COMP9313: Big Data Management

Introduction to Big Data Management

What is big data?

Dan Ariely

January 6, 2013 at 6:17pm ·

Big data is like teenage sex: everyone talks about it, nobody really knows how to do it, everyone thinks everyone else is doing it, so everyone claims they are doing it...

Like Comment Share

Tweeted by Prof. Dan Ariely, Duke University

What is big data?

- No standard definition!
- Wikipedia:
  - Big data is a field that treats ways to analyze, systematically extract information from, or otherwise deal with data sets that are too large or complex to be dealt with by traditional data-processing application software.
- Amazon:
  - Big data can be described in terms of data management challenges that – due to increasing volume, velocity and variety of data – cannot be solved with traditional databases.

What is big data?

Word could which is generated from the top-20 results when search “what is big data” in Google.

What is big data?

- A set of data
- Special characteristics
  - Volume 体积
  - Variety 品种
  - Velocity 速度
  - …
- Traditional methods cannot manage
  - Store
  - Analyse
  - Retrieve 检索?
  - Visualization 可视化
  - …

That’s why we need this course

Big Data Definitions Have Evolved Rapidly

- 3 V's
  - In a research report by Doug Laney in 2001
  - Volume, Velocity and Variety 体积，速度和变化
- 4 V's
  - In Hadoop – big data tutorial, 2006
  - Veracity 准确性
- 5 V's
  - Around 2014
  - Value
- 7 V's, 8 V's, 10 V's, 17 V's, 42 V's, …

Major Characteristics of Big Data

体积

Volume

多样性

Variety

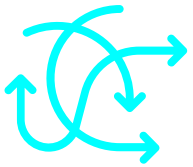
准确性

Veracity

Big Data

速率

Velocity



可变性；易变性

Variability

价值

Value

Visibility

可视性

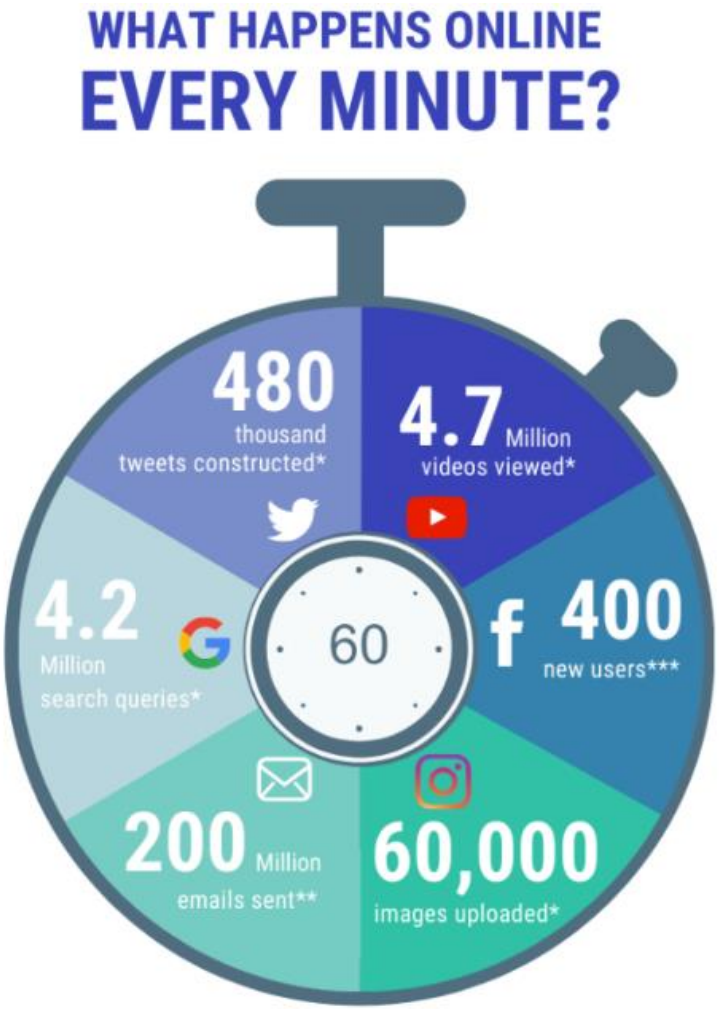
Volume (Scale)

- Quantity of data being created from all sources
- The fundamental of big data

大数据的基础

- 18 Zetabytes (ZB) of data in 2018, will grow to 175 ZB in 2025
  - 1 zettabyte ≈ 10^3 exabytes ≈ 10^9 terabytes
  - Source: https://www.seagate.com/files/www-content/our-story/trends/files/idc-seagate-dataage-whitepaper.pdf

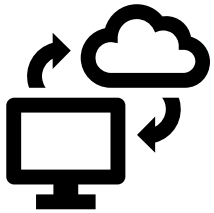
Volume



Source: https://www.nodegraph.se/how-much-data-is-on-the-internet/

Volume – Why Challenging?

Model	RAM	Disk	Data
Macintosh Classic (1990)	1MB – 4MB	0 – 40MB	
Power Mac G4 (2000)	256MB – 1.5GB	20GB – 60GB	5 EB in 2003
iMac (mid 2010)	4GB – 16GB	500GB – 2TB	1 ZB in 2012
iMac (early 2019)	8GB – 64GB	1TB – 3TB	~40 ZB



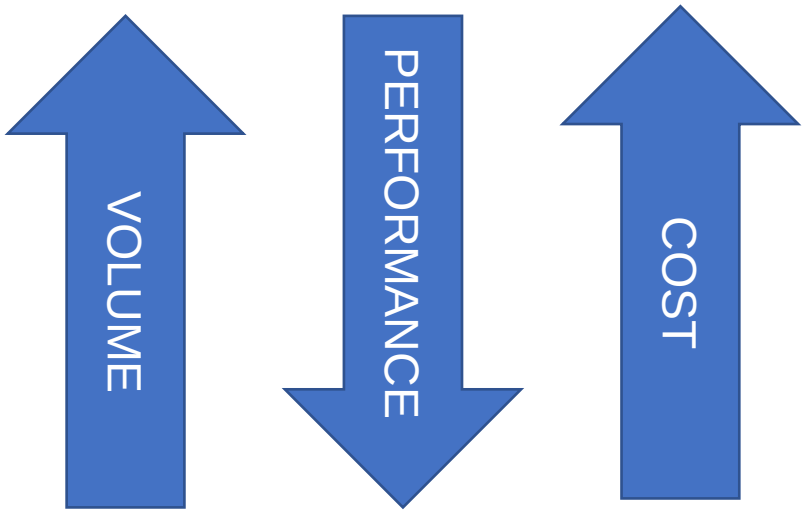
?

1990s 2000s 2010s future

DBMS Storage

Volume – Why challenging?

- Time complexity
  - Sort algorithms: O(N logN)
  - Merge join: O(N logN + M logM)
  - Shortest path: O(V logV + E logV)
  - Nearest neighbor search: O(dN)
  - NP hard problems



Variety (Diversity)

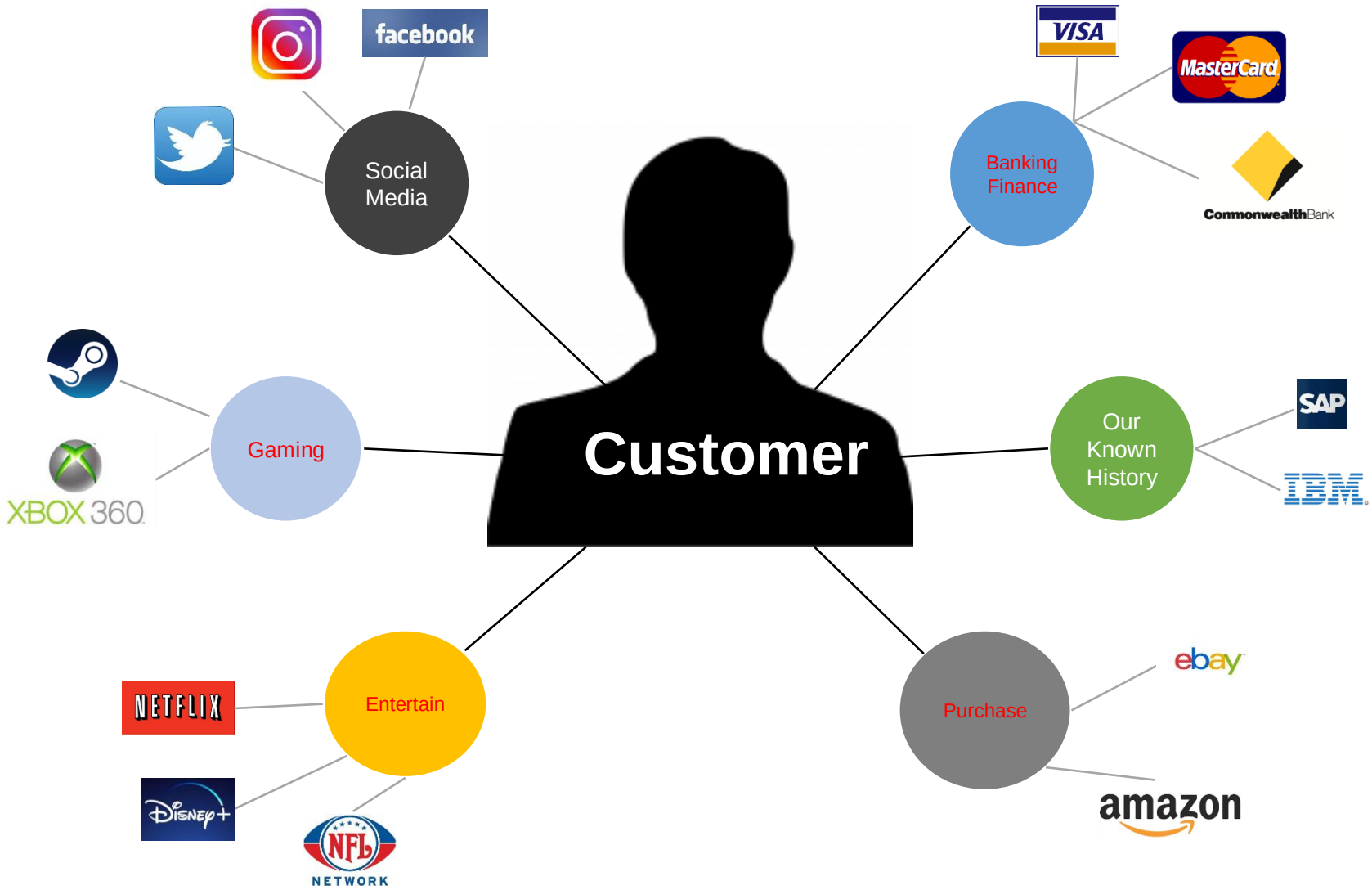
- Different Types
  - Relational data (tables/transactions)
  - Text data (books, reports)
  - Semi-structured data (JSON, XML)
  - Graph data (social network, RDF)
  - Image/video data (Instagram, Youtube)
- Different sources
  - Movie reviews from IMBD and Rotten Tomatoes
  - Product reviews from different provider websites
  - Personal information from different social apps

Variety

- A single application can be generating or collecting multiple types of data
  - Email
  - Webpage

- If we want to extract knowledge, then all the data with different types and sources need to be linked together

Variety - A Single View to the Customer



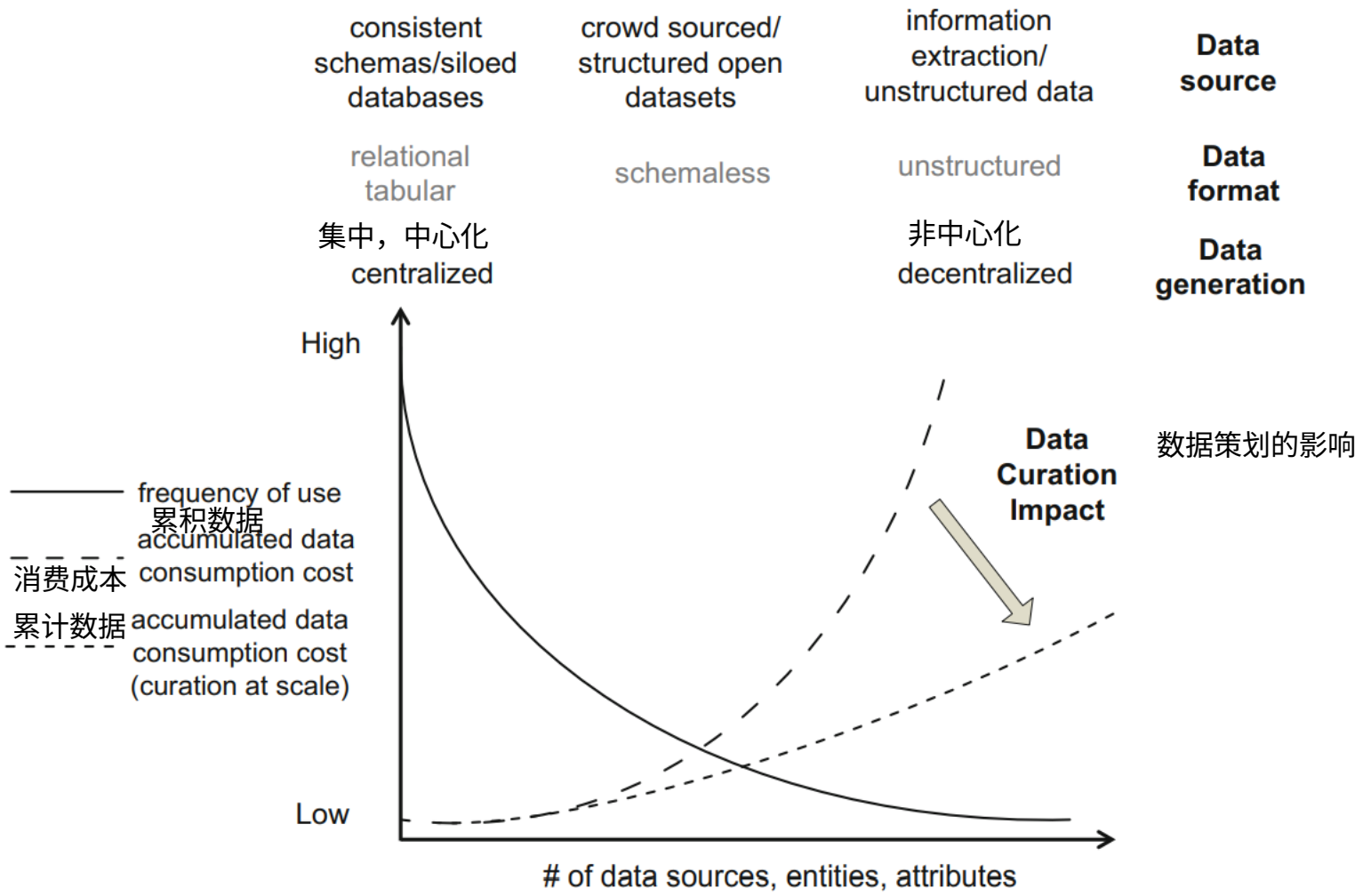
Variety – Why Challenging?

数据整合

- Data integration
  - Heterogeneous 由很多种类组成的；各种各样的
    - Traditional data integration relies on **schema mapping**, the difficulty and time complexity is directed related to the level of heterogenity and data sources 异质性和数据源
  - Record linkage in variety data 在品种数据中记录联动
    - needs to identify if two records refer to the same entity. How to make use of different types of data/information from different sources？ 需要确定两个记录是否引用同一实体。 如何利用不同来源的不同类型的数据/信息？
- Data curation 数据保管
  - Organization and integration of data collected from various sources
  - Long tail of data variety

数据多样性和数据整理的漫长尾巴

The Long Tail of Data Variety and Data Curation



Source: Curry, E., & Freitas, A. (2014). Coping with the long tail of data variety.

速度

Velocity (Speed)

- Data is being generated fast, thus need to be
  - stored fast
  - processed fast
  - analysed fast

- Every second
  - **8,991** Tweets sent
  - **994** Instagram photos uploaded
  - **4,683** Skype calls
  - **93,508** GB of Internet traffic
  - **83,165** Google searches
  - **2,915,385** Emails sent

Source: http://www.internetlivestats.com/one-second/

Velocity

- Reason of growth
  - Users:
    - 16 million in 1995 to 3.4 billion in 2016
  - IoT:
    - sensor devices, surveillance cameras
  - Cloud computing:
    - $26.4 billion in 2012 to $260.5 billion in 2020
  - Website:
    - 156 million in 2008 to 1.5 billion in 2019
  - Scientific data:
    - weather data, seismic data

高速

Velocity

现在，数据以连续的方式实时流传输到服务器，并且只有在延迟非常短的情况下，结果才有用

- Data is now streaming into the server in real time, in a continuous fashion and the result is only useful if the delay is very short.

- Many application need immediate response
  - Fraud detection
  - Healthcare monitoring
  - Walmart's real-time alerting

速度 速度

Velocity – Why Challenging?

- Batch processing

- Real time processing

- Transmission
  - Transferring data becomes a prominent issue in big data
  - Balancing latency/bandwidth and cost
  - Reliability of data transmission