Oracle Big Data Spatial and Graph
An Overview

July, 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.
Agenda

1. Introduction to Big Data Spatial and Graph
2. Big Data – Graph Features
3. Big Data – Spatial Features
4. Resources
5. Q & A
Oracle’s Spatial and Graph Strategy

Enable Spatial and Graph use cases on every Big Data platform

Oracle Big Data Spatial and Graph

Oracle Database Spatial and Graph

Spatial and Graph in Cloud Offerings
Oracle Big Data Spatial and Graph

Property Graph for Analysis of:
• Social Media relationships
• Internet of Things interactions
• Cyber-Security

Spatial Analysis Features for:
• Location Data Enrichment
• Proximity and containment analysis
• Preparation of digital map and imagery data sets
Conventional database or Big Data technologies

Typical technical decision criteria

- Tooling maturity
- Stringent Non-Functionals
- ACID transactional requirement
- Security
- Variety of data formats
- Data sparsity
- ETL simplicity
- Ingestion rate
- Cost effectively store low value data
- Straight Through Processing (STP)

Hadoop
Relational
The Big Picture – Oracle Big Data Management System

<table>
<thead>
<tr>
<th>DATA RESERVOIR</th>
<th>DATA WAREHOUSE</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cloudera Hadoop</td>
<td>Oracle Database</td>
</tr>
<tr>
<td>Oracle Big Data SQL</td>
<td>In-Memory, Multi-tenant</td>
</tr>
<tr>
<td>Oracle NoSQL</td>
<td>Oracle Industry Models</td>
</tr>
<tr>
<td>Oracle R Distribution</td>
<td>Oracle Advanced Analytics</td>
</tr>
<tr>
<td>Oracle Big Data Spatial and Graph</td>
<td>Oracle Spatial and Graph</td>
</tr>
<tr>
<td>Big Data Appliance</td>
<td>Exadata</td>
</tr>
<tr>
<td>Oracle Event Processing</td>
<td>Oracle Data Integrator</td>
</tr>
<tr>
<td>Apache Flume</td>
<td>Oracle GoldenGate</td>
</tr>
<tr>
<td>Oracle GoldenGate</td>
<td>Oracle Event Processing</td>
</tr>
</tbody>
</table>

SOURCES
Program Agenda

1. Introduction to Big Data Spatial and Graph
2. Big Data – Graph Features
3. Big Data – Spatial Features
4. Resources
5. Q & A
Why Graph Databases Now?

Rise of social networking
Google, Yahoo, Twitter, Facebook, Linked In

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
- **Health Care**: CRM, fraud detection

Modeling complex relationships as graphs is efficient
- Improves performance
- Simplifies queries, traversal, search and analytics
Graph Data Models

**Property Graph Model**
- Graph Data Management
- Social Network Analysis
- Entity analytics

- National Intelligence
- Public Safety
- Social Media search
- Marketing - Sentiment

**Network Data Model**
- Network path analysis
- Transportation modeling

- Logistics
- Transportation
- Utilities
- Telcos

**RDF Data Model**
- Data federation
- Knowledge representation
- Semantic Web

- Life Sciences
- Health Care
- Publishing
- Finance

---

Use Case | Graph Model | Industry Domain
---|---|---

**Social Network Analysis**

**Spatial Network Analysis**

**Linked Data / Metadata Layer**
Graph for Social and Unstructured Data Analysis

Graph is a powerful tool for Data Analysis

By representing your data as a graph with relationships between data entities

When analyzing such a graph, you are using explicit relationships to find implicit information about your data

Without computing multiple joins

Individual relationships are represented as links
Graph Analysis Examples

• Attribute searching (Get people with a given name)
• Node/edge adjacency (Get people that like a given Web page)
• Fixed-length paths (Get the friends of the friends of a given person)
• Reach-ability (Is there a “friend” connection between two people?)
• Pattern matching (Get the common friends between two people)
• Aggregates (Get the number of friends of a given person)
Common Graph Analysis Use Cases

- **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

**Product Recommendation**
- Purchase Record
- Customer items

**Influencer Identification**
- Communication Stream (e.g. tweets)

**Community Detection**
- Graph patterns

**Graph Pattern Matching**
- Graph patterns
Property graph model
- Dynamic construction of IP network
- The graph includes metadata as well as events/enriched data
- Extensible by other data source (add properties, relations)
- Search – Text search on graph DB properties
Graph Solution Workflows and Characteristics

• Graph Data Management
  – Raw business data is converted to graph format and persisted as HDFS
  – Graph queries on HDFS or NoSQL using Java REST APIs

• Analysis and Exploration (in-memory analysis engine)
  – 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, …
Enterprise Requirements

**Performance**
- Distributed Processing: Hbase, NoSQL load, index, query, search
- Parallel, in-memory graph analytics

**Scalability**
- Horizontal scalability
- Concurrency: multiple users and analytic operations on one or more in-memory graphs
- Filtering to refine in-memory graph requirements

**Ease of Programming**
- Easy to use parallel & distributed analytics
- Ease of graph data modeling
- Popular open source Java and REST APIs

**Manageability**
- Integration with Hbase and Oracle NoSQL database features for indexing, sharding, fault tolerance, availability, installation, and management
- Option to Oracle BDA
Big Data Graph Detail
Graph Data Model

What is a graph?
- A set of links and nodes (and optionally attributes)
- A graph is simply **linked data**

Why do we care?
- Graphs are everywhere
  - Social networks/Social Web (Facebook, Linkedin, Twitter, Baidu, Google+,...)
  - Cyber networks, power grids, protein interaction graphs
  - Knowledge graphs (IBM Watson, Apple SIRI, Google Knowledge Graph)
- Graphs are intuitive and flexible
  - Easy to navigate, easy to form a path, natural to visualize
  - Do not require a predefined schema
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
Big Data Spatial and Graph (BDSG)

**Property Graph Features**

- Highly scalable graph database and analytics engine
- Implemented on Apache HBase and Oracle NoSQL Database
- Rich developer APIs
  - Blueprints, REST, Java graph plus support for Groovy, Python, PHP, Perl, Ruby, and JavaScript
- Fast, scalable suite of social network analysis functions
  - Ranking, centrality, recommender, community detection, path finding...
  - Targeted to address main industry requirements
- Manageability
  - Bulk load
  - Console to execute Java and Gremlin APIs
Big Data Graph Architecture

Graph Analytics
In-memory Analytic Engine

Graph Data Access Layer API
Blueprints & SolrCloud / Lucene

Scalable and Persistent Storage
Property Graph Support on
Apache HBase and Oracle NoSQL

Java APIs

Python, Perl, PHP, Ruby,
Javascript, ...

REST Web Service

Java APIs
Support for Open Source TinkerPop Graph Tool Stack

Oracle Big Data Spatial and Graph Blueprints API implementation provides support for the de-facto graph database standard TinkerPop component stack.

These include query language, dataflow, REST APIs, and others.
Data Format Support

• GML, GraphML, GraphSON

• Oracle-defined Property Graph flat files
  – Vertex file, Edge file
  – Support basic data types + Date with Timezone + Serializable objects
  – Allow multiple data types to be associated with one key
  – UTF8 based

```
1,name,1,Barack%20Obama,,
1,age,2,,53,
1,likes,1,scrabble,,
1,likes,5,,2009-01-20T00:00:00.000-05:00
1,occupation,1,44th%20president%20of%20United%20States%20of%20America,,
```
Text Search through Apache Lucene/Solr

• Integration with Apache Lucene/Solr
• 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*”
In-Memory Graph Analysis Framework

• Large graph analysis is time-consuming because ...
  – The computation typically involves touching most nodes and edges in the graph
  – The data-access pattern is random
• **In-memory, parallel** framework for fast graph analytics
• Exploits the architecture of modern servers
  – The computation is parallelized using multiple CPU cores
  – The non-sequential data-access is mitigated with large DRAMs
• J2EE container support (WLS, Tomcat, Jetty)
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)
Unique Graph Filtering Operations

• Graph analysis engine reads graph into memory from HBase or NoSQL
  – Can reach memory limit for huge graphs
  – Subgraph mechanism to address this
• Data Access Layer filtering used to create subgraph for Property Graph Engine analytics
• Persisted graph can still be modified through Java and REST APIs
  – Changes can be propagated to in-memory graph
Program Agenda

1. Introduction to Big Data Spatial and Graph
2. Big Data – Graph Features
3. Big Data – Spatial Features
4. Resources
5. Q & A
What is Spatial Data

Integral part of almost every database

• Business data that contains or describes location
  – Geographic features (roads, rivers, parks, etc.)
  – Assets (pipe lines, cables, transformers
  – Sales data (sales territory, customer registration, etc.)
  – Street and postal address (customers, stores, factories, etc.)

• Anything associated with a physical location
• Described by coordinates or implicitly as text (place name), ...
• Location is a “universal key” relating otherwise unrelated entities
Oracle Big Data **Spatial** Overview

- Oracle Big Data Spatial and Graph on Apache Hadoop is a framework that uses the MapReduce programs and analytic capabilities in a Hadoop cluster to store, access, and analyze the spatial data.

- The spatial features provide a **schema** and **functions** that facilitate the storage, retrieval, update, and query of collections of spatial data.

- The spatial data is loaded for query and analysis by the Spatial Server and the images are stored and processed by an Image Processing Framework.
Oracle Big Data – Spatial Features

• Geo-enrichment for Data Harmonization
  – Resolution of location-related information
  – Determination of location hierarchies

• Categorization and filtering
  – Tracking, proximity analysis, geo-fencing and categorization based on location

• Data preparation
  – Large scale geoprocessing for cleansing, preparation of imagery, sensor data, and raw data input

• Data visualization
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)
Oracle Big Data **Spatial** Use Cases

<table>
<thead>
<tr>
<th>Industry</th>
<th>Usage</th>
</tr>
</thead>
<tbody>
<tr>
<td>Utilities</td>
<td>Smart grids, dynamic demand response, network utilization</td>
</tr>
<tr>
<td>Financial services, Insurance</td>
<td>Fraud analysis, origin-destination/flow, site planning, demographic analysis, risk assessment</td>
</tr>
<tr>
<td>Telecomm</td>
<td>Network monitoring and planning, location based advertising, tracking/location based services</td>
</tr>
<tr>
<td>Transportation, logistics</td>
<td>Asset tracking, fleet management, service planning</td>
</tr>
<tr>
<td>Retail</td>
<td>Site planning, location-based marketing</td>
</tr>
</tbody>
</table>
Use Cases: Geo-Enrichment

(i) Wireless network performance (dropped calls, utilization): aggregate and display geocoded Call Data Records, sensor data, other sources for analysis

(ii) Transportation origin-destination analysis: combine transit card/payment info and other sources to determine where (and how many) people travel to, starting from any station on a transit network

(iii) Geotagged Twitter: where are the tourists and locals tweeting
Use Cases: Categorization, Filtering, Aggregation

The one million tweet map

Irvington, NJ: Cluster and Outlier Analysis
Risk Cluster Types
- All other places
- High-High
- Low-Low
Use case: Data preparation

Mosaic images

Terrains and contours

Pyramiding: layers at different resolution

Shaded reliefs
Overview of Spatial Features

• Vector Data Processing
  – Support spatial processing of data stored in HDFS
  – Commonly used operations like pointInPolygon, buffer creation, distance calculations, anyinteract operations, etc.
  – Supports both Geodetic and Cartesian data
  – Data enrichment services using GeoNames and geometry hierarchy data
  – Map visualization API (HTML5)

• Raster Data Processing
  – GDAL to load raster data onto HDFS
  – Raster processing operations: Mosaic and sub-set operations
  – MapReduce framework for raster analysis operations (for example, calculate the slope at each pixel based on the DEM)
Big Data Spatial and Graph
Spatial Vector Processing Framework

Mapper and Reducer Classes

Customer Application

MapReduce Framework, templates

Sample Application

S&G Java API

Spatial Operators, Functions

Spatial Enrichment, Categorization API

GeoJSON, JGeometry format

GeoNames and Hierarchy data

RecordReader class

HDFS

Geospatial Vector data (any format)

Enrichments, Categorizations results

Customer data

Generated data

Oracle Provided

Customer code
Oracle Big Data Spatial Features (Vector Services) cont.

Build Map Visualization using HTML5-based API

• Includes an HTML5-based map client API for developers
  – Default world boundary data provided as JSON files
  – The map view can display any data stored in GeoJSON files
  – Browser based, rich interaction
Vector spatial data storage in HDFS

• Customers load their data into HDFS using a loader of their choice
  – We do not require the data to be in a format that we specify
  – This makes it easy for customers to use any data format their applications prefer
  – And the data can have other business data and not just spatial data

• We require the customer to provide a RecordReader class
  – This class reads the customer data record and produces an instance of JGeometry or a GeoJSON instance
  – With this model we can support any data format customers uses for their data
Oracle Big Data Spatial Features (Raster Services)

• HDFS storage for the image or raster files
  – We can support dozens of file formats
  – Images are georeferenced
  – Images can be in different coordinate systems and resolutions

• Three main capabilities
  – GDAL-based loader to load raster data from NFS to HDFS
  – Mosaic and subset operations
  – Image processing framework for raster analysis
Oracle Big Data Spatial Features (Raster Services) cont.

• Console to view the set of images that are available
  – Map displays all available images using image footprints
  – Users can zoom into different areas of the world to see available images for any region
  – Information about the source and other spatial information for each image can also be displayed
  – Ability to group sets of images into groups for further processing
  – Images can be given priority in a group based on date, resolution, etc.
Create Index on spatial data in HDFS
Console

Run Map Reduce job to perform categorization based on spatial hierarchy
Results in Console
“Tweets in May by State”
Program Agenda

1. Introduction to Big Data Spatial and Graph
2. Big Data – Spatial Features
3. Big Data – Graph Features
4. Resources
5. Q & A
Resources

• Oracle Big Data Spatial and Graph on Oracle.com: https://www.oracle.com/database/big-data-spatial-and-graph


• Blog (technical examples and tips): https://blogs.oracle.com/bigdataspatialgraph/

• Big Data Lite Virtual Machine (a free sandbox environment to get started): http://www.oracle.com/technetwork/database/bigdata-appliance/oracle-bigdatalite-2104726.html
Q&A
Hardware and Software
Engineered to Work Together