Graph and Spatial Analytics for Built for Big Data Platforms

SF DAMA Day
December 14, 2016

Jean Ihm, Product Manager
Zhe Wu, Architect

Oracle Spatial and Graph Technologies
Safe Harbor Statement

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.
Agenda

1. Introduction
2. Spatial Technologies
3. Graph Technologies
4. Resources, Q + A
Graph and Spatial Analysis – It is about relationships

- Are things in the same location? Who is the nearest? What tax zone is this in? Where can deliver in 35 minutes? What is in my sales territory? Is this built in a flood zone?

- Which supplier am I most dependent upon? Who is the most influential customer? Do my products appeal to certain communities? What patterns are there in fraudulent behavior?
The Big Picture – Oracle Information Management System

DATA RESERVOIR

- Cloudera Hadoop
- Oracle Big Data SQL
- Oracle NoSQL
- Oracle R Distribution
- Oracle Big Data Spatial and Graph

Big Data Appliance

DATA WAREHOUSE

- Oracle Database
- In-Memory, Multi-tenant
- Oracle Industry Models
- Oracle Advanced Analytics
- Oracle Spatial and Graph

Oracle Big Data Connectors
Oracle Data Integrator
Oracle Event Processing

SOURCES

Oracle Data Integrator
Oracle GoldenGate
Oracle Event Processing

Copyright © 2016 Oracle and/or its affiliates. All rights reserved.
Oracle’s Spatial and Graph Strategy

Enable Spatial and Graph use cases on every platform

Oracle Database Spatial and Graph

Oracle Big Data Spatial and Graph

Spatial and Graph in Oracle Cloud

Database 12c

Big Data Appliance

Oracle Big Data Cloud Service

Oracle Database Cloud Service
Oracle Spatial and Graph: A Proven Platform

Location-Enabled Business Applications

Specialist Geospatial Applications

- Geocoding
- Spatial searches
- Routing
- Mapping
- Geometry Topology
- GeoRaster Networks
- Linear Referencing
- Geodetic
- Long Transactions
- 3D (Point clouds, LIDAR)
Oracle Spatial and Graph
Location and graph analysis with secure storage for enterprise data

Deployable Services | Geocoding | Routing | Web Services (OGC)

Polygons | SQL | ORACLE | Raster
Lines | Networks | RDF Graphs | Property Graphs (new in 12.2 on Oracle Cloud)
Points | 3D / LIDAR | Topologies
Big Data Challenges

Requires more development resources and data scientists

Build your own environment from commodity hardware and open source software.
Big Data Solution

Make developers and data scientists more productive

Optimized, pre-configured Big Data Appliance and Cloud platforms

Pre-built, parallel, MapReduce and Spark Spatial Algorithms

Raster and Vector processing Frameworks
Oracle Big Data Spatial and Graph

Spatial Analysis:
• Location Data Enrichment
• Proximity and containment analysis, Clustering
• Spatial data preparation (Vector, Raster)
• Interactive visualization

Property Graph for Analysis of:
• Social Media Relationships
• eCommerce Targeted Marketing
• Cyber-Security, Fraud Detection
• IoT, Industrial Engineering

Multimedia Analysis:
Framework for processing video and image data, such as facial recognition
Insurance Industry

86% of Insurance companies agree that analyzing multiple data sources together is crucial to making accurate predictions.

88% Agree that linking information by location is key to combining disparate sources of Big Data.

Source: “The big data: How data analytics can yield underwriting gold.”
Survey conducted by Ordnance Survey and Chartered Insurance Institute, 25 April 2013.

Use Case: Linking Information by Location
Data Harmonization: Linking information by location

Are these data points related?

• **Tweet**: sailing by #goldengate

• **Instagram image subtitle**: 골든게이트 교*

• **Text message**: Driving on 101 North, just reached border between Marin County and San Francisco County

• **GPS Sensor**: N 37°49′11″ W 122°28′44″

• Now find all data points around Golden Gate Bridge ...

* Golden Gate Bridge (in Korean)
GeoSocial Analysis Use Cases: “Nearest Friends”

“It’s 11:30. Want to meet Jon, Melli, and Albert for lunch @Milano’s today at Noon?”

URGENCY WITH LIMITED AVAILABILITY: Lunch promotion to targeted potential “table of 4” who know each other within 1 km.

“We know you and your old college buddies love Elton John. Get together with Tom, Dick and Hari and the rest of the frat next month.”

BROADER SOCIAL REACH; WIDER DISTANCE. NO TIME CONSTRAINT: Concert promotion.

“Great seats @ Cinema 18 for 7:30 show of new Avengers movie tonite. Free popcorn and soft drink for you and Mary. Text her at 555-1234.”

TIME CONSTRAINED: Target people within 20 minute drive 30 minutes before the show.
Use case: Data preparation

- Mosaic images
- Terrains and contours
- Shaded reliefs

Pyramiding: layers at different resolution
What problems can Big Data Spatial analysis address?

Data Harmonization using any location attribute (address, postal code, lat/long, placename, etc).

Categorization and filtering based on location and proximity.

Preparation, validation and cleansing of Spatial and Raster data.

Visualizing and displaying results on a map.

Spatial querying and analysis of Hadoop data with SQL.
What features does Big Data Spatial have?

Data enrichment service API using GeoNames and geometry hierarchy data

MapReduce routines for distance calculations, PointInPolygon, buffer creation, Categorization, KMeansClustering, Binning

Spatial processing of data stored in HDFS, NoSQL or Spark. Raster processing operations: Mosaic and sub-set operations. Geodetic and Cartesian data

HTML5 Map Visualization API

Hive SQL API
Query from Oracle DB with Big Data SQL & Oracle SQL Connectors for Hadoop

Copyright © 2016 Oracle and/or its affiliates. All rights reserved.
Demo

Spatial analysis on global security incidents

Using Big Data Discovery + Big Data Spatial Analytics
Summary
Oracle Big Data Spatial and Graph

- Commercial, supported software
- Componentry to **boost efficiency** of data scientists and developers – save time on custom development
- Bring location context to big data – harmonize and offer **new insight** into customers, assets, organizations
- **Complement** Relational and Big Data processing and analysis
- Oracle Big Data Spatial and Graph offers
  - Dozens of pre-built algorithms/enrichment services, and map visualization
  - Scaleable storage and parallel processing on Hadoop /Oracle NoSQL/Spark
  - Runs on commodity hardware or BDA, both on-premise or in the Cloud
Overview of Property Graph Support in Oracle Big Data Spatial and Graph (BDSG) and Oracle Database 12.2
Relational Model vs. Property Graph Model

- Relational Model

- Graph Model

Courtesy: Tom Sawyer 2016
The Property Graph Data Model

- 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 (or links)
  - 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.

https://github.com/tinkerpop/blueprints/wiki/Property-Graph-Model
How is graph analysis important to business?

• What patterns are there in fraudulent behavior?
• Which supplier am I most dependent upon?
• Who is the most influential customer?
• Do my products appeal to certain communities?
• What targeted products or services do I recommend to customers?
Graph Analysis in Business

- **Product Recommendation**: Recommend the most *similar* item purchased by *similar* people.
- **Influencer Identification**: Find out people that are *central* in the given network – e.g. influencer marketing.
- **Community Detection**: Identify group of people that are close to each other – e.g. target group marketing.
- **Graph Pattern Matching**: Find out all the sets of entities that match to the given pattern – e.g. fraud detection.
Oracle Big Data Spatial and Graph Property Graph Architecture

Graph Analytics
- Parallel In-Memory Graph Analytics (PGX)

Access Layer
- Apache Blueprints & Lucene/SolrCloud

Oracle Big Data Spatial and Graph
- Apache HBase
- Oracle NoSQL Database

Java APIs

REST/Web Service/Notebooks

Property graph formats supported
- GraphML
- GML
- Graph-SON
- Flat Files
- CSV
- Relational Data Sources
Graph Construction: Convert from Relational to Flat Files

- Two Key Java APIs:
  - OraclePropertyGraphUtils.convertRDBMSTable2OPV (E)
  - ColumnToAttrMapping

- Key Steps:
  - Column Mapping
  - Data Type Definition
  - Conversion

<table>
<thead>
<tr>
<th>EMPID</th>
<th>hasName</th>
<th>hasAge</th>
<th>hasSalary</th>
</tr>
</thead>
<tbody>
<tr>
<td>101</td>
<td>Jean</td>
<td>20</td>
<td>120.0</td>
</tr>
<tr>
<td>102</td>
<td>Mary</td>
<td>21</td>
<td>50.0</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
</tbody>
</table>

Example output .opv file

1101,name,1,Jeain,,
1101,age,2,,20,
1101,salary,4,,120.0,
Data Access (APIs)

- Blueprints 2.3.0, Gremlin 2.3.0, Rexster 2.3.0
- Groovy shell for accessing property graph data
- REST APIs (through Rexster integration)
- PGQL (Property Graph Query Language)
Text Search through Apache Lucene/SolrCloud

- Integration with Apache Lucene & SolrCloud
- Support manual and auto indexing of Graph elements
  - Manual index:
    - `oraclePropertyGraph.createIndex("my_index", Vertex.class);`
    - `indexVertices = oraclePropertyGraph.getIndex("my_index", Vertex.class);`
    - `indexVertices.put("key", "value", myVertex);`
  - Auto Index
    - `oraclePropertyGraph.createKeyIndex("name", Edge.class);`
    - `oraclePropertyGraph.getEdges("name", "*hello*world");`
  - Enables queries to use syntax like "*oracle* or *graph*"
Support for Cytoscape Open Source Visualization
Integration with Tom Sawyer Perspectives via property graph REST APIs
Integration with Tom Sawyer Perspectives (2) via property graph REST APIs
Integration with Tom Sawyer Perspectives (3) via property graph REST APIs
Program Agenda with Highlight

1. Graph Data Management and Analysis
2. Graph Data Model and BDSG Architecture
3. In-memory Analyst (PGX)
4. What’s New
5. Demos
Parallel In-Memory Graph Analyst

- An in-memory, parallel framework for fast graph analytics
  - Read a graph from NoSQL or HBase
  - Handles **analytic** workloads while the data access layer handles transactional workloads
  - Supports multiple users/graphs
  - Dozens of graph analysis functions
### Social Network Analysis Algorithms (1)

<table>
<thead>
<tr>
<th>Structure Evaluation</th>
<th>Ranking</th>
</tr>
</thead>
<tbody>
<tr>
<td>• Conductance</td>
<td>• closenessCentralityUnitLength</td>
</tr>
<tr>
<td>• countTriangles</td>
<td>• degreeCentrality</td>
</tr>
<tr>
<td>• inDegreeDistribution</td>
<td>• eigenvectorCentrality</td>
</tr>
<tr>
<td>• outDegreeDistribution</td>
<td>• Hyperlink-Induced Topic Search (HITS)</td>
</tr>
<tr>
<td>• partitionConductance</td>
<td>• inDegreeCentrality</td>
</tr>
<tr>
<td>• partitionModularity</td>
<td>• nodeBetweennessCentrality</td>
</tr>
<tr>
<td>• sparsify</td>
<td>• outDegreeCentrality</td>
</tr>
<tr>
<td>• K-Core computes</td>
<td>• pagerank</td>
</tr>
<tr>
<td>Community Detection</td>
<td>• personalizedPagerank</td>
</tr>
<tr>
<td>• communitiesLabelPropagation</td>
<td>• randomWalkWithRestart</td>
</tr>
</tbody>
</table>

- personalizedPagerank
- randomWalkWithRestart
- approximatePagerank
- weighted Pagerank
Social Network Analysis Algorithms (2)

**Pathfinding**
- fattestPath
- shortestPathBellmanFord
- shortestPathBellmanFordReverse
- shortestPathDijkstra
- shortestPathDijkstraBidirectional
- shortestPathFilteredDijkstra
- shortestPathFilteredDijkstraBidirectional
- shortestPathHopDist
- shortestPathHopDistReverse

**Recommendation**
- salsa
- personalizedSalsa
- whomToFollow

**Classic - Connected Components**
- sccKosaraju
- sccTarjan
- wcc
“No Coding” Graph Analysis

Degree Centrality

Page Rank

Betweenness Centrality

Community Detection

\[
\text{heroInfluence} = \text{analyst.inDegreeCentrality}()
\]

\[
\text{heroPR} = \text{analyst.pageRank().topK}(15)
\]

\[
b = \text{analyst.betweenness().topK}(15)
\]

\[
\text{comic_coms} = \text{analyst.communities}()
\]
Computational Analytics: Built-in Package

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 Specification

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

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

- 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

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

... and parallel graph mutation operations

- The original graph
  - Left Set: “a,b,e”
  - Create Bipartite Graph
  - Create Undirected Graph
  - Filter-Expression
  - Sort-By-Degree (Renumbering)
  - Simplify Graph

- Filtered Subgraph
  - e b d a i a j c g h
Program Agenda with Highlight

1. Introduction
2. Graph Data Model and BDSG Architecture
3. In-memory Analyst (PGX)
4. What’s New
5. Demos
## What’s New: Property Graph Features
### Big Data Spatial and Graph 2.0

**Faster, more powerful and scalable**

| • Integration with Apache Spark | • Vertex Label Support |
| • PGQL: Declarative Graph Query Language | • Node.js Client Support |
| • Distributed In-memory Graph Analysis | • Many new SNA Algorithms |
| • Hortonworks 2.4; Apache Solr 5.2x | • Data type support: long, char, byte, short, spatial |
| • Conversion of CSV & Relational data to Graph | • and many more… |
Oracle Spatial and Graph Property Graph Architecture

Graph Analytics
- Parallel In-Memory Graph Analytics (PGX)

Access Layer
- Apache Blueprints & Lucene/SolrCloud

Oracle Database 12.2

Java APIs

REST/Web Service

Java, Groovy, Python, ...

Property graph formats supported
- GraphML
- GML
- Graph-SON
- Flat Files
- CSV
- Relational Data Sources

Java APIs/JDBC/SQL/PLSQL

Parallel In-Memory Graph Analytics (PGX)
Property Graph on Oracle Database
Oracle Spatial and Graph: Database 12.2

• In-Memory Analyst (PGX) bundled with Oracle Spatial and Graph
• Oracle Text and Apache Lucene/Solr Cloud integration
• SQL-based navigation and graph query
• In-database graph analytics
  – Sparsification, shortest path, page rank, triangle counting, WCC, sub graph generation...
• RDBMS benefits: Parallel load, parallel query, B-tree index, compression (columnar, ACO, etc), Label security and authentication, Database In-memory
• Spatial support in property graphs
• Integration with Oracle R, Oracle Data Miner, Oracle Business Intelligence
Oracle Differentiators -- Graph

• **Complete, Supported, Graph Solution:**
  – **Storage**: NoSQL, Hbase, RDBMS back-ends
  – **Data Access**: Blueprints, Java, Property Graph Query Language (PGQL)
  – **Rich Graph Analytics**: 40 pre-built, in-memory graph algorithms

• **Scalable:**
  – Analyze 20-30 billion edge graph in memory on single BDA node
  – Persist extremely large graphs on disk

• **Security**: Secure NoSQL, Kerberos CDH, Label Security on Oracle Database 12c

• **10-50x Faster** than graph analysis competitors
Program Agenda with Highlight

1. Introduction
2. Graph Data Model and BDSG Architecture
3. In-memory Analyst (PGX)
4. What’s New
5. Demos
Demo: Tax Fraud Detection and Visualization with Oracle Big Data Spatial and Graph (BDSG) Property Graph
BIWA + Spatial Summit 2017
Jan. 31-Feb. 2, 2017, Oracle HQ, Redwood Shores, CA

- Premier educational event – spans Big Data, Analytics, Warehousing, Spatial + Graph, Cloud, IoT technologies
- 3 days / 6 tracks of technical sessions & hands on labs from Oracle developers and partners, customer use cases
- Sessions include
  - Architecture Live - Designing an Analytics Platform for the Big Data Era
  - Use Oracle Big Data SQL to Analyze Data Across Oracle Database, Hadoop, and NoSQL – Hands On Lab
  - Oracle Streaming Big Data and Internet of Things Driving Innovation
  - Analysing the Panama Papers with Oracle Big Data Spatial and Graph
  - A Shortest Path to Using Graph Technologies: Best Practices in Graph Construction, Indexing, Analytics and Visualization
  - Getting Started with Maps in OBIEE, BI Cloud Service and Data Visualization Desktop
  - Bringing Location Intelligence To Big Data Applications on Spark, Hadoop, and NoSQL – Hands On Lab
  - Using Machine Learning to unlock the Business Value in Big Data |Taking R to new heights for scalability & performance
  - ... and many more...

Visit www.biwasummit.org
Discount Registration Code for SFDAMA: BIWADAMA
Resources

• Oracle Big Data Spatial and Graph OTN product page: www.oracle.com/technetwork/database/database-technologies/bigdata-spatialandgraph
  – White papers, software downloads, documentation and videos

• Oracle Big Data Lite Virtual Machine - a free sandbox to get started: www.oracle.com/technetwork/database/bigdata-appliance/oracle-bigdatalite-2104726.html

• Hands On Lab for Big Data Spatial: tinyurl.com/BDSG-HOL

• Blog – examples, articles & tips: blogs.oracle.com/bigdataspatialgraph

• Oracle By Example tutorials: www.oracle.com/goto/oll
  (search “Big Data Spatial and Graph”)

• @OracleBigData, @SpatialHannes, @JeanIhm   Oracle Spatial and Graph Group
Q&A
Announcing Oracle Database 12c Release 2 on Oracle Cloud

Available now

• Database Cloud Services
• Exadata Cloud Service
• Exadata Express Cloud Service

Oracle is presenting features for Oracle Database 12c Release 2 on Oracle Cloud. Features and enhancements related to the on-premises version of Oracle Database 12c Release 2 are not being announced at this time.
Spatial and Graph comparison: Spatial Features

**Big Data Spatial**
- Batch processing (MapReduce Java functions) for data categorization, filtering and data preparation
- Data Enrichment Services for any data with location attributes
- 2D, 3D, and Raster Analysis
- Spatial index for data access
- Supports application data in HDFS
- HTML5 Map Visualization

**Oracle Database Spatial and Graph**
- Full Oracle ACID transactions
- High performance concurrent parallel SQL query, update, insert
- Support for 2D, 3D, and Raster Analysis, Topology, Network Analysis, Geocoding, Routing, OGC Web Services, Linear referencing; Supported by all major GIS tools
- Spatial data types and indexes
- Exadata optimizations