Scalable Forensics with TSK and Hadoop

Jon Stewart
CPU Clock Speed
Hard Drive Capacity

The graph shows the trend in hard drive capacity from 1985 to 2010. The capacity increased significantly over this period, indicating exponential growth. The x-axis represents the years, while the y-axis represents the capacity in GB.
The Problem

• CPU clock speed stopped doubling

• Hard drive capacity kept doubling

• Multicore CPUs to the rescue!

• ...but they're wasted on single-threaded apps

• ...and hard drive transfer speeds might be too slow for 24 core machines (depending)
Solution:
Distributed Processing

• Split the data up
• Process it in parallel
• Scale out as needed

• Sounds great!!
Typical Distributed Processing == Storage Bottleneck

Copy Data Back and Forth
(every time you need to run something)
Stop contributing to Larry Ellison’s Island
These guys have a lot of data...
...how do they do their processing?
Apache Hadoop

Apache Hadoop is an open source software suite for distributed processing and storage, primarily inspired by Google's in-house systems.

http://hadoop.apache.org
Hadoop Distributed File System

Collected Data

Ingest Once + Replication

Storage Cluster (Hadoop HDFS)
HDFS
Just like any other file system, but better

- Metadata master server ("NameNode")
  - One server acts as FAT/MFT
  - Knows where files & blocks are on cluster
- Blocks on separate machines ("DataNodes")
  - Automatic replication of blocks (default 3x)
  - All blocks have checksums
  - Rack-aware
  - DataNodes form p2p network
  - Clients contact directly for reading/writing file data
- Not general purpose
  - Files are read-only once written
  - Optimized for streaming reads/writes of large files
MapReduce

Storage + Processing Cluster
(Hadoop HDFS + MapReduce)
MapReduce
Bring the application to the data

• Code is sent to every node in the cluster

• Local data is then processed as "tasks"
  – Support for custom file formats
  – Failures are restarted elsewhere, or skipped

• One nodes can process several tasks at once
  – Idle CPUs are bad

• Output is automatically collated

• Parallelism is transparent to programmer
HBase

Logical Table

Row A
Row B
Row C
Row D
Row E
Row F
Row G
Row H
Row I
Row J
Row K
Row L
Row M
Row N
Row O

Regions

Row A
Row B
Row C
Row D
Row E
Row F
Row G
Row H
Row I
Row J
Row K
Row L
Row M
Row N
Row O

Region Servers

Client

Diagram showing the structure of HBase with logical tables, regions, and region servers.
# HBase
Random access, updates, versioning

<table>
<thead>
<tr>
<th>Conventional RDBMS (MySQL)</th>
<th>HBase</th>
</tr>
</thead>
<tbody>
<tr>
<td>- Limited throughput</td>
<td>- High write throughput</td>
</tr>
<tr>
<td>- Not distributed</td>
<td>- Distributed automatically</td>
</tr>
<tr>
<td>- Fields always allocated</td>
<td>- Null values not stored</td>
</tr>
<tr>
<td>- varchar[255]</td>
<td>- A row is Name:Value list</td>
</tr>
<tr>
<td>- Strict schemas</td>
<td>- Schemas not imposed</td>
</tr>
<tr>
<td>- ACID integrity/transactions</td>
<td>- Rows sorted by key</td>
</tr>
<tr>
<td>- Support for joins</td>
<td>- No joins, no transactions</td>
</tr>
<tr>
<td>- Indices speed things up</td>
<td>- Good at table scans</td>
</tr>
<tr>
<td></td>
<td>- Fields are versioned</td>
</tr>
</tbody>
</table>
Let's do forensics on Hadoop!

- Lightbox designed and created a proof-of-concept
- Army Intelligence Center of Excellence funded an open source prototype effort
- In collaboration with 42six Solutions, Basis Technologies, and Dapper Vision
Table Schema Design

• Images
  – pkey is md5 of image, calculated on ingest
  – Columns for image details

• Hashes
  – pkey is hash value, currently either sha1 or md5
  – Store hash value + entry ID (fkey -> Entries) so dupes are always near each other
  – Columns for hash sets

• Entries
  – pkey is (image md5 + file path md5 + dir index #)
  – Columns for every piece of metadata
Ingest:
File Extraction / Hash Calculation

/ dd, E01

Lightbox technologies, inc
fsrip

The Sleuth Kit

md5sum
sha1sum

hadoop + APACHE HBASE
Processing Tasks

- Hash set analysis
- Keyword searching
- Text extraction
- Document clustering
- Face detection
- Graphics clustering
- Video analysis
- Not limited – parse *anything* with plugin interface
Processing Plugins

• Plugin interface for community members to extend functionality
• Python + <other languages>
• Plugins can return data to the system:
  – PST file -> Emails -> HBase rows
  – Internet History records -> HBase rows
  – Extracting zip files -> HBase rows
Graphics Clustering

Cluster 13 (100 images)

Cluster 0 (100 images)
How Do I Run It?

- Spin up Amazon EC2 instances (start with 5)
- Install Cloudera Manager (http://www.cloudera.com/products-services/tools/)
- Deploy Hadoop & Hbase with CM
Wrap Up

• Scalable forensics processing
  – No dongles, fancy hardware, or Oracle necessary

• Swimming in CPUs
  – Add many machines
  – Accomplish more & better analysis

• Ready for primetime soon...
One Last Surprise!

- **Lightgrep** to be open sourced Q4 – 2012
- Will be integrated with bulk_extractor in cooperation with Naval Postgraduate School
- Provides:
  - Perl syntax
  - Full Unicode support
  - Millions of keywords
  - Automated QA with millions of tests
  - GPL license
https://github.com/jonstewart/sleuthkit-hadoop

http://www.sleuthkit.org/tsk_hadoop/

Jon Stewart  | jon@lightboxtechnologies.com