MUSIC

&

BIG DATA

-

October 2017 - Ecole Centrale de Nantes

Guillaume Gardey

PLANNING

- Session 1
  - Talk & QA: *Music & Web - Architecture & Technology Overview*
  - Lab 1: Working with APIs

- Session 2
  - Talk & QA: *Music & Big Data - Overview of challenges & technologies*
  - Lab 2: Introduction to Data Processing - Python/Pandas

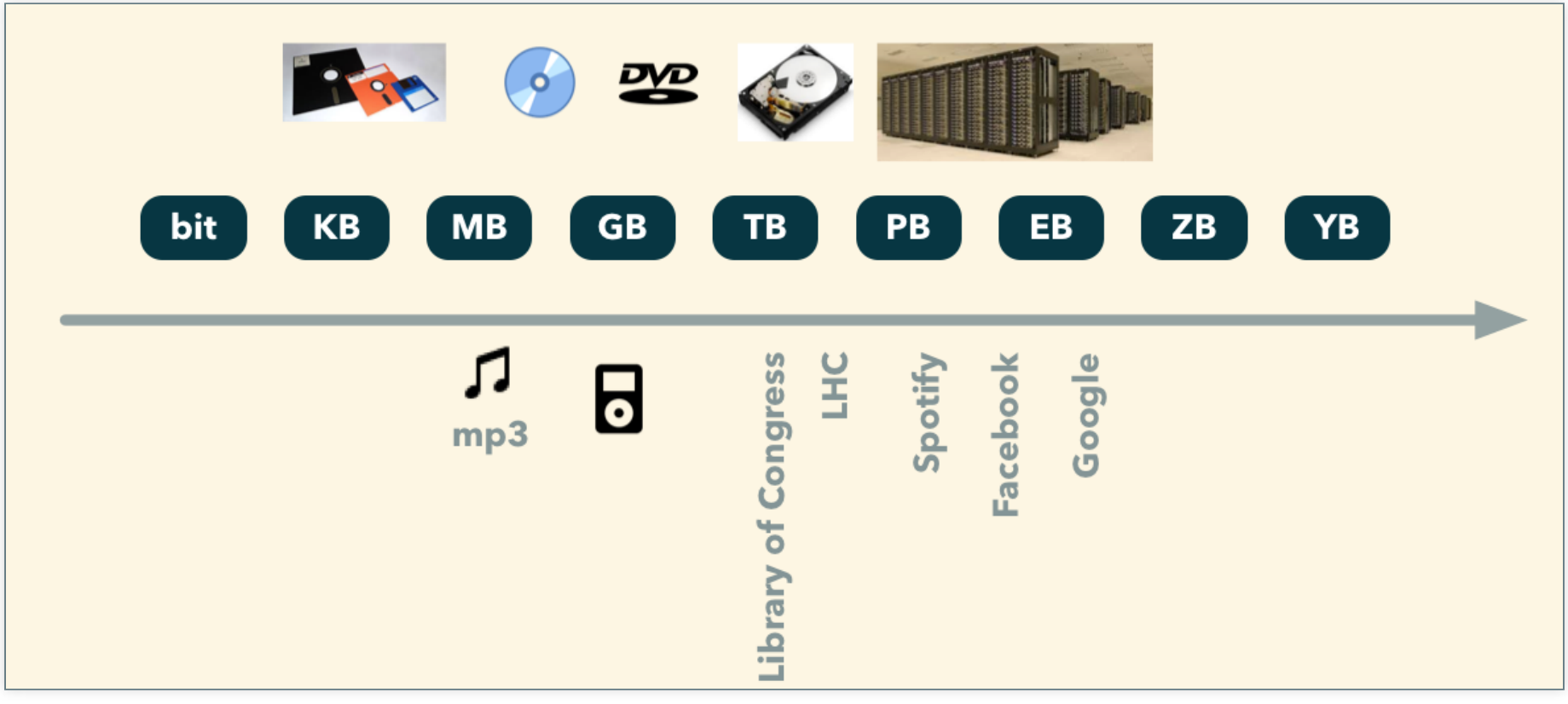
BIG DATA

describes large amount of data (structured or unstructured) that are difficult to process using traditional database and software

BIG DATA

Big data usually includes data sets with sizes beyond the ability of commonly used software tools to capture, curate, manage, and process data within a tolerable elapsed time.

BIG DATA


DVD
bit
KB
MB
GB
TB
PB
EB
ZB
YB
mp3
Library of Congress
LHC
Spotify
Facebook
Google

BIG DATA - THE 3 VS... (AND MORE)

VOLUME

- large data sets
- information is not sampled

VELOCITY

- rapidly changing
- available in real-time

VARIETY

- different type: text, images, audio, video, ...
- structured: JSON, XML, ...
- (un|semi) structured: email, images, audio, music, text

... AND MORE VS

- **Veracity**
  - how much trust can be put in the data
- **Value**
  - eventually drives revenues or new features for companies
- **Variability**
  - no fixed data or schema
  - evolution in time

BIG DATA IN MUSIC

-

WHERE ?

CONTENT

- Audio
- Metadata
- Lyrics

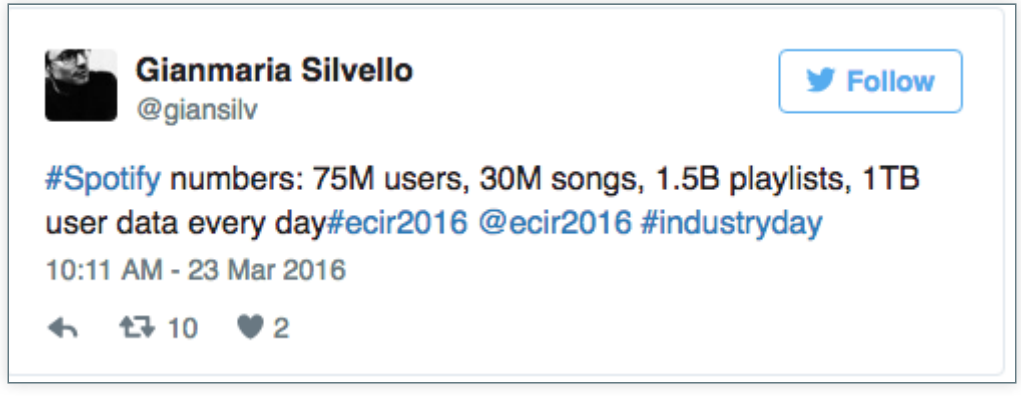
EVENTS

- Listening patterns
- Application events
- User activity
- Social media data

DERIVED DATA

- Crowd sourced data
- Recommendations
- Playlists
- User content

EXAMPLE: BIG DATA @ SPOTIFY



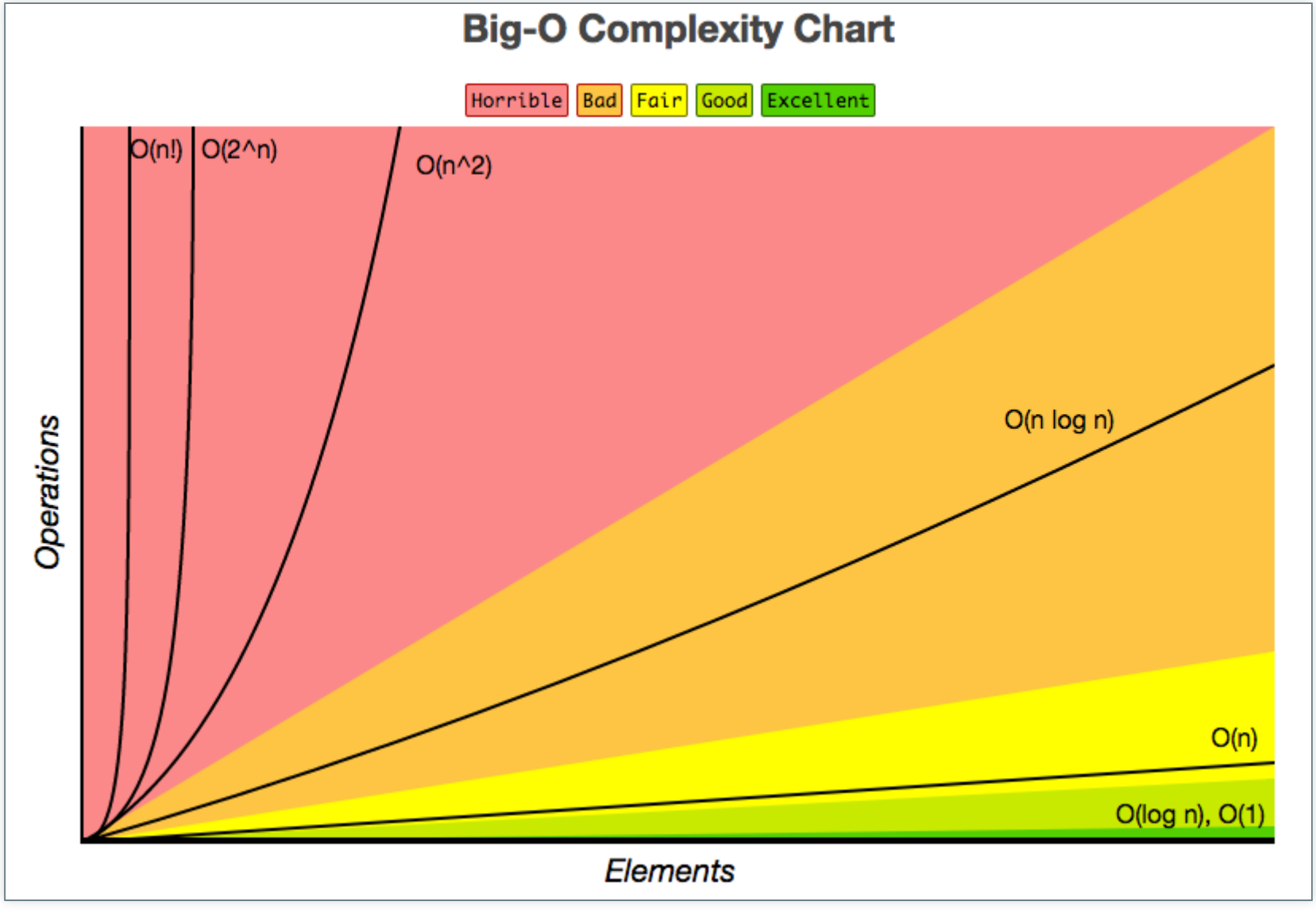
- 42PB Storage
- 200TB data generated / day
- 1300 Servers

HOW TO SCALE?

ALGORITHMS & DATA STRUCTURES

- Algorithmic & Complexity
- Data Structures

COMPLEXITY


Big-O Complexity Chart
Horrible
Bad
Fair
Good
Excellent
O(n!)
O(2^n)
O(n^2)
O(n log n)
O(n)
O(log n), O(1)
Operations
Elements

DATA STRUCTURES

Common Data Structure Operations

Data Structure	Time Complexity								Space Complexity
	Average				Worst				Worst
	Access	Search	Insertion	Deletion	Access	Search	Insertion	Deletion	
Array	Θ(1)	Θ(n)	Θ(n)	Θ(n)	O(1)	O(n)	O(n)	O(n)	O(n)
Stack	Θ(n)	Θ(n)	Θ(1)	Θ(1)	O(n)	O(n)	O(1)	O(1)	O(n)
Queue	Θ(n)	Θ(n)	Θ(1)	Θ(1)	O(n)	O(n)	O(1)	O(1)	O(n)
Singly-Linked List	Θ(n)	Θ(n)	Θ(1)	Θ(1)	O(n)	O(n)	O(1)	O(1)	O(n)
Doubly-Linked List	Θ(n)	Θ(n)	Θ(1)	Θ(1)	O(n)	O(n)	O(1)	O(1)	O(n)
Skip List	Θ(log(n))	Θ(log(n))	Θ(log(n))	Θ(log(n))	O(n)	O(n)	O(n)	O(n)	O(n log(n))
Hash Table	N/A	Θ(1)	Θ(1)	Θ(1)	N/A	O(n)	O(n)	O(n)	O(n)
Binary Search Tree	Θ(log(n))	Θ(log(n))	Θ(log(n))	Θ(log(n))	O(n)	O(n)	O(n)	O(n)	O(n)
Cartesian Tree	N/A	Θ(log(n))	Θ(log(n))	Θ(log(n))	N/A	O(n)	O(n)	O(n)	O(n)
B-Tree	Θ(log(n))	Θ(log(n))	Θ(log(n))	Θ(log(n))	O(log(n))	O(log(n))	O(log(n))	O(log(n))	O(n)
Red-Black Tree	Θ(log(n))	Θ(log(n))	Θ(log(n))	Θ(log(n))	O(log(n))	O(log(n))	O(log(n))	O(log(n))	O(n)
Splay Tree	N/A	Θ(log(n))	Θ(log(n))	Θ(log(n))	N/A	O(log(n))	O(log(n))	O(log(n))	O(n)
AVL Tree	Θ(log(n))	Θ(log(n))	Θ(log(n))	Θ(log(n))	O(log(n))	O(log(n))	O(log(n))	O(log(n))	O(n)
KD Tree	Θ(log(n))	Θ(log(n))	Θ(log(n))	Θ(log(n))	O(n)	O(n)	O(n)	O(n)	O(n)

PROGRAM OPTIMIZATION

- CPU
- Memory
- IO
- Network

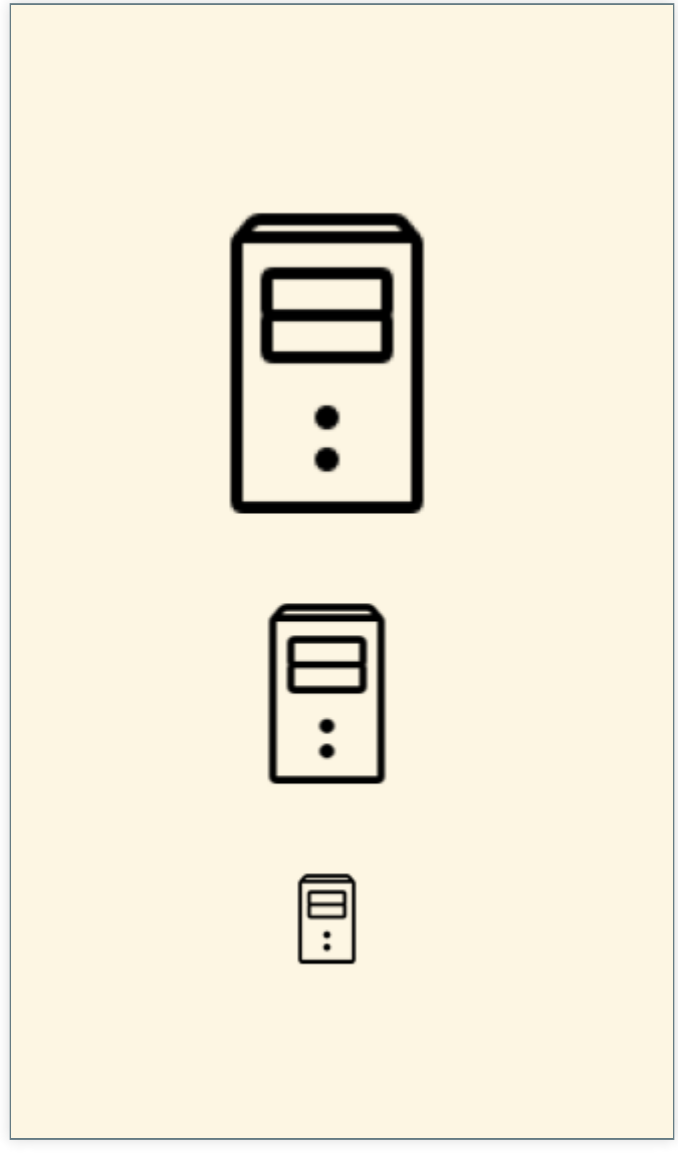
PARALLELISM

- Multi**threading**
- Multi**processing**

VERTICAL SCALING

- Same server
- More
  - CPU
  - Memory
  - Storage

VERTICAL SCALING



NEW PARADIGMS

- Dedicated hardware
  - GPU (Graphical Process Unit)
  - FPGA (Field Programmable Gate Array)

- New paradigm
  - DNA Computing
  - Quantum Computing

HORIZONTAL SCALING

Distribute resources and work to many computers

HORIZONTAL SCALING

HORIZONTAL SCALING

- Distributed Systems
- Clusters
- Sharding
- Share Nothing
- Cloud

BIG DATA & HADOOP

- Fundations
  - Google File System (2003)
  - Google MapReduce (2004)
  - Google BigTable (2005/2006)

- Open Source implementation
  - *Apache Nutch* (web crawler)
  - Development moved to the *Hadoop* project in 2006

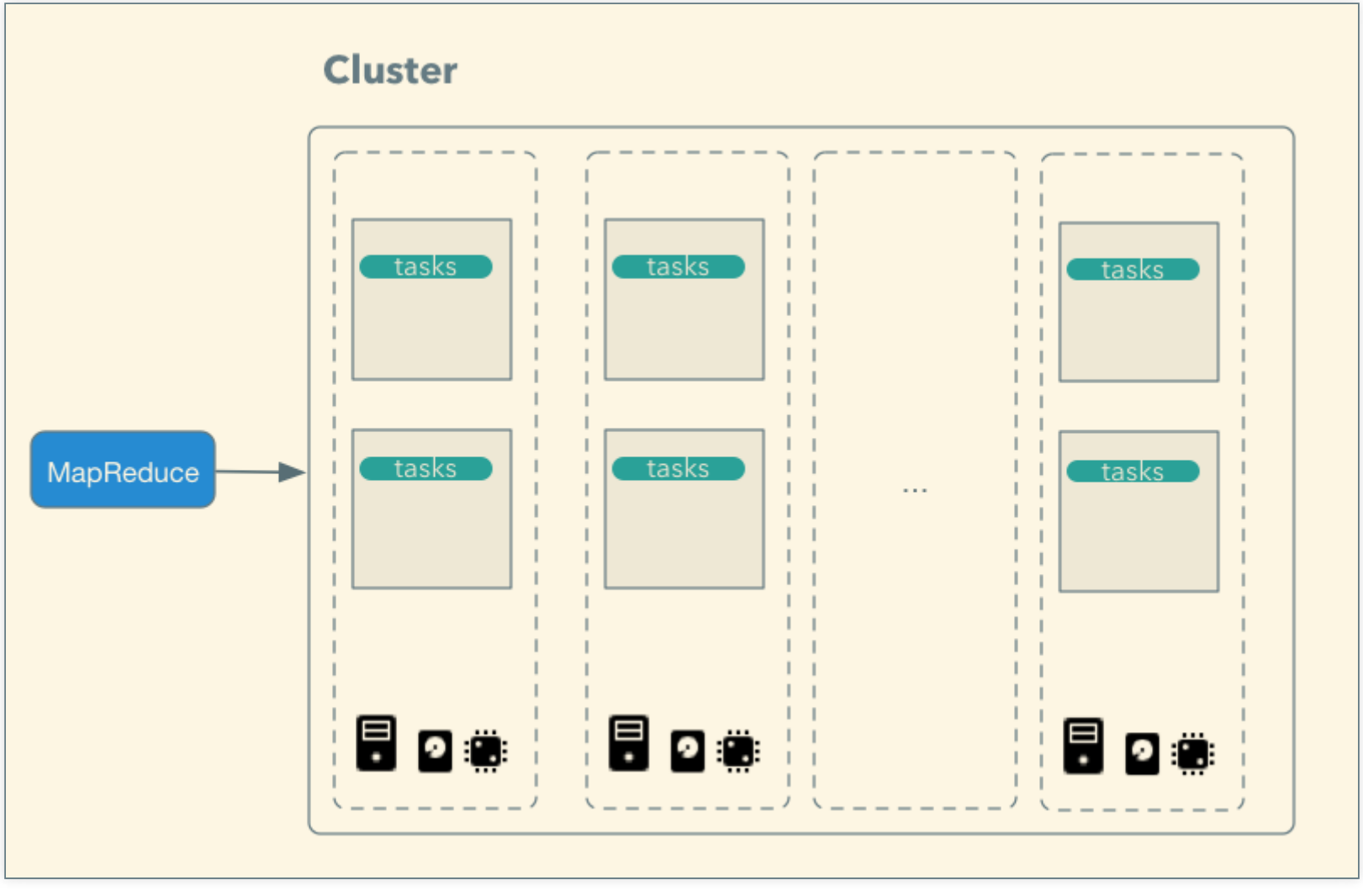
MAP-REDUCE

A programming model for processing and generating large data sets with a parallel, distributed algorithm on a cluster

Takes advantage of the locality of data, processing it near the place it is stored in order to reduce the distance over which it must be transmitted.

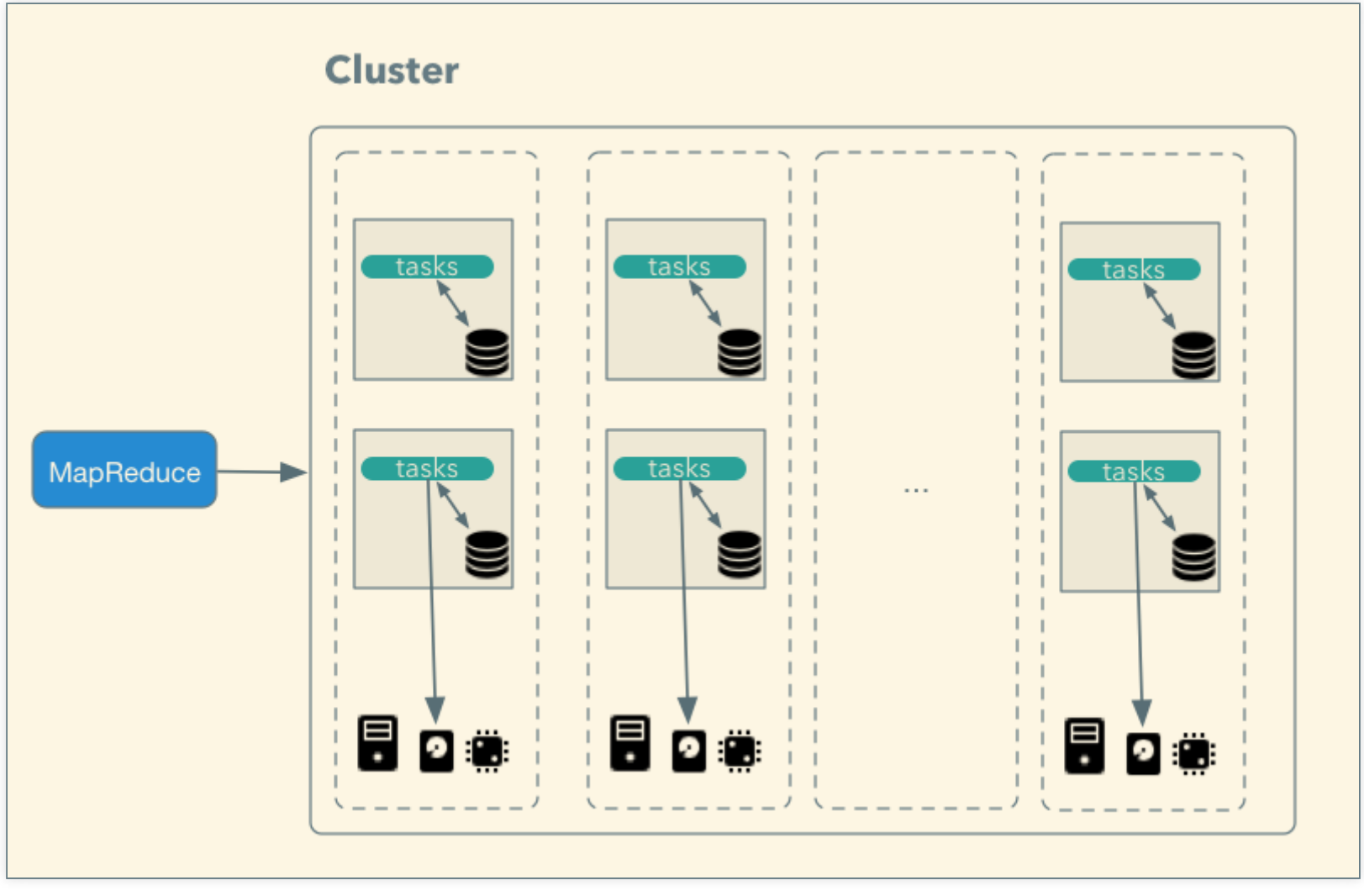
MAP-REDUCE

Parallel computations on a cluster

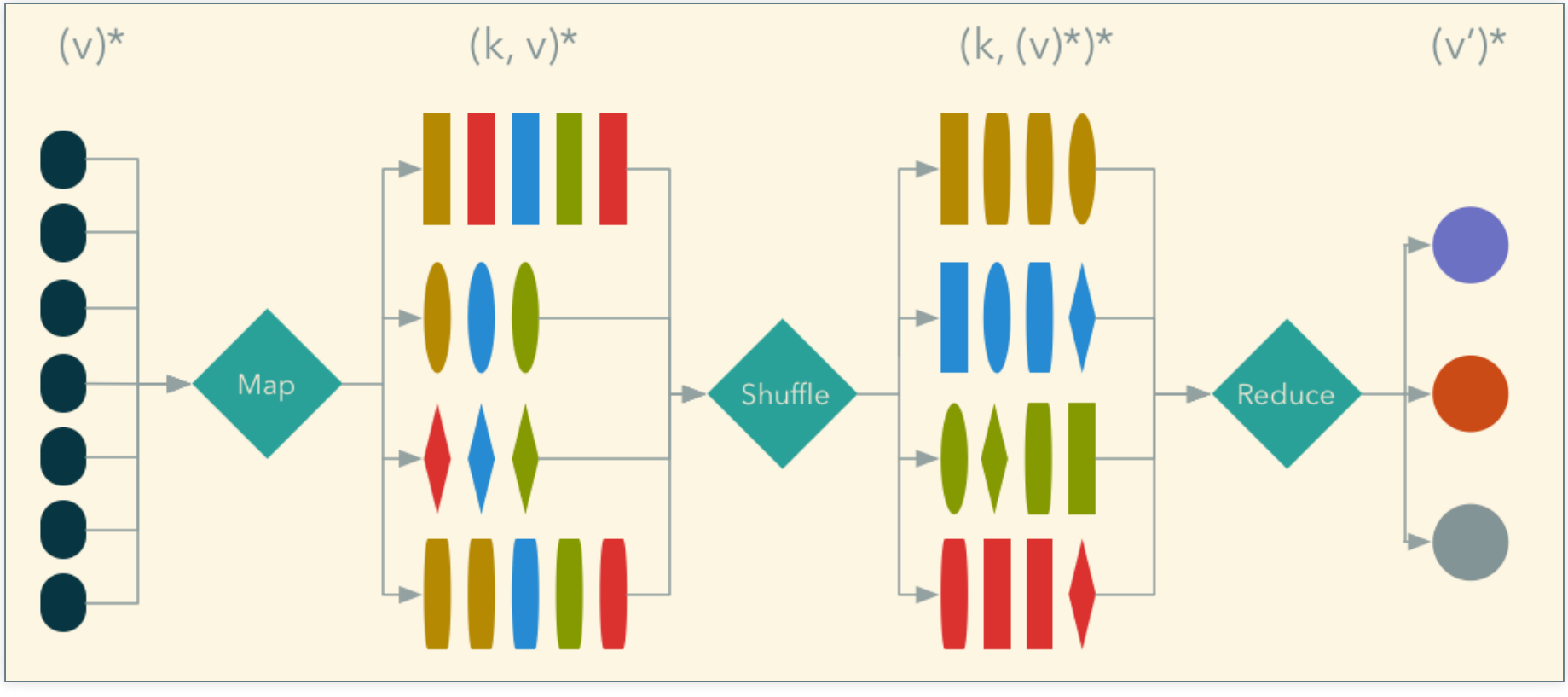


MAP-REDUCE

Data locality



MAP-REDUCE

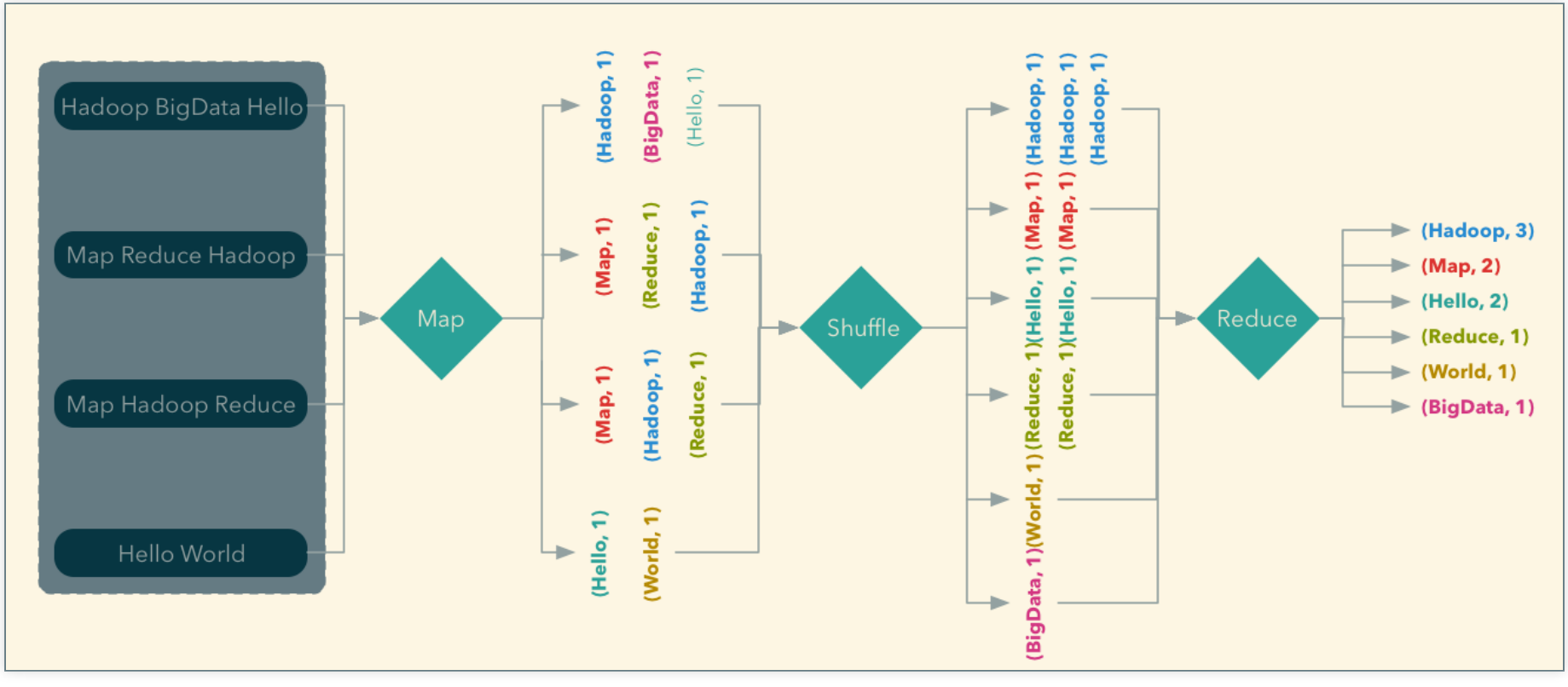


MAP-REDUCE - WORD COUNT

```
def map(document):
    for word in document:
        emit(word, 1)

def reduce(word, values):
    count = 0
    for value in values:
         count += value
    emit(word, count)
```

MAP-REDUCE - WORD COUNT



MAP-REDUCE - WHAT NOW?

Relatively simple computational model

but

Many problems can be translated/solved!

- SQL
- ETL (Extract / Transform / Load)
- Machine Learning
- Bespoke analysis
- ...

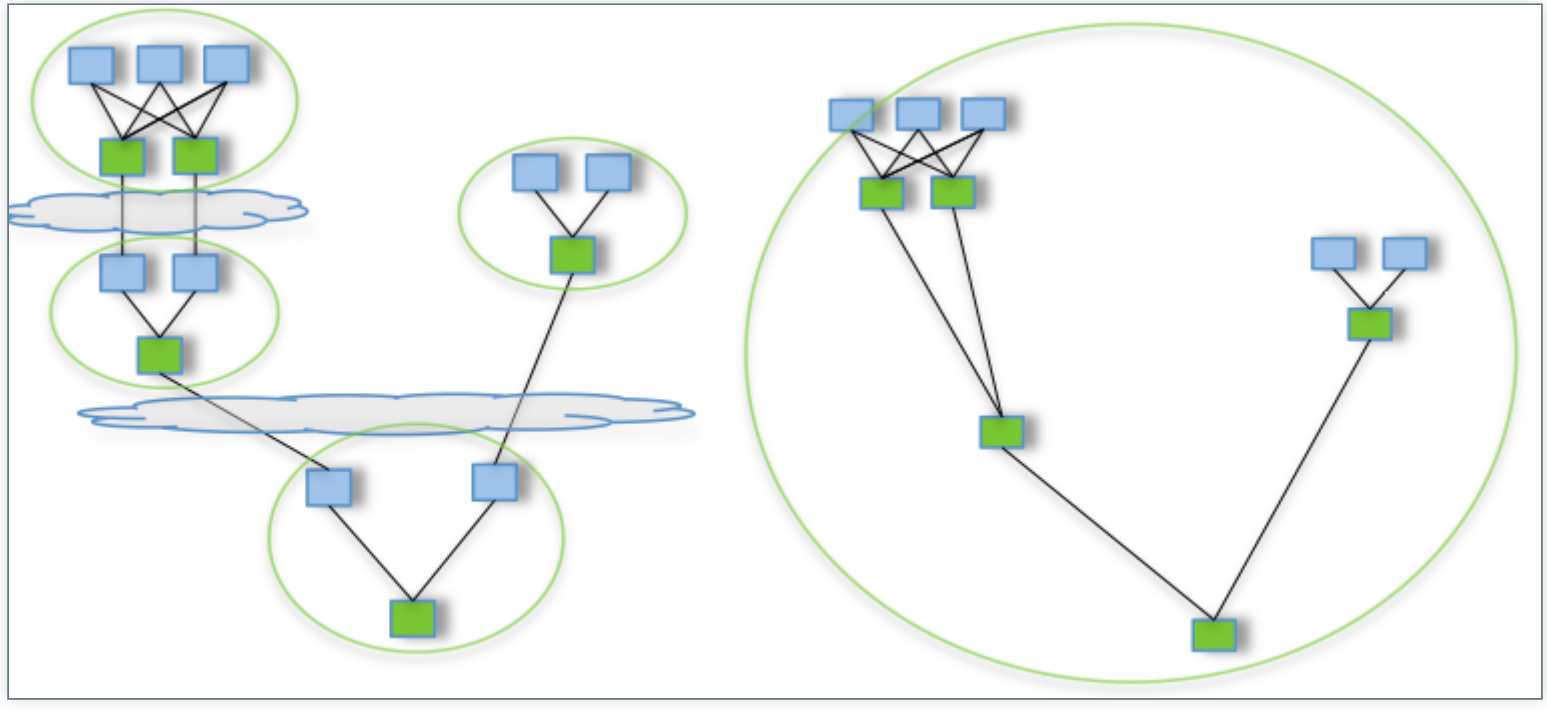
MAP-REDUCE - LIMITATIONS

- MapReduce jobs independent from each others
- Network & Disk IO intensive in some cases (shuffle)
- Lack of iterative/in-memory computation

BIG DATA - BEYOND MAP REDUCE

NEW FRAMEWORKS - DAG

Direct Acyclic Graph



NEW FRAMEWORKS - DAG

- Generalization of Map-Reduce concept
- Jobs are aware of all the tasks involved
- Allows global optimization
- Better use of resources

Spark, *Tez*, *Drill*, *Dremel*, *Spanner*, ...

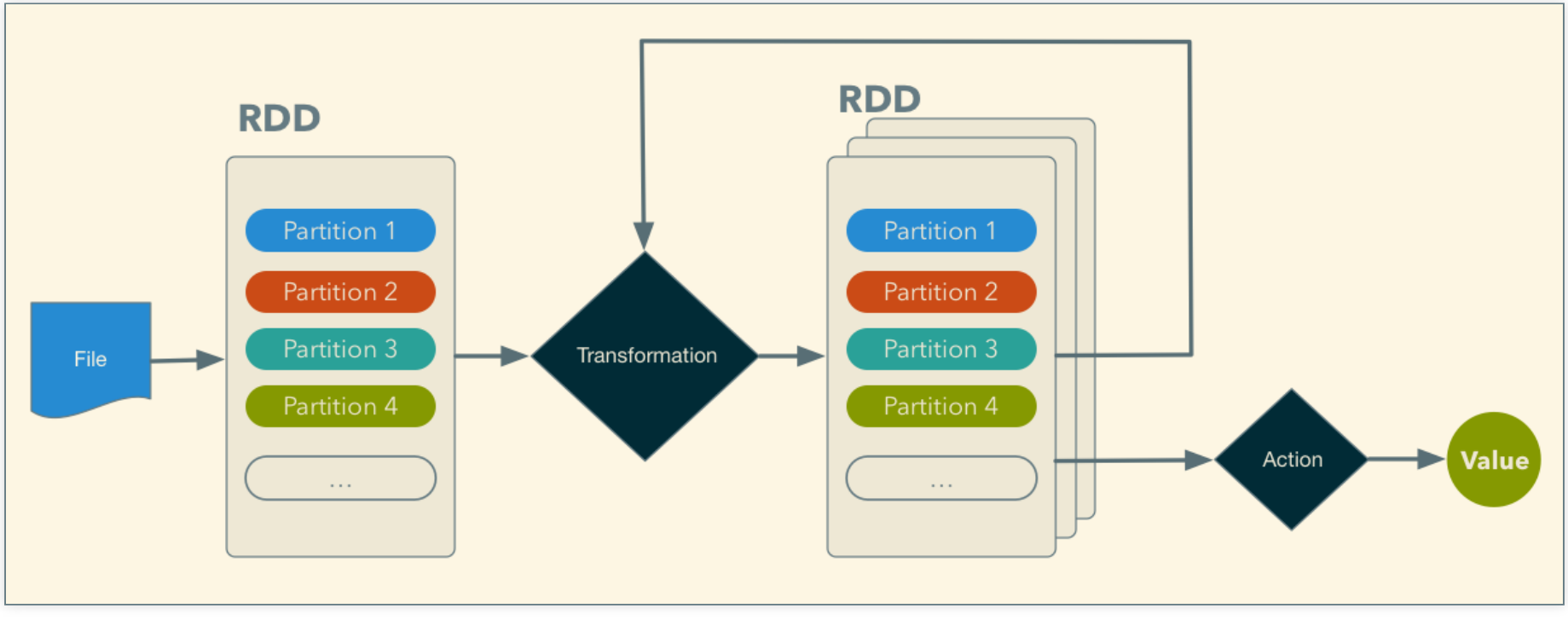
SPARK

- Fundations
  - Berkeley's AMPLab from 2009
  - Open sourced and moved as an Apache project in 2013

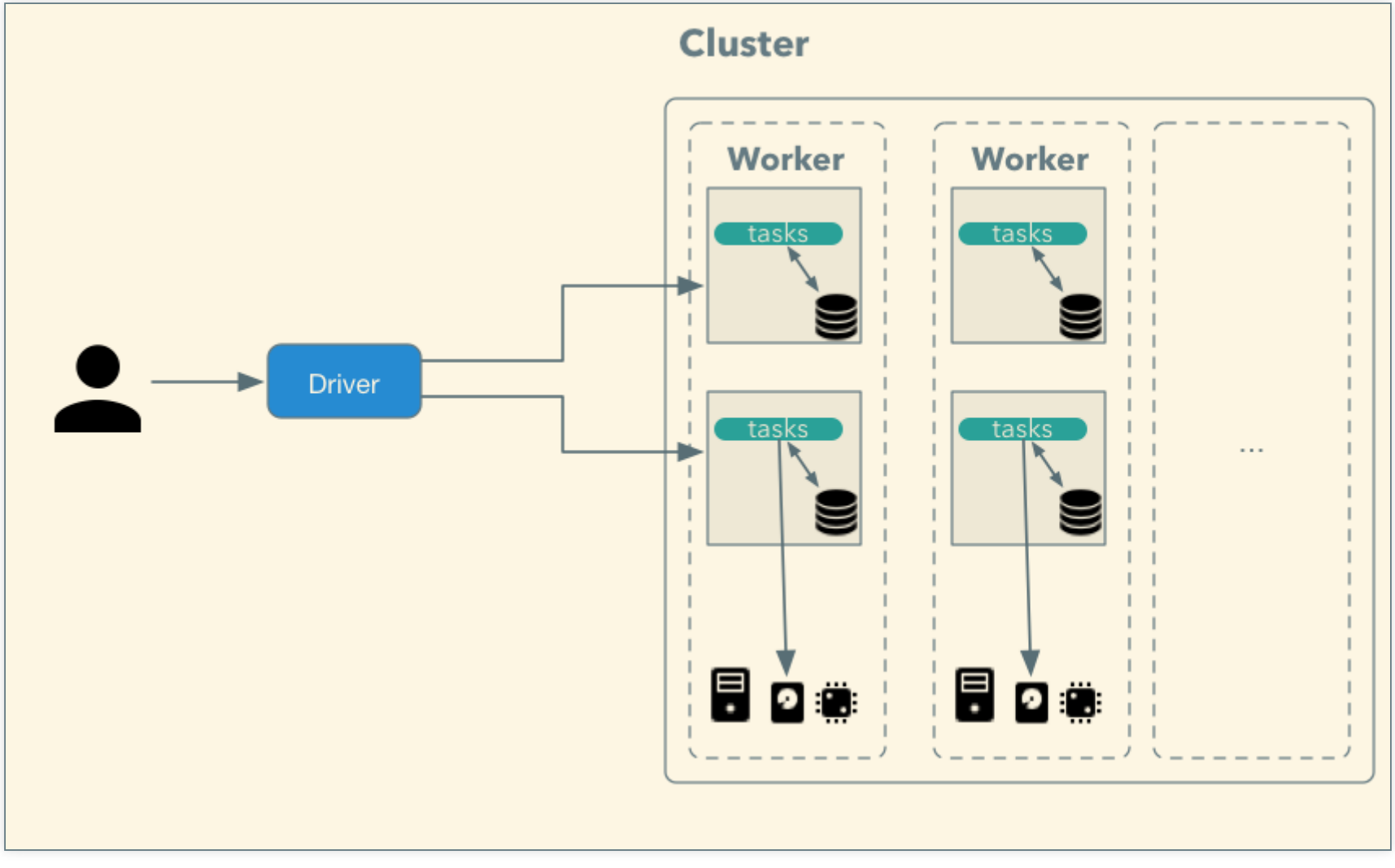
- Improvements on the Map-Reduce paradigm
  - In memory cluster computing
  - Iterative algorithms
  - Interactive & Exploratory analysis
  - Batch & Streaming

SPARK - RDD (RESILIENT DISTRIBUTED DATASETS)

a fault-tolerant collection of elements that can be operated on in parallel



SPARK - DRIVER & WORKERS



SPARK - HIGH LEVEL LIBRARIES

- SQL
- Streaming
- **Machine Learning**
- Graph

MACHINE & DEEP LEARNING

MACHINE LEARNING

Machine learning is a field of computer science that gives computers the ability to learn without being explicitly programmed.

MACHINE LEARNING

- Clustering / Classification
- Anomaly Detection
- Supervised / Unsupervised Learning
- Reinforcement Learning
- Neural Nets

MACHINE LEARNING & BIG DATA

- Vast quantities of Data
- Large data sets for training
- Improvement in software/hardware
  - GPU
  - High Level libraries

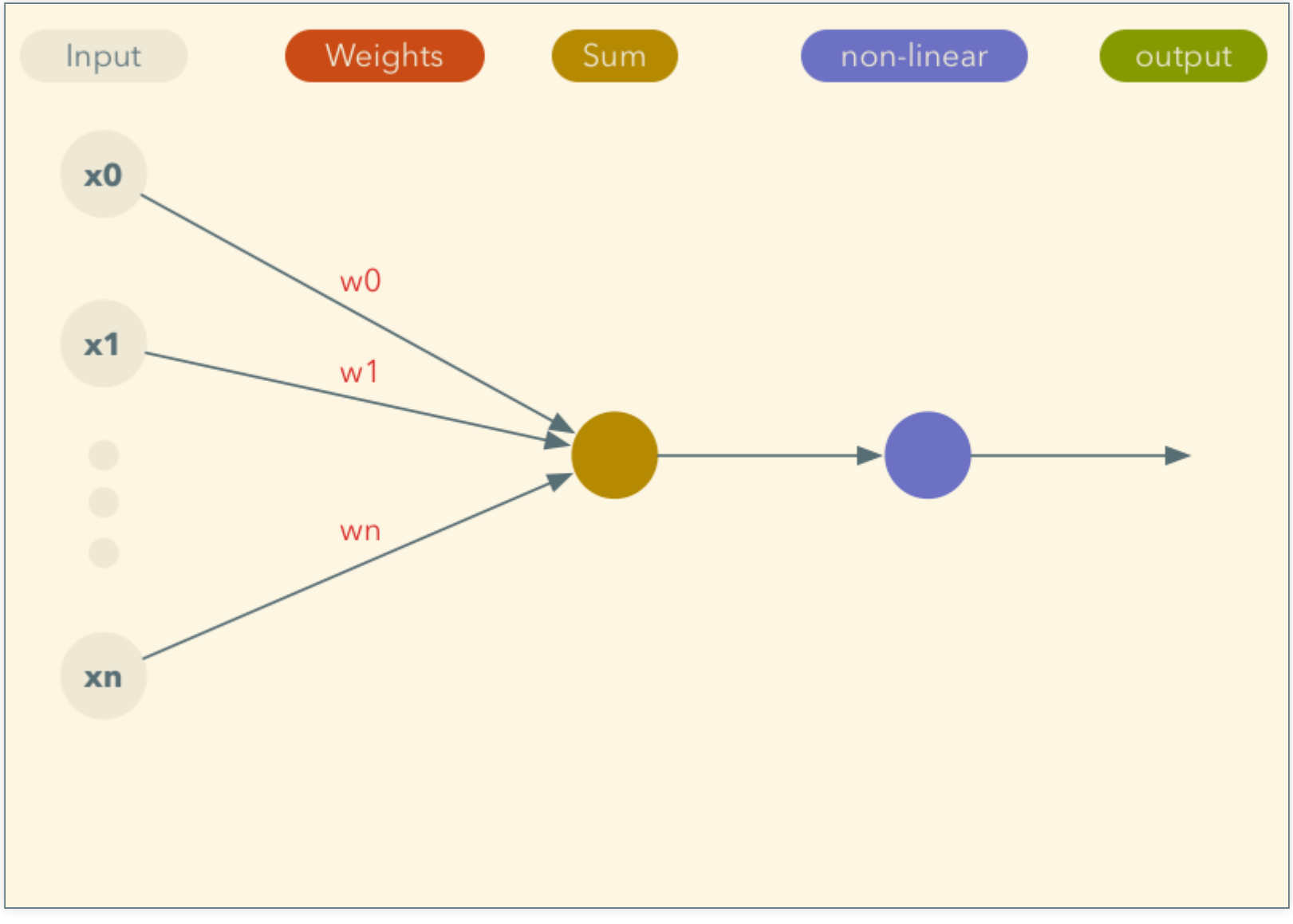
- Broadly accessible

DEEP LEARNING

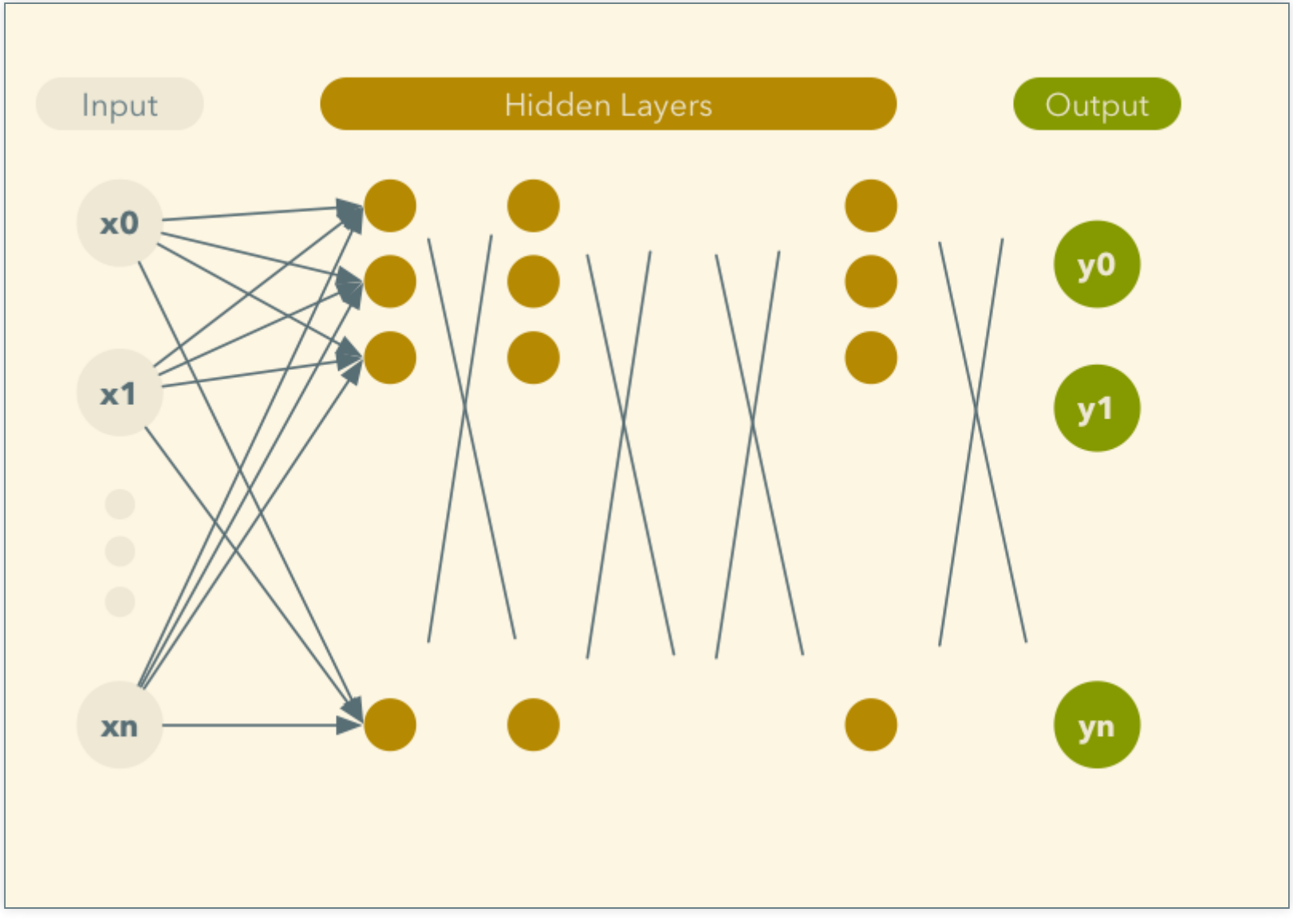
- Originated in the end of 50'
  - Perceptron
  - Frank Rosenblatt
  - Neurobiology

- Neural Networks

PERCEPTRON



DEEP NEURAL NETWORK



DEEP LEARNING

- Image classification
- Text Translation
- Speech Recognition
- Speech Synthesis
- Game
- ...

APPLICATION TO MUSIC

- **Recommendation**
  - Playlist & Marketing
- **Classification**
  - Genre, Mood, Tempo, Danceability
- **Music Generation**
  - Games, Ambient Music
- Techniques
  - Collaborative Filtering
  - Natural Language Processing
  - Deep Learning

QUESTIONS?