Big Data Analytics

By Metin Senturk

So far…

- Introduced few computer science concepts, introduction to python syntax
    - Inheritance, Abstraction, etc.
- Understand what parallelism means, using our stack effectively
    - One vs multiple CPUs, threads, processes, concurrency vs parallel, etc.

Outline

- Introduction to Big Data
- Clustering Systems
- Algorithms

Definition of Big Data

- The term “big data” appeared first in 1997 paper from NASA scientists [1].
  - “data sets are generally quite large, taxing the capacities of main memory, local disk, and even remote disk. We call this the problem of big data.
  - When data sets do not fit in main memory (in core), or when they do not fit even on local disk, the most common solution is to acquire more resources.”
- In 2001, Doug Laney introduced the concept of 3Vs.
  - 3-D Data Management: Controlling Data Volume, Velocity and Variety
  - (Today, people also add Variability, and Value, Veracity, etc.)
- 2008, big data popularized by the American computer scientists
  - “transform the activities of companies, scientific researchers, medical practitioners, and our nation’s defence and intelligence operations.”

[1] https://dl.acm.org/citation.cfm?id=266989.267068&coll=DL&dl=GUIDE

Definition of Big Data

- 2014, the definition of big data added to Wikipedia
  - Big data is a field that treats ways to analyse, systematically extract information from, or otherwise deal with data sets that are too large or complex to be dealt with by traditional data-processing application software.
- Followed by Wiki, Oxford English Dictionary defined Big Data as
  - Computing (also with capital initials) data of a very large size, typically to the extent that its manipulation and management present significant logistical challenges; (also) the branch of computing involving such data.[1]

[1] https://www.oed.com/view/Entry/18833

Three Vs

Definitions

1. Volume, or the total amount of data stored.
2. Velocity, or how often new data is created and needs to be stored
3. Variety, or how heterogeneity your data structures and sources are

People also added

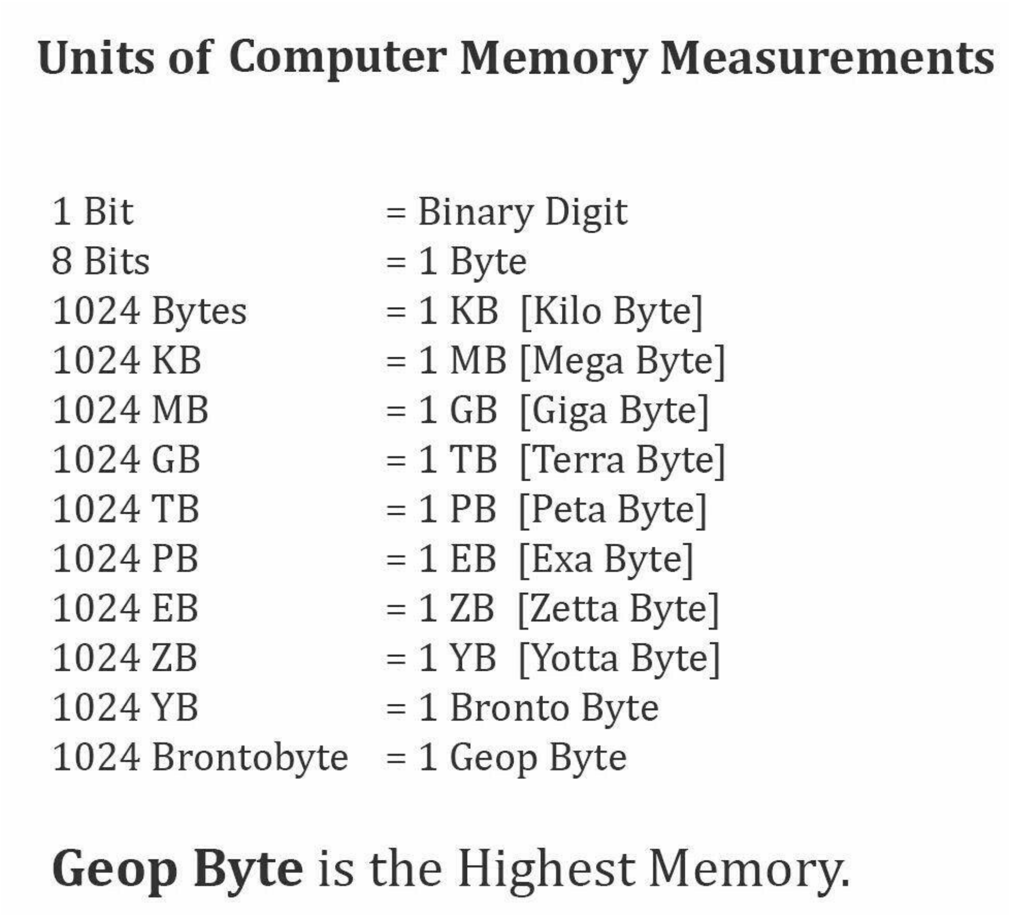
1. Veracity, or the “truthiness” and integrity of your data.
2. Value, or the significance your data to your business goals and the impact it has on the bottom line.

Exabyte (EB)

Zettabyte (ZB)

Yottabyte (YB)

Byte Scale

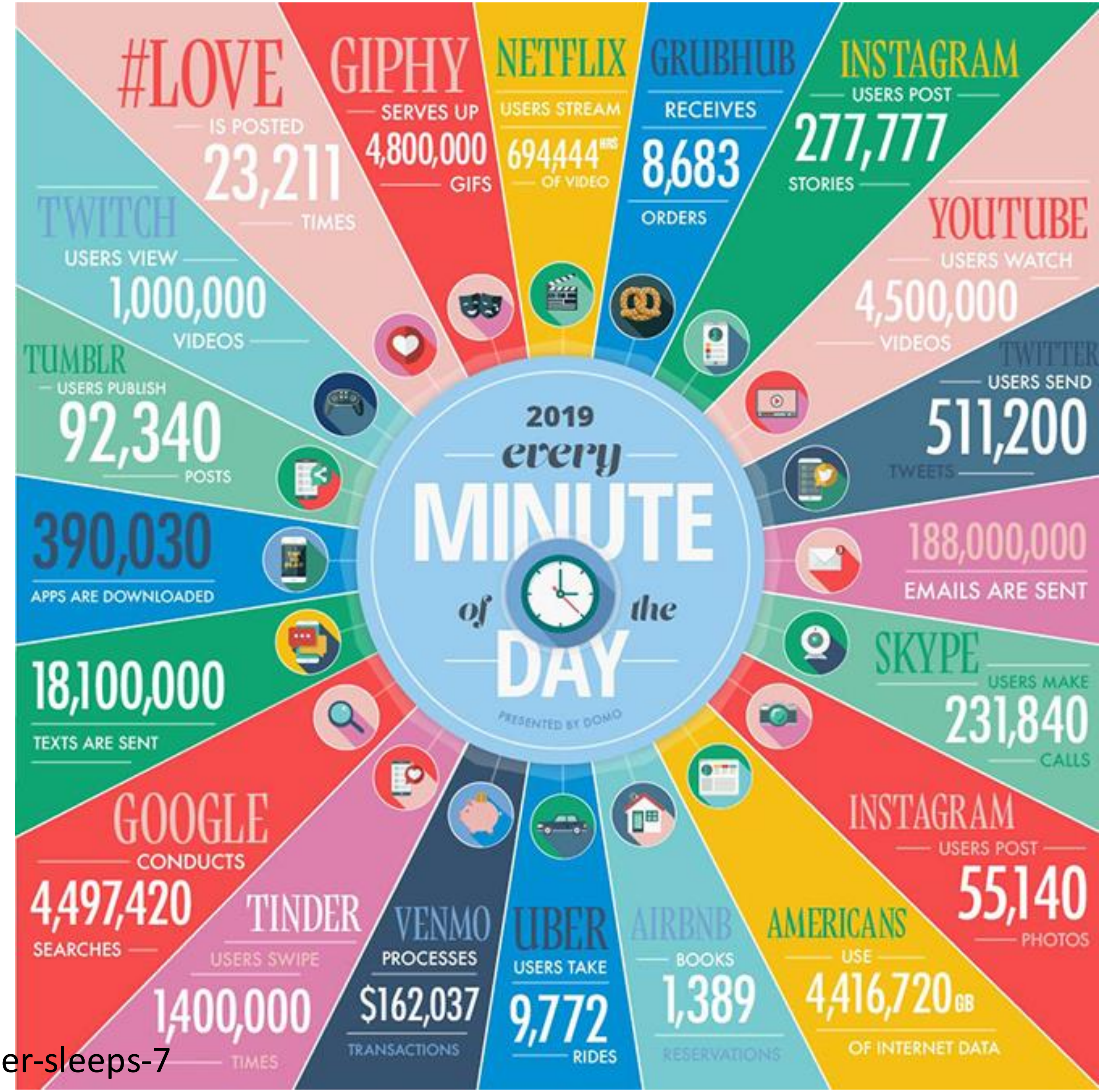


Units of Computer Memory Measurements

1 Bit	= Binary Digit
8 Bits	= 1 Byte
1024 Bytes	= 1 KB [Kilo Byte]
1024 KB	= 1 MB [Mega Byte]
1024 MB	= 1 GB [Giga Byte]
1024 GB	= 1 TB [Terra Byte]
1024 TB	= 1 PB [Peta Byte]
1024 PB	= 1 EB [Exa Byte]
1024 EB	= 1 ZB [Zetta Byte]
1024 ZB	= 1 YB [Yotta Byte]
1024 YB	= 1 Bronto Byte
1024 Brontobyte	= 1 Geop Byte

Geop Byte is the Highest Memory.

How much data generated every minute?



source: https://www.domo.com/learn/data-never-sleeps-7

How much data is generated?

- In 2020, 40 trillion gigabytes of data (40 zettabytes)
- IBM says, 90% of all data has been created in the last two years.
- 97.2% of organizations are investing in big data and AI.
- Using big data, Netflix saves $1 billion per year on customer retention.
- Wikibon says big data market is $49 billion.

Reference: https://techjury.net/stats-about/big-data-statistics/#gref

What Resources Needed for Big Data?

- Massive Parallelism
- Huge data Volume Storages
- Data Distribution
- High Speed Networks
- High Performance Computing
- Task and Thread Management

→

- Data Mining and Analytics
- Data Retrieval
- Machine Learning
- Data Visualization

Did you notice? How many of these concepts do we have already?

Big Data Use Cases

- BI, reporting, and searching.
  - Report generation, trend analysis, search optimization, etc. (Google Trends, Google, FB Ad tracking, etc.)
- Network Analysis
  - Social network relation analysis, Gene Co-expression Network, Chemical Compound Analysis, etc.
- Credit fraud detection, data profiling, recommendation engines, price modeling, energy network monitoring and optimization,

Big Data Tools and Techniques

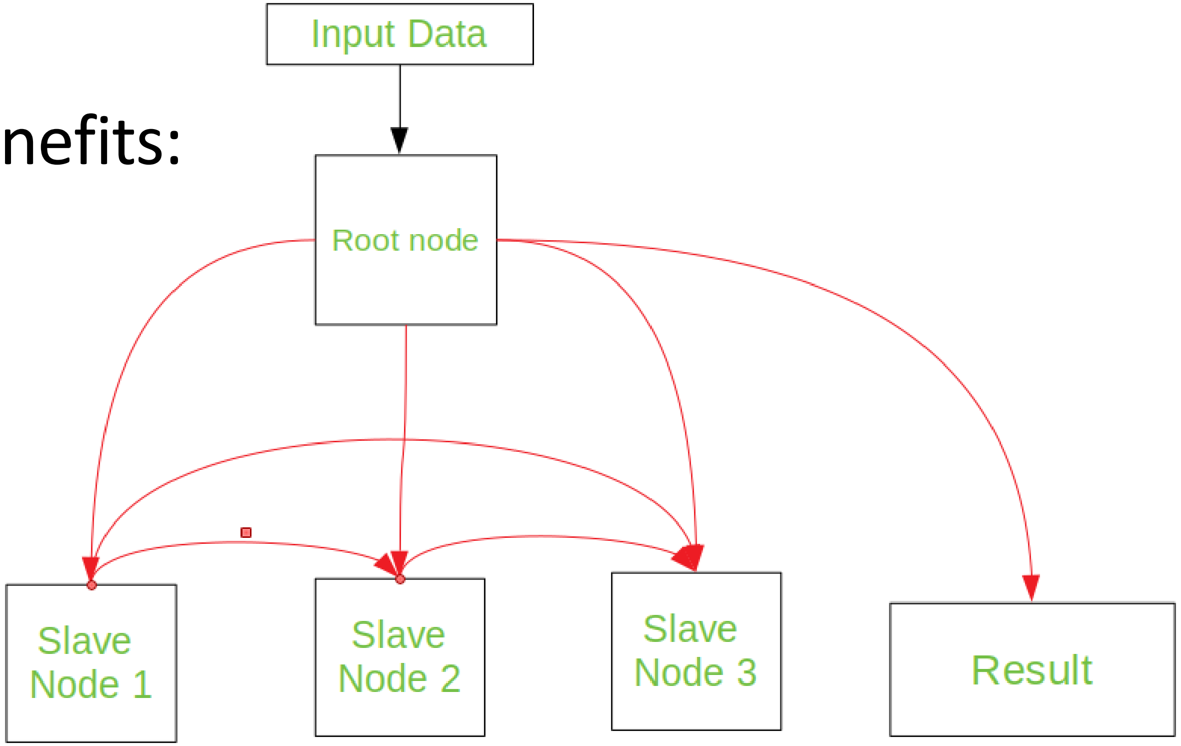
- Bigdata applications achieve their performance and scalability through deployment on a collection of storage and computing resources bound together within a runtime environment.
  - The architecture of the underlying computing platform, both from a hardware and more importantly from a software perspective.
- Four key resources:
  - 1. **Processing capability**, often referred to as a CPU, processor, or node.
  - 2. **Memory**, which holds the data that the processing node is currently working on. Most single node machines have a limit to the amount of memory.
  - 3. **Storage**, providing persistence of data—the place where datasets are loaded, and from which the data is loaded into memory to be processed.
  - 4. **Network**, which provides the “pipes” through which datasets are exchanged between different processing and storage nodes

Where to go?

- https://blog.microfocus.com/how-much-data-is-created-on-the-internet-each-day/#
- 4Vs of big data by IBM, https://www.ibmbigdatahub.com/infographic/four-vs-big-data
- Berkeley Uni. https://datascience.berkeley.edu/blog/what-is-big-data/
- Awesome book about clustering systems, http://dx.doi.org/10.1016/B978-0-12-417319-4.00007-7

What is a Clustering System?

- Computers (servers) and other resources that are connected through hardware, networks and software to behave as if they were a single system.
- Why adopt clustering systems? Benefits:
  - High availability
  - Load balancing
  - Parallel processing
  - Systems management
  - Scalability
  - Fault tolerancy

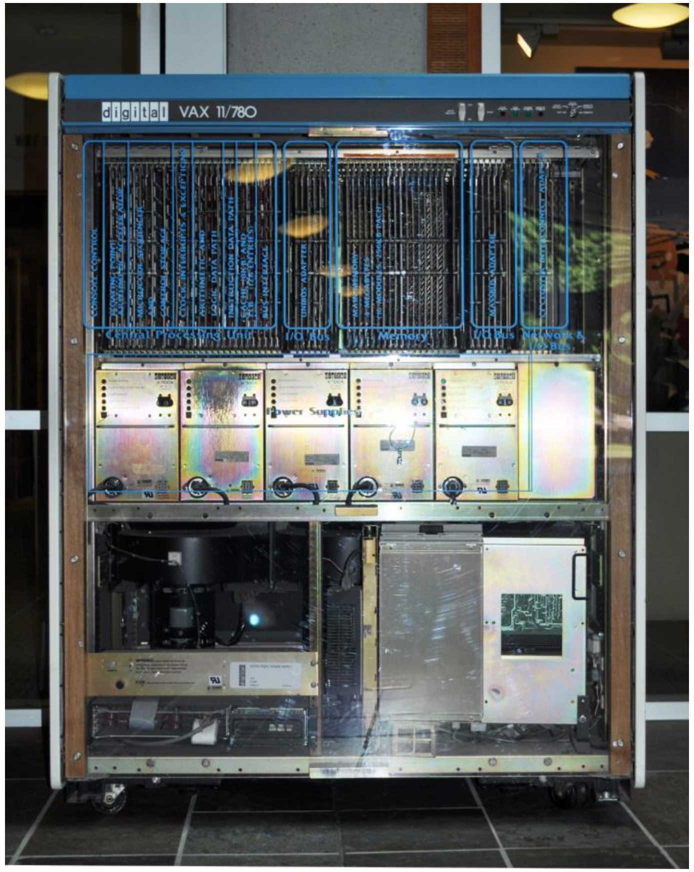


Type of Cluster Systems

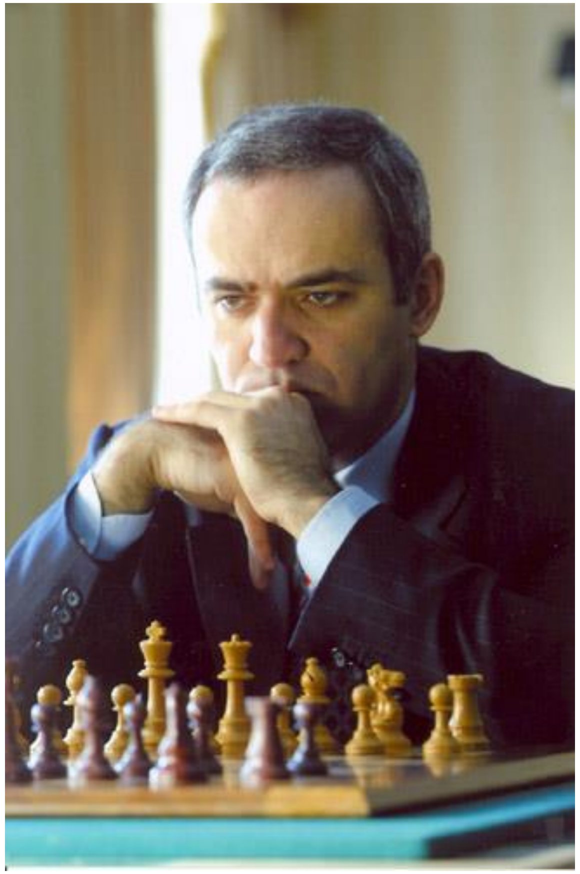
- Cluster computer systems can be used for any purpose, from business needs such as web-service support, FTP client, to computer intensive scientific calculations
- Three types of clusters:
    - Load-Balancing Clusters
        - the cluster checks which machine is the least busy and then sends the request to that machine.
    - Fail-Over Clusters
        - as soon as a service on one machine breaks down the other machines try to take over
    - High-Availability Clusters
        - the cluster checks which machine is the least busy and then sends the request to that machine.

First Computer Cluster!

?


digital VAX 11/780
CONSOLE CONTROL
FLOATING POINT ARITHMETIC ACCELERATOR
MICROCODE SEQUENCER AND CONTROL STORAGE
CLOCK, INTERRUPTS & EXCEPTIONS
ARITHMETIC AND LOGIC DATA PATH
INSTRUCTION DATA PATH
CACHE (8KB AND TLB (128 ENTRIES)
BUS INTERFACE
UNIBUS ADAPTER
MAIN MEMORY 2 MEGABYTES (16 MODULES, 256KB EACH)
MASSBUS ADAPTER
CI CLUSTER INTERCONNECT ADAPTER
Central Processing Unit
I/O Bus
Memory
I/O Bus
Network & I/O Bus
Power Supplies

A VAX 11/780, c. 1977



Garry Kasparov

A bit of history clustering computing systems

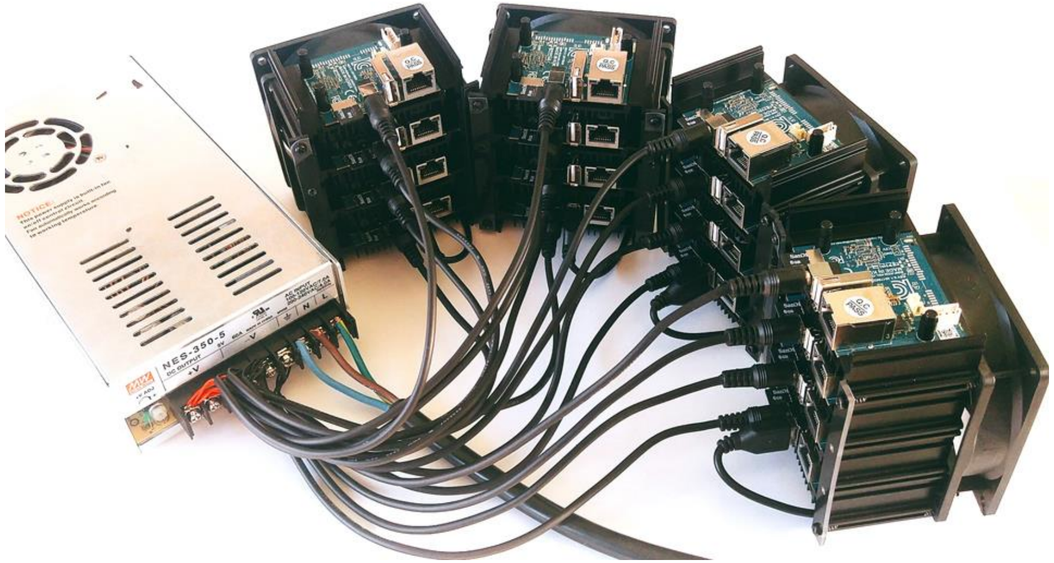
- The ARPANET project succeeded in creating in 1969 what was arguably the world's first commodity-network based computer cluster by linking four different computer centres.
  - Guess what ARPANET is now?
- 1977, The ARCnet and VAXcluster products not only supported parallel computing, but also shared file systems and peripheral devices
- 1997, a bunch of RS/6000s were clustered together to provide the needed scalability for the famous "Deep Blue" chess championship.
  - A match between Gary Kasparov and Deep Blue, Deep Blue won!

For more info, read wiki. https://en.wikipedia.org/wiki/History_of_computer_clusters

Some Clustering Systems

CRAY
TRINITY
Alliance for Computing at Extreme Scale
NNSA
Los Alamos
Sandia National Laboratories

Cray



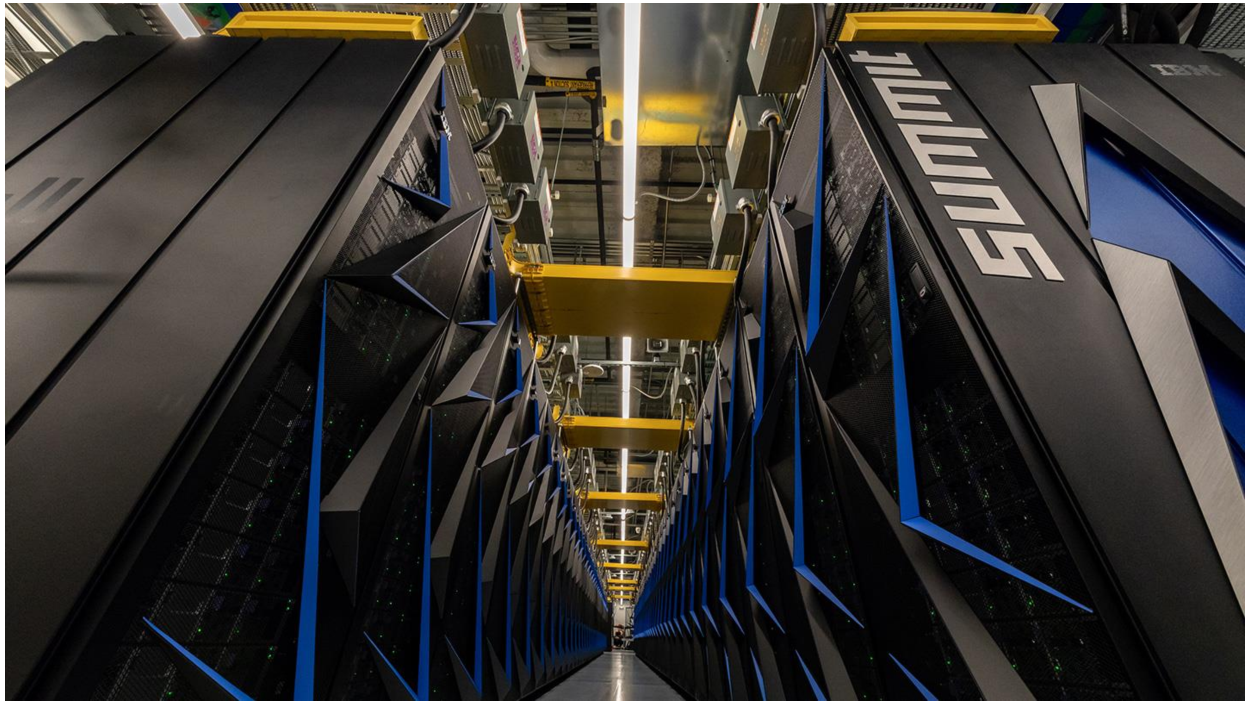
Raspberry Pi Cluster*

Network Attached Units

*https://www.hardkernel.com/shop/odroid-mc1-my-cluster-one-with-32-cpu-cores-and-8gb-dram/

The most advanced computer cluster?

- Summit, an IBM AC922 system, links
  - More than 27,000 NVIDIA Volta GPUs
  - More than 9,000 IBM Power9CPUs
  - A 200-petaflop machine
  - (By Space) Summit could fill two tennis courts.

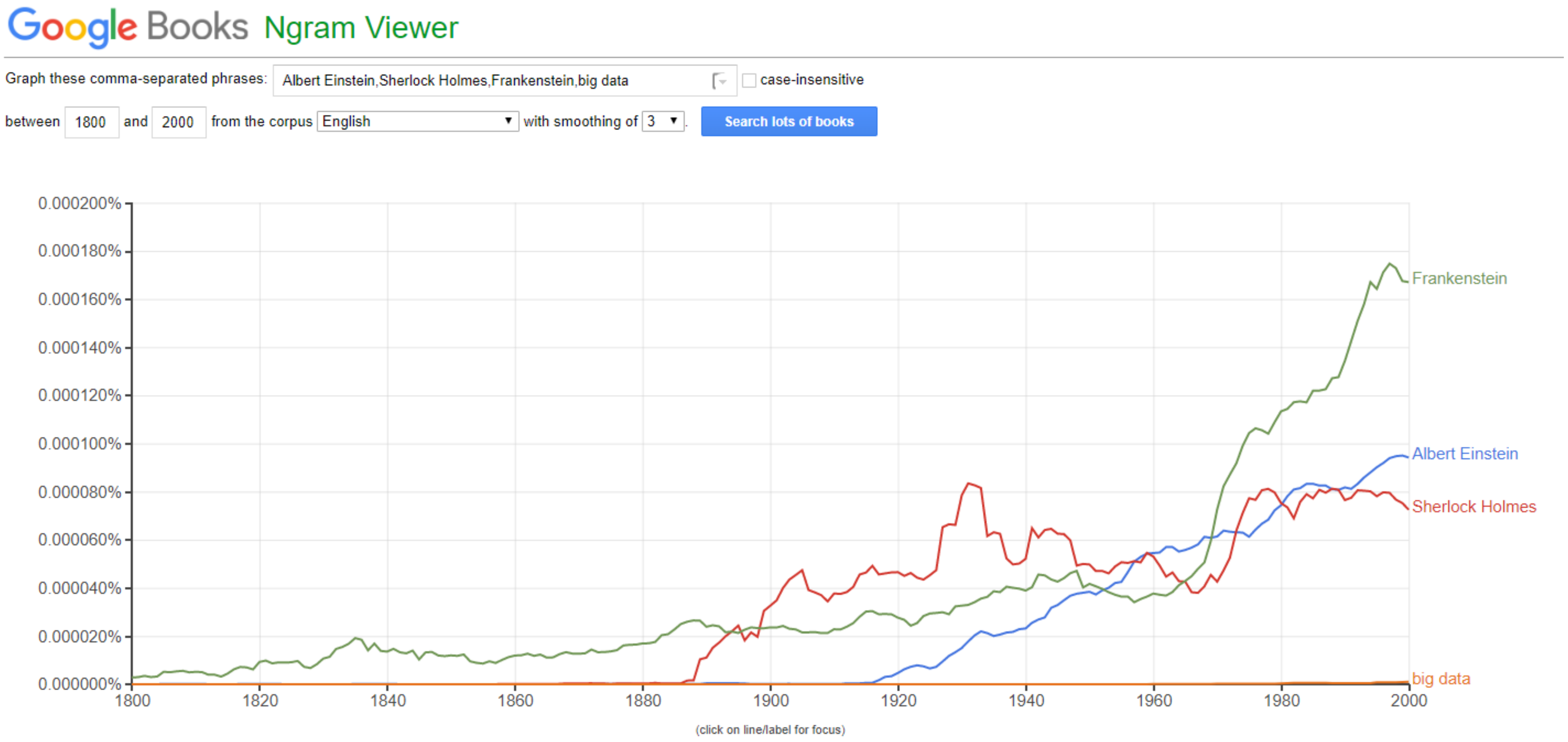


https://www.olcf.ornl.gov/summit/

Where to go?

- Quantum computing vs Summit, https://www.technologyreview.com/f/614416/google-researchers-have-reportedly-achieved-quantum-supremacy/
- Clustering 101, https://www.ibm.com/developerworks/aix/tutorials/clustering/clustering.html

Example usage (Google NGrams)



Google Books Ngram Viewer

Graph these comma-separated phrases: Albert Einstein,Sherlock Holmes,Frankenstein,big data □ case-insensitive

between 1800 and 2000 from the corpus English ▾ with smoothing of 3 ▾.

https://books.google.com/ngrams

Algorithms

- Map Reduce Algorithm

Thanks

Questions?