State of Geospatial BigData

Mansour Raad
thunderheadxpler.blogspot.com
mraad@esri.com
@mraad

“WHERE” IS UBIQUITOUS !

- ***Where*** is the closest ATM ?

- ***Where*** is the best location to place my store ?

- ***Where*** is UBL ?

- ***Where*** is the next Ebola/Zika outbreak ?

A BIT OF HISTORY…

With Esri Specifically :-)

~1990

Article Talk

Shapefile

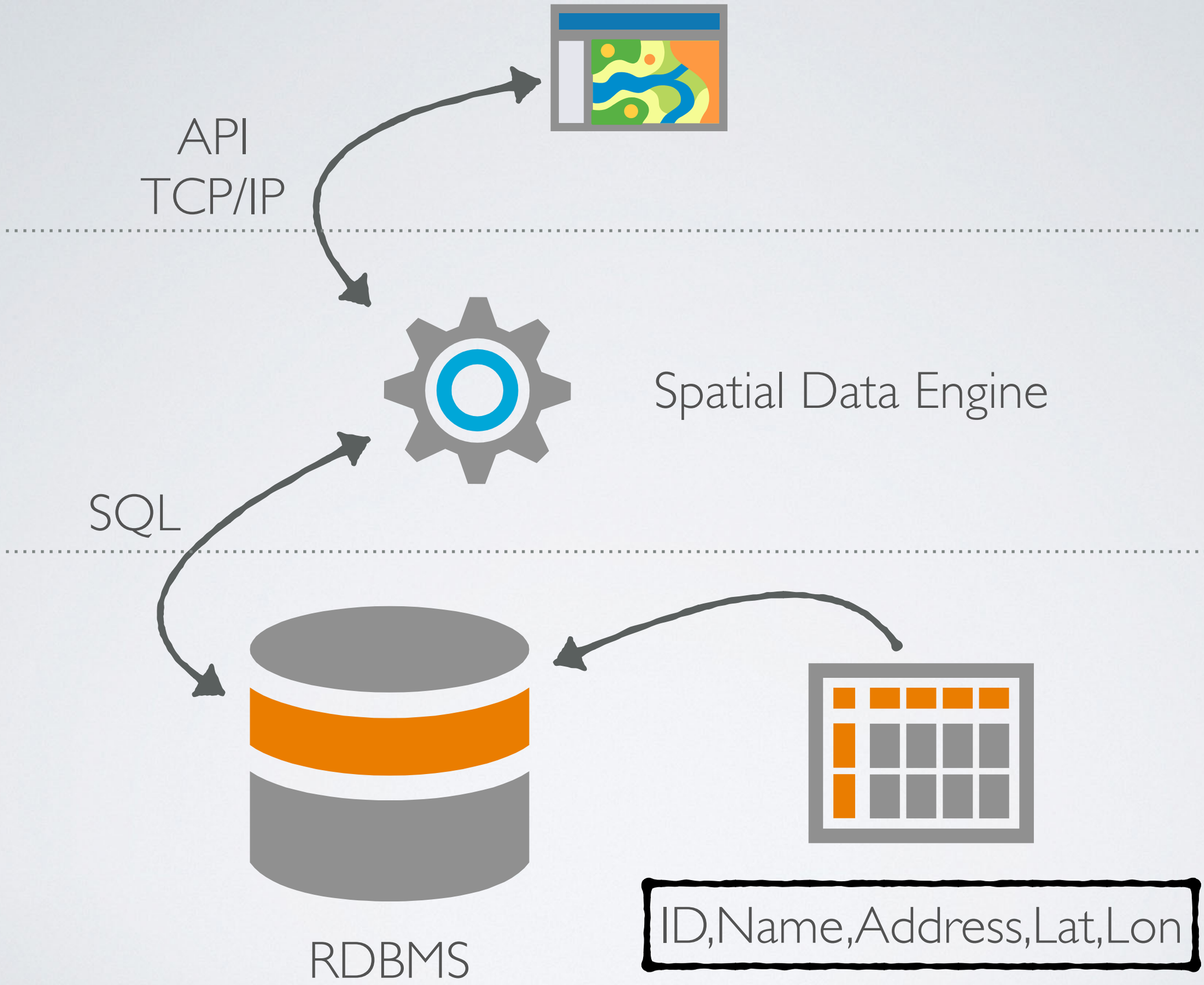
From Wikipedia, the free encyclopedia

The **Esri shapefile**, or simply a **shapefile**, is a popular geospatial vector data format for geog
developed and regulated by Esri as a (mostly) open specification for data interoperability amo
Shapefiles spatially describe vector features: points, lines, and polygons, representing, for exa
usually has attributes that describe it, such as *name* or *temperature*.

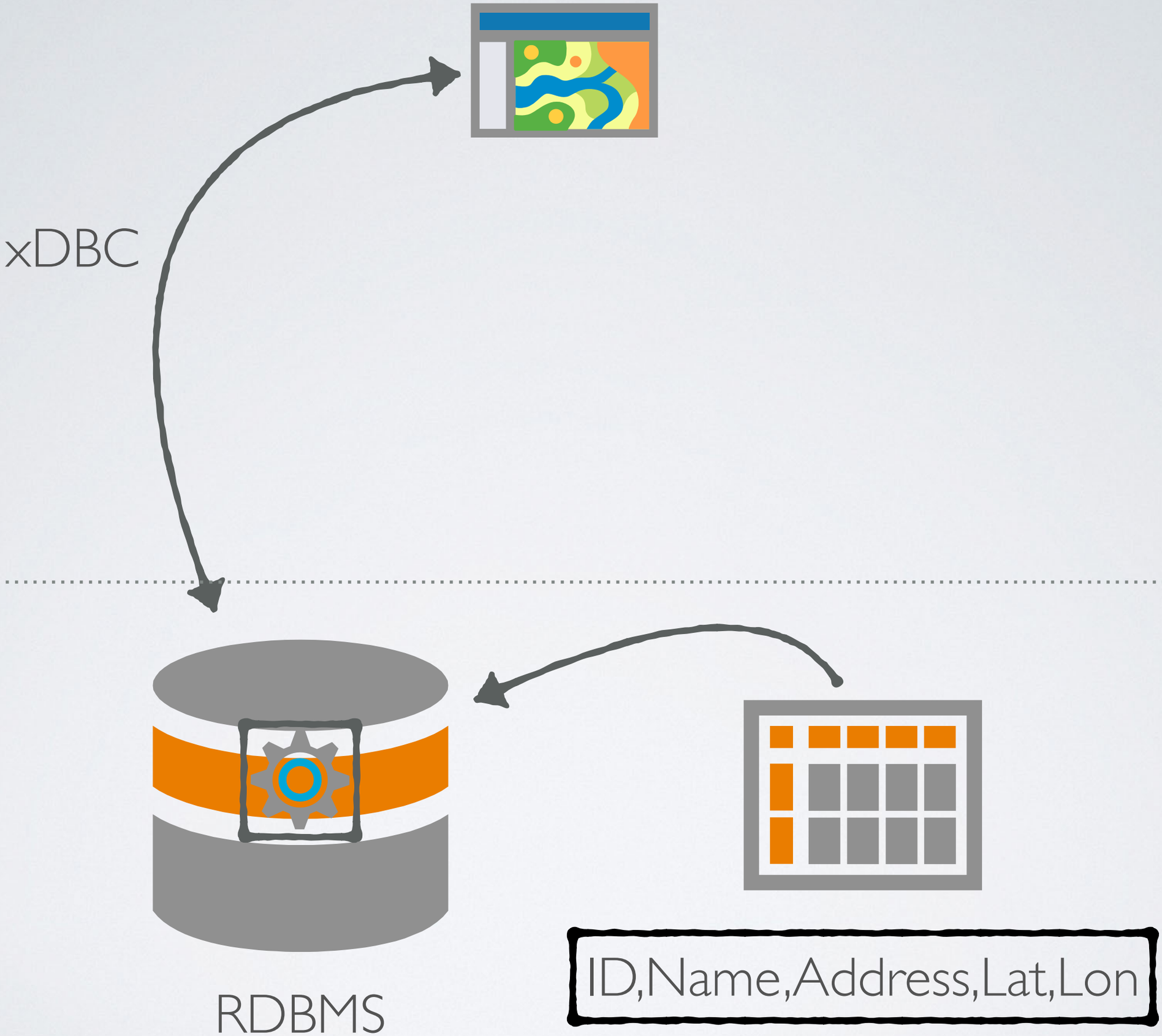
Contents [hide]

http://en.wikipedia.org/wiki/Shapefile

1995


API
TCP/IP
Spatial Data Engine
SQL
RDBMS
ID,Name,Address,Lat,Lon

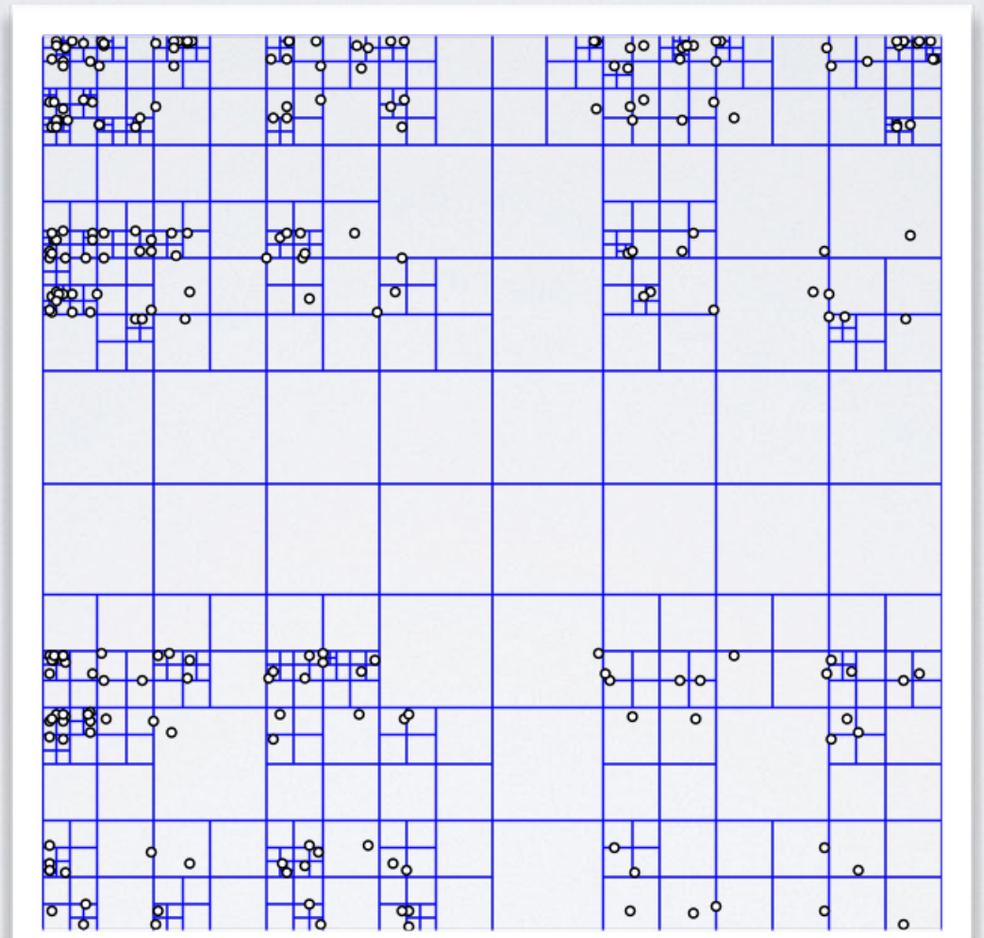
> 1996-2005


xDBC
RDBMS
ID,Name,Address,Lat,Lon

SPATIAL INDEXING

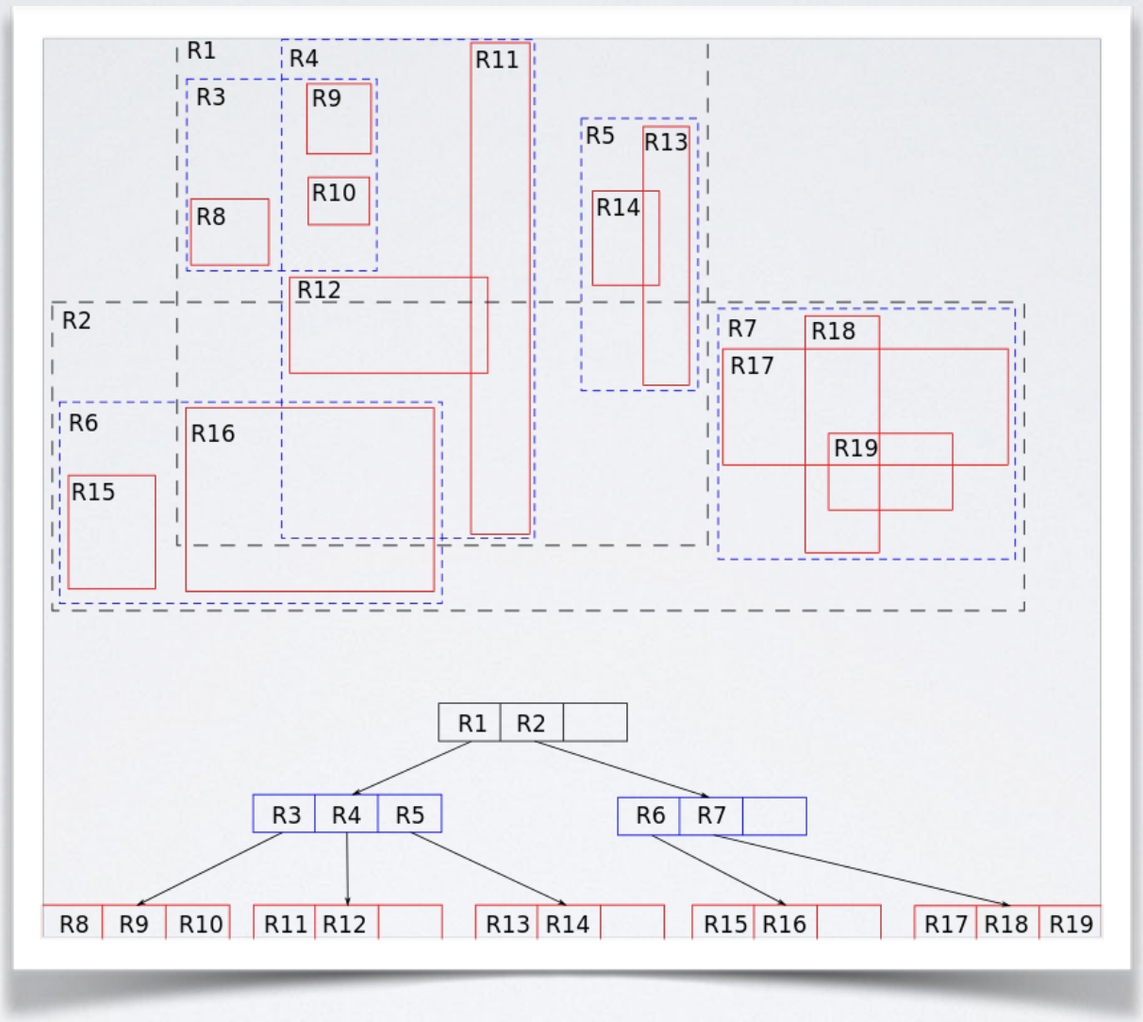
http://en.wikipedia.org/wiki/Spatial_database#Spatial_index

QUADTREE



http://en.wikipedia.org/wiki/Quadtree

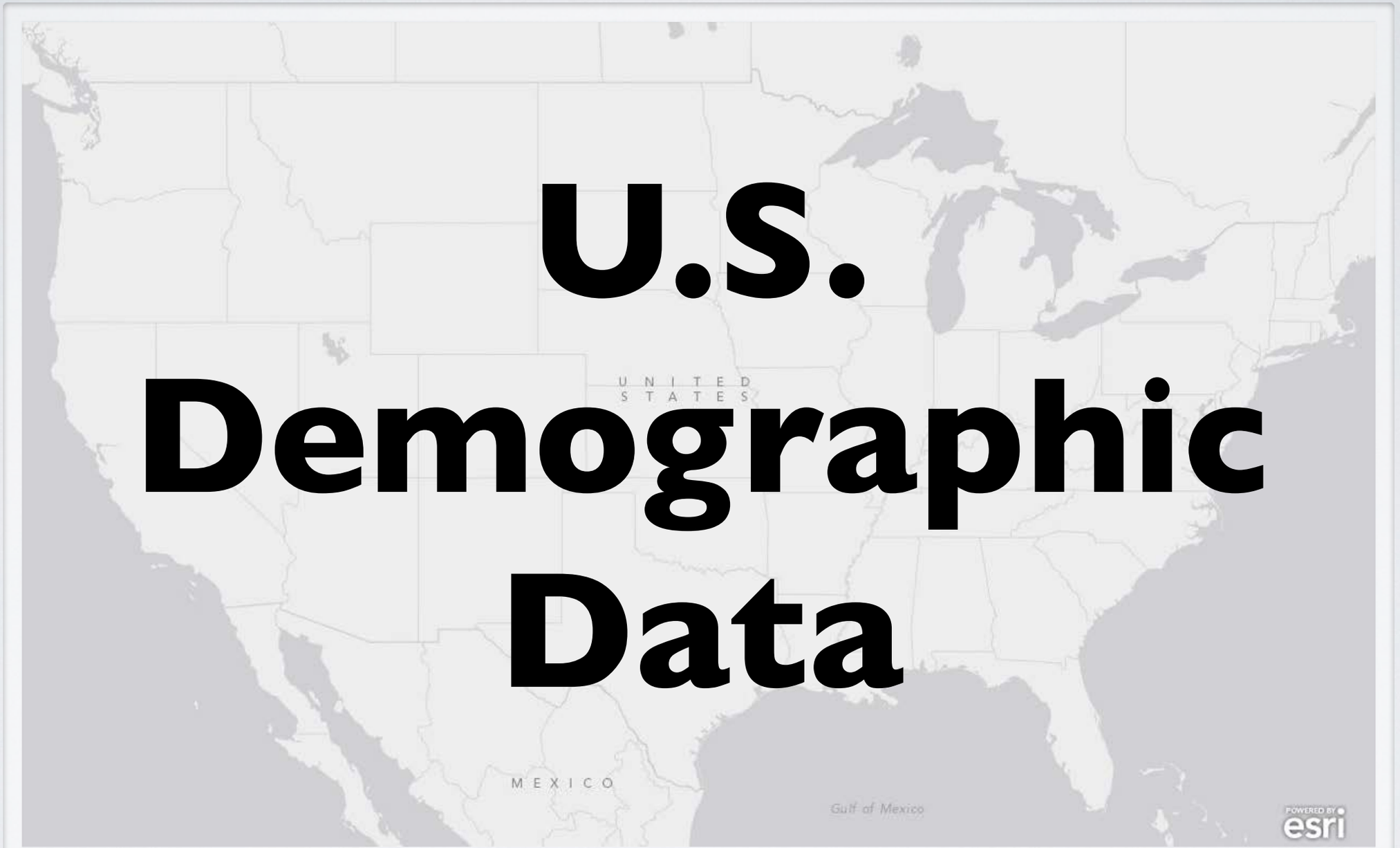
R-TREE



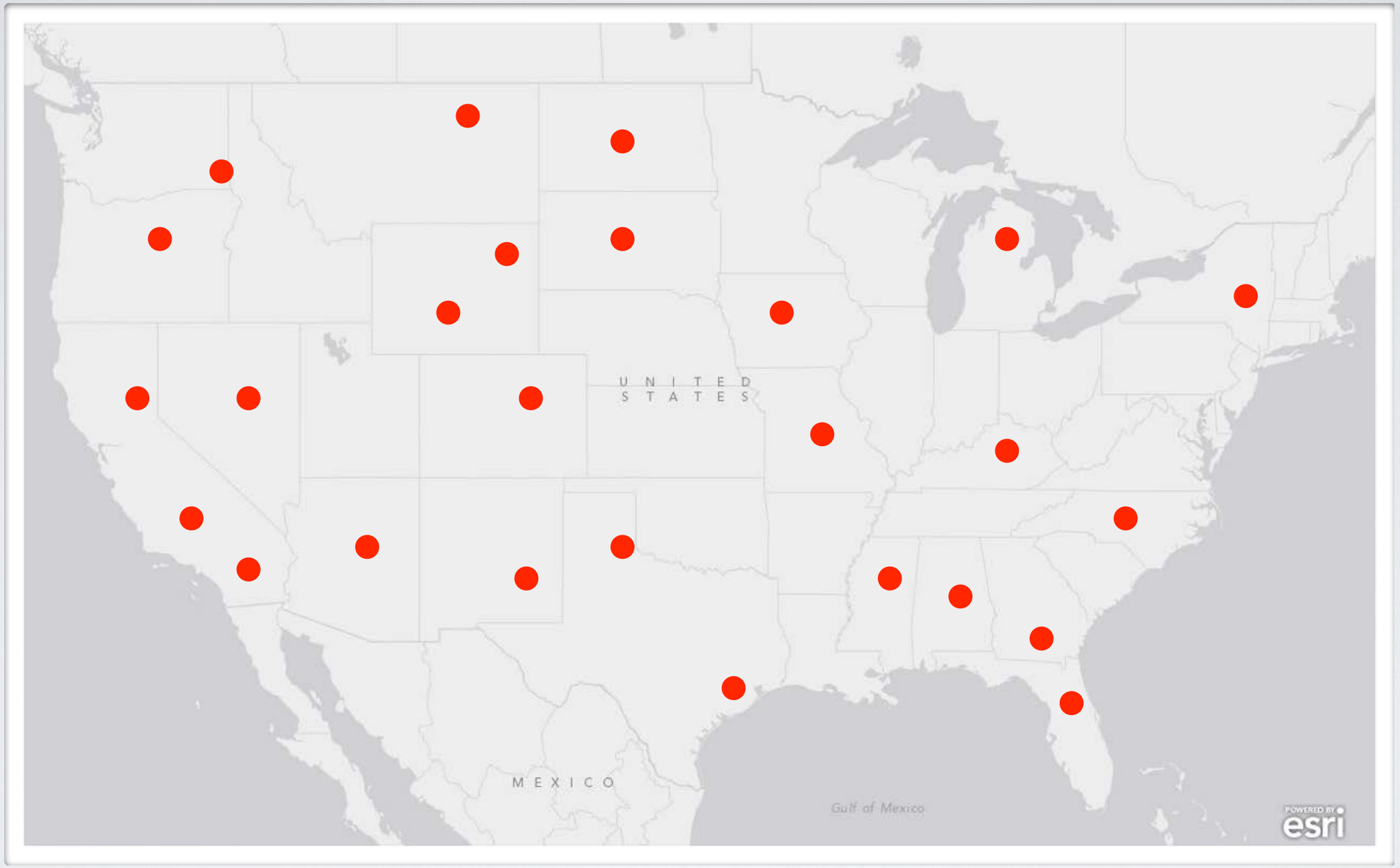
http://en.wikipedia.org/wiki/R-tree

(NOT SO) MODERN DAY…

STORY TIME…



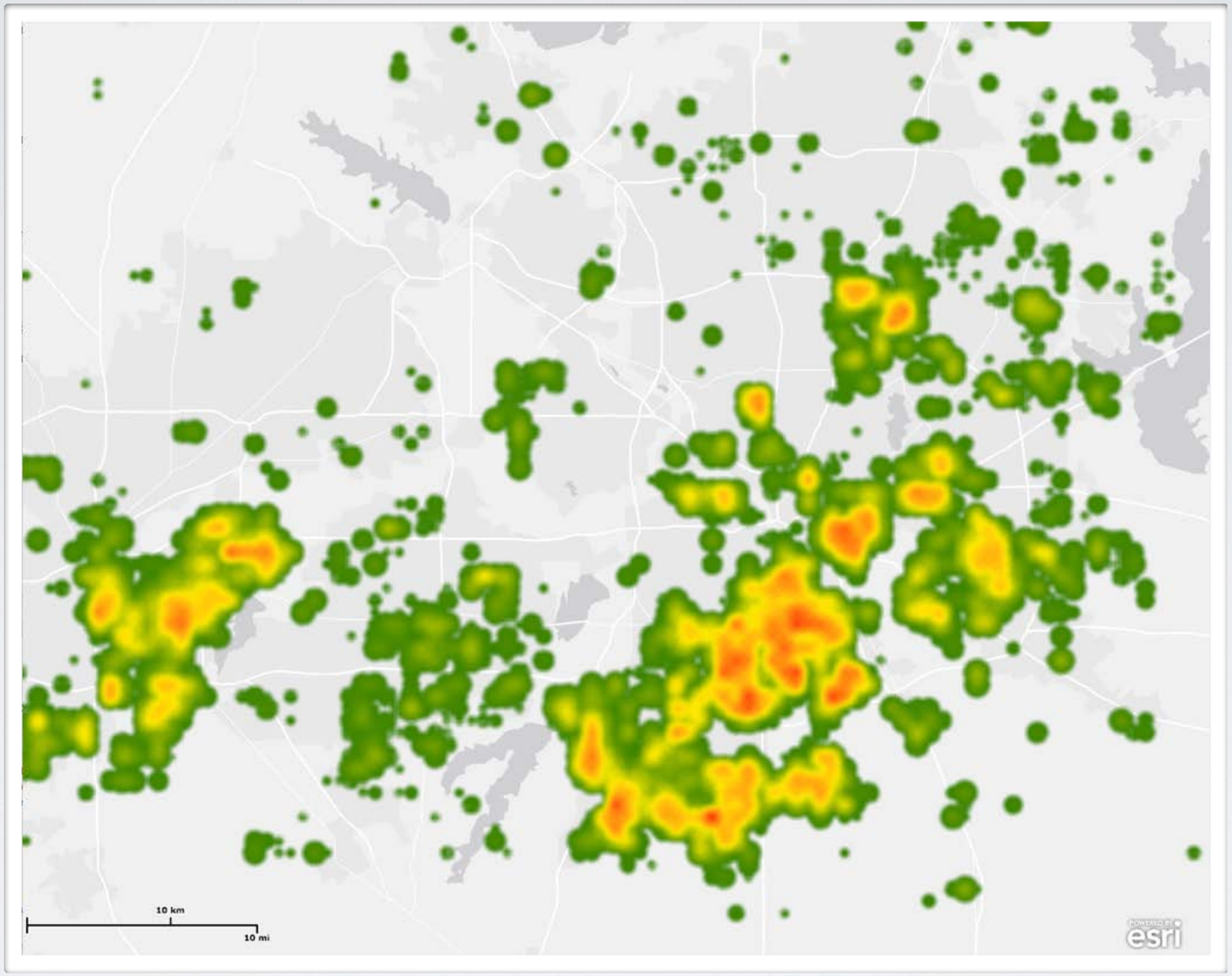
U.S. Demographic Data


UNITED
STATES
MEXICO
Gulf of Mexico
POWERED BY
esri

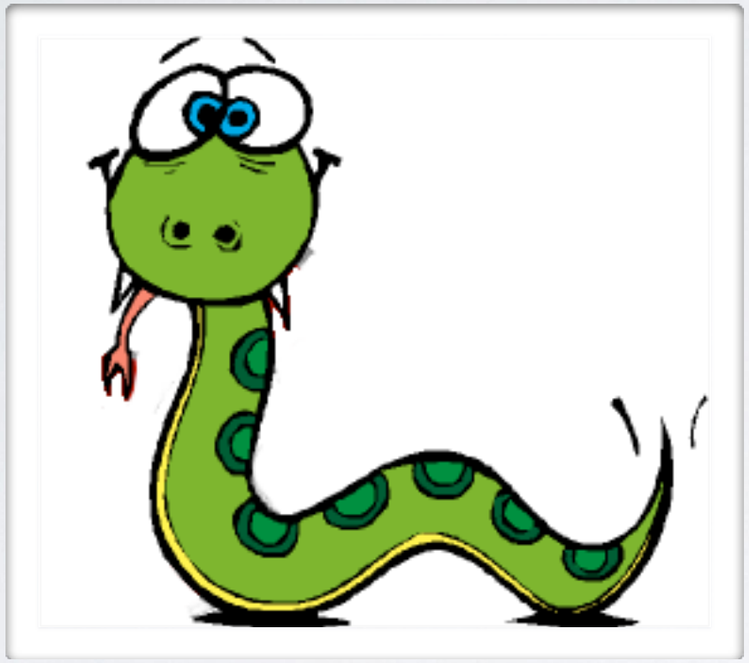
FOR EACH LOCATION
FOR EACH DEMOGRAPHIC

↓

50 MILE HEATMAP


10 km
10 mi
esri

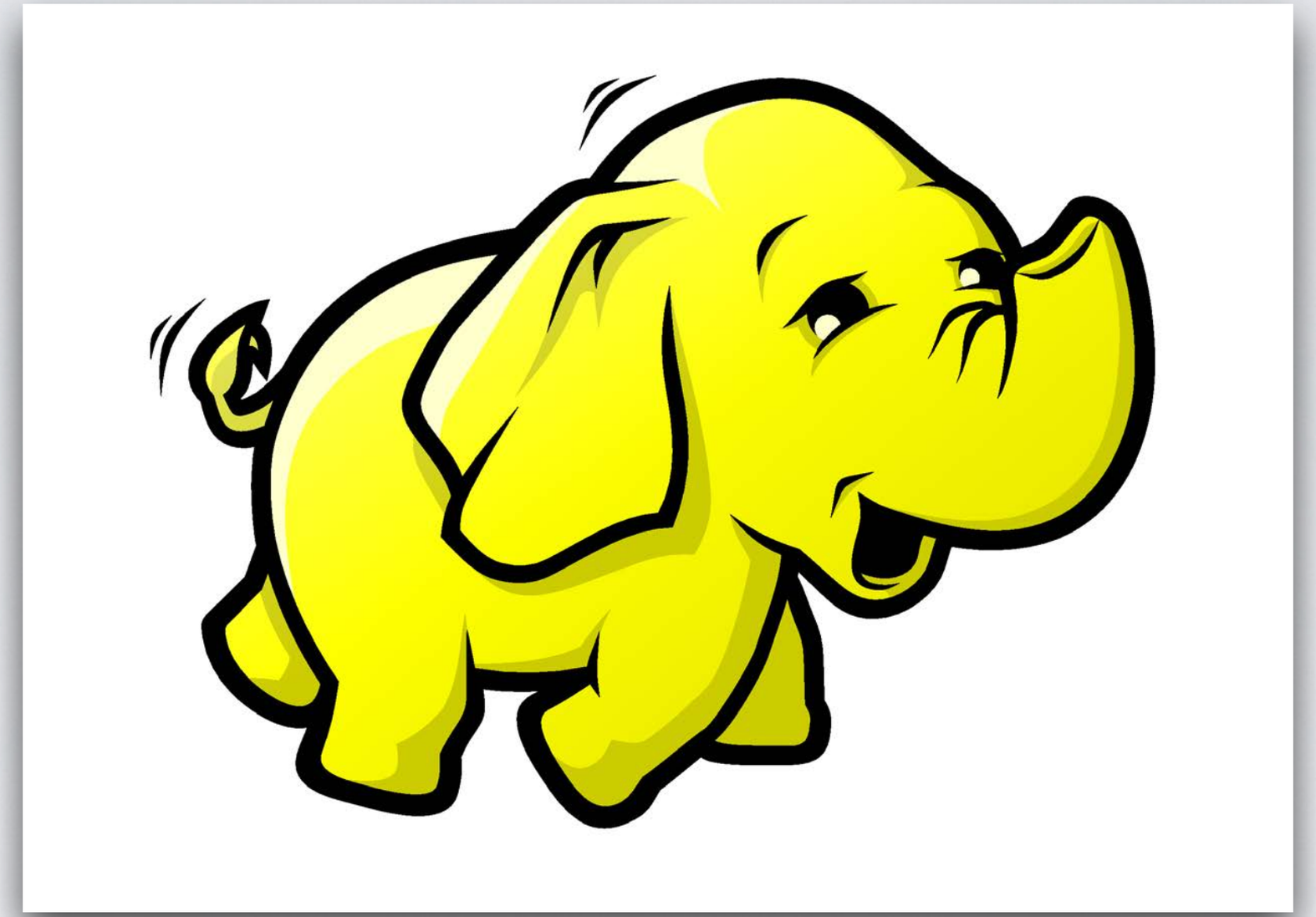
TRADITIONAL MEANS…



14 Days

850 GB Raster Files

BETTER WAY ?




amazon
web services™

2012-07-18 15:41 EDT	0 hours 2 minutes	21

Create a New Job Flow

Cancel

DEFINE JOB FLOW · SPECIFY PARAMETERS · CONFIGURE EC2 INSTANCES · ADVANCED OPTIONS · BOOTSTRAP ACTIONS · REVIEW

Please review the details of your job flow and click "Create Job Flow" when you are ready to launch your Hadoop Cluster.

Job Flow Name:	My Heat Map	
Type:	Custom Jar	Edit Job Flow Definition
Jar Location:	s3n://ags/MRHeatMap.jar	
Jar Arguments:	s3n://mraads3 s3n://mraadout/heatmap	Edit Job Flow Parameters
Master Instance Type:	m1.small	
Core Instance Type:	m1.large	**Instance Count:** 10
		Edit EC2 Configs
Amazon EC2 Key Pair:		
Amazon Subnet Id:		
Amazon S3 Log Path:	s3n://mraadlogs/mrheatmaplog	
Enable Debugging:	No	
Keep Alive:	No	
Termination Protected:	No	Edit Advanced Options
Bootstrap Actions:	No Bootstrap Actions created for this Job Flow	Edit Bootstrap Actions

‹ Back

Create Job Flow

Note: Once you click "Create Job Flow," instances will be launched and you will be charged accordingly.

Services ▾ Edit Shortcut ▾

Your Elastic MapReduce Job Flows

Region: US East (Virginia) ▾ | Create New Job Flow | Terminate | Debug

Viewing: All ▾

	Name	State	Creation Date	Elapsed Time	Normalized Instance Hours
☐	HeatMap4	COMPLETED	2012-07-19 07:19 EDT	0 hours 30 minutes	41
☐	HeatMap31	COMPLETED	2012-07-18 16:07 EDT		

Outstanding balance for this statement	$37.67

Details

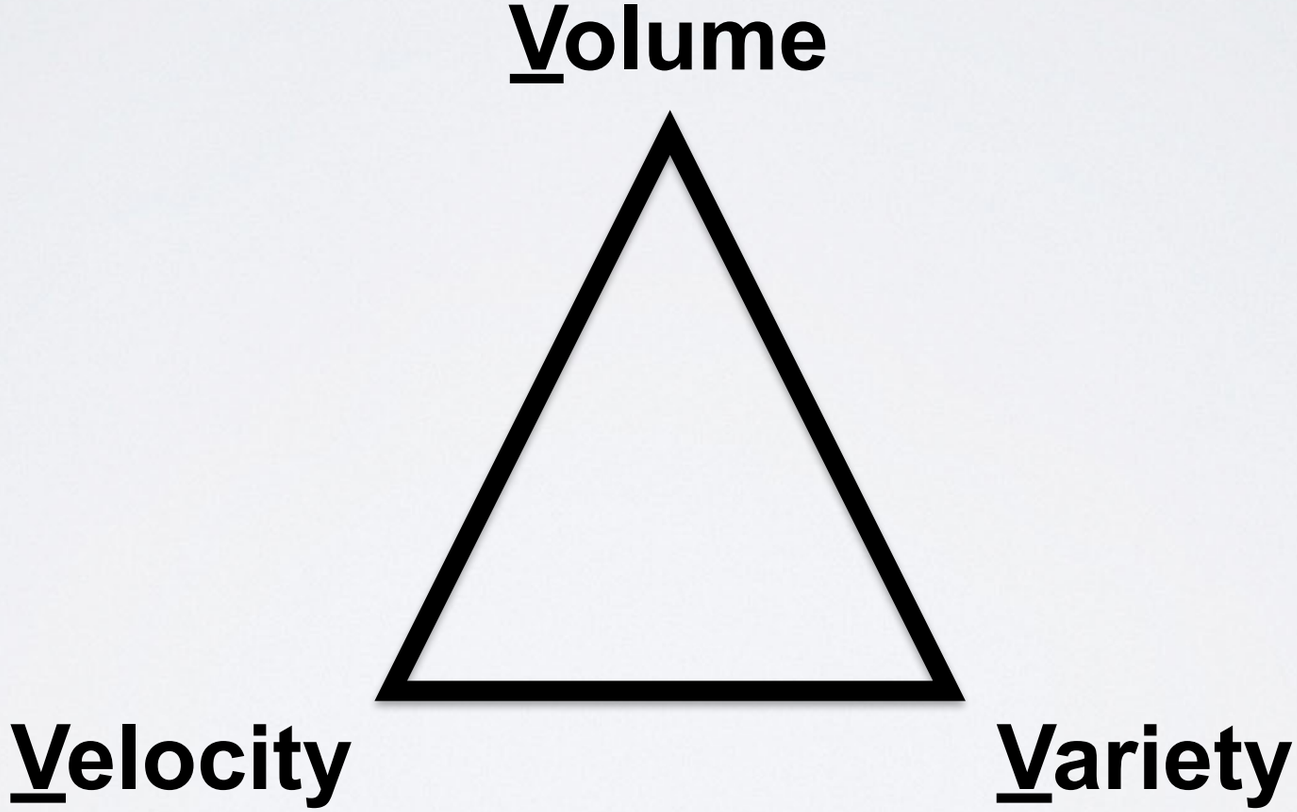
Expand All Services | Collapse All Services — Printer Friendly Version

AWS Service Charges	$37.67
Amazon Elastic Compute Cloud Download Usage Report »	$31.59
Amazon SimpleDB Download Usage Report »	$0.00
Amazon Simple Notification Service Download Usage Report »	$0.00
Amazon Simple Storage Service Download Usage Report »	$2.00
Amazon Elastic MapReduce Download Usage Report »	$4.07
AWS Data Transfer (excluding Amazon CloudFront)	$0.01
VAT to be collected	$0.00

† Usage and recurring charges for this statement period will be charged on your next billing date, August 1, 2012. Estimated charges shown on this page, or shown on any notifications that we send to you, may differ from your actual charges for this statement period. This is because estimated charges presented on this page do not include usage charges accrued during this statement period after the date you view this page. Similarly, information about estimated charges sent to you in a notification do not include usage charges accrued during this statement period after the date

BIG DATA ?

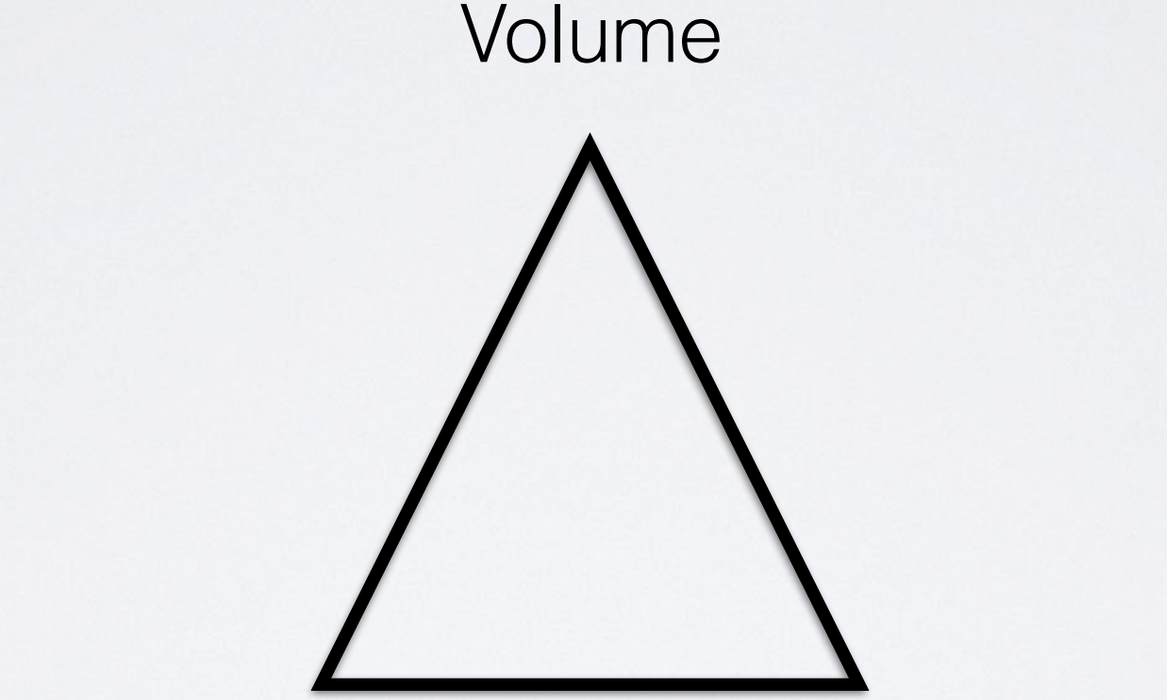
U R IN BIGDATA SPACE IF...



BUT THEN I'VE SEEN…

Volume → data at rest
Velocity → data in motion
Variety → many types
Veracity → data in doubt
Validity → data that is correct
Visualization → data in patterns
Vulnerability → data at risk
Value → data that is meaningful

I'M STICKING WITH…



NOSQL

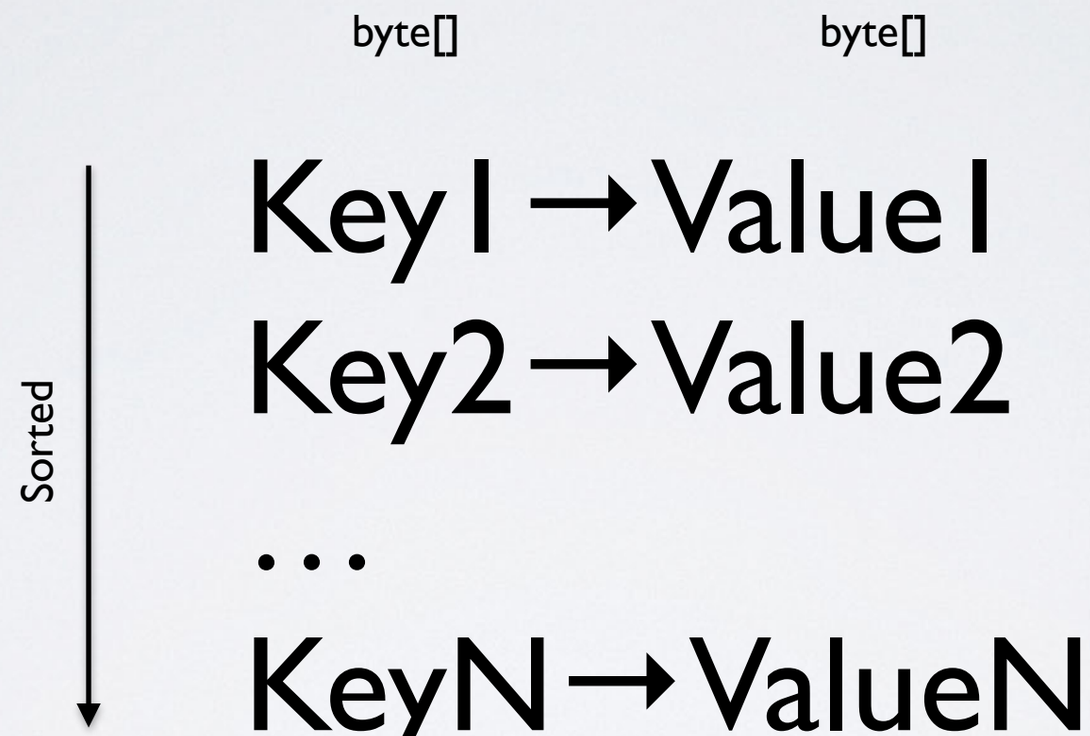
(NOT ONLY SQL :-)

GEOJSON

http://geojson.org/

```
{
  "type": "Feature",
  "geometry": {
    "type": "Point",
    "coordinates": [125.6, 10.1]
  },
  "properties": {
    "name": "Dinagat Islands"
  }
}
```

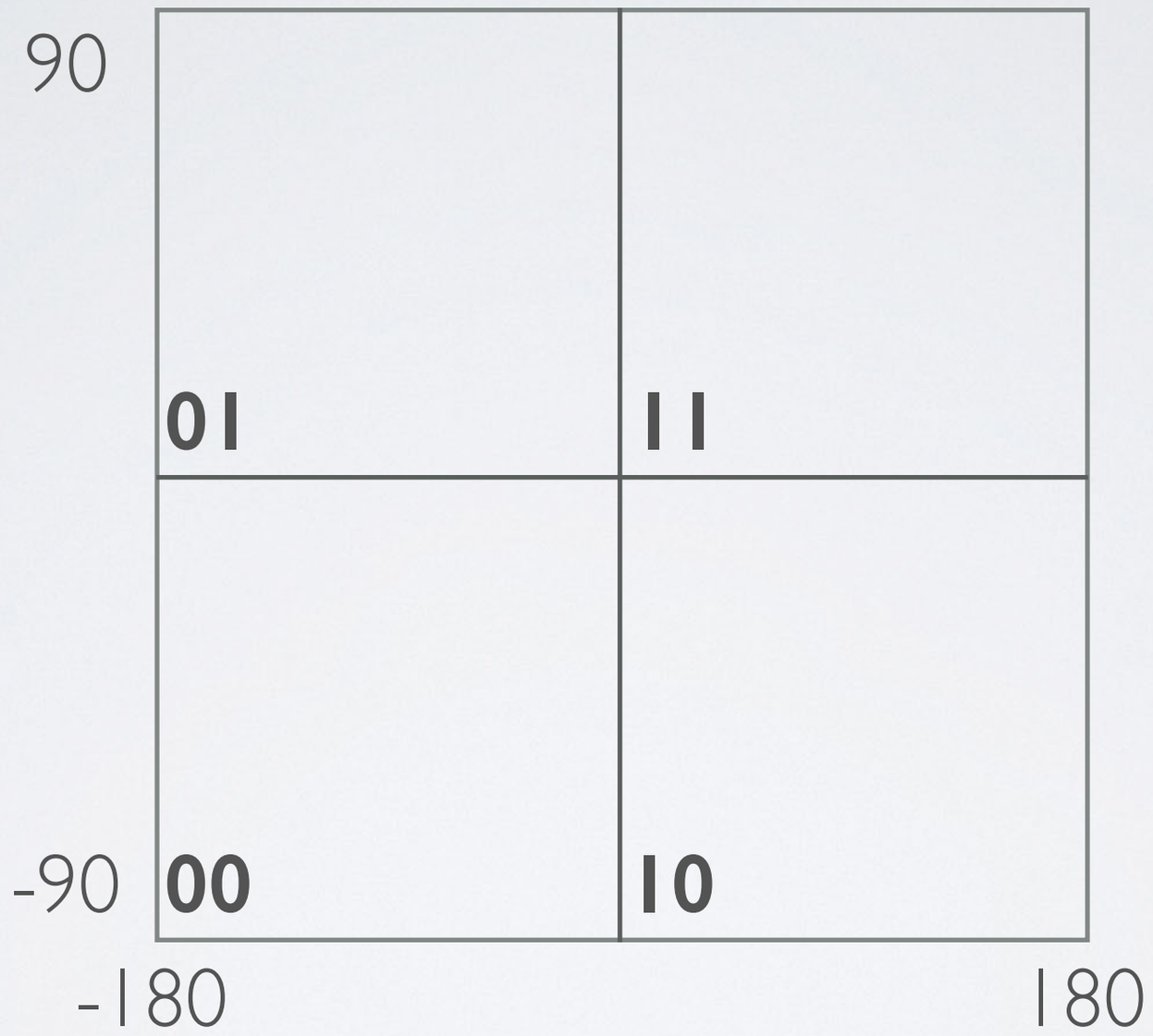
- Points
- Lines
- Polygons
- Multipoints
- Multilines
- Multipolygons
- Geometry Collection

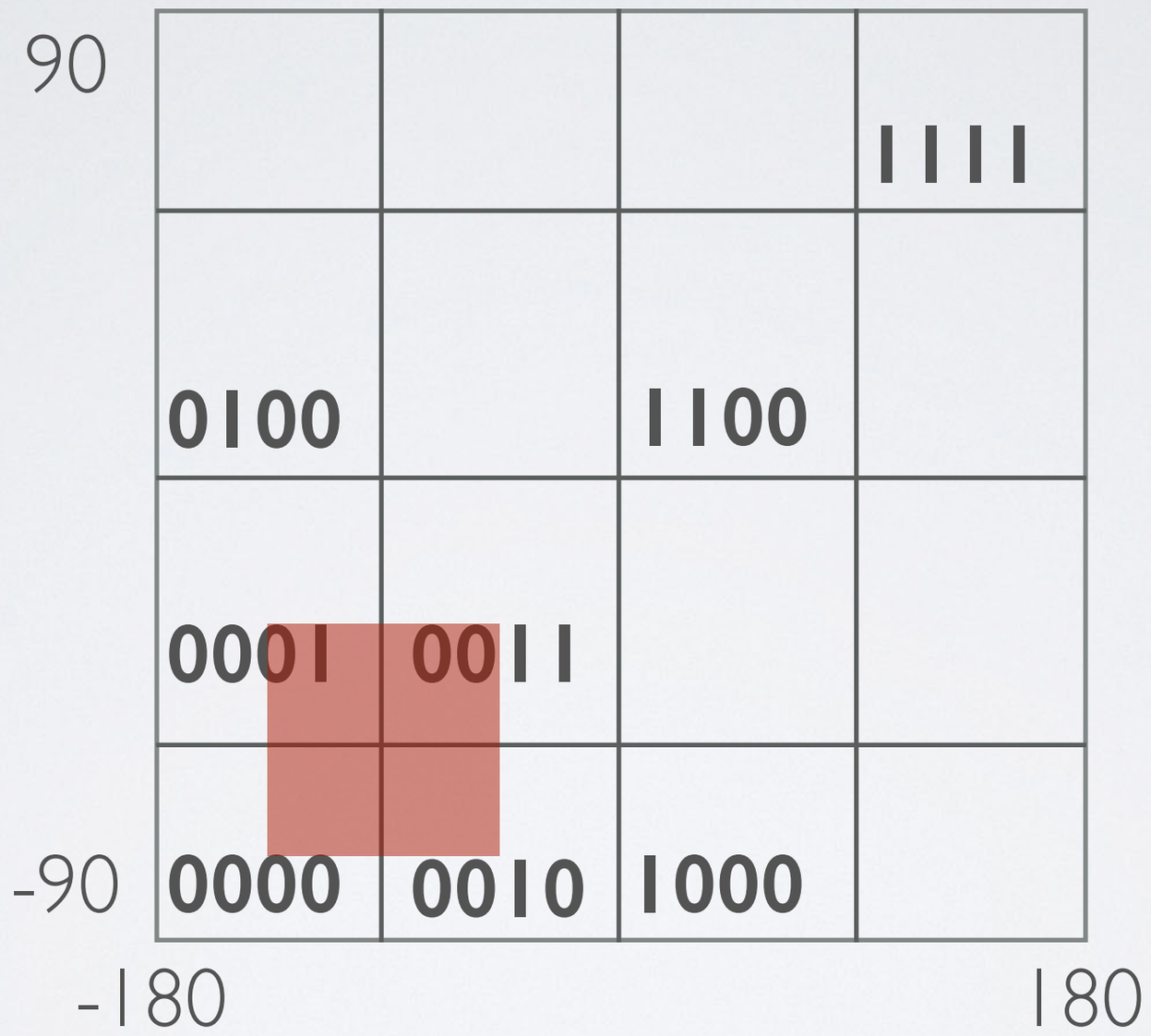

byte[]
byte[]
Key1 → Value1
Key2 → Value2
...
KeyN → ValueN
Sorted

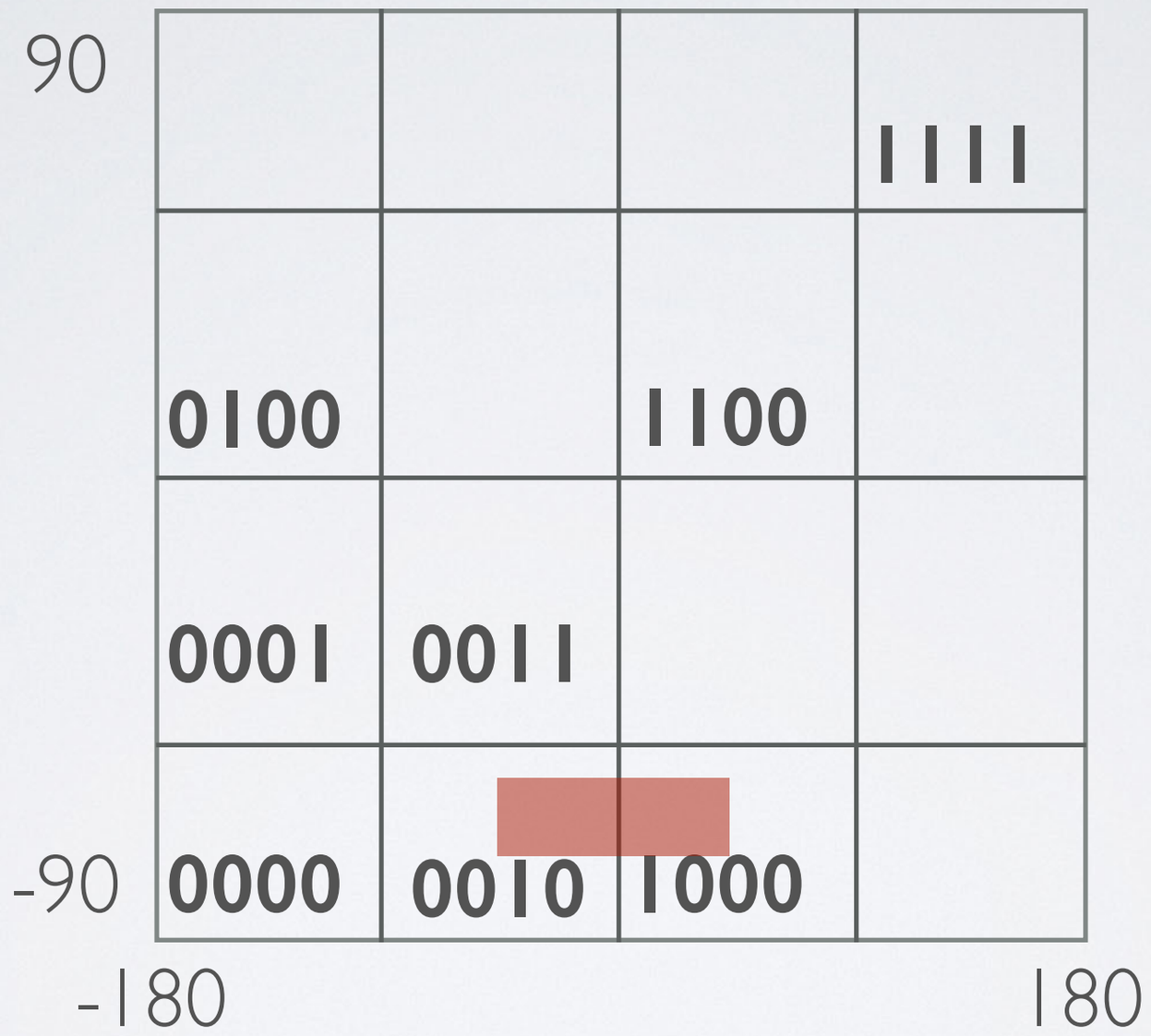
GEOHASH

http://en.wikipedia.org/wiki/Geohash

if left of vertical center set left bit to 0 else 1
if lower of horizontal center set right bit 0 else 1





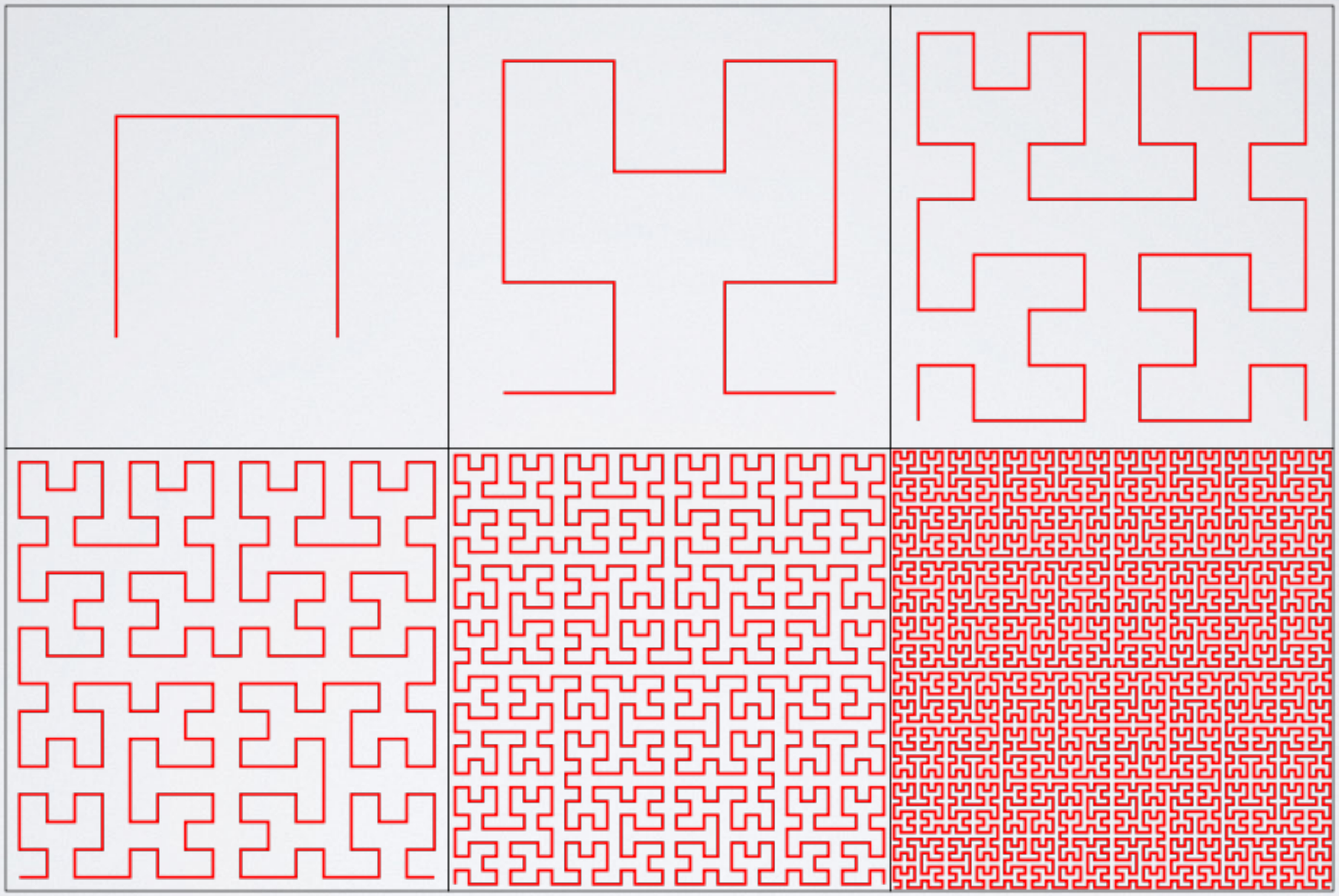


SPACE FILLING CURVES

~ 1880

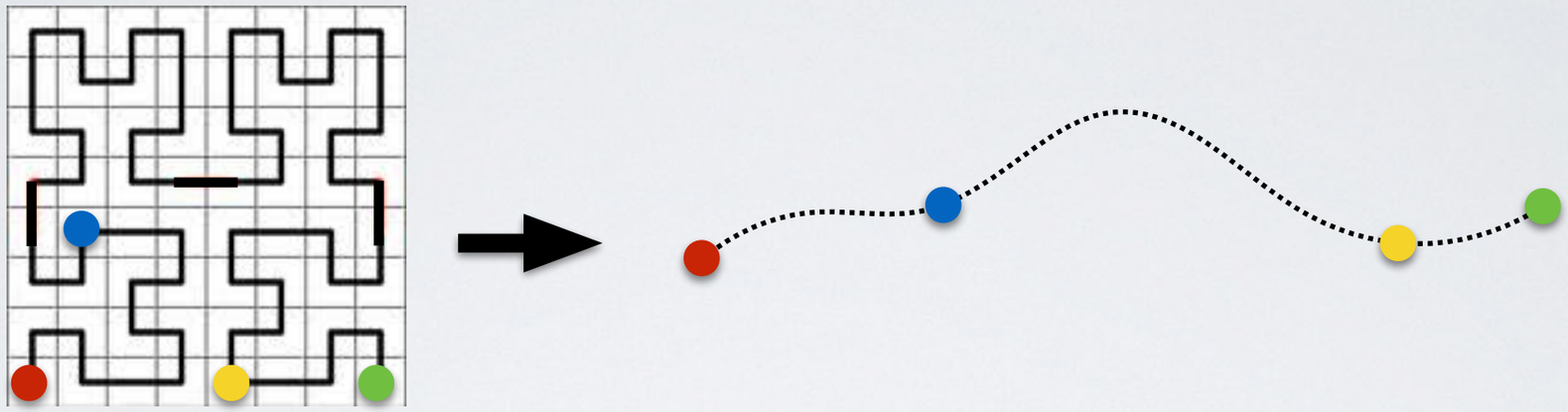
N DIM → 1 DIM

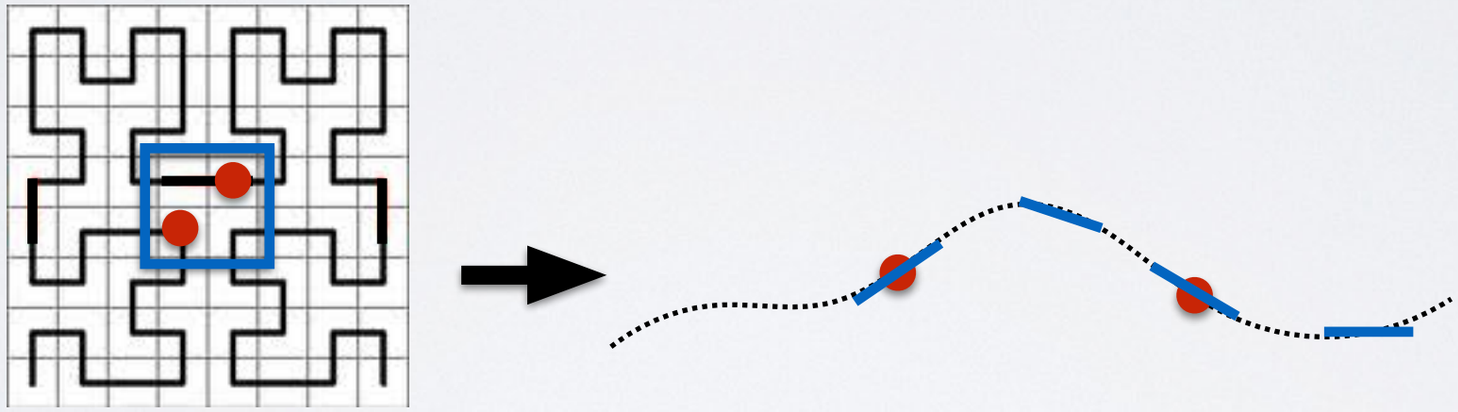
HILBERT CURVE

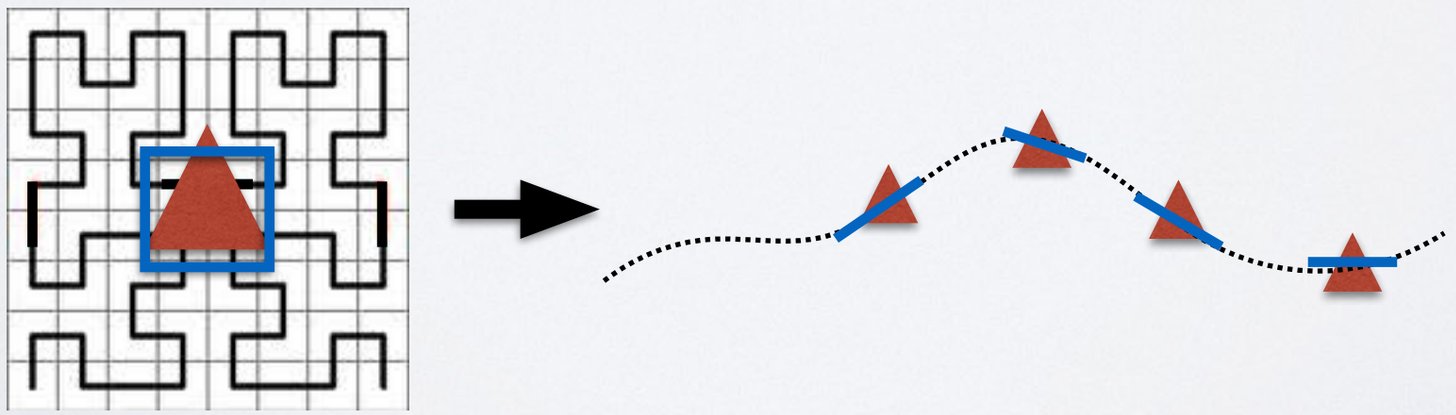


http://en.wikipedia.org/wiki/Space-filling_curve

SPACE LINEARIZATION







SPATIAL SUPPORT

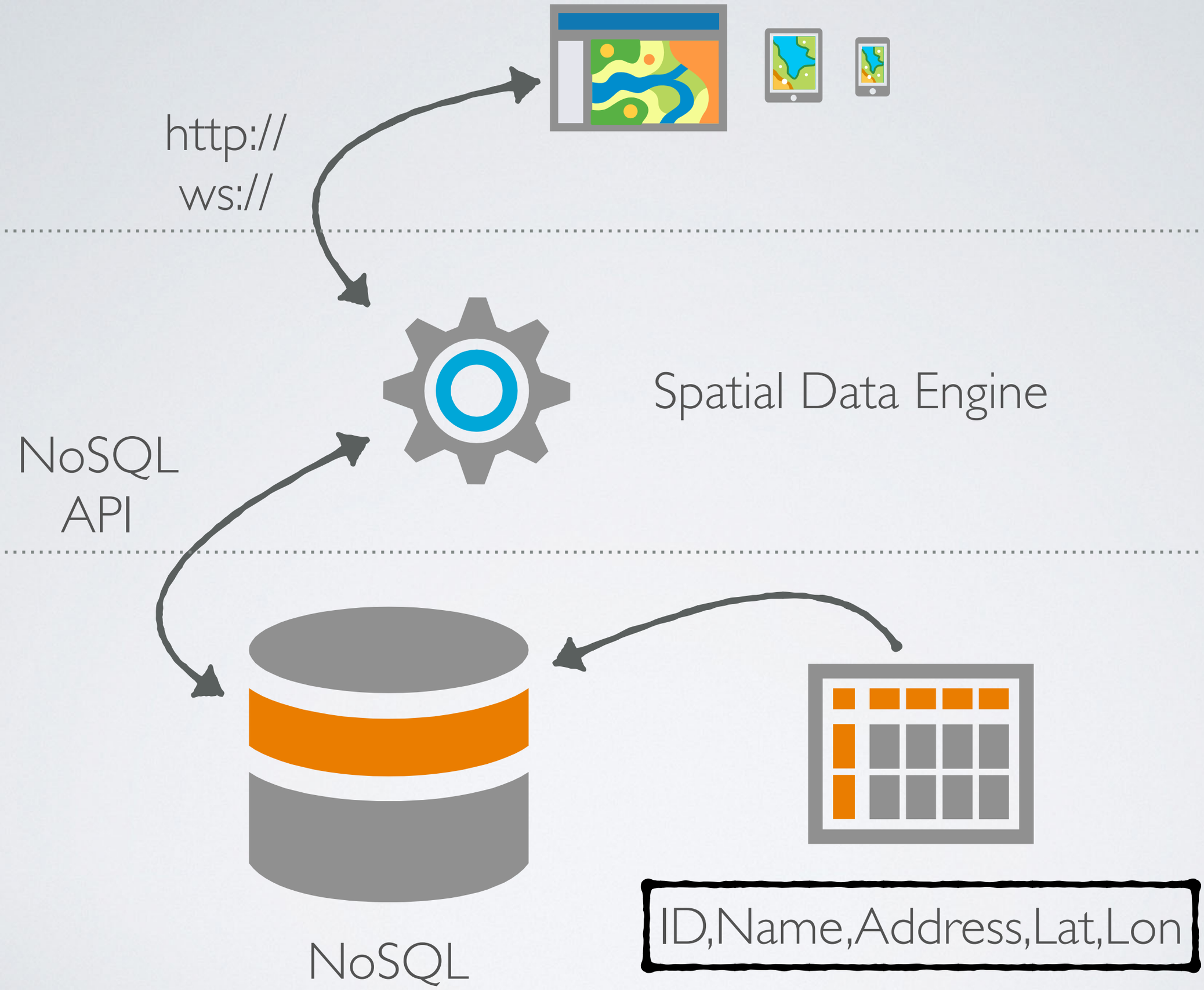
RTree

INDIRECT SUPPORT

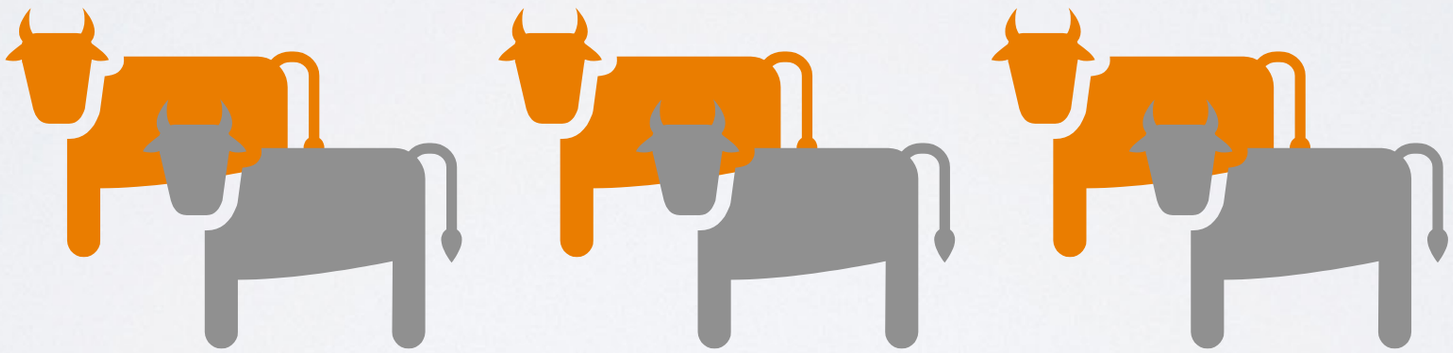
WHAT IS OLD…

IS NEW AGAIN !

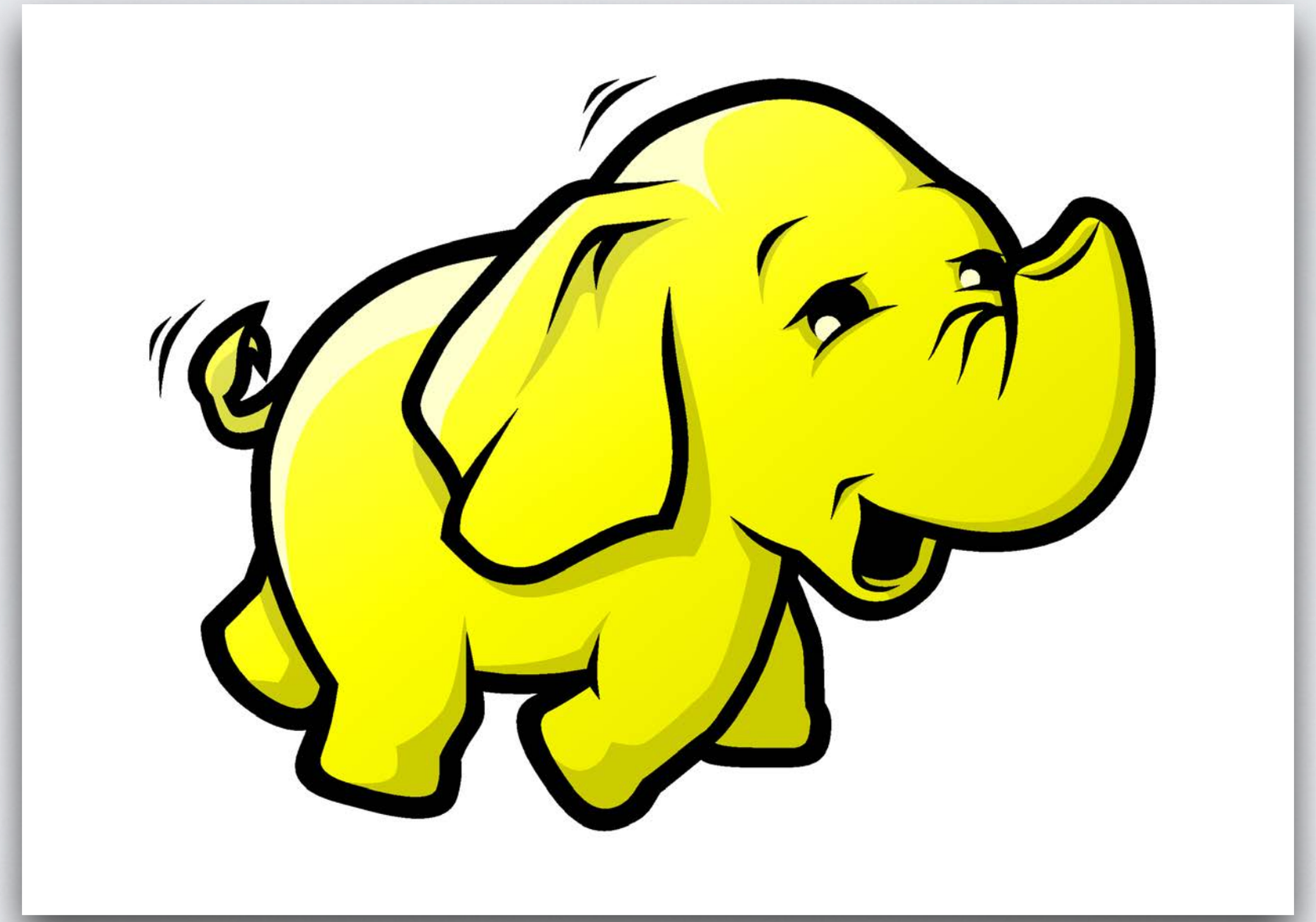
SPATIAL MIDDLEWARE


http://
ws://
Spatial Data Engine
NoSQL
API
NoSQL
ID,Name,Address,Lat,Lon

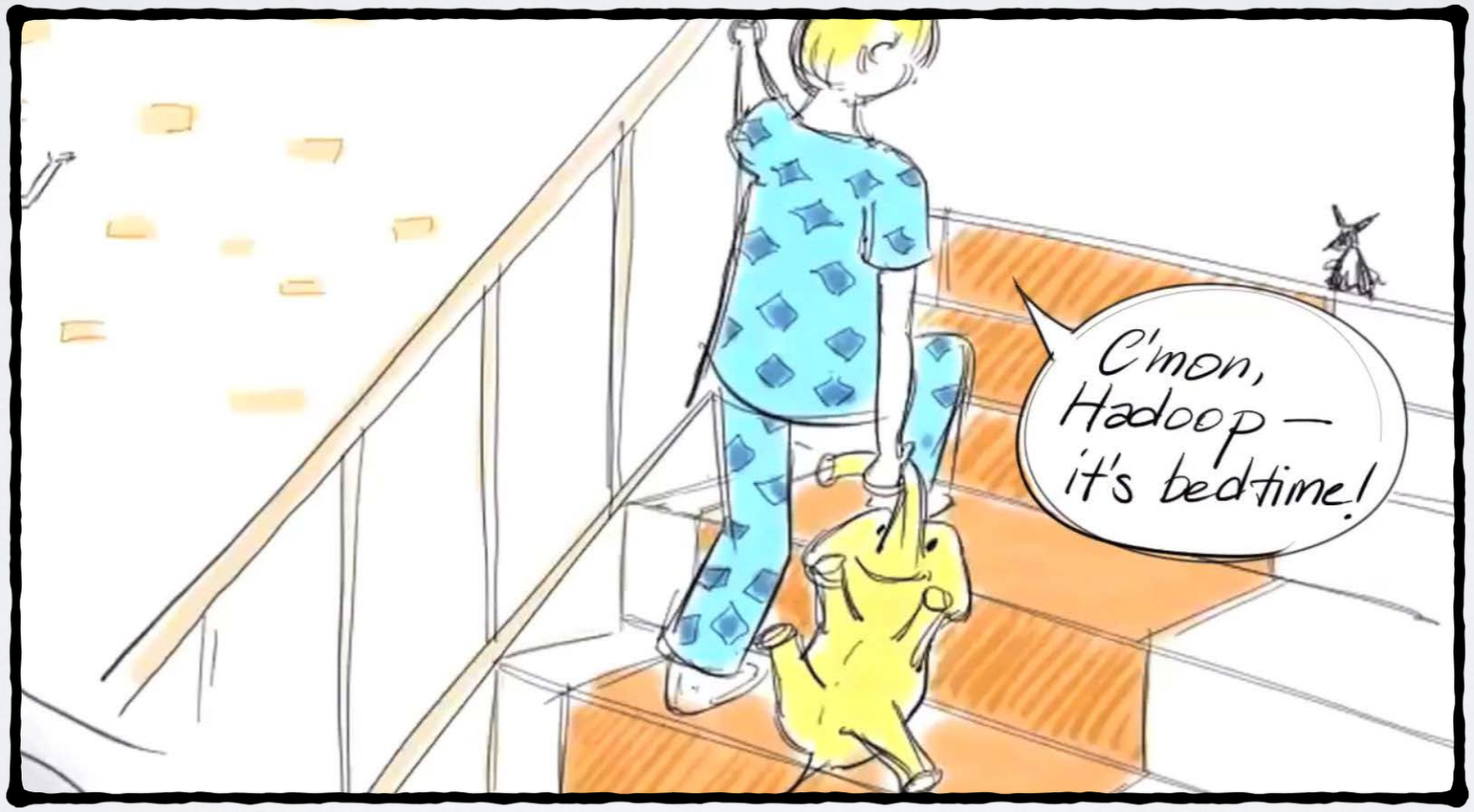
NOT A BIGGER OX…



HADOOP.APACHE.ORG



WHAT'S IN A NAME ?



http://blog.pivotal.io/pivotal/products/demystifying-hadoop-in-5-pictures

WHAT IS HADOOP ?

- Library / Framework
- Very Very Large Un/Structured Dataset
- Multi Node Distributed Processing
- Resilient To Commodity Hardware Failure

HADOOP BASIC STACK

MapReduce

Yet Another Resource Negotiator (YARN)

Hadoop Distributed File System (HDFS)

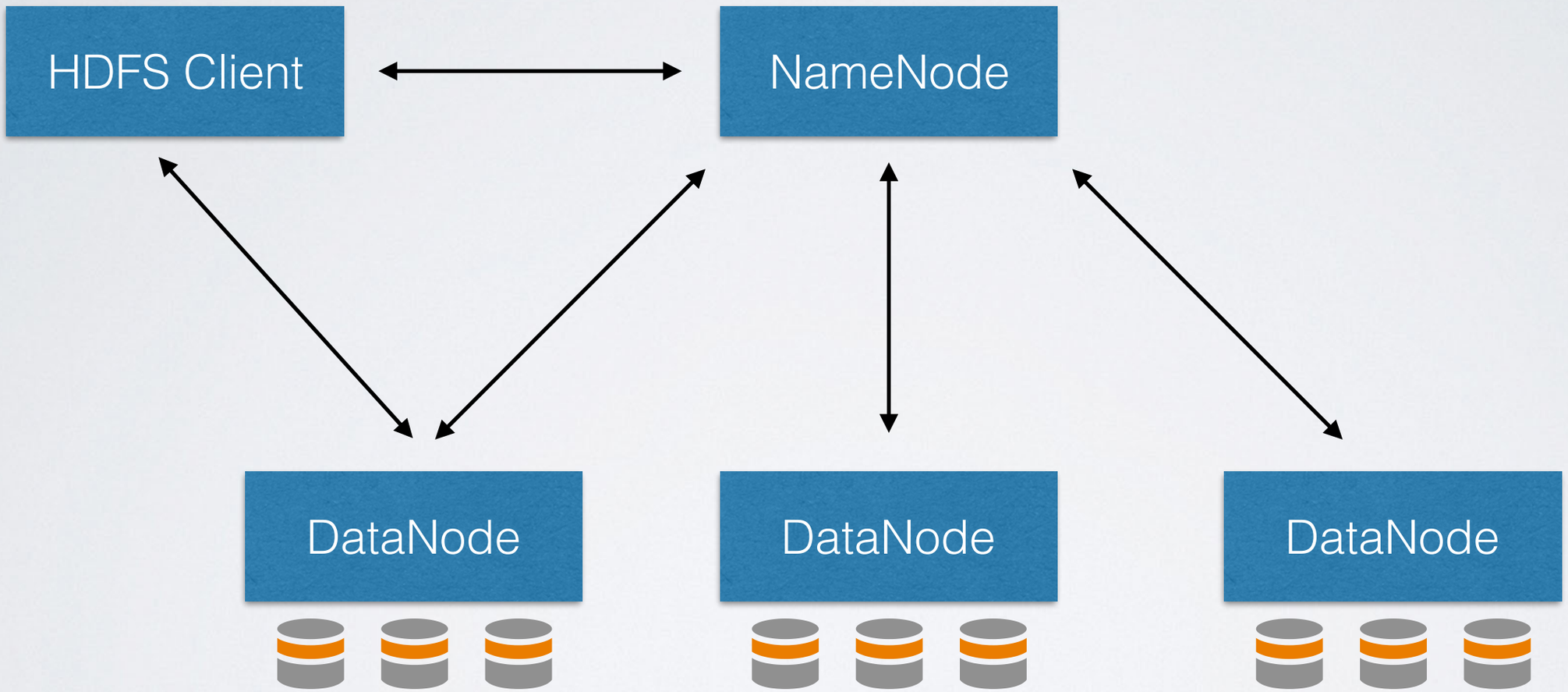
OTHER HADOOP PROJECTS

- Avro - Serialization / RPC System
- HBase - Distributed Columnar Database
- Hive - Ad Hoc "SQL" Interface
- Pig - Data Flow Parallel Execution (AML)
- ZooKeeper - Coordination Service
- More….

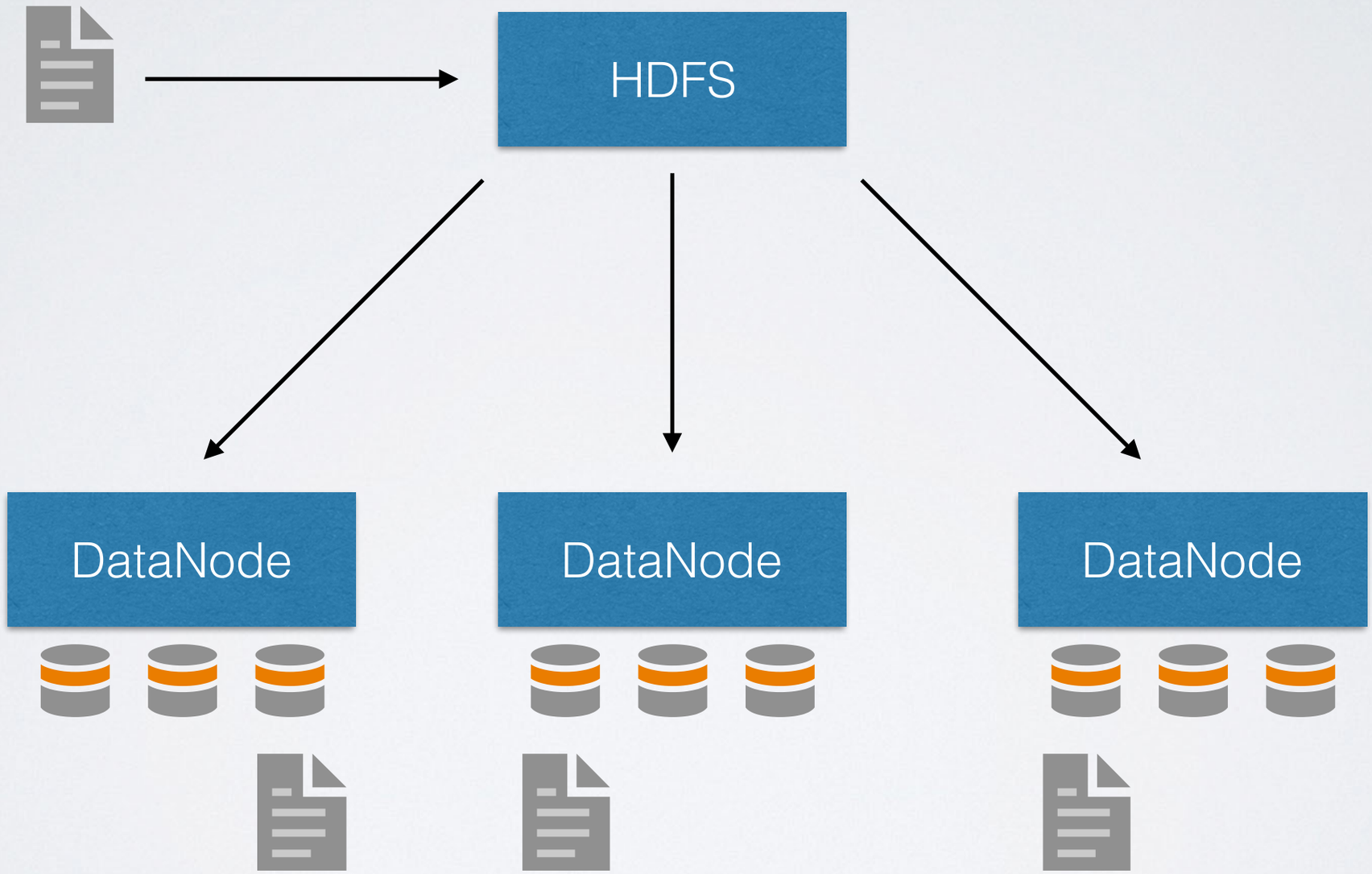
HDFS

- Distributed File System
- Lots and Lots of Commodity Drives
- Fault Tolerant
- Loves Big Files
- “POSIX” Like Interface

HDFS



HDFS Resilience !



Program

BigData

Program

BigData

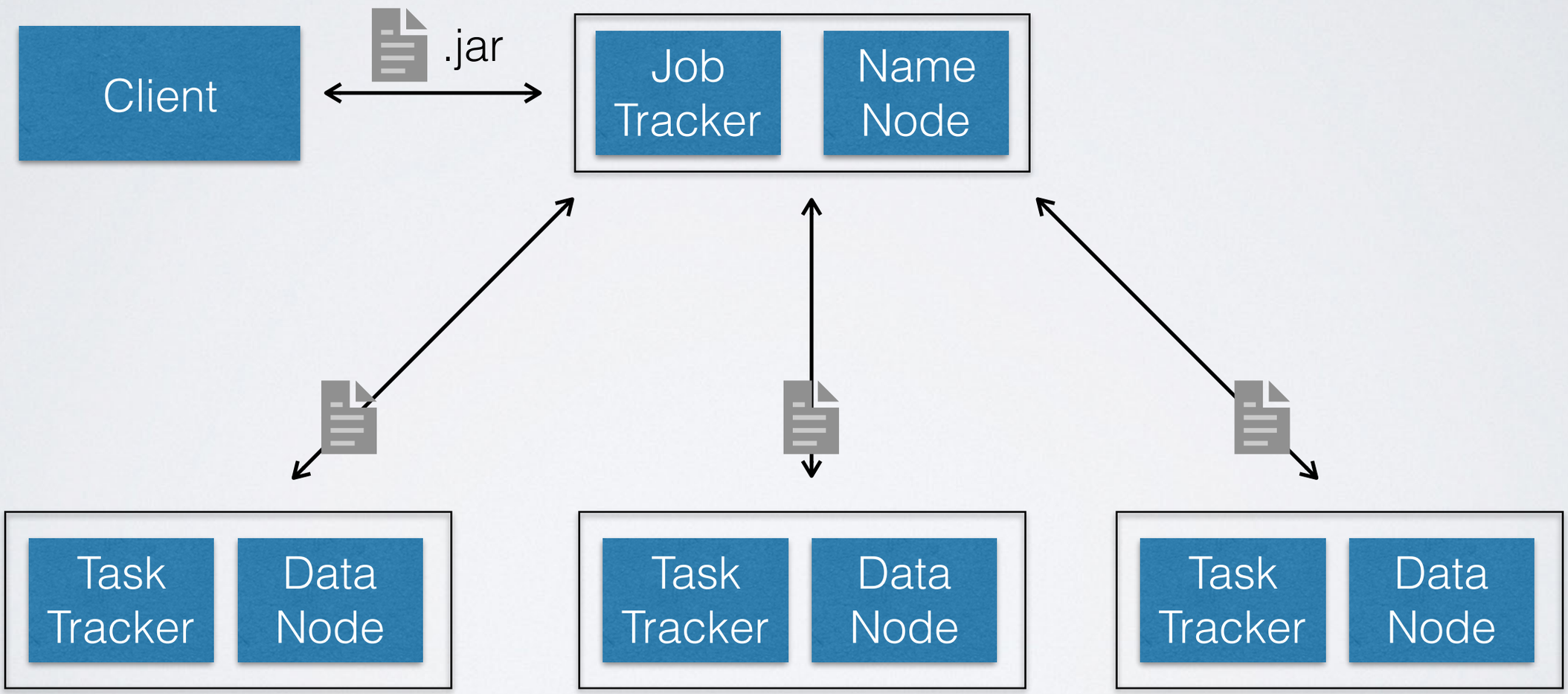
MAPREDUCE

http://hadoop.apache.org/docs/r1.2.1/mapred_tutorial.html

WHAT IS MAPREDUCE ?

- Parallel Fault Tolerant Framework
- Splits Large Input
- Invoke User Defined “Map” Function
- Shuffle and Sort
- Invoke User Defined “Reduce” Function

MAPREDUCE & HDFS



WRITING MR IS HARD…

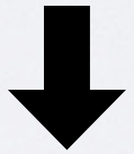
HOW ABOUT....

NO PROGRAMING ???


HIVE

APACHE HIVE

"SQL"



MapReduce Job

HQL

drop table if exists *logs*;

create external table if not exists *logs*(
ip **string**,
method **string**,
uri **string**,
status **string**,
bytes **int**,
time_taken **int**,
referrer **string**,
user_agent **string**
) **partitioned by** (*year* **int**, *month* **int**, *day* **int**, *hour* **int**)
row format delimited
fields terminated by '\t'
lines terminated by '\n'
stored as *textfile*
location ‘hdfs://hadoop:8020/logs/';

OTHER ADHOC ENGINES

- Cloudera Impala
- Facebook Presto
- SparkSQL
- Bypass MR generation / Direct HDFS Access

WHAT ABOUT SPATIAL ?

GIS Tools for Hadoop by Esri

GIS Tools for Hadoop by Esri

GIS Tools for Hadoop

Big Data Spatial Analytics for the Hadoop Framework

View project on GitHub

Looking at data without location, most of the time seems like looking at just part of a story. Including location and geography in analysis reveals patterns and associations that otherwise are missed. As Big Data emerges as a new frontier for analysis, including location in Big Data is becoming significantly important.

Data that includes location, and that is enhanced with geographic information in a structured form, is often referred to as Spatial Data. Doing Analysis on Spatial data requires an understanding of geometry and operations that can be preformed on it. Enabling Hadoop to include spatial data and spatial analysis is the goal of this Esri Open Source effort.

GIS Tools for Hadoop is an open source toolkit intended for Big Spatial Data Analytics. The toolkit provides different libraries:

- Esri Geometry API for Java: A generic geometry library, can be used to extend Hadoop core with vector geometry types and operations, and enables developers to build MapReduce applications for spatial data.
- Spatial Framework for Hadoop: Extends Hive and is based on the

is maintained by Esri.

This page was generated by GitHub Pages using the Architect theme by Jason Long.

Display a menu

GIS TOOLS FOR HADOOP

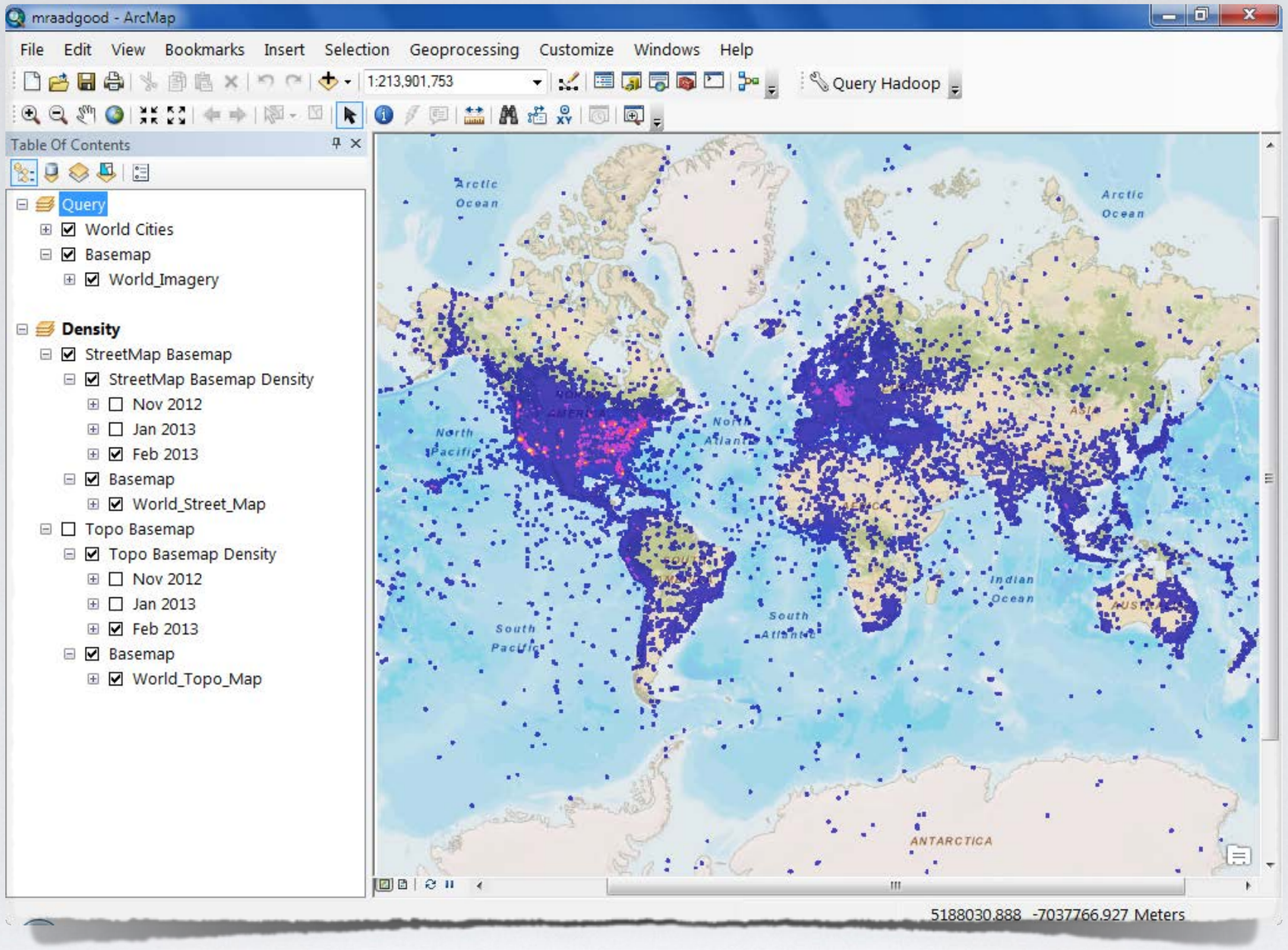
- Computational Geometry Library
- Hive Spatial UDF Functions
- GeoProcessing Extensions to ArcMap

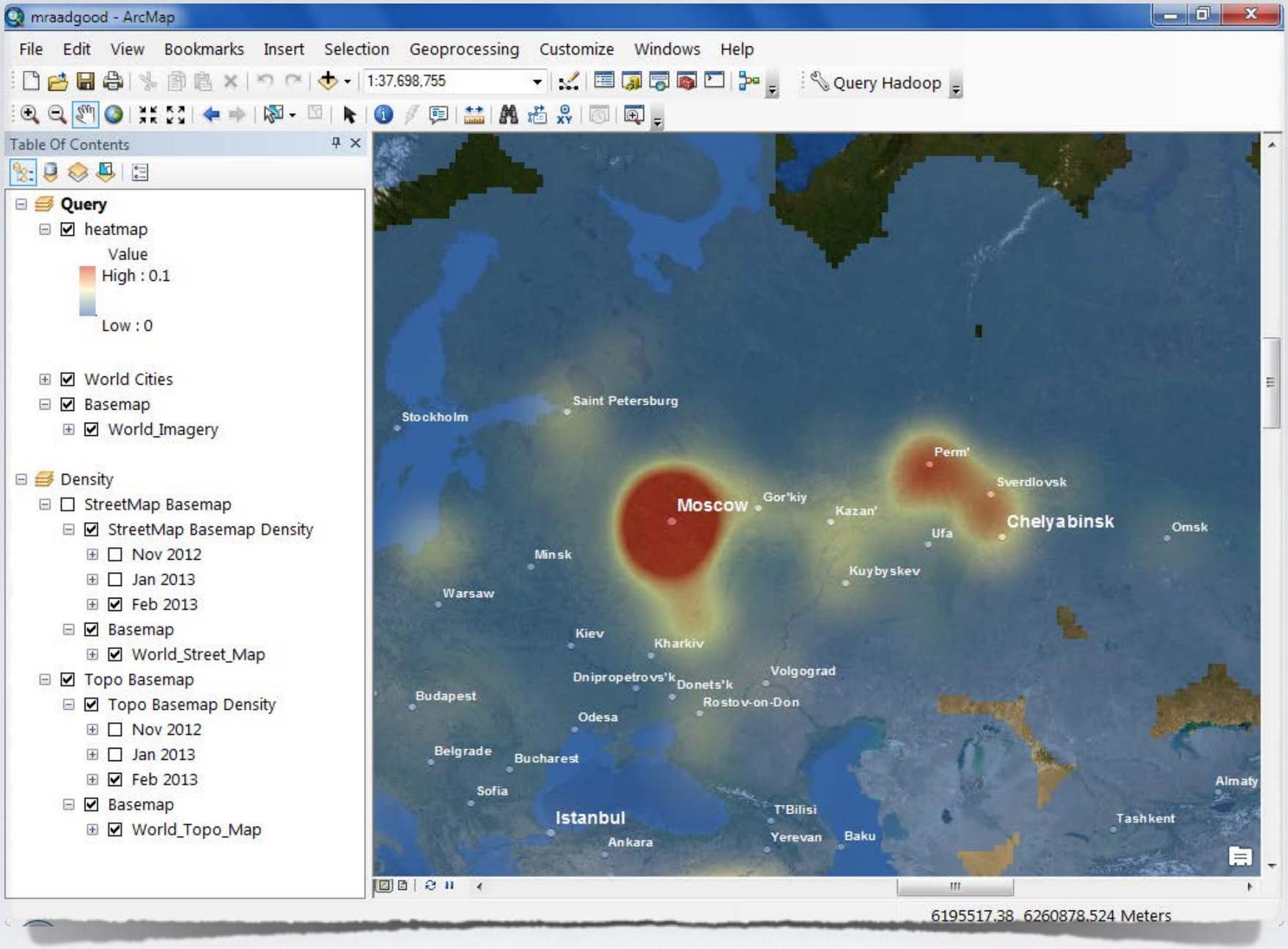
GEOMETRY LIBRARY

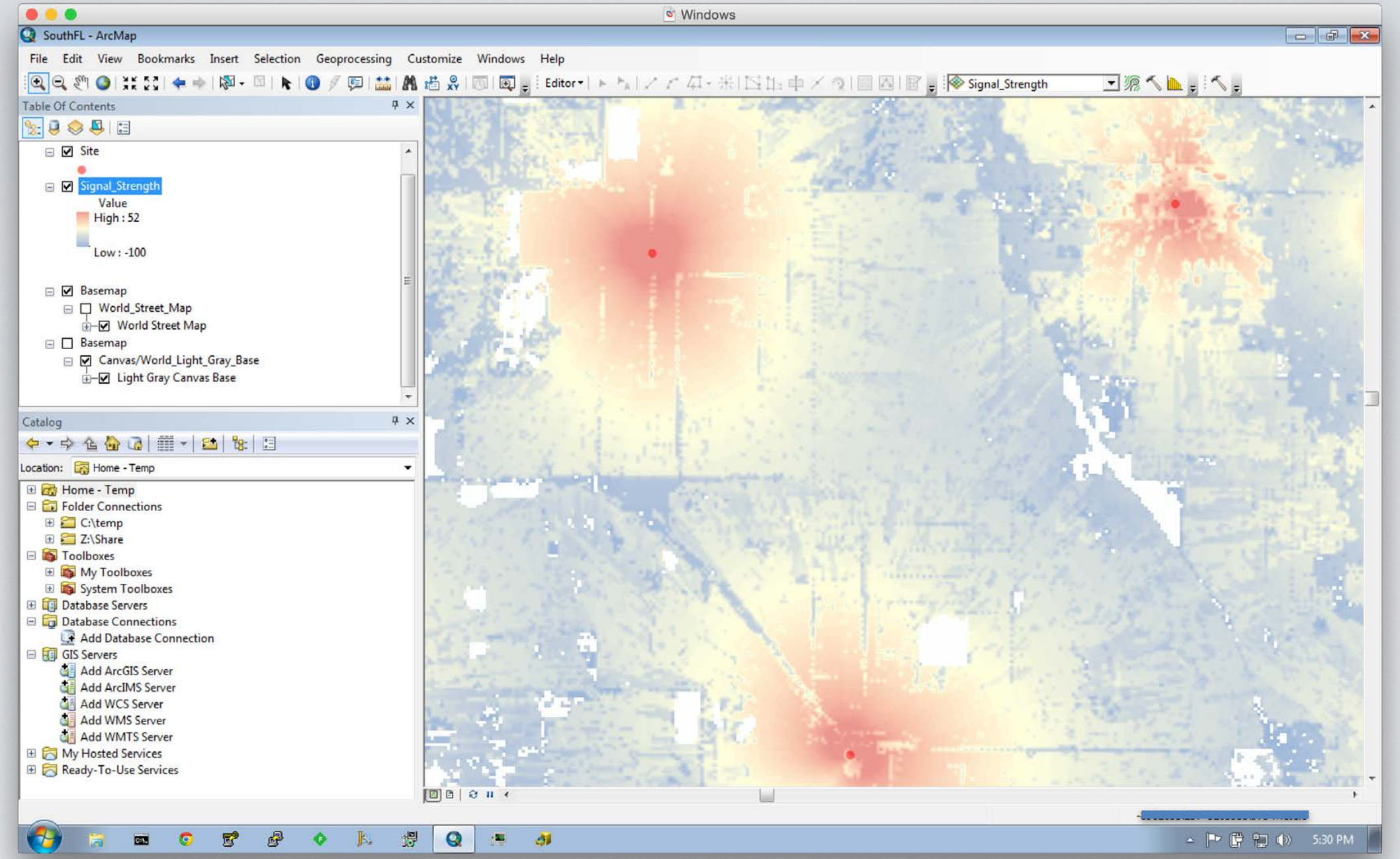
- Points / Lines / Polygons
- I/O (GeoJSON,WTK,WBT,Shape)
- Spatial Relations (inside, touches, intersects,…)
- Spatial Operations (buffer, cut, convex hull,…)
- In-Memory Spatial Index

API USAGE IN BIGDATA

- Map-only jobs - GeoEnrichment
  - Given set of locations
  - Given demographic area
  - Augment location with demographic attributes


mraadgood - ArcMap
File Edit View Bookmarks Insert Selection Geoprocessing Customize Windows Help
1:213,901,753
Query Hadoop
Table Of Contents
Query
World Cities
Basemap
World_Imagery
Density
StreetMap Basemap
StreetMap Basemap Density
Nov 2012
Jan 2013
Feb 2013
Basemap
World_Street_Map
Topo Basemap
Topo Basemap Density
Nov 2012
Jan 2013
Feb 2013
Basemap
World_Topo_Map
Arctic Ocean
Arctic Ocean
North Pacific
North Atlantic
Indian Ocean
South Pacific
South Atlantic
ANTARCTICA
5188030.888 -7037766.927 Meters


mraadgood - ArcMap
File Edit View Bookmarks Insert Selection Geoprocessing Customize Windows Help
1:37,698,755
Query Hadoop
Table Of Contents
Query
heatmap
Value
High : 0.1
Low : 0
World Cities
Basemap
World_Imagery
Density
StreetMap Basemap
StreetMap Basemap Density
Nov 2012
Jan 2013
Feb 2013
Basemap
World_Street_Map
Topo Basemap
Topo Basemap Density
Nov 2012
Jan 2013
Feb 2013
Basemap
World_Topo_Map
Stockholm
Saint Petersburg
Perm'
Sverdlovsk
Moscow
Gor'kiy
Kazan'
Ufa
Chelyabinsk
Omsk
Minsk
Kuybyskev
Warsaw
Kiev
Kharkiv
Volgograd
Dnipropetrovs'k
Donets'k
Budapest
Rostov-on-Don
Odesa
Belgrade
Bucharest
Almaty
Sofia
Istanbul
T'Bilisi
Tashkent
Yerevan
Baku
Ankara
6195517.38 6260878.524 Meters


Windows
SouthFL - ArcMap
File Edit View Bookmarks Insert Selection Geoprocessing Customize Windows Help
Editor
Signal_Strength
Table Of Contents
Site
Signal_Strength
Value
High : 52
Low : -100
Basemap
World_Street_Map
World Street Map
Basemap
Canvas/World_Light_Gray_Base
Light Gray Canvas Base
Catalog
Location: Home - Temp
Home - Temp
Folder Connections
C:\temp
Z:\Share
Toolboxes
My Toolboxes
System Toolboxes
Database Servers
Database Connections
Add Database Connection
GIS Servers
Add ArcGIS Server
Add ArcIMS Server
Add WCS Server
Add WMS Server
Add WMTS Server
My Hosted Services
Ready-To-Use Services
5:30 PM

TELCO

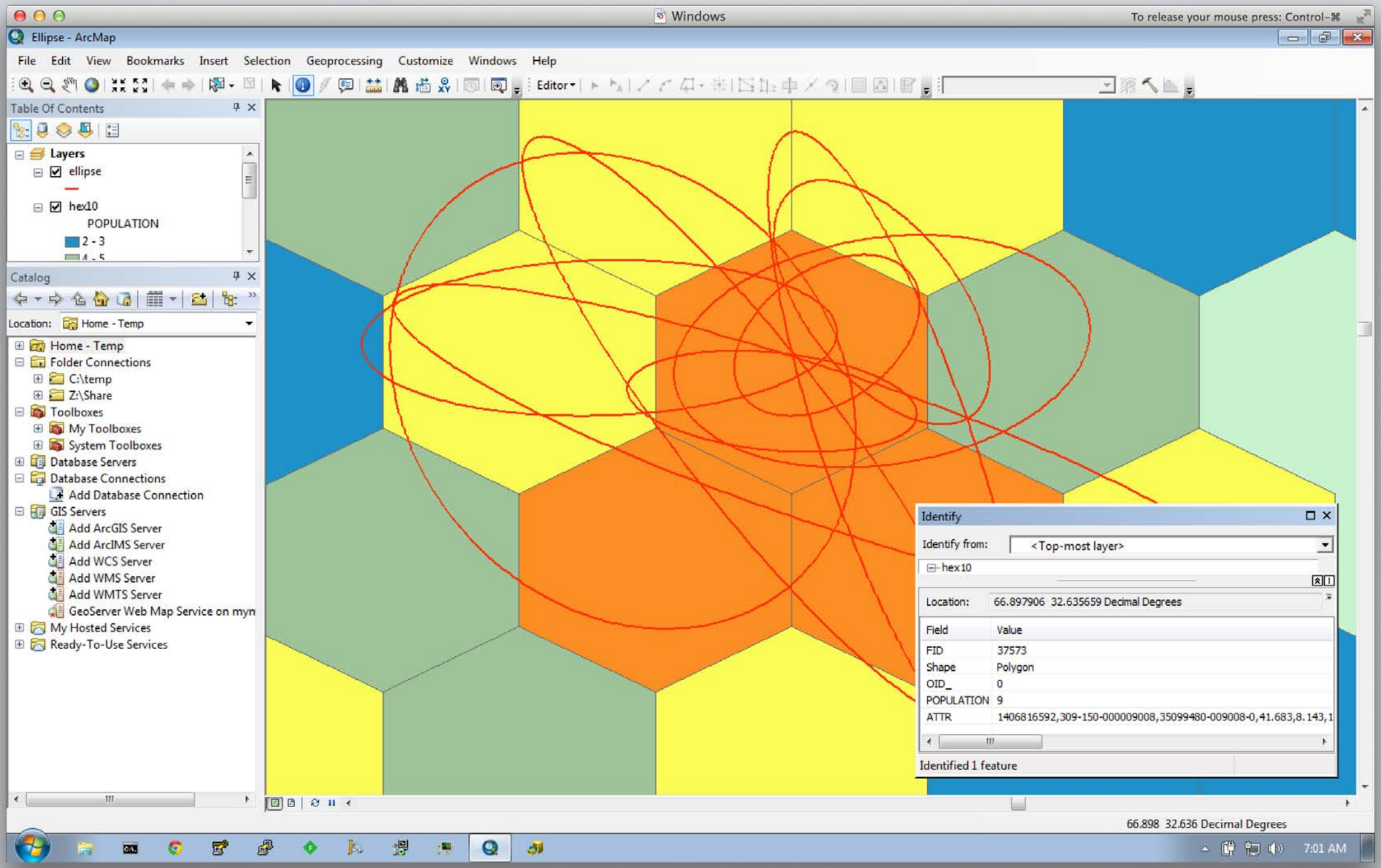
- CDR - Call Data Record
    - DateTime, UUID, LatLon, Duration, Status, etc…
- Drop Call Emerging HotSpot
- Street Traffic Condition
- Massive Spatial Join million x million polygon
- Overlay with Demographic polygons
- Overlay with Current Weather
- Overlay with Social Media

TELEMATICS

- Feedback to Engineering
- Car to Car Communication
- Street Condition Detection
- Best Route Prediction (EV)
- Overlay with weather

Insurance as you drive !

Observations


Windows
To release your mouse press: Control-⌘
Ellipse - ArcMap
File Edit View Bookmarks Insert Selection Geoprocessing Customize Windows Help
Editor
Table Of Contents
Layers
ellipse
hex10
POPULATION
2 - 3
Catalog
Location: Home - Temp
Home - Temp
Folder Connections
C:\temp
Z:\Share
Toolboxes
My Toolboxes
System Toolboxes
Database Servers
Database Connections
Add Database Connection
GIS Servers
Add ArcGIS Server
Add ArcIMS Server
Add WCS Server
Add WMS Server
Add WMTS Server
GeoServer Web Map Service on myn
My Hosted Services
Ready-To-Use Services
Identify
Identify from: <Top-most layer>
hex10
Location: 66.897906 32.635659 Decimal Degrees
Field Value
FID 37573
Shape Polygon
OID_ 0
POPULATION 9
ATTR 1406816592,309-150-000009008,35099480-009008-0,41.683,8.143,1
Identified 1 feature
66.898 32.636 Decimal Degrees
7:01 AM

Title

A Hadoop-enabled Ship Tracking Application for the Port of Rotterdam

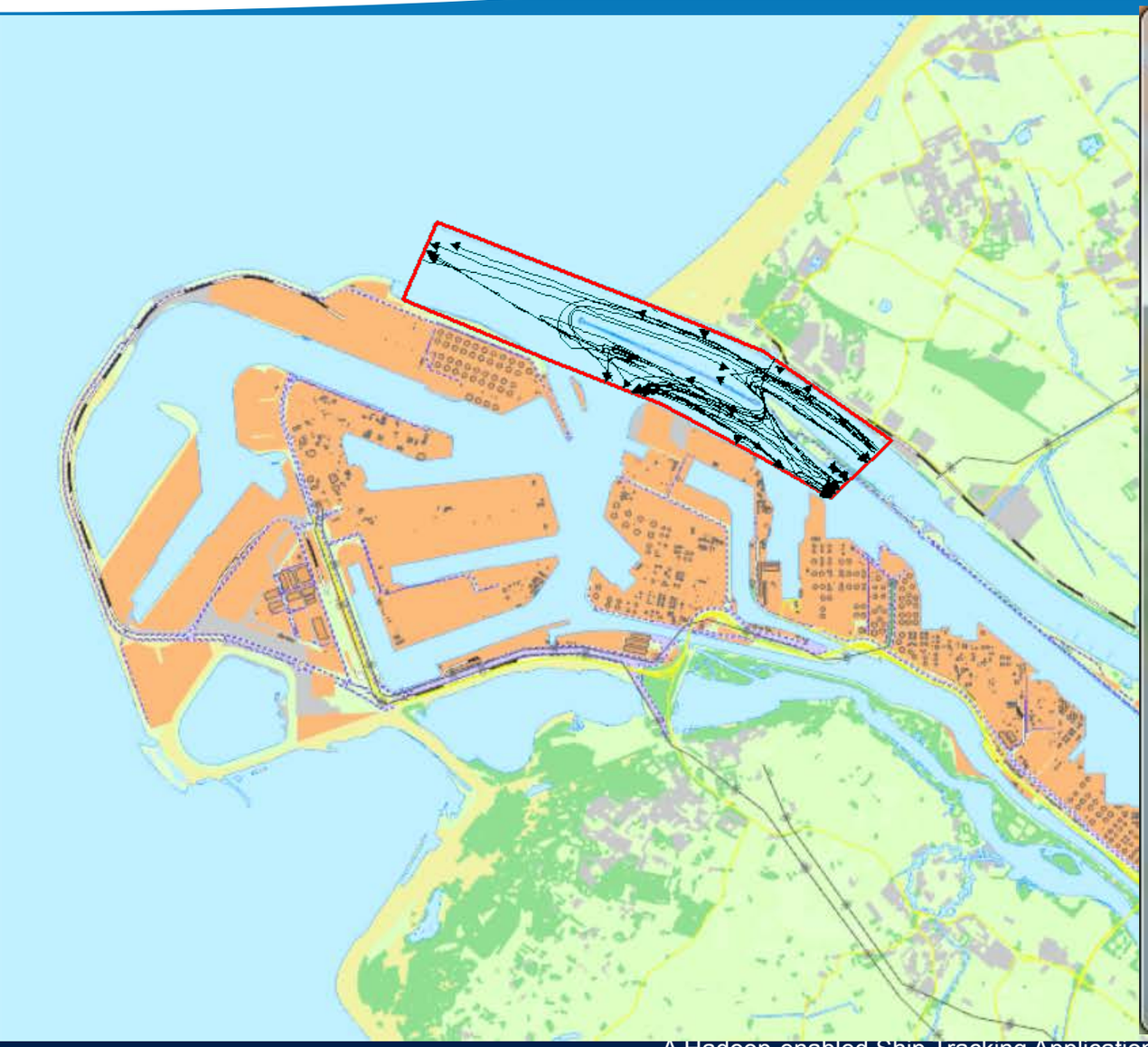
Hadoop Summit, Brussels, 15 April 2015

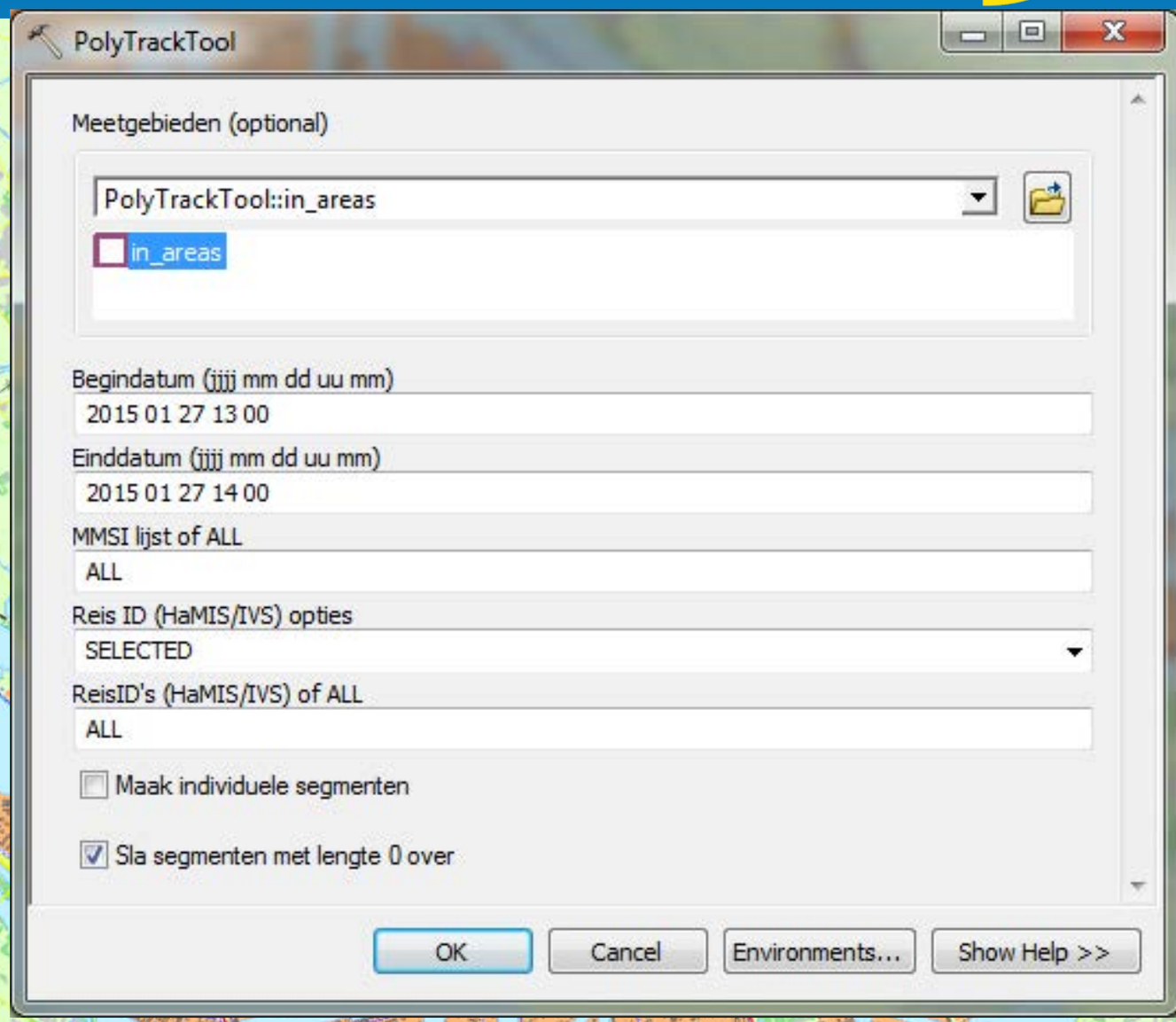
Frank Cremer (Geomatik)

Mansour Raad (ESRI)

Image

Access information in three clicks



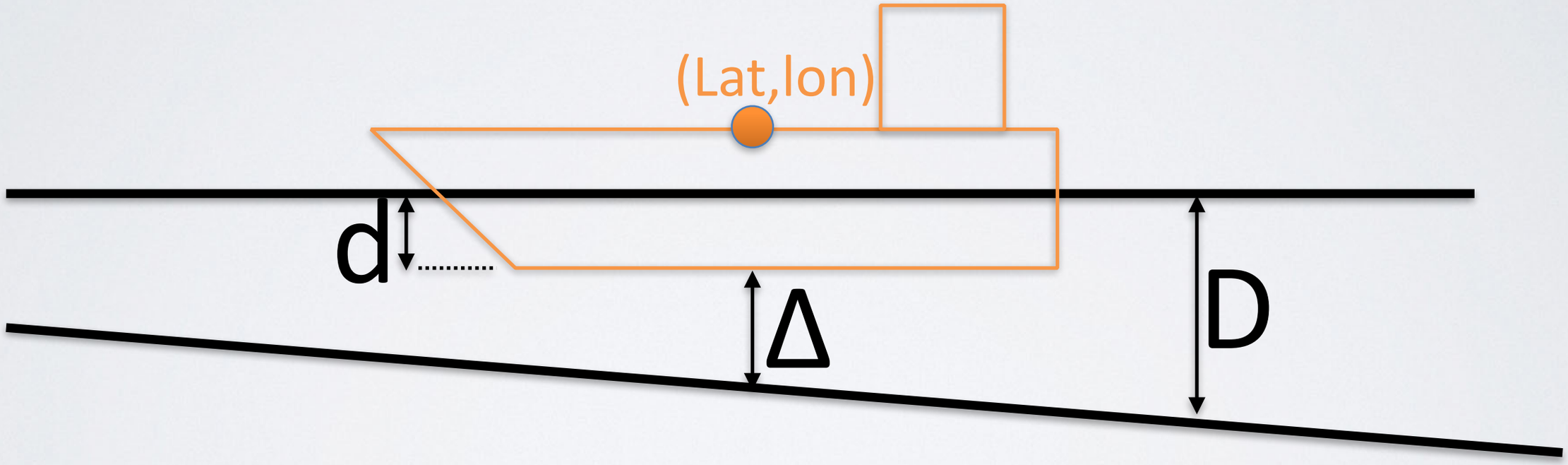


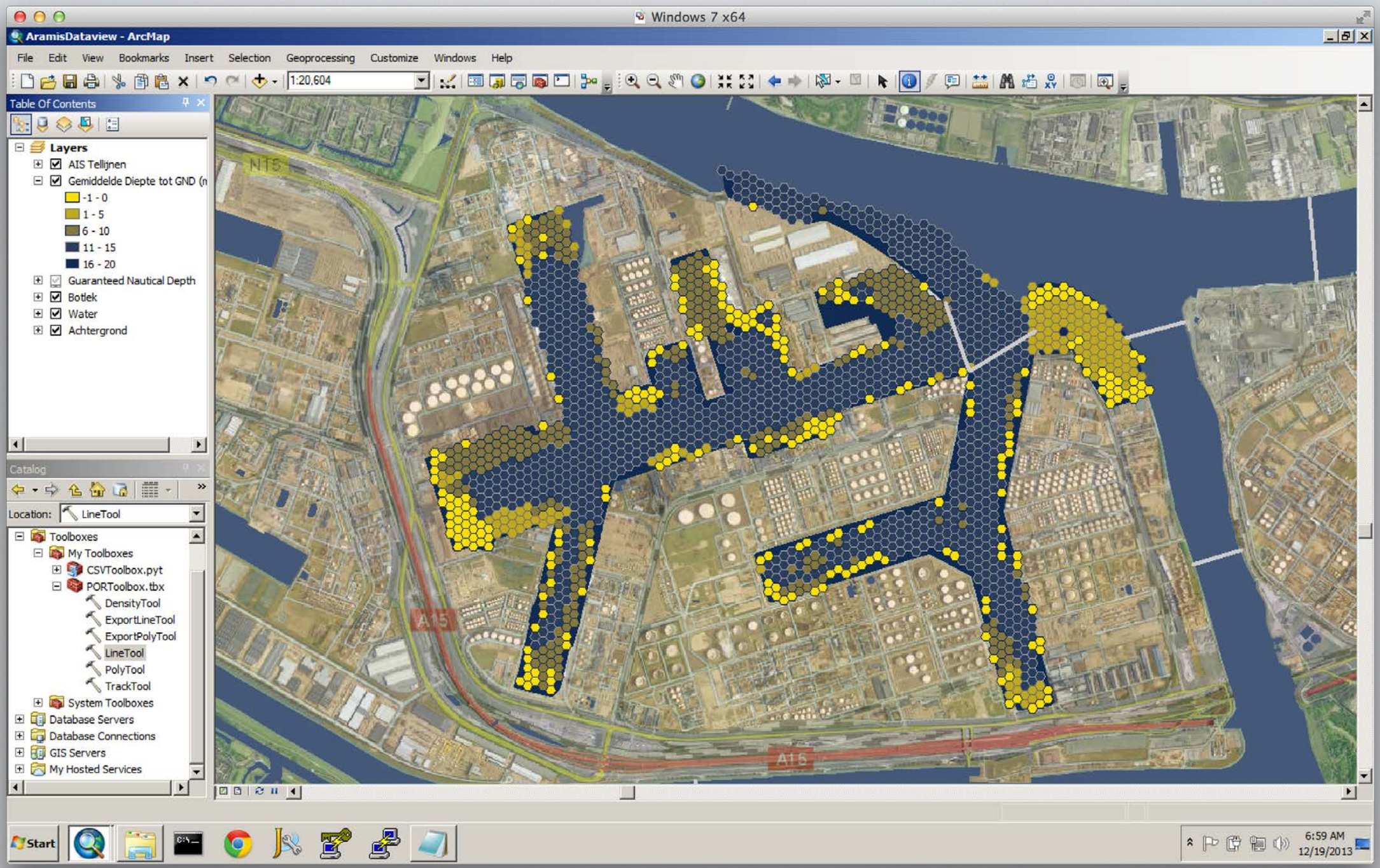
Text

Usage of ship position data

- **Harbour master**
  - Incident analysis
  - Safety checks
- **Capacity management**
  - Identifying bottlenecks
  - Planning decision support
- **Environmental management**
  - Pollution (NOx) calculations
  - Speed measures to reduce pollutions

Where is $\Delta \simeq 0$?




Windows 7 x64
AramisDataview - ArcMap
File Edit View Bookmarks Insert Selection Geoprocessing Customize Windows Help
1:20,604
Table Of Contents
Layers
AIS Tellijnen
Gemiddelde Diepte tot GND (n
-1 - 0
1 - 5
6 - 10
11 - 15
16 - 20
Guaranteed Nautical Depth
Botlek
Water
Achtergrond
Catalog
Location: LineTool
Toolboxes
My Toolboxes
CSVToolbox.pyt
PORToolbox.tbx
DensityTool
ExportLineTool
ExportPolyTool
LineTool
PolyTool
TrackTool
System Toolboxes
Database Servers
Database Connections
GIS Servers
My Hosted Services
N15
A15
A15
Start
6:59 AM
12/19/2013

FOILing NYC's Taxi Trip Da

chriswhong.com/open-data/foil_nyc_taxi/

Apps BDMap GPU

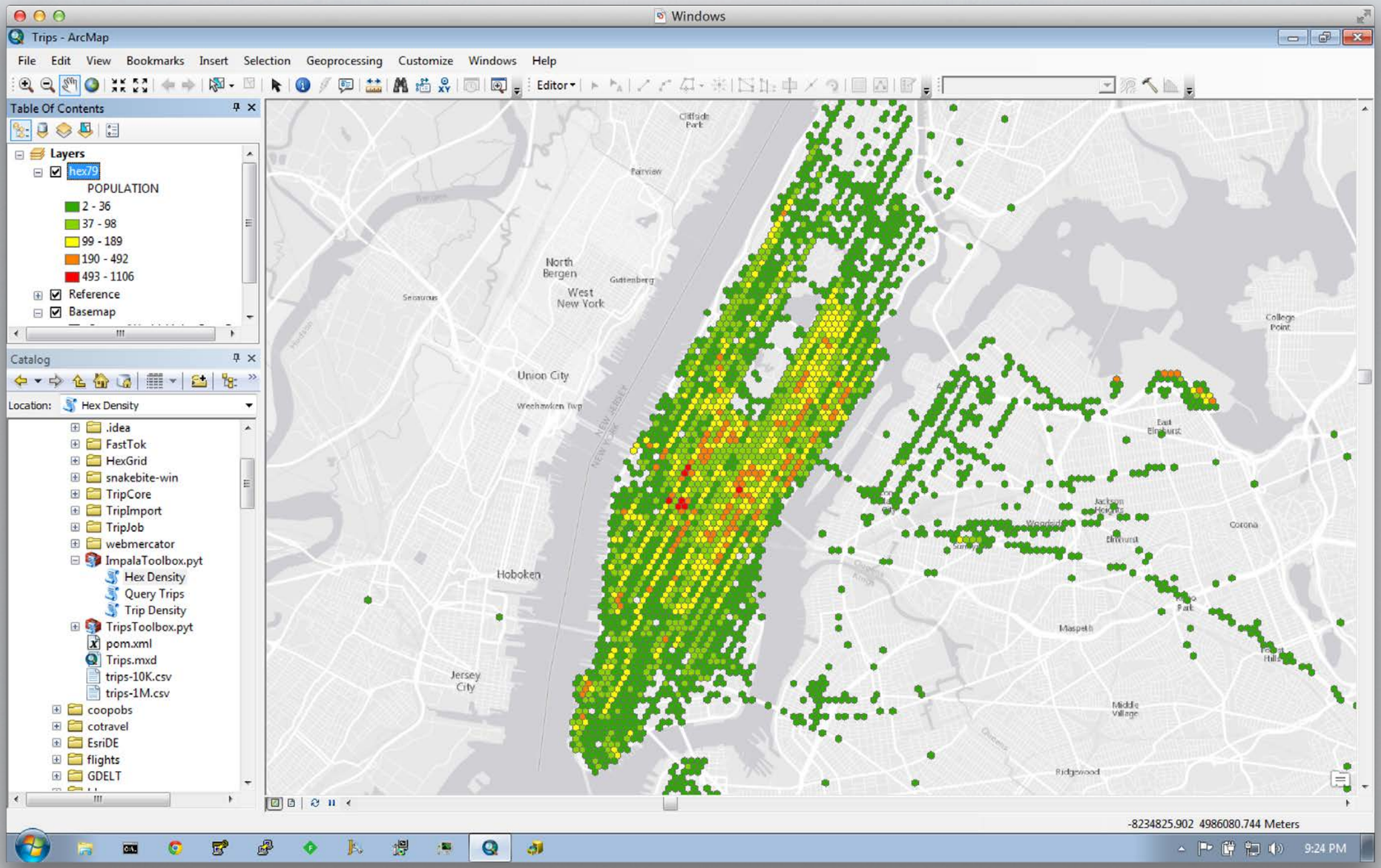
Chris Whong

Search this website...

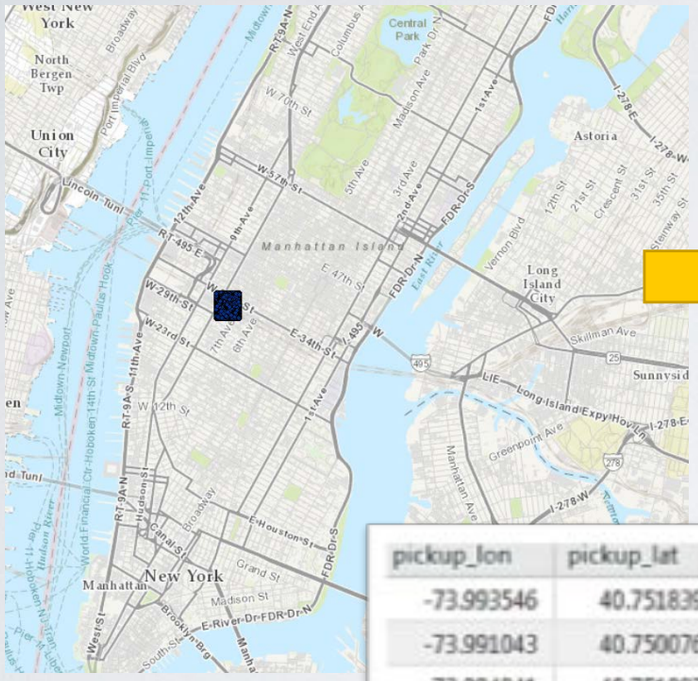
HOME DATA VISUALIZATION TRANSPORTATION NYC

FOILing NYC's Taxi Trip Data

Chris March 18, 2014 Data Visualization, Mapping, NYC, Open Data, Transportation

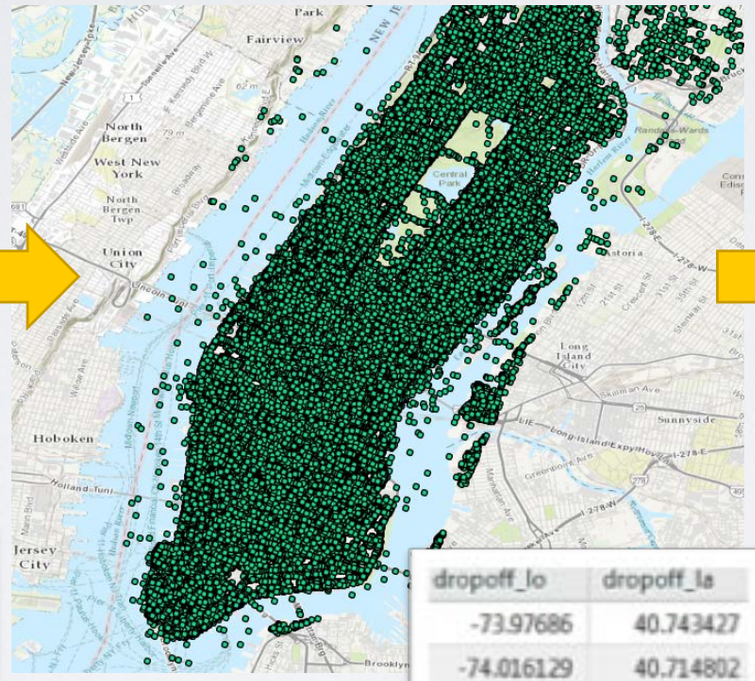

Windows
Trips - ArcMap
File Edit View Bookmarks Insert Selection Geoprocessing Customize Windows Help
Editor
Table Of Contents
Layers
hex79
POPULATION
2 - 36
37 - 98
99 - 189
190 - 492
493 - 1106
Reference
Basemap
Catalog
Location: Hex Density
.idea
FastTok
HexGrid
snakebite-win
TripCore
TripImport
TripJob
webmercator
ImpalaToolbox.pyt
Hex Density
Query Trips
Trip Density
TripsToolbox.pyt
pom.xml
Trips.mxd
trips-10K.csv
trips-1M.csv
coopobs
cotravel
EsriDE
flights
GDELT
Cliffside Park
Fairview
North Bergen
Guttenberg
West New York
Secaucus
Union City
Weehawken Twp
Hoboken
Jersey City
College Point
East Elmhurst
Jackson Heights
Corona
Maspeth
Middle Village
Ridgewood
-8234825.902 4986080.744 Meters
9:24 PM

Selected pickups



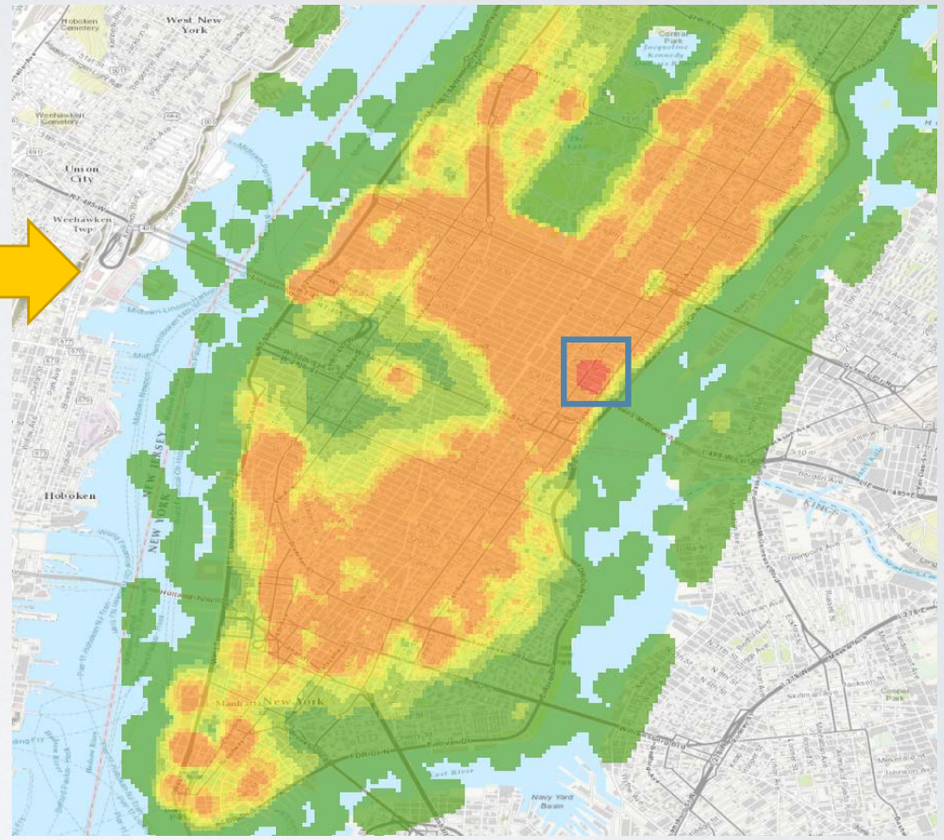
pickup_lon	pickup_lat
-73.993546	40.751839
-73.991043	40.750076
-73.994041	40.751087
-73.991463	40.750256

Corresponding drop-offs

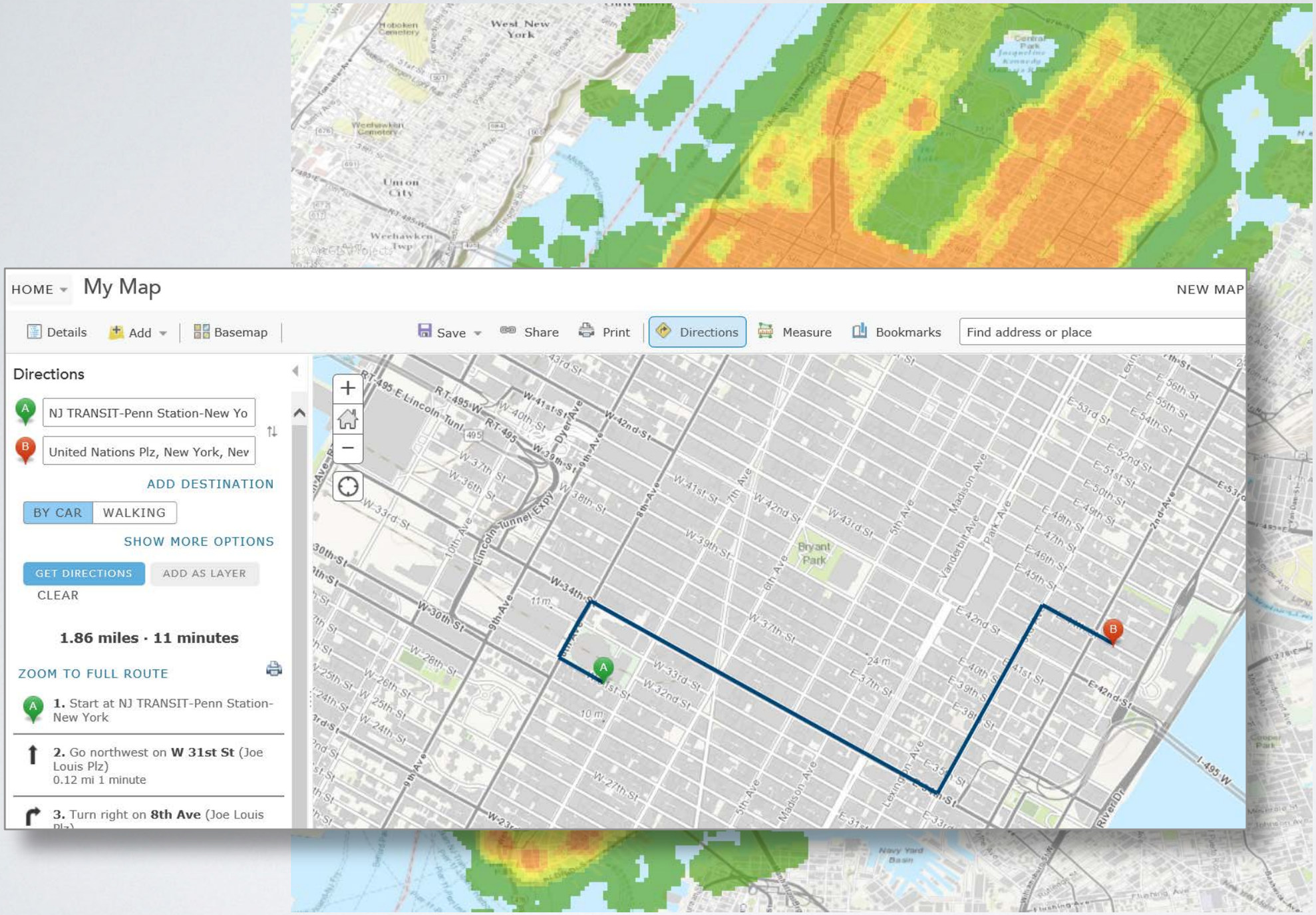


dropoff_lo	dropoff_la
-73.97686	40.743427
-74.016129	40.714802
-73.982033	40.739262
-73.980934	40.730453

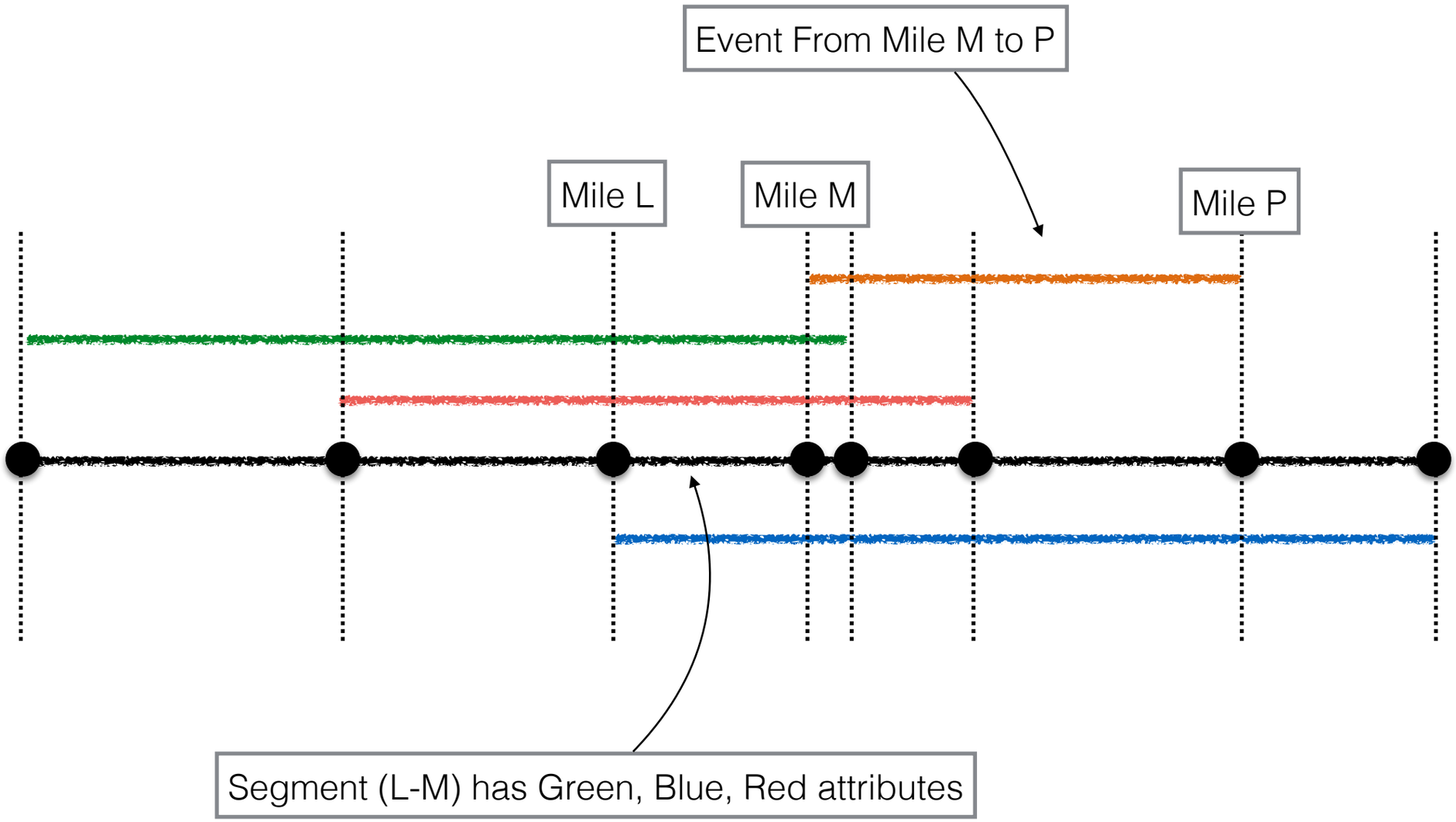
Density of passenger drop-offs



Turtle Bay – UN


HOME
My Map
NEW MAP
Details
Add
Basemap
Save
Share
Print
Directions
Measure
Bookmarks
Find address or place
Directions
NJ TRANSIT-Penn Station-New Yo
United Nations Plz, New York, Nev
ADD DESTINATION
BY CAR
WALKING
SHOW MORE OPTIONS
GET DIRECTIONS
ADD AS LAYER
CLEAR
1.86 miles · 11 minutes
ZOOM TO FULL ROUTE
1. Start at NJ TRANSIT-Penn Station-New York
2. Go northwest on W 31st St (Joe Louis Plz)
0.12 mi 1 minute
3. Turn right on 8th Ave (Joe Louis
Shuttle Locations
Pickup
Drop-off

Mission: Fast Dynamic Segmentation for Linear Referencing


Event From Mile M to P
Mile L
Mile M
Mile P
Segment (L-M) has Green, Blue, Red attributes

Step 1 - Clean and Bulk Load Roads

BulkLoad Feature - {json}

Roads.gdb
Features

ElasticSearch
{
roadid:”state_code_road_id”,
shape: “wkt-string”
}

Step 2 - New Dyn-Seg GeoProcess

Clean Roads

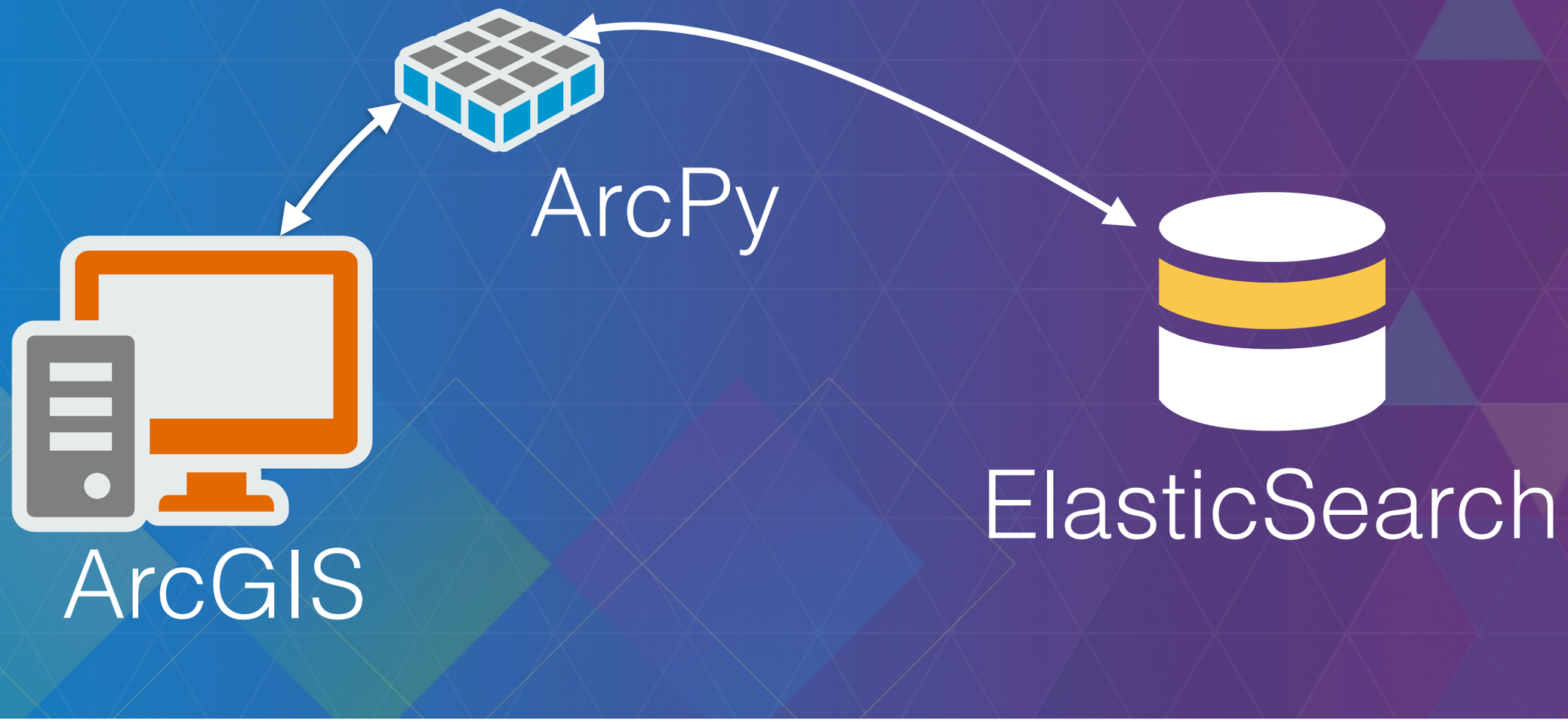
10M Events
year
statecode
routeid
mile1
mile2
key
val

Parallel
Distributed
Share-Nothing
“GeoProcessing”

{json}

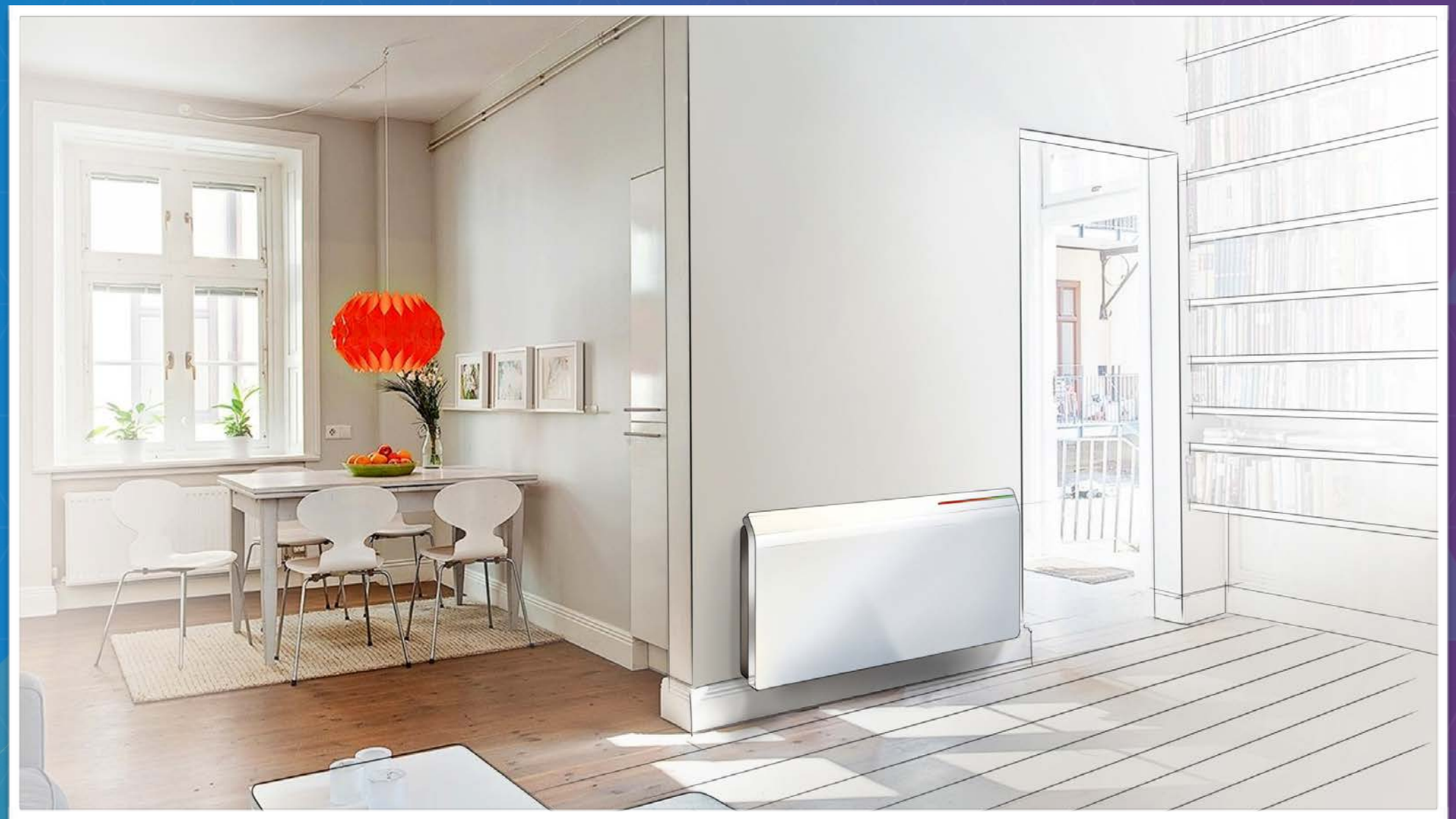
ElasticSearch
Index Every Document/Field
{
roadid=“xxx”
wkt=“multilinestring(((x y,…)
IRI=1
F_SYSTEM=2
….
}

Step 3 - Query and Display

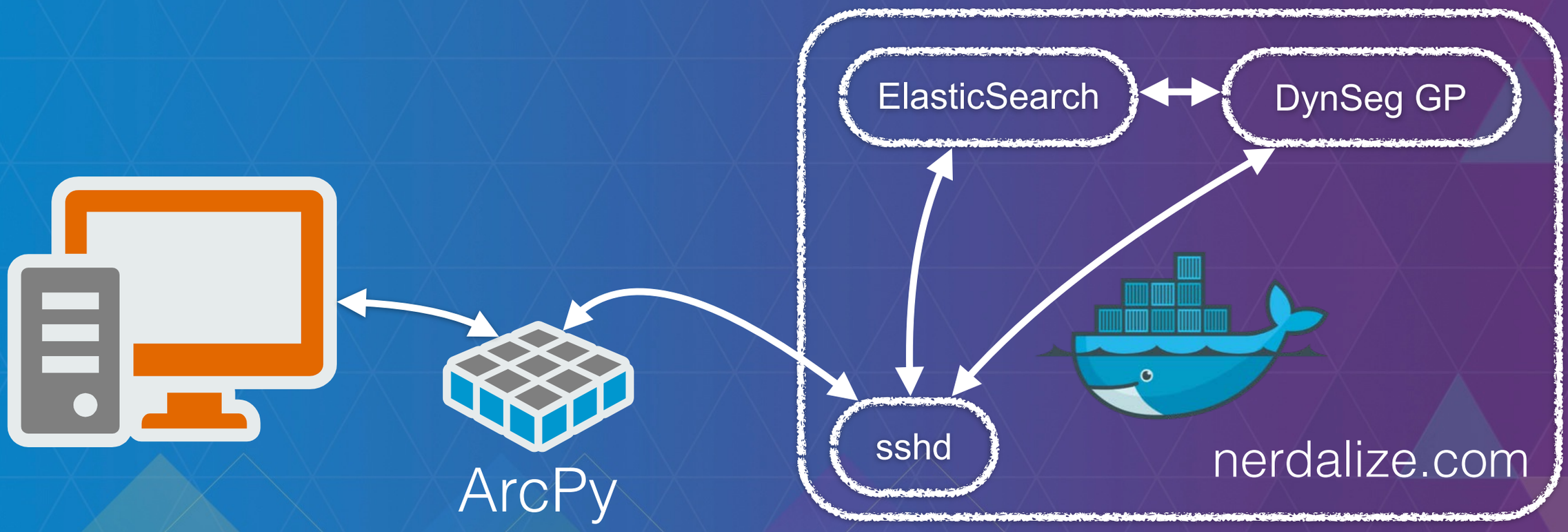


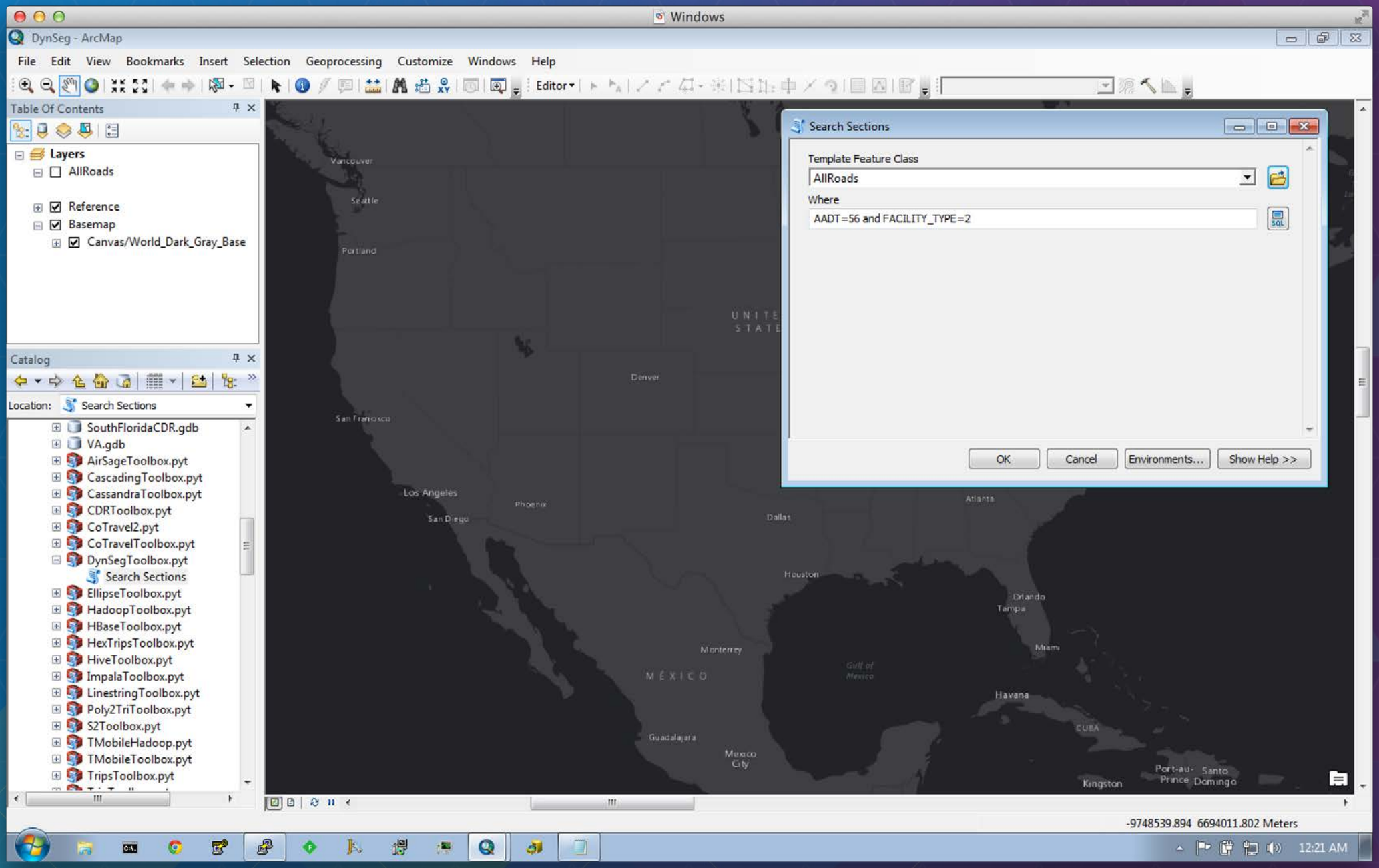
DataCenters

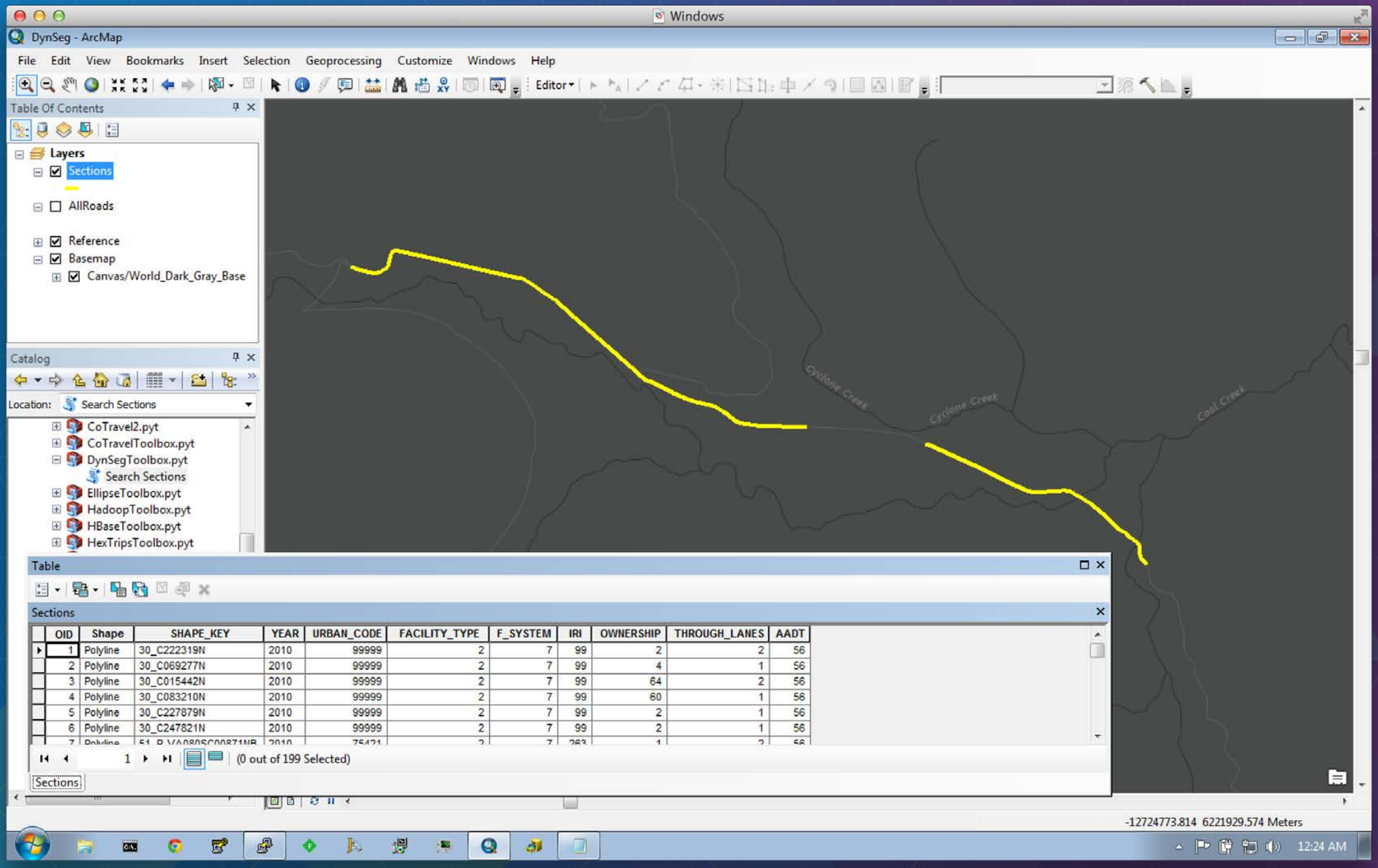

Nerdalize



ArcGIS → Nerdalize




Windows
DynSeg - ArcMap
File Edit View Bookmarks Insert Selection Geoprocessing Customize Windows Help
Editor
Table Of Contents
Layers
AllRoads
Reference
Basemap
Canvas/World_Dark_Gray_Base
Catalog
Location: Search Sections
SouthFloridaCDR.gdb
VA.gdb
AirSageToolbox.pyt
CascadingToolbox.pyt
CassandraToolbox.pyt
CDRToolbox.pyt
CoTravel2.pyt
CoTravelToolbox.pyt
DynSegToolbox.pyt
Search Sections
EllipseToolbox.pyt
HadoopToolbox.pyt
HBaseToolbox.pyt
HexTripsToolbox.pyt
HiveToolbox.pyt
ImpalaToolbox.pyt
LinestringToolbox.pyt
Poly2TriToolbox.pyt
S2Toolbox.pyt
TMobileHadoop.pyt
TMobileToolbox.pyt
TripsToolbox.pyt
Search Sections
Template Feature Class
AllRoads
Where
AADT=56 and FACILITY_TYPE=2
SQL
OK
Cancel
Environments...
Show Help >>
Vancouver
Seattle
Portland
San Francisco
Los Angeles
San Diego
Phoenix
Denver
Dallas
Houston
Atlanta
Orlando
Tampa
Miami
Monterrey
MÉXICO
Gulf of Mexico
Havana
CUBA
Guadalajara
Mexico City
Kingston
Port-au-Prince
Santo Domingo
-9748539.894 6694011.802 Meters
12:21 AM


Windows
DynSeg - ArcMap
File Edit View Bookmarks Insert Selection Geoprocessing Customize Windows Help
Editor
Table Of Contents
Layers
Sections
AllRoads
Reference
Basemap
Canvas/World_Dark_Gray_Base
Catalog
Location: Search Sections
CoTravel2.pyt
CoTravelToolbox.pyt
DynSegToolbox.pyt
Search Sections
EllipseToolbox.pyt
HadoopToolbox.pyt
HBaseToolbox.pyt
HexTripsToolbox.pyt
Cyclone Creek
Cyclone Creek
Coal Creek
Table
Sections
OID Shape SHAPE_KEY YEAR URBAN_CODE FACILITY_TYPE F_SYSTEM IRI OWNERSHIP THROUGH_LANES AADT
1 Polyline 30_C222319N 2010 99999 2 7 99 2 2 56
2 Polyline 30_C069277N 2010 99999 2 7 99 4 1 56
3 Polyline 30_C015442N 2010 99999 2 7 99 64 2 56
4 Polyline 30_C083210N 2010 99999 2 7 99 60 1 56
5 Polyline 30_C227879N 2010 99999 2 7 99 2 1 56
6 Polyline 30_C247821N 2010 99999 2 7 99 2 1 56
1 (0 out of 199 Selected)
Sections
-12724773.814 6221929.574 Meters
12:24 AM

YOU CAN DO IT TOO !

hortonworks.com/products/hortonworks-sandbox/

Blog | Partners | Contact | | | Login

HORTONWORKS

Why Hortonworks | Products | Customers | Solutions | Training | Services | Developers | Get Started

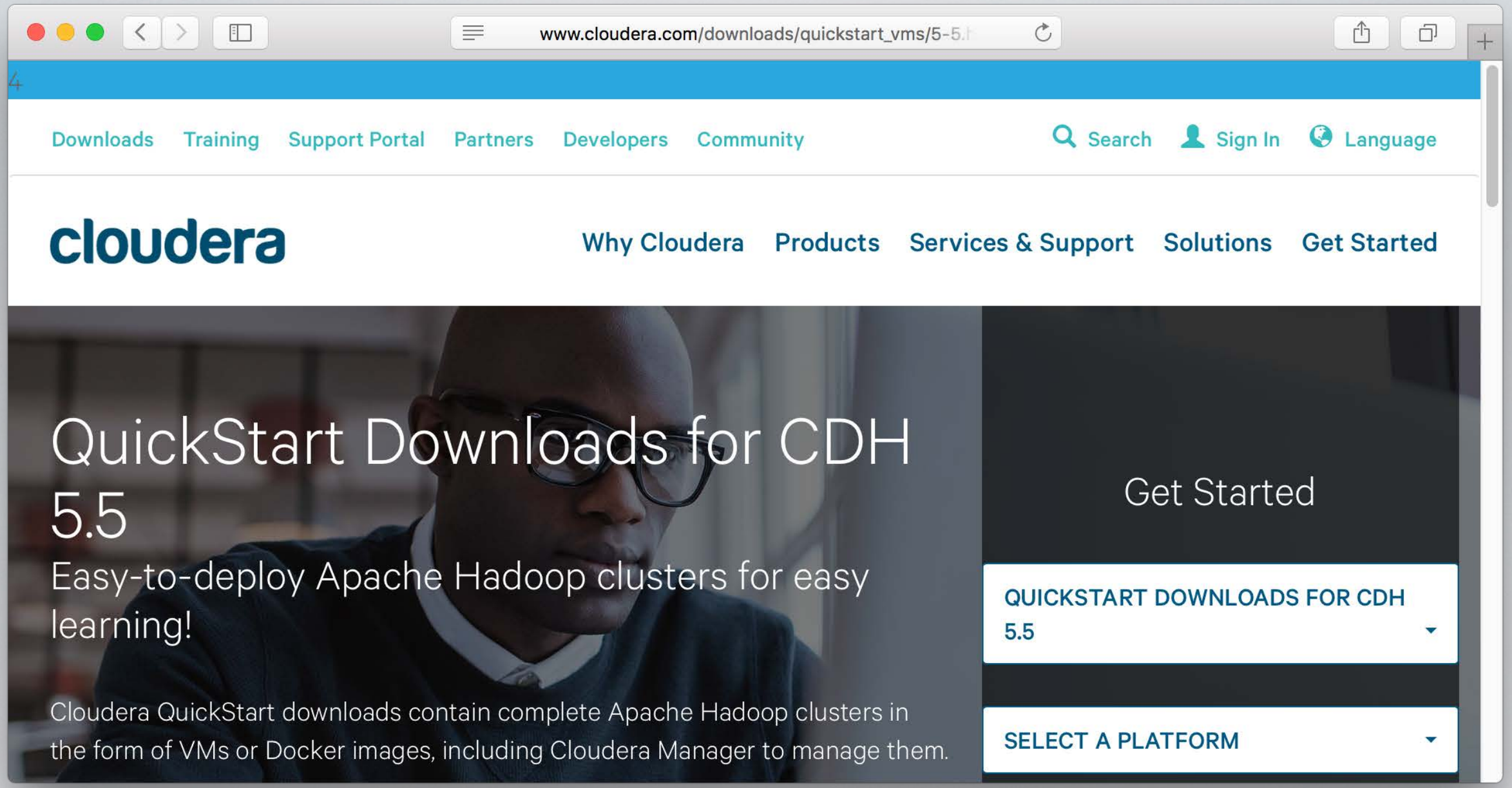
PRODUCTS » SANDBOX

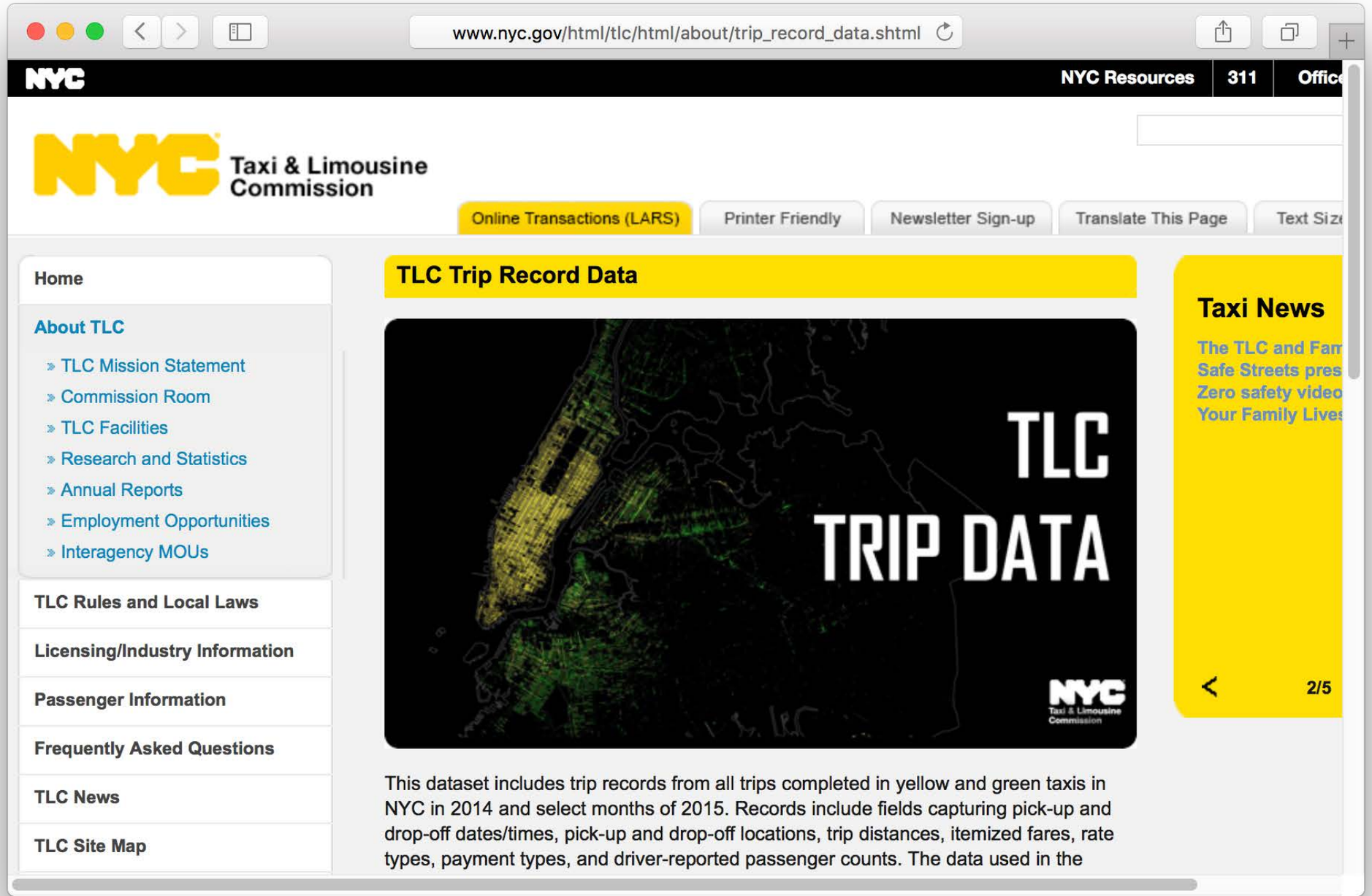
Hortonworks Sandbox

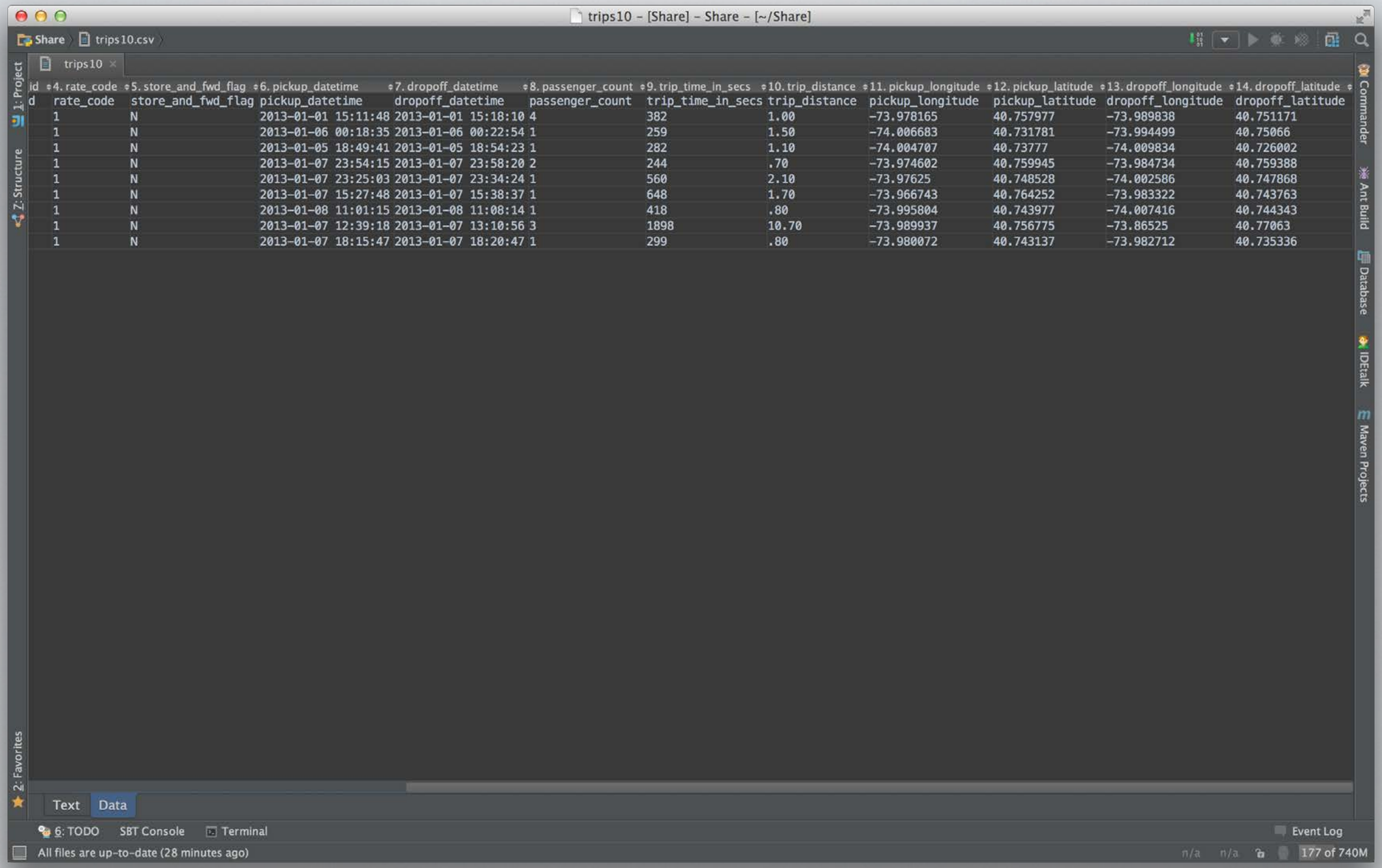
The easiest way to get started with Enterprise Hadoop

Overview | Download & Install | Tutorials | Archive | Q&A Forum

The easiest way to get started with Enterprise Hadoop

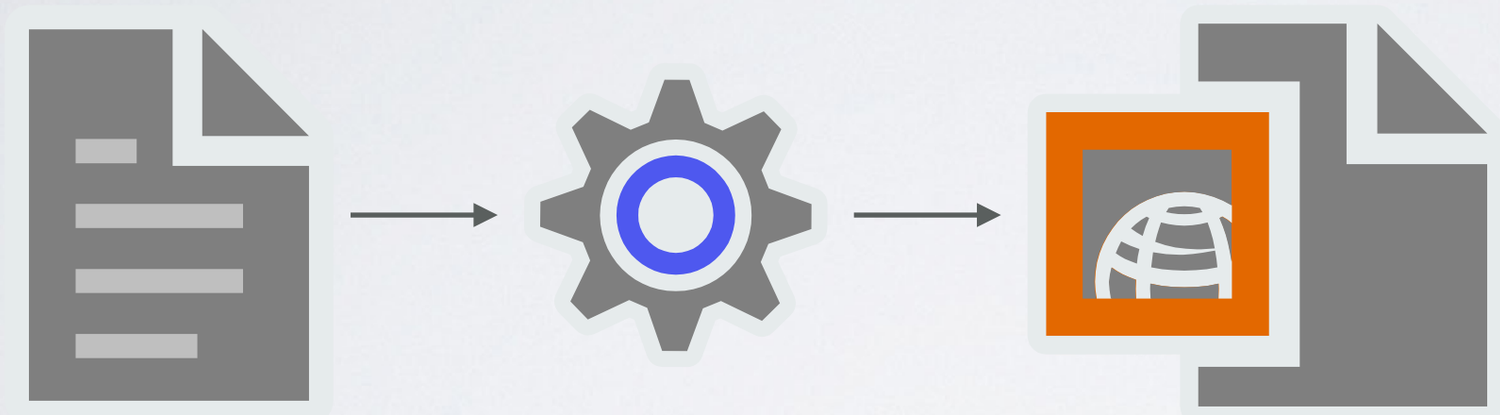

www.cloudera.com/downloads/quickstart_vms/5-5.
Downloads
Training
Support Portal
Partners
Developers
Community
Search
Sign In
Language
cloudera
Why Cloudera
Products
Services & Support
Solutions
Get Started
QuickStart Downloads for CDH 5.5
Easy-to-deploy Apache Hadoop clusters for easy learning!
Cloudera QuickStart downloads contain complete Apache Hadoop clusters in the form of VMs or Docker images, including Cloudera Manager to manage them.
Get Started
QUICKSTART DOWNLOADS FOR CDH 5.5
SELECT A PLATFORM


www.nyc.gov/html/tlc/html/about/trip_record_data.shtml
NYC
NYC Resources
311
Office
NYC Taxi & Limousine Commission
Online Transactions (LARS)
Printer Friendly
Newsletter Sign-up
Translate This Page
Text Size
Home
About TLC
TLC Mission Statement
Commission Room
TLC Facilities
Research and Statistics
Annual Reports
Employment Opportunities
Interagency MOUs
TLC Rules and Local Laws
Licensing/Industry Information
Passenger Information
Frequently Asked Questions
TLC News
TLC Site Map
TLC Trip Record Data
TLC
TRIP DATA
NYC
Taxi & Limousine Commission
Taxi News
The TLC and Fam
Safe Streets pres
Zero safety video
Your Family Lives
2/5
This dataset includes trip records from all trips completed in yellow and green taxis in NYC in 2014 and select months of 2015. Records include fields capturing pick-up and drop-off dates/times, pick-up and drop-off locations, trip distances, itemized fares, rate types, payment types, and driver-reported passenger counts. The data used in the



~180 million entries - small :-)

IMPORT BY TIME/SPACE



*.csv MRImport hdfs://…/yyyy/mm/dd/hh/uuid.csv

```
create external table if not exists trips (
pickupdatetime string,
dropoffdatetime string,
pickupx double,
pickupy double,
dropoffx double,
dropoffy double,
passengercount int,
triptime int,
tripdist double,
rc25 string,
rc50 string,
rc100 string,
rc200 string
) partitioned by (year int, month int, day int, hour int)
row format delimited
fields terminated by '\t'
lines terminated by '\n'
stored as textfile;
```

GitHub, Inc. [US] https://github.com/cloudera/impyla

GPU LingPipe BigDataIl

This repository Search Explore Gist Blog Help

cloudera / impyla

Watch 23

Python client and Numba-based UDFs for Impala

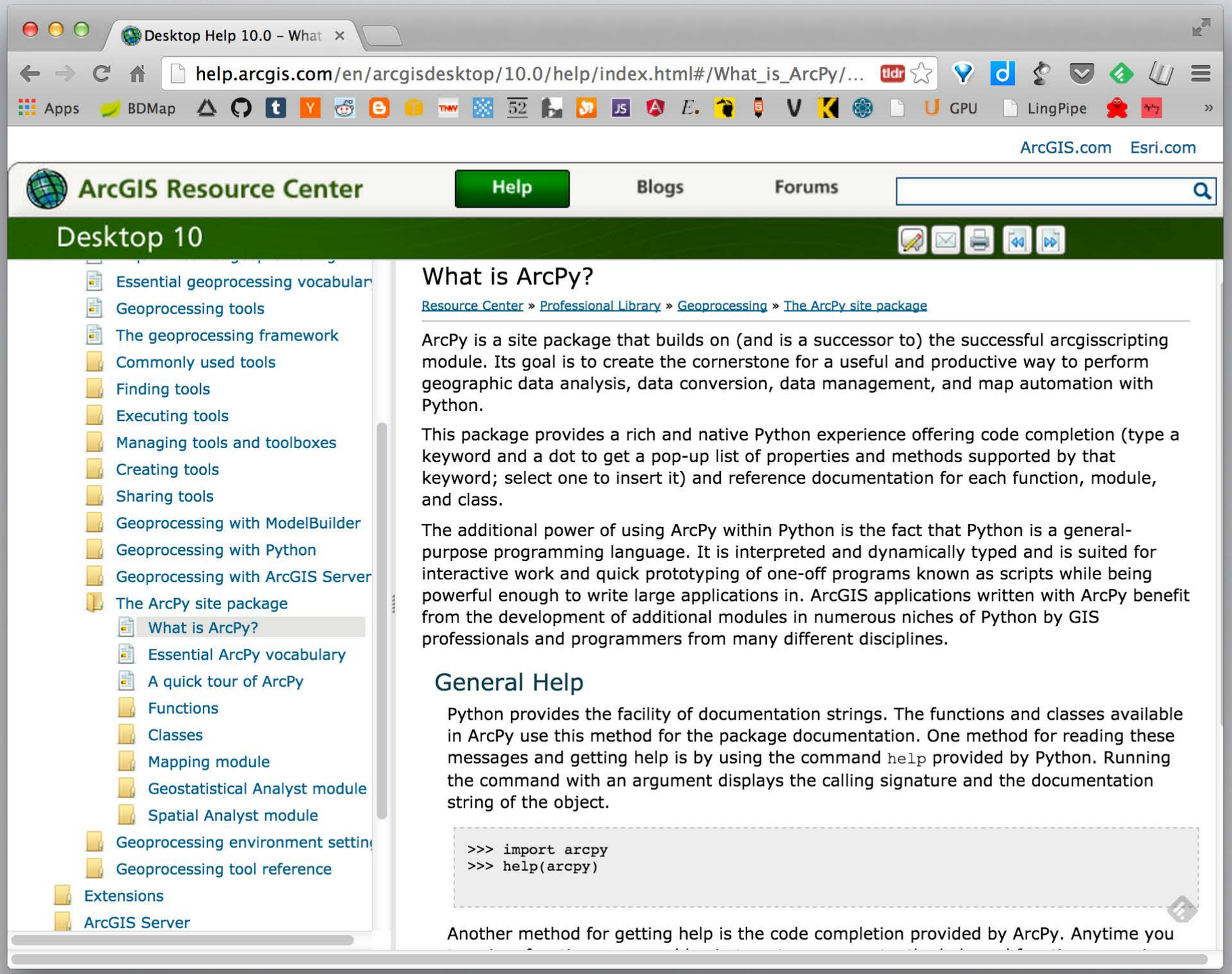
152 commits **3** branches **2** releases **5** contributors

branch: **master** impyla / +

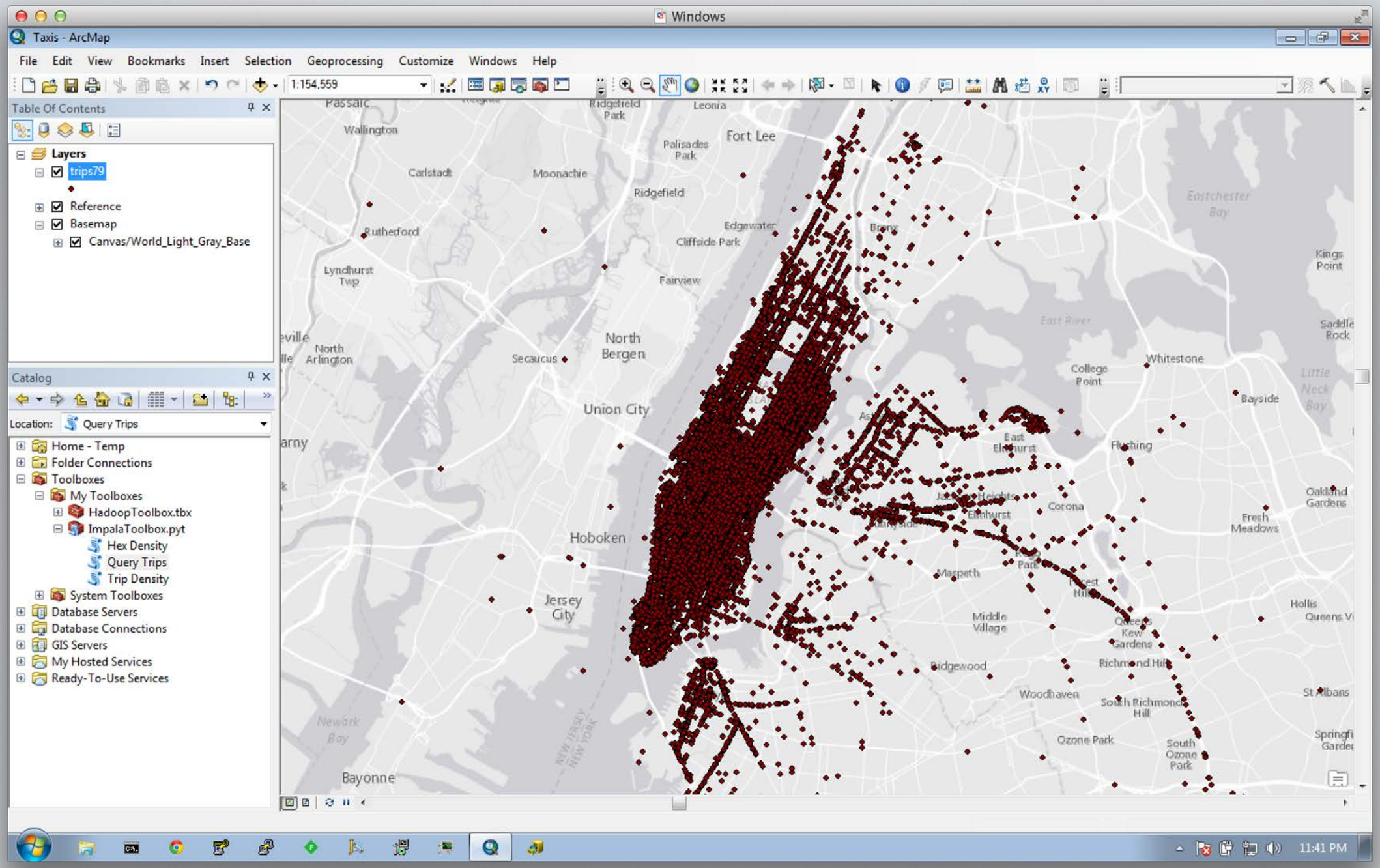
Got sklearn back into working shape as well ···

laserson authored 2 days ago latest commit 9df5e6f35e

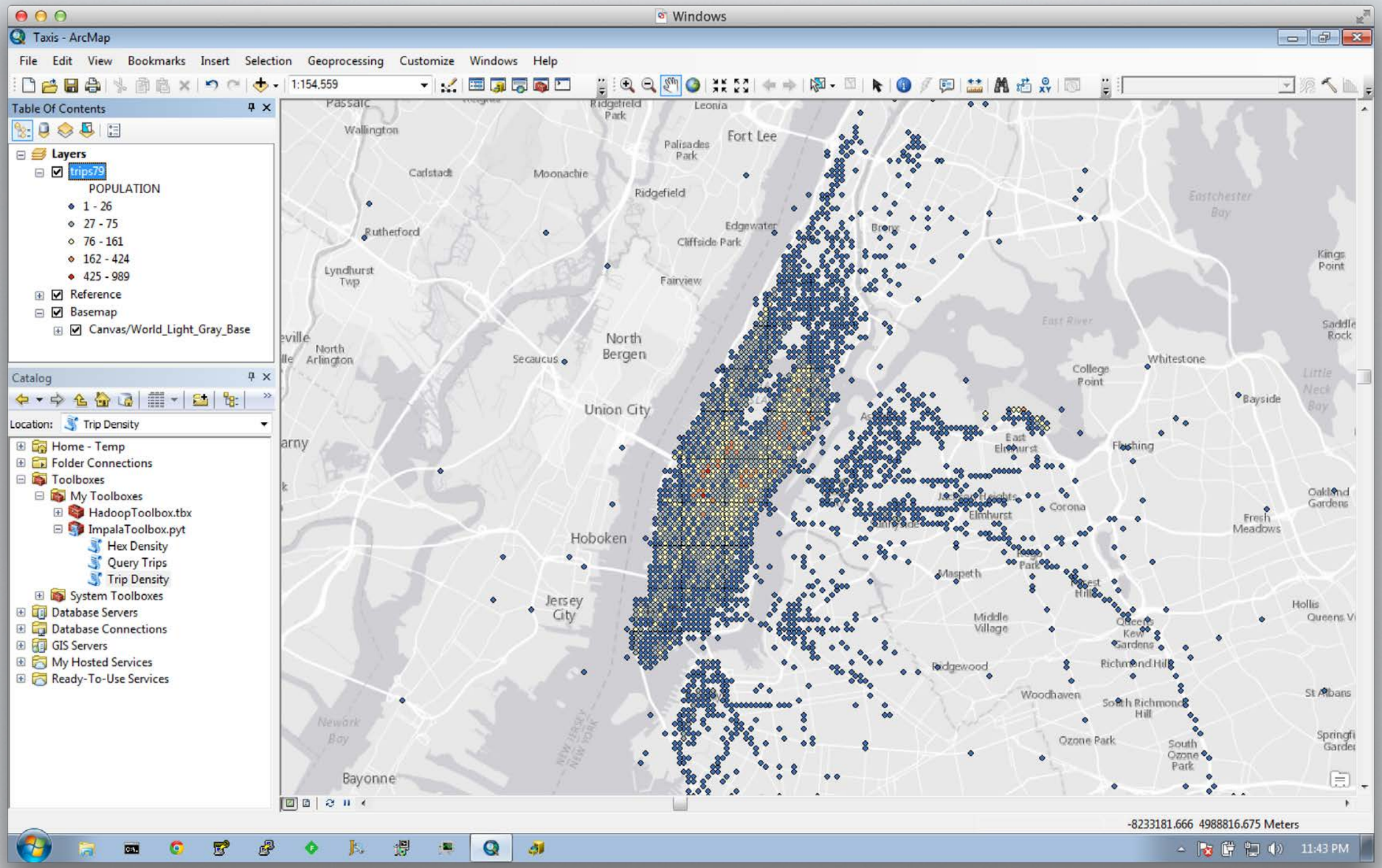
bin	Fixed tab to space issues after large merges	22 days ago
examples	Fixed tab to space issues after large merges	22 days ago
impala	Got sklearn back into working shape as well	2 days ago
jenkins	Added Jenkins test script	5 days ago
thrift	Added support for HiveServer2 protocol V6	8 days ago

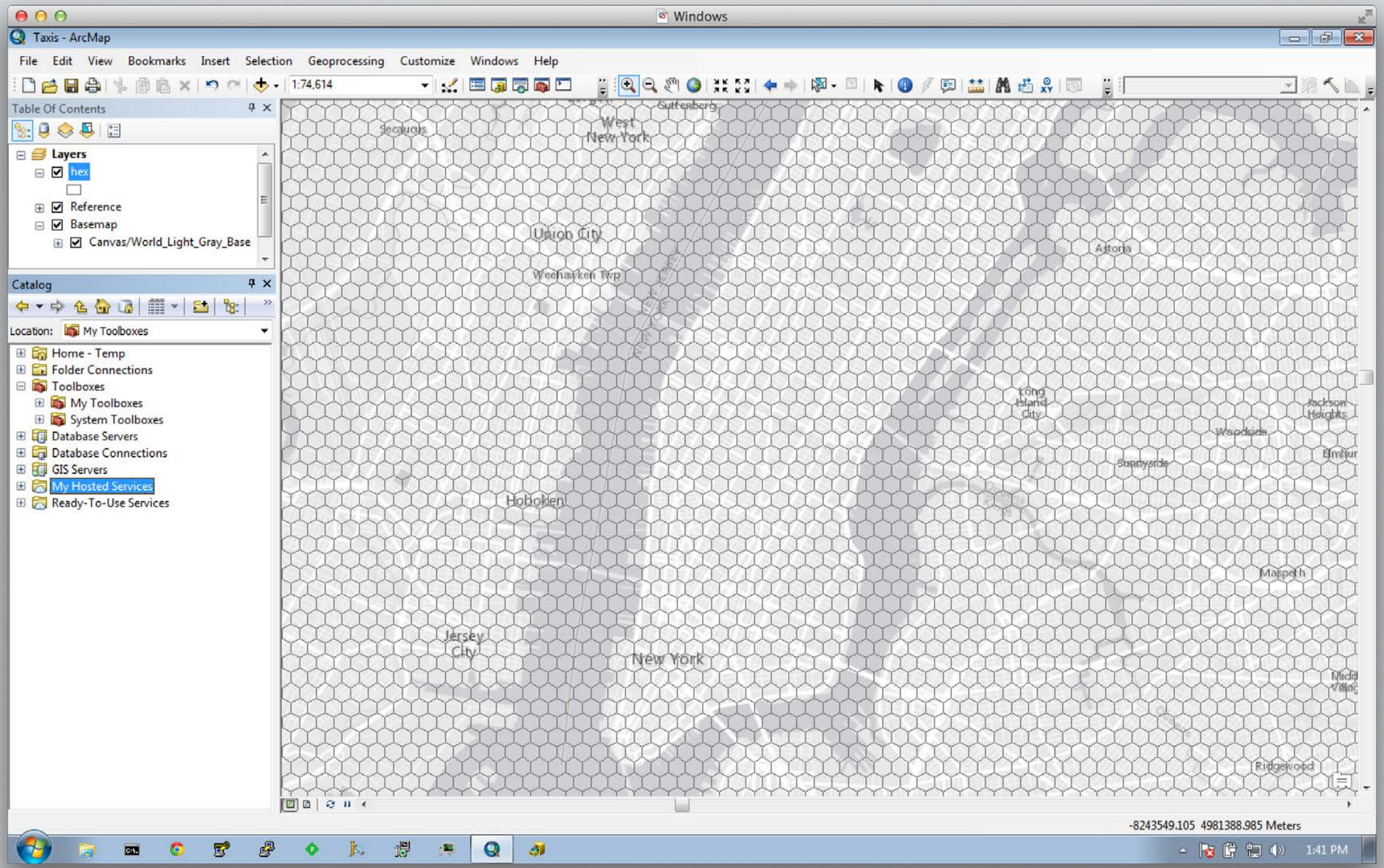

Desktop Help 10.0 – What
help.arcgis.com/en/arcgisdesktop/10.0/help/index.html#/What_is_ArcPy/...
Apps
BDMap
GPU
LingPipe
ArcGIS.com
Esri.com
ArcGIS Resource Center
Help
Blogs
Forums
Desktop 10
Essential geoprocessing vocabular
Geoprocessing tools
The geoprocessing framework
Commonly used tools
Finding tools
Executing tools
Managing tools and toolboxes
Creating tools
Sharing tools
Geoprocessing with ModelBuilder
Geoprocessing with Python
Geoprocessing with ArcGIS Server
The ArcPy site package
What is ArcPy?
Essential ArcPy vocabulary
A quick tour of ArcPy
Functions
Classes
Mapping module
Geostatistical Analyst module
Spatial Analyst module
Geoprocessing environment settin
Geoprocessing tool reference
Extensions
ArcGIS Server
What is ArcPy?
Resource Center » Professional Library » Geoprocessing » The ArcPy site package
ArcPy is a site package that builds on (and is a successor to) the successful arcgisscripting module. Its goal is to create the cornerstone for a useful and productive way to perform geographic data analysis, data conversion, data management, and map automation with Python.
This package provides a rich and native Python experience offering code completion (type a keyword and a dot to get a pop-up list of properties and methods supported by that keyword; select one to insert it) and reference documentation for each function, module, and class.
The additional power of using ArcPy within Python is the fact that Python is a general-purpose programming language. It is interpreted and dynamically typed and is suited for interactive work and quick prototyping of one-off programs known as scripts while being powerful enough to write large applications in. ArcGIS applications written with ArcPy benefit from the development of additional modules in numerous niches of Python by GIS professionals and programmers from many different disciplines.
General Help
Python provides the facility of documentation strings. The functions and classes available in ArcPy use this method for the package documentation. One method for reading these messages and getting help is by using the command help provided by Python. Running the command with an argument displays the calling signature and the documentation string of the object.
>>> import arcpy
>>> help(arcpy)
Another method for getting help is the code completion provided by ArcPy. Anytime you

```python
class TripTool(BaseTool):
    def __init__(self):
        super(TripTool, self).__init__()
        self.label = "Query Trips"
        self.description = "Tool to query trips table using Impala"
        self.canRunInBackground = False

    def getParameterInfo(self):
        paramWhere = self.getParamString(name="in_where", displayName="Where", value="hour between 7 and 9")
        return [paramWhere, self.getParamName(value="trips79"), self.getParamFC()]

    def execute(self, parameters, messages):
        try:
            name = parameters[1].value
            fc = "in_memory/" + name
            self.deleteFC(fc)
            spref = arcpy.SpatialReference(102100)
            arcpy.management.CreateFeatureclass("in_memory", name, "POINT", spatial_reference=spref)
            arcpy.management.AddField(fc, "PICKUP_DT", "TEXT")
            arcpy.management.AddField(fc, "PASS_COUNT", "SHORT")
            arcpy.management.AddField(fc, "TRIP_TIME", "SHORT")
            arcpy.management.AddField(fc, "TRIP_DIST", "FLOAT")
            conn = impala.dbapi.connect(host='quickstart', port=21050)
            rows = conn.cursor()
            hql = """select
                pickupx,
                pickupy,
                pickupdatetime,
                passengercount,
                triptime,
                tripdist
                from trips
                where {w}""".format(w=parameters[0].value)
            hql = re.sub(r'\s+', ' ', hql)
            rows.execute(hql)
            with arcpy.da.InsertCursor(fc,
                                       ['SHAPE@XY', 'PICKUP_DT', 'PASS_COUNT', 'TRIP_TIME', 'TRIP_DIST']) as cursor:
                for row in rows:
                    cursor.insertRow([(row[0], row[1]), row[2], row[3], row[4], row[5]])
                del row
            del rows
            parameters[2].value = fc
        except:
            arcpy.AddMessage(traceback.format_exc())
```


Windows
Taxis - ArcMap
File Edit View Bookmarks Insert Selection Geoprocessing Customize Windows Help
1:154,559
Table Of Contents
Layers
trips79
Reference
Basemap
Canvas/World_Light_Gray_Base
Catalog
Location: Query Trips
Home - Temp
Folder Connections
Toolboxes
My Toolboxes
HadoopToolbox.tbx
ImpalaToolbox.pyt
Hex Density
Query Trips
Trip Density
System Toolboxes
Database Servers
Database Connections
GIS Servers
My Hosted Services
Ready-To-Use Services
Passaic
Wallington
Ridgefield Park
Leonia
Fort Lee
Palisades Park
Carlstadt
Moonachie
Ridgefield
Rutherford
Edgewater
Cliffside Park
Bronx
Eastchester Bay
Kings Point
Lyndhurst Twp
Fairview
East River
Saddle Rock
North Arlington
Secaucus
North Bergen
College Point
Whitestone
Little Neck Bay
Bayside
Union City
East Elmhurst
Flushing
Oakland Gardens
Corona
Elmhurst
Fresh Meadows
Hoboken
Maspeth
Middle Village
Hollis
Queens Vi
Kew Gardens
Ridgewood
Richmond Hill
Woodhaven
South Richmond Hill
St Albans
Ozone Park
South Ozone Park
Springfi Garde
Jersey City
Newark Bay
NEW JERSEY NEW YORK
Bayonne
11:41 PM

```
    def execute(self, parameters, messages):
        try:
            cell1 = float(parameters[0].value)
            cell2 = cell1 * 0.5
            where = parameters[1].value
            name = parameters[2].value

            fc = "in_memory/" + name
            self.deleteFC(fc)
            spref = arcpy.SpatialReference(102100)
            arcpy.management.CreateFeatureclass("in_memory", name, "POINT", spatial_reference=spref)
            arcpy.management.AddField(fc, "POPULATION", "LONG")
            conn = impala.dbapi.connect(host='quickstart', port=21050)
            rows = conn.cursor()
            hql = """select
                T.X*{c1}+{c2} as X,
                T.Y*{c1}+{c2} as Y,
                count(*) AS POPULATION from (
                select
                cast(floor(pickupx/{c1}) as int) as X,
                cast(floor(pickupy/{c1}) as int) as Y
                from trips where {w}) T
                group by T.X,T.Y
            """.format(w=where, c1=cell1, c2=cell2)
            hql = re.sub(r'\s+', ' ', hql)
            rows.execute(hql)
            with arcpy.da.InsertCursor(fc, ['SHAPE@XY', 'POPULATION']) as cursor:
                for row in rows:
                    cursor.insertRow([(row[0], row[1]), row[2]])
                del row
            del rows
            parameters[3].value = fc
        except:
            arcpy.AddMessage(traceback.format_exc())
```


Windows
Taxis - ArcMap
File Edit View Bookmarks Insert Selection Geoprocessing Customize Windows Help
1:154,559
Table Of Contents
Layers
trips79
POPULATION
1 - 26
27 - 75
76 - 161
162 - 424
425 - 989
Reference
Basemap
Canvas/World_Light_Gray_Base
Catalog
Location: Trip Density
Home - Temp
Folder Connections
Toolboxes
My Toolboxes
HadoopToolbox.tbx
ImpalaToolbox.pyt
Hex Density
Query Trips
Trip Density
System Toolboxes
Database Servers
Database Connections
GIS Servers
My Hosted Services
Ready-To-Use Services
Passaic
Wallington
Ridgefield Park
Leonia
Fort Lee
Palisades Park
Carlstadt
Moonachie
Ridgefield
Rutherford
Edgewater
Cliffside Park
Bronx
Eastchester Bay
Kings Point
Lyndhurst Twp
Fairview
East River
Saddle Rock
North Arlington
Secaucus
North Bergen
College Point
Whitestone
Little Neck Bay
Bayside
Union City
East Elmhurst
Flushing
Oakland Gardens
Corona
Elmhurst
Fresh Meadows
Hoboken
Maspeth
Middle Village
Hollis
Queens Village
Kew Gardens
Ridgewood
Richmond Hill
Woodhaven
South Richmond Hill
St Albans
Ozone Park
South Ozone Park
Springfield Gardens
Jersey City
Newark Bay
Bayonne
-8233181.666 4988816.675 Meters
11:43 PM


Windows
Taxis - ArcMap
File Edit View Bookmarks Insert Selection Geoprocessing Customize Windows Help
1:74,614
Table Of Contents
Layers
hex
Reference
Basemap
Canvas/World_Light_Gray_Base
Catalog
Location: My Toolboxes
Home - Temp
Folder Connections
Toolboxes
My Toolboxes
System Toolboxes
Database Servers
Database Connections
GIS Servers
My Hosted Services
Ready-To-Use Services
Secaucus
Guttenberg
West New York
Union City
Weehawken Twp
Astoria
Long Island City
Jackson Heights
Woodside
Sunnyside
Hoboken
Maspeth
Jersey City
New York
Ridgewood
-8243549.105 4981388.985 Meters
1:41 PM

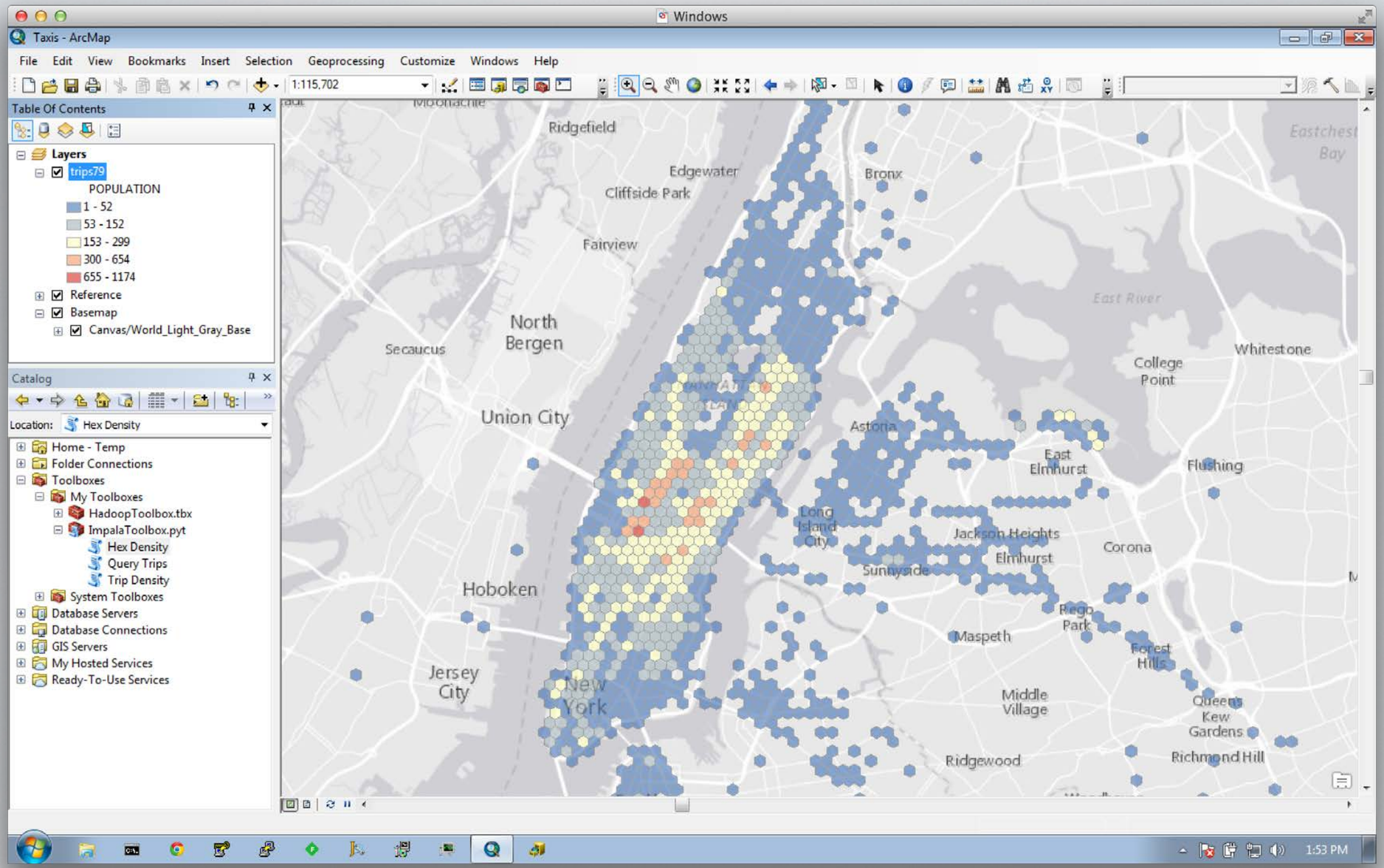
cloudera-quickstart-vm-5.1.0-1-vmware

Applications Places System | Home - Cloudera Man... | cloudera@quickstart:~ | Wed Oct 15, 10:48 AM cloudera

cloudera@quickstart:~

File Edit View Search Terminal Help

```
Welcome to the Impala shell. Press TAB twice to see a list of available commands.

Copyright (c) 2012 Cloudera, Inc. All rights reserved.

(Shell build version: Impala Shell v1.4.0-cdh5-INTERNAL (e801bd8) built on Sat Jul 12 06:45:04 PDT 2014)
Query: invalidate metadata

Returned 0 row(s) in 3.75s
[quickstart.cloudera:21000] > describe trips;
Query: describe trips
+-----------------+--------+---------+
| name            | type   | comment |
+-----------------+--------+---------+
| pickupdatetime  | string |         |
| dropoffdatetime | string |         |
| pickupx         | double |         |
| pickupy         | double |         |
| dropoffx        | double |         |
| dropoffy        | double |         |
| passengercount  | int    |         |
| triptime        | int    |         |
| tripdist        | double |         |
| rc25            | string |         |
| rc50            | string |         |
| rc100           | string |         |
| rc200           | string |         |
| year            | int    |         |
| month           | int    |         |
| day             | int    |         |
| hour            | int    |         |
+-----------------+--------+---------+
Returned 17 row(s) in 14.18s
[quickstart.cloudera:21000] > select pickupx,pickupy,rc25,rc100 from trips limit 5;
Query: select pickupx,pickupy,rc25,rc100 from trips limit 5
+--------------------+-------------------+-----------+----------+
| pickupx            | pickupy           | rc25      | rc100    |
+--------------------+-------------------+-----------+----------+
| -8235510.550453553 | 4977585.413571722 | 4736|1489 | 1184|372 |
| -8231890.885890919 | 4975433.207053881 | 4678|1573 | 1170|393 |
| -8236461.775502382 | 4976817.795186556 | 4715|1467 | 1179|366 |
| -8214370.645193438 | 4960174.878029009 | 4271|1977 | 1068|494 |
| -8234755.581666993 | 4977389.204397635 | 4730|1507 | 1183|376 |
+--------------------+-------------------+-----------+----------+
Returned 5 row(s) in 1.06s
[quickstart.cloudera:21000] >
```


Windows
Taxis - ArcMap
File Edit View Bookmarks Insert Selection Geoprocessing Customize Windows Help
1:115,702
Table Of Contents
Layers
trips79
POPULATION
1 - 52
53 - 152
153 - 299
300 - 654
655 - 1174
Reference
Basemap
Canvas/World_Light_Gray_Base
Catalog
Location: Hex Density
Home - Temp
Folder Connections
Toolboxes
My Toolboxes
HadoopToolbox.tbx
ImpalaToolbox.pyt
Hex Density
Query Trips
Trip Density
System Toolboxes
Database Servers
Database Connections
GIS Servers
My Hosted Services
Ready-To-Use Services
Ridgefield
Edgewater
Cliffside Park
Bronx
Fairview
North Bergen
Secaucus
Union City
Astoria
East Elmhurst
Flushing
Long Island City
Jackson Heights
Elmhurst
Corona
Sunnyside
Hoboken
Rego Park
Maspeth
Forest Hills
Jersey City
New York
Middle Village
Queens
Kew Gardens
Ridgewood
Richmond Hill
College Point
Whitestone
East River
1:53 PM

STATISTICAL SIGNIFICANCE ?

HOTSPOT ANALYSIS

The Getis-Ord local statistic is given as:

$$G_i^* = \frac{\sum_{j=1}^{n} w_{i,j}x_j - \bar{X}\sum_{j=1}^{n} w_{i,j}}{S\sqrt{\frac{\left[n\sum_{j=1}^{n} w_{i,j}^2 - \left(\sum_{j=1}^{n} w_{i,j}\right)^2\right]}{n-1}}} \quad (1)$$

where x_j is the attribute value for feature j, $w_{i,j}$ is the spatial weight between feature i and j, n is equal to the total number of features and:

$$\bar{X} = \frac{\sum_{j=1}^{n} x_j}{n} \quad (2)$$

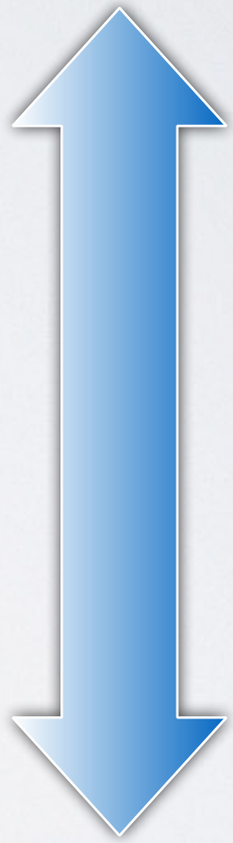
$$S = \sqrt{\frac{\sum_{j=1}^{n} x_j^2}{n} - \left(\bar{X}\right)^2} \quad (3)$$

The G_i^* statistic is a z-score so no further calculations are required.

http://resources.arcgis.com/en/help/main/10.1/index.html#//005p0000001000000

PROCESSING EVOLUTION

- Transaction - Batch
- Operational - Dashboard
- Analytics - Exploration
- Intelligent - Realtime / Predictive



WHAT IS NEXT ?

- In Memory
- Native Spatial Index In NoSQL DB
- Native Spatial Types (Point, Line, …)
- Out-of-the-box Spatial Operators / Operations
- Distributed/Disconnected GPU Integration
- Visualization via Gamification

Q&A

Mansour Raad
thunderheadxpler.blogspot.com
mraad@esri.com
@mraad