- Chapter 1 -
Big Data Foundations

Dr. Heba Ismail

The slides have been slightly updated by Prof. Laiali. Almazaydeh

"The future belongs to those who rule the data"

Agenda

- Understanding the Basics
  - Why Big Data?
  - Big Data Characteristics
  - Big Data explosion
- Big Data Application Domains
  - The Promise of Big Data
  - Applications for Big Data analytics
    - Science
    - Internet
    - Intelligent Transportation Systems
- Big Data value chain
  - Storage
  - Processing
  - Analytics
  - Visualization
- Big data opportunities for Developing Countries
  - Financial
  - Healthcare
  - Education
  - Agriculture
  - Others
- Big IT Infrastructure for Big Data
  - Big Data Landscape
  - Value of Big Data Analytics
  - Big Data will transform IT infrastructure
- Big Data Challenges and Open Research Questions
- Conclusion
- References

Understanding the Basics

New Age of Big Data
How it is generated

- The world has gone mobile
  - 5 billion cellphones produce daily data
- Social Media

  - Facebook generates **4 petabytes** of data per day which is around million gigabytes

  - Around **65,000 photos** are shared on Instagram every minutes

  - Twitter produces **200M** tweets a day
- Emails

  - Around **300 billion** emails are sent every day
- Transactions in Banking sector

  - More than **100K** transactions are done per second
- E-Commerce

  - Amazon, … and all ecommerce websites generates large volume of data per day from which users buying trends can be traced.

What comes next?

- Kilobyte (KB) – 10^3 bytes
- Megabyte (MB) –10^6 bytes
- Gigabyte (GB) – 10^9 bytes
- Terabyte (TB) –10^{12} bytes
- Petabyte (PB) – 10^{15} bytes
- Exabyte (EB) – 10^{18} bytes
- Zettabyte (ZB) – 10^{21} bytes
- Yottabyte (YB) – 10^{24} bytes

The Model Has Changed…

- **The Model of Generating/Consuming Data has Changed**

Old Model: Few companies are generating data, all others are consuming data

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

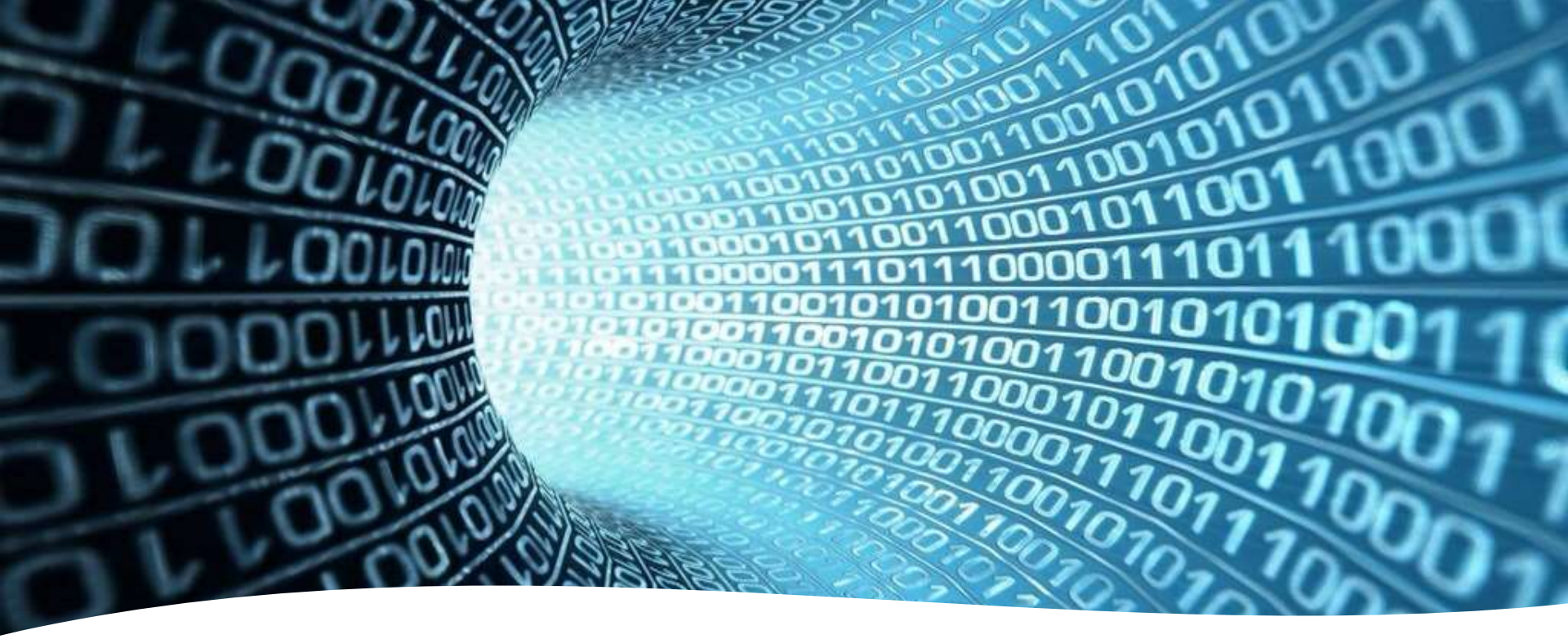
Why are they collecting all this data?

Target Marketing

- To send you catalogs for exactly the merchandise you typically purchase.
- To suggest medications that precisely match your medical history.
- To "push" television channels to your set instead of your "pulling" them in.
- To send advertisements on those channels **just for you**!

Targeted Information

- To know what you need before you even know you need it based on past navigation!
- To notify you of your expiring driver's license or credit cards or last refill on a Rx, etc. **Do you use TAMM Applications?**
- To give you turn-by-turn directions to a shelter in case of emergency. **Have you ever received emergency alert for AD Police?**



How Can You Avoid Big Data?

- Pay cash for everything!
- Never go online!
- Don’t use a telephone!
- Don’t fill any prescriptions!
- Do not use any device!
- Never leave your house!

LIFE HAS CHANGED – BIG DATA HAS BECOME PART OF OUR LIFE → LET'S EMBRACE IT!

Can you do it?

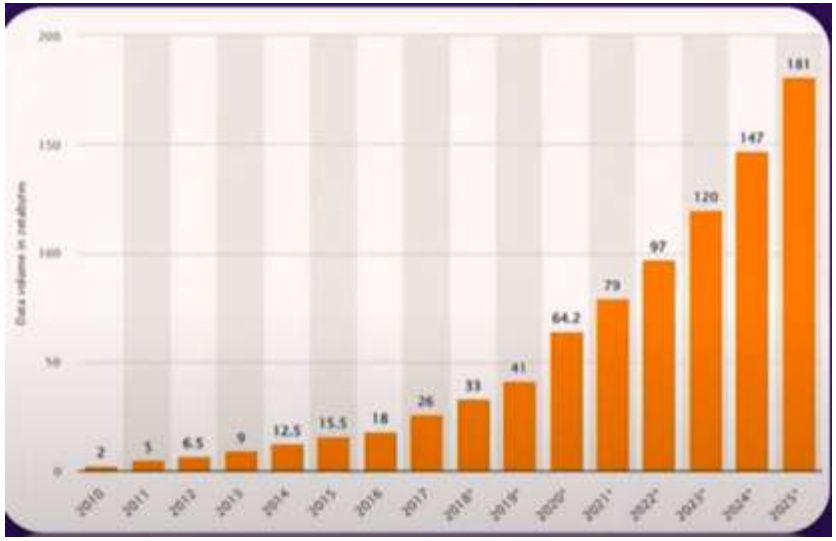
Big Data Definition

Big Data Characteristics (the 5 v's)

Characteristics of Big data are given by 5V's as follows:

1. Volume

- Related to enormous size
- Data generates from different sources such as social media platforms, ecommerce, business processes and many more



2. Variety

- Data collected from different sources can be structured, unstructured or semi-structured.
  - Structured data is the data that is ready for modeling and also for analyzing
  - Unstructured data is very much scattered data which cannot be straight away used for analyzing or reporting
  - Semi-structured data lies in between structured and unstructured

- Data collected is of different formats as follows:

✓ Text files (PDFs, docs, PPTs, etc.)
✓ Audio files (MP3, WAV, DAT, etc)
✓ Video files (MP4, MOV, AVI, MKV, etc.)
✓ Image files (PNG, JPG, TIFF, PCX, etc.)
✓ Emails (with different attachments)

3. Value

- It is the benefit that the organization may get from the data

- For any data that is considered to be valuable, following questions will have a positive answer:

✓ Will the data be helpful to achieve company's aim?

✓ Will it help to magnify the growth of the company?

4. Velocity

- Velocity refers to the rate of speed with which data is generated
- Data generally flows from different sources such as social media
- This determines the potential of data that how fast and continuously the data is extracted and then processed to satisfy the demands of the organization.
- Example: Number of clicks and action of users active on an ecommerce website are traced constantly and rapidly.

5. Veracity

- This refers to the reliability and correctness of the data
- Usually while data is captured many uncertainty and inconsistencies are found in it.
- Because of this processing becomes difficult
- Hence, Veracity determines the quality of data

Big Data Types

1. Structured data
- Well organized
- Can be stored, processed and accessed in a fixed format
  - Call detail records
  - Point of sale records
  - Claims data

2. Unstructured data
- Lacks a specific format or structure for storage and handling
- Difficult to process and analyze
  - Video, Audio,
  - Images, Text

3. Semi-structured data
- Combination of structured and unstructured big data
- Have a variable key-value pair structure
- Doesn't require structured query language hence called as NoSQL data
  - Web logs
  - Sensor data
  - Email, Twitter
  - JSON (JavaScript Object Notation)
  - XML (eXtensible Markup Language)

Big Data Types

Unstructured data

The university has 5600 students.
John's ID is number 1, he is 18 years old and already holds a B.Sc. degree.
David's ID is number 2, he is 31 years old and holds a Ph.D. degree. Robert's ID is number 3, he is 51 years old and also holds the same degree as David, a Ph.D. degree.

Semi-structured data

```
<University>
 <Student ID="1">
  <Name>John</Name>
  <Age>18</Age>
  <Degree>B.Sc.</Degree>
 </Student>
 <Student ID="2">
  <Name>David</Name>
  <Age>31</Age>
  <Degree>Ph.D. </Degree>
 </Student>
 ....
</University>
```

Structured data

ID	Name	Age	Degree
1	John	18	B.Sc.
2	David	31	Ph.D.
3	Robert	51	Ph.D.
4	Rick	26	M.Sc.
5	Michael	19	B.Sc.

Where do we see “Big Data”?

Application of Big Data

1. Entertainment

Industries collect feedback and track the activity of users on the content and analyze it to:

- Create more content for the target audience
- Recommend some trending content
- Gain more profit by focusing on the popular shows

Companies like Netflix are using big data terminologies

2. Education

- Track the interests of student
- Track the time spent by student on different pages
- Recommend courses based on the interest
- Teacher's performance can be traced
- Student's and teacher's feedback can be analyzed to take further decisions on updating the system or flow

Application of Big Data

3. Healthcare

- Helps doctors to recommend the best solutions by old patient's feedback
- Track the improvements of the patient's health
- Easily record the spreading rate of chronic disease
- Patient can easily find the nearest and the best reviewed hospitals

4. IoT

- Sensors installed on specific location captures data
- This data is analyzed, visualized and after this patterns and hidden insights are traced
- IoT sensors can help to collect big data from the following sector:
  - Agriculture
  - Air Quality
  - Traffic, etc

Application of Big Data

5. Transportation

- Traffic control
- Best and shortest route planning
- Intelligent systems for transportation
- Travel arrangements for tourists

6. Customer shopping behavior

- To track frequently purchased products and it’s brand
- Help seller to readily manufacture subsequent quantity of products based on the purchasing trends
- Helps to provide different offers on items to attract customers
- Ultimately helps to avoid the customer churn

Application of Big Data

7. Self-driving vehicles

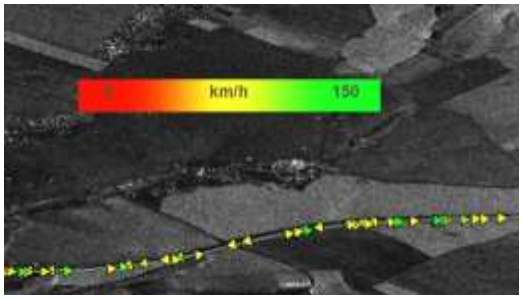
- Big data helps vehicles to drive without human intervention
- Sensors installed on different locations of vehicles captures surrounding data information
- Example: How much angle to rotate, when to slow down or when to stop

8. Energy

- Smart meter readers collect data at definite intervals
- This data is used to analyze the consumption of utilities better
- Optimizes the power generation and planning
- Ultimately helps in economy

Big Data – Intelligent Transportation Systems

The future lies in integration, mining and analytics of BIG DATA

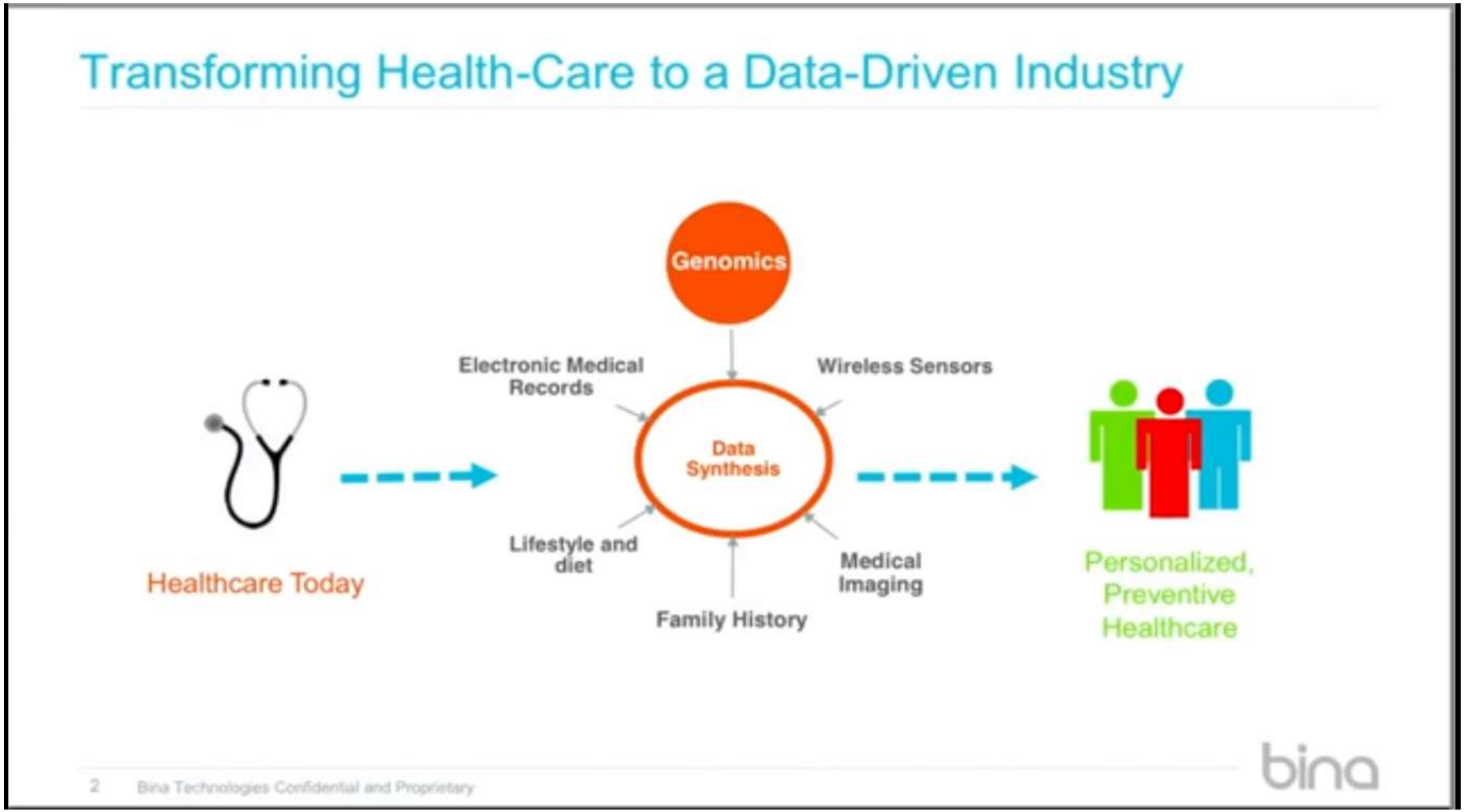


From the sky or space

From the ground

From the vehicles

Big Data – Healthcare, transforming Healthcare to a Data-Driven Industry

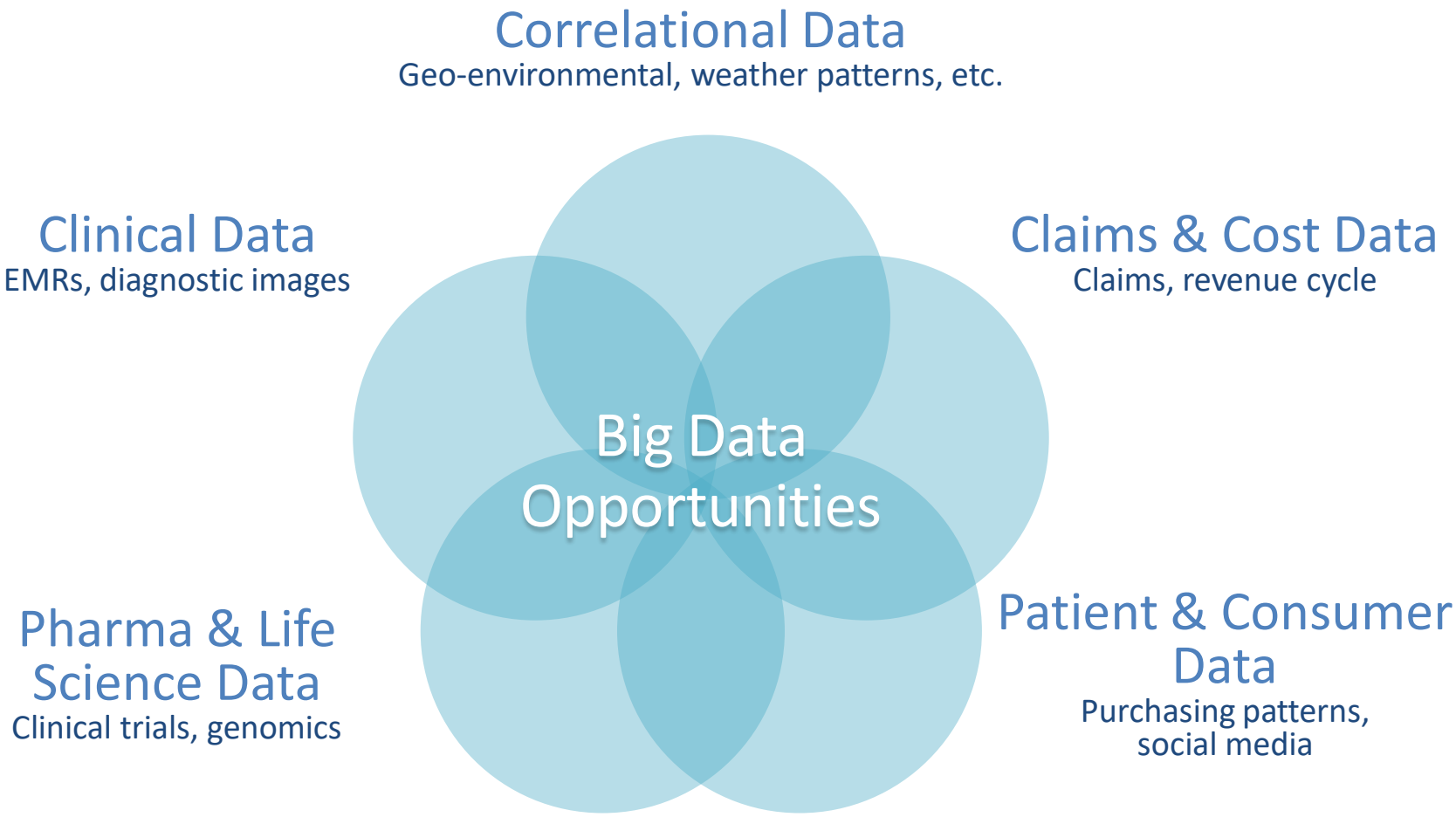


Big Data Opportunities: Healthcare Example

Big Data – Opportunities

- Big Data presents unprecedented opportunities to
  - Accelerate scientific discovery and innovation
  - Lead to new fields of inquiry that would not otherwise be possible
  - Improve decision making
  - Understand human and social processes
  - Promote economic growth
  - Improve health and quality of life

Big Data opportunities in Health

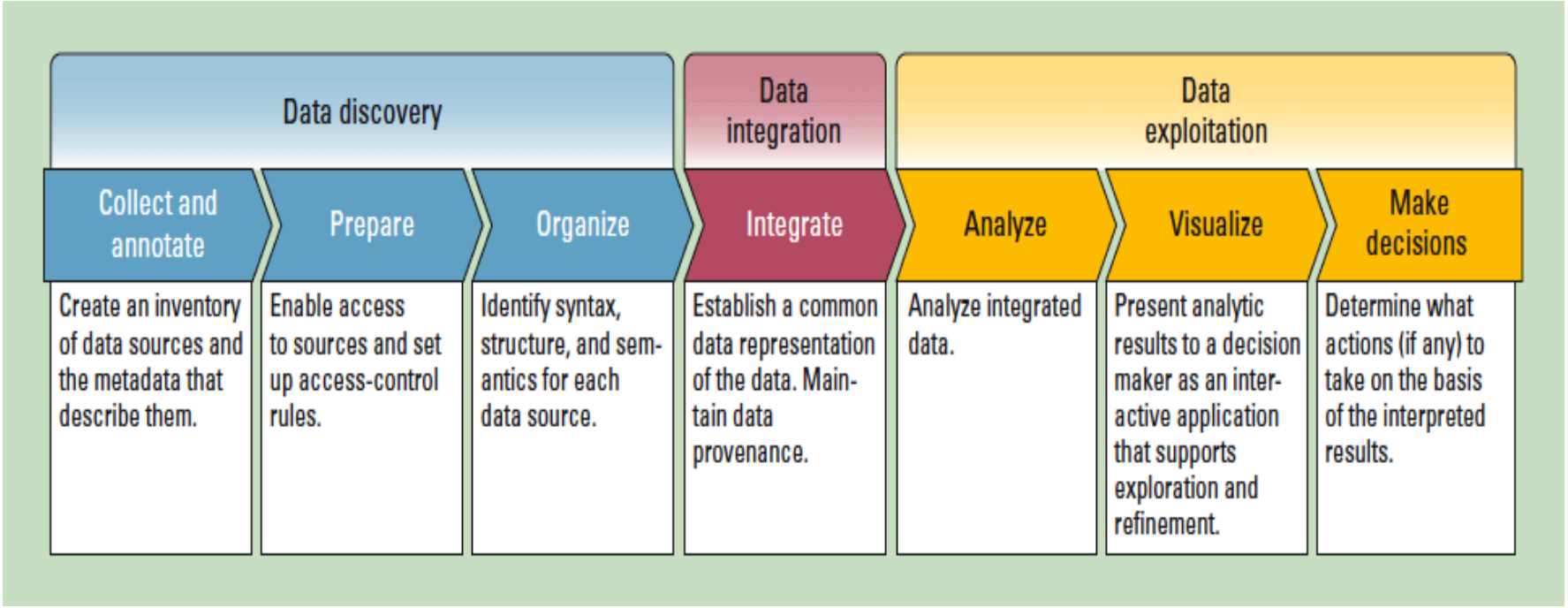


Big Data Lifecycle/Value Chain

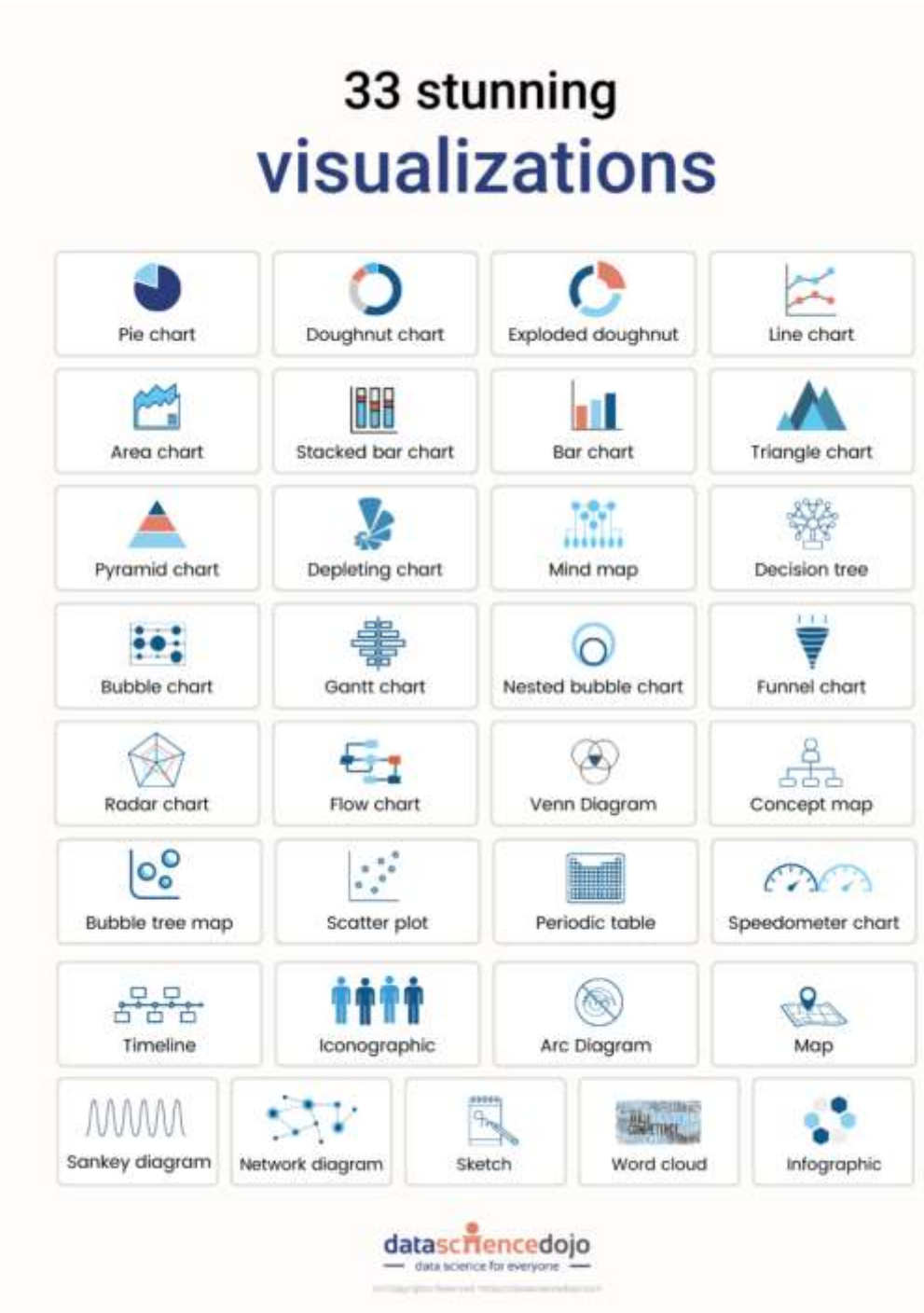
Big Data Value Chain

The Big Data value chain describes all the activities that **create value** from Big Data

Lets check this dataset out



- Many visualization options
- Each visual can show a different aspect about the dataset



Why Visualize Data?

Lets explore the visualization of this Netflix Dataset: link

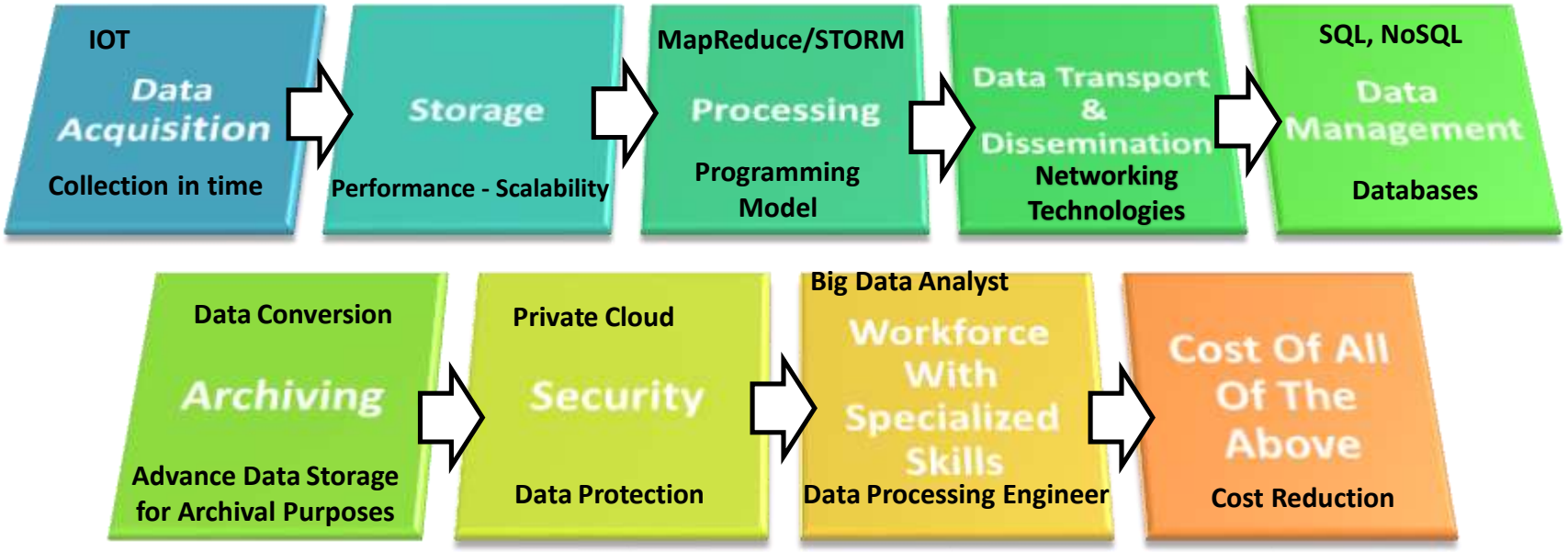
- Can you, by looking at the data, tell the percentage or count TV shows or movies?

- Can you tell by looking at the visual?

Day Activities Visualization

Big Data Challenges

Top 9 challenges About Big Data



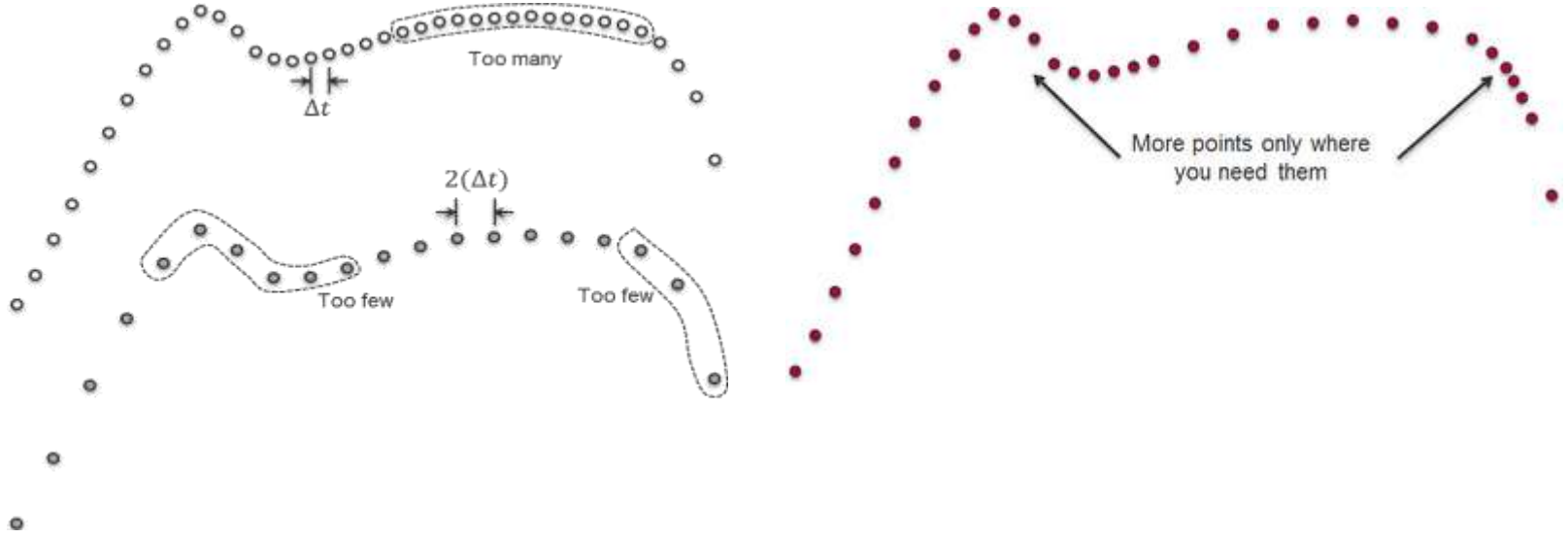
Challenge: Collection

Where does the data come from?

- Input from humans, instruments/sensors, existing datasets, etc.
- Potentially many sources
- Transport data from source to repository

Data Acquisition

- Data acquisition is the process of sampling signals that measure real world physical conditions and converting the resulting samples into digital numeric values that can be manipulated by a computer – Wiki
- The question is how can you acquire closely spaced data only when you need so that the data file doesn't become too large with unnecessary data points.



Challenge: Organization

How is the data structured?

- Data needs to be labeled, sorted, etc.
- Relationships may exist between pieces
- Exclude inaccurate or unknown data

Challenge: Storage

How do we store large volumes of data?

- Need space for 100s of Terabytes of data (modern hard drive holds 1 TB)
- Data needs to be *efficiently* accessed by servers doing computation

Storage

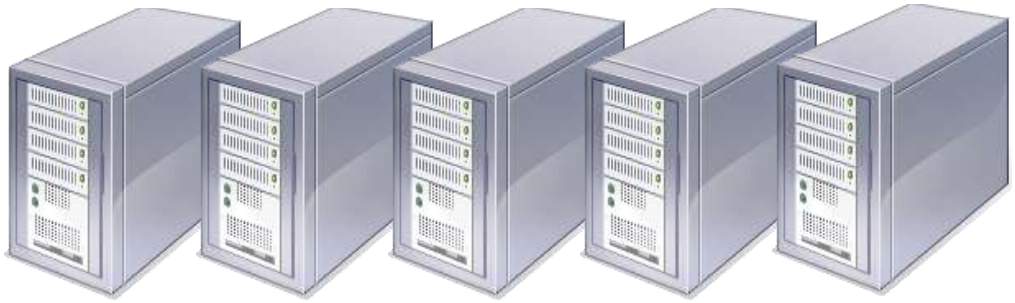
Big Data storage for active repositories should meet the following requirements:

- High performance (very low latency)
  - Average reads < 5ms, writes < 10ms
- Seamless scalability
  - No table or throughput limits
  - Live repartitioning (no downtime)
- High durability (availability)
- Predictable performance

Challenge: Computation

How is the data processed to obtain desired information?

- Algorithms determine actions to perform
- Need computers to run the algorithms
- May be constrained by time, space, etc.



Processing

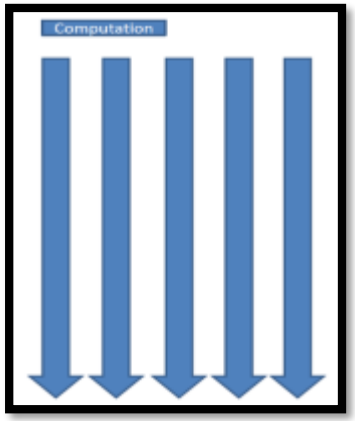
"We have terabytes of click-stream data, what can we do with it?"

- Very large data repositories
- Complex data analysis
- Distributed and parallel data processing

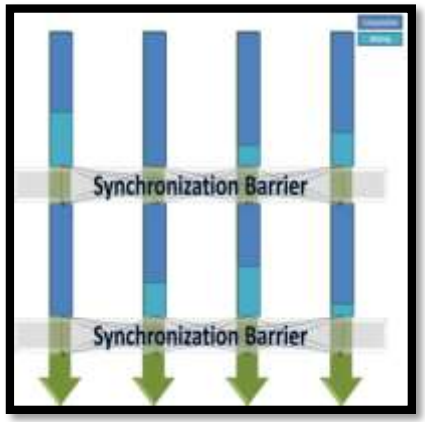
Processing

Computation patterns in parallel data processing

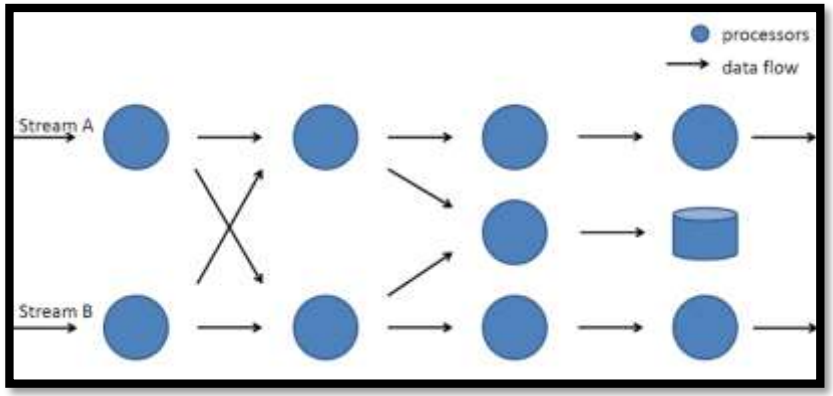
Embarrassingly parallel



Loosely synchronous



Stream processing



Suitable programming models

MapReduce /Bulk Synchronous Processing **STORM**

- **Do we need new programming models?**

Not likely, the existing ones are adequate. The challenge is to make them more efficient.

Data management

- The question is how can we store, organize, and query our data.
  - High performance
  - High scalability
  - High availability

Challenge: Visualization

How is the data (or results) presented?

- Seek clear, concise representation of the data
- Emphasize desired information
- May require many related visualizations

Security

- Data confidentiality
- Data Integrity
- Data Accuracy

Workforce with specialized skills

- A need for more specialized and highly skilled workforce to help us deal with Big data.
  - A new Job Title: Data scientist, Data Consultant, etc…

Class Activity Cont.

- Based on the explanation related to Big Data in Healthcare, identify the following with your teammate:
  - Big Data Challenges in the domain of healthcare.
  - It will be nice to support your inputs with some references, statistics, screenshots, links, … etc. (if applicable)

- Use your class Jamboard to sketch your idea and discuss them with your classmates ☺
- https://jamboard.google.com/d/16XMlUtVsWS-kDLUbKihFxU6VfZcULsm5Kq70QBAdYPg/edit?usp=sharing

Conclusion

- Big data is a very promising research and development area.
- Many applications domains are generating Big data and requires storage, analytics, and quality evaluation.
- The explosion in data is creating challenges and prompting innovation in computer storage and processing, in terms of software, hardware and data center architecture.
- Big data is extremely important in terms of social welfare, productivity, and competiveness.
- The success of Big Data requires Cloud infrastructure, large scale database and distributed file system, and advanced data analytics.
- Many research challenges were not addressed yet.
- Research on big data is in it early stage.

References

[1]"Visual Networking Index (VNI) - VNI Forecast Highlights", *Cisco*, 2016. [Online]. http://www.cisco.com/c/en/us/solutions/service-provider/visual-networking-index-vni/vni-forecast.html.

[2]"A sea of sensors", *The Economist*, 2010. [Online]. Available: http://www.economist.com/node/17388356

[3]"IBM big data platform - Bringing big data to the Enterprise", *Www-01.ibm.com*, 2016. [Online]. http://www-01.ibm.com/software/data/bigdata/.

[4]"HubbleSite - The Telescope - Hubble Essentials - Quick Facts", *Hubblesite.org*, 2016. [Online]. http://hubblesite.org/reference_desk/facts_.and._figures/quick_facts/quick_facts_2.shtml#data_stats.

[5]"Hortonworks : Open and Connected Data Platforms", *Hortonworks*, 2016. [Online]. http://hortonworks.com/.

[6]"Welcome to Apache™ Hadoop®!", *Hadoop.apache.org*, 2016. [Online]. Available: https://hadoop.apache.org/.

[7] ”Amazon Web Services ”,Amazon, 2016. [Online]. https://aws.amazon.com/.

[8]"What Is Big Data? - Blog", *Datascience.berkeley.edu*, 2014. [Online]. https://datascience.berkeley.edu/what-is-big-data/.

[9]"The Evidence Is In: People Want to Collaborate with Their Doctors and Co-Produce Their Clinical… — Tincture", *Medium*, 2016. [Online]. https://medium.com/tincture/the-evidence-is-in-people-want-to-collaborate-with-their-doctors-and-co-produce-their-clinical-8c02069ab965#.xbfm74tj5.

[10]"Strata + Hadoop World", *Conferences.oreilly.com*, 2016. [Online]. Available: http://conferences.oreilly.com/strata.

[11]"Do you know big data's top 9 challenges? -- Washington Technology", *Washingtontechnology.com*, 2013. [Online]. https://washingtontechnology.com/articles/2013/02/28/big-data-challenges.aspx.

[12]T. challenge, "Bluehill User Community: The data acquisition challenge", *Bluehillusercommunity.blogspot* [Online]. http://bluehillusercommunity.blogspot.tw/2011/07/data-ccquisition-challenge.html.

[13]2016. [Online]. http://www.ijcaonline.org/volume10/number7/pxc3872013.pdf.

[14]*Iza.org*, 2016. [Online]. http://www.iza.org/conference_files/eddi10/EDDI10_Presentations/EDDI10_P1_PeterWittenburg_Slides.ppt.

[15] The overview of The information technology Industry Chain in Big Data era: http://link.springer.com/chapter/10.1007%2F978-3-642-55038-6_66