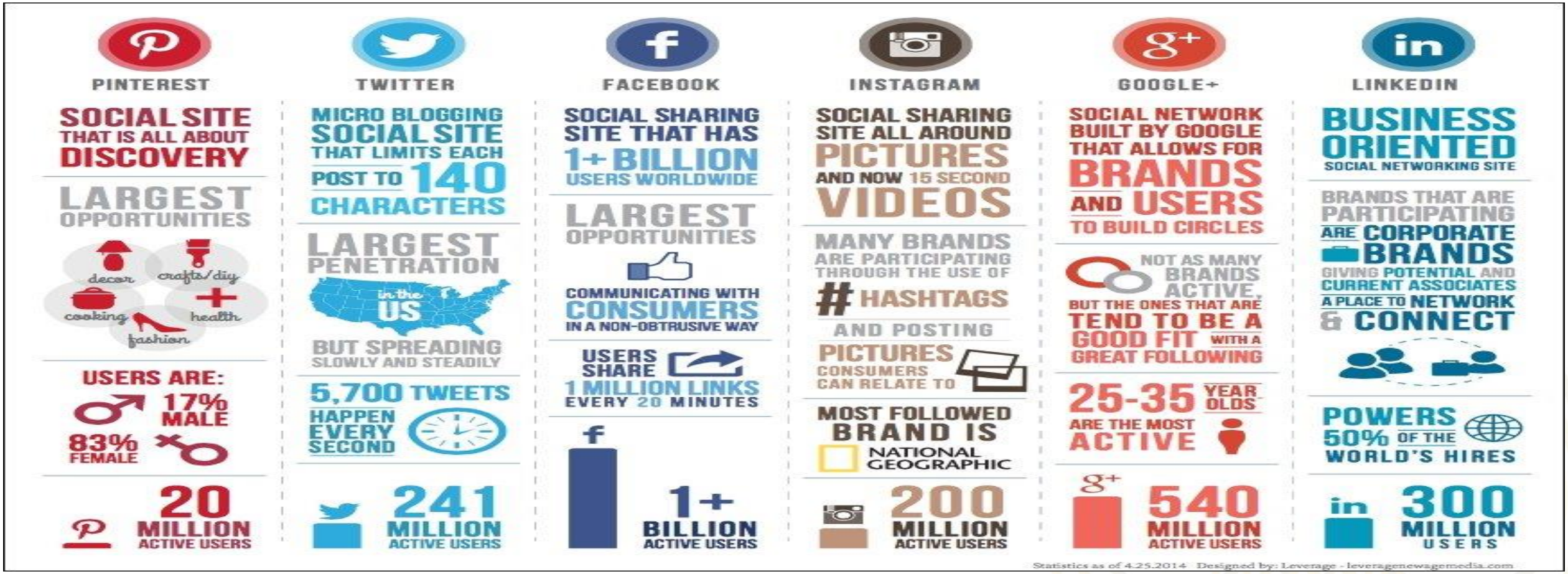
Mahidol University
Wisdom of the Land

BIG DATA PROCESSING

Week1-summary

Mahidol University
Wisdom of the Land

Big data generated by people



Characteristic of human genarated: Unstructured

- Text
- Image

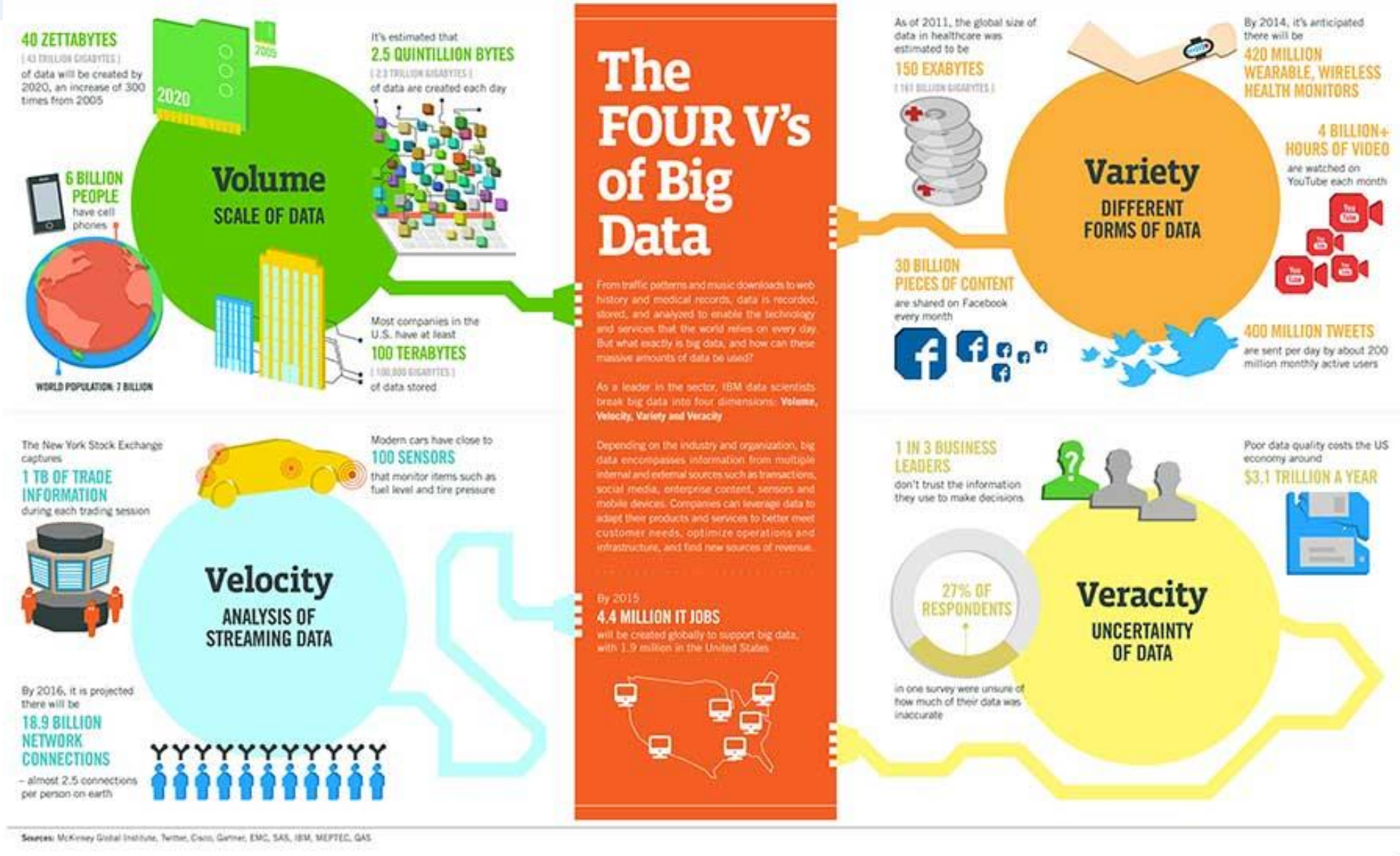
- Huge grown

- Challenge

Company	Data Processed Daily
eBay	100 Petabytes (PB)
Google	100 PB
Facebook	30+ PB
Twitter	100 Terabytes(=.1PB)
Spotify	64 Terabytes

Mahidol University
Wisdom of the Land

Characteristics of Big Data

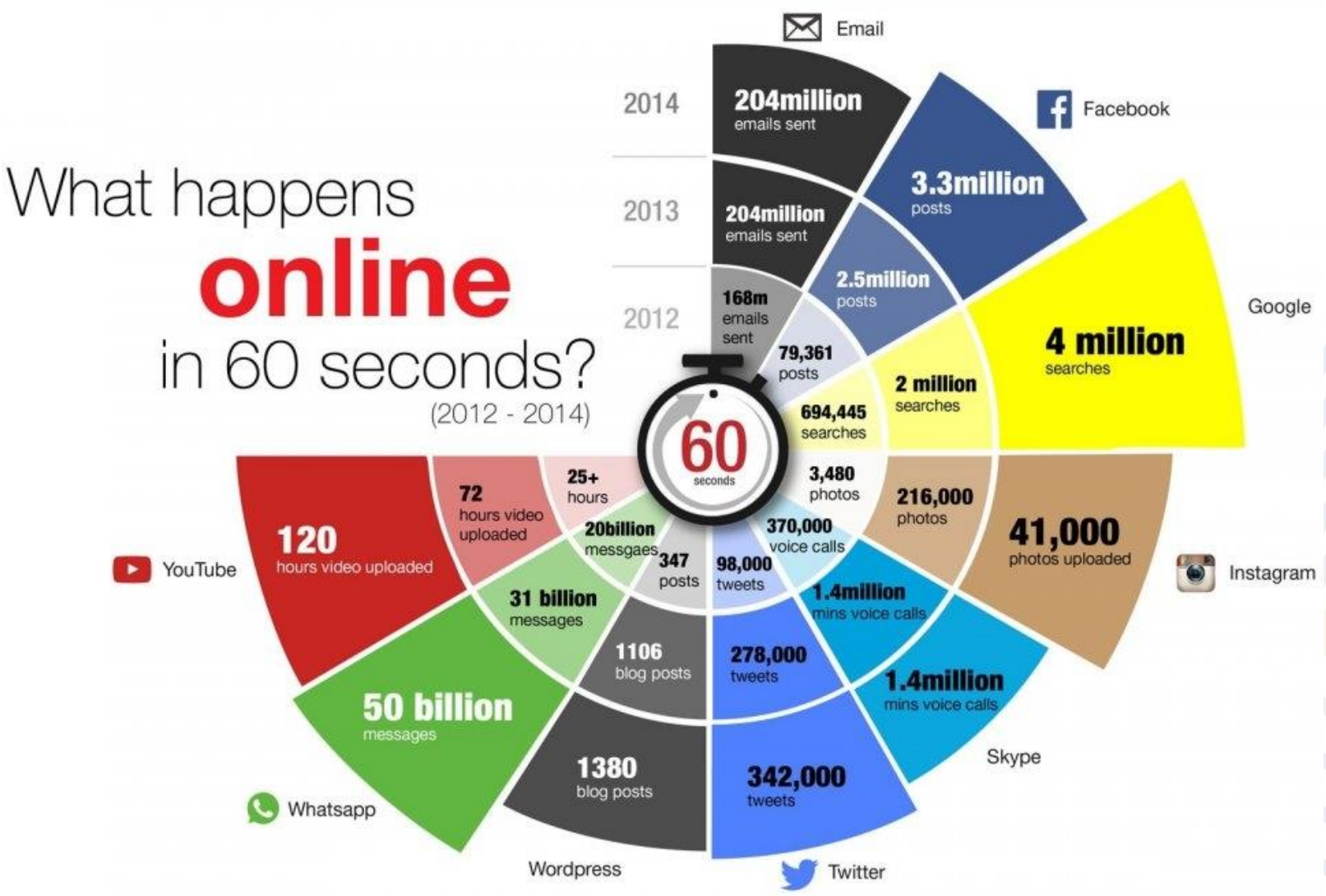


All V's

- Volume
- Velocity
- Variety
- Veracity
- Valence (connectedness)
- **Value**

V- Volume

- VOLUME = SIZE



What about now??

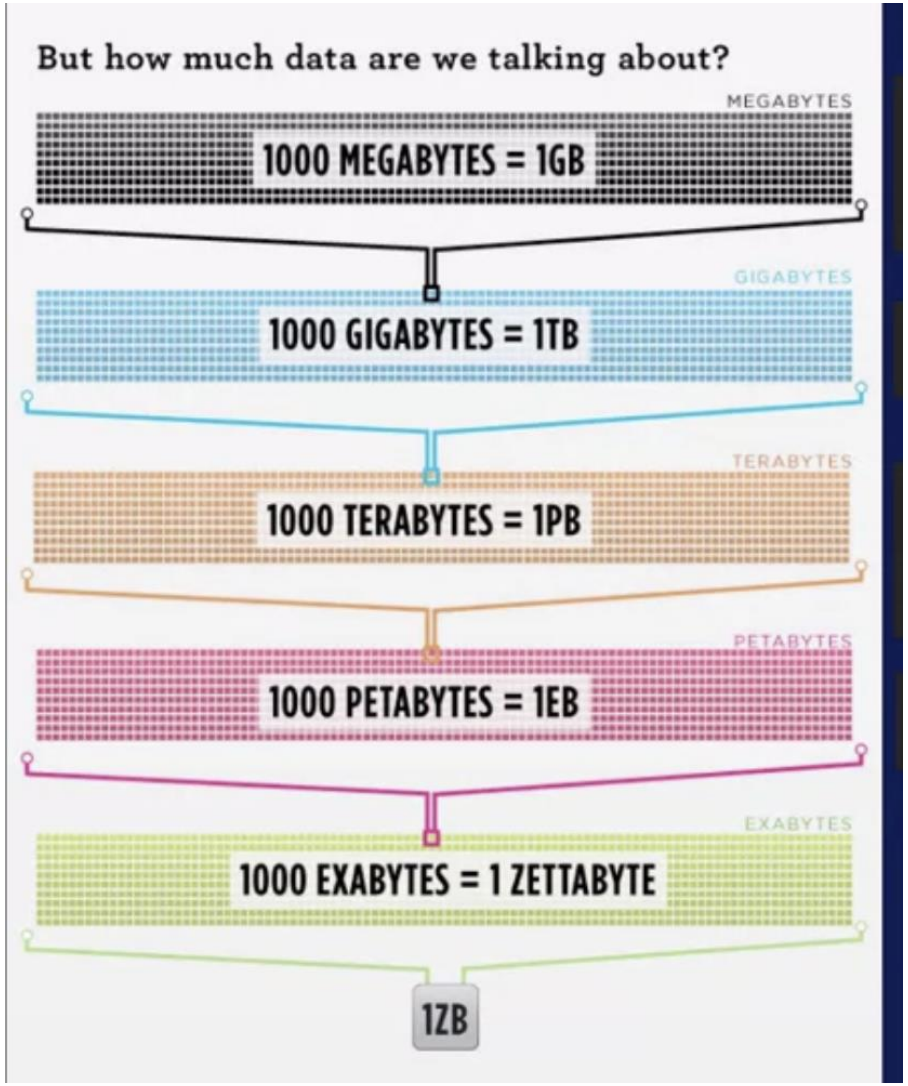
- FACEBOOK

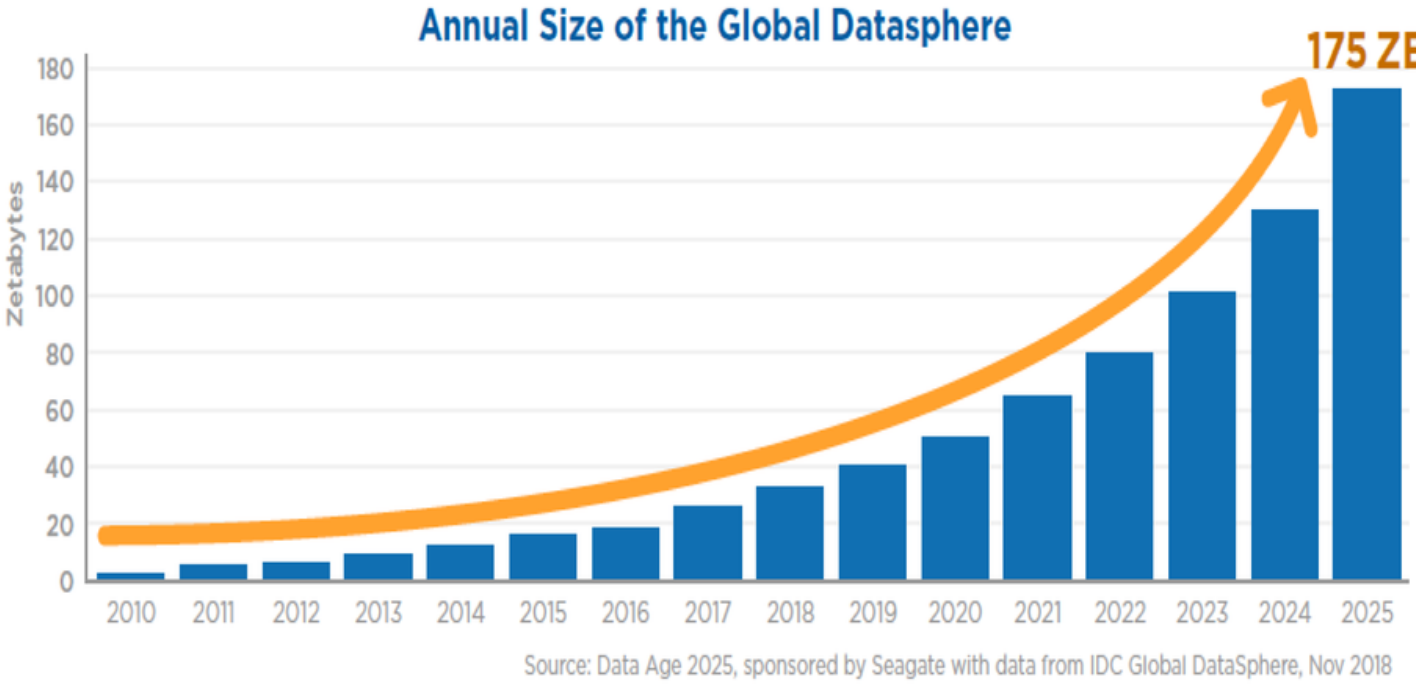
  Twitter

- Tiktok

V-Volume

Twitter data





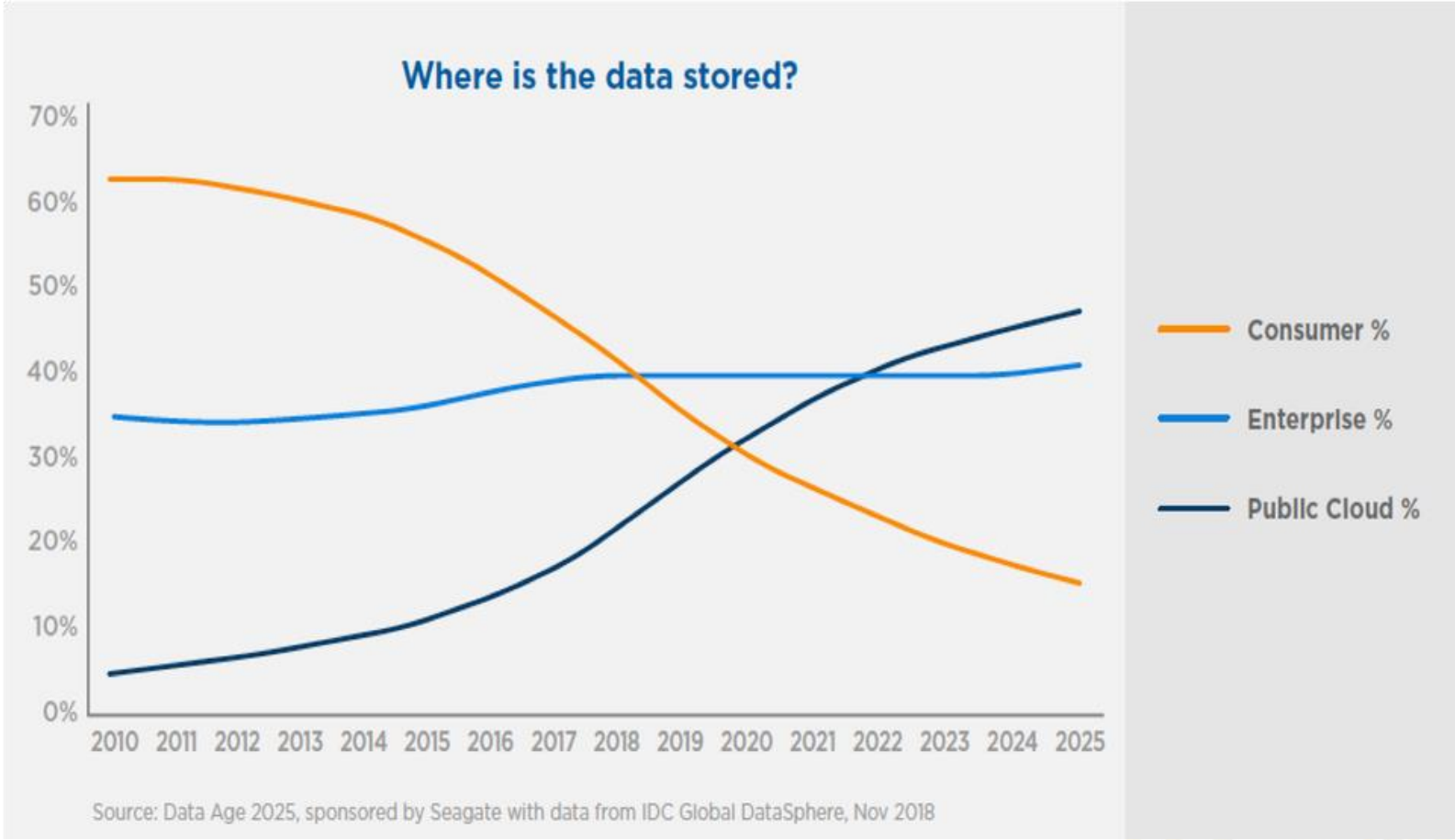
Volume = Challenge

- Storage → Processing
  - Cost of storage
  - Cost of speed/Process
  - I/O needs

- Challenge
  - Cost
  - Performance may drop
  - Scalability
  - performance to storage, access and processing

Mahidol University
Wisdom of the Land

Volume= SIZE



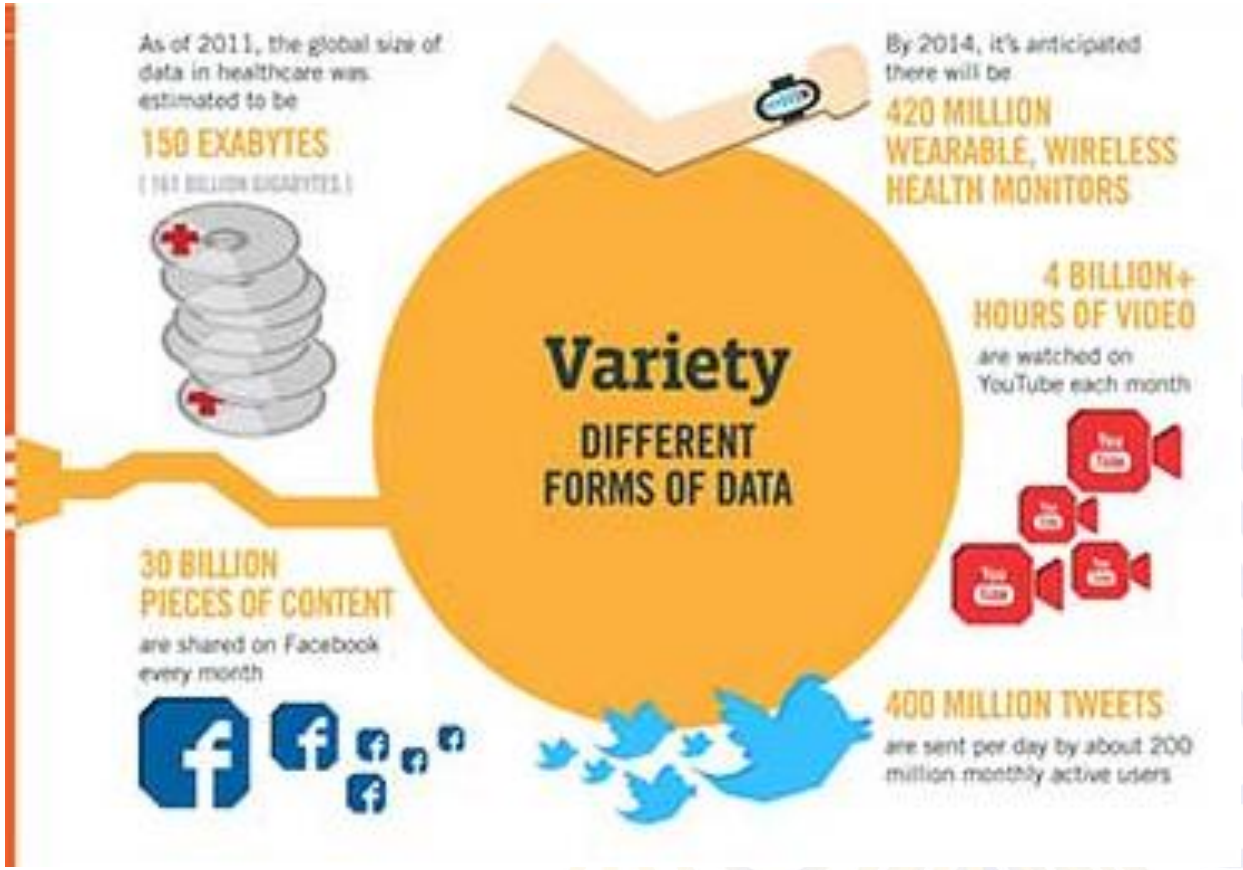
Power of Ten

Astronomical Scale

https://www.youtube.com/watch?v=0fKBhvDjuy0

V=Variety

- Table structure data
- Health data
- Sattelite data
- Social Media Data
- Youtube Vdos



Mahidol University
Wisdom of the Land

Traditionl ETL

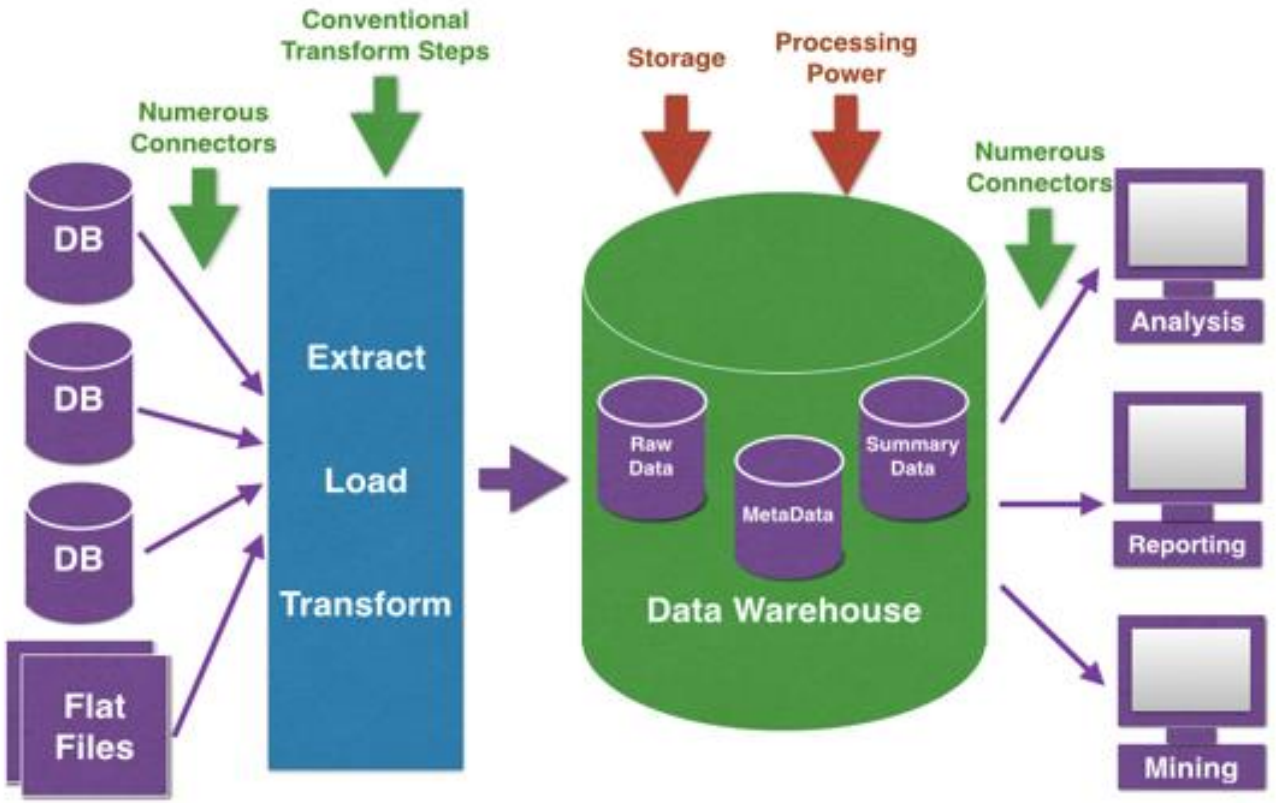


Figure 1 - Traditional Data Integration

Traditional Structured data

- Example: Google Form

Berghotel
Grosse Scheidegg
3818 Grindelwald
Familie R.Müller

Rech.Nr. 4572 30.07.2007/13:29:17
Bar Tisch 7/01

2xLatte Macchiato	à	4.50	CHF	9.00
1xGloki	à	5.00	CHF	5.00
1xSchweinschnitzel	à	22.00	CHF	22.00
1xChässpätzli	à	18.50	CHF	18.50

Total : CHF 54.50

Incl. 7.6% MwSt 54.50 CHF: 3.85

Entspricht in Euro 36.33 EUR
Es bediente Sie: Ursula

MwSt Nr.: 430 234
Tel.: 033 853 67 16
Fax.: 033 853 67 19
E-mail: grossescheidegg@bluewin.ch

Unstructured data

- Media records
- Social media
- Business documents – email , presentation, etc
- Communication– chat
- Survey response
- Publication
- Web pages

80-90% unstructured

Dimension of data variety

- Structural variety
  - formats,scale, magnitudes
- Media variety
  - Vdo/audio /Subtitle
- Semantic variety
  - Qualitative, Quantity values
- Availability
  - real-time/ itermittently

V=Variety

- Media Variety
  - *Youtube vdo requires picture/sound/transcript*
- Data object variety
  - News coverage→ sequences of items
- Semantic variety
  - *1 Venti capuccino with 2 pumps of vanilla syrup*

V=Variety => time frame

- Real time data/ store data
  - *Sensors, heart rate*
  - health history

- Hybrid of data
  - *Email*
    - *Senders, reciver, subject, date*
    - *Body of the email*
    - *Attachment*
    - *Who send to whom*

Hybrid of structured/ unstructured data

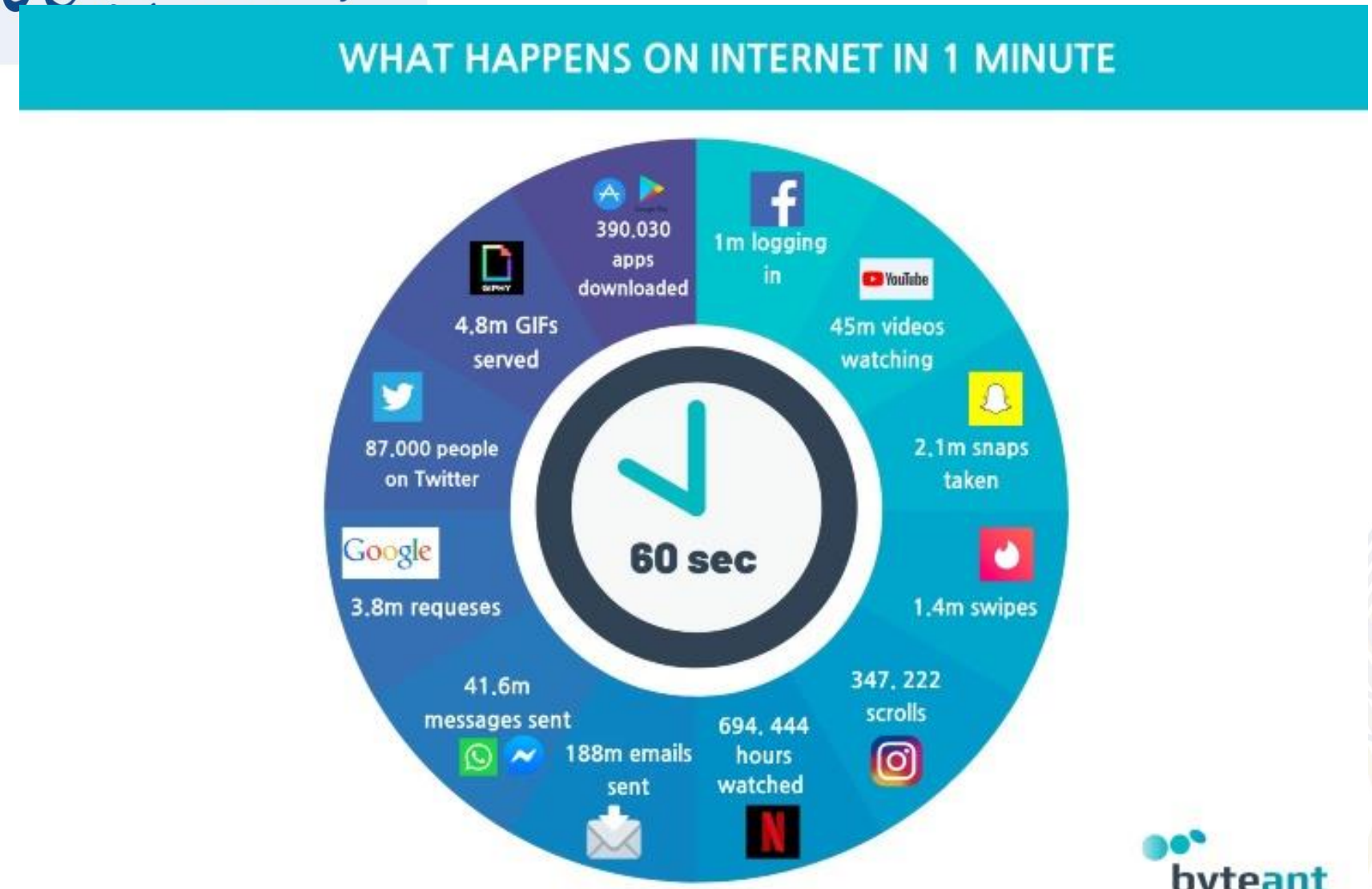
- Email

V=velocity

- Velocity = Speed
  - To Create
  - To Store
  - To Analyze

Mahidol University

Data flow



Mahidol University
Wisdom of the Land

V=velocity= real time

Batch Processing

- Collect -> Clean -> Feed in Chunks -> **Wait** -> Act

Real-Time Processing(Online)

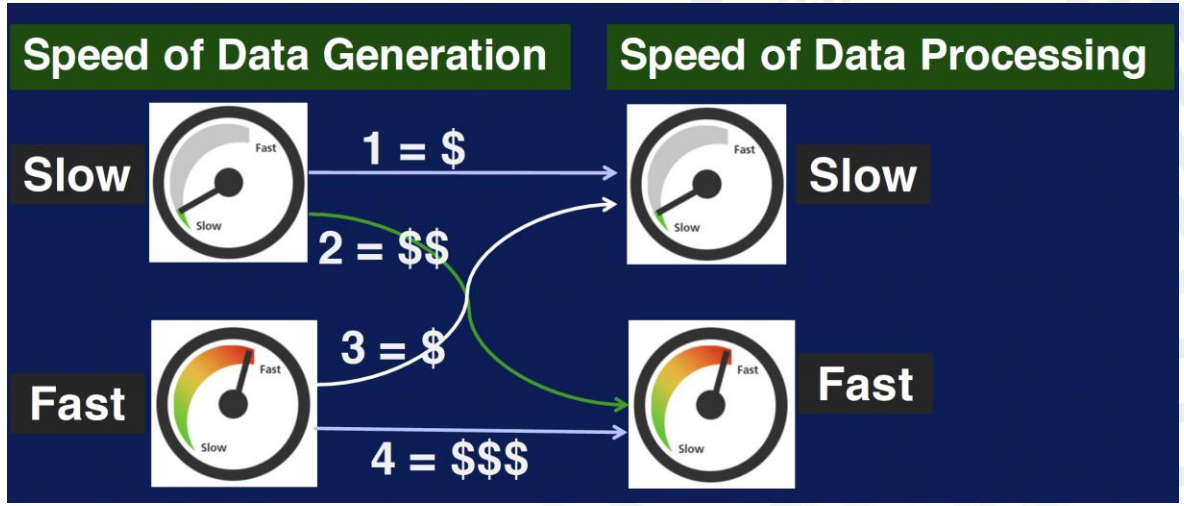
- Instant capture -> Feed in realtime -> Process in realtime ->Act

V=velocity= Challenge

- Rate of data driven action

  v.s.

- Rate of data generation
- Rate data processing

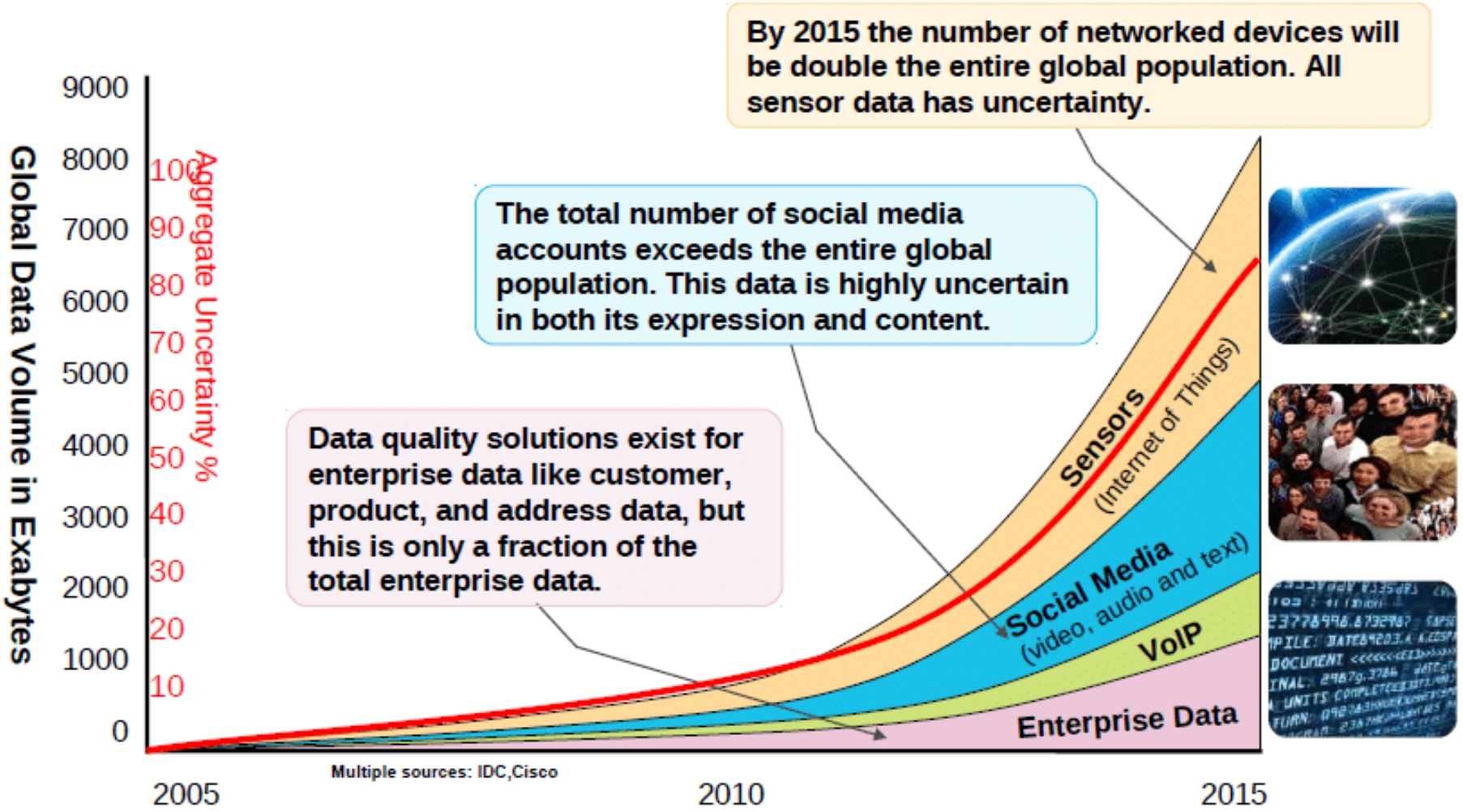


Mahidol University
Wisdom of the Land

Velocity

- Streaming data
  - What happening NOW
  - What's the surrounding now
- Act
- Make accurate decisions

Uncertainty of the data



ref: https://www.researchgate.net/publication/329755755_Big_Data_A_Trouble_or_A_Real_Solution/download

Mahidol University
Wisdom of the Land

Amazon Review

Hutzler

Hutzler 571 Banana Slicer

4.5 out of 5 stars ▾ 5,702 customer reviews | 706 answered questions

Amazon's Choice for "banana slicer"

Price: $5.23 ✓prime

FREE Shipping on orders over $25—or get **FREE Two-Day Shipping** with Amazon Prime

In Stock.

Want it tomorrow, Jan. 24? Order within **6 hrs 52 mins** and choose **One-Day Shipping** at checkout.

Details

Ships from and sold by Amazon.com. Gift-wrap available.

Package Quantity: **1**

Size: **11.25**

- Faster, safer than using a knife
- Great for cereal
- Plastic, dishwasher safe
- Slice your banana with one quick motion
- Kids love slicing their own bananas

Compare with similar items

Mrs Toledo

5 out of 5 stars **Saved my marriage**

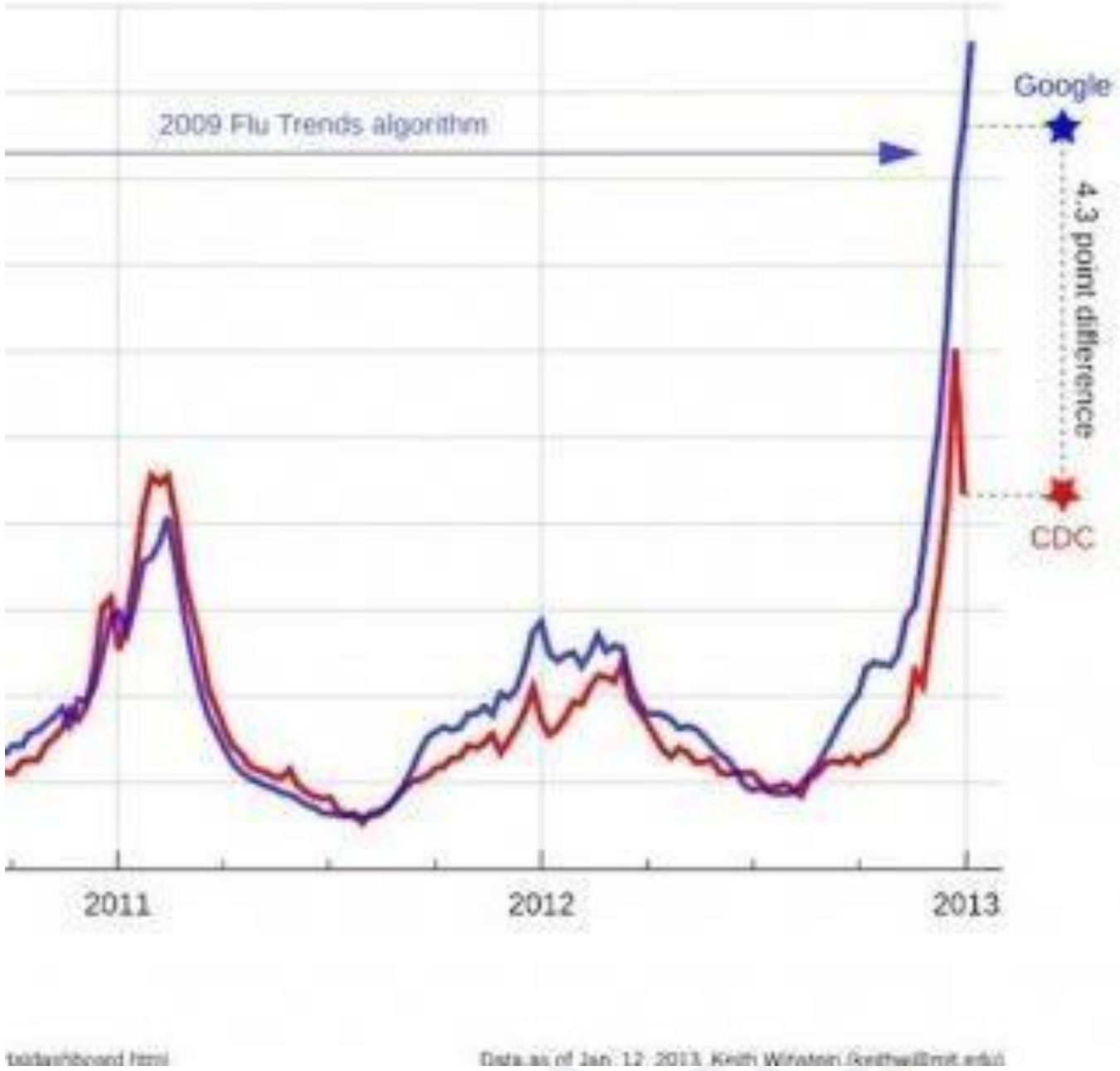
July 30, 2012

What can I say about the 571B Banana Slicer that hasn't already been said about the wheel, penicillin, or the iPhone.... this is one of the greatest inventions of all time. My husband and I would argue constantly over who had to cut the day's banana slices. It's one of those chores NO ONE wants to do! You know, the old "I spent the entire day rearing OUR children, maybe YOU can pitch in a little and cut these bananas?" and of course, "You think I have the energy to slave over your damn bananas? I worked a 12 hour shift just to come home to THIS?!" These are the things that can destroy an entire relationship. It got to the point where our children could sense the tension. The minute I heard our 6-year-old girl in her bedroom, re-enacting our daily banana fight with her Barbie dolls, I knew we had to make a change. That's when I found the 571B Banana Slicer. Our marriage has never been healthier, AND we've even incorporated it into our lovemaking. THANKS 571B BANANA SLICER!

Comment | 34,702 people found this helpful. Was this review helpful to you? Yes No Report abuse

Mahidol University
Wisdom of the Land

Google Flu trend



V=veracity = Challenge

- Accuracy
- Where it came from?
- How has it gone to?
- What transfomation does it go to?

V= Valence

- Valene= connectedness
- Facebook
  - Two Facebook users
- Companies
  - Company group
- Employe → Company

V= Valence =challenges

- Valence increases over time
- More complex data exploration
  - Dense valence make data anlysis more complex
- Model and predict of valence change
- Group event detection
- Emergent behaviour analysis

Mahidol University
Wisdom of the Land

V = Value

- What we can do about the big data
- Pull value from the big data
- Derive the value of th big data
- The solution is uniques

Mahidol University
Wisdom of the Land

Examples

- Twitter uses sentiment analysis
  - https://monkeylearn.com/blog/sentiment-analysis-of-twitter/
- Amazon uses recommendation system

- Diaster management

Mahidol University
Wisdom of the Land

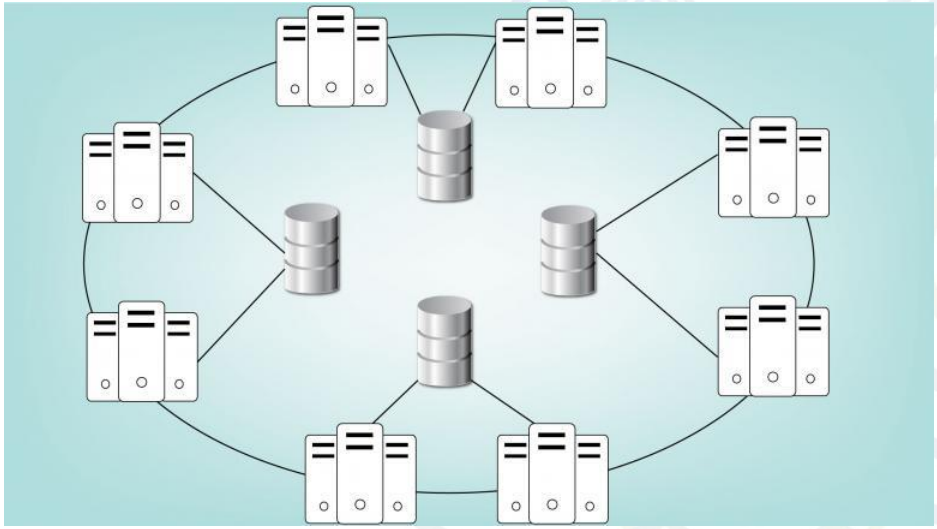
Single File system problems

- Not enough space
  - Put in different places
  - Get a bigger space
- Many different types
  - one type of databases/computer might not be enough
- How to find / process data?

Mahidol University
Wisdom of the Land

Distributed File System:Definition

- Data set can be distributed in many nodes
- The analysis of the data can be moved to those nodes
- The can help replicate the data between racks/ regions



Mahidol University
Wisdom of the Land

Distributed System:Benefit

- Fault tolerant
- Scaling the access to many users (Scalability)
- With highly parallelized replications, each reader can have its own nodes
- Increase performance

Mahidol University
Wisdom of the Land

Distributed System:Difficulty

- High concurrency → Low consistency
- Takes time to update
- More suitable for longer term storage
- Most big data is for data appending

Mahidol University
Wisdom of the Land

Parallel computers

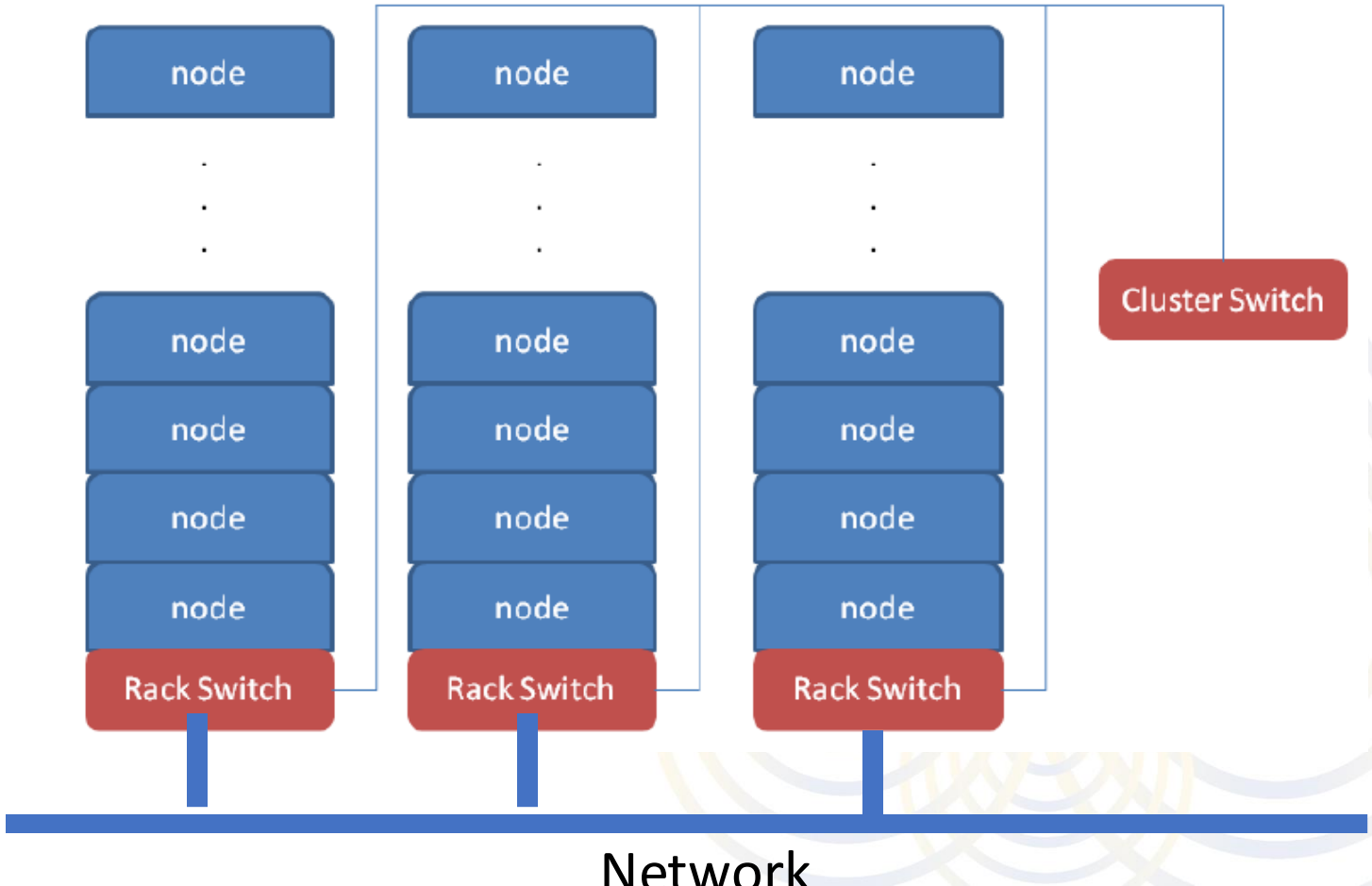
- Parallel computer
  - Big
  - Many cores
  - Expensive
  - Most scientenfic calculations

- Commodity cluster
  - Average number of computing nodes
  - Affordable
  - Less-specialization

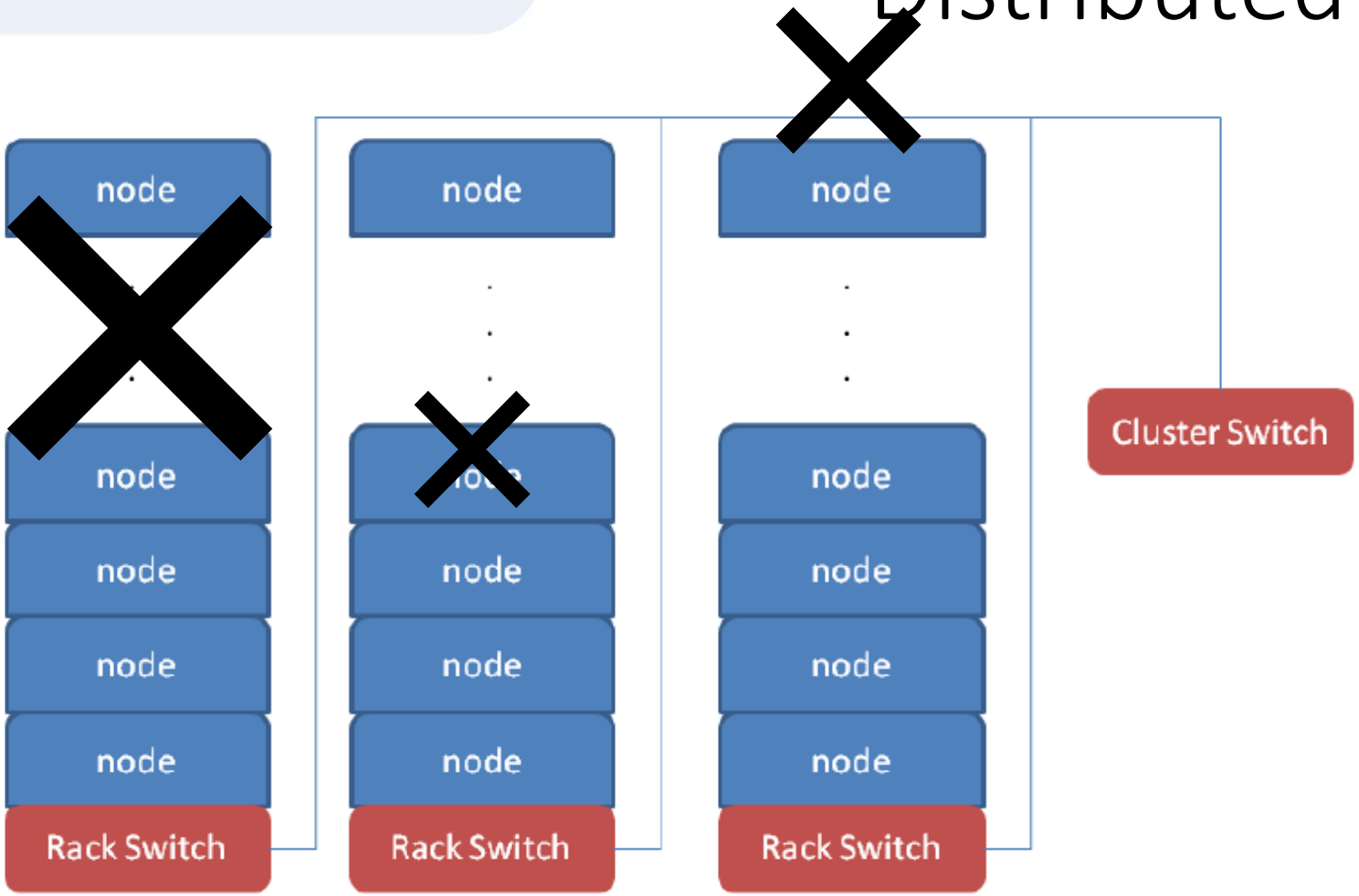
Mahidol University
Wisdom of the Land

Commodity Cluster

- Mostly used for “Distributed computing”
- Reduce cost
- Create ‘job-level parallelism / data parallelism



Distributed computing



Distribute over the internet called **'Data Parallelism'**

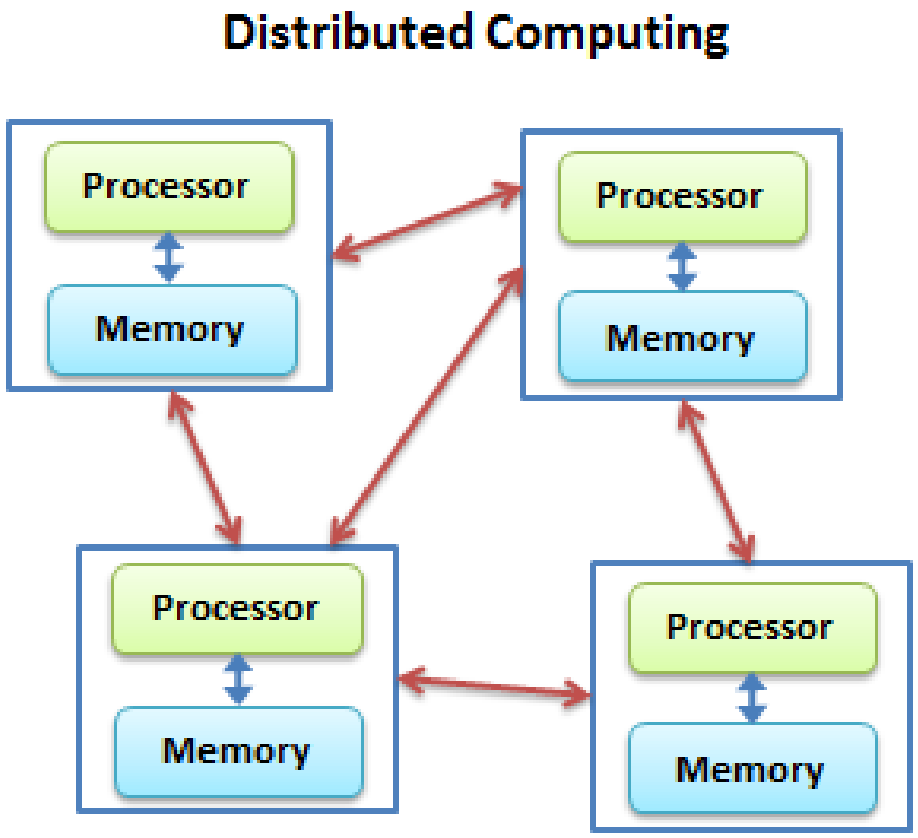
Solutions

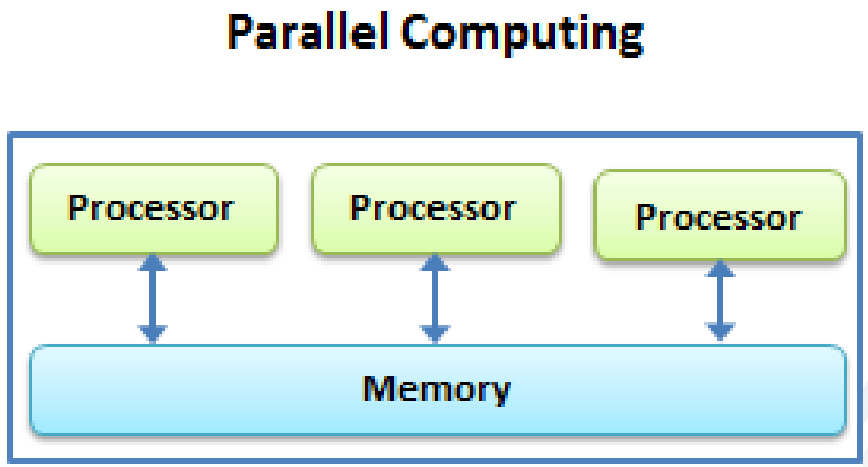
'Data redundant storage'

'Data parallel job restart'

Mahidol University
Wisdom of the Land

Paralle File System





Mahidol University
Wisdom of the Land

Advantages of distributed system

- Reliability and availability
- Improved performance
- Query processing time reduced
- Ease of growth/scale
- Continuing reliability

Mahidol University
Wisdom of the Land

What is Hadoop?

- Set of open-source program and procedures
- Used for processing large amount of data
- Hadoop cluster is a collection of computer working together to perform tasks
- Ecosystem not database

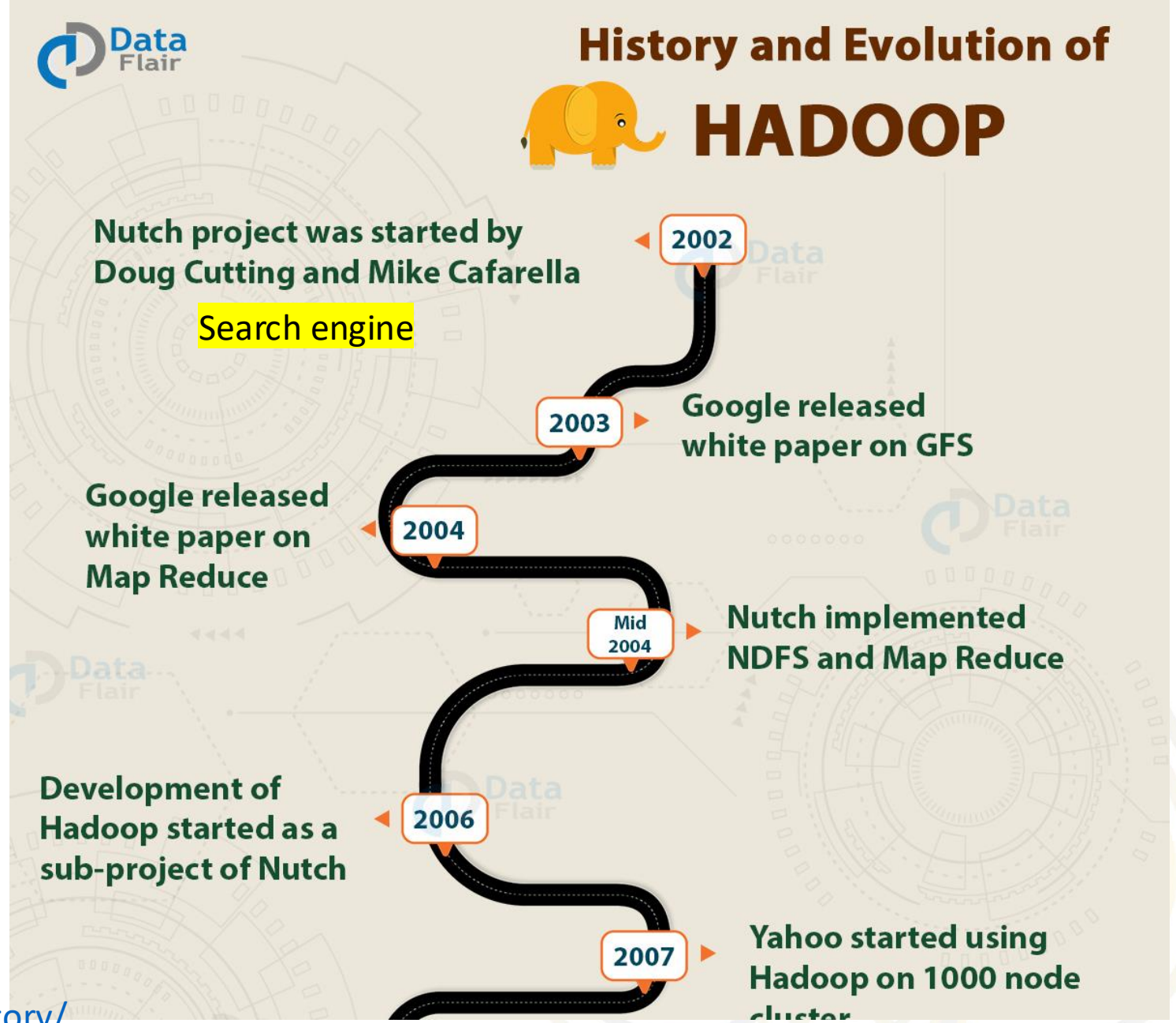
Mahidol University
Wisdom of the Land

Hadoop

- Model processing large data
- Split complications into different parallel tasks
- Make efficient use of large commodity clusters and distributed file systems.
- Abstract out the details of parallelization
- Full tolerance, data distribution, monitoring and load balancing.

Mahidol University
Wisdom of the Land

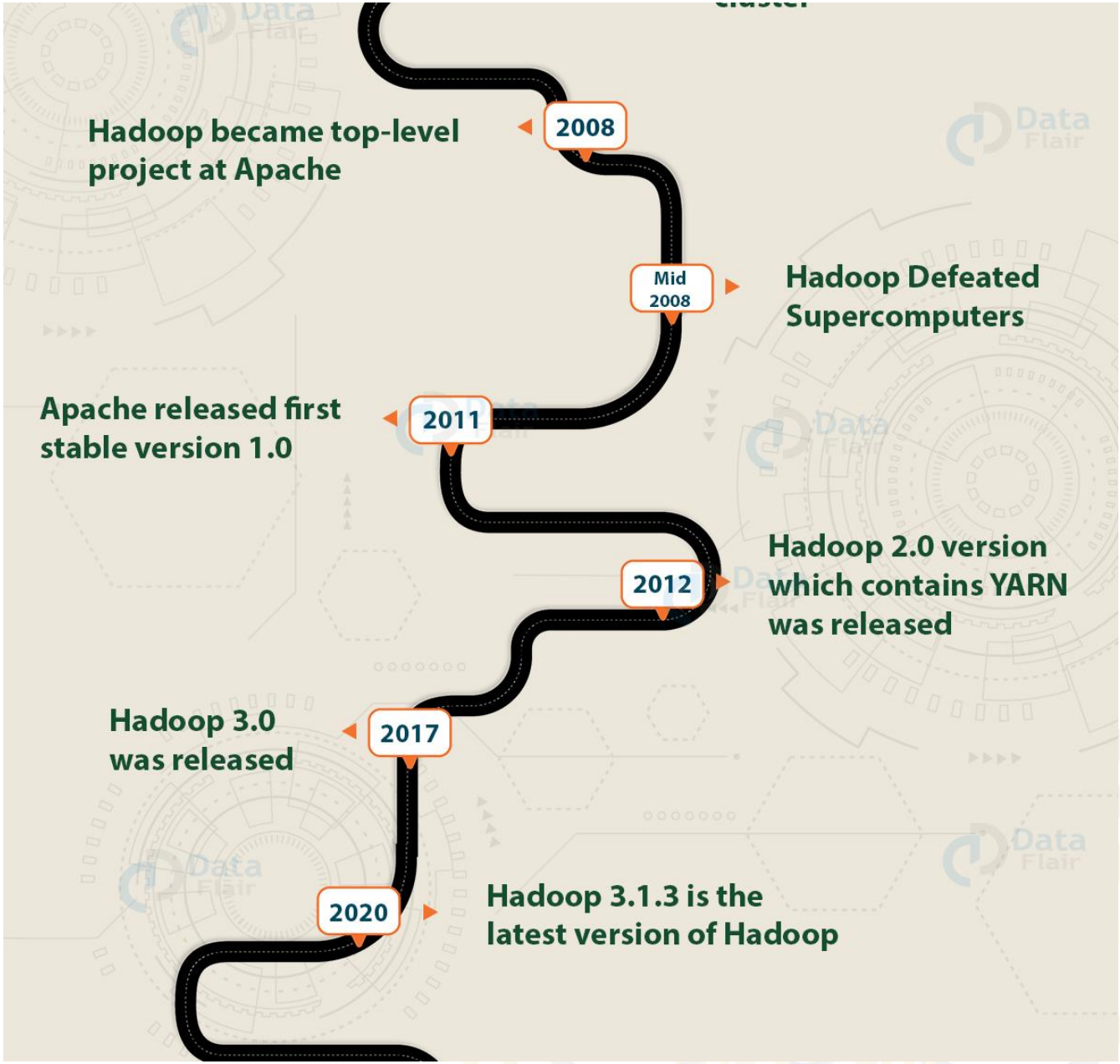
Hadoop History



ref: https://data-flair.training/blogs/hadoop-history/

Mahidol University
Wisdom of the Land

Hadoop history



Mahidol University
Wisdom of the Land

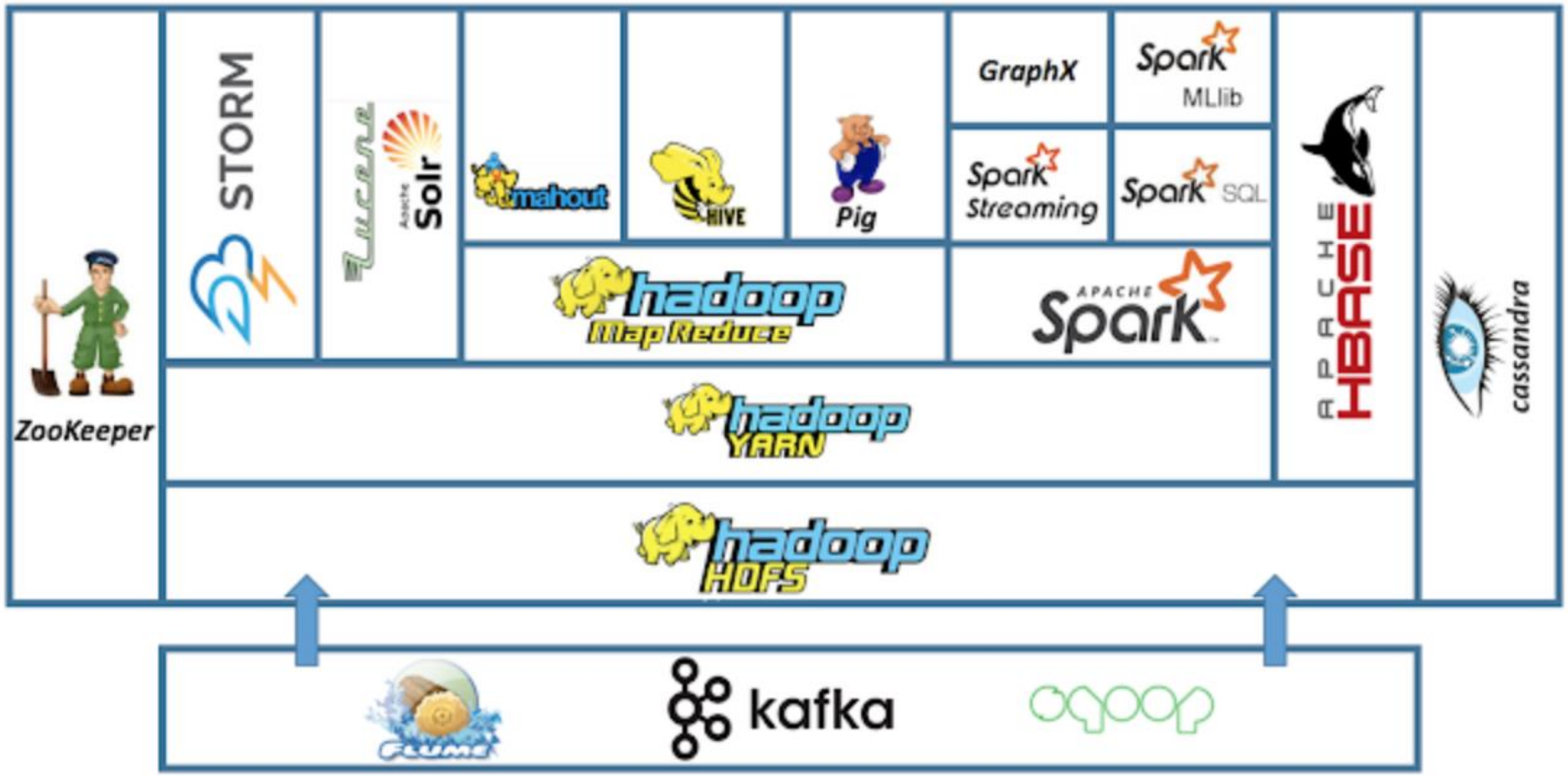
Hadoop

Hadoop Ecosystem

- Enable Scalability
- Handle Fault Tolerance
- Optimized for a variety data types: ’structured’, semi-structured and ‘unstructured data’
- Faciliate a shared environment

Mahidol University
Wisdom of the Land

Hadoop Ecosystem



Mahidol University
Wisdom of the Land