Graph and Spatial Analytics
Built for Big Data Platforms

Jim Steiner
Vice President
Server Technologies
October 28, 2015
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.
Oracle’s Spatial and Graph Strategy

Enable Spatial and Graph use cases on every platform

Oracle Big Data Spatial and Graph

Oracle Database Spatial and Graph

Spatial and Graph in Cloud Offerings

Big Data:
Single Model Data Store

Database 12c:
Polyglot (Multi-model) Data Store
Hadoop
- Data in files, NoSQL
- Parallel programming model
- Scalable processing
- Ideal for high volume, unstructured data storage and processing

NoSQL
- Data organized for fast query
- Simple programming models
- Elastic scale out
- Little or no fixed data model
- Access specific record
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?
Big Data Spatial and Graph

Big Data often needs an Organizing principle – Data Harmonization

Big Data analysis is often about relationships not aggregation

Big Data platform is economically compelling for working with massive data sets found in spatial and graph workflows
Oracle Big Data Spatial and Graph (BDSG)

Spatial Analysis Features

Property Graph Database
Graph Analysis
Who is most important? 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!
What Big Data problems can Graphs address?

- **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

---

Copyright © 2015, Oracle and/or its affiliates. All rights reserved.
Modeling and Analyzing The Internet of Things

- Massive scale – Trillions of Connections
- Dynamic Network
- Includes metadata as well as events/enriched data
- Extensible by other data source
Modeling and Analyzing The Internet of Things

Cyber-Security
- Critical / Alternate Path Analysis
- Community Detection
- Network Monitoring
- Predictive Analysis

Multiple System Impact Analysis
- Transportation
- Utilities
- Finance
Big Data Spatial and Graph
Property Graph
Big Data Graph Architecture

Lightning-Fast In-Memory Analytics
- YARN Container
- Standalone Server
- Embedded

Massively Scalable Graph Store
- Oracle NoSQL
- HBase
Scales to Trillions of Edges

- Leverage HBase or Oracle NoSQL for massive storage scalability
- Use underlying engine for node/edge access and graph filtering
In-Memory Analyst

1. Analyst engine runs on one or more servers
2. Users request loads subgraph of interest into engine
3. Users call simple functions for complex analyses
   Write results back or share on-the-fly

Analyst A

Graph Database

Analyst B
Multiple Interfaces for Many Kinds of Users

In-Memory Analyst

Graph Database

SQL
35 Graph Functions

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

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

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

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

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

Other Classics
- Vertex Cover
- Minimum Spanning-Tree(Prim’s)
Coding Graph Analysis Can Be Complicated

Find the size of the 2-hop network of superheroes (Gremlin+Python)

```python
sum([v.query() \ .direction(blueprints.Direction.OUT).count() \ for v in OPGIterator(v0.query() \ .direction(blueprints.Direction.OUT) \ .vertices().iterator())])
```

Computing “importance” is more nebulous and harder to code
Importance as **Degree Centrality**

- The more **edges** a vertex has, the higher its **degree**

- The greater the degree, the more important the vertex is

- This is one way to look at importance

- Is your most connected customer most important?
Degree Centrality in Big Data Graph

Code

```python
heroInfluence = analyst.inDegreeCentrality()
```
Importance as **Page Rank**

- Importance can flow **through** a graph

- A node connected to by important nodes is **also** important

- This is importance as a measure of
  - Trust
  - Prominence

- Thinking about customers in a graph requires multiple definitions of importance
Page Rank in Big Data Graph

Code

```java
heroPR = analyst.pageRank().topK(15)
```
Importance as “Between-ness”

- Importance can be how often you’re on the critical path

- **Betweenness** is the number of shortest paths a node is part of
Betweenness Centrality in Big Data Graph

Code

\[ b = \text{analyst.betweenness().topK(15)} \]
Communities Are Just Special Subgraphs

- Community
  - A subgraph in which
    - Nodes are more connected to each other

- Communities provoke interesting questions
  - How many?, How large?
  - How do they relate to each other?

- Analyses can be performed on a community
  - Who’s the most valuable customer in a community?
Community Detection in Big Data Graph

Code
comic_coms = analyst.communities()
Differentiators -- Graph

• Commercial, supported software
• “Best of Both Worlds” Graph DB
  – In-memory graph analysis algorithms – Like Neo4J
  – Distributed graph database model – Like Datastax Titan
• Dozens of pre-built in-memory graph analysis algorithms
• 10-50x faster analytics than competitors’ offerings
• Analyze 20-30 Billion edge graph in memory on a single BDA node
Motivation for Oracle Big Data Spatial

- Emergence of Hadoop for spatial analysis in business and spatial workflows
- Existing Hadoop-based Spatial technologies are GeoSpatial-centric not Application-centric
- No significant commercial offerings
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.
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)
Use Case: Linking Information by Location

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.

Data Products for Rate Structures Underwriting/Risk Analysis

Source: “The big data: How data analytics can yield underwriting gold.” Survey conducted by Ordnance Survey and Chartered Insurance Institute, 25 April 2013.
What features does Big Data Spatial have?

- **Data enrichment service API using GeoNames and geometry hierarchy data**
- **Spatial processing of data stored in HDFS. Raster processing operations: Mosaic and sub-set operations. Geodetic and Cartesian data**
- **MapReduce routines for distance calculations, PointInPolygon, buffer creation, Categorization, KMeansClustering, Binning, etc.**
- **HTML5 Map Visualization API**
Store any business data with spatial information in HDFS

Oracle JGeometry

USER-PROVIDED InputFormat/RecordReader Class

LOAD ANY FORMAT

ANY BUSINESS DATA

HDFS
## Vector Data Processing API Functions

<table>
<thead>
<tr>
<th>Single Geometry</th>
<th>Geometry Pairs</th>
<th>Categorization and Enrichment</th>
</tr>
</thead>
<tbody>
<tr>
<td>• Length</td>
<td>• Range Queries</td>
<td>• Associate a data set with a known geometry or named hierarchy</td>
</tr>
<tr>
<td>• Area</td>
<td>• Point in Polygon</td>
<td>– Process all Tweets for a period of time and count how many are associated with each city, county, state, etc.</td>
</tr>
<tr>
<td>• Buffer</td>
<td>• Touch, Overlap, Intersect, Contains, Any Interaction</td>
<td></td>
</tr>
<tr>
<td>• Simplify</td>
<td>• Join Queries</td>
<td></td>
</tr>
<tr>
<td></td>
<td>• Interactions on sets of data</td>
<td></td>
</tr>
<tr>
<td></td>
<td>• E.g.: Find all the dropped cell calls in all coverage areas</td>
<td></td>
</tr>
</tbody>
</table>
Data Categorization Services

Any hierarchical geometry data set for reference

Customers choose a set of layers For example, they can select (continents, countries, cities) or (countries, states, counties) as the hierarchy

Big Data Spatial map-reduce job processes the customer data and produces a result file
Spatial Index for Spatial Queries

Spatial Index stored as a MapFile

HDFS

MapReduce Job with Index

Copy the index to distributed cache

Mapper reads the index data for the corresponding HDFS block

Process only those records that return hits from the index search
How is Spatial Indexing used?

• Performance of some of the Spatial processing operations can be improved by using spatial indexes.

• When the map-reduce job is created, user can specify whether to use or not use the spatial index.

• Spatial index data is placed in Hadoop Distributed Cache for fast access.
Spatial Server Console

- A sample J2EE application that can be deployed in Jetty
  - Create spatial indexes on data that is already loaded into HDFS
  - Run Hadoop jobs to do spatial processing
    - This will create and run the categorization MapReduce job that can use a provided spatial index or read each input record
  - Sample map application that can display results of a categorization job
Image Server

• HDFS storage for the image or raster files
  – We can support dozens of file formats (GDAL supported formats)
  – Images are geo-referenced
  – Images can be in different coordinate systems and resolutions

• Three main capabilities
  – Loader to load raster data from NFS to HDFS
  – Mosaic and subset operations based on a virtual mosaic
  – Image processing framework for raster analysis
Big Data Spatial and Graph
Spatial Raster Processing Framework

Raster Analysis Application
Sample Console Application

Raster MapReduce framework

S&G Java API
Analysis Algorithms

Subset
Mosaic
Raster Analysis API

HDFS

Raster catalog
Derived Rasters

Rasters

GDAL Loader (GDAL formats)

Raster Data Files on NFS

Customer data
Generated data
Oracle Provided
Customer code

ORACLE
Why Spatial and Graph Together?
GEO-SOCIAL NETWORKS (GeoSNs)

Social network functionality + Location-based services = Geo-Social Query

My Friends in range

Courtesy: 14th International Symposium on Spatial and Temporal Databases
Drimitris Papadias (HKUST, Hong Kong)
CHALLENGE

Social Relations (Friendships)

Online Task
Geo-Social Query Processing

Geographical Information (current check-ins)

Large ☹
Complex ☹
Relatively Static 😊

Small 😊
Simple 😊
Dynamic ☹

Courtesy: 14th International Symposium on Spatial and Temporal Databases
Drimitris Papadias (HKUST, Hong Kong)
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.
Q&A
Multimedia Analytics for Big Data Applications
Video and Image
Multimedia Analytics v1
Feature of Oracle Big Data Spatial and Graph

• Toolkit to process and analyze video and image data in Hadoop

• High speed processing that leverages Hadoop parallelism

• Massive storage and InfiniBand network of Oracle Big Data Appliance enables processing and management of petabytes of multimedia data
Feature Benefits

• Scalable, high speed processing

• Flexible framework for plug-in of customer/partner software
  – Enables custom analysis of video and images

• Easy downstream integration with analytics and BI tools, with database
  – Enable inclusion of multimedia data sources to Big Data applications
  – Comprehensive analysis and richer insights in Big Data applications
Multimedia Analytics v1

• Detect and identify faces
  – OOTB with OpenCV

• Integrate third party software for
  – License plate recognition
  – Object recognition
  – ... and other custom applications

• Enable partners to build solutions
Potential Use Cases

• Face recognition (OOTB)
  – Law enforcement
  – Security
• With custom code
  – Traffic management
  – Drone video analysis
  – Retail
    • Customer in-store behavior
  – Oil and Gas
    • Ex: Automated analysis of images of underwater pipes
Example: Face Recognition (On four nodes)

Identify faces in video

Video Analysis Output: Face labels, timestamps

Big Data Discovery

OBIEE

Oracle Big Data Appliance

OpenCV face recognition module

Other modules can be plugged in
Meet the Experts Sessions
Full list at tinyurl.com/oowspatialgraph

+ Meet the Experts: Oracle Spatial and Graph [MTE9565]
Daniel Geringer, Spatial Solutions Architect, Oracle
Siva Ravada, Senior Director of Development, Oracle
Marco Mantovani, ICONSULTING
Nick Salem, Distinguished Engineer, Neustar
Xavier Lopez, Senior Director, Product Management, Oracle
Tuesday, Oct 27, 6:15 p.m. | Moscone South—306

+ How to Use Graph and Spatial Technologies: Practitioners and Experts Speak [CON10317]
Hassan Chafi, Director, Research & Advanced Development, Oracle
Steven Pierce, CEO, Think Huddle, LLC
Jayant Sharma, Director, Product Mgmt, Oracle
James Steiner, Vice President, Product Management, Oracle
Alan Wu, Architect, Oracle
Thursday, Oct 29, 12:00 p.m. | Moscone South—254
Demos
Full list at tinyurl.com/ooowspatialgraph

**MONDAY & TUESDAY**
10:15 – 6,
**WEDNESDAY**
10:15 – 4:15

- **Graph Analytics and Data Management for Big Data and Oracle Database**
  Workstation ID: SLD-040
  Venue: Moscone South, Upper Left, Database

- **Location Analytics and Spatial Data Management for Big Data and Oracle Database**
  Workstation ID: SLD-043
  Venue: Moscone South, Upper Left, Database

- **Big Data Graph Analytics for the Enterprise**
  Workstation ID: SBD-018
  Venue: Moscone South, Upper Right, Big Data Showcase
The Spatial & Graph SIG User Community

- Connect & exchange knowledge via online communities, conferences & events
- **Join us online**
  - Find us at [LinkedIn](https://www.linkedin.com), [Google+](https://plus.google.com) & [IOUG SIG](http://www.ioug.org)
  - Search for “Oracle Spatial Community”

- **Meet us at OOW** – Join us for drinks and appetizers before the concert!
  - Wednesday at 4:30 to 6 p.m. at Sam’s Grill (374 Bush St. at Kearny – a brief walk north of Moscone)
  - RSVP with a board member or Oracle team
Oracle Spatial Summit at BIWA 2016
Jan. 26-28 | Oracle HQ Conference Center

Business Intelligence, Warehousing, and Analytics + Spatial & Graph
IOUG Special Interest Groups (SIGs)

*Register by November 15th for Early Bird rates

*Final Call for Speaker Abstracts – Submit by November 2

www.biwasummit.org
Session Surveys

Help us help you!!

• The organizing committee OW would like to invite you to take a moment to give us your session feedback. Your feedback will help us to improve your conference.

• Please be sure to add your feedback for your attended sessions by using the Mobile Survey or in Schedule Builder.