Semantics for Risk Analysis in Reference Data Distribution

Dr Barbara Hughes, PolarLake
Amir Halfon, Oracle
Topics

- Introduction
- Using Semantics to Achieve Quality Data
- Semantics in Data Management
- The Oracle Semantics Platform
- Case studies
- Summary
**PolarLake**
Comprehensive and Open Reference Data Framework
Purpose-built Reference Data Applications
Experienced domain specific and integration specialists

**The Benefits**
Reduced project risk
Quicker implementations
Less Total Cost of Ownership

*6 of the top 10 Investment Banks*
*2 of the top 5 Prime Brokers*
*2 of the top 10 Asset managers*
Efficient Data Supply Chain is central to Good Risk Management

**Feed On-Boarding & Search Service**
Central Sourcing of Vendor Price and Reference Data

**Price Processing**
Pricing Exception Management and Price Arbitration

**Security Master**
Building up Consolidated, Validated or “Cleansed Records”

**Reference Data Distribution**
Getting the Right Data to the Right Place at the Right Time in the Right Format

**Cost Optimization**
Optimizing and Controlling Pricing and Reference Data Spend Across the Firm
Complete Supply Chain for Reference Data

Integrated End to End Business Operation Monitoring

Rapid New Product Onboarding
11 x Faster Loading Into Soft Repository

Validation, Exceptions, Data Lineage, Identification Resolution, Routing, Transformation, Classification, Normalisation, Arbitration, Workflow, Decision Tables, Sequencing, Assembly, Monitoring

Multiple Formats, Protocols and Interaction Models Entitlements and Quota.
Subscriptions, Request/Reply, Deltas, In Memory Cache Distribution

Analyst Sandbox: Semantic Enabled
Semantic Models

Acquisition
Manufacturing
Distribution
Data Management – The traditional way

**Traditional Stages:**
1. Design Canonical Model
2. Map Sources to Model
3. Distributed from Model

- **Vendor A**
  - Commercial / Home grown Data Store
    - Rigid Canonical Models
    - Never complete – ‘semantic’ mismatch
    - Source feeds change and new ones arrive
    - Poor at conflicting data and classifications
    - Cause of Data Loss

- **Vendor B**
  - Commercial / Home grown Data Store
    - Rigid Canonical Models
    - Never complete – ‘semantic’ mismatch
    - Source feeds change and new ones arrive
    - Poor at conflicting data and classifications
    - Cause of Data Loss
PolarLake Semantic Repository and Data Store

Vendor A Data
Vendor B Data
Internal Data

Vendor A Metadata
Vendor B Metadata

Linkages
Metadata imports
Classification imports

Custom Models
Semantic Store
Benefits: Soft Models, Semantics and Policy Engine

Data Suppliers:
Reference, Price, Trade Data

Vendor Feeds

Internal Sources

Apps
Trade SWIFT, FpML, CSV

Data Consumers

Back Office
Clearing and Settlement

Middle Office
Risk Management

Front Office
Trading

Business/Data Analyst
Rule Management Configuration
Scenario Analysis

Business Management
Audit, Compliance, Exceptions,
Reports and Analytics

Business Operations Dashboard

Business Analyst Workbench

Data Distribution API

Data Supplier Management

Semantic Data Policy Engine

Data Consumer Management

Semantic Models

Dynamic Data Store and Cache

Consumer/Supplier – Connectivity Subsystem
Oracle 11g Semantic Technologies

- Only leading commercial database with native semantic data management
- Scalable & secure platform scales to repositories w/ billions of triples
- RAC & page-level compression support
- Choice of SQL or SPARQL query
- Native inferencing and 3rd party reasoner support e.g., PelletDB
- Industry leading 3rd party & open source tools, services, apps support
- W3C standards-based technologies

Key Capabilities:

**Load / Storage**
- Native RDF graph data store
- Manages billions of triples
- Fast load - bulk, incremental

**Query**
- SPARQL-Jena/Joseki, Sesame
- SQL: SEM_Match
- Ontology assisted query of relational data

**Reasoning**
- RDFS, OWL 2 RL support
- User-defined SWRL-like rules
- Plug-in architecture
<table>
<thead>
<tr>
<th>Reasoning and Discovery</th>
<th>Data Integration</th>
</tr>
</thead>
<tbody>
<tr>
<td>• Persistent RDFS / OWL inferencing</td>
<td>• Jena &amp; Sesame distributed SPARQL queries</td>
</tr>
<tr>
<td>• User-defined inference rules</td>
<td>• Integration w/ top 3rd party NLP entity extraction engines: e.g., OpenCalais</td>
</tr>
<tr>
<td>• Plug-in architecture for inference engines such as PeltetDB, OntoBroker</td>
<td>• Semantic Indexing for documents</td>
</tr>
<tr>
<td>• Inferencing proofs and explanations</td>
<td>• Versioning</td>
</tr>
<tr>
<td>• SPARQL &amp; mixed SQL DB queries</td>
<td></td>
</tr>
</tbody>
</table>
Oracle Database Provides:

<table>
<thead>
<tr>
<th>Scalability</th>
<th>Security</th>
</tr>
</thead>
<tbody>
<tr>
<td>• Compact storage of RDF data – removes RDF redundancy and verbosity</td>
<td>• Virtual Private Database declarative constraints based on RDF data char. &amp; app. / user context</td>
</tr>
<tr>
<td>• 60% storage reduction w/ page compression</td>
<td>• Oracle Label Security restricts RDF data access to users having compatible access labels</td>
</tr>
<tr>
<td>• SQL*Loader direct-path bulk loading</td>
<td></td>
</tr>
<tr>
<td>• RAC, Exadata, partitioning</td>
<td></td>
</tr>
<tr>
<td>• Incremental &amp; parallel inferencing, large scale owl:sameAs handling</td>
<td></td>
</tr>
<tr>
<td>• Optimized queries, hinting</td>
<td></td>
</tr>
</tbody>
</table>
Semantic Search and Indexing for Documents

• Create a semantic index on documents using third-party entity extractors.
  – Out-of-the-box support for OpenCalais from Thomson Reuters & General Architecture for Text Engineering (GATE).
  – Extensible framework to plug-in any information extractor.

• Embed SPARQL graph pattern queries in SQL to find documents of interest and return relevant information
  E.g., Find documents that refer to some financial institution in the context of Bailout.
Fine-Grained Access Control mechanisms restrict access to data within the most critical database objects.

- Uses Virtual Private Database feature of Oracle Database to apply declarative constraints to RDF classes and properties.
- Intercept and rewrite the user query to restrict the result set using additional predicates.
- Restricts access to parts of the RDF graph based on certain characteristics of the instance data and application/user context.
  - E.g: Only a manager of a contract can access its monetary value
Change Management for Semantic Data

• Manage public and private versions of semantic data in database workspaces
• Collaborate with multi-user workspaces
• Isolate a group of changes in a workspace
  – Data in multiple valid states: e.g., Current – Planned - History
  – Publish all, some or none of the changes made in a workspace
• Create multiple scenarios in different workspaces
  – “What if” analysis
  – Allow multiple application testers to use the same data set
• Provides efficient data storage and querying
  – New versions created only for changed data
  – Queries are workspace-specific
  – Uses the Workspace Manager feature of Oracle Database
## Query Performance on Oracle Exadata

### Ontology
**LUBM25K**
- 3.3 billion & 2.7 billion inferred

### OWLPrime & new inference components

<table>
<thead>
<tr>
<th>Query</th>
<th>Q1</th>
<th>Q2</th>
<th>Q3</th>
<th>Q4</th>
<th>Q5</th>
<th>Q6</th>
<th>Q7</th>
</tr>
</thead>
<tbody>
<tr>
<td># answers</td>
<td>4</td>
<td>2528</td>
<td>6</td>
<td>34</td>
<td>719</td>
<td>260M</td>
<td>67</td>
</tr>
<tr>
<td>Complete?</td>
<td>Y</td>
<td>Y</td>
<td>Y</td>
<td>Y</td>
<td>Y</td>
<td>Y</td>
<td>Y</td>
</tr>
<tr>
<td>Time (sec)</td>
<td>0.01</td>
<td>20.65</td>
<td>0.01</td>
<td>0.01</td>
<td>0.02</td>
<td>23.07</td>
<td>4.99</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Query</th>
<th>Q8</th>
<th>Q9</th>
<th>Q10</th>
<th>Q11</th>
<th>Q12</th>
<th>Q13</th>
<th>Q14</th>
</tr>
</thead>
<tbody>
<tr>
<td># answers</td>
<td>7790</td>
<td>6.8M</td>
<td>4</td>
<td>224</td>
<td>15</td>
<td>0.11M</td>
<td>197M</td>
</tr>
<tr>
<td>Complete?</td>
<td>Y</td>
<td>Y</td>
<td>Y</td>
<td>Y</td>
<td>Y</td>
<td>Y</td>
<td>Y</td>
</tr>
<tr>
<td>Time (sec)</td>
<td>0.48</td>
<td>203.06</td>
<td>0.01</td>
<td>0.02</td>
<td>0.02</td>
<td>2.40</td>
<td>19.45</td>
</tr>
</tbody>
</table>

- **Setup**: Full Rack Sun Oracle Data Machine and Exadata Storage Server (8 node cluster)
- **Auto DOP is used. Total # of answers 465,849,803 in less than 5 minutes**
Case Study 1: Disruptive Technology with Massive Performance Improvements

- **11X Faster Load Time**
- **17X More Securities Queried Per Second**
- **20X Faster Feed On-boarding**

Bloomberg, S&P, Fitch, Moodys Loaded, Linked and Searchable in 3 weeks
Case Study 2: Managing Ratings Data for a large Buy Side firm.

Managing Reference Data and Ratings Data Throughout the supply chain.

✓ Compressing a 3.5 year program to 5 months
✓ On-boarding over 300 data formats
Case Study 3: Semantic Policy Based Data Management

Financial Institution

- Investment Banking Division
- Private Banking Division
- Asset Management Division

120+ applications
1,000 requests per day

Vendors

- Ad hoc Service
- Bulk Data

PolarLake RDAM

BUSINESS OPERATIONS DASHBOARD

- Authorise & Authenticate
- Policy Layer
- Usage Monitoring
- Per-Division Data Order Mgmt
- Entitlements, Accounting, Authentication
- Semantic Data Store

Ad hoc Integration

Request

Pre Stock

Out of Stock -> In Stock

Message Audit & Reporting

Operations Support: Resolve business user enquiries, disputes

Market Data Management Usage
Reports by instrument, user, division, etc.

Equity / Equity Funds: T&C, EODP
Fixed Income: T&C, EODP, ratings
Issuer & Rating data
Summary

Using semantics we have developed an efficient supply chain that is robust and scalable.

Supply quality data – the raw material of risk management.

Diverse case studies prove the success of our semantic approach.
Thank you