Bringing Data Integration to a New Horizon by SOA, GRID Computing, and Integrated Metadata Server

MDM – METADATA – GRID - SOA

Timothy G. Davis, Executive Director
Center of Excellence for Data Integration (CEDI)
IBM Software
The Typical Enterprise Today

- Fragmented tools and different approaches
- Redundant business rules and “multiple versions of the truth”
- Incomplete information flows
- Inaccurate, incomplete, inconsistent data
- Inability to keep pace with growing data volumes and velocity
- Fragile, complex, hard-coded infrastructure

Root Causes:
- SMP/Shared Memory Systems Lack Scaleability
- Results In Proliferation of “Server Farms” in the enterprise.
- Results in Proliferation of Point to Point (P2P) interfaces in the enterprise
- Results in multiple silos of information, and multiple version of the truth.
Data Integration is the Key to Enterprise Integration

- Merging of Batch, Real Time, and Transaction Processing Systems
- Data Quality, Alignment, Harmonization
- Synchronize data to deliver single view of business operations
- Consolidate redundant applications & systems
- GRID Computing
- High Availability Computing
Building An Enterprise Data Integration Capability

Why Most Companies Fail To Make This Transition:

- This is not an ETL problem, an EAI problem, or a BPM problem.
- This is an Enterprise Data Integration problem.
- Making this transition requires implementing an Integrated Infrastructure.
- Implementing an Integrated Infrastructure has involved organizational and technical challenges beyond the capabilities of most large corporations.
  - Architecture, Methodology, Software Technology, Best Practices
Mizuho Brokerage firm accidentally offered to sell 610,000 shares of a small recruiting company for one yen, instead of one share at 610,000 yen.

Total loss from this botched stock transaction is over $335 Million.

“Mizuho Securities Co. will pay a hefty premium to buy back shares in a small recruiting company that Mizuho sold my mistake.”

“Nikkei Index took a 2% tumble on the day of Mizuho’s mistake.”

**Accurate Reference Data, Master Data, and Business Rules Could Have Prevented This.**
Building An Enterprise Data Integration Capability

The Solution:

- Master Data Management - MDM
- Integrated End To End Metadata
- Architected Information GRID
- GRID/Shared Nothing Systems Provide Unlimited Scaleability
- Prevents the proliferation of “Server Farms”.
- Reduces Point to Point (P2P) Interfaces.
- Provides a Framework for Enterprise Data Virtualization, IOD, and SOA.
Data Integration For Master and Reference Data is the Key

Master Data Management Applications

- Configured End User Interface
- Cataloging
- Reporting
- Search & Review
- Key Management
- Index Management
- Hierarchy Management
- Master Data Model
- Entitlement
- Ongoing Data Quality Monitoring
- Data Synchronization to Applications
- Messaging

Enterprise Data Quality Integration Services

- Source Data Profiling
- Ongoing Data Quality Monitoring
- Standardization
- Global Address Verification & Certification
- Data Enrichment
- Matching & Record Linkage
- Transition, Connection & Loading
- Metadata Management

IBM IIS

- Implementation and Methodology Resources
- Database Management System
- Operating System
- Scalable Hardware Platform

Harmonized

Enterprise Inconsistent Dirty Data

IBM INFORMATION ON DEMAND 2006
## Fundamentals Of Enterprise Data Integration

### Analysis & Mapping
- Understand physical and logical data characteristics
- Determine data consistency and cleanliness
- Discover related data elements
- Facilitate gap analysis

### Standardization & Quality
- Match and reconcile related data elements
- Remove duplicate, redundant data
- Reengineer data to match single corporate standard
- Quality processing to ensure data integrity

### Transformation Harmonization
- Enable creation of data
- Load to EDW, Datamart, ODS, ERP, CRM Applications
- Interface with risk models & reporting applications
- Maintain audit trail

### Transaction Synchronization
- Enable Synchronization of Legacy and New Systems
- Transaction Replication with Commit and Roll Back.
- Enhance Real Time Delta Processing
- Enables Parallel Run

---

**IBM INFORMATION ON DEMAND 2006**

**TAKE BACK CONTROL**
Data Integration Architecture In Action!

Source Data

Data Extraction

Stage

Data Assessment
- Discover Validate

Requirements
- Business and Technical Specifications

Data Alignment
- Translate Standardize (Split Domains)

Translation Tables

Aligned Data Structure

Data Harmonization
- Integrate Standardize Consolidate

Cross Ref & Match Index Tables

Harmonized Data Structure

Solution Implementation
- Prepare Transform Load

OLTP

EDW

Other

Flow of Data

Information Flow

Extended Teams

Core Team

Business Solution Level
- system movement

Integrated Across Sources
- many to 1 movement

Individual Source Level
- 1 to 1 movement

Integrated

Harmonized

Data Structure

Edw

Flow of Information

Integrated Across Sources
- many to 1 movement

Individual Source Level
- 1 to 1 movement

Extended Teams

Core Team

IBM INFORMATION ON DEMAND 2006
The Importance Of Standards For Enterprise-Wide Metadata

- A Business Case for Metadata Management
  - Corporate functions that benefit from Metadata Management
    - Compliance Auditing and Reporting (Sarbanes-Oxley & BASEL II)
    - Business Intelligence
    - Financial Reporting
A Business Case for Metadata Lineage

Tracing the Data Lineage

This shows the flow of data from the source to what is stored and ultimately displayed. In order to provide Lineage on Total Sales Amt we must follow the path this column traveled and discover what occurred to it.

Data Lineage

Historical processing and the data inputs used to create a particular source data set.
A Business Case for Metadata Lineage

Why is it Important?

– Senior management is under scrutiny lately to be accountable for the numbers in their company’s financial reports. ---- Surviving Audits !!!!!!

– Tracking change, tracing metadata lineage, and corporate compliance enforcement is mission-critical, it’s all about Data Governance!
A Business Case for Metadata Lineage

How Compliant is your Enterprise Data?

Today inconsistent sales total amounts, data names, are in some cases reconciled by maintaining spreadsheets, manually trying to trace back Lineage, etc.
A Business Case for Metadata Lineage

- Corporate IT departments are the ones typically saddled with the responsibility of meeting regulatory requirements.

- From an auditing point of view lineage is a very powerful reconciling capability for auditors.

- It helps auditors discover sources and targets of data, and how the data came to have the value shown by documenting data elements, business rules (logic), and source to target mappings.
The Replacement For MetaStage Is Here Today!

Basel II Metadata Architecture

This part of the process is only long as customer is using Data
Evolution of Computing Technology

New GRID Platforms Available Now

- Uni-processor
- LAN Server
- SMP
- Massively Parallel Grids
- Cluster
How Data Integration Grid Works

Before Grid

ProfileStage Project 1

DataStageEE Project 2

QualityStage Project 3

DataStageEE Project 4

IBM Software Project N

“Siloed” architecture & proliferation of SMP servers:

• Higher capital costs through limited pooling of IT assets across silos
• Higher operational costs
• Limited responsiveness due to more manual scheduling and provisioning
• Inherently more vulnerable to failure
How Data Integration Grid Works

**After Grid**

- **ProfileStage Project 1**
- **DataStageEE Project 2**
- **QualityStage Project 3**
- **DataStageEE Project 4**
- **IBM Software Project N**

DataStage EE Multi-Process Grid Framework

Eliminates The SMP nightmare, Allows Unlimited Scaleability
The Evolution of Software Technology
From Multi-Threaded to Multi-Process

**Dedicated Disk**
- Disk
- CPU
- Memory

**Shared Disk**
- Disk
- CPU
- CPU
- CPU
- CPU
- Shared Memory

**Shared Nothing**
- Disk
- Disk
- Disk
- Disk
- CPU
- CPU
- CPU
- CPU
- Memory
- Memory
- Memory
- Memory

**Uniprocessor**
- PC
- Workstation
- Single processor server

**SMP System**
- IBM, Sun, HP, Compaq
- 2 to 64 processors
- Majority of installations

**Grid and MPP**
- 2 to hundreds of processors
- MPP: IBM and NCR Teradata
- each node is a uniprocessor or SMP
The “GRID Blade” Data Integration Paradigm

Application Assembly: One Dataflow Graph Created With the DataStage GUI

Source Data → Transform → Clean → Load → Data Warehouse

Source → Transform → Clean → Load → Target

Application Execution: Sequential or Parallel

32 4-way nodes

Hardware Platform

Uniprocessor

SMP System (Symmetric Multiprocessor)

128 Processor GRID Blade System
Enterprise Grid Computing

Machine assignments for this job

Scheduler and Resource Management

Resource Scheduler

Batch Requests

Real-time Requests

Batch

Requests

Appl Logic

Storage

DB Write

Appl Logic

Storage

DB Write

Appl Logic

Storage

DB Write

Appl Logic

Storage
Enterprise Grid Computing

Grid Hardware

High Availability Cluster Configuration

Servers are in use, or down, or available

Farm of servers

Different computers provisioned for different activities
Common Infrastructure: Batch & On-Demand

Workload/Resource Usage

% Utilization

Hour of Day

Batch

On-Demand
Cost Effectiveness of Grid

- For the same $$$ Grid hardware has much greater throughput!
### Why Information Integration Grid Now?

#### Large Data Service Provider Gains From Grid Versus SMP

- 60X increase in processing throughput
- 20X to 40X increase in hardware price/performance
- $150 million annual capital budget cut by 75 to 80%
- Solve new, high value problems for customers
- >$1 billion in new revenue contracts generated

### Acxiom efficiency gains

<table>
<thead>
<tr>
<th></th>
<th>Former approach (SMP)</th>
<th>Customer information infrastructure</th>
</tr>
</thead>
<tbody>
<tr>
<td>661% increase to batch processing capacity</td>
<td>7 million records/hour</td>
<td>54 million records/hour</td>
</tr>
<tr>
<td>99% time reduction for analytical processing (21.5 million records)</td>
<td>8 hours</td>
<td>5 minutes</td>
</tr>
<tr>
<td>83% reduction in days required to post new data</td>
<td>22 days</td>
<td>3 days</td>
</tr>
<tr>
<td>89% update time reduction</td>
<td>7.7 million records/hour, 27 days</td>
<td>70 million records/hour, 3 days</td>
</tr>
<tr>
<td>50% cycle time reduction</td>
<td>8 weeks</td>
<td>4 weeks</td>
</tr>
<tr>
<td>400% increase of input capacity</td>
<td>1 billion records at once</td>
<td>5 billion records at once</td>
</tr>
<tr>
<td>14% improvement for data match rates</td>
<td>83%</td>
<td>95%</td>
</tr>
<tr>
<td>10% content increase</td>
<td>10%</td>
<td></td>
</tr>
<tr>
<td>• 668 million records processed</td>
<td>5 days</td>
<td>5 hours</td>
</tr>
<tr>
<td>Internal InfoBase Product Factory build</td>
<td>26 days (8 CPU SMP)</td>
<td>3 days (grid)</td>
</tr>
</tbody>
</table>
How To Deploy Your Own Information GRID

Specialized CEDI Workshops
Grid Computing Discovery, Architecture and Planning Workshop

- **Grid Readiness Assessment**
  - Customer requirements and goals
  - Existing standards and infrastructure
  - Administrative and Developer skill sets

- **Technical Overview of the IBM IIS Grid Enablement Toolkit**
  - Architecture
  - Hardware and Software Requirements
    - including Operating System and Grid Resource Manager requirements
  - Administrative and Support Considerations

- **Deliverable: Grid Deployment Plan**
  - Overview of IIS Grid offering
  - Hardware and O/S recommendations
  - Software requirements
  - Infrastructure and support recommendations
  - Estimated scope for Grid Install / Deploy Workshop

- **Duration:**
  - 2 days on-site, 1 week offsite to prepare report
Grid Computing Installation and Deployment Workshop

- **Installation of software components**
  - IBM IIS Grid Toolkit
  - (optional) hardware and Operating System configuration
  - (optional) “Build Your Own Grid” Toolkit for NIS configuration, PXE boot processing, Resource Manager

- **Configuration and Testing of components**
  - Grid Resource Manager
  - IIS Grid Enablement Toolkit
  - IIS Enterprise Edition products (DS/EE, QS/EE, PS/EE)

- **Job design review**
  - Adapting and Optimizing to run in the Grid environment

- **Grid Toolkit mentoring**
  - Management and Administration
  - Performance Considerations

- **Duration:**
  - 2 - 4 weeks on-site, depending on scope
**Information Grid**

A grid infrastructure to create an Enterprise Data Integration Hub and unified view of data enabling increased levels of collaboration and business insight.

**New Solution From IBM**

- Four component layers
  - Multi-Process Virtualization
  - Access virtualization
  - File system virtualization
  - Physical storage virtualization
- Any layer or combination of layers can be implemented to realize business value.
- An entry level offering that can expand to a complex based on customer need.

**Believe huge opportunity**

**Focus**
- Technical Architecture
- Solution Design
- Market Research

**Technologies:**
- IBM
- SUN
- HP
- INTEL
- LINUX
- Oracle
- DB2
- Teradata
- Others
Leveraging a service oriented architecture to make information more available, consistent, and trustworthy.
Architecting The Right SOA Solution

DSEE Framework
- Transformations
- Matching
- Standardization

RTI, SOA, Web Services
- XML

Under 300 Message/Sec
- MQ
- Persistent Store
- Unit of Work Operator
- 2 Phase Commit
- DSEE Framework
- Batch Processing
- Batch
- 1,000 to 10,000 Message/Sec and Up

Real Time Processing
- MQ
- Sockets
- Pipes
- RTI Service
- WS
- EJB
- JMS
- Gigabytes to Terabytes and Up

Real Time
- 1,000 to 10,000 Message/Sec and Up
- DSEE Framework
- Transformations
- Matching
- Standardization

Targets
- RDBMS
- MQ
- RDBMS

Sources
IBM - Center of Excellence For Data Integration
IBM - CEDI

- Iterations 2 Methodology
- Standard Practices
- Architecture and Design
- Education and Mentoring
- Virtual Services
- Certification
The Center of Excellence for Data Integration

- What is a CEDI?
  - A competency center within a company that is the core nucleus of knowledge and experience around data integration and in particular around our tool suite

- What does it include?
  - Software
  - Architecture
  - Methodology
  - Standards and best practices
  - Education
  - Mentoring
Why the CEDI?

Because there are many roads and many ways to fail, but only a few right ways to succeed.
A Menu of Offerings to Facilitate Enterprise Deployments

All IBM CEDI Components are delivered as services engagements usually in a workshop format.
CEDI Impact On Your Enterprise

Center of Excellence

- Technical Architecture
- Skills Assessment
- Custom Curriculum
- Best Practices
- Competency Plan
- Methodology

Impact on Productivity

Repeatability/Scalability

Project Implementation

Time

Productivity
Leveraging Enterprise Data Across an Information GRID

MDM – METADATA – GRID - SOA

Timothy G. Davis, Executive Director
Center of Excellence for Data Integration (CEDI)
IBM Software