You have Terabytes Worth of Triples, Now What?
-- Mining Insights from Your Semantic Data Store

Xavier Lopez
Director, Product Management
Oracle Corporation

Zhe Wu
Consulting Member of Technical Staff
Oracle Corporation

Semantic Tech & Biz Conference- 5 June 2012
THE FOLLOWING IS INTENDED TO OUTLINE OUR GENERAL PRODUCT DIRECTION. IT IS INTENDED FOR INFORMATION PURPOSES ONLY, AND MAY NOT BE INCORPORATED INTO ANY CONTRACT. IT IS NOT A COMMITMENT TO DELIVER ANY MATERIAL, CODE, OR FUNCTIONALITY, AND SHOULD NOT BE RELIED UPON IN MAKING PURCHASING DECISION. THE DEVELOPMENT, RELEASE, AND TIMING OF ANY FEATURES OR FUNCTIONALITY DESCRIBED FOR ORACLE'S PRODUCTS REMAINS AT THE SOLE DISCRETION OF ORACLE.
Outline

• Characteristics of a RDF Triple Store
• Challenge: Need for Powerful User-friendly Tools
• Performance and Scalability of Oracle RDF store
• Integration of Business Analytics, Data Mining, and R
• Summary
Basic Characteristics of a Triple Store

• Standards Compliance – W3C
  – RDF, RDFS, OWL 2, SKOS, SPARQL, ...

• Fast loading of triple
  – Incremental and bulk loading

• Indexing of triples for fast access
  – Incrementally maintained

• Inferencing
  – Pre-computed inferences (forward chaining)
  – Run-time inferences (backward chaining)

• Querying
  – Allow multiple RDF graphs in SPARQL queries
  – Query execution planning for optimal performance
Enterprise Capabilities of a Triple Store

• Scalability to 100s of billions of triples and more
• Integrated access to Relational and RDF data
  – SPARQL query (embedding) in SQL
  – Join SPARQL results with ubiquitous relational data
  – Rich SQL operators (such as aggregates) on triples
• Semantic indexing
  – Index on a source document is an RDF graph with 1 named graph per doc
  – Triples are extracted from a document using NLP and entity/concept extraction
  – RDF graph incrementally updated as new documents entered
• Security: Fine-Grained Access Control (for each triple)
• Querying Text, Spatial and temporal data using SPARQL
Other Enterprise Features of a Triple Store

• User-defined rules

• Better inferencing
  – Higher order logic beyond OWL 2
  – Incremental inferencing for higher availability

• Tools that need to work with RDF data
  – Navigation and visualization of RDF graphs
  – Graph creation and manipulation
  – Reporting and traditional charting of selected RDF data
  – Exploring & analyzing (testing conjectures)
  – Automated discovery (mining) & predictive analysis
Capabilities Overview of Oracle Database Release 11.2

NLP engines, Tools, Editors, Complete DL reasoners, ...

SQL/PLSQL APIs & JAVA APIs (Jena, Sesame)

INFER
- RDF/S
- OWL/SKOS
- User defined rules

QUERY
- Query RDF/OWL data and ontologies
- Ontology-Assisted Query of Enterprise Data

STORE
- Incr. DML
- Batch-Load
- Bulk-Load

Built-in Security and Versioning for semantic data
- RDF/OWL data
- Ontologies & rule bases

Relational data

ORACLE
Role of Semantic-aware Ontology in Intelligence Domain

Data Sources
- Contents Repository
- Databases
- Web resources
- Blogs, Mails, news, RSS feeds

Information Extraction
- Categorization, Feature/term Extraction

Extracted Entities & Relationships
- RDF
- SQL/SPARQL

Search, Presentation, Report, Visualization, Query

National Intelligence Scenario

Person: Abduwali Abdukhadir Muse
- Nationality: Somalian
- Country: UK
- Group: Al Shabab
- Ideology: Islamist

Person: ?
- Nationality: Pakistani
- Country: Pakistan
- Group: ?

Person: Chehab Abdouljamid Bouy Aly
- Country: Morocco
- Group: al Qaeda
- Currently resides
- Member of
- Supports

Link ?

Member of
- Currently resides
- Has
- Has
Software Tooling is Key to Adoption

• Need for tools
  – Navigation and visualization of RDF graphs
  – Graph creation and manipulation
  – Reporting and traditional charting of selected RDF data
  – Exploring & analyzing (testing conjectures)
  – Automated discovery (mining) & predictive analysis
Semantic Technologies Partners:

Ontology Engineering
- TopQuadrant
- protégé
- Ontoprise

Reasoners
- clarkparsia, llc
- Ontoprise

NLP Entity Extractors
- EXPERT SYSTEM
- LYMBA

Open Source Frameworks
- jena
- openRDF.org
- Sesame
- Joseki

Standards
- OGC
- RDF
- Semantic Web

Applications
- PolarLake
- MedTrust

SI / Consulting
- infoMENTUM
- Raytheon
- NKA-DECKER

Oracle
Navigation and Visualization of RDF graphs

Open Source

• **Adapt** Cytoscape to work with very large RDF graphs

Commercial Products

• Tom Sawyer’s Perspective now supports RDF
Graph Creation and Manipulation

Open Source

• Protege ontology editor

Commercial Products

• Top Quadrant Composer
Reporting and Charting Tools for RDF

• Native RDF tools unavailable
  – for BI style reporting, charting and interactive refinement, eg. Oracle’s BI Dashboard, available today primarily for Relational and XML Database
RDF Tools for Exploring & Analyzing
Tools for guided pattern discovery & statistical analysis

### Supervised data mining, eg. Oracle Data Miner

<table>
<thead>
<tr>
<th>Problem Classification</th>
<th>Sample Problem</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Classification</strong></td>
<td>Given demographic data about a set of customers, predict customer response to an affinity card program</td>
</tr>
<tr>
<td><strong>Regression</strong></td>
<td>Given demographic and purchasing data about a set of customers, predict customers' age</td>
</tr>
<tr>
<td><strong>Attribute Importance</strong></td>
<td>Given customer response to an affinity card program, find the most significant predictors</td>
</tr>
</tbody>
</table>

### Statistical analysis, eg. Oracle R Enterprise

- Open source language and environment
- Statistical computing and graphics
- Easily produces publication-quality plots
- Highly extensible with open source R packages

---

*Copyright © 2012, Oracle and/or its affiliates. All rights reserved.*
Automated Discovery & Predictive Analysis
Oracle Data Miner

<table>
<thead>
<tr>
<th>Problem Classification</th>
<th>Sample Problem</th>
</tr>
</thead>
<tbody>
<tr>
<td>Anomaly Detection</td>
<td>Given demographic data about a set of customers, identify customer purchasing behavior that is significantly different from the norm</td>
</tr>
<tr>
<td>Association Rules</td>
<td>Find the items that tend to be purchased together and specify their relationship – market basket analysis</td>
</tr>
<tr>
<td>Clustering</td>
<td>Segment demographic data into clusters and rank the probability that an individual will belong to a given cluster</td>
</tr>
<tr>
<td>Feature Extraction</td>
<td>Given demographic data about a set of customers, group the attributes into general characteristics of the customers</td>
</tr>
</tbody>
</table>
Web Mapping with GeoSPARQL
Piping Big Data to RDF Analytics

MapReduce:
- Extract entities
- Annotate with RDF

Unstructured Documents

Load RDF triples

Oracle RDF

InfiniBand

Semantic Analytics

Stream | Acquire | Organize | Analyze & Visualize
How to Integrate RDF with Enterprise Reporting, Analysis and Discovery Tools

Zhe Wu
Consulting Member of Technical Staff
Oracle Corporation
Billions of Triples, Terabytes of Storage
Scalability is High, and Getting Better

- LUBM 25K tested
  - 3.4 Billion triples in the model
  - 2.7 Billion triples inferred
  - Storage space including indexes: over 1 Terabytes
  - Load speed: 109 minutes on Sun M8000
  - Inference speed: 160 minutes
  - Query throughput: 0.5 Billion in 9 minutes

- **Balanced hardware**, high parallelism for responsiveness
  - Sun M8000, 512GB RAM, 2TB Flash array, 128 Threads
Parallel Execution Performance on M8000

- LUBM 25K local inference on Sun M8000
- 6.1B+ quads (3.4B asserted, 2.7B inferred)

Oracle’s Parallel Execution is completely transparent!
- Cross CPUs/Cores on a single node
- Cross multiple nodes in a cluster
Inference Performance on Exadata V2

<table>
<thead>
<tr>
<th>Data Set (# triples)</th>
<th>Triples Inferred</th>
<th>Time</th>
<th>Degrees of Parallelism</th>
</tr>
</thead>
<tbody>
<tr>
<td>LUBM 100K (13B)</td>
<td>5B</td>
<td>1h, 58'</td>
<td>DOP = 32</td>
</tr>
<tr>
<td>LUBM 25K (3.3B)</td>
<td>2.7B</td>
<td>4h, 7'</td>
<td>DOP = 32</td>
</tr>
<tr>
<td>LUBM 8K (1.1B)</td>
<td>869M</td>
<td>46'</td>
<td>DOP = 64</td>
</tr>
</tbody>
</table>

1 Preliminary result: 1 round of OWLPrime (OWL Horst semantics)

2 Inference: OWLPrime + components: INTERSECT, INTERSECTSCOH, SVFH, THINGH, THINGSAM, UNION

Setup:

- **Hardware**: Full Rack Sun Oracle Database Machine X2-2 (8 nodes, 72GB RAM per node), and Exadata Storage Server

- **Storage required**: LUBM8K: 330GB or LUBM25K 1TB + 110GB temp table space

- **Software**: Oracle Database 11.2.0.1.0 + Patch 9819833: SEMANTIC TECHNOLOGIES 11G R2 FIX BUNDLE 2
  Each node: SGA_TARGET=32G and PGA_AGGREGATE_TARGET=31G
Query Performance on Exadata V2

Auto DOP used. 465,849,803 answers generated for LUBM 25K in 274.2 sec.

<table>
<thead>
<tr>
<th>Ontology</th>
<th>LUBM 25K</th>
<th>3.3 billion triples &amp; 2.7 billion inferred</th>
</tr>
</thead>
<tbody>
<tr>
<td>Query</td>
<td>Q1</td>
<td>Q2</td>
</tr>
<tr>
<td># answers</td>
<td>4</td>
<td>2528</td>
</tr>
<tr>
<td>Complete?</td>
<td>Y</td>
<td>Y</td>
</tr>
<tr>
<td>Time (sec)</td>
<td>0.01</td>
<td>20.65</td>
</tr>
</tbody>
</table>

OWLPrime & new inference components

<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>
A Strategy to Introduce Analytic Tools

• Great need for RDF analytic tools
  – But the field of analytic tools for RDF databases is a barren one

• In contrast, the mature field of relational & XML analytics is abundant
  – It is still a major undertaking to add native RDF/Sparql support to relational/XML analytic and mining tools

• Our strategy
  – Provide a simple way to make target RDF data available in XML or relational
Use BI Tool against Semantic Data

• Make the semantic data available to a BI tool in an appropriate format
  – Turn a semantic data store into yet another data source for BI tool
    • Logical inference can be used to “enrich” asserted facts
  – Relational and XML are popular formats

• Static versus dynamic data transformation
  – Static data transformation is acceptable if data seldom updates
  – Dynamic data transformation is crucial if data frequently updates
    • Eliminates synchronization
    • Maintains a single source of truth, better security
Use BI Tool: Semantic Data to Relational

- Data transformation: semantic data ➔ relational
  - Bindings from a SPARQL query can naturally be viewed as “columns”

```sparql
select ?agency_name ?label ?budget ?quarter
where {
  ?agency :hasQuarter ?quarter .
  ?agency :hasBudgetAmount ?budget .
  ?agency :hasAbbrev ?agency_name
}
```
Use BI Tool: Semantic Data to Relational

• Data transformation: semantic data → relational
  - `create view AGENCY_BUDGET as select * from` (table(sem_match(’
    select ?agency_name ?label ?budget ?quarter
    where {
      ?agency :hasQuarter ?quarter .
      ?agency :hasBudgetAmount ?budget .
      ?agency :hasAbbrev ?agency_name
    }’ ... )))

• Using materialized view is possible
• Add post-processing (e.g. un-escaping) of column values if necessary
Using BI Tool: Semantic Data to XML

- XML is the default format of SPARQL Query response from a web service endpoint. Transformation is necessary
  - BI Tool may require a different XML format
  - May need to remove namespaces, data type URIs, etc.

<table>
<thead>
<tr>
<th>SPARQL Query Response XML</th>
<th>OBIEE expected XML</th>
</tr>
</thead>
<tbody>
<tr>
<td>&lt;sparql xmlns=&quot;http://www.w3.org/2005/sparql-results#&quot;&gt;</td>
<td></td>
</tr>
<tr>
<td>&lt;head&gt;</td>
<td></td>
</tr>
<tr>
<td>&lt;variable name=&quot;agency_name&quot;/&gt;</td>
<td></td>
</tr>
<tr>
<td>&lt;results&gt;</td>
<td></td>
</tr>
<tr>
<td>&lt;result&gt;</td>
<td></td>
</tr>
<tr>
<td>&lt;binding name=&quot;agency_name&quot;&gt;</td>
<td></td>
</tr>
<tr>
<td>&lt;literal&gt;DEF. ADV. RESEARCH PROJ. &lt;/literal&gt;</td>
<td></td>
</tr>
<tr>
<td>&lt;/binding&gt;</td>
<td></td>
</tr>
<tr>
<td>&lt;/head&gt;</td>
<td></td>
</tr>
<tr>
<td>&lt;results&gt;</td>
<td></td>
</tr>
<tr>
<td>&lt;result&gt;</td>
<td></td>
</tr>
<tr>
<td>&lt;binding name=&quot;agency_name&quot;&gt;</td>
<td></td>
</tr>
<tr>
<td>&lt;literal&gt;DEF. ADV. RESEARCH PROJ. &lt;/literal&gt;</td>
<td></td>
</tr>
<tr>
<td>&lt;/binding&gt;</td>
<td></td>
</tr>
<tr>
<td>&lt;/sparql&gt;</td>
<td></td>
</tr>
</tbody>
</table>
Semantic Data to XML via SPARQL Gateway

• SPARQL Gateway is a feature of Jena Adapter
Using BI tool: Create Business Model

Relational, XML, … or RDF data sources
Use BI Tool against Semantic Data

• Tie it all together
  – Turn a semantic data store into yet another data source to BI
  – Perform conventional BI modeling
  – Define presentation layer, build report/dashboard
  – This is one example of what you may get:
Using Data Mining Tool against Semantic Data

• Make the semantic data available to a data mining tool in an appropriate format
  – Turn a semantic data store into yet another data source for DM tool

```sql
create view N_COUNTRY_BD_RATE as select name
    , to_number(brate) as brate
    , to_number(drate) as drate
    , to_number(popu) as population
    , to_number(mig) as net_migration_rate
    , to_number(imr) as infant_mortal_rate
    , to_number(leab) as life_expectancy
from table(sem_match('{
  }' ... ))
```
Using Data Mining Tool against Semantic Data

• Tie it all together
  – Turn a semantic data store into yet another data source to DM
  – Follow the conventional DM process:
    • Data preparation, build/evaluate model, deployment
    • This is one example of what you may get:
Using Data Mining Tool against Semantic Data

- Tie it all together
  - Turn a semantic data store into yet another data source to DM
  - Follow the conventional DM process:
    - Data preparation, build/evaluate model, deployment
    - Some Mining results can be saved back as RDF into Oracle database

---

Anomaly Detection output in SQL

Convert into RDF

:AbnormalCase1 :hasSubject :Dominica

:AbnormalCase1 :probability "0.54"
Using Oracle R Enterprise with Semantic Data

• Make the semantic data available to ORE in relational format
  – Turn a semantic data store into yet another data source for R tool

```sql
create view COUNTRY_INFO as select name,
  to_number(brate) as brate,
  to_number(drate) as drate,
  to_number(popu) as population,
  to_number(mig) as net_migration_rate,
  to_number(imr) as infant_mortality_rate,
  to_number(leab) as life_expectancy
from table(sem_match('{
}'))
```
Using Oracle R Enterprise with Semantic Data

• Tie it all together
  – Turn a semantic data store into yet another data source to ORE
  – Use R to analyze semantic data

• This is one example of what you may get:
Using Oracle R Enterprise with Semantic Data

- Tie it all together
  - Turn a semantic data store into yet another data source to ORE
  - Use R to analyze semantic data

- This is one example of what you may get:
Using Oracle R Enterprise with Semantic Data

• Tie it all together
  – Turn a semantic data store into yet another data source to ORE
  – Use R to analyze semantic data

• This is one example of what you may get:

```
qplot
```
Using Oracle R Enterprise with Semantic Data

• Tie it all together
  – Turn a semantic data store into yet another data source to ORE
  – Use R to analyze semantic data

• This is one example of what you may get:

```
library(ggplot2)

df <- read.csv('data.csv')

ggplot(df, aes(x = birth_rate, y = death_rate)) + geom_point()
```
Using Oracle R Enterprise with Semantic Data

- Tie it all together
  - Turn a semantic data store into yet another data source to ORE
  - Use R to analyze semantic data

- This is one example of what you may get:
Summary

• Oracle delivers enterprise-class semantic data management
  – Manage RDF data w/ Oracle’s scalability, performance, availability and triple-level label security
  – Incorporate popular open source technologies:
    • Jena, Sesame, Pellet, GATE, Cytoscape, Protégé
  – Exploit W3C and OGC standards
    • Query using SPARQL 1.1, SQL, GeoSPARQL
    • Inference in the database w/ OWL2 RL, EL+, SKOS rules
    • Develop in Java, PL/SQL
     – Semantically index documents and unstructured text
     – Use leading commercial tools: TopQuadrant, Tom Sawyer, Lymba…

• Oracle provides enterprise-class tools to help you mine insight from semantic data
  – OBIEE
  – Oracle Data Mining
  – Oracle R Enterprise
For More Information

Oracle RDF

Xavier.Lopez@oracle.com
Alan.Wu@oracle.com

oracle.com
Oracle **In-Database** Advanced Analytics

**Comprehensive Advanced Analytics Platform**

**Oracle R Enterprise**
- Popular open source statistical programming language & environment
- Integrated with database for scalability
- Wide range of statistical and advanced analytical functions
- R embedded in enterprise apps & OBIEE
- Exploratory data analysis
- Extensive graphics
- Open source R (CRAN) packages
- Integrated with Hadoop for HPC

**Oracle Data Mining**
- Automated knowledge discovery inside the Database
- 12 in-database data mining algorithms
- Text mining
- Predictive analytics applications development environment
- Star schema and transactional data mining
- Exadata "scoring" of ODM models
- SQL Developer/Oracle Data Miner GUI