Agile Data Warehousing with Production Sandboxing

Stephen Brobst
Chief Technology Officer
Teradata Corporation
stephen.brobst@teradata.com

Oliver Ratzesberger
Senior Director
eBay
oratzesberger@ebay.com
Agenda

• The Business Need
• Implementation Options
• Governance
• eBay Case Study
• Discussion
The biggest enemy to sustained success in data warehousing is stagnation.

Continued delivery of value from a data warehouse demands that organizations aggressively encourage new and creative methods in their use of information and analytics.

If an organization does not evolve and improve its data warehouse, value will diminish over time.

---

**Value of a New Analytic Capability**

- **Initial Deployment**
- **Repeated Use**

![Graph showing the value over time for a new analytic capability]
Value of a New Analytic Capability

When a report or analytic capability is first introduced, it offers the potential for new insight and ways of exploiting information.

- Value initially increases as adoption of the new capability takes place.
- As time goes on those once groundbreaking insights become old news.
- The value of the report or analytic capability begins to decrease over time after the point where the organization has incorporated the insights into its standard operating procedures.

Over time, the data warehouse will be burdened with generating lots and lots of reports that eventually deliver value that is less than the cost of maintaining them.

The Requirement for Innovation

Organizations must engage in two actions to ensure sustained success of a data warehouse:

1. Constantly innovate and develop new capabilities.
2. Prune older capabilities where the value no longer justifies the cost.

It is not enough to encourage and sponsor innovation...organizations must provide an appropriate platform and environment to facilitate innovation in the area of information exploitation.
Marketing has an outside source of data that it wants to bring into the data warehouse, but they are not yet sure if it has high value or not...so we need to experiment with the data before signing the purchase contract.
The Manifesto for Agile Software Development puts forth the following principles:

- Value individuals and interactions over processes and tools,
- Value working software over comprehensive documentation,
- Value customer collaboration over contract negotiation, and
- Value response to change over following a plan.

While it is recognized that there is value in the items on the right, the items on the left are valued even more in the agile development methodology.
The underlying philosophy that drives these values in the context of data warehousing is to put the highest priority on satisfying end user (knowledge worker) requirements through early and continuous delivery of analytic capability.

This does not mean that requirements documents, design documents, entity-relationship diagrams, data dictionaries, etc. are not important - but it does mean that the emphasis is much more on delivery than process.

Goal: Same day availability of data into the analytic environment.
Agile Data Warehousing

Need a way to get data into the data warehouse without the overhead of a full blown development methodology:

• Allow for "load and go" analytics.

• Non-certified content to be used in cooperation with content in the enterprise data warehouse (EDW).

• Limited users and limited use.

Implementation Options
Option 1: Separate Development System

- Deploy the "experimental" data in the development/test environment.

- Configure the development/test environment to the full size of the production environment.

- Demonstrate the value of the data and then (assuming positive ROI) use best practices to bring data into production DW.

Objection: Costs too much to have development/test environment at full size and kept fully up-to-date with production environment.
Option 2: Downsized Development System

- Deploy the "experimental" data in the development/test environment.
- Configure the development/test environment with "sampling" from the production environment.
- Demonstrate the value of the data and then (assuming positive ROI) use best practices to bring data into production DW.

Objection: More difficult to implement prototype with sampled data sets and more difficult to make ROI case.
Option 3: Federated Development System

- Deploy the “experimental” data in the development/test environment.
- Join across the development and production environments to perform the analysis.
- Demonstrate the value of the data and then (assuming positive ROI) use best practices to bring data into production DW.

Objection: Performance sucks when joining large data sets across a network (even when both systems are Teradata!).
Option 4: Production Sandbox

- Deploy the "experimental" data directly into the production environment.
- Separate Teradata "databases" for sandbox data with joins allowed to the production data all on a single system.
- Demonstrate the value of the data and then (assuming positive ROI) use best practices to bring data into production DW.

Objection: End users manage their own space and have create table privileges on the production system.
Option 4: Production Sandbox

Objection: End users manage their own space and have create table privileges on the production system.

Best Answer!

Governance
Controls

- No "production" reporting from the sandbox environment.

- Automated resource governors to prevent "runaway" queries in the sandbox area.

- Data residency no more than XX days.

Promoting Content into the EDW

- Monitoring and "promoting" data content in the sandbox.

- Once the value is proven, use "proper" methodologies to integrate content into the enterprise data warehouse.

- Encourage refinement of requirements as part of the sandbox experience.
The eBay Experience

eBay Analytics Technology Highlights

>50 TB/day of new, incremental data

>100k data elements

>50 PB/day Processed

>50^10 new records/day

>50k chains of logic

>5000 business users & analysts

Active/Active

turning over a TB every 5 seconds

24x7x365 Always online

Millions of queries/day

99.9+% Availability

Near-Real-time
Design for the Unknown

>85% of eBay analytical workload is **NEW & Unknown**

**Exploration** is the core of an analytical company

The metrics you know are 'cheap'

The metrics you don’t know are expensive but also high in potential ROI

Design can’t be **static** or dependent on specific questions or dimensions
Proliferation of Analytics

**Decentralized** Analytics

Shortened **Time to Market** Requirements

**Adhoc Exploration**

**Departmental**

**Data**

Prototyping - Can’t wait for EDW

"We Need Data Marts!"

---

Proliferation of Analytics

**Hub and Spoke** Architecture

[Table showing survey results]

<table>
<thead>
<tr>
<th>Architecture</th>
<th># of Respondents</th>
<th>% of Respondents</th>
</tr>
</thead>
<tbody>
<tr>
<td>Data warehouse with dependent data marts (i.e. &quot;hub and spoke&quot;)</td>
<td>50</td>
<td>42.25%</td>
</tr>
<tr>
<td>Independent data marts (no consistent design)</td>
<td>40</td>
<td>32.14%</td>
</tr>
<tr>
<td>Central warehouse only</td>
<td>30</td>
<td>23.68%</td>
</tr>
<tr>
<td>&quot;Centrally&quot; data mart (two-tier design)</td>
<td>10</td>
<td>8.11%</td>
</tr>
<tr>
<td>Virtual data warehouse (i.e., pull data dynamically from source systems when needed)</td>
<td>5</td>
<td>4.08%</td>
</tr>
<tr>
<td>Not Answered</td>
<td>0</td>
<td>0%</td>
</tr>
<tr>
<td>Total Respondents</td>
<td>118</td>
<td>100%</td>
</tr>
</tbody>
</table>

Source: MIT Enterprise Forum Technology Survey, San Diego Conference, August 2004
Data Mart Dilemma

Total Cost of Ownership (TCO)
Fully loaded cost staggering $500k++
Biggest drivers are
- Maintaining separate databases
- weekly/daily/hourly data transfers
- Data inconsistencies
- Data redundancy
- Increased complexity
- Loss of lineage over time

A Data Mart cannot be 'cheap' enough to justify its existence

Analytics as a Service

Massive scale Analytical Utility Computing

Bring your data - Perform your Analytics

From Simple Web based data upload
...to fully private Utility access

Combine custom data and code with ALL existing data
Analytics as a Service

From a simple web based table upload:

<table>
<thead>
<tr>
<th>Field Type</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>CREATE TABLE SQL</td>
<td>Enter your CREATE TABLE SQL here</td>
</tr>
<tr>
<td>Data File</td>
<td>This is the file containing the data to be uploaded. The schema is: &quot;table_name [field_1, field_2, ...]&quot;</td>
</tr>
<tr>
<td>TB Server</td>
<td>&quot;Select&quot; the server to which the data is to be uploaded. If left blank, the data will be uploaded to the primary production server (Teradata)</td>
</tr>
<tr>
<td>Expiration Date</td>
<td>&quot;Select&quot; the expiration date for the data. If left blank, the data will not expire.</td>
</tr>
</tbody>
</table>

Note: The data file must be in CSV format. If the file name ends in .sql, the data will be treated as SQL. If the file name ends in .csv, the data will be treated as CSV.
We call them PET (Prototyping Environment in Teradata)

More than 75 active right now

In most cases they are small (100GB-5TB)

since all the main data is already in the EDW

They are free to the business units

Improved Time To Market - Days/Weeks versus Months.

Enable the business to do agile prototyping.

Enable the users to "Fail Fast" - Make it easy to try out new ideas.

Eliminate stray Data Marts.
Governance Rule 1

- Keep the production data clean.
  - The data life cycle methodology is there for a reason.
  - Do not “pollute” production data with data of unknown source and validation.
    - Equivalent to a viral injection...and you may not recover.

- Do not inject prototype data into “core” DW data:
  - Data ingest (ETL/ELT) does NOT have access to sandbox.
    - Not even to populate the sandbox.
  - Strictly and conceptually enforced on both Batch and User accounts.

Governance Rule 2

- Prototypes written by experienced personnel:
  - PETs assigned to NAMED personnel.
  - Previous Experience and Training Required.

- Prototype personnel are typically former DW developers who transitioned into a business unit.
  - Speed of implementation.
  - Knowledge of DW processes and methodologies.
  - Knowledge of data.
Governance Rule 3

- Sunset dates must be applied:
  - Hold a post mortem.
  - Retire it or promote it.

- The prototype must not become a "black market" production application.
  - Business cannot depend on them.
  - DW cannot give them appropriate support.

Key Process 1

Pre-defined methods, templates, and rules for setup and teardown:

- Well-defined rules for usage.
- Defined, named owners.
- Pre-defined security templates.
- Pre-defined Help Desk responses to add/drop users.
Key Process 2

Help Desk support is critical.

- Direct access for PET personnel to the most senior architectural and technical personnel.

- "Bidirectional" mentoring:
  - Best and brightest technical resources get closer to the business...
  - Business gets closer to fast and effective implementations.
  - It does not take long for PET personnel to become self-sufficient.

Key Learning

- In-place processes enable "time-to-market" benefits.
  - Put the processes and security in place first.

- Failure = Learning
  - Do so with great effectiveness...
  - Fail fast, fail early.

- Most business units now maintain a permanent sandbox.
  - Complex analysis and decision making within a business day!
Questions?

References

