Intranet Mediators

M. Saelee (team leader)
R. Ahuja, S. Beitzel, E. Jensen, A. Pilotto, D. Schmeltzer, D. Yap
D. Grossman (faculty advisor)

Information Retrieval Laboratory
Department of Computer Science
ENPRO 307
http://mediator.iit.edu

ILLINOIS INSTITUTE
OF TECHNOLOGY

Business Need

- Provide single, correlated query to multiple datatypes (e.g.; structured data, text, image, video, sound).
- Example queries
    - List articles in trade magazines whose byline is within 100 miles of headquarters and written by former employees.
    - Show all internal planning reports with sales projections and match these up with actual sales.
    - What are the three best restaurants in Chicago.

2

Agenda

- Overview
- Current Market
  - Structured
    - Data Warehouse
    - OLAP Tools
    - ERP (Enterprise Resource Planner)
  - Unstructured
    - EIP (Enterprise Information Portal)
    - Search Engines
  - Semi-structured Systems (XML Database Engines)
- Mediator
- Barriers to Entry
- Data Suppliers
- Customers
- Future Plans
  - Marketing
  - Deployment
- Demo

3

Business Plan

- Develop and sell products and services to use a next generation search to provide access to both structured, semi-structured, and unstructured data on an intranet.
- This is a viable plan as the "business intelligence" market includes data warehouse, portals, search engines etc.
- Competition does not give an answer to questions.

4

Overview

Query: **What are the three best restaurants in Chicago?**

- Answering this requires:
    - Structured data (list of restaurant ratings)
    - Unstructured data (printed reviews)
    - Semi-structured (Restaurant menu)
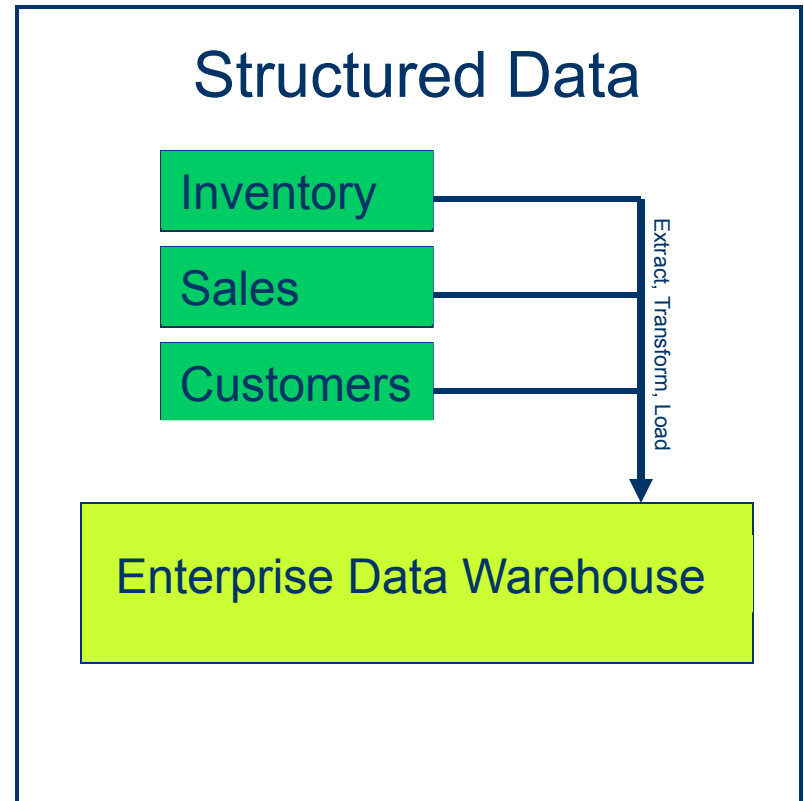
Advantages over the Competition

- Provides an "answer" to the query
- Lacks need for training
- Simple natural language query interface
- Integration of structured,semi-structured, and unstructured data repositories

Market

- Structured Systems
  - Data Warehouse
  - OLAP Tools
  - Enterprise Resource Planners
- Unstructured Systems
  - Information Retrieval (general search)
  - Question Answering
  - Enterprise Information Portals

Data Warehouse

- Structured applications feed the data warehouse periodically.

- ETL (Extract transform and load) tools are used to migrate data to the warehouse.

- All of this is *structured* information.

Structured Data

| Inventory |
| Sales |
| Customers |

Extract, Transform, Load
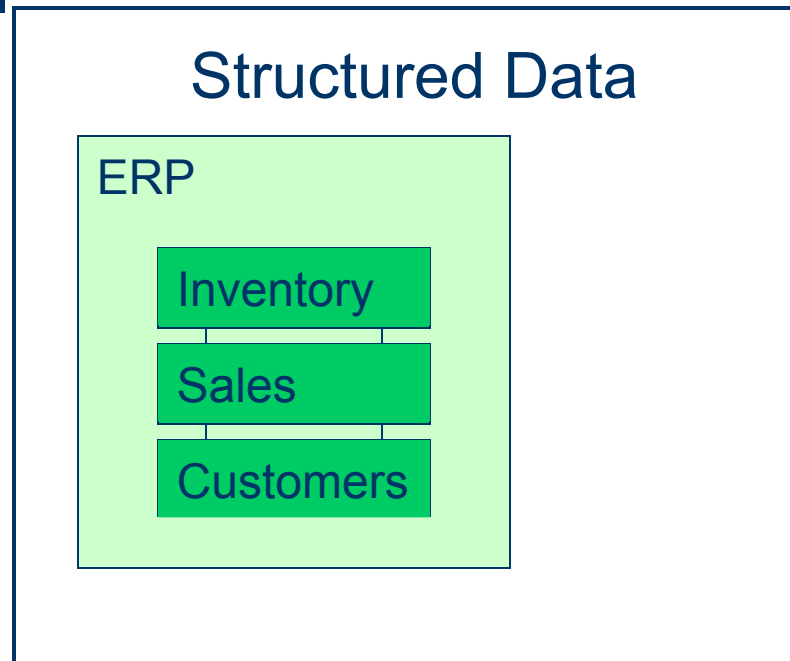
Enterprise Data Warehouse

Data Warehouse Products

- Query Tools
  - Brio
  - Business Objects
  - Forest and Trees
- ETL Tools
  - Informatica
  - Ardent
- Market size is 7 billion dollars (IDC).

Enterprise Resource Planners (ERP)

- Single commercial solution for numerous corporate applications.
- Again, a *structured* data repository.

Structured Data

ERP

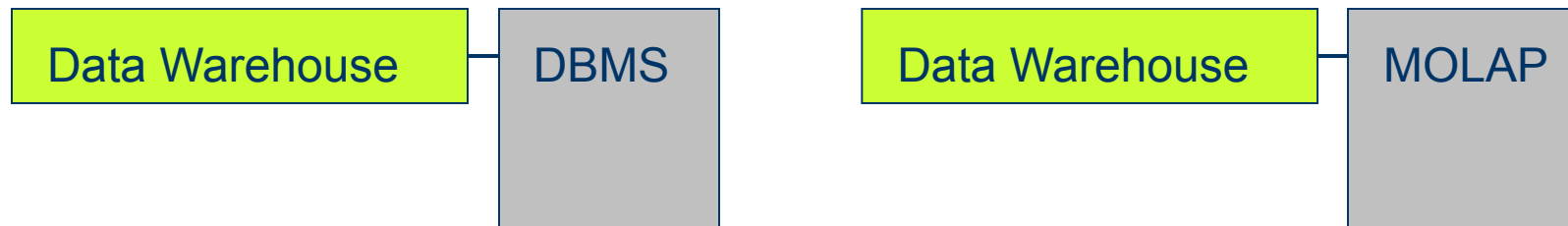Inventory

Sales

Customers

ERP Vendors

- SAP
- J.D. Edwards
- Baan
- PeopleSoft
- Oracle
- Lawson

- Estimated Size of Market: 9.7 Billion in 2000 (Merrill Lynch)

OLAP Tools

- On-line analytical processing
  - Allows users to slice and dice data through various dimensions. View all products by date, by customer, by region, etc.
  - New type of database engine
- Simply another data warehouse repository built only for *structured* data

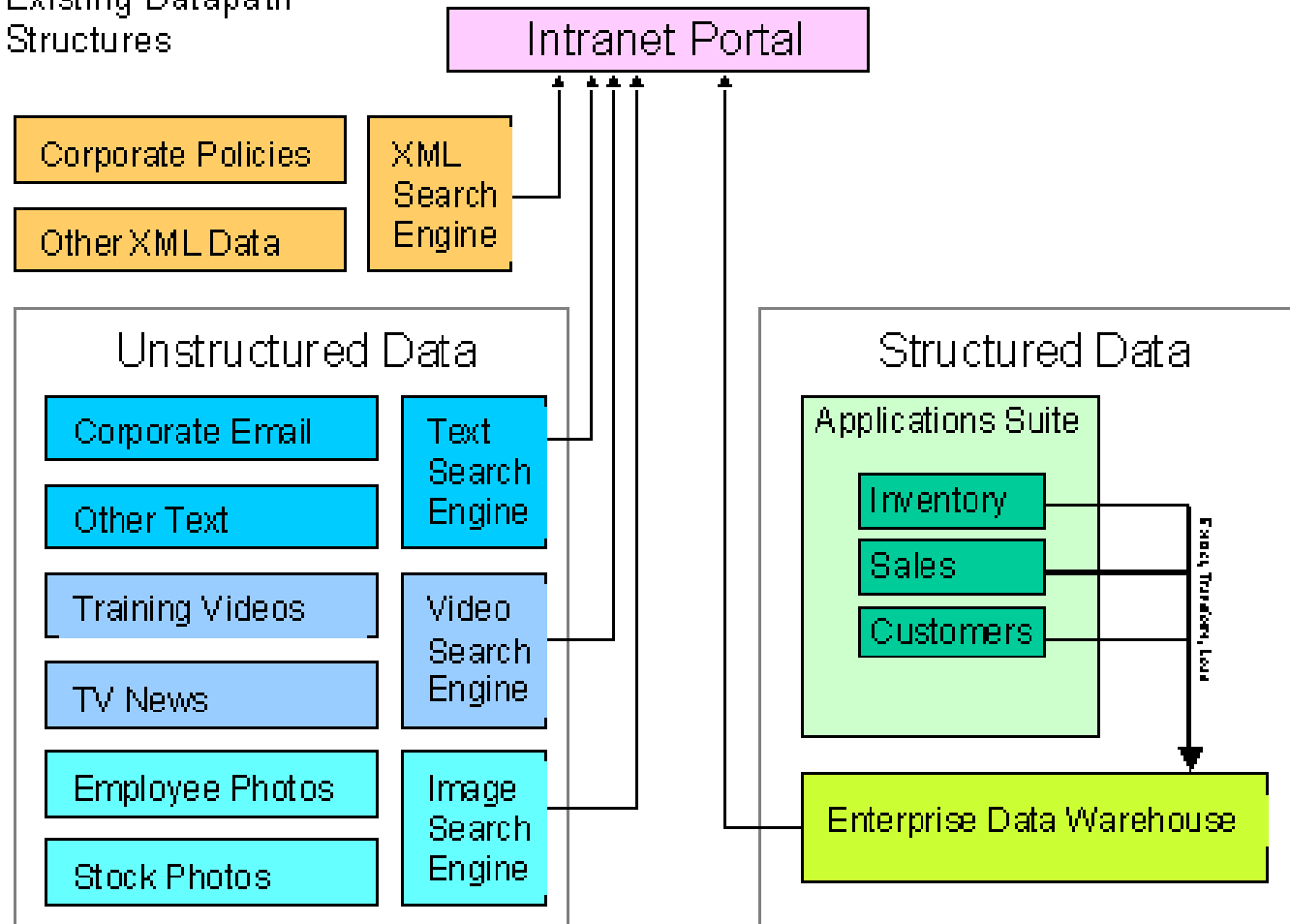| Data Warehouse | DBMS | | Data Warehouse | MOLAP |

OLAP Products

- Business Objects
- Clime
- Cognos Powerplay
- Crystal Info
- Oracle Express
- Gentia
- Holos
- Hyperion Essbase
- iTM1
- Knowsys ProClarity & Max 3.0
- MetaCube
- MIS Alea & DeltaMiner
- PowerOLAP
- Pulse Suite
- SQL Server - Analysis Services
- SuperModeller
- Visual OLAP
- Size of Market: 1.8 billion in 1998 (InformationWeek)
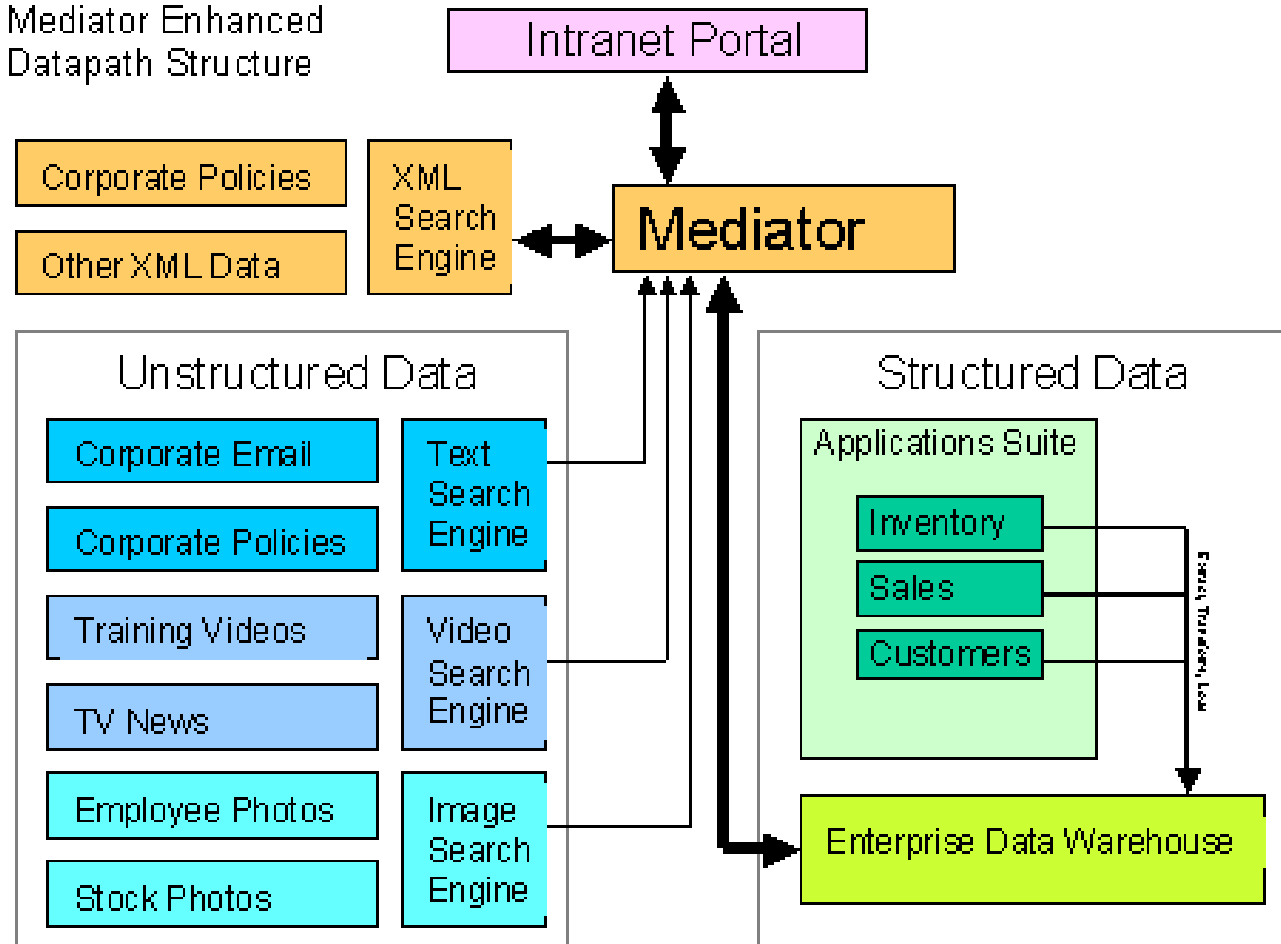
Enterprise Information Portals

Portal Products

- ExcelWriter
- Octane8
- IntraView Corporate Intranets
- Enlighten.Net
- tmsINTRANET
- Business Collaborator
- Aptrix
- Instant Intranet Builder
- Web 500 Intranet
- InStranet 2000
- Intranet Connections
- Adapting Intranet
- Active Intranet
- ZOPE
- SixCMS
- intraline
- IntraMedia
- OfficeClip-In-A-Box
- Illumination
- tmsINTRANET
- KnowledgeGear IMS
- Interfacer
- Company Intranet
- Decision.Control
- SenseNetII Office
- Statistics Server Live Stats

- Roxen Platform
- CyberSERV
- Netobjects Authoring Server 2000
- SenseNet II Office
- Merlin
- Intranet Connections
- Tango
- simpleWebGraph4
- Ding! Enterprise 2.0
- Swwwing Intranet in a Box
- Allure 4.1
- Decosoft Direction
- Business in a Box 4.6
- insideRoom 2.0
- Intranets.com
- Planet-Intra
- Mindbridge
- OfficePack 1.1
- Taskforce 1.0
- FirstClass Intranet Server 5.5
- **Accent 2.0**

- Portal market will reach $1.2 billion by 2003 – Yankee Group

Next Generation Search Engine
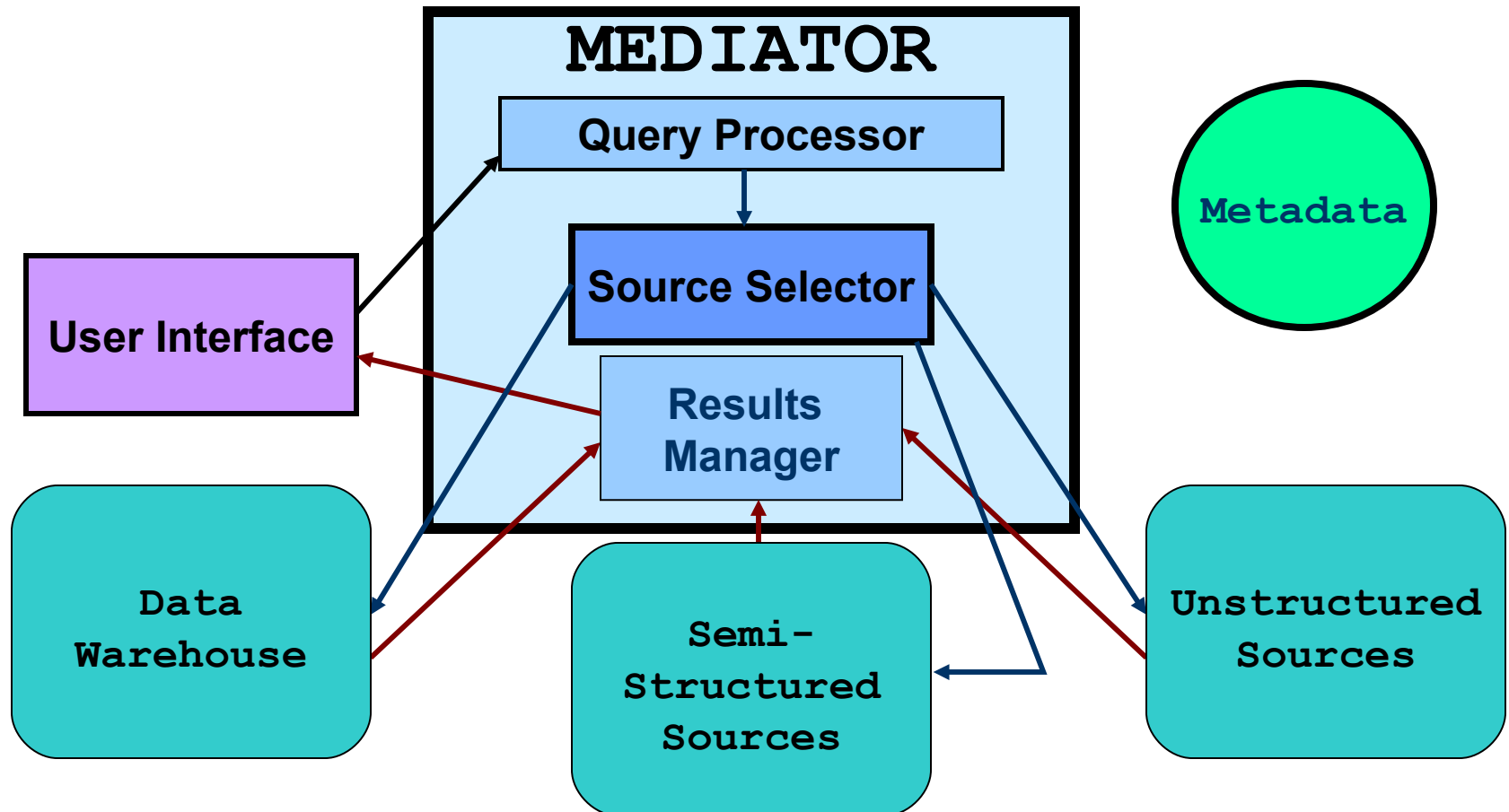
Intranet Mediators Technical Overview

- An Intranet mediator provides unified access to disparate data.
- Deployed on top of an existing data warehouse
  – structured data are reconciled off-line prior to query time.
- Mediator accepts the query and selects the best possible information sources.
- One query can be posed to structured, semi-structured, and unstructured data sources.
- Use of data warehouse makes the mediator *much* faster and *much* more accurate.

General Mediator Algorithm

- We query disparate data by:
  - Extracting concepts from the query.
  - Selecting appropriate data sources.
  - Posing the query to each source.
  - Presenting the user with an answer to their query, along with a collection of related information in case our answer was not enough.

Architecture

Barriers to Entry

- IIT has filed a Patent Application
- Existing development
    - Core mediator
    - XML Taggers
- Vendors unlikely to respond until they see initial successes from the mediator and interest in the product from their key customers.

Data Suppliers

- Internal data from the customer
- External data sources
    - Clickstream Data
    - Credit Rating
    - Demographic Information
    - News Service
    - Stock Market Data
    - Weather Service

Potential Customers

- Fortune 500 Companies
  - Estimates range from 5 to 30 billion dollars are spent on business-intelligence software

- Government Organizations

- Key Requirements
  - Need integrated response to natural language questions.
  - Have an existing data warehouse.

Plans for Phase II (2 semesters)

- Complete pilot at IIT (semester 1)
  - We already have undergraduate bulletin, would add graduate and course offerings as well
- Identify a corporate intranet to use as an initial test site for the mediator (semester 2).
- Present results at key industry forums (e.g. data warehouse institute), academic conferences, etc.

Risks

- Adding data to the mediator may be more complicated than we think.
- The mediator may not scale to the terabyte range
- Some organizations may not have a working data warehouse.

Risk Mitigation

- We have added XML data, structured data and built search engines which routinely access large volumes of text.

- The data warehouses that exist often scale to the terabyte range, we will be leveraging off of this asset.

- Most Fortune 500 companies have at least one data warehouse.

Summary

- Niche market
- Business plan shows that this will integrate structured data management products, semi-structured and unstructured products
- Patent application filed.