UE Large Scale Processing @ ESIGELEC
2019/2020

03 – Big Data & the data explosion

Abdel Dadouche
DJZ Consulting

adadouche@hotmail.com
@adadouche

The concept behind “Big Data”

A Quick Definition

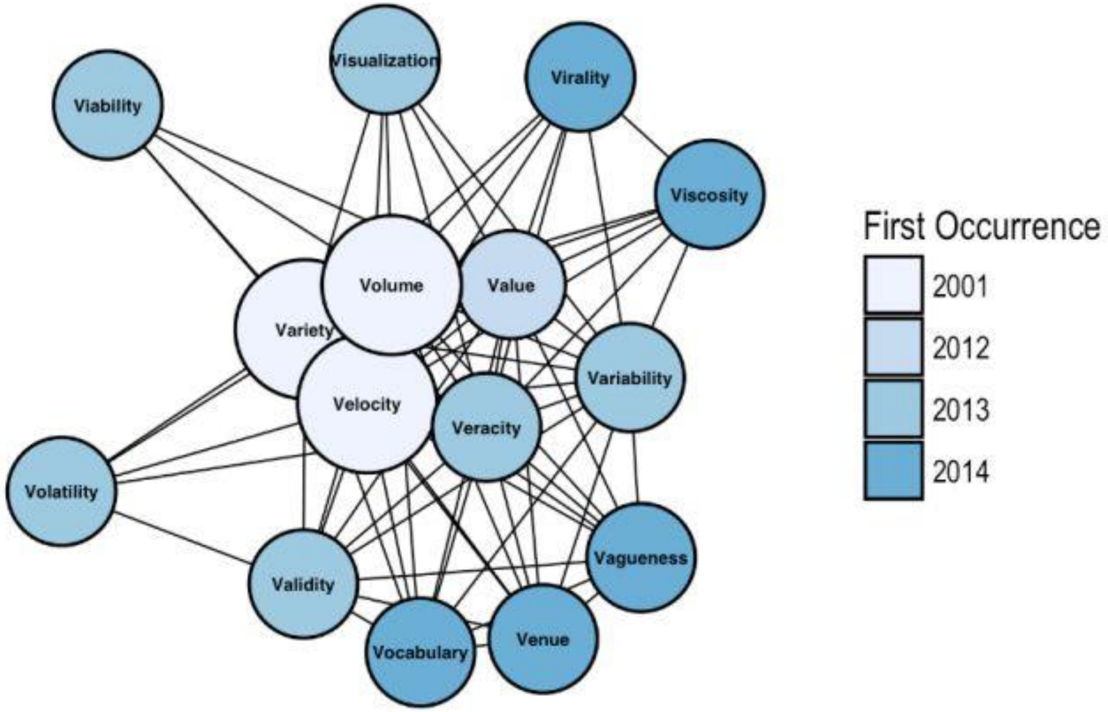
- "Big data" usually relates to data sets with sizes beyond the ability of commonly used software tools to capture, curate, manage, and process data within a tolerable elapsed time.
- These datasets can be unstructured, semi-structured or structured
- The "size" of "Big data" is constantly moving
- The term has been in use since the 1990s

Source: https://en.wikipedia.org/wiki/Big_data

A History of V’s

- The initial 3 V’s of Big Data
  - **Volume**: The quantity of generated and stored data
  - **Velocity**: The speed at which the data is generated and processed
  - **Variety** : The type and nature of the data
- Then 7 V’s:
  - **Value**
  - **Veracity**
  - **Variability**
  - **Visualization**
- And some added 3 more:
  - **Validity**
  - **Vulnerability**
  - **Volatility**

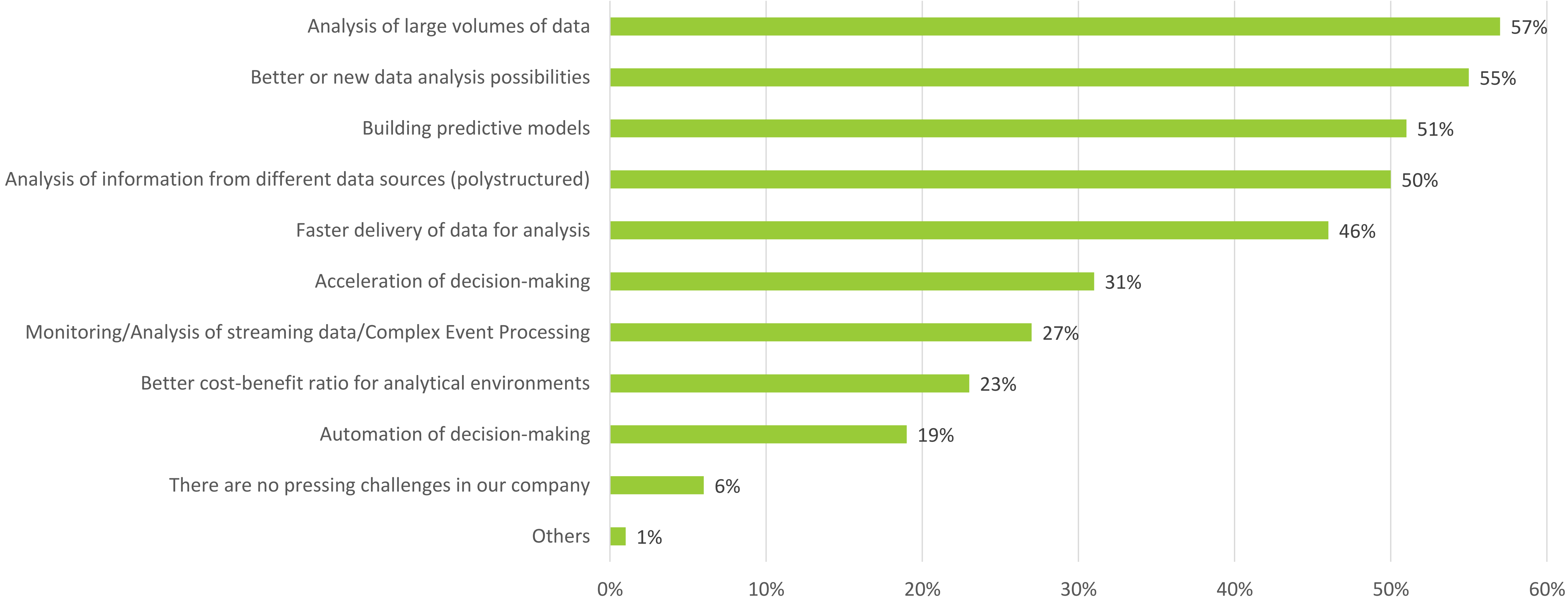
Now, up to the 42 V’s of Big Data & Data Science by Tom Shafer, Elder Research, Inc.



Source: https://en.wikipedia.org/wiki/Big_data
https://www.kdnuggets.com/2017/04/42-vs-big-data-data-science.html

“Big Data”: Challenges & Benefits

Big Data Challenges



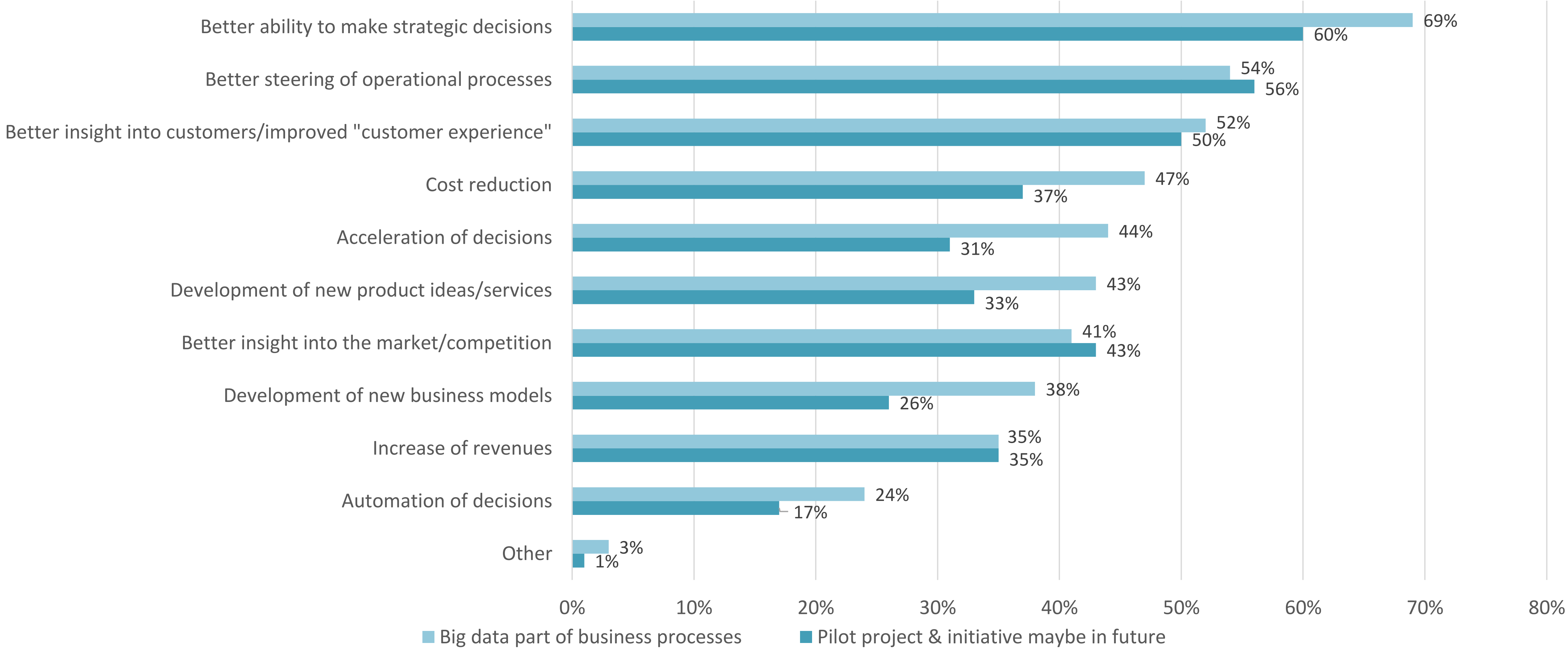
Source: http://barc-research.com/research/big-data-use-cases-2015/

With great power comes great responsibility!

- You must setup a “Data Governance”:
  - Don’t let everyone access your data (GDPR)
  - Don’t let everyone put data (TCO)
  - Make sure ingested data follow predefined « guardrails »
  - Keep “one version of the truth” and enforce it across the organization with Master Data Management

➔Authentication, Authorization & Integrity!

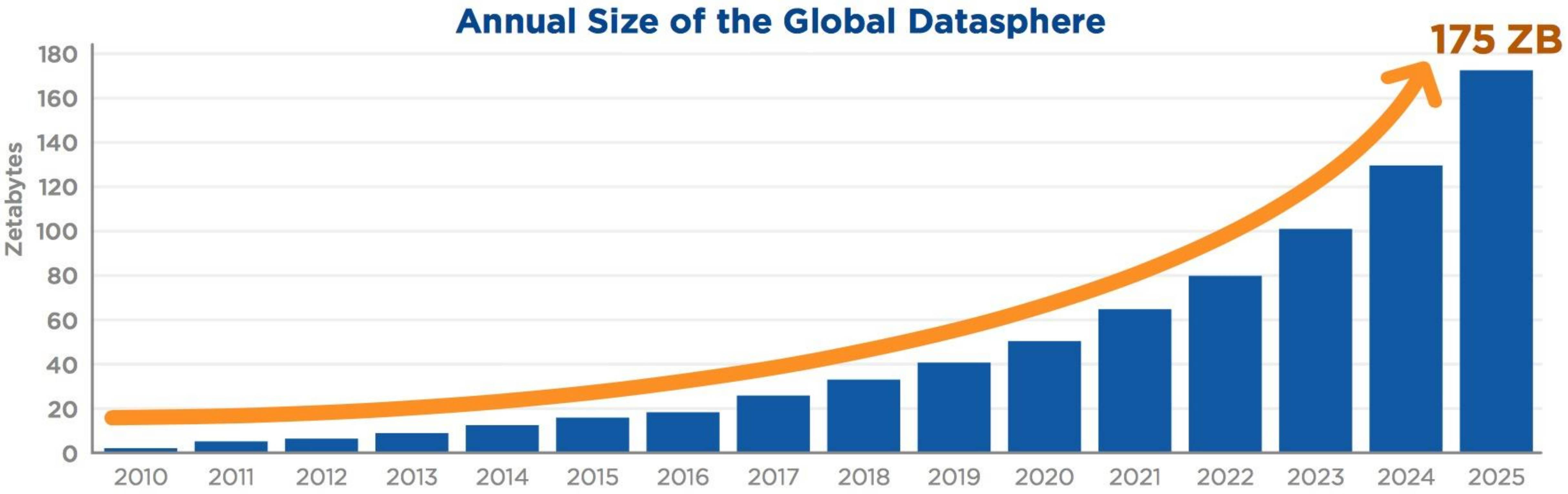
Big Data Benefits



Source: http://barc-research.com/research/big-data-use-cases-2015/

“Big Data”: a remedy to data explosion ?

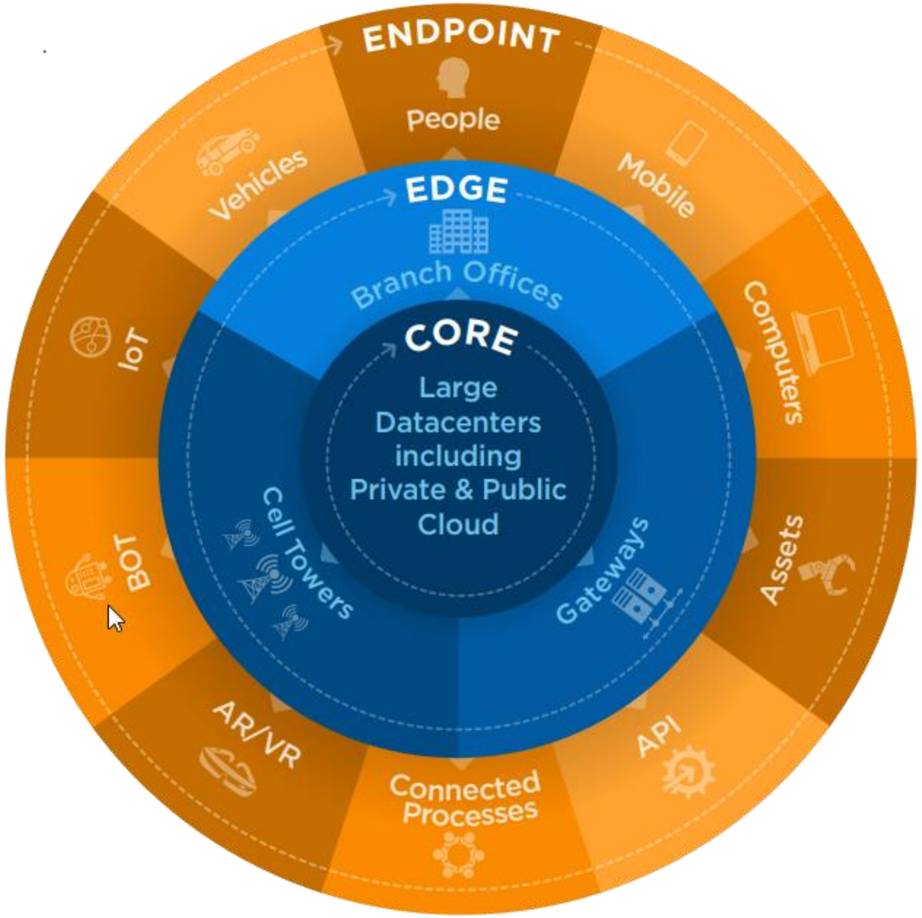
The Global Datasphere



Source: Data Age 2025, sponsored by Seagate with data from IDC Global Datasphere, Nov 2018
https://www.seagate.com/files/www-content/our-story/trends/files/idc-seagate-dataage-whitepaper.pdf

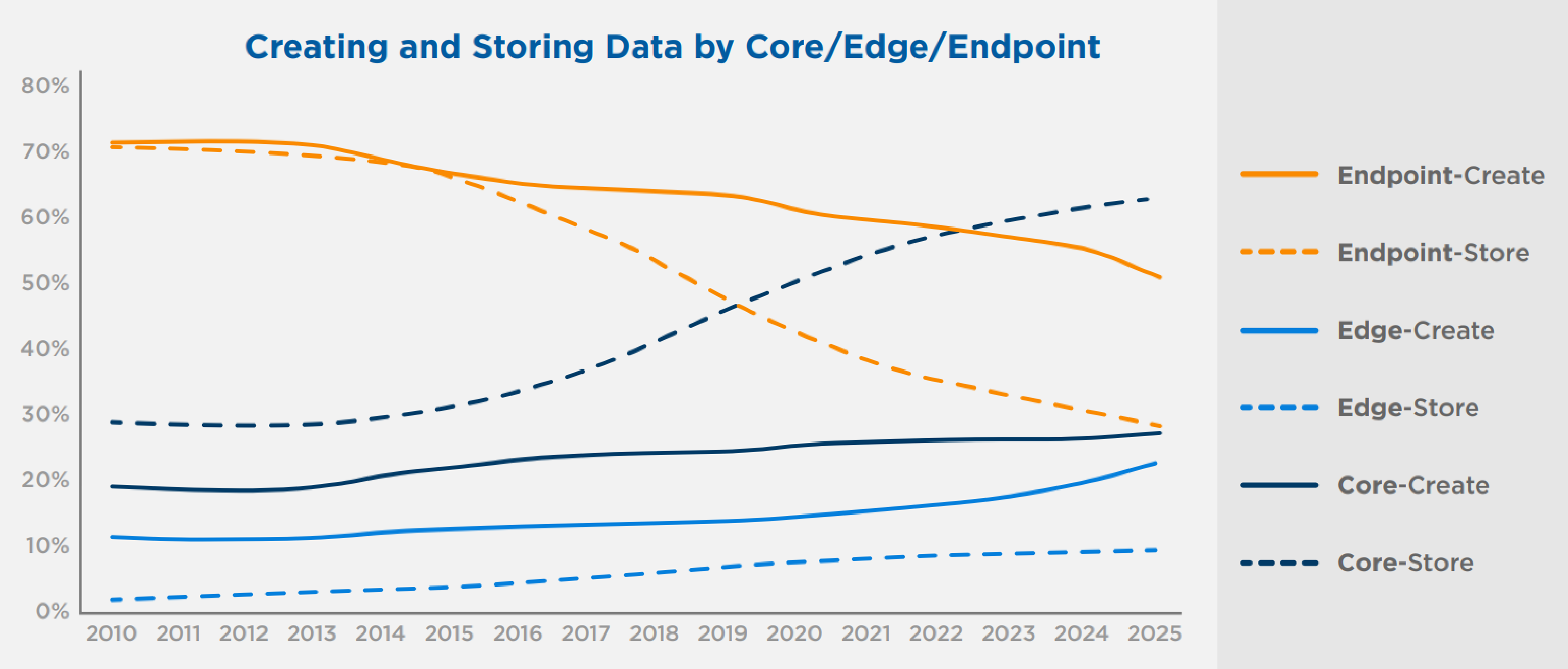
How is these data propagated?

- **Endpoint**: All devices at the edge of the network, including PCs, phones, industrial sensors, connected cars, and wearables
- **Edge**: Computing datacenter not in the core datacenters
- **Core**: Computing datacenter (private, public or hybrid)



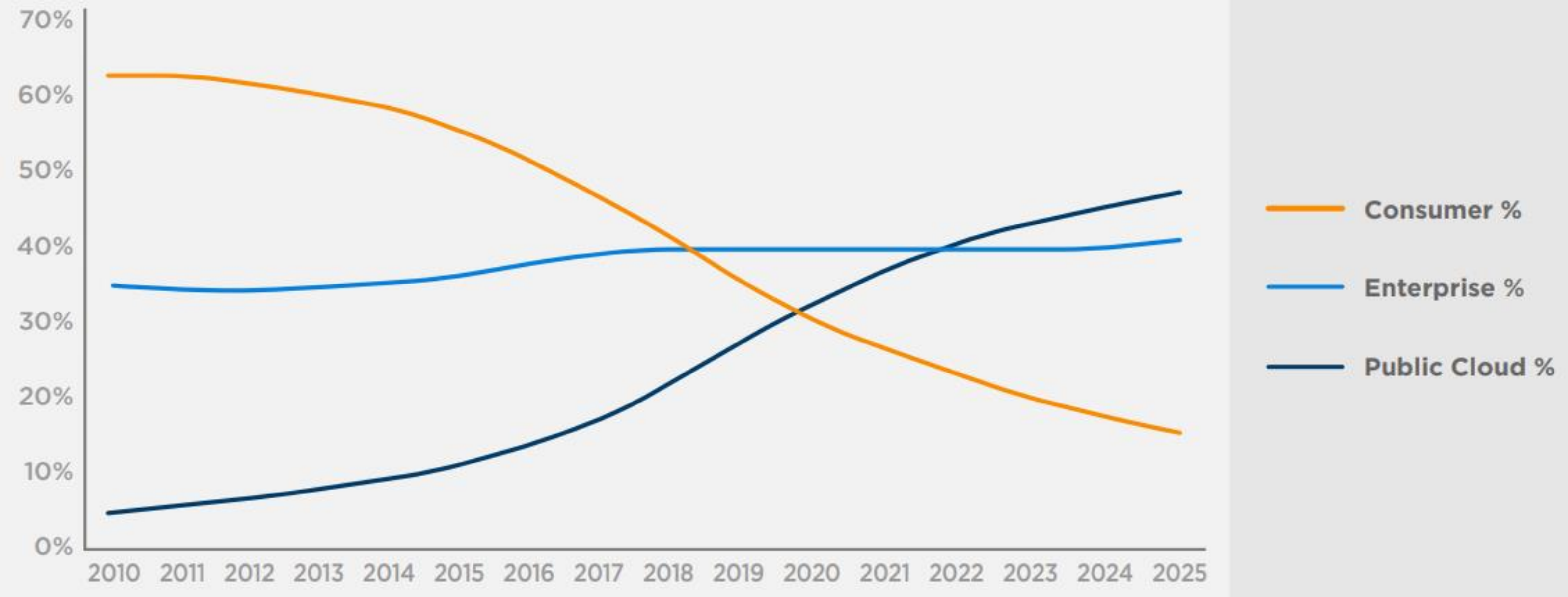
Source: Data Age 2025, sponsored by Seagate with data from IDC Global Datasphere, Nov 2018
https://www.seagate.com/files/www-content/our-story/trends/files/idc-seagate-dataage-whitepaper.pdf

Who is creating and storing this data?



Source: Data Age 2025, sponsored by Seagate with data from IDC Global Datasphere, Nov 2018
https://www.seagate.com/files/www-content/our-story/trends/files/idc-seagate-dataage-whitepaper.pdf

Where is the data stored?



Source: Data Age 2025, sponsored by Seagate with data from IDC Global Datasphere, Nov 2018
https://www.seagate.com/files/www-content/our-story/trends/files/idc-seagate-dataage-whitepaper.pdf

The Big Data Answer to : How do I get this working?

Distributed Computing

- How do I leverage each machine to process my data as if it was one machine?
- How can I scale up or down the computing capabilities without downtime?
- How can I make sure the result will always be correct?

Distributed Data Storage

- How do I partition the data across multiple heterogeneous machines?
- How do I access specific portions of my data?
- What happens if one or multiple machines crash?
- How do I make sure my data is safe?

The rise of "Distributed Computing"

Beowulf Clusters (1998)

- A cluster of identical, commodity-grade computers networked into a small local area network with identical libraries and programs installed and running a Unix-like operating system
- The cluster "Server" node (master) assigned tasks (with data) to the cluster "Client" nodes (slaves) which have no other purpose than server the "Server" node
- The result is a high-performance parallel computing cluster from inexpensive personal computer hardware

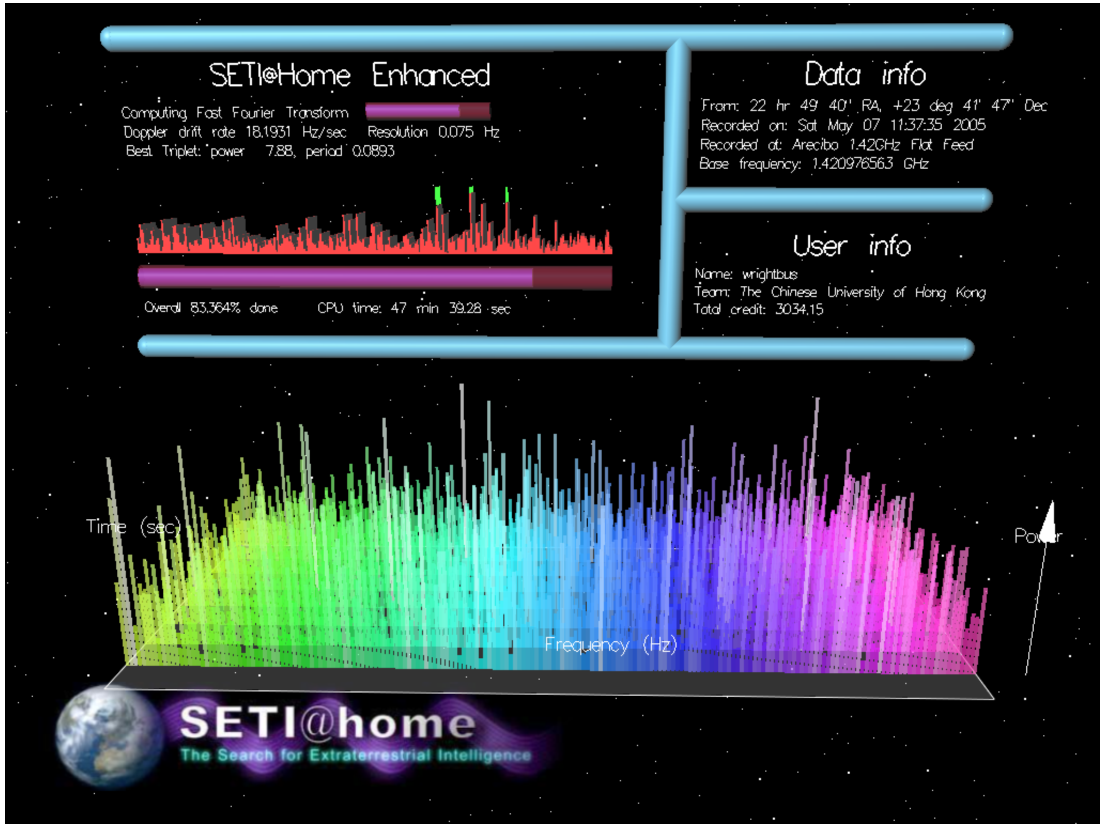
Source: https://en.wikipedia.org/wiki/Beowulf_cluster

SETI@home (1999)

- An Internet-based public volunteer computing project to search for possible evidence of radio transmissions from extraterrestrial intelligence

- Using distributed computing, SETI@home sends millions of chunks of data to be analyzed off-site by home computers, and then have those computers report the results.

- It uses the BOINC software platform from Berkeley SETI Research Center and hosted by the Space Sciences Laboratory, at the University of California, Berkeley.



SETI@Home Enhanced

Computing Fast Fourier Transform
Doppler drift rate 18.1931 Hz/sec Resolution 0.075 Hz
Best Triplet: power 7.88, period 0.0893

Overall 83.364% done CPU time: 47 min 39.28 sec

Data info

From: 22 hr 49 40' RA, +23 deg 41' 47' Dec
Recorded on: Sat May 07 11:37:35 2005
Recorded at: Arecibo 1.42GHz Flat Feed
Base frequency: 1.420976563 GHz

User info

Name: wrightbus
Team: The Chinese University of Hong Kong
Total credit: 3034.15

BOINC : Berkeley Open Infrastructure for Network Computing

Source: https://en.wikipedia.org/wiki/SETI@home

IBM Blue Gene (1999)

- An IBM project, started in 1999, aimed at designing supercomputers that can reach operating speeds in the peta-FLOPS
- In 2004, first commercial version of Blue Gene/L (for light) was released as a 16-rack system, with 1,024 compute nodes per rack and a compute capacity of 70.72 TFLOPS
- This is one of the first example of a commercial massively parallel computer

FLOPS : floating point operation per seconds

Source: https://en.wikipedia.org/wiki/IBM_Blue_Gene

Conclusion

- BOINC provided the first “large scale” distributed computing framework but with the amount of data to be processed and the regulation around it, this approach is not applicable to many use cases

- Super computers like « Blue Gene » are really expensive and over-sized for many use cases

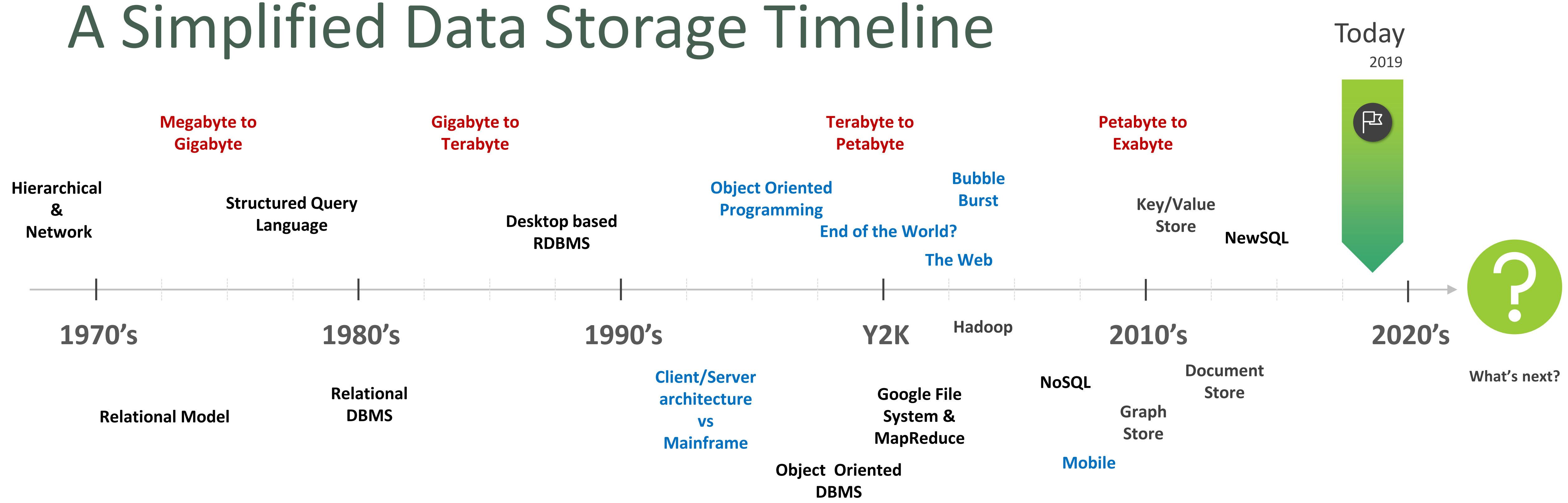
- And Beowulf clusters still requires a significant investment to acquire, setup and maintain

The evolution of "Data Storage"

So why not “just” a Relational Database for Big Data?

- Traditional Databases addresses mostly structured data and process mostly locally stored data (file or memory)
- Despite process parallelization and query optimization, it is almost impossible to aggregate, join, merge or process such large volumes in a reasonable amount of time and cost using relational databases
- Scaling up or out a relational database is technically complex & expensive

A Simplified Data Storage Timeline



Today

2019

Megabyte to Gigabyte

Gigabyte to Terabyte

Terabyte to Petabyte

Petabyte to Exabyte

Hierarchical & Network

Structured Query Language

Desktop based RDBMS

Object Oriented Programming

End of the World?

Bubble Burst

The Web

Key/Value Store

NewSQL

1970's

1980's

1990's

Y2K

Hadoop

2010's

2020's

Relational Model

Relational DBMS

Client/Server architecture vs Mainframe

Object Oriented DBMS

Google File System & MapReduce

NoSQL

Mobile

Graph Store

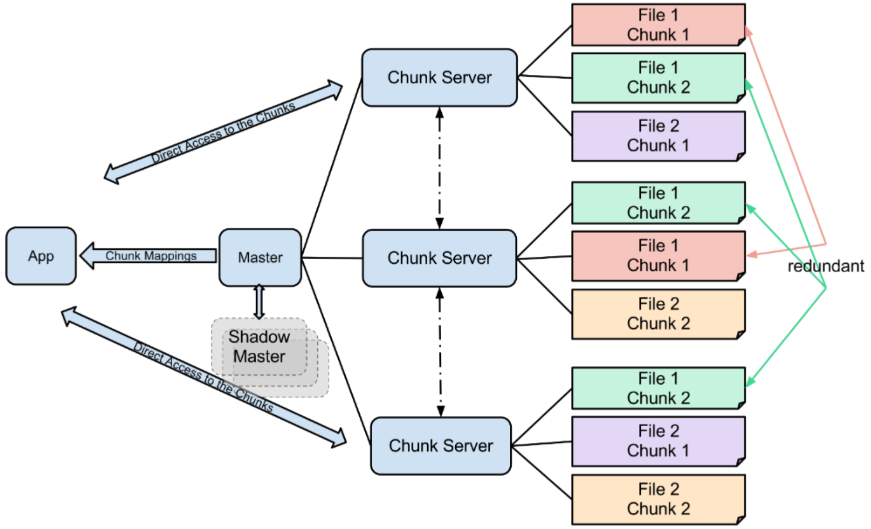
Document Store

What's next?

The key milestones in distributed computing & data storage

Google File System (2003)

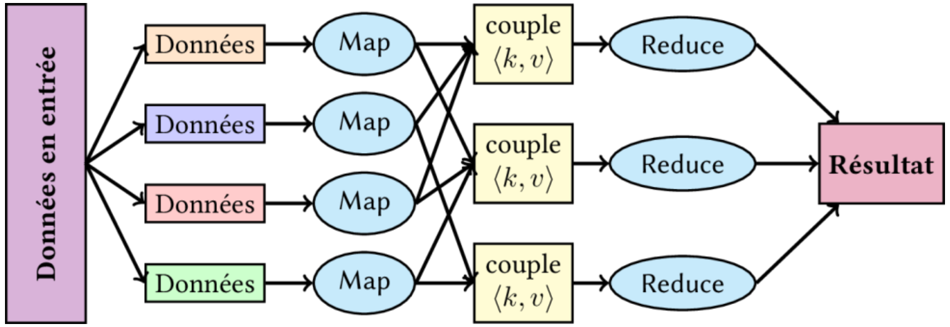
- A proprietary distributed file system developed by Google in
- It guarantee efficient and reliable access to data using large clusters of commodity hardware



Source: https://en.wikipedia.org/wiki/Google_File_System

MapReduce (2004)

- A programming model for processing large data sets with a parallel & distributed algorithm on a cluster
- Composed of :
  - a **map** method performing filtering and sorting
  - a **reduce** method performing a summary operation (aggregation)



Données en entrée
Données
Données
Données
Données
Map
Map
Map
Map
couple $\langle k, v \rangle$
couple $\langle k, v \rangle$
couple $\langle k, v \rangle$
Reduce
Reduce
Reduce
Résultat

Source: https://en.wikipedia.org/wiki/MapReduce

Summary

- The concept of Big Data is not new (despite the association of a name)
- A History of V’s (I stopped at 5 myself)
- The “Data” explosion is having a huge on innovation and adoption
- Big Data help solves many organizations challenges
- Thanks Google for the GFS & MapReduce whitepapers!