OMAN GEOSPATIAL FORUM™

10-11 APRIL, 2017
Oman Convention Centre
Muscat, Oman

ORACLE
BIG DATA

Big Data Technologies and Geospatial Data Processing:

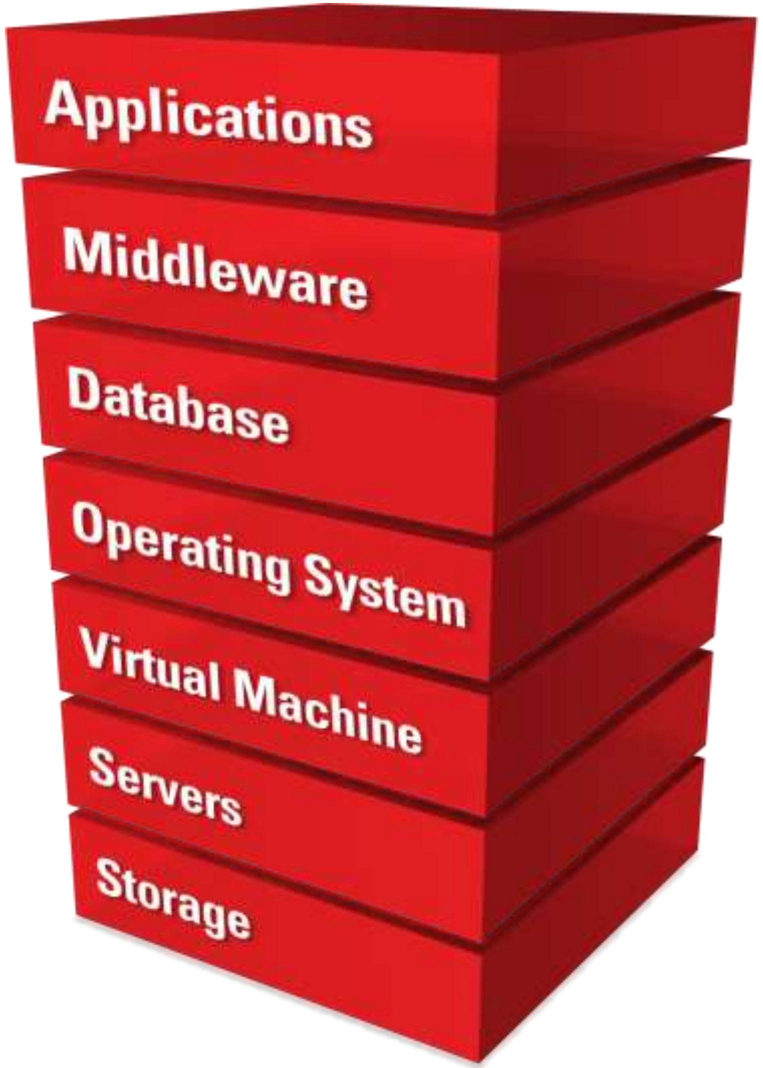
A perfect fit

Albert Godfrind
Spatial and Graph Solutions Architect
Oracle Corporation

ORACLE

Agenda

1. The Data Explosion
2. Big Data ?
3. Big Data and Geo Data Processing
4. Graphs and More…

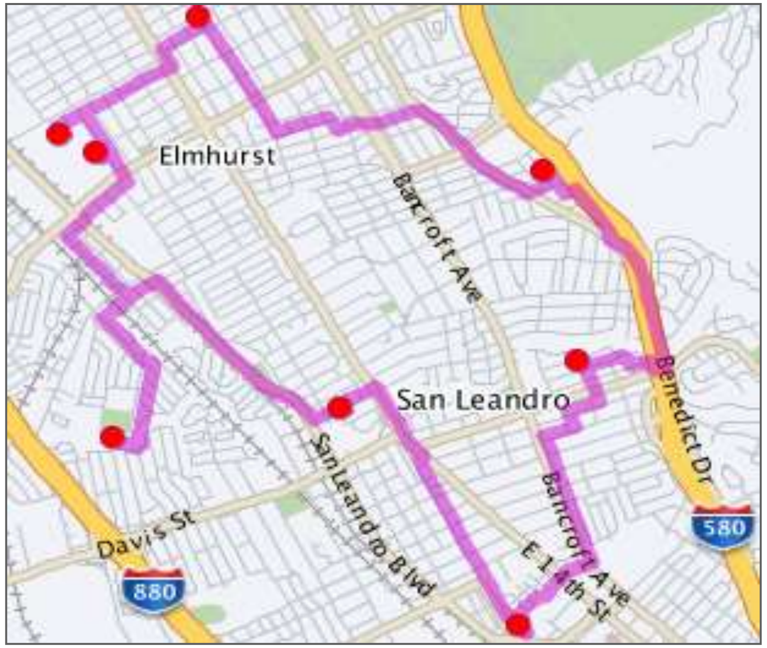


- Complete End-to-End Offering
- Software, Hardware, Services
- Cloud Services
- Engineered Systems
- Custom deployments

Spatial Data Processing Needs are Exploding

Track and Trace

- Vehicle tracking, guidance, traffic sensors,



Rasters

- Satellite imagery, climate data, statistics, extraction, calculations

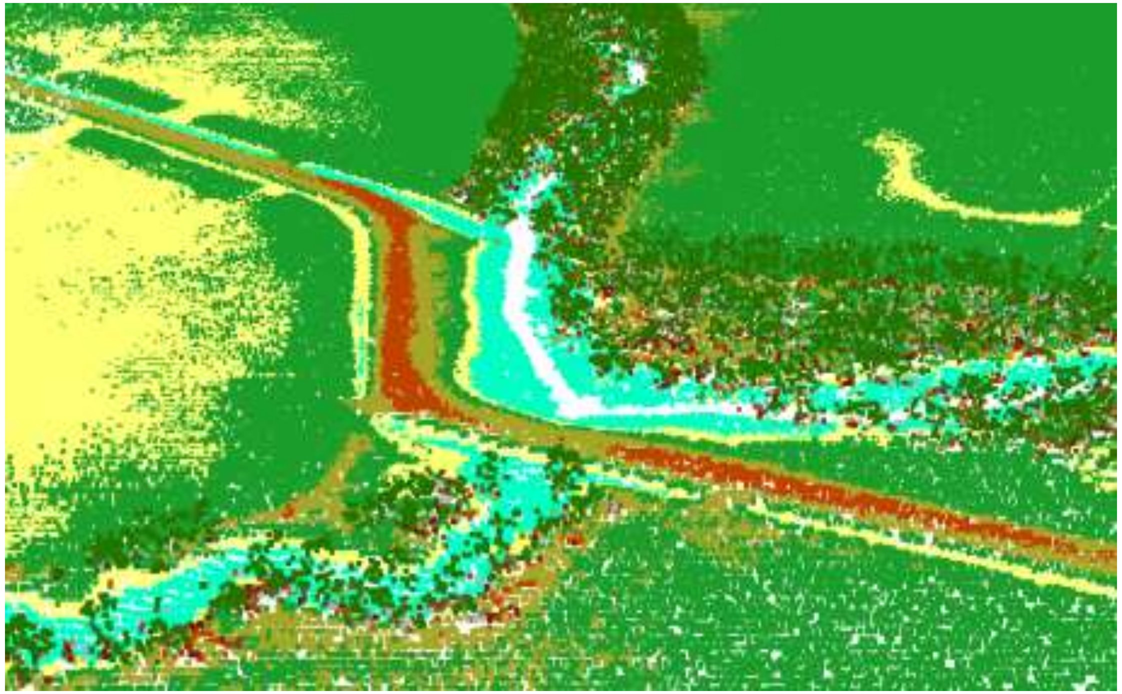
Spatial Data Processing Needs are Exploding …

3D

- Urban landscapes, virtual visits, infrastructure planning

Point Clouds

- Billions of points captured by sensors. Change detection, object recognition



implementation
increasing
social
factor
structure
sources
state
management
size
billion
drive
terms
business
nearly
processes
second
storage
scientific
flow
statistics
per
distributed
mapreduce
large
times
people
development
science
online
hand
many
simulations
databases
private
various
search
companies
technologies
research
institute
paradigm
applications
learning
high
limits
shared
international
integration
information
time
trends
algorithms
include
based
relevant
intelligence
records
market
analytics
sets
software
tools
used
process
needed
funding
new
economic
day
across
well
program
critiques
future
processing
year
queries
sensor
million
big
systems
out
visualization
traffic
determine
between
critique
years
complex
networks
even
real
exabytes
every
san
rate
insight
results
variety
amount
approaches
ability
privacy
terabytes
effective
hypothesis
huge
now
massive
uses
nas
end
manage
higher
making
countries
analysis
center
past
set
data
framework
infrastructure
national
government
petabytes
related
database
users
cost
replication
use
most
handle
internet
one
technology
definition
volume
worldwide
successful
using
initiative
parallel
world
project
would
less
challenges
work

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

ORACLE

APACHE hadoop
hadoop YARN
APACHE HBASE
HDFS
HIVE
APACHE Spark
HUE
OOZIE
APACHE kafka
A distributed streaming platform
ZOO
Lucene
FLUME
ORACLE
NOSQL DATABASE
Apache Solr
sqoop

Big Data Appliance

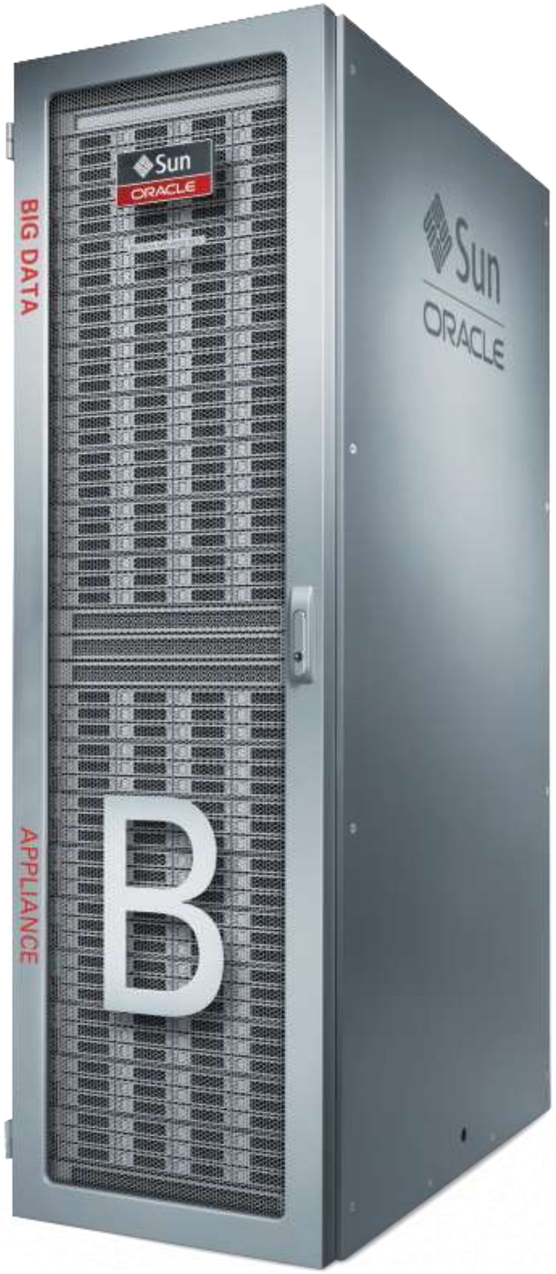
Sun Oracle X6-2L nodes with (**per node**):

- 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



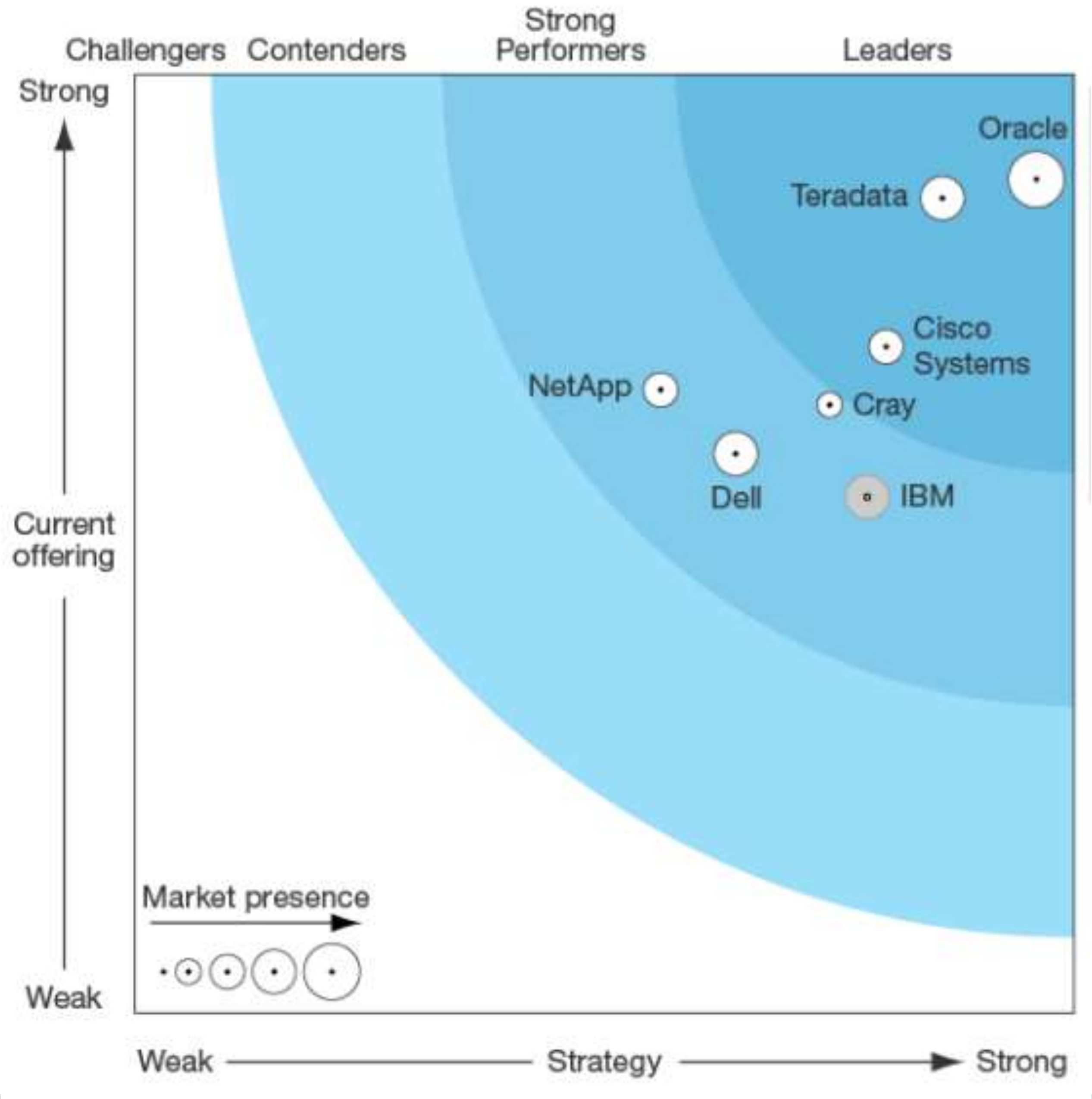
Big Data Cloud Service

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.

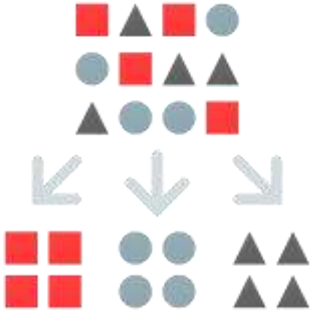
- Oracle Big Data Connectors
  - Oracle SQL Connector for Hadoop Distributed File System
  - Oracle Loader for Hadoop
  - Oracle XQuery for Hadoop
  - Oracle Data Integrator Enterprise Edition
- Oracle R Advanced Analytics for Hadoop
- Oracle Big Data Spatial and Graph

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).

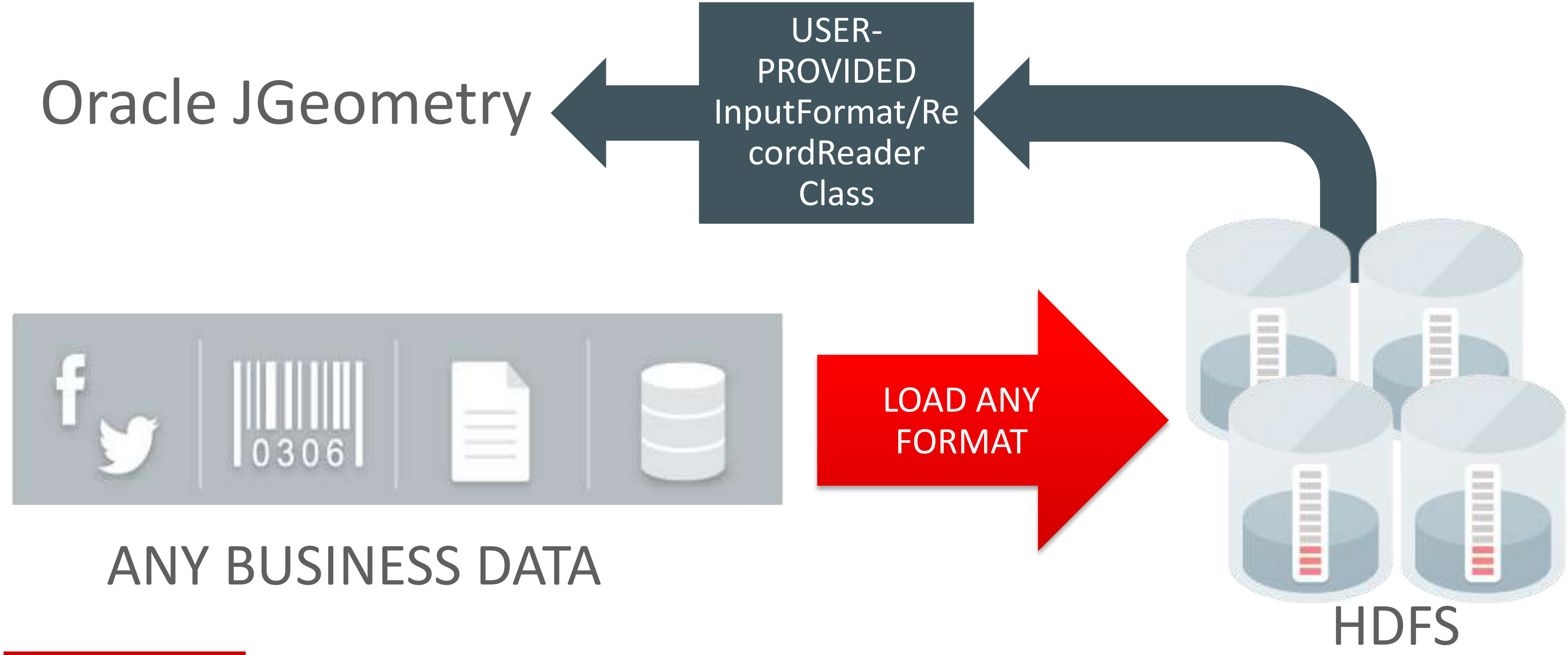


Categorization and filtering based on location and proximity

Preparation, validation and cleansing of Spatial and Raster data

Visualizing and displaying results on a map

Store any data with spatial information in HDFS



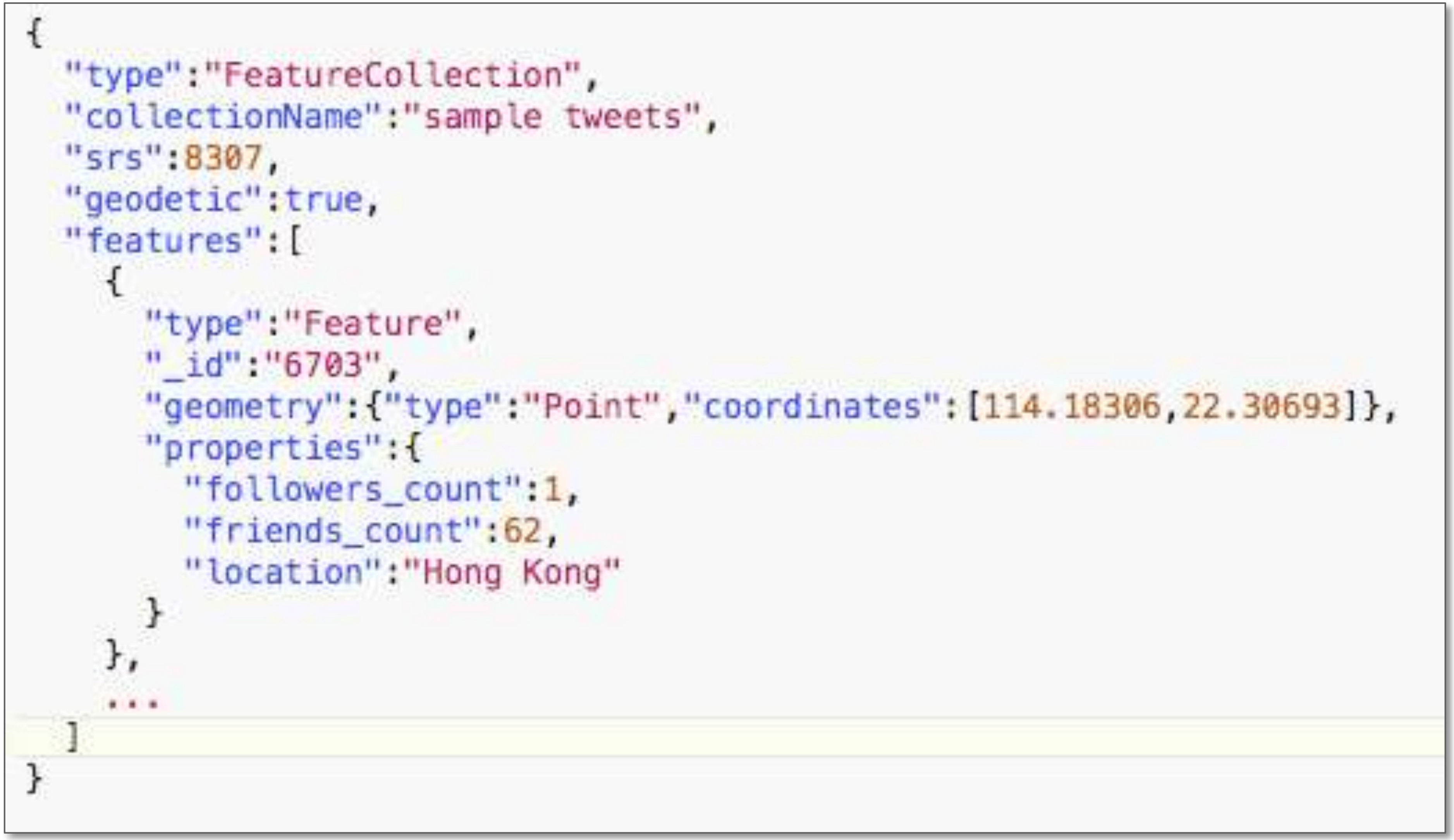
Supports All Vector Data

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

- 2D and 3D structures
- Projected and Geodetic

- Topological and distance operations
  - Anyinteract, inside, distance, length, simplify, buffer, PointInPolygon

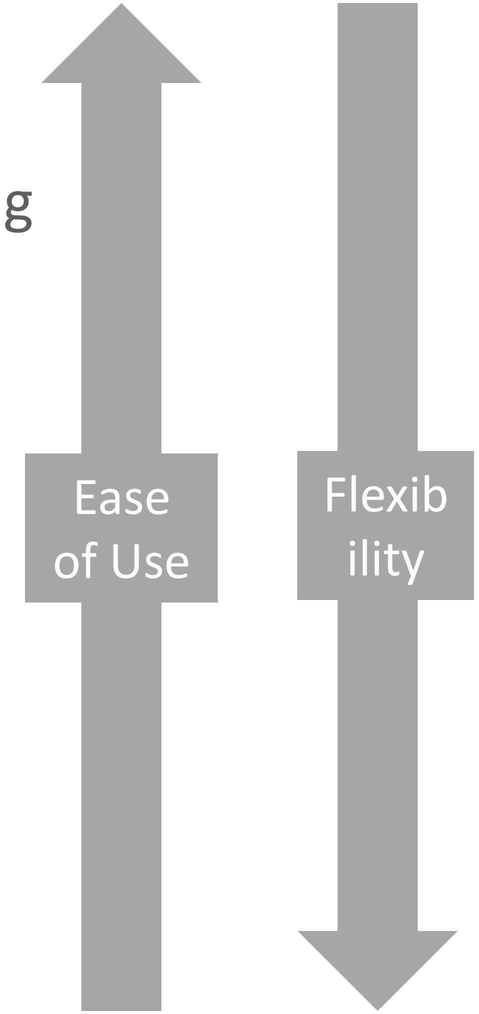
GeoJSON



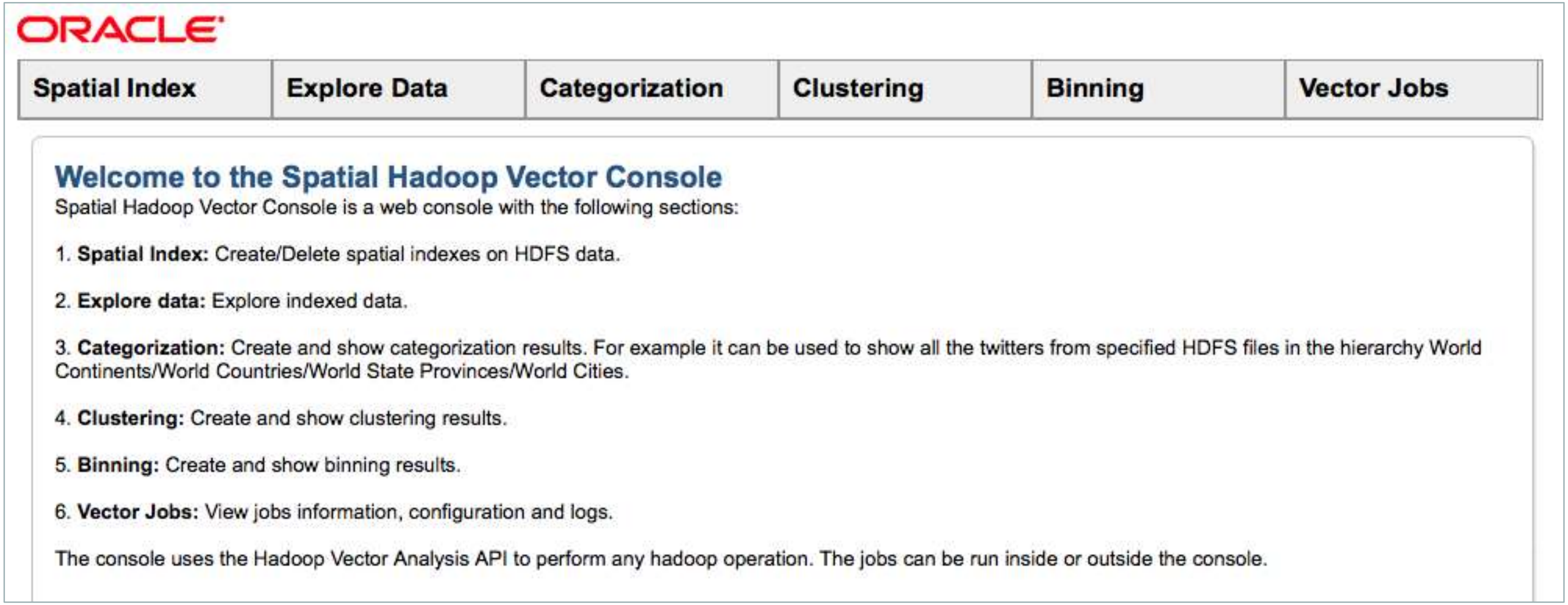
```
{
  "type":"FeatureCollection",
  "collectionName":"sample tweets",
  "srs":8307,
  "geodetic":true,
  "features":[
    {
      "type":"Feature",
      "_id":"6703",
      "geometry":{"type":"Point","coordinates":[114.18306,22.30693]},
      "properties":{
        "followers_count":1,
        "friends_count":62,
        "location":"Hong Kong"
      }
    },
    ...
  ]
}
```

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



Spatial Console



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

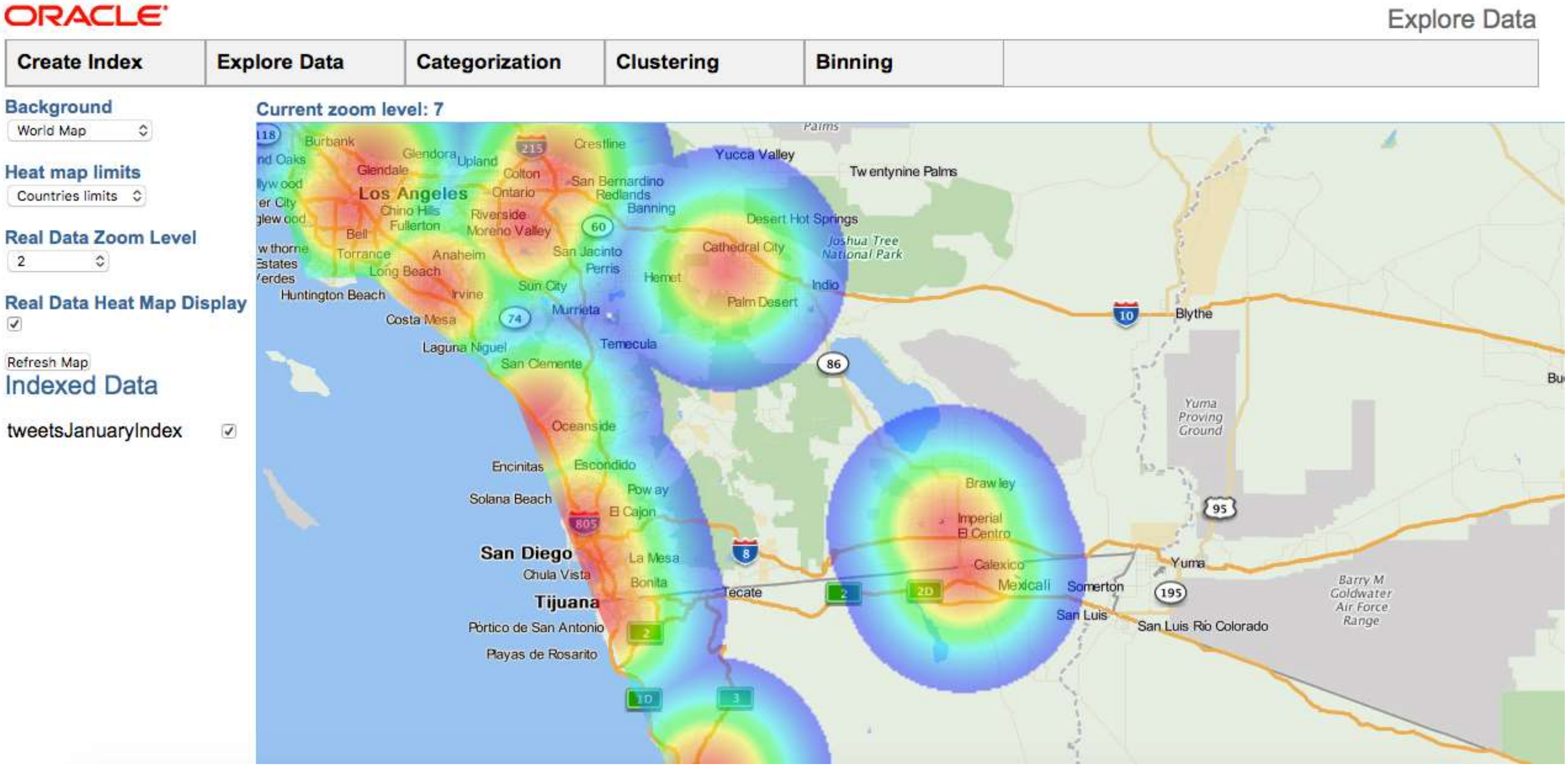
Are these data points related?

- 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 ...

Golden Gate Brg
101
Marine Dr

- Uses the **Geonames** data set

* Golden Gate Bridge (in Korean)



Explore Data

Create Index | Explore Data | Categorization | Clustering | Binning

Background
World Map

Heat map limits
Countries limits

Real Data Zoom Level
2

Real Data Heat Map Display
☑

Refresh Map

Indexed Data

tweetsJanuaryIndex ☑

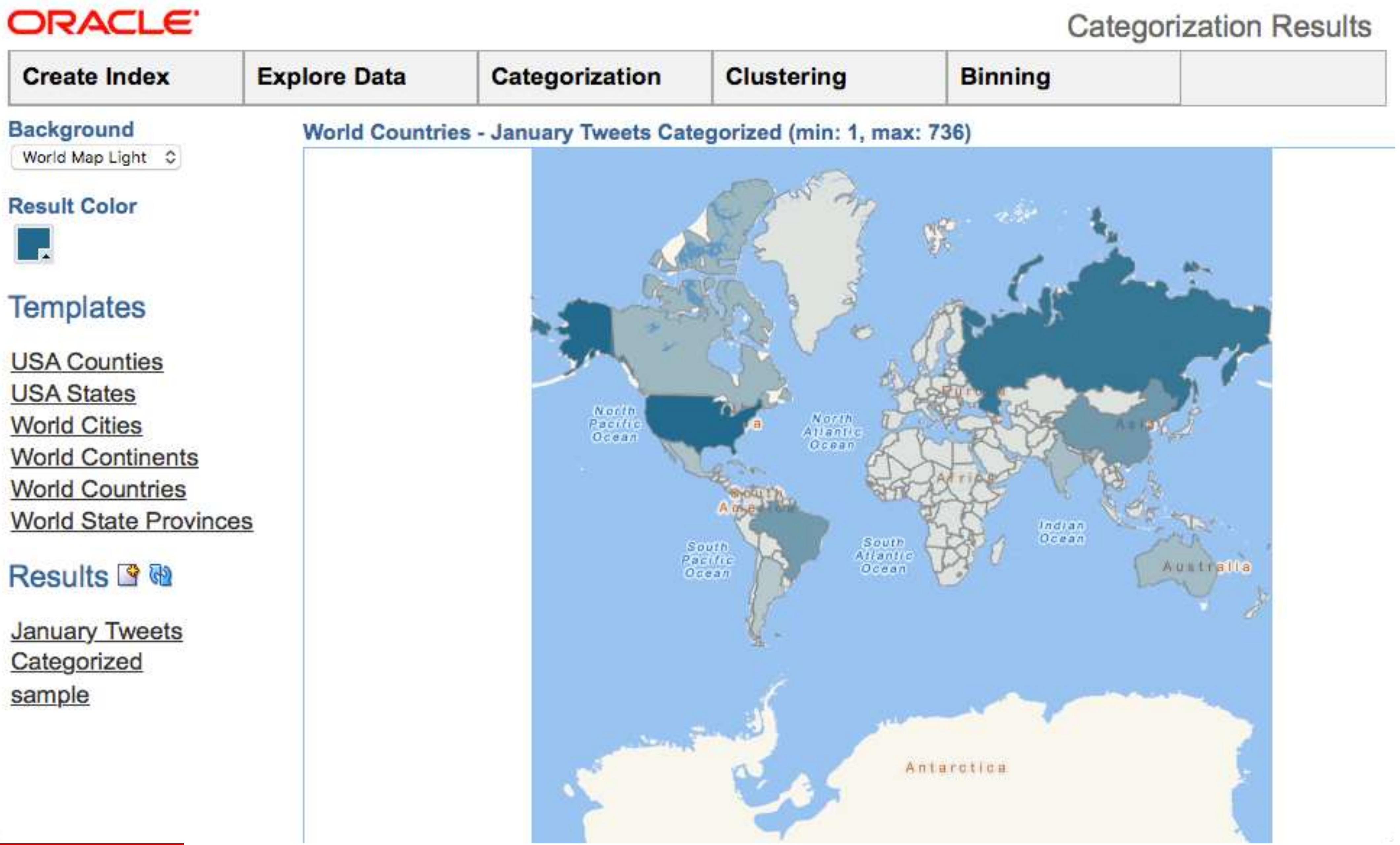
Current zoom level: 7

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



ORACLE

Categorization Results

Create Index | Explore Data | Categorization | Clustering | Binning

Background

World Map Light

Result Color

🟦

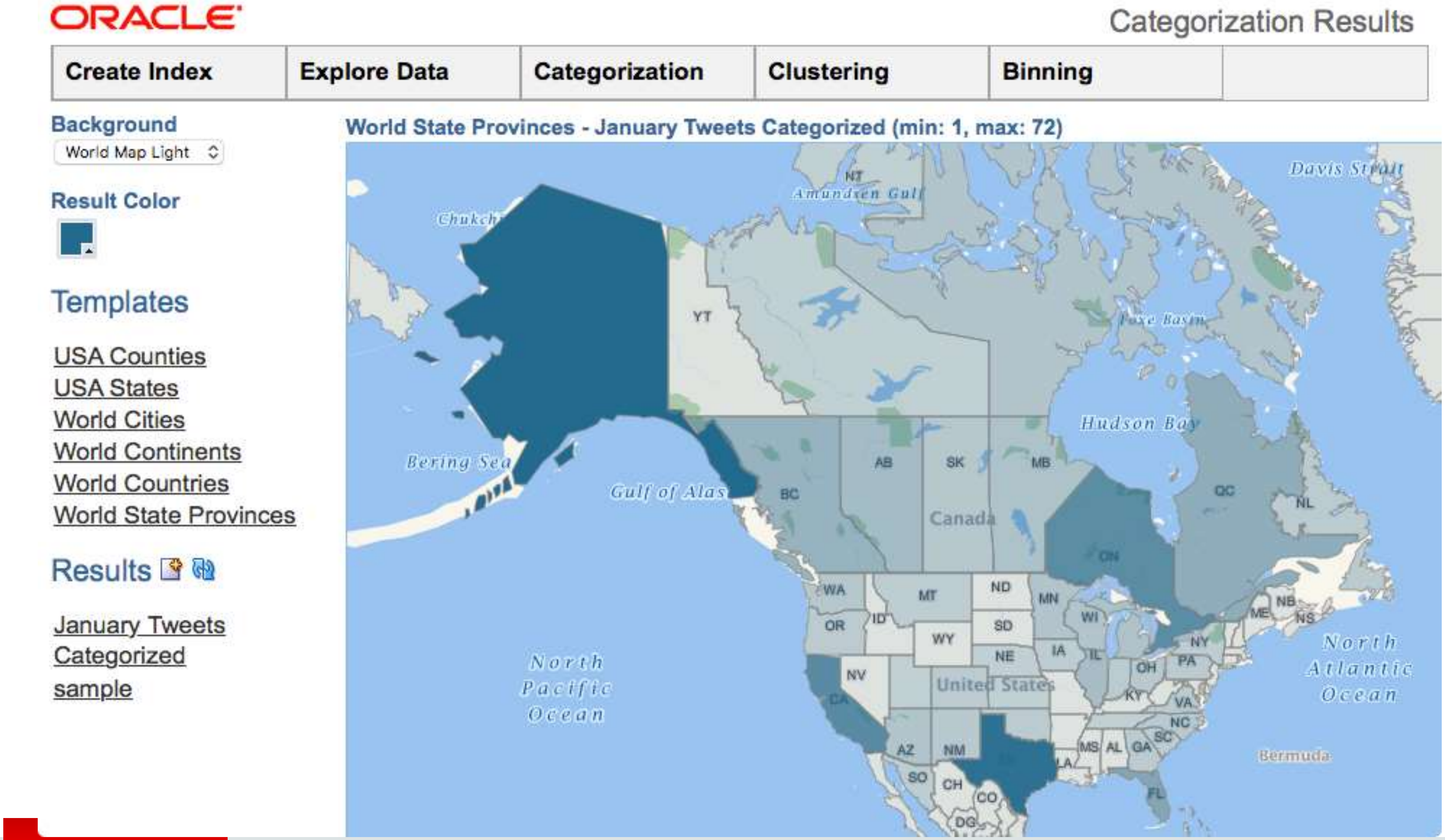
Templates

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

Results

January Tweets Categorized sample

World Countries - January Tweets Categorized (min: 1, max: 736)



ORACLE

Categorization Results

Create Index	Explore Data	Categorization	Clustering	Binning

Background

World Map Light

Result Color

Templates

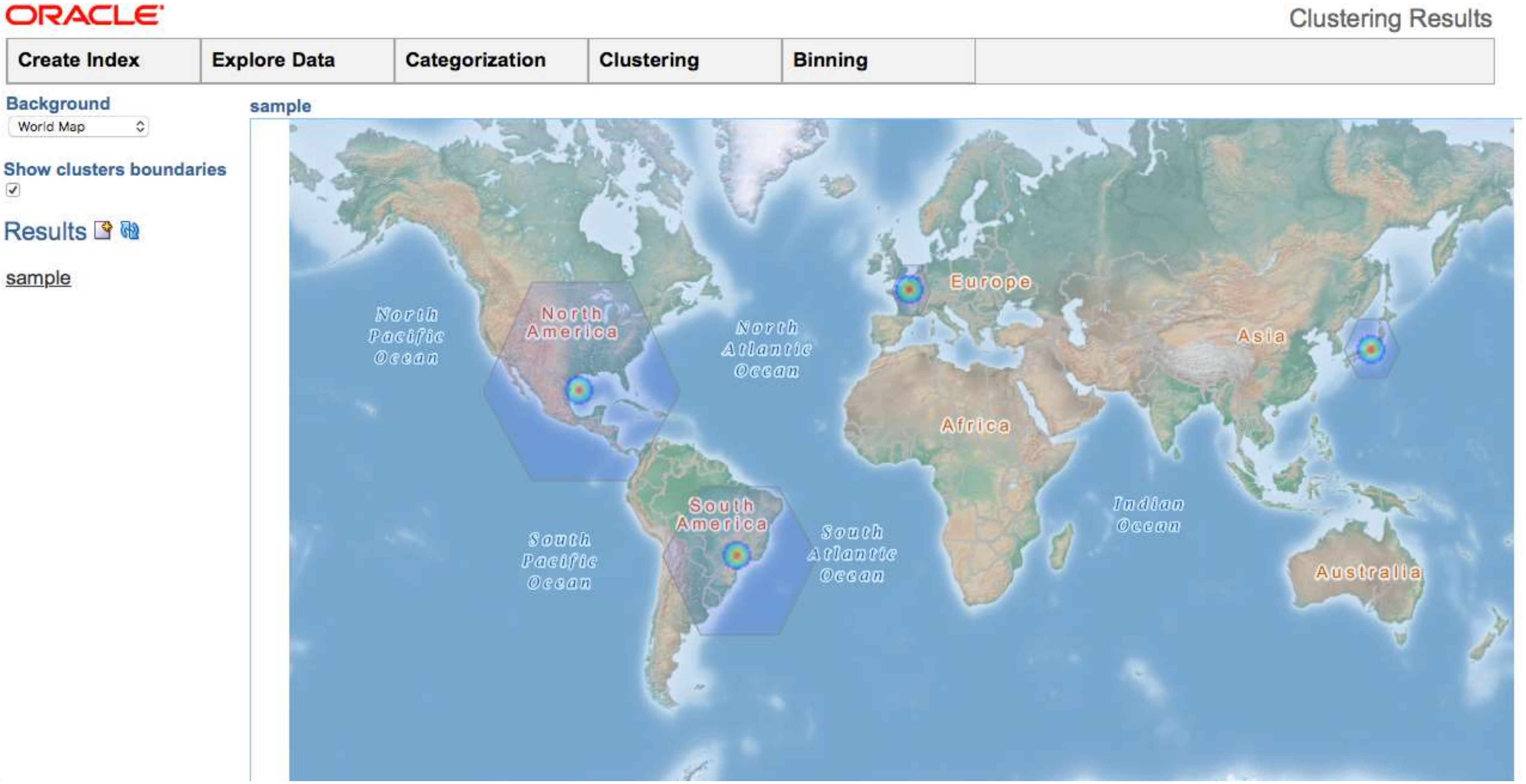
- USA Counties
- USA States
- World Cities
- World Continents
- World Countries
- World State Provinces

Results

- January Tweets Categorized sample

World State Provinces - January Tweets Categorized (min: 1, max: 72)

Spatial Clustering



Clustering Results

Background

World Map

Show clusters boundaries

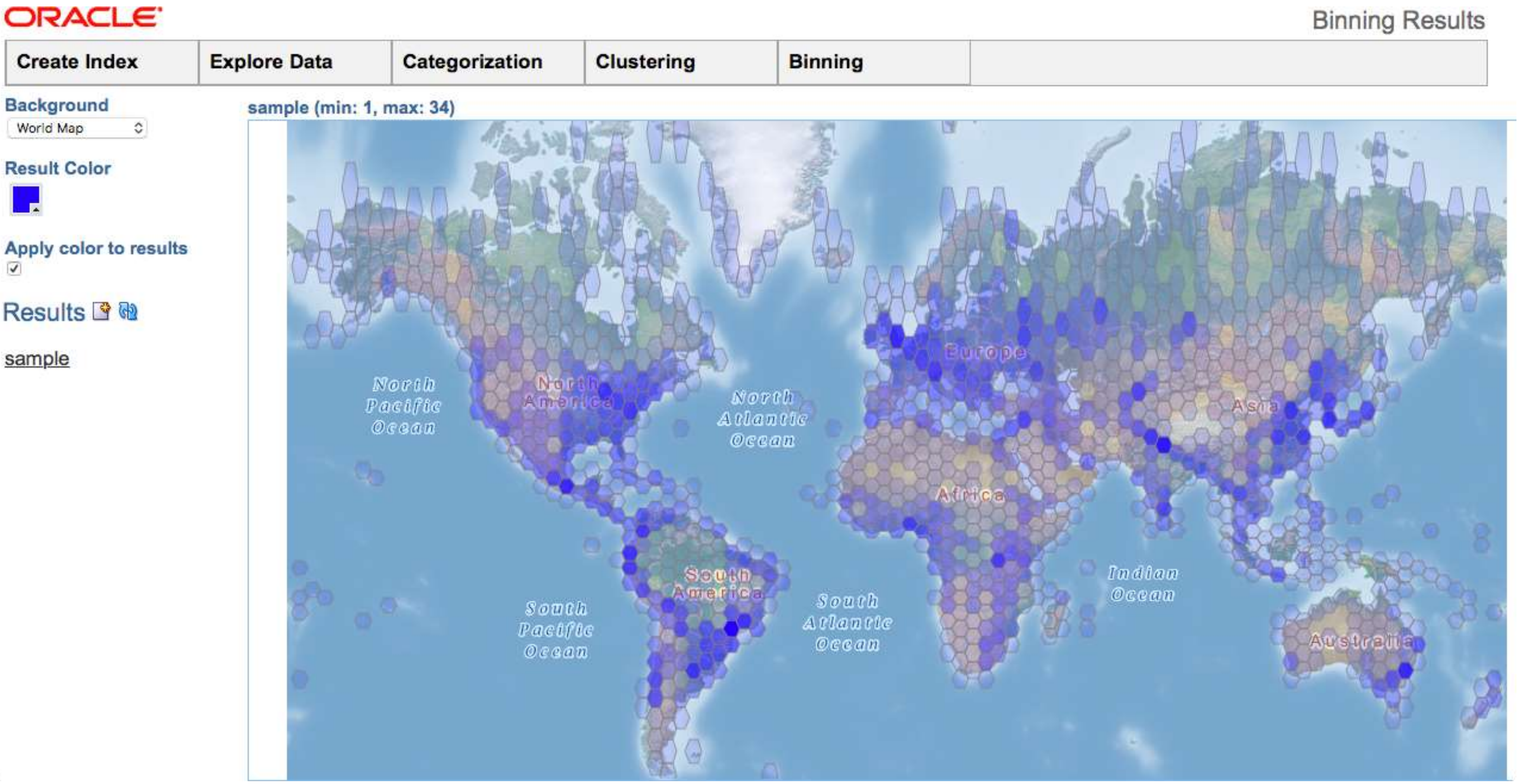
☑

Results

sample

sample

Spatial Binning



Background

World Map

Result Color

Apply color to results

☑

Results

sample

sample (min: 1, max: 34)

Run a Spatial Processing Job

```
hadoop jar $API_LIB_DIR/sdohadoop-vector.jar oracle.spatial.hadoop.vector.mapred.job.Categorization \
  -libjars $HADOOP_LIB_JARS \
  spatialOperation=IsInside \
  input=/user/oracle/HOL/tweets.json \
  output=/user/oracle/HOL/catOutputEuro \
  inputFormat=oracle.spatial.hadoop.vector.geojson.mapred.GeoJsonInputFormat \
  recordInfoProvider=oracle.spatial.hadoop.vector.geojson.GeoJsonRecordInfoProvider \
  srid=8307 geodetic=true tolerance=0.5 \
  hierarchyInfo=hol.EuroHierarchyInfo \
  hierarchyIndex=/user/oracle/HOL/hierarchyIndex \
  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

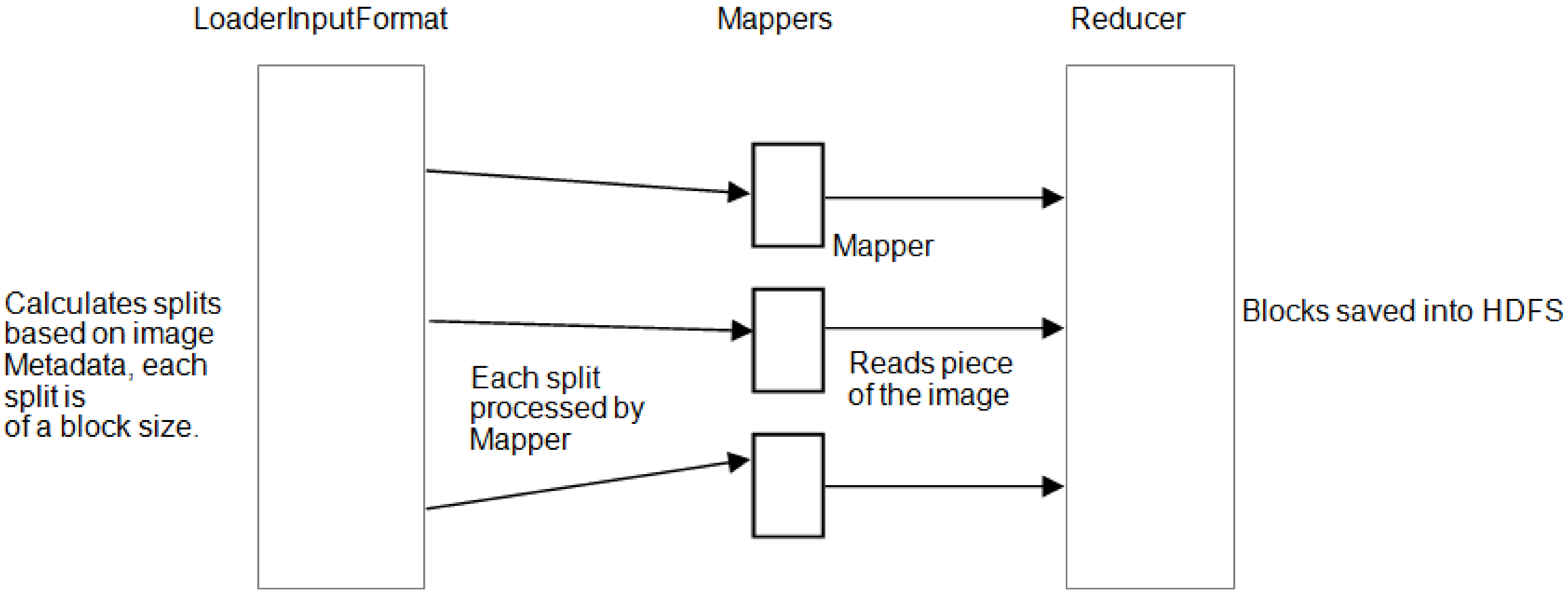
- **HDFS storage** for the image or raster files
  - We can support dozens of file formats (GDAL supported formats)
  - Images are geo-referenced
  - Images can be in different coordinate systems and resolutions
- **Raster Processing**
  - **Loader** to load raster data from NFS to HDFS
  - **Mosaic** 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.

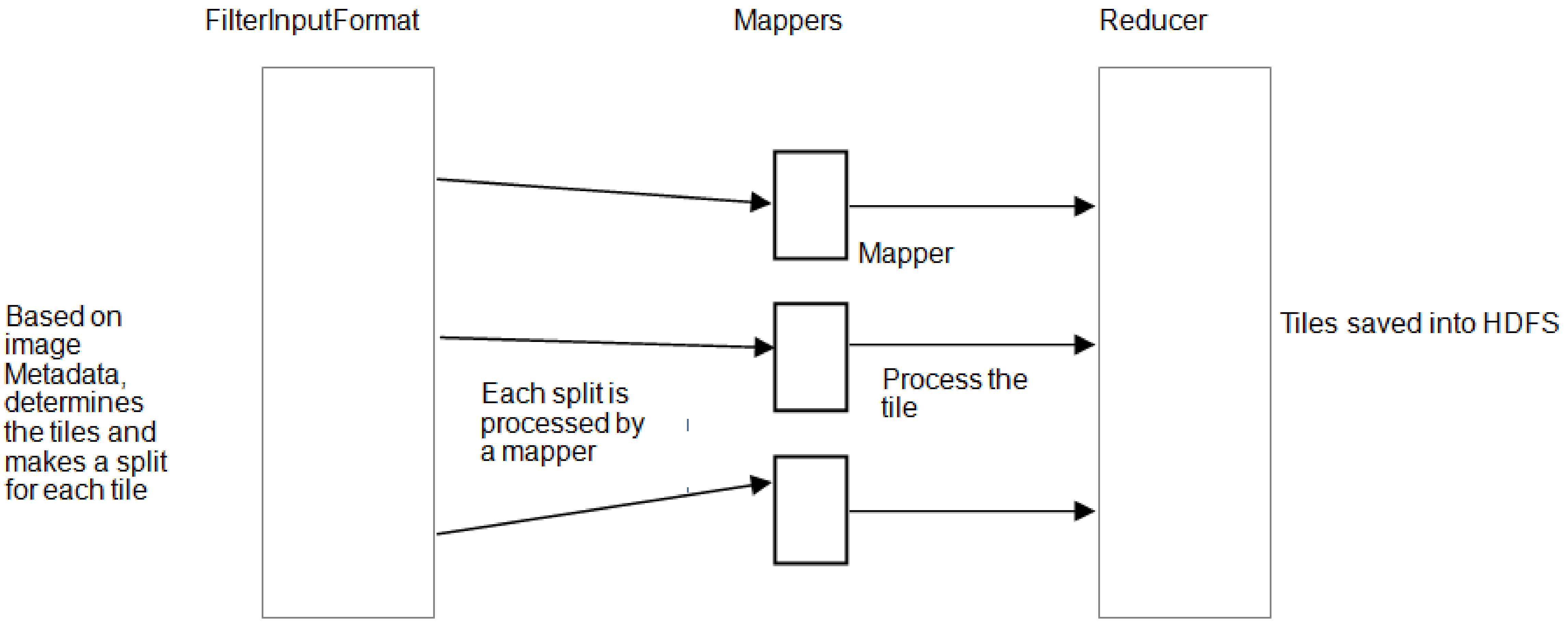
Each split processed by Mapper

Mapper

Reads piece of the image

Blocks saved into HDFS

Raster Processing Map/Reduce Job


FilterInputFormat
Mappers
Reducer
Mapper
Based on image Metadata, determines the tiles and makes a split for each tile
Each split is processed by a mapper
Process the tile
Tiles saved into HDFS

Subset / Process / Mosaic Operation

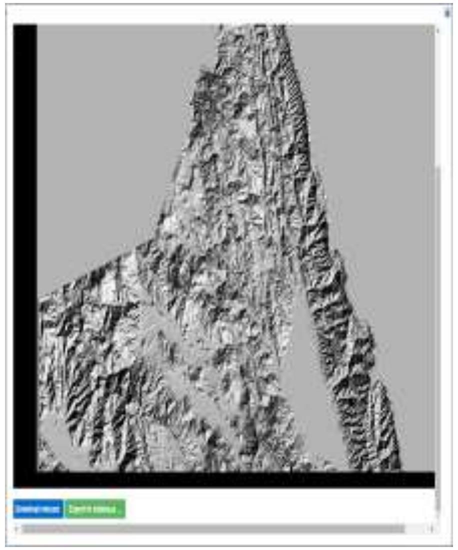
- Find the set of images from a given catalogue covering a user specified region
- The new images have user-specified resolution and coordinate system
- Apply pixel-level processing (“raster algebra”)
- Mosaic the input images to deal with gaps and overlaps
- Create a new image with the chosen file format

Raster Algebra Processing

Local Map algebra operations

localnot
localif
localadd
localsubstract
localmultiply
localdivide
localpow
localsqrt
localround
locallog
locallog10
localfloor
localceil
localnegate
localabs
localsin
localcos
localtan
localsinh
localcosh
localtanh
localasin
localacos
localatan
localdefined
localundefined

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

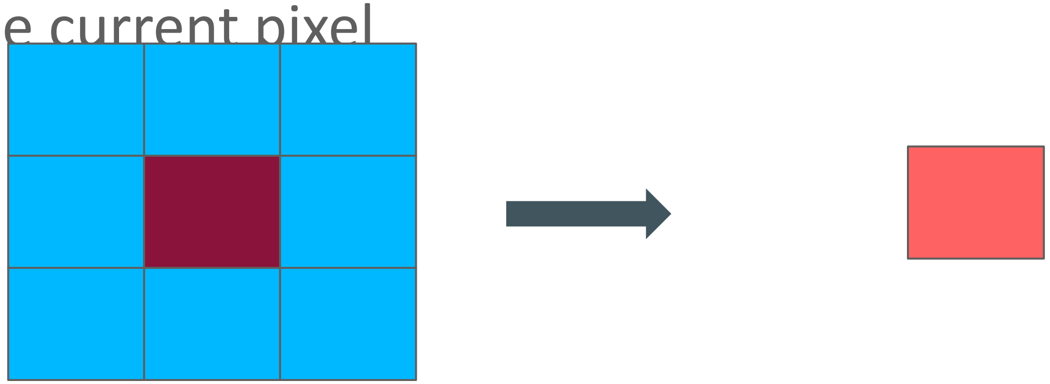
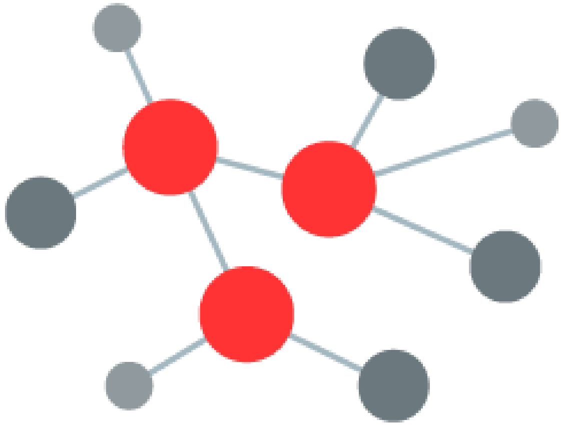


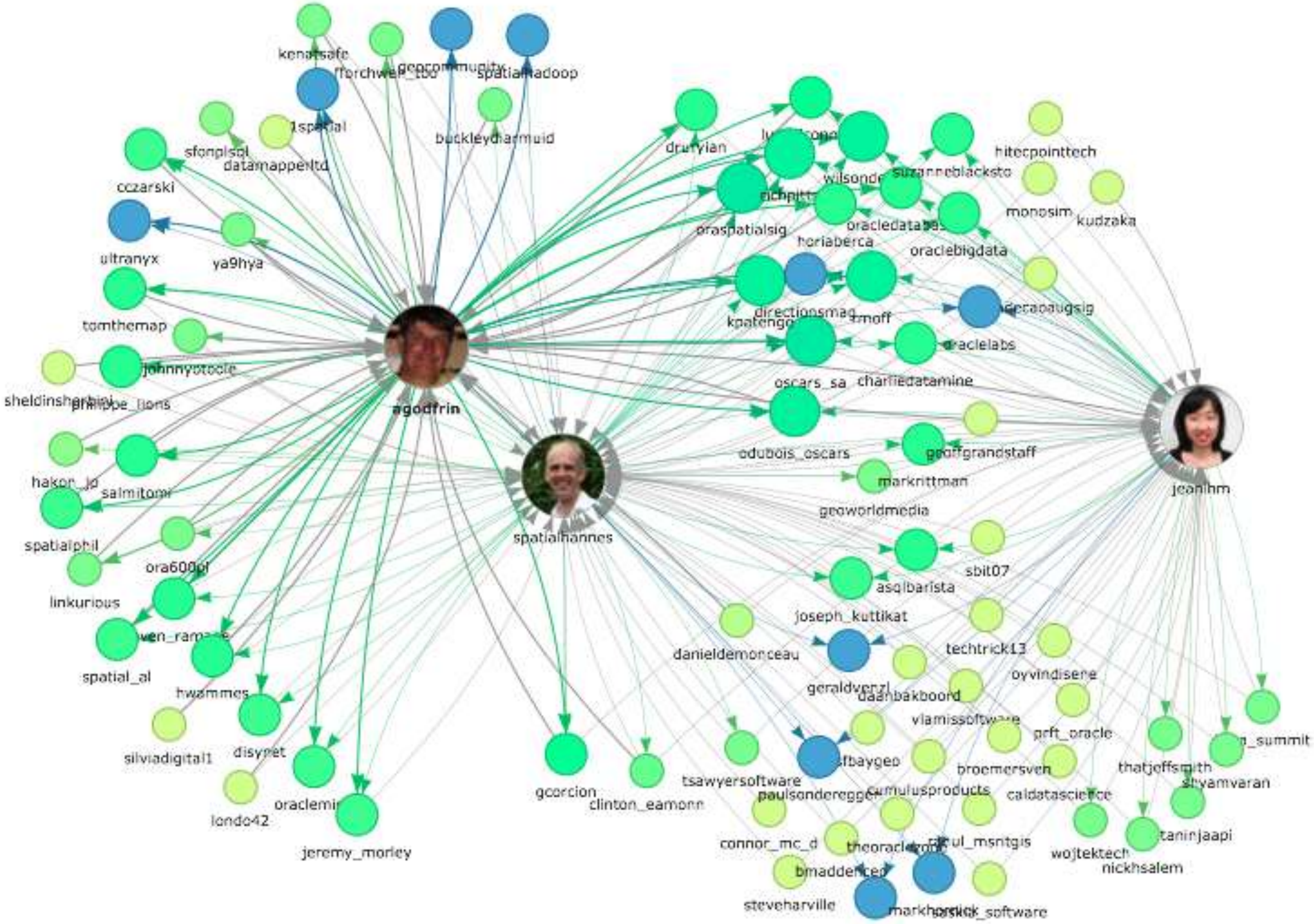
Image Server Console

- **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**


kenatsafe
fforchwell_too
geocommunity
spatialhadoop
1spatial
buckleydiarmuid
sfonplsql
datamapperltd
cczarski
ultranyx
ya9hya
tomthemap
johnnyotoole
sheldinsherbini
philippe_lions
hakon_jo
salmitomi
spatialphil
ora600pl
linkurious
ven_ramaeke
spatial_al
hwammes
silviadigital1
disynet
londo42
oraclemi
jeremy_morley
agodfrin
spatialhannes
gcorcion
clinton_eamonn
druryian
wilsonde
suzanneblacksto
richpitts
oraspatialsig
oracledatabase
horiaberca
oraclebigdata
hitecpointtech
monosim
kudzaka
directionsmag
kpatengo
rmoff
oraclelabs
oscars_sa
charliedatamine
edubois_oscars
geoffgrandstaff
markrittman
geoworldmedia
sbit07
asqlbarista
joseph_kuttikat
danieldemorceau
techtrick13
geraldverzl
daanbakboord
oyvindisene
vlamissoftware
prft_oracle
fbaygeo
tsawyersoftware
broemersven
paulsonderegger
cumulusproducts
caldatascience
thatjeffsmith
shyamvaran
_summit
connor_mc_d
theoraclemo
paul_msntgis
wojtektech
taninjaapi
nickhsalem
bmaddenced
steveharville
markhornick
_software
jeanihm

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?

Tabular questions: Well-suited to SQL-like tools

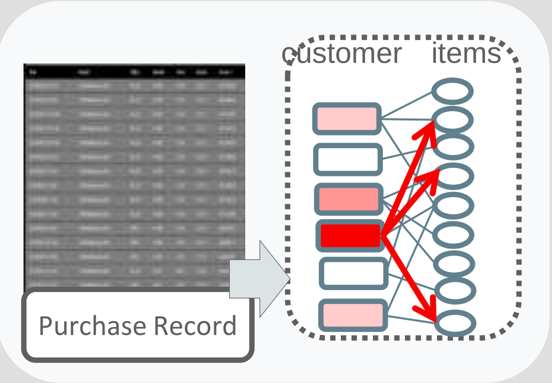
- Answers from **Connectivity**
  - Who's most influential?
  - Which supplier do I depend on the most?
  - What is the most critical link in my power grid ?

Graph questions: We need something different!

Common Graph Analysis Use Cases

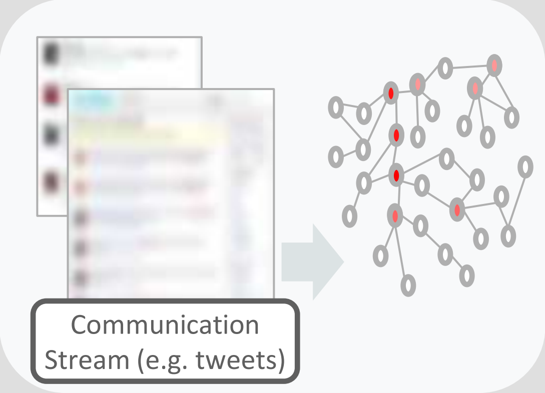
Recommend the most *similar* item purchased by *similar* people

Product Recommendation



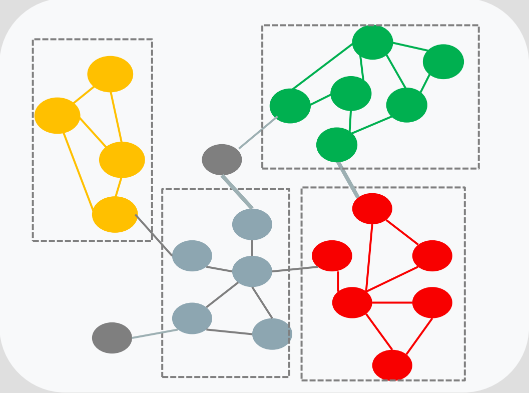
Find out people that are *central* in the given network – e.g. influencer marketing

Influencer Identification



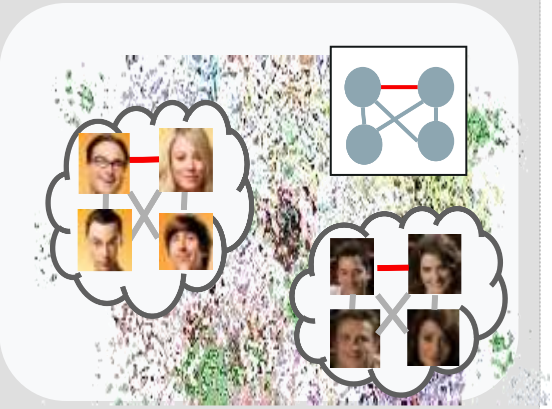
Identify group of people that are close to each other – e.g. target group marketing

Community Detection



Find out all the sets of entities that match to the given pattern – e.g. fraud detection

Graph Pattern Matching



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

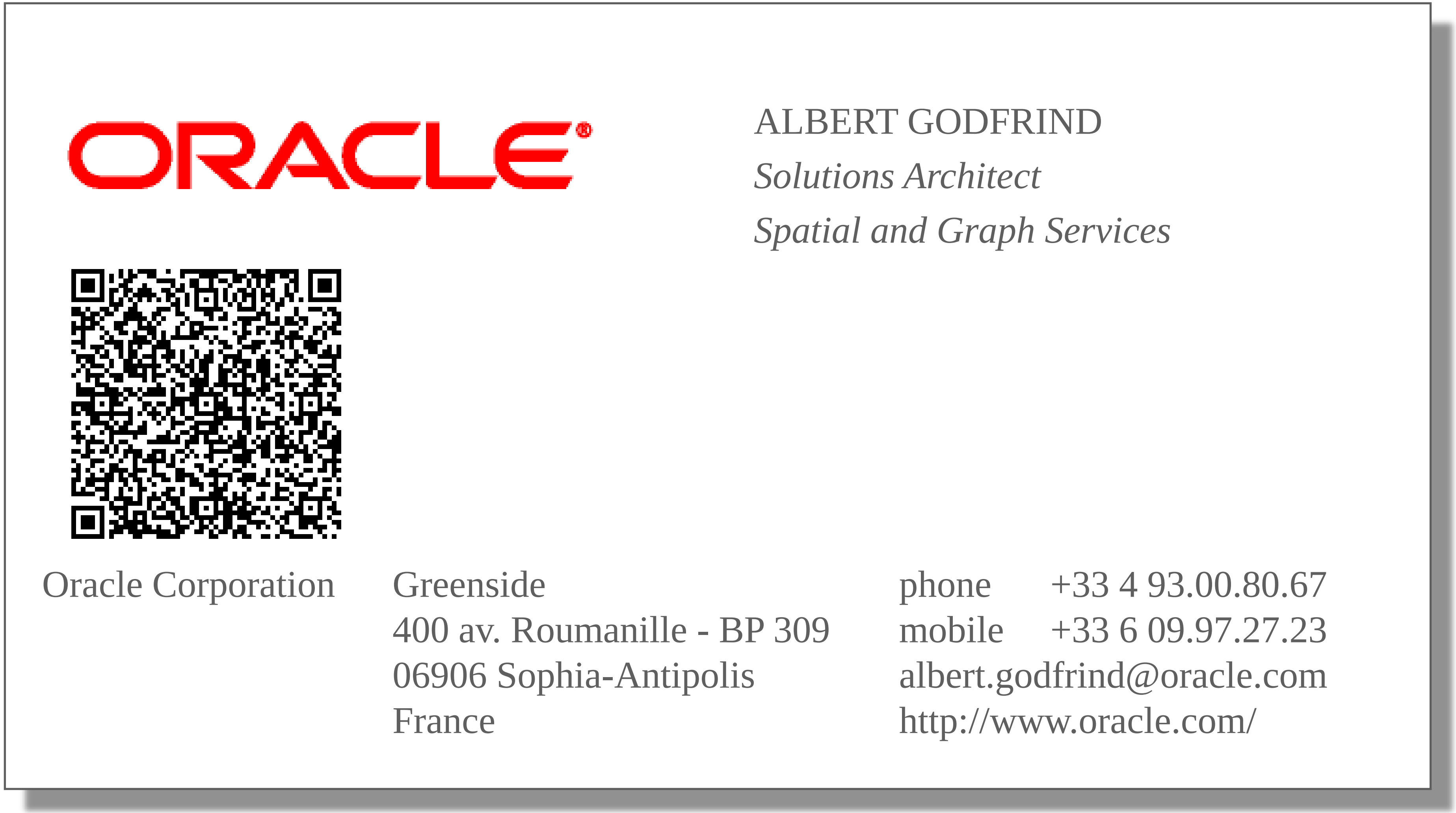
http://www.oracle.com/database/big-data-spatial-and-graph

http://www.oracle.com/technetwork/database/database-technologies/bigdata-spatialandgraph

https://blogs.oracle.com/bigdataspatialgraph

- “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

http://www.oracle.com/technetwork/database/bigdata-appliance/oracle-bigdatalite-2104726.html



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

ORACLE®