CSE 550: Systems for all

Au 2021

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

<table>
<thead>
<tr>
<th>Parallelizable (Data parallel)</th>
<th>Latency</th>
<th>Throughput</th>
</tr>
</thead>
<tbody>
<tr>
<td>Search, KV lookup (NoSQL)</td>
<td></td>
<td>Word counting (MapReduce)</td>
</tr>
<tr>
<td>Intertwined</td>
<td>Transactions (Traditional DBs)</td>
<td>Drug simulations (HPC systems)</td>
</tr>
</tbody>
</table>

Inspired by Bill Howe and Michael Isaard
Over to Prashant and Wentao