Federal GIS Conference 2014

February 10–11, 2014 | Washington DC

BigData And the Zoo

Mansour Raad

http://thunderheadxpler.blogspot.com/

mraad@esri.com

@mraad

What is **BigData** ?

”

Big data is like teenage sex: everyone talks about it, nobody really knows how to do it, everyone thinks everyone else is doing it, so everyone claims they are doing it…

- Dan Ariely

No…but seriously !

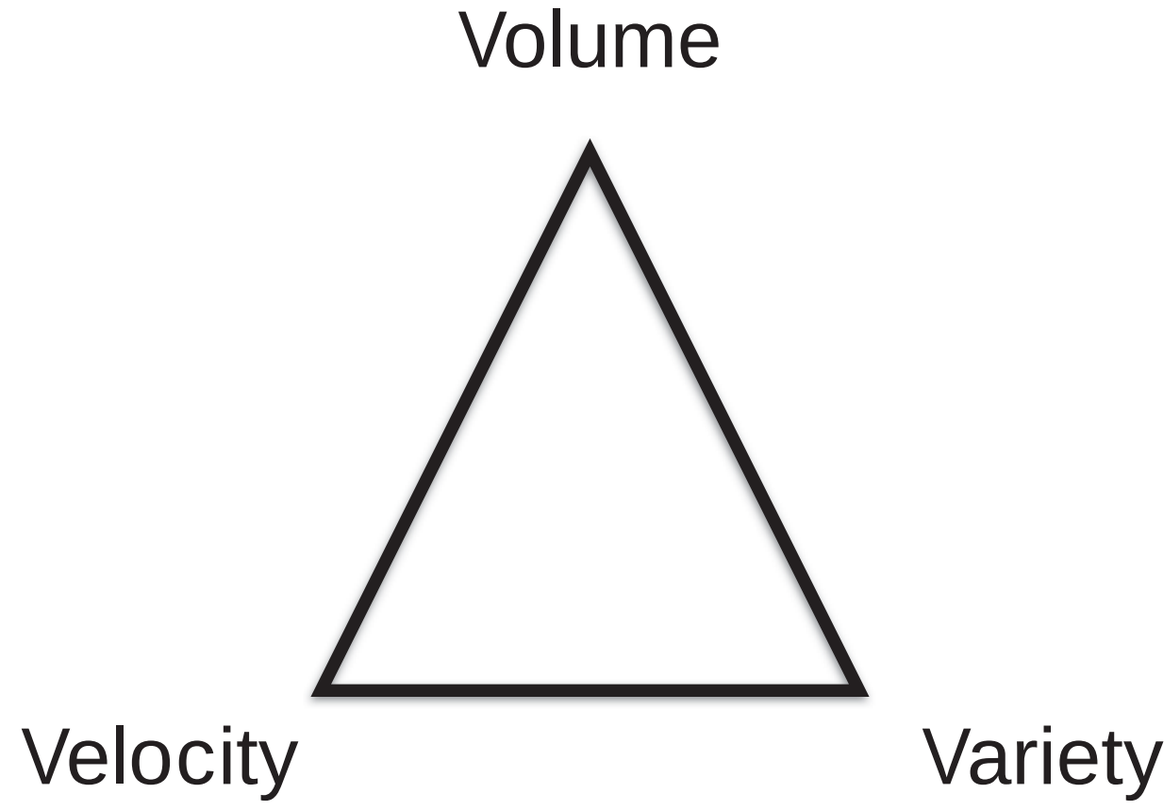
Academics

- **V**olume
- **V**elocity
- **V**ariety

But then I’ve seen…

- **V**olume → data at rest
- **V**elocity → data in motion
- **V**ariety → many types, forms and structures or no structures
- **V**eracity → data in doubt
- **V**alidity → data that is correct
- **V**isualization → data in patterns
- **V**ulnerability → data at risk
- **V**alue → data that is meaningful

I’m sticking with…

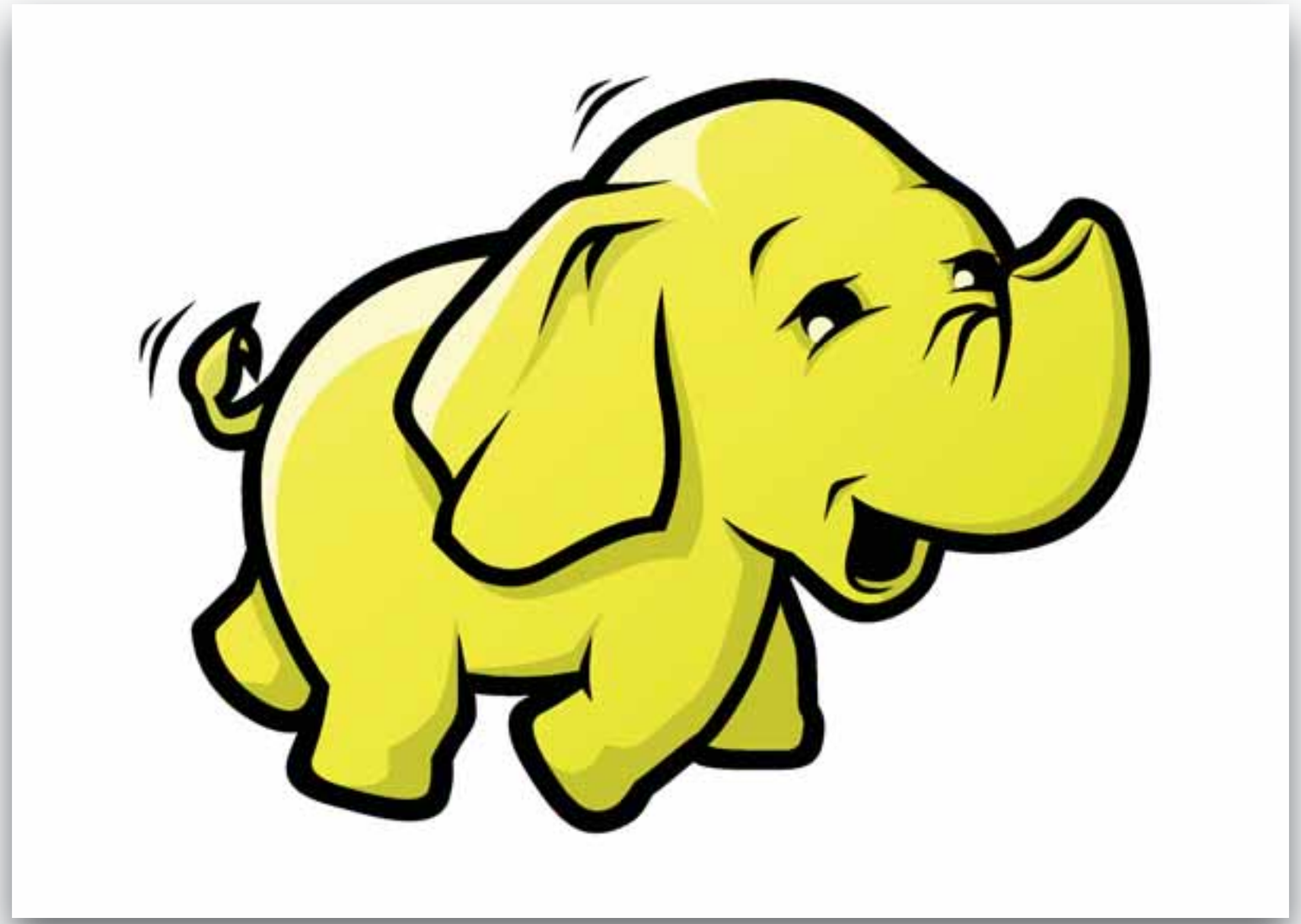


“When the traditional means are failing you”

-Anonymous

What are the new means?

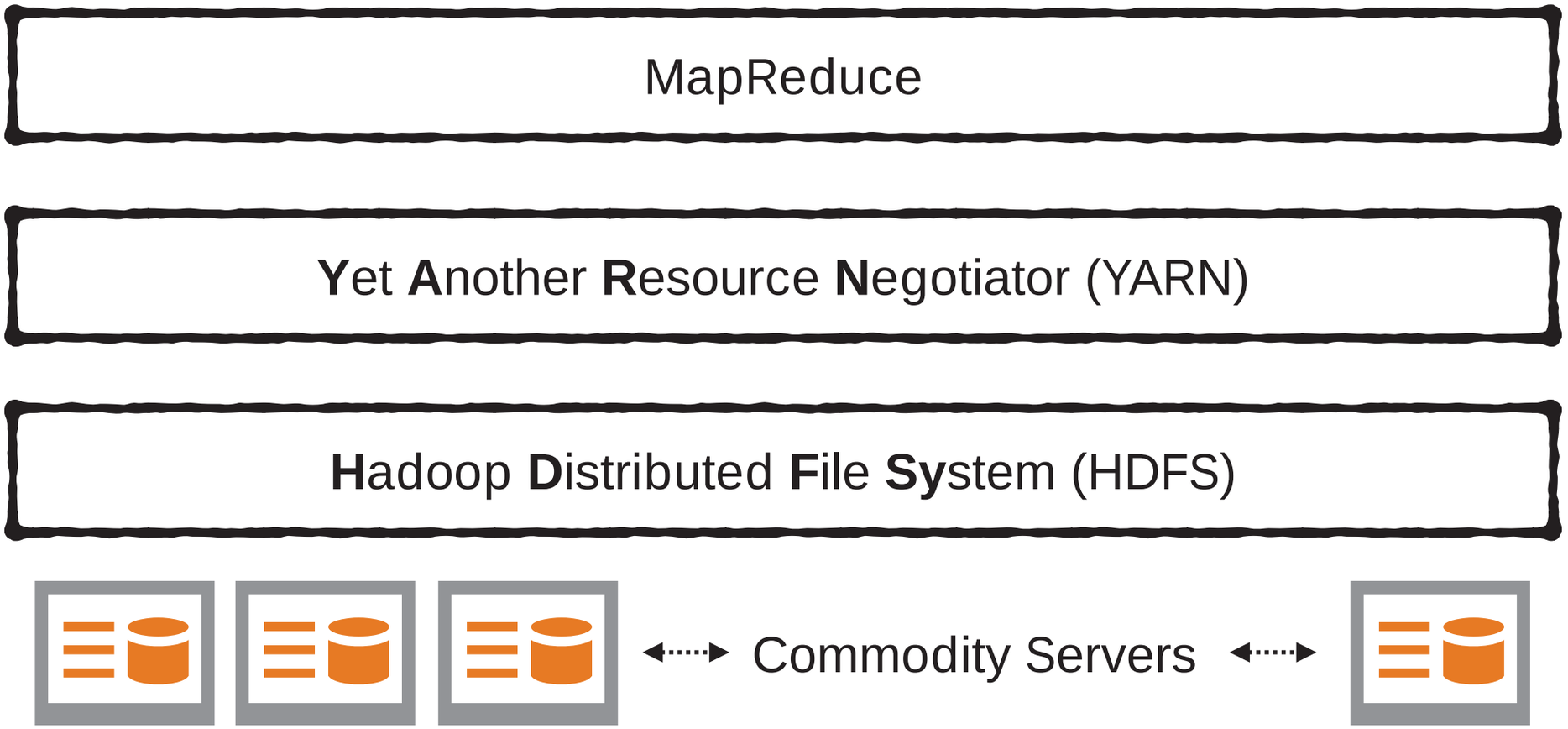
http://hadoop.apache.org



What Is Hadoop ?

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

Hadoop Basic Stack



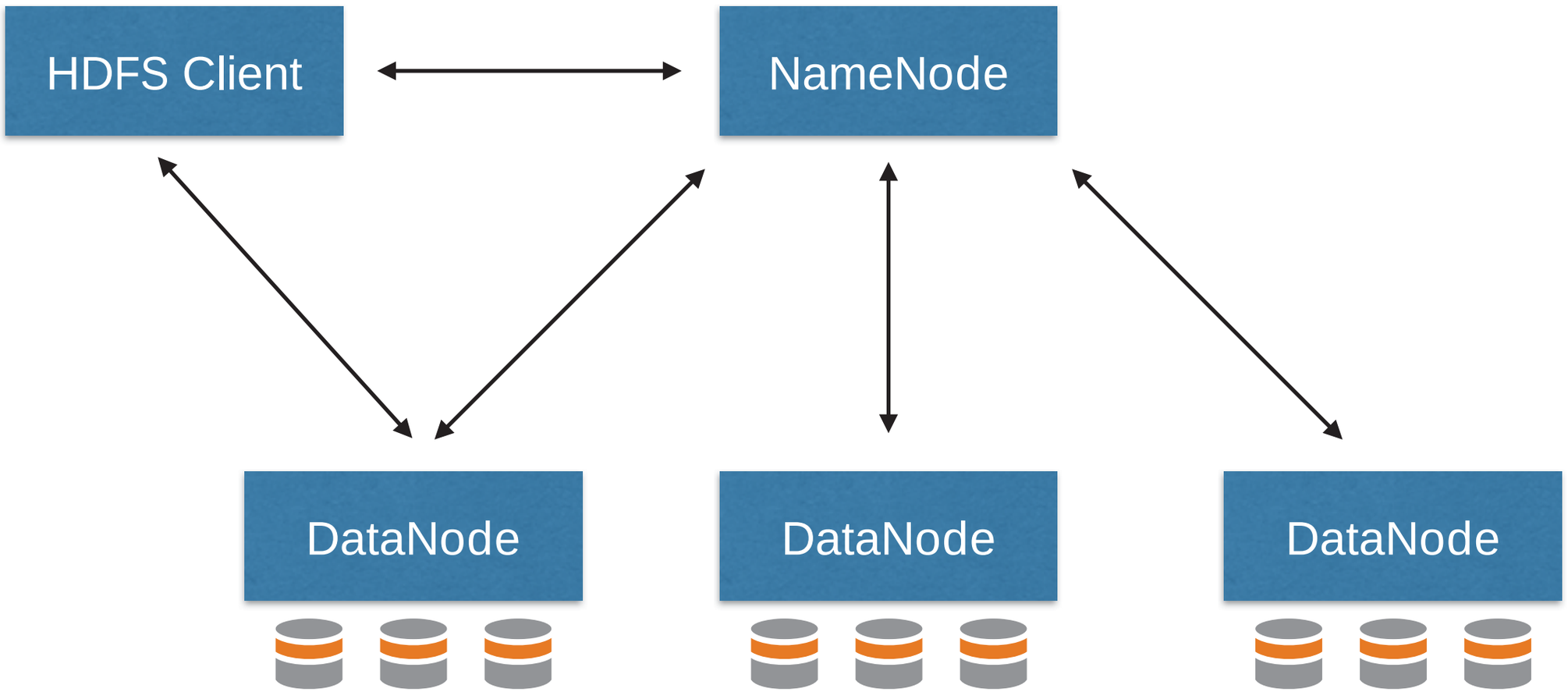
Other Hadoop Projects

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

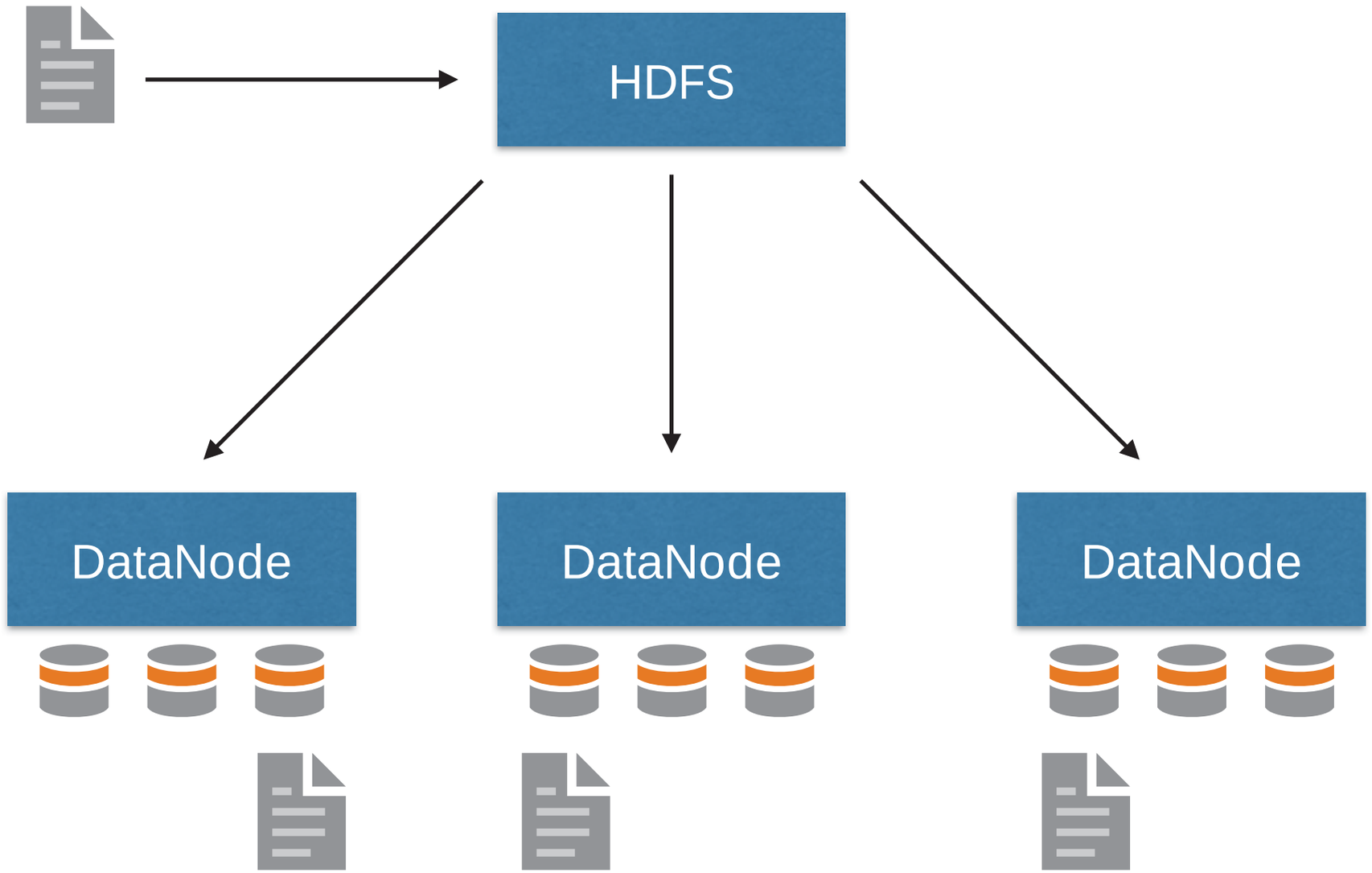
HDFS

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

HDFS



HDFS Resilience !



Program

BigData

Program

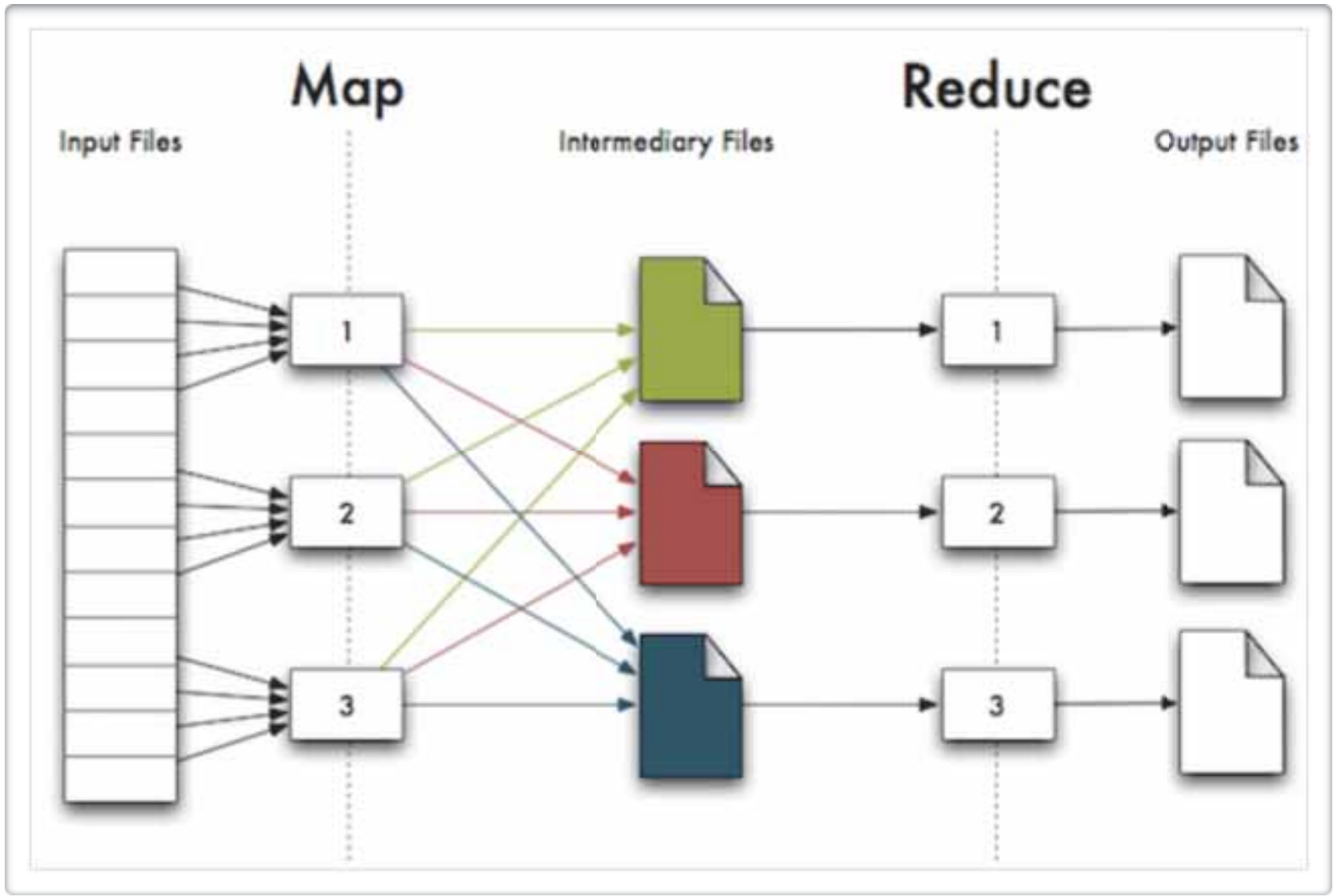
BigData

MapReduce

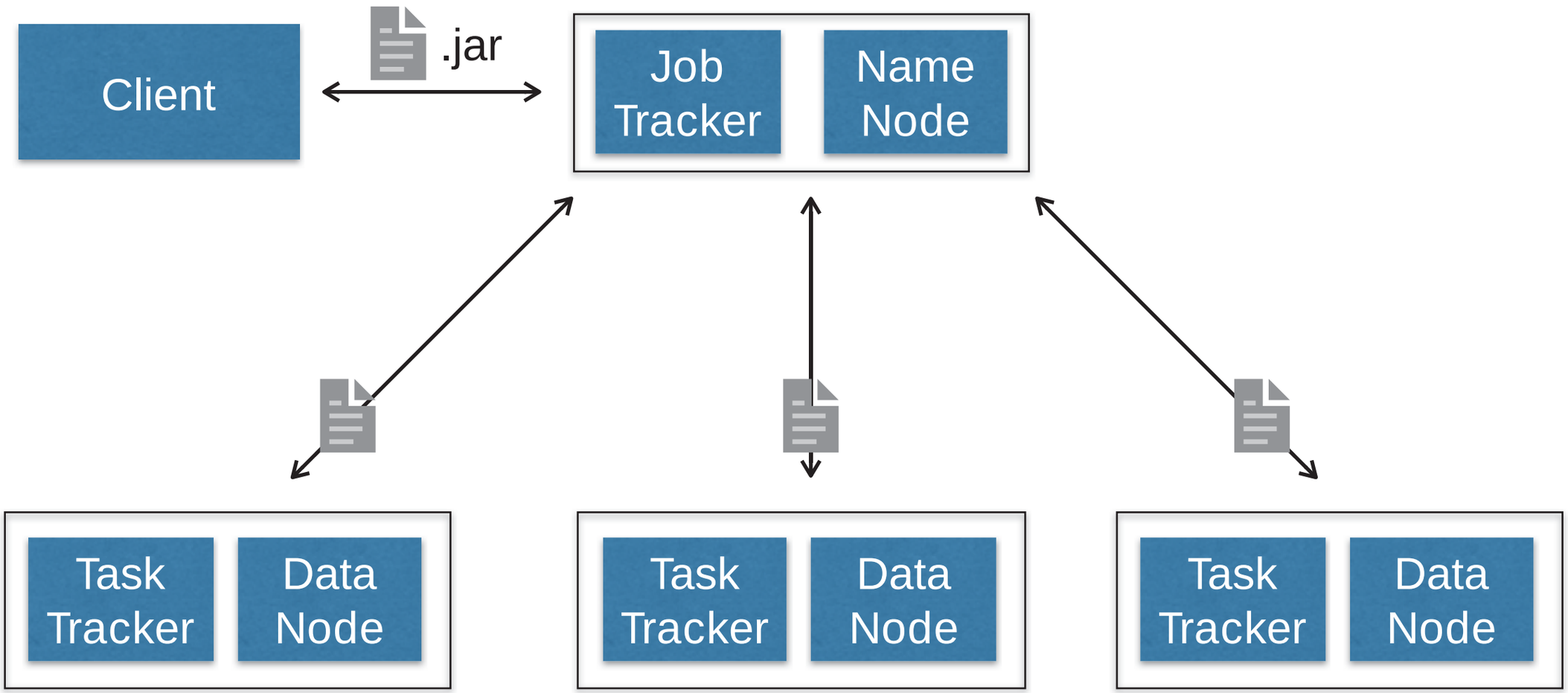
What Is MapReduce ?

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

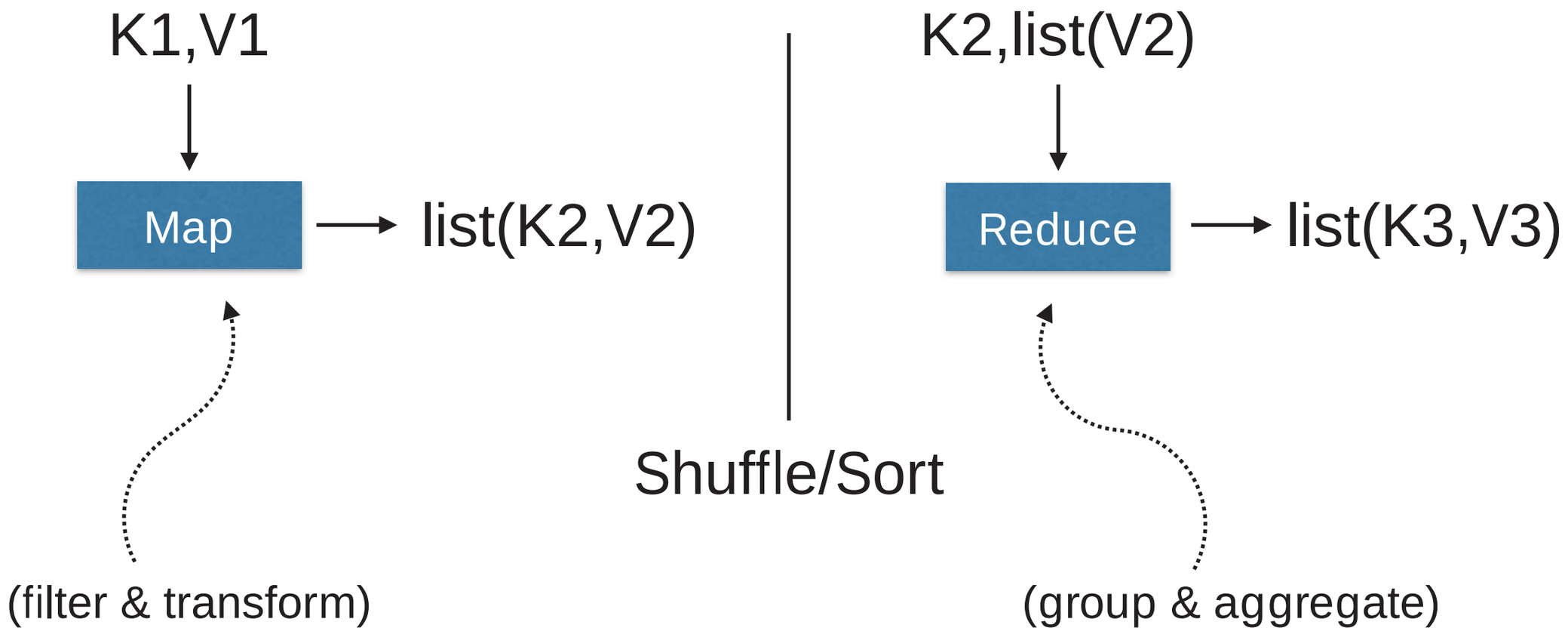
MapReduce


Map
Reduce
Input Files
Intermediary Files
Output Files
1
2
3
1
2
3

MapReduce & HDFS



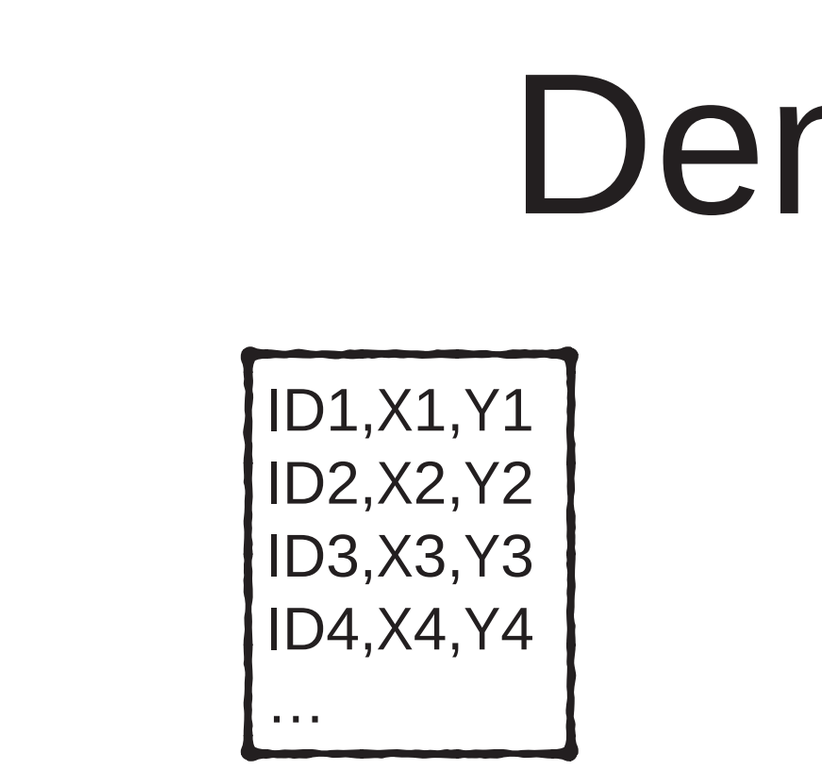
Thinking In MR

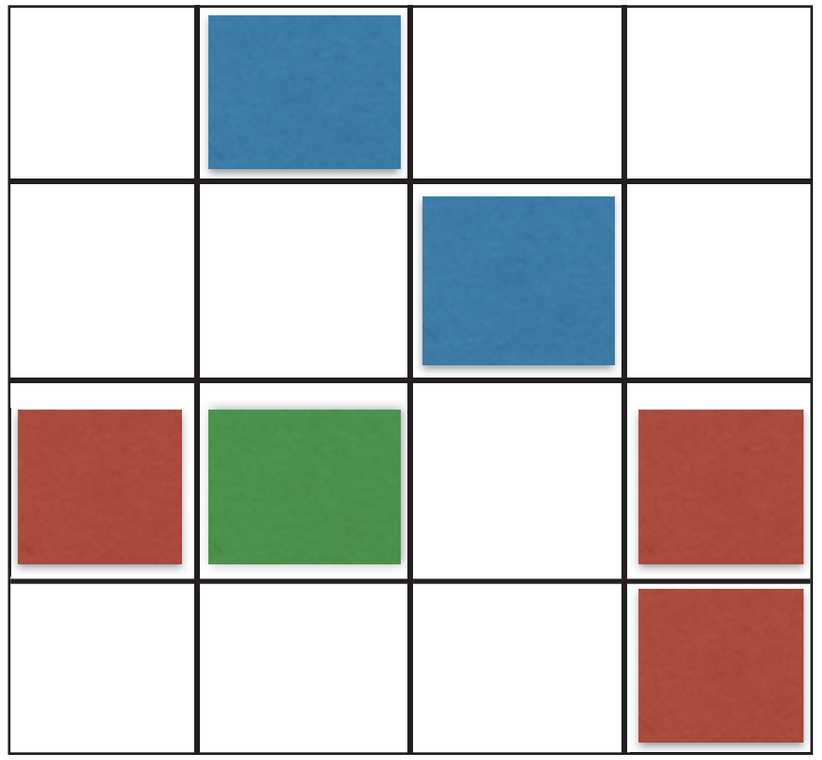


Hello MapReduce !

DensityMap

ID1,X1,Y1
ID2,X2,Y2
ID3,X3,Y3
ID4,X4,Y4
…





DensityMap

```
function map(lineno,text)
{
   tokens = text.split(‘,’)
   cell = toCell(tokens[1],tokens[2])
   emit( cell, 1)
}

function toCell(x,y)
{
  // some math !!
  return cell
}
```

```
function reduce(cell,iterator)
{
  sum = 0
  for( one : iterator)
    sum += one
  emit( cell, sum)
}
```

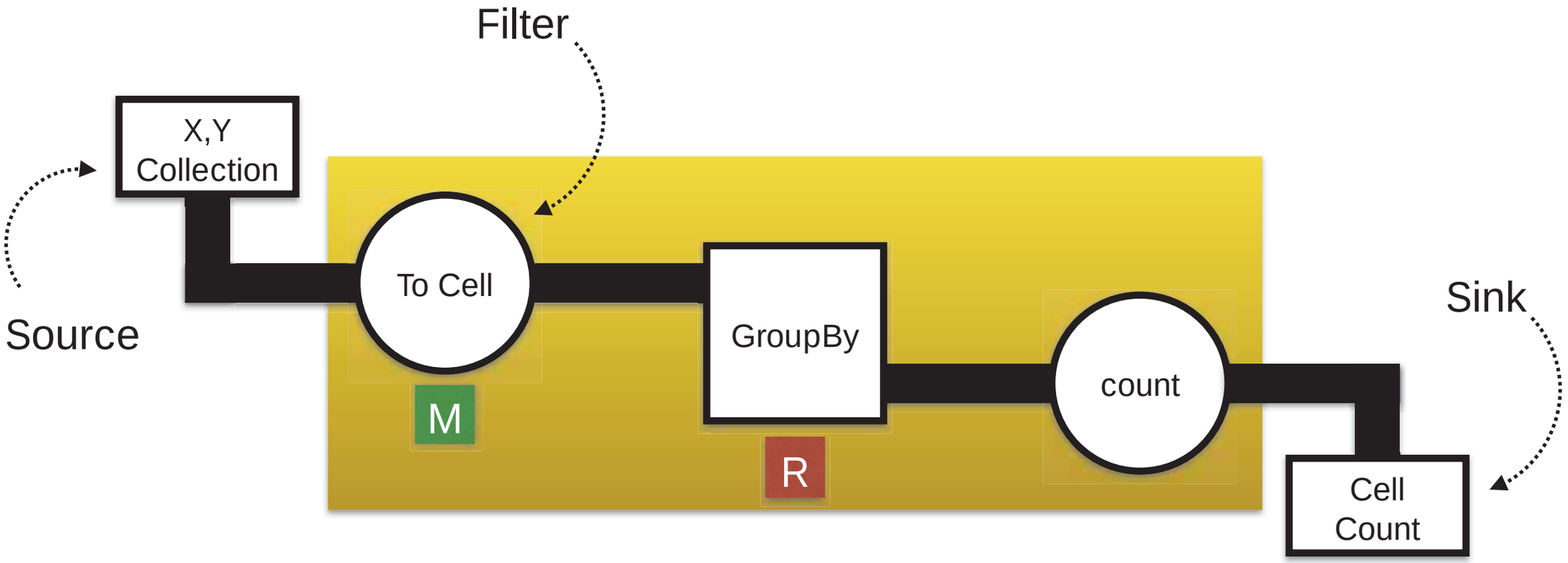
http://thunderheadxpler.blogspot.com/2013/03/bigdata-kernel-density-analysis-on.html

Writing MR Is Hard…

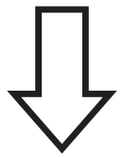
http://www.cascading.org

Think of Data as Water In Pipes

Workflow Pipeline



Cascading pipeline



MapReduce Job

Cascading Pipe

// Pipe tap x,y input fields into spatial function
Pipe pipe = new Each("start", new Fields("X", "Y"), new SpatialDensity());
// Group by emitted ‘cell’ value
pipe = new GroupBy(pipe, “cell”);
// Count by group and name count ‘POPULATION’
pipe = new Every(pipe, Fields.GROUP, new Count(new Fields("POPULATION")));

http://thunderheadxpler.blogspot.com/2014/01/cascading-workflow-for-spatial-binning.html

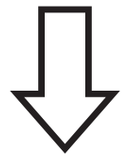
How About….

No Programing ???

Apache 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/';

HQL

```
$ hive
hive> select hour,count(hour)
from logs
where year=2014
and month=01
and uri = “http://mybog.com/mypage.html”
group by hour
order by hour;
```

Other AdHoc Engines

- Cloudera Impala
- Facebook Presto
- Amplab Shark
- 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

- Open Source / Github
- Apache 2.0 License
- Geometry API
- Spatial Framework Hive
- GeoProcessing Tools

Geometry API

- Shapes
- Points
- Polylines
- Polygons
- Envelopes

Geometry API

- Geometry Operations
- Contains / Intersects / …
- Union / Difference / …
- Buffer / ConvexHull
- Spatial Index

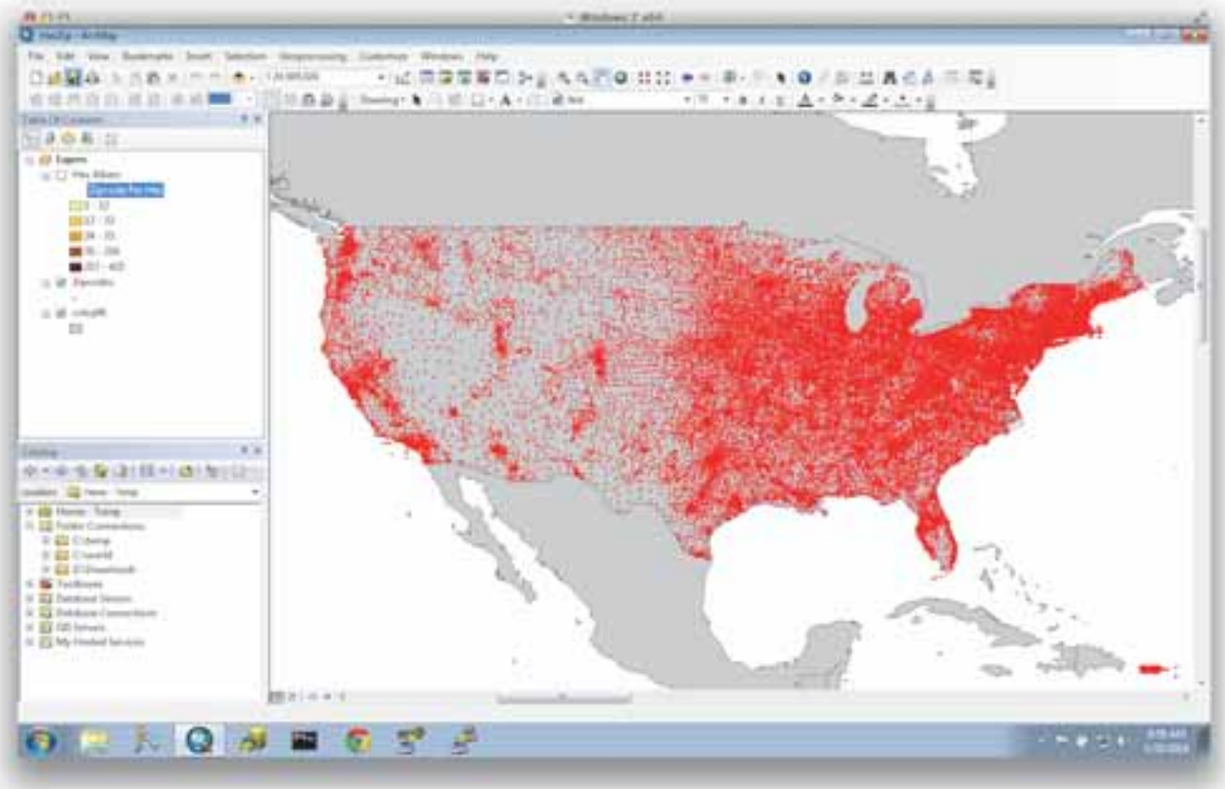
Geometry API

- I/O Operations
- WKT
- OGC
- GeoJSON
- Shape (bin)

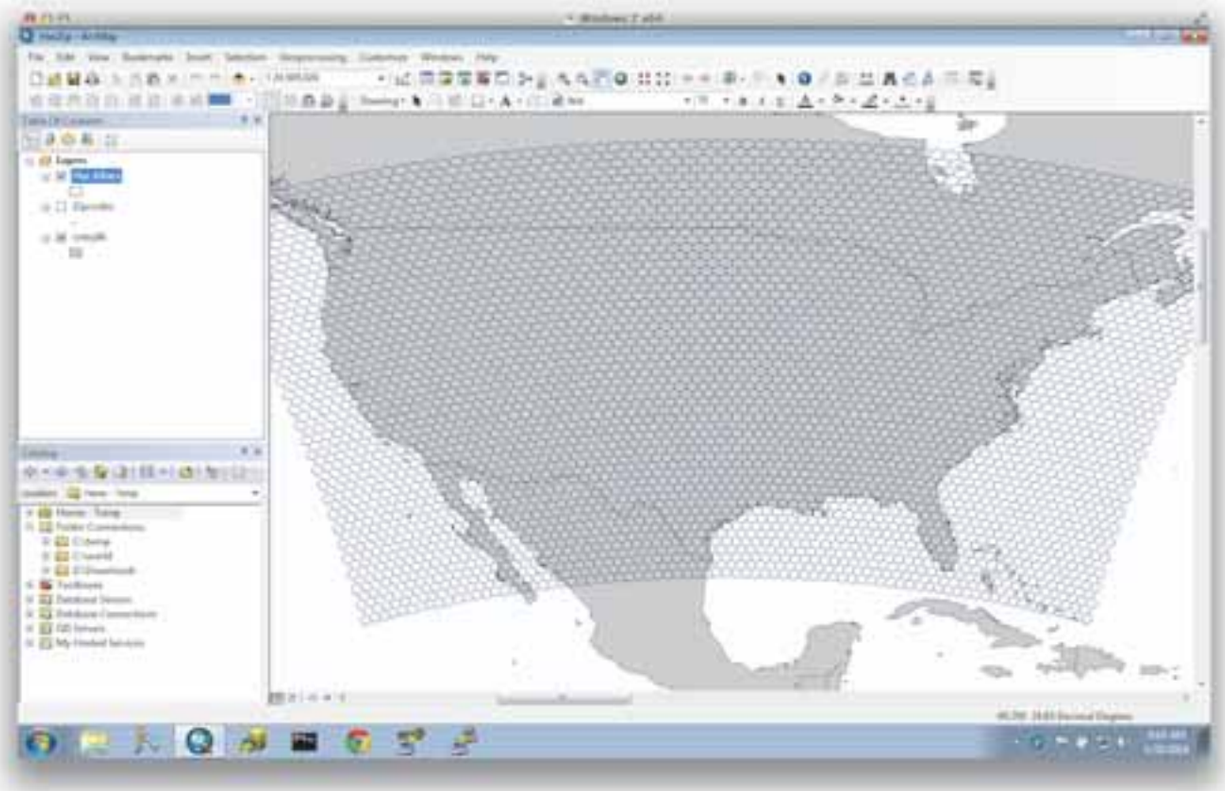
API Usage in BigData

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

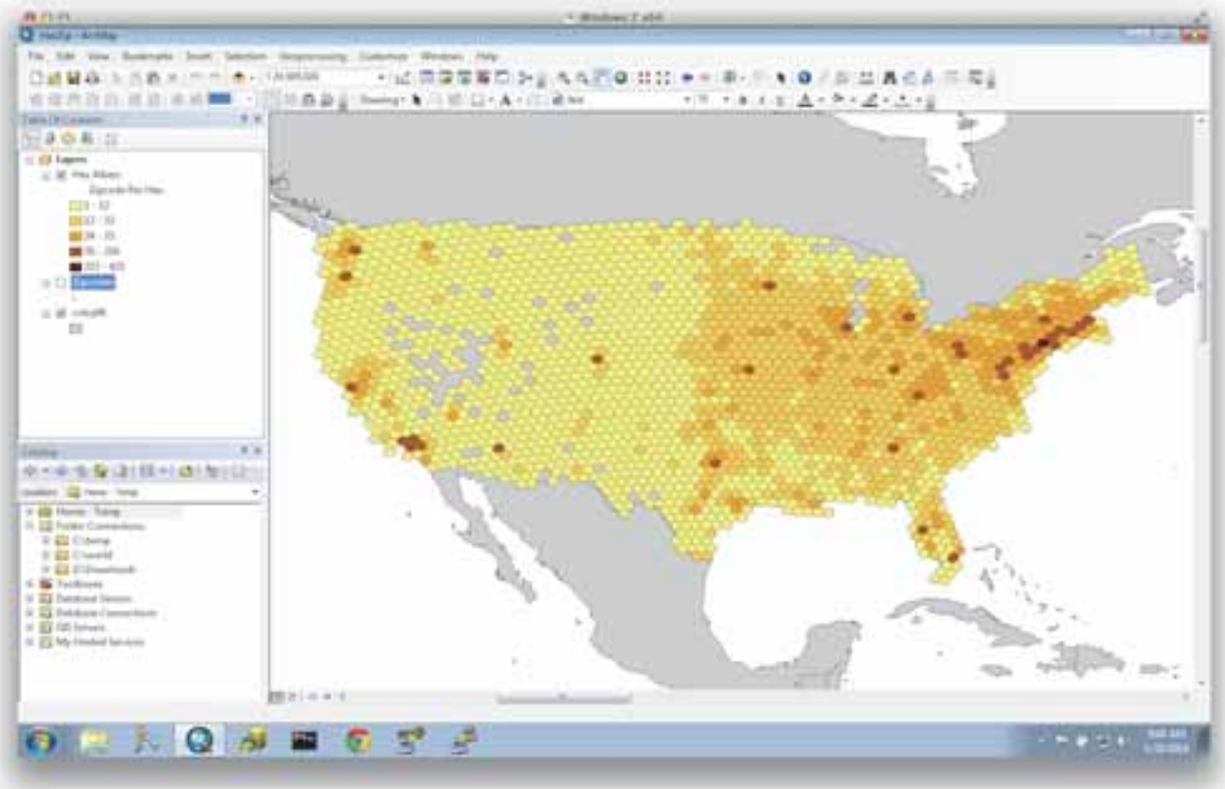
BigData Binning



BigData Binning



BigData Binning



- BigData Spatial Join
- kNN
- Range Queries
- Custom Input Format

SQL Is Still King !

Spatial Hive UDF

(User Defined Functions)

Spatial Hive UDF

- Uses Geometry API
- Constructor
  - ST_POINT / ST_GeomFromGeoJSON
- Relations
  - ST_Contains / ST_Buffer
- Accessor
  - ST_Distance, ST_Area

Spatial Hive

```
SELECT counties.name, count(*) cnt
FROM counties
JOIN earthquakes
WHERE ST_Contains(counties.boundaryshape,
ST_Point(earthquakes.longitude, earthquakes.latitude))
GROUP BY counties.name
ORDER BY cnt desc;
```

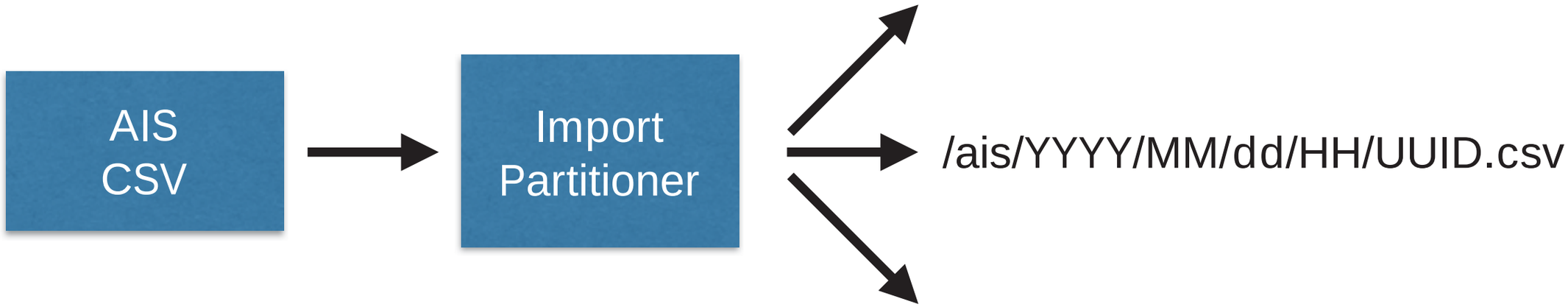
AIS Demo

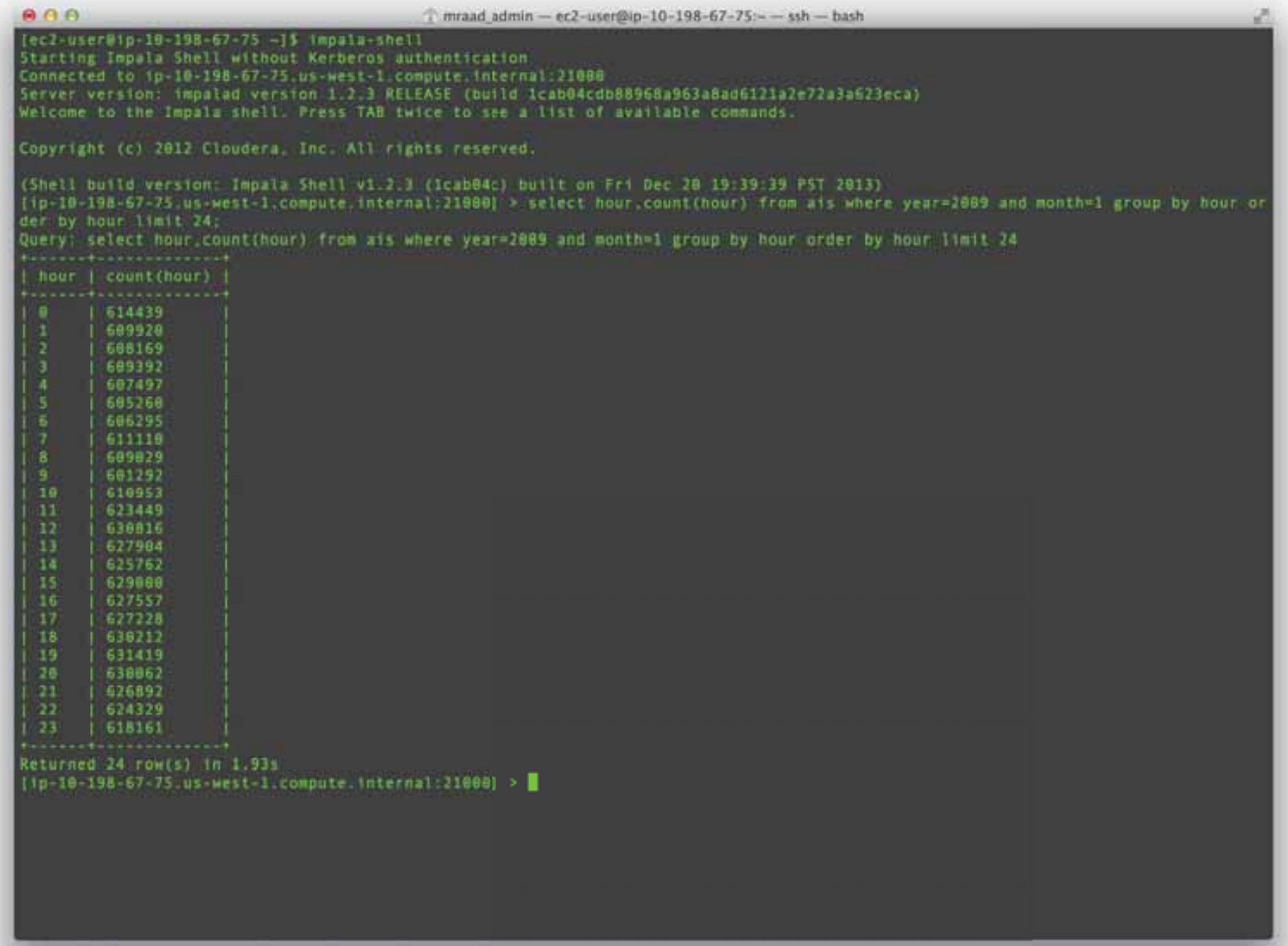
AIS Data

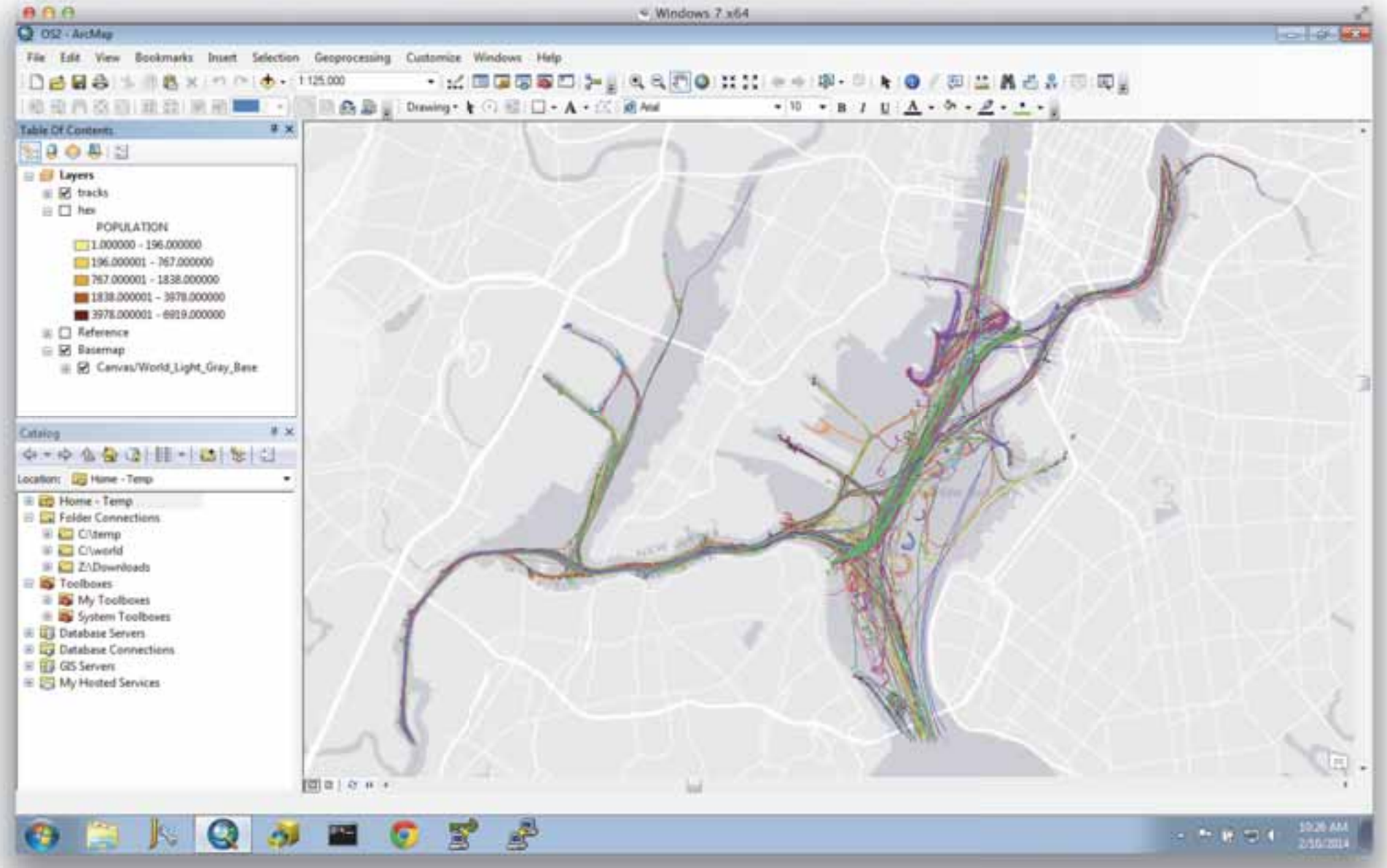
- 14.8 million GPS information
  - 1 month
  - Zone 18 (North East / NY Area)
- MMSI, ZuluTime, Lon/Lat, VoyageId, Draught

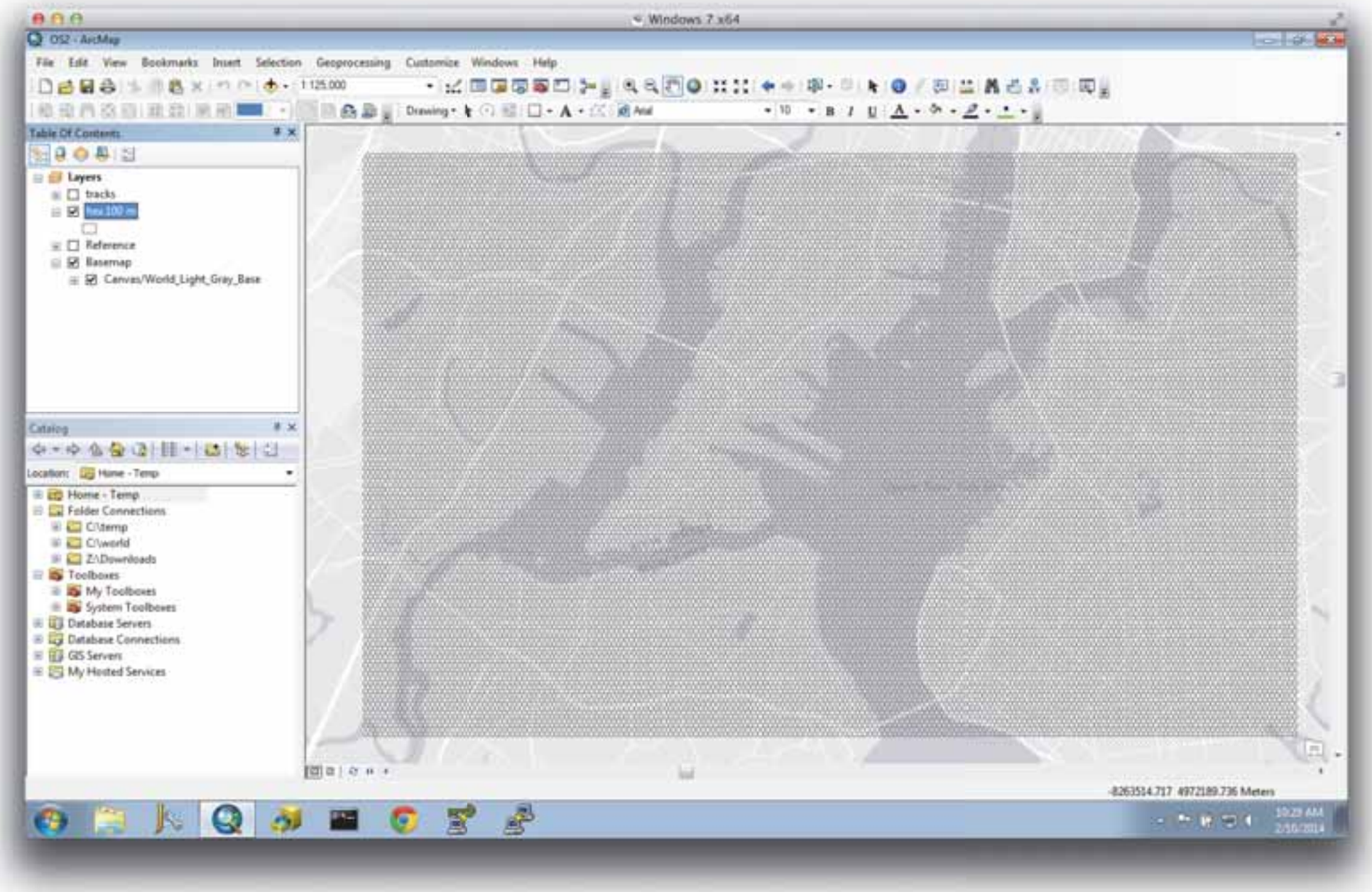
Demo Steps

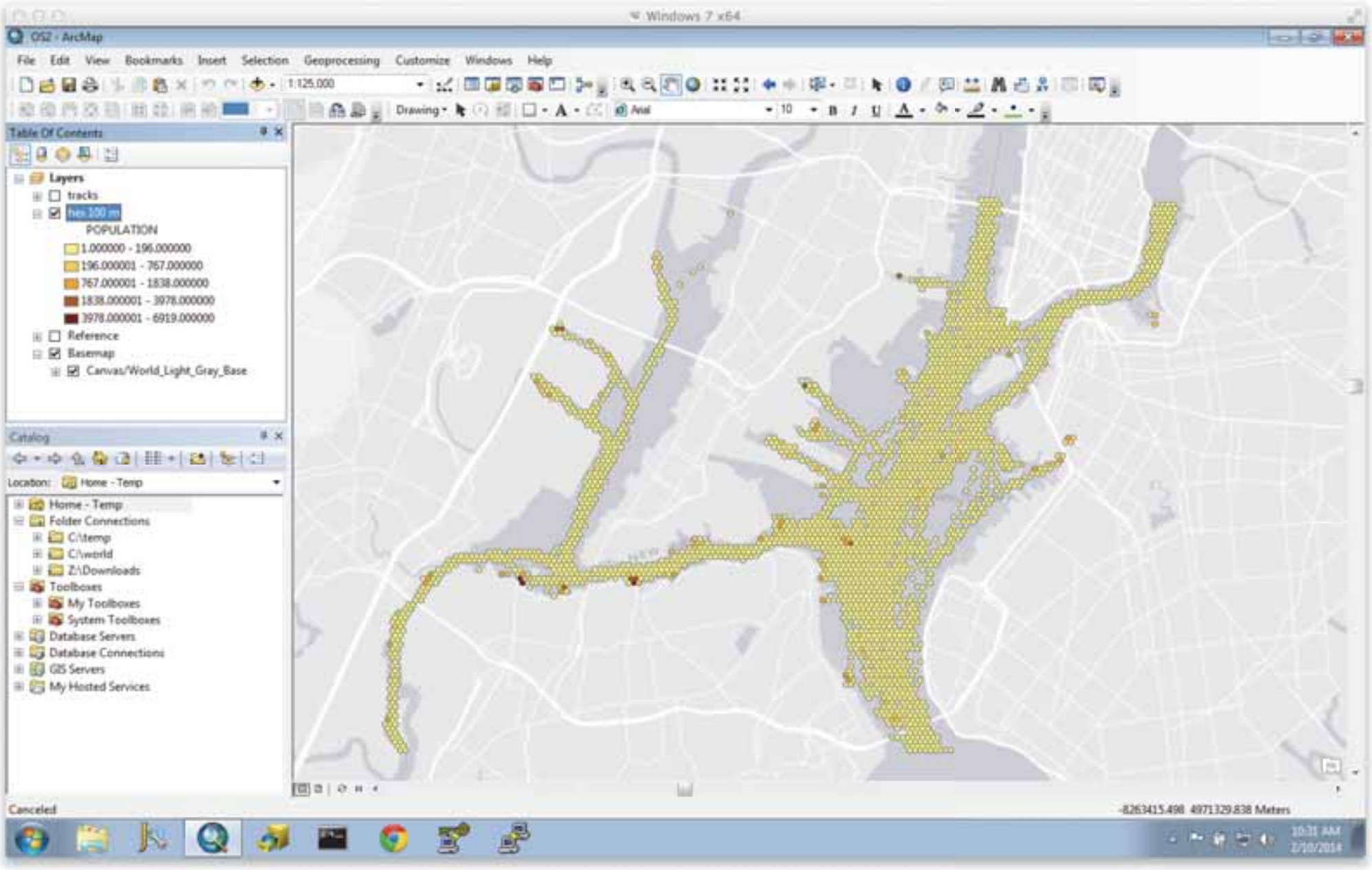
- GP Toolbox
- Track Assembly From Targets
- Hex Generation
- Density Analysis


AIS
CSV
Import
Partitioner
/ais/YYYY/MM/dd/HH/UUID.csv
HDFS
MapReduce


mraad_admin — ec2-user@ip-10-198-67-75:~ — ssh — bash
[ec2-user@ip-10-198-67-75 ~]$ impala-shell
Starting Impala Shell without Kerberos authentication
Connected to ip-10-198-67-75.us-west-1.compute.internal:21000
Server version: impalad version 1.2.3 RELEASE (build 1cab04cdb88968a963a8ad6121a2e72a3a623eca)
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 v1.2.3 (1cab04c) built on Fri Dec 20 19:39:39 PST 2013)
[ip-10-198-67-75.us-west-1.compute.internal:21000] > select hour,count(hour) from ais where year=2009 and month=1 group by hour order by hour limit 24;
Query: select hour,count(hour) from ais where year=2009 and month=1 group by hour order by hour limit 24
hour	count(hour)
0	614439
1	609920
2	608169
3	609392
4	607497
5	605260
6	606295
7	611110
8	609029
9	601292
10	610953
11	623449
12	630016
13	627904
14	625762
15	629000
16	627557
17	627228
18	630212
19	631419
20	630062
21	626892
22	624329
23	618161
Returned 24 row(s) in 1.93s
[ip-10-198-67-75.us-west-1.compute.internal:21000] >

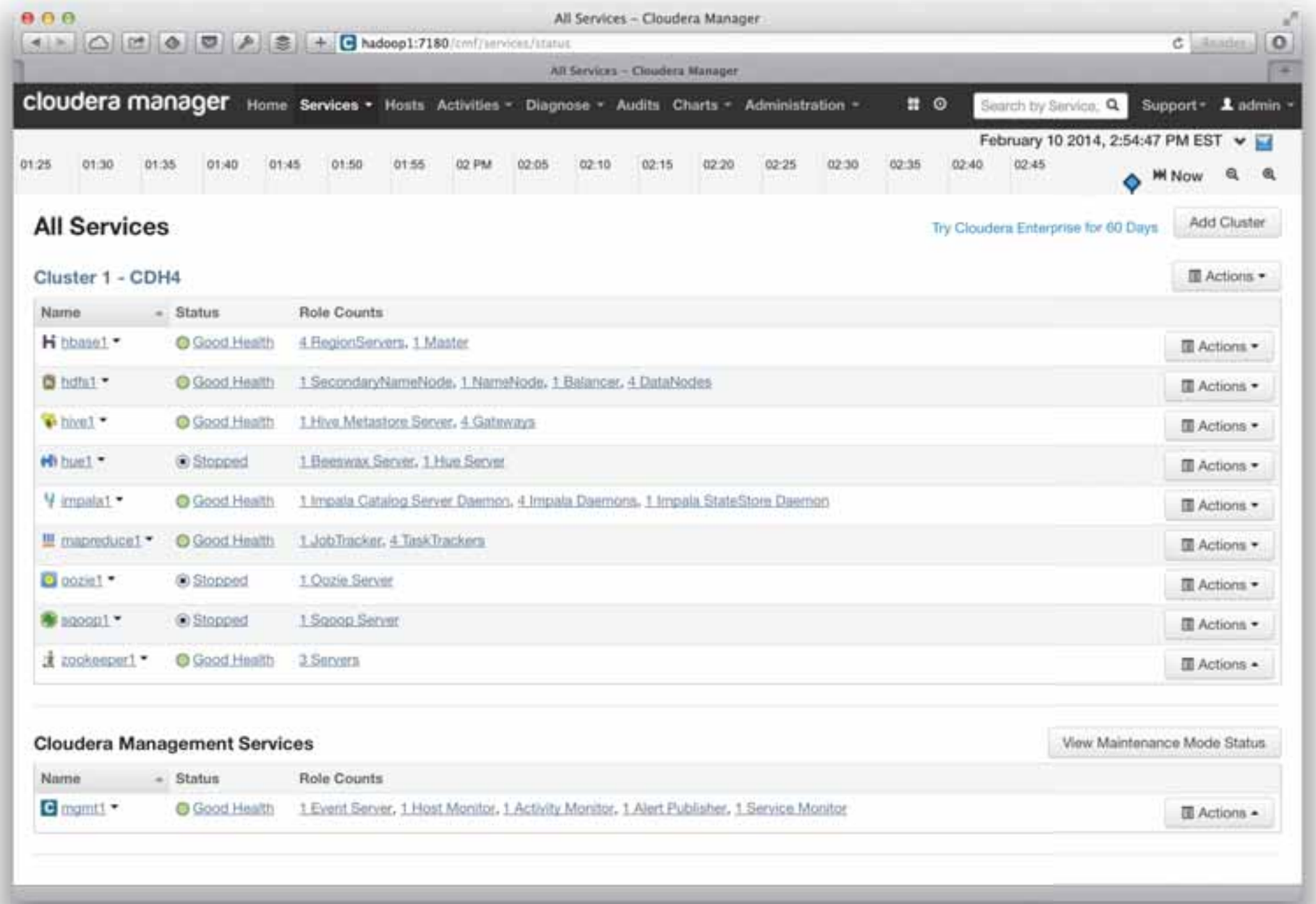

Windows 7 x64
OS2 - ArcMap
File Edit View Bookmarks Insert Selection Geoprocessing Customize Windows Help
1:125,000
Drawing
Arial
10
Table Of Contents
Layers
tracks
hex
POPULATION
1.000000 - 196.000000
196.000001 - 767.000000
767.000001 - 1838.000000
1838.000001 - 3978.000000
3978.000001 - 6919.000000
Reference
Basemap
Canvas/World_Light_Gray_Base
Catalog
Location: Home - Temp
Home - Temp
Folder Connections
C:\temp
C:\world
Z:\Downloads
Toolboxes
My Toolboxes
System Toolboxes
Database Servers
Database Connections
GIS Servers
My Hosted Services


Windows 7 x64
OS2 - ArcMap
File Edit View Bookmarks Insert Selection Geoprocessing Customize Windows Help
1:125,000
Drawing
Arial
10
Table Of Contents
Layers
tracks
hex 100 m
Reference
Basemap
Canvas/World_Light_Gray_Base
Catalog
Location: Home - Temp
Home - Temp
Folder Connections
C:\temp
C:\world
Z:\Downloads
Toolboxes
My Toolboxes
System Toolboxes
Database Servers
Database Connections
GIS Servers
My Hosted Services
-8263514.717 4972189.736 Meters
10:29 AM
2/16/2014


Windows 7 x64
OS2 - ArcMap
File Edit View Bookmarks Insert Selection Geoprocessing Customize Windows Help
1:125,000
Drawing
Arial
10
Table Of Contents
Layers
tracks
hex 100 m
POPULATION
1.000000 - 196.000000
196.000001 - 767.000000
767.000001 - 1838.000000
1838.000001 - 3978.000000
3978.000001 - 6919.000000
Reference
Basemap
Canvas/World_Light_Gray_Base
Catalog
Location: Home - Temp
Home - Temp
Folder Connections
C:\temp
C:\world
Z:\Downloads
Toolboxes
My Toolboxes
System Toolboxes
Database Servers
Database Connections
GIS Servers
My Hosted Services
Canceled
-8263415.498 4971329.838 Meters

How to get started ?

Cloudera QuickStart VM



Book That I Recommend

- Hadoop - The Definitive Guide
- Hadoop In Action
- HBase In Action
- Hadoop - Real World Solution Cookbook
- MapReduce Design Pattern

Q&A

http://thunderheadxpler.blogspot.com

mraad@esri.com

@mraad