

Introduction to Data Science Course

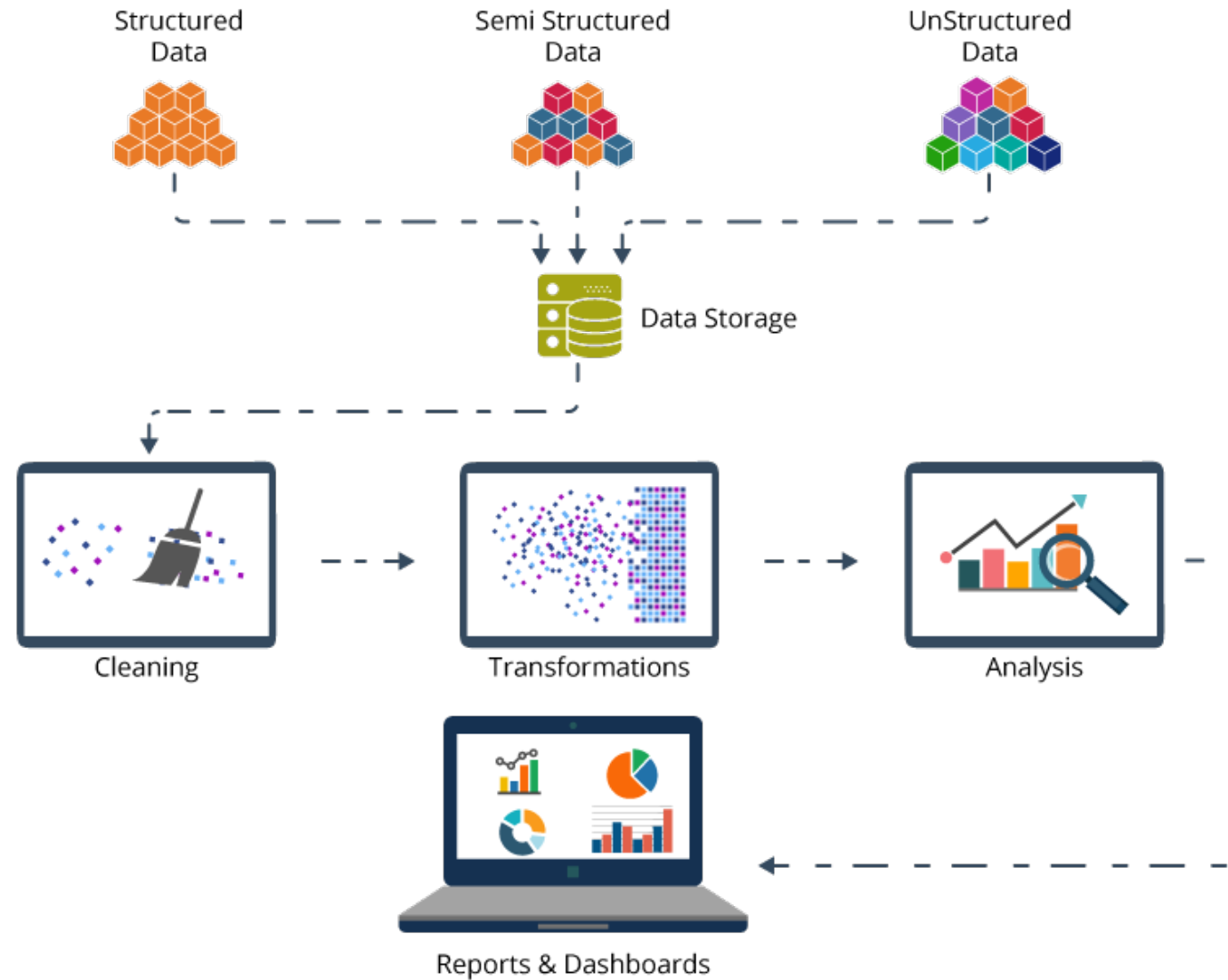
# Big Data Parallel and Distributed Computing

Le Ngoc Thanh  
Inthanh@fit.hcmus.edu.vn  
Department of Computer Science

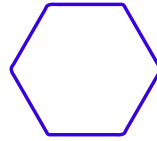
# Contents

- ◎ **Introduction to Big Data**
- ◎ Big data architecture
- ◎ Big data and data science
- ◎ Parallel and distributed computing

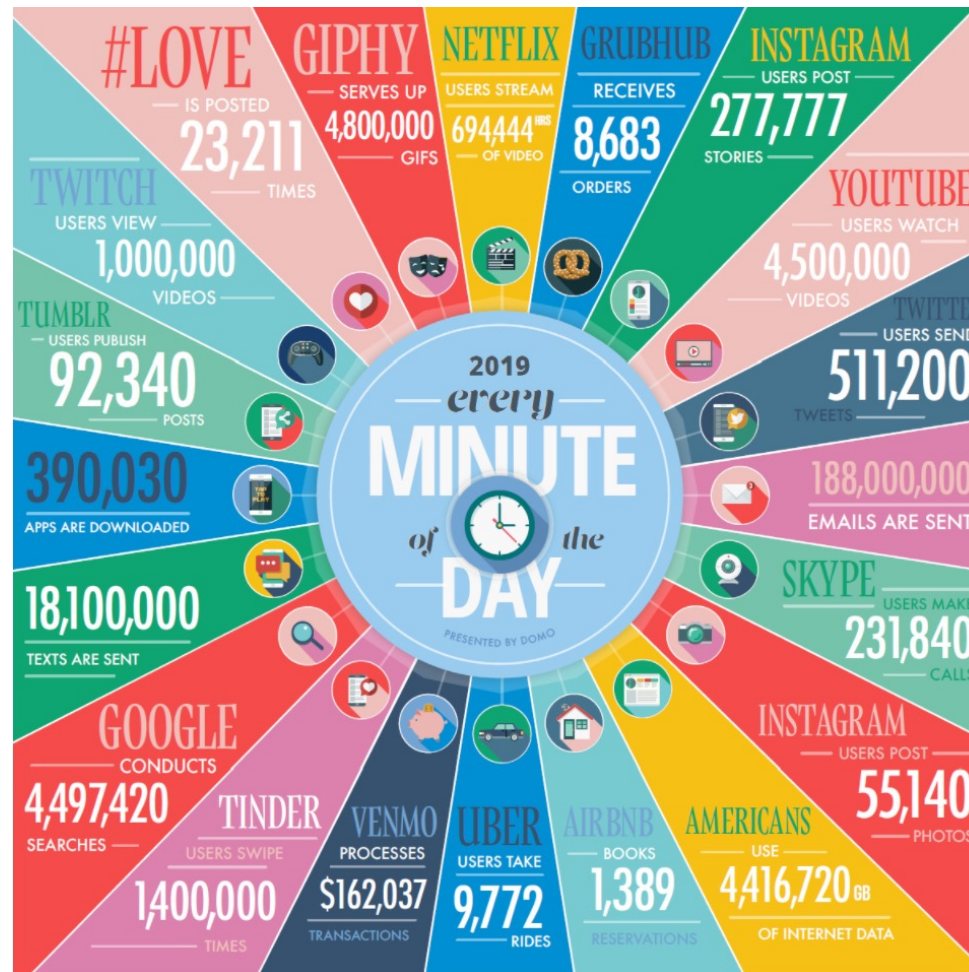
# Data Science Process



# Data Never Sleeps

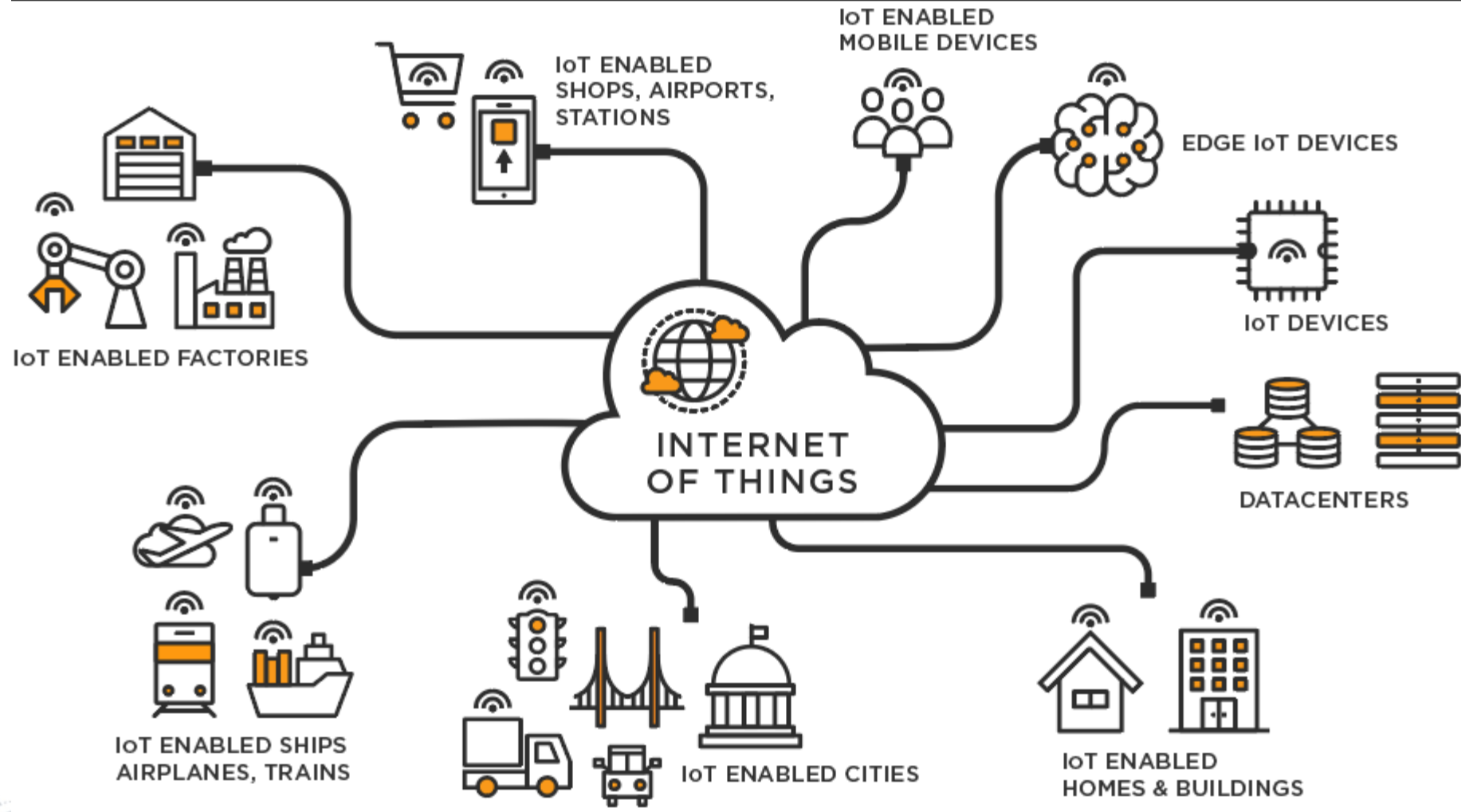


◎ How much data is generated every minute?

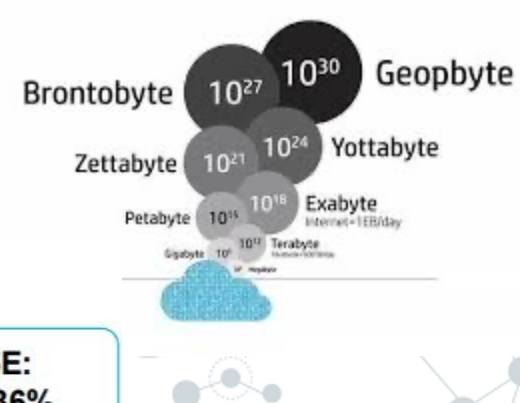


9

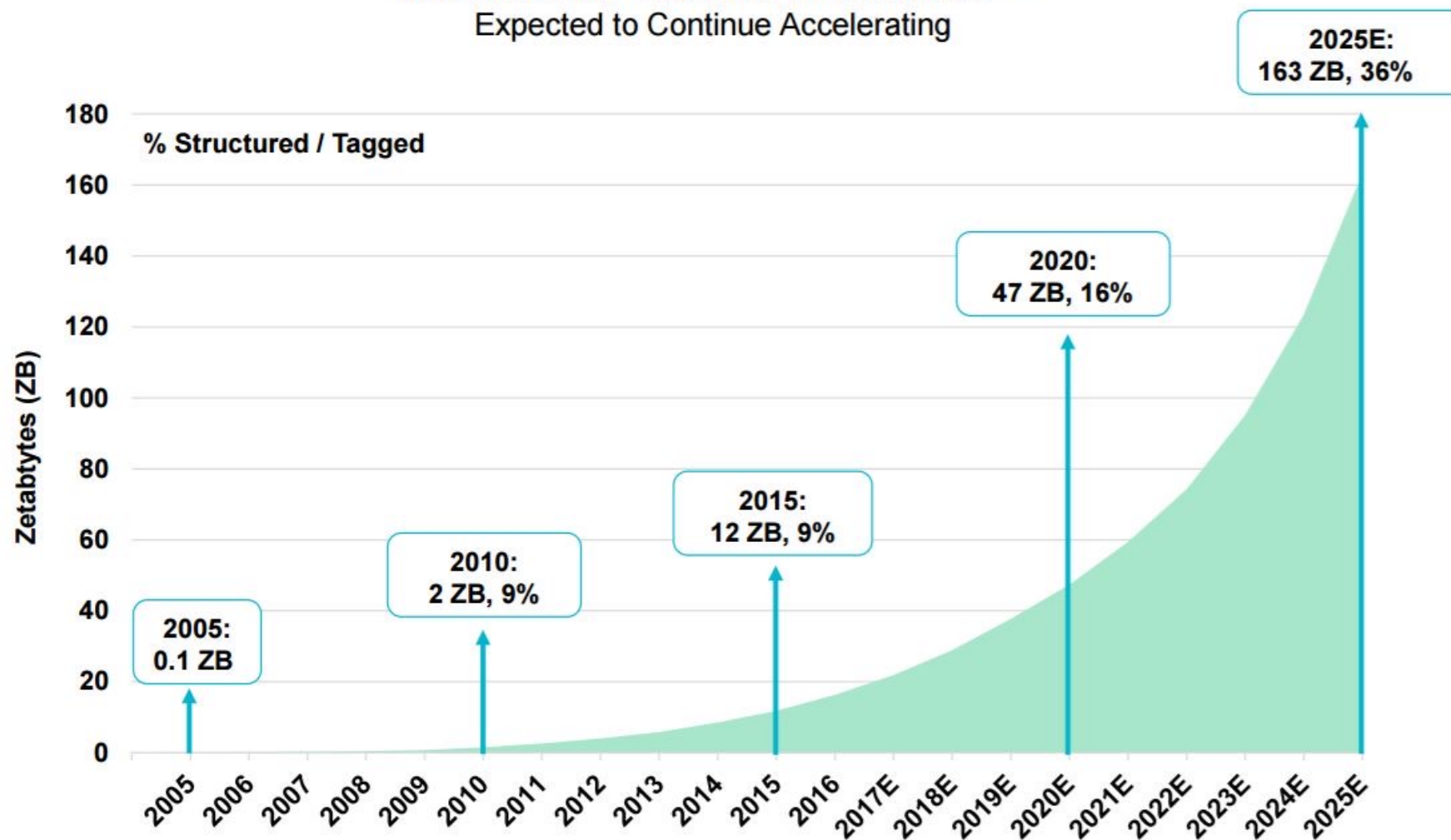
# Data Never Sleeps



# Data Growth



**Information Created Worldwide =**  
Expected to Continue Accelerating





# What is Big Data

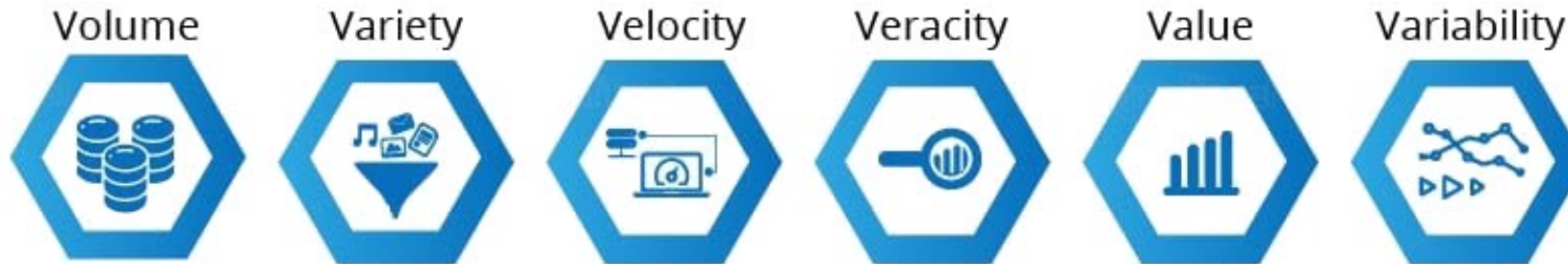
- © **Big data** is term used to describe the **massive volume** of both structured and unstructured data that is so large it is **difficult to process** using traditional techniques.



# Characteristics of Big data

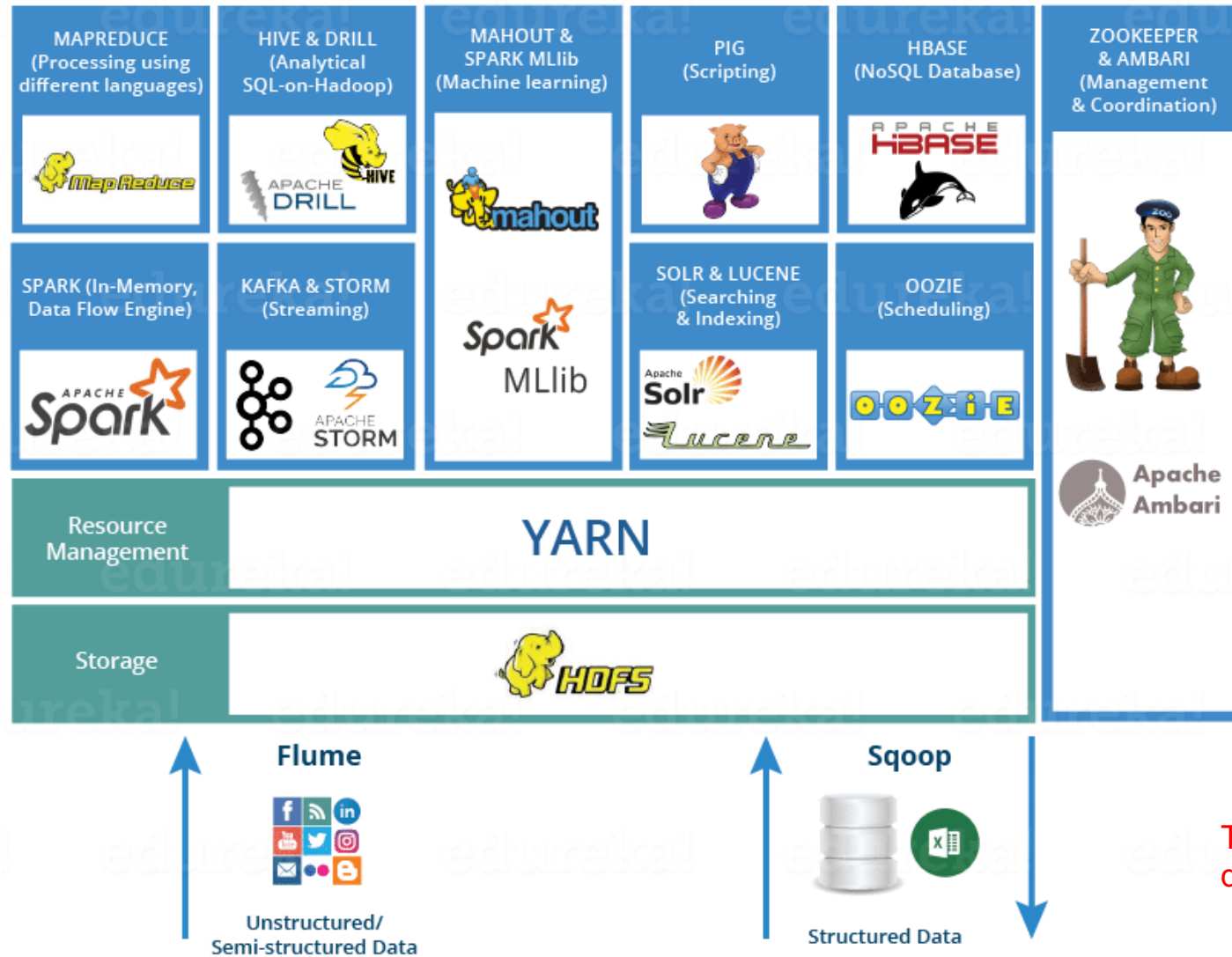
- ◎ The characteristics of Big data are characterized by the V's.

## 6 Vs of Big Data





# Big data ecosystem

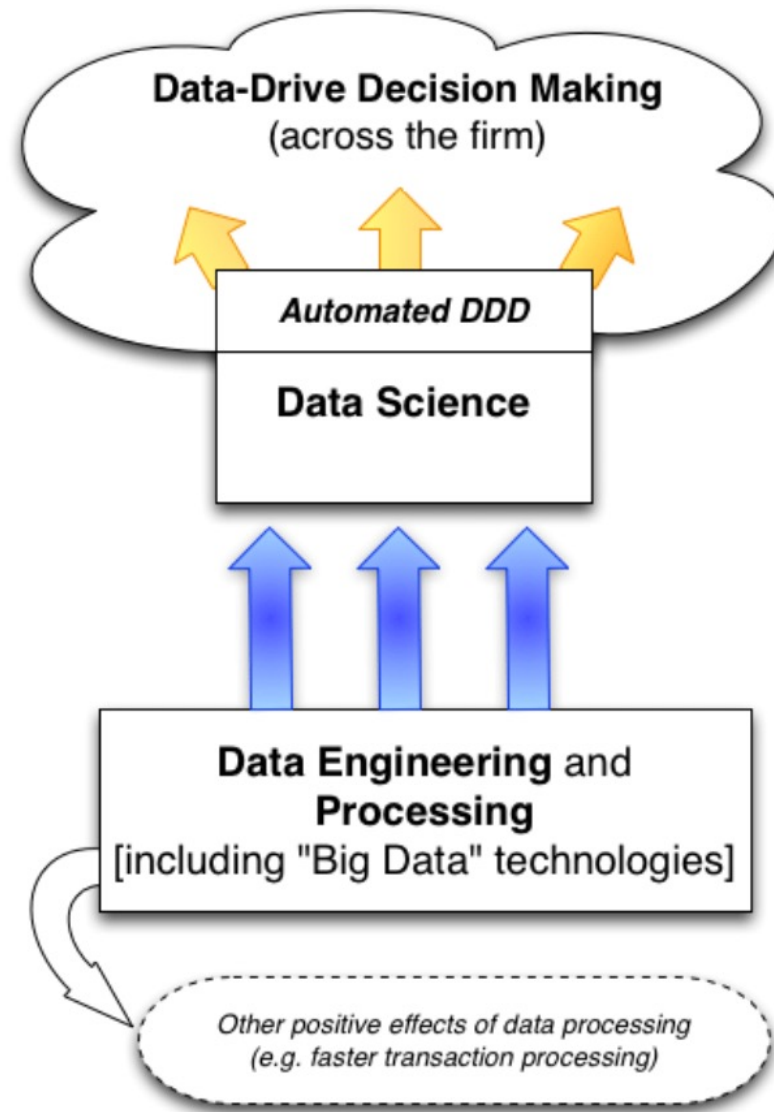


To learn about big data in more detail, enroll in the big data course

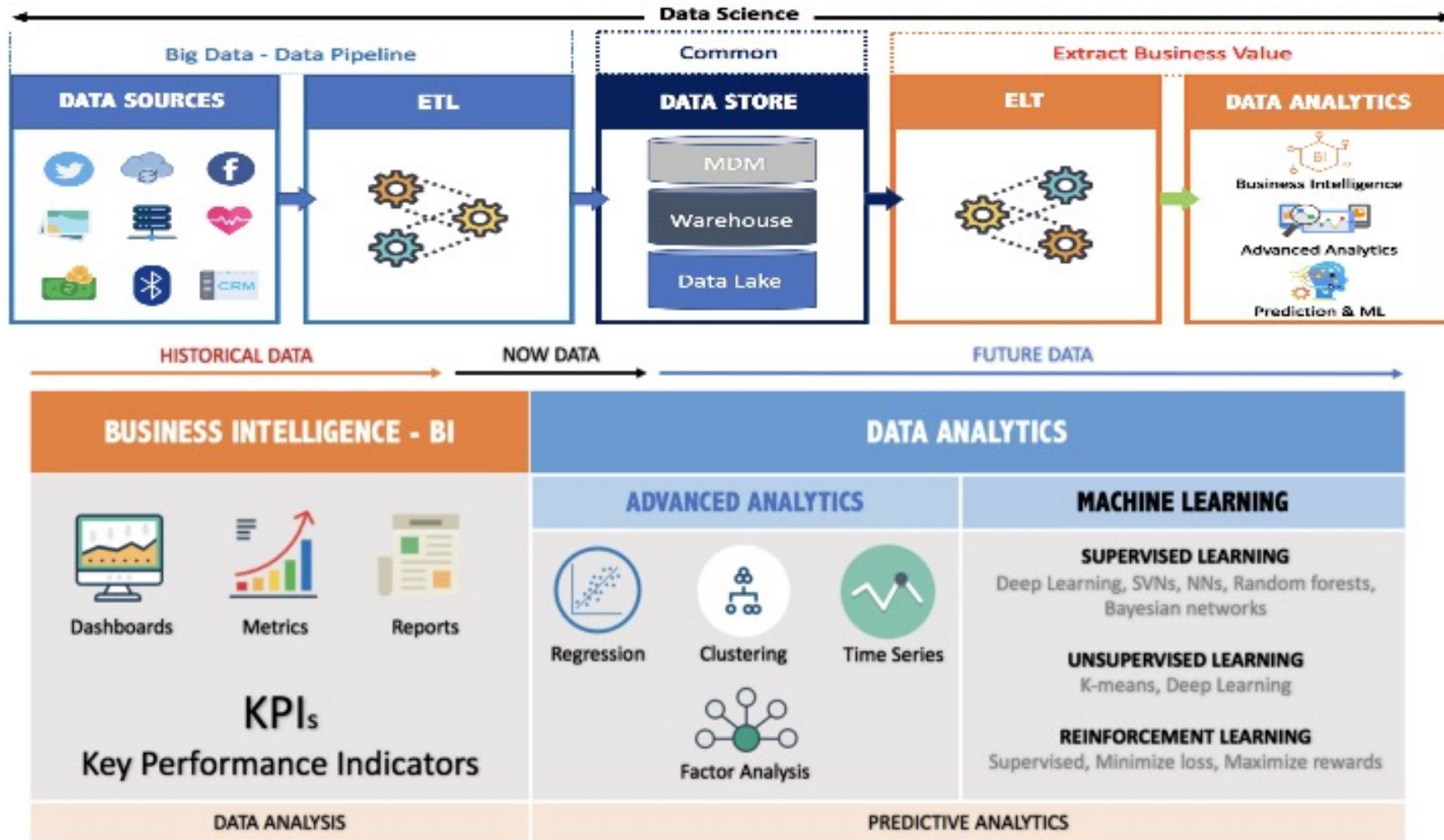
# Contents

- © Introduction to Big Data
- © Big data architecture
- © **Big data and data science**
- © Parallel and distributed computing

# Big Data and Data Science



# Big Data and Data Science



# Contents

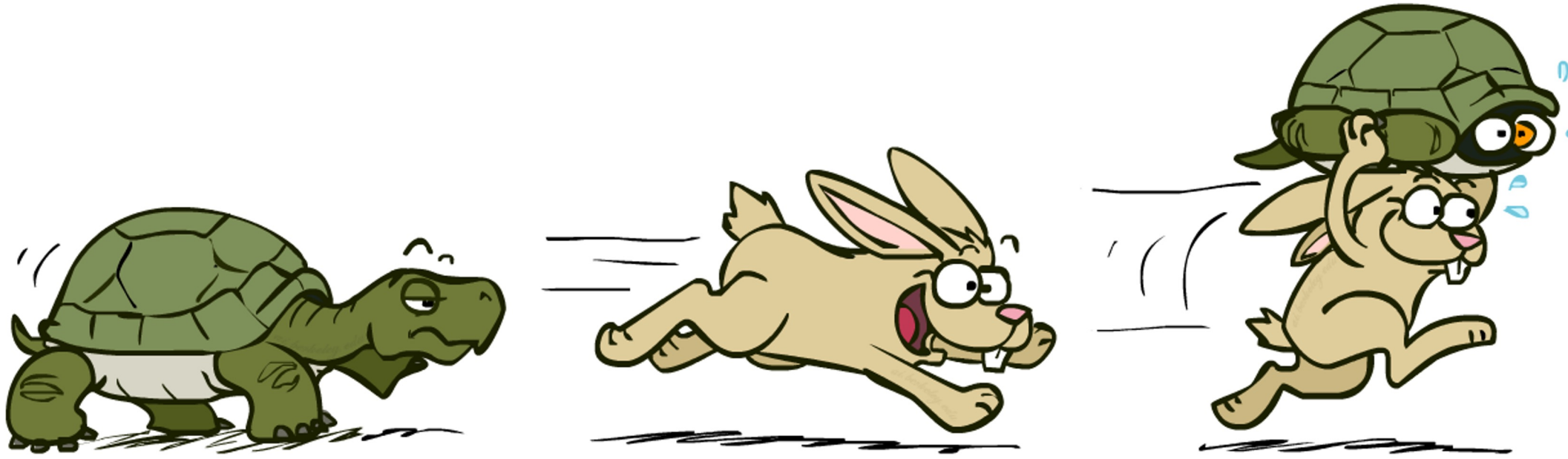
- ◎ Introduction to Big Data
- ◎ Big data architecture
- ◎ Big data and data science
- ◎ **Parallel and distributed computing**



# Massive Data Analyzing Problem

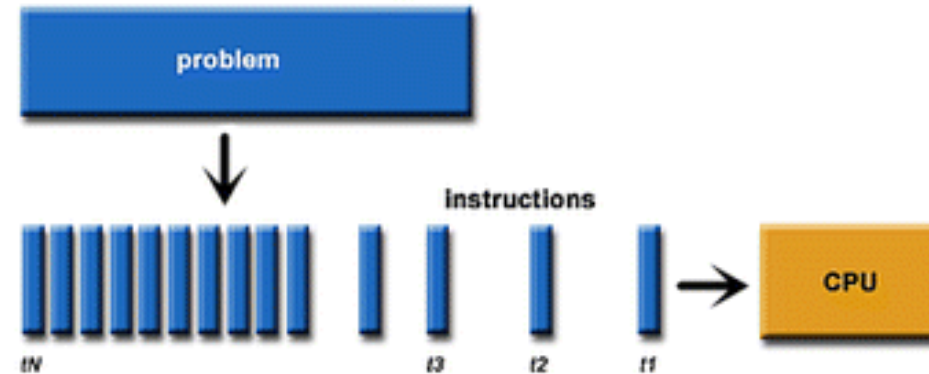


# Parallel and distributed computing

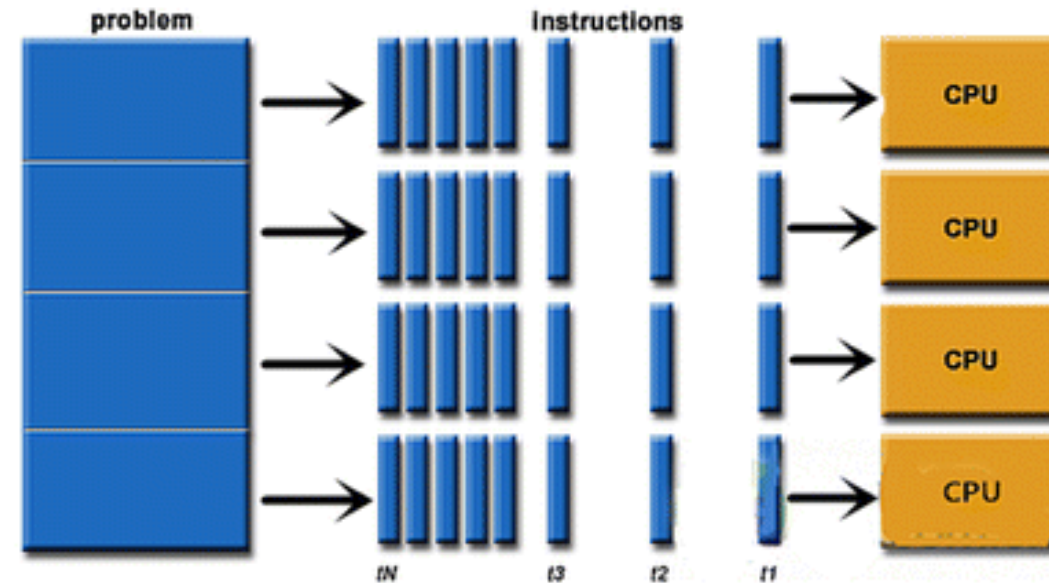


# Parallel computing

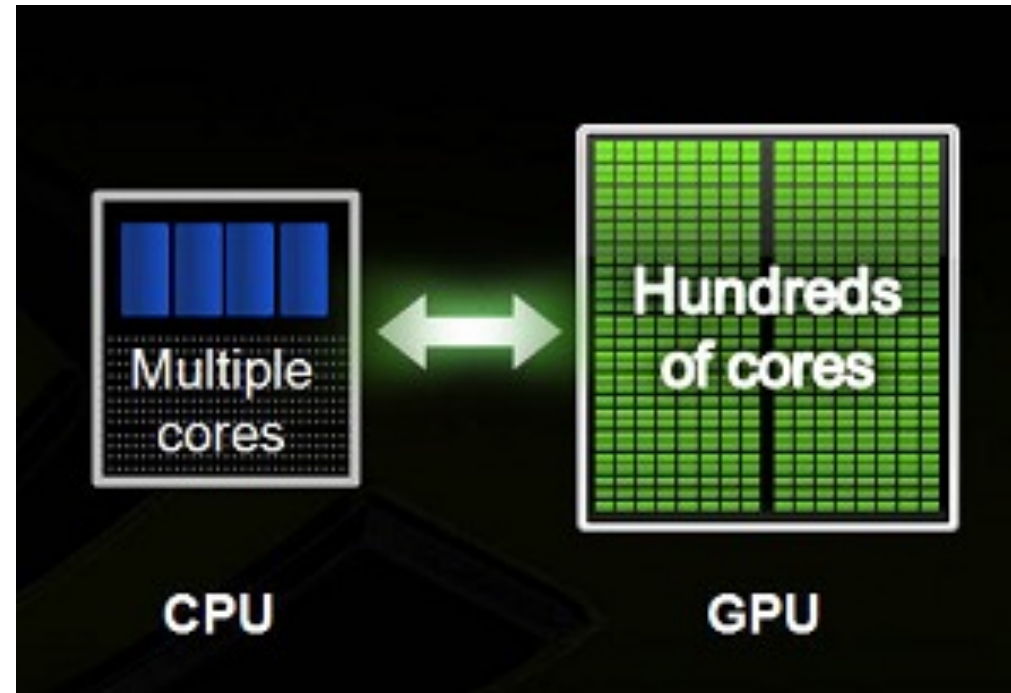
Serial operation schematic diagram



Parallel computing



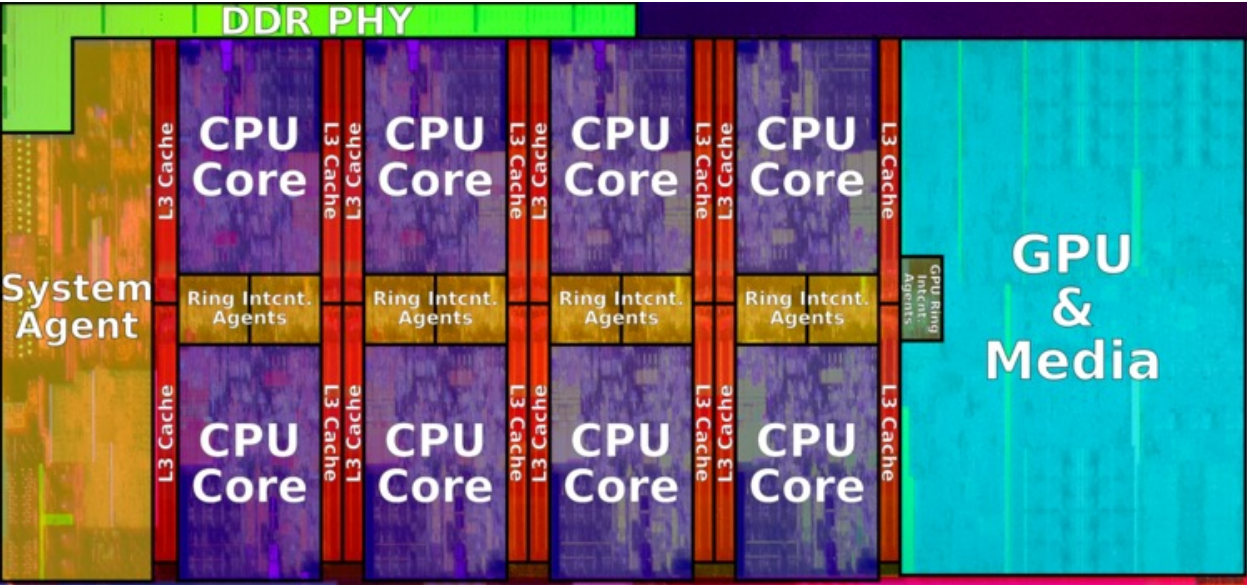
# Parallel computing with GPU



To learn parallel programming, enroll in the parallel programming with GPU course



# Limitations of parallel processing



Intel Core i9 – 9900K

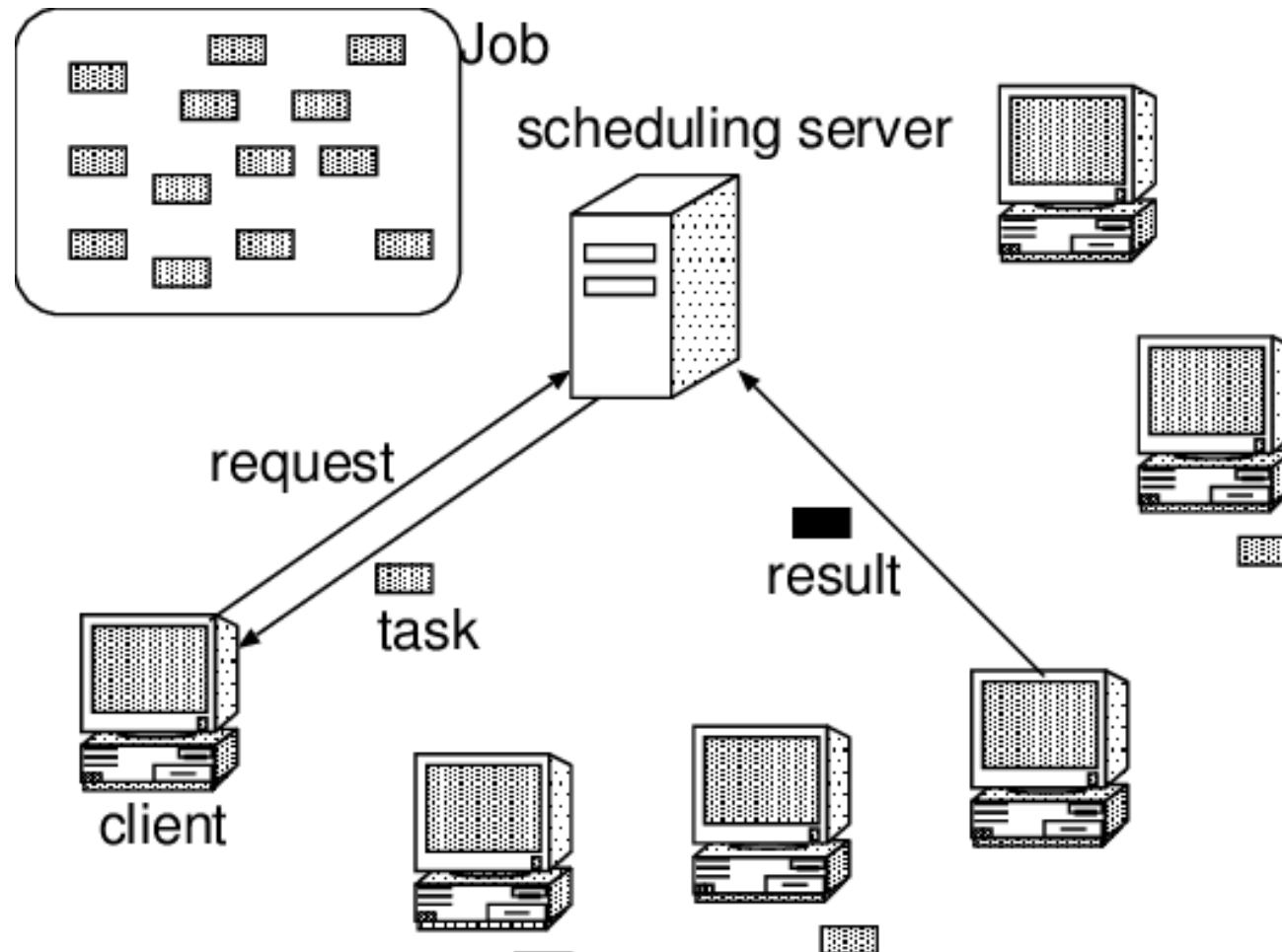


	Peak Performance
Transistor Count	54 billion
Die Size	826 mm <sup>2</sup>
FP64 CUDA Cores	3,456
FP32 CUDA Cores	6,912
Tensor Cores	432
Streaming Multiprocessors	108
FP64	9.7 teraFLOPS
FP64 Tensor Core	19.5 teraFLOPS
FP32	19.5 teraFLOPS
TF32 Tensor Core	156 teraFLOPS   312 teraFLOPS*
BFLOAT16 Tensor Core	312 teraFLOPS   624 teraFLOPS*
FP16 Tensor Core	312 teraFLOPS   624 teraFLOPS*
INT8 Tensor Core	624 TOPS   1,248 TOPS*
INT4 Tensor Core	1,248 TOPS   2,496 TOPS*
GPU Memory	40 GB
GPU Memory Bandwidth	1.6 TB/s
Interconnect	NVLink 600 GB/s PCIe Gen4 64 GB/s
Multi-Instance GPUs	Various Instance sizes with up to 7MIGs @5GB
Form Factor	4/8 SXM GPUs in HGX A100
Max Power	400W (SXM)

GPU Tesla A100

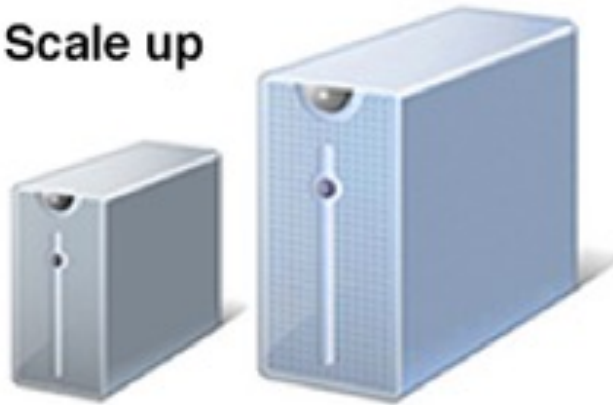


# Distributed computing



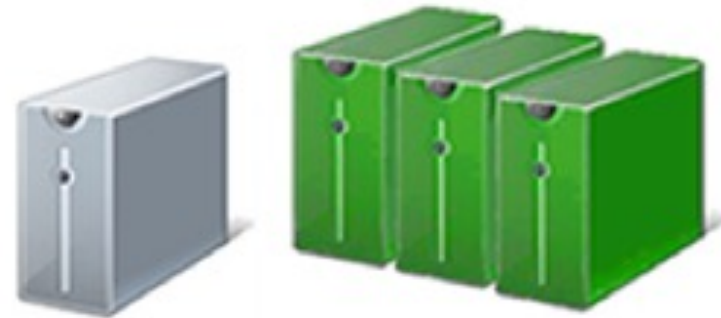
# Distributed computing

**Scale up**



**Get a larger server  
or larger data arrays**

**Scale out**

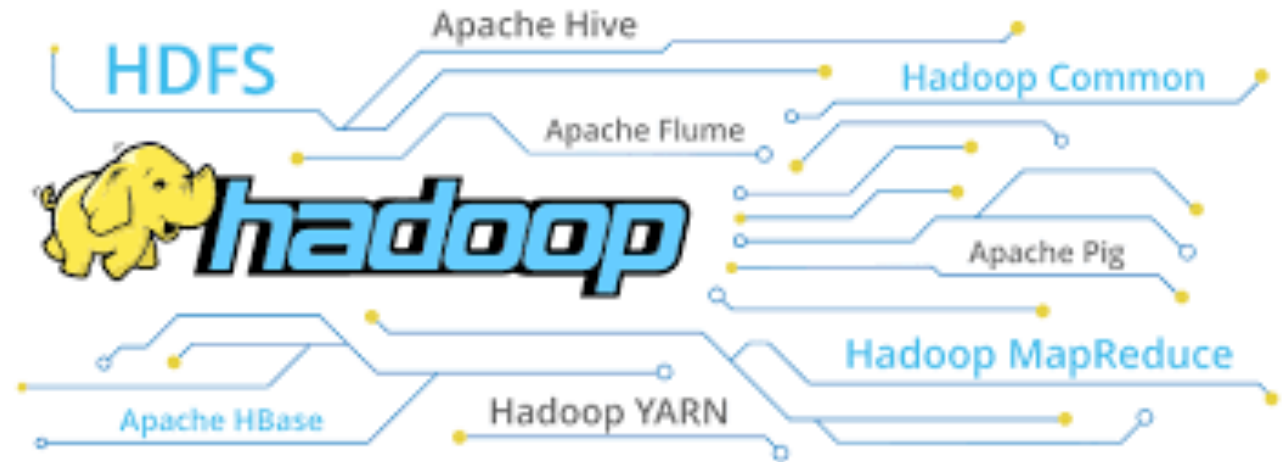


**Distribute the data and workload  
over several servers**

# Distributed computing

◎ Some terms are related to:

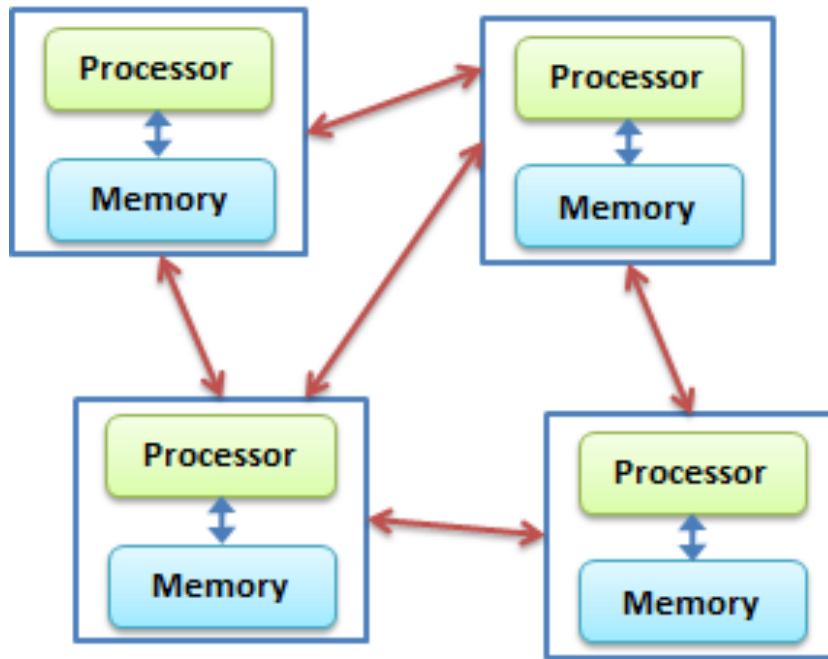
- Cloud computing
- Grid computing
- Cluster computing
- Network computing
- Edge computing
- Fog computing



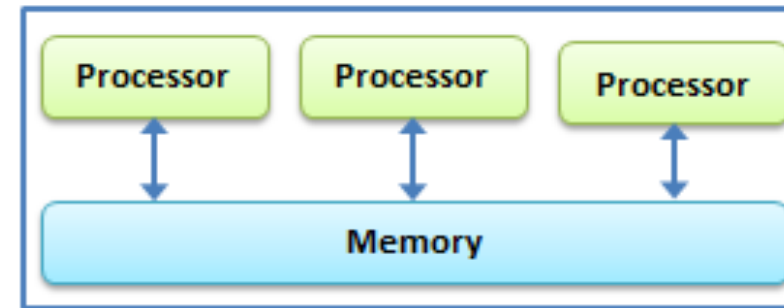
To learn distributed programming,  
enroll in Big Data course, Distributed  
computing course, ...

# Distributed vs Parallel Computing


**Distributed Computing**



**Parallel Computing**



Cooperate



*The End*