Big Data Technologies and Geospatial Data Processing:
A perfect fit

Albert Godfrind
Spatial and Graph Solutions Architect
Oracle Corporation
Agenda

1. The Data Explosion
2. Big Data?
3. Big Data and Geo Data Processing
4. Graphs and More...
• Complete End-to-End Offering
• Software, Hardware, Services
• Cloud Services
• Engineered Systems
• Custom deployments
Spatial Data Processing Needs are Exploding

**Track and Trace**
- Vehicle tracking, guidance, traffic sensors,

**Rasters**
- Satellite imagery, climate data, statistics, extraction, calculations
Spatial Data Processing Needs are Exploding ... 

<table>
<thead>
<tr>
<th>3D</th>
<th>Point Clouds</th>
</tr>
</thead>
<tbody>
<tr>
<td>Urban landscapes, virtual visits, infrastructure planning</td>
<td>Billions of points captured by sensors. Change detection, object recognition</td>
</tr>
</tbody>
</table>

![3D urban landscape](image1.png)

![Point Clouds](image2.png)
Big data refers to the analysis of large volumes of unstructured data using new technologies, methods, and systems.

Key terms:
- Petabytes
- Data storage
- Search
- Analytics
- Internet
- Government
- National
- Analysis
- Results
- Future
- Learning
- Social
- Large
- Implementation
- Scientific
- Software
- Research
- Complex
- Visualization
- Intelligence
- Processing
- Based
- Approaches
- Amount
- Records
- Relevant
- Management
- Processes
- Second
- Business
drive
- Hybrid
- Systems
- Data
- Set
- Search
- Every
- Across
- People
- Innovation
- Development
- Progress
- Data
- Science
- Companies
- Analysis
- Result
- Estimation
- Size
- Billion
- Economic
- High
- Various
- Integration
- Various
- Different
- Models
- Frameworks
- Successful
- Users
- Worldwide
- World
- Various
- Challenges
- Using
Distributed Data Processing

✓ Distribute the **processing**
  – Many servers
  – Scheduling, coordination, monitoring, recover from failures

✓ Distribute the **data**
  – Many servers
  – High availability, no data loss, recover from failures

“Big Data” is all about making this easier
Big Data Appliance

Sun Oracle X6-2L nodes with (per node):

• 2 * 22 Core (2.2GHz) Intel Xeon E5-2699 v4 Processors
• 256 GB DDR4-2400 Memory
• 96TB Disk space

Included Software:

• Oracle Linux 6.7
• Cloudera Distribution of Apache Hadoop 5.7 – EDH Edition
• Cloudera Manager 5.7
• Oracle R Distribution
• Oracle NoSQL Database Community Edition

• Starter Rack = 6 nodes, Full Rack = 18 nodes
Big Data Cloud Service

http://docs.oracle.com/cloud/latest/bigdata-cloud/

• Oracle Linux operating system

• Cloudera Distribution:
  – Apache Hadoop, HDFS, MapReduce engine (YARN)
  – Cloudera Manager
  – Apache projects: Hive, Pig, Oozie, ZooKeeper, HBase, Sqoop, and Spark
  – Cloudera applications: Impala, Search, Navigator.

• Oracle Big Data Connectors
  – Oracle SQL Connector for Hadoop Distributed File System
  – Oracle Loader for Hadoop
  – Oracle XQuery for Hadoop
  – Oracle Data Integrator Enterprise Edition

• Oracle R Advanced Analytics for Hadoop

• Oracle Big Data Spatial and Graph
The Forrester WaveTM: Big Data Hadoop-Optimized Systems, Q2 ’16
Oracle Big Data Spatial and Graph

- Data Harmonization using any location attribute (address, postal code, lat/long, placename, etc).
- Preparation, validation and cleansing of Spatial and Raster data.
- Visualizing and displaying results on a map.
- Categorization and filtering based on location and proximity.
Store any data with spatial information in HDFS

Oracle JGeometry

USER-PROVIDED InputFormat/RecordReader Class

LOAD ANY FORMAT

ANY BUSINESS DATA

HDFS
Supports All Vector Data

• Points, Lines, Polygons, Collections
  – Including Arcs, compound line strings, NURBs, compound polygons, etc.

• 2D and 3D structures
• Projected and Geodetic

• Topological and distance operations
  – Any interact, inside, distance, length, simplify, buffer, PointInPolygon
GeoJSON

```json
{
    "type": "FeatureCollection",
    "collectionName": "sample tweets",
    "srs": 8307,
    "geodetic": true,
    "features": [
        {
            "type": "Feature",
            "id": "6703",
            "geometry": {"type": "Point", "coordinates": [114.18306, 22.30693]}
        },
        ...
    ]
}
```
How to Do Vector Processing ...

• Option 1: Use the **spatial console**
  – Use it to run **categorization**, **clustering** and **binning** jobs, also creating indexes and viewing the data on a map.

• Option 2: Use the **command line**
  – Use the “hadoop jar” command to submit predefined jobs for categorization, clustering and binning, or creating indexes.

• Option 3: Use **SQL**
  – Use hive to run SQL queries over hadoop

• Option 4: Write **custom map-reduce code**
  – Use spatial’s java APIs in custom Map/Reduce code
Welcome to the Spatial Hadoop Vector Console

Spatial Hadoop Vector Console is a web console with the following sections:

1. **Spatial Index**: Create/Delete spatial indexes on HDFS data.

2. **Explore data**: Explore indexed data.

3. **Categorization**: Create and show categorization results. For example, it can be used to show all the twitters from specified HDFS files in the hierarchy World Continents/World Countries/World State Provinces/World Cities.

4. **Clustering**: Create and show clustering results.

5. **Binning**: Create and show binning results.

6. **Vector Jobs**: View jobs information, configuration and logs.

The console uses the Hadoop Vector Analysis API to perform any hadoop operation. The jobs can be run inside or outside the console.
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 ...

• Uses the Geonames data set

* Golden Gate Bridge (in Korean)
Spatial Categorization

Any hierarchical geometry data set for reference

Customers choose a template. For example (continents, countries, cities) or (countries, states, counties)

Big Data Spatial map-reduce job processes the customer data and produces a result file
World Countries - January Tweets Categorized (min: 1, max: 736)
World State Provinces - January Tweets Categorized (min: 1, max: 72)

Templates
- USA Counties
- USA States
- World Cities
- World Continents
- World Countries
- World State Provinces

Results
- January Tweets Categorized sample
Spatial Clustering

Clustering Results

<table>
<thead>
<tr>
<th>Create Index</th>
<th>Explore Data</th>
<th>Categorization</th>
<th>Clustering</th>
<th>Binning</th>
</tr>
</thead>
</table>

Background:
- World Map

Show clusters boundaries

Results

- sample
Spatial Binning
Run a Spatial Processing Job

```
hadoop jar $API_LIB_DIR/sdohadoop-vector.jar oracle.spatial.hadoop.vector.mapred.job.Categorization \
    -libjars $HADOOP_LIB_JARS \n    spatialOperation=IsInside \n    input=/user/oracle/HOL/tweets.json \n    output=/user/oracle/HOL/catOutputEuro \n    inputFormat=oracle.spatial.hadoop.vector.geojson.mapred.GeoJsonInputFormat \n    recordInfoProvider=oracle.spatial.hadoop.vector.geojson.GeoJsonRecordInfoProvider \n    srid=8307 geodetic=true tolerance=0.5 \n    hierarchyInfo=hol.EuroHierarchyInfo \n    hierarchyIndex=/user/oracle/HOL/hierarchyIndex \n    hierarchyDataPaths=file:///opt/oracle/oracle-spatial-graph/spatial/vector/HOL/data/eurozone_countries.json,file:///opt/oracle/oracle-spatial-graph/spatial/vector/HOL/data/eurozone_provinces.json
```
Use SQL For Spatial Processing

```
SELECT id, followers_count, friends_count, location
FROM hive_tweets
WHERE ST_Contains(
    ST_Polygon(
        '{"type": "Polygon",
        "coordinates":
           [[[106, 25], [106, 30], [104, 30], [104, 25], [106, 25]]],
        4326
    ),
    ST_Point(geometry, 4326),
    0.5
)
and followers_count > 50;
```

- Implemented as Hadoop of Spark jobs
## Vector Data Processing Functions

### Single Geometry
- Length
- Area
- Buffer
- Simplify

### Geometry Pairs
- Range Queries
  - Point in Polygon
  - Touch, Overlap, Intersect, Contains, Any Interaction
- Join Queries
  - Interactions on sets of data
  - E.g.: Find all the dropped cell calls in all coverage areas

### Categorization and Enrichment
- Associate a data set with a known geometry or named hierarchy
  - Process all Tweets for a period of time and count how many are associated with each city, county, state, etc.
Big Data Raster Capabilities

• **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

• **Raster Processing**
  – **Loader** to load raster data from NFS to HDFS
  – **Mosaic** and **subset** operations based on a virtual mosaic
  – **Image processing** framework for raster analysis

• **Console** for viewing, loading and processing rasters
Loading Raster Data

• 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

• Many formats supported by GDAL
Raster Loading Map/Reduce Job

LoaderInputFormat → Mappers → Reducer

Calculates splits based on image Metadata, each split is of a block size.

Each split processed by Mapper → Reads piece of the image → Blocks saved into HDFS
Raster Processing Map/Reduce Job

Based on image metadata, determines the tiles and makes a split for each tile.

Each split is processed by a mapper.

Tiles saved into HDFS.
Subset / Process / Mosaic Operation

• Find the set of images from a given catalogue covering a user specified region
• The new images have user-specified resolution and coordinate system
• Apply pixel-level processing (“raster algebra”)
• Mosaic the input images to deal with gaps and overlaps
• Create a new image with the chosen file format
Raster Algebra Processing

Local Map algebra operations

- localnot
- localmultiply
- localround
- localceil
- localcos
- localtanh
- localdefined
- localif
- localdivide
- localround
- localnegate
- localtan
- localundefined
- localadd
- localdivide
- locallog
- localtanh
- localundefined
- localpow
- locallog10
- localabs
- localsin
- localsinh
- localsubstract
- localsqrt
- localfloor
- localsin
- localcosh
- localatan
Example: Shaded Relief calculation

• **Input**: NxM pixels where each pixel is a floating point number denoting elevation
• Find the shaded relief from the DEM

• **Algorithm**
  – Look at the values of 8 neighbors and the current pixel value and generate a new pixel
  – Needs the neighboring pixel values to calculate the new pixel value corresponding to the current pixel
Image Server Console

- **Load data** into HDFS from NFS
- **Create catalogues** from existing images on HDFS
- **Run** Hadoop jobs to do **mosaic** operations
  - Input rasters can be in any resolution or coordinate system
- **Run** Hadoop jobs to do **subset** operations
  - This will create and run the map-reduce job to the specified subset operation including changing resolution, changing coordinate system, etc.
- **Run** Hadoop jobs to do **raster** analysis
  - This will create and run the map-reduce job to the specified raster analysis operations
  - Users will need to specify the java class that is used to process the pixels and produce new pixel values for the output image
And now, something completely different!

Big Data Spatial and Graph
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 most critical link in my power grid?

**Tabular questions:**
Well-suited to SQL-like tools

**Graph questions:**
We need something different!
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
Resources ...

http://www.oracle.com/big-data
http://www.oracle.com/technetwork/topics/bigdata

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


https://blogs.oracle.com/bigdataspacialgraph
• “Big Data Appliance” in a box ... and more
  – Cloudera Hadoop, NoSQL, Big Data Spatial and Graph, Big Data Discovery
  – Big Data Connectors, Oracle NoSQL

• But also ...
  – Oracle Database 12c, Oracle Data Integrator, GoldenGate, SQL Developer, Oracle R

ALBERT GODFRIND
Solutions Architect
Spatial and Graph Services

Oracle Corporation
Greenside
400 av. Roumanille - BP 309
06906 Sophia-Antipolis
France

phone +33 4 93.00.80.67
mobile +33 6 09.97.27.23
albert.godfrind@oracle.com
http://www.oracle.com/
Integrated Cloud
Applications & Platform Services