Horizontal Virtualization
On Commodity Hardware

Stefan Groschupf
sg@datameer.com
Laws Of Physics

<table>
<thead>
<tr>
<th>Storage Type</th>
<th>Random Values/Sec</th>
<th>Sequential Values/Sec</th>
</tr>
</thead>
<tbody>
<tr>
<td>Disk</td>
<td>316</td>
<td>53,200,000</td>
</tr>
<tr>
<td>SSD</td>
<td>1,924</td>
<td>42,200,000</td>
</tr>
<tr>
<td>Memory</td>
<td>36,700,000</td>
<td>358,200,000</td>
</tr>
</tbody>
</table>

Adam Jacobs
The Pathologies of Big Data
Laws Of Physics

1. $F_z = 100 \text{ N}$
   - $s = 10 \text{ cm}$

2. $F_z = 50 \text{ N}$
   - $s = 20 \text{ cm}$

3. $F_z = 33\frac{1}{3} \text{ N}$
   - $s = 30 \text{ cm}$

4. $F_z = 25 \text{ N}$
   - $s = 40 \text{ cm}$

Hadoop(mergesort)

DB(b-tree/index)
Thesis, Antithesis, Synthesis

**Thesis:**
SQL/ACID solves all your data problems.

**Antithesis:**
Nobody needs SQL/ACID, let's throw it all out.

**Synthesis:**
By carefully considering data integrity constraints we can find a more optimal data management solution for a particular problem.

voidpointer:
http://news.ycombinator.com/item?id=1163516
PUBLIC ENEMY

DON'T BELIEVE THE HYPE
Considerations

- Keep it simple!
  => Simple to distribute.

- What is really the problem to solve?
  • Serve Data or Analyze Data?

- Sequential Write/Read performs best.
  • Data aggregation challenges.

- Index only if really necessary.
  • For data serving.
Distribute on low cost HW

- Distributed Storage, Computation.
- Distributed Index.
- Horizontal Virtual Operating System.
- Open Source Platform.
Meet the ...
www.datameer.com
Free for .edu!
Street Cred

- Long time open source contributor

http://github.com/sgroschupf/zkclient
http://github.com/sgroschupf/aws-tasks