FEATURE OVERVIEW

Oracle Big Data Spatial and Graph: Spatial Features

ORACLE BIG DATA

For over a decade, Oracle has offered leading spatial and graph analytic technology for Oracle Database. Oracle is applying this spatial and graph expertise to Big Data workloads on Hadoop and NoSQL. Location can be used as a universal key across the disparate data commonly used in Hadoop-based analytic solutions. Oracle Big Data Spatial and Graph includes a range of spatial capabilities: a geo-enrichment service to enable data harmonization based on location, location analysis functions for categorizing and filtering data, and the ability to perform raster data cleansing and image processing. This package of commercial-grade components allows developers and data scientists to obtain deeper insights into Big Data workloads – while reducing complexity and simplifying development.

"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 added value 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

Oracle Big Data Spatial and Graph: Spatial Features

Oracle Big Data Spatial and Graph provides spatial and graph processing in a single enterprise-class Big Data platform. It provides a wide range of spatial vector and raster analysis functions and services, and visualization tools, to deliver insights and uncover patterns in business data in Hadoop systems.

This document introduces the spatial features of Oracle Big Data Spatial and Graph. For an overview of the complete product, including the graph features, please see the *Oracle Big Data Spatial and Graph Data Sheet*.

Data Enrichment and Categorization Services

Big Data workloads often include unstructured and semi-structured data from a wide variety of sources. Location can be useful to correlate, associate, and categorize this disparate data. Oracle Big Data Spatial and Graph provides services that take place names, addresses, zip codes, longitude and latitude, and other location identifiers, and enriches this data with known geographic context.

You can use these services to associate existing data sets with known location identifiers, or with named geometric hierarchies. For example, incoming Twitter log feeds can be analyzed and displayed using thematic maps to show how many tweets originate from each city, county, and

ORACLE

KEY BUSINESS BENEFITS

- Manage your most challenging spatial and raster data processing in a single enterprise-class Big Data platform
- Gain deeper insights into Big Data workloads through commercial-grade spatial algorithms and map visualization
- Enrich and categorize social data using location to harmonize disparate data sets
- Discover relationships and visual patterns based on location
- Store and process large volumes of satellite imagery and spatial sensor data using the low-cost, parallel Hadoop platform
- Reduce the complexities and simplify implementation of spatial processing in the Hadoop environment
- Optimized for Oracle Big Data Appliance

NEW FEATURES RELEASE 1.2

Vector Data Processing

- Hive support – SQL queries for spatial analysis and processing of data stored in HDFS
- Spatial joins – join two spatial data sets to find all interacting pairs of geometries

RELEASE 1.1.2

Vector Data Processing

- Spatial clustering and binning

Raster Processing

- Image loader – support for multi-band images

state. Text search can find the word "BOSTON" in a Twitter feed and associate it with the hierarchy of "BOSTON -> SUFFOLK COUNTY -> MASSACHUSETTS -> UNITED STATES". Or if the Twitter feed contains geographic latitude/longitude data, the service can associate the point with the relevant city, county, state, country, etc. where the point lies.

Oracle Big Data Spatial and Graph includes a library of geographic hierarchical boundary data covering worldwide countries, states, counties, and cities, as well as named hierarchical data sets for text matching. You can select a data set and template to use with the geographic hierarchy of your choice. You can also create and use custom data sets in the hierarchy, such as customer sales regions, in combination with packaged boundary regions.

MapReduce jobs provide results of these services in GeoJSON format on the Hadoop File System (HDFS), which are available for further processing. You can also build a map application visualizing these results with the provided HTML5 map visualization API.

Spatial Data Processing Features

The spatial features in Oracle Big Data Spatial and Graph allow the scalable parallel processing characteristics of the Big Data platform to be applied to a number of traditional geospatial workloads. Working with spatial data may involve format conversion, data cleansing, and preparation and processing of raw data into a final-use data product.

For **vector data (2D and 3D digital map data)**, commonly used spatial operations such as POINT-IN-POLYGON, BUFFER, DISTANCE, and ANYINTERACT are provided as MapReduce jobs to filter and analyze any spatial data stored in HDFS. Developers may also use SQL to perform spatial filtering and analysis, through support for the Hive framework.

Oracle Big Data Spatial and Graph also offers **raster-processing operations** to work with large volumes of geospatial imagery and gridded data sets. It includes operations such as MOSAIC (to align and stitch together different imagery) and SUBSET (to produce a single object containing all cells of a given subset of the image based on a window, layer or band numbers, and pyramid level). It also has a MapReduce framework for raster analysis operations, such as calculating the slope at each pixel based on a digital elevation model (DEM).

Working with Spatial Vector Data

The vector features support the steps in a typical workflow, and include

- Loading data into HDFS for storage, or identifying existing data sets to be analyzed
- Creating indexes (if desired)
- Performing spatial analysis and processing, either through Map Reduce or Hive SQL
- Visualizing spatial data and analysis results on a map

The workflow steps for using both MapReduce and the Hive SQL

KEY FEATURES

Vector Data Processing

- Support spatial processing and analysis for data stored in HDFS, through MapReduce or SQL
- Support for Hive framework – use SQL to analyze and process data on HDFS (**NEW FOR RELEASE 1.2**)
- Data type support for text, 2D, and 3D geospatial formats
- Geodetic and Cartesian data model support
- Enrichment service to associate documents or data with location
- Built-in gazetteer of geographic names (cities, states, countries, etc.) and text matching services
- Service to associate latitude/longitude with worldwide administrative hierarchies
- Spatial analysis operations including ANYINTERACT, CONTAINS, WITHIN DISTANCE, DISTANCE AND LENGTH CALCULATIONS, BUFFER, POINT-IN-POLYGON
- Spatial binning and clustering for fast analysis and discovery (**NEW FOR RELEASE 1.1.2**)
- Spatial joins – join two spatial data sets to find all interacting pairs of geometries (**NEW FOR RELEASE 1.2**)
- Spatial indexing for fast retrieval of data

framework are described below.

Loading Data into HDFS

You can use the loader of your choice to load data into HDFS – there are no format requirements for data. You may use any data format appropriate for your application – data does not have to be organized by a geospatial attribute. This ensures that your Big Data application can easily combine location information with business data. If you already have existing data in HDFS, you can use the spatial framework and algorithms on top of that data as well.

The GeoJSON and Esri Shapefile data formats are natively supported, and spatial queries will operate directly on data in those formats. For data in other formats, you need to provide an InputFormat class that reads your data records and produces a JGeometry instance at query runtime.

This Big Data approach allows organizations to incorporate spatial analysis directly into their existing Hadoop processes – instead of a spatial-centric approach that silos or separates spatial data.

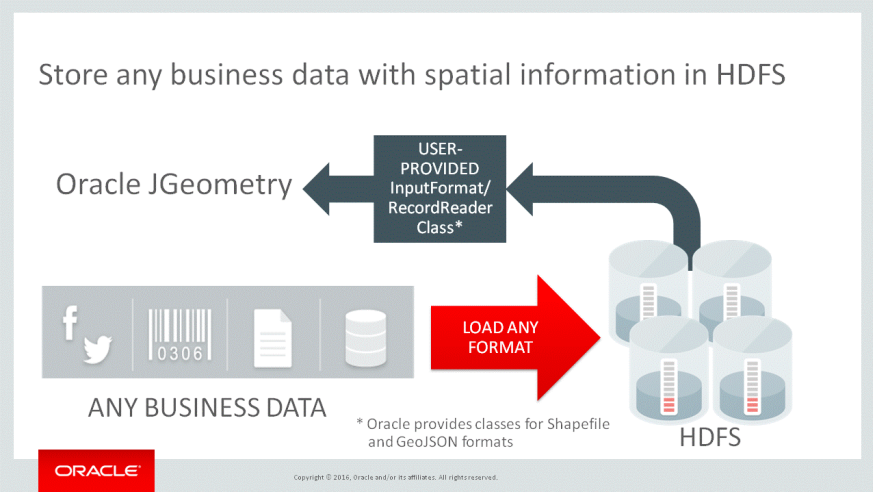

Figure 1. Spatially-Enabling Business Data in HDFS

Performing Spatial Analysis

Oracle Big Data Spatial and Graph includes a Java API and a set of spatial functions packaged as Java methods. The API supports commonly used spatial queries and calculations including:

- Operations on single geometries (such as BUFFER, SIMPLIFY, LENGTH, and AREA)
- Operations on pairs of geometries (such as POINT-IN-POLYGON, and ANYINTERACT)
- Spatial binning and clustering: quickly process large numbers of records into bins or clusters that can be visualized to identify areas of interest for further analysis (New for Release 1.1.2)
- Joins: detecting spatial interactions between records of two data sets (INSIDE, ANYINTERACT, WITHINDISTANCE) (New for Release 1.2)

"Big Data systems are increasingly being used to process large volumes of data from a wide variety of sources. With the introduction of Oracle Big Data Spatial and Graph, Hadoop users will be able to enrich data based on location and use this to harmonize data for further correlation, categorization and analysis. For traditional geospatial workloads, it will provide value-added spatial processing and allow us to support customers with large vector and raster data sets on Hadoop systems."

STEVE PIERCE
CEO
THINK HUDDLE

You can write a MapReduce job in your application that calls Java methods such as buffer or point-in-polygon, and that executes these operations very quickly. You can specify query results to be written either to HDFS or a different file system.

Spatial binning and clustering analysis can quickly process large numbers of records, such as millions of tweets, into bins or clusters that can be visualized into a thematic map. You can then very quickly see which areas have "hot" and "cold" levels of activity – and identify points of interest for further drilldown and analysis. The MapReduce framework allows for fast processing of large data sets to obtain insights.

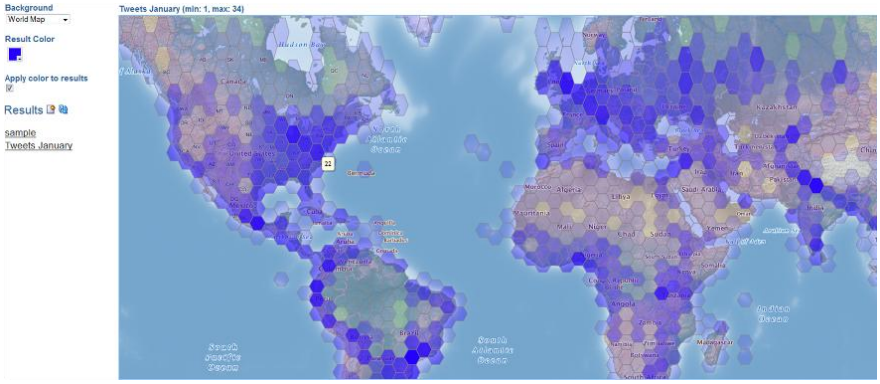

Figure 2. Spatial binning of worldwide Twitter data

Spatial joins are a powerful way to determine all spatial interactions between two different data sets. In a regular spatial query, you can ask "Find all the tweets that occurred in zip code boundary 94065". A spatial join allows you to ask "Find all the tweets that occurred in every zip code in the US zip codes data set". For joins, the data sets are often large, and the calculations time-intensive. Oracle Big Data Spatial and Graph provides a spatial partitioning mechanism that leverages Hadoop parallelism to perform the join. You can simply use a single Java function call to execute the join.

Creating Spatial Indexes

Spatial indexing provides fast query performance in a Hadoop environment. Local spatial indexes on each node maximize the parallel processing capabilities of MapReduce architectures. This minimizes index creation time, latency, and single-node bottlenecks, and can quickly process large query volumes for demanding applications. Performance of range queries, such as POINT-IN-POLYGON and ANYINTERACT, is significantly improved by avoiding unnecessary secondary filter operations.

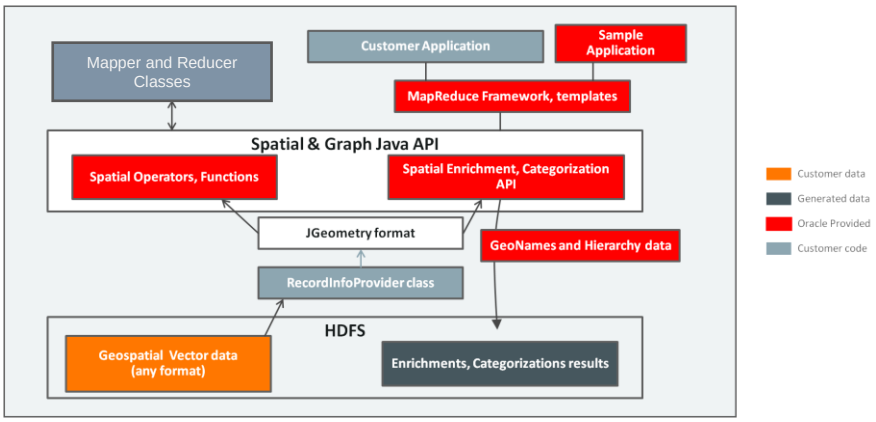

Figure 3. Vector Processing Framework in Oracle Big Data Spatial and Graph

Using the Hive Spatial Vector API to Perform Spatial Processing and Analysis with SQL

Instead of writing MapReduce jobs, some developers may prefer to use SQL to perform spatial processing and analysis. The new Hive spatial framework eliminates the need to write your own MapReduce jobs for commonly used spatial functions. The Hive syntax will be very familiar to those who are accustomed to writing SQL-based applications.

Hive is an open source framework that allows developers to issue SQL queries on a Hadoop cluster. Hive SQL provides MapReduce interfaces to HDFS, so users can write SQL queries. Oracle Big Data Spatial and Graph generates all the MapReduce jobs required to execute those queries across a Hadoop cluster.

Oracle Big Data Spatial and Graph provides Hive support for:

- 2D and 3D spatial data types (such as ST_POINT, ST_LINE, ST_POLYGON)
- Spatial functions (such as AREA, INTERSECT, CONTAINS) within Hive's User Defined Function (UDF) framework
- A de-serializer that reads file formats into Hive

To enable spatial features using Hive, you need to provide an InputFormat class that reads your data records from the file system, and converts it to JSON records, in this case. From there, the Oracle framework will convert JSON records into Hive geometries.

Then, you need to create an external table interface to the HDFS file system (a standard step for most Hive SQL implementations). In the <CREATE TABLE> statement, you specify the location of your data file on HDFS, along with input and output formats. External table columns can be defined based on the data stored in your record in HDFS, for example, twitter data with tweet ID, number of followers, and location.

Once the table definition is in place, you can then write SQL statements to perform spatial analysis, such as finding all the tweets that are contained

in a specific zip code boundary.

Spatial index creation within Hive is also supported. Using indexes is an optional step for spatial processing, and can improve performance significantly.

Sample scripts for the Spatial Vector Hive API, and a complete list of supported types and functions, are included in the product documentation.

KEY FEATURES

Spatial Server Console

- J2EE sample application deployable in Jetty and other application servers
- Explore, categorize, view data in a variety of formats, coordinate systems
- Manage vector and raster processing workflows
- Use a map visualization API (HTML5-based) to build map applications

Spatial Server Console and Map Visualization API

A convenient Java user interface is provided to manage spatial data processing workflows. This is a sample J2EE application that can be deployed in a Jetty, Tomcat, WebLogic, or other supported Java application server. From the console, you can create spatial indexes on data already loaded into HDFS. You can also run Hadoop jobs to do spatial processing. The console creates and runs the MapReduce job, such as categorizing tweets by city, state, and country.

To view spatial data and analyze results on a map (such as a United States map indicating number of tweets by state), you can use the HTML5-based map visualization API.

The API allows you to apply styles (such as colors and patterns) to themes or data layers (such as countries, states, and tweet origin locations), and to render a map as an image for display on a webpage. Maps may have several themes representing political entities (such as city and state boundaries) or physical entities (such as highways and rivers). When the map is rendered, each theme represents a layer in the complete image.

The HTML5 map visualization API takes advantage of the capabilities of modern browsers. Features include:

- Built-in support to retrieve background maps from various third-party map services
- Rich client-side rendering of geospatial data with on-the-fly application of rendering styles and effects, such as gradients, animation, and drop-shadows
- Auto clustering of large numbers of points and client-side heat map generation
- Client-side feature filtering based on attribute values and spatial predicates (query windows)
- A rich set of built-in map controls and tools, including a customizable navigation bar and information windows, configurable layer control, and tools for redlining (user-defined features of interest) and distance measurement

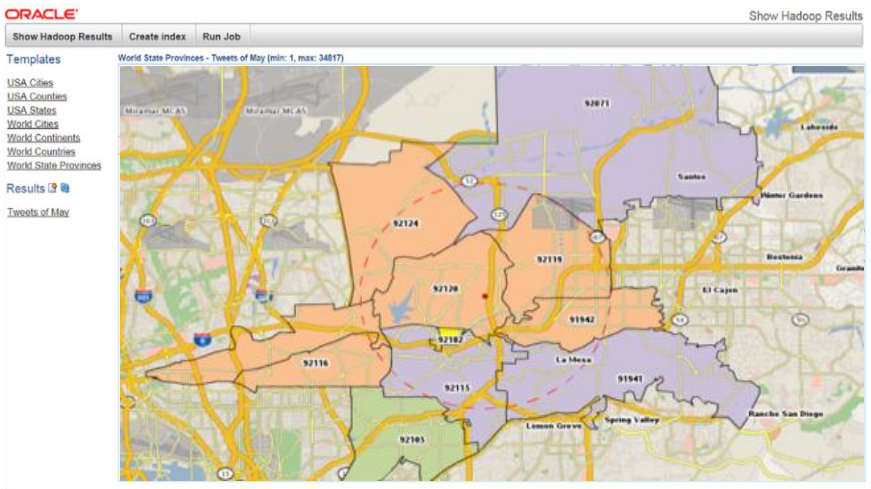

Figure 4. Map of categorized tweets in US, using Spatial Server Console sample map visualization application

KEY FEATURES

Raster Imagery Data Processing

- GDAL-based loading of raster data onto HDFS from other file systems
- Support for many file formats including georeferenced images, 3 band, single band, and multi-band images
- Raster processing operations such as MOSAIC and SUBSET
- MapReduce framework for large scale raster analysis operations

Raster Data Processing

Oracle Big Data Spatial and Graph supports data preparation services for raster imagery. For example, source imagery may be georeferenced, and stored in different coordinate systems or resolutions. Hadoop environments are ideally suited to efficiently carry out basic raster processing jobs for cleansing and preparing data within a workflow, on a very large scale. Oracle Big Data Spatial and Graph provides HDFS storage for image or raster files, with support for many GDAL-supported formats.

Raster support includes:

- Loading and transforming raster data formats from traditional file systems into HDFS for storage
- Raster analysis: mosaicking and subsetting
- Image processing framework for further analysis, such as pyramiding, and terrain and contour generation
- Image server console with sample J2EE application for managing raster processing workflows

Loading Imagery Data into HDFS

In most Big Data scenarios, large volumes of raster data are generated by a variety of sensors. This raw data is usually streamed into file systems for storage and follow-on processing. Oracle Big Data Spatial and Graph provides a GDAL-based loader to import data into HDFS in a manner optimized for MapReduce processing jobs. Many data formats are supported: 3 band images, single band images with float and byte data types, and multi-band images.

When raster data is loaded into HDFS, for optimal processing it should be organized so that a MapReduce job can process it with a minimum amount of data transfer between nodes. The GDAL loader can be configured to

RELATED PRODUCTS

The following are related products available from Oracle

- Oracle Big Data Appliance
- Oracle NoSQL Database
- Oracle Big Data SQL
- Oracle Big Data Connectors
- Oracle Exadata
- Oracle Spatial and Graph

support alternative HDFS storage models to ensure pixel data is properly partitioned over different HDFS blocks. For example, optimized HDFS storage for processing a shaded relief map from digital elevation model (DEM) data is likely to be different from the storage model for raster analysis. The storage virtual layers also ensure that all imagery data is properly georeferenced.

Raster Analysis: Mosaic and Subset Operations

Mosaic and subset operations are based on the concept of a **virtual mosaic**, where you can logically combine a certain number of images into a catalog. This allows you to store imagery in different coordinate systems and resolutions – all of which can be mosaicked on the fly. A subset operation allows you to find a set of images from a given catalog covering a user-specified region and generate a new image file (in the specified file format) from the original source files. A follow-on mosaic process cleans up any gaps and overlaps in the imagery.

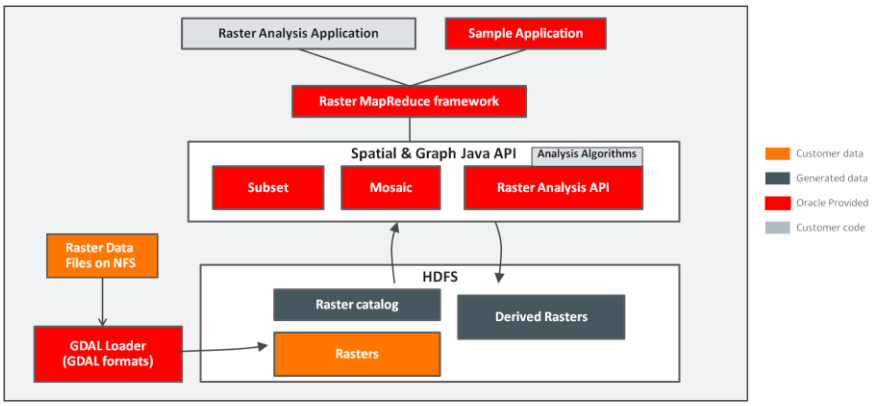

Figure 4. Raster Processing Framework in Oracle Big Data Spatial and Graph

Image Processing Framework

You can also use the provided MapReduce framework to write and carry out further image processing or raster analysis operations. For example, you can write a map algebra routine to calculate the slope at each pixel, based on a digital elevation model (DEM).

Spatial Server Console Support for Raster Data

The sample Spatial Server Console allows you to manage raster data processing workflows. The console's interface supports loading data from a network file system into HDFS, creating catalogs from existing images on HDFS, running Hadoop subset jobs, and running Hadoop raster analysis jobs.

RESOURCES

For more information on Oracle Big Data Spatial and Graph, visit

Oracle Technology Network
Software downloads, documentation, tutorials, white papers:
www.oracle.com/technetwork/database/database-technologies/bigdata-spatialandgraph

Oracle.com
Product overviews, videos, press:
www.oracle.com/database/big-data-spatial-and-graph

Blog
Technical tips, code samples:
blogs.oracle.com/bigdataspatialgraph

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/overview/index.html .

ORACLE

CONTACT US

For more information, visit oracle.com or call +1.800.ORACLE1 to speak to an Oracle representative.

CONNECT WITH US

blogs.oracle.com/oracle
facebook.com/oracle
twitter.com/oracle
oracle.com

Hardware and Software, Engineered to Work Together