TOC

INDEX

VIEW

B10 Information Integration Technical Overview

Dan Wolfson, Senior Technical Staff Member, IBM

This talk will provide a technical overview of the information integration initiative within Data Management. Information integration is one of the five styles of integration publicized in WebSphere's Business Integration announcement this May. Information integration addresses two areas: how to integrate heterogeneous data from diverse sources in a meaningful way and how to make the resulting information available through both existing and emerging programming models and clients. II incorporates many technologies - in this overview we will review our work in XML, Federation, Replication, Web Services, Messaging, and programming models.

B10

Information Integration Overview

Dan Wolfson

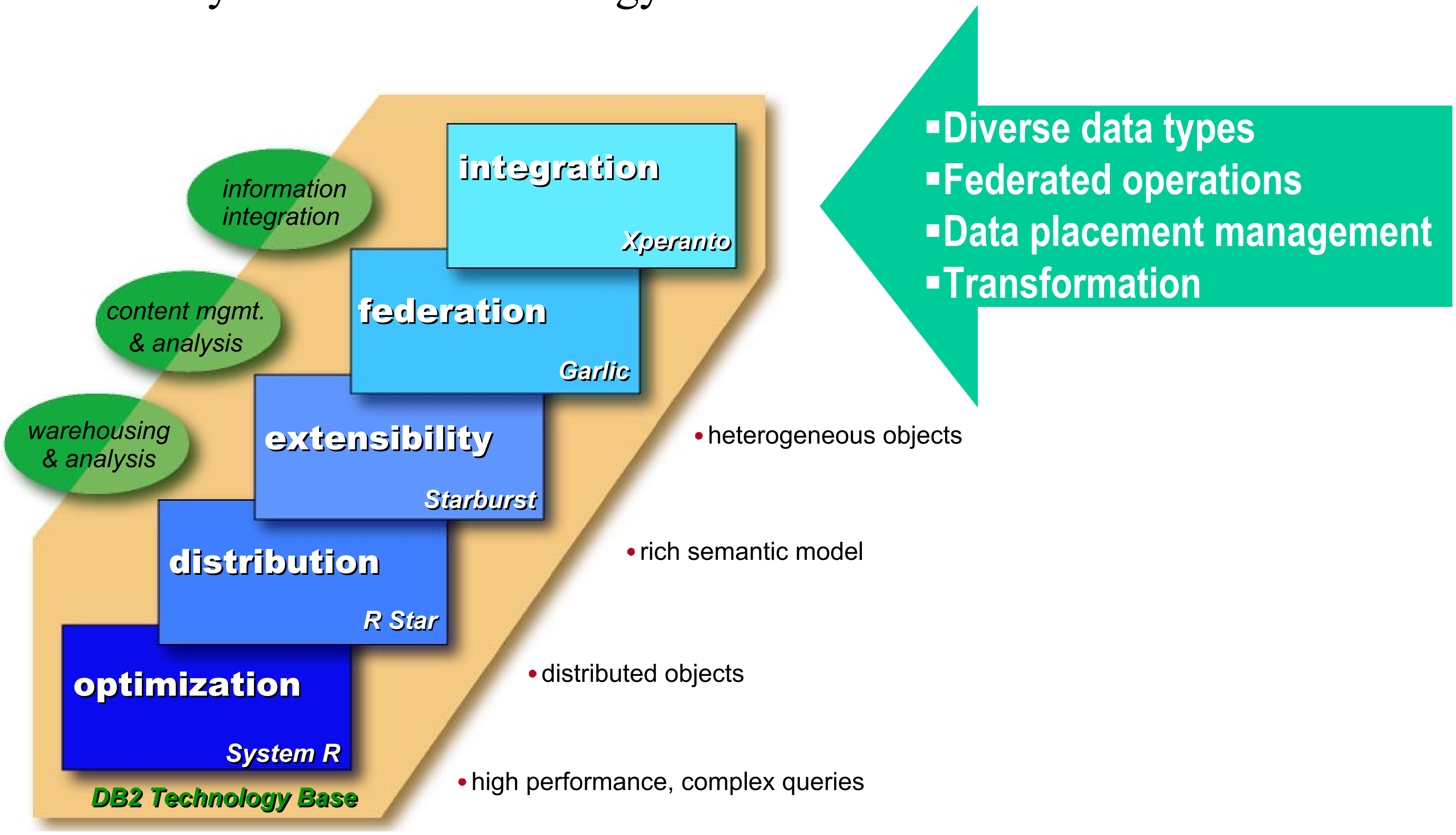
IBM Data Management Technical Conference

Anaheim, CA **Sept 9 - 13, 2002**

RETURN TO INDEX

Information Integration

- IBM’s work to address customer requirements for integrating information
- Evolutionary from DB2 technology investments



- Diverse data types
- Federated operations
- Data placement management
- Transformation

information integration

content mgmt. & analysis

warehousing & analysis

integration *Xperanto*

federation *Garlic*

extensibility *Starburst*

distribution *R Star*

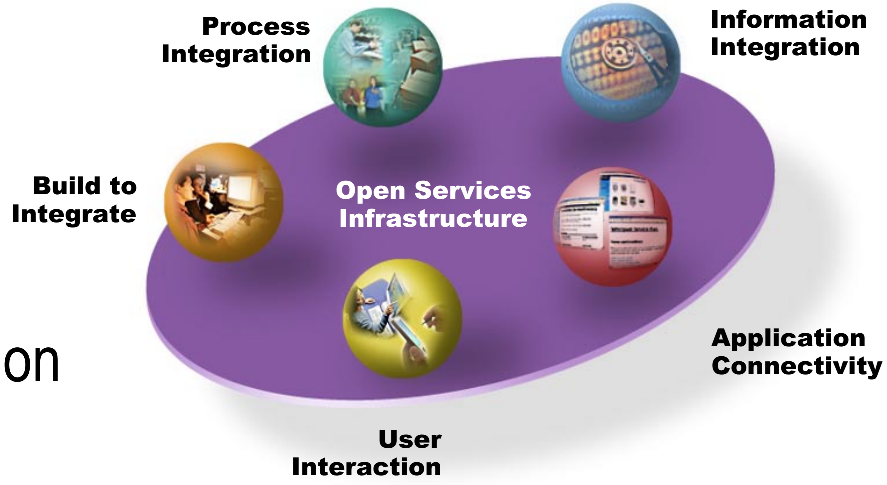
optimization *System R*

DB2 Technology Base

- heterogeneous objects
- rich semantic model
- distributed objects
- high performance, complex queries

What is Information Integration?

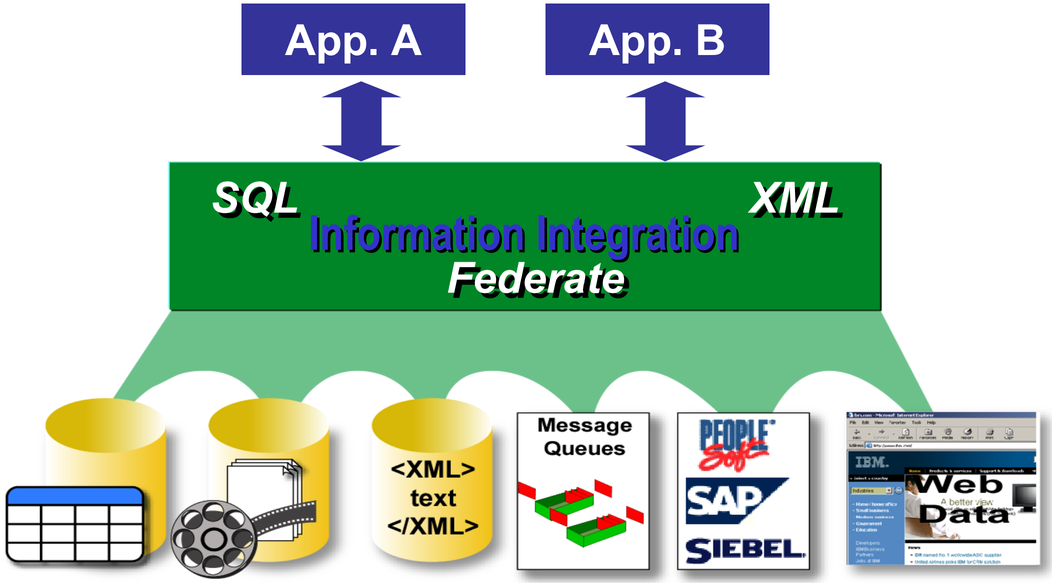
II technology enables the **integration of traditional and emerging data sources** to provide real-time read/write access, transformation to meet the needs of business analysis, and data placement management for performance, currency, and availability.



Business Integration

Information Integration Key Elements

II technology enables the **integration of traditional and emerging data sources** to provide real-time read/write access, transformation to meet the needs of business analysis, and data placement management for performance, currency, and availability.



App. A

App. B

SQL

XML

Information Integration

Federate

<XML>
text
</XML>

Message Queues

PEOPLE Soft

SAP

SIEBEL

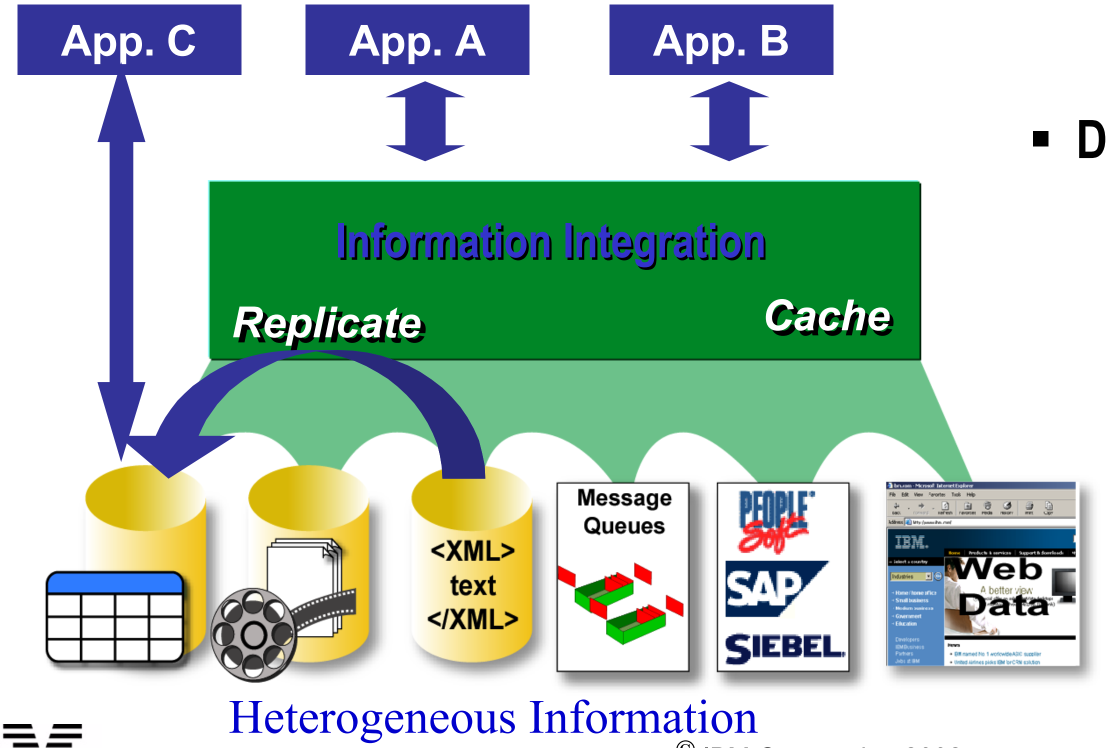
Web Data

Heterogeneous Information

- **Read/Write Access to Diverse Content Sources**
  - ➢Structured, semi-structured, and unstructured
  - ➢Legacy, packaged,and Web
  - ➢Real-time and point-in-time

Information Integration Key Elements

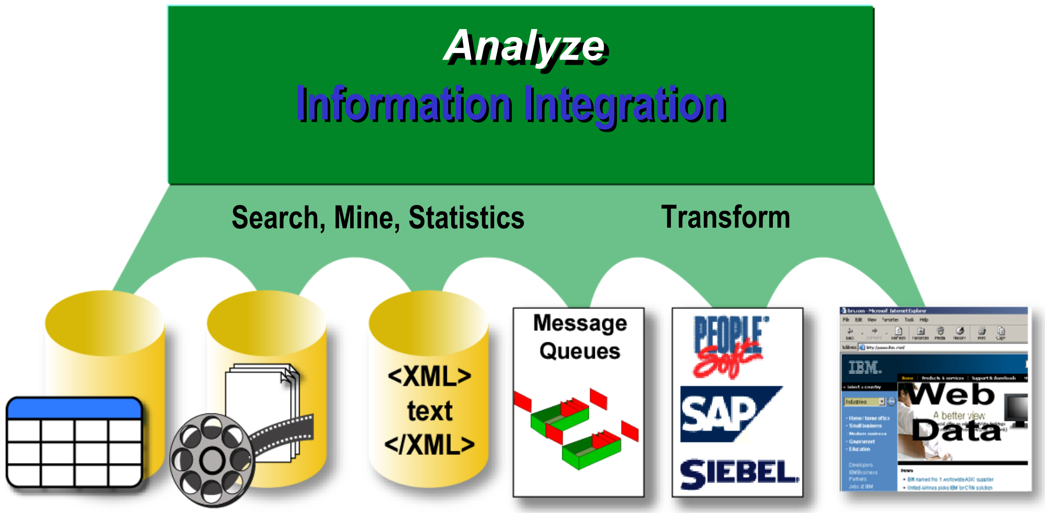
II technology enables the **integration of traditional and emerging data sources** to provide real-time read/write access, transformation to meet the needs of business analysis, and data placement management for performance, currency, and availability.



- Read/Write Access to diverse content sources
  - ➢Structured, semi-structured, and unstructured
  - ➢Legacy, packaged,and Web
  - ➢Real-time and point-in-time
- **Data Placement Management**
  - ➢Replication, ETML, Caching, Over Heterogeneous Information

Information Integration Key Elements

II technology enables the **integration of traditional and emerging data sources** to provide real-time read/write access, transformation to meet the needs of business analysis, and data placement management for performance, currency, and availability.

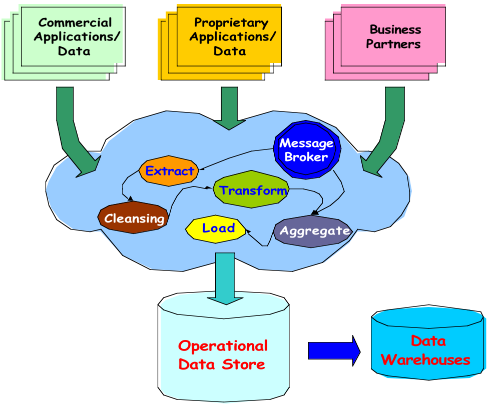


Heterogeneous Information

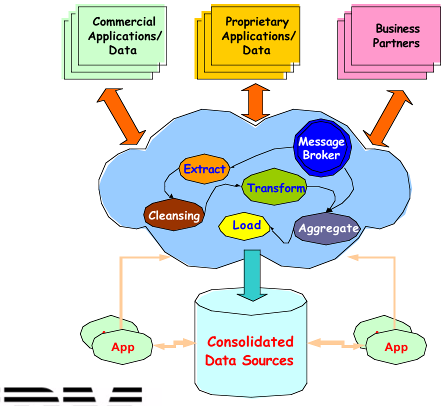
- Read/Write Access to diverse content sources
  - Structured, semi-structured, and unstructured
  - Legacy, packaged,and Web
  - Real-time and point-in-time
- Data Placement Management
  - Replication, ETML, Caching, Over Heterogeneous Information
- **Transformation for analysis**
  - Search, Mining, Statistics, …
  - SQL, XML
  - Metadata management

Today’s Challenges

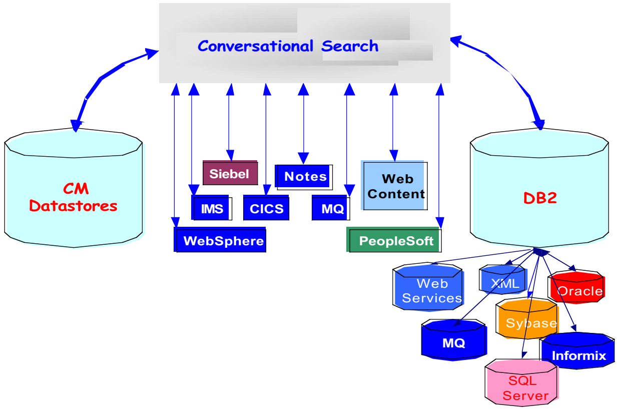
Information Consolidation



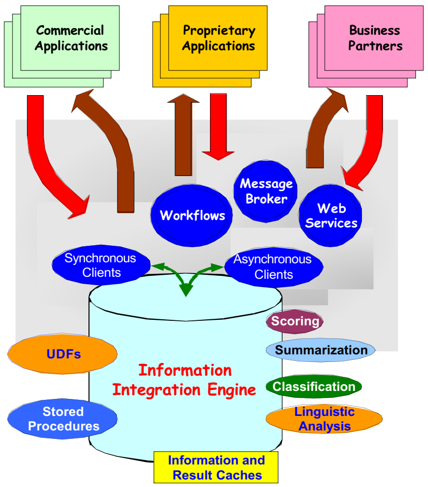
Information Convergence


Commercial Applications/ Data
Proprietary Applications/ Data
Business Partners
Message Broker
Extract
Transform
Cleansing
Load
Aggregate
App
Consolidated Data Sources
App

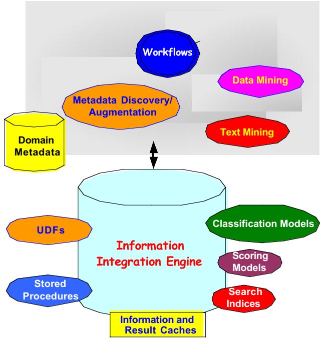
Information Federation



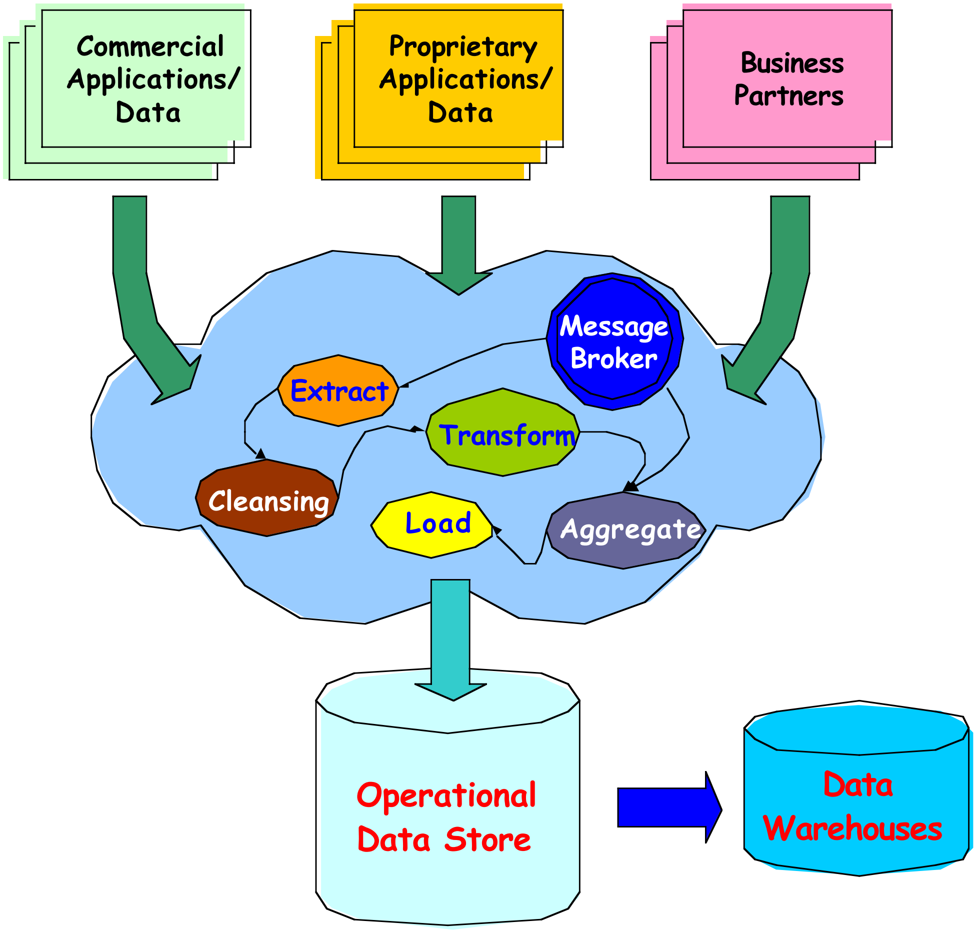
Active Analysis



Deep Analysis

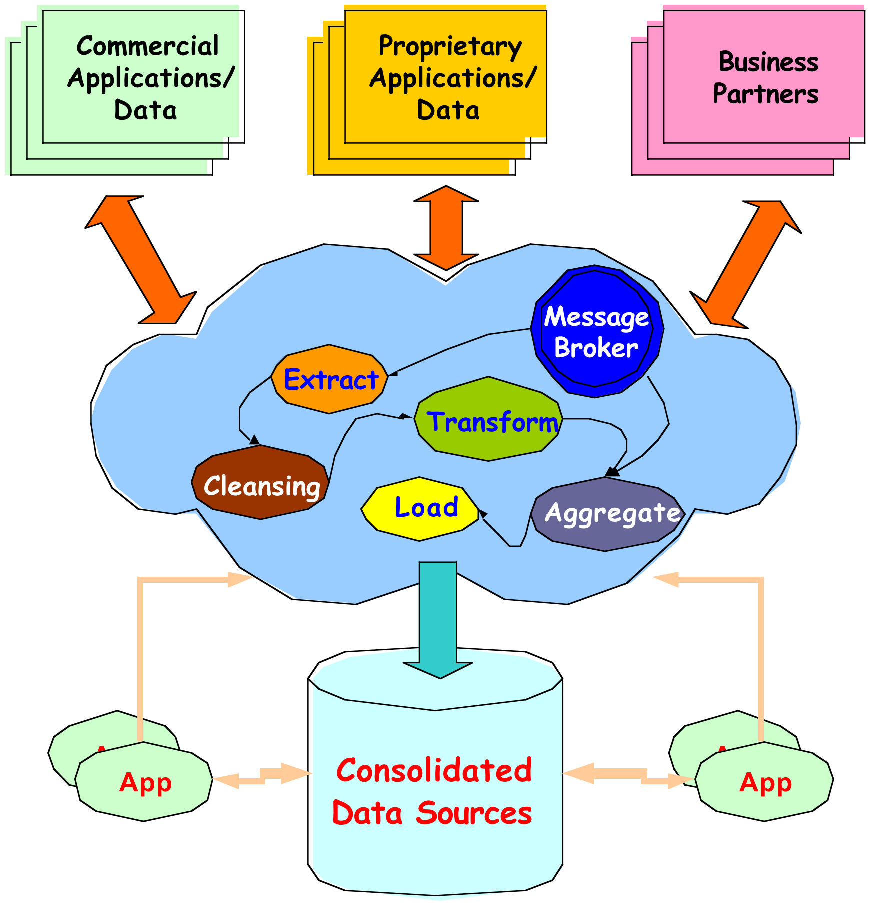


Information Consolidation



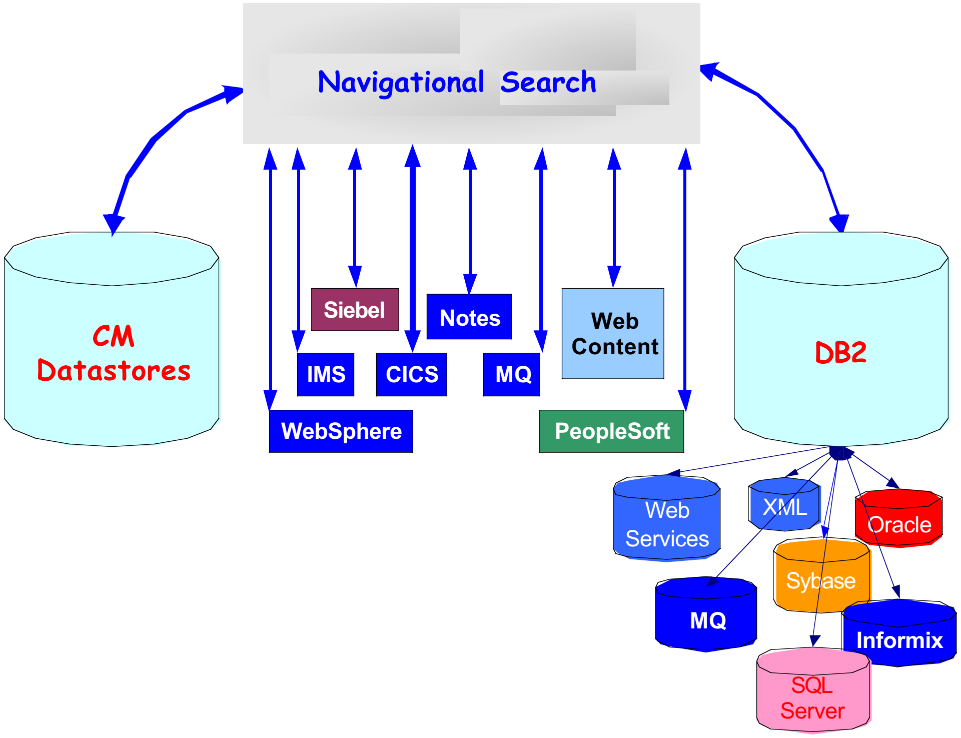
- Collect information from multiple heterogeneous sources and load into a central repository
- Consolidated information is Read-Only
- Central repository based on a (new) common schema
- Information must be significantly transformed and cleansed before being introduced into the central repository
- Information may be received in batches or near-realtime, synchronously or not
- New applications, warehouses, and marts are constructed on top of this central repository

Information Convergence


Commercial Applications/ Data
Proprietary Applications/ Data
Business Partners
Message Broker
Extract
Transform
Cleansing
Load
Aggregate
App
Consolidated Data Sources
App

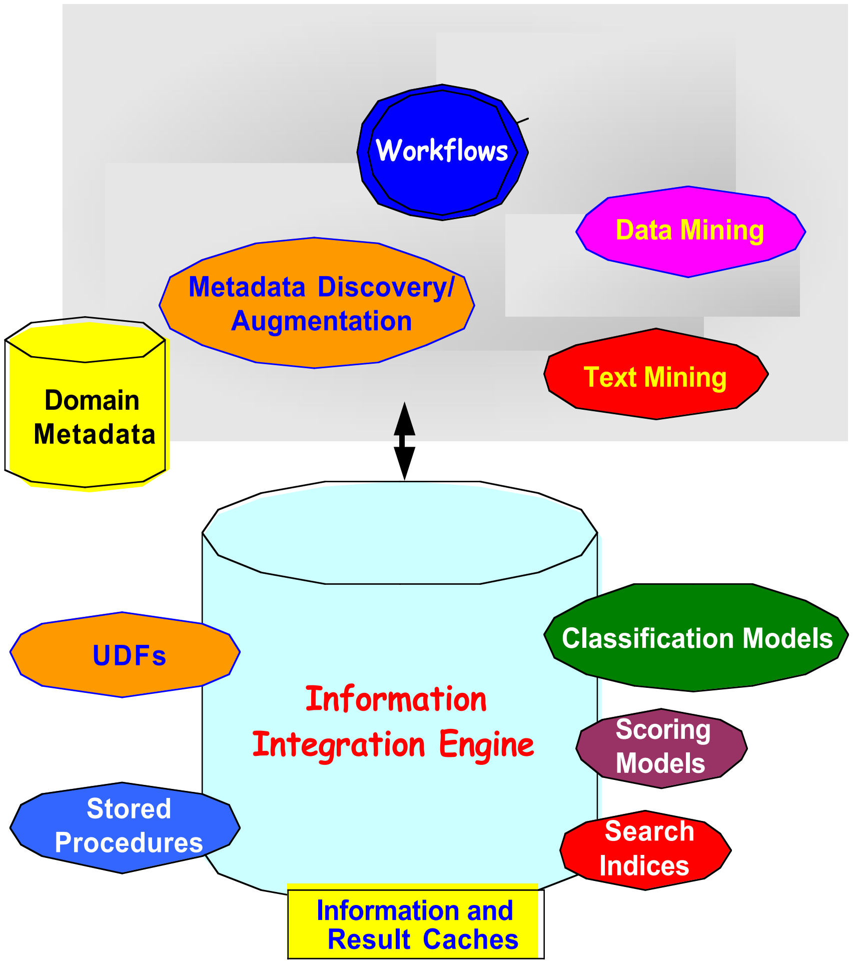
- Collect information from multiple heterogeneous sources, cleanse and load into consolidated repository.
- Consolidated repository based on a (new) common schema
- New applications constructed using consolidated repository for both read and write operations.
- Convergent Consistency Model
  - Update operations on repository are pushed back to information sources.
  - Information may be received in batches or near-realtime, synchronously or not
  - If all updates stop, system will converge on a consistent state.
  - The system of record is always with respect to a particular application.
- New applications, warehouses, and marts are constructed on top of this central repository

Information Federation



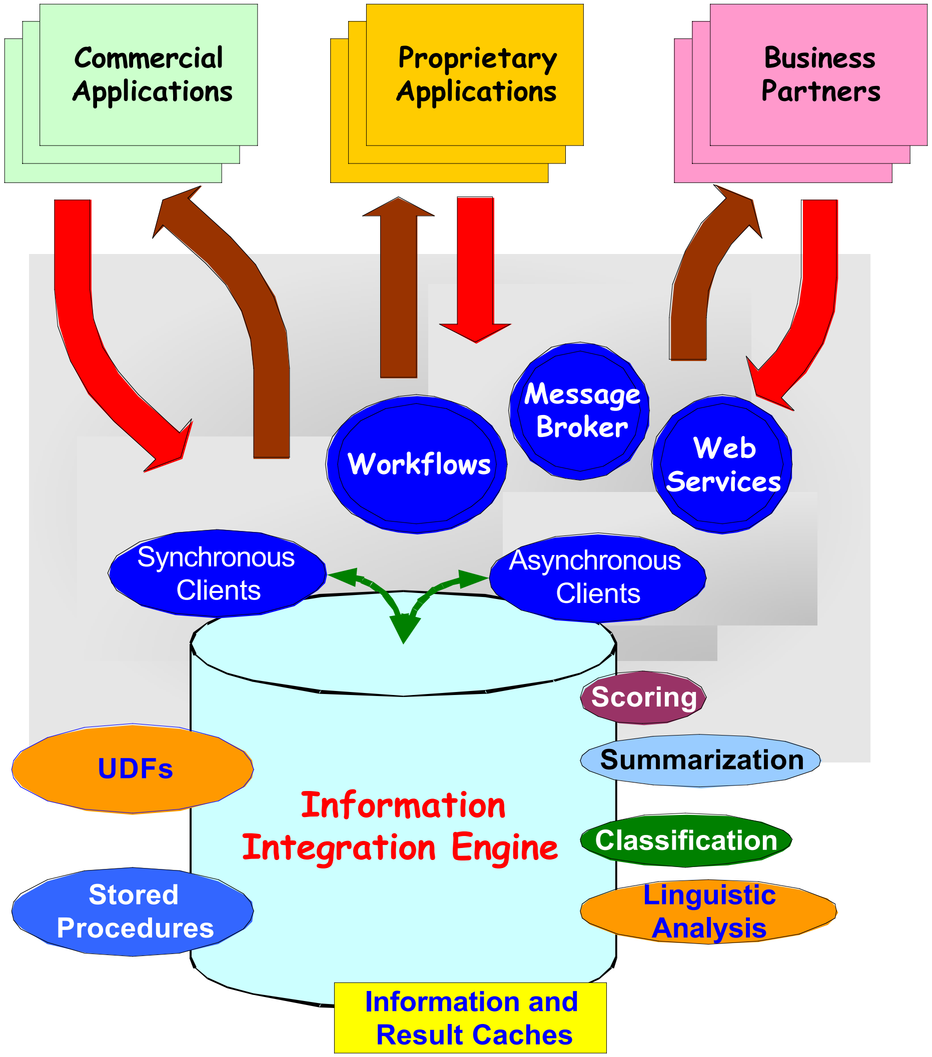
- Access to data without moving it.
  - Access structured and unstructured data
  - Always access current data
  - Ownership and privileges maintained
  - Little transformation and cleansing
  - Query over potentially sparse or imprecise content
- Two techniques
  - Aggregation by navigational search using EIP provides access to a wide variety of information
  - Federation with DB2 Relational Connect provides optimized query execution over heterogeneous sources

Deep Analysis



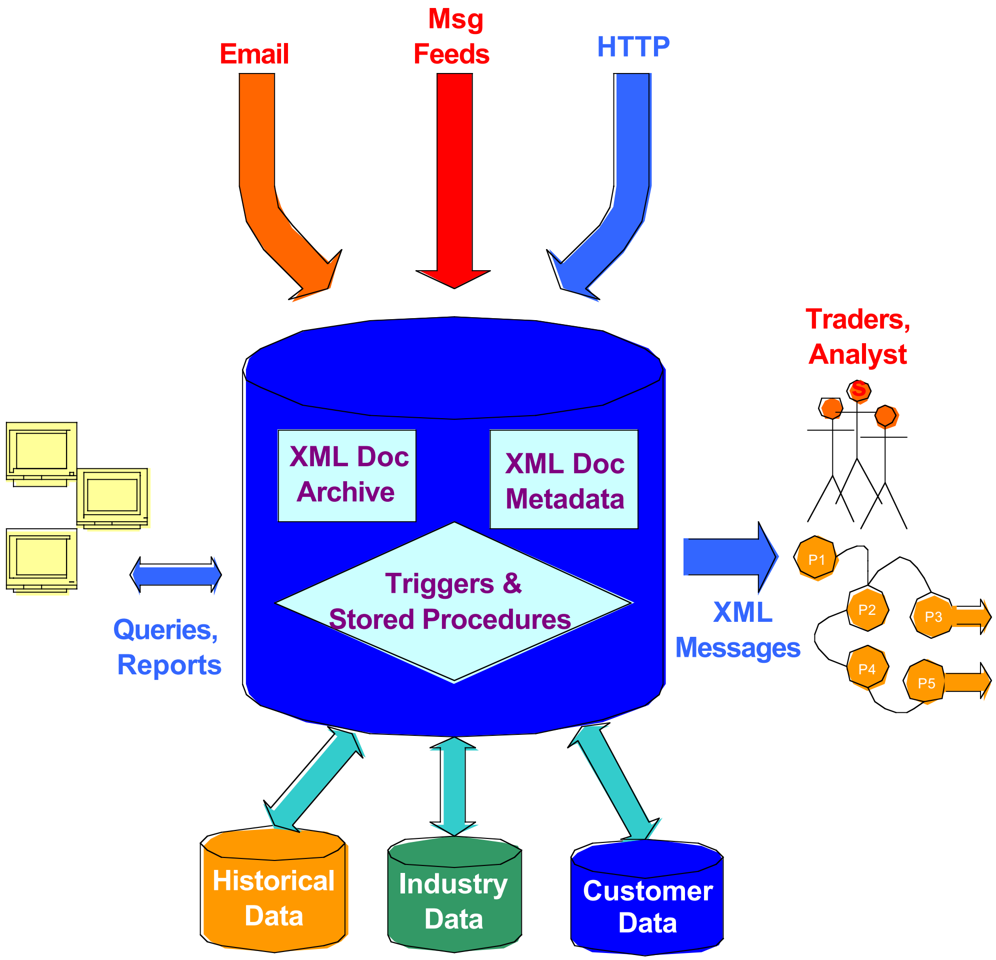
- This is about making performing complex analysis, typically over very large quantities of data
  - Mining, classification, metadata discovery, scoring models..
- Analysis may be performed over structured and/or unstructured data
- Analysis does not happen at transactional speeds

Active Analysis



- This is about making near real-time / online decisions triggered by changing data and events
- In active analysis the timeliness of the answer is a key component of the quality
- Set oriented processing optimized by database; sequential flows executed by flow engines
  - Combine synchronous and asynchronous processing
  - Incorporate current and historical information

Financial Research Management and Processing



- Information Consolidation
  - Gather raw research, transforming to RIXML if needed.
- Information Federation
  - Federate historical, current, customer data for Portals, Active Analysis
  - Federate information from newsfeeds, market and research sites
- Deep Analysis
  - Construct classification models, historical trends, research accuracies
  - Text mining and analysis
- Active Analysis, Active Data
  - If important market information (scoring), Email notifications to key customers
  - Notify traders, correlate to other core information and report
- Portal
  - Provide internal and external customers with a query interface to provide alerts and allow for search over federated metadata and structured/unstructured sources.

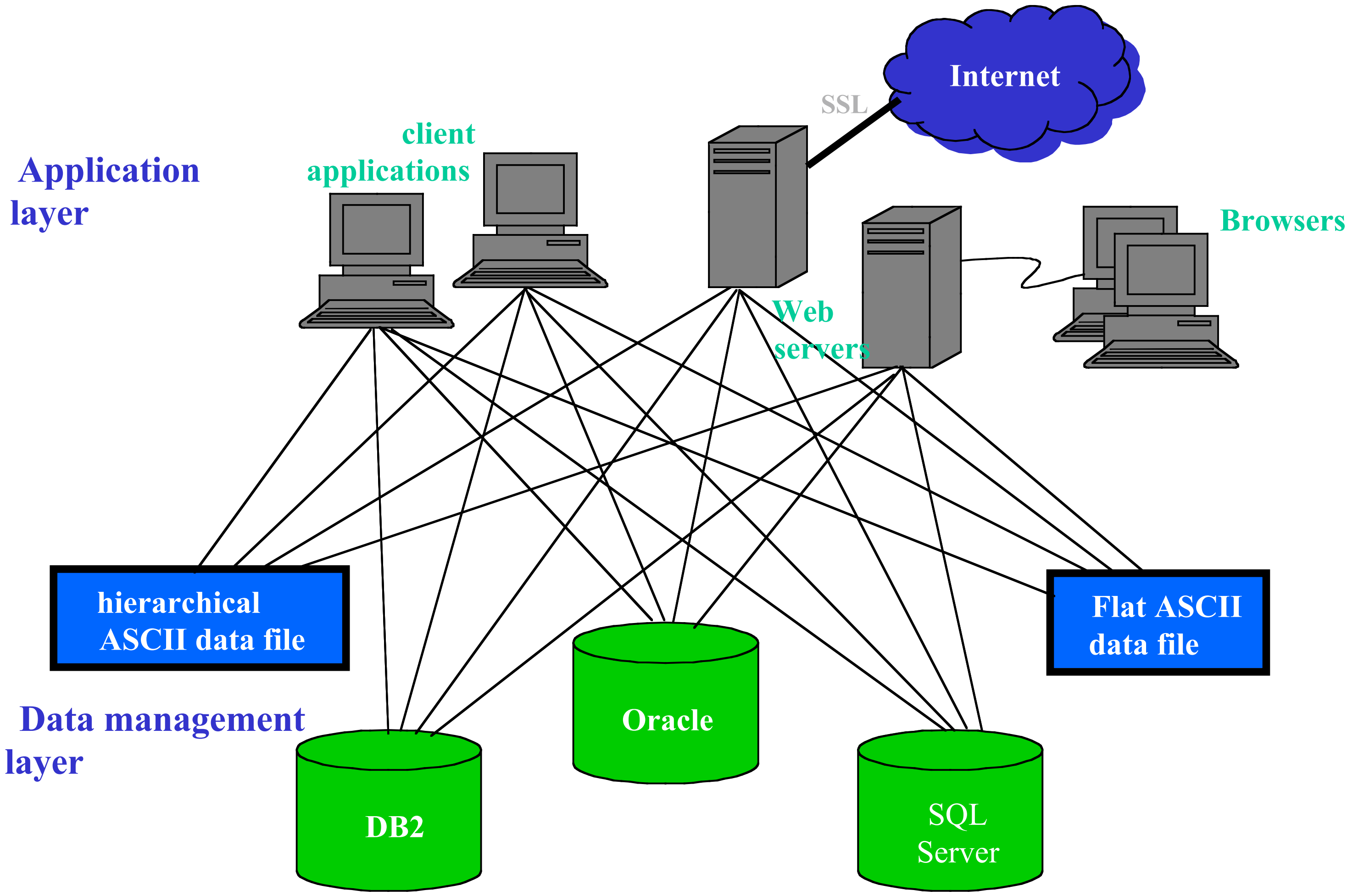
Information Integration Requirements

- Requirements
  - Consistent, coherent access and management across all information assets
  - Streamlined flow of information between people, processes, applications
  - Leverage existing application and infrastructure investment
  - Speed of deployment
- Challenges
  - Managing and leveraging the information explosion
    - Online data volumes are huge, rate of change is increasing
  - Degree of diversity of information
    - Not just heterogeneous relational -- flat files, documents, Web content, specialized applications
  - Taxonomy and metadata management
    - Cultural and organizational aspects are the larger problem
  - Changing environment

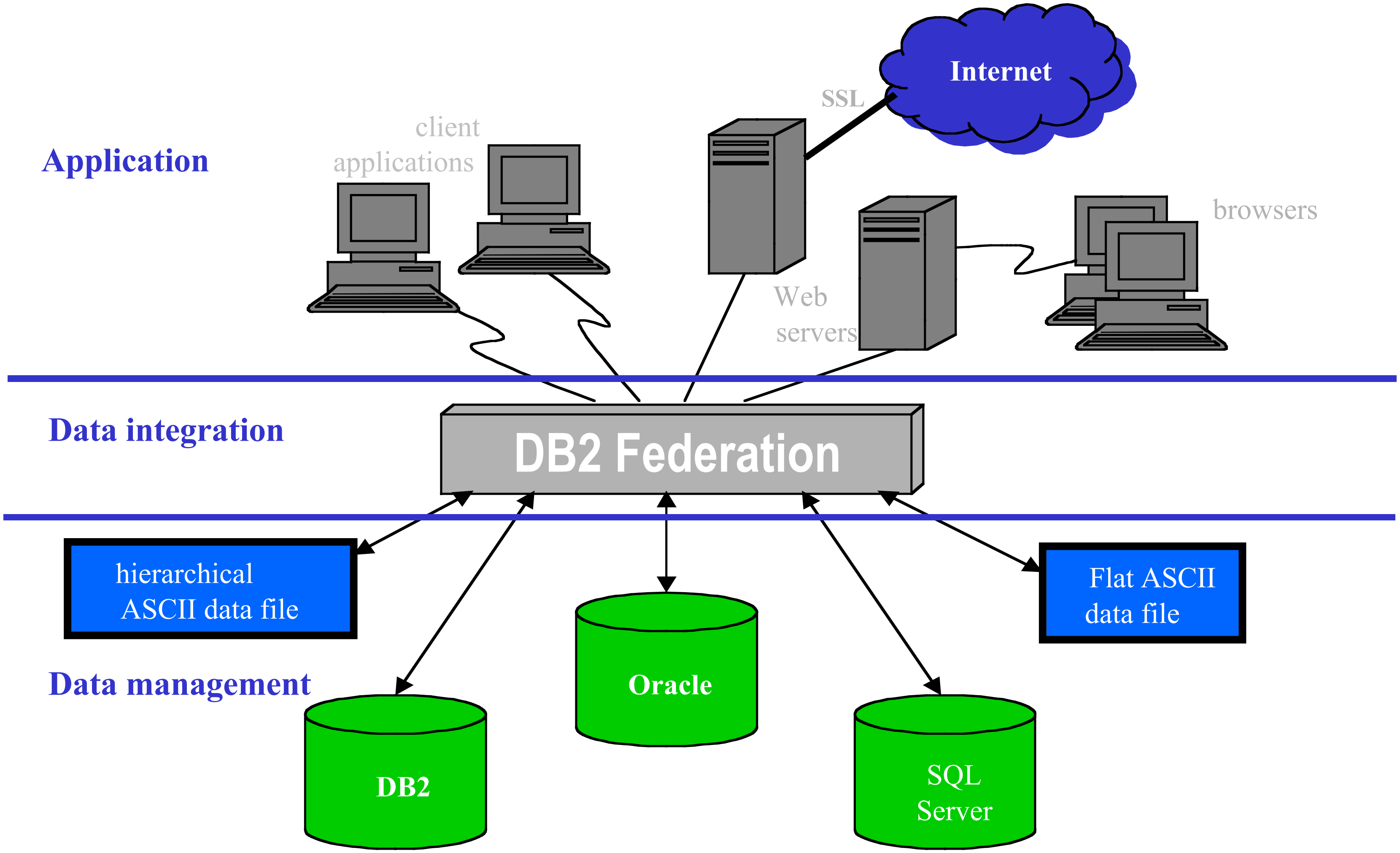
The State of the Business World

Computer
Generated
Output
Images
Audio & Video
XML
Hierarchical,
Files
Relational
Text
Scanned
Documents
Dances with databases
SPATIAL
DB2
Spatial
Temporal
Web Content
BAXTER BAY BANK

Without data integration

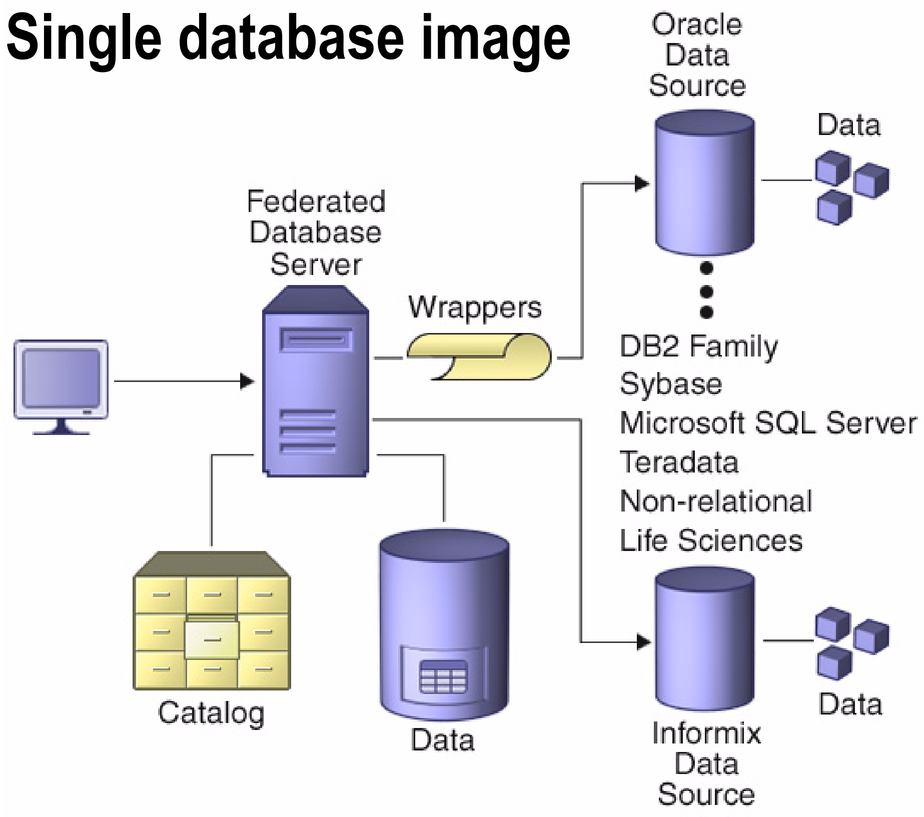


Federation Approach



Federation Technology

Single database image



- Transparency
  - Appears to be one source
- Heterogeneity
  - Integrates data from diverse sources
  - Structured, XML, unstructured, messages, Web, …
- Function and Extensibility
  - SQL/XQuery plus backend specific functions
- Autonomy
  - Non-disruptive to data sources
- Performance
  - Distributed optimization

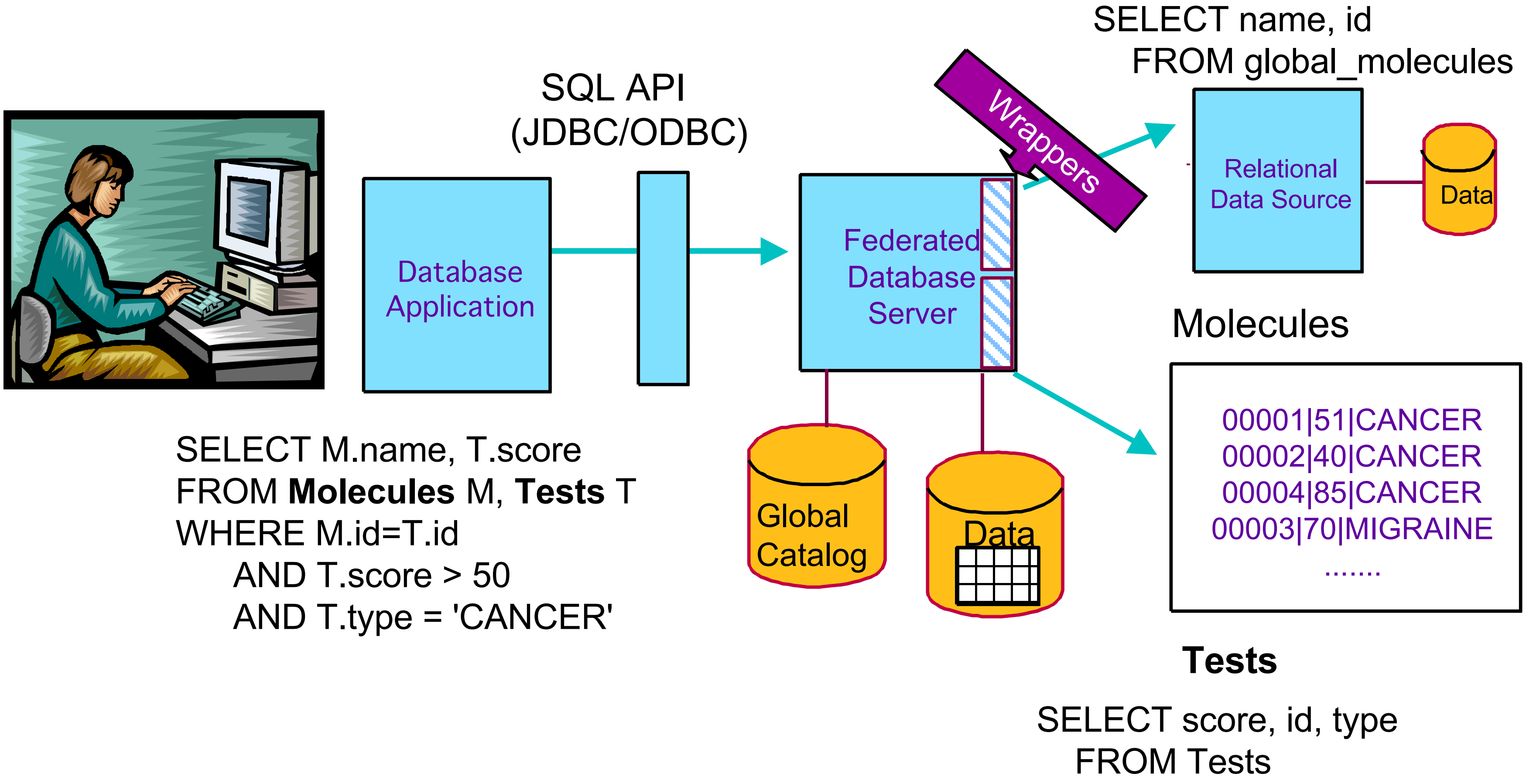
Why federate at the database level?

- Single query interface to access heterogeneous data
  - SQL, SQLJ, JDBC, ODBC
- Exploit the features of the DB2 Engine!
  - Heterogeneous join
  - Cost-based optimization can be used to determine best query execution plan
  - Exploit latest database features over federated data
    - Summary tables
    - Caching
    - XML features
    - OLAP functions
    - Analytical functions
    - and many more…

The DB2 Federated Approach

- Incorporate data sources using wrappers
  - Clean, simple interface for wrapper provider
    - Subclass C++ templates
    - Java wrappers support in plan
    - Library of support functions
  - "Thin" wrapper philosophy
- Powerful query processing engine in federated server
  - Decomposes and distributes queries
  - Cost-based optimizer chooses query plan
    - Primitive costing in Version I; will improve subsequently
  - Query execution engine drives wrappers, combines results
  - Compensates for missing function in data source
- Object-relational modeling capabilities

DB2 Federated Database Architecture



```
SELECT M.name, T.score
FROM Molecules M, Tests T
WHERE M.id=T.id
      AND T.score > 50
      AND T.type = 'CANCER'
```

```
SELECT name, id
   FROM global_molecules
```

```
00001|51|CANCER
00002|40|CANCER
00004|85|CANCER
00003|70|MIGRAINE
.......
```

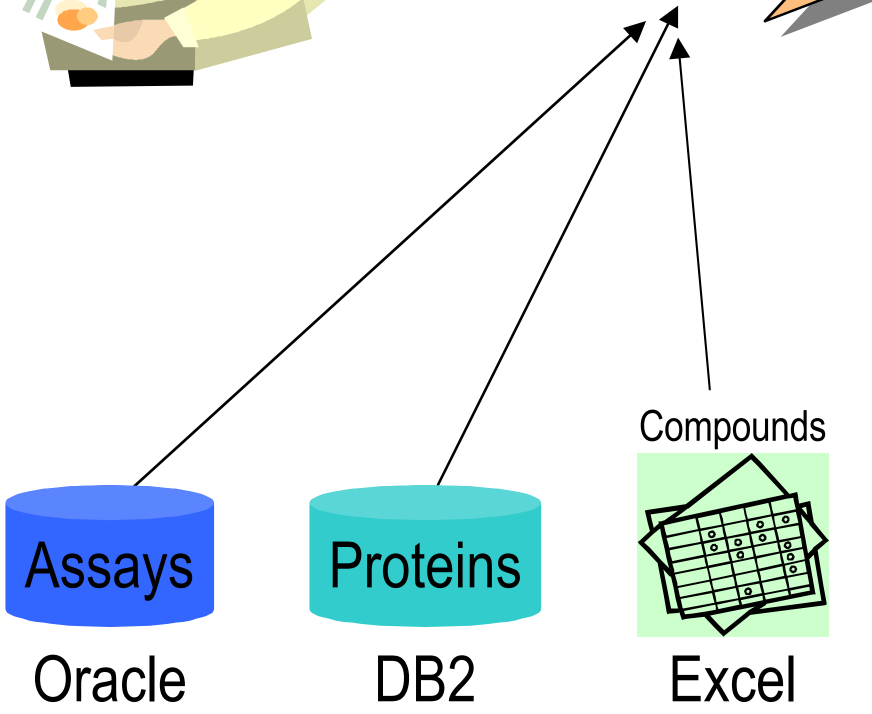
```
SELECT score, id, type
   FROM Tests
```

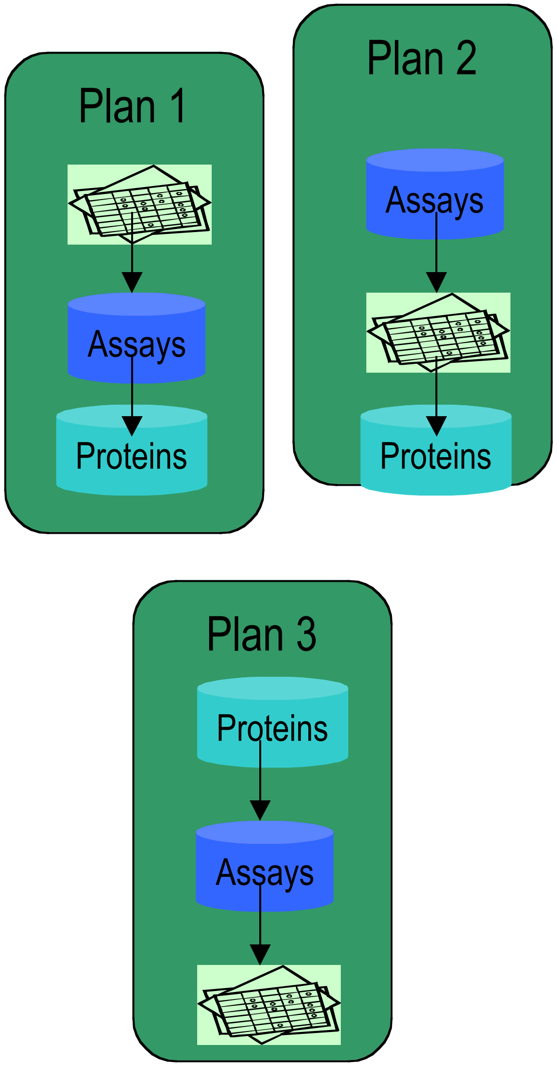
Wrapper Architecture

- Wrapper provides key services
  - Modeling data as tables
  - Query planning
  - Query execution
  - Default code provided for each service
  - Implement only what's needed for each source
- Benefits
  - Startup cost to write a wrapper is small
  - Wrappers can evolve over time
  - Diverse data sources can be wrapped
  - New wrappers can be added at any time
  - Query optimization is supported for performance

DB2 Federation: Cost-based query optimization

“Find compounds that have been tested against serotonin receptors and have IC50 values between 4-9.”





DB2 Federation: Which plan is best?

- Depends on many factors
  - The number of compounds
  - Number of compounds with low IC50 values
  - The number of proteins in serotonin family
  - Number of assays for each compound
- DB2 query optimizer is essential to executing cross-source queries with good performance

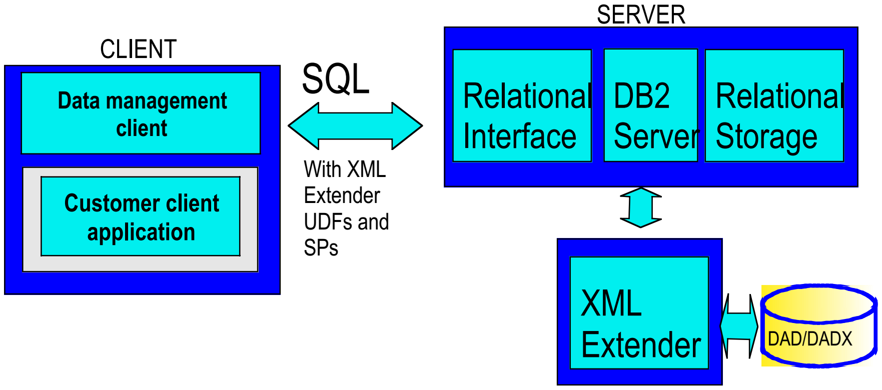
Summary: DB2 Federated Technology

- Infrastructure for extracting information from data
  - Wrappers allow access to key data sources
  - Query engine provides cross-source queries
- An extension of industrial-strength DB2 technology
  - Robust, high function, high performance
  - Benefit from R & D in relational DBMS
  - Benefit from experience in content management
- Eases application development
  - Transparency, heterogeneity, high function
  - APIs for modern environments (XML, J2EE, Web Services)
- Application Autonomy
  - Does not disrupt existing applications

See Session B13,B14!

XML Technology

- Object–relational implementation
  - Store, retrieve, compose, decompose, validate, extract, transform
  - Storage options
    - Store intact
    - Store as a collection of columns
- Hybrid XML-relational store
  - SQL or XQuery
  - XML specific storage, query, indexing, privileges, transformation, schema, interfaces, search
  - DB2 engine core attributes: scalability, availability, reliability, manageability



CLIENT

Data management client

Customer client application

SQL

With XML Extender UDFs and SPs

SERVER

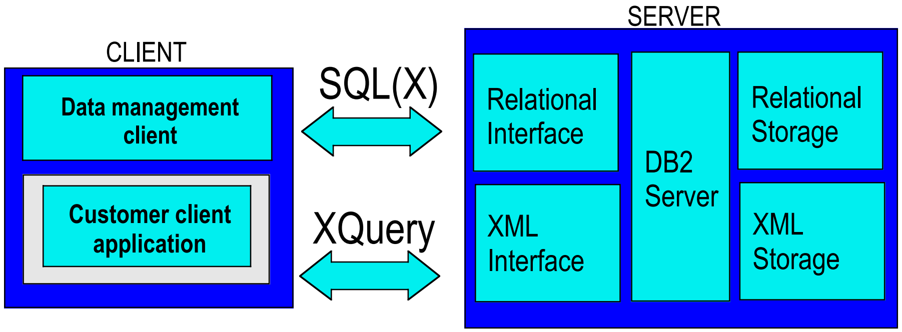
Relational Interface

DB2 Server

Relational Storage

XML Extender

DAD/DADX



CLIENT

Data management client

Customer client application

SQL(X)

XQuery

SERVER

Relational Interface

DB2 Server

Relational Storage

XML Interface

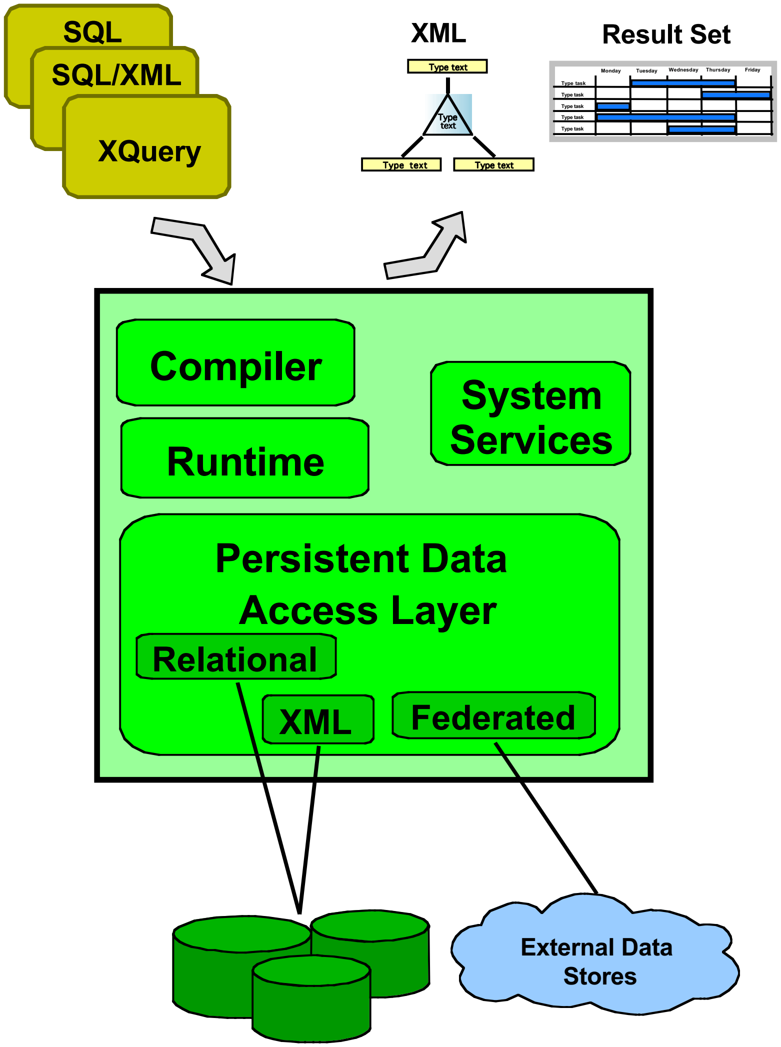
XML Storage

XML Storage Today

- Store/retrieve whole XML documents
  - As XML column (entire document) or collection of fields
- Compose or decompose and store/retrieve portions
  - Mapping over relational tables
- DB2 support:
  - Document Access Definition (DAD)
    - Shape
    - Scope
  - XML data types
    - XMLVARCHAR, XMLCLOB, XMLFILE (file name)
  - Search: fast and powerful search/indexing on XML

XML in DB2 Future Support

- Deeper engine exploitation
- "Feels" relational and/or XML
- SQL, SQl/XML/XQuery
- XML *is* DB2 internals
- Performance, performance, performance



XML Query

- Potential to be the language interface of choice
  – querying and manipulating XML
- Why not SQL?
  – Extensions to SQL are being developed in SQLX Consortium
  – Developers who are well-versed in XML may not want to learn SQL
  – Traversing XML tree documents with SQL may be awkward
- XQuery contains XML-natural constructs
  – X-path
  – powerful query mechanisms borrowed from SQL
- Like SQL, XQuery is based on a formal algebra
- Goal:
  – human-readable and
  – machine-readable forms of the syntax
- IBM is co-submitter of XML Query specification:
  – http://www.w3.org/TR/xquery/

XQuery: FLWR based - For Let Where Return

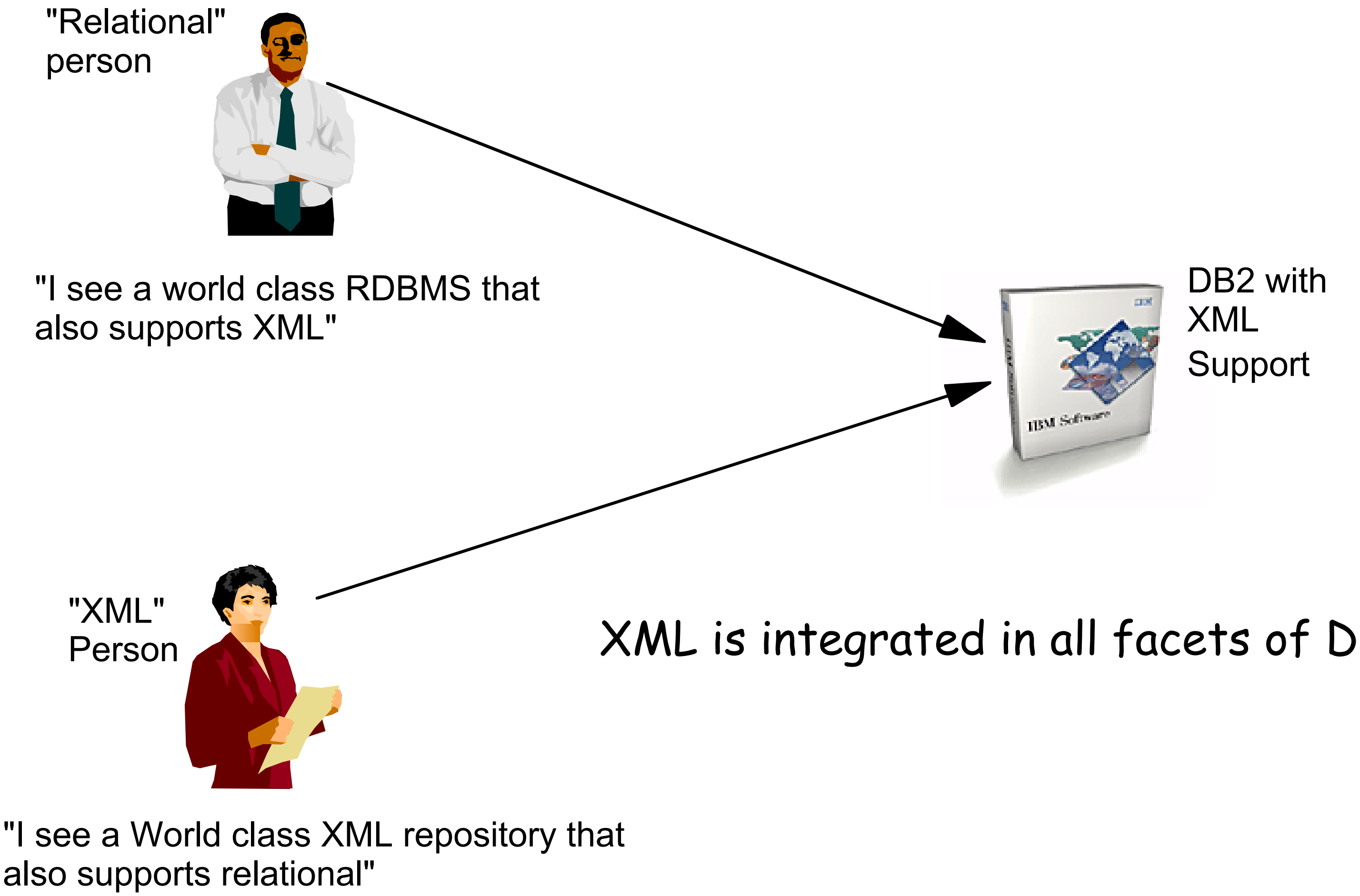
Query:

```
     FOR $book IN document("bib.xml")//book
     WHERE $book/publisher = "Addison-Wesley"
     RETURN
   <book>
         {
         $book/title
          }
   </book>
```

Results:

```
      <book>
            <title>TCP/IP Illustrated</title>
      </book>
      <book>
           <title>Inside XML</title>
      </book>
```

XML “Vision”



"Relational" person

"I see a world class RDBMS that also supports XML"

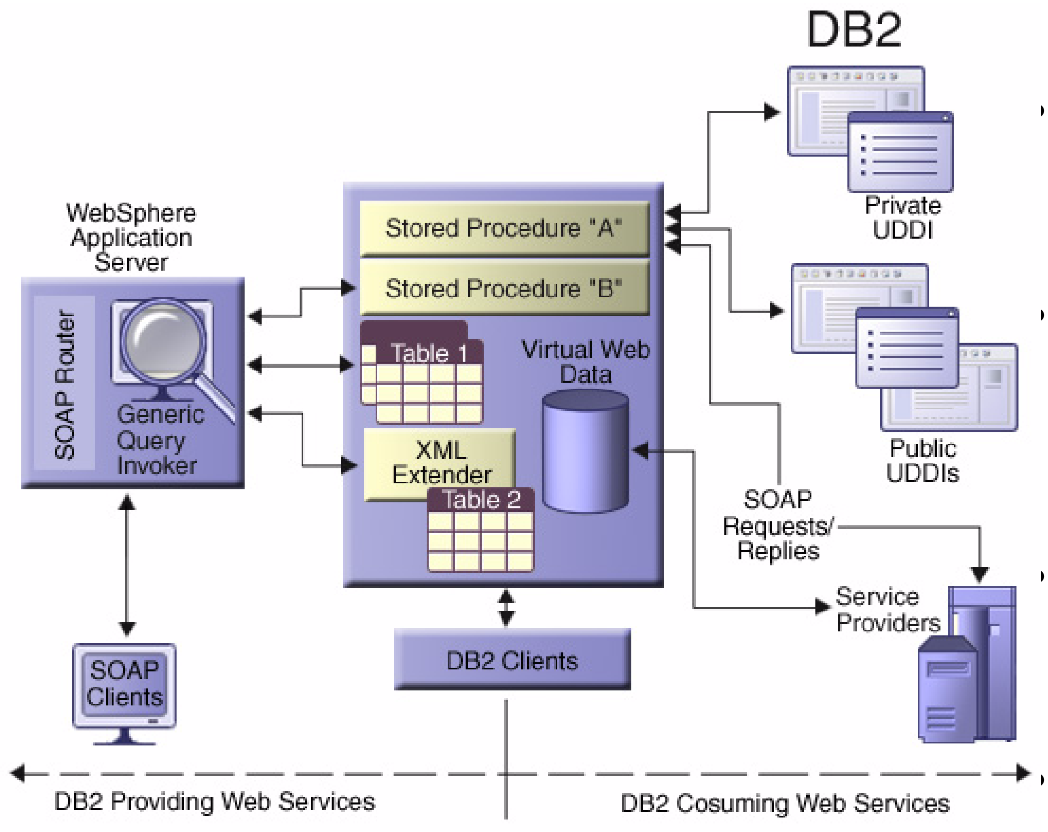
"XML" Person

"I see a World class XML repository that also supports relational"

DB2 with XML Support

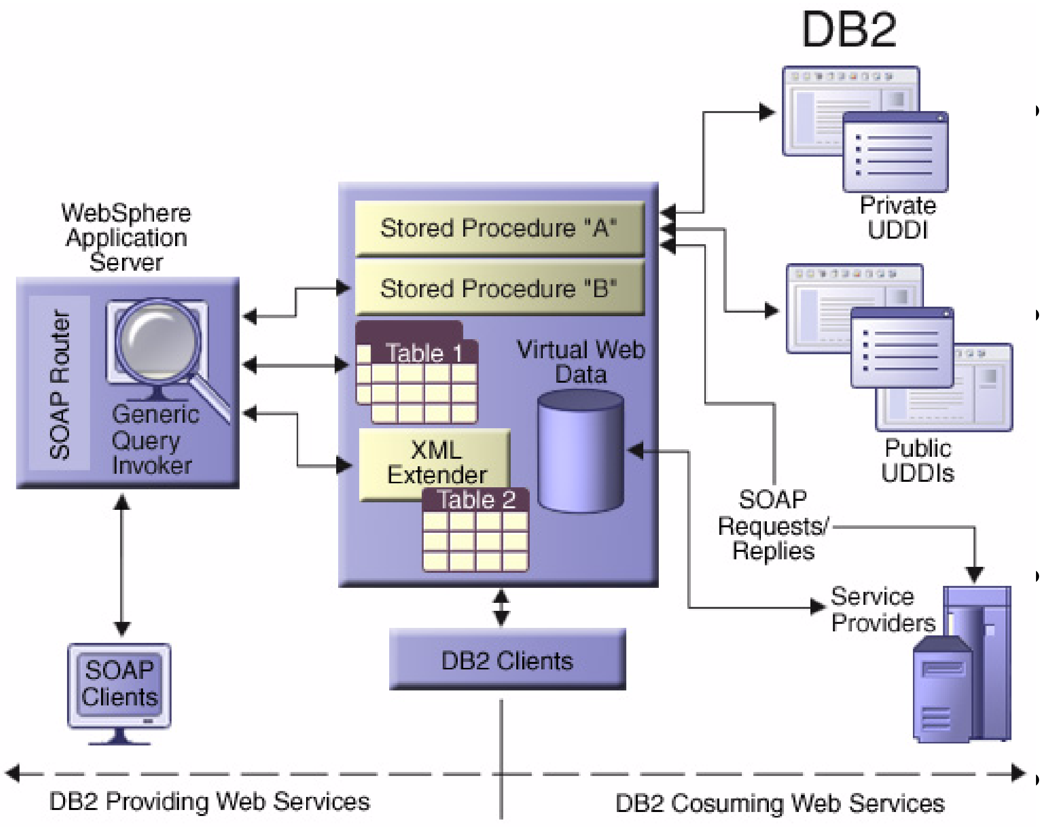
XML is integrated in all facets of DB2!

Web Services Technology



- Provider support
  - Access resources from Web clients
- Consumer support
  - Extend reach of database to non-traditional, 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…

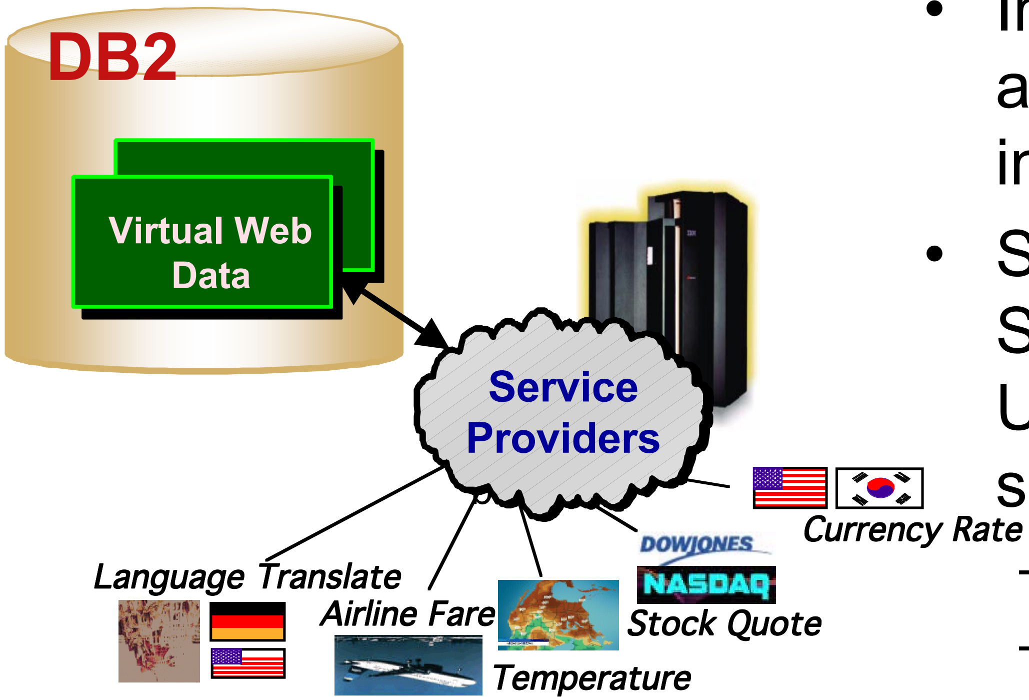
Web Services Technology



- Provider support
  - Access resources from Web clients
- Consumer support
  - Extend reach of database to non-traditional, 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…

– See Session B06, B07, F07!

Accessing Web Services from DB2

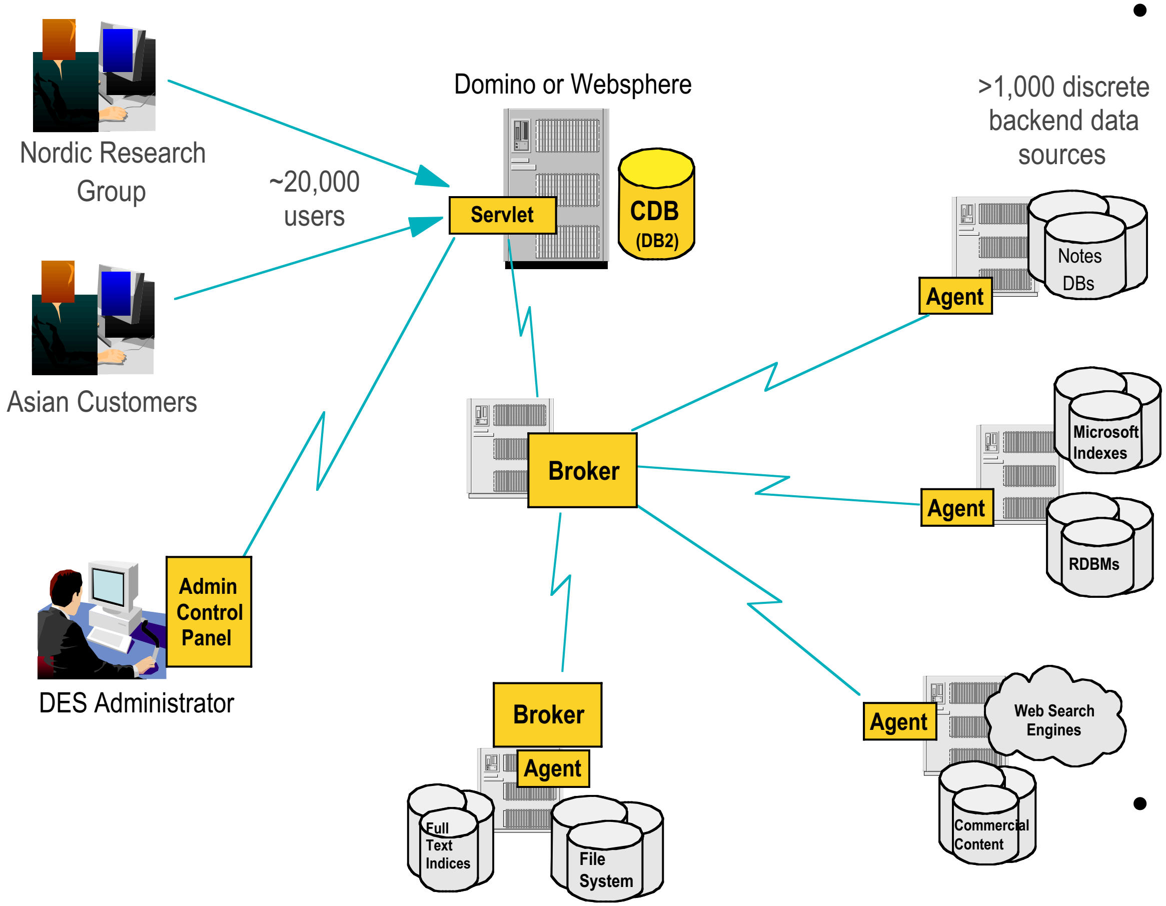


- Integrate SQL statements and Web Service invocations
- Support for generating SQL scalar and table UDFs based on wsdl web service description
  - Command line version
  - Tooling integrated into Web Sphere Studio

SELECT city, GetTemperature(city)

FROM location

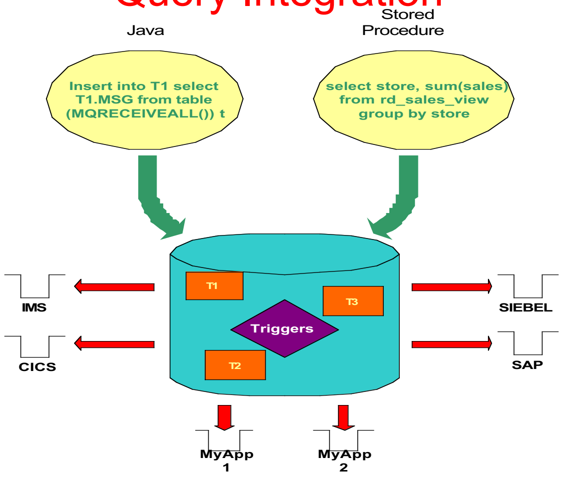
Search Technology



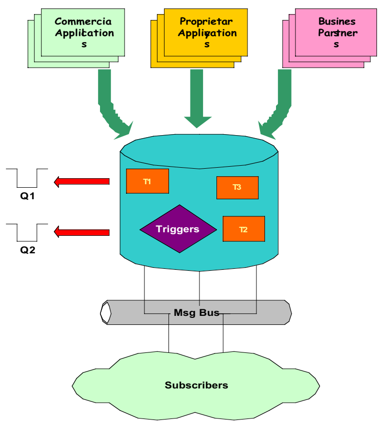
- IBM Lotus Extended Search
  - Brokered search architecture for searching thousands of existing data sources
  - Results are aggregated, ranked, and returned in a single hit list
  - Easily embeddable into any application
  - Lotus databases, document systems, full text indexes, email, directories, WWW, syndicated content, relational, file systems
- Combined with Federation
  - Generated search arguments
  - Sophisticated ordering
  - XML document generation

A Sampling of Message Integration Scenarios

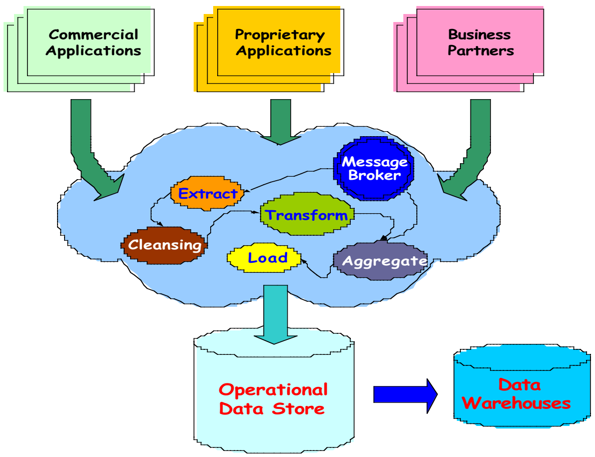
Query Integration


Java
Stored Procedure
Insert into T1 select T1.MSG from table (MQRECEIVEALL()) t
select store, sum(sales) from rd_sales_view group by store
T1
T3
Triggers
T2
IMS
CICS
SIEBEL
SAP
MyApp 1
MyApp 2

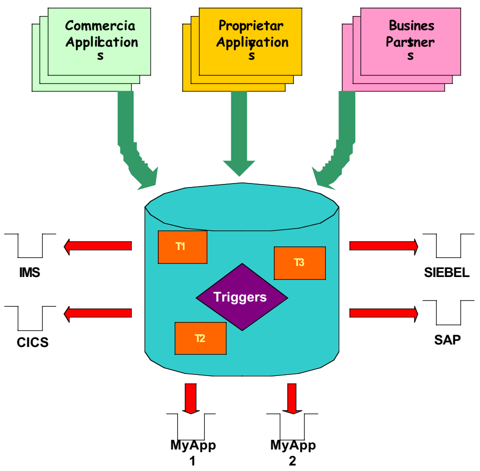
Event Notification



Information Aggregation

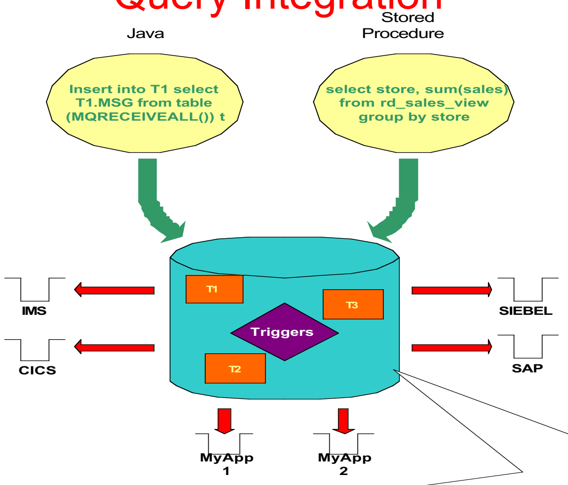


Information Forwarding



A Sampling of Scenarios

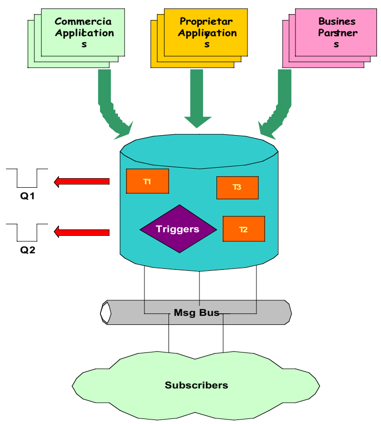
Query Integration



```
Insert into T1 select T1.MSG from table (MQRECEIVEALL()) t
```

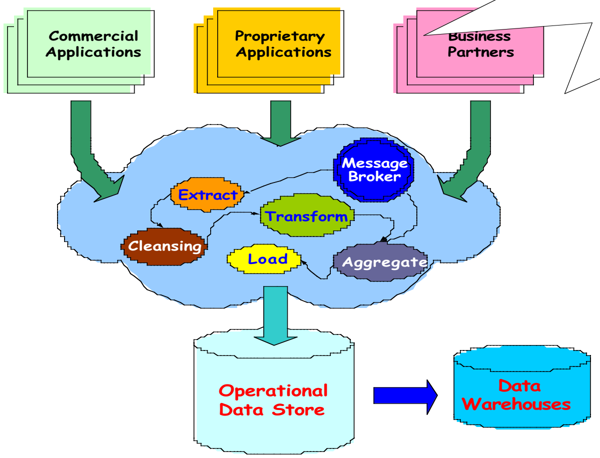
```
select store, sum(sales) from rd_sales_view group by store
```

Event Notification

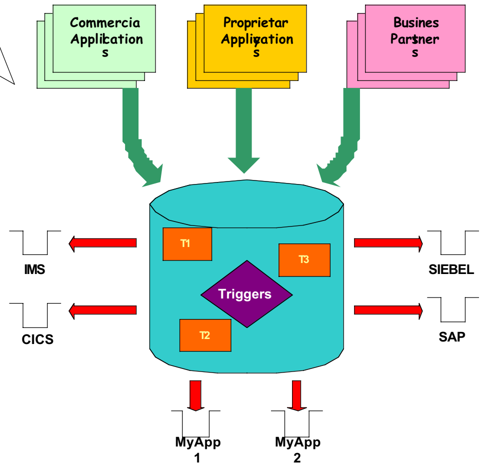


See Session F07, F08!

Information Aggregation

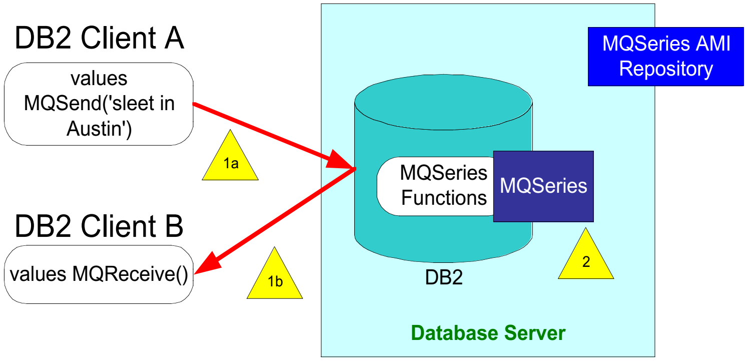


Information Forwarding

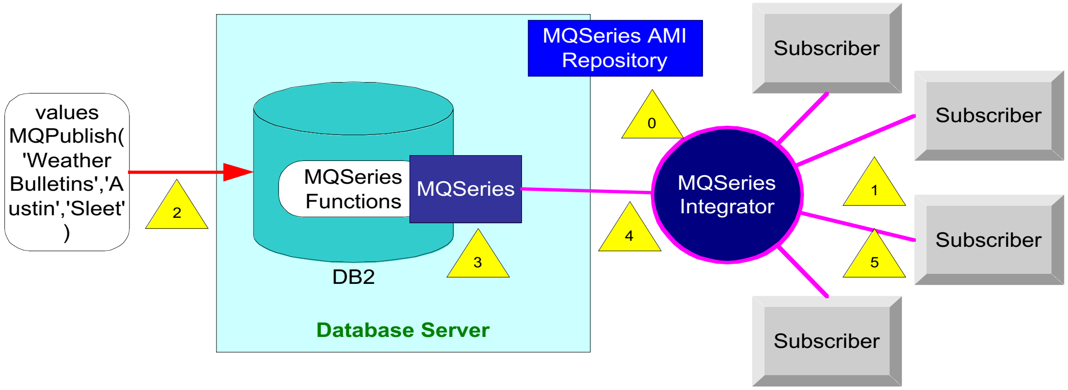


MQSeries Integration

- MQSeries server executes on the same machine as the DB2 Server
- DB2 clients may be local or remote.
  - ➡standalone applications, WebSphere servlets or EJBs
  - ➡local stored procedures



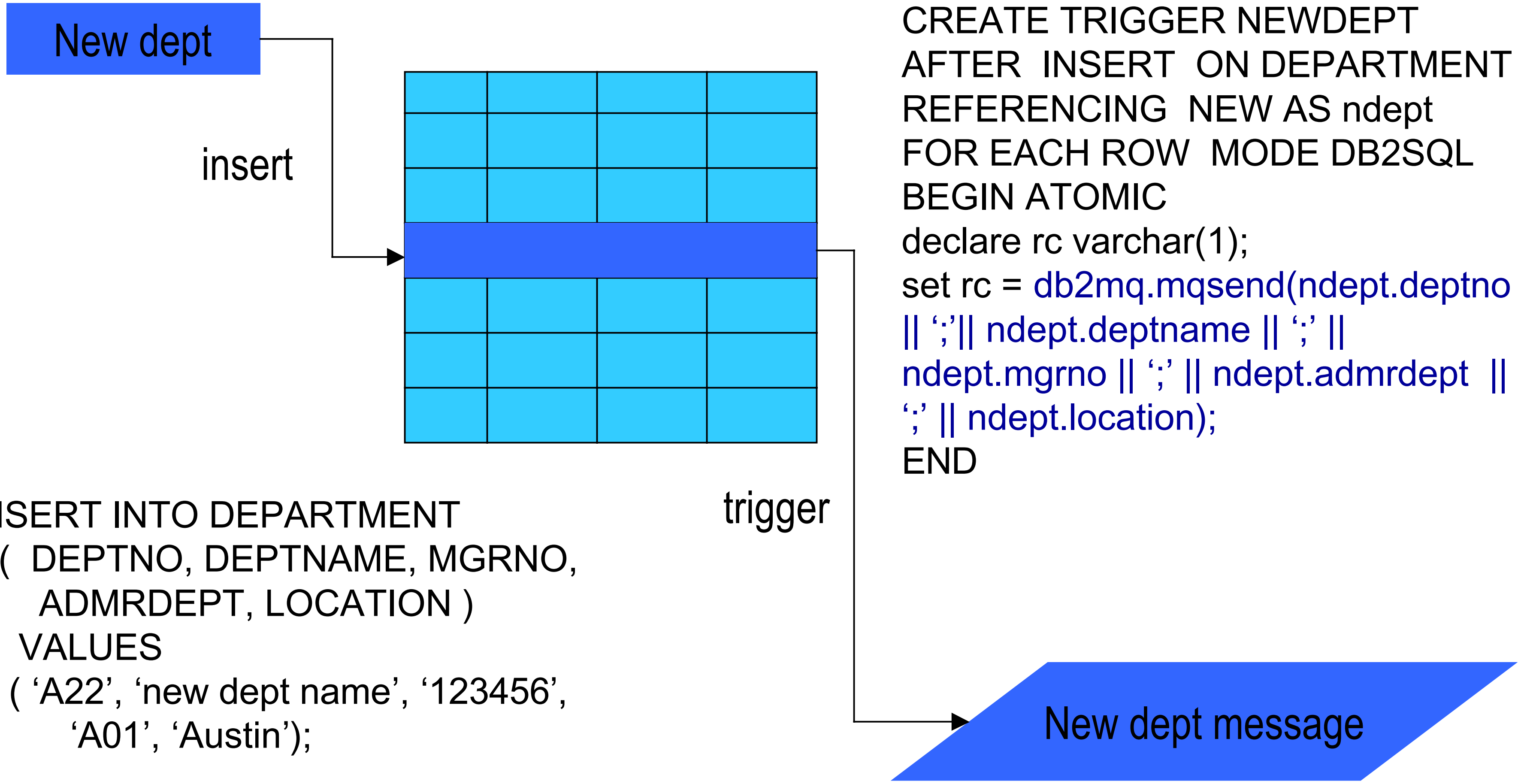
Basic Local Messaging



Simple Data Publication

- Supports both MQSeries publish/subscribe and MQSI V2
- Publication may be on-demand or automated through triggers
- Subscribers can register one or more topics

Easy Notification of Database Changes



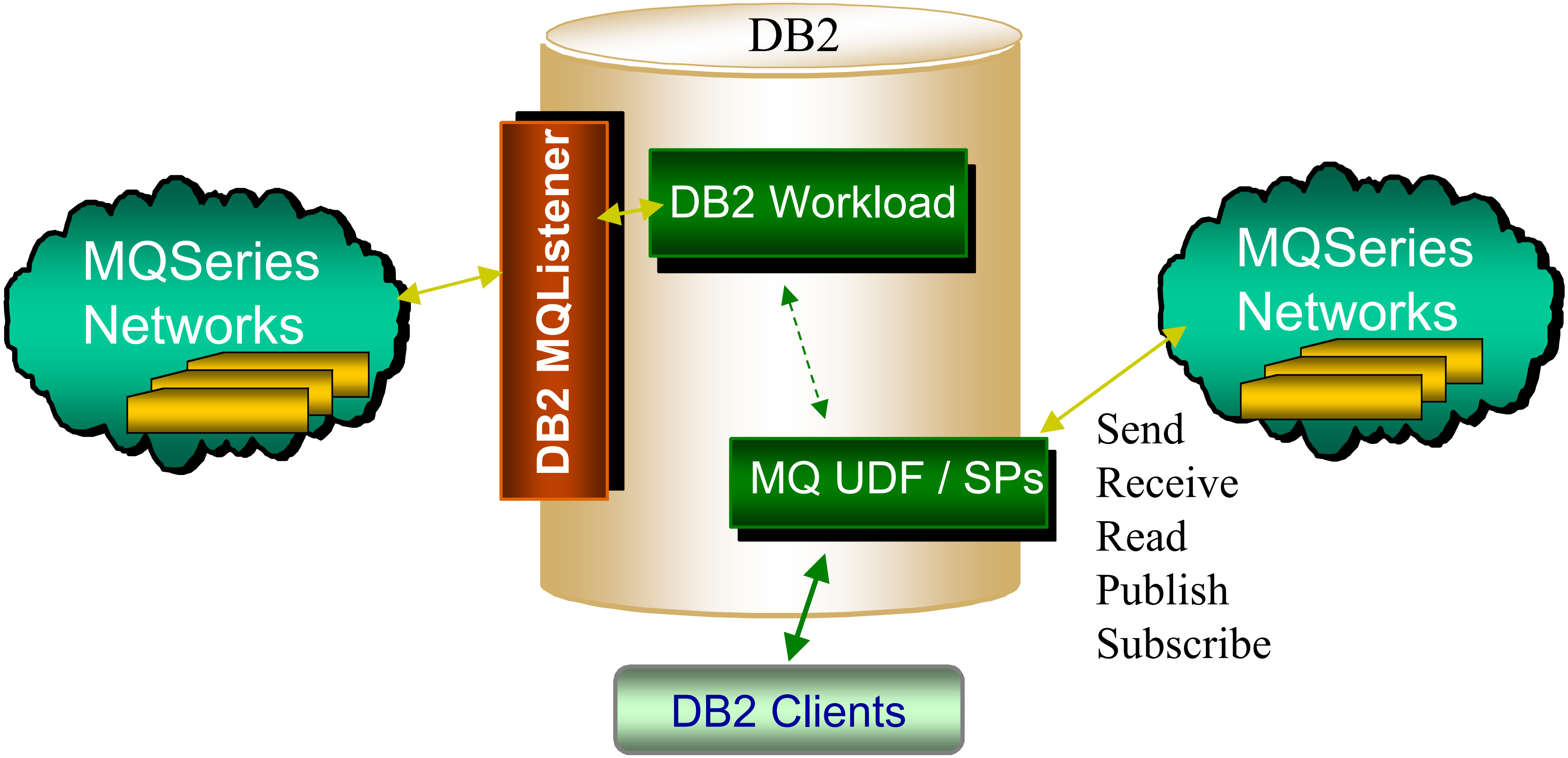
```
CREATE TRIGGER NEWDEPT
AFTER  INSERT  ON DEPARTMENT
REFERENCING  NEW AS ndept
FOR EACH ROW  MODE DB2SQL
BEGIN ATOMIC
declare rc varchar(1);
set rc = db2mq.mqsend(ndept.deptno
|| ‘;’|| ndept.deptname || ‘;’ ||
ndept.mgrno || ‘;’ || ndept.admrdept  ||
‘;’ || ndept.location);
END
```

```
INSERT INTO DEPARTMENT
  (  DEPTNO, DEPTNAME, MGRNO,
     ADMRDEPT, LOCATION )
   VALUES
  ( ‘A22’, ‘new dept name’, ‘123456’,
      ‘A01’, ‘Austin’);
```

DB2 / MQ Integration

- Provide a simple way to access MQSeries from database applications
  - Leverage the set-oriented nature of SQL with messaging
  - Provide a new kind of data source, leveraging the heterogeneous, federated capabilities of DB2.
  - Simplify the creation of operational data stores and data warehouses for business intelligence
  - Publish database changes
- Provide a gentle introduction to the MQSeries Family to database programmers and administrators
  - Integration with MQSeries, MQSeries Pub/Sub, MQSI

DB2 / MQ Integration Futures

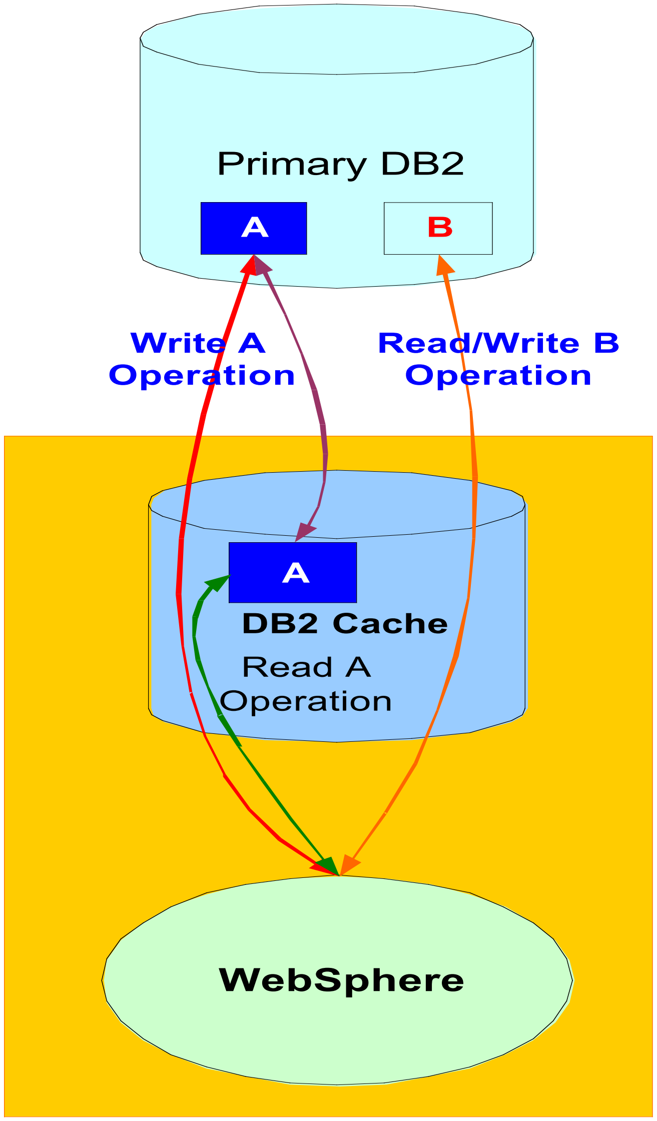


Send
Receive
Read
Publish
Subscribe

Replication and Caching

- Policy based data movement
- Applications
  - Warehouse and ODS applications
  - Consolidation and distribution
  - Application integration
  - Availability management
- Heterogeneous replication
  - DB2, Oracle, Sybase, Informix, Microsoft
- Table-based or transaction-consistent
- Point-in-time or continuous operation
- Embedded transformation
- Future directions
  - Publish over MQ
  - Transaction-based replication over MQ
  - Replication and AST linkage

Relational Database Caching



Improving Access Performance - Stages

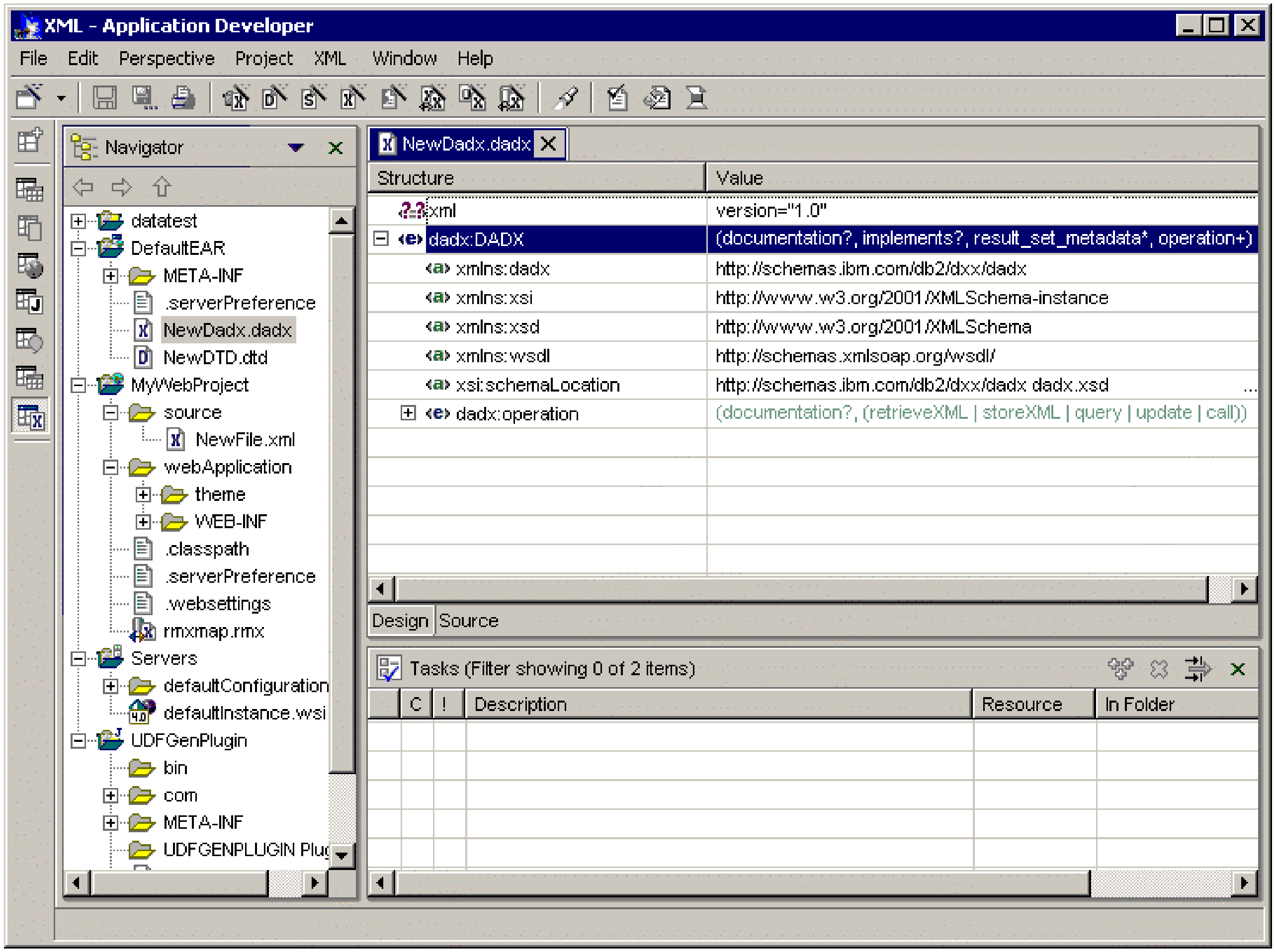
- Automated Summary Tables
  - Pre-computed values
  - Application indicates ability to use cache
  - Transparent to application
- Cached replica
  - Identical tables are created in the cache
  - Updates are made to the primary database
  - Replication manages cache refresh
- Cached snapshots over federated sources
  - Subsets of tables or federated views can be cached
  - Selectable update semantics
  - Replication manages cache refresh

Application Development

Three part approach:

- Core database tooling (e.g. Control Center, Development Center, …)
  - Administrative tasks, simple development of server-side functionality (SPs, UDFs, wrapper configuration, etc)
- WebSphere Studio tooling
  - Premier development environment, extended with data specific plugins
  - Web services tooling, XML tooling, SP/UDF tooling, etc.
- Microsoft Visual Studio tooling
  - Primary focus will be to ensure easy exploitation of DB2 from the Microsoft development environment

WebSphere Studio Plug-ins



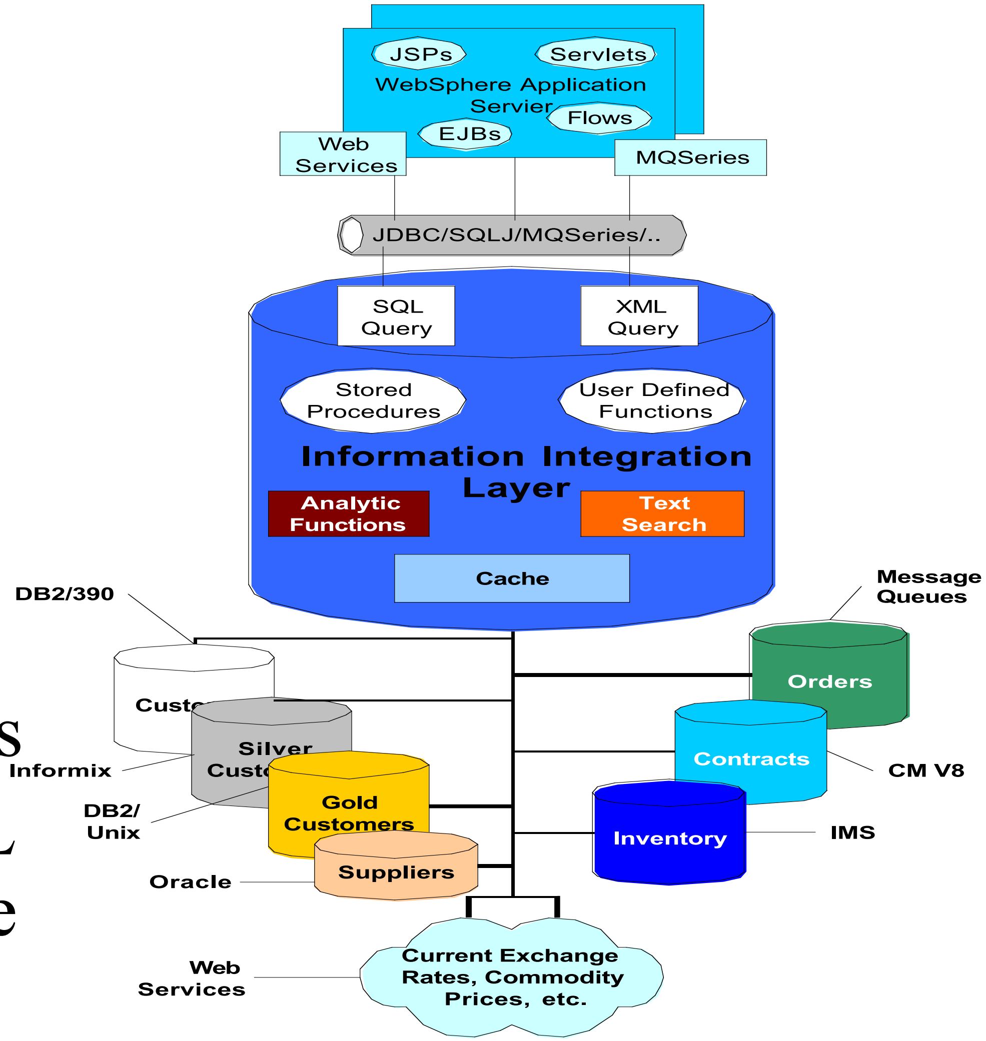
Structure	Value
?-? xml	version="1.0"
⟨e⟩ dadx:DADX	(documentation?, implements?, result_set_metadata*, operation+)
⟨a⟩ xmlns:dadx	http://schemas.ibm.com/db2/dxx/dadx
⟨a⟩ xmlns:xsi	http://www.w3.org/2001/XMLSchema-instance
⟨a⟩ xmlns:xsd	http://www.w3.org/2001/XMLSchema
⟨a⟩ xmlns:wsdl	http://schemas.xmlsoap.org/wsdl/
⟨a⟩ xsi:schemaLocation	http://schemas.ibm.com/db2/dxx/dadx dadx.xsd ...
⟨e⟩ dadx:operation	(documentation?, (retrieveXML | storeXML | query | update | call))

C	!	Description	Resource	In Folder

- Available today:
  - XML extender tooling - DAD
  - JSP DataTags
  - **WebService data access - DADX**
  - SP/UDF development
  - **Web service UDF client & wizard**
  - **MQ UDF client and wizards**
- Under development:
  - Workflow UDF client & wizard
  - Advanced data access
- Under design:
  - Data workflow nodes

Programming Model Integration

- Facilitate DB2 access and exploitation from all key programming environments
- Leverage information integration capabilities from WebSphere and MQSeries environments
- Develop seamless XML query, handling, storage



Summary

- Information integration is one (key) aspect of IBM's overall approach to delivering a business integration infrastructure
- Information integration enables integration of traditional and emerging data sources for
  - Real-time read and write access
  - Transformation for business analysis
  - Data placement for performance, currency, and availability
- Information integration is the natural evolution and next major challenge for data management software
- Xperanto is the code name for IBM's efforts to address customers information integration requirements
- IBM is well positioned to lead the market

RETURN TO INDEX