GEOSS Platforms: Combining Geodata, Big Data, Semantics, Graphs, Supporting Collaborations & Knowledge Bases

Steven Hagan
Vice President
Oracle Database Server Technologies
December, 2016
Overview

• Trends in Hardware, Software, Data Volumes/Rates, Costs
• Polyglot Persistence, Multi-model Databases and effects
• Semantics, Ontologies, Metadata effects
• Graph databases and Analytics
Global Digital Data Growth: Exceeds Storage Mfg
Growing leaps and bounds by 40+% YoY!

2009 = .8 Zetabytes
= .08 ZB Structured Data
= .72 ZB Unstructured Data

2020 = 35 Zetabytes
= 3.5 ZB Structured Data
= 31.5 ZB Unstructured Data*
(1 Zetabyte = 1 Trillion Gigabytes)

• Chart conservatively assumes a constant 9:1 ratio of unstructured data vs. structured data (based upon IDC’s estimate that 90% of all digital data is unstructured).
• Chart does not reflect IDC’s projection that unstructured data is currently growing twice as fast as structured data at the rate of 63.7% vs. 32.3% CAGR.

Source: IDC Digital Universe Study, A Digital Universe Decade – Are Your Ready?, 2010
Human Race is Generating Data Vastly faster than Making Computer Storage: You Cannot Keep it ALL! Deletion Policy Needed

- Yottabytes
- Zettabytes
- Exabytes
- Petabytes

YEARS 2020 ->

DATA

STORAGE
Data Volume & Variety Generation Explosion Continues – Terabytes, Petabytes, Exabytes, Zettabytes

- VIDEO: UAVs, DRONES, SURVEILLANCE
- IMAGERY/Raster: (Satellites, Medical)
- Sensors (IOT), LIDAR, 3D, RFID
- Social Media, Web Scraping, Mobile Phones
- New data products for: Land and Water mgmt, Agriculture, Environment Transportation, Terrain and City Models, SDIs for planning, maintenance, Emergency response, Defense, Intelligence, Consumers, Healthcare
- Location is a Powerful Organizing Principle
- Semantics, Ontologies --
- Wearable Technologies
- Genomics (DNA Sequencing), Astronomy
- MULTIPLE VERSIONS OF THE ABOVE
IT TRENDS: Next 5 years or so

• Computer System Performance –
  – Hardware - **EVOLUTIONARY** – Moore’s law still holding
  – New possibilities at Research Level – not yet proven
    • DNA for Storage; 3D Glass, Holography; Carbon Nanotubes, Graphene
  – Software – **DISRUPTIVE** – Parallelism enables clusters of **10,000+** computers, CLOUD

• Software – **FLEXIBILITY** - Now Supporting Diverse Data types in Databases
  – Databases/persistent stores can handle all types of data – **POLYGLOT PERSISTENCE**
  – Software – Graph Storage, Semantics – Add all types of data and build new relationships
    • Without disruptive upgrades & no schema changes
  – Stream data arriving; Filter the data; Keep what matches your requirements; aggregate it
  – Deletions: immediately or eventual
  – **NOTE:** TEXT AND NUMBERS ARE NOT THE SPACE PROBLEM!
SPECIAL DATA TYPES: SEVERAL POPULAR DATA MODELS: But Unique separate persistent stores results in: MANY databases to secure & manage

- RELATIONAL TABLES ROWS
- NoSql Key-Value
- HADOOP Sharding
- GRAPHS: PROPERTY RDF/OWL SEMANTICS
- REAL TIME IN MEMORY
- SPATIAL IMAGING VIDEO AUDIO COLUMNAR DOCUMENT
For National / UN / GEO: MULTI-MODEL Database is Best -
Many Different Data Models Supported as Native Data Types in

**ONE SHARED STORE**

- Parallel Database Server has multiple models
- Unified Security Approach
- Highly Available
- Disaster Tolerant
- Shares Main Memory; more efficient
- Shares Disks, Flash Storage: more efficient
- Managed as a single entity: more efficient
- *(ORACLE HAS THIS TODAY)*
Future Trends

- Technology Trends in Data Creation, Maintenance, and Management
- Reliance on ‘big data’ technologies
- The right information at the right time
- Machine-processable descriptions of data.
- Semantic technologies will play an important role
- Skills and Training: train the individuals is at least five years

Requirement for enhanced Data Management Systems
Today: More HW/SW Efficiencies: But **Labor Costs Growing**

Innovative Systems for Statistics Needed

![Graph showing the increase in maintenance costs over time with decreasing hardware and software costs.](image-url)
Guidance: Do Not Build Your Knowledge Solutions From Scratch

Long Term Cost of Ownership rises with custom construction & Open Source

UN-GGIM: “train the individuals is at least five years”

Time to Build
Optimizations
Maintenance
Managing All Spatial and Graph Data – in One Store

Location and Graph analysis with Secure, scalable storage for enterprise data

Oracle Spatial and Graph 12c

SQL access
Web Services (OGC)
SPARQL End Point

40+ Graph Analysis Functions (PGX)

Property Graphs
Network Graphs
RDF Semantic Graphs
Topologies

Rasters
3D, point clouds (LiDAR)

Spatial Vector Acceleration
Geocoding
Routing
Inferencing

Copyright © 2015, Oracle and/or its affiliates. All rights reserved.
Spatial: Open and Interoperable
GEOSS Data Collaboration / Repurposing: Ontology-driven Enable Shared, Actionable Knowledge

- Simple Features
- GeoRaster
- Topology
- Networks
- Gazetteers

- Data Integration
- National Map schemas
- Geographic names
- Temporal
- Naïve Geography
- …
Semantics: Harmonizing the Electronic Health Care Ecosystem

Enterprise-wide, Patient-centric, longitudinal Record System

Domain Ontologies
(business metadata + Ontologies)

Data Servers

Lab/clinical Care
Research
Content Mgmt
Billings/Claims
Reporting/BI

Data Sources / Data Types

Social Media
Medical Devices
Lab Information Systems
Subscription Services
Legacy Patient Records
Linked Open Data: Connecting With other Services and Clouds
Support Breadth of National & UN Data

Data arrives, is filtered, stored data is available to all GEO Organizations

GUIDANCE: THIS IS AN ARCHITECTURE TO SUPPORT ONE SHARED MULTIPURPOSE NATIONAL / UN STORE
## Semantic & Graph Technology

**What terms to look for:** Buzzwords For Apps & Workflows using

<table>
<thead>
<tr>
<th>Semantic Web</th>
<th>Property Graphs</th>
</tr>
</thead>
<tbody>
<tr>
<td>W3C RDF/OWL/SPARQL</td>
<td>Taxonomy/Terminology Mgmt</td>
</tr>
<tr>
<td>Graph Data Management</td>
<td>Faceted Search</td>
</tr>
<tr>
<td>Social Network Analysis (SNA)</td>
<td>Inferencing / Reasoning</td>
</tr>
<tr>
<td>Knowledge Discovery</td>
<td>Sentiment Analysis</td>
</tr>
<tr>
<td>Knowledge Mining</td>
<td>Text Mining</td>
</tr>
<tr>
<td>Big Data</td>
<td>NoSQL Database</td>
</tr>
<tr>
<td>Schema-less Data</td>
<td></td>
</tr>
</tbody>
</table>
Oracle: Linked Data support: on-premise or in the Cloud

Included in Oracle Database-as-a-Service Cloud Offering

- Highly scalable, secure triple store based on RDF (Resource Description Framework)
  - 1 TRILLION TRIPLE BENCHMARK, leading Triple Store: W3.org
    - 1.13 million triples per second query performance

- SPARQL and SPARQL in SQL support
  - Apache Jena and OpenRDF Sesame pre-integrated
  - SPARQL endpoint enhanced with query control
  - GeoSPARQL support (classes, properties, datatypes, query functions)

- Forward-chaining based inferencing engine in the database
  - Various native rulebases (RDFS, OWL2 RL, SKOS, ...), integration with OWL2 reasoners (TrOWL, Pellet)

- RDB to RDF mapping on relational data aligned with RDB2RDF standard
What data model is most important to You? Depends on your questions - There Are Lots of Answers.

• Answers from **Aggregation**
  – Who spends the most?
  – Who buys the highest margin goods?
  – Who is most consistently a top contributor?

• Answers from **Connectivity**
  – Who’s most influential?
  – Which supplier do I depend on the most?
  – What is the right product mix for millennials?

**Tabular questions:**
Well-suited to SQL-like tools

**Graph questions:**
We need something different!
RDF Graph v. Property Graph

RDF Semantic Graph

- Typical Use Case:
  - Linked data management, semantic integration
- Analytics:
  - pattern matching, Inferencing
- URIs as node IDs
  - Unique across graphs, can be used for integration
- Edges only have a label (a predicate URI)
  - An edge is not unique, represents a relationship
  - Complex reification to represent unique edges
- RDFS, OWL, etc for semantics

Property Graph

- Typical Use Case:
  - Social & intelligence network analysis
- Analytics:
  - Clustering, centrality, page rank, path finding
- Simple IDs
  - Not intended for graph fusion
- Edges have unique IDs and labels, and also properties and values
- No built-in semantics
Computational Analytics: Built-in Graph Package in S&G

Rich set of built-in parallel graph algorithms

Detecting Components and Communities
- Tarjan’s, Kosaraju’s, Weakly Connected Components, Label Propagation (w/ variants), Soman and Narang’s Spacification

Evaluating Community Structures
- Conductance, Modularity, Clustering Coefficient (Triangle Counting), Adamic-Adar

Link Prediction
- SALSA (Twitter’s Who-to-follow)

Other Classics
- Vertex Cover
- Minimum Spanning-Tree (Prim’s)

Ranking and Walking
- Pagerank, Personalized Pagerank, Betweenness Centrality (w/ variants), Closeness Centrality, Degree Centrality, Eigenvector Centrality, HITS, Random walking and sampling (w/ variants)

Path-Finding
- Hop-Distance (BFS), Dijkstra’s, Bi-directional Dijkstra’s, Bellman-Ford’s

... and parallel graph mutation operations

The original graph

Create Bipartite Graph

Sort-By-Degree (Renumbering)

Filter-Expression

Filtered Subgraph

Create Undirected Graph

Simplify Graph

The original graph
Left Set: “a,b,e”
Why Are Graph Databases Popular Now?

Rise of social networking

Google, Yahoo, Twitter, Facebook, Linked In:
Enterprises want to mine public graphs and also build their own networks

Enterprise applications increasingly need to model data relationships

Telecoms: Network & Data center management, identity management
Financial Services: Fraud detection; cross-selling
Media & Publishing: Social apps, recommendation, sentiment analysis
Health Care: CRM, fraud detection, Doctor, patient, diagnosis, treatment analysis;

Modeling complex relationships as a graph is efficient

Improves performance
Simplifies queries, traversal, search and analytics

**ORACLE IS THE ONLY MAJOR PLAYER!**

No IBM, No Microsoft – only small open-source startups – we can scale and are secure ....
## Oracle Spatial and Graph Partners: Semantics / RDF
### Integrated Tools and Solution Providers

<table>
<thead>
<tr>
<th>Ontology Engineering &amp; Visualization</th>
<th>Reasoners</th>
<th>NLP Entity Extractors</th>
</tr>
</thead>
<tbody>
<tr>
<td>TopQuadrant</td>
<td>clarkparsia, LLC</td>
<td>lyamba, semantic system</td>
</tr>
<tr>
<td>Tom Sawyer</td>
<td>ontoprise</td>
<td>expert system</td>
</tr>
<tr>
<td>protégé</td>
<td>ontoprise</td>
<td>general architecture</td>
</tr>
<tr>
<td>Cytoscape</td>
<td>trOWL</td>
<td>for text engineering</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Open Source Frameworks</th>
<th>Standards</th>
<th>Applications &amp; Tools</th>
</tr>
</thead>
<tbody>
<tr>
<td>jena</td>
<td>OGC</td>
<td>Callimachus</td>
</tr>
<tr>
<td>openRDF.org</td>
<td>RDF</td>
<td>IO</td>
</tr>
<tr>
<td>Joseki</td>
<td>Semantic Web</td>
<td>INFORMATICS</td>
</tr>
<tr>
<td>Sesame</td>
<td>W3C</td>
<td>Mondeca</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>SI / Consulting</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Sesame</td>
<td>infoMENTUM</td>
<td>arms</td>
</tr>
<tr>
<td>trivadis</td>
<td>computas</td>
<td>SAIC</td>
</tr>
<tr>
<td>TenForce</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
National / UN GEO: one Multi-Model Store

External Data Sources:
- Transactional & Operational Systems
- Contents Repository
- Databases
- Mobile Devices, Web resources
- Blogs, Mails, news
- Satellite Imagery, UAVs

Real-time Data Streams

Search, Presentation, Report, Visualization, Query

Multi-Model Data Management Infrastructure
- GeoSpatial
- Documents
- Secured
- Historical Records
- POIs
- Demographics
- Customer Data
- Call Records

Automatically Responses and Publishing
- SMS
- Console Alerts
- EV Grid Management
- Workflow Initiation
- Real-time Dashboards
Oracle Geospatial Platform and Cloud Data Platforms
Support any data, any scale, on-premises or in the Cloud
In-Memory / Flash Based / Disk Based – Scale to Many Petabytes

Relational Store
• Relational
• Spatial
• Imagery, Video
• Graph – RDF and Property
• Documents, Text
• Real-time Analytics
• Multi-Model

NoSQL Store
• Key-value
• Graph – RDF
• Graph – Property
• Document

Big Data Store
HADOOP, Spark
• Logs
• Streaming
• Archive
• Spatial
• Graph – RDF and Property
• Web Analytics

Data Integration
Change Capture and Apply, ETL, and Federated SQL
You Enhance Innovation & Sharing By Using **STANDARDS**
e.g. – The Spatial / Semantics Data Domains

- **ISO**
  - TC 211; TC 204
- **Open Geospatial Consortium**
  - Simple Features; GML; Web Services
- **De-facto Standards**
  - SHP, MGE, DXF, KML
- **Professional Standards**
  - ISPRS, FIG, WMO
- **Java, .NET, Flash**
- **W3C: RDF, OWL, SPARQL, GeoSPARQL**
- **TAGGED METADATA** – agree on tags

**SQL3/MM Spatial**
Accessible Shared Data: **CYBERSECURITY** is Major Challenge Requires Information Security and Privacy

- Monitoring
  - Configuration Management
  - Audit Vault, Total Recall

- Access Control
  - Database Vault
  - Label Security

- Encryption & Masking
  - Advanced Security
  - Secure Backup, Data Masking
  - Silicon based Security
Public Clouds, Private Clouds: Platforms

- Used by multiple tenants on a shared basis
- Hosted and managed by cloud service provider

**Public Clouds**
- IaaS
- PaaS
- SaaS

**Private Cloud**
- IaaS
- PaaS
- SaaS

**Trade-offs**
- Lower *upfront* costs ↔ Lower *total* costs
- Outsourced management ↔ Greater control over security, compliance, QoS
- OpEx ↔ CapEx & OpEx

ELASTICITY is key value of Clouds

YOU MAY NEED A CLOUD IN EACH COUNTRY ---DEPENDS ON THEIR LAWS

---

Oracle Technology Supplies both Public and Private clouds

---

**YOU MAY NEED A CLOUD IN EACH COUNTRY ---DEPENDS ON THEIR LAWS**
Collaboration / Knowledge Bases: Success Enhanced with MULTI-MODEL DATABASE PLATFORM

**Big & Fast Data**
- Volunteered Geographic Statistical Information
- Sensors Streaming Data
- Geo-referenced Video, 3D, LiDAR Satellites

**Simplify Statistics IT**
- Support for Open Standards
- Spatial Database, Application Server, BI, tools
- Support by Leading Partner solutions
- Multi-Model Engineered Systems

**Deep Analytics**
- Real-time Complex Event Processing
- Dense Visualization
- Spatial Analysis Graph Analytics

**On Premise, On Cloud, Shared Services**
- Shared GeoSpatial Services Location Aware Everything

*Fully Parallel and Secure*

Copyright © 2015, Oracle and/or its affiliates. All rights reserved.
Evolution of Enabling Workflows

- HENRY FORD
- 100 YEARS AGO
- Industrialized the Manufacturing Workflow
- Popularized the use of
  - Assembly Line in Manufacturing
- One Automated Machine Driven Platform
  - for the entire car assembly process
  - Used STANDARD, SHAREABLE, INTERCHANGEABLE parts
  - Lowered the priced of a car by factor of 10, in 3 years.
- His success led to widespread adoption of Workflows