CS315: Principles of Database Systems
Big Data

Arnab Bhattacharya
arnabb@cse.iitk.ac.in

Computer Science and Engineering,
Indian Institute of Technology, Kanpur
http://web.cse.iitk.ac.in/~cs315/

2nd semester, 2013-14

Tue, Fri 1530-1700 at CS101

Big Data

- What is big data?

Big Data

- What is big data?
- Data which is big
- How big is “big”?

Big Data

- What is big data?
- Data which is big
- How big is “big”?
- For sociologists, 10 subjects is big

Big Data

- What is big data?
- Data which is big
- How big is “big”?
- For sociologists, 10 subjects is big
- For social networks, terabytes is normal

Big Data

- What is big data?
- Data which is big
- How big is “big”?
- For sociologists, 10 subjects is big
- For social networks, terabytes is normal
- For astrophysicists, terrabytes is a day’s work

Big Data

- What is big data?
- Data which is big
- How big is “big”?
- For sociologists, 10 subjects is big
- For social networks, terabytes is normal
- For astrophysicists, terrabytes is a day’s work
- So, no absolute definition or threshold

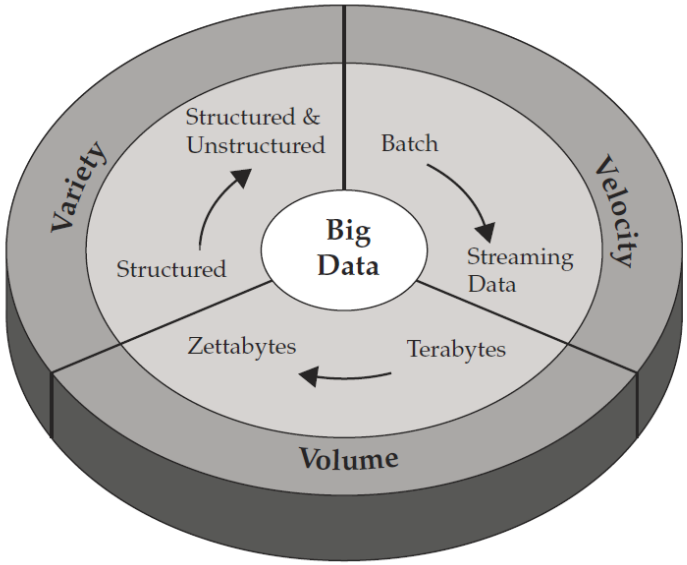
Big Data

- What is big data?
- Data which is big
- How big is “big”?
- For sociologists, 10 subjects is big
- For social networks, terabytes is normal
- For astrophysicists, terrabytes is a day’s work
- So, no absolute definition or threshold
- When data is bigger than most machines can store or most algorithms can handle

Characterization of big data

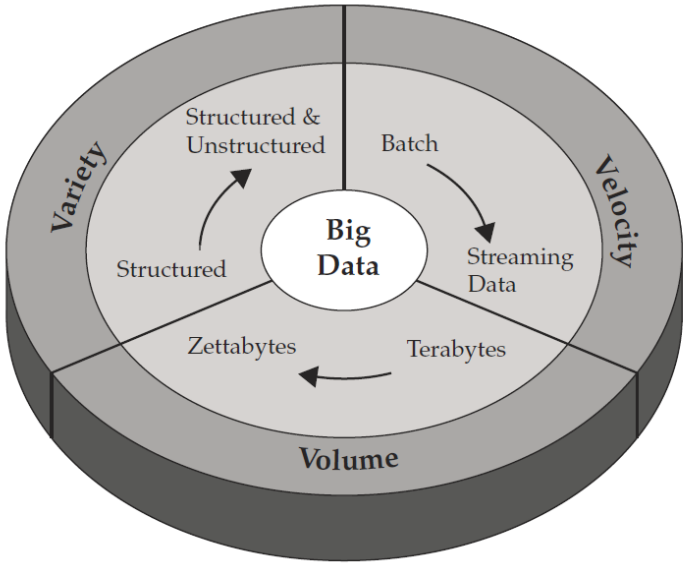
- Having large volumes of data requires
  - Newer techniques
  - Newer tools
  - Newer architectures
- Allows solving newer problems
  - Can also solve older problems better

Properties of big data



- 3 V's: volume, variety, velocity
- Volume: When data is extremely large in size, how to load it, index it or query it
- Variety: Data can be semi-structured or unstructured as well; how to query
- Velocity: Data can arrive at real time and can be streaming

Properties of big data



- 3 V's: volume, variety, velocity
- Volume: When data is extremely large in size, how to load it, index it or query it
- Variety: Data can be semi-structured or unstructured as well; how to query
- Velocity: Data can arrive at real time and can be streaming
- Extended V's: veracity, validity, visibility, variability

Enablers of big data

- Increased storage volume and type
- Increased processing power
- Increased data
- Increased network speed
- Increased capital
- Increased business

Too much data can lead to spurious results

- Is it sensible to try and detect possible terror links among people?

Too much data can lead to spurious results

- Is it sensible to try and detect possible terror links among people?
- Setting: assume terrorists meet at least twice in a hotel to plot something sinister
- Government method: scan hotel logs to identify such occurrences

Too much data can lead to spurious results

- Is it sensible to try and detect possible terror links among people?
- Setting: assume terrorists meet at least twice in a hotel to plot something sinister
- Government method: scan hotel logs to identify such occurrences
- Data assumptions
  - Number of people: 10^9
  - Tracked over 10^3 days (about 3 years)
  - A person stays in a hotel with a probability of 1%
  - Each hotel hosts 10^2 people at a time

Too much data can lead to spurious results

- Is it sensible to try and detect possible terror links among people?
- Setting: assume terrorists meet at least twice in a hotel to plot something sinister
- Government method: scan hotel logs to identify such occurrences
- Data assumptions
  - Number of people: 10^9
  - Tracked over 10^3 days (about 3 years)
  - A person stays in a hotel with a probability of 1%
  - Each hotel hosts 10^2 people at a time
- Deductions
  - A person stays in hotel for

Too much data can lead to spurious results

- Is it sensible to try and detect possible terror links among people?
- Setting: assume terrorists meet at least twice in a hotel to plot something sinister
- Government method: scan hotel logs to identify such occurrences
- Data assumptions
  - Number of people: 10^9
  - Tracked over 10^3 days (about 3 years)
  - A person stays in a hotel with a probability of 1%
  - Each hotel hosts 10^2 people at a time
- Deductions
  - A person stays in hotel for 10 days
  - Total number of hotels is

Too much data can lead to spurious results

- Is it sensible to try and detect possible terror links among people?
- Setting: assume terrorists meet at least twice in a hotel to plot something sinister
- Government method: scan hotel logs to identify such occurrences
- Data assumptions
  - Number of people: 10^9
  - Tracked over 10^3 days (about 3 years)
  - A person stays in a hotel with a probability of 1%
  - Each hotel hosts 10^2 people at a time
- Deductions
  - A person stays in hotel for 10 days
  - Total number of hotels is 10^5

Too much data can lead to spurious results

- Is it sensible to try and detect possible terror links among people?
- Setting: assume terrorists meet at least twice in a hotel to plot something sinister
- Government method: scan hotel logs to identify such occurrences
- Data assumptions
  - Number of people: 10^9
  - Tracked over 10^3 days (about 3 years)
  - A person stays in a hotel with a probability of 1%
  - Each hotel hosts 10^2 people at a time
- Deductions
  - A person stays in hotel for 10 days
  - Total number of hotels is 10^5
    - Each day, 10^7 people stay in a hotel
    - Per hotel, 10^2 people stay

Bonferroni's principle

- In a day, probability that person A and B stays in same hotel is

Bonferroni's principle

- In a day, probability that person A and B stays in same hotel is 10^{-9}

Bonferroni's principle

- In a day, probability that person A and B stays in same hotel is 10^{-9}
  - Probability that A stays in a hotel that day is 10^{-2}
  - Probability that B stays in a hotel that day is 10^{-2}
  - Probability that B chooses A's hotel is 10^{-5}

Bonferroni's principle

- In a day, probability that person A and B stays in same hotel is 10^{-9}
  - Probability that A stays in a hotel that day is 10^{-2}
  - Probability that B stays in a hotel that day is 10^{-2}
  - Probability that B chooses A's hotel is 10^{-5}
- Probability that A and B meet twice is

Bonferroni's principle

- In a day, probability that person A and B stays in same hotel is 10^{-9}
  - Probability that A stays in a hotel that day is 10^{-2}
  - Probability that B stays in a hotel that day is 10^{-2}
  - Probability that B chooses A's hotel is 10^{-5}
- Probability that A and B meet twice is 10^{-18}
  - Two independent events: $10^{-9} \times 10^{-9}$

Bonferroni's principle

- In a day, probability that person A and B stays in same hotel is 10^{-9}
  - Probability that A stays in a hotel that day is 10^{-2}
  - Probability that B stays in a hotel that day is 10^{-2}
  - Probability that B chooses A's hotel is 10^{-5}
- Probability that A and B meet twice is 10^{-18}
  - Two independent events: $10^{-9} \times 10^{-9}$
- Total pairs of days is

Bonferroni's principle

- In a day, probability that person A and B stays in same hotel is 10^{-9}
  - Probability that A stays in a hotel that day is 10^{-2}
  - Probability that B stays in a hotel that day is 10^{-2}
  - Probability that B chooses A's hotel is 10^{-5}
- Probability that A and B meet twice is 10^{-18}
  - Two independent events: $10^{-9} \times 10^{-9}$
- Total pairs of days is (roughly) 5×10^5
  - Any 2 out of 10^3: $\binom{10^3}{2}$

Bonferroni's principle

- In a day, probability that person A and B stays in same hotel is 10^{-9}
  - Probability that A stays in a hotel that day is 10^{-2}
  - Probability that B stays in a hotel that day is 10^{-2}
  - Probability that B chooses A's hotel is 10^{-5}
- Probability that A and B meet twice is 10^{-18}
  - Two independent events: $10^{-9} \times 10^{-9}$
- Total pairs of days is (roughly) 5×10^5
  - Any 2 out of 10^3: $\binom{10^3}{2}$
- Probability that A and B meet twice in some pair of days is

Bonferroni's principle

- In a day, probability that person A and B stays in same hotel is 10^{-9}
  - Probability that A stays in a hotel that day is 10^{-2}
  - Probability that B stays in a hotel that day is 10^{-2}
  - Probability that B chooses A's hotel is 10^{-5}
- Probability that A and B meet twice is 10^{-18}
  - Two independent events: $10^{-9} \times 10^{-9}$
- Total pairs of days is (roughly) 5×10^5
  - Any 2 out of 10^3: $\binom{10^3}{2}$
- Probability that A and B meet twice in some pair of days is 10^{-13}
  - $10^{-18} \times 5 \times 10^5$

Bonferroni's principle

- In a day, probability that person A and B stays in same hotel is 10^{-9}
  - Probability that A stays in a hotel that day is 10^{-2}
  - Probability that B stays in a hotel that day is 10^{-2}
  - Probability that B chooses A's hotel is 10^{-5}
- Probability that A and B meet twice is 10^{-18}
  - Two independent events: $10^{-9} \times 10^{-9}$
- Total pairs of days is (roughly) 5×10^5
  - Any 2 out of 10^3: $\binom{10^3}{2}$
- Probability that A and B meet twice in some pair of days is 10^{-13}
  - $10^{-18} \times 5 \times 10^5$
- Total pairs of people is

Bonferroni's principle

- In a day, probability that person A and B stays in same hotel is 10^{-9}
  - Probability that A stays in a hotel that day is 10^{-2}
  - Probability that B stays in a hotel that day is 10^{-2}
  - Probability that B chooses A's hotel is 10^{-5}
- Probability that A and B meet twice is 10^{-18}
  - Two independent events: $10^{-9} \times 10^{-9}$
- Total pairs of days is (roughly) 5×10^5
  - Any 2 out of 10^3: $\binom{10^3}{2}$
- Probability that A and B meet twice in some pair of days is 10^{-13}
  - $10^{-18} \times 5 \times 10^5$
- Total pairs of people is (roughly) 5×10^{17}
  - Any 2 out of 10^9: $\binom{10^9}{2}$

Bonferroni's principle

- In a day, probability that person A and B stays in same hotel is 10^{-9}
  - Probability that A stays in a hotel that day is 10^{-2}
  - Probability that B stays in a hotel that day is 10^{-2}
  - Probability that B chooses A's hotel is 10^{-5}
- Probability that A and B meet twice is 10^{-18}
  - Two independent events: $10^{-9} \times 10^{-9}$
- Total pairs of days is (roughly) 5×10^5
  - Any 2 out of 10^3: $\binom{10^3}{2}$
- Probability that A and B meet twice in some pair of days is 10^{-13}
  - $10^{-18} \times 5 \times 10^5$
- Total pairs of people is (roughly) 5×10^{17}
  - Any 2 out of 10^9: $\binom{10^9}{2}$
- Expected number of suspicions, i.e., number of people meeting twice on any pair of days is

Bonferroni's principle

- In a day, probability that person A and B stays in same hotel is 10^{-9}
  - Probability that A stays in a hotel that day is 10^{-2}
  - Probability that B stays in a hotel that day is 10^{-2}
  - Probability that B chooses A's hotel is 10^{-5}
- Probability that A and B meet twice is 10^{-18}
  - Two independent events: $10^{-9} \times 10^{-9}$
- Total pairs of days is (roughly) 5×10^5
  - Any 2 out of 10^3: $\binom{10^3}{2}$
- Probability that A and B meet twice in some pair of days is 10^{-13}
  - $10^{-18} \times 5 \times 10^5$
- Total pairs of people is (roughly) 5×10^{17}
  - Any 2 out of 10^9: $\binom{10^9}{2}$
- Expected number of suspicions, i.e., number of people meeting twice on any pair of days is 2.5×10^5
  - $5 \times 10^{-13} \times 5 \times 10^{17}$
- Bonferroni's principle: if you look in more places for interesting patterns than your amount of data supports, you are bound to "find" something "interesting" (most likely spurious)

Tools for Big Data

- Hosting: Distributed servers or Cloud
  - Amazon EC2

Tools for Big Data

- Hosting: Distributed servers or Cloud
  - Amazon EC2
- File system: Scalable and distributed
  - HDFS, Amazon S3

Tools for Big Data

- Hosting: Distributed servers or Cloud
  - Amazon EC2
- File system: Scalable and distributed
  - HDFS, Amazon S3
- Programming model: Distributed scalable processing
  - Map-reduce framework, Hadoop

Tools for Big Data

- Hosting: Distributed servers or Cloud
  - Amazon EC2
- File system: Scalable and distributed
  - HDFS, Amazon S3
- Programming model: Distributed scalable processing
  - Map-reduce framework, Hadoop
- Database: NoSQL
  - HBase, MongoDB, Cassandra

Tools for Big Data

- Hosting: Distributed servers or Cloud
  - Amazon EC2
- File system: Scalable and distributed
  - HDFS, Amazon S3
- Programming model: Distributed scalable processing
  - Map-reduce framework, Hadoop
- Database: NoSQL
  - HBase, MongoDB, Cassandra
- Operations: Querying, indexing, analytics
  - Data mining, Information retrieval
  - Machine learning: Mahout on top of Hadoop

Discussion

- Emerging term: data science

Discussion

- Emerging term: data science
- Big data does not always need large investment
- Many open-source tools
- Cloud, etc. can be rented

Discussion

- Emerging term: data science
- Big data does not always need large investment
- Many open-source tools
- Cloud, etc. can be rented
- Most applications do not require big data

Discussion

- Emerging term: data science
- Big data does not always need large investment
- Many open-source tools
- Cloud, etc. can be rented
- Most applications do not require big data
- “Big Data” is currently too hyped