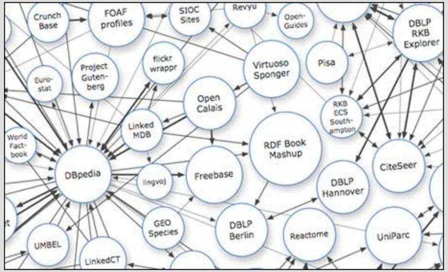
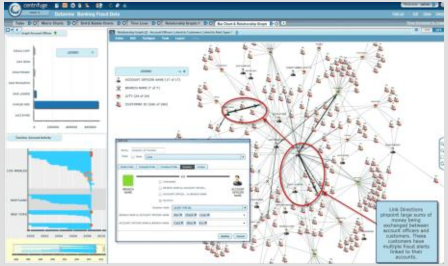
Graph Database Performance:
An Oracle Perspective

Xavier Lopez, Ph.D.

Senior Director, Product Management





Program Agenda

- Broad Perspective on Performance
- Graph Technology Enhancements at Oracle
- Performance: Database 11g
- Concluding Topics / Discussion

ORACLE

A Broad Perspective on Performance & Features

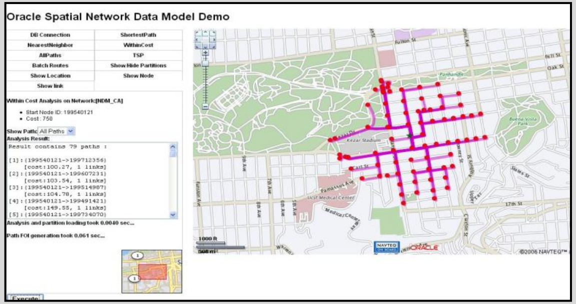
- Hardware:
  - Microprocessor-specific graph optimizations
  - Disc based storage
- Database:
  - RDF and NDM graph models, SPARQL language, optimizer, query engine, text search…
- Big Data Appliance:
  - RDF for NoSQL; HBase
- Middleware:
  - Jena,Sesame adapters; Protégé plug-in, Cytoscape plug-in, graph API
- Tools / Applications:
  - Oracle Business Intelligence, BPMN

ORACLE

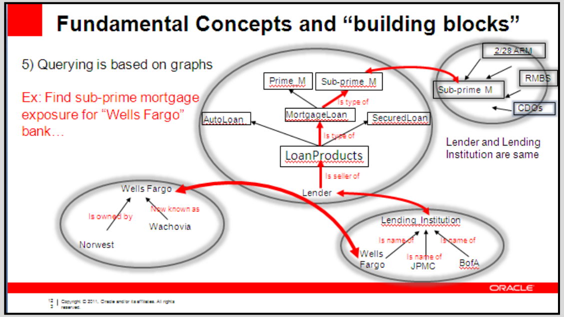
Oracle Spatial and Graph option

Two Graph Data Models

- Network Data Model graph
  - Manages logical / spatial networks in database
  - Persists link/node structure, connectivity and direction
  - Supports constraints at link and node level
  - Logically partitioning network graphs for scalability



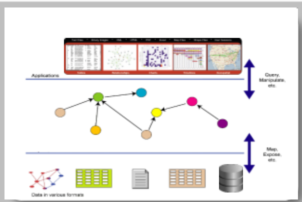
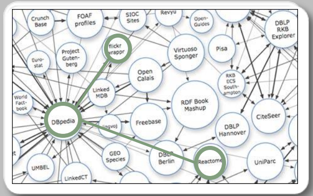
- RDF Semantic graph (triple store)
  - Enterprise class RDF Graph Database
  - Scales to petabytes of triples – by exploiting Exadata, RAC, SQL*Loader , Parallelism, Label Security
  - W3C standards support
  - SQL, PL/SQL APIs and Java APIs (Jena/Sesame)



RDF Graph (Triple Store) Use Cases

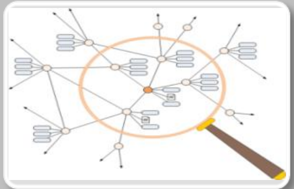
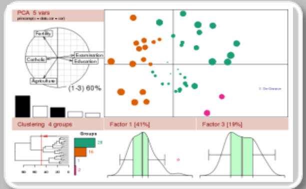
Semantic Metadata Layer

- Unified content metadata for federated resources
- Validate semantic and structural consistency

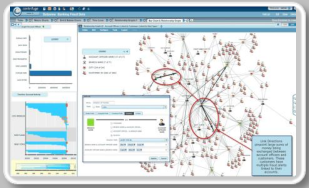
Text Mining & Entity Analytics

- Find related content & relations by navigating connected entities
- “Reason” across entities

Social Media Analysis

- Analyze social relations using curated metadata
  - Blogs, wikis, video
  - Calendars, IM, voice

 |

ORACLE

Metadata driving Federation & Integration

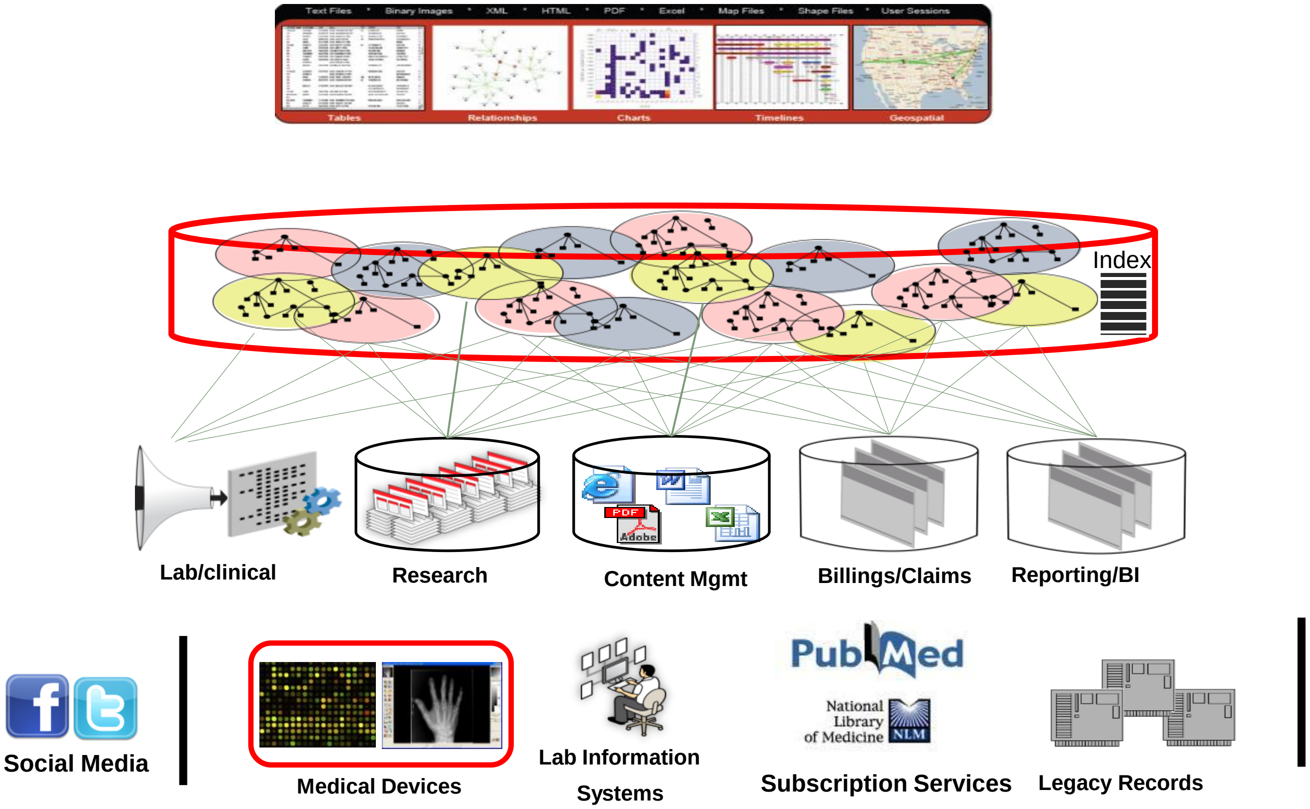
Domain applications

Domain Vocabularies

(integrated graph metadata)

Data Servers

Data Sources / Data Types



ORACLE

Industries Have Already Adopted the Concept

Industries

- Life Sciences
- Finance
- Media / Publishing
- Networks & Communications
- Defense & Intelligence
- Public Sector

Swiss Institute of Bioinformatics

Hutchinson 3G Austria

Thomson Reuters

RDF Semantic Graph Technologies Partners:
Integrated Tools and Solution Providers

Ontology Engineering
TopQuadrant
protégé
ontoprise — know how to use Know-how

Reasoners
clark&parsia, llc — thinking clearly
TrOWL
ontoprise — know how to use Know-how

NLP Entity Extractors
LYMBA — the power to answer
EXPERT SYSTEM — SEMANTIC INTELLIGENCE
GATE — general architecture for text engineering
CALAIS — Powered by Thomson Reuters
SAIL LABS TECHNOLOGY

Open Source Frameworks
Jena — semantic web framework
Joseki
openRDF.org
Sesame

Standards
OGC® Technical Committee Member
RDF
W3C Semantic Web

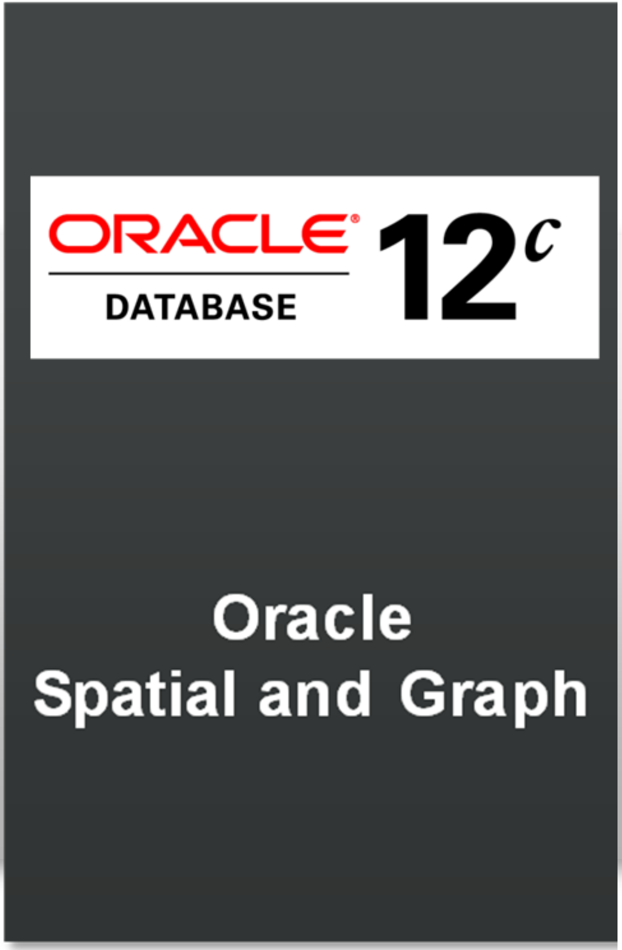
Applications & Tools
Bloomberg PolarLake
Callimachus Enterprise
IO INFORMATICS
MONDECA
Tom Sawyer® SOFTWARE

SI / Consulting
infoMENTUM
computas
NKA-DECKER
trivadis — makes IT easier.
arms Developments
SAIC® — From Science to Solutions
TenForce — The Pragmatic Company

ORACLE

 |

RDF DATABASE FEATURES



Oracle Database 11g RDF Triple Store

- Scalable to billions of triples
- RAC & Exadata scalability
- Compression & partitioning
- SQL*Loader direct path load
- Parallel load, inference, query
- High Availability
- Triple-level label security
- Choice of SPARQL or SQL
- Native inference engine
- Growing ecosystem of 3rd party tools

Key Capabilities:

Load / Storage
- Native RDF graph data store
- Manages billions of triples
- Optimized storage architecture

Query
- SPARQL-Jena/Joseki, Sesame
- SQL/graph query, b-tree indexing
- Ontology assisted SQL query

Reasoning
- RDFS, OWL2 RL, EL+, SKOS
- User-defined SWRL-like rules
- Incremental, parallel reasoning
- Plug-in architecture

"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."

New functions in Oracle Database Release 12.1

- Native SPARQL 1.1 query support
  - 40+ new query functions/operators: IF, COALESCE, STRBEFORE, REPLACE, ABS,
  - Aggregates: COUNT, SUM, MIN, MAX, AVG, GROUP_CONCAT, SAMPLE
  - Sub-queries
  - Value Assignment: BIND, GROUP BY Expressions, SELECT Expressions
  - Negation: NOT EXISTS, MINUS
  - Improved Path Searching with Property Paths

- GeoSPARQL Support
  - Leverages native spatial database feature in Oracle
  - Provide foundation for qualitative spatial reasoning

New functions in Oracle Database Release 12.1

- RDF views on relational tables (through W3C RDB2RDF)
  - RDF views can be created on a set of relational tables and/or views
  - SPARQL queries access data from both a relational and RDF store
  - Allows filtering of data in a relational store based upon ontology
  - Support RDF view creation using
    - Direct Mapping: simple and straightforward to use
    - R2RML Mapping: customizations allowed

ORACLE

New functions in Oracle Database Release 12.1

- Inference
  - Native OWL 2 EL inference support
    - Useful for expressing large biomedical ontologies (SNOMED CT)
  - User defined inferencing
    - Allows generation of new RDF resources
    - Temporal reasoning, Spatial reasoning
  - Ladder Based Inference
    - Fine grained security for inference graph
  - Performance optimization for user defined rules
  - Integration with TrOWL*, an external OWL 2 reasoner
    - TrOWL is a transformation based, tractable reasoner for OWL 2
    - Pellet was supported in 11g

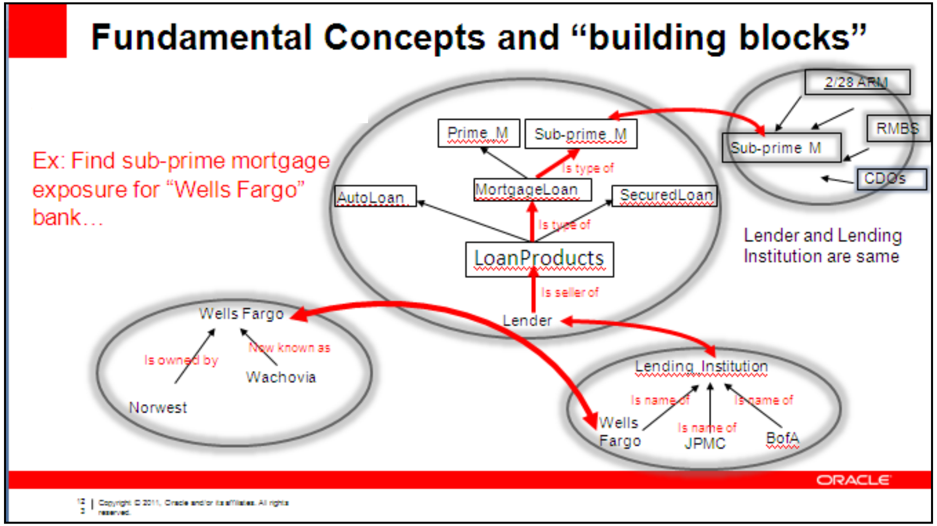
ORACLE

* http://trowl.eu/

RDF & SPARQL for Oracle NoSQL Database

RDF Graph Feature for NoSQL

- **RDF support in Oracle NoSQL Database Enterprise Edition**
- **High performance Key Value store**
- **Standard access to graph data: SPARQL 1.1**
- **Jena & Joseki SPARQL endpoint Web Services**
- **Massive horizontal scalability – petabytes of triples**
- **Support for World Wide Web Consortium (W3C) Semantic Web standards**

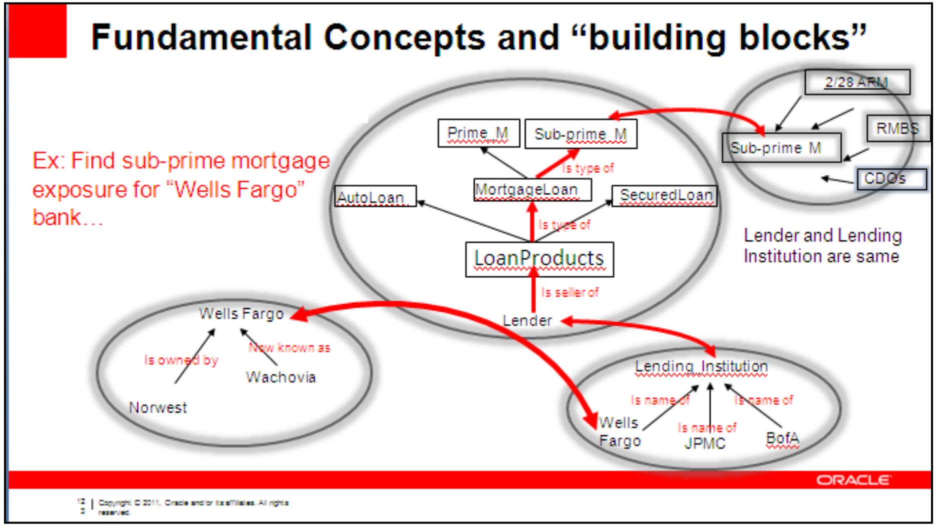


When to Consider a NoSQL Database

For horizontal scalability, lower query latency/cost, ease of install & management

RDF Graph Feature for NoSQL

- **Scale-out requirements**
- **High volume, simple queries**
- **Queries aggregating over most of the graph (e.g. what are the hobbies of the 100 most popular people in the network)**
- **Frequent, large-scale updates**
- **Open Linked Data applications**



LUBM PERFORMANCE ORACLE DATABASE 11G

- LOAD
- INFERENCE
- QUERY

Oracle Spatial and Graph - LUBM 200K on 3-Node RAC Sun Server X2-4

Load Performance

Data Set	Quads Loaded	Time	Degrees of Parallelism
LUBM200K Load into Staging Table: Load into the RDF graph:	 27.4 billion Quads (with duplicates) 26.6 billion Quads (unique quads)	 2 hrs 6 min. 22 hrs 23 min.	 DOP = 66 DOP = 80

- Data loading included de-duplication and building of two indexes on the quads. A significant portion (11 hrs 18 minutes) of the total load time was spent in building the two indexes.
- Loading from the 198 compressed N-Quad formatted files was done by defining an External Table (with *gunzip* preprocessor) on those files and then using sem_apis.LOAD_INTO_STAGING_TABLE
- Load flags => parse mbv_method=shadow parallel=80 parallel_create_index DEL_BATCH_DUPS=USE_INSERT

Setup:

Hardware: Sun Server X2-4, 3-node RAC

- Each node configured with 1TB RAM, 4 CPU 2.4GHz 10-Core Intel E7-4870)
- Storage: Dual Node 7420, both heads configured as: Sun ZFS Storage 7420 4 CPU 2.00GHz 8-Core (Intel E7-4820)
  256G Memory 4x SSD SATA2 512G (READZ) 2x SATA 500G 10K. Four disk trays with 20 x 900GB disks @10Krpm, 4x SSD 73GB (WRITEZ)

Software: Oracle Database 11.2.0.3.0, SGA_TARGET=750G and PGA_AGGREGATE_TARGET=200G

Note: Only one node in this RAC was used for performance test. Test performed in April 2013.

Oracle Spatial and Graph - LUBM 200K on 3-Node RAC Sun Server X2-4

Inference Performance

Data Set (# quads)	Quads Inferred	Time	Degrees of Parallelism
LUBM 200K (27.4B)	21.4 billion	17 hrs 56 min.	DOP = 80

Inference included building 2 indexes on the inferred triples that took a little over 5 hrs.

Inference Semantics: OWLPrime + the following components:

INTERSECT, INTERSECTSCOH, SVFH, THINGH, THINGSAM, UNION

Inference Options: RAW8=T, Dynamic Sampling level 1

Setup:

Hardware: Sun Server X2-4, 3-node RAC

- Each node configured with 1TB RAM, 4 CPU 2.4GHz 10-Core Intel E7-4870)
- Storage: Dual Node 7420, both heads configured as: Sun ZFS Storage 7420 4 CPU 2.00GHz 8-Core (Intel E7-4820) 256G Memory 4x SSD SATA2 512G (READZ) 2x SATA 500G 10K. Four disk trays with 20 x 900GB disks @10Krpm, 4x SSD 73GB (WRITEZ)

Software: Oracle Database 11.2.0.3.0, SGA_TARGET=850G and PGA_AGGREGATE_TARGET=150G

Note: Only one node in this RAC was used for performance test. Test performed in April 2013.

Oracle Spatial and Graph - LUBM 200K on 3-Node RAC Sun Server X2-4
Query Performance

Ontology **LUBM 200K – 48B quads** **27.4 billion asserted quads** **26.6 billion inferred quads**		**LUBM Benchmark Queries**						
OWLPrime & new inference components	**Query**	**Q1**	**Q2**	**Q3**	**Q4**	**Q5**	**Q6**	**Q7**
	# answers	4	494.5M	6	34	719	2.067B	67
	Time (sec)	0.01	1160	0.01	609.22	0.04	1105.07	712.48
	Query	**Q8**	**Q9**	**Q10**	**Q11**	**Q12**	**Q13**	**Q14**
	# answers	7790	53.86M	4	224	15	926088	1.568B
	Time (sec)	1228.95	3139.28	0.01	0.01	1.2	208.88	946.01

DOP = 40, Dynamic sampling level = 6. **4.18 Billion answers generated in 2.53 hrs on a single node.**

Setup:

Hardware: Sun Server X2-4, 3-node RAC

- Each node configured with 1TB RAM, 4 CPU 2.4GHz 10-Core Intel E7-4870)
- Storage: Dual Node 7420, both heads configured as: Sun ZFS Storage 7420 4 CPU 2.00GHz 8-Core (Intel E7-4820) 256G Memory 4x SSD SATA2 512G (READZ) 2x SATA 500G 10K. Four disk trays with 20 x 900GB disks @10Krpm, 4x SSD 73GB (WRITEZ)

Software: Oracle Database 11.2.0.3.0, SGA_TARGET=850G and PGA_AGGREGATE_TARGET=150G

Note: Only one node in this RAC was used for performance test. Test performed in April 2013.

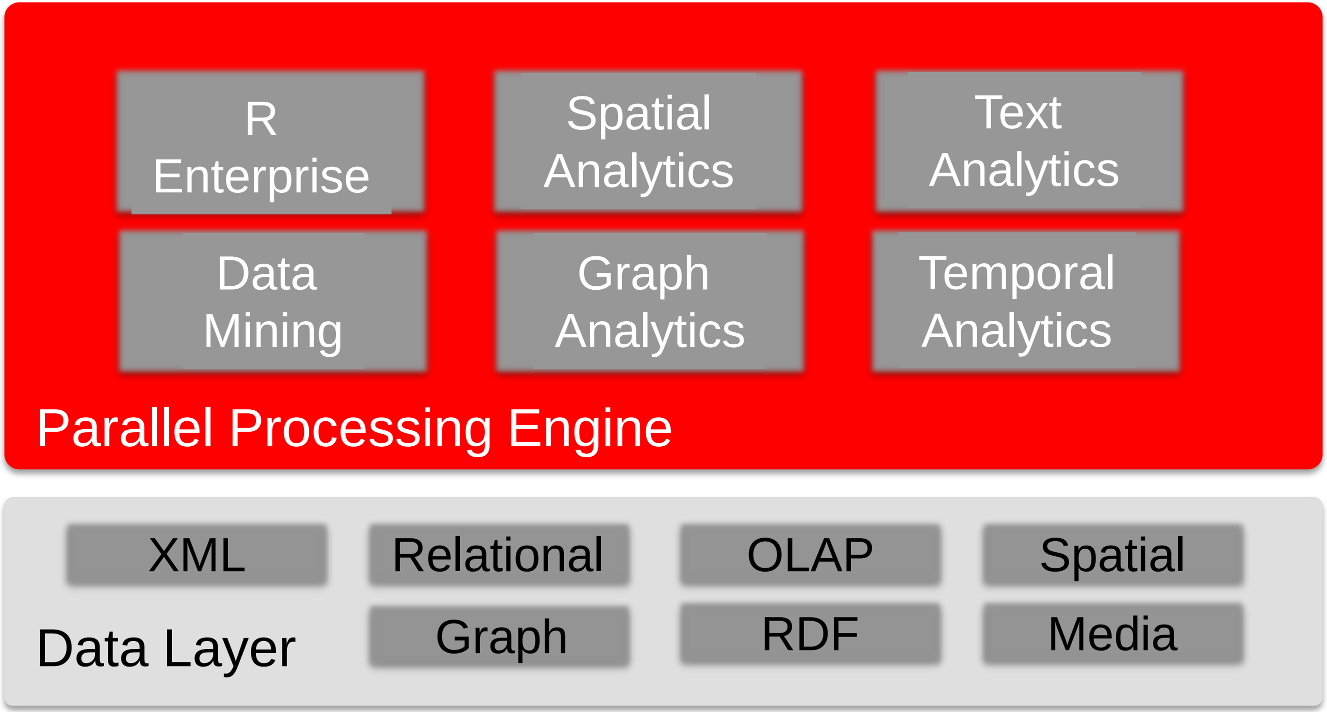
USING RDF GRAPHS FOR MINING SOCIAL MEDIA

ORACLE DATABASE 12c

Oracle
Spatial and Graph

ORACLE

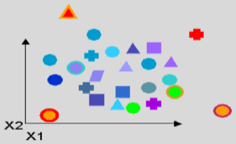
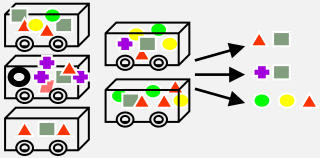
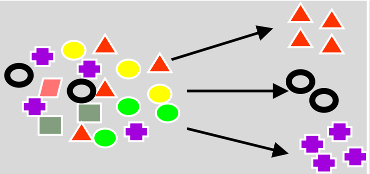
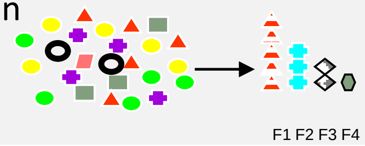
Oracle In-Database Analytics Platform



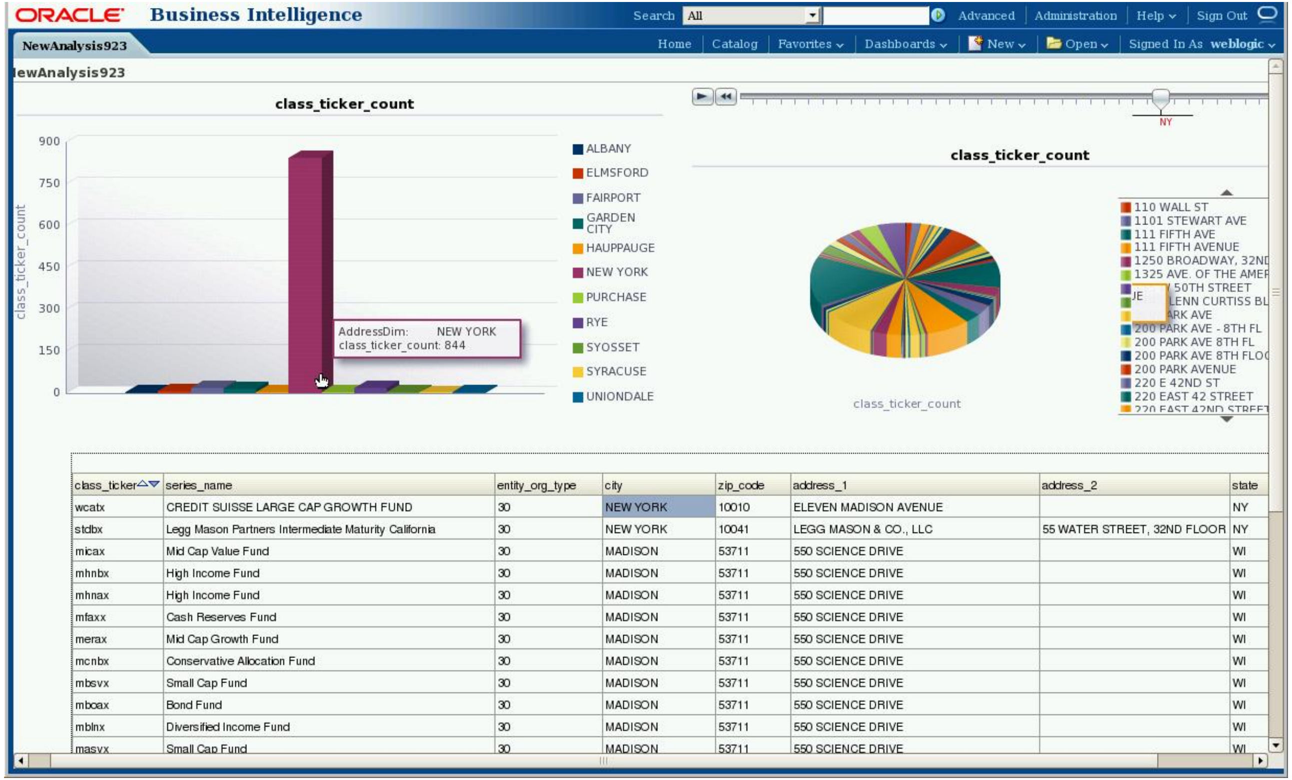
ORACLE

Tools: Discovery & Predictive Analysis

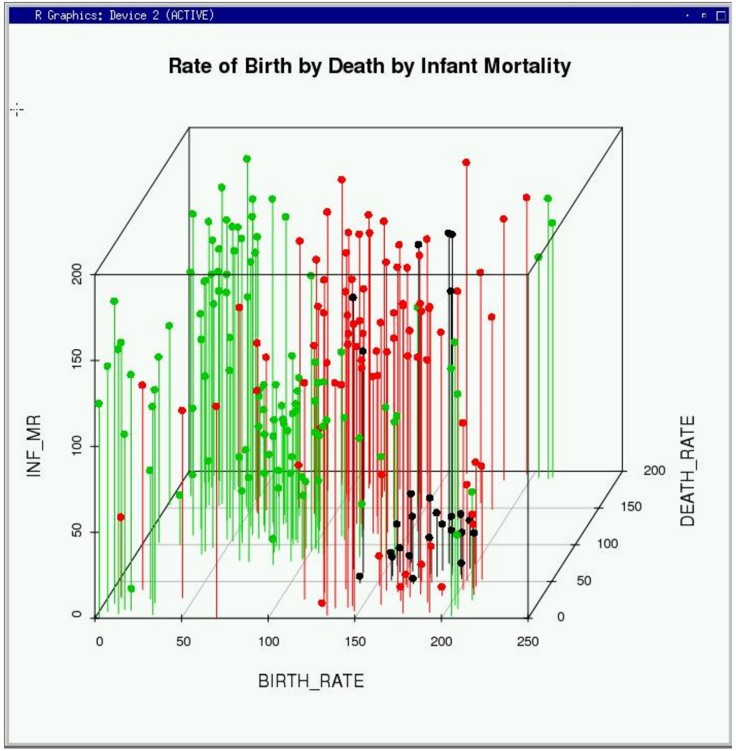
Oracle Data Mining

Problem Classification	Sample Problem
Anomaly Detection 	Given demographic data about a set of customers, identify customer purchasing behavior that is significantly different from the norm
Association Rules 	Find the items that tend to be purchased together and specify their relationship – market basket analysis
Clustering 	Segment demographic data into clusters and rank the probability that an individual will belong to a given cluster
Feature Extraction 	Given demographic data about a set of customers, group the attributes into general characteristics of the customers

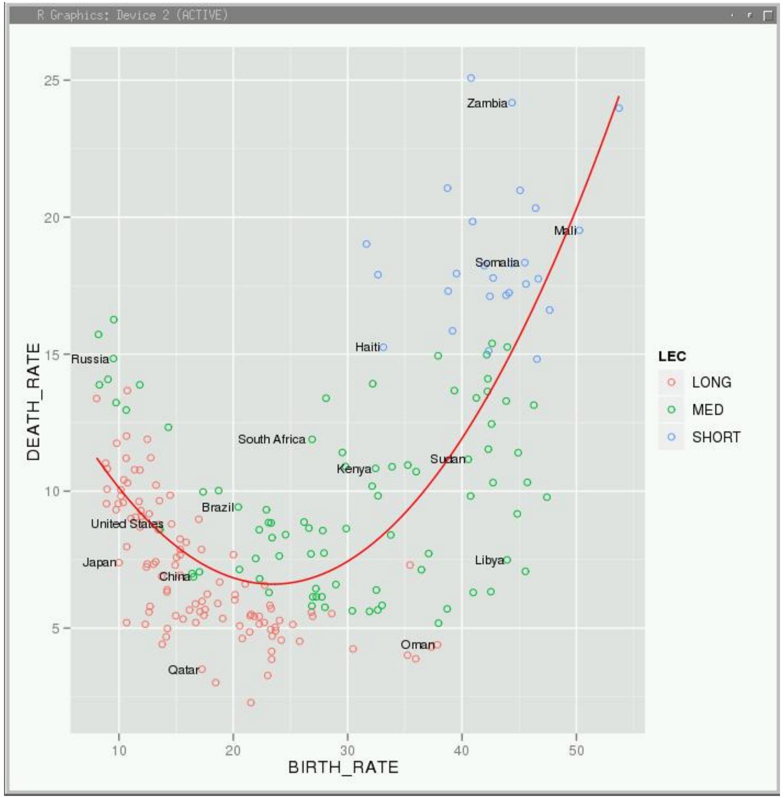
Finance Data: Visualizing RDF in OBIEE



Charting RDF data: Oracle R Graphics



Charting RDF data: Oracle R Graphics (2)



 |

CONCLUDING DISCUSSION TOPICS

Some topics to consider…

- Excellent work identifying customer RDF "pain-points"!!
  - Challenge: translating to repeatable database benchmarks
  - Pre-processing, loading, inferencing, querying
- Keep options open for <u>explanatory benchmarks</u>
  - Hardware, database, middleware, applications
- Better definition of "<u>graph models"</u>
  - LDBC is evaluating "RDF" and "graph" models. Please define each carefully
  - Distinguishing the two graphs via best practices and use cases might be useful