# Oracle Big Data Spatial and Graph

Oracle Big Data Spatial and Graph offers a set of analytic services and data models that support Big Data workloads on Apache Hadoop, Oracle NoSQL Database, and Spark technologies.  For over a decade, Oracle has offered leading spatial and graph analytic technology for the Oracle Database.  Oracle is now applying this expertise to work with social network data and to exploit Big Data architectures.

Oracle Big Data Spatial and Graph has three components: a distributed property graph with 40 high-performance, parallel, in-memory analytic functions; a wide range of spatial analysis functions and services to evaluate data based on how near or far something is to one another, whether something falls within a boundary or region, or to process and visualize geospatial map data and imagery; and, a multimedia framework for processing video and image data in Apache Hadoop, such as facial recognition.

Using the graph features, analysts can discover relationships and connections among customers, organizations, and assets.  With the spatial capabilities, users can achieve insight into location-based patterns and trends across big data volumes, harnessing inherent location relationships in disparate data sources through harmonization and enrichment.

## KEY BUSINESS BENEFITS

- Reduces the complexities of Hadoop development and time to implementation
- Commercial-grade spatial and graph algorithms enable deeper insights into Big Data workloads
- Adds new dimensions to discovery of relationships and patterns among customers and prospects using social network data

## Graph Data Management and Analysis

Much of the Big Data generated these days contains inherent relationships between the collected data objects. For example, important relationships and patterns are found in social network data from Facebook, a listener's music preferences from an online music service like Spotify, online shopper behavior on eBay, and bloggers and their relationship to followers and other bloggers. These relationships can be readily structured as a set of interconnected objects in a graph.  Graphs use data structures called vertexes and edges, and associated properties or attributes to model relationships.  Some graphs have an inherent spatial relationship, such as networks of roads, telecommunications, water, and other utilities.  Different graphs may have other kinds of relationships, for instance the connections among entities in the Internet of Things, and within biological pathways and social networks.

As described in Figure 1 below, graphs are easy to represent, and readily

intuitive to visualize. More importantly, a variety of machine-driven analytic processes can be applied to discover underlying relationships, yielding important insights from big data.
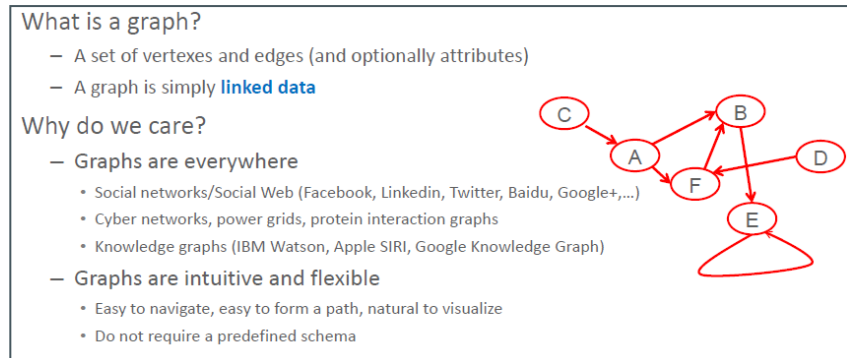


Figure 1. Simple Graph Data Model

## Graph Capabilities in Oracle Big Data Spatial and Graph

Oracle Big Data Spatial and Graph provides optimized data storage, querying and analysis of property graphs, a common graph model. It includes a data access layer and an in-memory analyst. A choice of databases (Oracle NoSQL and Apache HBase) provides distributed, scalable and secure graph management.

Oracle Big Data Spatial and Graph has a powerful and efficient in-memory analyst engine. Graph analysis traditionally has been time consuming because it routinely involves touching most of the nodes in the graph in a non-sequential (random) fashion. The in-memory analyst addresses this performance challenge by performing graph analysis in-memory, applying the parallelism inherent in modern system architectures and distributing, as needed a particular analysis across multiple instances of the in-memory analyst running on different nodes in the cluster.

Unlike other products that offer a short list of algorithms and primitives to code your own, the in-memory analyst comes pre-integrated with a rich set of nearly forty built-in social network analysis (SNA) algorithms that address typical graph analysis needs. Examples include graph traversal, recommendations, finding communities and influencers, and other pattern matching. Graphs in memory can also be queried for patterns with a declarative language.
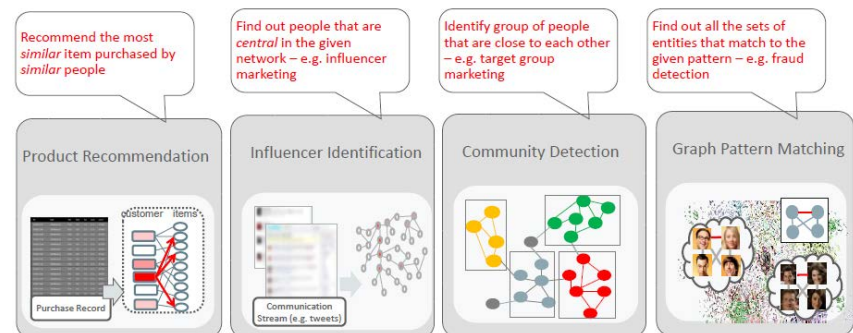


Figure 2. Common Graph Analysis Use Cases

For scalability and performance, the in-memory analyst engine relies on the data access layer to provide efficient scanning of the distributed graph database and filtering to identify nodes, edges and properties of interest that are read into memory for analysis.
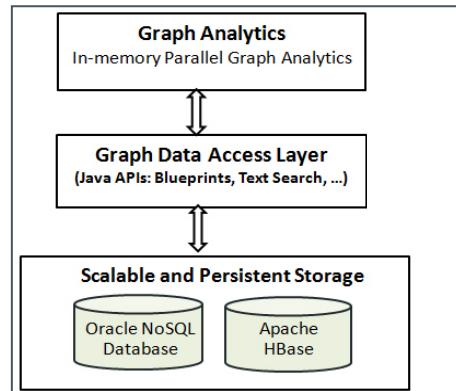


Figure 3. Oracle's Property Graph Architecture

Developers can use a variety of Java APIs to access graph data and perform graph operations. Other features include optimized support for parallel bulk load and export of property graph data, along with text search integration with Apache Lucene and SolrCloud.

## Spatial Analysis and Services

Spatial data is any data that represents or includes information about the location, size, or shape of something found in a geographic or geometric space. This can include features like parks, neighborhoods, addresses, land parcels, country, city, county, and postal boundaries, or GPS coordinates.  Indoor locations, such as shops inside a shopping mall, cubicles in office buildings, and stadium and arena layouts are also examples of spatial data. This kind of information is often displayed on maps.

There are two major categories of spatial data – vector and raster.  Vector data is represented by two-dimensional points, lines and polygons (for things like addresses, roads and boundaries respectively) or three-dimensional point and polygon data to represent surfaces, building outlines, and other objects that include a height dimension.

Raster data comprises a set of cells (or pixels) stored in a grid where each cell contains a value.  These values can represent the intensity of bands of light, moisture values, sound values, temperature and the results of sensor readings. Commercial usage of raster products includes satellite imagery, digital aerial photographs, agricultural soil reading and weather readings.

 Location information is a common element of Big Data, whether it's a text address, city, or landmark name in a Twitter feed, or latitude/longitude coordinates from a GPS sensor. Businesses can use this data as the basis for associating and linking disparate data sets, a concept known as data harmonization. Location information can also be used to track and categorize entities based on proximity to another person, place, or object, or on their

- Management interfaces and map visualization tool
- Support for large-scale spatial data processing (vector and raster)
- Support for spatial vector processing on Oracle NoSQL Database
- Hive support for spatial analysis and processing – developers can use SQL to analyze and process data on HDFS, NoSQL Database, or Apache HBase
- Access spatial big data sources from Oracle Database using Big Data SQL or Oracle SQL Connectors for Hadoop
- Spatial joins for vector processing
- Spatial clustering and binning
- Raster loader: support for multi-band images
- Raster simulator framework to simplify development of raster analysis classes without HDFS

*"With the explosion of Hadoop environments, the need to spatially-enable workloads has never been greater, and Oracle could not have introduced Oracle Big Data Spatial and Graph at a better time. This exciting new technology will provide value-add to spatial processing and handle very large raster workloads in a Hadoop environment. We look forward to exploring how it helps address the most challenging data processing requirements."*

**KEITH BINGHAM**
CHIEF ARCHITECT AND TECHNOLOGIST
BALL AEROSPACE

presence within a particular area. Location information enables a technique known as geo-fencing to support location-based advertising -- offering a shopping promotion to customers entering a nearby area.



Figure 4: Using location to relate diverse big data sources

Imagery and sensor data can be analyzed to support a variety of business benefits. Distributed sensors are generating vast amounts of raster imagery data in raw data formats that require large-scale geoprocessing for cleansing and preparation. Hadoop environments are ideally suited to storing and processing these high data volumes quickly, in parallel across MapReduce nodes.

## Spatial Capabilities in Oracle Big Data Spatial and Graph

Oracle Big Data Spatial and Graph gives developers and users a wide range of features and services to enable the use of the Hadoop data processing system, Oracle NoSQL Database, or Apache Spark for spatial data analysis.

The spatial features include support for data enrichment of location information; filtering and categorization based on distance and location-based analysis; spatial querying and analysis of Hadoop data with SQL; and vector and raster processing for data sets such as digital maps, sensor-generated information, and satellite and aerial imagery. The product also includes a rich set of APIs for map visualization.



Figure 5: Spatial binning analysis of social media data

Specific services for data enrichment, filtering and categorization, querying and analysis, and vector data processing include:

- Ability to associate documents and data containing place names, such as city or state names, or longitude/latitude information, with real-world location definitions and associated default administrative boundaries; support for custom geographic regions such as sales territories
- Native support for text-based 2D and 3D geospatial formats, including GeoJSON files, Shapefiles, GML, and WKT. For other non standard formats, a user defined RecordReader class can be used to read any data containing a geographic component.
- A Vector Analysis API (Java and Scala) for Apache Spark and Apache Spark SQL to create spatial RDDs (Resilient Distributed Datasets) for performing spatial transformations and actions, including IsInside, Contains, AnyInteract, WithinDistance, MBR, and Nearest Neighbors
- An HTML5-based map client API and a sample console to explore, categorize, view data in variety of formats, and coordinate systems
- Topological and distance operations: Anyinteract, Contains (or Inside), Distance calculation, Length calculation, Within Distance, Buffer, Point-in-polygon and others
- Spatial indexing for fast retrieval of data
- Spatial clustering and binning analysis
- Spatial join queries to detect spatial interactions between all records of two data sets (e.g., find all of the dropped cell phone calls in all coverage areas)
- Hive support for spatial analysis and processing, enabling developers to use SQL to analyze and process data on HDFS
- Spatial query and analysis from Oracle Database on HDFS, NoSQL or Apache HBase data using Oracle Big Data SQL or Oracle SQL Connectors for Hadoop

In addition, it has a set of services for working with large volumes of spatial raster data:

- Support for dozens of image file formats supported by GDAL and image files stored in HDFS
- A sample Java console to view raster images and manage raster data processing workflows
- Raster operations including subset (finding a set of images from a catalog covering a user-specified region), georeferencing (associating geographic coordinates to image data), mosaic (virtually combining input images to deal with gaps and overlaps), and format conversion
- A Raster Processing API (Java) for Apache Spark for performing parallel operations using dataframes

## Multimedia Analytics

The multimedia analytics feature of Oracle Big Data Spatial and Graph provides a framework for processing video and image data in Apache Hadoop. The framework enables distributed processing of video and image data. Features of the framework include:

- APIs to process and analyze video and image data in Apache Hadoop and Spark
- Scalable, high speed processing, leveraging the parallelism of Apache Hadoop
- Distributed with OpenCV libraries
- Ability to install and implement custom video/image processing (for example, license plate recognition) to use the framework to run in Apache Hadoop

## Support for Oracle Big Data Appliance and Other Hadoop Platforms

Oracle Big Data Spatial and Graph can be deployed on Oracle Big Data Appliance, an open, multi-purpose engineered system for Hadoop and NoSQL processing, as well as other supported Hadoop and NoSQL systems.  For details on supported platforms, please visit http://www.oracle.com/technetwork/database/database-technologies/bigdata-spatialandgraph/.

## Summary

Oracle has taken nearly two decades of experience working with spatial and graph technologies to deliver a new Big Data platform designed to handle the most demanding workloads. With the introduction of Oracle Big Data Spatial and Graph, developers and data scientists can manage their most challenging graph, spatial, and raster data processing in a single enterprise-class Big Data platform.

Oracle Big Data Spatial and Graph represents an opportunity for the Big Data platform.  With its out of the box spatial enrichment services and analysis functions, as well as dozens of the most popular graph analysis functions, analysts and developers can now more easily apply commercial-grade algorithms to their big data workloads to achieve valuable new insights.

**Hardware and Software, Engineered to Work Together**

Oracle is committed to developing practices and products that help protect the environment