Models for massive data

centralized

distributed

static/one-shot

continuous/real-time
Models for massive data

How to deal with:

Fast/small main memory
+ larger/slower external memory?
Models for massive data

The external memory (I/O) model
(Aggarwal and Vitter 1988)

Captures:
- memory hierarchy

Cost: \# I/Os

Disks

RAM

CPU

one block each I/O

blocks

distributed

centralized

static/one-shot

continuous/real-time
Models for massive data

The external memory (I/O) model
(Aggarwal and Vitter 1988)

Captures:
- memory hierarchy

Central issue:
- exploit the data locality

Distributed vs. centralized

Static/one-shot vs. continuous/real-time

Disk

RAM

CPU

one block
each I/O
blocks

cost: \# I/Os
Models for massive data

The external memory (I/O) model
(Agarwal and Vitter 1988)

Captures:
- memory hierarchy

Central issue:
- exploit the data locality

Applications:
- central in modern DB systems.
  e.g., external hashing, B-tree, ...

The diagram shows a hierarchical model with three levels:
- Disk
- RAM
- CPU

The cost is measured as the number of I/Os (I/Os).

The model is used in modern DB systems, such as external hashing and B-trees.
Models for massive data

- Distributed
- Centralized

- Static/one-shot
- Continuous/real-time

I/O - model
Models for massive data

What if we are incapable of storing everything?
Models for massive data

The streaming model
(Alon, Matias and Szegedy 1996)

Captures:
- limited storage

I/O - model

RAM

CPU

centralized

distributed

static/one-shot

continuous/real-time
Models for massive data

The streaming model
(Alon, Matias and Szegedy 1996)

Captures:
- limited storage

Game Time!
Models for massive data

The streaming model
(Alon, Matias and Szegedy 1996)

Captures:
- limited storage
Models for massive data

The streaming model
(Alon, Matias and Szegedy 1996)

Captures:
- limited storage

I/O - model

RAM

CPU

distributed
centralized
static/one-shot
continuous/real-time
Models for massive data

The streaming model
(Alon, Matias and Szegedy 1996)

Captures:
- limited storage

I/O - model

RAM

CPU
Models for massive data

The streaming model
(Alon, Matias and Szegedy 1996)

Captures:
• limited storage

Centralized vs. Distributed

Static/One-shot vs. Continuous/Real-time

I/O - model

CPU

RAM
Models for massive data

The streaming model
(Alon, Matias and Szegedy 1996)

Captures:
• limited storage
Models for massive data

The streaming model
(Alon, Matias and Szegedy 1996)

Captures:
- limited storage

Centralized vs. distributed

Static/one-shot vs. continuous/real-time

I/O - model

CPU

RAM
Models for massive data

The streaming model
(Alon, Matias and Szegedy 1996)

Captures:
- limited storage

RAM

CPU

I/O - model

centralized
distributed

static/one-shot
continuous/real-time
Models for massive data

The streaming model
(Alon, Matias and Szegedy 1996)

Captures:
- limited storage

The I/O - model

RAM

CPU

centralized

distributed

static/one-shot

continuous/real-time

29
Models for massive data

The streaming model
(Alon, Matias and Szegedy 1996)

Captures:
- limited storage

I/O - model

centralized

static/one-shot

continuous/real-time

distributed
Models for massive data

The streaming model
(Alon, Matias and Szegedy 1996)

Captures:
- limited storage

RAM

CPU

I/O - model

centralized

distributed

static/one-shot

continuous/real-time
Models for massive data

The streaming model
(Alon, Matias and Szegedy 1996)

Captures:
- limited storage

I/O - model

RAM

CPU

centralized
distributed

static/one-shot
continuous/real-time
Q: What’s the median?

Models for massive data

The streaming model
(Alon, Matias and Szegedy 1996)

Captures:

- limited storage

I/O - model

RAM

CPU

centralized
distributed

static/one-shot

continuous/real-time
Models for massive data

The streaming model
(Alon, Matias and Szegedy 1996)

Captures:
- limited storage

Q: What's the median?

Answer: 33

Q: What's the median? Answer: 33
Models for massive data

The streaming model
(Alon, Matias and Szegedy 1996)

Captures:
- limited storage

Central issue:
- space-approximation tradeoff

I/O - model

RAM

CPU

static/one-shot

centralized

distributed

continuous/real-time
Models for massive data

The streaming model
(Alon, Matias and Szegedy 1996)

Captures:
- limited storage

Central issue:
- space-approximation tradeoff

Systems:
- NiagaraCQ, Stanford Stream, Aurora, Telegraph, ...
- Companies e.g., Streambase

RAM
CPU
I/O - model

static/one-shot
continuous/real-time
distributed
centralized
Models for massive data

- I/O - model
  - Streaming model

- Distributed
- Centralized

- Static/One-shot
- Continuous/Real-time
Models for massive data

What if data is stored distributively?

I/O - model

Streaming model

centralized

distributed

static/one-shot

continuous/real-time
Models for massive data

Among others, **MapReduce** (Google 2004)

Captures:
- distributed computation

Input → Map → Shuffle → Reduce → Output

I/O - model

Streaming model
Models for massive data

Among others, MapReduce (Google 2004)

Captures:
- distributed computation

Central issue:
- communication and round complexity

I/O model

Streaming model
Models for massive data

Among others, **MapReduce** (Google 2004)

Captures:
- distributed computation

Central issue:
- communication and round complexity

Systems:
- Hadoop
- Google MapReduce

I/O - model

Streaming model
Models for massive data

- **centralized**
  - static/one-shot
  - I/O - model
  - Streaming model

- **distributed**
  - continuous/real-time
  - MapReduce
Models for massive data

How to deal with data comes as multiple streams distributed at different locations?

MapReduce

I/O - model

Streaming model

centralized

distributed

static/one-shot

continuous/real-time
Models for massive data

- **Static/one-shot**
  - Centralized
  - Distributed

- **Continuous/real-time**
  - Streaming model
  - Distributed streaming model (next page)

- **I/O - model**
  - MapReduce

- **Streaming model**