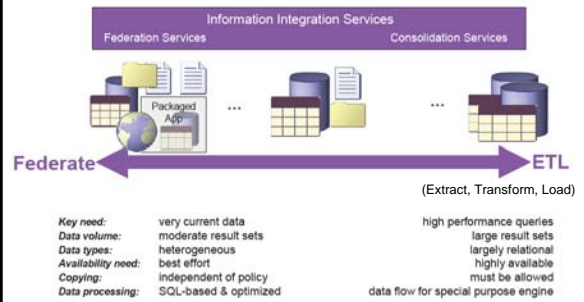


Industrial Information Integration Tools

Jose Luis Ambite

Based in part on slides by Mike Carey, Laura Haas, IBM, BEA

Information Integration Approaches : Virtual (Federated) vs Materialized (ETL)



Industrial Integration Tools: Design Choices

- Inside Out: Extend a DBMS engine to work with data not managed by the DBMS itself
 - Examples:
 - SQL user defined functions (scalar & table)
 - Informix DataBlades, *DB2 Federation Server*, Sybase Federation Server
- Outside In: Build middleware to combine data from multiple sources
 - Examples:
 - Application-level programming
 - ODBC/JDBC drivers
 - ETL/Warehousing systems: Informatica, Microsoft Data Warehouse
 - SOA-style: *BEA AquaLogic Data Services*, IBM Information Server
 - 'Outside the DBMS' federation engines: Composite, Oracle Sunopsis

[based on slide by Mary Roth]

Data Integration Tools

- Research in data integration has transitioned to industry:
 - from industrial research labs, universities
 - through acquisitions of (university-originated) start-ups
- Two significant examples:
 - WebSphere Federation Server** (part of **IBM Information server**)
 - GAV* Relational integration core
 - Research: Garlic project, 1995-1999, IBM Almaden
 - Laura Haas et al. (& Mike Carey) at IBM
 - Related research: TSIMMIS (Stanford)
 - <http://www-306.ibm.com/software/data/integration/>
 - BEA AquaLogic Data Services Platform**
 - GAV* XML/Web Services integration core
 - Research: Mike Carey et al. at BEA
 - Related research: Yannis Papakonstantinou, Vasilis Vassalos, Stanford graduates from the TSIMMIS group, found Enosys Software, acquired by BEA in 2003 (BEA acquired by Oracle in Jan 2008)
 - <http://www.bea.com/dataservices>

4

Overview of IBM Information Server

ON DEMAND BUSINESS

© IBM Corporation

Beauty Meets the Beast: the Commercial World of Integration

When	Research Highlights	Research Projects	Products
Late '70's to early '80's	Distributed compilation, P2P, data conversion	R*, Multibase, EXPRESS, ...	Oracle, SQL/DS, DB2
80's	Storing, indexing, querying non-relational data with relational	Starburst, Postgres, ...	Informix, Oracle, ...
90's	Query & optimization over heterogeneous sources, wrappers; data integration	Garlic, Tsimmis, DISCO, Information Manifold, ...	UniSQL, Illustra, ..., Oracle dblink, DB2 DataJoiner, ...
'98 - '05	Schema mapping & mgmt, text analytics, data exchange, ontology	Clio, UIMA, ...	DiscoveryLink, Info Integrator, Callixa, Composite, Ascential, Informatica
Today	Integrating structured & unstructured, discovering & exploiting metadata	MAUI, CIM, ...	Information Server

Integration is a process

- Understand
 - What data is available
 - Important properties and values
 - Meaning or intent
 - Standardize/Cleanse
 - Schema, field level, terminology and abbreviations
 - How to identify information about the same object
 - How to handle missing or inconsistent values
 - Specify/Transform
 - Choose integration engine(s) and style
 - Execute
- Iterate!!

Marketing: IBM Information Server
Delivering information you can trust

IBM Information Server

Unified Deployment

Understand Discover, model, and govern information structure and content	Standardize Standardize, merge, and correct information	Specify Describe how to combine and restructure information for new uses	Execute Synchronize, virtualize and move information for in-line delivery
--	---	--	---

Unified Metadata Management

Parallel Processing
Rich Connectivity to Applications, Data, and Content

IBM is the acknowledged industry leader for vision and execution in information integration

IBM

Reality: Lots of products, lots of choices

IBM Information Server

Unified Deployment
WebSphere Information Services Director

Understand WebSphere Information Analyzer WebSphere Business Glossary Business Architect	Standardize WebSphere QualityStage Business Glossary	Specify and Data Architect WebSphere Federation Server WebSphere DataStage WebSphere Replication Server WebSphere Data Event Publisher	Execute
--	---	--	----------------

Unified Metadata Management
WebSphere Metadata Server

Parallel Processing
Rich Connectivity to Applications, Data, and Content

+ additional products for content federation, search, and special architectures or sources

IBM is the acknowledged industry leader for vision and execution in information integration

- We can provide an end-to-end solution
- Otherwise, the problem is much worse – many vendors, mismatched products

IBM

IBM Software Group

IBM Information Server
Delivering information you can trust

Support for Service-Oriented Architectures

Understand Discover, model, and govern information structure and content	Cleanse Standardize, merge, and correct information	Transform Combine and restructure information for new uses	Deliver Synchronize, virtualize and move information for in-line delivery
--	---	--	---

Platform Services

Parallel Processing	Connectivity	Metadata	Administration	Deployment
---------------------	--------------	----------	----------------	------------

IBM

IBM Software Group

The IBM Solution: IBM Information Server
Delivering information you can trust

IBM Information Server

Understand

Information Analyzer Business Glossary
Data profiling for understanding what data you have and how it relates to other data, plus data analysis for measuring and monitoring ongoing data quality. Management of Physical and Business Metadata

Parallel Processing
Rich Connectivity to Applications, Data, and Content

IBM

IBM Software Group

Information Analyzer

- **Source System Analysis**
 - Provides the key understanding of the source data
 - Column & Domain analysis
 - Table/Primary Key analysis
 - Foreign Key analysis
 - Cross-Domain analysis
- **Iterative Analysis**
 - Leverages the analysis to facilitate iterative tests
 - Baseline analysis

Primary Key Analysis

Column Analysis

Source 1 Source 2

Foreign Key & Cross-Domain Analysis

IBM

IBM Software Group

Information Analyzer: Column Analysis, Table

13

IBM Software Group

Information Analyzer: Column Analysis, Chart

- Frequency Distribution
 - View Frequency Distribution either in Tabular or in Graph
 - Add user defined values to Frequency Distribution
 - Generate Reference Tables
 - Sort and Filter Frequency Data

14

IBM Software Group

Information Analyzer: Column Analysis, Properties

- Properties
 - Six property values are inferred for each column: Data Type, Length, Precision, Scale, Nullability and Cardinality Type.
 - Distribution of data types, lengths, precisions and scales is displayed graphically

15

IBM Software Group

Information Analyzer: Primary Key Analysis

- Reviewing Duplicates
 - View Summary of Distinct and Duplicated Values
 - Display list of all Primary Key values and %/ Duplicated.

16

IBM Software Group

Information Analyzer: Cross Domain Analysis

17

IBM Software Group

IBM Business Glossary

Web-based tool for authoring, managing, and sharing an enterprise vocabulary & classification system

Database = DB2
Scheme = NAACCT
Table = DLYTRANS
Column = TAXVL
data type = Decimal (14,2)
Derivation: SUM(TRNXTXAMT)

Term: Tax Expense
Full Name: Tax to be paid on Gross Income
"The expense due to taxes"
(John Walsh is responsible for updates. 90% reliable source)
Status: CURRENT

Achieve a common vocabulary between business & technical users!

18

IBM Software Group IBM

The IBM Solution: IBM Information Server Delivering information you can trust

WebSphere QualityStage
Data cleansing, standardization, matching, and survivorship for enhancing data quality and creating coherent business views

19

IBM Software Group IBM

WebSphere QualityStage

Data Cleansing

- Standardize, correct source data fields
 - Ex: global postal verification
- Match records across sources
 - Blocking rules
 - Probabilistic Matching
- Consolidate (de-duplicate) information
- Visual workflow specification (integrated with DataStage)

20

IBM Software Group IBM

Standardization - Example

Input File:

Address Line 1	Address Line 2
639 N MILLS AVENUE	ORLANDO, FLA 32803
306 W MAIN STR, CUMMING, GA 30130	
3142 WEST CENTRAL AV	TOLEDO OH 43606
843 HEARD AVE	AUGUSTA-GA-30904
1139 GREENE ST ACCT #1234	AUGUSTA GEORGIA 30901
4275 OWENS ROAD SUITE 536 EVANS	GA 30809

Result File:

House #	Dir	Str. Name	Type	Unit	No.	NYSIIS	City	SOUNDEX	State	Zip	ACCT#
639	N	MILLS	AVE			MAL	ORLANDO	O645	FL	32803	
306	W	MAIN	ST			MAN	CUMMING	C552	GA	30130	
3142	W	CENTRAL	AVE			CANTRAL	TOLEDO	T430	OH	43606	
843		HEARD	AVE			HAD	AUGUSTA	A223	GA	30904	
1139		GREENE	ST			GRAN	AUGUSTA	A223	GA	30901	1234
4275		OWENS	RD	STE	536	ON	EVANS	E152	GA	30809	

21

IBM Software Group IBM

Probabilistic Record Linkage

Are these two records a match?

WILLIAM J KAZANGIAN 128 MAIN ST 02111 12/8/62

WILLIAM JOHN KAZANGIAN 128 MAINE AVE 02110 12/8/62

- Fields are evaluated for degree-of-match
 - Take into account frequency of value
- Weights are summed to derived a total score
- ~Statistical probability of a match

22

IBM Software Group IBM

Matching Rules

Comparison	Description
ABS_DIFF%	Absolute difference comparison
ASC_SORT	Double alphanumeric left/right intervals
ASC_INTERVAL	Alphanumeric odd/even intervals
CHAR%	Character comparison
CNT_DIFF%	Counting errors in columns
D_INT	Left/right interval comparison
D_LIST%	Left/right United States Postal Service interval
D_LIST%	Date comparison in the form of YYYYMMDD
DELTA_PERCENT%	Delta percentage comparison
DISTANCE%	Geometric distance comparison
INT_TO_INT%	Interval-to-Interval comparison
INTERVAL_COMPAR	Interval comparison
INTERVAL_PARTY	Interval comparison with parity
LR_CHAR	Left/right character string comparison
LR_UNCERT	Left/right uncertainty character comparison
MULT_EXACT%	Multi-word exact
MULT_RANGE	Multi-address range
MULT_UNCERT%	Multi-word uncertainty
NAME_UNCERT%	First name uncertainty string comparison
NUMERIC%	Numeric comparison
PREFIX%	Prefix comparison for truncated columns
PROBATEP%	Probated comparison
TIME%	Time comparison
UNCERT%	Uncertainty character comparison
USPS	United States Postal Service ranges
USPS_DIST	United States Postal Service double interval comparison
USPS_INT%	United States Postal Service interval comparison

Left-right string equal

Max date difference in days

Building 5 Apartment 4B - Apartment 4-B Building 5

AL - ALBERT
W - WILLIAM

23

IBM Software Group IBM

Survivorship - Example

Survivorship Input (Match Output)

Group	Legacy	First	Middle	Last	No.	Dir.	Str. Name	Type	Unit	No.
1	D150	Bob		Dixon	1500	SE	ROSS CLARK	CHR		
1	A1367	Robert		Dickson	1500			CHR		
23	D689	Ernest	A	O'Brian	5901	SW	74TH	ST	STE	202
23	A436	Ernie	Alex	O'Brian	5901	SW	74TH	ST		
23	D352	Ernie		O'Brian	5901		74	ST	#	202

Consolidated Output

Group	Legacy	Group	First	Middle	Last	No.	Dir.	Str. Name	Type	Unit	No.
1	D150	1	Robert		Dickson	1500	SE	ROSS CLARK	CHR		
1	A1367										
23	D689	23	Ernie	Alex	O'Brian	5901	SW	74TH	ST	STE	202
23	A436										
23	D352										

24

IBM Software Group IBM

The IBM Solution: IBM Information Server

Delivering information you can trust

WebSphere DataStage
Complex transformation for simplified data exchange and reduced coding

Parallel Processing
Rich Connectivity to Applications

25

IBM Software Group IBM

WebSphere DataStage

- Complete ETL functionality
 - Deals with very large volumes of data
 - Parallel execution
 - Supports batch & real-time operations
- Visual design of data flows
 - hundreds of built-in transformation functions
- Supports connectivity to large range of sources

26

IBM Software Group IBM

Connectivity Ensures Data Access

Enterprise Applications JD Edwards Oracle Applications PeopleSoft SAP BW (BAPI, IDOC) SAP R/3 (ABAP, BAPI, IDOC) Siebel	Business Exchange Formats XMLS EDI FIX SWIFT HIPAA	Flat File and General Access VSAM VSAM CICIS IDMS C-ISAM Sequential File Complex Flat File File Set Data Set Named Pipe FTP (standard, secure) Compressed / Encoded Data External Command Call Parallel Wrap 3 rd party applications
RDBMS IBM DB2 IBM MIS VSAM Oracle Informix Redbrick SQL Server Sybase Teradata U2 (Universe, UniData) Tandem NON-STOP SQL SAS	Real-Time WebSphere MQ SeeBeyond Java Messaging Services Java (Client & Transformer) XML (Read / Write) XSL-T XSL-T Transformer Web Services (SOAP) Enterprise Java Beans	...And many more!

27

IBM Software Group IBM

IBM Information Server *Delivering information you can trust*

Platform Services

- Parallel Processing
- Connectivity
- Metadata
- Administration
- Deployment

Federation Server
Dynamically integrate information from disparate data and content sources in a single query as if it were coming from one system

28

IBM Software Group IBM

WebSphere Federation Server

- Provides query access to heterogeneous and distributed sources as if they were one system
 - Define integrated views across diverse distributed data
 - Query optimization
 - Access to many source types
 - Both read & write capabilities

29

IBM Software Group IBM

WebSphere Federation Server Basic Concepts

Federated server: a DB2 database enabled for federation.

Wrapper: a library allowing access to a particular class of data sources or protocols.

Server: represents a specific data source

Nickname: a local alias to data on a remote server (mapped to rows and columns). Appears as a DB2 table

30

Federating Data Through Wrappers

- Supply DBMS with custom code to access data sources and function
 - 1 reusable wrapper code module per data source type
- Can appear where ever tables can appear in an SQL statement
- Can 'push down' as much function as possible
- Can provide optimizer with cost and statistical information
- Can leverage DBMS to apply more operations to the data

```
SELECT C.name, A.URL
FROM Compounds C, Experiments E, Articles A
WHERE E.result < 1.1e-9 and E.id = C.id
and search("subject", c.name) > 0
```

