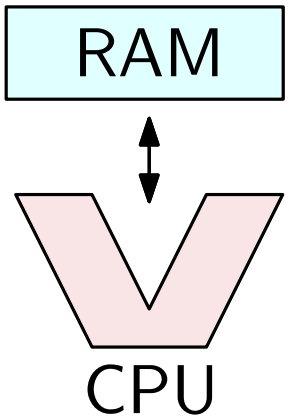
RESEARCH HORIZONS 2016

Qin Zhang

Efficient Algorithms for Querying Noisy Disturbed/Streaming Datasets

INDIANA UNIVERSITY
SCHOOL OF INFORMATICS AND COMPUTING

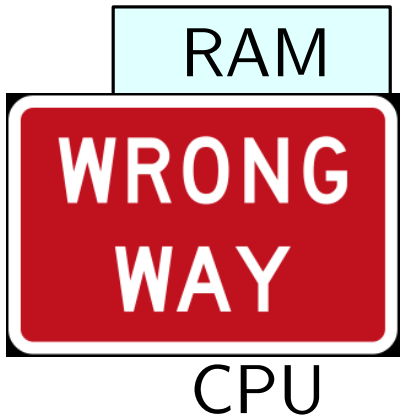
Classical computation model does not fit "big data"



The random-access memory (RAM) model

- A processor and an unbounded size memory
- Centralized, cost is in terms of # memory cells read/written

Classical computation model does not fit "big data"



The random-access memory (RAM) model

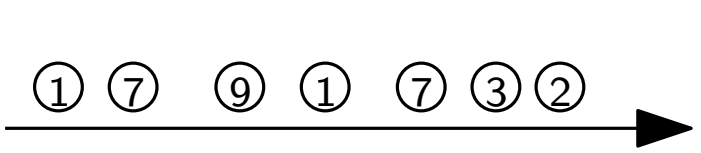
- A processor and an unbounded size memory
- Centralized, cost is in terms of # memory cells read/written

Big data doesn't fit!

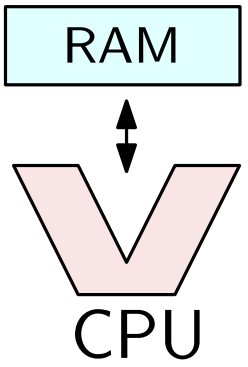
- High-speed online data : incapable of storing everything
- Data is distributed in different machines

Big data computation model: streaming model

Streaming Model



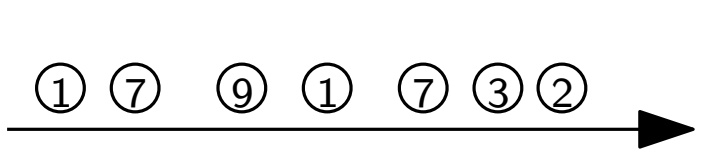
E.g., what is the number of distinct elements?



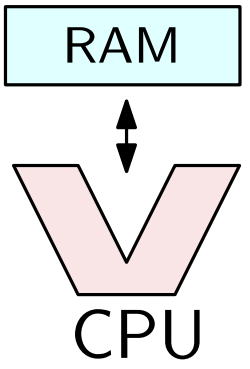
- small RAM space
- small update time
- small query time

Big data computation model: streaming model

Streaming Model

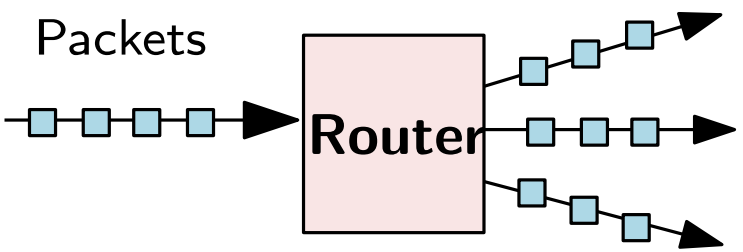


E.g., what is the number of distinct elements?

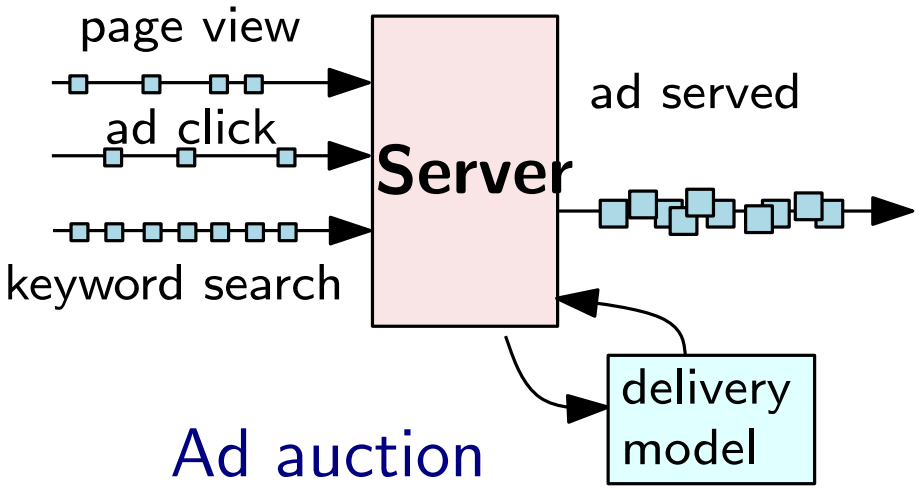


- small RAM space
- small update time
- small query time

Captures many settings, e.g.,



Network routers



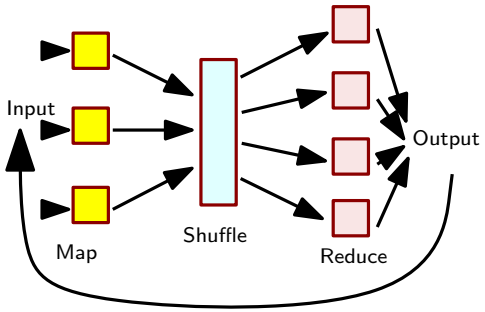
Ad auction

And flight logs, telephone switches, etc.

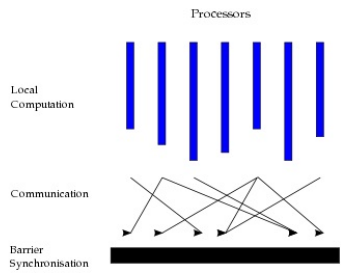
Big data computation model: k-site model

Communication $\rightarrow$ time, energy, bandwidth, ...

Sensor networks

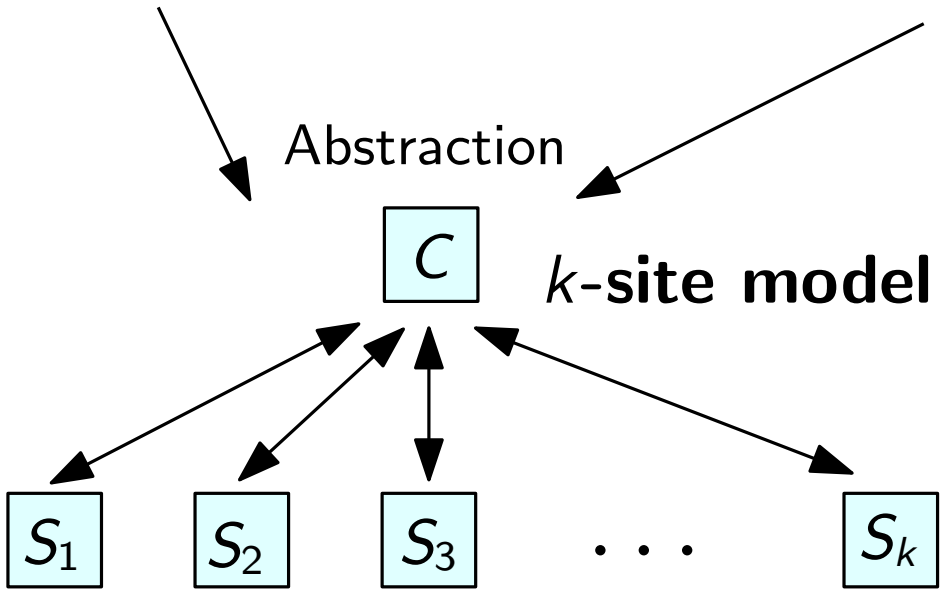


The MapReduce model.



The BSP model.

Cloud computation



- small comm. cost
- small # comm. rounds
- small local computation time

Data is noisy

Real-world datasets are often noisy.
Imprecise references to the same real-world entities are ubiquitous in scientific and commercial databases.

- 

  music, images, ...
  after compressions, resize, reformat, etc.

Data is noisy

Real-world datasets are often noisy.
Imprecise references to the same real-world entities are ubiquitous in scientific and commercial databases.

- 

  music, images, ...
  after compressions, resize, reformat, etc.

- 

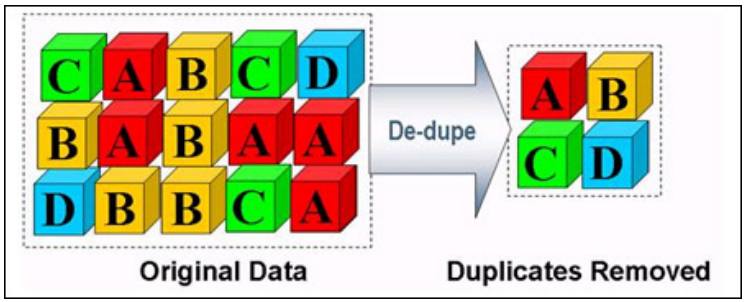
  "IUB"
  "Indiana University Bloomington"
  "Hoosier University"

Queries of the same meaning sent to Google

Challenges

Cannot do a comprehensive data deduplication in the streaming/k-site model using small space/communication!

Cannot



With my PhD students, we are working on the design of distributed and streaming algorithms that
run directly on the noisy datasets,
resolve the noise "on the fly", and
retain communication and space efficiency

project funded by NSF

RESEARCH HORIZONS 2016

THANK YOU

CONTACT INFORMATION

Email: qzhangcs@Indiana.edu
http://homes.soic.indiana.edu/qzhangcs/

INDIANA UNIVERSITY
SCHOOL OF INFORMATICS AND COMPUTING