The Role of the Analytic Database in a Modern Data Warehouse Architecture

Philip Russom
Research Director, TDWI
September 27, 2012
Speakers

Philip Russom
TDWI Research Director,
Data Management

Chris Selland
VP of Marketing,
Vertica
Agenda

• Background and Definitions
  – Trends driving up ADBMS use
  – Analytic Database Mgt Systems (ADBMSs)

• Data Warehouse (DW) Architectures
  – Workloads and Edges
  – Monoliths and Hybrids

• Uses of ADBMSs
  – Methods for advanced analytics
  – Departments and discovery analytics
  – Business analysts and data scientists

• The Future & How to Prepare for It
Background

- TDWI 2011 Big Data Analytics Survey
  - 33% contemplating replacement of data warehouses, analytic databases, and similar platforms
  - To adjust to new business requirements for analytics, real time, big data

- An increasingly common reaction is…
  - Users are turning to specialized analytic database management systems (ADBMSs)

- TDWI Technology Survey in May 2012
  - Half of organizations surveyed have already deployed one or more standalone analytic databases
  - Another third plan to deploy their first within three years
Why the increase in standalone Analytic Databases?

- Change is rampant in business
  - Analytics helps discover what changed & how to react

- There are opportunities to leverage
  - Advanced analytics is the best way to find new customer segments, best suppliers, products of affinity, sales seasonality, etc.

- Even organizations with a mature EDW find that analytics requires:
  - A database mgt system purpose built for analytics
  - A separate platform for analytic workloads, to avoid impacting traditional DW workloads

- Increasing focus on departmental BI
  - Analytics is inherently departmental
  - Some depts can afford own ADBMS
DEFINITIONS

Analytic Database, etc.

• Database management system (DBMS)
  – Vendor-built software for managing databases

• Database
  – A collection of data managed by a DBMS
  – Or colloquial for database management system

• Analytic database management system (ADBMS)
  – Vendor-built DBMS designed to manage data for analytics & DW
  – Or just call an ADBMS an “analytic database”

• Vendors package or market their ADBMSs a variety of ways
  – DW appliances, columnar DBMSs, analytic accelerators, in-memory DBMSs, multi-tool analytic platforms, cloud- or SaaS-based analytic applications or platforms.

• Most ADBMSs are relational, focused on SQL-based analytics
  – Alternatives: NoSQL or Hadoop for multi-structured data
DEFINITION

Advanced Analytics

OLAP & its Variants
• Users have this
• They’ll keep & grow it
• OLAP won’t go away

Advanced Analytics
• Discovery oriented
• Excels with Big Data
• Demands special platforms: ADBMS
• Experiencing strong adoption by users

– Online Analytic Processing (OLAP)
  – It’s somewhat rudimentary, but required.
  – Demands multidimensional data modeling, but works well with most EDWs.
  – There are multiple approaches to OLAP.

– Extreme SQL
  – Uses well-known SQL-based tools & techniques.
  – Relies on long, complex SQL statements to define recent business events.

– Predictive Analytics
  – Uses data mining and/or statistics to anticipate future events.
  – Requires special tools and training.

– Other Analytic Methods
  – Visualization, artificial intelligence, natural language processing.
  – ADBMS functions: columnar data stores, DW appliances, MapReduce, etc.
THE CATCH

“Square Peg” Analytic Workloads may not fit “Round Hole” DW Architectures

• Most data warehouses were designed and optimized for common deliverables and methods:
  – Standard reports, dashboards, performance mgt, online analytic processing (OLAP)
  – This is a design and architectural decision made by users, not a failing of vendor platforms

• Can/should all DW & analytic workloads run on your EDW?
  – If your EDW can handle multiple mixed concurrent workloads with performance and without impeding other workloads, then run all workloads (including analytics) on the EDW, for simplicity’s sake
  – If not, you may need additional platforms for some workloads, including an ADBMS for analytic workloads
Workloads for Analytics etc Affect DW Architecture

• System on the Side (SOS) or Edge System
  – A workload and its data that’s deployed on a platform separate from the EDW
  – Usually integrates with EDW via shared data or data models

• Long-standing tradition of SOSs w/EDWs
  – Data marts, operational data stores (ODSs), data staging areas, file systems (for flat files, documents, logs)
  – Workload types: analytics, real-time, detailed source data, unstructured data

• Trend: Analytics is driving up SOSs
  – Each analytic method (or even each analytic application) may need its own SOS

• Outcome
  – For analytic workloads on the edge of a distributed DW architecture, users are turning more and more to standalone ADBMSs
Monolithic EDWs vs Distributed Architectures

• Monolithic DW Architecture – EDW
  – All or most BI/DW workloads via a single DBMS instance for the EDW
    • Usually involves mart/ODS consolidation; sometimes a change of DBMS platform for the EDW; “Green field” EDWs may start with a single DBMS
    • Requires a hefty DBMS platform and a great user design to handle so-called “mixed workloads” = multiple, diverse, concurrent DW workloads

• Distributed DW Architecture – EDWE
  – Users deploy separate DBMS instances outside and alongside the EDW for nonstandard workloads
  – Warning: If not controlled, data marts, ODSs, analytic databases may proliferate. Complexity increases, which deters standards, tuning, etc.

• Hybrid DW Architecture
  – Monolith managing core reporting/OLAP data, plus most workloads
  – Only a few workloads are deployed on separate systems
    • Offload invasive or unpredictable analytic workloads, like extreme SQL
Benefits of Complementing DW with ADBMS

What are the potential benefits of complementing an EDW with an ADBMS? Select all that apply.

<table>
<thead>
<tr>
<th>Benefit</th>
<th>Percentage</th>
</tr>
</thead>
<tbody>
<tr>
<td>Optimized for more analytic workloads than our EDW</td>
<td>58%</td>
</tr>
<tr>
<td>Enables the &quot;analytic sandboxes&quot; that many users need</td>
<td>57%</td>
</tr>
<tr>
<td>Provides analytic and data mgt capabilities that complement our EDW</td>
<td>56%</td>
</tr>
<tr>
<td>Isolates ad hoc analytic work that might degrade EDW performance</td>
<td>46%</td>
</tr>
<tr>
<td>Manages multi-Tb raw source data for analytics better than EDW</td>
<td>33%</td>
</tr>
<tr>
<td>Handles real-time data feeds for analytics better than EDW</td>
<td>29%</td>
</tr>
<tr>
<td>Takes analytic processing to Big Data, instead of reverse</td>
<td>22%</td>
</tr>
<tr>
<td>Other</td>
<td>3%</td>
</tr>
</tbody>
</table>

SOURCE: TDWI Technology Survey in May 2012. 219 responses from 72 respondents.
Benefits of ADBMSs over Traditional DBMSs

What are the benefits of data warehousing with an ADBMS instead of a DBMS designed for OLTP? Select all that apply.

- Purpose-built for data warehousing and analytics: 71%
- Fast query performance or model rescores: 68%
- Enables new business practices based on analytics: 58%
- MPP computing architecture instead of SMP: 29%
- Cheap license compared to leading OLTP DBMSs: 7%

Use Cases for Analytic Databases

• ADBMSs excel with exploratory analytics. Discover:
  – Customer base segments
  – Customer behaviors and their meaning
  – Forms of churn and their root causes
  – Relationships among customers and products
• Analyze big data you’ve hoarded. Finally understand:
  – Web site visitor behavior
  – Product quality based on robotic data from manufacturing
  – Product movement via RFID in retail
• Analytics is inherently departmental. And some departments can afford their own ADBMS and staff:
  – Customer analytics in marketing & sales
  – Supplier analytics in procurement
  – Product development analytics in the R&D departments of firms for pharmaceuticals, petroleum, consumer package goods, automotive, etc.
• ADBMSs bring performance to older analytic apps:
  – Fraud detection, Risk management, Actuarials
  – “Fork lifts” of old data onto new ADBMS work well
• Empower business analysts and data scientists
  – ADBMS is powerful & affordable for small user count
• Put your entire EDW on an Analytic DBMS:
  – A few user organizations have done this successfully
A Look Into the Future of Analytic Databases

- The definition of “analytic database” will evolve
  - As user requirements for analytics continue to evolve
- ADBMSs will become a common part of DW archs.
  - But usually as secondary platforms, akin to data marts & ODSs today
- EDWs on Analytic DBMSs exist today
  - These will eventually will be common, but not norm
- Most ADBMSs can handle big data volumes
  - Half-petabyte analytic databases be common in 3 yrs
- Vendor products for ADBMSs will keep delivering greater scale and speed
  - Expect more analytic functions in ADBMSs for statistics, mining, predictive, graphs
- Users’ slow trend toward distributed DW architectures will continue
  - Driven by new, demanding workloads for real-time, big data, & multi-structured data – not just analytics
  - Yet, monolithic EDWs won’t go away
- Advanced analytics is a new competency
  - Organizations will hire & train, plus acquire new tools & platforms – including standalone Analytic DBMSs
Recommendations

• Choose the analytic approaches that you need
  – OLAP is good and it’s not going away; complement it with discovery oriented methods & advanced analytics
  – Select database platforms and analytic tools that are appropriate to methods chosen

• Foster your EDW as a killer platform for reporting and OLAP.
  – Consider offloading analytic workloads to a DWA or analytic DBMS.

• Be open to alternative architectures
  – Systems on the Side (SOSs) have a place, but you must control them
  – Both DW and DI architectures need adjustments to accommodate analytics

• Be open to new or alternative DW platforms, not just traditional ones.
  – New ADBMS types and brands give us more options, so at least consider them:
    • Analytic DBMSs, Data Warehouse Appliances, Columnar databases, No-SQL, etc.
  – Also: Hadoop, Map Reduce, Clouds for DW/BI & analytics, SaaS

• Take command of your architecture(s).
  – Analytics is driving up DW architecture complexity
  – Know the biz/tech requirements per analytic app & design arch accordingly
Turning data into answers

HP Vertica, Big Data & Real-Time Analytics

Chris Selland - HP Vertica
cselland@vertica.com
The explosion of data is not news to anyone …

Every 60 seconds

- 98,000 tweets
- 23,148 +apps downloaded
- 400,710 ads requests
- 208,333 minutes Angry Birds played
- 2000 lyrics played on Tunewiki
Neither are the challenges

50%  98%  34%  35%

Do not have an effective information strategy in place

Cannot deliver the right information at right time

Say half their information is unused

Are not effective at accessing enterprise information

* Source: Coleman Parkes, October 2011
Legacy architectures were built for a different world

Yesterday’s data warehouse and analytic infrastructure

- Proprietary
- Expensive
- Centralized, monolithic
- Process laden
- Batch
- Summary
- Slow
Vertica, purpose built for answers in near real time

50x-1000x faster performance at 30% the cost, proven by more than 600 customers

Volume  

Variety  

Velocity

Value

© Copyright 2012 Hewlett-Packard Development Company, L.P. The information contained herein is subject to change without notice.
Designed for answers from the very first line of code, Vertica technology makes the difference

**Columnar storage and execution**
Achieve best data query performance with unique Vertica column store

**Clustering**
Add resources on the fly with linear scaling on the grid, commodity hardware

**Compression**
Store more data, provide more views, 90% less storage required

**Continuous performance**
Query and load 24x7 with zero administration

**Database design**
Automated performance tuning

**Advanced analytics**
Time-series, geospatial, click-stream and an SDK for more
More than 600 customers in multiple industries are finding answers

- Promotional Testing
- Claims Analyses
- Patient Records Analyses
- Clinical data Analyses
- Fraud Monitoring
- Financial tracking
- Tick data back-testing

- Behavior Analytics
- Click Stream Analyses
- Network Analyses
- Customer Analytics
- Compliance Testing
- Loyalty Analysis
- Campaign Management
Medical innovation saves lives—and money

BlueCross BlueShield Association

54M lives, 3.5B man-months of health care data
• Unpredictable customer demands, long batch queues
• With Vertica, response time reduced to minutes/seconds, batch wait time is 0
• Now being used for modeling of healthcare costs, chronic illness prevention, detection of claims fraud, off-brand drug use, and more

“Started small with one research effort, result set came back so fast that we thought the queries had failed!
Today we’ve moved the entire analytics stack to Vertica! Thrilled to be a Vertica customer!”

Doug Porter, Senior Vice President and CIO of BlueCross BlueShield Association

Did you know? Vertica is often used by data scientists in conjunction with tools like SAS or R for deep predictive modeling
Real-time engagement marketing

Zynga, Inc.

• Personalized gaming experience drives reach, revenue and retention.

• Monetizing individual customer behavior – finding the influencers

• One of the largest Vertica implementations with over 3TB loaded every day

“With over 40 million players, 3TB of data loaded every day and 230 nodes spread across two clusters Zynga columnar data warehouse from Vertica is no analytical windup toy.”

- Ken Rudin, VP Analytics

Did you know? Vertica is a scalable platform for graph analytics
Retail sales insights in real time

GUESS, Inc.

- Replaced Oracle-based POS data warehouse that wouldn’t scale
- 100x faster performance using Vertica, 24x7 loads, minimal admin overheads
- Every store manager has access to the data on their mobile device using Microstrategy

Retail moves at lightning speed so we needed a high-performance analytics platform that could handle our fast-paced requirement for information.”

- Mike Relich, CIO, Guess, Inc.

Did you know? Vertica connects to a broad ecosystem of BI products via standard interfaces like ODBC/JDBC and ADO.NET

© Copyright 2012 Hewlett-Packard Development Company, L.P. The information contained herein is subject to change without notice.
Vertica and Hadoop – Best Tools for the Job

- Designed for Performance
- Interactive Analytics
- A Rich SQL Ecosystem

- Designed for Fault Tolerance
- Batch Analytics
- A Rich Programming Model

Both purpose-built scalable analytics platforms

Evaluate Vertica!

Enterprise Edition
• Free 30 day evaluation

Community Edition
• Free Download 1TB, 3 nodes

www.vertica.com
Questions?
Contact Information

If you have further questions or comments:

Philip Russom, TDWI
prussom@tdwi.org

Chris Selland, Vertica
cselland@vertica.com