CIT650
Introduction to Big Data

Principles of Big Data

Today's Agenda

- **Big Data Phenomena**
- **Big Data 1.0 Systems**
- **Big Data 2.0 Systems**

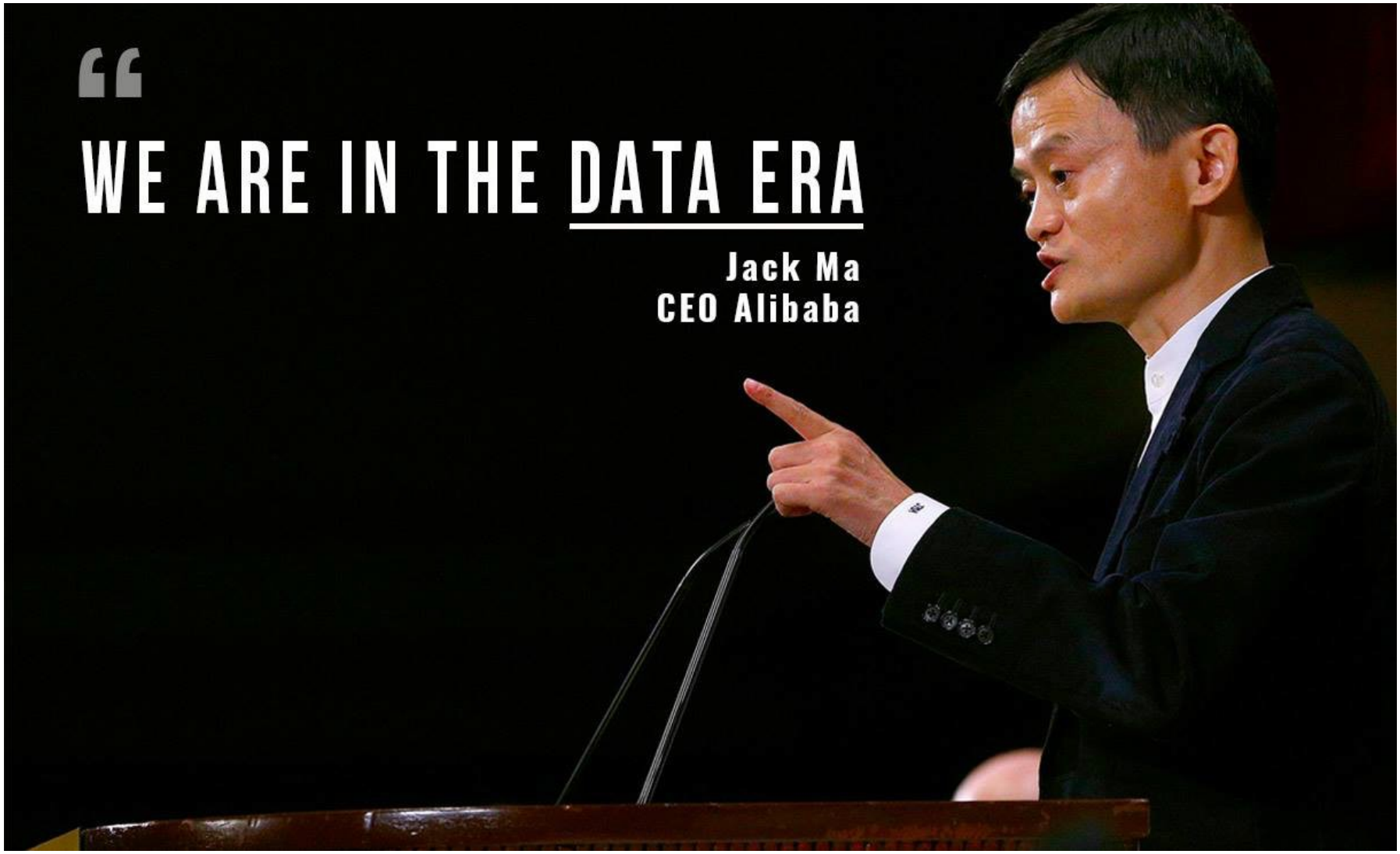
Part I

Big Data Phenomena

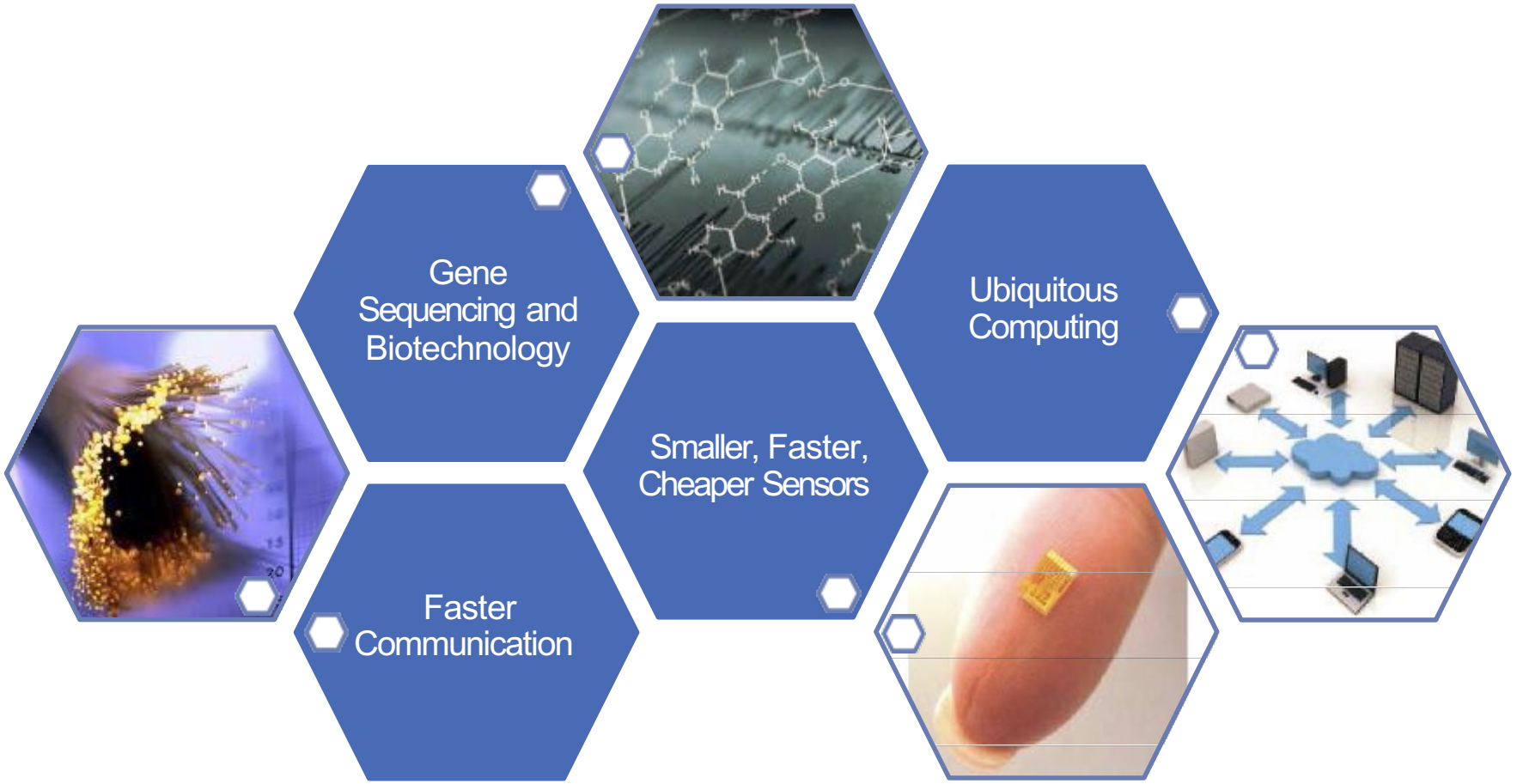
Big Data

- Data is key resource in the modern world.
- According to IBM, we are currently creating 2.5 quintillion bytes of data everyday.
- IDC predicts that the world wide volume of data will reach 40 zettabytes by 2020.
- The radical expansion and integration of computation, networking, digital devices and data storage has provided a robust platform for the explosion in big data.

Big Data



On the Verge of A Disruptive Century: Breakthroughs



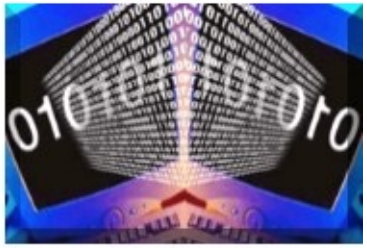
Big Data Applications are Everywhere

Smarter Healthcare

Multi-channel sales

Finance

Log Analysis



Homeland Security

Traffic Control

Telecom

Search Quality

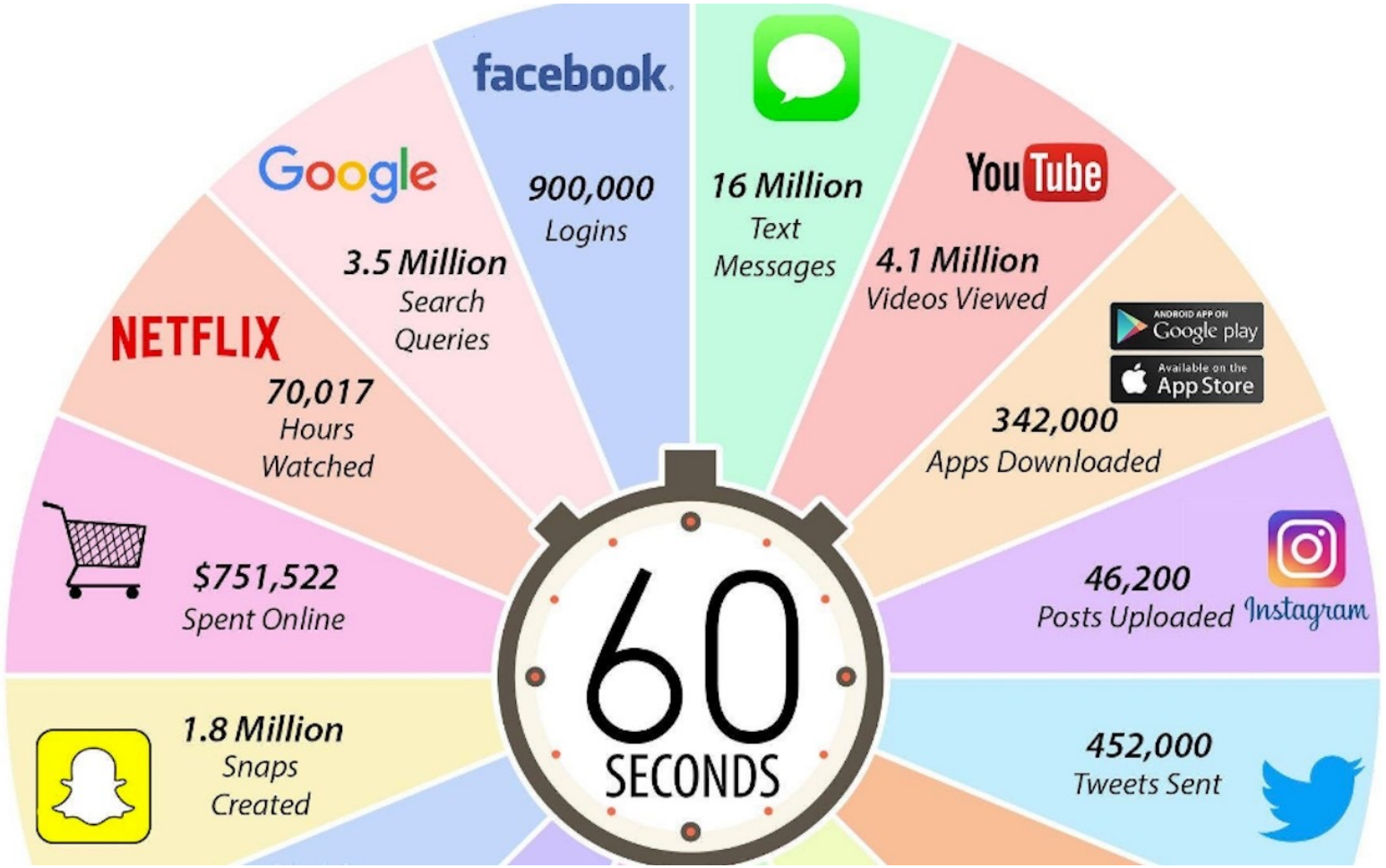
Manufacturing

Trading Analytics

Fraud and Risk

Retail: Churn, NBO

Big Data: What Happens in the Internet in a Minute?



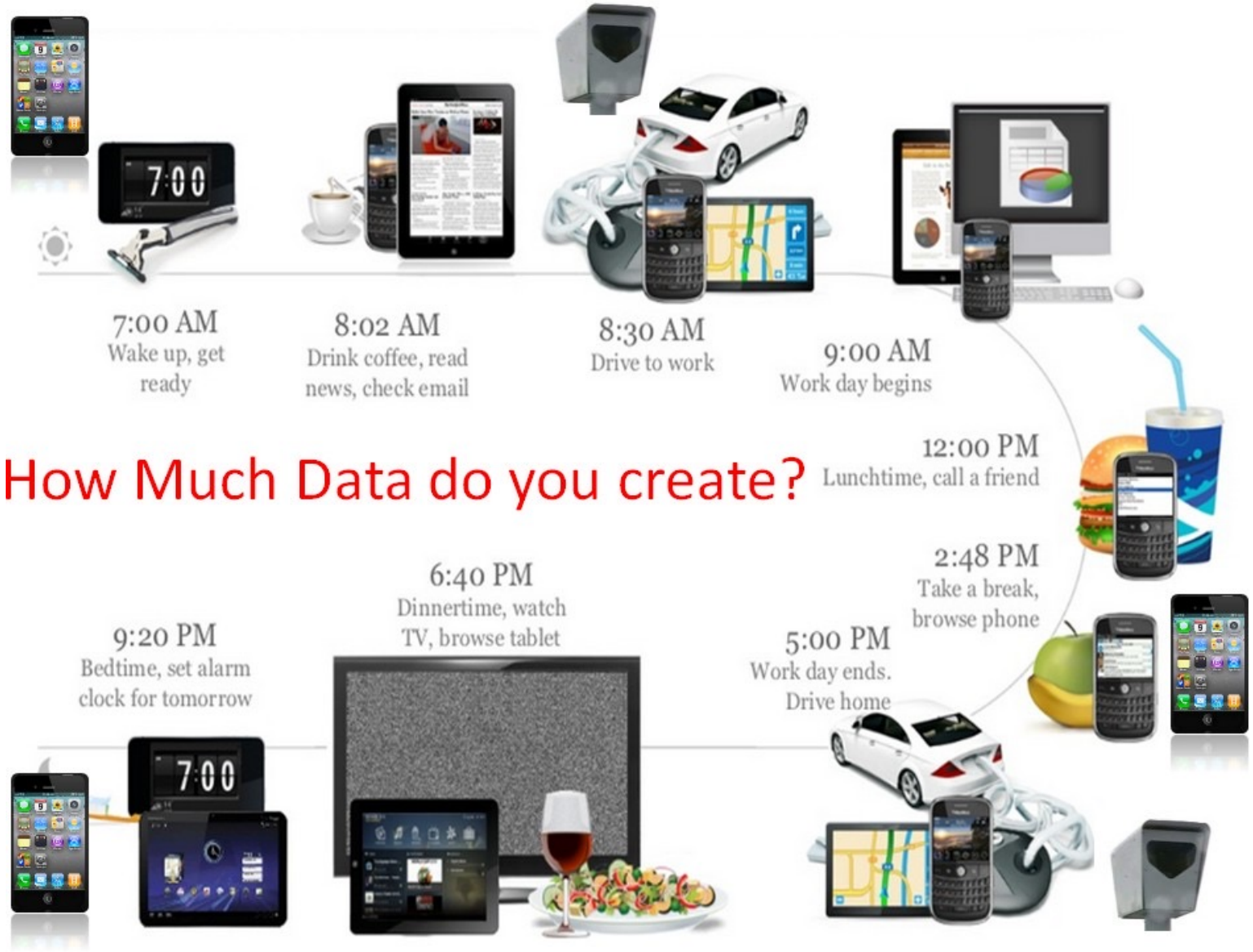
Data Generation and Consumption Model is Changing

Data Generation and Consumption Model is Changing

- **Old Model**: Few companies (producers) are generating data, all others are consuming data

- **New Model**: All of us are generating data, and all of us are consuming data

Big Data


7:00
7:00 AM
Wake up, get ready
8:02 AM
Drink coffee, read news, check email
8:30 AM
Drive to work
9:00 AM
Work day begins
How Much Data do you create?
12:00 PM
Lunchtime, call a friend
2:48 PM
Take a break, browse phone
5:00 PM
Work day ends. Drive home
6:40 PM
Dinnertime, watch TV, browse tablet
9:20 PM
Bedtime, set alarm clock for tomorrow
7:00

Big Data

- Data generation and consumption is becoming a main part of people's daily life especially with the pervasive availability and usage of Internet technology and applications.

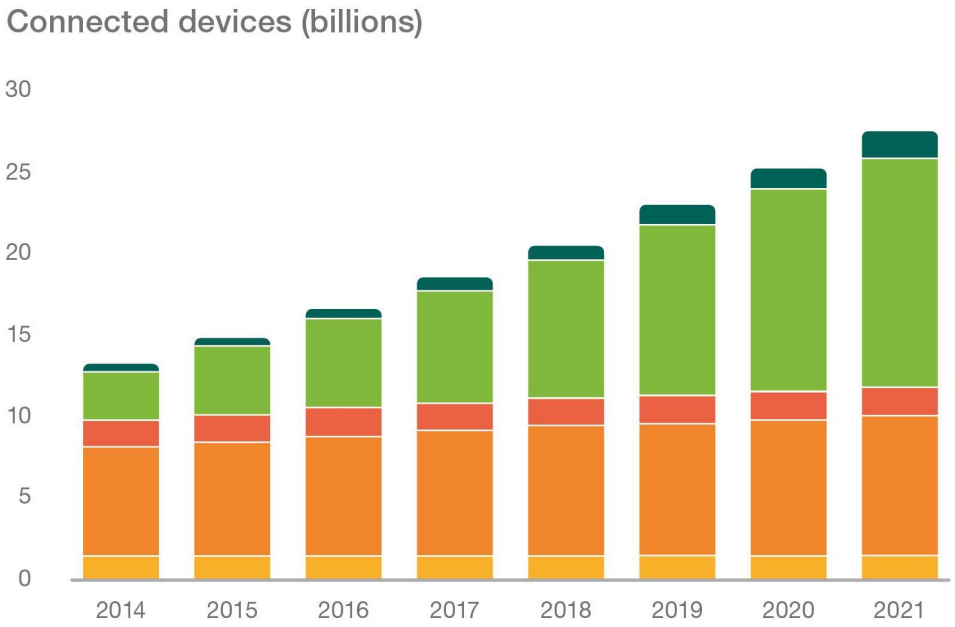
Your Smart Phone is now Very smart

Internet of Things (IoT)

- A network devices, connect directly with each other to capture, share and monitor vital data automatically through a SSL that connects a central command and control server in the cloud
- Enabling communication between devices, people & processes to exchange useful information & knowledge that create value for humans
- A global Network Infrastructure linking Physical & Virtual Objects
  - Infrastructure: Internet and Network developments
  - Specific object identification, sensor, and connection capability

Big Data: Internet of Things

Prediction of IoT Usage[1]



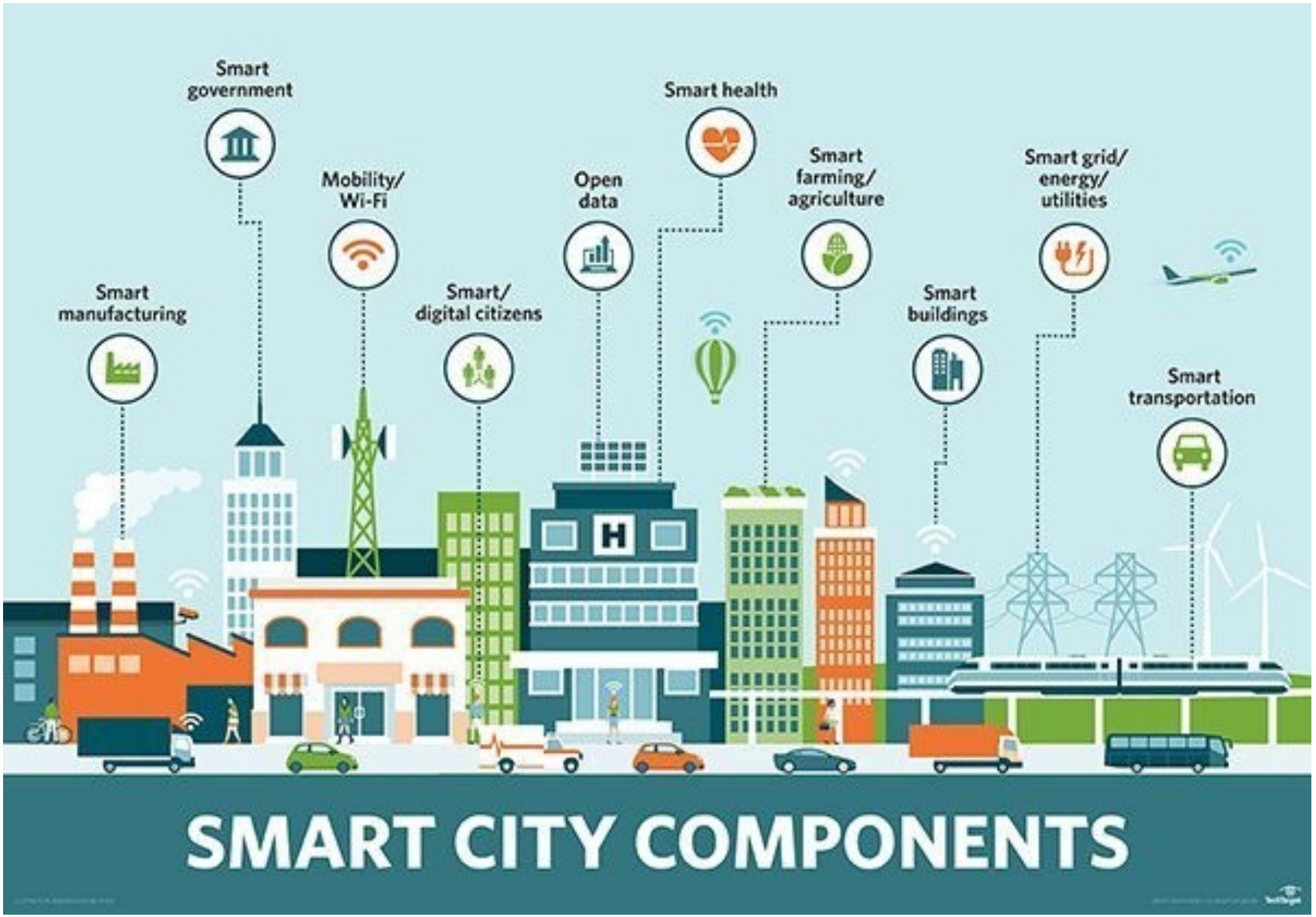
	15 billion	28 billion	CAGR 2015–2021
Cellular IoT	0.4	1.5	27%
Non-cellular IoT	4.2	14.2	22%
PC/laptop/tablet	1.7	1.8	1%
Mobile phones	7.1	8.6	3%
Fixed phones	1.3	1.4	0%
	2015	2021	

[1] https://www.ericsson.com/

Why IoT opportunity is growing now?

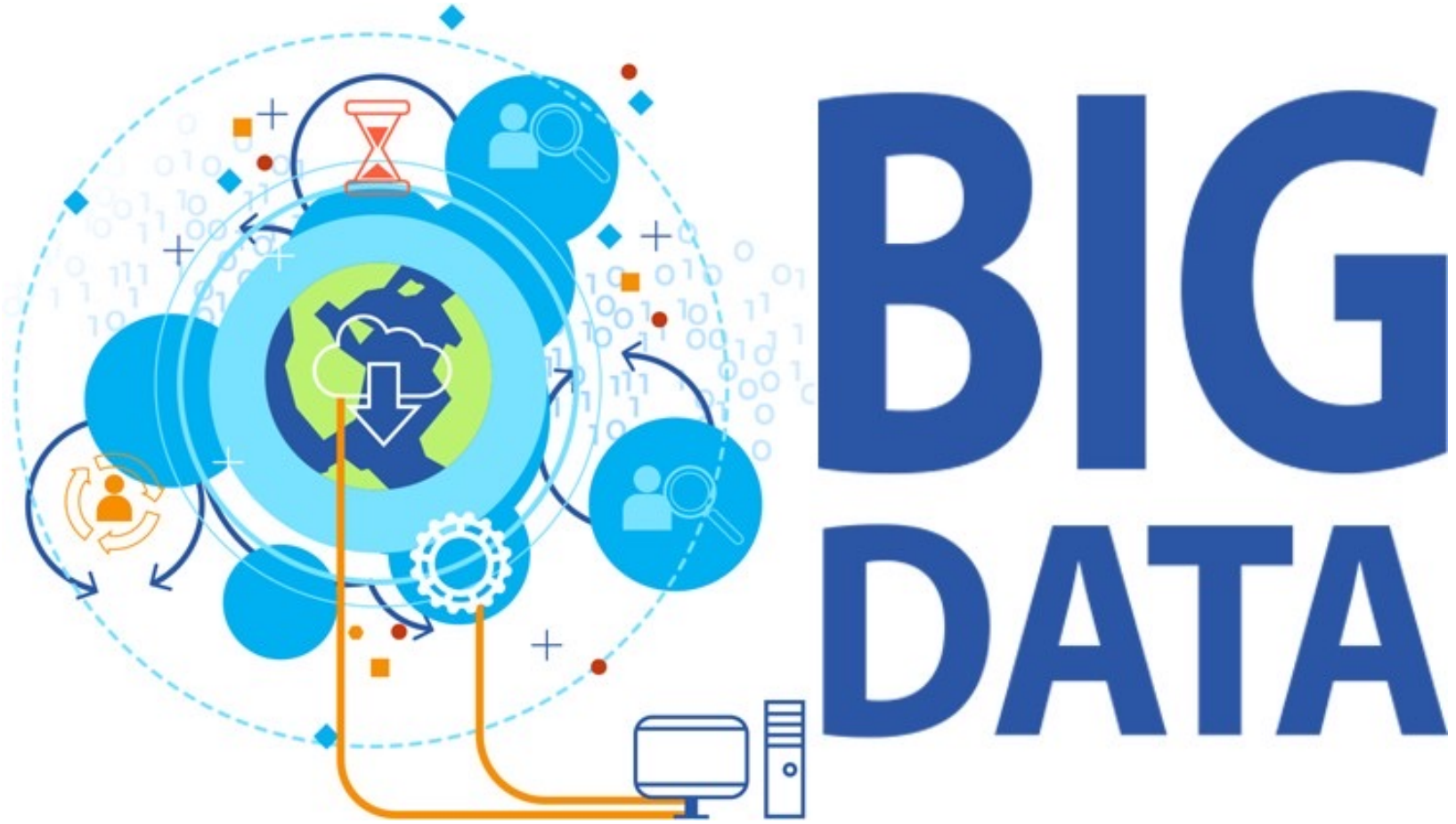
- **Affordable hardware**: Costs of actuators & sensors have been cut in half over last 10 years
- **Smaller, more powerful hardware**: Form factors of hardware have shrunk to millimeter or even nanometer levels
- **Ubiquitous & cheap mobility**: Cost for mobile devices, bandwidth and data processing has declined over last 10 years
- **Availability of supporting tools**: Big data tools & cloud based in-frastructure have become widely available

Smart X Phenomena



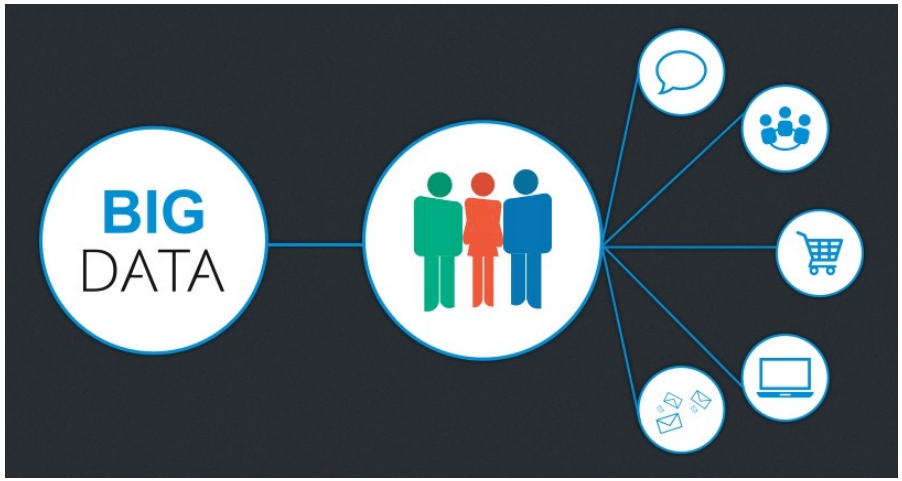
What it all produce?

Data … Data … Data



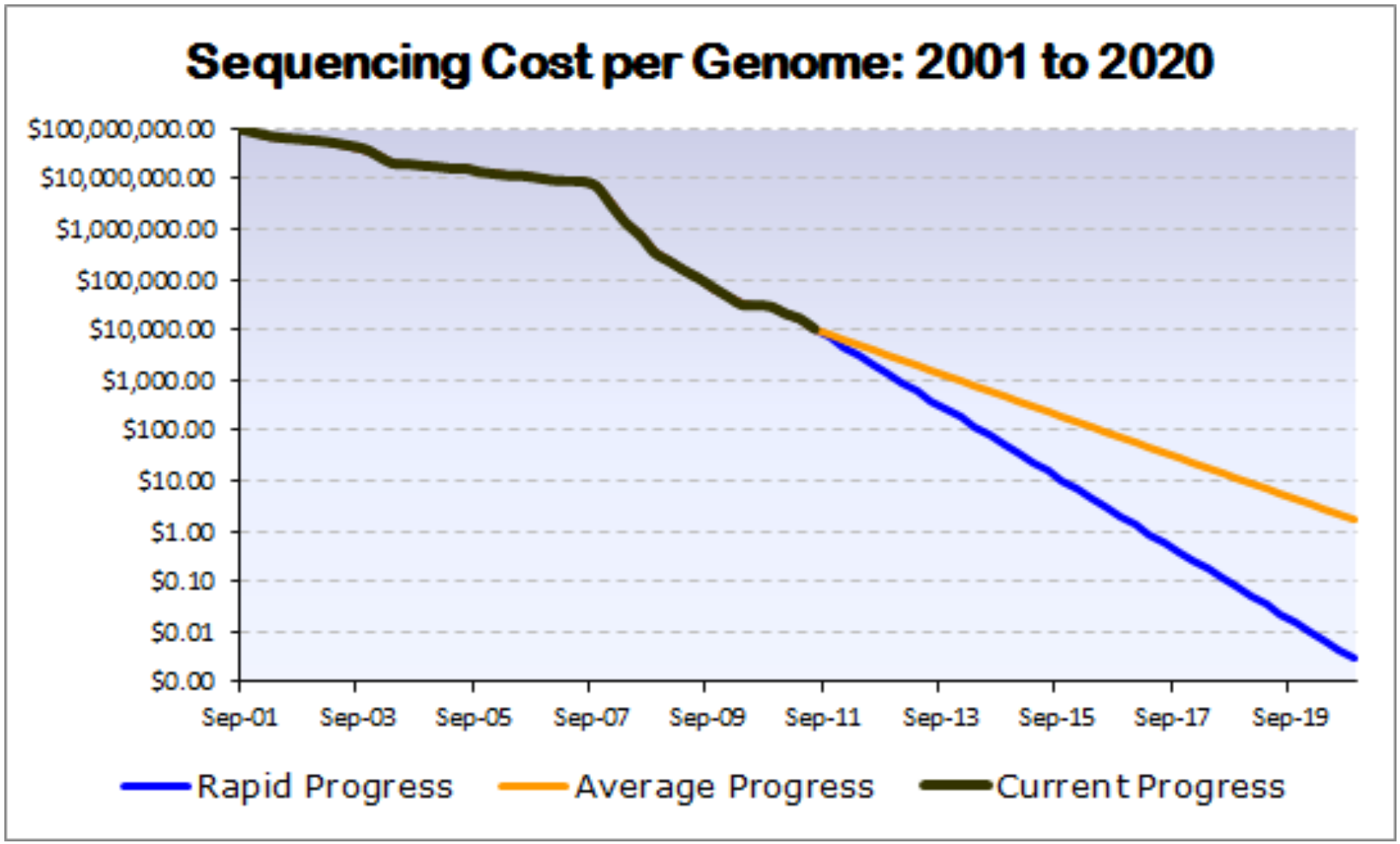
Big Data: Activity Data

- Simple activities like listening to music or reading a book are now generating data.
- Digital music players and eBooks collect data on our activities.
- Your smart phone collects data on how you use it and your web browser collects information on what you are searching for.
- Your credit card company collects data on where you shop and your shop collects data on what you buy.
- **It is hard to imagine any activity that does not generate data.**



Big Data

- The cost of sequencing one human genome has fallen from $100 million in 2001 to $1K in 2015



New Types of Data

Sentiment

understand how customers feel about your brand and products -right wow

Clickstream

Capture and analyze website visitors' data trails and optimize your website

Sensors

Discover patterns in data streaming automatically from remote sensors and machines

Geographic

Analyze location based data to manage operations where they occur

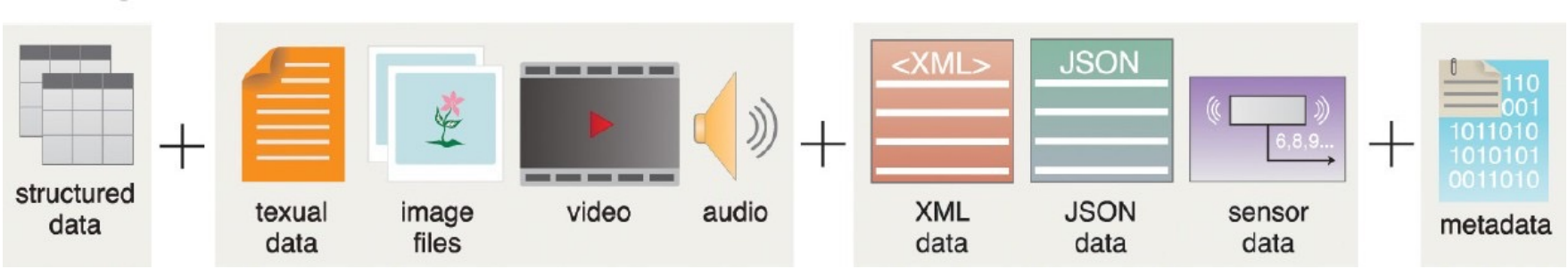
Server Logs

Research logs to diagnose process failures and prevent security breaches

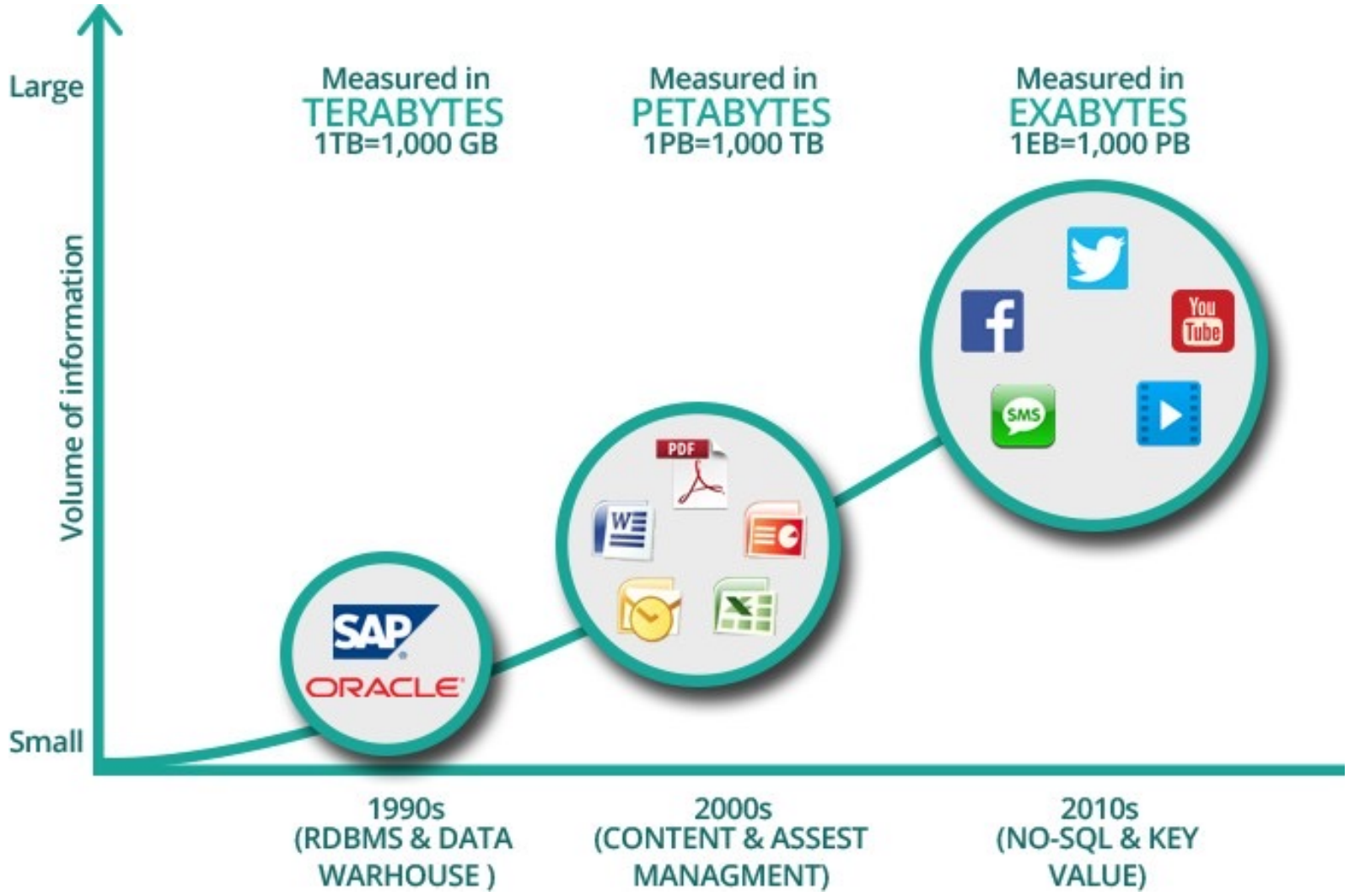
Unstructured

Understand patterns in files across millions of web pages, emails, and documents

New Types of Data

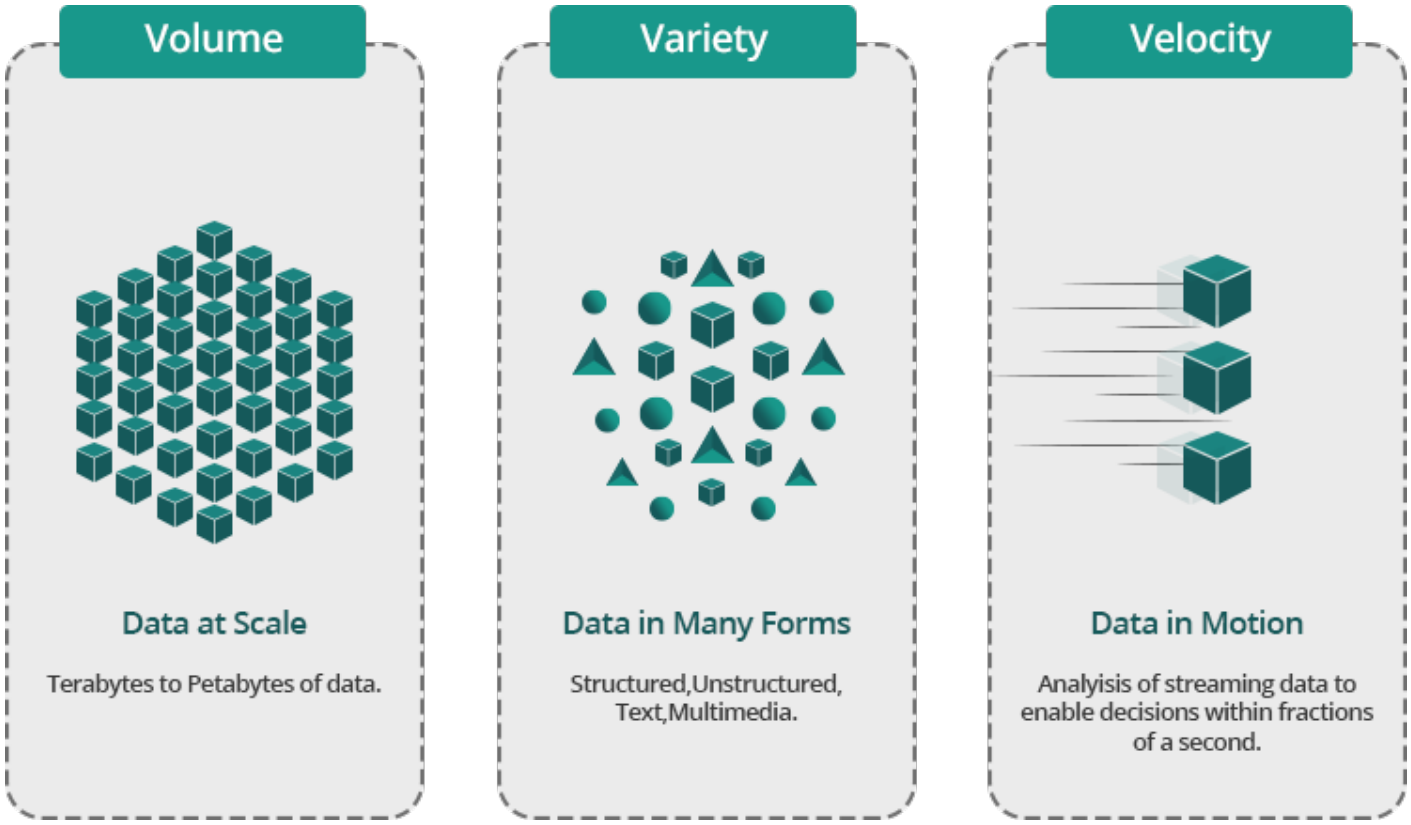


The Data Structure Evolution Over the Years

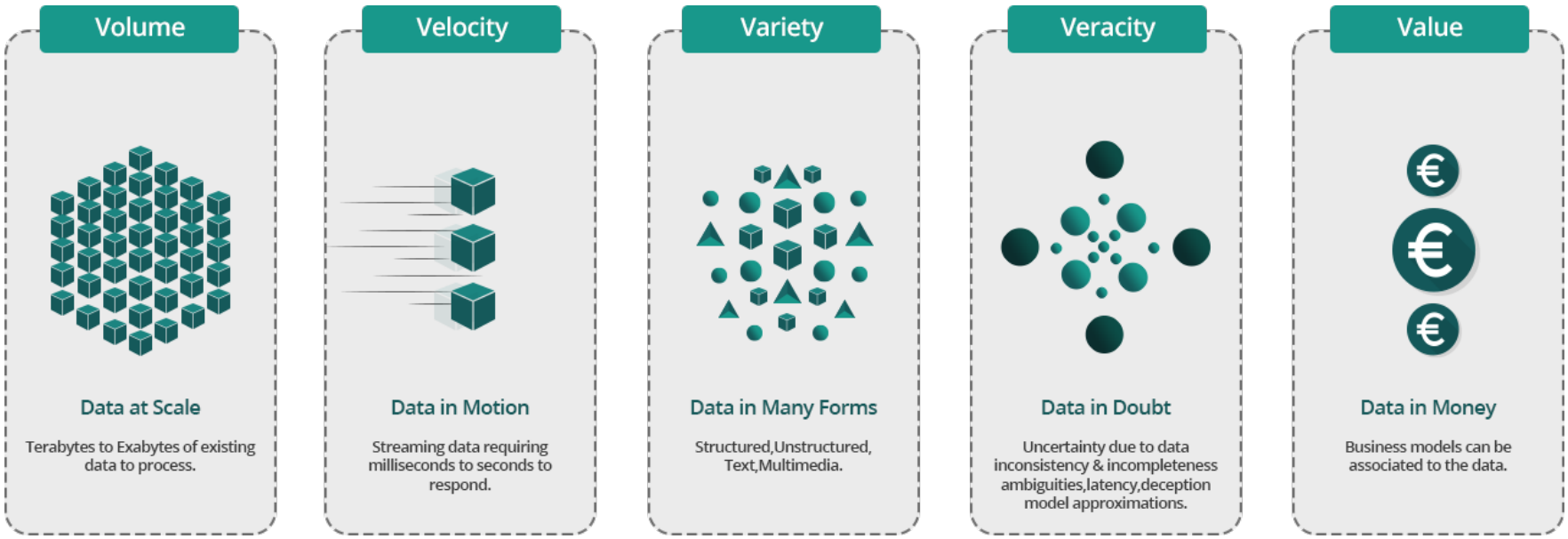


What Means **Big Data**?

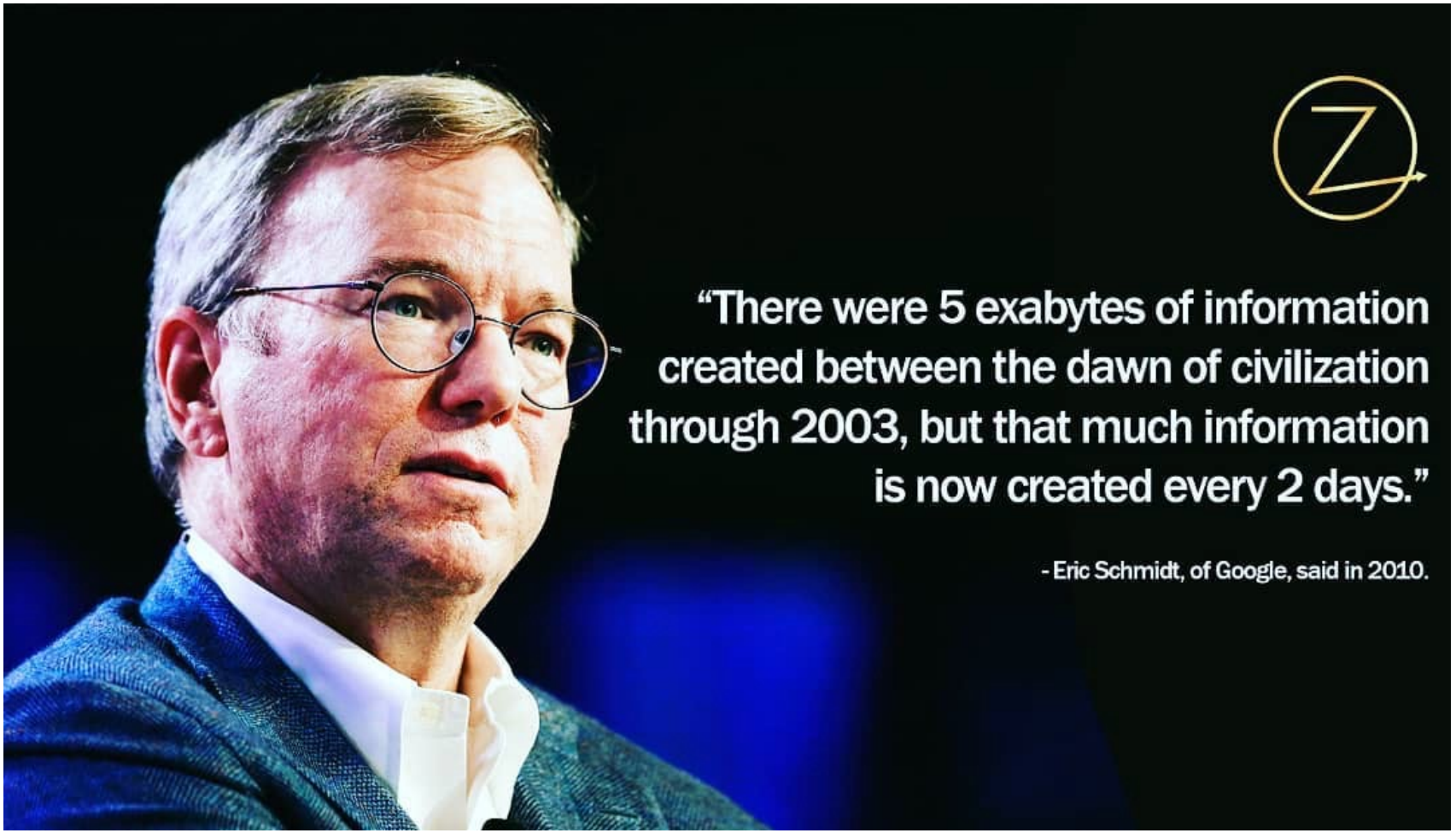
Big Data (3V)


Volume
Data at Scale
Terabytes to Petabytes of data.
Variety
Data in Many Forms
Structured,Unstructured,
Text,Multimedia.
Velocity
Data in Motion
Analyisis of streaming data to
enable decisions within fractions
of a second.

Big Data (5V)


Volume
Data at Scale
Terabytes to Exabytes of existing data to process.
Velocity
Data in Motion
Streaming data requiring milliseconds to seconds to respond.
Variety
Data in Many Forms
Structured,Unstructured, Text,Multimedia.
Veracity
Data in Doubt
Uncertainty due to data inconsistency & incompleteness ambiguities,latency,deception model approximations.
Value
Data in Money
Business models can be associated to the data.

Big Data


"There were 5 exabytes of information created between the dawn of civilization through 2003, but that much information is now created every 2 days."
- Eric Schmidt, of Google, said in 2010.

Big Data Definition

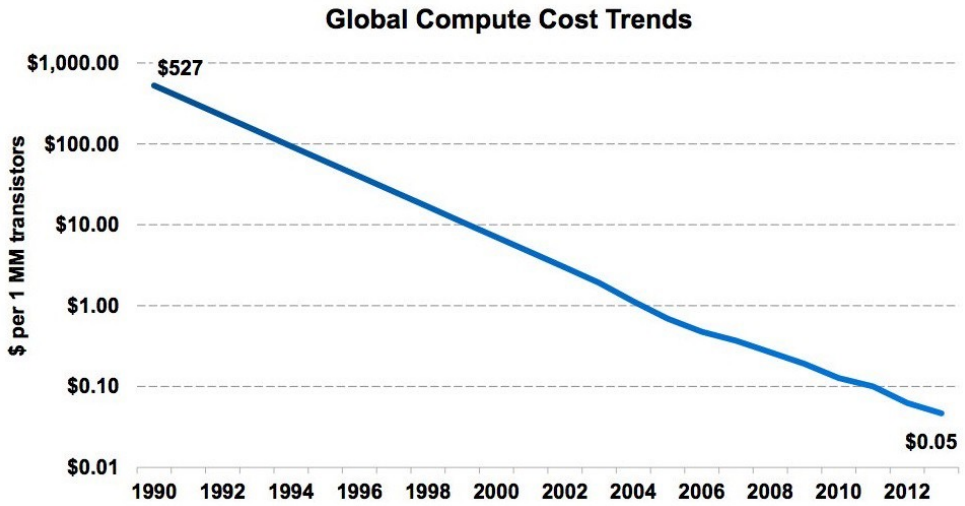
- McKinsey global report described big data as *the next frontier for innovation and competition*.
- The report defined big data as ”*Data whose scale, distribution, diversity, and/or timeliness require the use of new technical architectures and analytics to enable insights that unlock the new sources of business value*”

Big Data Revolution

IBM 5MB Hard Disk ;-)

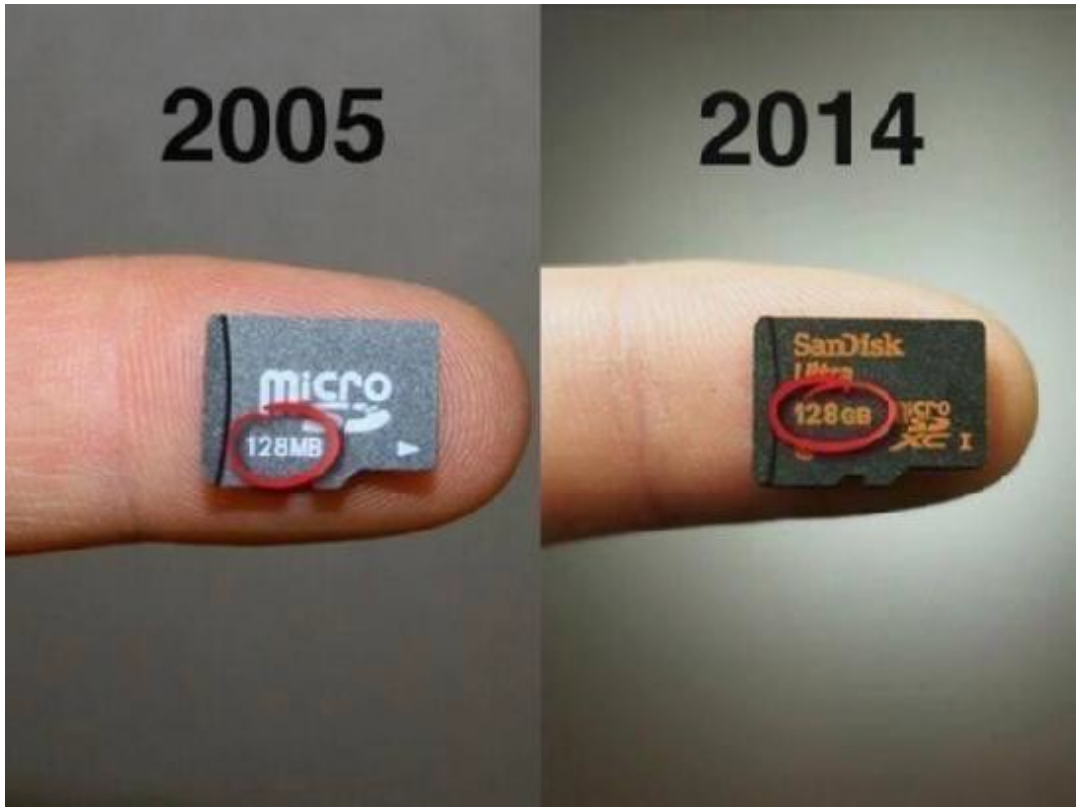
Recent Advances in Computational power

- Cheaper, larger, and faster **disk storage**
  - You can now put all your large database on disk
- Cheaper, larger, and faster **memory**
  - You may even be able to accommodate it all in memory
- Cheaper, more capable, and faster **processors**
- **Parallel computing architectures**:
  - Operate on large datasets in reasonable time
  - Try exhaustive searches and brute force solutions



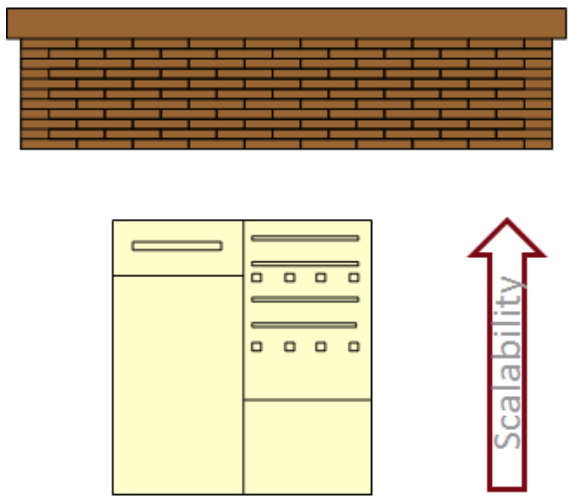
Big Data

- **Moore's Law**: The information density on silicon integrated circuits double every 18 to 24 months
- Users expect more sophisticated information

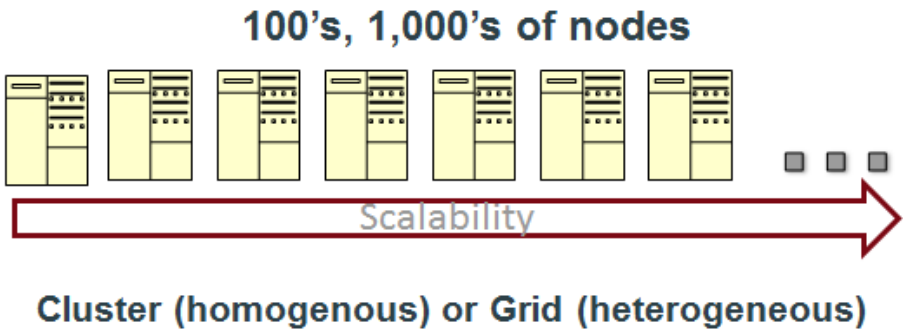


Your Pocket Size Terabytes Hard Disk

Hardware Advancements Enable Big Data Processing

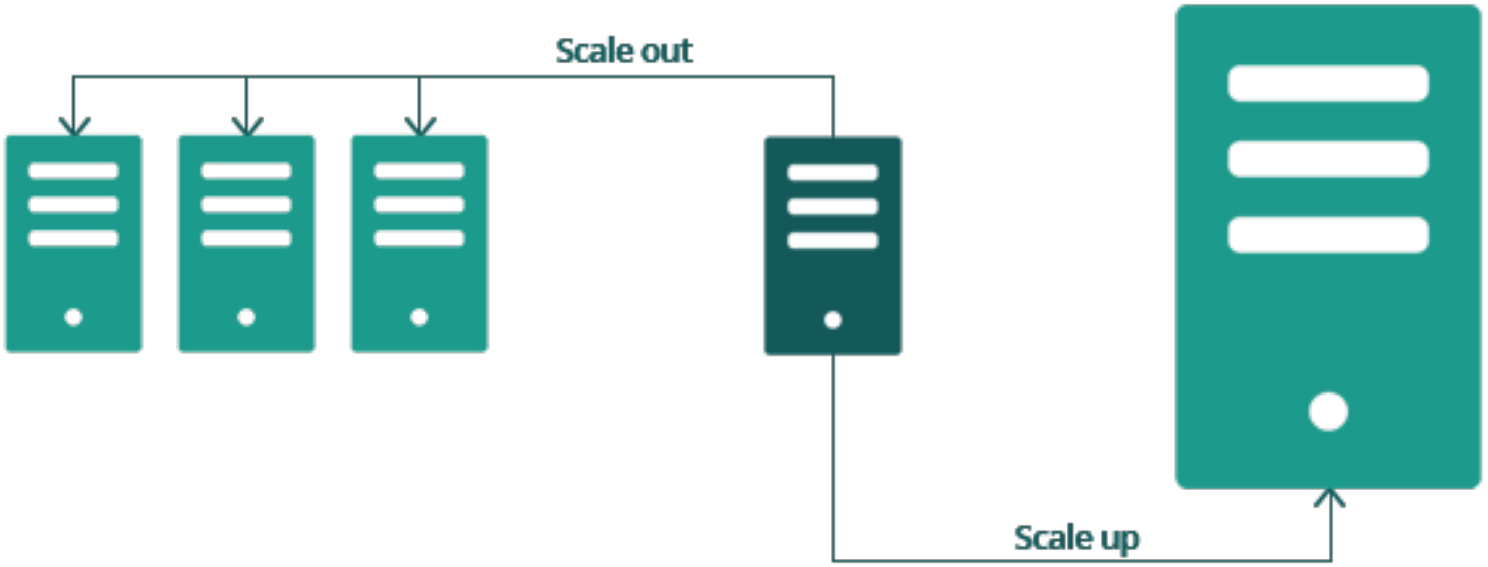


SMP – Symmetric Multiprocessing
"Shared Everything"
CPU, memory, disk (SAN, NAS)



MPP – Massively Parallel Processing
"Shared Nothing"
Nodes do not share
CPU, memory, disk (DAS)

Scale Up VS Scale Out



Scale out : Run your solutions on several servers .

Scale up : Run your solutions on bigger server.

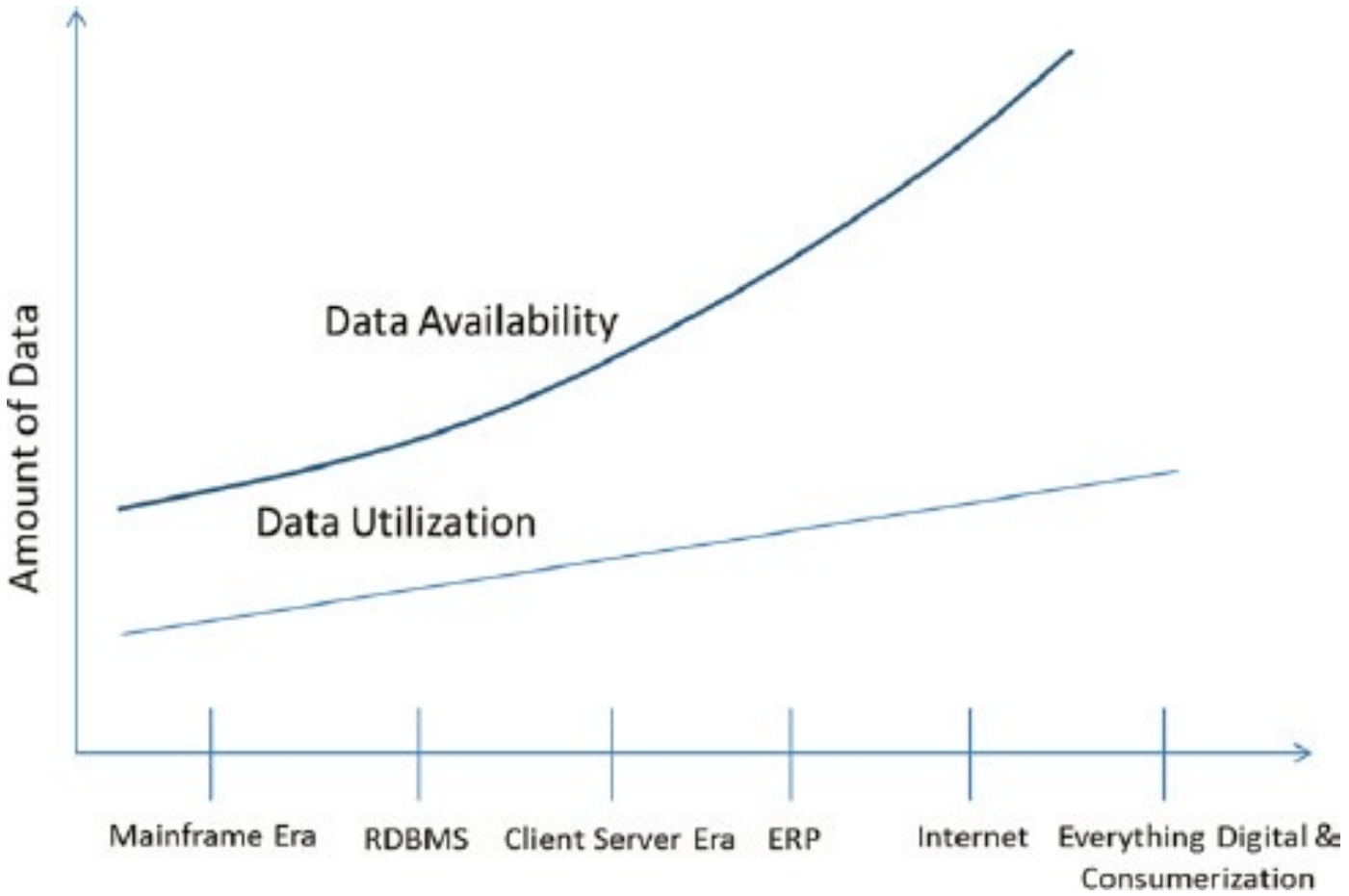
Scale Up	Scale Out
Vertical expansion/Upgrade to more powerful server configuration	Horizontal expansion through a grid or cluster of commodity servers
More expensive hardware Eventually hits a limit	Less expensive hardware Less likely to hit a limit

The Data Overload Problem

The Data Overload Problem

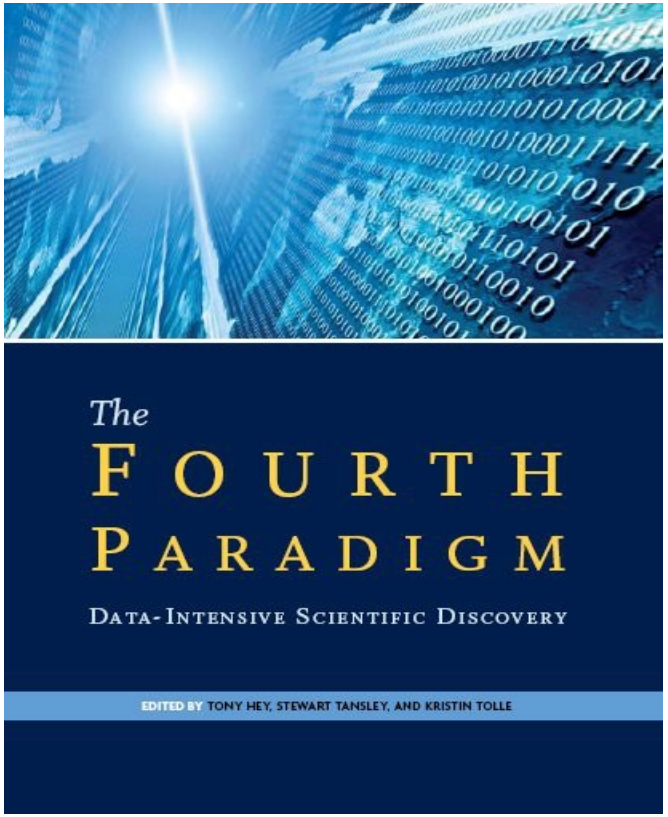
- Data is growing at a phenomenal rate. It has become massive, operational, and opportunistic
- **Drowning in data but starving for knowledge**
- The hidden information and knowledge in these mountains of data are really the most useful

The Data Overload Problem

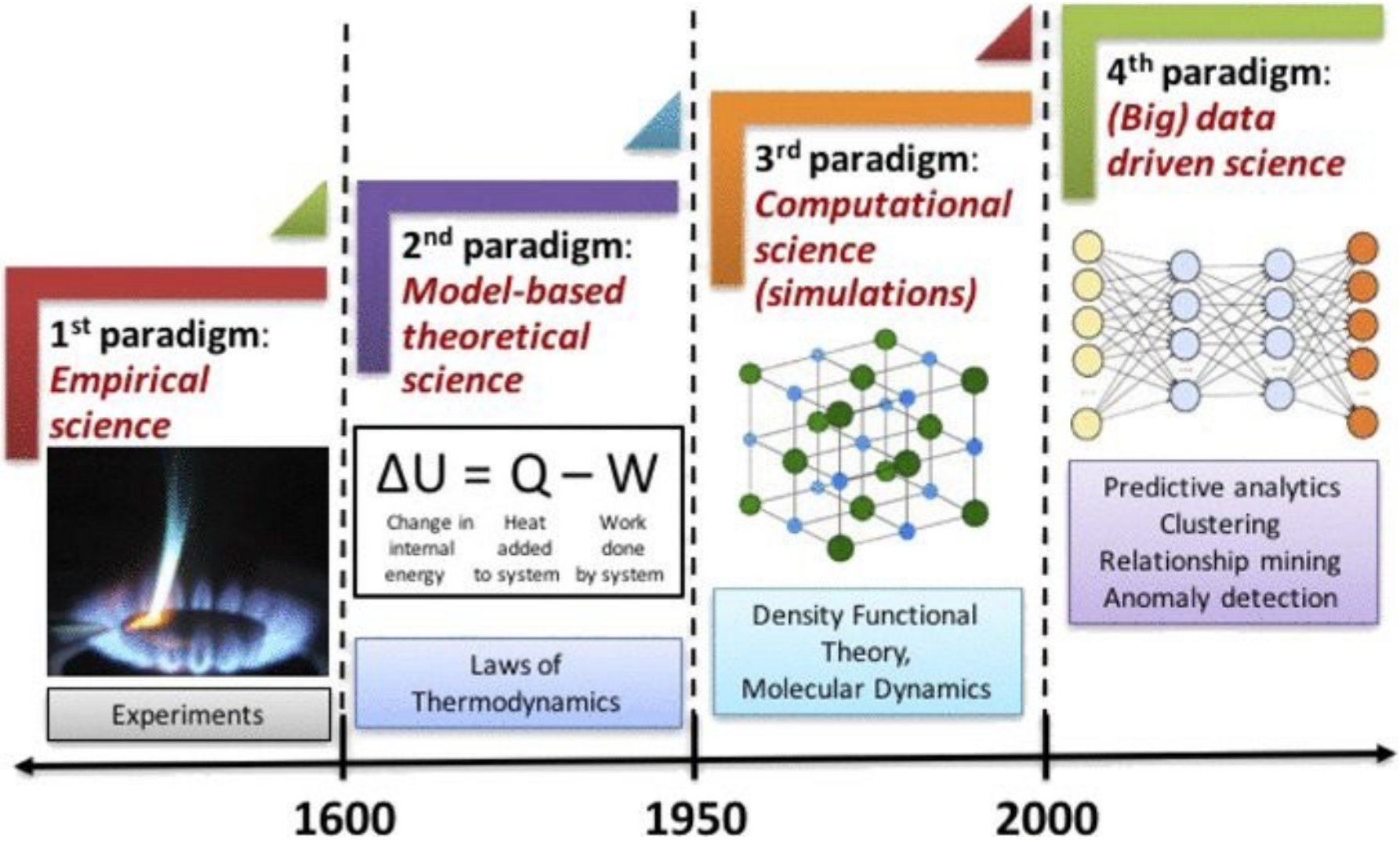


Fourth Paradigm

- Jim Gray, a database pioneer, described the big data phenomena as the **Fourth Paradigm** and called for a paradigm shift in the computing architecture and large scale data processing mechanisms.
- The first three paradigms were *experimental*, *theoretical* and, more recently, *computational science*



Fourth Paradigm



Fourth Paradigm

- Thousand years ago - **Experimental Science**
  - Description of natural phenomena
- Last few hundreds years - **Theoretical Science**
  - Newton's laws, Maxwell's equation , ...
- Last few decades - **Computational Science**
  - Simulations of complex phenomena
- Today - **Data-Intensive Science**
  - Scientists overwhelmed with datasets from many different sources
  - Data captured by instruments
  - Data generated by simulations
  - Data generated by sensor networks

Computing Clusters

- Many racks of computers, thousands of machines per cluster.
- Limited bisection bandwidth between racks.

Data Centers

Big Data is a Competitive Advantage

Big Data is a Competitive Advantage

”It’s not who has the best algorithm that wins, It’s who has the most data”

Andrew Ng

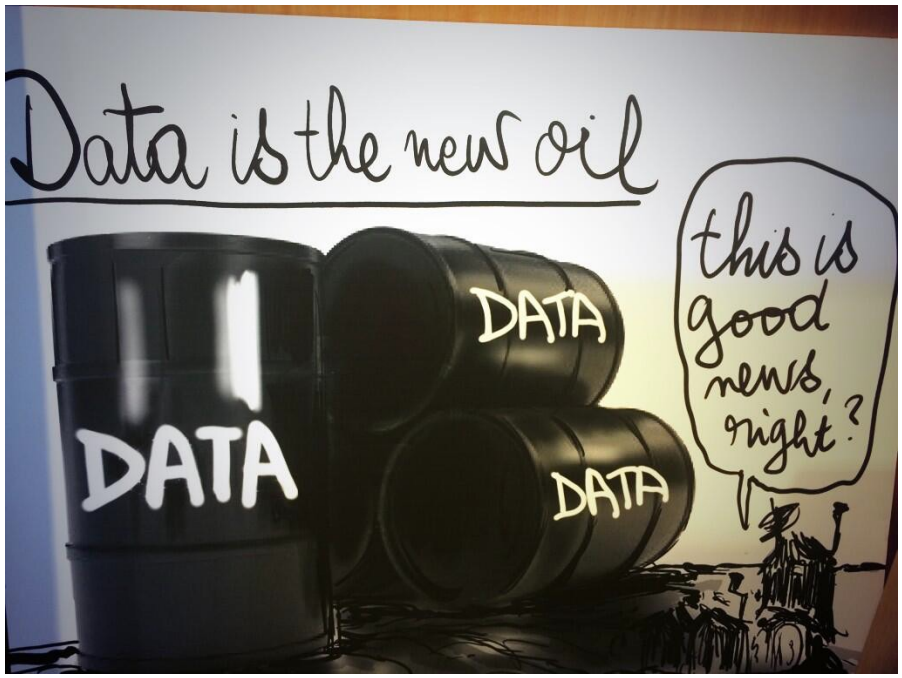
Data is the new Oil/Gold

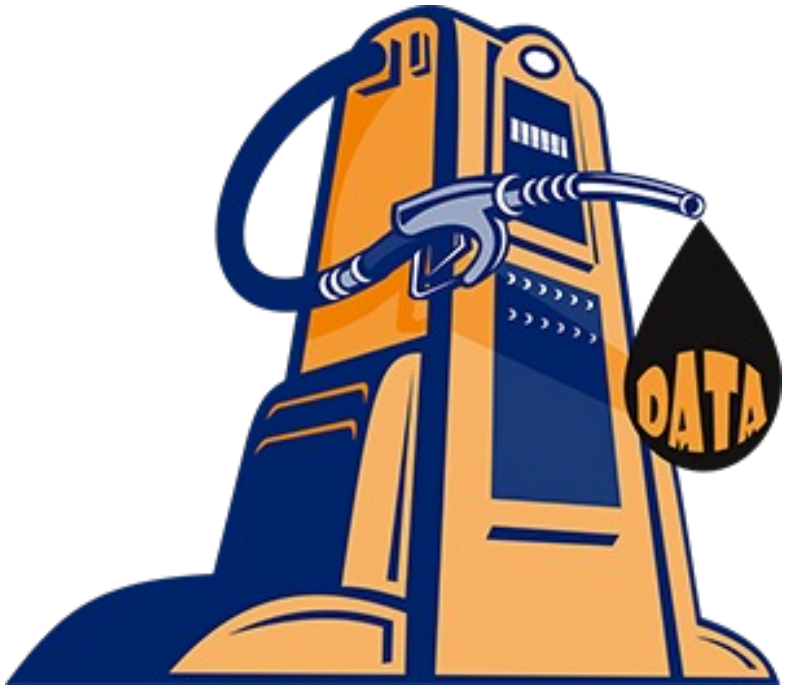
Big Data Processing Systems

Big Data is the New Oil

and

Big Data Processing Systems is the Machinery





Part II

Big Data 1.0 System: The Hadoop Decade

A Little History: Two Seminal contributions

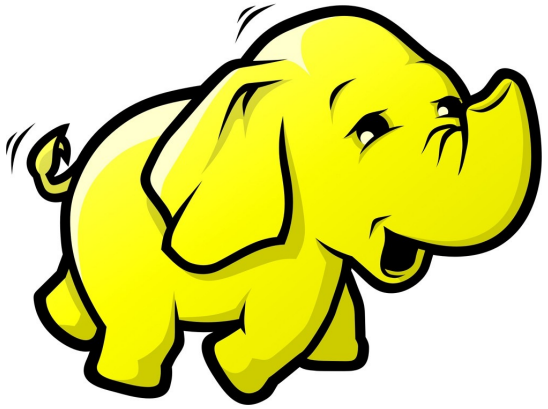
- **"The Google File System"**[2]
  - Describes a scalable, distributed, fault-tolerant file system tailored for data-intensive applications, running on inexpensive commodity hardware, delivers high aggregate performance
- **"MapReduce: Simplified Data Processing on Large Clusters"**[3]
  - Describes a simple programming model and an implementation for processing large data sets on computing clusters.

[2] S. Ghemawat, H. Gobioff, S. Leung. *The Google file system*. SOSP 2003

[3] J. Dean, S. Ghemawat. *MapReduce: Simplified Data Processing on Large Clusters*. OSDI 2004

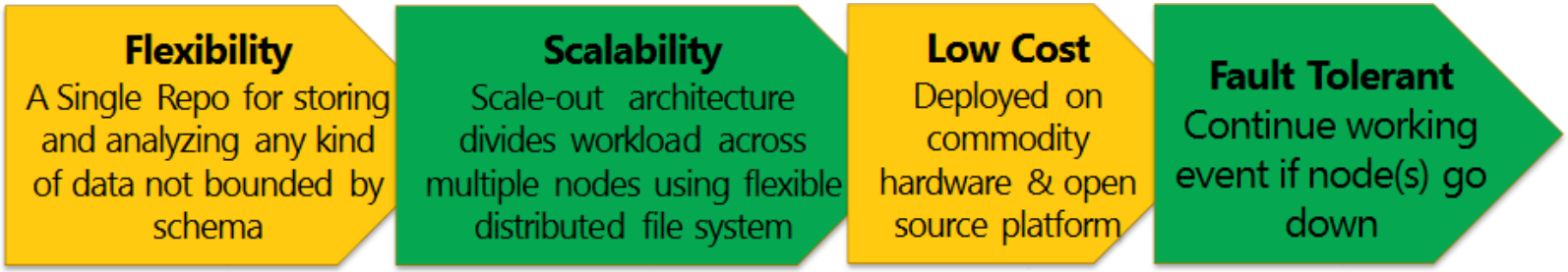
Hadoop[4]: A Star is Born

- Hadoop is an **open-source** software framework that supports data-intensive distributed applications and **clones** the Google's MapReduce framework.
- It is designed to process very large amount of unstructured and complex data.
- It is designed to run on a large number of machines that don't share any memory or disks.
- It is designed to run on a cluster of machines which can put together in relatively lower cost and easier maintenance.



[4] http://hadoop.apache.org/

Key Aspects of Hadoop

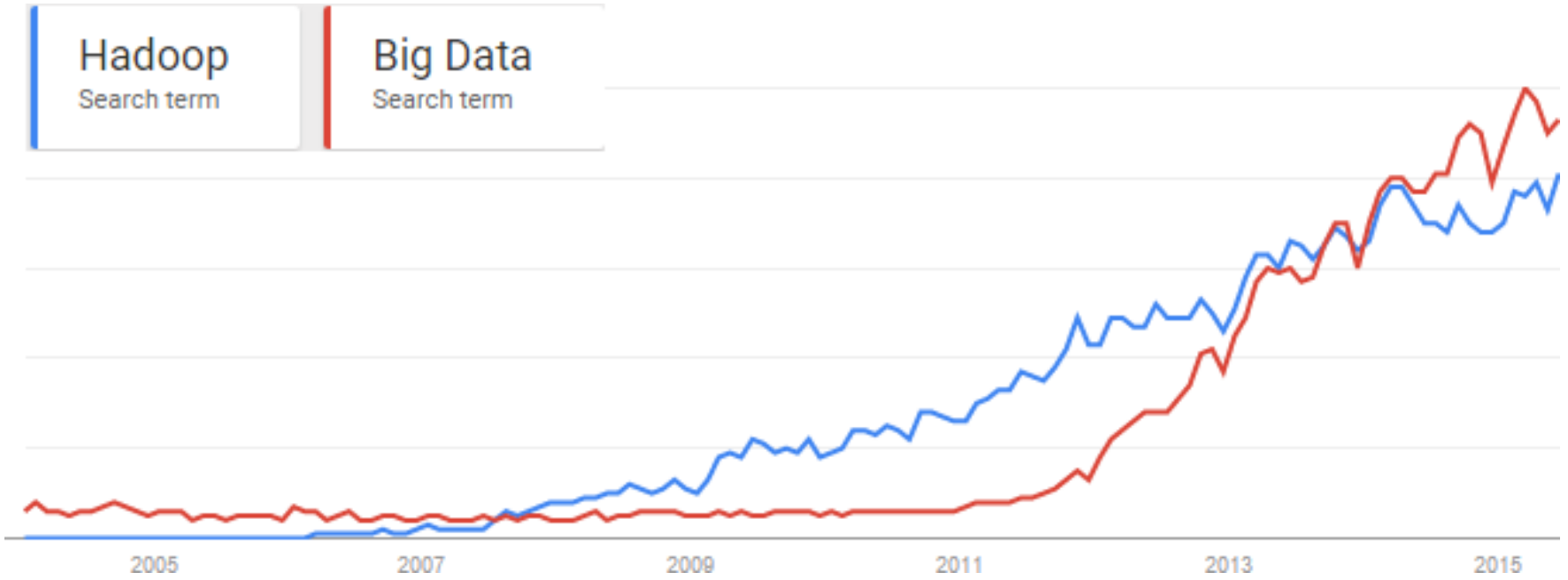


Hadoop's Success

Big Data 1.0 = Hadoop

Hadoop's Success[5]

Big Data 1.0 = Hadoop



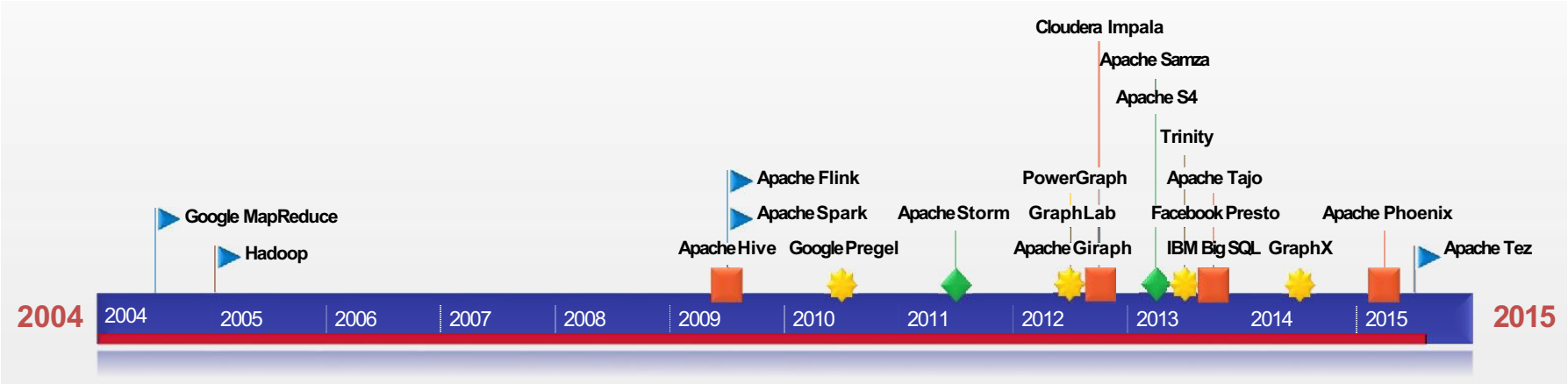
[5] https://www.google.com/trends/

The Always Dilemma: Does One Size Fit All?!

Big Data 2.0 Processing Systems

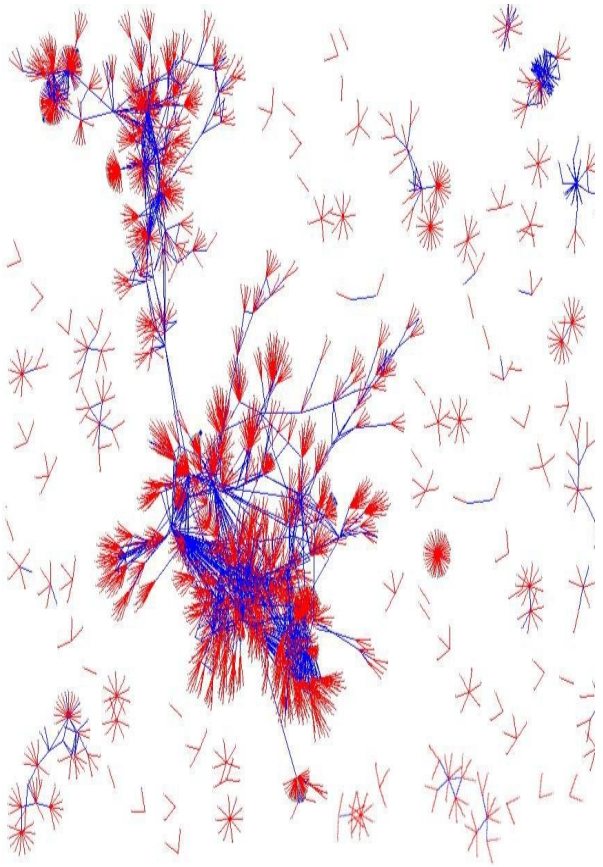
Big Data 2.0 != Hadoop

Domain-specific, optimized and vertically focused systems



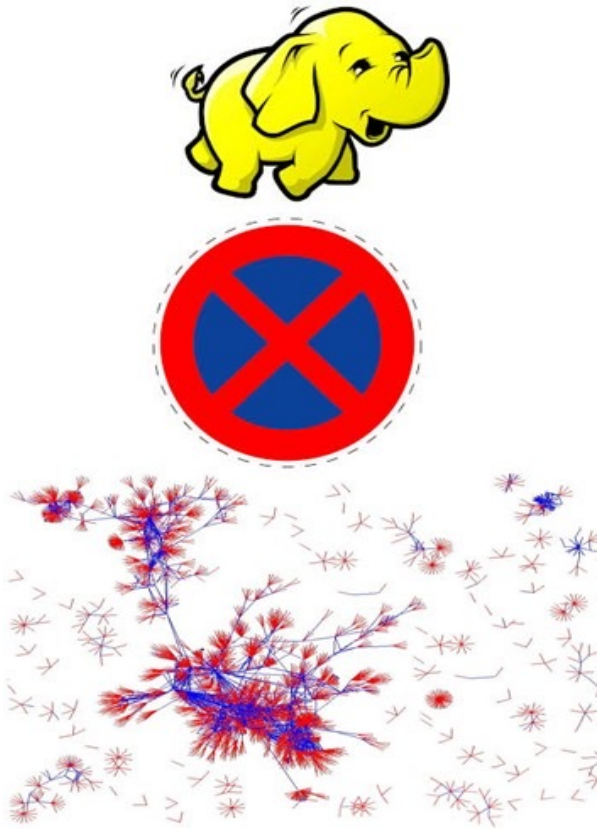
Big Graphs

- Google estimates that the total number of web pages exceeds 1 trillion; experimental graphs of the World Wide Web contain more than 20 billion nodes and 160 billion edges.
- Facebook reportedly consists of more than a billion users (nodes) and more than 140 billion friendship relationships (edges) in 2012.
- The LinkedIn network contains almost 260 million nodes and billions of edges.
- Linked data contains about 31 billion triples.



Hadoop for Big Graphs?!

- Popular graph query/analysis operations such as: Page rank, Pattern matching, Shortest path, Clustering (e.g. Max clique, triangle closure), Community detection,..., etc. are **iterative** in nature.
- MapReduce programming model does not directly support iterative data analysis. Programmers may implement iterative programs by manually issuing multiple MapReduce jobs and orchestrating their execution using a driver program which wastes I/O, network bandwidth and CPU resources.
- *It is not intuitive to think of graphs as key/value pairs or matrices*.



Pregel/Giraph[6]

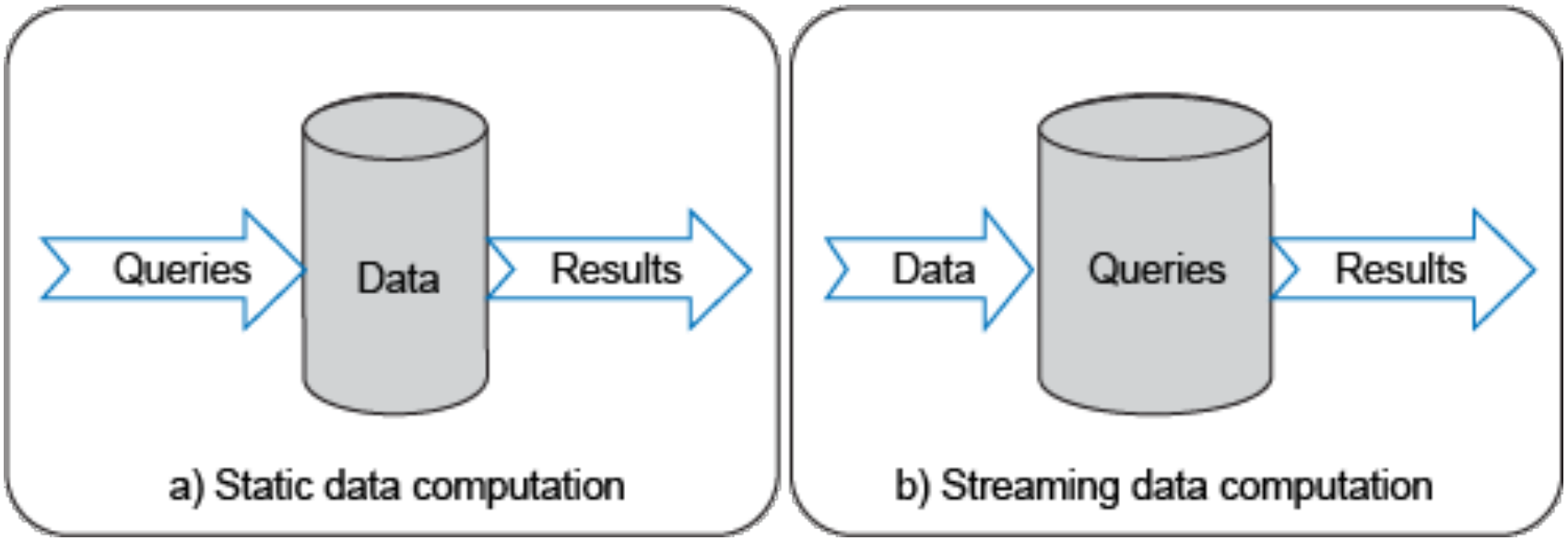
- In 2010, Google introduced the **Pregel** system as a scalable platform for implementing graph algorithms.
- Pregel relies on a vertex-centric approach and is inspired by the Bulk Synchronous Parallel (BSP) model.
- In 2012, **Apache Giraph** was launched as an open source project that clones the concepts of Pregel and leverages the Hadoop infrastructure.
- Other Projects: *Spark GraphX* (Apache), *GoldenOrb* (Apche), *GraphLab* (CMU) and *Signal/Collect* (UZH).

[6]https://giraph.apache.org/

Big Streaming Data

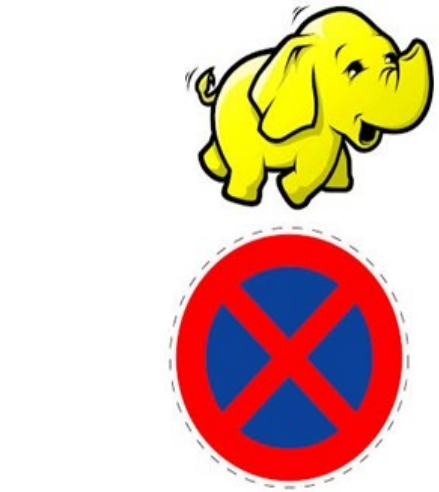
- Every day, Twitter generates more than 12 TB of tweets.
- New York Stock Exchange captures 1 TB of trade information.
- About 30 billion radio-frequency identification (RFID) tags are created every day.
- Hundreds of millions of GPS devices sold every year.

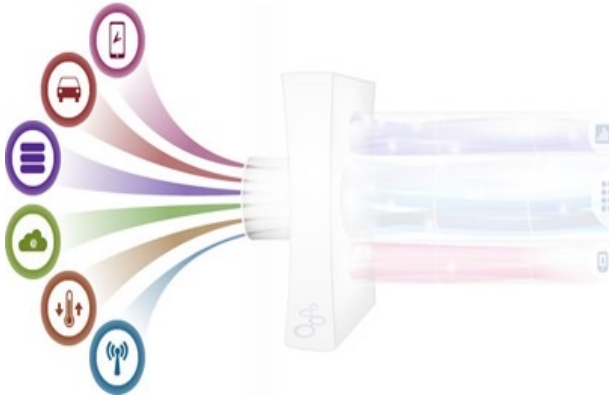
Static Data Computation vs Streaming Data Computation



Hadoop for Big Streams?!

- From the stream-processing point of view, the main limitation of the original implementation of the MapReduce framework is that it was designed so that the entire output of each map and reduce task is **materialized** into a local file before it can be consumed by the next stage.
- This materialization step enables the implementation of a simple and elegant checkpoint/restart fault-tolerance mechanism. But it causes significant delay for jobs with real-time processing requirements.
- Some Hadoop-based Trials include: *MapReduce Online* and *Incoop*.





Twitter Storm[7]

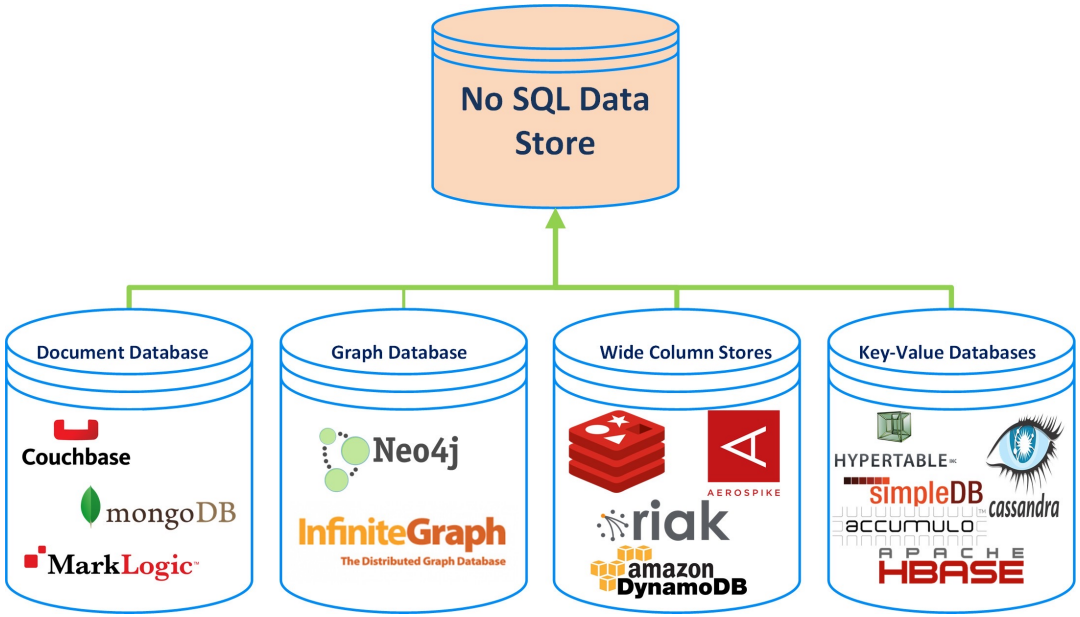
- Open source project developed by Nathan Marz and acquired by Twitter in 2012.
- Storm is a distributed stream-processing system with the following key design features: horizontal scalability, guaranteed reliable communication between the processing nodes, fault tolerance and programming-language agnosticism.
- A Storm cluster is superficially similar to a Hadoop cluster. One key difference is that a MapReduce job eventually finishes, whereas a Storm job processes messages forever (or until the user kills it).
- Other Projects: *Flink*, *Apex*, *Spark Streaming* and *Kafka Streams*.

[7] http://storm.apache.org/

Massively Parallel Processing (MPP) Optimized SQL query engines

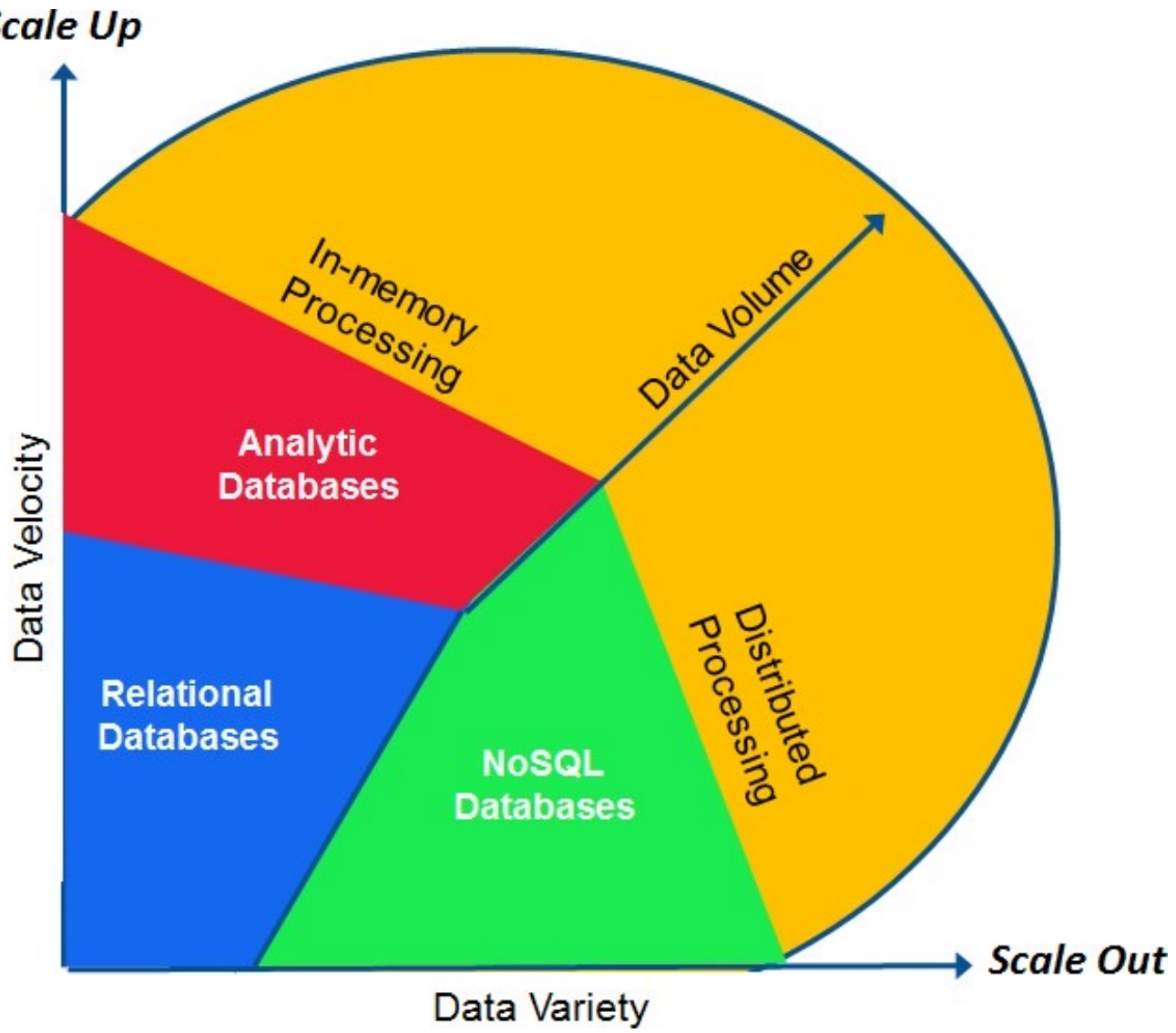
NoSQL Databases[8]

- NoSQL database systems represent a new generation of low-cost, high-performance database software which is increasingly gaining more and more popularity.
- These systems promise to simplify administration, be fault-tolerant and able to scale on commodity hardware (Scale out).



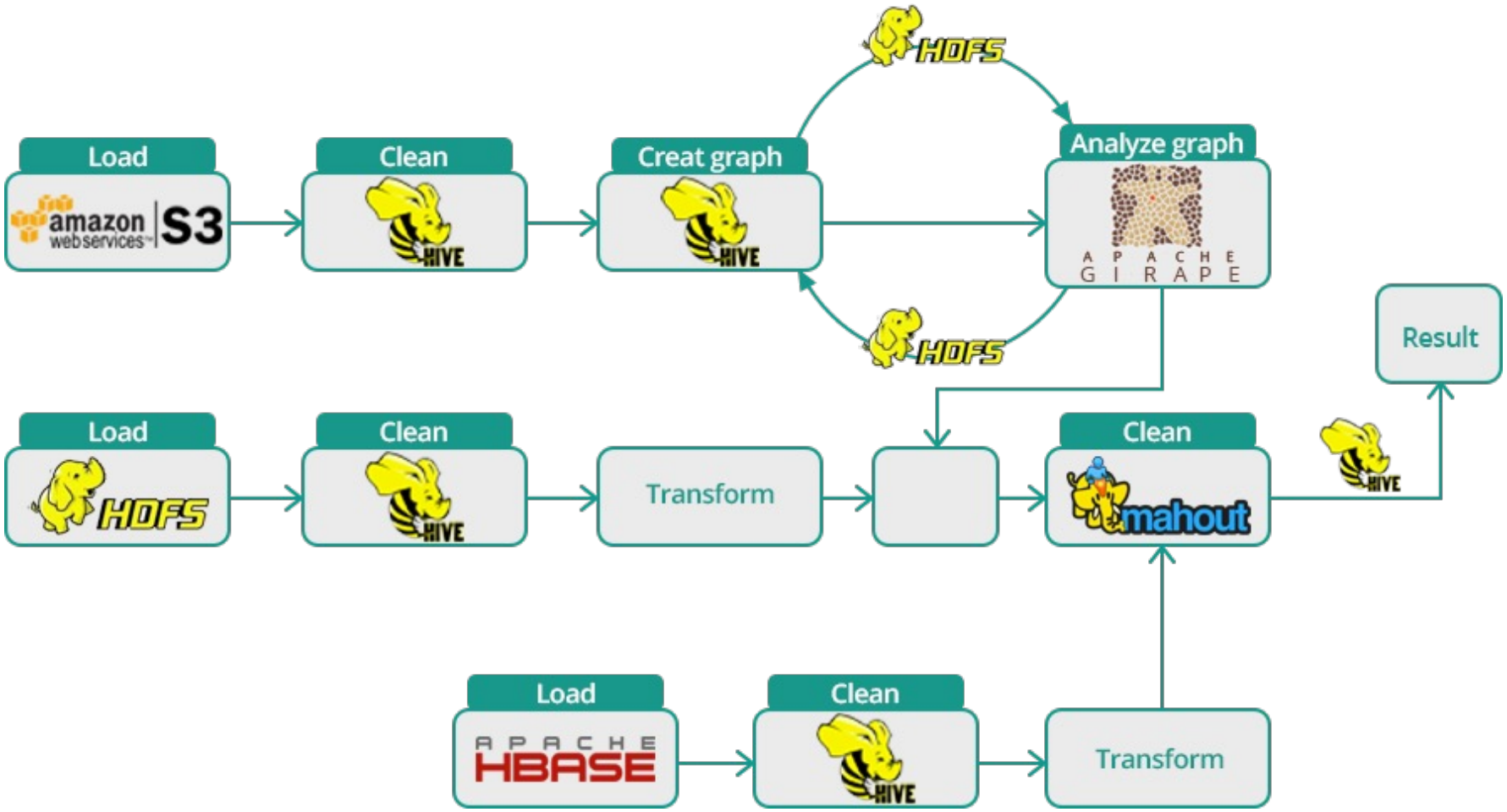
[8] http://nosql-database.org/

Data Storage Options

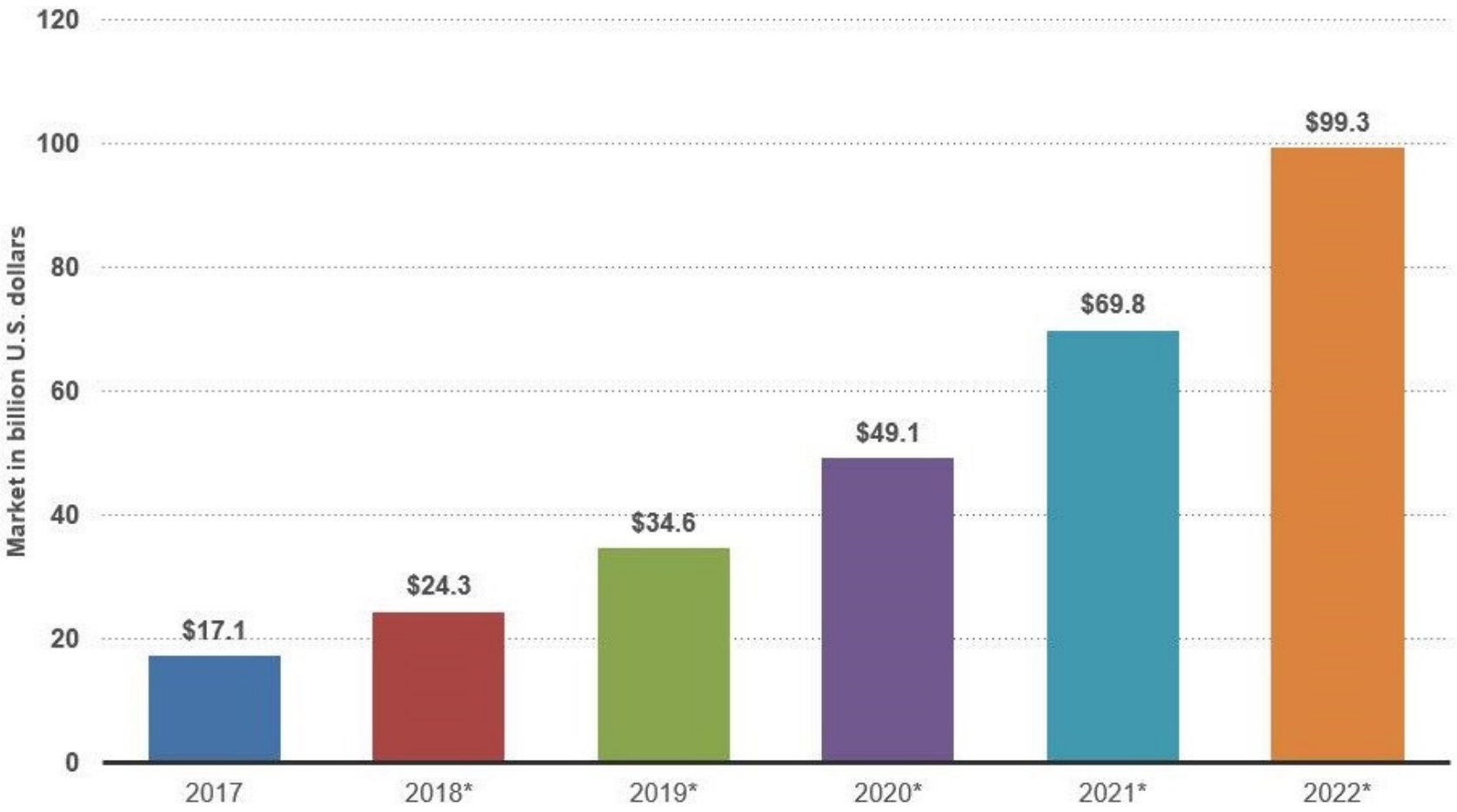


Big Data Landscape

Big Data Landscape



Big Data Market Size[9]



[9] https://www.statista.com/

The End


Thank You
Mahalo
Kiitos
Tack
Toda
Grazie
Thanks
Obrigado
Takk
Merci
Gracias