CSE 550: Systems for all

Au 2022

Ratul Mahajan

“Big Data” problems

Lots of data

Semi-structured data

- Web logs
- Documents

Ad hoc computations

- But naturally parallelizable

Throughput-oriented (not latency-oriented)

Consistency not so important

- Or, writes are uncommon

Why not DBs?

Design space for data processing systems

	Latency	Throughput
Parallelizable (Data parallel)	Search, KV lookup (NoSQL)	Word counting (MapReduce)
Intertwined	Transactions (Traditional DBs)	Drug simulations (HPC systems)

Inspired by Bill Howe and Michael Isaard

General approach to high-performance data processing

1. Build a data processing graph
   a. Figure out parallelism
   b. Figure out processing dependencies

2. Speed up processing “kernels”
   a. CPU, GPU, TPU

3. Orchestrate data across kernels
   a. Storage, memory, networking

Over to Benedikt and Matthew