TDWI WEBINAR SERIES

The Pros and Cons of Data Warehouse Appliances

Philip Russom
Senior Manager of Research and Services
TDWI: The Data Warehousing Institute
prussom@tdwi.org
www.tdwi.org
Agenda

• Data Warehouse Appliances
  – Definitions
  – Strategies
  – Pros
  – Cons
• Conclusions
• Recommendations
Data Warehouse Appliance Definitions

What do you think a data warehouse appliance is?

- 53% = Server hardware and database software built specifically to be a data warehouse platform
- 13% = Either definition
- 14% = Any server hardware and database software bundled to create a data warehouse platform
- 19% = Don't know

Respondents feel they know what a data warehouse appliance is:
- Most think the hardware and software are built for data warehousing
- But a minority think it could be any hardware-software bundle

Appliance and Bundle Vendors

- **Data warehouse appliances**
  - “Server hardware and database software built specifically to be a data warehouse platform”
  - Netezza, DATAllegro, Calpont, and Teradata (maybe)

- **Data warehouse bundles**
  - “Any server hardware and database software bundled to create a data warehouse platform”
  - IBM DB2 Integrated Cluster Environment (ICE) for Linux, Sun-Sybase iForce Solutions, Unisys ES7000 BI Solutions

- **Miscellaneous appliances** (outside data warehousing)
  - Network Appliance (storage), Google Search Appliance, Thunderstone Search Appliance
Data Warehouse Appliance Strategies

What’s your group’s strategy for data warehouse appliances?

- **Most respondents don’t plan to evaluate a DWA.**
- **A surprisingly high number (18%) say they’ve already deployed one.**
- **One-fifth are evaluating one now or in the near future.**

Sweet Spot – Large Data Marts

• Most appliances/bundles support “large data marts”:
  – Vendors describe their customers/prospects this way.
  – Users describe their implementations this way.
  – Mart enables an analytic app, typically focused on analysis of call-level detail, customers, shopping baskets, etc.

• The large data mart is a “sweet spot” for appliances/bundles:
  – Users have succeeded with this kind of project.
  – Strategy: Isolated project to prove appliance before expanding.

• Other sweet spots are coming.
  – A few users with appliances (and many with similar bundles) have deployed an enterprise data warehouse.
  – Eventually, the EDW will be another sweet spot.
Data Warehouse Appliance Pros

What do you think is the leading benefit of a data warehouse appliance?

- 38% = Pre-tuned for data warehousing
- 15% = Fast query performance
- 13% = Reduced system integration
- 13% = Fast installation
- 8% = Low cost
- 7% = Easy incremental expansion
- 6% = Other


- Users want guaranteed performance:
  - Users felt pre-tuning and fast queries were most compelling benefits.
- Few users expect low cost, though vendors compete on this point.
Sweet Spot – 1TB to 10TB of Data

- Most appliances today manage between 1TB and 10TB of data:
  - This is another way to describe the “sweet spot,” whether large data mart or enterprise data warehouse.
- Sweet spot will shift with data growth:
  - TDWI data: mid/large marts/DWs ~15% annual growth.
  - Appliance users start with 1-3TB, grow toward 10TB.
  - Some appliances now deployed with >10TB.
- Vendors offer appliances of great capacity:
  - Netezza 8650 (27TB), 10400 (>50TB), 10800 (>100TB)
  - DATAllegro C25 (25TB)
- At this rate of growth, very large data warehouses (>10TB) will be common on data warehouse appliances within 2 years.
Data Warehouse Appliance Cons

What do you think is the leading problem with a data warehouse appliance?

- 44% = Proprietary platform that makes migration off it difficult
- 27% = Single-use hardware that cannot be re-allocated to non-warehouse use
- 10% = Black box that resists tuning
- 9% = New training for the new platform
- 7% = Don’t Know
- 3% = Other


- Users’ main concern was migrating off an appliance.
  - *My perspective: No harder than other relational migrations.*
- Single-use hardware is the reality of all appliances.
Open Source in Data Warehousing

- Adoption of open source in data warehousing is low.

  Data warehouses on open source operating systems:
  - 41% = No plans
  - 9% = Deployed

  Data warehouses on open source databases:
  - 64% = No plans
  - 8% = Deployed

- Open source databases and operating systems are typical components of DW appliances and similar bundles.

- Lack of interest in open source among data warehouse professionals could be a barrier to appliance adoption.
Conclusions

- Definition of a data warehouse appliance:
  - “Server hardware and database software built specifically to be a data warehouse platform”
  - But similar bundles offer similar advantages.
- Pros:
  - Pre-tuned for DW, fast queries, less system integration
- Cons:
  - Proprietary platform, single-use hardware
- Strategy – Succeed with Sweet Spots:
  - 1-10 terabyte databases (but growing)
  - Terabyte-size data marts (but enterprise DWs are possible)
- Expect 10TB+ enterprise data warehouses to be common on data warehouse appliances in 2 years or so.
Recommendations

- Consider a data warehouse appliance for:
  - Analytic app with terabyte-size data mart
  - Intense queries not appropriate to EDW
  - Short dev/deployment time
  - Sponsor wanting low price per TB
  - Budget needing minimal SI cost
  - Apps without a FTE as administrator

- Other users have succeeded in these situations – you can, too.
Stuart Frost, CEO
What is a DATAAllegro DW Appliance?

- Complete DW solution
  - “From SQL to storage”
- Modular, rack-based appliances
  - True commodity hardware platform
  - Standard Intel motherboard
  - Western Digital enterprise disks
  - Patent-pending architecture
  - Linear scaling
  - Fault tolerant
  - Leverage open source Linux & Ingres
- DATAAllegro P2™ & P5™
  - Very high performance
  - 2TB user data - $250k
  - 5TB user data - $450k
- DATAAllegro C25™
  - Good performance
  - 25TB user data - $450k
  - $18k per TB
- Encrypted versions available
Commodity HW & RDBMS

Open Source DBMS highly tuned for DSS

Commodity components chosen for speed & reliability

Standard 2U chassis

Flexible CPU/disk balance rides multi-core wave

800MBps sustained read/write speed per node

Ingres® r3

6 GB RAM

6 x SATA – RAID0

6 x SATA – RAID0

Flexible CPU/disk balance rides multi-core wave
Redundant Array of Inexpensive DW (RAIDW™) - MPP with FT & Low Cost

20Gbps redundant Infiniband network

Master

Slave array

Failover pair
Ultra Shared Nothing (USN) – Multi-level Partitioning + Replication

- Multi-level hash and/or range partitioning within nodes
- Hash partitioning across nodes

FAST LOAD

ETL flat file >300GB per hour via dual GbE

Tables can be partitioned and/or replicated
- Speeds joins & reduces net traffic

Partition locking allows real-time updates with minimal impact on queries
Direct Data Streaming (DDS) – Sequential I/O, no tuning or indexes

Master breaks query into steps that run efficiently on Ingres with minimal/no tuning or indexes

ODBC/JDBC
SQL92 with many SQL99/Oracle/Teradata extensions

Individual steps run on slaves with 98% sequential I/O - Faster, more reliable
DW Appliance Pros

<table>
<thead>
<tr>
<th>Percentage</th>
<th>Feature</th>
<th>✔️</th>
<th>✔️</th>
<th>✔️</th>
<th>✔️</th>
<th>✔️</th>
<th>✔️</th>
<th>✔️</th>
<th>✔️</th>
<th>✔️</th>
</tr>
</thead>
<tbody>
<tr>
<td>38%</td>
<td>Pre-tuned</td>
<td>✔️</td>
<td>✔️</td>
<td>✔️</td>
<td>✔️</td>
<td>✔️</td>
<td>✔️</td>
<td>✔️</td>
<td>✔️</td>
<td>✔️</td>
</tr>
<tr>
<td>15%</td>
<td>Fast query performance</td>
<td>✔️</td>
<td>✔️</td>
<td>✔️</td>
<td>✔️</td>
<td>✔️</td>
<td>✔️</td>
<td>✔️</td>
<td>✔️</td>
<td>✔️</td>
</tr>
<tr>
<td>13%</td>
<td>Reduced system integration</td>
<td>✔️</td>
<td>✔️</td>
<td>✔️</td>
<td>✔️</td>
<td>✔️</td>
<td>✔️</td>
<td>✔️</td>
<td>✔️</td>
<td>✔️</td>
</tr>
<tr>
<td>13%</td>
<td>Fast installation</td>
<td>✔️</td>
<td>✔️</td>
<td>✔️</td>
<td>✔️</td>
<td>✔️</td>
<td>✔️</td>
<td>✔️</td>
<td>✔️</td>
<td>✔️</td>
</tr>
<tr>
<td>8%</td>
<td>Low cost</td>
<td>✔️</td>
<td>✔️</td>
<td>✔️</td>
<td>✔️</td>
<td>✔️</td>
<td>✔️</td>
<td>✔️</td>
<td>✔️</td>
<td>✔️</td>
</tr>
<tr>
<td>7%</td>
<td>Easy expansion</td>
<td>✔️</td>
<td>✔️</td>
<td>✔️</td>
<td>✔️</td>
<td>✔️</td>
<td>✔️</td>
<td>✔️</td>
<td>✔️</td>
<td>✔️</td>
</tr>
</tbody>
</table>
Performance Comparison

Teradata vs. DATAllegro P3 Benchmark Results

Query Type

DATAllegro Timings
Legacy Platform Timings
**DW Appliance Cons**

- **44%** Proprietary = lock-in
  - Standard interfaces (ODBC etc.)
  - DA uses commodity HW

- **27%** Single use HW
  - Typ. used for new DM etc.
  - Low cost mitigates

- **10%** Resists tuning
  - DA allows indexes, some tuning
  - Performance limits need for tuning

- **9%** New training
  - Some training required, but far less than traditional platforms
Summary

- DW appliances are here to stay:
  - Price
  - Performance
  - Scalability
  - Ease of installation and use
  - Two successful vendors

- DATAllgro addresses most concerns raised by TDWI community