Best Practices, Optimized Interfaces, API’s designed for Storing Massive Quantities of Long Term Retention Data

Stacy Schwarz-Gardner

Spectra Logic, Active Archive Alliance
Strategic Technical Architect
Impact of Big Data Explosion

Growth in unstructured data far outpacing IT resources

Global Digital Information
48% Growth
2.7 Zettabytes in 2012
- IDC

2010 2020

50x

1.5x
Evidence Of The Problem in the “Machine Room”

Survey of Fortune 1000 Storage Professionals’ Pain Points

Managing Storage Growth
Proper Capacity Forecasting and Storage Reporting
Backup Administration and Management
Managing Costs
Managing Complexity
Lack of Integrated Tools
Storage Provisioning
Data Mobility
Archiving and Archive Management
Dealing with Performance Problems
Regulatory Compliance
Vendor Management
Power Management
Application Recoveries / Backup Retention
Data Migration
Managing Storage Equipment
Securing Storage
Disaster Recovery
Technology Complexity
Storage Monitoring
Staff Retention
Integration Complexity
Other

CIOs rank data growth as their top concern, and making sense of all that data as their number one priority.

(2012 Gartner CIO Agenda).

“For many organizations, unstructured content is fundamentally out of control,”

Tom Eid, Research Vice President for Gartner
### Example Customer Environments Using Active Archive Architectures

<table>
<thead>
<tr>
<th>Environment</th>
<th>Storage Capacity</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cosmological Research</td>
<td>100.0 PB</td>
</tr>
<tr>
<td>Aerospace agency (40 GB/sec) (21 years online)</td>
<td>60.0 PB</td>
</tr>
<tr>
<td>National Weather Analysis (300 TB/day I/O – 105GB/s NFS throughput)</td>
<td>60.0 PB</td>
</tr>
<tr>
<td>Movie visual effects – (1.8 Billion files)</td>
<td>24.0 PB</td>
</tr>
<tr>
<td>Aerospace research (21 years online)</td>
<td>20.0 PB</td>
</tr>
<tr>
<td>University – Genomics Research</td>
<td>20.0 PB</td>
</tr>
<tr>
<td>Sports League Active Archive (~40TB/day ingest)</td>
<td>18.0 PB</td>
</tr>
<tr>
<td>Scientific Research (21 yrs in prod, always online)</td>
<td>8.0 PB</td>
</tr>
<tr>
<td>Media and Entertainment Film Library</td>
<td>7.0 PB</td>
</tr>
<tr>
<td>Weather prediction (+13TB/day) – With Lustre</td>
<td>6.5 PB</td>
</tr>
<tr>
<td>European National Institute for Audio &amp; Video</td>
<td>4.5 PB</td>
</tr>
<tr>
<td>National Computing and Network Services (Europe)</td>
<td>4.0 PB</td>
</tr>
<tr>
<td>European National Research Agency</td>
<td>4.0 PB</td>
</tr>
<tr>
<td>European Space Research</td>
<td>4.0 PB</td>
</tr>
<tr>
<td>European Oil and Gas Customer</td>
<td>3.0 PB</td>
</tr>
<tr>
<td>Multinational Oil and Gas Customer</td>
<td>2.7 PB</td>
</tr>
<tr>
<td>Aircraft Manufacturer</td>
<td>2.0 PB</td>
</tr>
<tr>
<td>Earth Data Research</td>
<td>1.7 PB</td>
</tr>
<tr>
<td>Super Computing Center</td>
<td>1.6 PB</td>
</tr>
</tbody>
</table>
Massive data growth forcing industry-wide changes

- **Ever-increasing sizes of data and usage** that have made it difficult to manage storage assets while maintaining performance standards

- **Larger data volumes and disks** strain conventional storage infrastructures

- Infrastructure must be modular and flexible to reduce costs and allow **non-disruptive migration** to new hardware as technology evolves
Inefficient use of deployed storage assets
  - Silos & infrastructure fragmentation
  - Monolithic solutions (Vendor) lock-in…

Continuous need for administrator intervention (rebalancing loads)

Expensive and time-consuming upgrade/replacement cycles

Difficulties in managing information in dynamic environments such as the cloud

Addressing the limitations of existing storage environments requires the deployment of new classes of storage solutions that are optimized for high volume and often unpredictable data ingest, storage, and access. - IDC, Dec. 2012
Archive Driving Principle

“You never know what you will need and when you will need it!”
Inactive Data Consumes Active Primary Storage

Data needs to be online and available…

But does it always need to live on the most expensive medium?

Eighty-Five percent of production data is Inactive
- 68% not accessed in 90 days

According to Forrester Research

Over 66% of files are re-opened once and 95% fewer than 5 times.

NSF-funded U.C. Santa Cruz Study on Data Access Patterns

The Challenge:
How to manage this?
Options Have Tradeoffs

- Just buy more primary disk
  - No reduction in storage costs, silos expand, backup volumes grow
  - Run out of DC space

- Use primary-class storage to archive data
  - Storage costs remain high
  - Problem continues to escalate

- Leverage cloud storage strategies
  - With no pre-defined Data Management or Rationalization Strategy
    - Uncontrolled growth and cost
    - Unpredictable costs
    - No search, organizational capabilities
    - Chaos in the cloud
Options Have Tradeoffs

- Backup Applications
  - Not designed for long term data preservation
  - Separate application – search and eDiscovery challenges
  - Does not address original datasets still sitting on primary storage
### Backup vs Archive

Backup and Archive are often confused.

<table>
<thead>
<tr>
<th>Backup</th>
<th>Active Archive</th>
</tr>
</thead>
<tbody>
<tr>
<td>Short-term data protection of day-to-day data</td>
<td>Long-term protection and accessibility to all data</td>
</tr>
<tr>
<td>Copies data</td>
<td>Moves data off primary storage</td>
</tr>
<tr>
<td>Backup data typically overwritten</td>
<td>Matches data to the right storage tier for long-term access/protection</td>
</tr>
<tr>
<td>Not a target for regulatory compliance</td>
<td>Addresses retention, retrieval, availability and data protection</td>
</tr>
<tr>
<td>No historic relevance</td>
<td>Data is of high value, available when needed</td>
</tr>
<tr>
<td>Not easily searched – It’s offline!</td>
<td>Easily searched – it’s online!</td>
</tr>
</tbody>
</table>
# Backup vs Archive

Backup and Archive are often confused.

<table>
<thead>
<tr>
<th>Backup</th>
<th>Active Archive</th>
</tr>
</thead>
<tbody>
<tr>
<td>Short-term data protection of day-to-day data</td>
<td>Long-term protection and accessibility to all data</td>
</tr>
<tr>
<td>Copies data</td>
<td>Moves data off primary storage</td>
</tr>
<tr>
<td>Backup data typically overwritten</td>
<td>Matches data to the right storage tier for long-term access/protection</td>
</tr>
<tr>
<td>Not a target for regulatory compliance</td>
<td>Addresses retention, retrieval, availability and data protection</td>
</tr>
<tr>
<td>No historic relevance</td>
<td>Data is of high value, available when needed</td>
</tr>
<tr>
<td>Not easily searched – It’s offline!</td>
<td>Easily searched – it’s online!</td>
</tr>
</tbody>
</table>

-- Data value is long term  --
- Hardware lifespan is relatively short term -
All data needs online access –
But at offline costs….

Design criteria for an active architecture:
- Online access to all data, all the time for users and applications
- Contains and/or reduces infrastructure and management costs
- Seamless optimization of data across storage types
- Ideally includes strategy for next generation infrastructure
“We can effectively keep a stable budget and still handle exponential growth with the technology of our active archive - and that's exciting!”

–Jason Hick  Storage System Group Lead  National Energy Research Scientific Computing Center (NERSC)
Active Archive Strategies
Break Down Barriers to Data

Active Online Storage Fabric:
- Data is no longer tied to a particular device or tech…
- Optimize acquisition & operational costs without impact to users…
- Enables proactive IT, not reactive – when managing data growth…
  - No user interruption, even as new technology becomes available…
  - Enables proactive strategy for migration to next generation infrastructure – seamless optimization of data across storage types
  - Ideally includes strategy for next generation infrastructure
Data Touch Layers

- **Data Movers** – Front End Utilities
  - Meta Data Layer (dataset attributes)
  - Open Platform

- **Connectors** – Back End Layer
  - Block, NAS, MAID, Dedupe, Object, Tape

- **Interfaces/API’s**
  - CIFS/NFS
  - HTTP/REST
  - S3
  - CDMI
  - Bulk Object (tape oriented)

- **File Systems**
  - Multiple Disk - Proprietary Options (CXFS, StorNext etc.)
  - Linear Tape File System (LTFS)
  - Archive eXchange Format (AXF)
**Index:** XML data that holds the file metadata and the mapping from files to Data Extents.

**Data Extents:** Set of one or more sequential logical blocks used to store file data.

- lto.org
AXF Tape File Layout

- openaxf.org
How Active Archiving is Done

- **Classify & Index Data**
  - Know what it is, whether it is active or not, what its time value is

- **Index and archive “retention” data...** i.e.: The stuff you really need to keep
  - Know what it is, whether it needs to be active or not

- **Backup only active data**
  - Keep backup & restore times lower

- **Search and Retrieve from archive when it needs to be active**

- **Restore from backup when problems arise with transactional data**
How Active Archiving is Done

- **Classify & Index Data**
  - Know what it is, whether it is active or not, what its time value is

- **Index and archive “retention” data**... i.e.: The stuff you really need to keep
  - Know what it is, whether it needs to be active or not

- **Backup only active data**
  - Keep backup & restore times lower

- **Search and Retrieve from archive when it needs to be active**

- **Restore from backup when problems arise with transactional data**
How Active Archiving is Done

- **Classify & Index Data**
  - Know what it is, whether it is active or not, what its time value is

- **Index and archive “retention” data**... i.e.: The stuff you really need to keep
  - Know what it is, whether it needs to be active or not

- **Backup only active data**
  - Keep backup & restore times lower

- **Search and Retrieve from archive when it needs to be active**

- **Restore from backup when problems arise with transactional data**
How Active Archiving is Done

- **Classify & Index Data**
  - Know what it is, whether it is active or not, what its time value is

- **Index and archive “retention” data**... i.e.: The stuff you really need to keep
  - Know what it is, whether it needs to be active or not

- **Backup only active data**
  - Keep backup & restore times lower

- **Search and Retrieve from archive when it needs to be active**

- **Restore from backup when problems arise with transactional data**

**Keep data where it is most cost effective according to its time-value**
Benefits of Active Archive

Key Benefits to Active Archive Architectures:

- Reduces overall cost:
  - More data available in user-accessible “online state” at significantly lower cost / TB.
  - By making lower power storage solutions available, energy & cooling costs are reduced.
  - Enables IT to better manage data growth proactively, rather than needing to be reactive to spikes with the most expensive solution.

- Increases data protection & availability:
  - Most active archive solutions decouple the data from a particular type of hardware. This means IT can be plan for forward migration with no disruption to current workflow.
Multiple Solution Choices

- There is no one-size fits all solution:
  - Software choices:
    - HSM (Tier Virtualization)
    - Hybrid solutions
    - Digital Asset Management
  - Storage Medium choices:
    - Flash Technologies
    - Disk technologies
      - Block, File, Object
    - MAID – Power-down drive technologies
    - Tape as “online storage”
      - Virtualized with disk in many ways
      - Self-describing technologies; AXF, LTFS...
    - Cloud (public & private cloud methodologies)

Most Active Archive solutions will include some or all of these.
Thank you

- For more visit the Active Archive Alliance website:
  http://www.activearchive.com

- Or contact me directly:
  Stacy Schwarz-Gardner
  stacys@spectralogic.com