“Applying a Unique Approach in a USJFCOM Joint Experimentation (J9) Rapid Assessment Project for Operational Net Assessment (ONA) Data Integration”

Gavin Robertson, CTO, WhamTech, Inc.
Agenda

1. What is WhamTech’s virtual data integration product (called EIQ Server®)?
2. EIQ Server vs. other data integration methods
3. USJFCOM Project Description
4. Conclusions
5. Recommendations
6. Questions
EIQ Server

- SCHEMA SAME AS DATA SOURCE FOR STRUCTURED QUERIES

Middleware
- Indexes
- Indexes
- Indexes

Queries resolved 100% in the middleware level

Structured
- Structured
- Structured

Data Source

Application

STANDARD DATABASE DRIVERS,
WEB SERVICES, AND SQL
**EIQ Server**

**PROS**
- Data warehouse performance
  - Find almost 100% of data vs. “up to 50% not found” in federated systems
  - Indexes/results clean and usable
  - Complete control over indexing and query processing
  - Consistent and multiple indexes across disparate data sources
- Data remains at source
- No major data and schema transforms
- No federated adapters or specialized connectors
- Almost any data source
- Highly flexible
- Security and metadata managed in middleware
- Almost no load on data source system

**PROS (continued)**
- Data source system/owner unaware of queries
- Index monitoring agents feed subscriptions
- Fast
- Connect to data sources “as user”
  - Security
  - Data updates – two-way

**CONS**
- Establishing index updates
- Indexes require storage
The project lasted over three months in 2004

- Five data sources selected out of eleven candidates
- DOD XML Metadata Registry was used for metadata
- Indexes were built and maintained external to the data sources
- SQL queries based on metadata including JOINs, range queries, and text search, were executed against the external indexes
- Pointers to result-set data were isolated, and...
- Results retrieved from data sources, integrated and presented in a standard format
Included Data Sources

1. DS1 – ONA SQL Server relational database and associated Word documents (structured and unstructured – static)

2. DS3 – TRACES, (stripped) patient medical records in an Excel spreadsheet (semi-structured – static)

3. DS6 – SEAS PMESII model results of simulating effects of a biological attack in two XML files (semi-structured static)

4. DS10 – Web documents from ONA-provided news Web sites (unstructured – batch/incremental update)

5. DS11 – RSS news feeds, including ONA-provided news Web sites (semi-structured – near real-time)
Excluded Data Sources

1. GTN database
2. ACTD Rosetta
3. Census data
4. NGA Fortune Cookie
5. FBIS Web site
6. Others
Reasons Data Sources Excluded

1. Access difficulties
   - Even though unclassified, they resided on limited access systems

2. Owners were reluctant to allow WhamTech to parse and index content
   - Not necessarily a copyright issue, but more of a process issue

3. Data was so disparate that there was little or no commonality
Metadata

- Access to the DOD XML Registry
  - 30 separate metadata repositories
  - Largest was TBD – “To Be Determined” - with over 14,000 data elements
  - Of the 30,000+ data elements, a lot of redundancy (overlap)
  - Able to use some (~25%) from COAL, GMI, INT, PER, and TBD
Basic Configuration Process

- Register a data source
- Build an index
- Create a Virtual Data Source
  – An index and registered data source pair
- Create a Superschema metadata result-set table containing a list of the data and information of interest
- Map data source fields to Superschema metadata
Example Combined Structured Data Query and Unstructured Text Search
Example Result Set in HTML
(other options were Excel and XML)
WhamTech Conclusions

- As an unclassified experiment, access to data sources was restricted or not a high priority
  - Should not be the case in deployment
- Cultural barriers to sharing reflected in a few data source owner’s responses
- Within DOD, a plethora of metadata dictionaries
  - None for ONA
- More than one metadata dictionary needs to be mapped to same data
- More than one metadata mapping WITHIN same metadata dictionary
- Need to accommodate variations in so-called standard DOD data
- Could probably use results level indexing instead of data level indexing with ONA and other complex or restricted access data sources
- Novel approach to Excel and XML files, enabling standard driver and SQL access to data as though database tables
JFCOM J9 Project Alpha Conclusions

- Unique approach to data integration
- Well suited to the ONA process, EBO, JC2, CIE, TIA, HF, and outside DOD in DHS, Intel and law enforcement agencies
- Advantages over data warehousing and federated database approaches
- Despite constraints, able to integrate disparate data sources in real-time
  - Represents an opportunity for ONA analysts to focus on analysis than data and information gathering
- Real benefits go beyond time savings...allows the analyst to accomplish more than current processes allow
WhamTech Recommendations (1 of 2)

- Develop a “best of” global common metadata dictionary for data integration and sharing, e.g. Esperanto
  - Don’t force all to adopt – allow applications, organizations, and countries to continue with own metadata and language
    - Map data sources to it
    - Map applications to it
- Need for ONA metadata dictionary and/or terms
WhamTech Recommendations (2 of 2)

- In an integrated ONA system:
  - ONA database as source of query/search terms
  - Other systems used as source for the ONA database
  - Entity extraction extremely valuable to ONA
  - Other KM tools such as semantic reasoning, categorization, and summarization
  - Closer to near real-time Assessment -> Planning -> Execution -> Assessment loop (EBO)
  - Real-time, interactive visualization could add significant value to ONA
  - Closer tie-in to business process management
  - Multiple ONA systems could be integrated and shared at a higher level
  - The communities of interest (COI) approach would seem to lend itself to ONA and EIQ Servers running ONA
  - Multiple COI systems could be integrated and shared at a higher level
FCOM J9 Project Alpha Recommendation

- Novel nature of EIQ Server warrants further investigation and integration
- EIQ Server approach for other areas than ONA
- Assistant Secretary for Defense for Network and Information Integration to include EIQ Server as part of HF and a future Quantum Leap proof-of-concept experiment
- WhamTech seek accreditation for EIQ Server use with classified data sources
Future Plans

• Included in several federal and state agency, and commercial project proposals
• Build on existing implementations as a turbo charger for RDBMSs and as a much-improved adapter in federated information sharing systems
• Aim to allow almost any application to work with almost any data source
  – Universal metadata management
  – Universal interoperability
  – Ultimate goal: universal semantic interoperability
• Reviewing inclusion of Latent Semantic Indexing (LSI)
• Improved entity extraction
• Link Indexes
  – Direct and indirect link analysis in middleware
Acknowledgements

Dr. Russell Richards, Dr. Kevin Brandt, Paul (Tom) Fernan and Christian Grant, USJFCOM, J9, Project Alpha
Questions?

Gavin Robertson
CTO & Senior VP
WhamTech, Inc.
4450 Sojourn Dr., Suite 200
Addison, TX 75001, USA
+1 (972) 380-4645 x223
gavin.robertson@whamtech.com
www.whamtech.com
Backup Slides

- Different Approach to Metadata Management
- Index Updates
- Current EIQ Server System Architecture
- Future EIQ Server System Architecture

Back to Questions?
Different Approach to Metadata Management

**Legacy Application**
- **QUERY IN LEGACY SCHEMA AND DATA**
  - **LEGACY SCHEMA TO COMMON SCHEMA TRANSFORMATION** (to be developed)
  - **LEGACY DATA TO COMMON DATA TRANSFORMATION**
  - **RESULTS TRANSFORMATION** (opposite of legacy data to common data transformation)
- **Application written for common schema and legacy data**

**Application written for common schema and legacy data**
- **QUERY IN COMMON SCHEMA AND DATA**
- **DATA SOURCE TO COMMON SCHEMA AND DATA TRANSFORMATION**
- **RESULTS IN COMMON SCHEMA AND DATA**

**Extended Information Sharing Approach**
- **Application written for common schema and data**
- **QUERY IN COMMON SCHEMA AND DATA**
- **QUERY IN LEGACY SCHEMA AND DATA**
- **LEGACY SCHEMA TO COMMON SCHEMA TRANSFORMATION**
- **LEGACY DATA TO COMMON DATA TRANSFORMATION**
- **RESULTS IN COMMON SCHEMA AND DATA**

**Standard Information Sharing Approach**
- **Adapter or EIQ Server**
- **DATA SOURCE**
- **STANDARD INFORMATION SHARING SYSTEM**
Index Updates

- At least seven methods:
  - **Data level indexing:**
    - Batch
      - Complete refresh
    - Incremental
      - Batch updates
    - Transaction or change logs
      - Usually not on data source system
    - Triggers
      - Usually install on data source system
    - Message Queues
      - Tap into
    - Existing replication/backup software
      - Use target as source
  - **Results level indexing**
    - Update rate and route depends on system
Future EIQ Server System Architecture

**APPLICATIONS**

- APPLICATIONS /CLIENTS

**APPLICATION-LEVEL: ENTERPRISE ADMINISTRATION TOOLS**

- Meta Data Management Tool
- Security Configuration Tool
- Configuration Tool
- Diagnostics Tool
- Query Tool
- ETI Tool
- Update Server Configuration Tool

**WEB SERVICES SERVER**

- UDDI Registry
- WSDL Registry

**META DATA SERVER**

- Meta Data Repository

**SPAPs SERVER**

- Security and Privacy Access Profiles

**QUERY SERVER**

- Rules Server
- EIQ Server
- Relational Query Engine

**INDEX SERVER**

- Index Engine
- EIQ Server Indexes

**DATA SOURCE SYSTEM(S)**

- Data Source
- Index Update

**WEB SERVICES**

- ODBC, JDBC, OLEDB w/ADO, WhamNAPI and JAVA WhamNAPI

**APPLICATION-LEVEL TOOLS**

- Diagnostics Tool
- Query Tool
- Security Configuration Tool
- Configuration Tool
- ETI Tool
- Update Server Configuration Tool

**APPLICATIONS /CLIENTS**

- Diagnostics Tool
- Query Tool
- Index Engine
- EIQ Server
- Security Configuration Tool
- Configuration Tool
- ETI Tool
- Update Server Configuration Tool

**META DATA Server**

- Meta Data Repository

**SPAPs Server**

- Security and Privacy Access Profiles