Big Data. Scalability and models

The value of data -- the need for computing

In science since long ago -- generalized in last years

Scalability

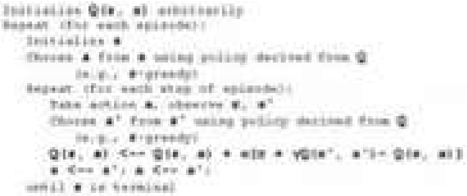
How is it achieved? At what cost? What do we gain?

What hardware architectures? What programming models?

Scaling computing

Infraestructura de cómputo

Algoritmos/Aplicaciones

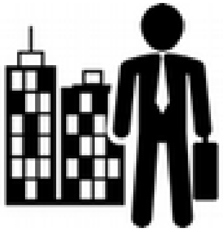


```
Initialize Q(s, a) arbitrarily
Repeat (for each episode):
    Initialize s
    Choose a from s using policy derived from Q
         (e.g., ε-greedy)
    Repeat (for each step of episode):
        Take action a, observe r, s'
        Choose a' from s' using policy derived from Q
             (e.g., ε-greedy)
        Q(s, a) <-- Q(s, a) + α[r + γQ(s', a')- Q(s, a)]
        s <-- s'; a <-- a';
    until s is terminal
```

Datos

Big Data

HPC



Recurso humano

map reduce

```
2012-01-01 09:08 BOG Libros 88.56 Discover
2012-01-01 09:09 BGA Libros 337.71 Efectivo
2012-01-01 09:52 BGA Libros 62.41 Discover
2012-01-01 10:08 MED Musica 93.37 Visa
2012-01-01 10:22 BGA Musica 369.94 MasterCard
2012-01-01 10:58 MED Musica 119.12 Efectivo
2012-01-01 11:36 BOG Musica 296.76 Discover
```

map reduce

```
2012-01-01 09:08 BOG Libros 88.56 Discover
2012-01-01 09:09 BGA Libros 337.71 Efectivo
2012-01-01 09:52 BGA Libros 62.41 Discover
2012-01-01 10:08 MED Musica 93.37 Visa
2012-01-01 10:22 BGA Musica 369.94 MasterCard
2012-01-01 10:58 MED Musica 119.12 Efectivo
2012-01-01 11:36 BOG Musica 296.76 Discover
```

1

```
BOG 88.56
BGA 337.71
BGA 62.41
MED 93.37
BGA 369.94
MED 119.12
BOG 296.76
```

map reduce

```
2012-01-01 09:08 BOG Libros 88.56 Discover
2012-01-01 09:09 BGA Libros 337.71 Efectivo
2012-01-01 09:52 BGA Libros 62.41 Discover
2012-01-01 10:08 MED Musica 93.37 Visa
2012-01-01 10:22 BGA Musica 369.94 MasterCard
2012-01-01 10:58 MED Musica 119.12 Efectivo
2012-01-01 11:36 BOG Musica 296.76 Discover
```

1

```
BOG 88.56
BGA 337.71
BGA 62.41
MED 93.37
BGA 369.94
MED 119.12
BOG 296.76
```

2

```
BOG 88.56 296.76

BGA 337.71 62.41 369.94

MED 93.37 119.12
```

map reduce

```
2012-01-01 09:08 BOG Libros 88.56 Discover
2012-01-01 09:09 BGA Libros 337.71 Efectivo
2012-01-01 09:52 BGA Libros 62.41 Discover
2012-01-01 10:08 MED Musica 93.37 Visa
2012-01-01 10:22 BGA Musica 369.94 MasterCard
2012-01-01 10:58 MED Musica 119.12 Efectivo
2012-01-01 11:36 BOG Musica 296.76 Discover
```

1

```
BOG 88.56
BGA 337.71
BGA 62.41
MED 93.37
BGA 369.94
MED 119.12
BOG 296.76
```

2

```
BOG 88.56 296.76

BGA 337.71 62.41 369.94

MED 93.37 119.12
```

3

```
BOG 385.32

BGA 770.06

MED 212.49
```

map reduce

```
2012-01-01 09:08 BOG Libros 88.56 Discover
2012-01-01 09:09 BGA Libros 337.71 Efectivo
2012-01-01 09:52 BGA Libros 62.41 Discover
2012-01-01 10:08 MED Musica 93.37 Visa
2012-01-01 10:22 BGA Musica 369.94 MasterCard
2012-01-01 10:58 MED Musica 119.12 Efectivo
2012-01-01 11:36 BOG Musica 296.76 Discover
```

1 map

```
BOG 88.56
BGA 337.71
BGA 62.41
MED 93.37
BGA 369.94
MED 119.12
BOG 296.76
```

map (k, v)

2 shuffle

```
BOG 88.56 296.76

BGA 337.71 62.41 369.94

MED 93.37 119.12
```

3 reduce

```
BOG 385.32

BGA 770.06

MED 212.49
```

reduce (k, [v_1, …])

map reduce

```
2012-01-01 09:08 BOG Libros 88.56 Discover              host A
2012-01-01 09:09 BGA Libros 337.71 Efectivo
2012-01-01 09:52 BGA Libros 62.41 Discover
2012-01-01 10:08 MED Musica 93.37 Visa                  host B
2012-01-01 10:22 BGA Musica 369.94 MasterCard
2012-01-01 10:58 MED Musica 119.12 Efectivo
2012-01-01 11:36 BOG Musica 296.76 Discover             host C
```

1 map

```
BOG 88.56
BGA 337.71
BGA 62.41
MED 93.37
BGA 369.94
MED 119.12
BOG 296.76
```

map (k, v)

2 shuffle

```
BOG 88.56 296.76

BGA 337.71 62.41 369.94

MED 93.37 119.12
```

3 reduce

```
BOG 385.32

BGA 770.06

MED 212.49
```

reduce (k, [v_1, …])

map reduce

```
2012-01-01 09:08 BOG Libros 88.56 Discover
2012-01-01 09:09 BGA Libros 337.71 Efectivo
```

host A

```
2012-01-01 09:52 BGA Libros 62.41 Discover
2012-01-01 10:08 MED Musica 93.37 Visa
2012-01-01 10:22 BGA Musica 369.94 MasterCard
```

host B

```
2012-01-01 10:58 MED Musica 119.12 Efectivo
2012-01-01 11:36 BOG Musica 296.76 Discover
```

host C

1 map

```
BOG 88.56
BGA 337
BGA
ME
     9.94
    119.12
OG 296.76
```

programmer

map (k, v)

2 shuffle

```
BOG 88.56

BGA        62.41 369.94

MED   .37 119.12
```

framework

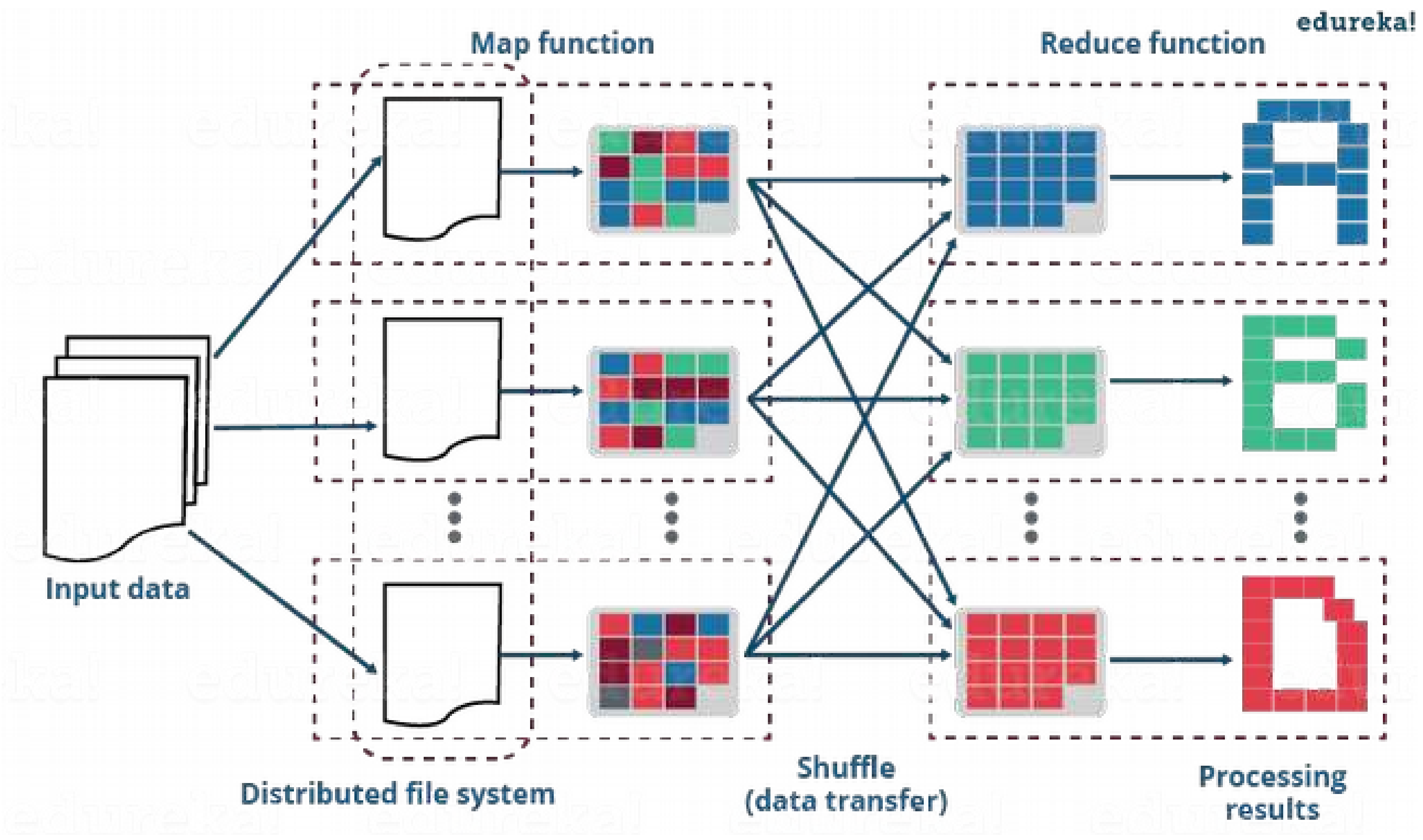
3 reduce

```
BOG 38

        .06

  D 212.49
```

programmer

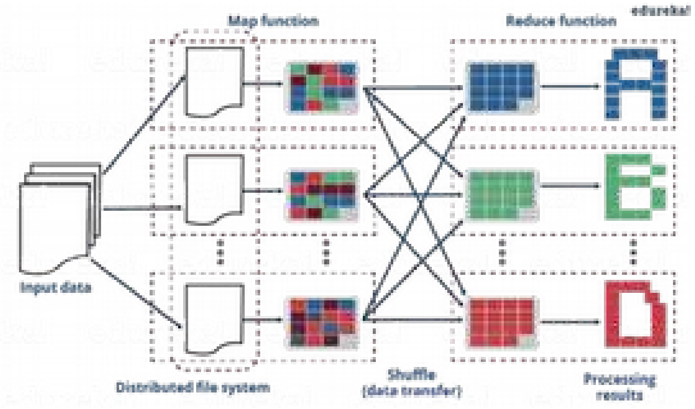
reduce (k, [v_1, …])

map reduce



DATA TRANSFER ONLY IN SHUFFLE

map reduce



Data ALREADY exists in nodes

MR Programmer → forget about paralellism

Framework developer → optimize coordination and comms

RESTRICTED PROGRAMMING MODEL

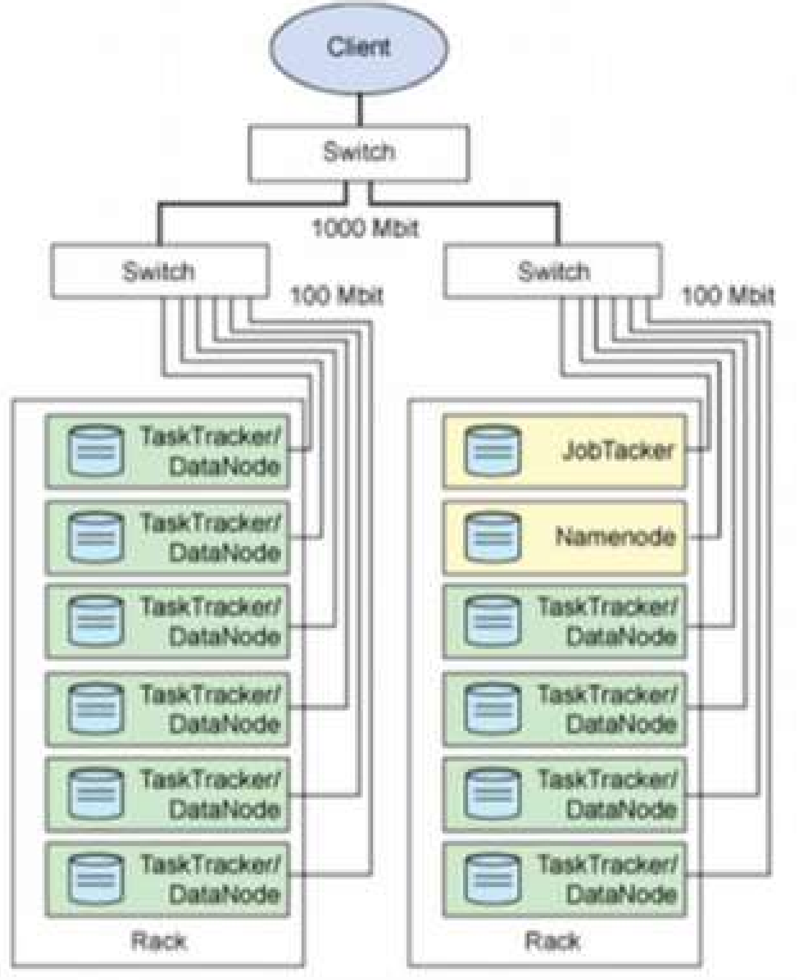
How to crunch 1PB

Lots of disk spinning all time
Redundancy, disks die
Lots of CPUs, working all time
Retry, since errors happen

Design qualities

Scalable – many servers
Reliable – redundant storage
Fault-tolerant – auto retry, self-healing

Computing to Data

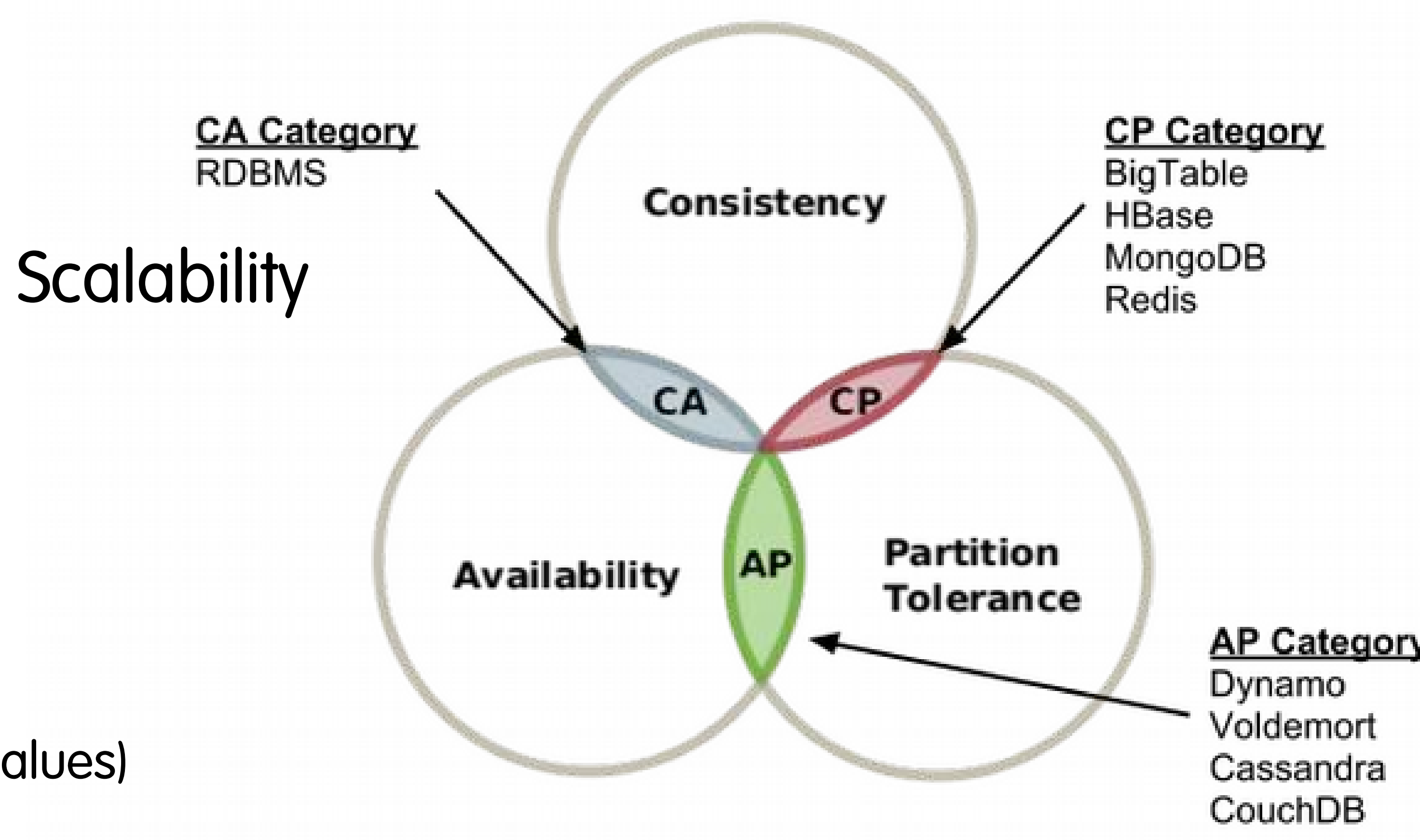


data goes to computing

computing goes to data

noSQL

Expressivity SQL vs. Scalability



Simpler data model (key, values)

Simpler operations

- Scan/access per key, basic transactions (check&put)
- No joins, no SQL language

Simple failover and scale up

Big table, Hbase, DynamoDB, Azure, Cassandra, etc.

Why Big Data

Data growing faster than computation speeds

Storage and network bottlenecks

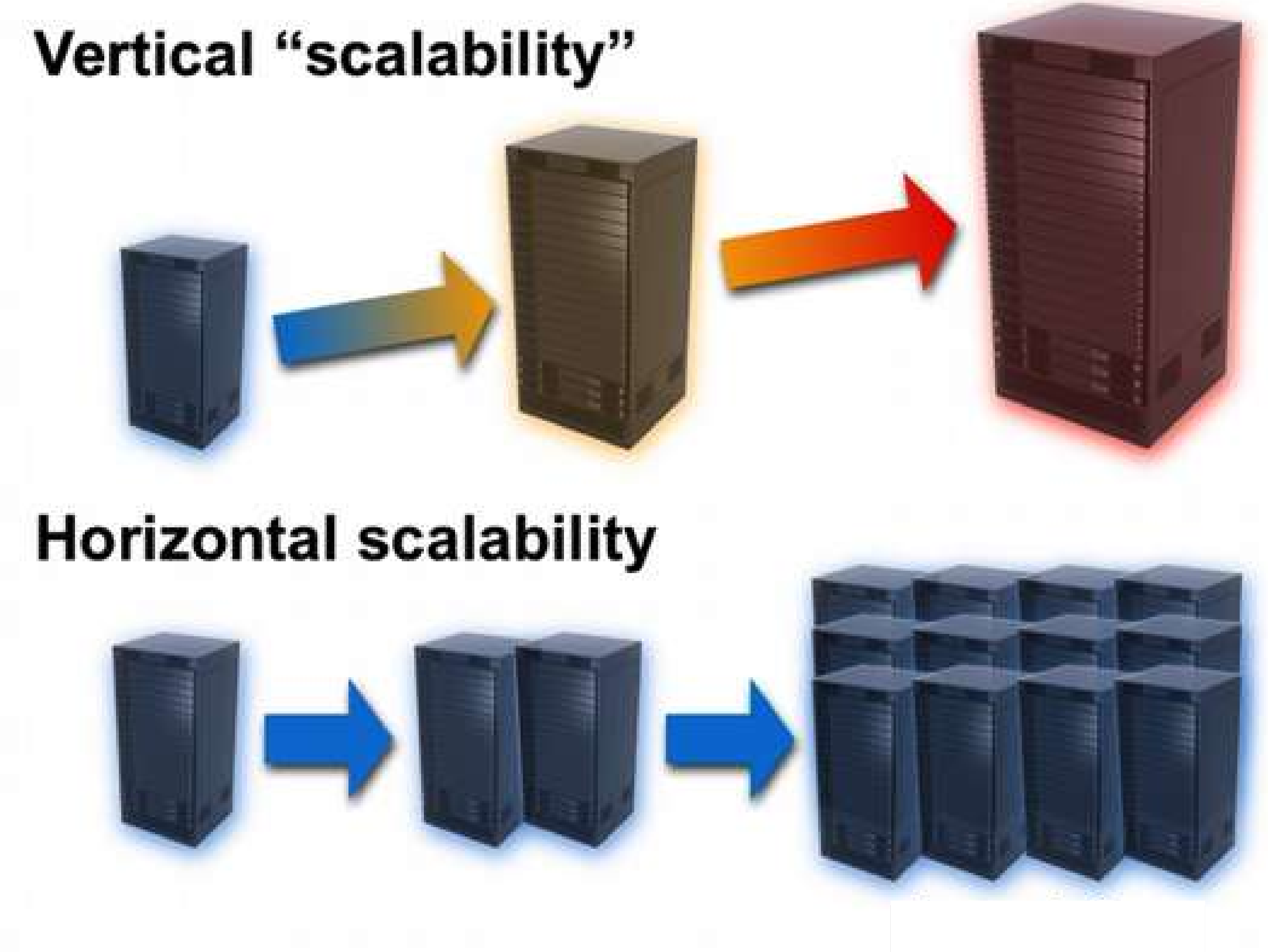
Facebook daily logs: 60TB

1000 genomes project: 200TB

Cost of 1TB of disk: USD 30

Time to read 1TB from disk: 3 hrs (100 MB/s)

Scalability in Big Data



scale up
traditional DBs

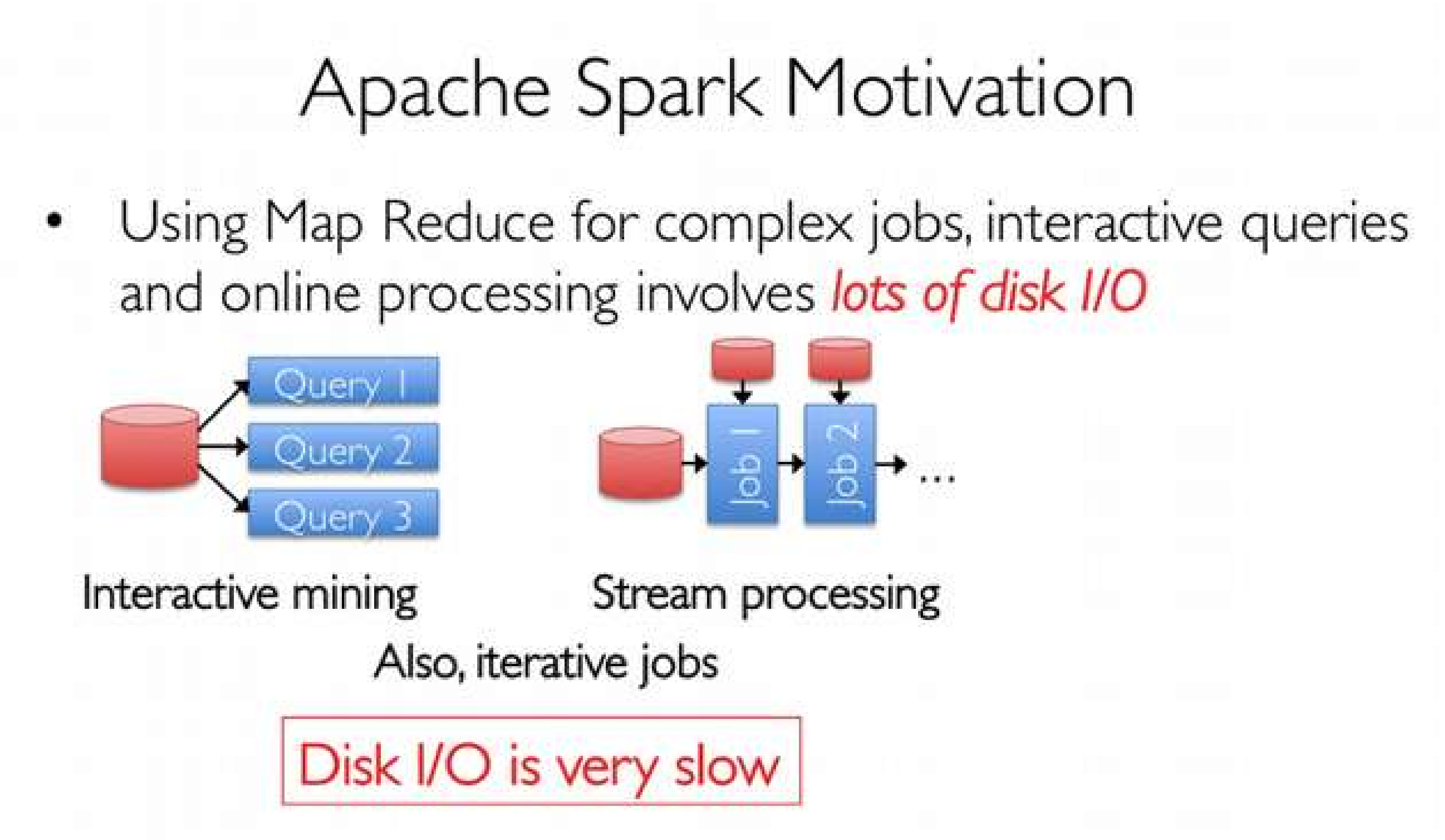
scale out
noSQL

seek "triviality" for appropriate
sw+hw architectures

recent technologies (virtualization, etc.) tend to favor the cost of scaling out

Spark

Lightning-Fast Cluster Computing



Apache Spark Motivation

- Using Map Reduce for complex jobs, interactive queries and online processing involves *lots of disk I/O*

Query 1
Query 2
Query 3
Job 1
Job 2
…

Interactive mining

Stream processing

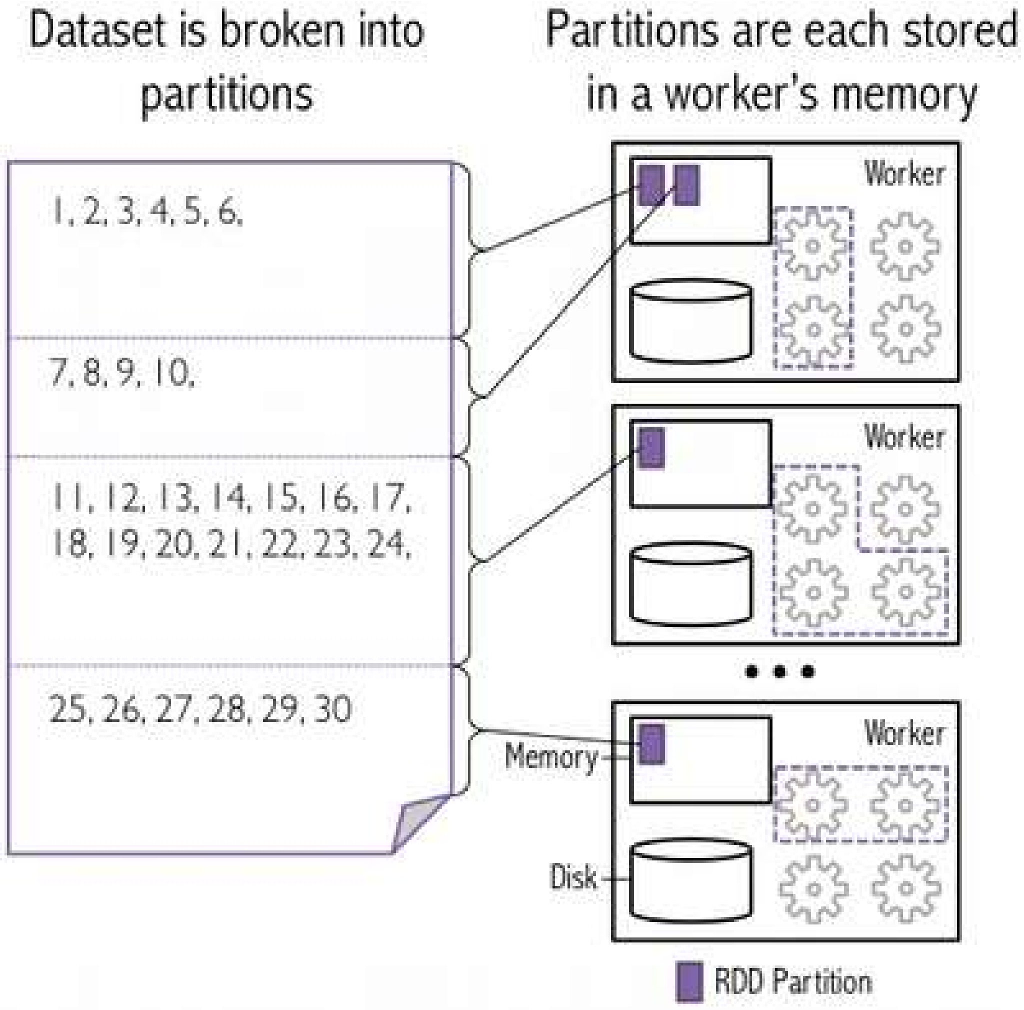
Also, iterative jobs

Disk I/O is very slow

Slide from Scalable Machine Learning course at EDX.com

Spark computing model

Resilient Distributed Datasets … **ON MEMORY**



Spark computing model

A program is a set of transformations on RDDs
(to/from the distributed memory)

Here is an operation, run it on all the data.
- Don’t care where or even run it twice!!!

Large set of distributed primitives
- M/R, groupby, etc.

Still computing goes to data!!!

Big Data

Focused on data
Scalability for the masses
Tradeoff to scale
More cloud oriented

Coarse grained parallelism

independent tasks, localized synch
machine based partitioning

Targets average performance

how much data is big data?

HPC

More science driven
Closer to hardware
Cutting edge algorithms
Better defined problems

Fine grained parallelism

intercommunicating synchronized tasks
also CPU/GPU based partitioning

Targets peak performance

let’s get to the guts of your code

Big Data and HPC

HIVE
mongoDB
slurm
workload manager
OPEN MPI
hadoop
OpenMP
OpenCL
APACHE
Spark
cassandra
OmpSs
cloudera
redis
Hortonworks
NVIDIA.
CUDA
lustre
File System
docker
amazon
web services
Microsoft
Azure
openstack
OpenNebula

Big data challenges for HPC

What kind of Big Data problems can be addressed with traditional HPC resources?

Big Data clusters (Spark/noSQL) managed very differently from job scheduling based clusters.

What is the "customer" base? Final users? Programmers?

What is the cost of using Big Data/HPC solutions?
$$$, people, opportunity?

Technological / Non functional requirements (security, streaming data, data delivery SLAs, etc.)

Big data approaches for HPC

consider container based management (i.e. Openstack).

consider Big Data models (spark) for parallelizing scientific software.

HPC community is strong in algorithmics → programming/deployment models for Big Data

complementarities: CPU/GPU software running on Spark clusters