Cold Data Archival

Sanjay Joshi
EMC² Isilon Storage Division
"...because as we know, there are known knowns; there are things that we know that we know.

We also know there are known unknowns; that is to say we know there are some things we don’t know.

But there are also unknown unknowns, the ones we don’t know we don’t know."

Donald Rumsfeld, United States Secretary of Defense 1975-1977, 2001-2007
The Reality of MTBF
40 ZB by 2020

Source: EMC Digital Universe Study, 2014
Research and Analysis by IDC

The Digital Universe
Monty Python, 1972. © BBC
Life: RNA came first?

Complex organic molecules produced by random chemistry

Lipid-RNA hypothesis


The first "write" process in the cell?

char(3*10^9) human_genome

3 gigabases \[ \frac{(3 \times 10^9) \times 2}{8} = \sim 750MB \]

with overlaps, \sim 1 GB per cell


DNA Sizing as Storage
Source: Church et al., Science, Vol 337, p1628, Sep 2012
2 PB/gram cheaper than tape by ~2025?

DNA as Storage: Process

Oligo library
ink-jet printer
Parallel Flow Cell Reactor 1 error/500 bases
Purify, mag beads PCR amplify
15 ng/μL

Derived from: Goldman et al., Nature (Jan 2013) doi:10.1038/nature11875

Note: Images from Agilent, GE and Life Technologies for example purposes only.
**DNA Sequencer**

Derived from: Goldman et al., Nature (Jan 2013) doi:10.1038/nature11875

Note: Image from Oxford Nanopore sequencer for example purposes only.
DNA Storage Scale & Cost

Derived from: Goldman et al., Nature (Jan 2013) doi:10.1038/nature11875

- DNA synthesis: 10× cheaper
- DNA synthesis: 100× cheaper
- Current costs for 1 MB of data
- Break-even time for tape rewritten every 5 yrs
- Break-even time for tape rewritten every 10 yrs
- Tape less expensive (assuming tape rewritten every 5 yr)
- Tape less expensive (assuming tape rewritten every 10 yr)

Tape rewritten every: 5 yrs, 10 yrs

DNA Storage Scale & Cost

© Copyright 2014 EMC Corporation. All rights reserved. Do not redistribute without permission
Brewer’s Theorem

Consistency

ACID
Atomicity, Consistency, Isolation, Durability

Partition Tolerance

BASE
Basically Available Soft-state Eventual consistency

Acknowledgement

Source: Brewer E, "CAP Twelve Years Later: How the "Rules" Have Changed,",
http://www.infoq.com/articles/cap-twelve-years-later-how-the-rules-have-changed

Distributing it…
“Information moves at the speed of trust.”
Dr. Douglas Fridsma, CSO

“Non-reproducible single occurrences are of no significance to science.” Karl Popper, 1959

Big Data = V^6
Volume, Velocity, Variability, Visualization, Viscosity, Veracity

Control + Visibility + Transparency

<table>
<thead>
<tr>
<th>Public</th>
<th>Infrastructure Managed By²</th>
<th>Infrastructure Owned By²</th>
<th>Infrastructure Located³</th>
<th>Accessible and Consumed By²</th>
</tr>
</thead>
<tbody>
<tr>
<td>Third Party Provider</td>
<td>Third Party Provider</td>
<td>Off-Premise</td>
<td>Untrusted</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Private/Community</th>
<th>Organization</th>
<th>On-Premise</th>
<th>Trusted</th>
</tr>
</thead>
<tbody>
<tr>
<td>Third Party Provider</td>
<td>Third Party Provider</td>
<td>Off-Premise</td>
<td>Untrusted</td>
</tr>
</tbody>
</table>

| Hybrid            | Both Organization & Third Party Provider | Both Organization & Third Party Provider | Both On-Premise & Off-Premise | Trusted & Untrusted |

5.3 Days: length of security breach, worst for Healthcare/Pharma*

*Ponemon Institute Mar 2014

** BitSight Industry Security Report, May 2014

343 US incidents through Jul 2013
321 in 2012 *ITRC

30M data breaches since 2009 (avg. $2M fine per breach, 2 yrs)
US Govt. Office of Civil Rights, HHS
£1.79B fines in UK for NHS 2012 breaches

4 data points (GPS) determine identity
... Scientific Reports, 3, Mar 2013

4M Common Variants in trace DNA

Private & Community Cloud
One Word: “Polymers”

Inspired by Mr. McGuire in movie ‘The Graduate’, © Embassy Pictures, 1967

\[
(2 \times 23) = 50 \text{ ZB} @ 10^{21} \text{ cells}
\]

50 ZB @ 10^{21} cells

Thank You!