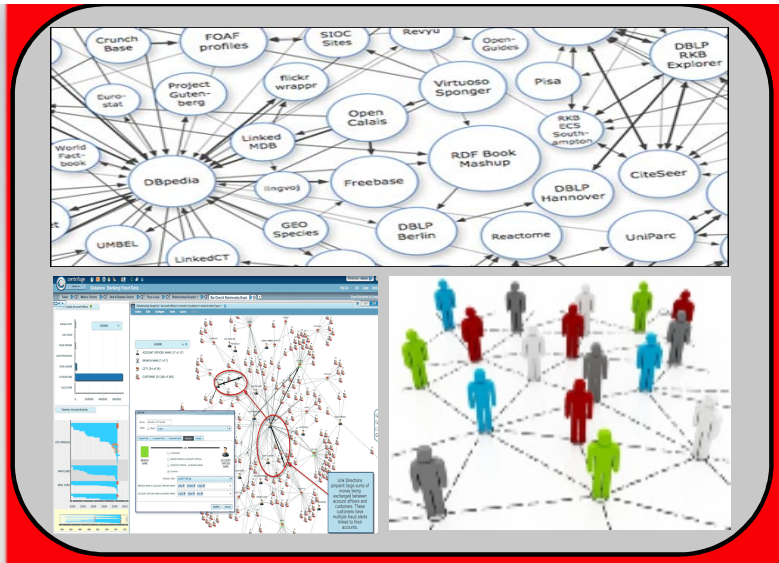
ORACLE

# Oracle Graph: Graph Features of Oracle Database 12c

**Zhe Wu**
alan.wu@oracle.com

Ph.D., Architect
Oracle Spatial & Graph

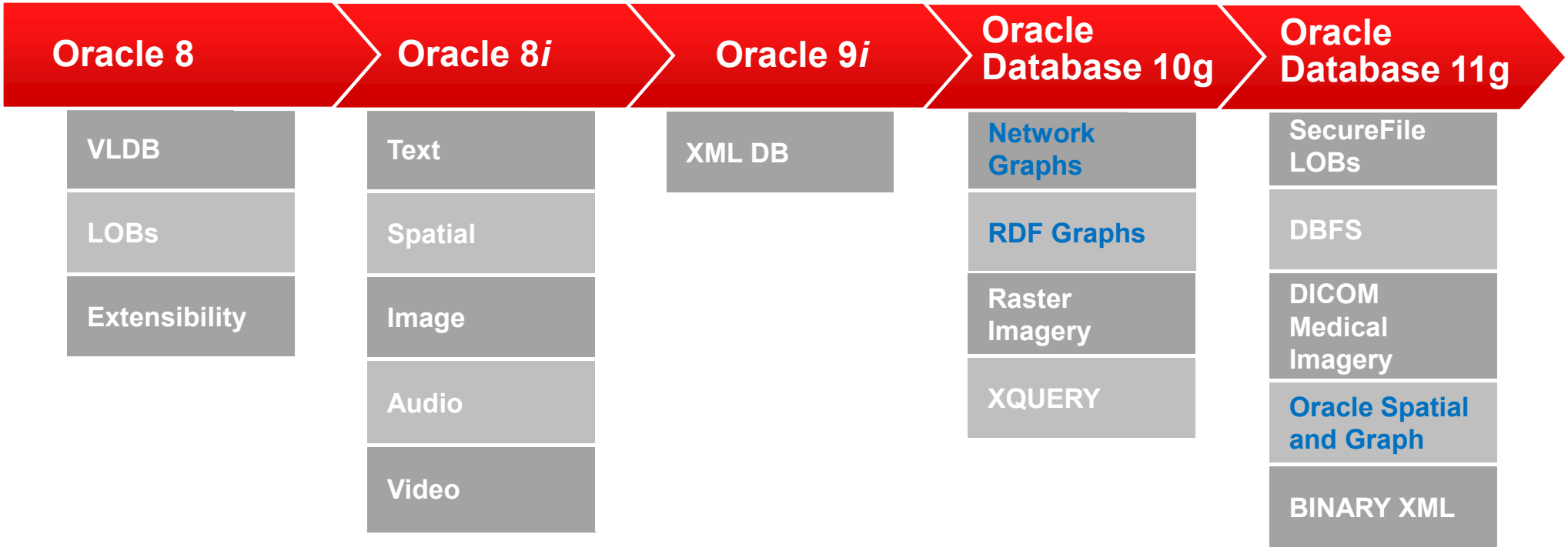
Feb, 2014



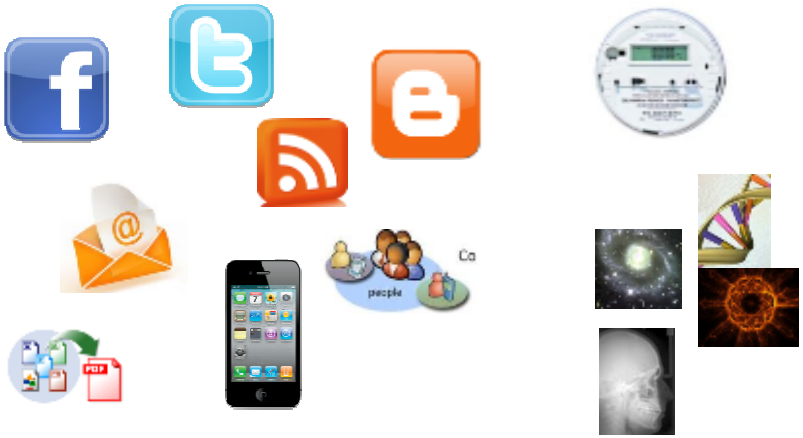

HARDWARE AND SOFTWARE ENGINEERED TO WORK TOGETHER

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 decisions.
The development, release, and timing of any features or functionality described for Oracle's products remains at the sole discretion of Oracle.

# Unstructured Data Support in Oracle Database



# Unstructured Data




- Device-generated data
- Documents
- Location data
- Audio, Video, Image
- Social Network and Interaction Models

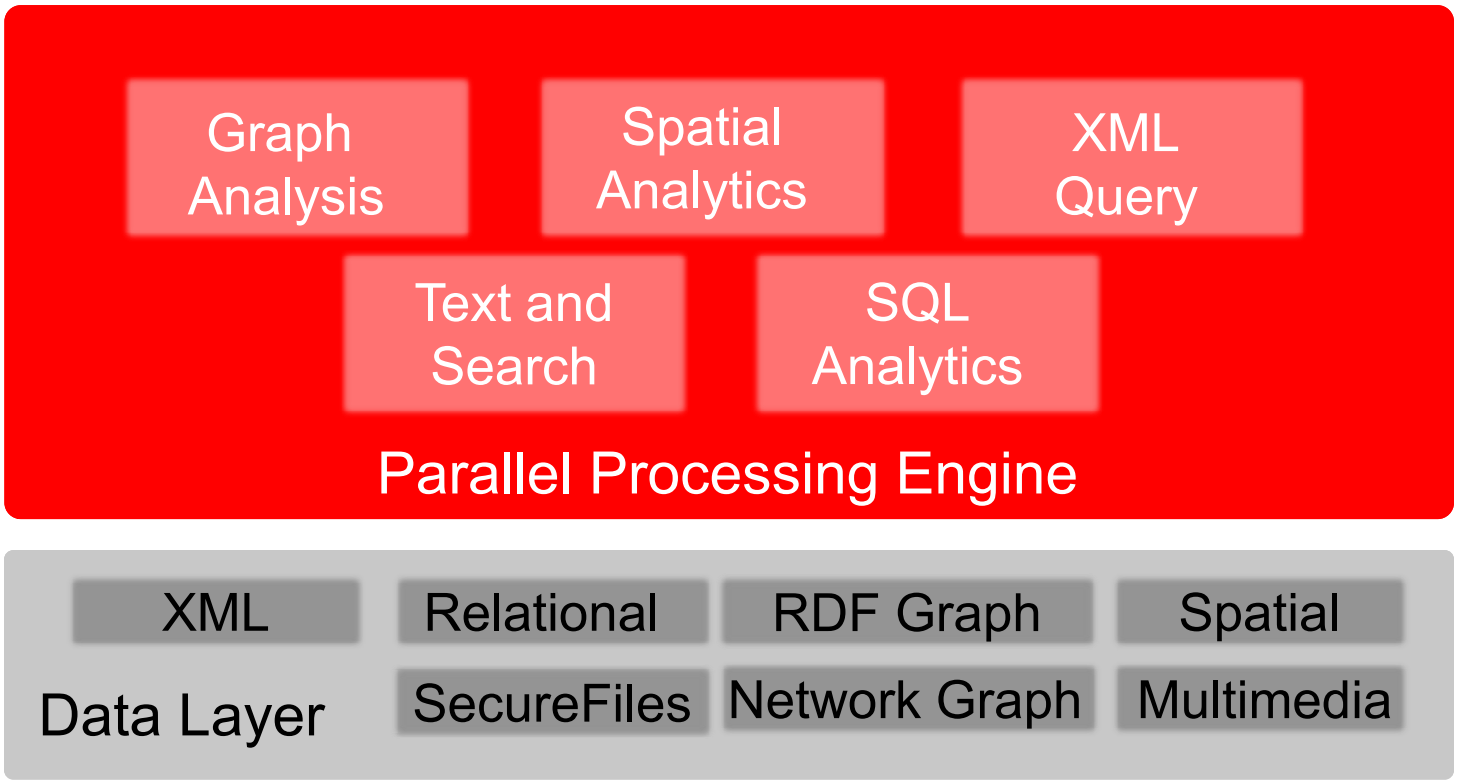
# Oracle Database Support for Unstructured Data

- Optimized Storage
- Specialized Data Types
- Administration & Management
- Indexing and Query
- Powerful Analytics

ORACLE®

DATABASE

# Oracle Database Support for Unstructured Data



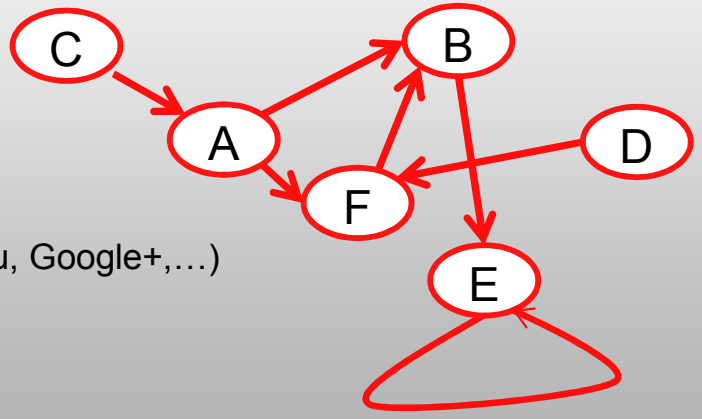
# Introduction to Graph Concept & Graph Data Models



7

# Overview of Graph

- What is a graph?
  - A set of vertexes and edges (and optionally attributes)
  - A graph is simply **linked data**
- Why do we care?
  - Graphs are everywhere
    - Road networks, power grids, biological networks
    - Social networks/Social Web (Facebook, Linkedin, Twitter, Baidu, Google+,…)
    - Knowledge graphs (RDF, OWL)
  - Graphs are intuitive and flexible
    - Easy to navigate, easy to form a path, natural to visualize
    - Do not require a predefined schema

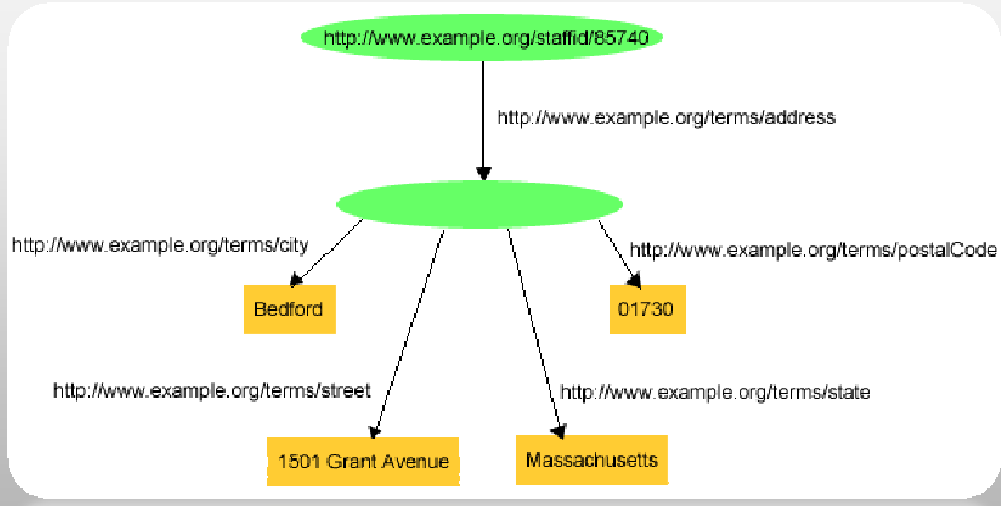


# Various Kinds of Graphs

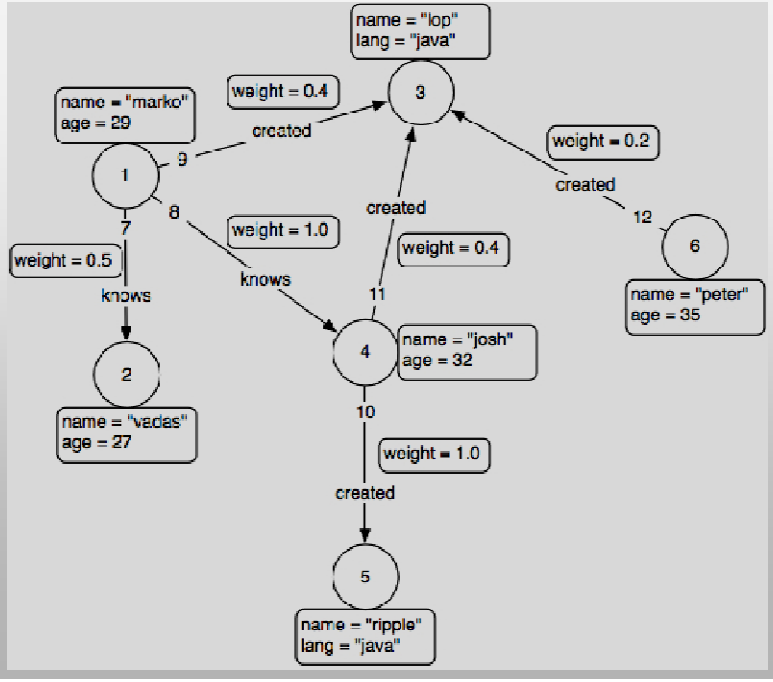
- There are more than one kind of graph
  - Simple graph, Weighted graph, Vertex-labeled graph, Edge-labeled graph, Directed graph (digraph), Undirected graph, Hypergraph, …
- Different application scenarios
  - Link-node graphs representing physical/logical networks used in transportation, utilities and telco
  - RDF Semantic Graphs modeling data as triples for social network, linked data and other semantic applications
  - Property Graphs allowing the association of K/V pairs with vertexes/edges

# Graph Data Model 1: RDF Semantic Graph

- Resource Description Framework
  - URIs are used to identify
    - Resources, entities, relationships, concepts
    - Data identification is a *must* for integration
- RDF Graph defines semantics
- Standards defined by W3C & OGC
  - RDF, RDFS, OWL, SKOS
  - SPARQL, RDFa, RDB2RDF, GeoSPARQL
- Implementations
  - Oracle, IBM, Cray, Bigdata ®
  - Franz, Ontotext, Openlink, Jena, Sesame, …



# Graph Data Model 2: Property Graph

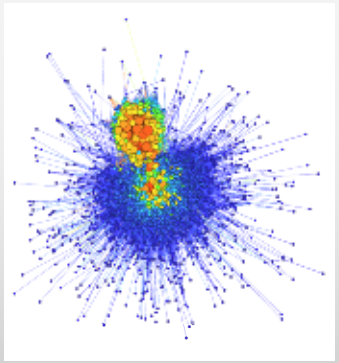


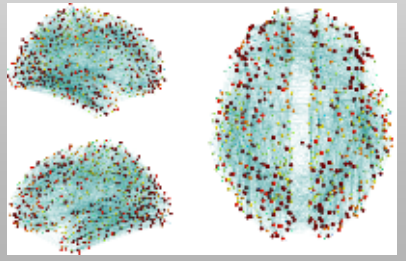

- A set of vertices (or nodes)
  - each vertex has a unique identifier.
  - each vertex has a set of in/out edges.
  - each vertex has a collection of **key-value** properties.
- A set of edges
  - each edge has a unique identifier.
  - each edge has a head/tail vertex.
  - each edge has a label denoting type of relationship between two vertices.
  - each edge has a collection of **key-value** properties.
- Blueprints Java APIs
- Implementations
  - Neo4j, Titan, InfiniteGraph, Dex, Sail, MongoDB …
- A property graph can be modeled as an RDF Graph

https://github.com/tinkerpop/blueprints/wiki/Property-Graph-Model

# Graphs Are Big
## and Are Getting Bigger

- Social Scale*
  - 1 Billion vertices, 100 billion edges
- Web Scale*
  - 50 billion vertices, 1 trillion edges
- Brain Scale*
  - 100 billion vertices, 100 trillion edges







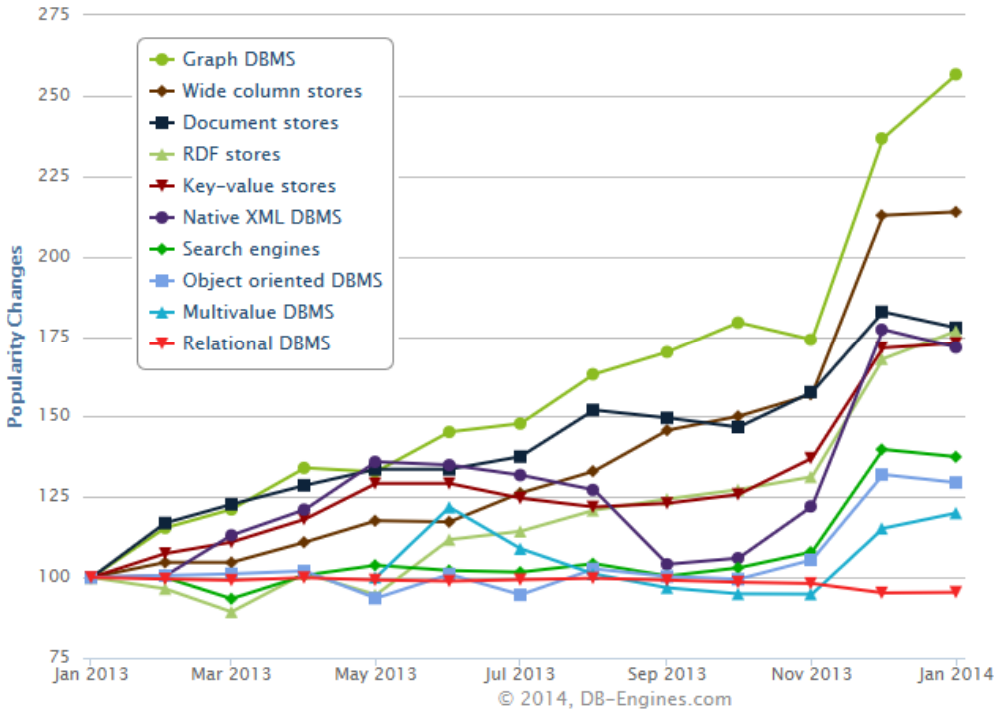

* An NSA Big Graph Experiment
http://www.pdl.cmu.edu/SDI/2013/slides/big_graph_nsa_rd_2013_56002v1.pdf

# Graph is an Emerging Market

Popularity based on Social Media mentions and sentiment analysis

## Graph Database Popularity

- Property Graph increased 250%
- RDF stores are tied at #3 with:
  - Document stores
  - Key value stores
  - Native XML stores

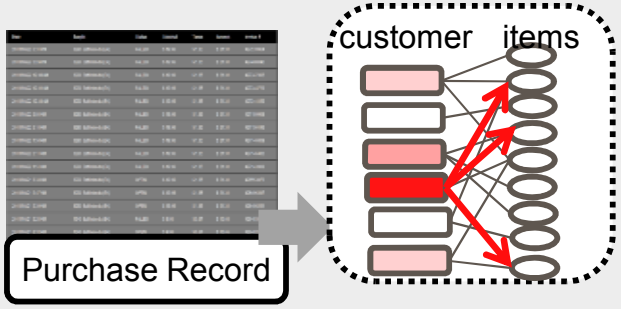



Source: DB-Engines.com, 2014

# Graph Analysis

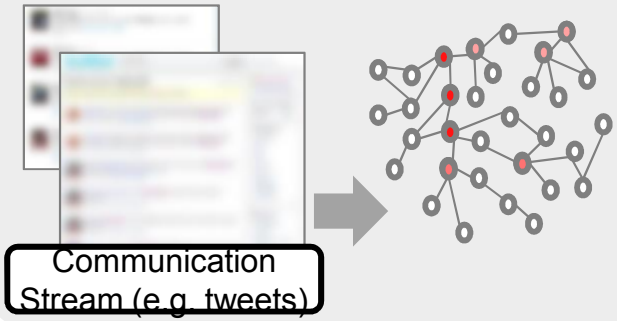
- Represent your data as a graph, analyze it and discover useful information
- By considering relationships between data entities

## Recommendation



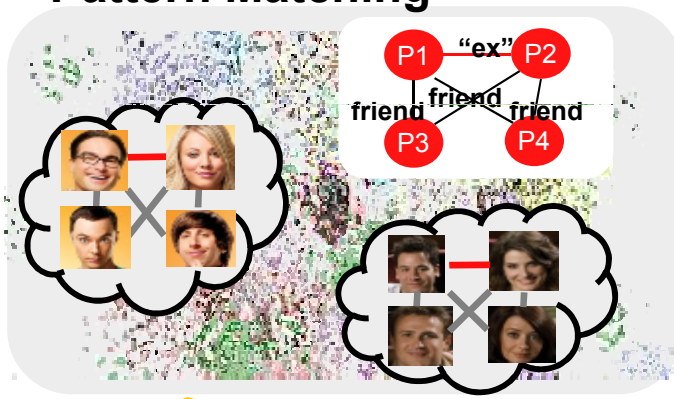

• Create recommendations for a specific customer from its *neighbors* in a purchase graph

## Influencer Identification




• Identify nodes that are critical to connectivity of an information flow graph

## Pattern Matching



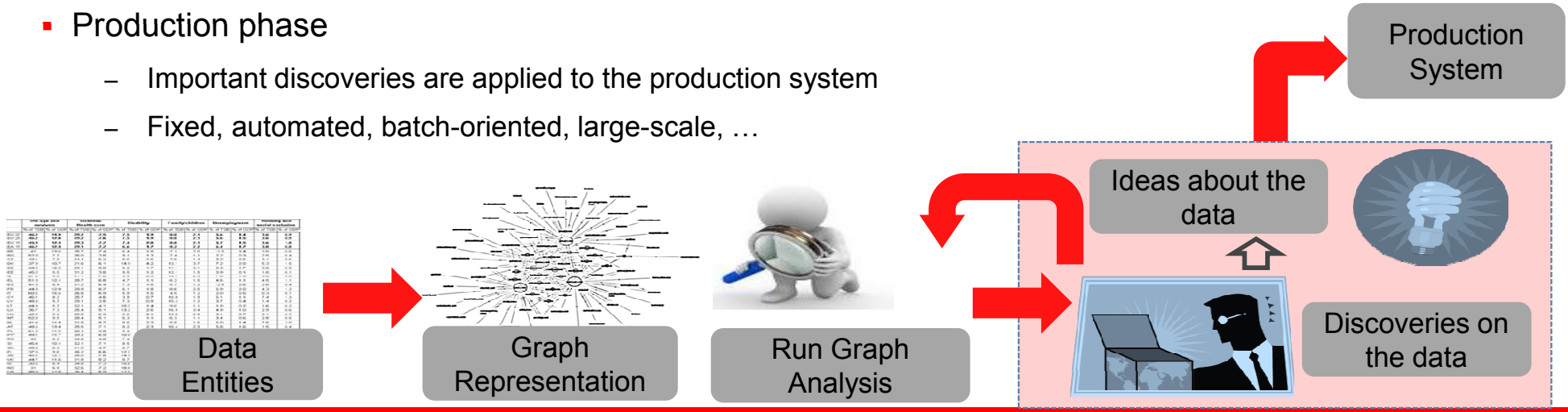

• Find sub-graphs that match a specified pattern

# Typical Graph Problems & Applications

- Data Integration
- Pattern matching
- Link analysis
- Recommender system
- Expert/Influencer Identification
- Clustering/Community Discovery
- Network bottleneck
- …

# Graph Analytics Workflow

- Basic flow
  - Represent raw data as graphs using a *graph data model (RDF or Property Graph)*
  - Run analysis algorithms on the graphs
- Exploration phase
  - Data scientists try different ideas (algorithms) on the data
  - Flexible, interactive, iterative, small-scale (sampled), ….
- Production phase
  - Important discoveries are applied to the production system
  - Fixed, automated, batch-oriented, large-scale, …



# Oracle Spatial and Graph

# Oracle Spatial and Graph

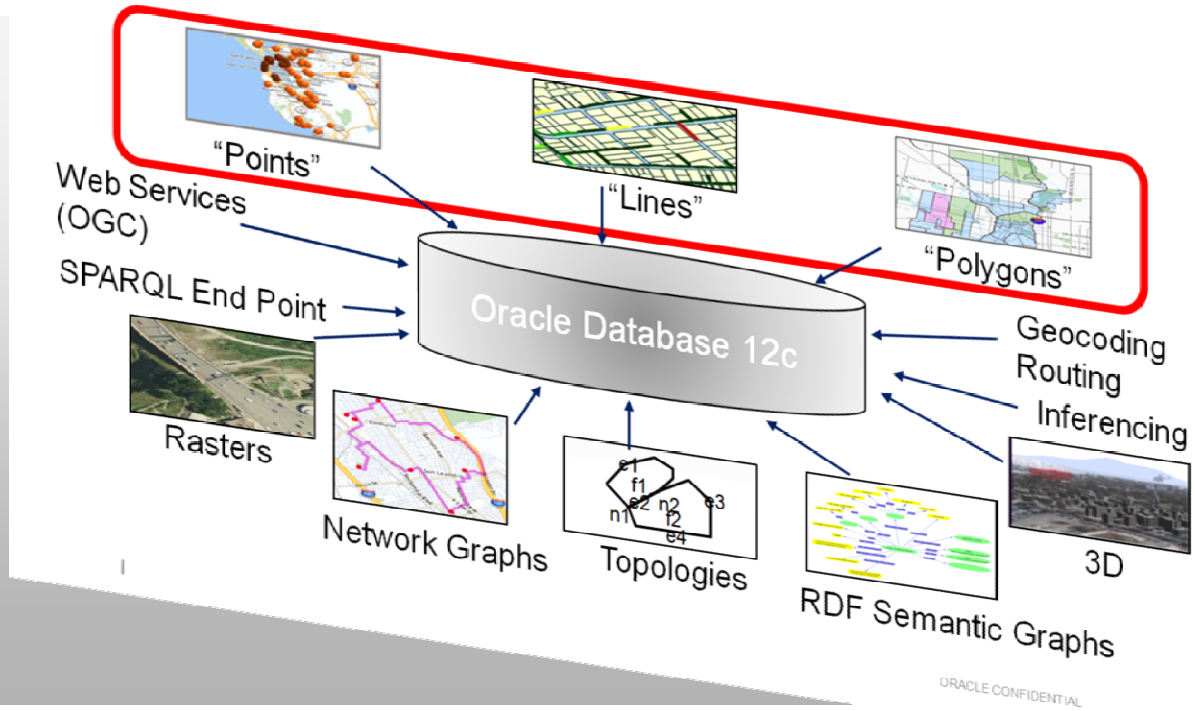
## In-Database Datatypes, Models and Analytics

**Complete**

**Open**

**Integrated**

**Most Widely Used**

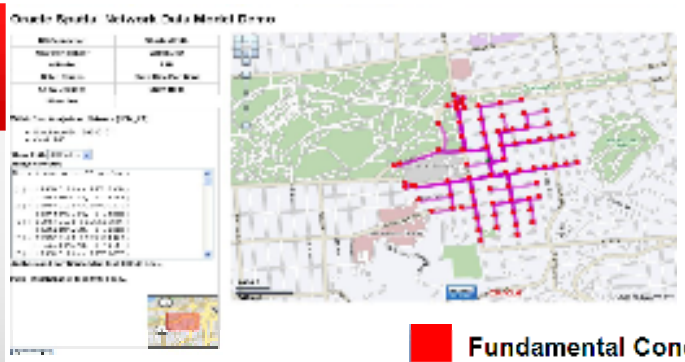


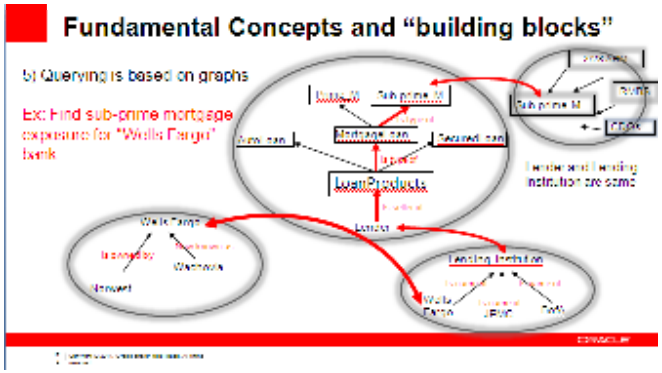
# Oracle Spatial and Graph

## Mature, Proven Graph Database Capabilities

### Graph Features

- Network Data Model graph
- W3C RDF Semantic graph

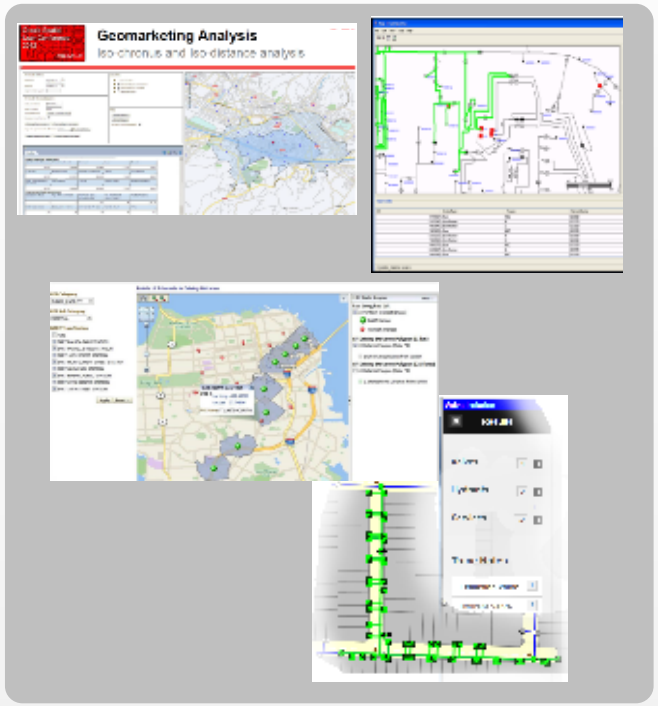




# Network Data Model Graph

## Use Cases


**Oracle Spatial and Graph**




- Transportation, Road and Multimodal Networks
- Drive Time Polygon Analysis
- Trade Area Management
- Service Delivery Optimization
- Water, Gas, Electric Utility, Network Applications

# Network Data Model Graph

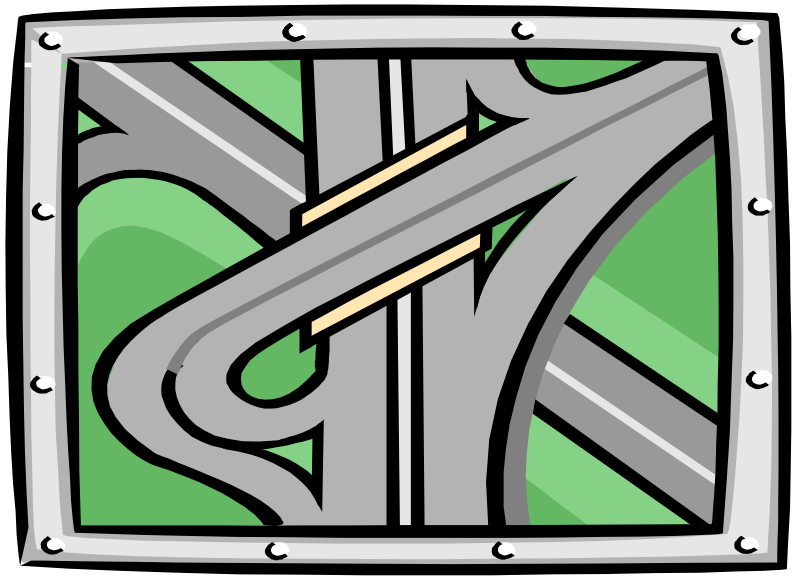
## Features

- A storage model to represent graphs and networks
- Graph tables consist of links and nodes
- Explicitly stores and maintains connectivity of the network graph
- Attributes at link and node level
- Logical or spatial graphs
- Can logically partition the network graph

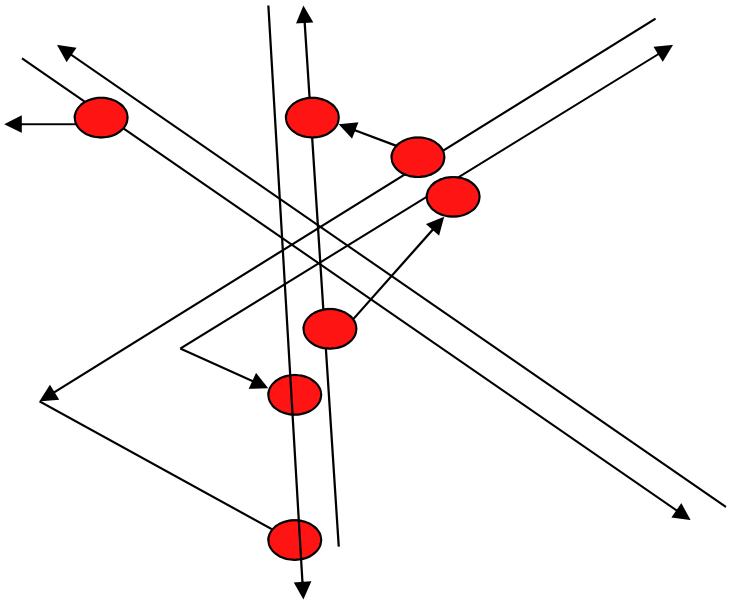
- Java API to perform Analysis in memory
- Loads and retains only the partitions needed
- Dynamic costs with real time input
- Shortest path, within cost, nearest neighbors
- Traveling salesman, spanning tree, ...
- Multiple Cost Support in Path/Subpath Analysis

# Real World Feature Modeling in NDM Graph

Feature Representation



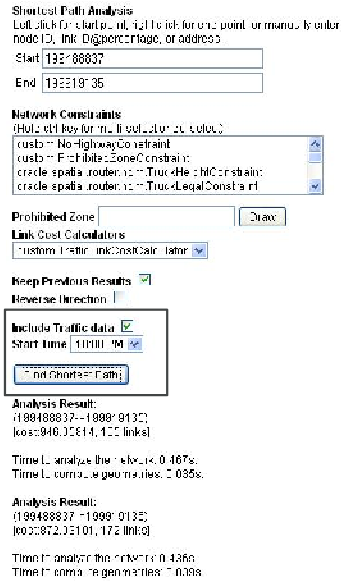
Network Representation

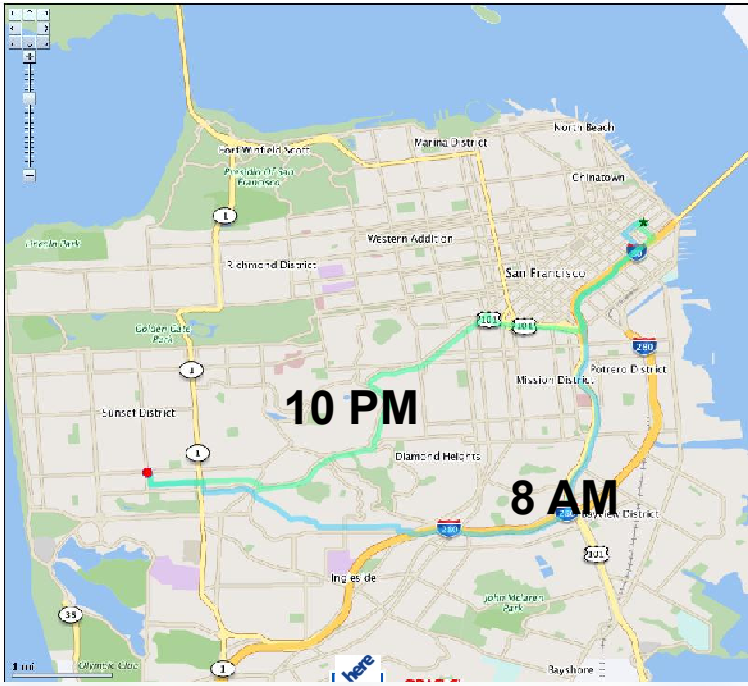


# Network Data Model Graph

## Temporal Modeling/Analysis

- Traffic Patterns
  - Record historical travel
  - Based on time of day and day of the week
- NDM can use traffic patterns to compute shortest paths
- Support Nokia/HERE Traffic Patterns format out of the box

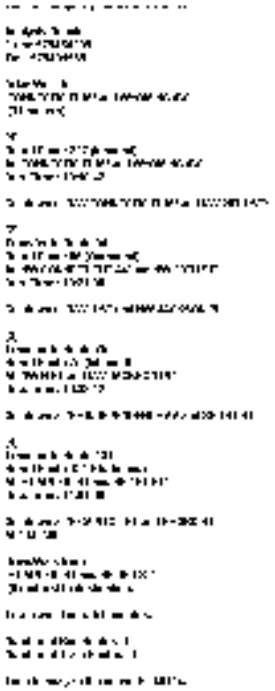


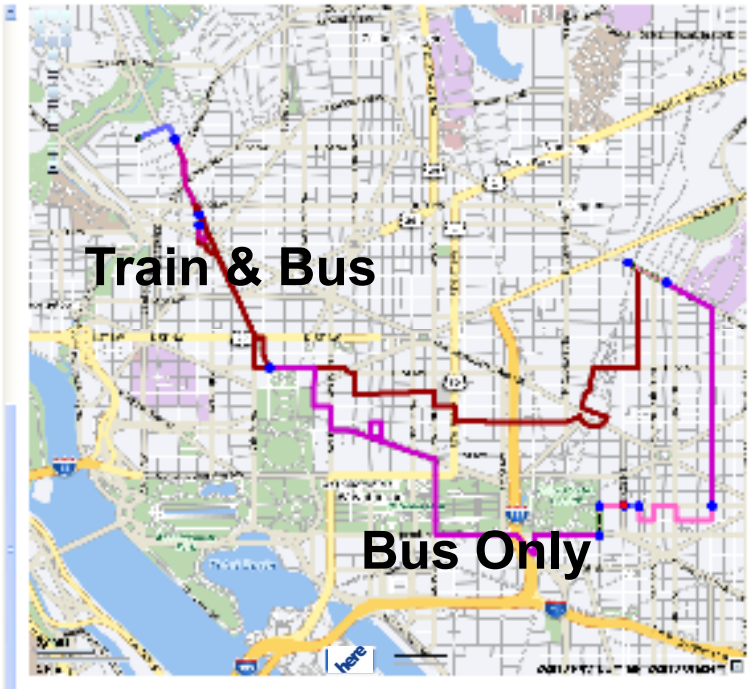



# Network Data Model Graph

## Multi-Modal Routing

- Each mode (car, bus, rail, bike, etc) modeled as a separate network
- Single logical network represents all modes of transportation
- Transition nodes where networks meet
- NDM APIs can specify the modes
- Out of the box support for transit data published by transit authorities





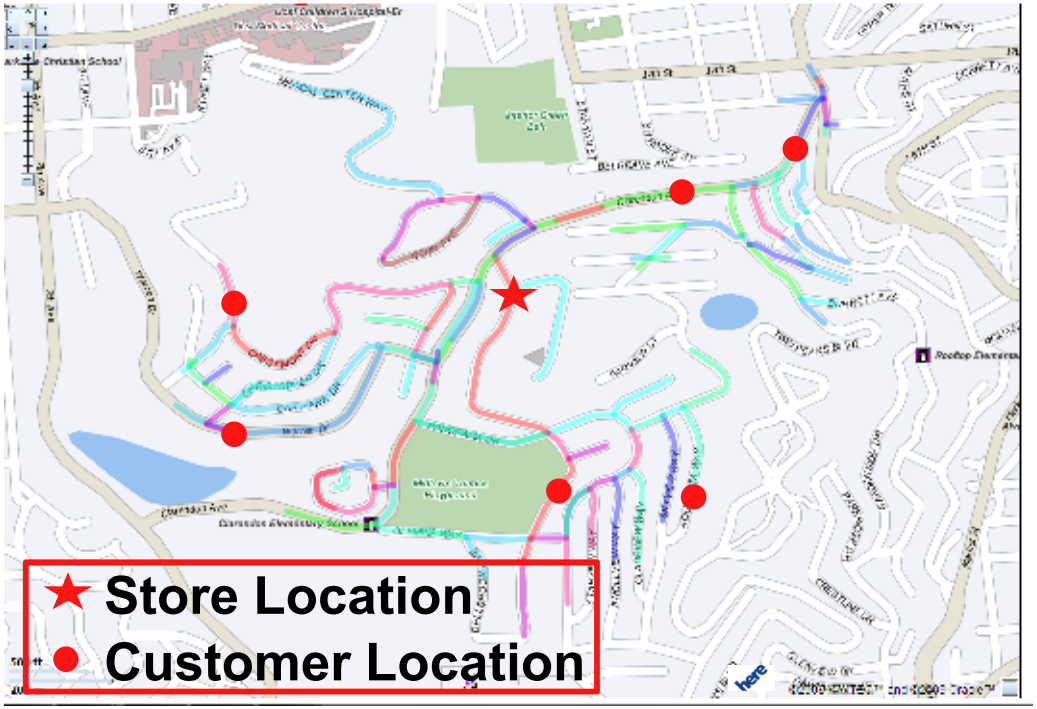

- http://www.oracle.com/technetwork/database/options/spatialandgraph/spatial-and-graph-wp-12c-1896143.pdf
- http://docs.oracle.com/cd/E16655_01/appdev.121/e17897/toc.htm

# Network Data Model Graph

## Large Scale Drive Time/Distance Analysis

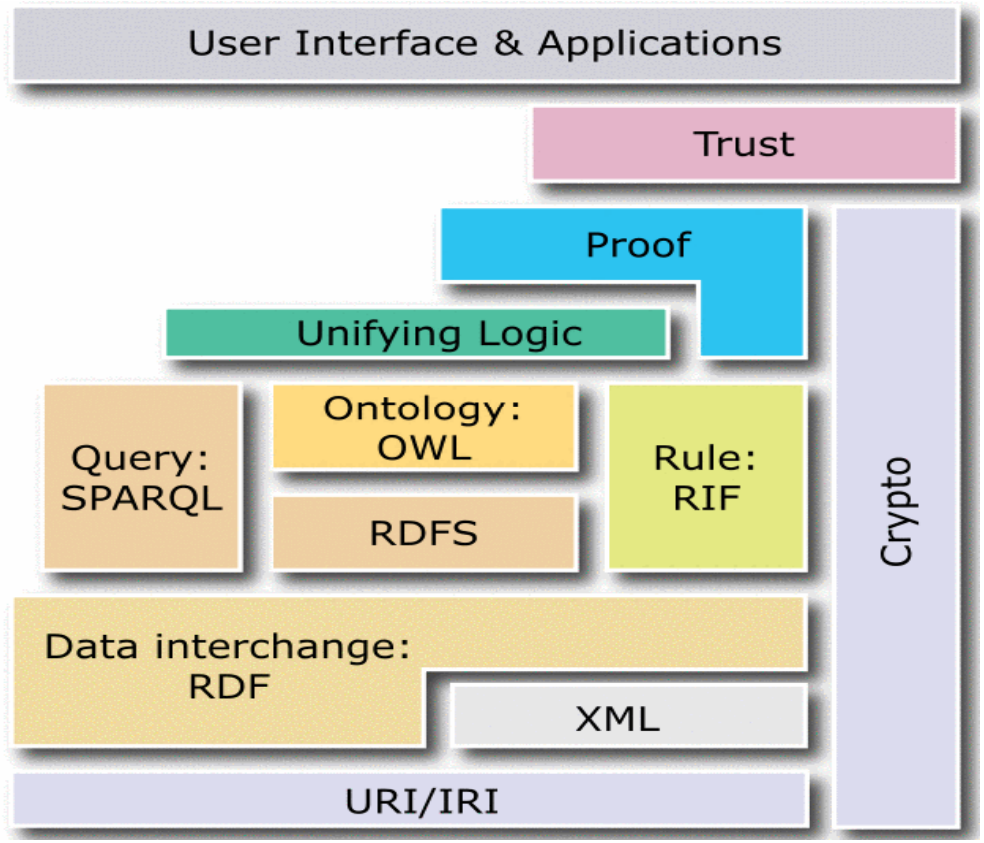
### Big Data Analysis

- Millions of customers, find closest store within a specified drive time
- Single database query to find closest store and drive time/distance for each customer
- Customers geocode as based on graph segment
- Network Buffer generates all possible paths



# W3C RDF Semantic Graph

# Semantic Technology Stack




- Core Technologies
  - *URI*
    - Uniform resource identifier
  - *RDF*
    - Resource description framework
  - *RDFS*
    - RDF Schema
  - *OWL*
    - Web ontology language

http://www.w3.org/2007/03/layerCake.svg

# What is RDF

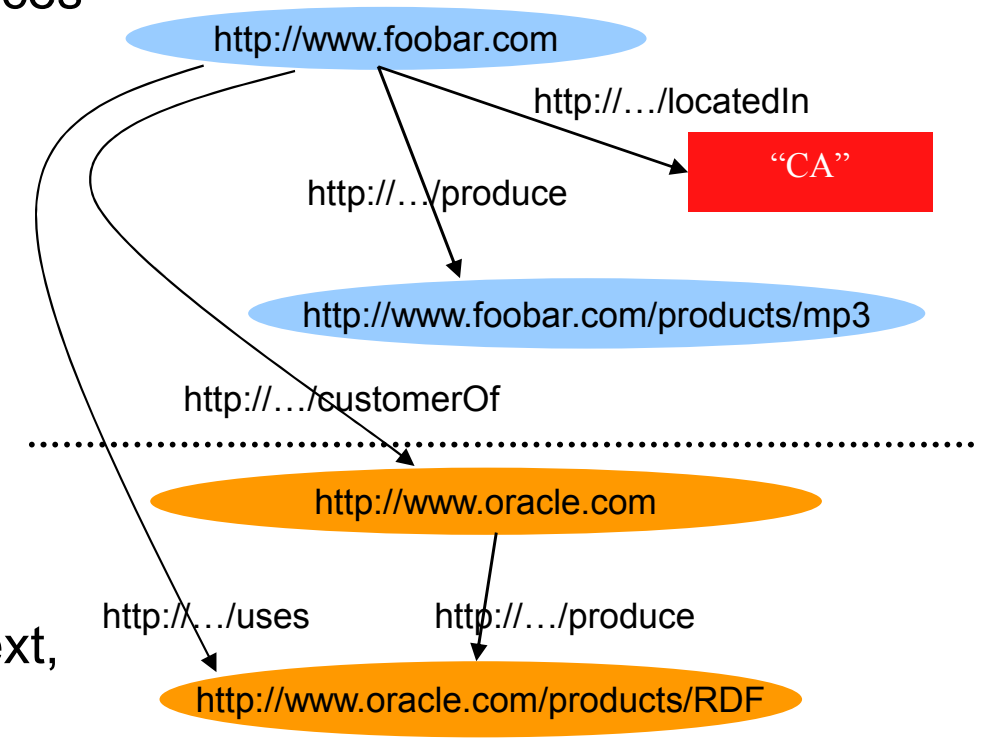
- A graph data model for web resources and their relationships
- The graph can be serialized into
  - RDF/XML, N3, N-TRIPLE, …
- Construction unit: **Triple**
  (or assertion, or fact)

<http://foobar> <:produces> <:mp3>

Subject Predicate Object

- Quads (**named graphs**) add context, provenance, identification, etc. to assertions

<http://foobar> <:produces> <:mp3 > **<:ProductGraph>**



# Basic Elements of RDF

- Instances

  E.g. :John, :MovieXYZ, :PurchaseOrder432

- Classes
  - Class represents a group/category/categorization of instances

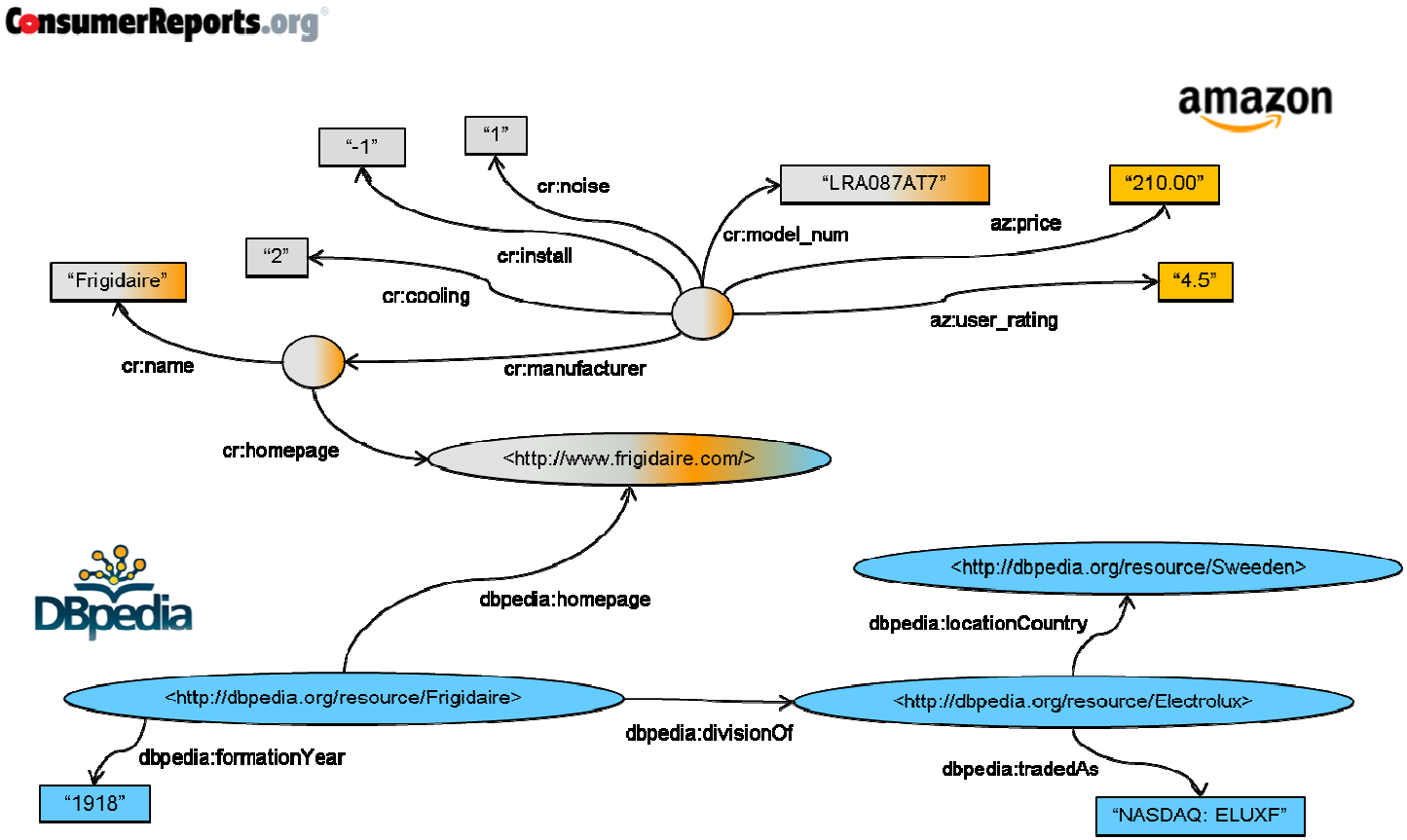
  E.g. :John rdf:type :Student

- Properties
  - Linking data together

  E.g. :John :brother :Mary,

  :John :hasAge “33”^^xsd:integer.

# A Small RDF Graph Linking Several Data Sources



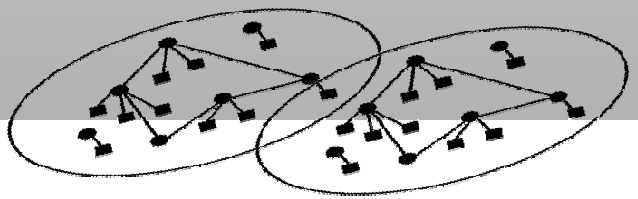
# Many Known Graphs and Vocabularies on the Web

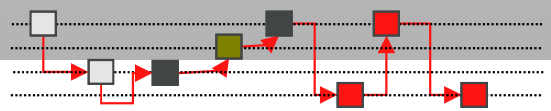
- DBPedia
- SIOC
- NCI
- SNOMED
- FOAF
- Geonames
- CIA World Fact Book
- DBLP
- UniProt
- UniParc
- CiteSeer
- Wordnet
- Drug Bank
- ACM
- Daily Med
- Linked CT
- Eurostat
- KEGG
- Data.gov.uk
- Music Brainz Data
- Semantic Tweet
- CO2 Emission
- Semanitc XBRL
- US Census
- YAGO
- Cyc/Open Cyc
- PubMed
- Freebase
- Gene Ontology
- UniRef
- Smart Link
- Reactome
- Diseasome

**And so much more !**

# RDF Graph Can Enrich Your Business Applications

- Flexible graph modeling adds agility
- Resource identification adds precision
- Integrate full breadth of enterprise content (structured, spatial, email, documents, web services)
- Reconcile differences in data semantics so that they can all “talk” and interoperate;
- Resolve semantic discrepancies across databases, applications
- Create consolidated “single” views across business applications
- Model and implement common Business Processes



ORACLE

# Oracle Spatial and Graph: Enterprise Graph Capability

**RDF Semantic Graph works well with these great technologies**

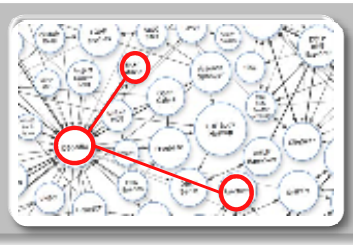
- **Relational**, XML, Spatial, Text, Security, Clustering, Compression, Data Guard …
  - Oracle's RDF/OWL support is ***native*** to the Database
- Web Services, SOA, BPMN, …
  - Support of popular Java APIs and ***standard*** complaint Web Service endpoint
- Advanced Analytics
  - Support integration with OBIEE, Oracle Data Mining, Oracle R Enterprise
- A rich set of third party tools including
  - Ontology editing, knowledge management, Complete DL reasoners
  - Graph/network visualization
  - NLP, text processing
  - …

# RDF Semantic Graph

## Use Cases

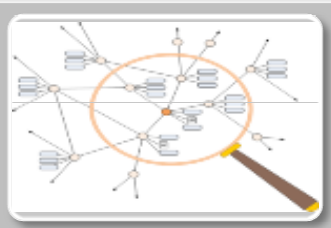
**Linked Data & Public Clouds**

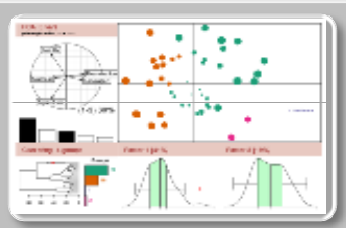
- Unified content metadata model for public clouds
- Validate semantic and structural consistency

**Text Mining & Entity Analytics**

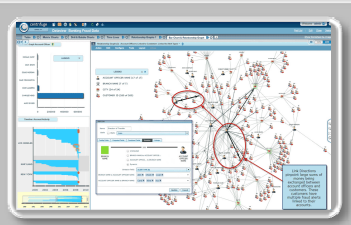
- Find related content & relations by navigating connected entities
- "Reason" across entities





**Social Media Analysis**

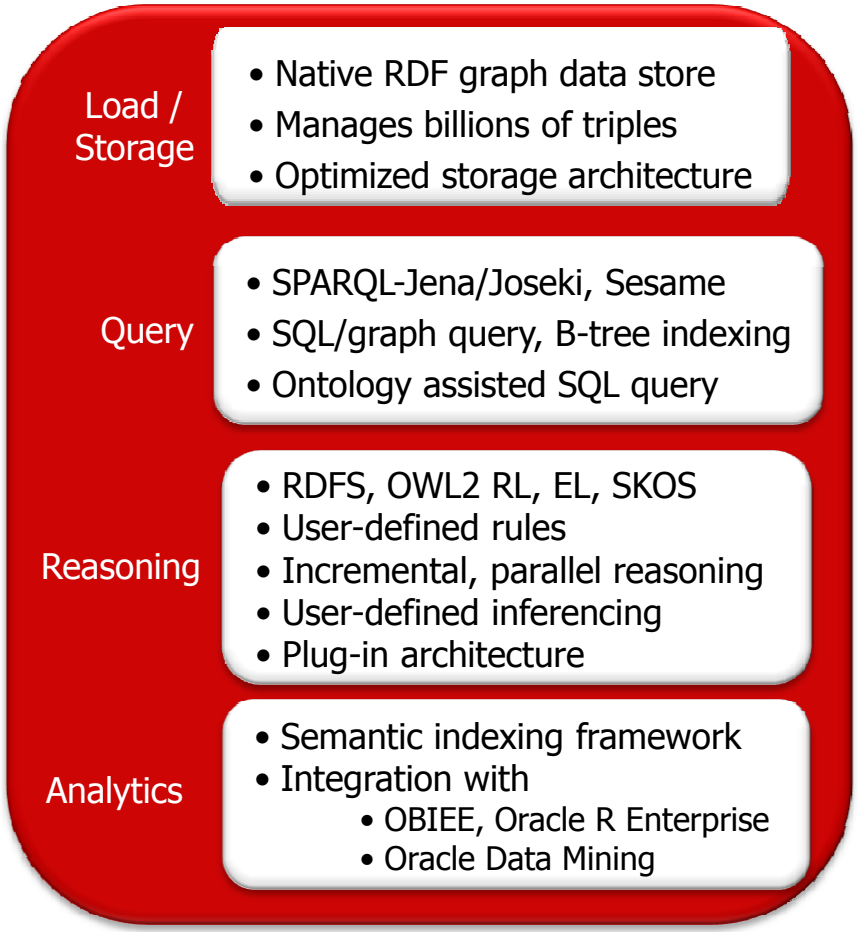
- Analyze content using integrated metadata
  - Blogs, wikis, video
  - Calendars, IM, voice



# Oracle Spatial and Graph
## RDF Triple Store

Leverages Oracle Manageability:

- RAC & Exadata scalability
- Compression & partitioning
- SQL*Loader direct path load
- Parallel load, inference, query
- High Availability
- Triple-level label security
  - Ladder based inference
- Choice of SPARQL, SQL, or Java
- Native inference engine
- Enterprise Manager



# New functions in Oracle Spatial and Graph

- Open Geospatial Consortium (OGC) GeoSPARQL
- 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
  - Subqueries
  - Value Assignment: BIND, GROUP BY Expressions, SELECT Expressions
  - Negation: NOT EXISTS, MINUS
  - Improved Path Searching with Property Paths

**On the fly inference: transitivity of rdfs:subClassOf**

```
SELECT ?c
WHERE {
  ?x rdf:type ?sc .
  ?sc rdfs:subClassOf* ?c }
```

**Social Networking: find all of John's friends**

```
SELECT ?c
WHERE {
 ?x foaf:name "John" .
 ?x (foaf:knows|foaf:friendOf)+ ?f .
 ?f foaf:name ?name }
```

# New functions in Oracle Spatial and Graph

- **RDF views on relational tables (through 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
  - Support RDF view creation using
    - Direct Mapping: simple and straightforward to use
    - R2RML Mapping: customizations allowed
- **Inference**
  - Native OWL 2 EL inference support
  - User defined inferencing
  - Ladder Based Inference
  - Performance optimization for user defined rules
  - Integration with TrOWL, an external OWL 2 reasoner



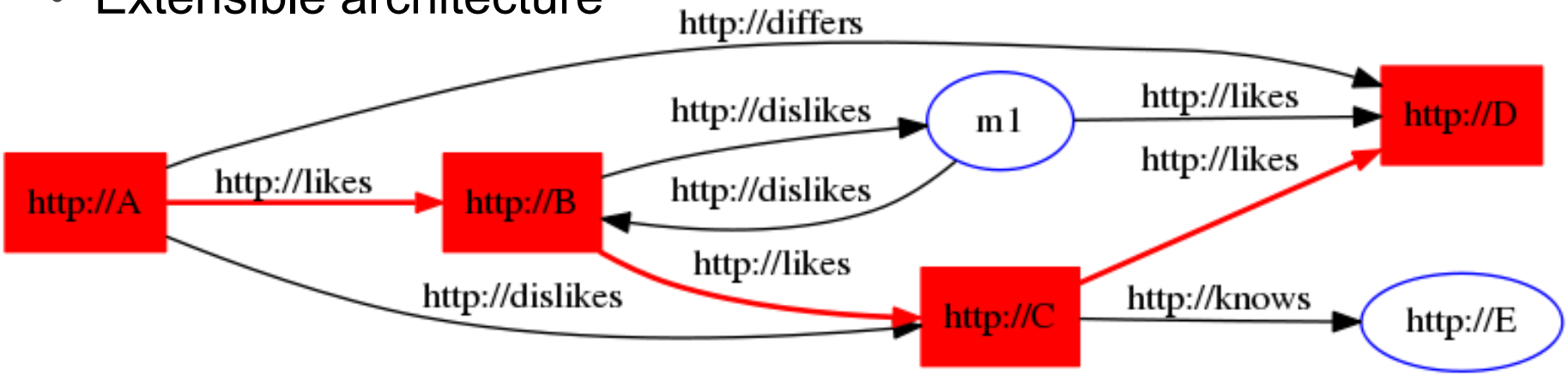

* http://trowl.eu/

# Jena Adapter for Oracle Database

- Requires Apache Jena 2.7.2, ARQ 2.9.2, Joseki 3.4.4, Oracle 11.2.0.3 or higher
- **SPARQL 1.1** compliance
- Named Graph (quads) support: DatasetGraphOracleSem
  - N-QUADS, TriG data format
  - Updated StatusListener interface
  - Named graph queries through Joseki web service endpoint
  - SPARQL Update through Joseki web service endpoint
- JSON output
- Named graph based local inference: Attachment

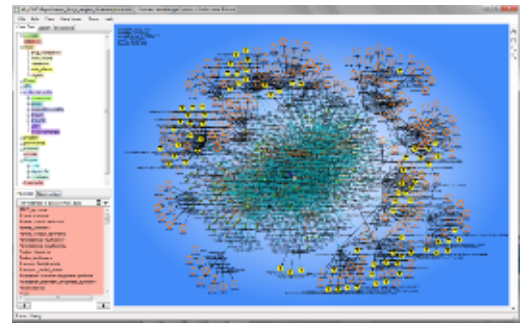
# Jena Adapter for Oracle Database (2)

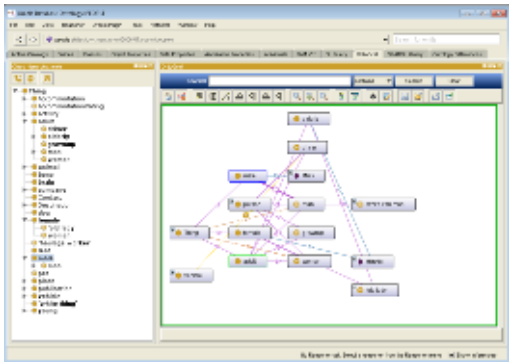
- Analytical functions for RDF data: **SemNetworkAnalyst**
  - Integrating Oracle Spatial and Graph network data model (NDM) with RDF Semantic Graph feature
  - Provides functions including shortest path, within cost, partitioning, …
  - Extensible architecture

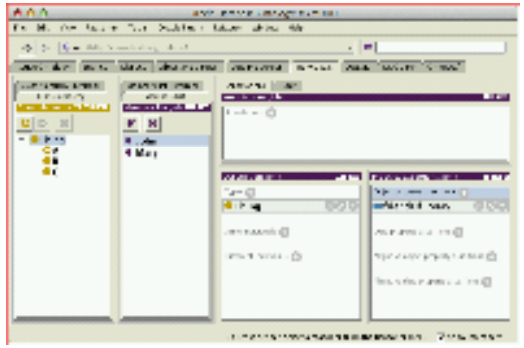


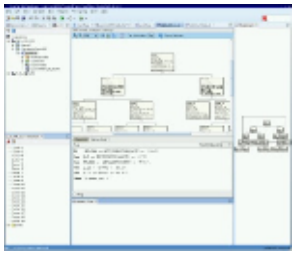
# Tools, Tools, & Tools!

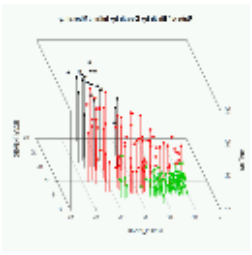
- Ontology editing
- Visualization
- Business Intelligence
- R
- Data Mining
- Knowledge management

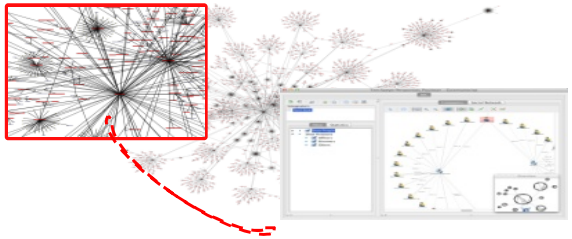












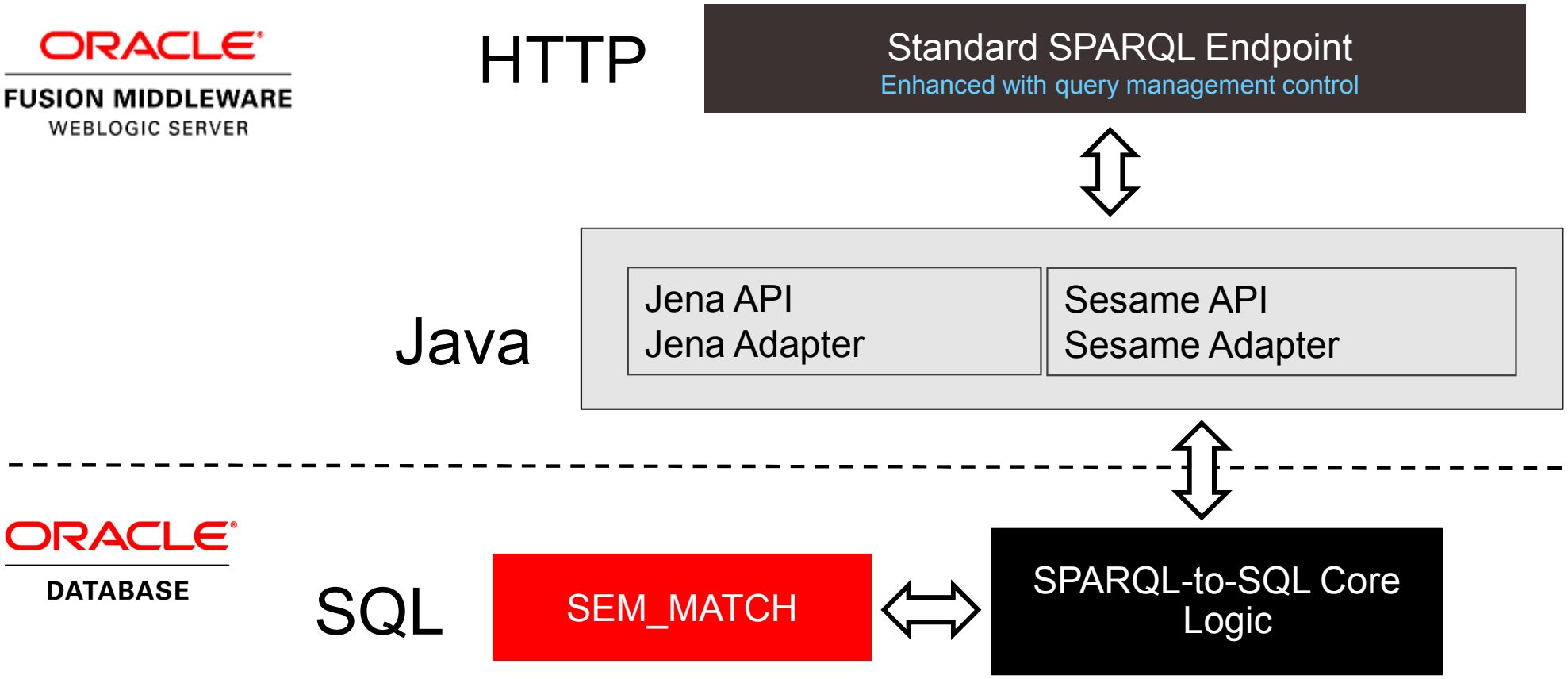
# Industries Have Already Adopted the Concept

## Industries

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

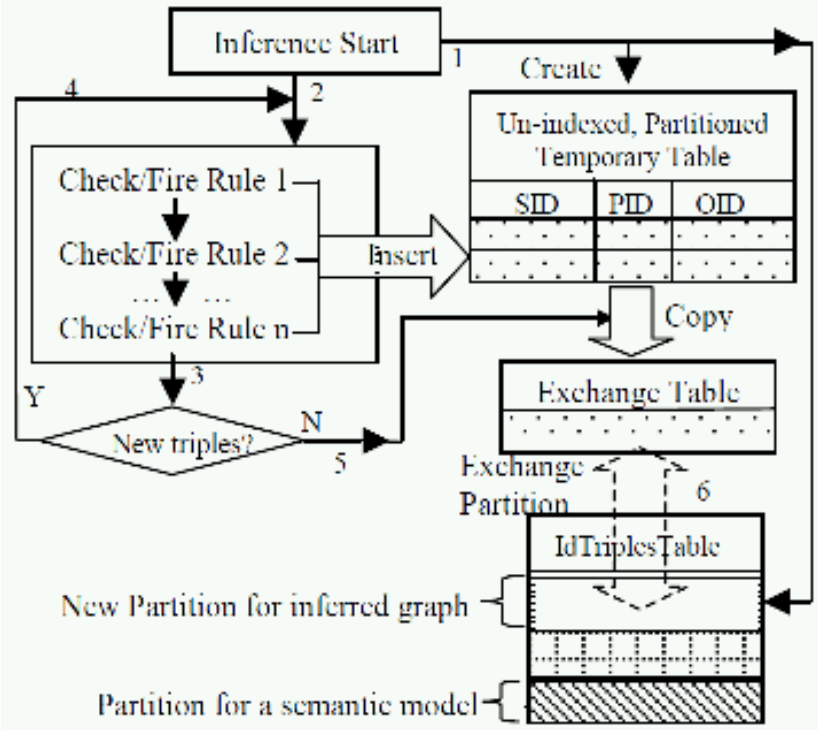
# How Things Work Under the Cover?

# SPARQL Query Processing



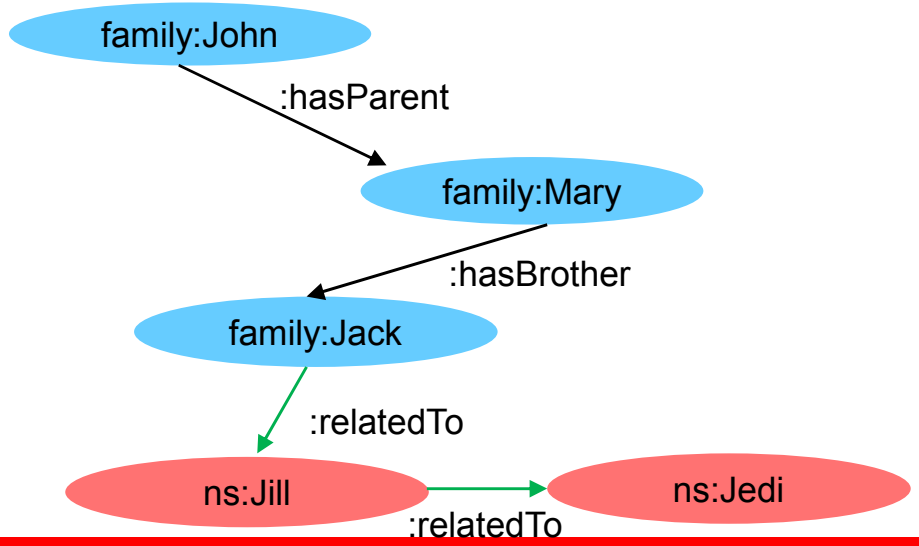
# Native Parallel Inference Engine in Oracle

**Parallel Execution** ➔



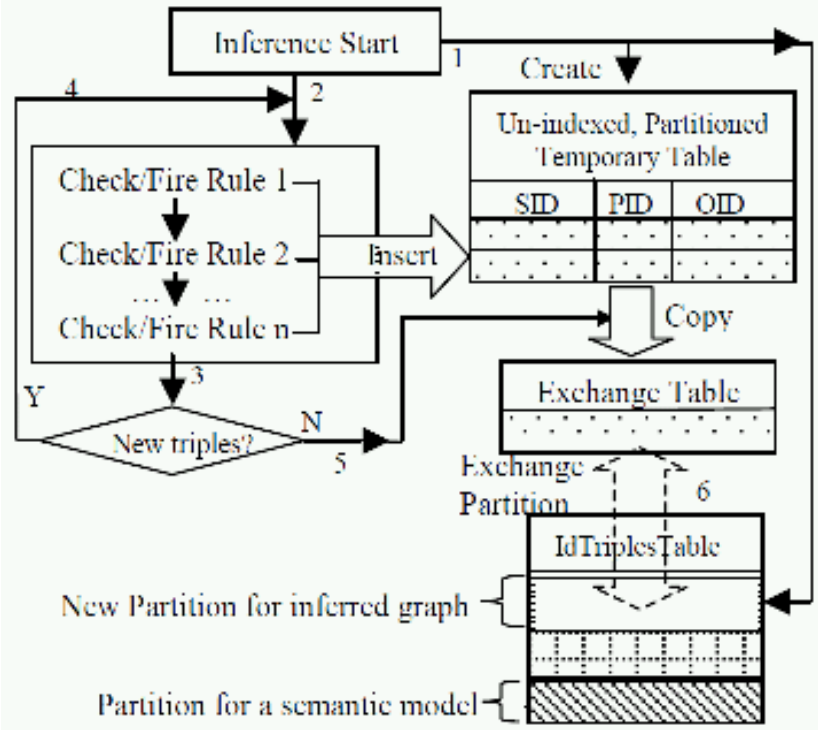

**Implementing an Inference Engine for RDFS/OWL Constructs, ICDE 2008**
**Optimizing Enterprise-scale OWL 2 RL Reasoning in a Relational Database System, ISWC 2010**
**Advancing the Enterprise-class OWL Inference Engine in Oracle Database, ORE 2012**
**Making the Most of your Triple Store: Query Answering in OWL 2 Using an RL Reasoner, WWW 2013**

- **Logical inference can discover hidden relationships with rules**
  - ?x :hasParent ?y . ?y :hasBrother ?z ➔ ?x :hasUncle ?z
  - ?x :relatedTo ?y . ?y :relatedTo ?z ➔ ?x :relatedTo ?z

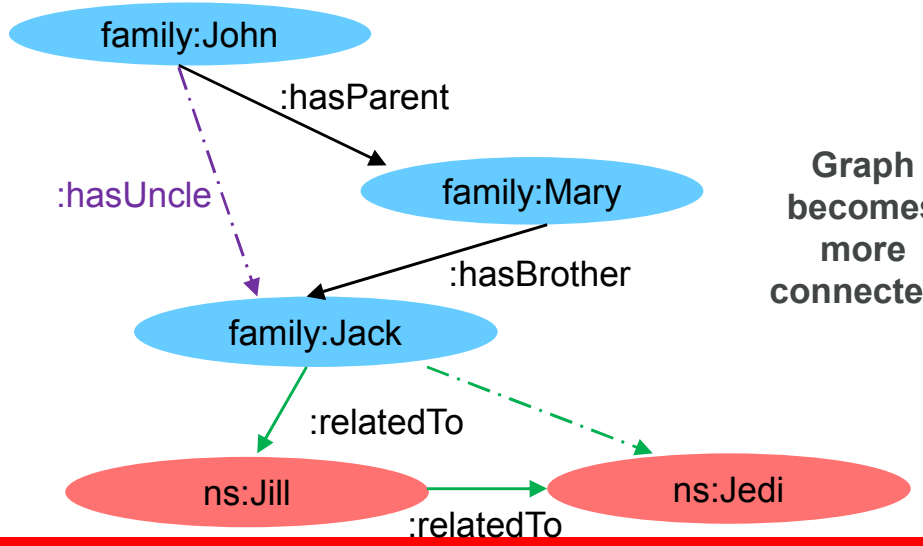


# Native Parallel Inference Engine in Oracle (2)

**Parallel Execution** ➔

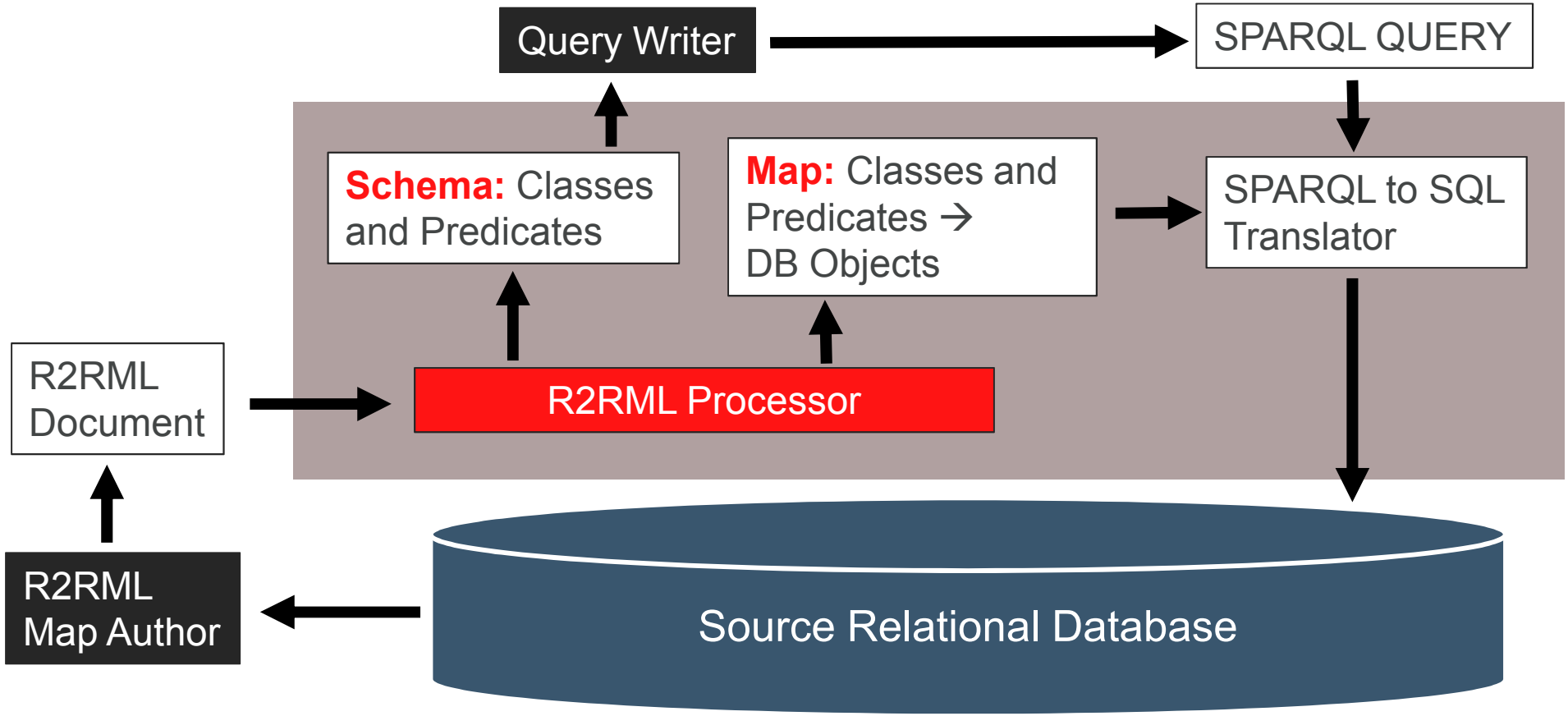



- **Logical inference can discover hidden relationships with rules**
  - ?x :hasParent ?y . ?y :hasBrother ?z ➔ ?x :hasUncle ?y
  - ?x :relatedTo ?y . ?y :relatedTo ?z ➔ ?x :relatedTo ?z




**Graph becomes more connected!**

# Using RDB2RDF (R2RML): Overall Flow



# Performance

# Setup for Performance

- Use a ***balanced*** hardware system for databases and mid-tier servers
  - A single, huge physical disk for everything is ***not*** recommended.
    - Multiple hard disks tied together through ASM is a good practice
  - A virtual machine for multiple databases and applications is ***not*** recommended
  - Make sure throughput of hardware components **matches** up

| Component | Hardware spec | Sustained throughput |
|---|---|---|
| CPU core | - | 100 - 200 MB/s |
| 1/2 Gbit HBA | 1/2 Gbit/s | 100/200 MB/s |
| 16 port switch | 8 * 2 Gbit/s | 1,200 MB/s |
| Fiber channel | 2 Gbit/s | 200 MB/s |
| Disk controller | 2 Gbit/s | 200 MB/s |
| GigE NIC (interconnect) | 2 Gbit/s | 80 MB/s* |
| Disk (spindle) | | 30 - 50 MB/s |
| MEM | | 2k-7k MB/s |

# Configure OS and Network

- Network configuration is important to data integration performance
  - Network MTU (TCP, Infiniband)
  - net core rmem_max, wmem_max
- Linux OS Kernel parameters
  - shmmax,
  - shmall,
  - aio-max-nr,
  - sem, …

# Configure Database

- Database parameters
  - SGA, PGA, filesystemio_options,
  - db_cache_size, auto dop, …
- Calibrate I/O performance
  - DBMS_RESOURCE_MANAGER.CALIBRATE_IO
- Gather statistics
- Run a ***typical*** workload on a ***typical*** data set
  - Check AWR report to see top waits
  - Check SQL Monitor report to find bottlenecks in SQL executions

# Configure Mid-Tier Server

- Understand bottleneck
  - Use tools, jstack/top for example, to identify top threads
- Set a proper JVM heap size
  - Pay close attention to GC activities and memory related settings
    - Try –XX:+UseParallelGC, -XX:+UseConcMarkSweepGC, :NewRatio, :SurvivorRatio, etc.
- For Java clients using JDBC (through Jena Adapter)
  - Network MTU, Oracle SQL*Net parameters including SDU, TDU, SEND_BUF_SIZE, RECV_BUF_SIZE,
  - Linux Kernel parameters: net.core.rmem_max, wmem_max, net.ipv4.tcp_rmem, tcp_wmem, …

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

## Load, Inference and Query Performance

- The LUBM 200K Graph has 48+ Billion triples (edges)
  - Original graph has 26.6 Billion unique triples (quads)
  - Inference produced another 21.4 Billion triples
- Data Loading Performance
  - Triples Loaded and Indexed Per Second (TLIPS): **273K**
- Inference Performance
  - Triples Inferred and Indexed Per Second (TIIPS): **327K**
- SPARQL Query Performance
  - Query Results Per Second (QRPS): **459K**

**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)

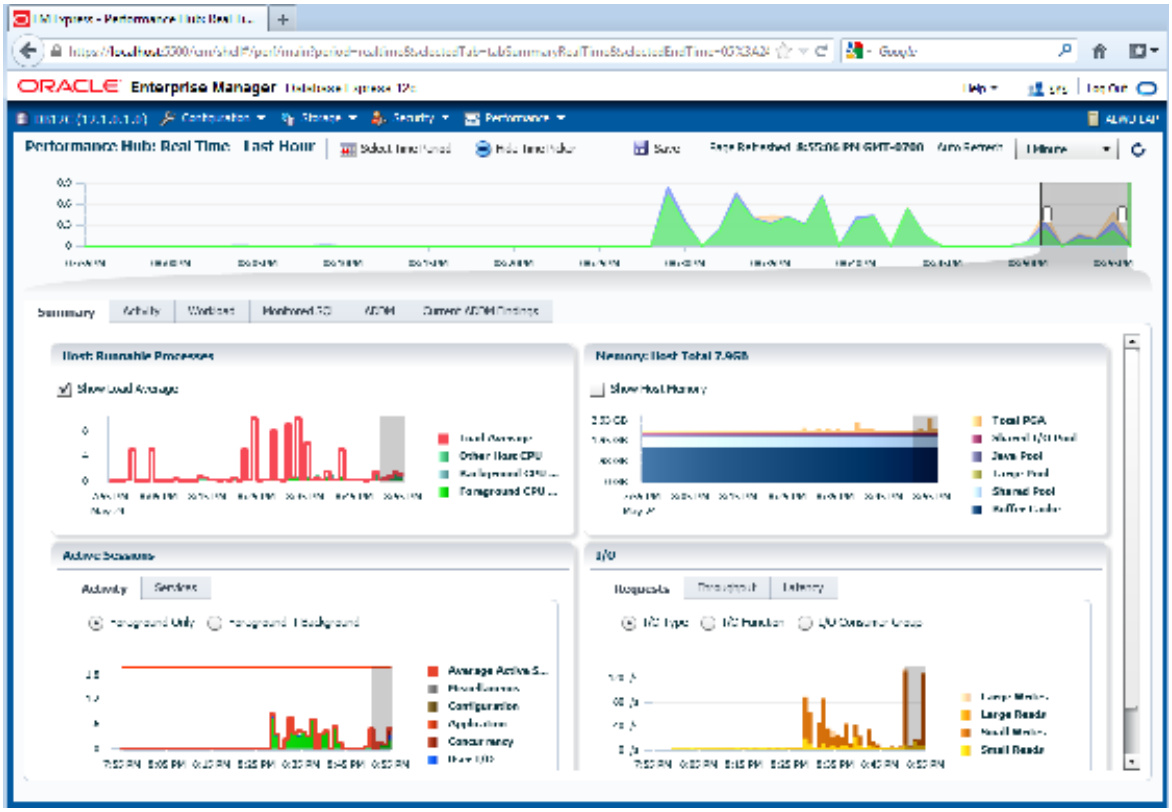
**Software:** Oracle Database 11.2.0.3.0, SGA_TARGET=750G and PGA_AGGREGATE_TARGET=200G

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

# Oracle Enterprise Manager

## Understand exactly what is the going on

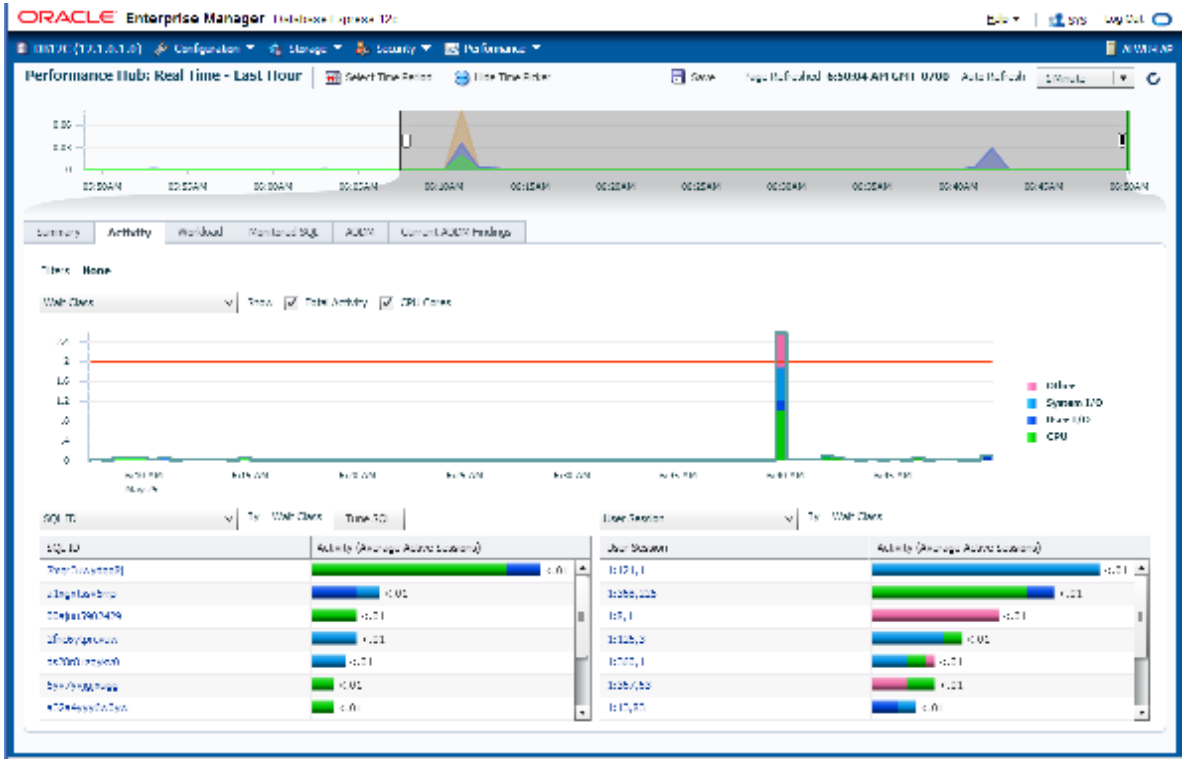
- Configuration
- Storage
- Security
- Performance
  - Real time monitor
  - CPU
  - Memory
  - I/O
  - Sessions
  - Activity
  - Workload
  - …
- …



# Oracle Enterprise Manager

## Understand exactly what is the going on

- **Configuration**
- **Storage**
- **Security**
- **Performance**
  - **Real time monitor**
  - **CPU**
  - **Memory**
  - **I/O**
  - **Sessions**
  - **Activity**
  - **Workload**
  - **…**
- **…**



# Summary of New Graph Features

- **Network Data Model graph**
  - Real World Feature Modeling
  - Multimodal Routing, Temporal Modeling and Analysis
  - Large Scale Drive Time/Distance Analysis
- **RDF Semantic Graph**
  - RDF views on relational tables
  - SPARQL 1.1, GeoSPARQL, SPARQL Gateway
  - Enhanced Reasoning and Security
  - Named Graphs
- **Future work on Large-Scale Parallel Graph Analytics**
  - Parallel In memory graph analytics, SQL-based graph analytics, Distributed in-memory analytics
  - Integration with Green-Marl

# Hardware and Software

ORACLE

# Engineered to Work Together


ORACLE®