Ontario Canada Informix User Group (OCIUG)

IBM DB2 Information Integrator - Product Overview

Ken Lee
IBM Certified Consulting IT Specialist
November 12, 2003
Agenda

- Information Integration Strategic Value
- Customer Topics
  - Scenario & Reference
  - Pain and Value
- Product Direction
- Product Content and Core Technologies
  - XML
  - Webservices
  - Federation
  - UDF's
  - Extended Search
  - Replication
  - Transformation
  - Application Development
- Getting More Information
Integration is a strategic priority -- and a major challenge

35% of CIOs list Integration as Strategic Priority
* Morgan Stanley CIO Survey 5/01/02

It is generally estimated that for each $1 spent for a packaged application, customers spend on average $5 to $9 on the labor for integration
Information Integration Approaches

- Centralizing data for consolidated access
  - Data warehouses
  - Operational data stores
  - Production applications
  - Typically managed by ETL (Extract, Transform, and Load) or replication technologies

- Integrated access to distributed sources
  - Real time data, e.g., stock quotes
  - Mixed format data, e.g., customer ODS plus related contract documents
  - Represents an emerging category of technology – EII (Enterprise Information Integration)
What is II?

Information integration is integration infrastructure middleware which lets applications access data as though it were in a single database, whether or not it is.

It enables the integration of data and content sources without having to move the data or change the platform:
- to provide real-time read and write access,
- to transform data for business analysis and data interchange, and
- to manage data placement for performance, currency, and availability.
EII Problem Domain

- Wide heterogeneity in data to be accessed
  - Data is distributed
  - Data in non-relational format
- Data changes rapidly
- Need real time data
- Need access to data source functions
- Copy costs cannot be business justified
  - Seldom-used information
  - Large volumes of data
- Data security or data ownership restricts data movement options
- Ad hoc analysis makes access patterns unpredictable
  - Impact to source systems can be contained
    - Queries are scheduled such as with Query Patroller
Customer Scenario:
Extending the Data Warehouse

- **Customer Challenge:**
  - Integrating information from multiple data and content stores for more comprehensive reporting
  - Performing analysis that combines historical and real-time or rapidly changing information
  - Rapidly deploying new applications that require data outside of existing data warehouses
Customer Scenario:
Extending the Data Warehouse

- **Customer value:**
  - Better decisions from more comprehensive and more up to date information
  - More responsive to business needs that require access to new data sources
  - Increased ROI from existing data warehouses

- **First Step:**
  - Use Information Integration to provide access to data sources
Customer Scenario:
Speeding Application Development

- **Customer Challenge:**
  - Integrating multiple data sources in a single application is complex and costly
  - Accessing nontraditional sources is too impractical to leverage their benefit
  - Time pressure to deploy new applications
  - People and skill shortages to develop new applications

![Diagram showing development effort to handle unique interfaces, joining data from varied sources, transformations, and correlating data](image)
Customer Scenario:
Speeding Application Development

- **Customer value:**
  - Reduce amount of coding by 2 times or more while using existing SQL development tools
  - Give applications access to all the relevant data sources
  - Reduce application maintenance costs
  - Deploy existing skills over wider ranging integration projects

- **First Step:**
  - Use Information Integration as data access service for application infrastructure

Application Developer handles:
- Interfaces for each data type
- Joining data from varied sources
- Transformation
- Correlating data
Information Integration Portfolio History

- Unchanged Application
- SQL
- Transparent
- Heterogeneous
- Optimized

DB2 Relational Connect

IBM DiscoveryLink
- Federated ORDBMS
- Bio-Data Genomes
- Text

DB2 DataJoiner

IBM Enterprise Information Portal
- Unified Search

Production Data
- Diverse Subsets

Library Server
- Object Server
- Object Server

DB2 Data Management Software
Vision for Information Integration

- **Data federation**
  - Read/write access across diverse data and content sources
  - Database programming model (SQL)
  - Content programming model (CM API)
  - XML programming model (XML API)

- **Data placement**
  - Caching and Replication over heterogeneous information

- **Data transformation**
  - SQL, XML, Web services
  - Advanced search and mining
  - Metadata management

- **Part of a complete business integration solution**
  - Complete XML support
  - WebSphere business integration
  - Open platform based on industry standards
The DB2 Information Integrator Products

- **IBM DB2 Information Integrator 8.1**
  - A new product offering based on DB2 technology
    - Both a federated data server and a replication server
  - Accessing primarily relational sources, but with the ability to also access XML, Web or content sources
  - Use by developers, IDEs, or analytical tools that leverage SQL

- **IBM DB2 Information Integrator for Content 8.2**
  - Repositioning of IBM Enterprise Information Portal
  - Accessing primarily IBM content repositories, but with the ability to also access relational, XML, Web or other content stores
  - Use by developers familiar with IBM Content Manager API
IBM DB2 Information Integration Products

**DB2 Information Integrator**
- SQL programming model
- Leverage SQL skills and tools
- Federated data server and replication server

**DB2 Information Integrator for Content**
- Content programming model
- Leverage CM skills and tools
- Federated data server, text mining, and workflow engine

The strategic information integration framework for access, manipulation, and integration of diverse, distributed and real-time data.
Core Technologies

Information Integration

Development Environment

Industry Standard APIs: SQL and XML

Federate
Search
Cache
Replicate
Transform

Data Access
Flexible Access with Standard API's

- **SQL**
  - Familiar language with widely deployed skills
  - Rich analytical capabilities
  - Traditional database clients
  - Extensions for XML data (SQL/XML)

- **XML**
  - Emerging standard for interchange
  - **XQuery - XML Query Language**
    - Based on a formal algebra
    - IBM is co-submitter of XML Query specification
      (http://www.w3.org/TR/xquery)
  - Exploit unique features within XML data model - hierarchy, sequence
  - Web services

---

Financial Services Sector to spend $8.3B (US) on XML and Web Services by 2005

– Zap Think, March 2002
XML Technology

- Object-relational implementation
  - Store, retrieve, compose, decompose, validate, extract, transform
  - Storage options
    - Store intact
    - Store as a collection of columns
Web Services

- **Provider support**
  - Access resources from Web clients

- **Consumer support**
  - Extend reach of database to nontraditional, real-time data sources

- **UDDI***
  - Catalog Web services for public of private use

- **XML Registry***
  - Manage XML artifacts such as XML schemas, style sheets, DTDs…

*Delivered by WebSphere*
Federation

- **Transparency:** Provides a single "virtual database" to applications
  - Appears to be one data source
  - Hides the differences in location, invocation, dialect, network, file/DBMS topology
  - Supports a high level query language with a robust API
  - SQL, stored procedures, User Defined Functions
  - Functional compensation and passthru

- **Heterogeneity:** Integrates data from different data sources
  - Diverse types of data
  - Diverse sources
  - Relational, flat files, XML, spreadsheets, proprietary

- **High function:** Capabilities of existing sources and of SQL
  - To search for and to manipulate data
  - Lose no functionality of source or of SQL language
  - One query can combine data from multiple sources
Federation (cont'd)

- Optimization: Getting the answer faster
  - Query re-write
  - Cost-based access path selection
  - Considers database, network, and processor capabilities and statistics

- Extensibility
  - *ANSI SQL/MED standard*
  - Retains optimization knowledge
  - Surfaces specialized source functions

- Autonomy: No perturbation of existing data, sources or applications
Federated Access to Diverse Data

Wide variety of Clients

Single virtual database view

“Single point-of-connect” for the end-users and applications
Wrappers

Four important tasks:

- Data modeling
  - Map data model to relational data model (tables with rows and columns)
  - Map functions into SQL operations
- Query Planning
  - Represent data source capabilities
  - Push down as much work to data source as sensible
  - Detect missing function at source (so engine can compensate)
  - Supply cost and cardinality information
- Connection and transaction management
- Query Execution and data retrieval
  - Execute parts of a user’s query for a specific data source
Wrapping it up

- Encapsulate data source information inside a wrapper
  - Models data as tables
  - Expose unique function
- Non-relational wrappers participate in query planning and execution
- Capability to add wrappers for unusual data sources

Oracle, MS SQL Server, Sybase, Informix, ODBC, Teradata, Flat file, XML, Excel, Documentum, Blast, Extended Search, and more coming
User Defined Functions Extend Access

- Provide additional logic which is invoked as part of SQL processing
- Can return either scalar, row, or table results
- Can be used to compose standard views
- Simple to develop and configure
- *Can exploit parallelism*
- Built-in UDFs for MQSeries, OLE DB

```
SELECT MQSend(followup.service,a.custid || ' ' || a.ordid)
FROM account a WHERE a.status = 'overdue'
```
IBM Lotus Extended Search: Unstructured Access

- IBM Lotus Extended Search
  - Brokered search architecture for searching thousands of existing data sources
  - Results are aggregated, ranked, and returned in a single hit list
  - Easily embeddable into any application
  - Lotus databases, document systems, full text indexes, e-mail, directories, WWW, syndicated content, relational, file systems

- Text mining
  - Summarization, categorization, feature extraction,…
  - Packaged with EIP
Heterogeneous Caching

- Improve query performance and availability
- Administrator defines Materialized Query Table
  - Precomputed or frequently used values
  - Any relational data from the federated system
  - Application indicates ability to use cache
  - Implicit or explicit use
- Developer enables application for cache use
  - If enabled, optimizer may use the cache, writes passed through to the source
  - If not, reads and writes passed through to source
- Cache refresh managed:
  - Manually
  - DPROP
- Flexible caching topologies supported
Replication

- Applications
  - Warehouse and ODS applications
  - Consolidation and distribution
  - Application integration
  - Availability management

- Heterogeneous replication
  - DB2, Oracle, Sybase, Informix, Microsoft, Teradata (apply only)

- Table-based or transaction-consistent
- Point-in-time or continuous operation
- Embedded transformation
Replication Architecture

Combined with the federation engine -> a powerful integration tool!
Transformation

- SQL features on rows
  - Statistics
  - OLAP
  - Mining
  - User-defined functions

- Views
  - Transforms physical DBMS schema to user logical schema
  - Can include SQL row transforms

- XML features
  - Compose, decompose, parse, and validate XML documents
  - XSL transformation
Application Development

- Improve developer productivity and application efficiency for integrating data
- Reduce ongoing administration and maintenance costs
- Leverage existing development and analytical tools

- Tool focus: Make IBM Information Integrator transparent to development tooling
  - DB2 Centers (Control Center, Development Center, etc)
    - Support for DBAs and core DB2 server-side developers
  - WebSphere Studio (Eclipse Platform)
    - Web developers, J2EE developers, Data developers, XML developers and more
  - DB2 Microsoft Visual Studio
    - Support for DB2 application developers
  - Business Intelligence Tools
    - Testing with partner tools e.g. Crystal Decisions, Business Objects,…
Administration

Discover
Wrappers
Data servers
Data objects
Transformations

Design
Views
Transformations
Topology

Deploy
Configure
Rollout
Administer/Tune
Monitor
Administration Tools: Dynamic Discovery

- Dynamic Discovery mechanism enables new sources to be discovered and automatically configured
- Wrapper writer optionally provides GUI and discovery routine for custom wrappers
Replication Administration

- **Definitions**
  - Manage control definitions for replication
  - Customize names and sizes of objects

- **Operations**
  - Start Capture, Apply, Monitor, Analyzer, and Trace
  - Issue commands such as STOP or STATUS

- **Monitoring**
  - Perform static and dynamic monitoring
DB2 Information Integrator for Content

- Define integrated views across diverse and distributed data
  - IBM Content Manager portfolio and other content repositories e.g. FileNET, Lotus databases, ODBC and JDBC compliant relational databases, and IBM Lotus Extended Search sources (LDAP directories, WWW, email databases,…)

- Search federated data
  - Search application uses the IBM Content Manager API

- Mine additional metadata from text documents
  - Identify document language
  - Extract entities like names or technical terms
  - Categorize documents based on a taxonomy
  - Group documents based on related content
  - Create a document synopsis

- Define workflows
IBM Leadership

- Leads the industry in integrating and exploiting Web services
  - First to deliver Web services provider & consumer support
  - More complete XML implementation than Oracle or Microsoft - Bloor report
- Only vendor to make data federation a real world option
- Only vendor to extend transformative and accelerating power of the relational database engine to the federated data
  - Statistical analysis, XML transforms, spatial transforms, UDFs, etc
  - Query optimization, query rewrite
  - Caching, Automatic summary tables
- IBM continues leadership role in industry standards organizations responsible for XML and Web services standards
Data Management Technology Evolution

- Twenty-five years of research and development
- Information On Demand

- **Warehousing & analysis**
  - heterogeneous objects

- **Information integration**
  - rich semantic model

- **Extensibility**
  - distributed objects
   - high performance, complex queries

- **Distribution**
  - System R

- **Optimization**
  - R Star

- **Federation**
  - Garlic

- **Integration**
  - Xperanto

- **DB2 Technology Base**
IBM Competitive Advantage

- More open
  - Supports many different platforms and open standards
  - Integrate and federate data rather than centralize

- More scalable
  - 25 years of data management experience
  - Cost based optimization, query rewrite

- More functions
  - Federated access to both structured and unstructured data
  - Both read and write

- More deployable
  - SQL-based solution now, XQuery in the future
  - Leverages widespread skills and tools infrastructure

- More to come
  - DB2 II a key investment area for IBM
  - Improve parallelism, automated caching, advance search, etc.
Customer II Site

http://www.ibm.com/software/data/integration/solution
For more information

http://xperanto.dfw.ibm.com/demo

Where can I find information about …..?

- **Information Integration**

- **II Technology Demo**

- **IBM's Data Replication Solution**

- **DB2 DProp**

- **IMS DProp**

- **DataRefresher**

- **DataJoiner**

- **DiscoveryLink Home Page**

- **Life Sciences Solutions: Data Management  SSR materials**

- **Customized Demos: Global e-business Solution Center**
  - [http://w3-1.ibm.com/support/stss/gesc.html](http://w3-1.ibm.com/support/stss/gesc.html)
Where can I find information about ..... ?

- **Technical Emphasis**
  - Technical article on federated
  - Paper in IBM Systems Journal
  - Garlic:
  - IBM Rebooks
    - [http://www.redbooks.ibm.com](http://www.redbooks.ibm.com)
      - Replication
      - SG24-6140 Migrating to the IBM Replication Solution
      - SG24-5463 My Mother Thinks I'm a DBA!
      - SG24-6217 IBM Replication for Pervasive Computing
      - SG24-6513 Building the ODS on DB2 Using Replication, MQ, WHMgr
      - SG24-5415 Getting Started with Data Warehouse and Bus Intelligence
      - Avail: 2002 Data Replication Using DPROP V8
      - Federated
      - SG24-6290 IBM Life Sciences: Turning Data into Discovery
      - XML
      - SG24-6285 Using XML on z/OS and OS/390 for App Integration
      - SG24-6130 Integrating XML with DB2 XML Extender & Text Extender
      - REDP0135 DB2 XML Extender Hints and Tips for iSeries
      - EIP
      - SG24-6101 IBM EIP: A Practical Approach
      - SG24-5749 IBM EIP: A Primer
      - SG24-6125 IBM EIP: A Cookbook
      - Info Integration
      - Avail: 2002 Data Integration Architecture