BIGDATA SI SCALABILITATE

Curs introductiv


hadoop


amazon
web services™

Terminologie

BigData

Scalabilitate

Performanta

Scalare pe verticala

Nestructurate

Scalabilitate pe orizontala

Big data

Zeci de TB

Exemple:

- Logurile de access la facebook
- Toate cartile lumii
- Conversatiile telefonice ale unei tari intr-o zi

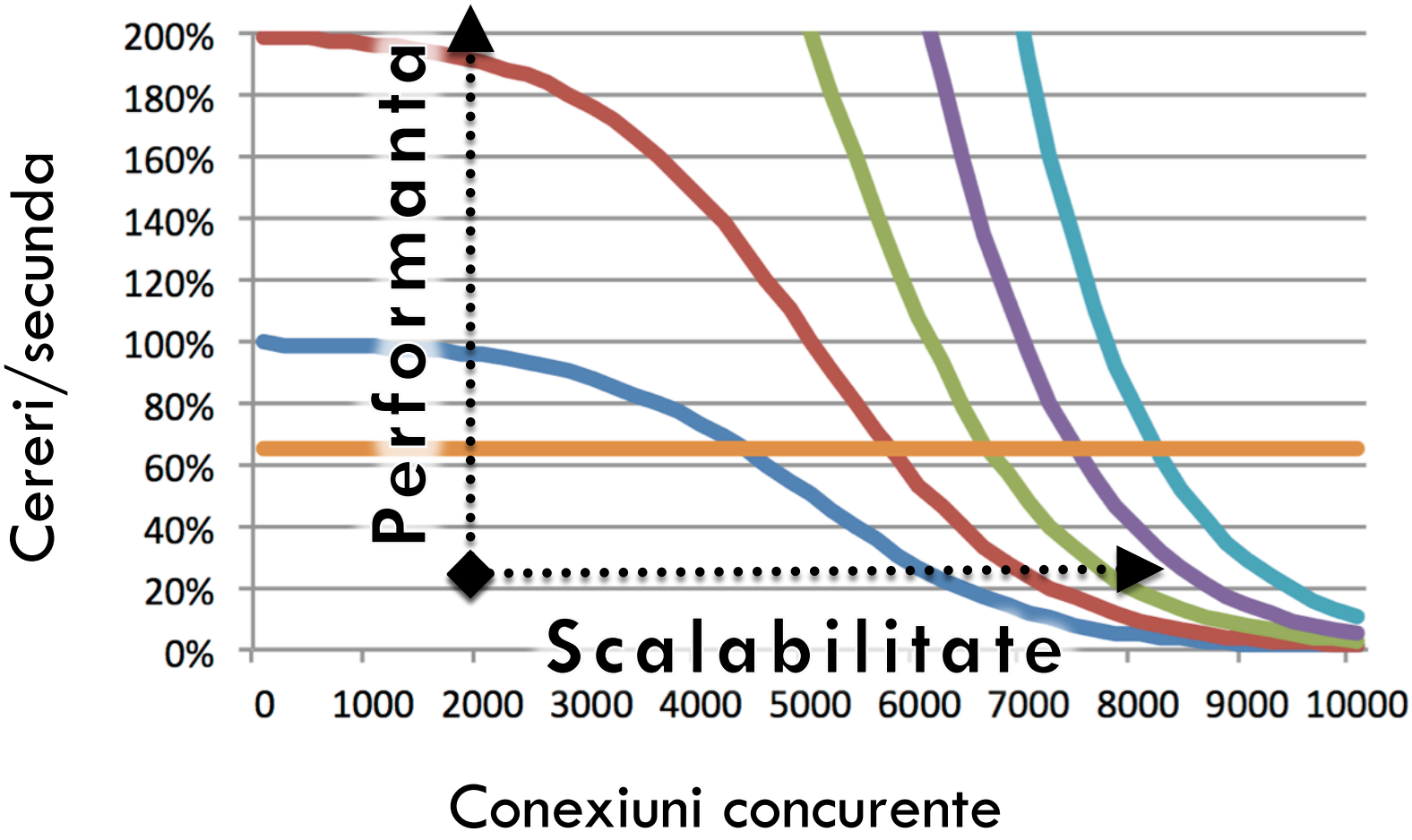
Scalabilitate si performanta

Sistem mai performant: capacitate ridicata de a procesa o singura operatie in timp mai scurt

Scalabilitate si performanta

Sistem scalabil: capacitatea de a procesa foarte multe operatii in paralel fara ca performanta sa scada drastic

Scalabilitate si performanta



Scalabilitate si performanta

Scalabilitate != performanta

Sunt ortogonale

Scalabilitate verticala si orizontala

- Verticala: Imbunatatirea hardware-ului pe care ruleaza un proces
  - Se bazeaza pe strategia: arunca cu bani in infrastructura
- Orizontala: Impartirea procesului pe mai multe masini

Date structurate si nestructurate

- Structurate: Date tabelare a caror forma este descrisa de o schema. Indexate.
  - MySQL
  - Oracle
  - etc
- Nestructurate: Restul!
  - Loguri
  - Pozele din vacanta
  - etc

Date structurate si nestructurate

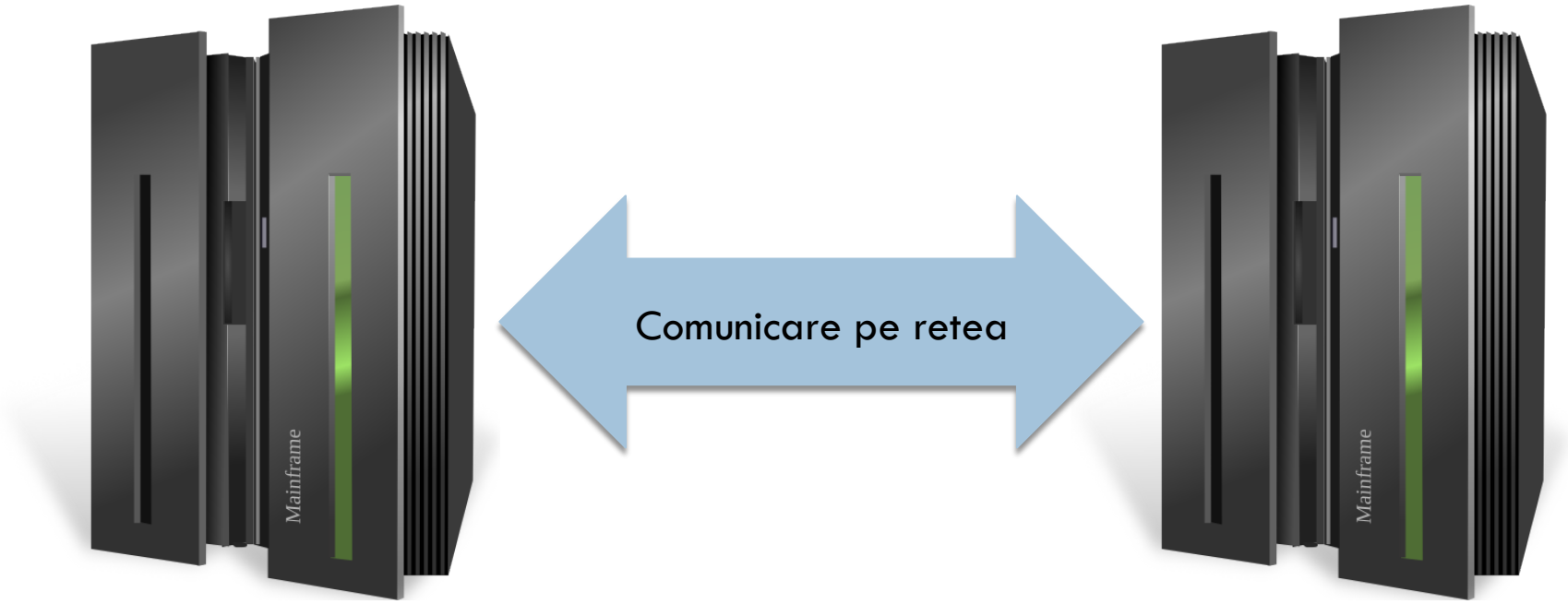
- Satisfac nevoi diferite
- Fiecare are avantaje si dezavantaje
- Alegerea depinde de natura problemei


hadoop

Modelul clasic

- Data warehouse
- Masini enterprise care tin datele structurat
  - Oracle
  - MySQL
  - etc
- Calculele ruleaza pe masini dedicate

Modelul clasic

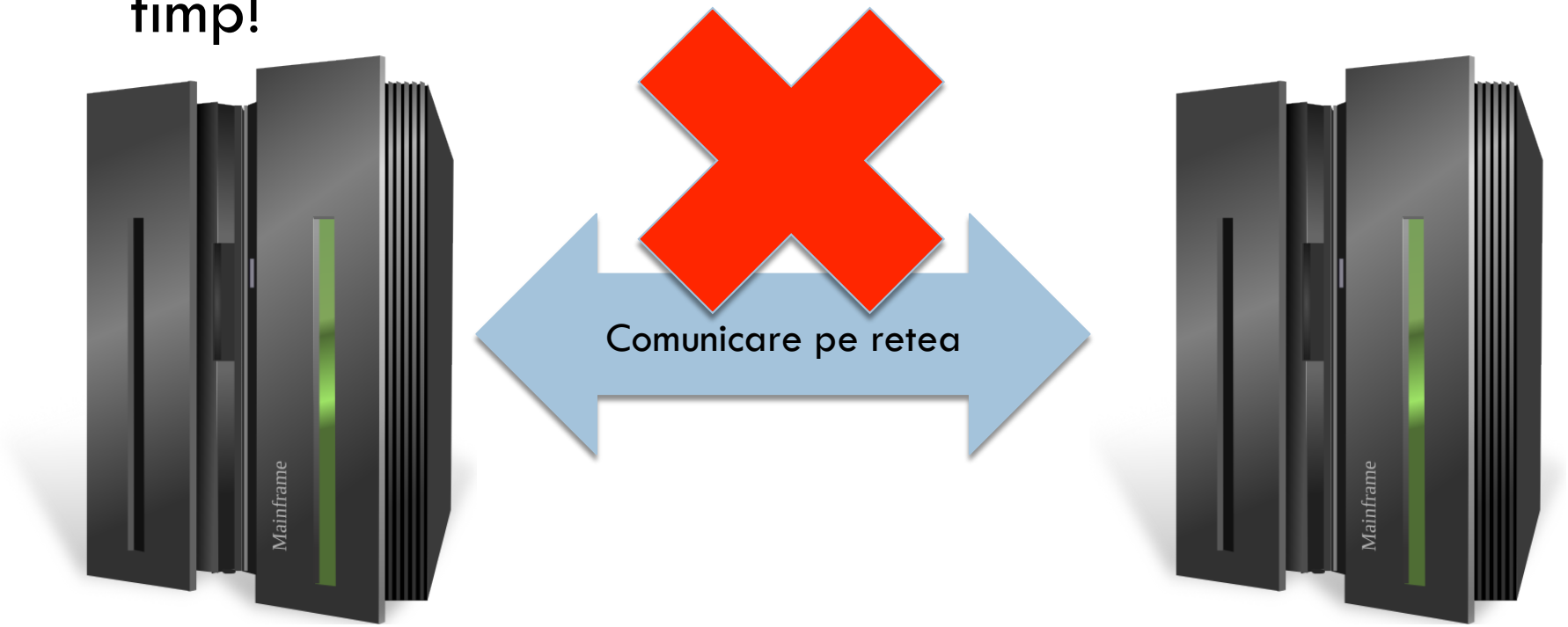


Baza de date

Masini de procesare

Modelul clasic

Transmiterea informatiei din baza de date catre masinile unde este procesata consuma foarte mult timp!



Baza de date

Masini de procesare

Modelul clasic

Solutie: Datele si prelucrarea lor trebuie sa coexiste pe aceeasi masina!

Cluster Hadoop

Date +
Prelucrare

Originile hadoop

- Bazat pe 2 paper-uri publicate de Google
  - Google Map Reduce
  - Google File System (GFS)
- Creat de catre Doug Cutting si Mike Cafarella
  - Cutting l-a denumit dupa numele elefantului de jucarie al fiului lui
- Hadoop este o implementare open source a celor 2 paper-uri

Componente hadoop

- Hadoop Distributed File System (HDFS)
- Hadoop Map reduce
- Functioneaza si independent

HDFS

- Sistem distribuit de fisiere construit peste sistemul de fisiere al sistemului de operare
- Redundanta ridicata a datelor
- Disponibilitate ridicata

HDFS: Tipuri de noduri

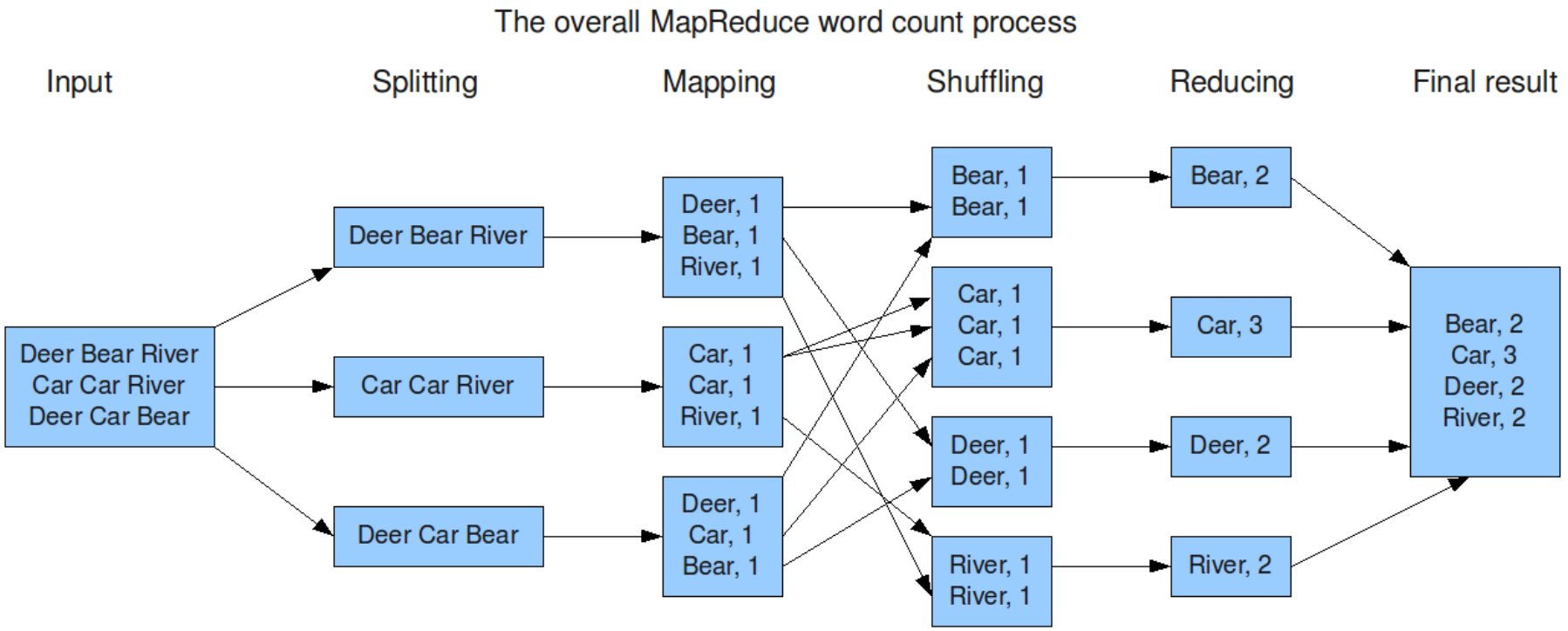
- Name node:
  - Tine minte lista de fisiere si masinile pe care se afla
  - Unul singur intr-un cluster
- Data node:
  - Stocheaza datele
  - Multiple masini intr-un cluster

HDFS: Exemplu

MapReduce

- Framework distribuit de procesare a datelor
- Are la baza 2 functii:
  - Map: Imparte informatia in bucati atomice (linii, cuvinte etc) si ii asociaza o cheie. Toate unitatile de informatie cu aceeasi cheie ajung sa fie trimise la aceeasi masina
  - Reduce: Primeste de la Map lista de informatie si cheile si o proceseaza.

MapReduce exemplu



Observatii

- Poate rula pe clustere formate din masini comerciale (pret redus)
- Este ca un tren (TGV)

Scalabilitate si AWS

Arhitectura unei aplicatii web

- O masina ce ruleaza:
  - Un server http: apache, tomcat, nginx
  - Aplicatia in containerul web
  - O baza de date: MySQL

Arhitectura unei aplicatii web

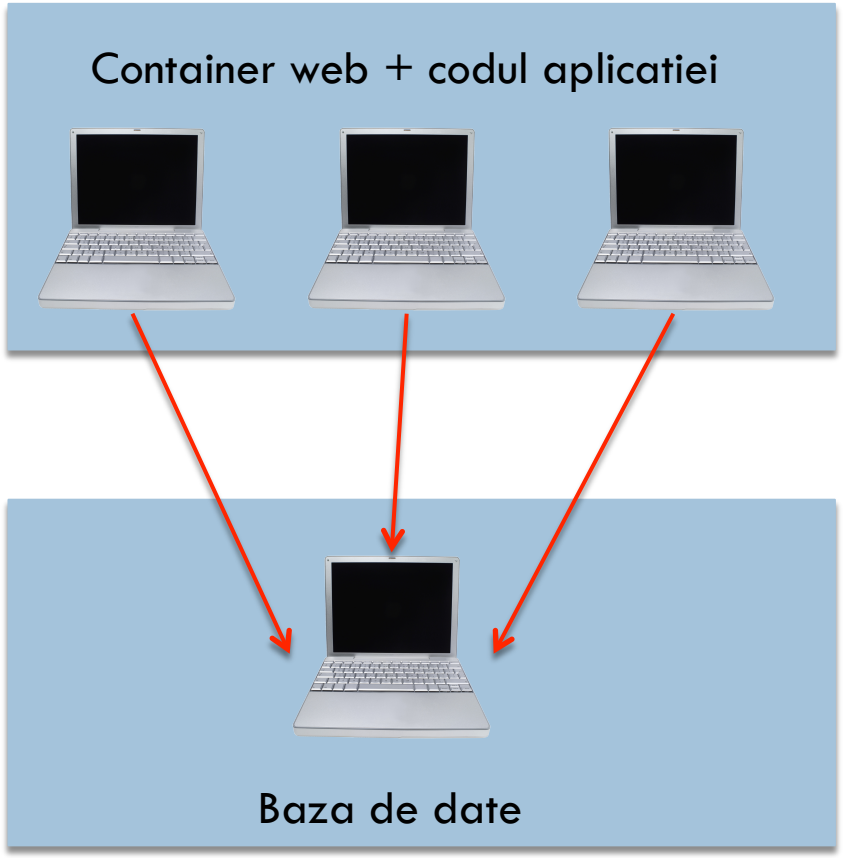
Oricat de mult hardware ai pune pe masina, serverul nu poate servi 1.000.000 de cereri pe secunda

Arhitectura unei aplicatii web

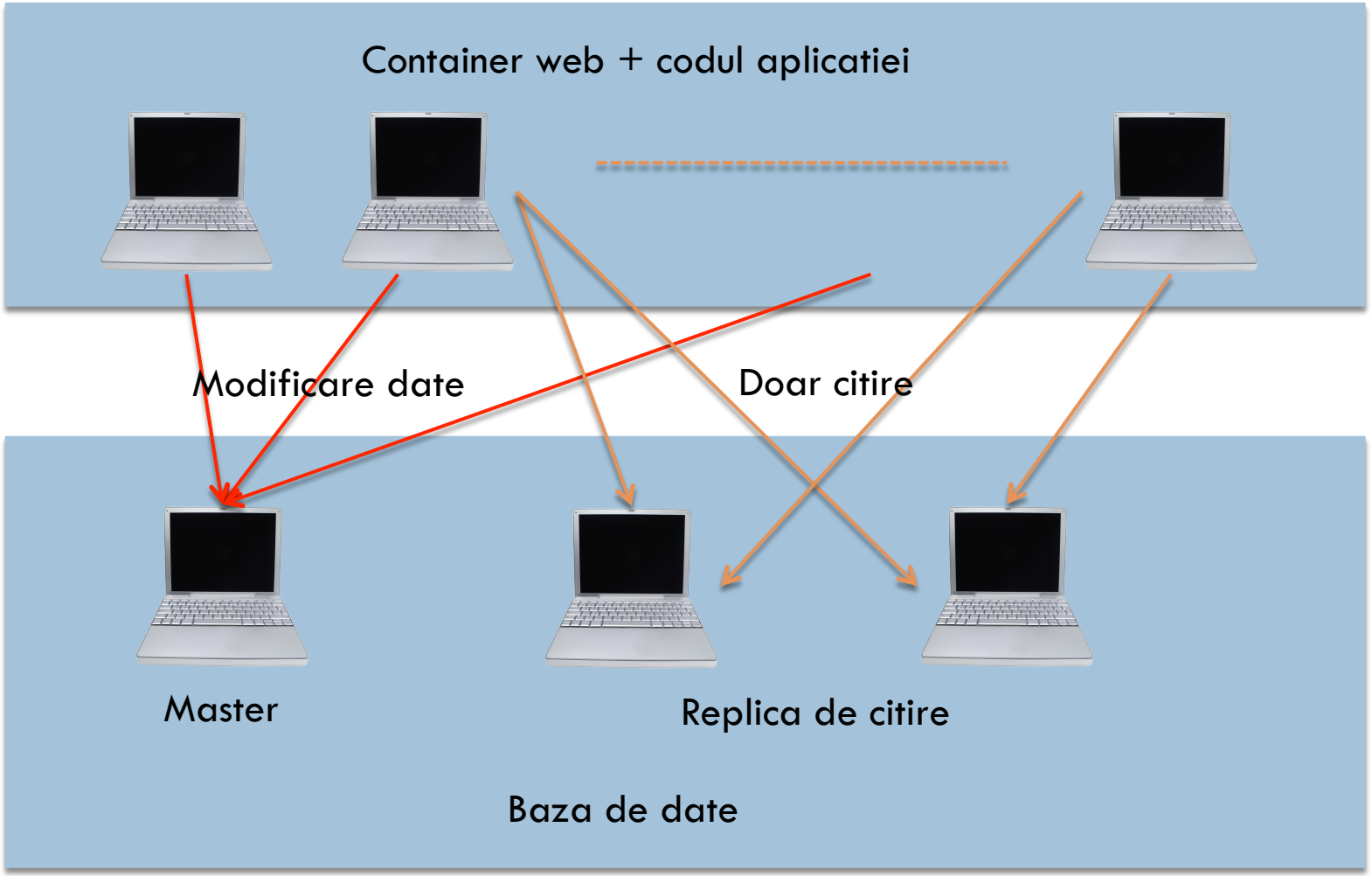
Solutie: Scalare pe orizontala.

Mai multe masini, care raspund la fel pentru aceeasi cerere.

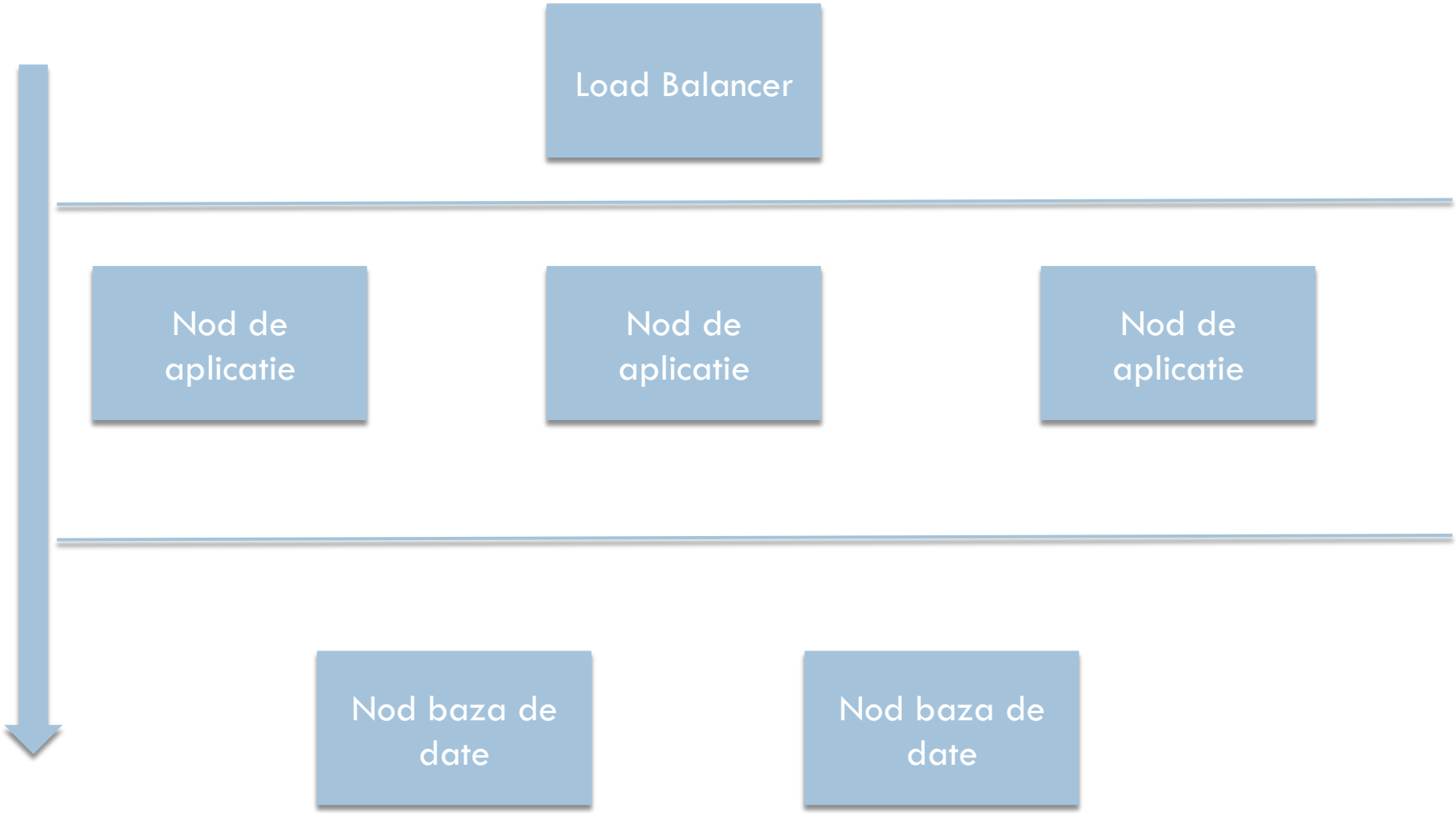
Layere in aplicatie



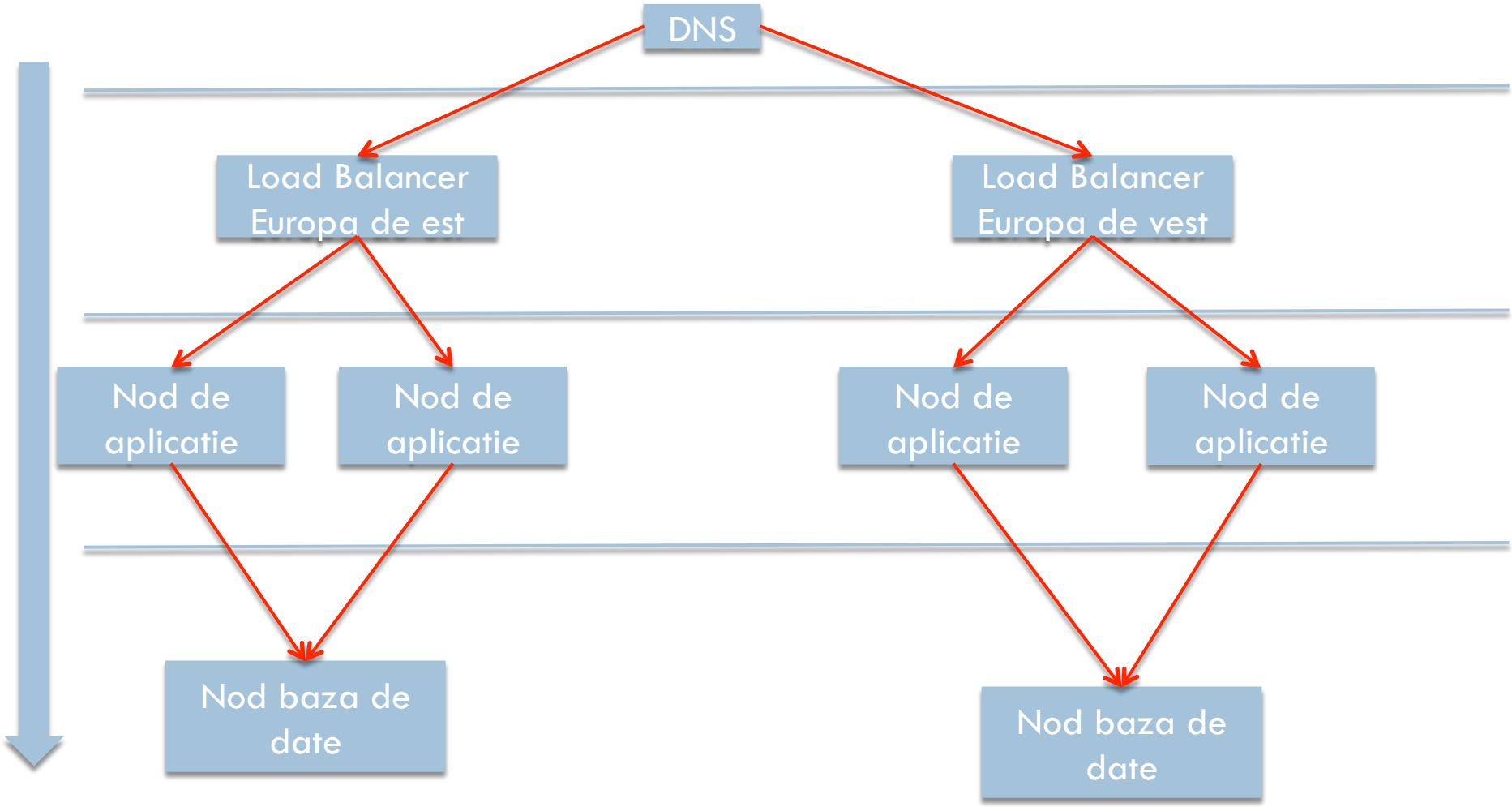
Imbunatatirea bazei de date



Load balancer



Controlul unei regiuni



Servicii AWS

Ofera componentele necesare pentru a implementa o arhitectura ca cea precedenta:

- EC2 – Elastic Cloud 2
- ELB – Elastic load balancer
- Route53 – DNS service
- S3 – Simple Storage Service
- SQS – Message Queue Service
- RDS – Managed Relational Database Service
- DynamoDB – Predictable and scalable NoSQL Store

Alte exemple de infrastructura

- Serviciu pentru conversie video
- Serviciu pentru schimb de fisiere

Alte caracteristici importante

- Viteza cu care sistemul de acomodeaza la noul trafic
- Disponibilitatea sistemului
- Securitate
- Arhitectura interna a aplicatiei

Intrebari