DB2 Information Management Software

Information Integration Solutions

Giulia Caliari
giulia_caliari@it.ibm.com

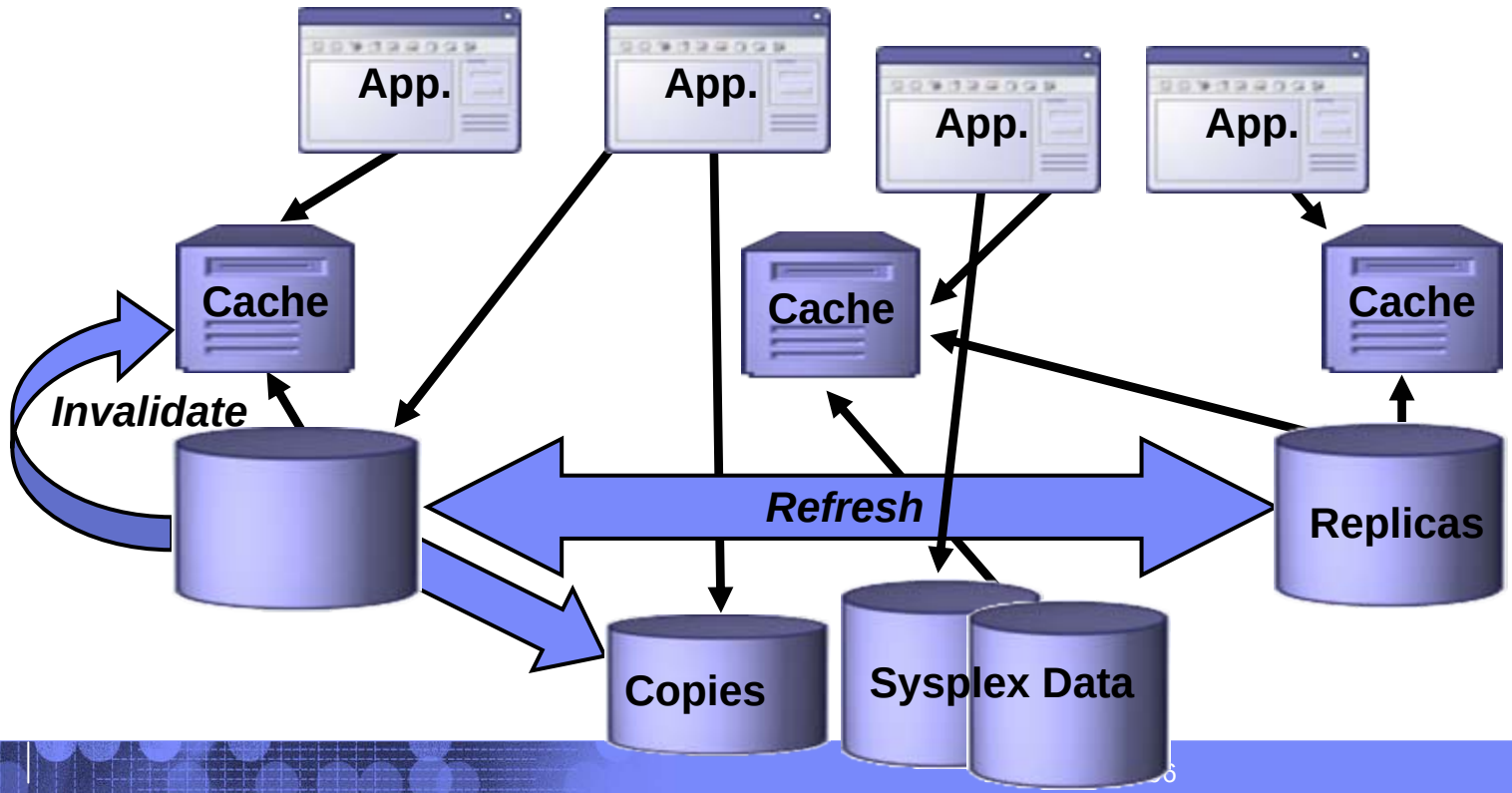
Bari, 20.09.2006

Agenda

- **Information Integration Overview**
- **Data Federation, Data Replication and Event Publishing**
- **Data analysis, cleansing and transformation**
  - ProfileStage
  - MetaStage
  - QualityStage
  - DataStage
- **What's coming next**

Today's World: Complex and Costly

- **Heterogeneous, distributed data**
- **Applications create and maintain caches and replicas of data**
- **Proliferation of copies**
- **30%- 50% design expenses go to copy management**
- **No guaranteed quality of service nor feedback**



Customer Business Issues

- Too much information and not knowing what's important
  - Not using demand signals to drive supply chain
  - Not using customer analysis to tailor marketing and sales
  - Not leveraging valuable unstructured information

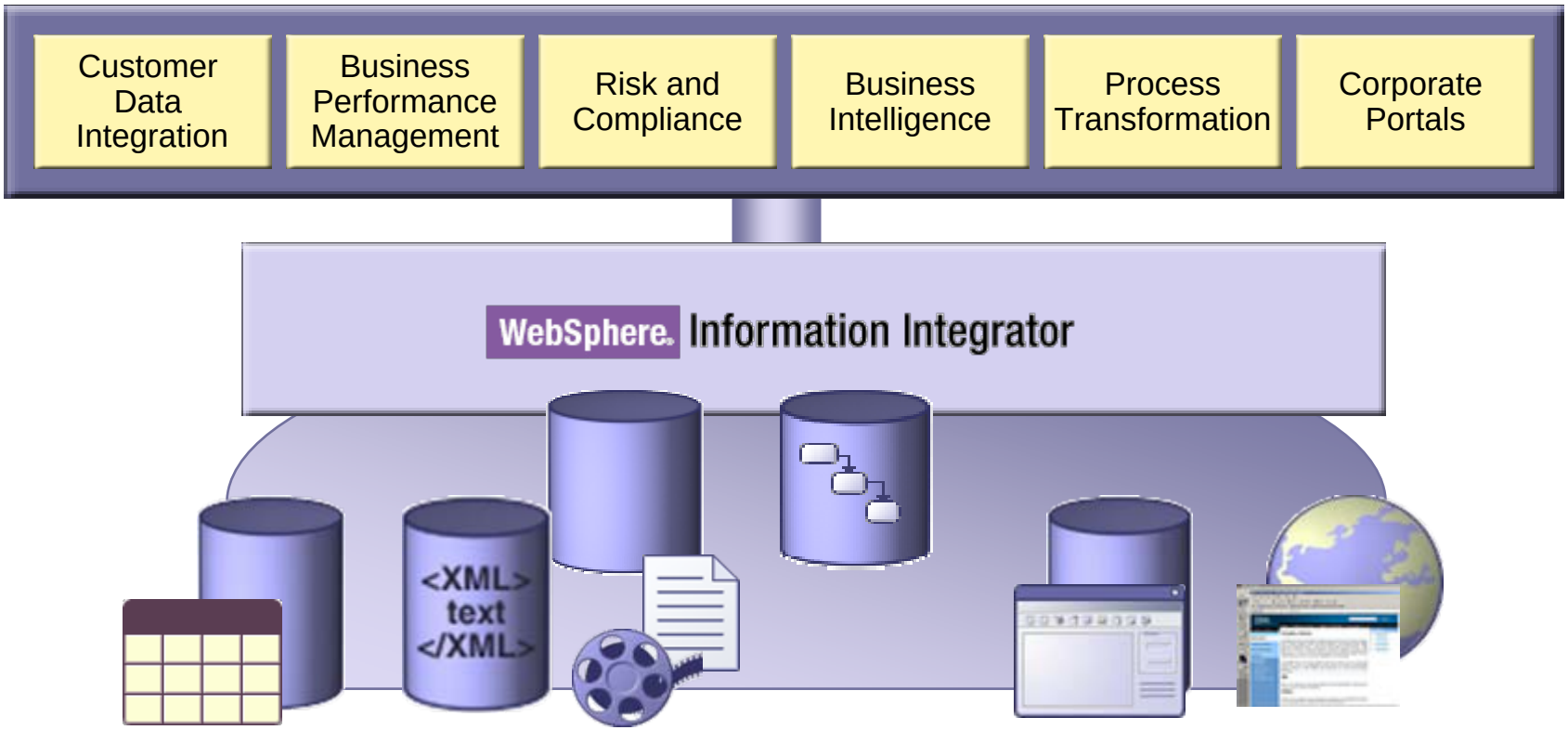
- Multiple versions of the truth
  - Problems managing customer, product and partner interactions
  - Regulatory compliance inhibited by poor transparency

- Lack of trusted information
  - Incomplete, out-of-date, inaccurate, misinterpreted data
  - Difficult to understand or control how information is used

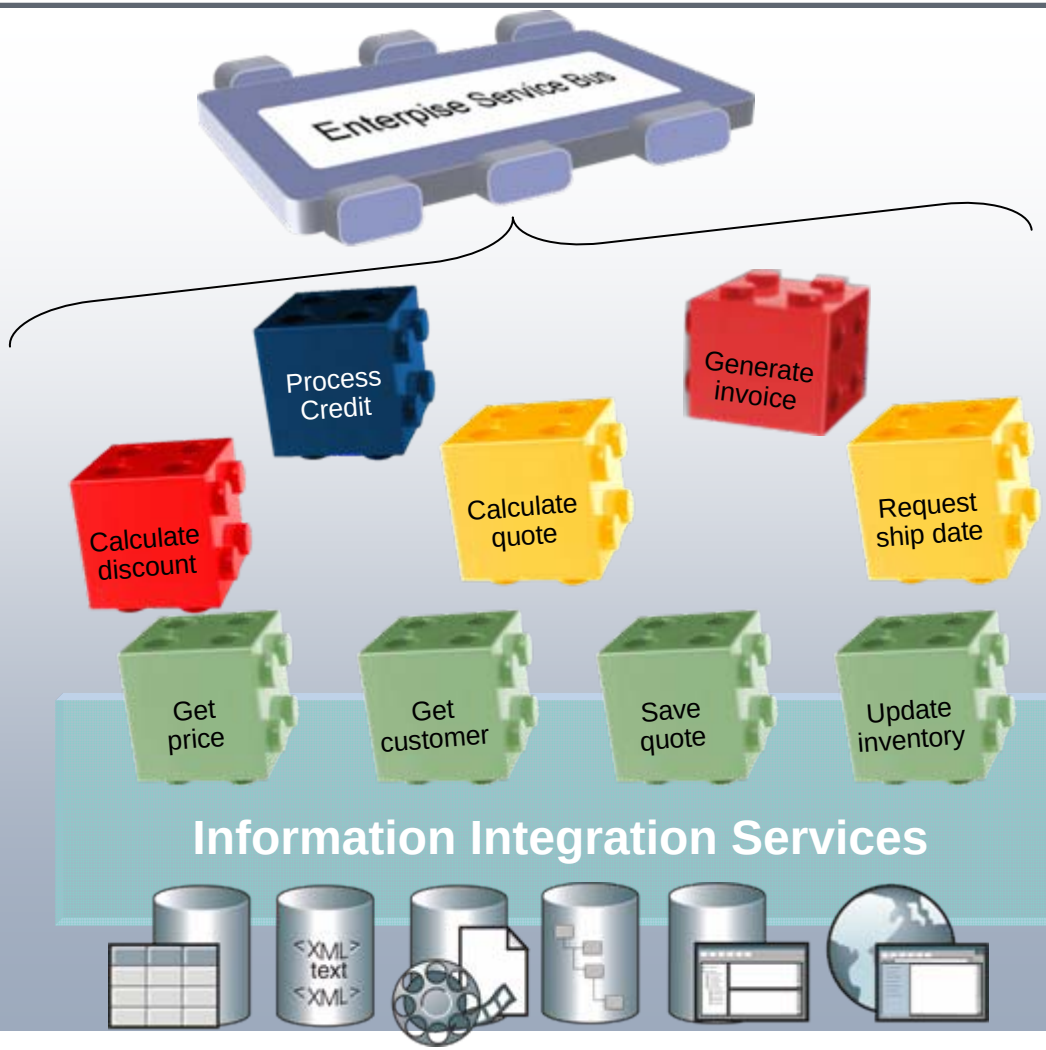
- Lack of agility
  - Inability to take advantage of opportunities for innovation
  - Escalating costs due to inflexible systems and changing needs

WebSphere Information Integrator

A strategic information integration platform to help enterprises become on demand businesses



Information Integration Services

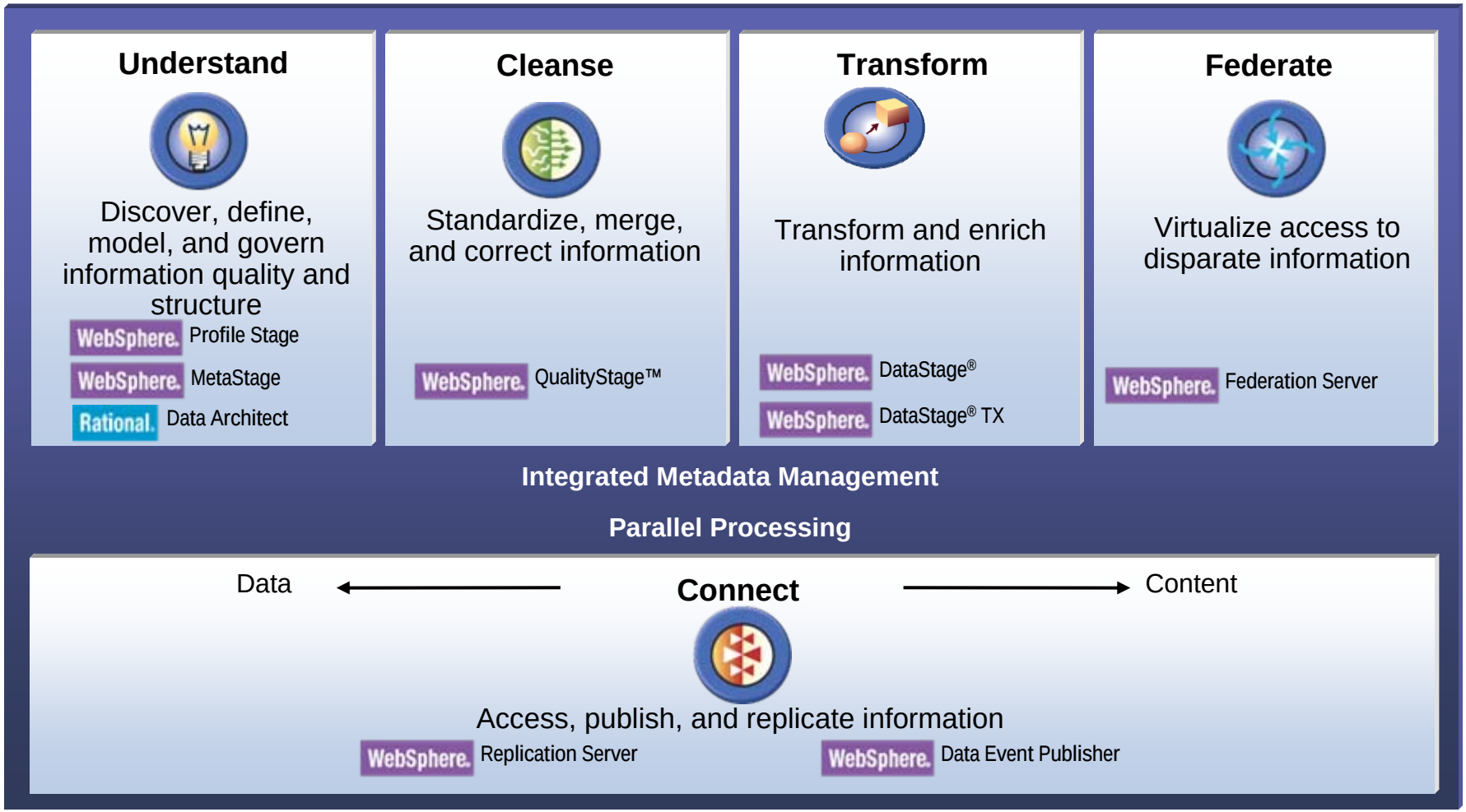


Information integration enables information to be made available as a service

Publishing consistent, reusable services for information that make it easier for processes to get the information they need from across a heterogeneous landscape.

- Select data from source 1
- Select data from source 2
- Match and link records
- Transform data to target

The IBM WebSphere Information Integration Platform



DB2 Information Management Software

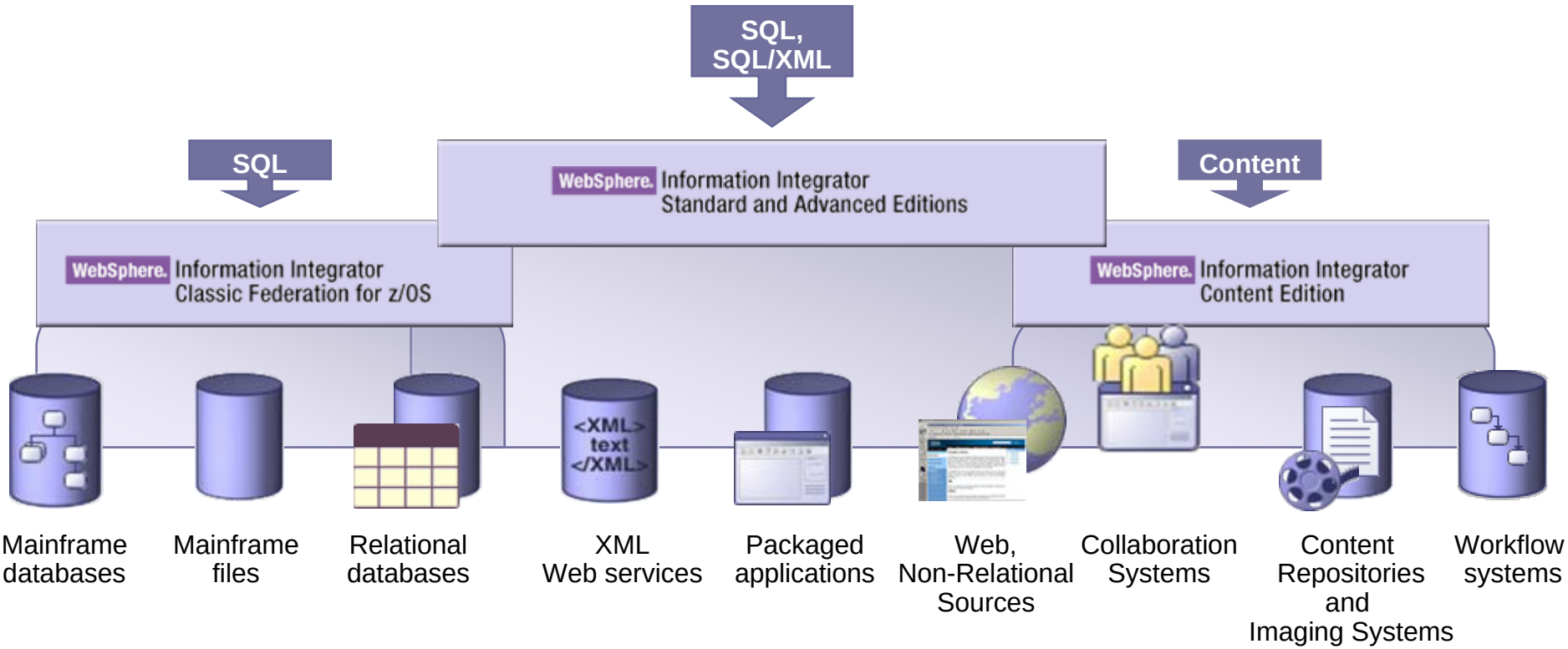
Data Federation, Replication and Event Publishing

Bari, 20.09.2006

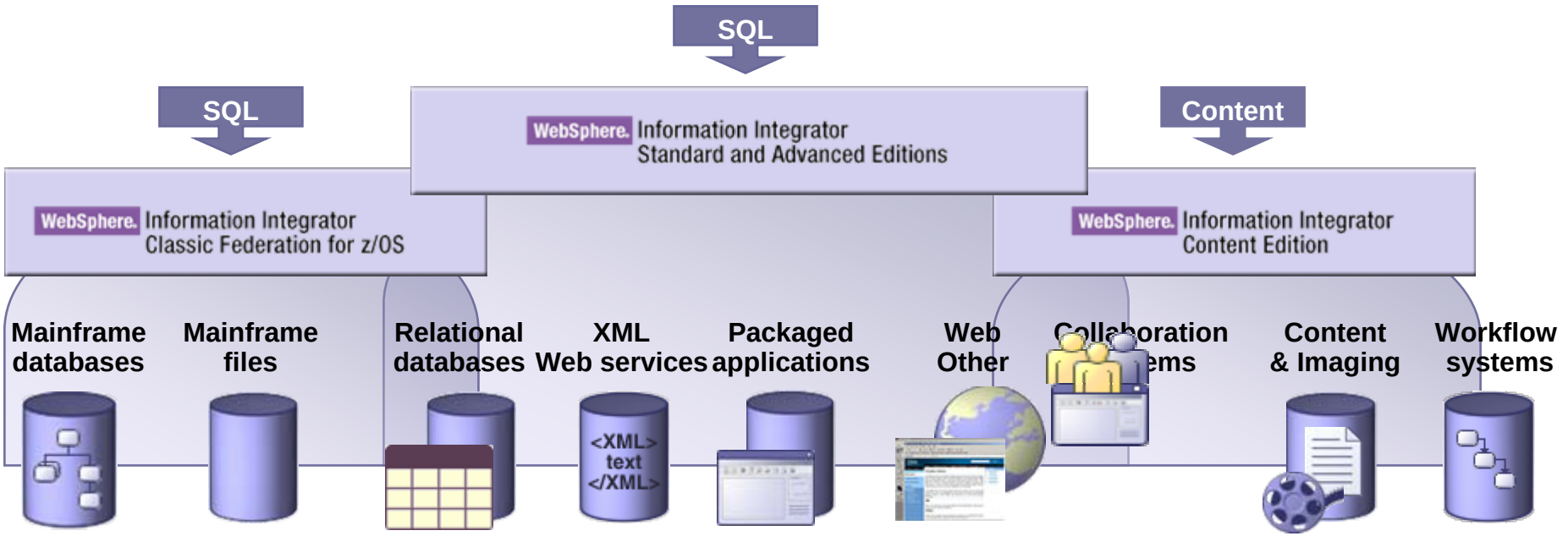
Federation: Virtualized Information Access

Access diverse and distributed information as if it were in one system

Single sign on – Unified views – Common language – Web services or Java API Query and update – Optimized access

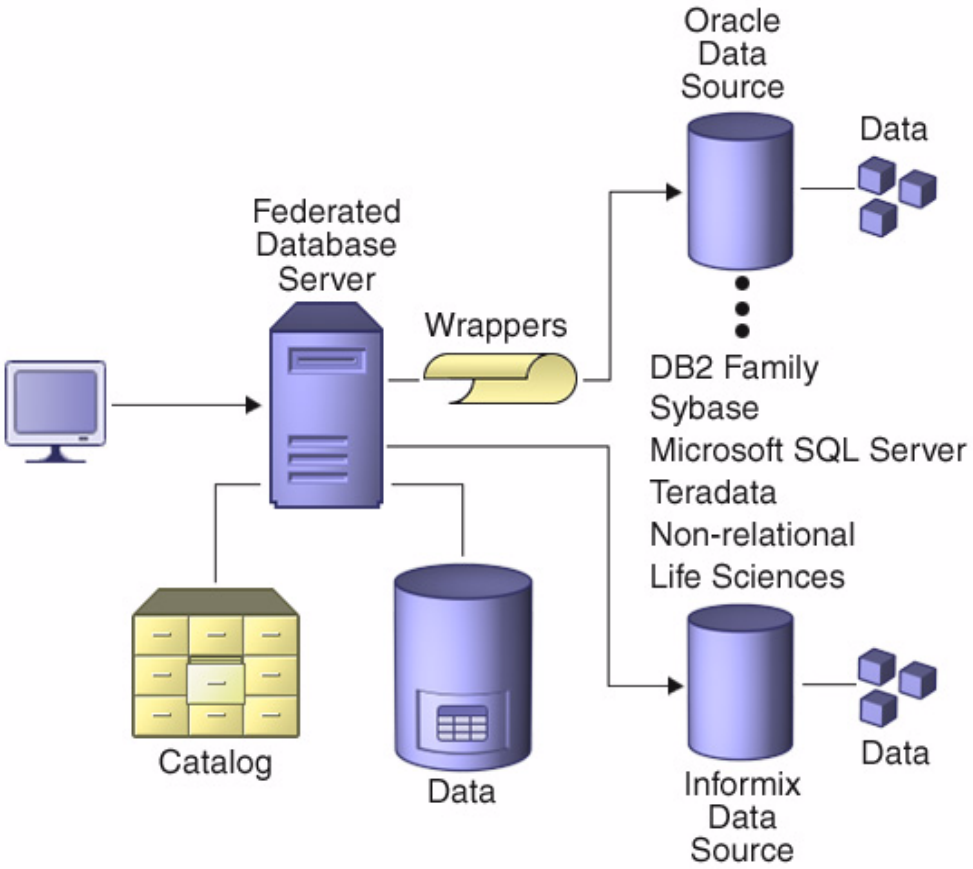


Federated Sources



Plus partner tools and custom-built connectors extend access to more sources

WebSphere Federation Server 9.1



Transparent
- Appears to be one source
- Independent of how and where data is stored
- Applications continue to work despite of any change in how data is stored

Heterogeneous
- Accesses data from diverse sources
- Relational, Structured, XML, messages, Web, …

Extensible
- Bring together almost any data source.
- Wrapper Development Toolkit

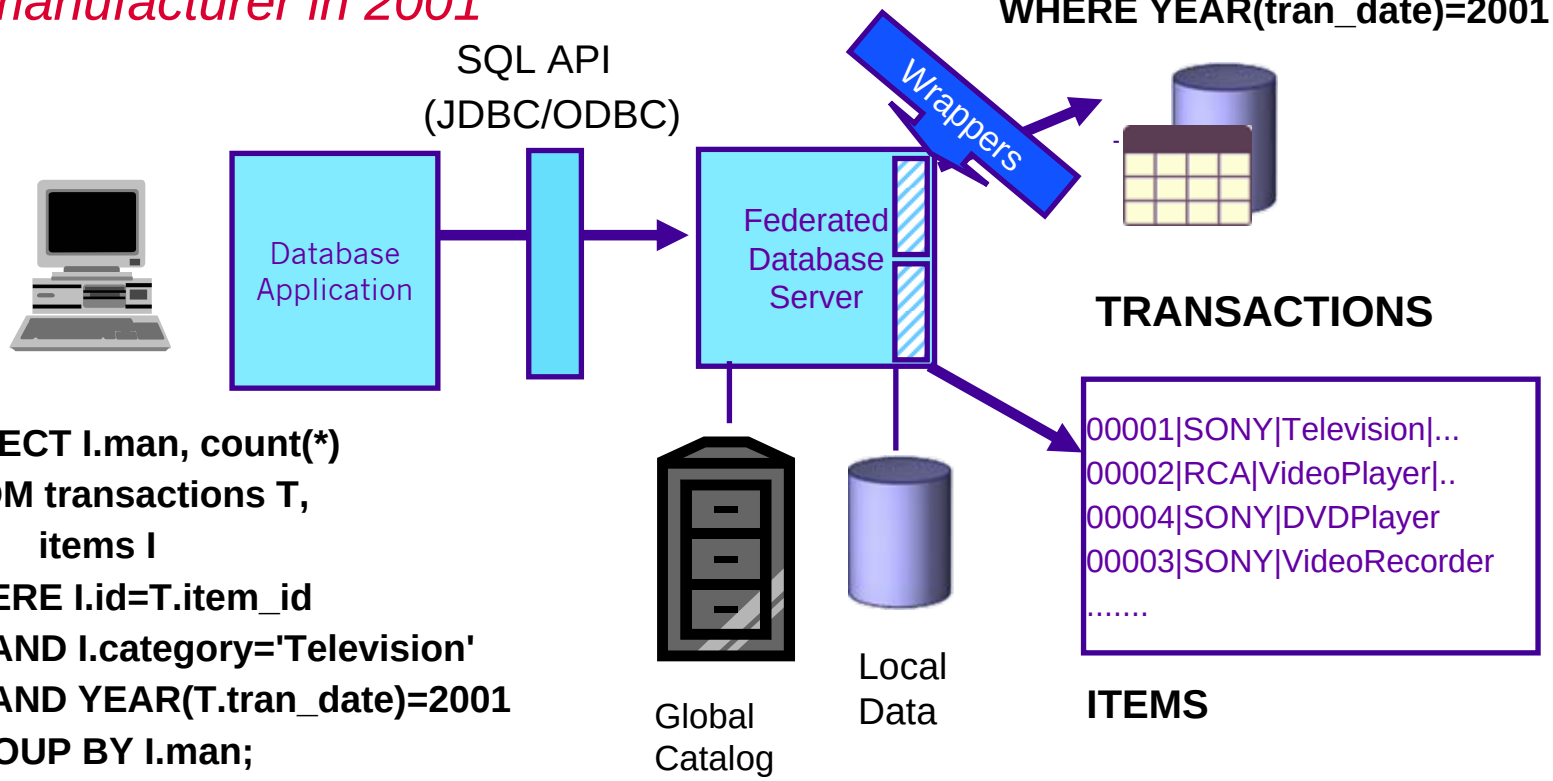
High Function
- Full query support against all data
- Capabilities of sources as well

Autonomous
- Non-disruptive to data sources, existing applications, systems.

High Performance
- Optimization of distributed queries

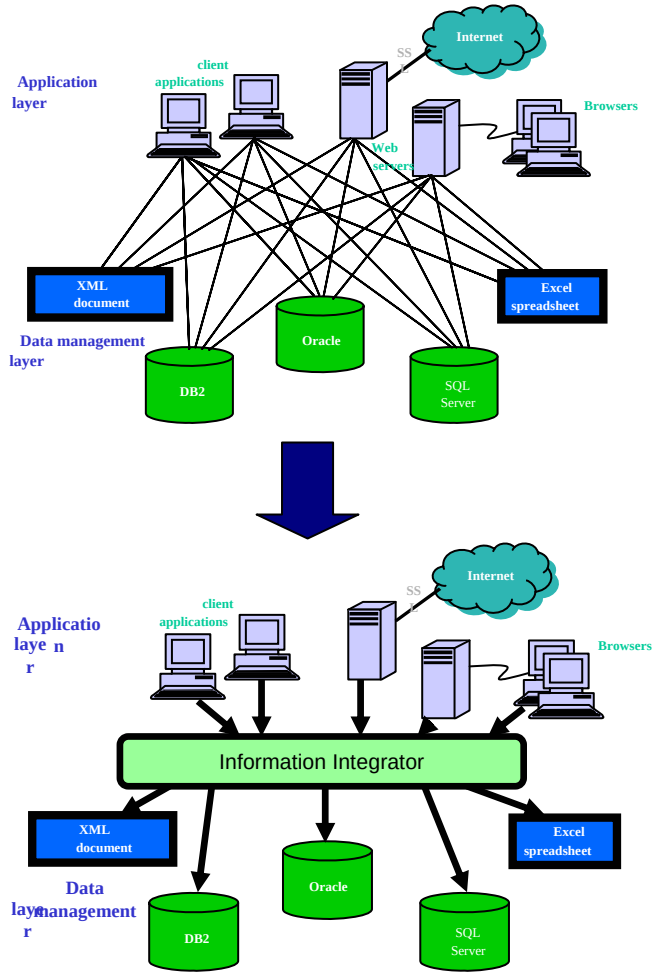
An Example of using Federated Database

List the number of TV sales per manufacturer in 2001



WebSphere Federation Server 9.1

- **Semplifica lo sviluppo applicativo**
  - Trasforma attività di sviluppo in attività di configurazione
- **Prestazioni elevate per operazioni (join) su (molte) tabelle distribuite e tecnologie diverse**
  - Capacità di ottimizzazione, push-down, statistiche sul sorgente, ecc.
  - No limite al n. di tabelle
- **Ricchezza di funzionalità**
  - Tutto l'SQL del DB2V8 più funzioni specifiche del sorgente
- **Approccio aperto, estendibilità**
- **Facilità di configurazione (tool grafici del DB2 Control Center)**



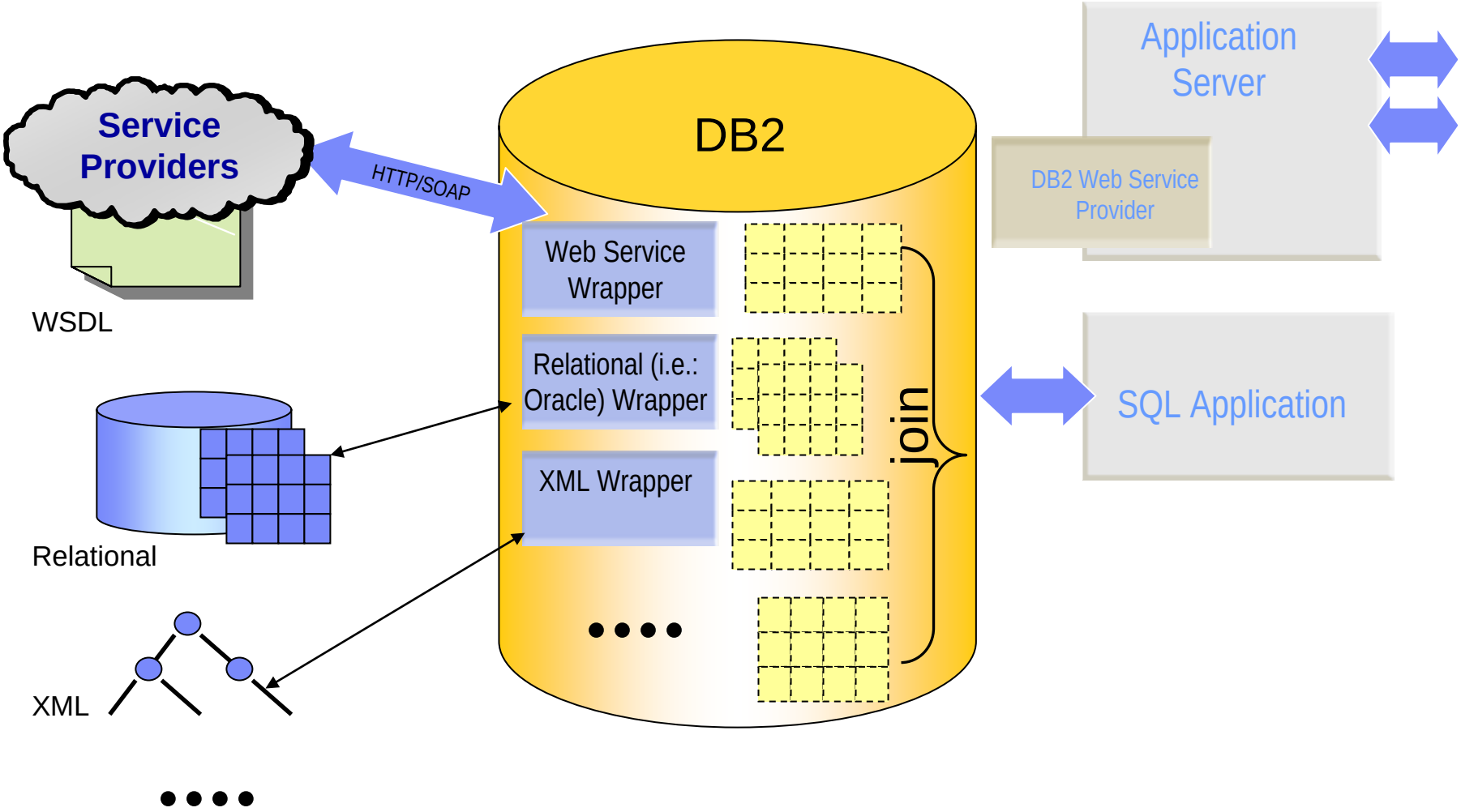
Federated Queries Make Integration as Easy as SQL

```
SELECT   parameters_return_billto_key as BILL_TO_KEY,
         billto_company_name,
         parameters_return_shipto_key as SHIP_TO_KEY,
         CASES_SHIPPED,
         GROSS_SALES,
         URL

FROM     GETKEYSSOAP_GETKEYSREALTIME_NN,

         GLOBAL_SALES_TRAN_NN,

         BILLTO_DIMENSION,

         URL_INVOICES

WHERE    getkeysrealtime_ship_to_number = '13546'
  and    getkeysrealtime_ship_to_number = URL_INVOICES.shipno
  and    ltrim(rtrim(translate(ship_to_number, ' ', x'0a')))
               = getkeysrealtime_ship_to_number
  and    parameters_return_billto_key = billto_key
  and    ltrim(rtrim(translate(sales_order_number, ' ', x'0a')))
               = URL_INVOICES.orderno
```

Single SQL Query Joins:

- ← **Web Service**
- ← **XML Documents**
- ← **Data Warehouse**
- ← **Unstructured Data**

Data Federation and Web Services



Value of Federation

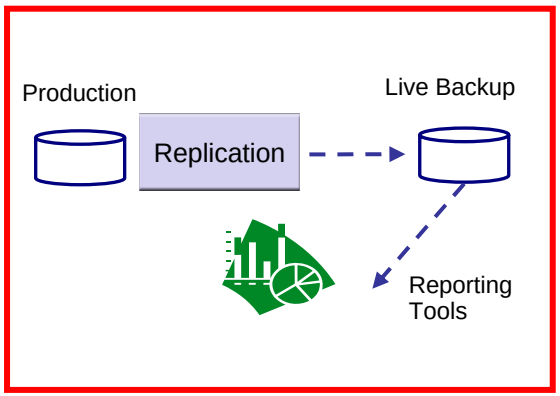
- **Speed time to market for new applications**
  - Simplify and enrich portal development
  - Reduce hand-coding by half
  - Reduce skills requirements
  - Use familiar SQL programming model and existing tools
  - Build on a standards-based, strategic integration platform
- **Enhance value and insight from existing assets and applications**
  - Work within your existing infrastructure
  - Extend existing warehouses
  - Combine existing data and content assets in new ways
  - Facilitate cross-divisional reporting
- **Increase control over IT costs**
  - Reduce need to rip and replace
  - Reduce need to manage redundant data

Data Replication

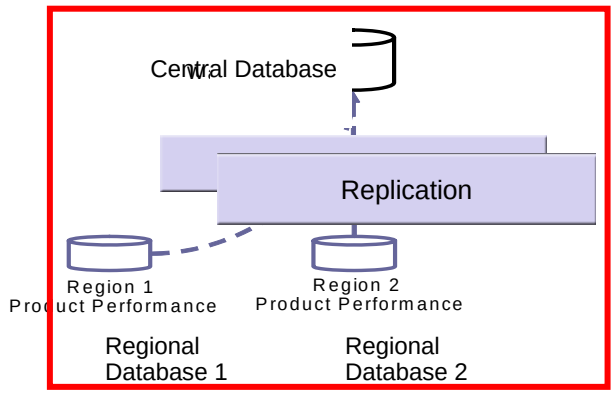
- **Replication can automatically keep multiple data locations consistent, and each target can be different to match the users needs. This includes different latency or differing timeliness of the data.**
  - replication can be by time interval, event driven or continous
  - different enhancements (derivations, summarization, transformations)
  - different formats to each target
- **Data Replication**
  - High availability of production applications
  - Distribution of data to other locations
  - Consolidation of data from other locations (Data Warehuse and ODS applications);
    - Data Replication as part of the ETL process
  - Bidirectional exchange of data with other locations
  - Some variation or combination of the above

Many Usage Scenarios For Replication

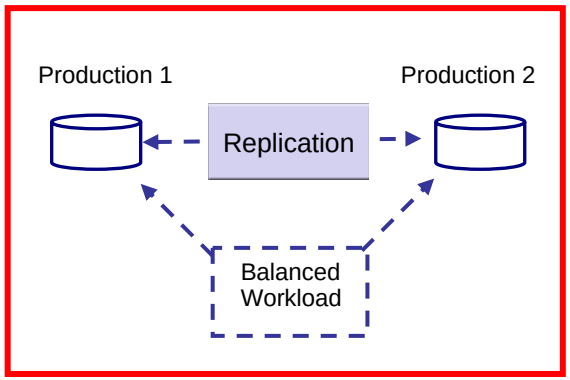
High Availability



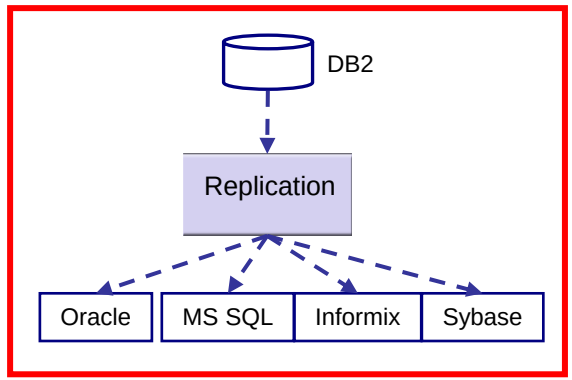
Rollup (many to 1)



Peer To Peer



Distribution (1 to many)



Websphere Replication Server: SQL Replication

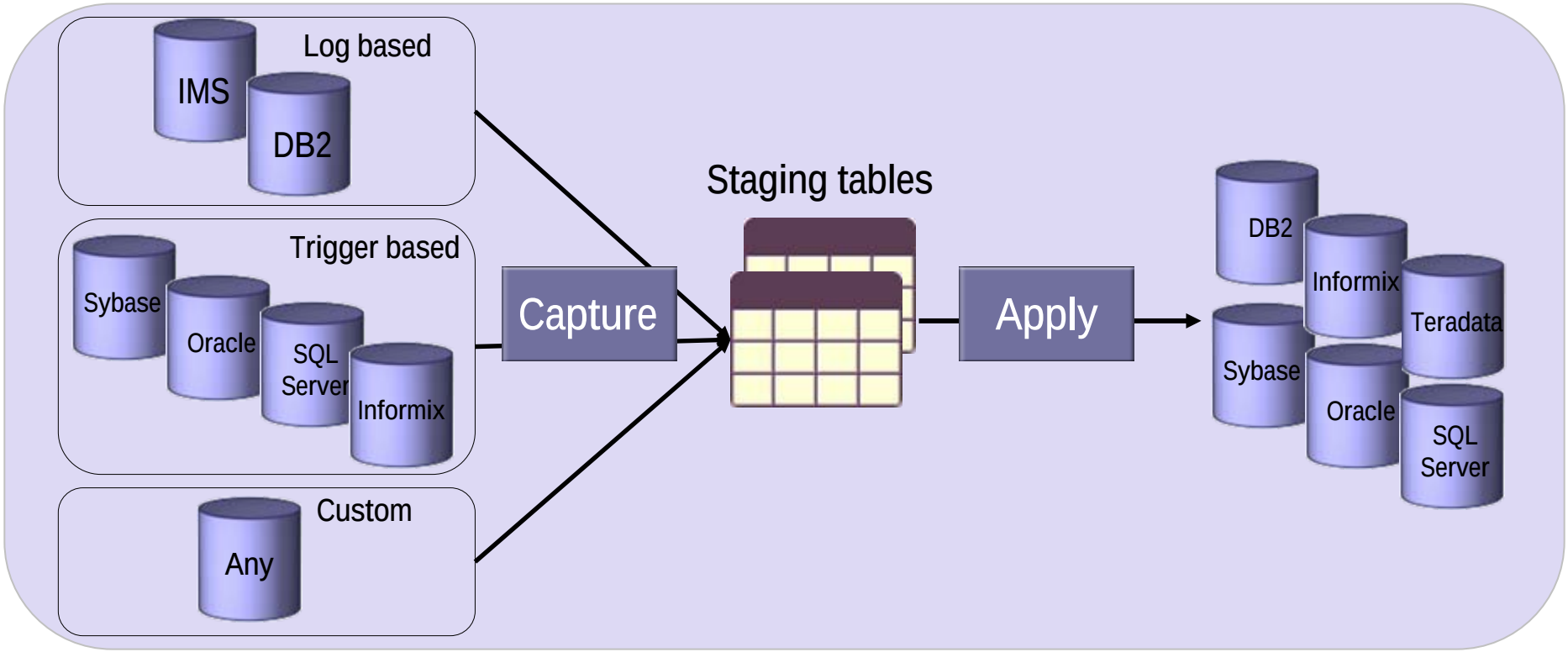
Flexible transformation and scheduling

Function

- Filter and transform, Apply by table or by transaction
- Choose latency by schedule, interval, event, or continuous
- Replicate point-to-point, for distribution, or for consolidation
- Maintain snapshots, simple copies, histories, or aggregates

Usage

- Business intelligence
- Distribution and consolidation
- Application integration



Websphere Replication Server: Q-Replication

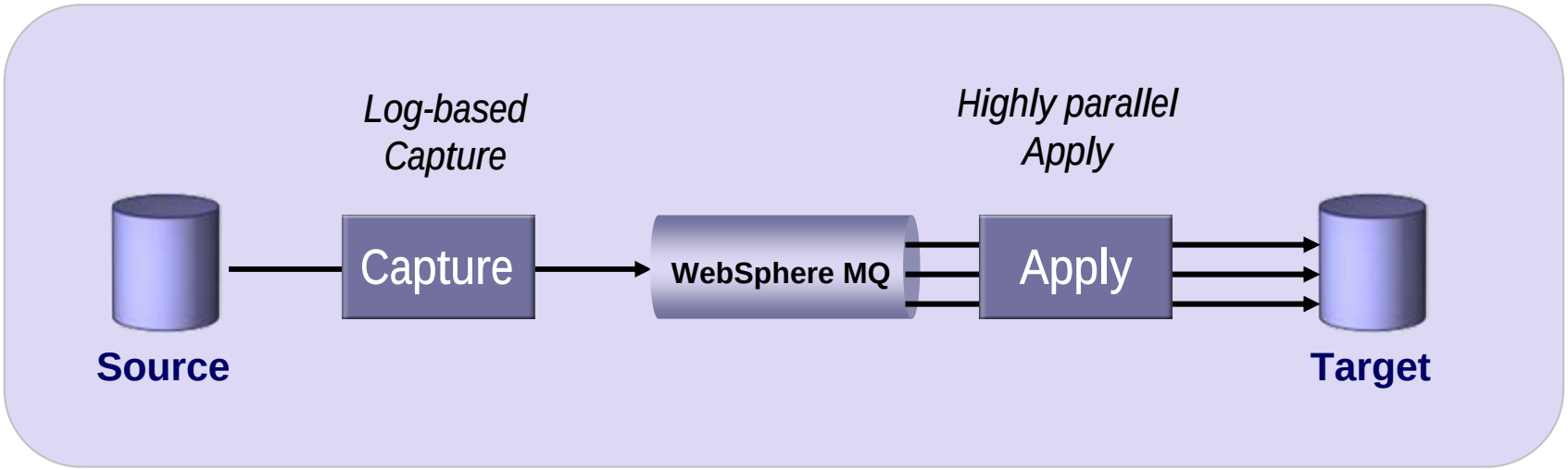
New replication architecture for delivering extremely low latency replication for peer-to-peer environments

Function

- **Replicate rows or transactions**
- **Filter and transform data**
- **Detect and resolve conflict**
- **Configure and monitor environment**

Usage

- **High availability**
- **Workload distribution**
- **Application integration**



Mazda

Improved Sales Process Via Up to the Minute Inventory Information

Challenge

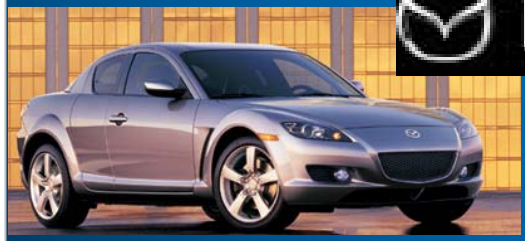
- Support 700 dealers in USA
- Trouble matching customer demand with available inventory
- More current data needed to track sales achievements with period-end goals

Solution

- **Sales and inventory information is replicated every minute to portal server**
- **Improved access to current data without changes to existing IT infrastructure**

"Within 5 weeks of receiving the [WebSphere] Information Integrator product we were able to implement it in our … environments. It now provides us up to the minute sales activity."

Joe Neria, Software Consultant. *Mazda*



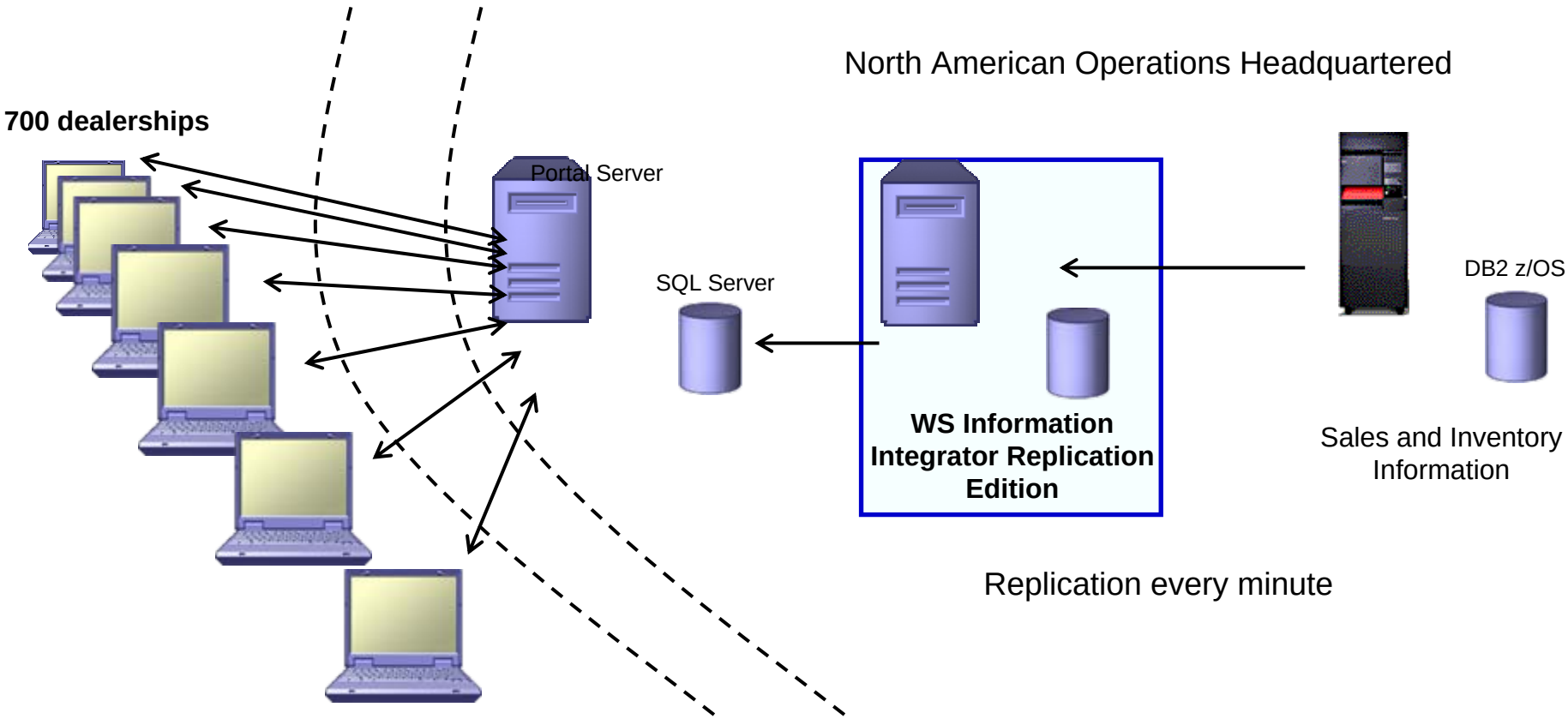
Business benefits

- Increased auto sales
- Improved dealer satisfaction
- Currency of information improved by 93%

Technology benefits

- Re-used existing application and data base infrastructure
- Decreased network load compared to full data refreshes 4 times an hour
- Ease and speed of deployment

Mazda North America



CitiStreet

Synchronizing User Access Data

citistreet
A State Street and Citigroup Company

Challenge

- Support single sign-on access through both Web and IVR applications ensuring 24x7 portal access for plan participants and sponsors

Solution

- **Support redundant, active single sign-on applications for failover processing replicating profile changes between them in real time.**

"Since nearly 10 million of CitiStreet customers are offered 24-hour access to their retirement accounts, the company can't afford downtime and must be able to replicate data changes when they happen. We fully replicate our database over redundancy data lines, so to us the stability and speed of that asynchronous replication is strategic for us."

Barry Strasnick , CIO
CitiStreet

Overview

- CitiStreet is one of the largest and most experienced global benefits providers servicing over 9 million plan participants across all markets. CitiStreet was formed in partnership between subsidiaries of State Street Corporation and Citigroup

Business benefits

- Ensure application availability for plan participants and sponsors
- The new solutions from IBM will improve data integrity with a reduced level of maintenance

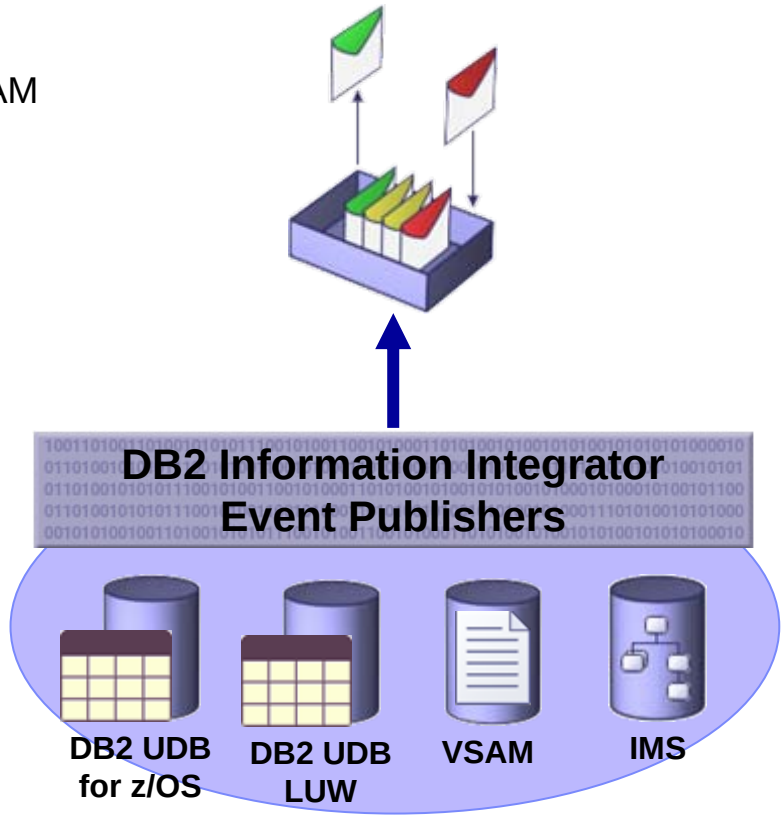
Technology benefits

- Maintain bi-directional synchronization of profile updates (approx 175,000 updates daily) in real time

Websphere Data Event Publishing

- **What is *Event Publishing* ?**
  - Capture changed-data from DB2, IMS and CICS/VSAM
  - Correlate by transactions within single database
  - Extract to consistent and documented XML format
  - “Publish” to WebSphere MQ queue
  - Received & Processed by any MQ “listener”

- **Why *Publish* data?**
  - Application to Application Messaging:
  - Event Notification
    - Stream changed data information to Web interface
    - Stream only particular events of interest (filter data)
  - Warehouse / Business Intelligence
    - Integrate captured changed data with an ETL tool
    - Perform very complex transformations
  - MQ provides guaranteed delivery
    - Works even when the target is not available



WebSphere Data Event Publishing

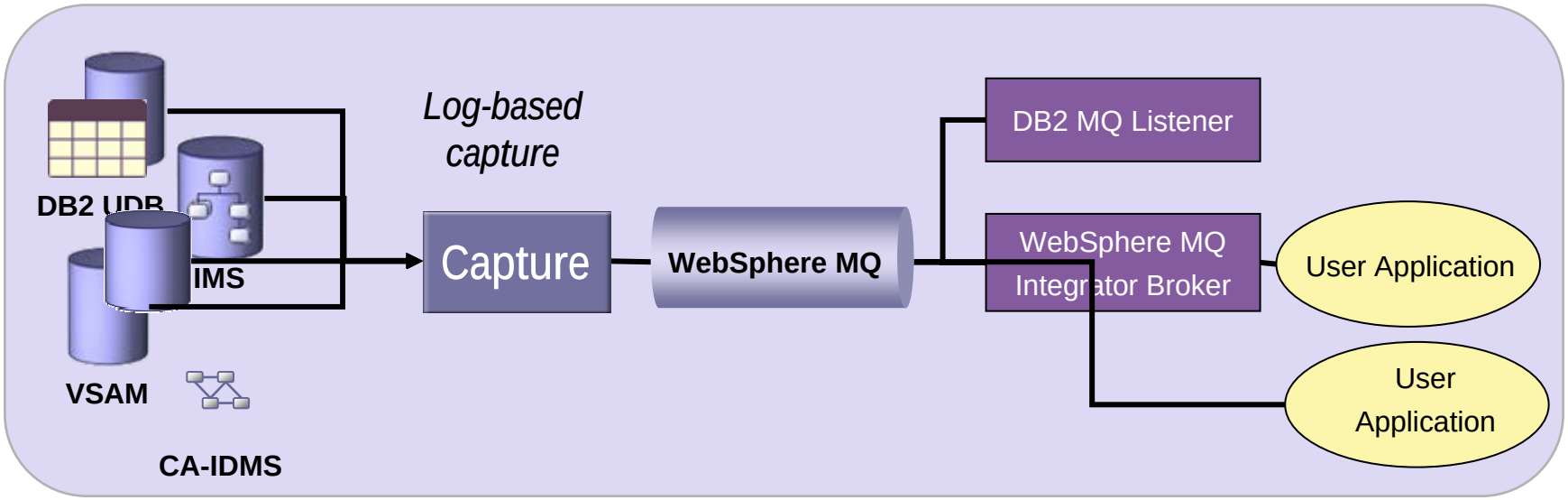
Capture database changes and publish them as XML messages to WebSphere MQ

Function

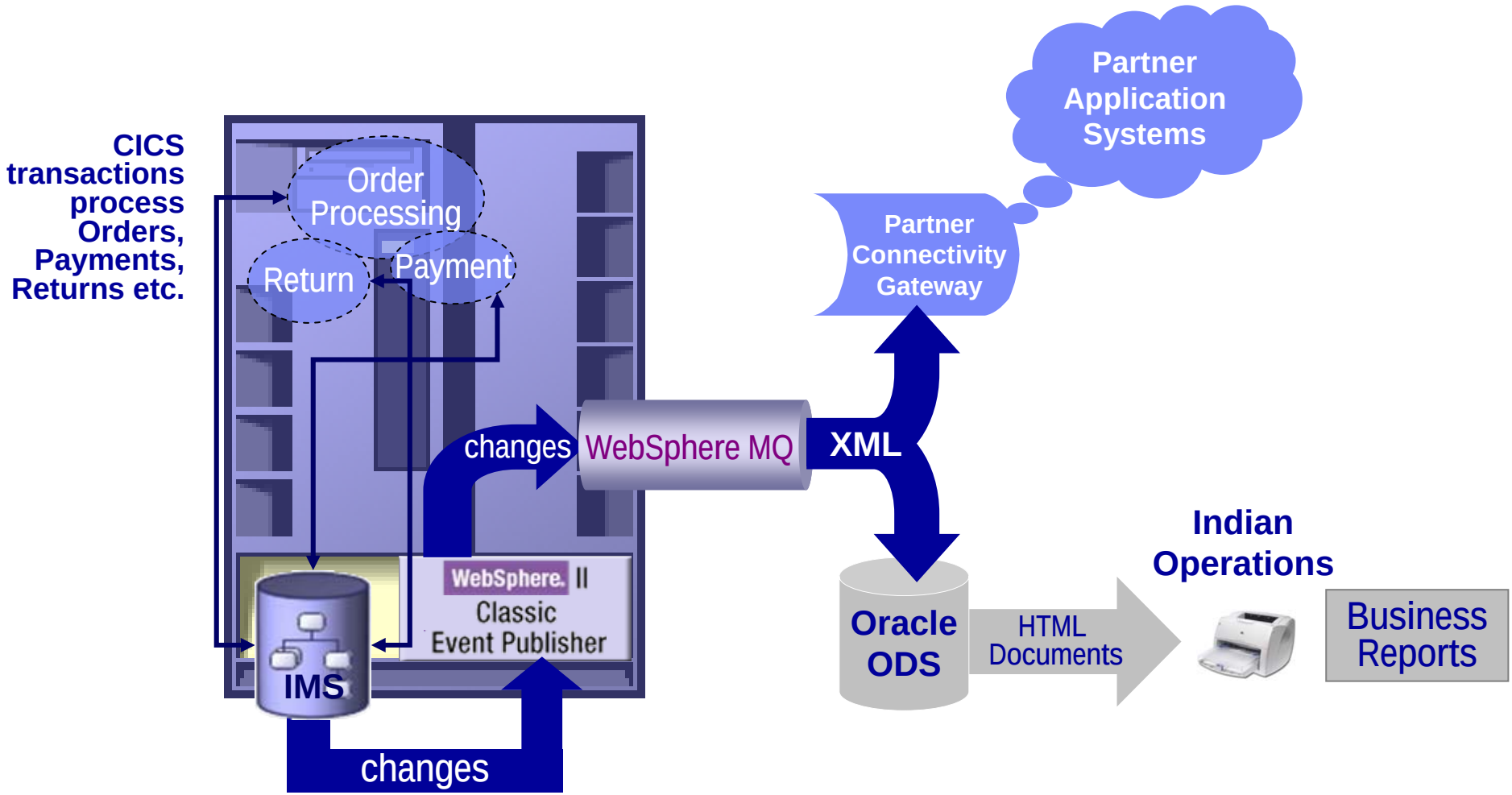
- **Publish events to a message queue**
- **XML self-describing format**
- **Wizard-driven configuration**

Usage

- **Application to application messaging**
- **Initiate business processes**
- **Source for ETL tool**



Data Event Publishing at a Major Technology Reseller

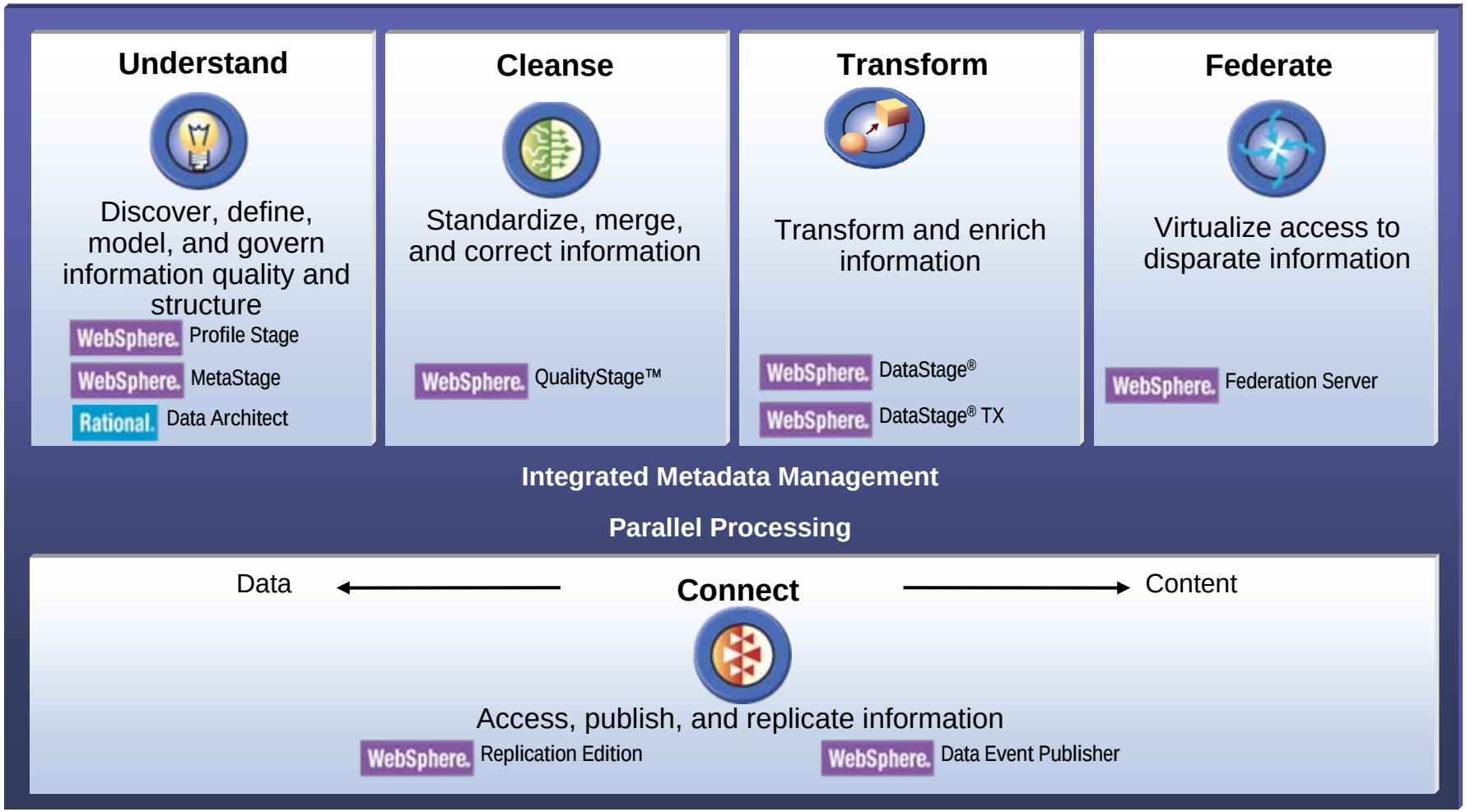


DB2 Information Management Software

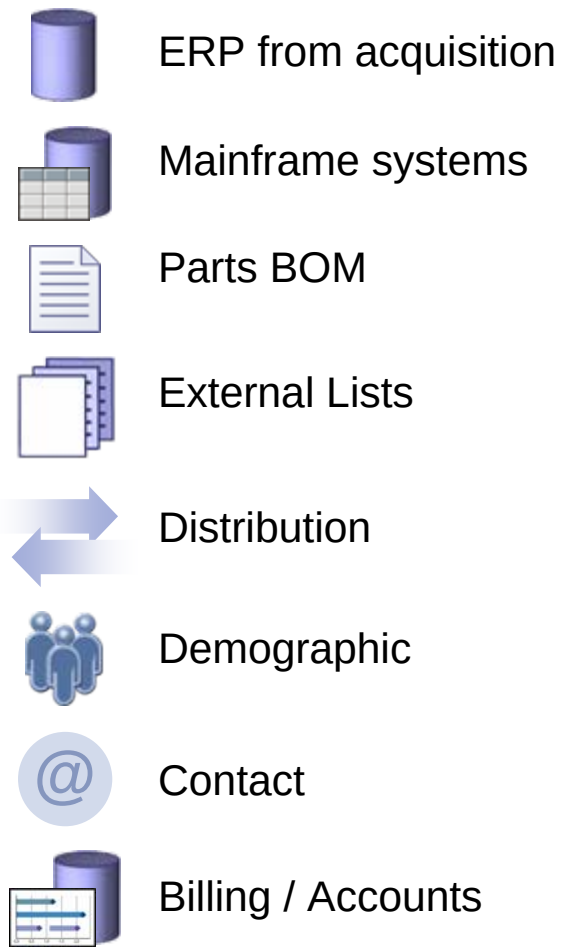
WebSphere Data Integration Suite: Understanding, Cleansing and Transforming

Bari, 20.09.2006

The IBM WebSphere Information Integration Platform



Data Profiling



Critical Problems

- **You don't know what data is really in your legacy systems**
- **Sources are new and unknown, or have changed**

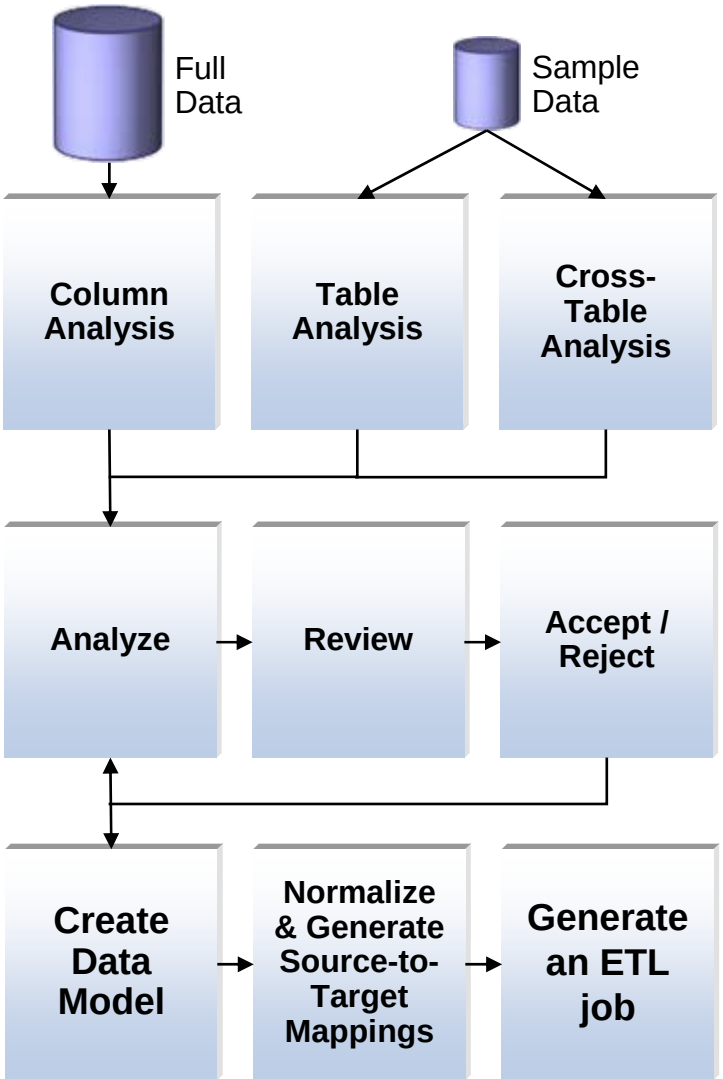
Why?

- **Data values and relationships are inconsistent and divergent from documented rules**
- **Documentation, if it exists, is incomplete**
- **Data sources are never static and frequently change without warning**

Typical Strategy

- **Labor intensive, resource devouring process**
- **Unable to review 100% of data elements**
- **Lacks maintenance infrastructure**
- **Lacks standardized approach across projects**
- **Narrow & shallow vs broad & deep**
- **First generation tools document but do not address the problem resolution**

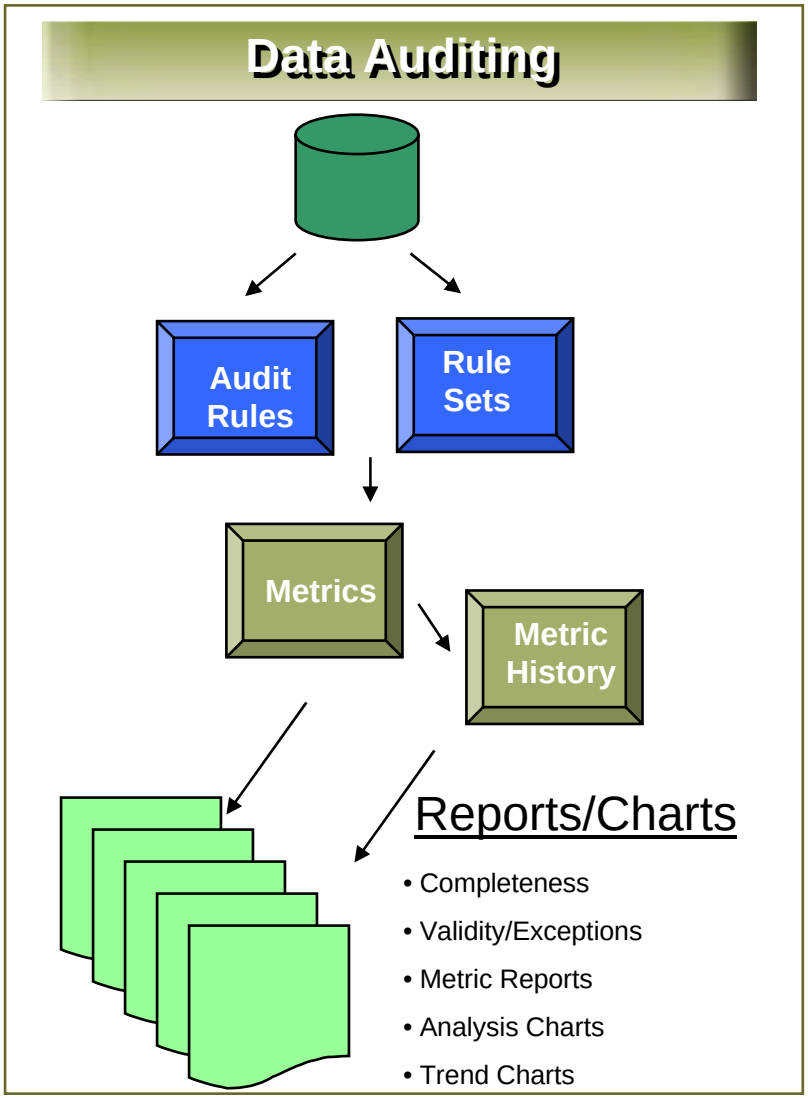
How WebSphere ProfileStage Works



Key Functionality

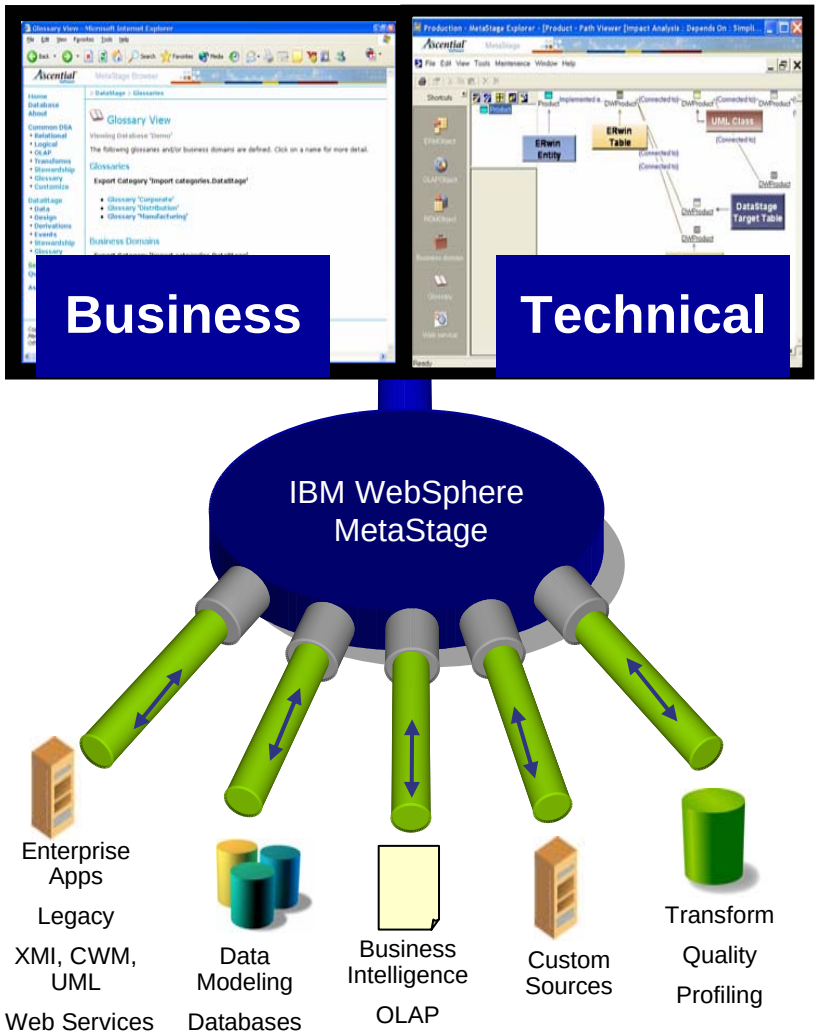
- **Enables you to understand your data before starting development**
- **Column analysis**
  - Generates frequency distributions for all values in all columns
  - Creates sample data for each table
- **Table analysis**
  - Within and across tables
  - Identifies primary and foreign key candidates
  - Correlation between columns within a table
- **Relationship and dependency analysis**
- **Duplicate analysis to identify and eliminate duplicate columns within and across tables**
- **Generates reports containing all the acquired information about your systems**
- **Generates normalized target database definition**
- **Creates a specification reflecting source to target mapping information**
- **Generates DDL and ETL job definition and metadata based on the specification**
- **Enables sharing of this information with modeling tools like ERWin**

WebSphere AuditStage



- Audit data over time
- Implement complex business rules in the audit process
- Demonstrable data consistency
- Comprehensive Data Quality Methodology
- Integrate Validation/Exceptions with ETL processes
- Increase ongoing data confidence
- Auditability – Validation or regulation requirements
- Monitor effectiveness of ongoing, implemented Data Quality strategies

MetaData Management



Critical Problems

- **Manage the definitions and relationships that are critical to the success of all data integration projects**
- **Establish common data definitions across business and IT in order to:**
  - Drive consistency throughout the data integration lifecycle
  - Provide regulatory data audit trail without coding
  - Deliver business and IT-oriented reporting
  - Enable the business to take ownership of their data
  - Provide enterprise visibility for change management
  - Easily extend to new, legacy and homegrown meta data sources

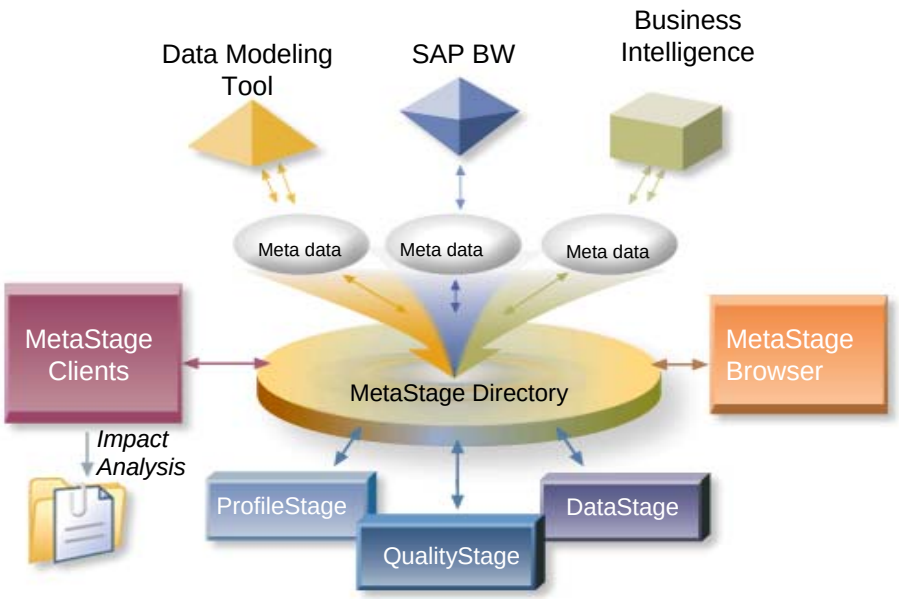
Why?

- Inability to respond and be flexible to changing data requirements
- Increased project costs due to lack of consistency and rampant redundancy
- Non compliance, stiff penalties and no data audit trail
- Lack of agreement and communication between IT and the Business
- Under utilized systems, higher support and training costs

Typical Strategy

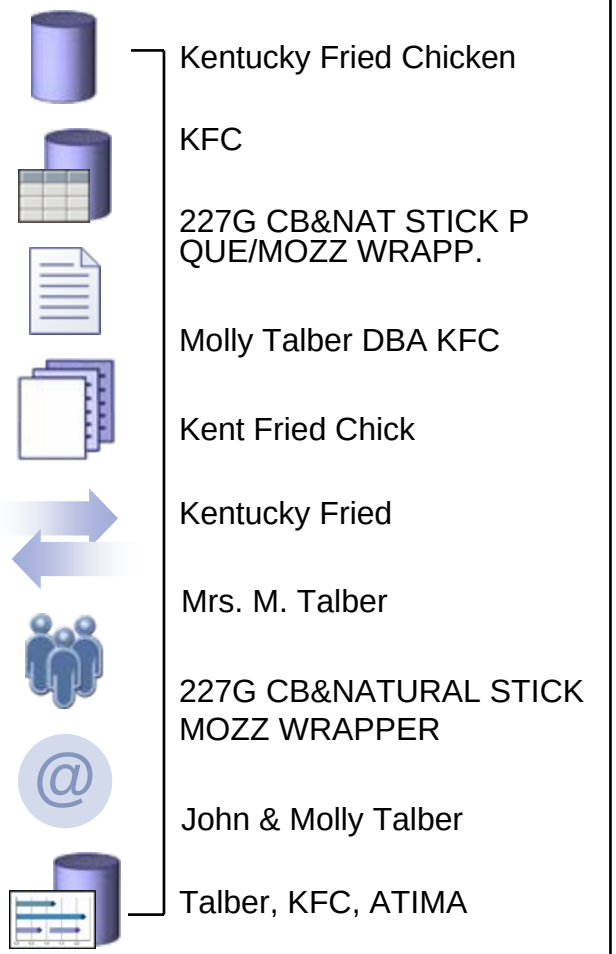
- Track meta data separately within each application
- Manual assessment of change and uncertainty
- Reliance upon least-common-denominator standards
- Asking technical users for business information
- Utilizing different custom applications: Business & IT

How WebSphere MetaStage Works



- **Meta data integration for all products used in the data integration lifecycle:**
  - Data Modeling/Case Tools
  - Business Intelligence applications
  - Databases and Data warehouses
  - Enterprise Applications
  - Enterprise Information Integration portfolio
- **Delivers cross-tool impact analysis and data lineage reporting and documentation**
- **Manage business glossaries, vocabularies and terms**
- **Assigns and maintains data stewardship**
- **Extensibility enabled through MetaArchitect**
- **Bi-directional meta data sharing and reuse via MetaBrokers®**
- **Imports business definitions from Analysts and Compliance Officers using MetaArchitect**
- **Receives DataStage/QualityStage design and execution information**
- **Shares meta data using OMG XMI, CWM and UML standards**

Data Quality



Critical Problems

- **Need to create & maintain 360 degree views of customers, suppliers, products, locations, events, etc.**
- **Need to leverage data: make reliable decisions, comply with regulations, meet service requirements, etc.**

Why?

- **No common standards across organization**
- **Unexpected values stored in fields**
- **Required information buried in free-form fields**
- **Fields evolve as they are used for multiple purposes**
- **No reliable keys for consolidated views**
- **Operational data degrades 2% per month**

Typical Strategies

- **Denial**
  - Problem misunderstood and ignored until too late (load & explode)
- **Hand-coding**
  - Clerical exception processing (time consuming/resource intensive)
- **Simplistic cleansing applications**
  - Evolved from direct marketing and list hygiene (lacks flexibility)

Why Should I Care About Cleansing Information?

- Lack of information standards
  - Different formats & structures across different systems
- Data surprises in individual fields
  - Data misplaced in the database
- Information buried in free-form fields
- Data myopia
  - Lack of consistent identifiers inhibit a single view
- The redundancy nightmare
  - Duplicate records with a lack of standards

```
Kate A. Roberts   416 Columbus Ave #2, Boston, Mass 02116

Catherine Roberts Four sixteen Columbus APT2, Boston, MA 02116

Mrs. K. Roberts   416 Columbus Suite #2, Suffolk County 02116
```

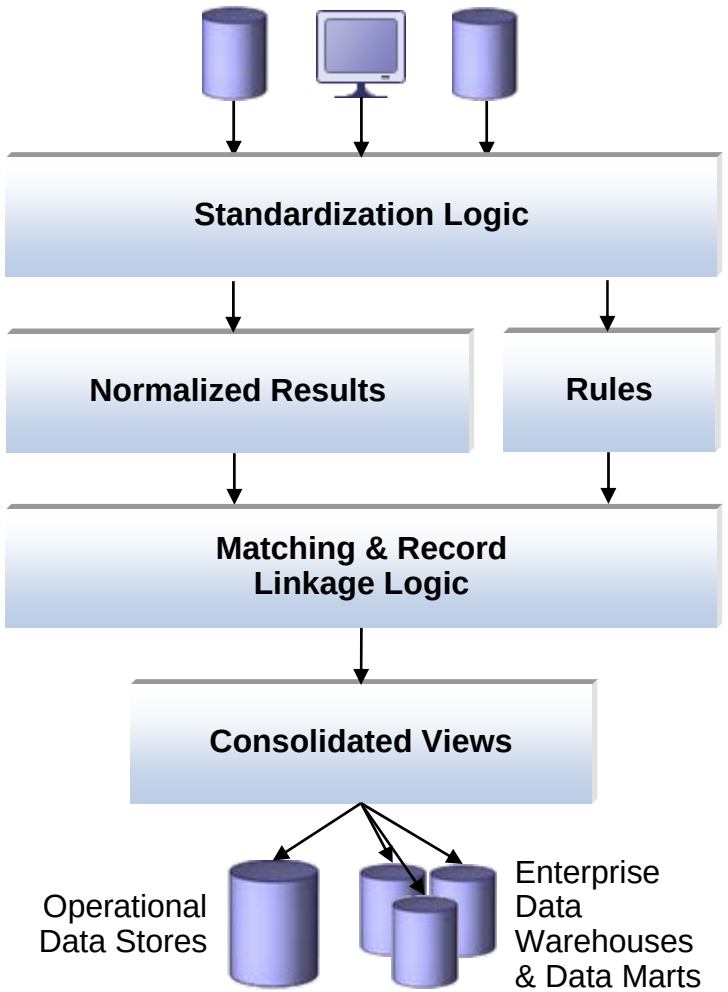
Name	Tax ID	Telephone
J Smith DBA Lime Cons.	228-02-1975	6173380300
Williams & Co. C/O Bill	025-37-1888	415-392-2000
1st Natl Provident	34-2671434	3380321
HP 15 State St.	508-466-1200	Orlando

```
WING ASSY DRILL 4 HOLE USE 5J868A HEXBOLT 1/4 INCH

WING ASSEMBY, USE 5J868-A HEX BOLT .25” - DRILL FOUR HOLES

USE 4 5J868A BOLTS (HEX .25) - DRILL HOLES FOR EA ON WING ASSEM

RUDER, TAP 6 WHOLES, SECURE W/KL2301 RIVETS (10 CM)
```

```
19-84-103    RS232 Cable 6' M-F CandS

CS-89641     6 ft. Cable Male-F, RS232 #87951

C&SUCH6      Male/Female 25 PIN 6 Foot Cable
```

```
90328574    IBM                  187 N.Pk. Str. Salem NH 01456
90328575    I.B.M. Inc.          187 N.Pk. St. Salem NH 01456
90238495    Int. Bus. Machines   187 No. Park St Salem NH 04156
90233479    International Bus. M. 187  Park Ave Salem NH 04156
90233489    Inter-Nation Consults 15 Main Street Andover MA 02341
90345672    I.B. Manufacturing   Park Blvd. Bostno MA  04106
```

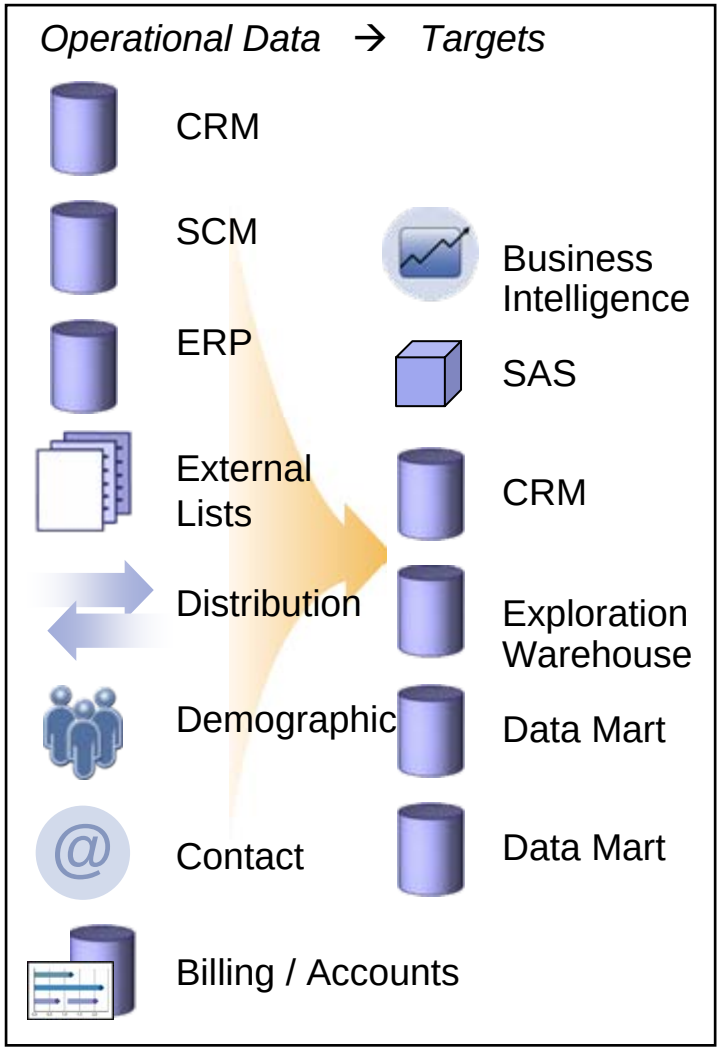
How WebSphere QualityStage Works



Key Functionality

- **Resolves format and context inconsistencies between source systems**
- **Investigates structure & content of free-form fields from any number of sources**
- **Uses a probabilistic matching engine with customized business rules for managing duplicates and creating “best-of-breed” systems of record**
- **Supports flexible survivorship rules to generates the most accurate and consistent set of data possible**
- **Same rules, same design executable in batch or real-time on mainframe, Windows, Unix, or Linux**
- **Performs parallel dataflow pipelining with in-flight data repartitioning for infinite scalability**
- **Multi-byte support for global deployment**
- **WebSphere QualityStage functions are callable from Web services or SOA applications**

Data Transformation



Critical Problems:

- Requirements are always evolving
- Custom coding is time consuming, doesn't scale, expensive to maintain
- Need to raise productivity by automating and streamlining process
- Regulatory compliance demands timely and accountable data integration

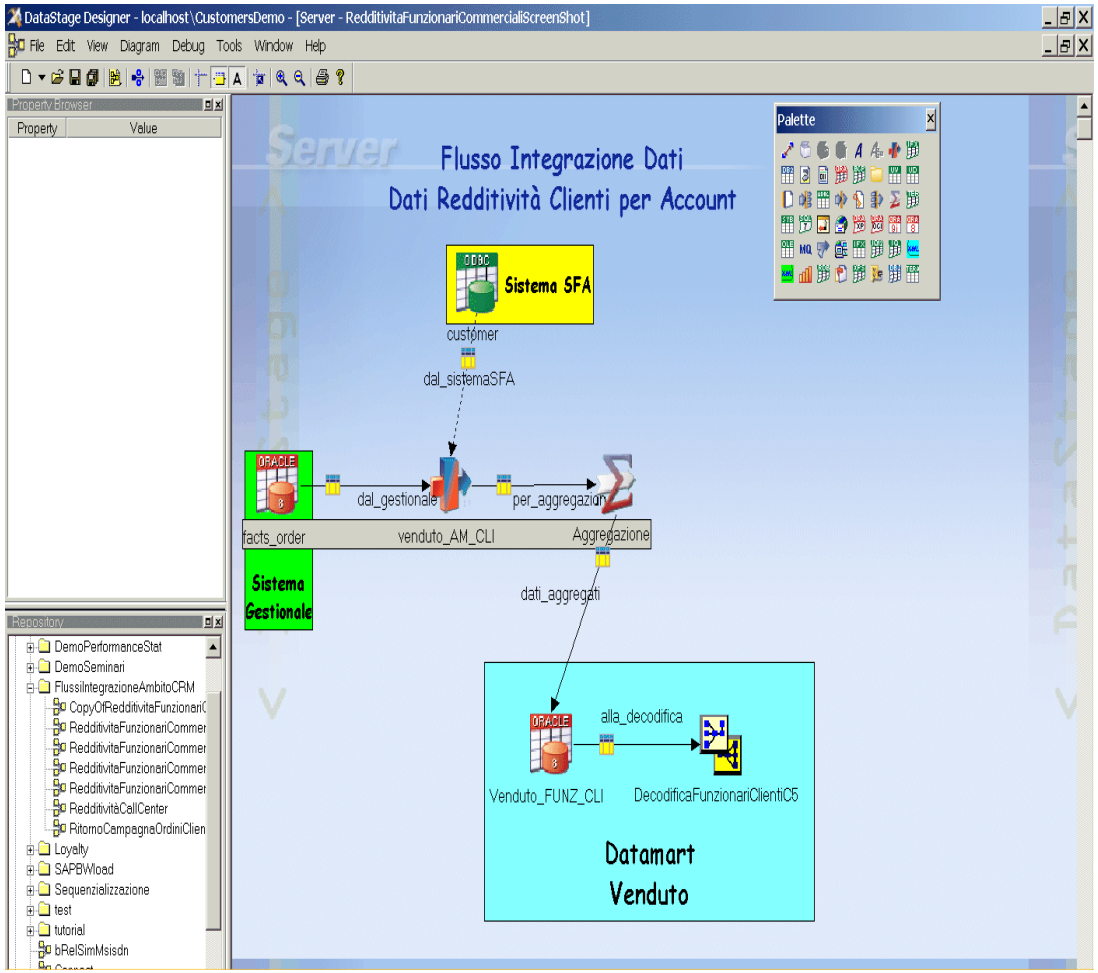
Why?

- Requirements are always evolving
- Custom coding is time consuming, doesn't scale, can rarely be reused, and is expensive to maintain (time is money)

Typical Strategies:

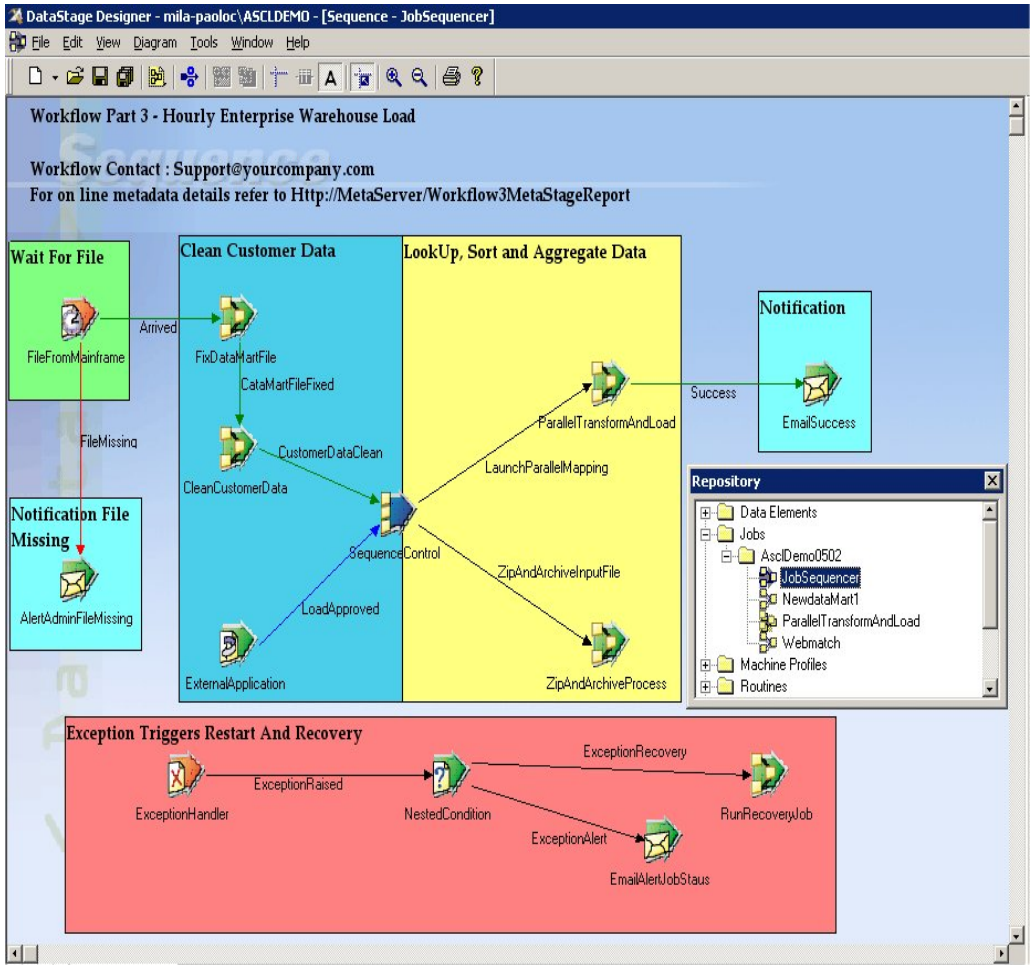
- Use a manual, labor intensive, resource devouring process
- Invest time and money integrating limited point solutions that don't scale
- Re-create the same transformation logic and Metadata across disparate tools

WebSphere DataStage: Graphical design metaphor



- Handles all transformations from simple to complex
- Complete development environment
  - One methodology, one skill set, one vendor
- Extensible, component based architecture
- Extensive re-use
- Built-In scripting language
- Built-In Debugger
- Rich support for application deployment
- Parameterization & version control
- Ubiquitous Connectivity
- External routine support

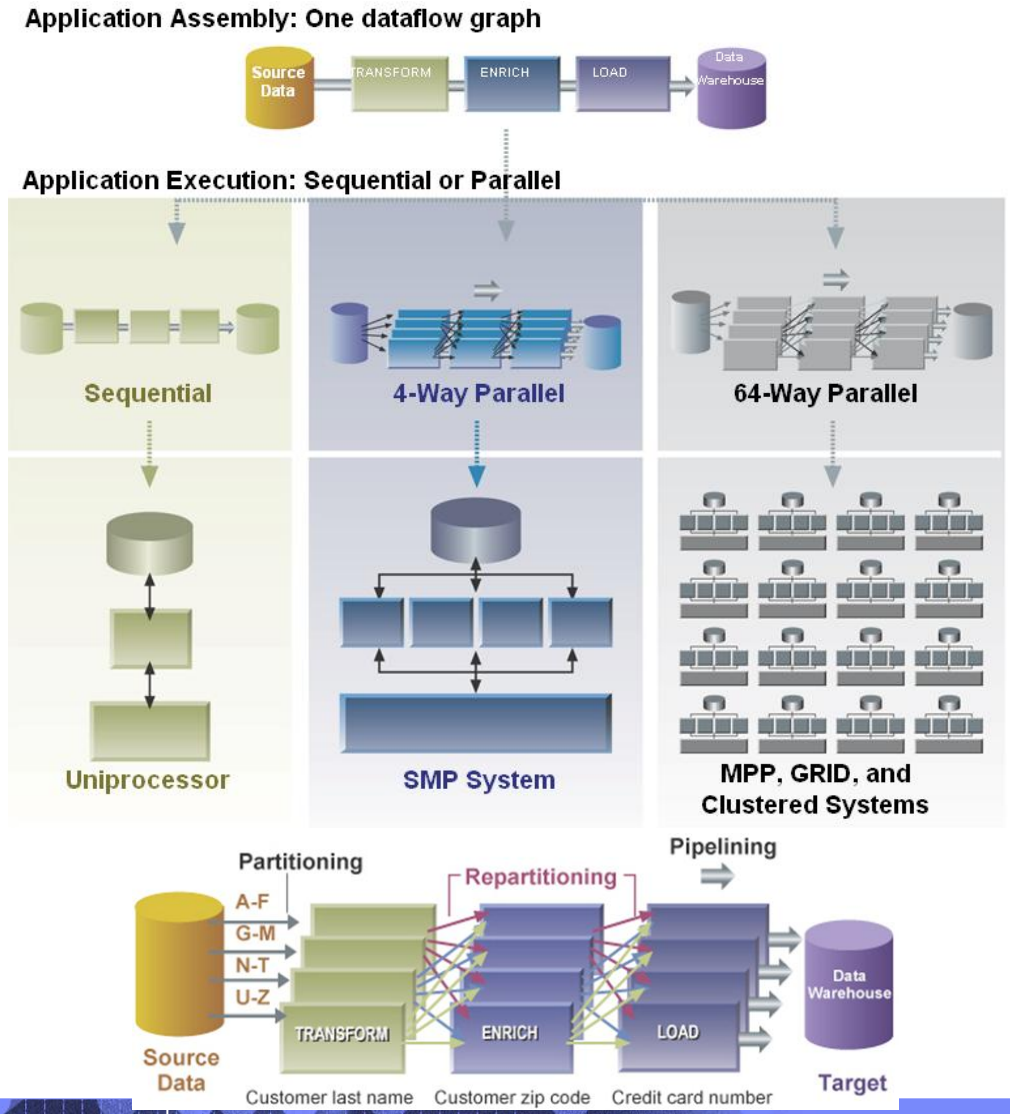
WebSphere DataStage: Graphical Workflow



- Launch job executions
- Manage job sequences and flows – i.e. job networks
- Manage global parameter passing
- Manage notifications and alerts (mail)
- Control job restart

NOT ONLY ETL for DWH

Performance and Scalability: Parallel Processing



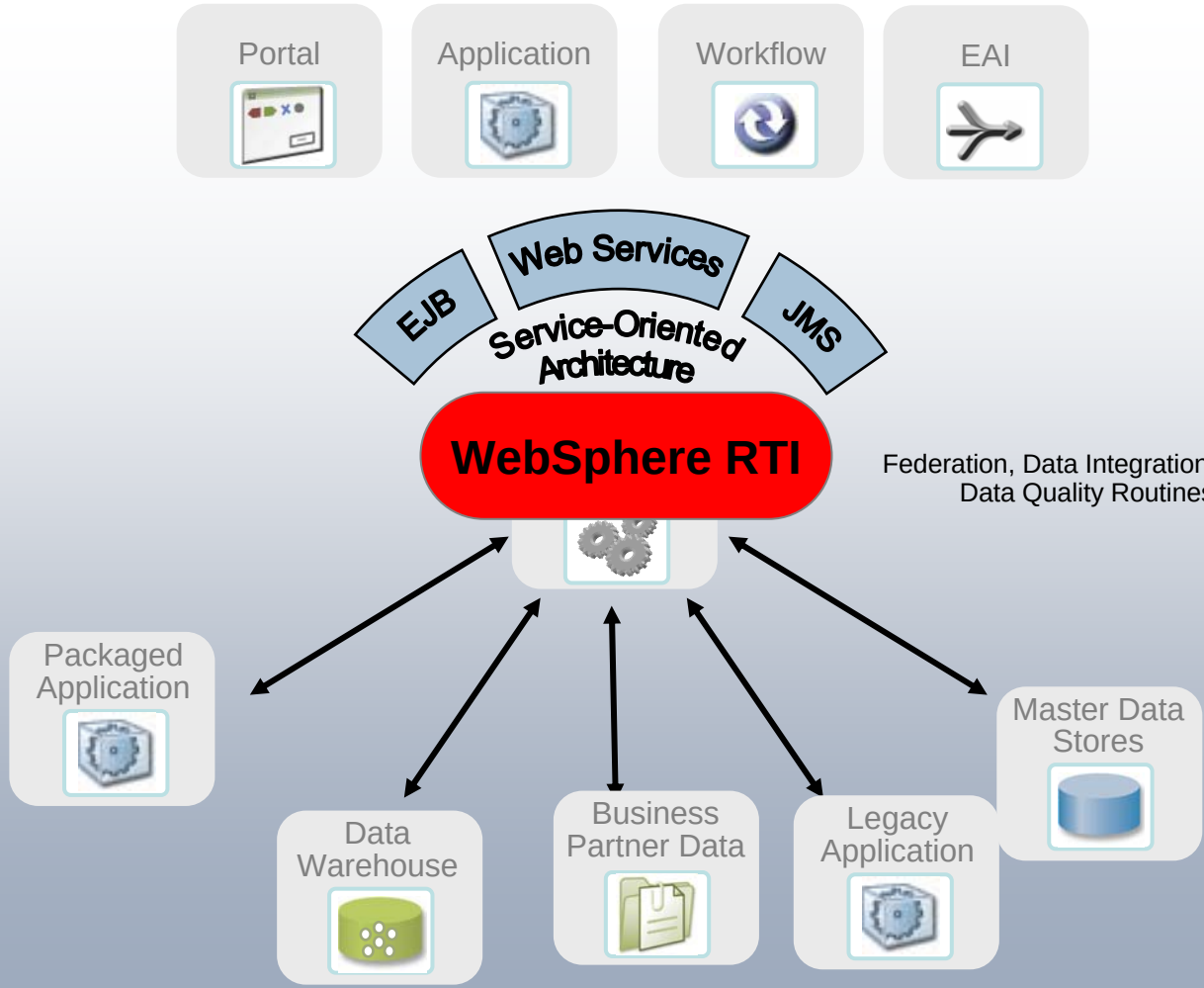
Why Enterprise Editions?

- Design sequentially, deploy in parallel
- Proven linear scalability
- Dynamic data partitioning and in-flight repartitioning of data
- Portable across SMP, Clustered, GRID, and MPP platforms
- Parallel RDBMS support
- Codeless parallelization
- Incorporate and parallelize existing applications into data integration process

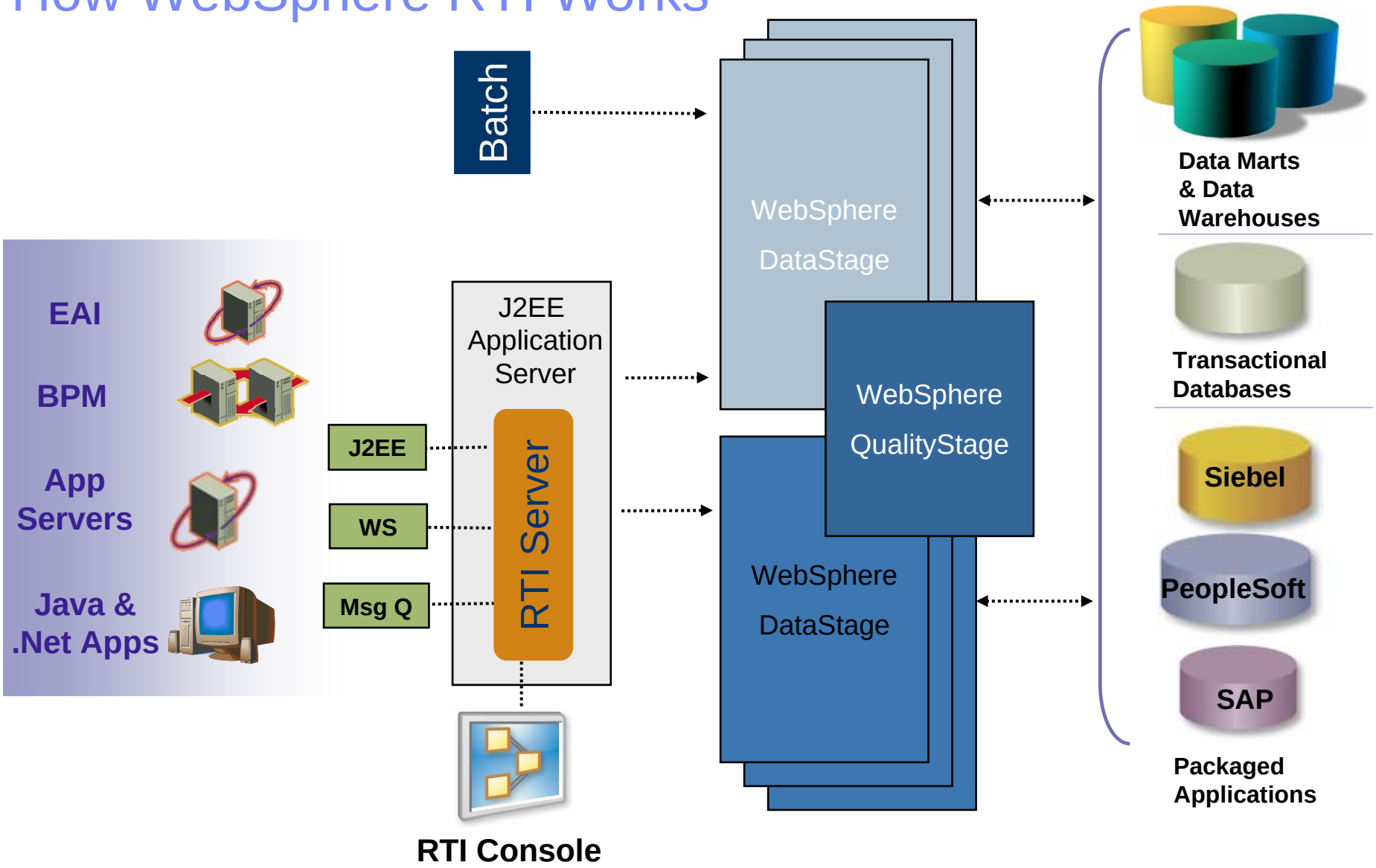
Business Benefits

- Meet business commitments through higher productivity
- Optimal hardware use
- Flexible execution options

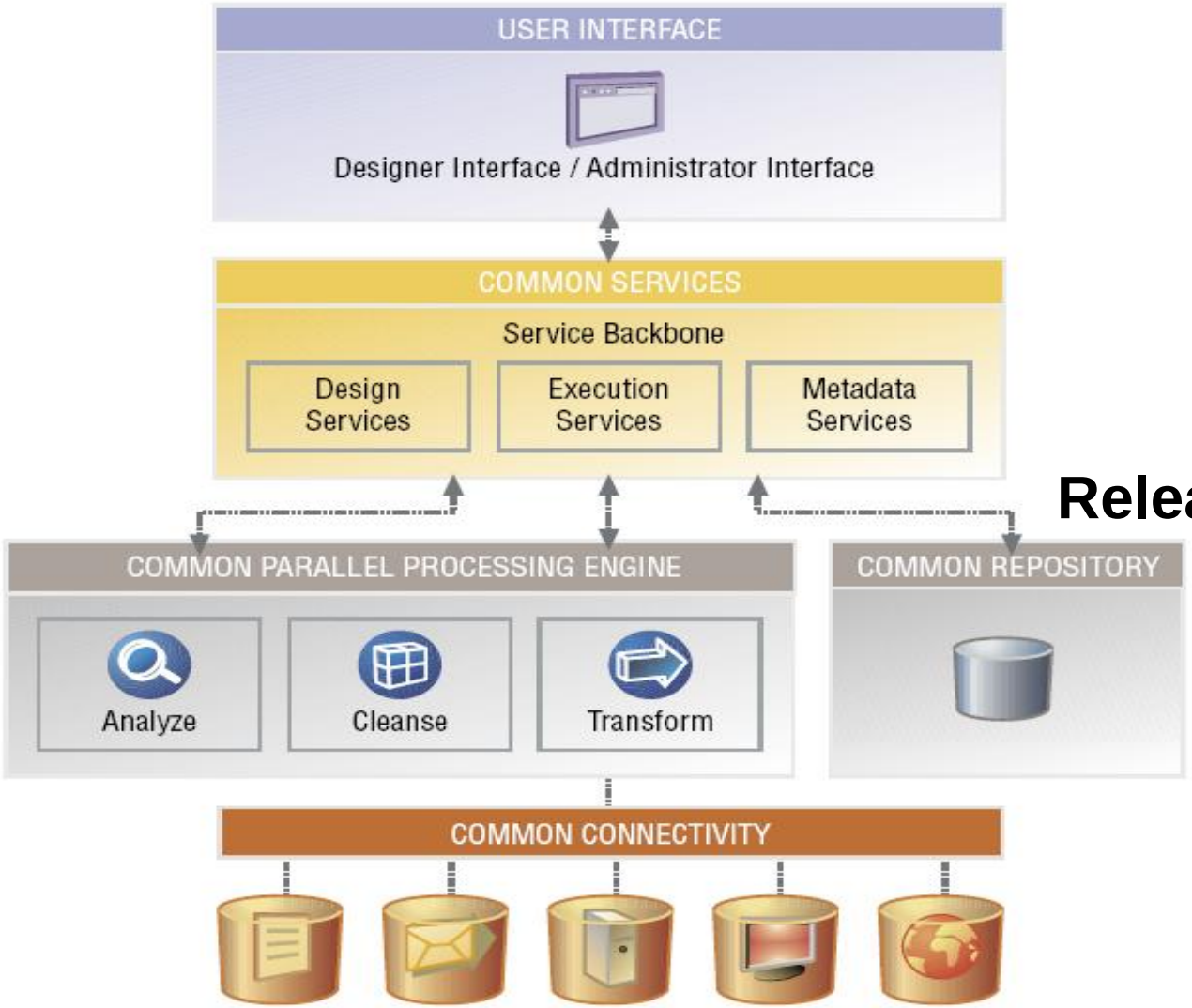
Information Integration and SOA: WebSphere RTI



How WebSphere RTI Works



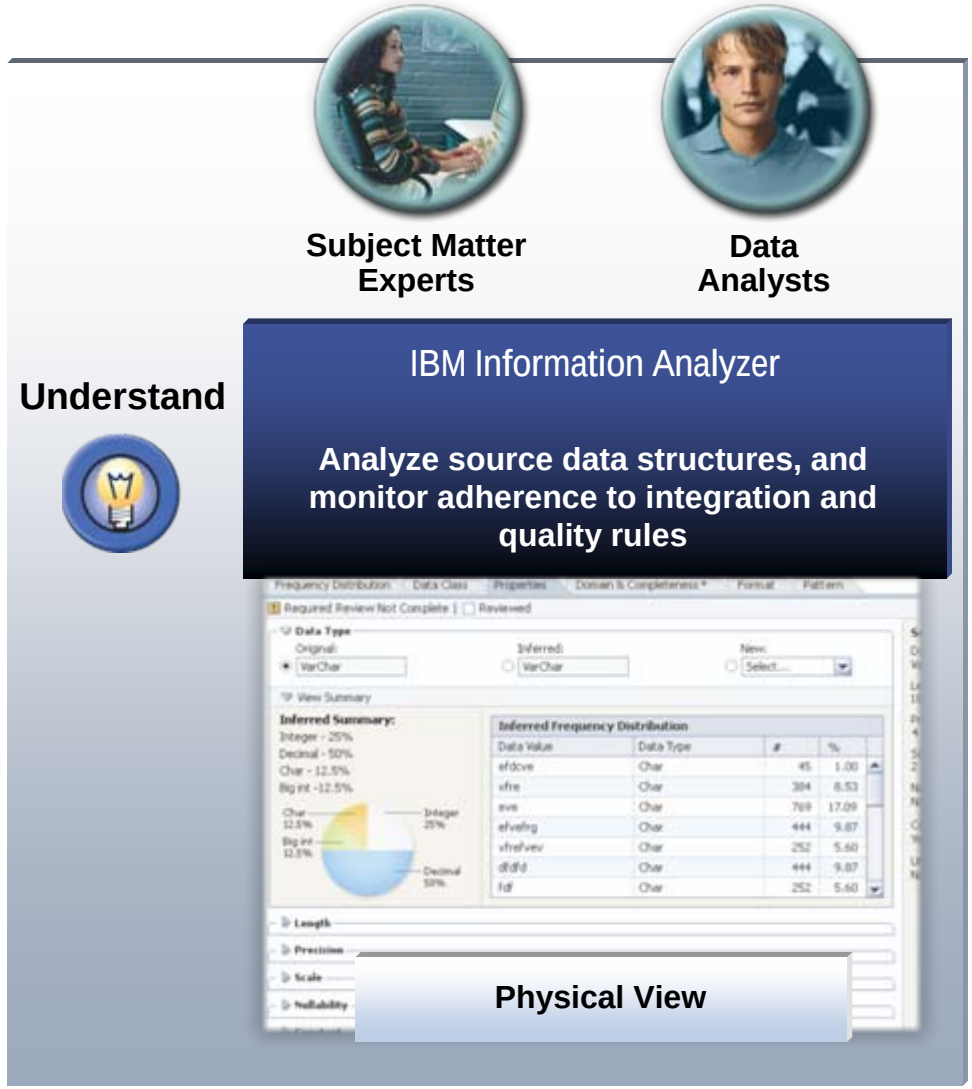
WebSphere Information Server Preview



Release 8.0

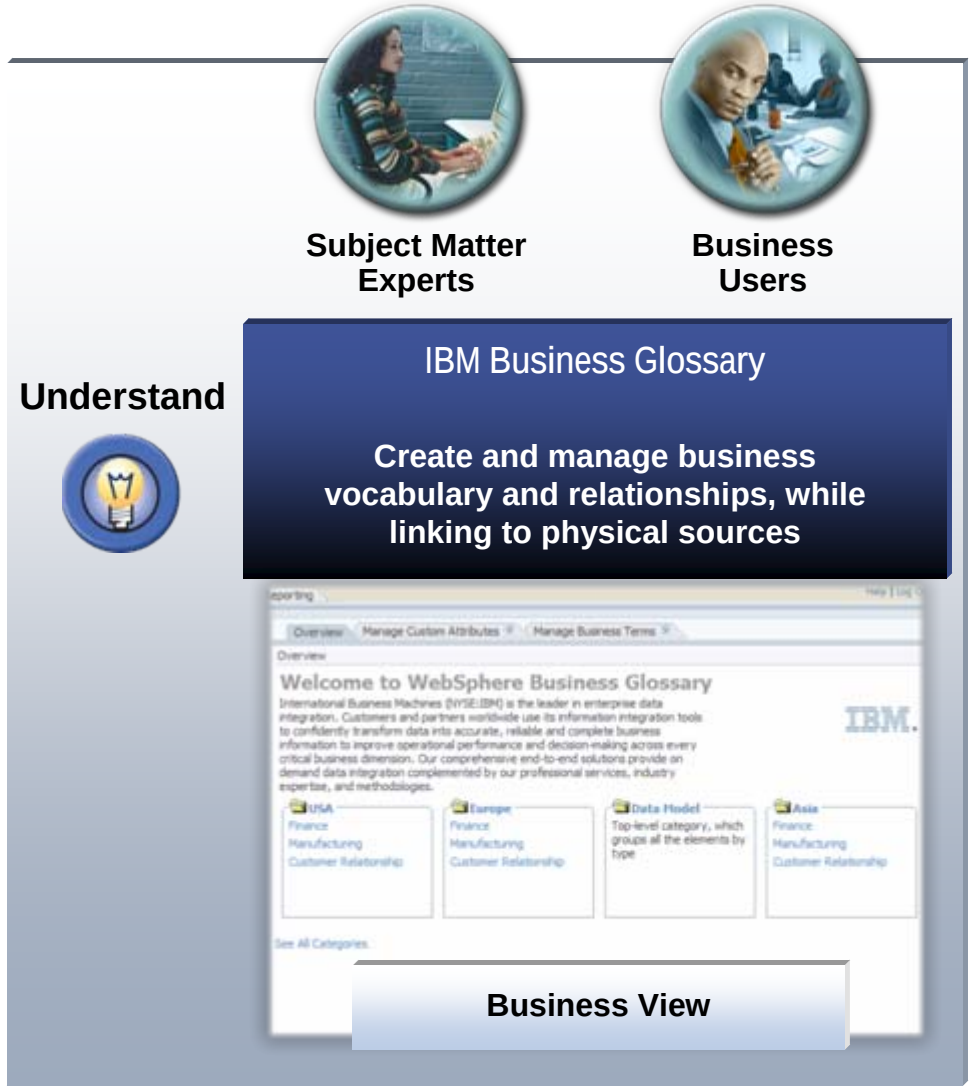
Physical Metadata: IBM Information Analyzer

- **Data-centric analysis of application, database and file-based sources**
- **Secure, detailed profiling of fields, across fields, and across sources**
- **Creation of metadata from profiling results**
- **Results instantly promotable across IBM Information Server**



Business Metadata: IBM Business Glossary

- **Web-based authoring, managing & sharing of business metadata**
- **Aligns the efforts of IT with the goals of the business**
- **Provides business context to information technology assets**
- **Establishes responsibility and accountability**



Database = DB2

Schema = NAACCT

Table = DLYTRANS

Column = ACCT_NO

data type = char(11)

Technical

Business

GL Account Number

The ten digit account number. Sometimes referred to as the account ID. This value is of the form L-FIIIIVVVV.

Logical Metadata: Rational Data Architect

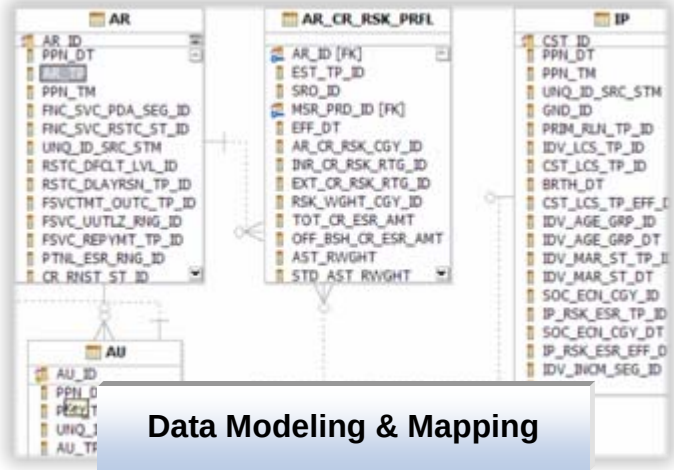
- **Data modeling for data structures and federations**
- **Federated data discovery**
- **Metadata relationship discovery & mapping**
- **Impact analysis, and synchronization across models**
- **SQL & XML generation capabilities**

Subject Matter Experts

Architects

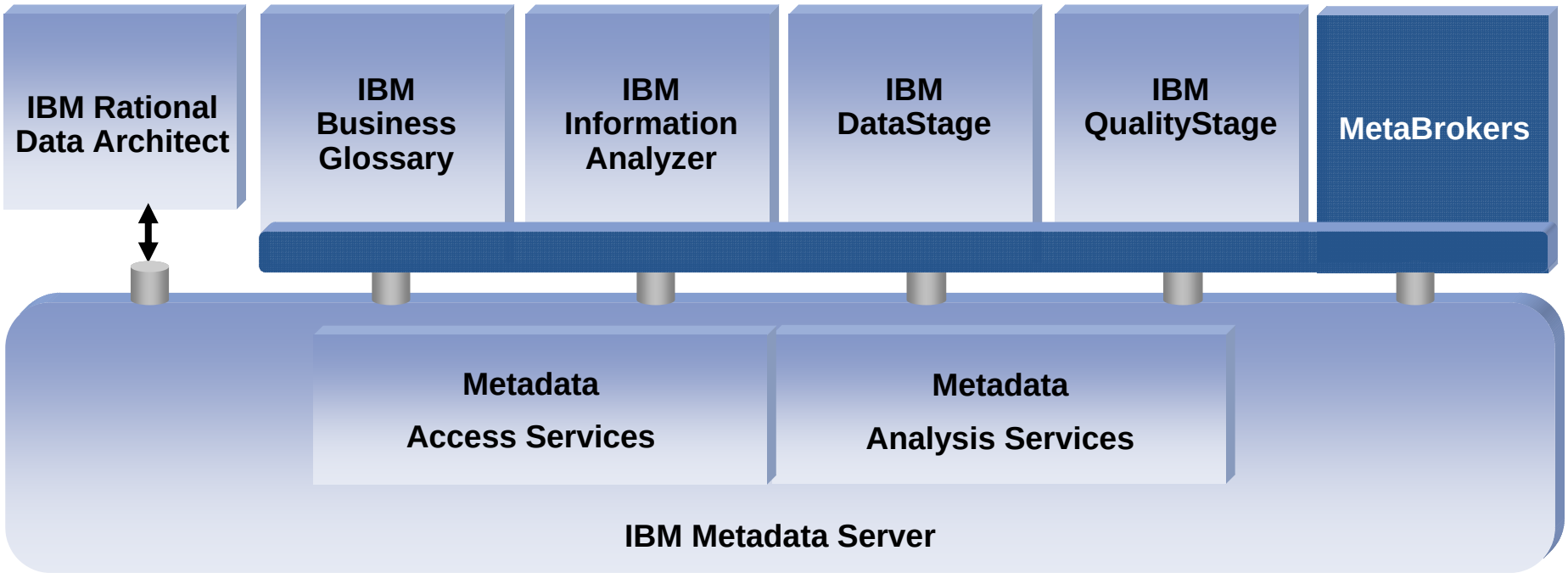
Rational Data Architect

Create and manage business vocabulary and relationships, while linking to physical sources

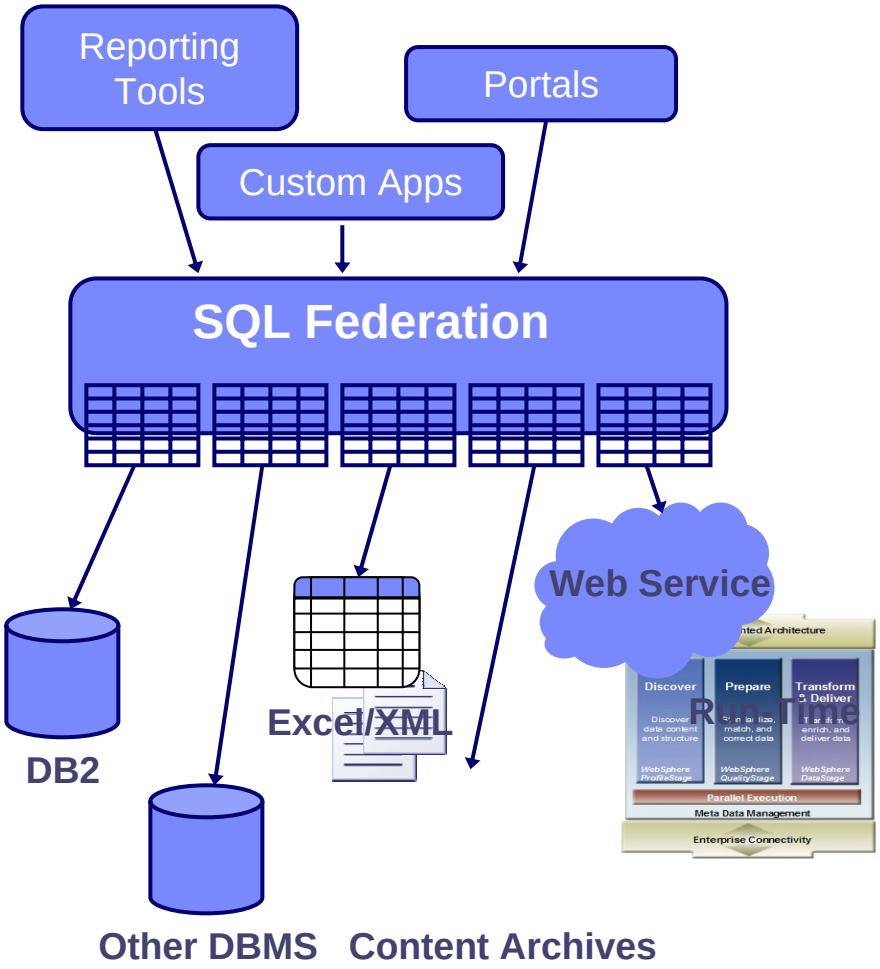


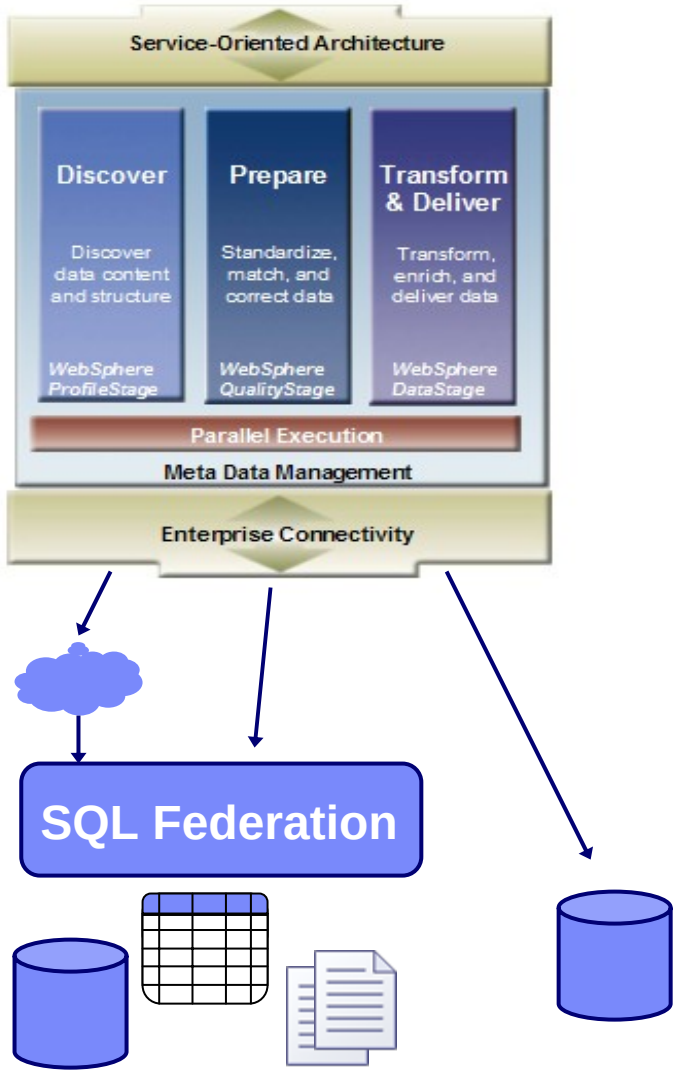
Data Modeling & Mapping

IBM Metadata Server – at the Core of IBM Information Server



Federation and Consolidation: can work together !!





Combine Event-Driven Processing & Transformation

Reduce latency for tactical decision making