1

04 January 2021 04:34 PM

Introduction to Big Data

21

Big Data: Definitions

- No single standard definition

According to *Gartner*: "Big data is an evolving term that describes any voluminous amount of structured, semistructured and unstructured data that has the potential to be mined for information" [1].

Wikipedia describe it as: "Big data an all-encompassing term for any collection of data sets so large and complex that it becomes difficult to process using on-hand data management tools or traditional data processing applications" [2].

Casari defines Big data as: "Big Data is data whose scale, diversity, and complexity require new architecture, techniques, algorithms, and analytics to manage it and extract value and hidden knowledge from it" [3].

According to *Forbes* Big data is new tools helping us find relevant data and analyze its implications [4].

structured
semi-S
U-S

SQL / RDBMS
NoSQL
Non-SQL

$O(N)$ $O(K)$

Screen clipping taken: 04-01-2021 09:49 PM

Definition of Big Data

- Big data is a collection of data sets so large and complex that it becomes difficult to process using on-hand database management tools or traditional data processing applications.

From wiki

22

Evolution of Big Data

- Birth: 1880 US census
- Adolescence: Big Science
- Modern Era: Big Business

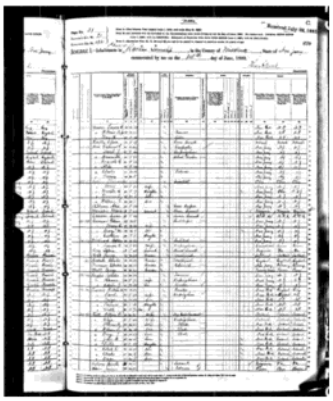
23

Birth: 1880 US census

24

The First Big Data Challenge

- 1880 census
- 50 million people
- Age, gender (sex), occupation, education level, no. of insane people in household

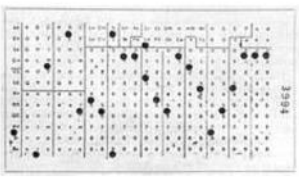


25

7 years.
↳ Report.
1887.

The First Big Data Solution

- Hollerith Tabulating System
- Punched cards – 80 variables
- Used for 1890 census
- 6 weeks instead of 7+ years



26

7 years

6 weeks.

Manhattan Project (1946 - 1949)

- $2 billion (approx. 26 billion in 2013)
- Catalyst for "Big Science"

27

1939.

1,30,000

Space Program (1960s)

- Began in late 1950s
- An active area of big data nowadays

28

Big Science vs. Big Business

- Common
  - Need technologies to work with data
  - Use algorithms to mine data
- Big Science
  - Source: experiments and research conducted in controlled environments
  - Goals: to answer questions, or prove theories
- Big Business
  - Source: transactions in nature and little control
  - Goals: to discover new opportunities, measure efficiencies, uncover relationships

29

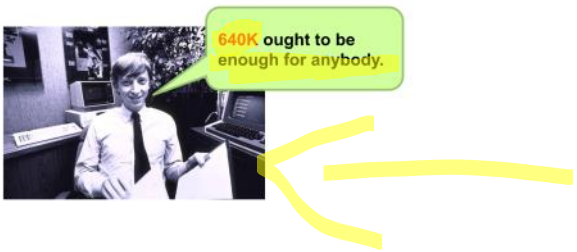
Big Data is Everywhere!

- Lots of data is being collected and warehoused
  - Web data, e-commerce
  - Purchases at department/ grocery stores
  - Bank/Credit Card transactions
  - Social Networks

30

How Big is Big ?

- 2008: Google processes 20 PetaByte per Day (Peta=10^{15})
- Apr 2009: Facebook has 2.5 PB user data + 15 TB/day
- May 2009: eBay has 6.5 PB user data + 50 TB/day
- 2011: Yahoo! Has 180-200 PB of data
- 2012: Facebook ingests 500TB/day



31

How many users and objects?

- Flickr has >6 billion photos
- Facebook has 1.15 billion active users
- Google is serving >1.2 billion queries/day on more than 27 billion items
- >2 billion videos/day watched on YouTube

2019

32

How much data?

- Modern applications use massive data:
  - Rendering 'Avatar' movie required >1 petabyte of storage
  - eBay has >6.5 petabytes of user data
  - CERN's LHC will produce about 15 petabytes of data per year
  - In 2008, Google processed 20 petabytes per day
  - German Climate computing center dimensioned for 60 petabytes of climate data
  - Someone estimated in 2013 that Google had 10 exabytes on disk and ~ 5 exabytes on tape backup
  - NSA Utah Data Center is said to have 5 zettabyte (!)
- How much is a zettabyte?
  - 1,000,000,000,000,000,000,000 bytes
  - A stack of 1TB hard disks that is 25,400 km high

33

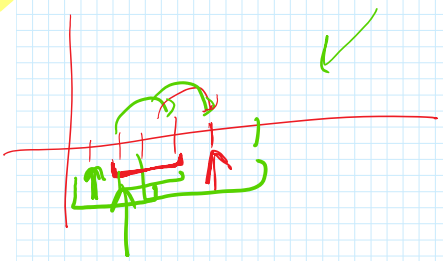
Characteristics of Big Data: 4V

Volume	Velocity	Variety	Veracity
From terabytes to exabyte to zetabytes of existing data to process	Batch data, real-time data, streaming data, milliseconds to seconds to respond	Structured, semi-structured, unstructured, text, pictures, multimedia	Uncertainty due to data inconsistency & incompleteness, ambiguities, deception, model approximation
IDC Analyze the Future	facebook		
8 billion TB in 2015, 40 ZB in 2020 5.2TB per person	New sharing over 2.5 billion per day new data over 500TB per day		

34

9V
13V
17V
$\boxed{41V}$

Figure 1: Characteristics of Big data



EB
PB
ZB.

How much computation?

- No single computer can process that much data
  - Need many computers!
- How many computers do

PB
2B.

How much computation?

- No single computer can process that much data
  - Need many computers!
- How many computers do modern services need?

  - Facebook is thought to have more than 60,000 servers
  - 1&1 Internet has over 70,000 servers
  - Akamai has 95,000 servers in 71 countries
  - Intel has ~100,000 servers in 97 data centers
  - Microsoft reportedly had at least 200,000 servers in 2008
  - Google is thought to have more than 1 million servers, is planning for 10 million (according to Jeff Dean)

35

What to do with More Data ?

- Answering factoid questions
  - Pattern matching on the Web
  - Works amazingly well

    Who shot Abraham Lincoln? --> ??? shot Abraham Lincoln
- Learning relations
  - Start with seed instances
  - Search for patterns on the Web
  - Using patterns to find more instances

Wolfgang Amadeus Mozart (1756 - 1791)
Einstein was born in 1879

Birthday-of(Mozart, 1756)
Birthday-of(Einstein, 1879)

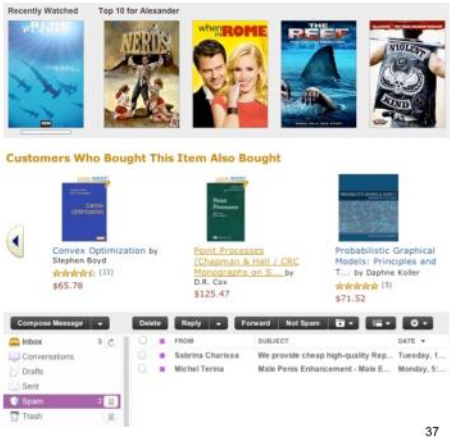
PERSON (DATE –
PERSON was born in DATE

(Brill et al., TREC 2001; Lin, ACM TOIS 2007)
(Agichtein and Gravano, DL 2000; Ravichandran and Hovy, ACL 2002; ...)

36

What to do with More Data ? (cont'd)
Personalization

- 100-1000M users
  - Spam filtering
  - Personalized targeting & collaborative filtering
  - News recommendation
  - Advertising



37

Big Data Analytics

DM

- Definition: A process of inspecting, cleaning, transforming, and modeling big data with the goal of discovering useful information, suggesting conclusions, and supporting decision making
- Hot in both industrial and research societies

Google | Find us on Facebook | Baidu 百度 www.baidu.com | YAHOO! | Business Partner IBM | amazon web services Partner Network | bing

38

Types of Analytics at eBay

- Basically measure anything possible - A **few** examples:

Marketing	Buyer Experience	Finance	Trust & Safety
Technology Operations	Customer Service	Loyalty	Information Security
Infrastructure	Finding	User Behavior	Seller Experience

40