Agenda

• Background and Overview
• Use Cases
• Oracle Database 11g Semantic Overview
• Performance
• Customers and Partners
Background and Overview
Brief Background

• Oracle Database 10g Release 2 - 2005
  • Beta adopters, W3C, Tim Berners-Lee apply Oracle Spatial Network Data Model “link-node” graph implementation to Semantic models.
  • Early adopters in life sciences, pharma, and intelligence community R&D use Oracle NDM to overcome (primarily) scalability, performance, and security issues faced in prototypes with other technologies.

• Oracle Database 11g Release 1 – 2007, 2008
  • Oracle introduces first commercial database with native RDF, OWL, ontology support, inferencing, and integration with leading open source and commercial semantic tools.
The Problem

- Discovery of data relationships across...
  - Structured data (database, apps, web services)
  - Unstructured data (email, office documents) Multi-data types (graphs, spatial, text, sensors)

- Text Mining & Web Mining infrastructure
  - Terabytes of structured & unstructured data

- Enable data reuse by associating more meaning (context) with the data
Semantic Technologies in a Nut Shell

• Model complex real-world relationships beyond Boolean in the data as a graph
  
  Allow schemas to continuously and dynamically evolve

• Inference among relationships with rules, standard concepts and terms to discover new information
  
  Enable machine-driven creation of new data relationships without restructuring the data model

• Query using graph patterns to provide more semantically complete information for decision-making
  
  Support discovery workflows
Adoption of Semantics-enabled Business Applications

- Intelligence, Law Enforcement:
  - Threat analysis, asset tracking, integrated justice
- Integrated BioInformatics & Health Care
  - Bio-Pathway analysis, protein interaction
- Health Care Informatics
  - Patient records, reporting, bio-surveillance
- Finance
  - Fraud detection, Compliance Management
- Web and Social Network Solutions
  - Recommender, Social Network Analysis, Activity Analysis
- Media, Games, Content Management
  - Media metadata, content re-purposing
Why Organizations use Oracle RDF Store

- Standards Compliance
  - RDF, RDFS, OWL; ontologies – SNOMED, SKOS, etc.

- Open Source Technology Support
  - Jena, Joseki, ARQ, TDB, SDB, Sesame, Pellet, D2RQ, Jetty, Cytoscape, GATE, Protégé…

- Scalability & Performance
  - storage, native inference engine, querying
  - Customers tell us other triple-stores routinely fail with large datasets
  - Leading Lehigh University Benchmark (LUBM) results

- Query RDF data using SPARQL or SQL
  - SEM_MATCH, SEM_CONTAINS, Ontology-assisted SQL queries

- Security
  - Virtual Private Database, Oracle Label Security to control access

- Semantic Indexing for Documents stored in Oracle Database
Life Science Example: Ontology Assisted Query

<table>
<thead>
<tr>
<th>ID</th>
<th>DIAGNOSIS</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Hand_Fracture</td>
</tr>
<tr>
<td>2</td>
<td>Rheumatoid_Arthritis</td>
</tr>
</tbody>
</table>

Patient has diagnosis :Upper_Extremity_Fracture

\[\text{rdfs:subClassOf} \Rightarrow \text{rdfs:subClassOf}\]

\[\text{:Hand}_\text{Fracture} \quad \text{:Arm}_\text{Fracture} \quad \text{:Upper}_\text{Extremity}_\text{Fracture}\]

\[\Rightarrow \text{:Hand}_\text{Fracture} \quad \text{:Upper}_\text{Extremity}_\text{Fracture}\]

NCI Ontology
Customer Use Cases
Extract, Model, Reason & Discover Workflow

Transform & Edit Tools
- Entity Extraction & Transform
  - OpenCalais
  - Linguamatics
  - GATE
  - D2RQ
- Ontology Eng.
  - TopQuadrant
  - Mondeca
  - Ontoprise
  - Protege
- Categorization
  - Cyc
- Custom Scripting

Load, Query & Inference
- RDF/OWL Data Management
- SQL & SPARQL
  - Sesame
  - Jena/Joseki
- Native Inferencing
  - PelletDB
- Semantic Rules
- Scalability, Security, Versioning
- Semantic Indexing

Applications & Analysis Tools
- BI, Analytics
  - Teranode
  - Metatomix
  - MetTrust
- Graph Visualization
  - Cytoscape
- Social Network Analysis
  - Metadata Registry
  - Faceted Search

Partner and Open Source Tools

Transaction Systems
Unstructured Content
RSS, email
Other Data Formats
Data Sources

Partner and Open Source Tools

Oracle Spatial
Entertainment Repository

- Oracle scales for millions of movie shot files, thousands of artists
- Sesame “out-of-memory” error sorting 10+ million rows
- RDF graph describing a movie shot allows sharing and reuse
- UI uses SPARQL graph pattern query to find movie shots
Cisco Enterprise Collaboration Platform - QUAD

Chose RDF as the data model for sharing ideas and following people, communities, and information across the enterprise.

Chose Oracle Database semantic store for its scalability, fine-grained security, incremental inferencing, and support for stds.

- Billions of relationships
- Sesame was unstable with large datasets
- Unifying RDF metadata model for
  - blogs, wikis, calendar, IM, WebEX, voice, and video
- Transactional workload requires incremental load & inference
- SPARQL graph queries
Medical Research: Cohort Identification

User interfaces
- Patient Data Entry
- Cyc natural language query
- Structured query
- Semantic wiki

Oracle SOA Suite 11g

Cyc ontologies

Oracle Database Semantic Technologies

Data-source adaptors

SPARQL

Operational System Data
- Patient registry
- Tagged literature, e.g., PUBMED
Using RDF/OWL for Cohort Identification

Why Oracle: Special purpose semantic store didn’t scale

<table>
<thead>
<tr>
<th>Added value of RDF/OWL</th>
<th>Relational Approach</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Searches:</strong></td>
<td></td>
</tr>
<tr>
<td>• Researchers’ criteria are defined in the data as RDF properties</td>
<td>• Criteria translated by DBA</td>
</tr>
<tr>
<td>• Queries are sub-graph patterns</td>
<td>• Requires foreign keys, joins, where-clauses, aggregation and text matching</td>
</tr>
<tr>
<td><strong>Schema evolution:</strong></td>
<td></td>
</tr>
<tr>
<td>• Relationship defined with one triple</td>
<td>• Relationship requires adding a new attribute column and/or foreign keys</td>
</tr>
<tr>
<td>• Asserted or inferred for all RDF data</td>
<td></td>
</tr>
<tr>
<td><strong>Discovery:</strong></td>
<td></td>
</tr>
<tr>
<td>• RDF graph structure is navigable</td>
<td>• Queries based on preexisting knowledge of the schema.</td>
</tr>
<tr>
<td>• Enables faceted search</td>
<td>• Relationships defined manually</td>
</tr>
<tr>
<td>• New relationships automatically discovered by OWL inferencing</td>
<td></td>
</tr>
</tbody>
</table>
Vocabulary Management

Intelligent Topic Manager (ITM)

- Manage
- Edit
- Maintain
- Search
- Control
- Import
- Export
- Audit

Oracle Database 11g
Ontology repository

Oracle RDF 11g
For graph based search

Oracle Text 11g
For text based search

ITM Features
- Web User Interfaces
- Multilingual
- Connectors to text mining
- Collaborative maintenance
- Import / export
- Scalability
- API & Web Services
- Java – J2EE – LDAP

Semantic Portal
Text Mining: National Intelligence

Web Resources

Information Extraction
Categorization, Feature/term Extraction

Processed Document Collection

RDF/OWL

Ontologies

Domain Specific Knowledge Base

SQL/SPARQL Query

Explore

Analyst

Browsing, Presentation, Reporting, Visualization, Query

Content Mgmt. Systems

News, Email, RSS

Ontology Engineering Modeling Process
Oracle 11g Semantic Technologies Overview
Oracle Database 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 batch, bulk and incremental load

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
Oracle Database Provides

Reasoning and Discovery
- Persistent RDFS / OWL inferencing
- User-defined rules for inferencing
- Plug-in architecture for inference engines such as PelletDB, OntoBroker
- Inferencing proofs and explanations
- SPARQL & mixed SQL DB queries

Data Integration
- Ontologically-assisted SQL queries
- Jena & Sesame distributed SPARQL queries
- Integration w/ top 3rd party NLP entity extraction engines: e.g., OpenCalais
- Semantic Indexing for documents
- Versioning
Oracle Database Provides

Scalability

• Efficient RDBMS storage of RDF data
• Support RAC, Exadata platform, partitioning, page compression, versioning
• Incremental & parallel inferencing
• SQL*Loader direct-path bulk loading

Security

• Virtual Private Database declarative constraints based on RDF data char. & app. / user context
• Oracle Label Security restricts RDF data access to users having compatible access labels
**Oracle Database Semantic Network**

**Semantic network**: models, rule bases, entailments, ...

- **SDO_RDF_TRIPLE_S**: A new object type for RDF.

- **Application Table**: Contains col of object type sdo_rdf_triple_s to allow loading and accessing RDF triples, and storing ancillary values.

- **Model**: A model holds an RDF graph and is associated with an sdo_rdf_triple_s column in an application table.

- **Rulebase**: A rulebase contains a set of rules used for inferencing.

- **Entailments**: An entailment stores triples derived via inferencing.

---

**Conceptual RDF to RDBMS Model**

- **Link Table**: Subject, Property, Object

- **Assertion/Entailments**: One row for each subject, property and object in the models w/ ID, value, type. IDs used in Triples Table

- **Value Table**: RDFS, OWL

- **Rulebases**: User Rule base..

- **Models Table**: Jena / Joseki, Sesame, SQL APIs
Oracle Database 11g OWL Support

Describes complex data models, vocabularies, logics

- RDFS++
  - rdfs:subClassOf, subPropertyOf, domain, range
  - RDFS plus owl:sameAs and owl:InverseFunctionalProperty
- OWLSIF (OWL with IF semantics)
  - Based on Dr. Horst’s pD* vocabulary¹
- OWL2 RL
  - owl:TransitiveProperty, SymmetricProperty, FunctionalProperty, InverseFunctionalProperty, inverseOf
  - owl:sameAs, differentFrom
  - owl:disjointWith, complementOf,
  - owl:hasValue, allValuesFrom, someValuesFrom
  - owl:equivalentClass, equivalentProperty
- Jointly determined with domain experts, customers and partners

¹ Completeness, decidability and complexity of entailment for RDF Schema and a semantic extension involving the OWL vocabulary
Semantic Indexing for Documents

- Links people – places – things – events to documents stored in Oracle Database though a semantic index
- Supports entity extraction tools such as OpenCalais from Thomson Reuters and GATE.
- Extends the power of Oracle Database to include semantic search in cross-domain queries.

Key Components
- Programmable API to plug-in 3rd party entity extractors
  - E.g. OpenCalais from Thomson Reuters
- SEM_CONTAINS Operator
- SEM_CONTAINS_SELECT Ancillary Operator
- SemContext Index type
Indiana authorities filed felony charges and a court issued an arrest warrant for a financial manager who apparently tried to fake his death by crashing his airplane in a Florida swamp. Marcus, 38, remained on the lam Tuesday afternoon, two days after authorities say he staged his disappearance and then rode out of a small Alabama town on a red motorcycle under cover of darkness.

**RDF/XML Output**

```xml
<rdf:rdf>
  <rdf:description rdf:about="http://..\Marcus">  
    <rdf:type rdf:resource="http://..\Person"/>
    <p:hasName .. >Marcus</p:hasName>
    <p:hasAge .. >38</p:hasAge>
    <p:hasGender .. >Male</p:hasGender>
  </rdf:description>
</rdf:rdf>

```
Fine-Grained Access Control for RDF

- Fine-Grained Access Control mechanisms restrict access to data within the most critical database objects.
- Intercept and rewrite the user query to restrict the result set using additional predicates.
- Restricted result set only includes the rows the user “needed to know”.
- Offers triple-level security for RDF data for defense, intelligence, and commercial regulatory environments.
Enterprise Security for Semantic Data

- Access control policies on semantic data
  - Uses Virtual Private Database feature of Oracle Database
  - Applies declarative constraints to RDF classes and properties
  - 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.*

- Data classification labels for semantic data
  - Uses Oracle Label Security option of Oracle Database
  - Assigns sensitivity labels to users and RDF data.
  - Restricts access to users having compatible access labels.
  - E.g: *A triple marked as “Sensitive” is only accessible to the users with clearance for “Sensitive” or “Highly Sensitive” information.*
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
Oracle Semantic Technologies Versioning

- An RDF graph may be logically versioned.
  - No duplication of data and all versions of data use a common physical storage.
  - Triples added/deleted in a workspace are private to the workspace until the workspace is merged.
  - Ability to detect conflicts with application logic.

- Physically versioned inferred data
  - Each workspace shares the inferred data with its child workspace until some data modifications are made in the child workspace.
  - Each workspace with some workspace-private changes maintains a private copy of the inferred data.

- SEM_MATCH queries on version-enabled RDF graphs are version aware.
Commitment to W3C Semantic Standards

- Our implementation entirely based on W3C standards (RDF, RDFS, OWL)
  - SPARQL support through Jena
- Members of following W3C Web Semantic Activities:
  - W3C Data Access Working Group (DAWG)
  - W3C OWL Working group
  - W3C Semantic Web Education & Outreach (SWEO)
  - W3C Health Care & Life Sciences Interest Group (HCLS)
  - W3C Multimedia Semantics Incubator group
  - W3C Semantic Web Rules Language (SWRL)
Considering Customer Requirements

- Enhance best-in-class semantic repository
  - Performance enhancements: data structures, caching, parallelism, hints, statistics and indexing, Exadata tuning
  - New versions of W3C stds. (e.g., SPARQL, OWL, RIF)
  - New versions of open source (e.g., Jena, Sesame, Pellet)
  - More inferencing: incremental delete, backward chaining
  - More utilities (visualization, querying) & 3rd party integration

- Provide building blocks for semantic BI applications
  - Graph analytics primitives for graph navigation & path querying
  - Semantic analysis of relational databases
  - Semantic / relational views and querying
    - Relational (SQL) views on RDF to integrate Oracle BI
    - RDF views and SPARQL queries on relational data
Performance & Scalability
Performance (testing in process)

LUBM 8K benchmark (1.06 billion triples asserted, 869 million inferred)
Exadata Version 2: full rack (8 database & 14 storage servers)
Parallelism used (DOP = Degrees Of Parallelism)

• Load
  • 8.5 hr. (DOP=128) from staging table
• Compression – affects performance & storage
  • 60% less space & I/O required w/ native page-level compression
• Inference performance
  • 46 min. W/ DOP=64 (1 hr. (DOP=32), 2.75 hr. (DOP=16))
• Query
  • Mostly sub-second
Customers and Partners
Some Oracle Database Semantics Customers

Life Sciences

- Lilly
- Pfizer
- Swiss Institute of Bioinformatics

Defense/Intelligence

- Geospatial-Intelligence National Security Agency

Clinical Medicine & Research

- The University of Texas Health Science Center at Houston
- Cleveland Clinic

Education

- The University of Michigan

Telecomm & Networking

- Hutchinson 3G Austria

Publishing

- Thomson Reuters
- Westlaw
## Oracle’s Partners for Semantic Technologies

### Integrated Tools and Solution Providers:

#### Ontology Engineering
- TopQuadrant
- Protégé

#### Query Tool Interfaces
- Jena
- openRDF.org
- Sesame
- Joseki
- cYcorp

#### Reasoners
- clarkparsia, llc
- Protégé

#### Standards
- W3C

#### NLP Entity Extractors
- CALAIS
- Linguamatics
- GATE

#### Applications
- Teranode
- MetatomiX
- Mondoca
- MedTrust

#### SI / Consulting
- Northrop Grumman
- Raytheon
- Orbis Technologies, Inc.
- Boeing
- McDonald Bradley
- Cognia
- Accenture
Summary: Oracle Database 11g Release 2

The only major relational database w/ native RDF/OWL support that is W3C standards compliant with

Reasoning and Discovery supporting std. ontologies persistent, native & 3rd party inference, and user-defined rules

Scalability to evolve schemas dynamically and grow to 100’s billions of triples, incremental & parallel inference

Data Integration to link structured & unstructured content, Loosely couple business silos

Security to protect data on a “need to know” basis

Integrated querying & managability SPARQL & SQL for RDF/OWL, relational, XML, text, location, & multimedia data
For More Information

http://search.oracle.com

Semantic Technologies

Google “Oracle RDF”

Or go to
oracle.com
ORACLE®