End-User Examples: High Performance Data Analysis (HPDA)

Steve Conway
IDC Research Vice President
HPC & High Performance Data Analysis
sconway@idc.com

July 2014
High Performance Data Analysis

Needs HPC resources
- Complex algorithms
- Near-real time (often)
- Data “long” and “wide”
- On premise or in cloud

Simulation & analytics
- Search, pattern discovery
- Iterative methods
- Established HPC users + new commercial users

Data of all kinds
- The 4 V’s: volume, variety, velocity, value
- Structured, unstructured
- Partitionable, non-partitionable
- Regular, irregular patterns
**Different Systems for Different Jobs**

**Partitionable Work**
- Most jobs are here
- **Search** (e.g., Jeopardy Watson)
- Global memory not so important
- Standard clusters + Hadoop, Cassandra, HPCC, etc.

**Non-Partitionable Work**
- Toughest jobs (e.g., graphing)
- **Dynamic pattern discovery** (SGI UV, YarcData Urika, medical Watson, et al.)
- Global memory important
- Systems turbo-charged for data movement

vs.
HPC Adoption Timeline (Examples)
67% of the sites perform HPDA work (data-intensive simulation and/or advanced analytics).

On average, HPDA consumes 30% of compute cycles.

29% of sites use Hadoop

Major pain points worth 10-15% premium pricing:
- Interconnects between nodes
- External I/O and storage
Big Science: Big Data Challenges Are Growing…

Some “Big Data” Grand Challenges

- How do we handle 700 TB/sec of data coming off the wire when we actually have to keep it around?
  - Required by the Square Kilometre Array

- Joe scientist says I’ve got an IDL or Matlab algorithm that I will not change and I need to run it on 10 years of data from the Colorado River Basin and store and disseminate the output products
  - Required by the Western Snow Hydrology project

- How do we compare petabytes of climate model output data in a variety of formats (HDF, NetCDF, Grib, etc.) with petabytes of remote sensing data to improve climate models for the next IPCC assessment?
  - Required by the 5th IPCC assessment and the Earth System Grid and NASA

- How do we catalog all of NASA’s current planetary science data?
  - Required by the NASA Planetary Data System
Total Revenue Protection Program

Processing Requirements

- Rate
  - 4 billion mail scans per day peak (74,000 per second)

- Geographic Scope
  - Incoming mail from 275 Processing and Distribution Centers
  - Outgoing mail to 33,000 postal operated facilities

Objective

- To find, track and reject mail pieces due to:
  - Duplicate postage
  - Short Pay
  - Ineligible Discounts
Why Real Time Fraud Detection?

Save time... print your postage online.

Print exact postage for letters and packages using just your PC and printer.

Print Postage Stamps
- Print any denomination
- Use for letters or packages
- Never run out of stamps again

Stamps.com... Your own personal Post Office open 24 hours a day.
Developed in conjunction with the United States Postal Service,™ Stamps.com is a revolutionary software-based service that allows you to calculate and print official USFS postage right from your PC.

NO ADDITIONAL HARDWARE REQUIRED. Stamps.com even keeps track of all your postal spending using your client codes, and can even recommend optimal delivery methods, formats and more. Plus, Stamps.com gives you postage discounts you can’t even get at the Post Office or with a postage meter.

© 2006 The Copyright in this document belongs to FedCentric Technologies LLC and no part of this document should be used or copied without their prior written permission.
### TRP Results using MCDB & TimesTen

<table>
<thead>
<tr>
<th>Pre-MCDB</th>
<th>With MCDB Deployed</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. 509 row inserts per second (RIPS)</td>
<td>1. 190,222 RIPS (3 Threads)</td>
</tr>
<tr>
<td>2. Direct path load option a partial solution (2000 RIPS)</td>
<td>2. 1,091,018 RIPS (18 Threads)</td>
</tr>
<tr>
<td>3. 275 Million Transactions per 15 hour processing window created backlog during peak processing windows</td>
<td>3. Processed 4 B Transactions in less than 6 hours</td>
</tr>
<tr>
<td>4. Revenue Protection performed as a batch data warehouse process, run 3 – 12 hours after Mailpiece scan</td>
<td><strong>Revenue Protection is performed in real-time upon first scan</strong></td>
</tr>
</tbody>
</table>

**MCDB = memory-centric database**
CMS: Government Health Care Fraud

- 5 separate databases for the big USG health care programs under Centers for Medicare and Medicaid Services (CMS)
- Estimated fraud: $150B-$450B. <$5B caught today)
- ORNL, SDSC have evaluation contracts to unify the databases and perform fraud detection on various architectures.
Use Case: PayPal

Fraud Detection / Internet Commerce

Slides and permission provided by PayPal, an eBay company
Detecting fraud in 'real time' as millions of transactions are processed between disparate systems at volume.

Finding suspicious patterns that we don’t even know exist in related data sets.

Ability to create and deploy new fraud models into event flows quickly and with minimal effort.

Provide environment for fraud modeling, analytics, visualization, M/R, dimensioning and further processing.
PayPal: Next Uses for HPC

- After this success, PayPal now plans further uses for HPC:
  - Managing the whole PayPal IT infrastructure
  - Affinity marketing to consumers ("Beacon" project)
- Parent company eBay is not using HPC yet.

“Clearly understand that HPC is not a mass consumption technology where we enable everyone in our organization with it. This is a deep engineering function. It's custom built and includes writing software to solve cutting-edge problems ... Think of HPC not as an IT function but as a competitive business advantage. There's a hard link between HPC and PayPal's top line and bottom line.”

PayPal CTO Jim Barrese (IDC interview, 2013)
Use Case: Network Security

Network Intrusion Detection
32,000 Items (src & dest. IP address, ports, days, times, activities, etc.)
2.57 million Transactions

No rules or queries required. Auto-detect intrusion patterns and surface suspicious activity.
Schrödinger: Cloud-based Lead Discovery for Drug Design

<table>
<thead>
<tr>
<th>Metric from March, 2012</th>
<th>Count</th>
</tr>
</thead>
<tbody>
<tr>
<td>Compute Hours of Work</td>
<td>109,927 hours</td>
</tr>
<tr>
<td>Compute Years of Work</td>
<td>12.55 years</td>
</tr>
<tr>
<td>Total # Cores/Servers</td>
<td>51132 cores, 6742 servers</td>
</tr>
<tr>
<td>Infrastructure Value</td>
<td>~ $20,000,000</td>
</tr>
<tr>
<td>AWS Regions</td>
<td>All (7: us-east, us-west1, us-west2, eu-west, sa-east, ap-northeast, ap-southeast)</td>
</tr>
</tbody>
</table>

Using CycleCloud & AWS: Impossible run in 3 hours for $4,828/hr
Today’s pricing < $1,000/hr

© 2013 IDC
Outcomes-Based Medical Diagnosis and Treatment Planning

- Enter the patient’s history and symptomology.

- While patient is still in the office, sift through millions of archived patient records for relevant outcomes.

- Provider considers the efficacies of various treatments for “similar” patients (but is not bound by the findings).

- Ergo, this functions as a powerful decision-support tool.

- Benefits: better outcomes + rein in costly outlier practices
Optum Labs: UHG-led Collaborative to Advance Big Data in Health Care

- $500 million center planned in Cambridge, MA
- Pre-competitive, open research
- Contributors sit on governance board (e.g., Mayo)
- Long-term goal: enable outcomes-based medicine
• Fast growth from a small starting point: $1.2B (€900M by 2016)
• HPDA ecosystem >$2B (€1.5B) in 2016
Storage is the fastest-growing HPC market (8.4% CAGR, 2011-16) and HPDA storage will grow even faster (18.1% CAGR).

### TABLE 2

<table>
<thead>
<tr>
<th>Worldwide High-Performance Data Analysis Storage Revenue, 2009–2016 ($M)</th>
</tr>
</thead>
<tbody>
<tr>
<td>----------</td>
</tr>
<tr>
<td>HPC storage</td>
</tr>
<tr>
<td>Share as total HPC server revenue (%)</td>
</tr>
<tr>
<td>HPDA storage</td>
</tr>
<tr>
<td>Big Data attach rate (%)</td>
</tr>
</tbody>
</table>

Source: IDC, 2013
Summary: HPDA Market Opportunity

- **HPDA: simulation + newer high-performance analytics**
  - IDC predicts fast growth from a small starting point

- **HPC and high-end commercial analytics are converging.**
  - Algorithmic complexity is the common denominator

- **Economically important use cases are emerging**
  - Which ones will become attractive markets?

- **No single HPC solution is best for all problems.**
  - Clusters with MR/Hadoop will handle most but not all work (e.g., graph analysis)

  - IDC believes our growth estimates could be conservative.
Questions?

Please email:

hpc@idc.com

Check out:

www.hpcuserforum.com
www.hpcuserforum.com/ROI
www.idc.com