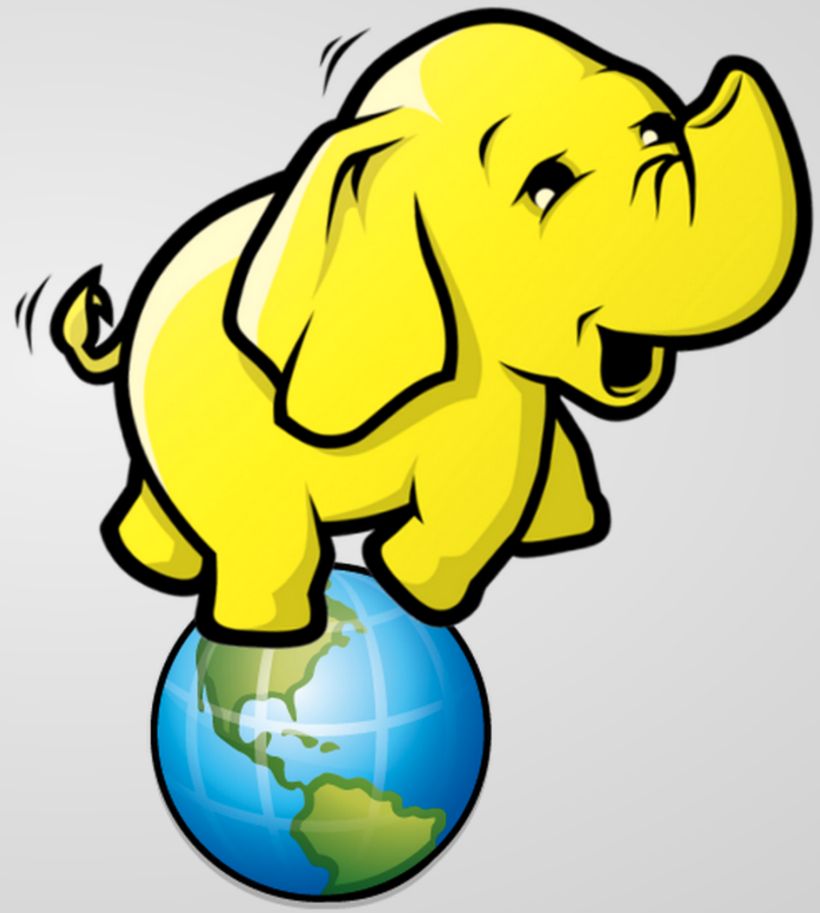
Big Data: Using ArcGIS with Apache Hadoop



David Kaiser @ddkaiser
Michael Park

Esri DevSummit 2013

Session Offering ID: 301

Follow along with this presentation

http://esri.github.com/gis-tools-for-hadoop/devsummit2013

Use WiFi network: Esri2013

Big Data: Using ArcGIS with Apache Hadoop

GIS To

Hadoop us
with limite

Esri has re
spatial-da

This enabl
that use bo

Why

ntation

nmit2013

e Hadoop

GIS Tools for Hadoop

Hadoop users often have data with spatial value, but with limited options for spatially analyzing this data

Esri has released an open-source framework to enable spatial-data processing in your Hadoop applications

This enables you, as a developer, to build analytical tools that use both Hadoop and ArcGIS.

Why is this important?

This enables you, as a developer, to build analytical tools that use both Hadoop and ArcGIS.

e Hadoop

Why is this important?

Your Hadoop applications can provide spatial analysis

...and your users can leverage your Hadoop applications from within the ArcGIS Geoprocessing environment

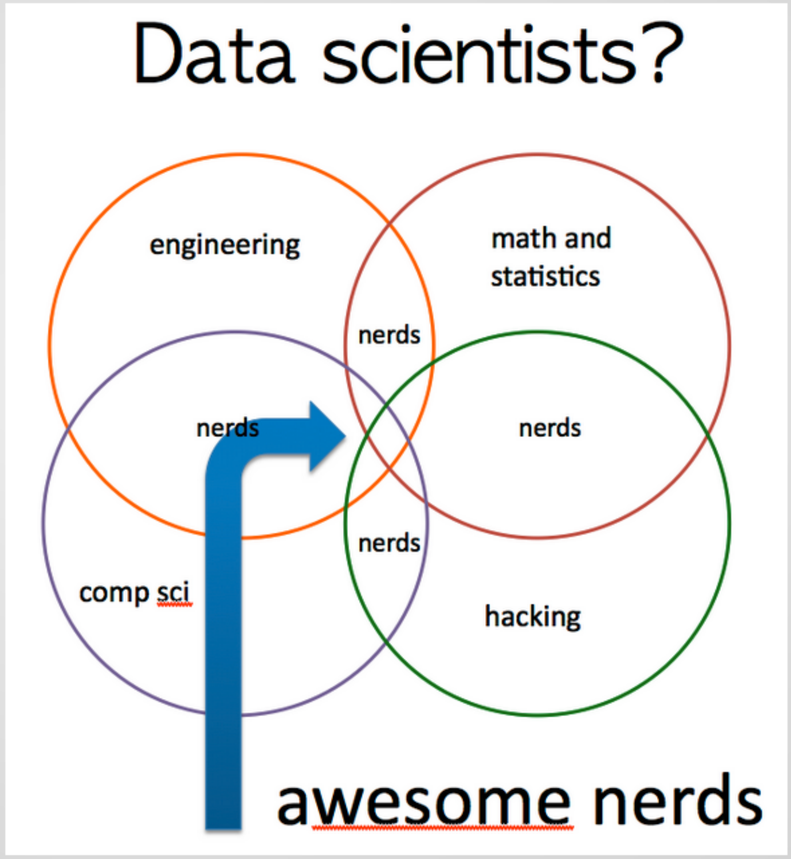
Who is a Data Scientist?

from within the ArcGIS Geoprocessing environment

Who is a Data Scientist?

Finding Your Data Scientist

Finding Your Data Scientist

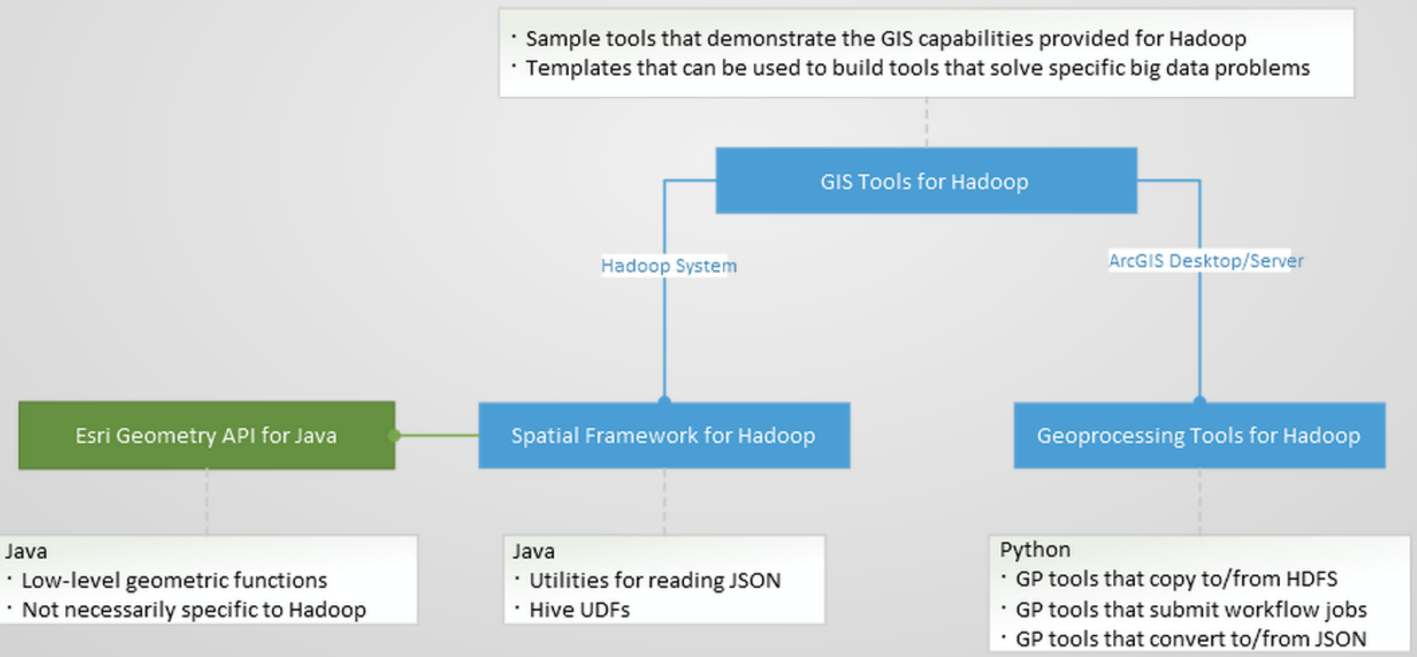


www.hilarymason.com

GIS Tools for Hadoop

The Hadoop 'Tools' are a combination of
custom Hadoop applications and ArcGIS GP Tools.

- Sample tools that demonstrate the GIS capabilities provided for Hadoop
- Templates that can be used to build tools that solve specific big data problems



E-mail:
David Kaiser <dkaiser@esri.com> @ddkaiser
Michael Park <mpark@esri.com>

- Send a Github 'pull request' so we can pull the tool back into our project

Let us know how you are doing
- Add your content to the Wikis on Github
- Troubles? Open new Issues in Github

http://github.com/Esri/spatial-framework-for-ha

Geoprocessing Tools for Hadoop
http://github.com/Esri/geoprocessing-tools-for-h

To build the source:
ant

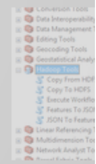
Download the GIS Tools Proje

Clone or Fork the project from Github
http://github.com/Esri/gis-tools-for-hado

Pre-built samples in the 'samples' director

Place your completed tools in the 'tools' directory if you want to share them

Geoprocessing Tools for Hadoop



Features To JSON, JSON To Features
- Provide serialization to and from JSON formats

Copy To HDFS, Copy From HDFS
- Moves files between ArcGIS and Hadoop

Execute Workflow
- Starts a workflow using the Hadoop Oozie workflow engine

Demo

Geometry API for Java

Simple API Functions for Java

com.esri.core.geometry.*

Relationship Analysis Operations

Spatial Framework for Hadoop

Enables developers to:
- spatially enable MapReduce applications

Enables Hadoop users to:

Developing Custom MapReduce Apps

en have data with spatial value, but
ons for spatially analyzing this data

an open-source framework to enable
essing in your Hadoop applications

as a developer, to build analytical tools
doop and ArcGIS.

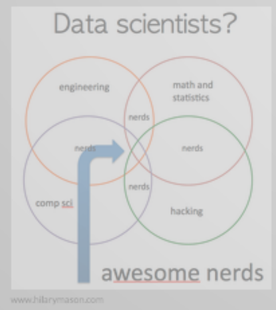
his important?

doop applications can
spatial analysis

an leverage your Hadoop applications
cGIS Geoprocessing environment

Data Scientist?

our Data Scientist



Geometry API for Java

Simple API Functions for Java

com.esri.core.geometry.*

Relationship Analysis

- equals
- disjoint
- touches
- crosses
- within
- contains
- overlaps

Operations

- buffer
- cut
- clip
- convexHull
- intersect
- union
- difference

Spatia

Enables
- spatial

Enables
- run spa

Provides
- JSON
- Hive U

Uses the

Spatial Framework for Hadoop

Enables developers to:

- spatially enable MapReduce applications

Enables Hadoop users to:

- run spatial Hive queries with ST_Geometry functions

Provides Java API's for:

- JSON Utility classes
- Hive UDF's

Uses the ***Geometry API for Java***

Uses the ***Geometry API for Java***

Spatial Data in Hadoop

JSON files store collections of 'features'

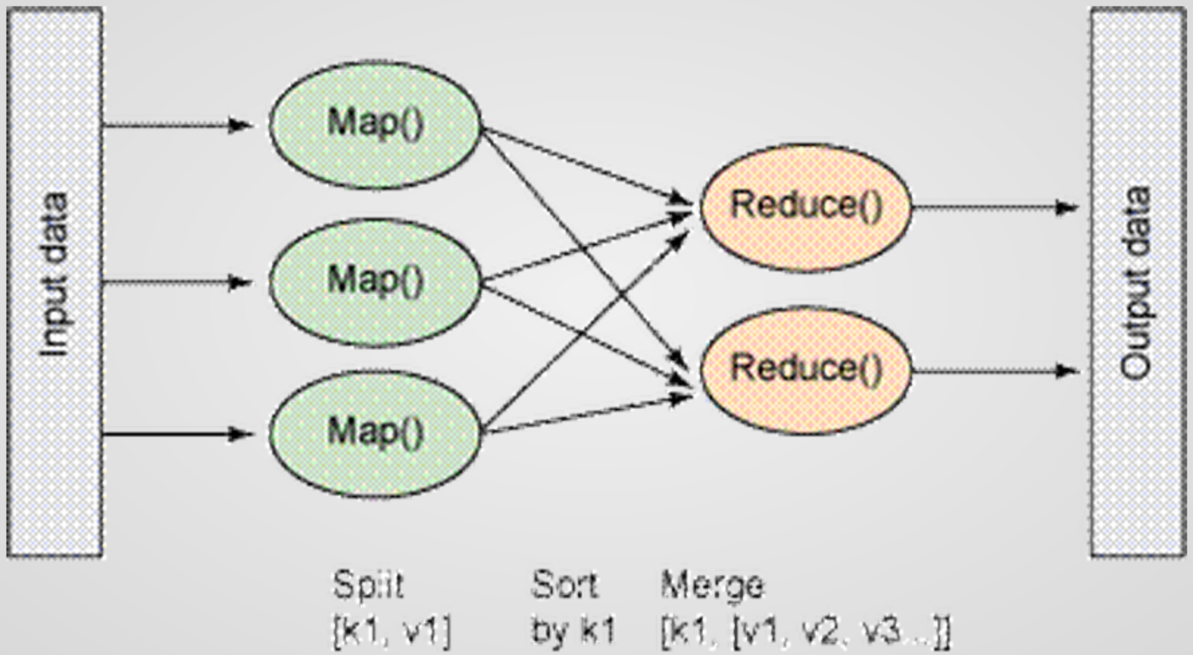
- **Unenclosed JSON** is the dominant style; simple and appendable
- **Enclosed JSON** can optionally be used as a 'feature class'
  (A collection that should be analyzed as a complete set)

Accessing geometries from Hadoop Data Sources

- **com.esri.hadoop.json** - access JSON data as arrays of 'features'
- **com.esri.core.geometry** - construct geometry from arguments

https://github.com/Esri/spatial-framework-for-hadoop/wiki/JSON-Formats

Developing Custom MapReduce Apps



GP tools that convert to/from JSON

ework for Hadoop

o:
apReduce applications

rs to:
eries with ST_Geometry functions

or:
es

PI for Java

Hadoop

ns of 'features'
e dominant style; simple and appendable
tionally be used as a 'feature class'
uld be analyzed as a complete set)

m Hadoop Data Sources
- access JSON data as arrays of 'features'
ry - construct geometry from arguments

al-framework-for-hadoop/wiki/JSON-Formats

JOIN earthquakes

WHERE **ST_Contains** (co
ST_Point (earthquakes.lo

GROUP BY counties.nam
ORDER BY cnt desc;

https://github.com/Esri/gis-tools-for-hadoop/

Simple MapRed

```
void setup() {
  iStream = hdfs.open(new P
  featureClass = EsriFeature
}

void Map(Long key, Text val
  float longitude = Float.pars
  float latitude =  Float.pars
  Geometry point = new Poi

  for (EsriFeature feature : f
    if (GeometryEngine.cont
      String name = feature.a
      context.write(new Tex
      found = true;
      break;
    }
  }
}
```

https://github.com/Esri/gis-to

Demo

Simple MapReduce Code Sample

```
void setup() {
  iStream = hdfs.open(new Path(config.get("input")));
  featureClass = EsriFeatureClass.fromJson(iStream);
}

void Map(Long key, Text value) {
  float longitude = Float.parseFloat(values[COL_LONG]);
  float latitude =  Float.parseFloat(values[COL_LAT]);
  Geometry point = new Point(longitude, latitude);

  for (EsriFeature feature : featureClass.features) {
    if (GeometryEngine.contains(feature.geometry, point) {
      String name = feature.attributes.get(LABEL_ATTR);
      context.write(new Text(name), data);
      found = true;
      break;
    }
  }
}
```

https://github.com/Esri/gis-tools-for-hadoop/tree/master/samples

Demo

ST_Geometry in Hive

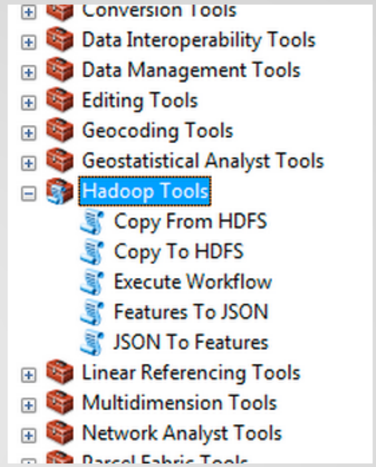
```
SELECT counties.name, count(*) cnt FROM counties

JOIN earthquakes
WHERE ST_Contains (counties.boundaryshape,
ST_Point (earthquakes.longitude, earthquakes.latitude))

GROUP BY counties.name
ORDER BY cnt desc;
```

https://github.com/Esri/gis-tools-for-hadoop/tree/master/samples/point-in-polygon-aggregation-hive

Geoprocessing Tools for Hadoop



Features To JSON, JSON To Features

- Provide serialization to and from JSON formats

Copy To HDFS, Copy From HDFS

- Moves files between ArcGIS and Hadoop

Execute Workflow

- Starts a workflow using the Hadoop Oozie workflow engine

Demo

Download the GIS Tools Project

Clone or Fork the project from Github

http://github.com/Esri/gis-tools-for-hadoop

Pre-built samples in the 'samples' directory

Place your completed tools in the 'tools' directory if you want to share them

Get the Source Code

Geometry API

```
http://github.com/Esri/geometry-api-java
```

Spatial Framework for Hadoop

```
http://github.com/Esri/spatial-framework-for-hadoop
```

Geoprocessing Tools for Hadoop

```
http://github.com/Esri/geoprocessing-tools-for-hadoop
```

To build the source:

```
ant
```

Contributing Your Work

Fork the **gis-tools-for-hadoop** project

- Hack on the code

  -> Make new tools

    -> Do awesome spatial analysis on big data

- Send a Github 'pull request' so we can pull the tool back into our project

Let us know how you are doing

- Add your content to the Wikis on Github
- Troubles? Open new Issues in Github

We want your feedback!

Session Feedback

http://esriurl.com/survey

Session Offering ID: 301

E-mail:

David Kaiser <dkaiser@esri.com> @ddkaiser

Michael Park <mpark@esri.com>

Follow along with this presentation

http://esrigithub.com/gis-tools-for-hadoop/devsummit2013

Use WiFi network: Esri2013

Big Data: Using ArcGIS with Apache Hadoop

David Kaiser @ddkaiser
Michael Park

Esri DevSummit 2013

Session Offering ID: 301

GIS Tools for Hadoop

Hadoop users often have data with spatial value, but with limited options for spatially analyzing this data

Esri has released an open-source framework to enable spatial-data processing in your Hadoop applications

This enables you, as a developer, to build analytical tools that use both Hadoop and ArcGIS.

Why is this important?

Your Hadoop applications can provide spatial analysis

...and your users can leverage your Hadoop applications from within the ArcGIS Geoprocessing environment

Who is a Data Scientist?

Finding Your Data Scientist

We want your feedback!

Session Feedback
http://esriurl.com/survey
Session Offering ID: 301

E-mail:
David Kaiser <**dkaiser@esri.com**> @**ddkaiser**
Michael Park <**mpark@esri.com**>

Contributing Your Work

Fork the **gis-tools-for-hadoop** project
- Hack on the code
  -> Make new tools
  -> Do awesome spatial analysis on big data
- Send a Github 'pull request' so we can pull the tool back into our project

Let us know how you are doing
- Add your content to the Wikis on Github
- Troubles? Open new Issues in Github

Get the Source Code

Geometry API
http://github.com/Esri/geometry-api-java

Spatial Framework for Hadoop
http://github.com/Esri/spatial-framework-for-hadoop

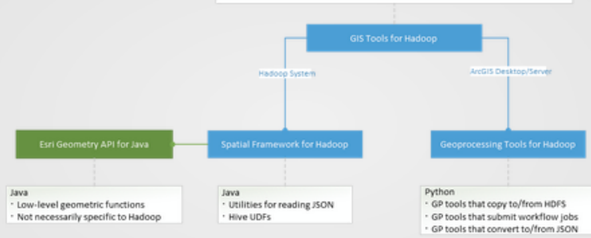
Geoprocessing Tools for Hadoop
http://github.com/Esri/geoprocessing-tools-for-hadoop

To build the source:
ant

GIS Tools for Hadoop

The Hadoop 'Tools' are a combination of custom Hadoop applications and ArcGIS GP Tools.

- Sample tools that demonstrate the GIS capabilities provided for Hadoop
- Templates that can be used to build tools that solve specific big data problems



Download the GIS Tools Project

Clone or Fork the project from Github
http://github.com/Esri/gis-tools-for-hadoop

Pre-built samples in the 'samples' directory

Place your completed tools in the 'tools' directory if you want to share them

Geoprocessing Tools for Hadoop

Features To JSON, JSON To Features
- Provide serialization to and from JSON formats

Copy To HDFS, Copy From HDFS
- Moves files between ArcGIS and Hadoop

Execute Workflow
- Starts a workflow using the Hadoop Oozie workflow engine

Demo

Geometry API for Java

Simple API Functions for Java

com.esri.core.geometry.*

Relationship Analysis	Operations
• equals	• buffer
• disjoint	• cut
• touches	• clip
• crosses	• convexHull
• within	• intersect
• contains	• union
• overlaps	• difference

Spatial Framework for Hadoop

Enables developers to:
- spatially enable MapReduce applications

Enables Hadoop users to:
- run spatial Hive queries with ST_Geometry functions

Provides Java API's for:
- JSON Utility classes
- Hive UDF's

Uses the **Geometry API for Java**

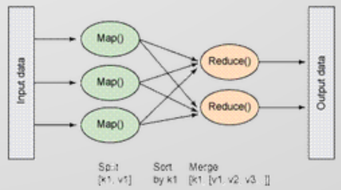
Spatial Data in Hadoop

JSON files store collections of 'features'
- **Unenclosed JSON** is the dominant style; simple and appendable
- **Enclosed JSON** can optionally be used as a 'feature class' (A collection that should be analyzed as a complete set)

Accessing geometries from Hadoop Data Sources
- **com.esri.hadoop.json** – access JSON data as arrays of 'features'
- **com.esri.core.geometry** – construct geometry from arguments

https://github.com/Esri/spatial-framework-for-hadoop/wiki/JSON-Formats

Developing Custom MapReduce Apps



ST_Geometry in 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;
```

https://github.com/Esri/gis-tools-for-hadoop/tree/master/samples/point-in-polygon-aggregation-hive

Simple MapReduce Code Sample

https://github.com/Esri/gis-tools-for-hadoop/tree/master/samples

Demo