Top 10 Things We Learned About Hadoop

Val Bercovici
Cloud & Big Data Czar
NetApp
What I hope to accomplish today...

Help you think beyond Big Data hype

&

Deliver a modern take on Hadoop
...and avoid this.
10. Both Kinds of Hadoop Customers

Data *is* their business

![Yahoo! Logo](image1)

![Facebook Logo](image2)

![eBay Logo](image3)

Data *supports* their business

![NetApp Logo](image4)

![Apple Logo](image5)

![Intel Logo](image6)
9. Technology Marches On

Disk and networking trends in Hadoop

MRPerf: A Simulator for MapReduce

© 2012 NetApp, Inc. All rights reserved.
8. SuperNodes to Dominate

http://www.slideshare.net/Hadoop_Summit/dynamic-namespace-partitioning-with-giraffa-file-system
7. C-Level Fear & Envy

- **2010**
  - “What is Hadoop?”
  - Google, Wikipedia, Slideshare

- **2011**
  - “How can Hadoop help us?”
  - “Let’s bring in a consultant”

- **2012**
  - “Our competition is using Hadoop”
  - “I can’t afford to be behind”
6. Cluster – Get Healthy, Stay Healthy

Network Overhead

- Availability and Resiliency
- Burst Handling and Queuing
- Oversubscription Ratio
- Data Node Network Speed
- Network Latency

Useful Work

Table 4. Drive Failure Recovery Results

<table>
<thead>
<tr>
<th>Test Scenario</th>
<th>Job Completion Time (hh:mm:ss)</th>
<th>Throughput (MB/sec)</th>
<th>Delta (vs. Healthy Cluster)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Healthy cluster</td>
<td>00:30:21</td>
<td>1,821</td>
<td>N/A</td>
</tr>
<tr>
<td>NetApp E2660 drive failure</td>
<td>00:32:06</td>
<td>1,486</td>
<td>-5.7%</td>
</tr>
<tr>
<td>Internal data node drive failure</td>
<td>01:12:13</td>
<td>660</td>
<td>-237.9%</td>
</tr>
</tbody>
</table>


5. It Takes a Village

Mobile Devices
Location/GPS
Logs
Sensors
Applications

Storage
(All other storage, i.e. internal DAS)

ETL
OLAP
OLTP
Reporting/Dashboard/Visualization
Infrastructure
Storage File Systems
Data Management
Content Repositories

Applications
Log/Dashboards/Visualization

IBM
ORACLE
SAP
Microsoft
Cloud Era

IBM
Teradata
Sas

PARACCEL
EXASOL
10gen
MAPR

INFORMAGEN
birst
Fuzzy Logix Systems

© 2012 NetApp, Inc. All rights reserved.
4. Hadoop has crossed Verticals

- BFSI
- Telecom
- Retail
- Consumer Durables
- Automobiles
- Healthcare
- Hospitality
- Travel
- IT
- Media
## 3. HFDS I/O Profile is Diverse

<table>
<thead>
<tr>
<th>Application</th>
<th>Map Input</th>
<th>Map Output</th>
<th>Reduce Output</th>
<th>Number of tasks</th>
<th>Cost (cycle/byte)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Grep</td>
<td>10GB</td>
<td>1MB</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>TeraGen</td>
<td>0KB</td>
<td>10GB</td>
<td>-</td>
<td>40</td>
<td>10</td>
</tr>
<tr>
<td>TeraSort</td>
<td>10GB</td>
<td>10GB</td>
<td>10GB</td>
<td>160</td>
<td>40</td>
</tr>
<tr>
<td>Join</td>
<td>10GB</td>
<td>1GB</td>
<td>10MB</td>
<td>160</td>
<td>40</td>
</tr>
<tr>
<td>Aggregate</td>
<td>10GB</td>
<td>100MB</td>
<td>10MB</td>
<td>160</td>
<td>10</td>
</tr>
<tr>
<td>Inverted Index</td>
<td>1GB</td>
<td>10GB</td>
<td>100MB</td>
<td>40</td>
<td>10</td>
</tr>
<tr>
<td>PageRank</td>
<td>1GB</td>
<td>10GB</td>
<td>1GB</td>
<td>40</td>
<td>100</td>
</tr>
<tr>
<td>Small</td>
<td>100KB</td>
<td>1MB</td>
<td>10KB</td>
<td>4</td>
<td>400</td>
</tr>
<tr>
<td>Summary</td>
<td>10GB</td>
<td>10MB</td>
<td>10KB</td>
<td>160</td>
<td>40</td>
</tr>
<tr>
<td>Compute</td>
<td>1GB</td>
<td>10GB</td>
<td>100MB</td>
<td>40</td>
<td>4000</td>
</tr>
</tbody>
</table>

© 2012 NetApp, Inc. All rights reserved.
2. Hadoop often over-provisioned
1. Big Brother = Big Compliance

Full Motion Video Capture

GORGON STARE
Current Reaper and Predator sensors offer a single field of view. New wide-area airborne sensors for the Reaper will record video from 12 cameras simultaneously.

Digital Video Surveillance
0. Fast Data is the new Big Data

Apache Drill
Thank you

Opportunity  Courage  Win  Propel
Innovate  Passion
Change  Inspire  Motivate
Grow  Adapt