University of Science, VNU-HCM
Faculty of Information Technology

Introduction to Data Science Course

Big Data
Parallel and Distributed Computing

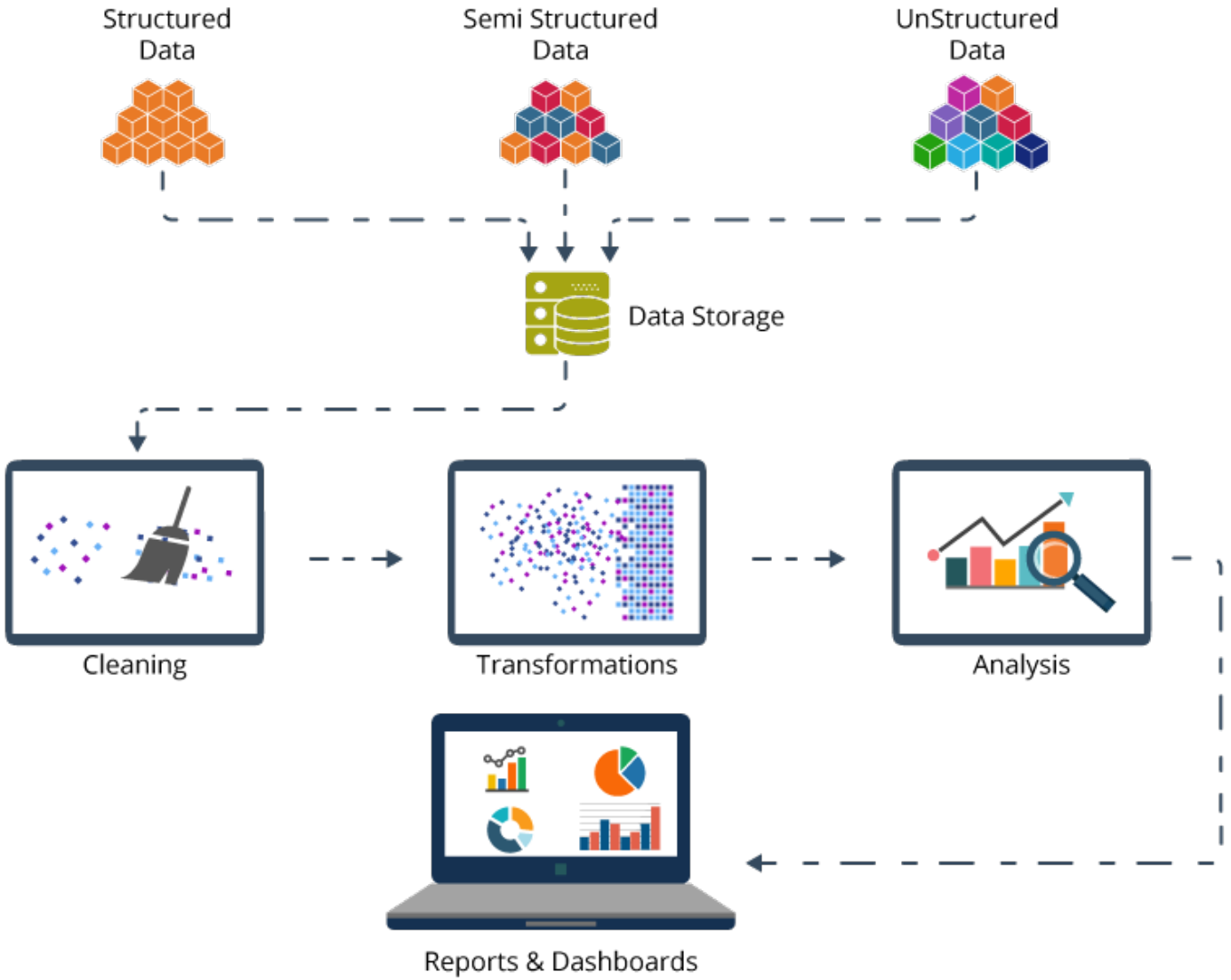
Le Ngoc Thanh
lnthanh@fit.hcmus.edu.vn
Department of Computer Science

Ho Chi Minh City

Contents

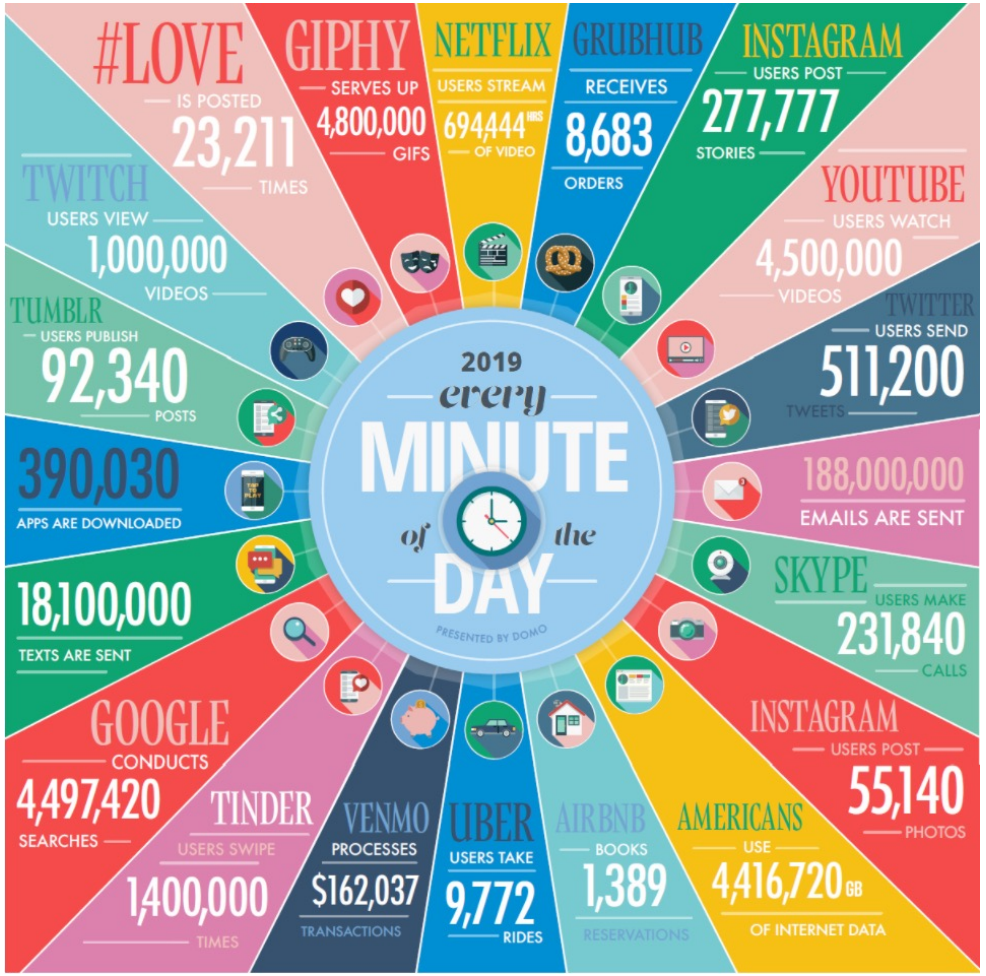
- **Introduction to Big Data**
- Big data architecture
- Big data and data science
- Parallel and distributed computing

Data Science Process

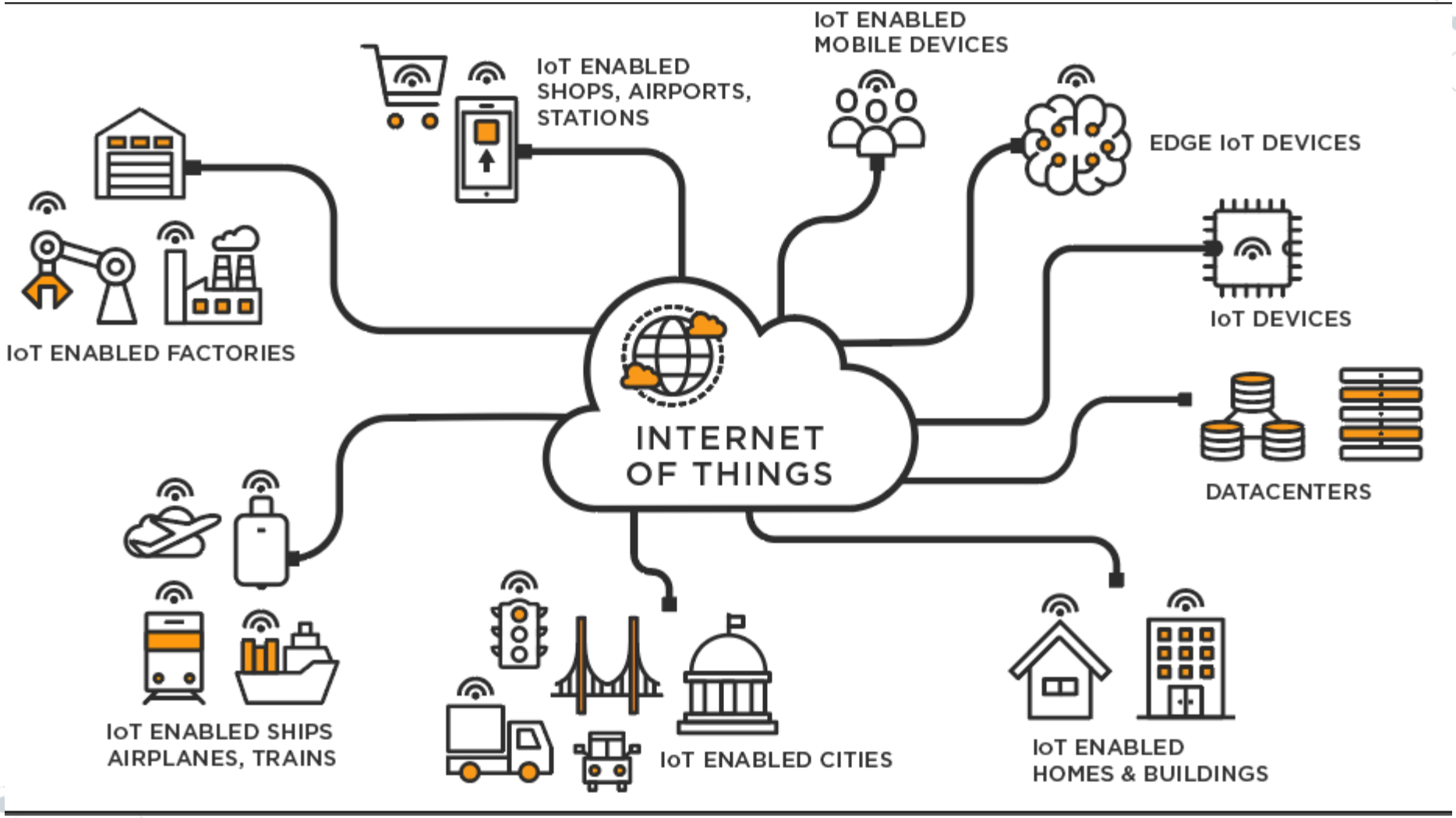


Data Never Sleeps

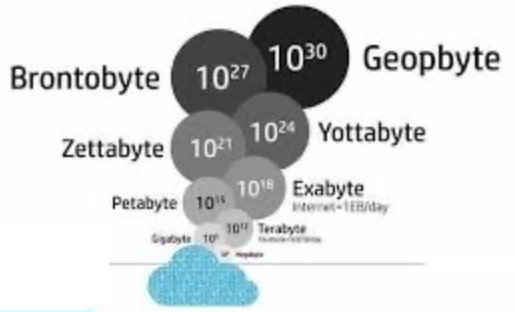
- How much data is generated every minute?

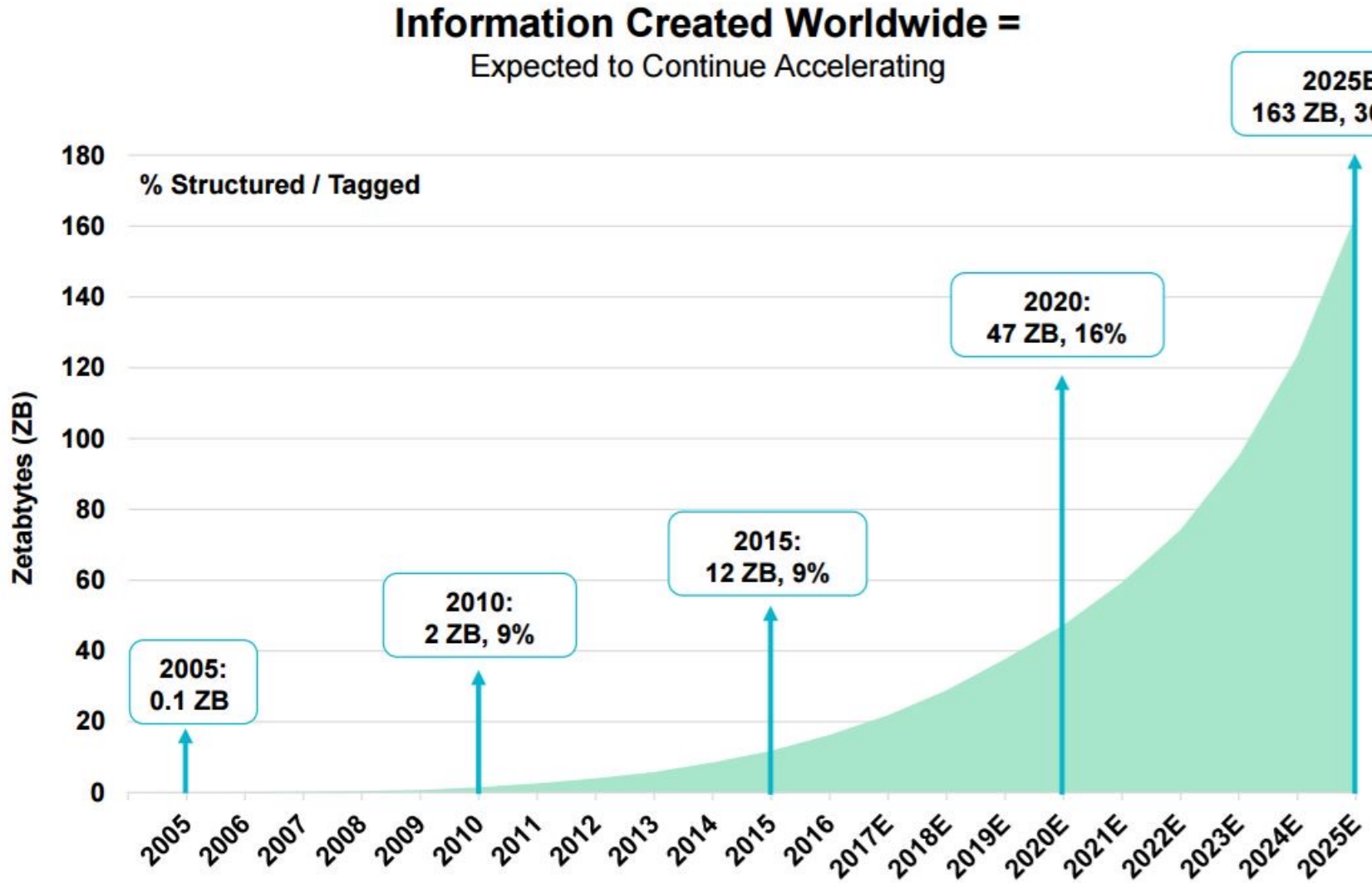


Data Never Sleeps



Data Growth


Brontobyte
10^{27}
10^{30}
Geopbyte
Zettabyte
10^{21}
10^{24}
Yottabyte
Petabyte
10^{15}
10^{18}
Exabyte
Terabyte

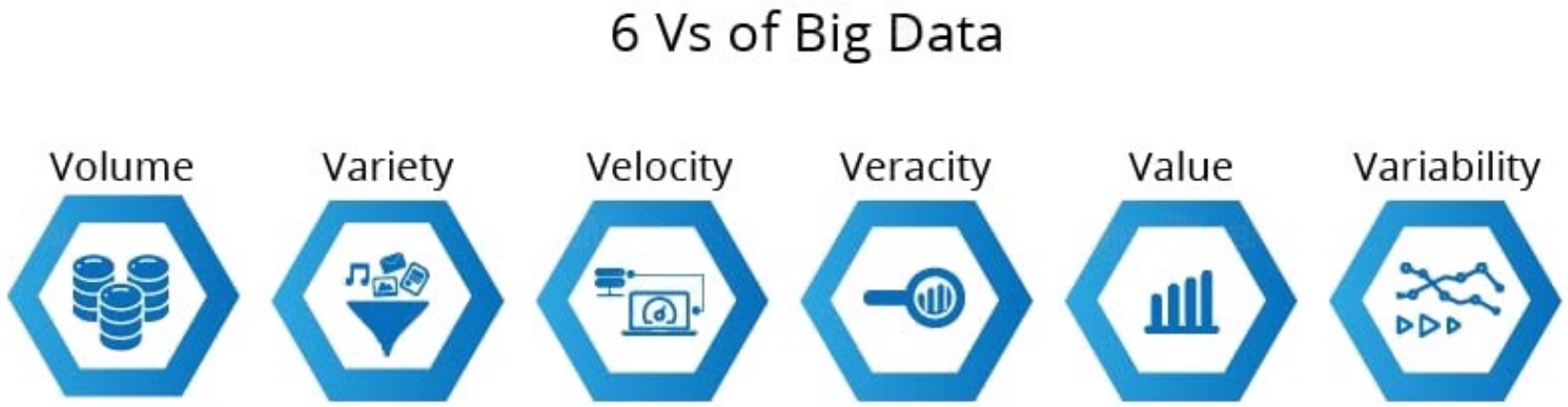

Information Created Worldwide =
Expected to Continue Accelerating
% Structured / Tagged
Zetabytes (ZB)
0
20
40
60
80
100
120
140
160
180
2005:
0.1 ZB
2010:
2 ZB, 9%
2015:
12 ZB, 9%
2020:
47 ZB, 16%
2025E:
163 ZB, 36%
2005
2006
2007
2008
2009
2010
2011
2012
2013
2014
2015
2016
2017E
2018E
2019E
2020E
2021E
2022E
2023E
2024E
2025E

What is Big Data

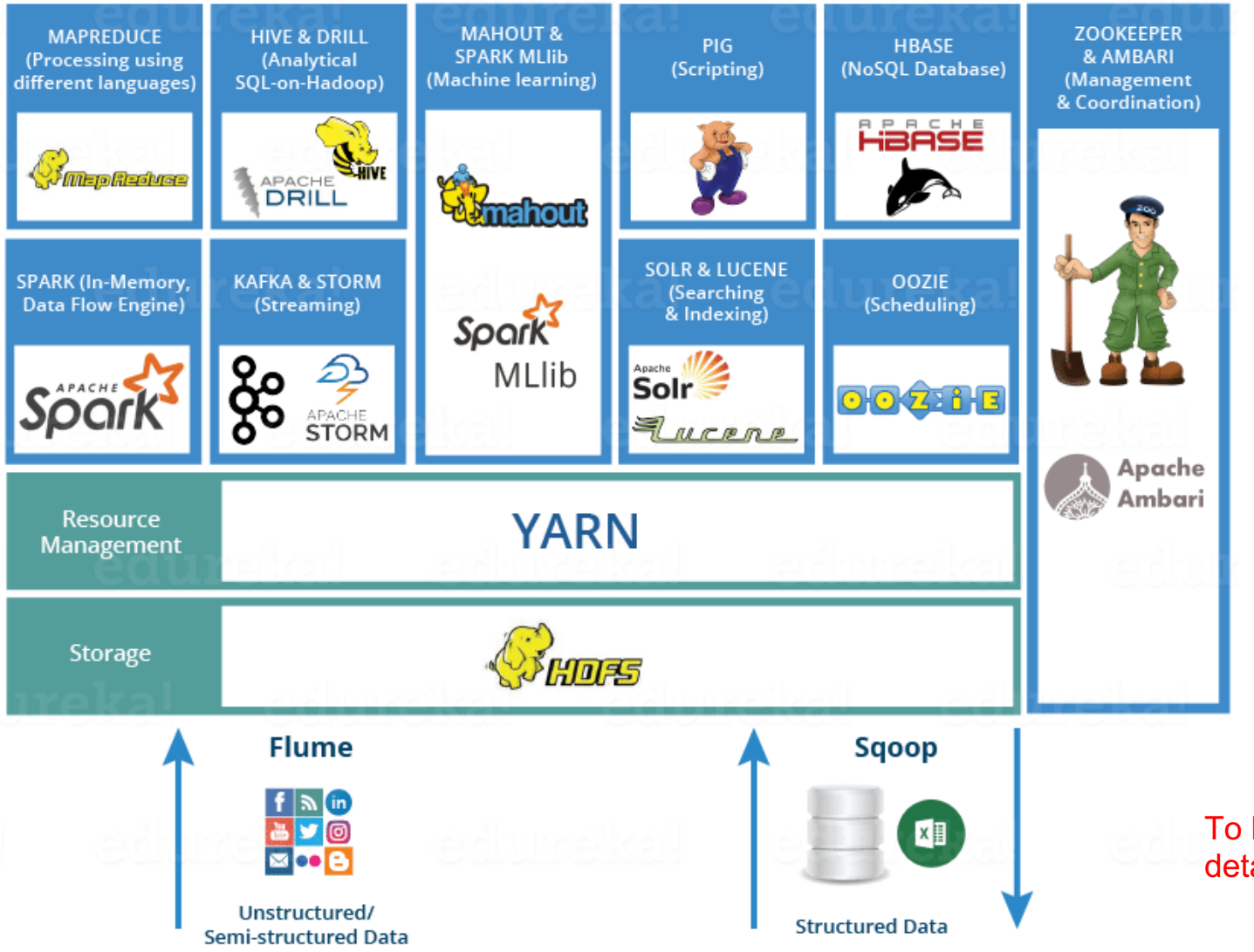
- Big data is term used to describe the massive volume of both structured and unstructured data that is so large it is difficult to process using traditional techniques.

Characteristics of Big data

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



Big data ecosystem

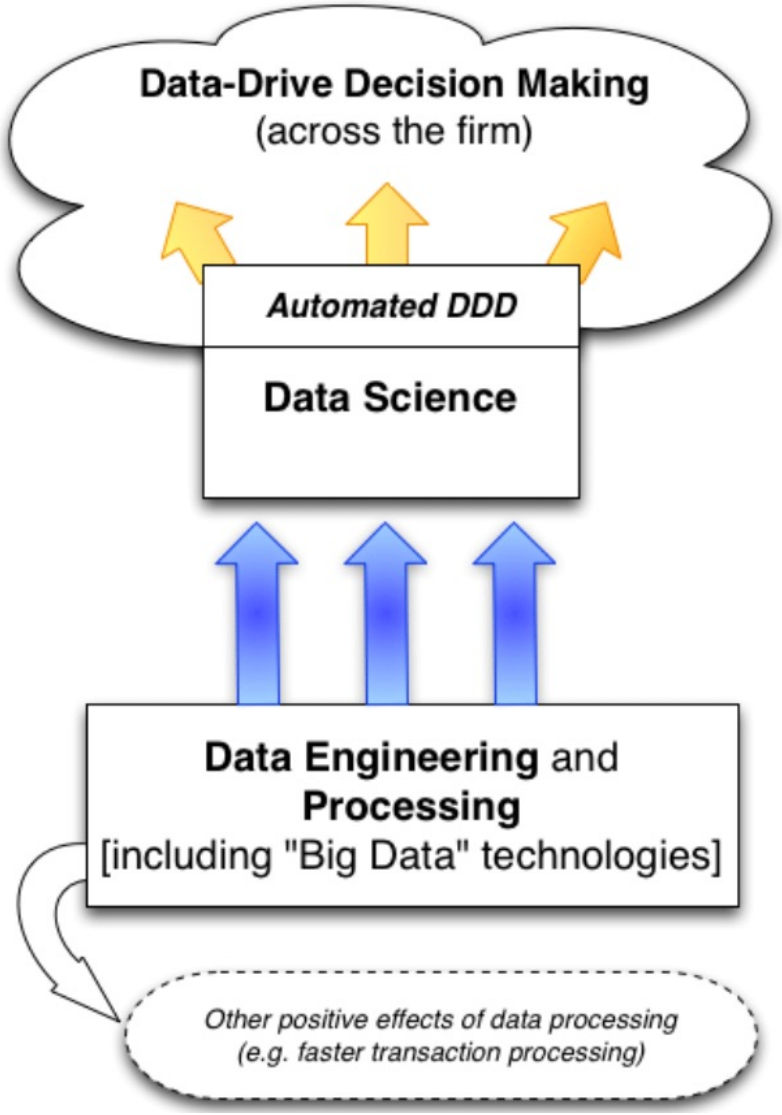


To learn about big data in more detail, enroll in the big data course

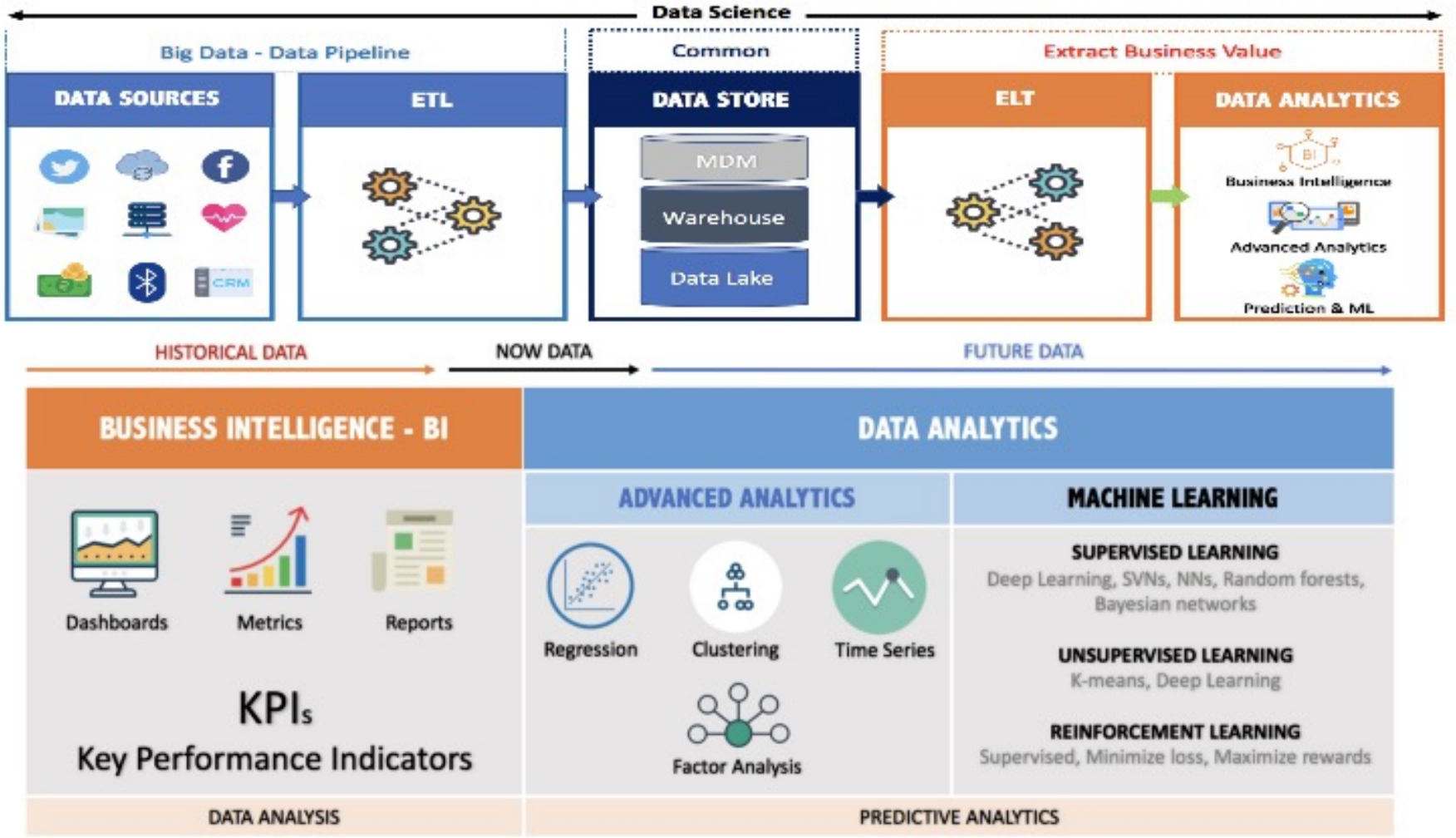
Contents

- 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

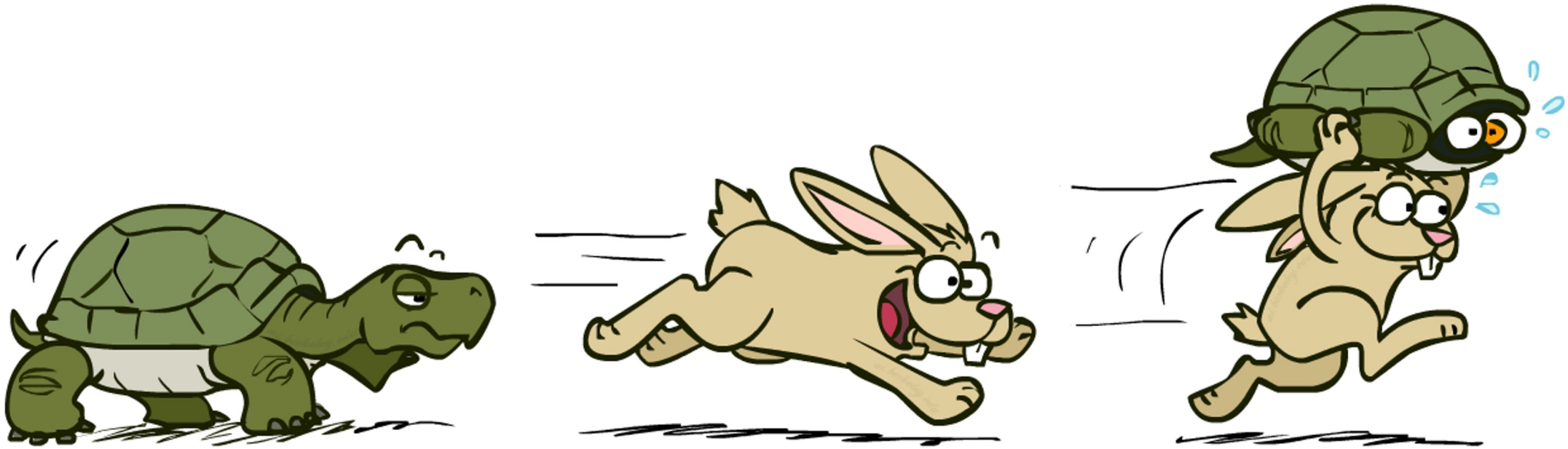


Contents

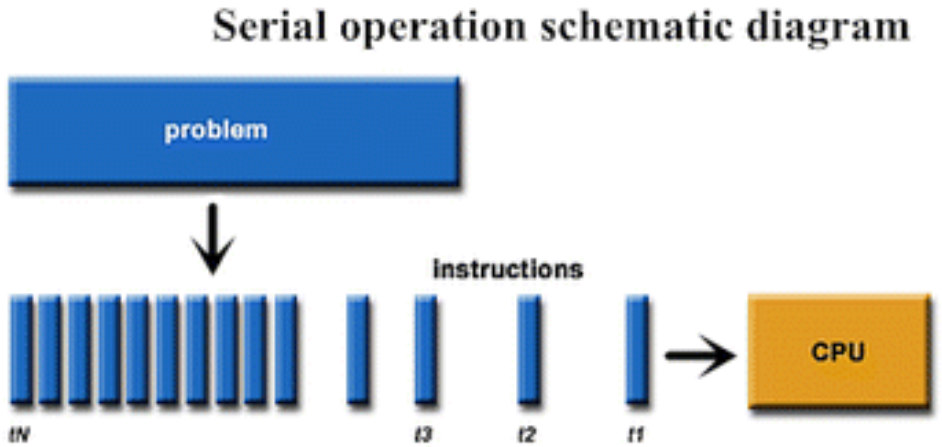
- Introduction to Big Data
- Big data architecture
- Big data and data science
- **Parallel and distributed computing**

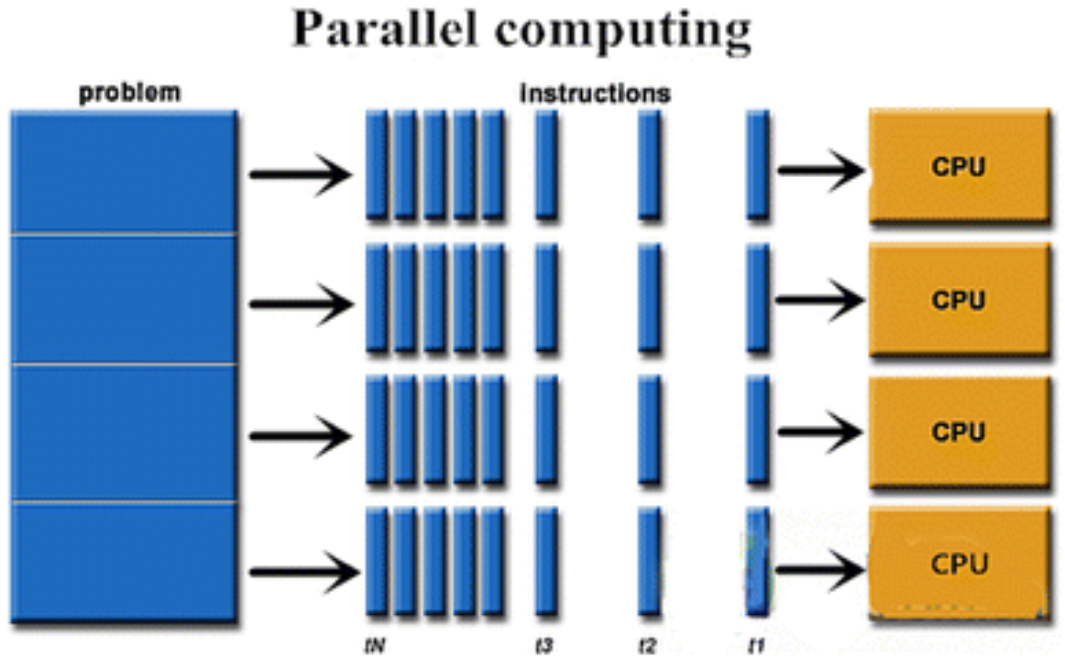
Massive Data Analyzing Problem

Parallel and distributed computing

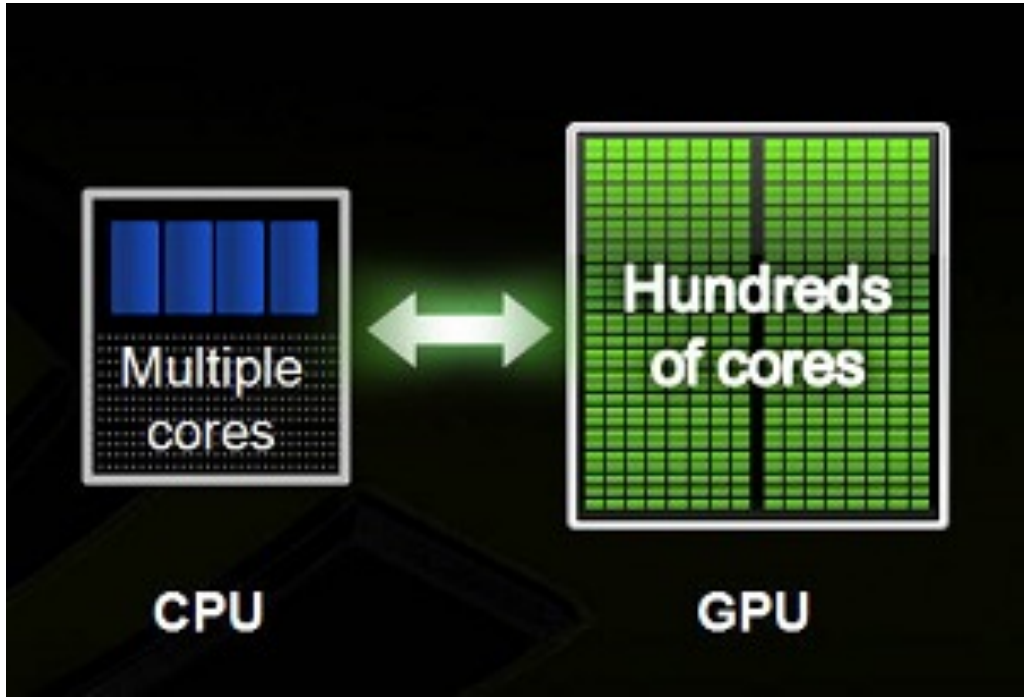


Parallel computing



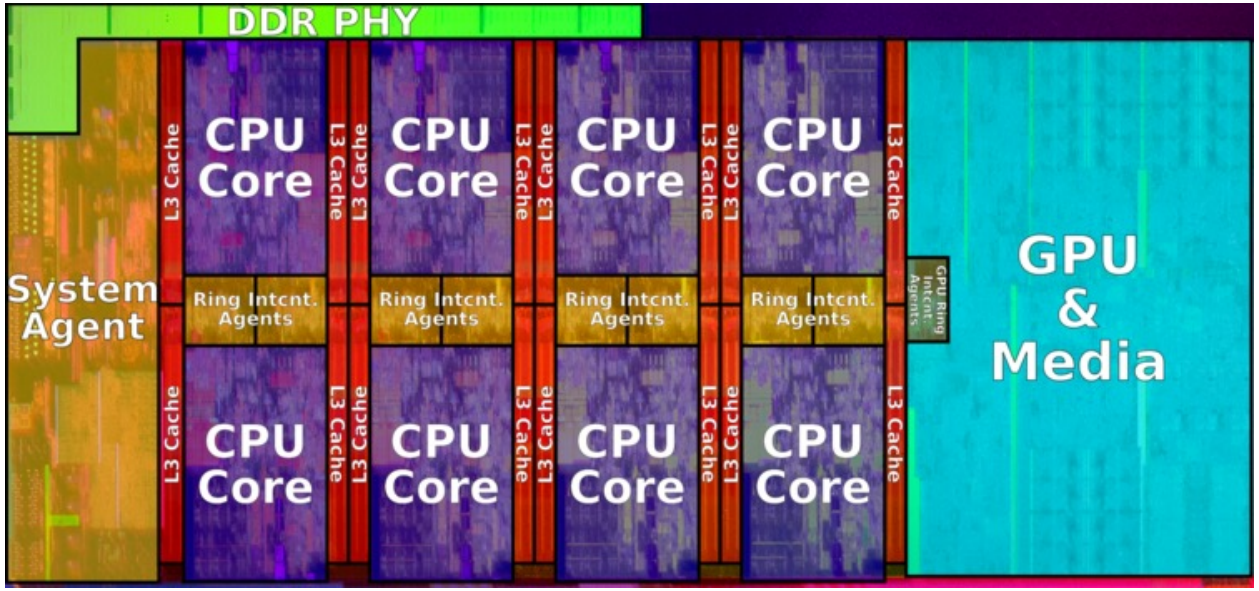


Parallel computing with GPU



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

Limitations of parallel processing

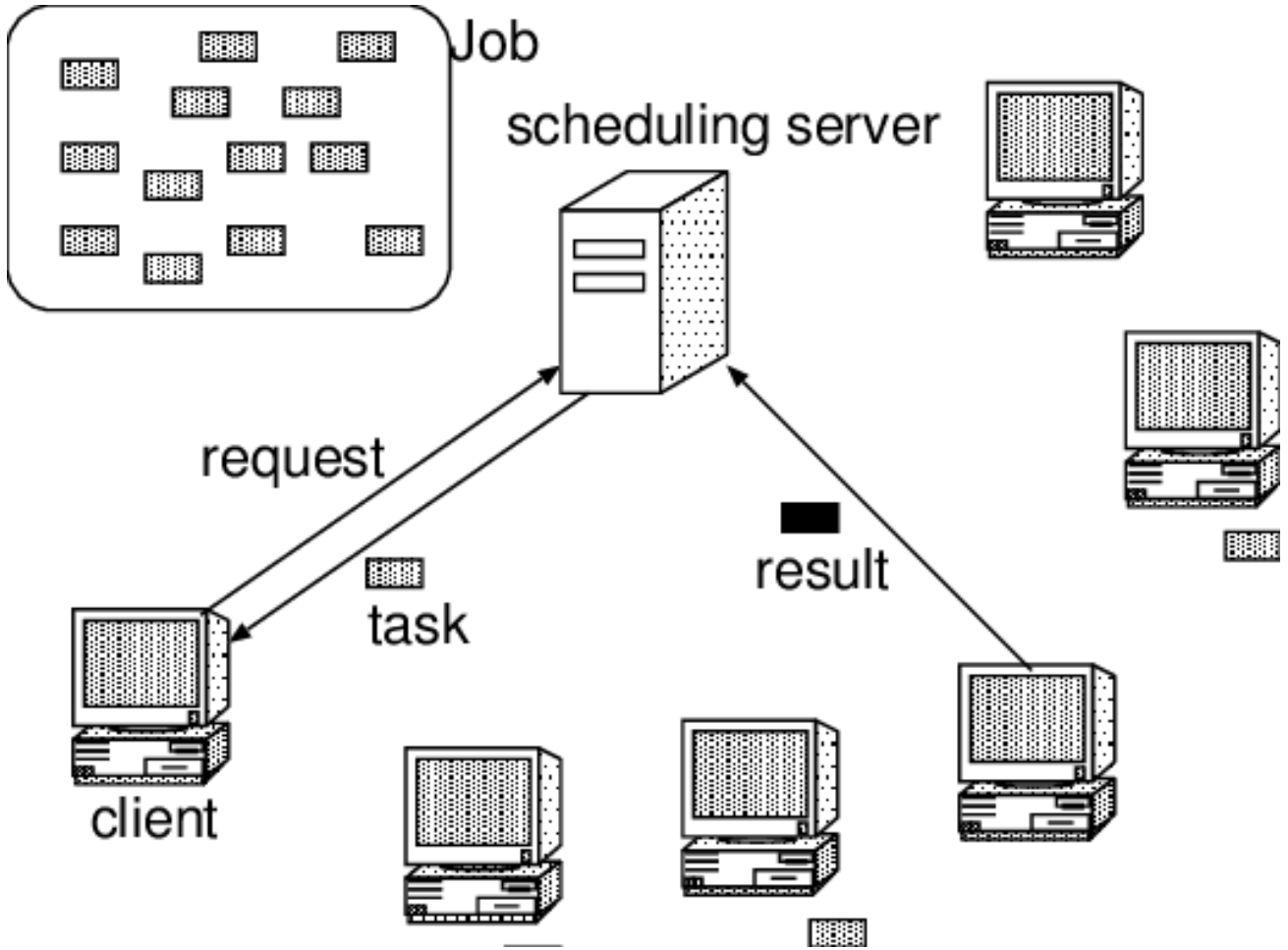


Intel Core i9 – 9900K

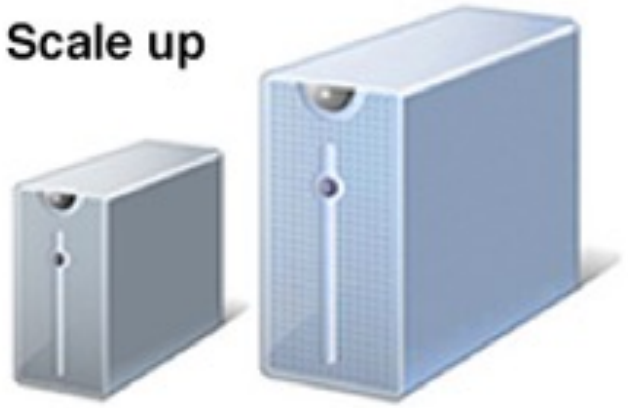
	Peak Performance
Transistor Count	54 billion
Die Size	826 mm^2
FP64 CUDA Cores	3,456
FP32 CUDA Cores	6,912
Tensor Cores	432
Streaming Multiprocessors	108
FP64	9.7 teraFLOPS
FP64 Tensor Core	19.5 teraFLOPS
FP32	19.5 teraFLOPS
TF32 Tensor Core	156 teraFLOPS | 312 teraFLOPS*
BFLOAT16 Tensor Core	312 teraFLOPS | 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

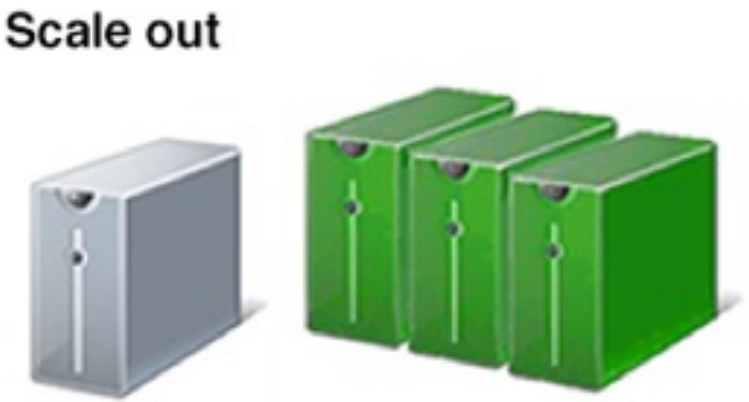
Distributed computing



Distributed computing



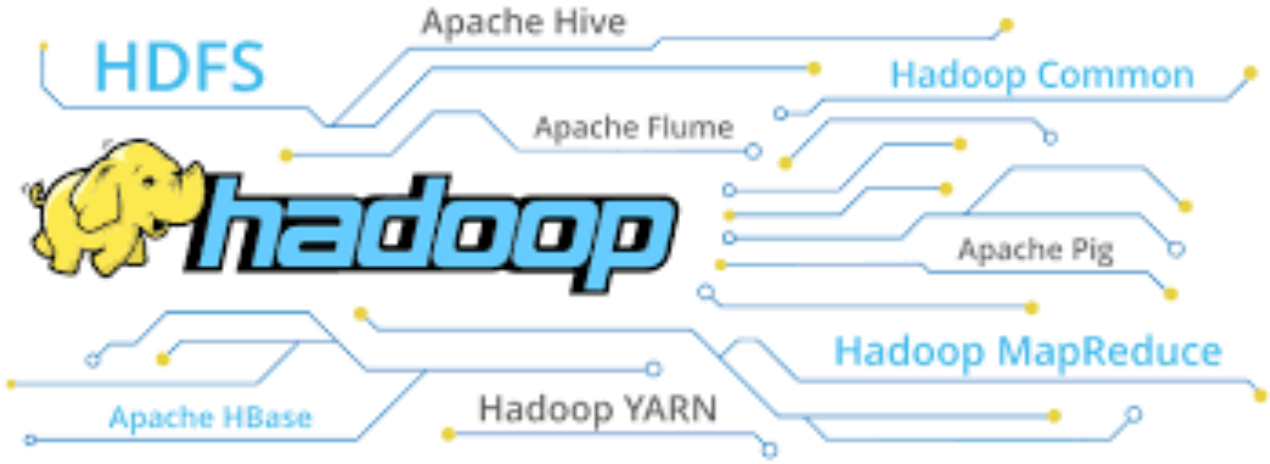
Get a larger server
or larger data arrays



Distribute the data and workload
over several servers

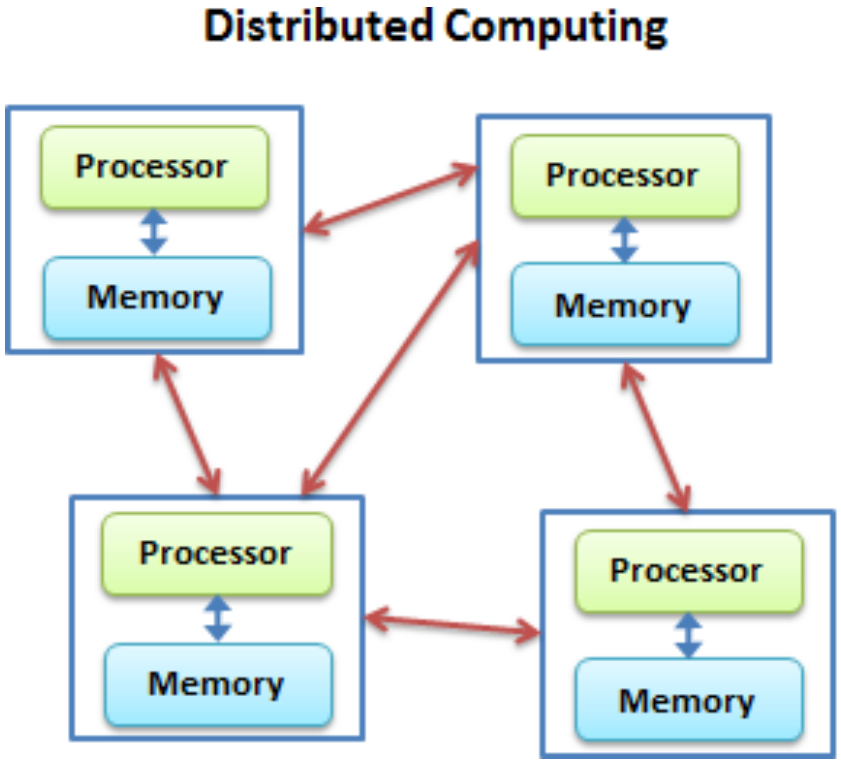
Distributed computing

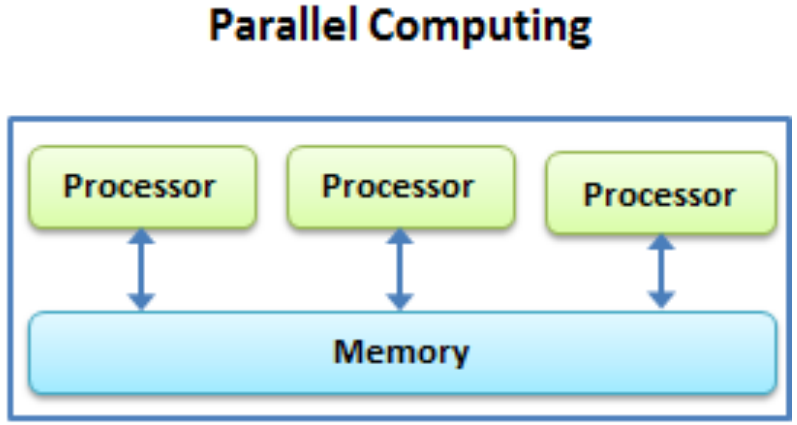
- Some terms are related to:
  - Cloud computing
  - Grid computing
  - Cluster computing
  - Network computing
  - Edge computing
  - Fog computing



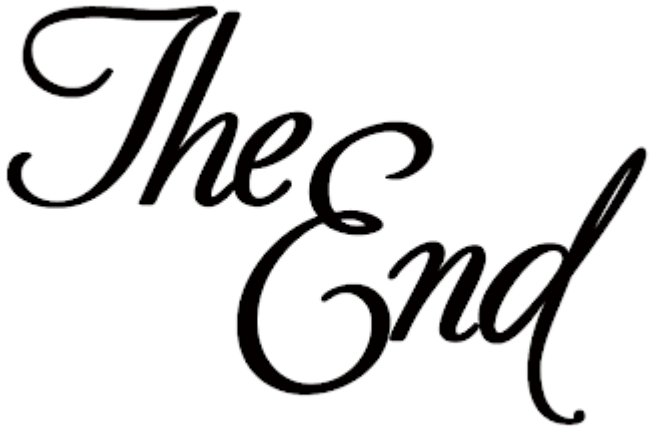
To learn distributed programming, enroll in Big Data course, Distributed computing course, …

Distributed vs Parallel Computing





Cooperate


The End