Federal GIS Conference 2014

February 10–11, 2014 | Washington DC

BigData - The Practice

Mansour Raad

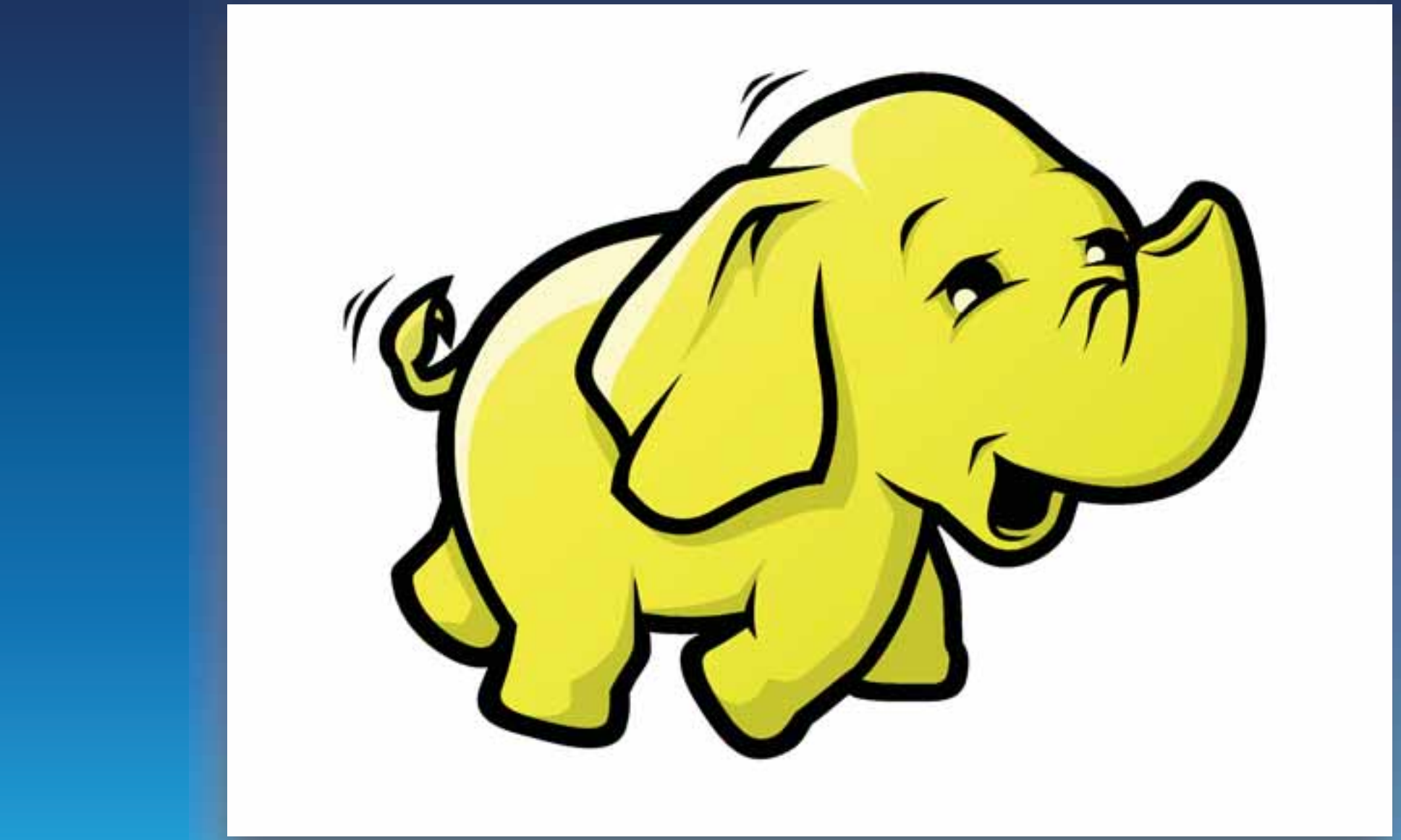
http://thunderheadxpler.blogspot.com/

mraad@esri.com

@mraad

On Today's Todo List:

- Run into store…
- Frantically ask "***What year is it ?***"
- When they reply
- Yell "**It Works, because of BigData !**"
- And run out



Hadoop Basic Stack

MapReduce

Yet Another Resource Negotiator (YARN)

Hadoop Distributed File System (HDFS)

The Zoo

- Hive - Ad Hoc Query - “SQL” to MapReduce
- Pig - High Level Data Analysis Language
- Impala - MPP SQL Engine
- Mahout - Machine Learning Toolbox
- HBase - Columnar KeyValue Database
- Cascading - Flow Data Analysis
- Avro - Data Serializer
- Zookeeper - Centralized State Management

GIS Tools for Hadoop

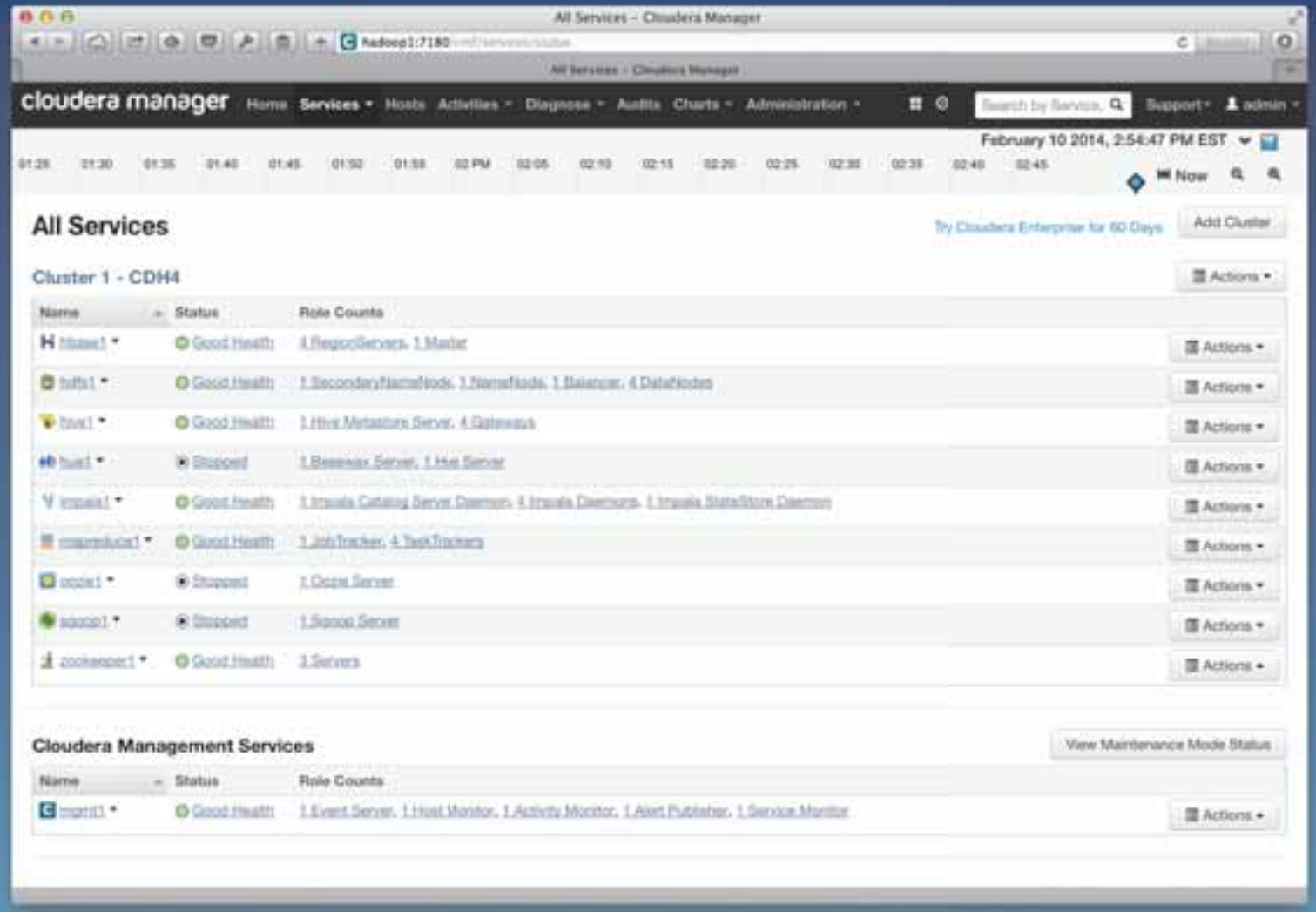
Big Data Spatial Analytics for the Hadoop Framework

View project on GitHub

GIS Tools For Hadoop

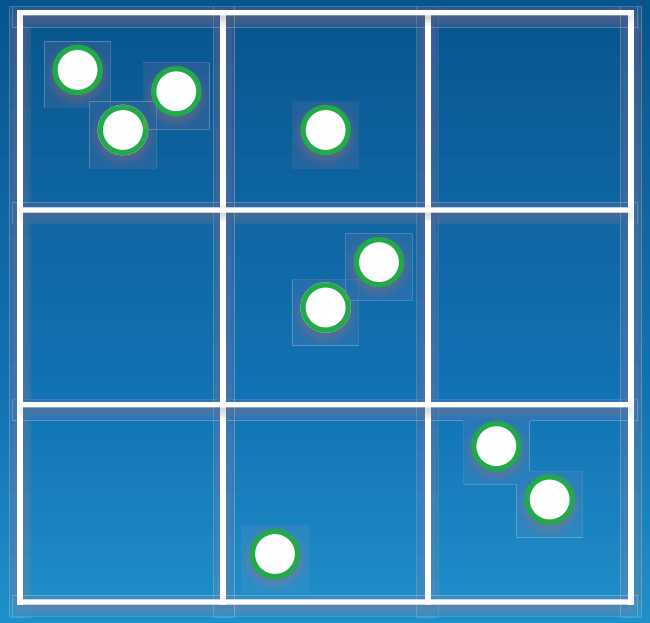
- Geometry API
  - Point / Line / Polygon
  - Operations - Contains, Intersect, Buffer
  - I/O - WKT, GeoJSON, Shape
- Hive Spatial UDF
  - ST_POINT, ST_CONTAINS
- GeoProcessing Extensions

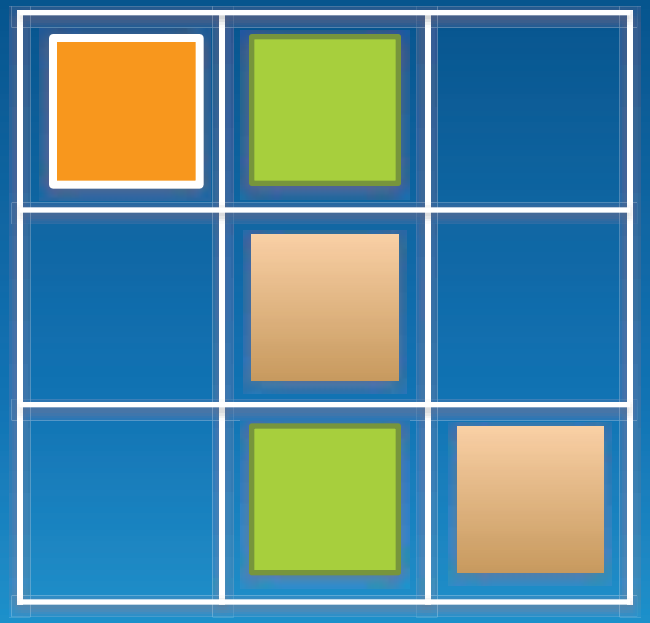
Cloudera Quick Start VM



Hello, MapReduce !

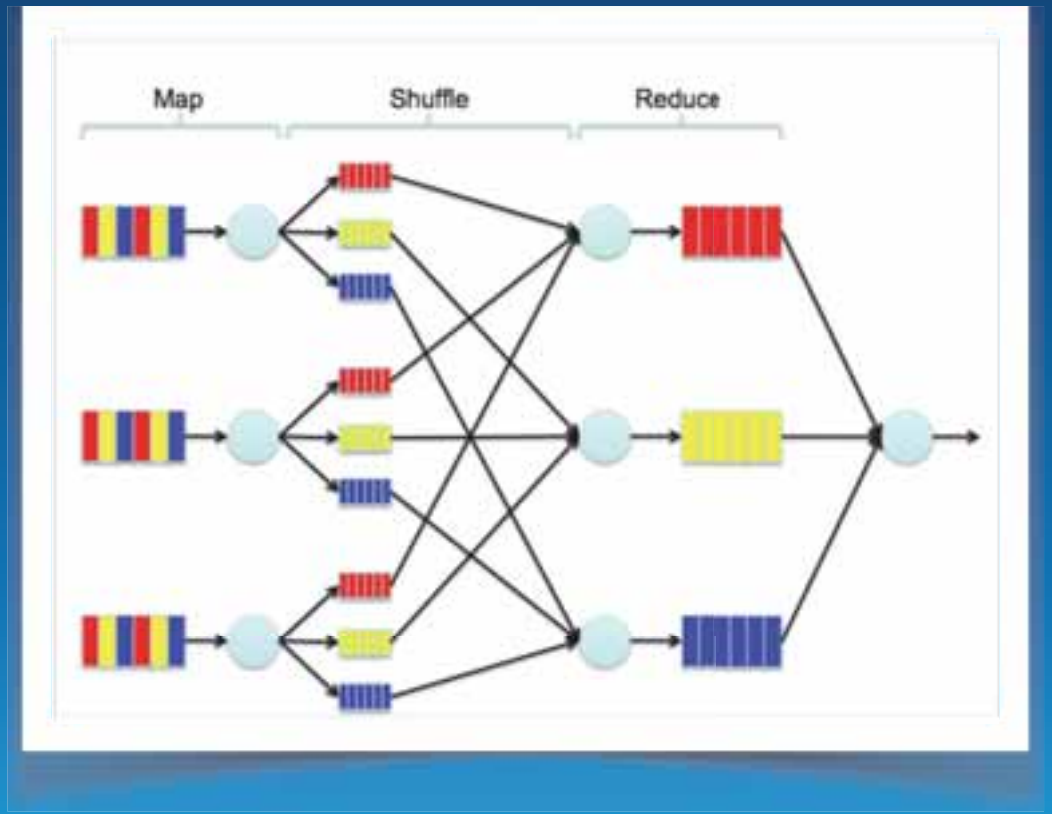
Density Analysis - Cell Count





MapReduce Recap

- Map
  - Extract
  - Filter
  - Transform
- Reduce
  - Group By
  - Aggregate



Cell Count

```
function map(lineno,text) {
    (x,y) = tokenize(text)
    if(inGrid(x,y)){
        (cellX,cellY) = toCell(x,y)
        emit((cellX,cellY),1)
    }
}
```

```
function reduce((cellX,cellY),iterator){
    sum = 0
    for( one in iterator){
        sum = sum + one
    }
    emit((cellX,cellY), sum)
}
```

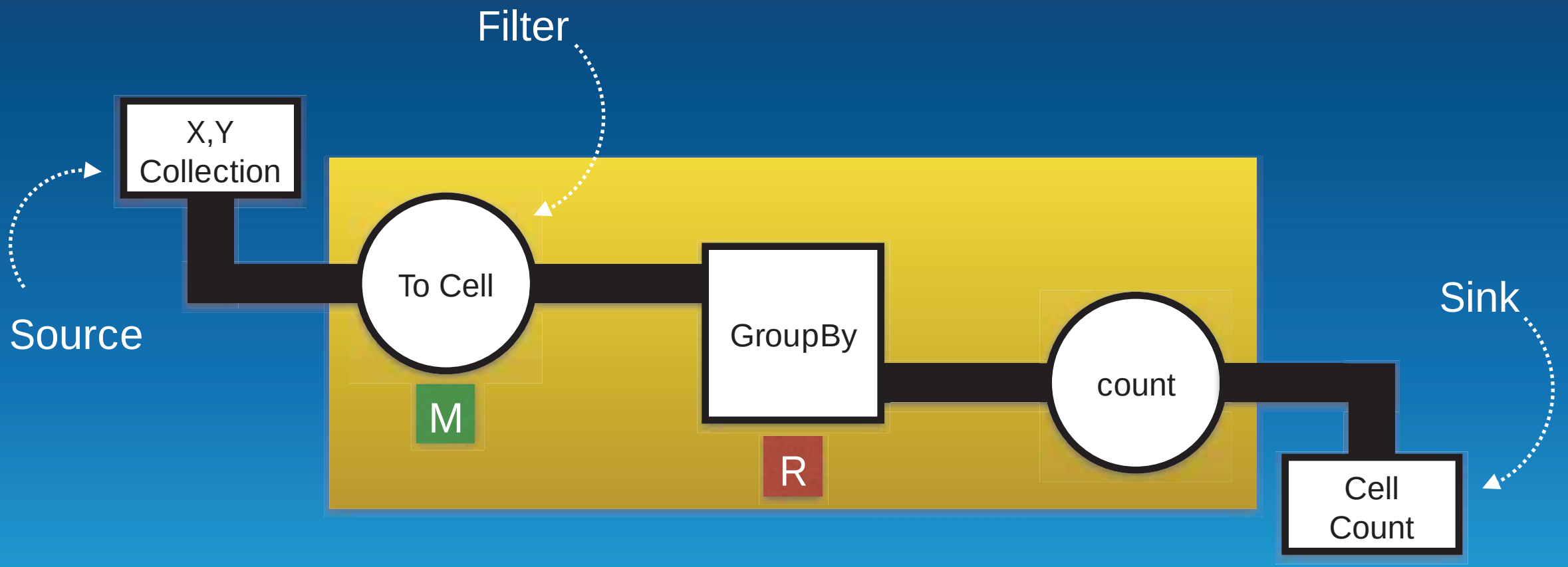
In Action Demo

MapReduce Is Hard…

Thinking Of Data As Water


cascading

Cascading Pipeline



```
33
34      final Fields inFields = new Fields("ID", "X", "Y").
35              applyTypes(long.class, double.class, double.class);
36
37      final Tap inTap = new Hfs(new TextDelimited(inFields, false, "\t"), args[0]);
38
39      final Tap outTap = new Hfs(new TextDelimited(true, ","), args[2], SinkMode.REPLACE);
40
41      final SpatialDensity spatialDensity = new SpatialDensity();
42
43      Pipe pipe = new Each("start", new Fields("X", "Y"), spatialDensity);
44
45      pipe = new GroupBy(pipe, spatialDensity.getFieldDeclaration());
46
47      pipe = new Every(pipe, Fields.GROUP, new Count(new Fields("POPULATION")));
48
49      final Properties properties = AppProps.appProps().
50              setJarClass(App.class).
51              buildProperties();
52
53      properties.put(SpatialDensity.KEY_SHP, args[1]);
54
55      final FlowConnector connector = new HadoopFlowConnector(properties);
56
57      final Flow flow = connector.connect(inTap, outTap, pipe);
58
59      flow.complete();
60  }
```

Cascading In Action

How About No Programming ?
What About SQL ?

Hive and Impala

```
drop table if exists zipcodes;

create external table if not exists zipcodes(
id int,
lon double,
lat double
) row format delimited
fields terminated by '\t'
lines terminated by '\n'
stored as textfile
location '/user/cloudera/zipcodes';
```

Cell Density in SQL

```
SELECT T.X-180+0.5 AS LON,T.Y-90+0.5 AS LAT,COUNT(*) AS POPULATION FROM (
SELECT FLOOR(LON+180) AS X,FLOOR(LAT+90) AS Y FROM ZIPCODES) T
GROUP BY T.X,T.Y;
```

Hive and Impala In Action

In Memory Spatial Index

In Memory Spatial Index

- Geometry API in GIS Tools For Hadoop
- new SpatialIndex(new Envelope2D(), depth);
- insert(new Envelope2D(), id)
- iterator = query(new Envelope2D())
- Use in mapper in “small” spatial joins

Spatial Index In Action

ArcGIS Desktop and Hadoop

esri

Understanding our world.

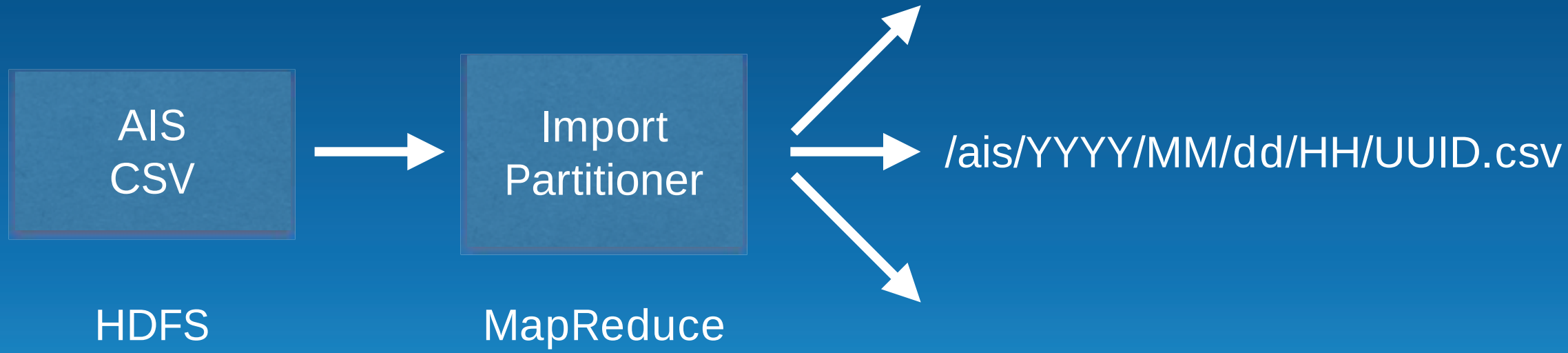
AIS DATA

- 14.8 Million data points
- 1 Month
- Zone 18 (North East / NY Area)
- MMSI, Zulu Time, Lat, Lon, Vessel ID, Draught

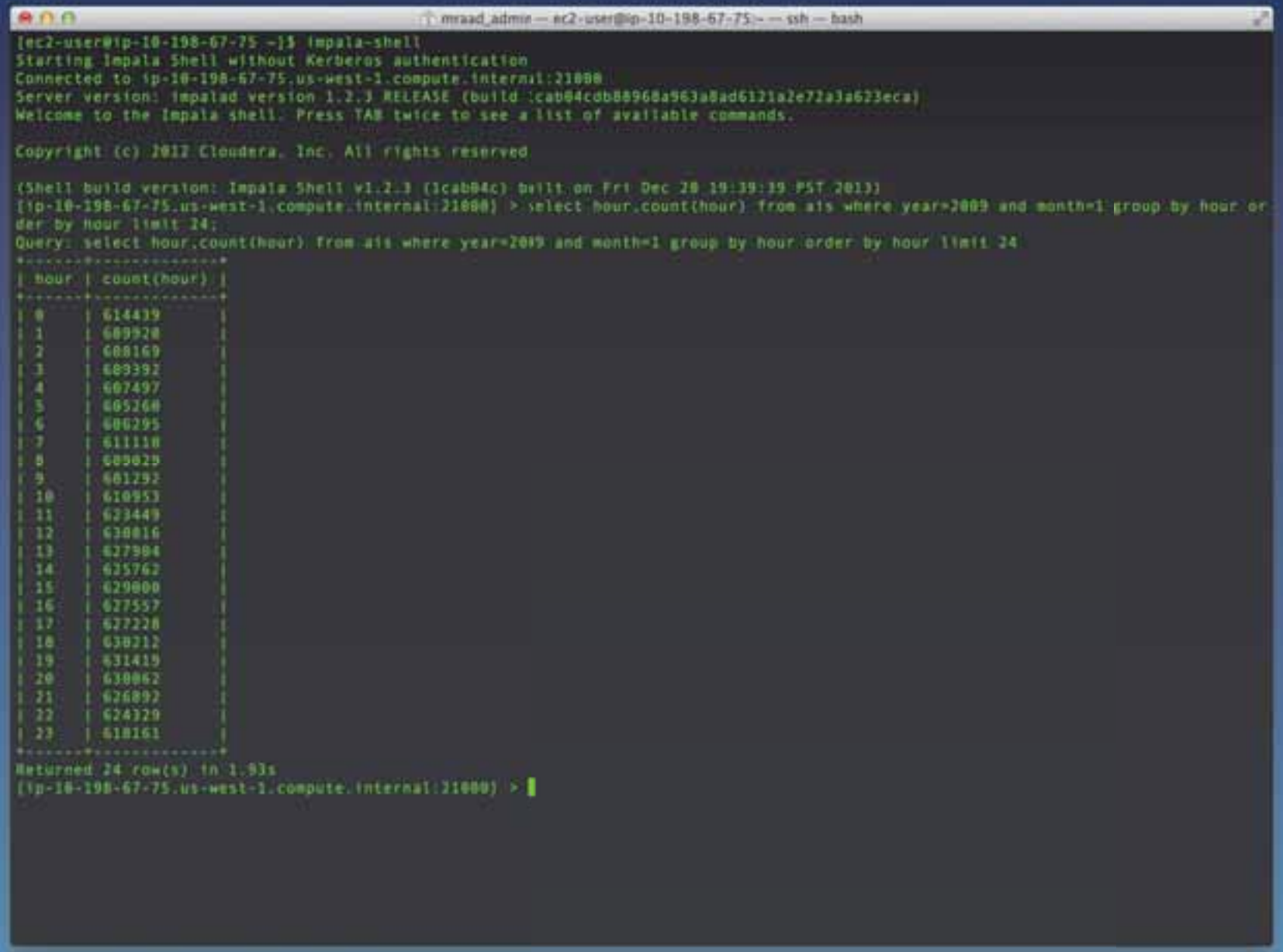
DEMO Steps

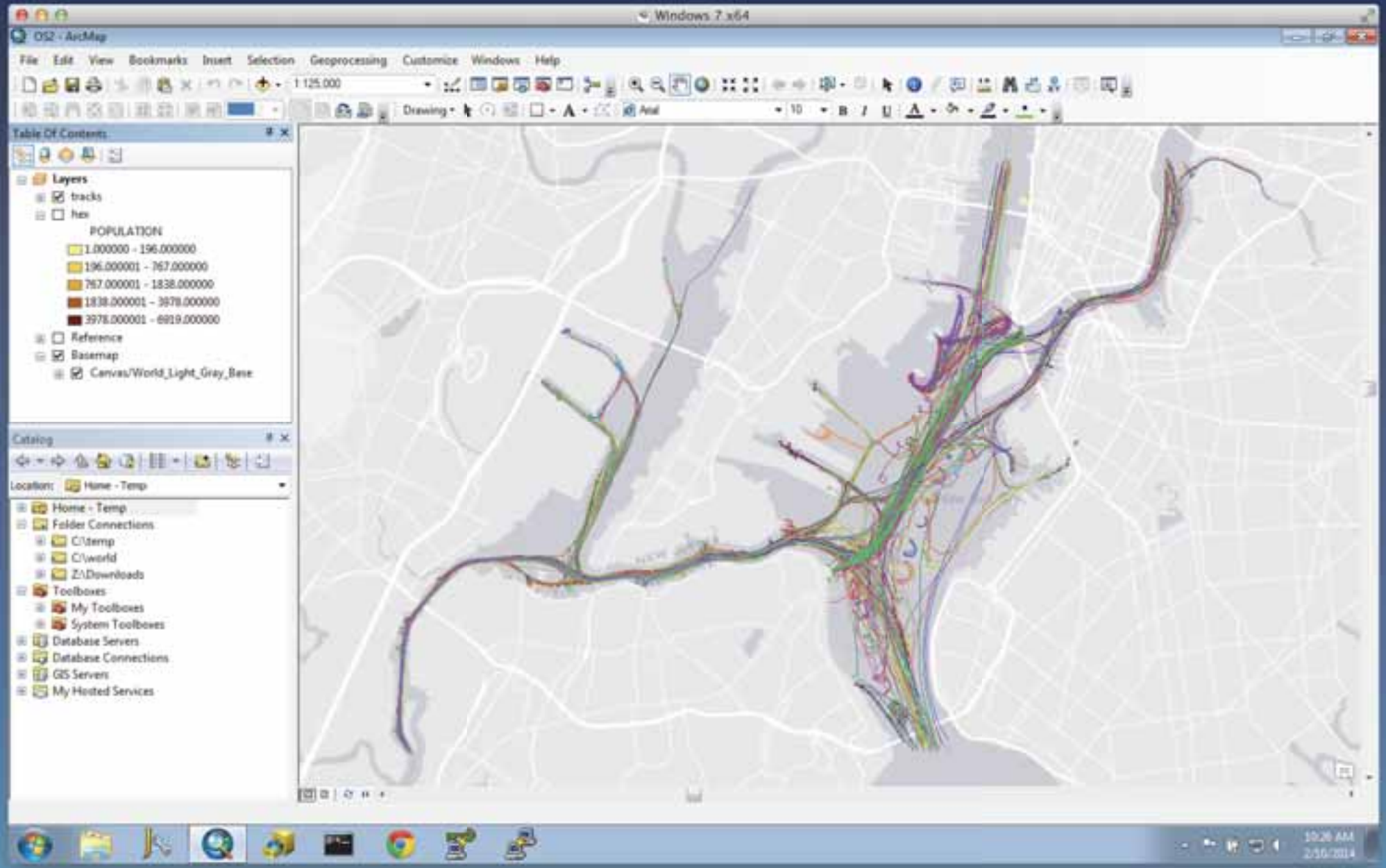
- GP Toolbox
- Track Assembly
- Hex Generation
- Density Analysis

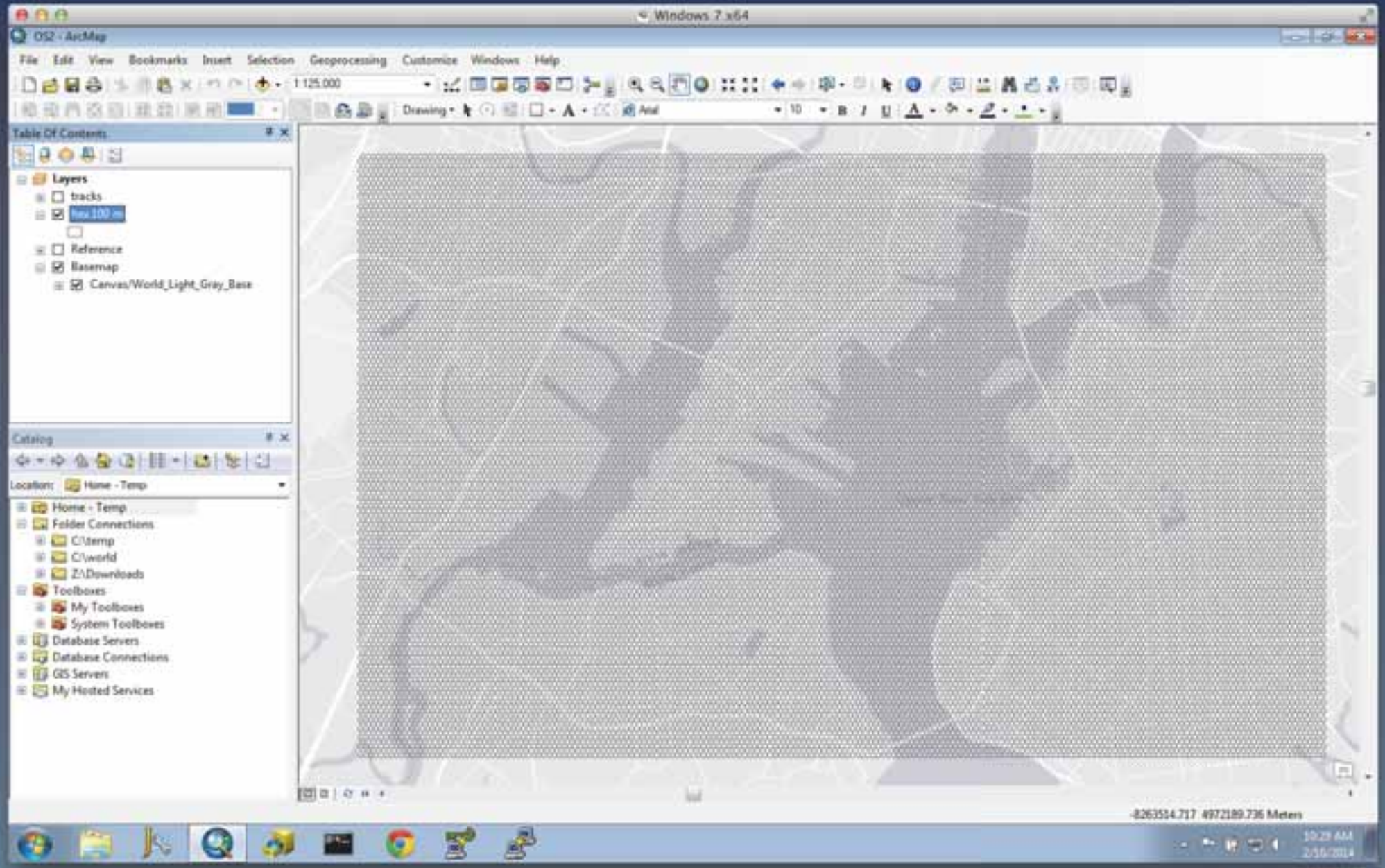
Import Job

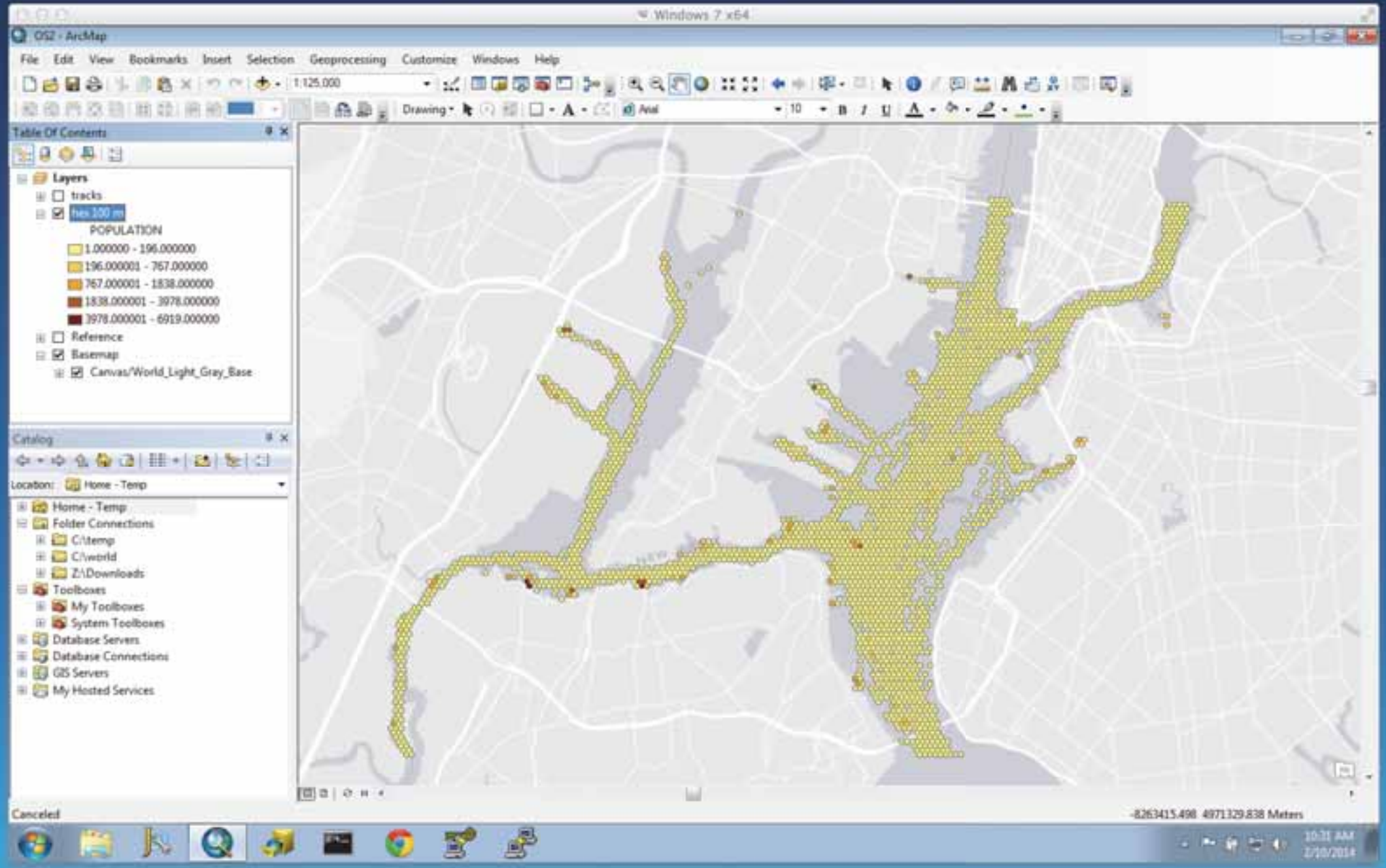

AIS
CSV
Import
Partitioner
/ais/YYYY/MM/dd/HH/UUID.csv
HDFS
MapReduce

```
 */
final class ImportReduce
        extends Reducer<DateHour, Text, NullWritable, Text>
{
    private MultipleOutputs<NullWritable, Text> m_mos;

    @Override
    protected void setup(final Context context) throws IOException, InterruptedException
    {
        m_mos = new MultipleOutputs<NullWritable, Text>(context);
    }

    @Override
    protected void reduce(
            final DateHour key,
            final Iterable<Text> values,
            final Context context) throws IOException, InterruptedException
    {
        final String outputPath = String.format("/por/%4d/%02d/%02d/%02d/%s",
                key.yy, key.mm, key.dd, key.hh, UUID.randomUUID().toString());
        for (final Text value : values)
        {
            m_mos.write(NullWritable.get(), value, outputPath);
        }
    }

    @Override
    protected void cleanup(final Context context) throws IOException, InterruptedException
    {
        m_mos.close();
    }
}
```


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	609528
2	608169
3	609392
4	607497
5	605260
6	606295
7	611110
8	609029
9	601292
10	610953
11	623449
12	630816
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
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
New 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


Windows 7 x64
OS2 - ArcMap
File Edit View Bookmarks Insert Selection Geoprocessing Customize Windows Help
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

Q&A

http://thunderheadxpler.blogspot.com

mraad@esri.com

@mraad