Oracle Big Data Spatial and Graph: Spatial Features

Roberto Infante
11/11/2015 – Latin America Geospatial Forum

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.
  - Index creation for fast data processing
  - Data categorization
  - Data enrichment services using GeoNames

Overview of Spatial Features

- Vector processing continued
  - Binning and clustering analysis
  - Map Visualization API (developed with HTML5 API)

- Raster Data Processing
  - GDAL to load raster data onto HDFS
  - Raster processing operations: Mosaic and sub-set operations

ORACLE

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 RecordInfoProvider class
  - This class translates the customer data record and produces an instance of JGeometry
  - With this model we can support any data format
  - The API provide RecordInfoProvider and InputFormats for the common formats GeoJSON and ESRI Shapefiles.

Vector Data Processing API

- Our API is broadly divided into three categories
  - Functions that operates directly on geometries
    - Buffer, simplify, length, area, pointInPolygon, AnyInteract etc.
  - Functions that categorize and enrich data
    - Associating a data set with a known geometry or named hierarchy
      - For example, process all tweets for a time period and count how many tweets are associated with each city, county, state, etc.
  - Functions that analyze the data
    - Binning and clustering
      - For example, create bins (regions) with the average number of followers of tweets data.

Spatial index

- Spatial Index
  - Metadatas are saved for each index
  - Customers control the data loading process
  - Spatial data are indexed and the records in the index can contain any non spatial data
  - The index can refer to original record
- Example
  - A large number of records/logs with lat/long information along with other attributes
  - Find all the records that are within 10 miles from the given lat/long and also contain the phrase "united nations"
- When the map-reduce job is created, we can specify whether to use or not use the spatial index
  - If spatial index is not used, each record is processed
  - If spatial index is used, the map-reduce job only processes those records that can potentially interact with the MBR of the query geometry

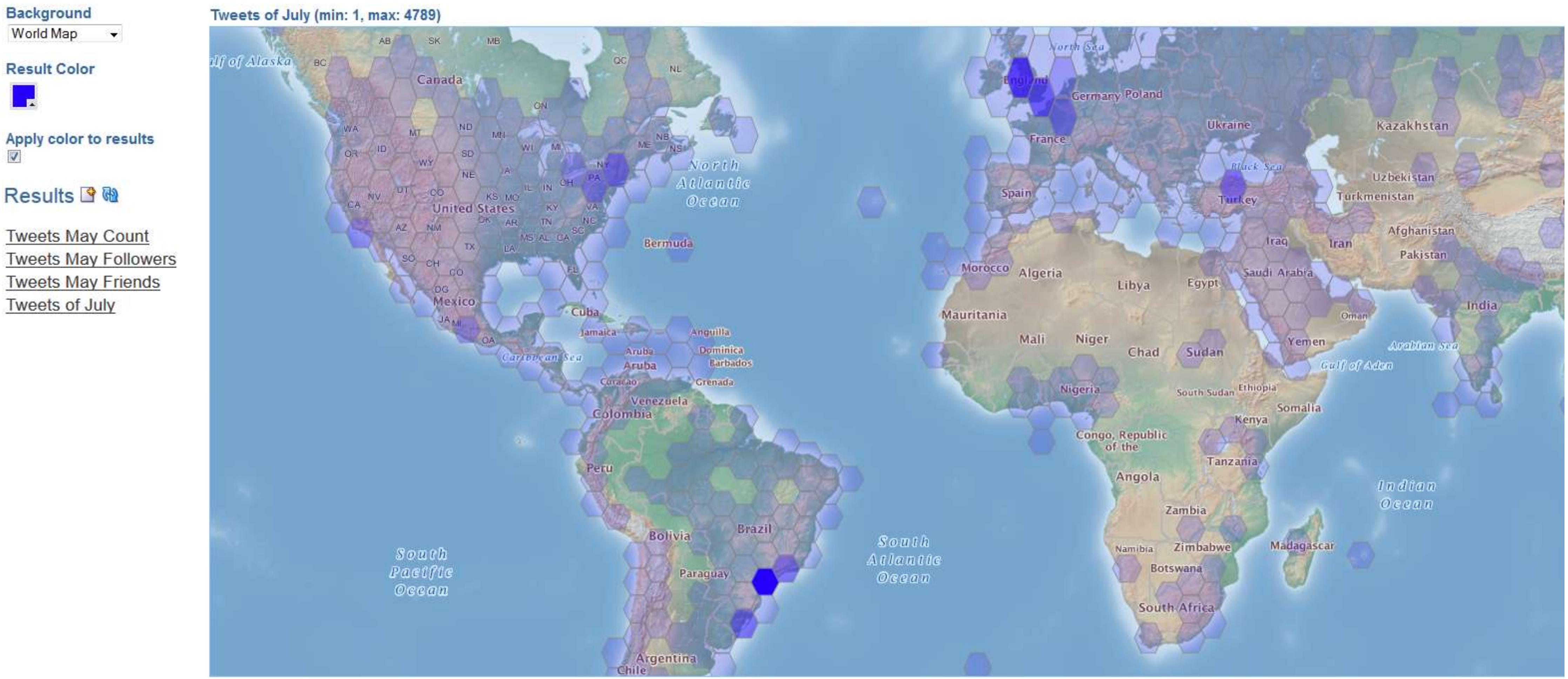
Data Categorization Services

- Any hierarchical geometry data set can be used as a reference data
  - We provide some well known geometry layers but the customers can use any sets of data layers
  - This data is stored as GeoJSON data
- Customers choose a set of layers they want to use
  - For example, they can select (continents, countries, cities) or (countries, states, counties) as the hierarchy
- We create a map-reduce job and process the customer data and produce a result file that can be used for further processing
  - For example, a summary file that can be used to visualize the categorized data on a map

Data Enrichment Services

- GeoNames data sets are used as reference data to enrich customer data with well known geographic named hierarchies
  - For example, customer data might haves references to “cambridge, ma” or “cambridge, ma, usa”
  - Both these data records are then fully qualified to say that it has a reference to “Cambridge, Massachusetts, United States”
- Customers provide a RecordInfoProvider class that reads the customer data record and extracts place names into a text string
  - Our service then takes those text strings and matches them against the reference data set to find the best match
- This enrichment is done as a map-reduce job

Vector Spatial Server Console (binning)



Background

World Map

Result Color

Apply color to results

☑

Results

Tweets May Count

Tweets May Followers

Tweets May Friends

Tweets of July

Tweets of July (min: 1, max: 4789)

Raster data Loader

- Customers usually have large volumes of raster data in traditional file systems
  - We provide a GDAL based loader to load the data into HDFS such that the resulting HDFS blocks are organized for map-reduce jobs
  - Any format supported by GDAL can be supported

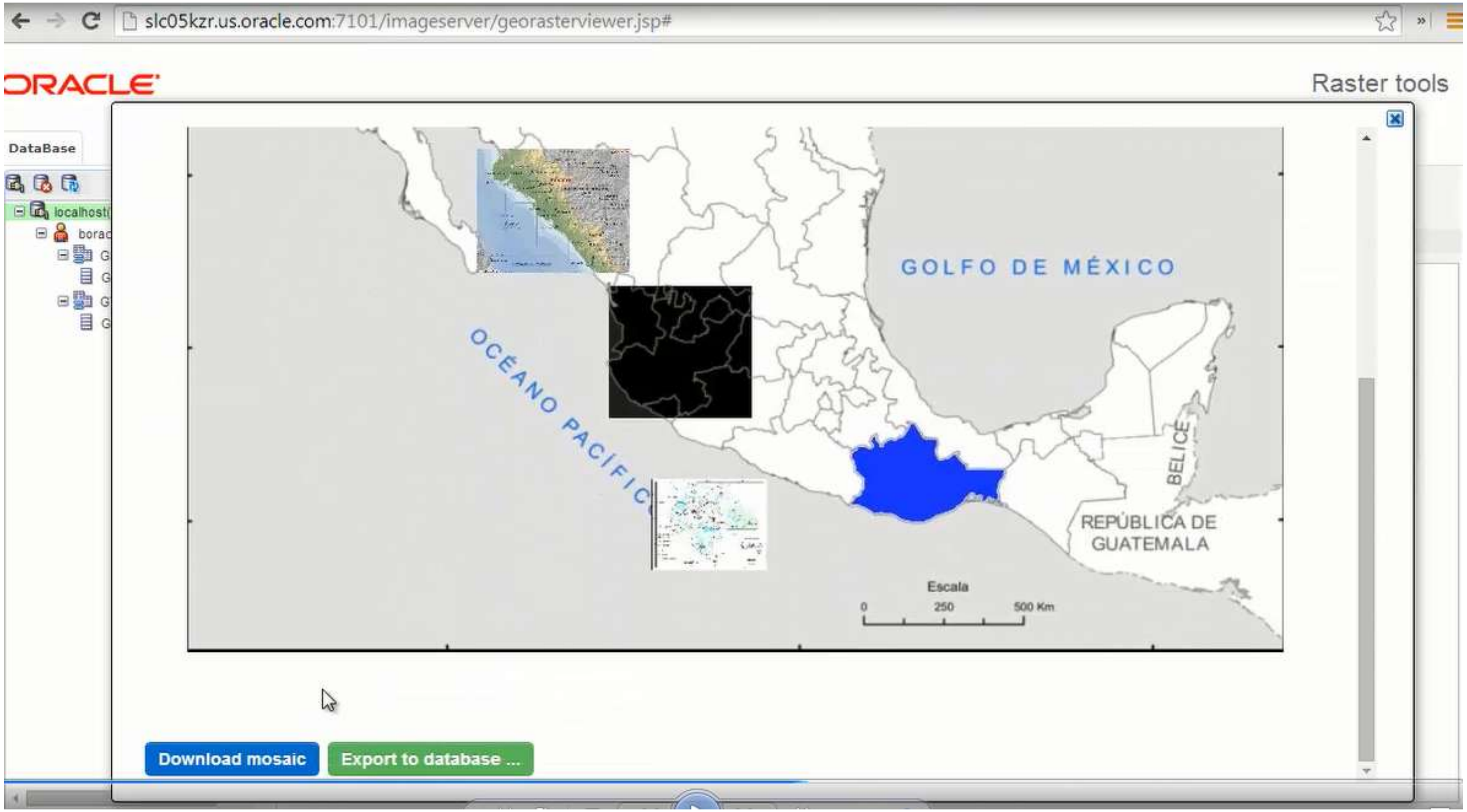
Raster Data Stored on HDFS

- When data is moved from traditional file system to HDFS, data should be organized in such a way that a map-reduce job can process it with minimum amount of data transfer between data nodes
- Data are process by blocks of pixels that contains also some boundaries pixels to make faster some operations that need the neighbors pixels (for example Slope operation that processes the pixels with their 8 neighbors).

Image Server Features

- Subset operation
  - Find the set of images from a given catalogue covering a user specified region and generate a new image from the source files
  - The new images will have user specified resolution and coordinate system
    - These can be different for different images in the source catalogue and the resulting image can have a different value
  - Mosaic the input images to deal with gaps and overlaps
  - Create a new file with the specified file format
  - 26 operations available. For example the user will be able to add/remove values from the pixels to get a new image.

Image Server Console



What’s next

- Vector Data Processing
  - Add support for Hive
  - Add support for Spark
  - Add support for Oracle Big Data SQL
- Raster Data Processing
  - Add more functionality to rasters processing (for example functions to average the pixels between several rasters).

Contact: roberto.infante@oracle.com