Big Spatial Data
Bruce Sanderson
TRY TO
READ THIS
IF YOU CAN THEN
YOU ARE A MUCH BETTER PERSON
THAN I CAN EVER HOPE TO BE IN THIS LIFETIME OR THE NEXT
GOODBYE CRUEL WORLD IT WAS NICE KNOWING YOU BUT NOW IT IS TIME TO GO FOR GOOD
THE PROBLEM WITH BIG DATA IS NOT THAT IT’S BIG
It’s that there is LOTS OF IT
IN ORDER TO SEE CLEARLY WE NEED BETTER TOOLS THAT BRING THINGS BACK INTO FOCUS
Agenda

• Our Hadoop implementation
• Big Spatial Data workflows
• What should you be doing?
SPATIAL DATA MANAGEMENT WINDOW too short
UNSTRUCTURED DATA
Drivers

- Volume of data
- Need to load/process quicker
- Data that doesn’t fit into RDBMS
How do you get started?

If you’re not a Big Data expert…

But have Big Data needs
Basic Hadoop Stack

MapReduce

Yet Another Resource Negotiator (YARN)

Hadoop Distributed File System (HDFS)

Commodity Servers
MapReduce

- *MapReduce* is a programming model and an associated implementation for processing and generating large data sets with a parallel, distributed algorithm on a cluster.

- Yuck!
Hive  Pig  Cascading  
Map  HBase  Impala  Spark  
Reduce  

Hadoop Distributed File System (HDFS)  

Sqoop  Flume
What Next?
Load data!

hadoop fs -put N:\07_2014_gps.txt \user\vehicle\data.txt
• Command line interface for transferring data between relational databases and HDFS

• Support joins and where clauses

• Quest Data Connector - Oracle
- SQL on Hadoop
- External tables
- Schema on read

```sql
hive> CREATE EXTERNAL TABLE well(id INT, api STRING, name STRING) ROW FORMAT DELIMITED FIELDS TERMINATED BY ',' LINES TERMINATED BY '\n' STORED AS TEXTFILE LOCATION '/home/admin/userdata';
```
• SQL on HDFS

• Bypasses MapReduce

• Most operations in-memory
• Up to 100x faster than Hadoop MapReduce

• Java, Scala, Python, R

• Access diverse data sources
  - HDFS, HBase, Hive, S3

• SparkSQL
Snakebite

- Pure Python client
- Allows us to use ArcPy!
Big Spatial Data workflows

• Vehicle Tracking Analysis
• Directional Survey Analysis
• Custom GeoAnalytics Interface
GeoJson

FME

SDE

Spark

HDFS

Sqoop

Directional Surveys

ArcMap
What did we learn?

- There is a need for *some* Big Data tools
- Mostly using *Spark* directly against HDFS
What does that mean for you?
10.4 Release

- Big Data *built-into* ArcGIS
- Push button deployment
- Any number of nodes
- *Spark* is part of framework!
• ‘Real-Time GIS: The Road Ahead’
• Room 14 B, Wed 1:30-2:45pm
Real-Time GIS
Integration & exploitation of high velocity & volume data

- Integrates real-time high velocity & volume data into ArcGIS
- Performs continuous processing & real-time analytics
- Sends updates & alerts to those who need it where they need it
Add features to a Spatiotemporal Big Data Store or Update features in a Spatiotemporal Big Data Store.
So you no longer need to be a Big Data expert to use it in your geospatial environment.
Hadoop

- Invest in Spark
- Learn Java, Scala, or Python

No Hadoop

- Wait for 10.4!
- Study Spark
- Learn Java, Scala, or Python
Summary

• Do you *need* Big Spatial Data Tools?
  - Yes – but probably worth waiting for 10.4

• Spark
Stay in touch:  

bruce.sanderson@apachecorp.com