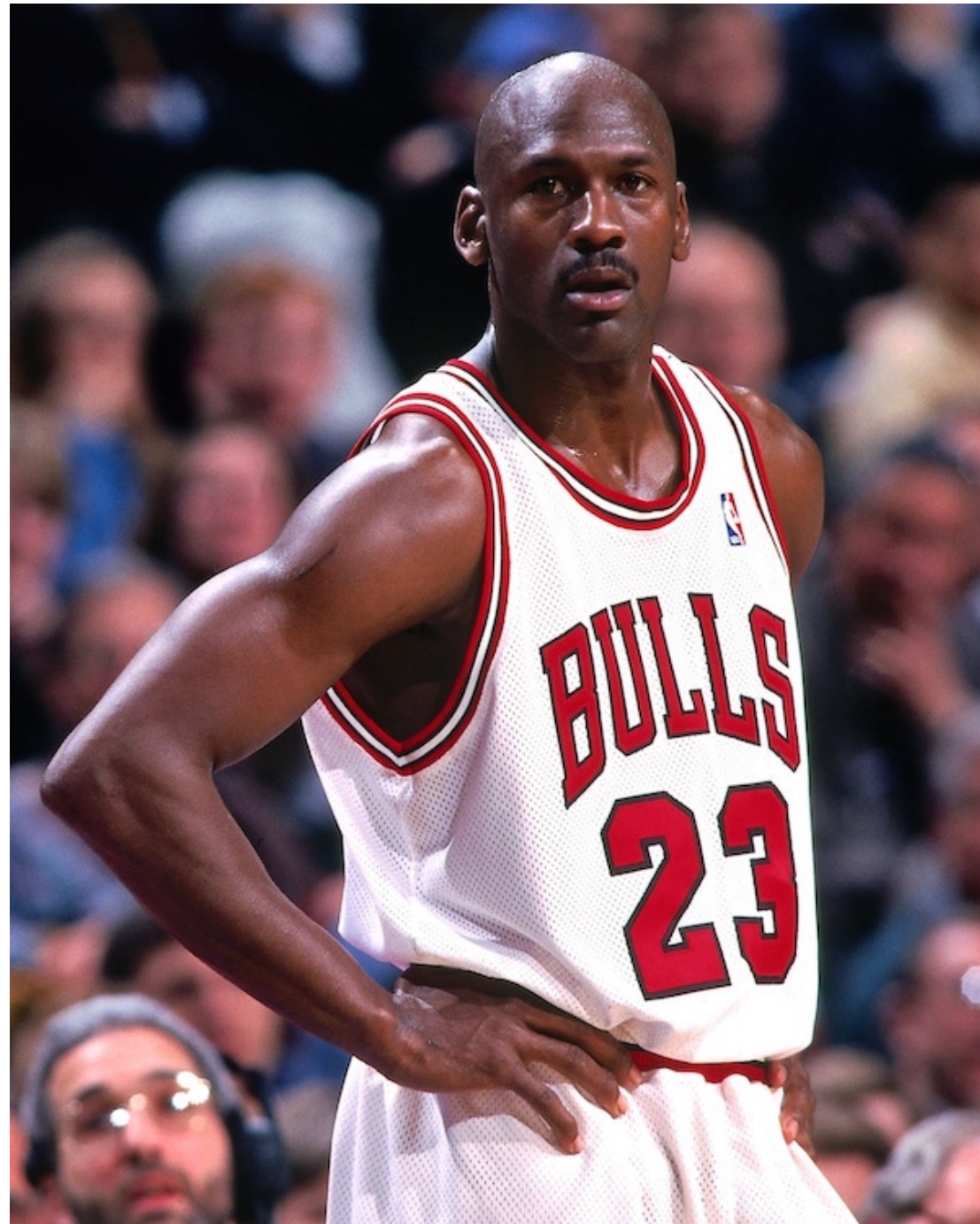
# Cloud computing

- **SaaS**  
software as a service
- **PaaS**  
platform as a service
- **IaaS**  
infrastructure as a service



**Highly Available / Highly Scalable**

# Mike orders a a basketball



Once clicked "order"

- Create order
- Check inventory
- Process payment
- Approve order
- Send to warehouse
- ...

## **System error**

fire / flood / electricity /  
hardware malfunction /  
software update...

# Possible outcomes

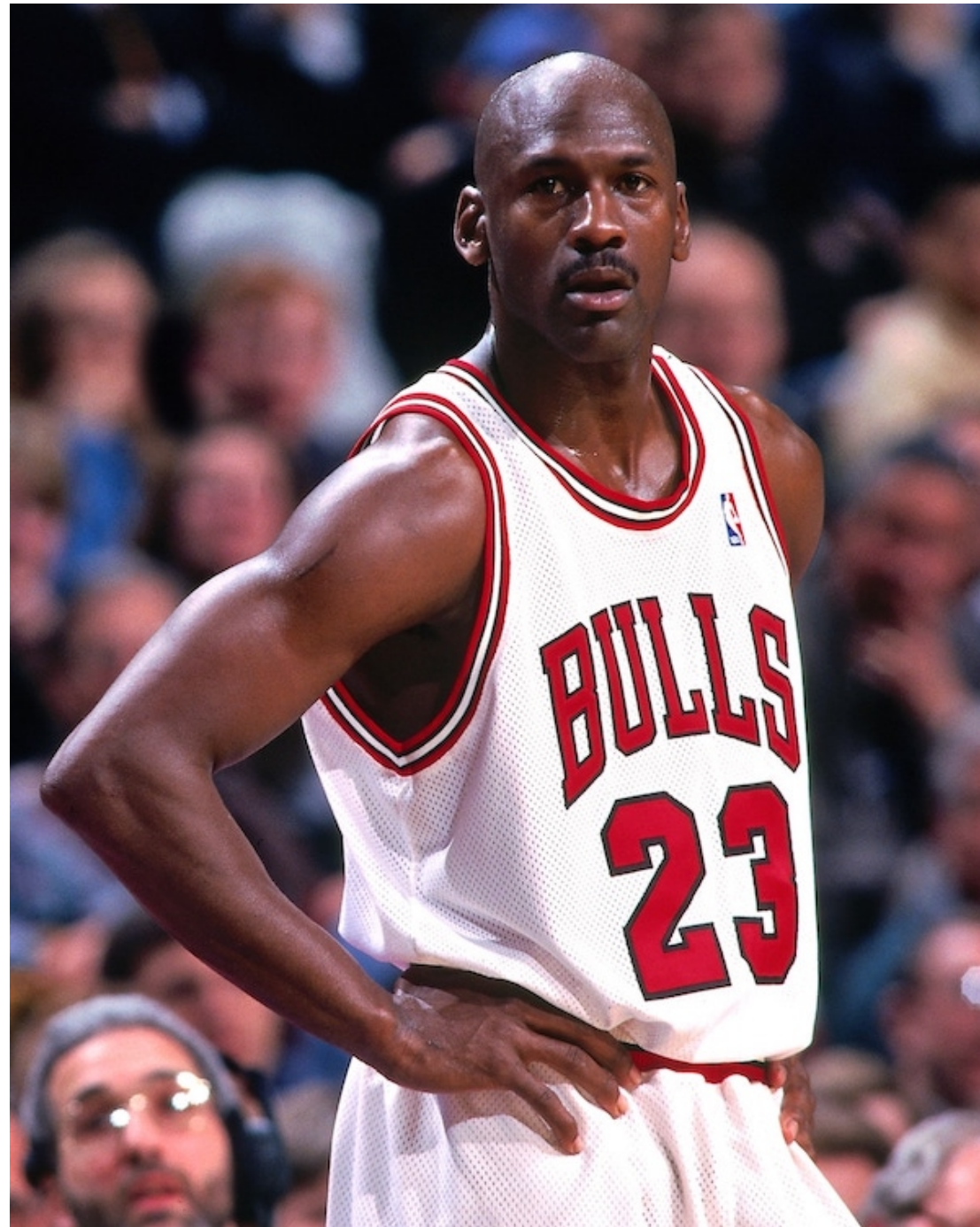
- Service disruption
- Data loss
- Data consistency
- Money lost (direct / reputation)
- A hard problem to solve for Databases  
disaster recovery:  
RTO (Recovery time) / RPO (Recovery point object)

# High availability

- “Nines”

<b>Availability</b>	<b>Downtime per day</b>	<b>Downtime per year</b>
<b>90%</b>	2.40 hours	36.53 days
<b>95%</b>	1.20 hours	18.26 days
<b>99%</b>	14.40 minutes	3.65 days
<b>99.9%</b>	1.44 minutes	8.77 hours
<b>99.99%</b>	8.64 seconds	52.60 minutes
<b>99.999%</b>	864.00 milliseconds	5.26 minutes
<b>99.9999%</b>	86.40 milliseconds	31.56 seconds

# Mike tweets about a basketball he bought



- Reach millions of users
- Millions of users try to buy the same basketball at the same time

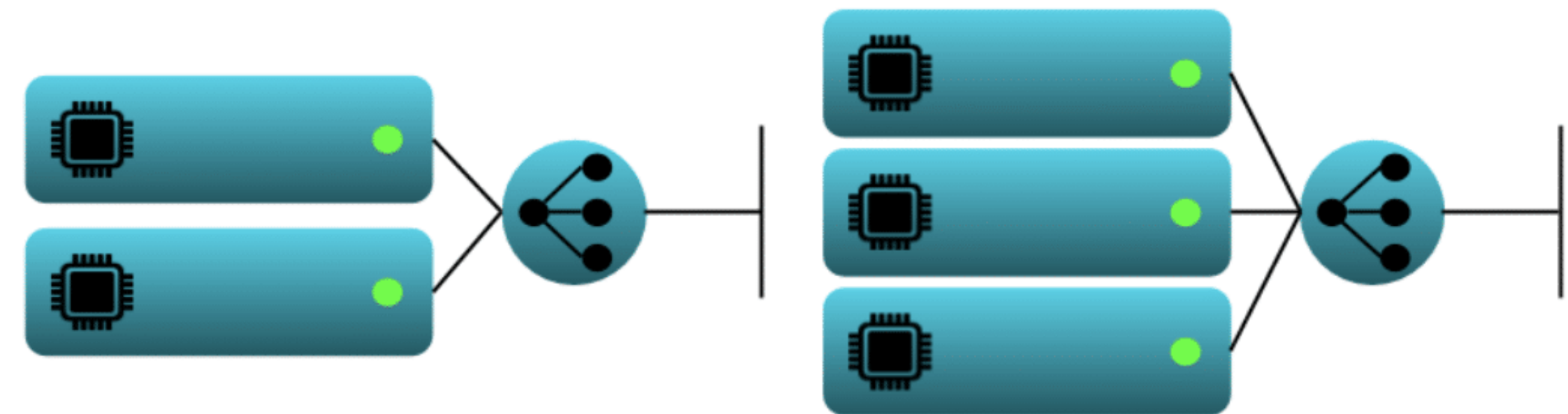
**System error**  
Too many requests

# High scalability

- Scale up vs scale out
- Commodity computing
- Stateless  
amazon's shopping cart is stateless?
- Microservices
- Sharding

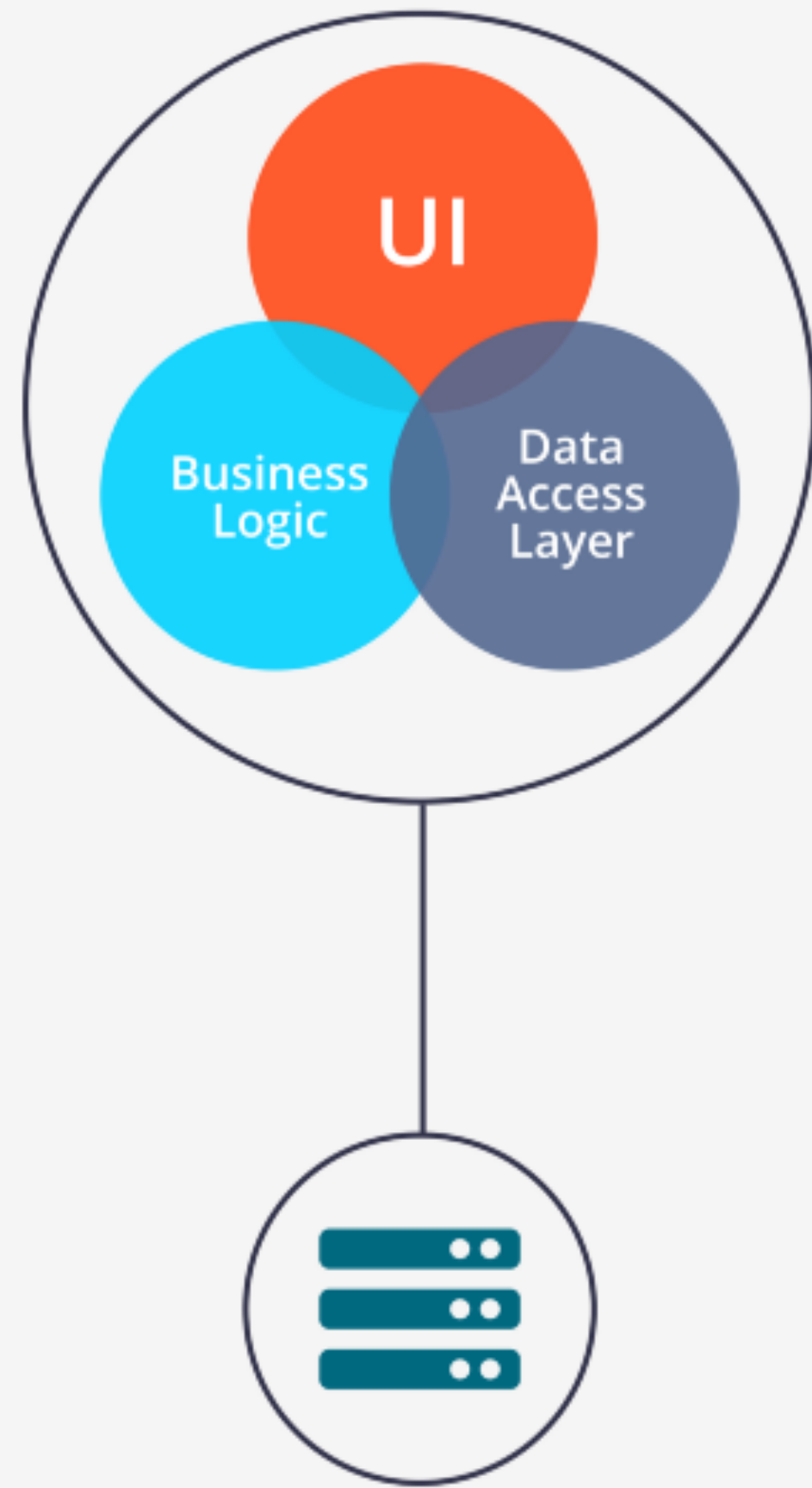


Scaling up from two to three CPUs

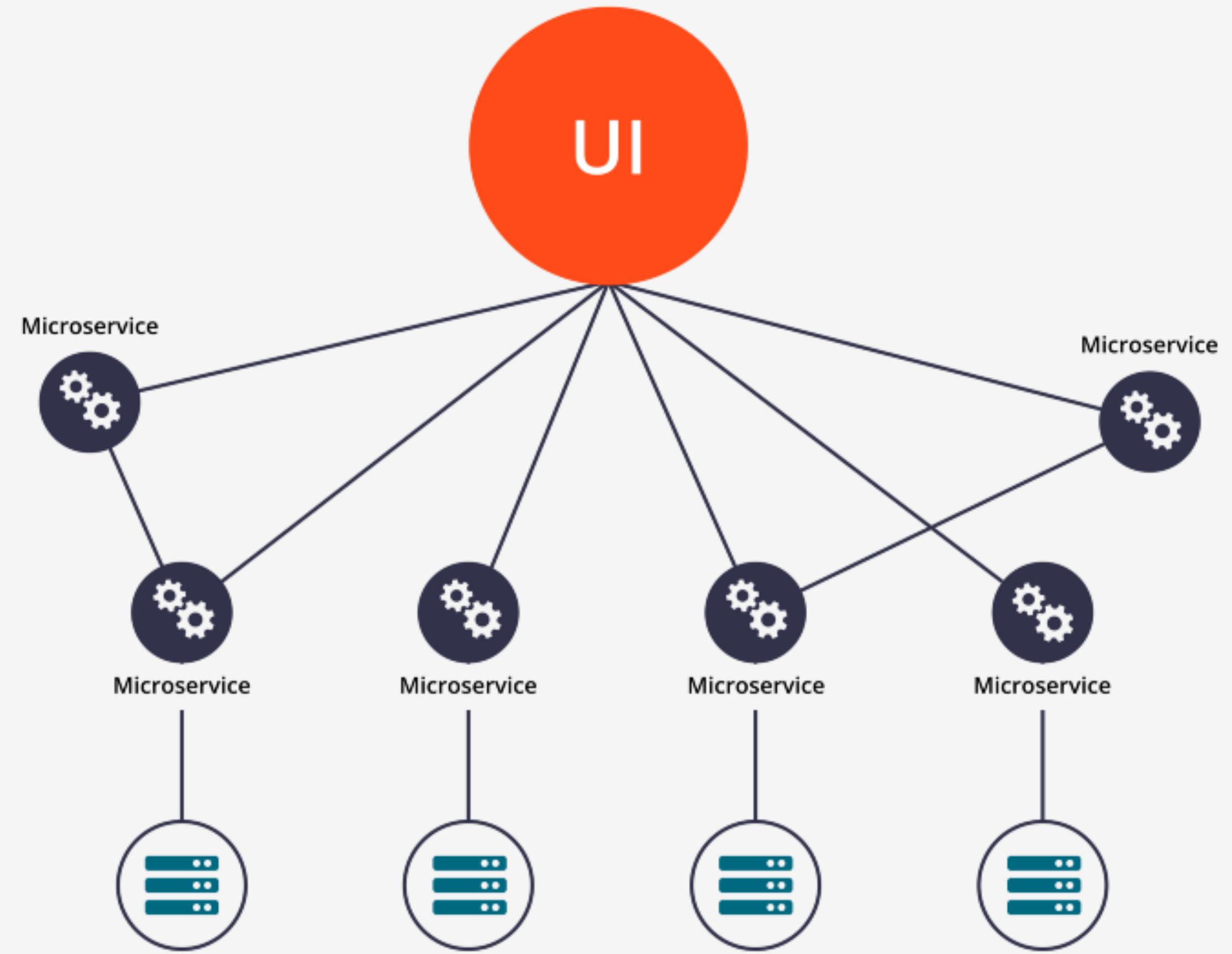


Scaling out from two to three CPU nodes

# Microservices



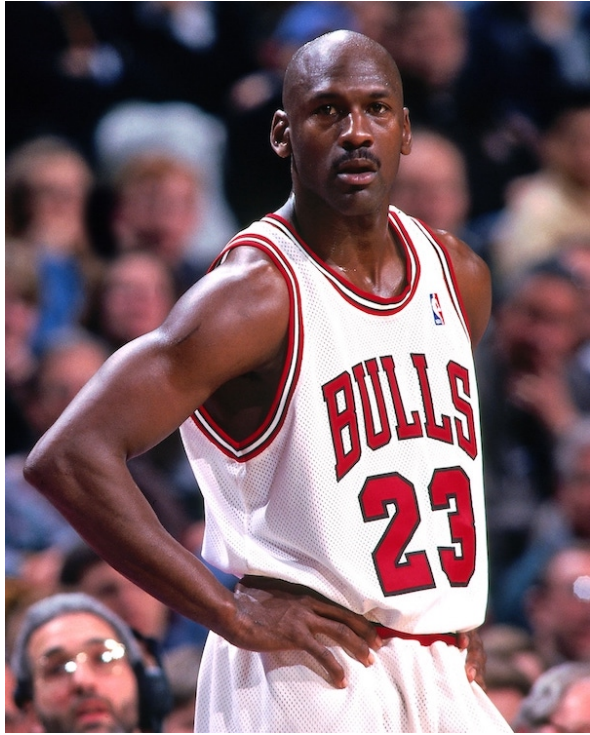
Monolithic Architecture



Microservice Architecture



# Ordering a basketball



Clicked order



Order creation

microservice  
highly scalable  
highly available

Inventory check

microservice  
highly scalable  
highly available

Process payment

microservice  
highly scalable  
highly available

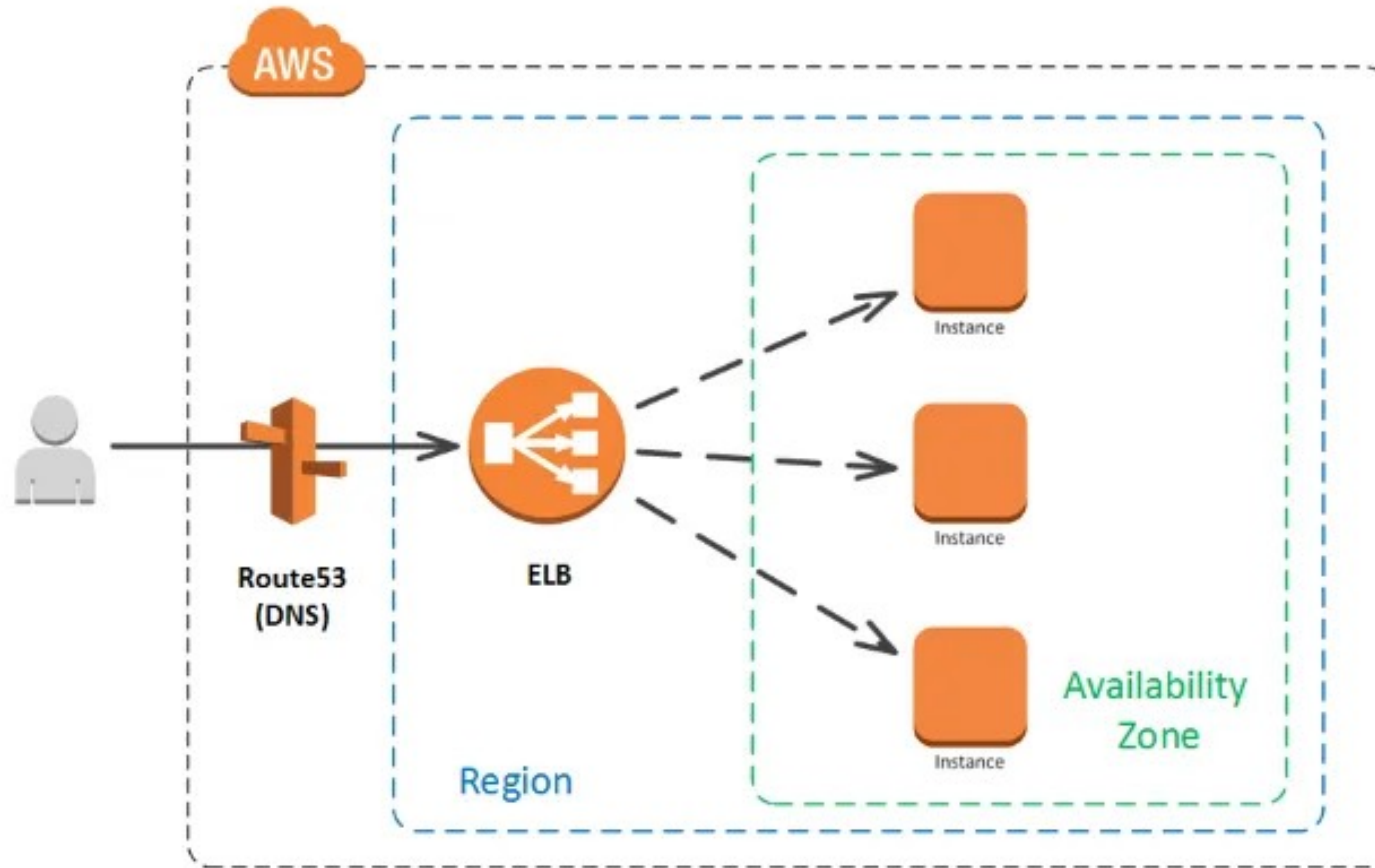
Send to warehouse

microservice  
highly scalable  
highly available

Order approve

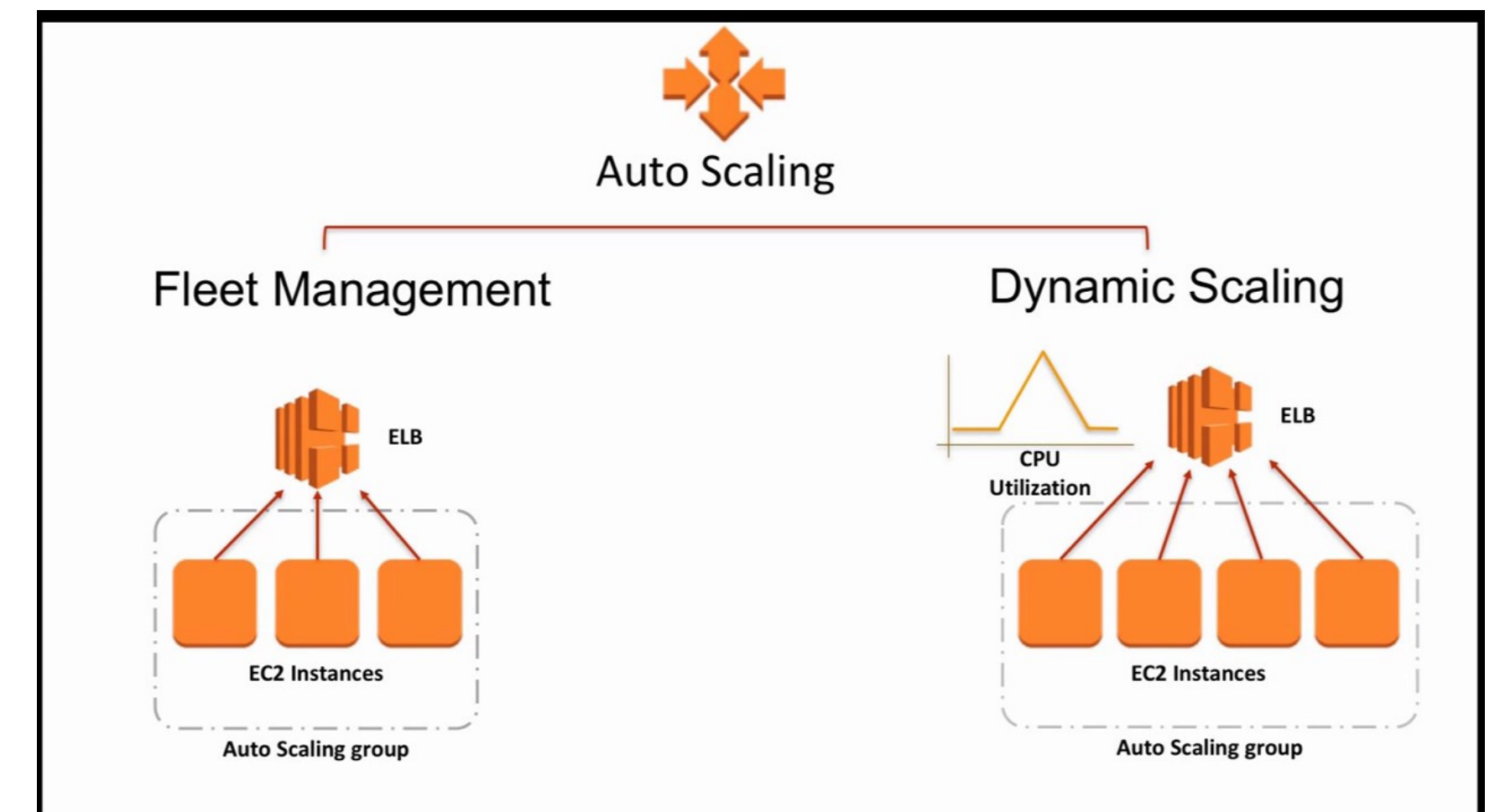
microservice  
highly scalable  
highly available

# Load balancer



# Auto scaling

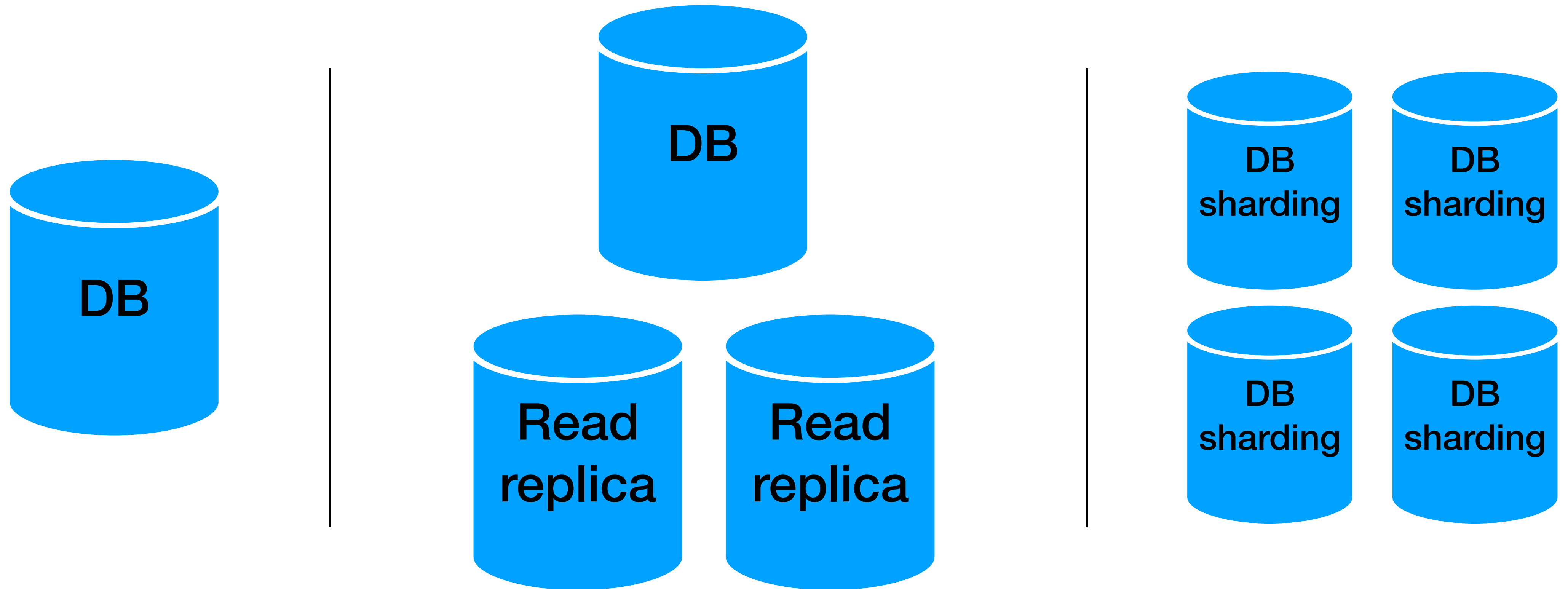
- When threshold occurs (hits / traffic / CPU...), create a new instance with the same logic and add to the load balancer
- When threshold drops, remove the from the load balancer and terminate the instance
- Usually requires stateless logic  
can Cassandra work with auto scale?



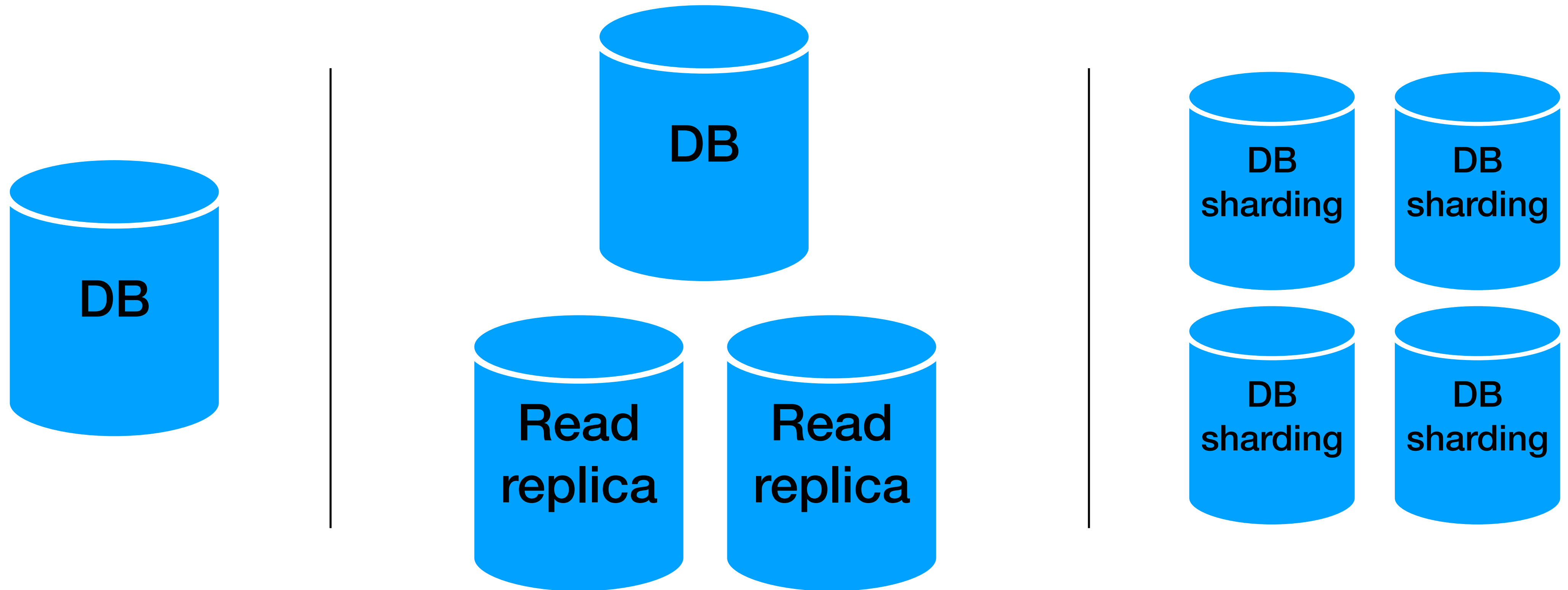
# Auto scaling - compute + storage?

- Some applications use both compute and storage (databases)
- Stateless?
- What happens when we scale down?

# Scaling databases



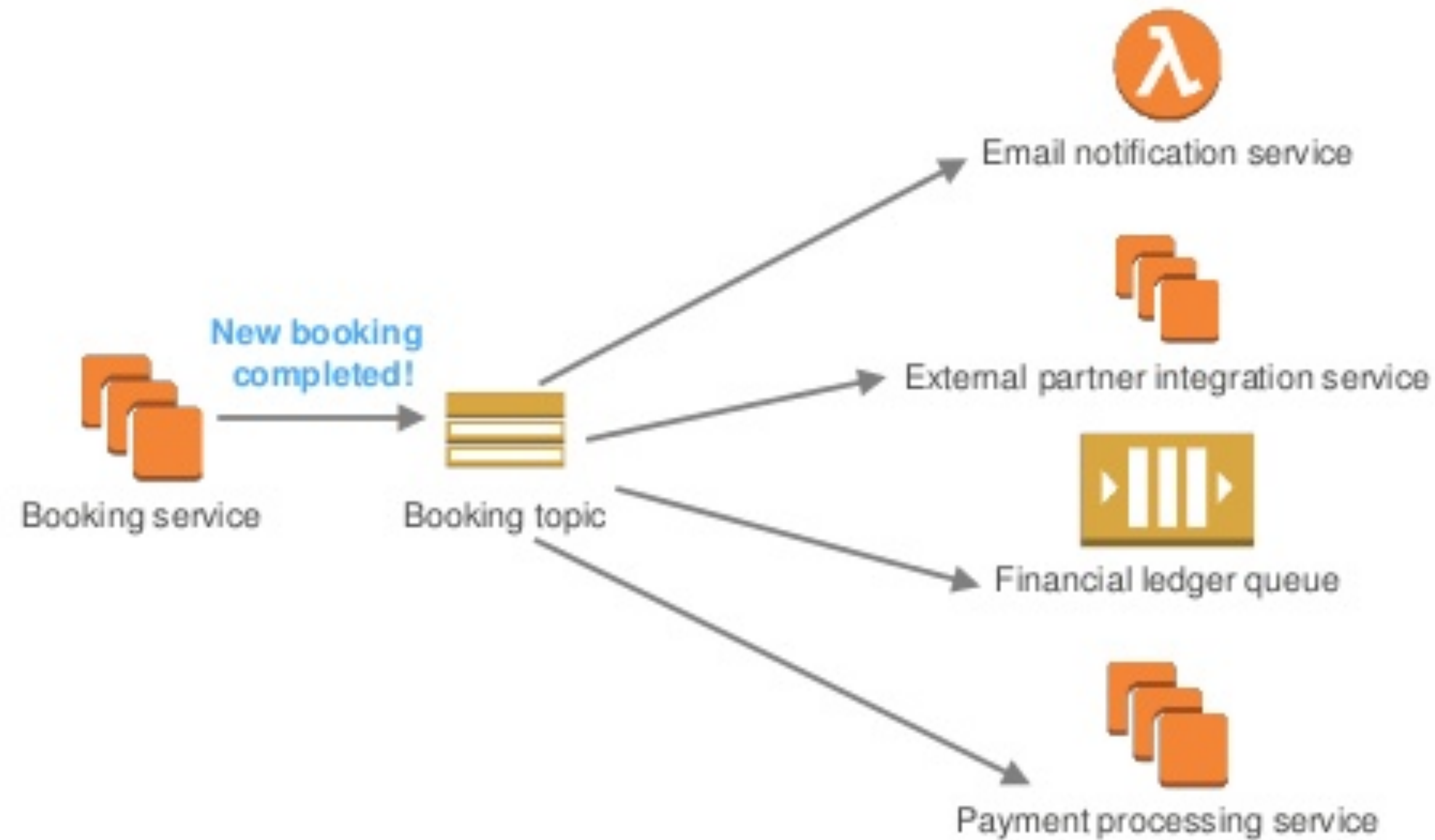
# Scaling databases



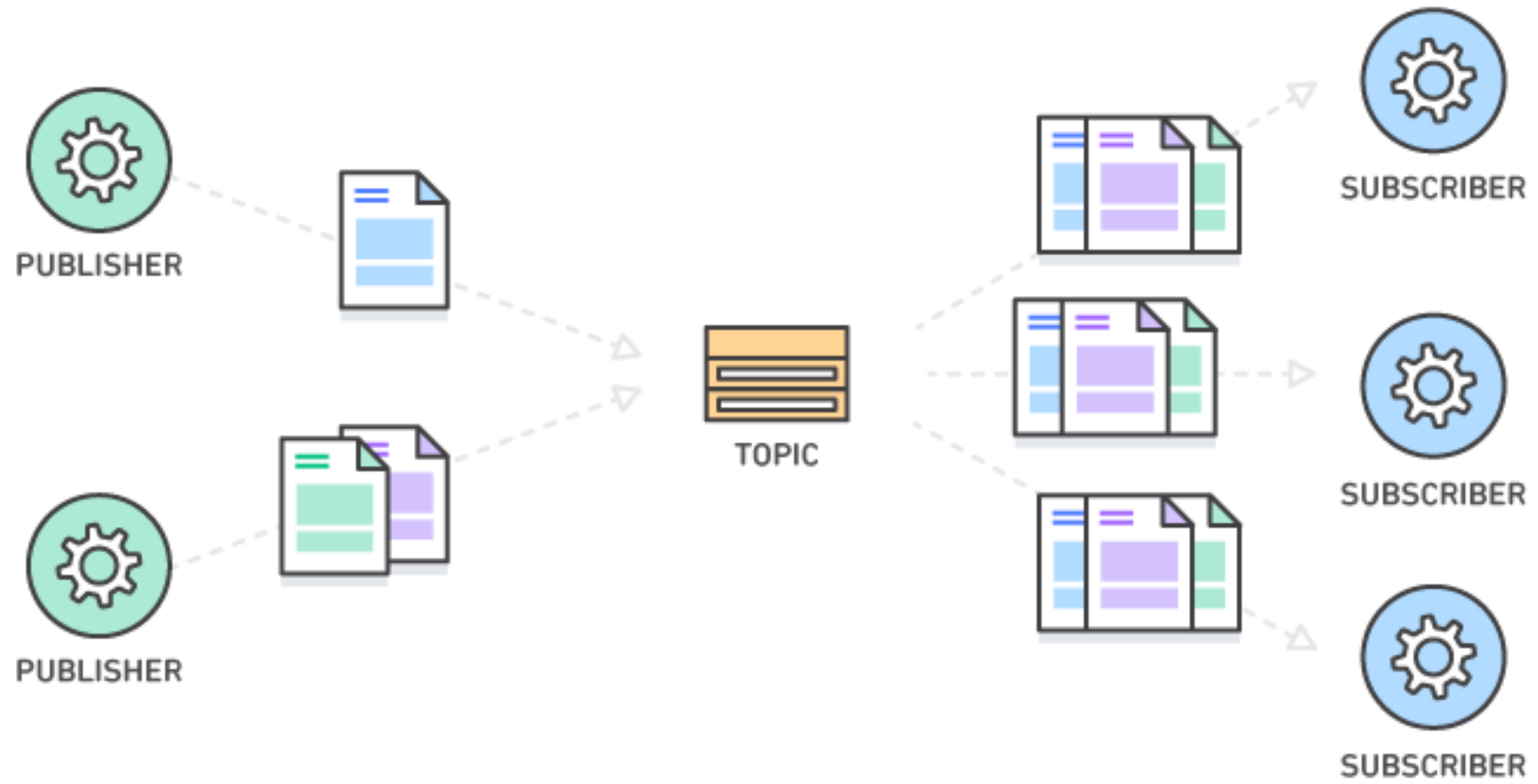
**Warning - we will talk about this a lot :)**

# Decoupling + event based services

- autonomous and unaware of each other services



# Pub sub





# Managed vs Unmanaged services

# Unmanaged service

You are responsible for everything!

- Choosing CPUs, storage, network...
- Installing OS, Java, core software, dependencies...
- Patches, updates
- Security
- Backup
- Monitoring
- Availability

# Unmanaged service (2)

Requires different skills

- System
- DevOps
- ...

# Managed service

- All the stuff we talked about before are managed for you out of the box
- Hardware utilization
- Focus on stuff that really matters for you
- Cost?

# Managed service cons

- **Cloud locked in**
- Slightly limited functionality
- Works only in the cloud
- **Cost?**  
(cheaper to go unmanaged on large scale, but a lot of headaches)

# In practice

- Some will be managed and some not
  - VMs
  - load balancers
  - network stuff
  - ...
- **To go managed or unmanaged with databases is a good question**

# Managed vs Unmanaged Databases

## Fully managed services on AWS

Spend time innovating & building new apps, not managing infrastructure

Self managed

You

Schema design  
Query construction  
Query optimization  
Automatic failover  
Backup & recovery  
Isolation & security  
Industry compliance  
Push-button scaling  
Automated patching  
Advanced monitoring  
Routine maintenance  
Built-in best practices

Fully managed

You

AWS

# But how managed service work?

- It is just someone else's software...
- Do we need to understand how it works behind the scenes?



**For databases, YES!**

# Big Data databases

- Managed big data databases are built on, well, big data databases
- **Data modeling is crucial.**  
(with bad modeling, nothing will work)

**To model data correctly,  
we need to understand the technology**  
(it is not just reading the API docs)

# Experiment (Sharding)

8:52 90

Live Forever  
Oasis

UP NEXT LYRICS RELATED

Playing from  
Live Forever Radio Save

All Familiar Discover Popular Deep cuts

- Live Forever  
Oasis • 4:37
- Mr. Brightside  
The Killers • 3:43
- Don't Look Back in Anger  
Oasis • 4:50
- Iris  
The Goo Goo Dolls • 4:50
- Learn to Fly  
Foo Fighters • 3:56
- Common People  
Pulp • 5:52
- High And Dry  
Radiohead • 4:18
- If I Had a Gun...  
Noel Gallagher's High Flying Birds • ...
- Snow (Hey Oh)

9:04 89

All Too Well (10 Minute...  
Taylor Swift

UP NEXT LYRICS RELATED

Playing from  
All Too Well (10 Minute Version) (Taylor's Version) [From The Vault] Save

All Familiar Discover Popular Deep cuts

- All Too Well (10 Minute Version)  
[Taylor's Version] [From The Vault]  
Taylor Swift • 10:13
- Rise  
Belle Mariano • 3:59
- You Are In Love (Taylor's Version)  
Taylor Swift • 4:28
- Falling  
Harry Styles • 4:01
- Let Her Go  
Passenger • 4:13
- We Don't Talk Anymore (feat. Selena Gomez)  
Charlie Puth • 3:38
- Everybody Rise (Acoustic)  
Amy Shark • 3:47
- You're Just A Boy (And I'm Kinda The Man)  
Maisie Peters • 3:06
- Your Song (Bonus Track)