Big Data and Analytics: Getting Started with ArcGIS

Mike Park
Erik Hoel
Agenda

• Overview of big data
• Distributed computation
• User experience
• Data management
Big data
What is it?

- Big Data is a loosely defined term used to describe data sets so large and complex that they become awkward to work with using standard software in a tolerable elapsed time
  - Big data "size" is a constantly moving target, ranging from a few dozen terabytes to many petabytes of data
  - In the past three years, 90% of all recorded data has been generated
- Every 60 seconds:
  - 100,000 tweets
  - 2.4 million Google searches
  - 11 million instant messages
  - 170 million email messages
  - 1,800 TB of data
NYC Taxis by Day

Manhattan Taxis Friday after 8pm
Big data
What techniques are applied to handle it?

- Data distribution – large datasets are split into smaller datasets and distributed across a collection of machines.
- Parallel processing – using a collection of machines to process the smaller datasets, combining the partial results together.
- Fault tolerance – making copies of the partitioned data to ensure that if a machine fails, the dataset can still be processed.
- Commodity hardware – using standard hardware that is not dependent upon exotic architectures, topologies, or data storage (e.g., RAID).
- Scalability – algorithms and frameworks that can be easily scaled to run on larger collections of machines in order to address larger datasets.

“Big data is not about the data.”
– Gary King
Harvard University
Director, Inst. For Quantitative Social Science

(Making the point that while data is plentiful and easy to collect, the real value is in the analytics)
ArcGIS users have big data

- Smart Sensors
  - Electrical meters (AMI), SCADA, UAVs
- GPS Telemetry
  - Vehicle tracking, smartphone data collectors, workforce tracking, geofencing
- Internet data
  - Social media streams, web log files, customer sentiment
- Sensor data
  - Weather sensors, stream gauge measurements, heavy equipment monitors, ...
- Imagery
  - Satellites, frame cameras, drones
GeoAnalytics Examples

- Aggregate vehicle locations into cells for each 10 minute period to reveal traffic patterns

- Aggregate 911 call logs into census blocks by hour to reveal call patterns

- Aggregate web logs of access to map tile servers to determine hotspots of customer interest

- Geocode large address sets in parallel using a geocoding service

- Enrich very large numbers of point locations with contextual data and then select subset of locations meeting certain criterion
Road ahead?
GeoAnalytics
What is it, and what does it enable me to do?

- GeoAnalytics will be a new *capability* of ArcGIS Server

- It provides me:
  - The ability to do fast batch analysis on large tabular / feature datasets
  - The ability to do fast batch analysis on large raster and image datasets
  - The ability to do fast batch analysis on large geo-event observation archives
GeoAnalytics

What does ‘batch’ analysis mean

• Batch analysis means the ability to run analysis jobs on large datasets
  - The input is a persisted standard or big dataset
  - The output is a persisted standard or big dataset

• Datasets
  - Standard geospatial data (geodatabases, files, services)
  - Big Data (databases, files, services)

• Key point:

  With suitably scaled GeoAnalytics, jobs that would take hours now take minutes
GeoAnalytics Extension for Server

• **Adds out of the box analytics to ArcGIS Server**
  - Analysis in ArcGIS Pro and Portal
  - Powered by a new Analysis Service / Toolbox in Server
  - Focused analysis for big data

• **Works with:**
  - Standard geospatial data (geodatabases, files, services)
  - Big Data (databases, files, services)
GeoAnalytics Extension for Server

Overview

• Users are able to manage, analyze, and visualize big data to derive valuable information

• Previously impossible or slow analytics are made possible by leveraging the power of distributed computation

• Analytics and complicated technologies are made easy by ArcGIS integration

• Ability to perform analysis on vector and raster data
Distributed computation

Integrated into ArcGIS Server

- Distributed analytics against distributed data
- Many frameworks/technologies exist for distributing computation
  - E.g., Hadoop, MapReduce, Spark
  - **Spark**: processes distributed data in memory
- ArcGIS Server integrates these technologies on a cluster to solve analytic problems
GeoAnalytics

Distributed analysis on distributed data

- Parallelized batch analytics on tabular, vector, raster, and imagery datasets (big and standard data)

- Supports data exploration via feature, map, and image layers
GeoAnalytics

Performance: minutes, not hours

- 16 nodes in the cluster
  - 4 cores per node
  - 8 – 16GB RAM per node

- Polygons (NYC Blocks) 40K
- Points (NYC Taxi) 170M
GeoAnalytics
User Experience - Analysis

• ArcGIS Pro:
  - Out of the box tools that run in Server and process services and registered data using a GP tool interface

• Tools are exposed through a REST-based interface that can be used by ArcGIS Pro or web clients
Initial release

ArcGIS 10.4 - Analysis

• Analysis capabilities patterned after the ArcGIS Online Spatial Analysis service
  - Contains a useful subset of the current tasks

• GeoAnalytics includes additional tools useful for a big data workflows
  - Move data to and from the client
  - Register and manage data resident in the Big Data Server’s directories
  - Addition of temporal capabilities
  - Ability to write to NetCDF
Analytic capabilities
ArcGIS 10.4 release

• Summarize Data
  - Aggregate Points by Polygon + time
  - Aggregate by Cell + time
  - Summarize Nearby + time
  - Summarize Within + time

• Find Locations
  - Find Existing Locations
  - Find Similar Locations

* New GeoAnalytics capabilities in orange
Analytic capabilities

**ArcGIS 10.4 release**

- **Analyze Patterns**
  - Calculate Density
  - Find Hot Spots + *time*

- **Use Proximity**
  - Create Buffers + *time*

- **Manage Data**
  - Extract Data
  - Field Calculator
  - Geocode Addresses

* New GeoAnalytics capabilities in orange
Data Stores
Management

• Both GIS data stores and big data stores are supported
  - Map and Feature services
  - ArcGIS SQL Data Store

• Directories of files (shapefiles, CSVs, etc.) serve as data stores
  - GIS file shares
    - Each file represents a single dataset
  - Big data file shares
    - Folder of sharded shapefiles or other file formats

• ArcGIS Big Data Store
Anatomy of a Feature

Not just spatial

Attributes

- Text
- Numbers
- Dates
- Binary
- ...

Geometry

- Point
- Polyline
- Polygon

Time

- Instant
- Interval
Why Time Is Important

Space/time relationship

- 10 meters
- 4 meters
- 5 minutes
- 30 minutes
- 5 minutes

X

Y

Time

X
Why Time Is Important

Summarization
Aggregation

Summary statistics

- **Numeric Statistics**
  - Count
  - Min
  - Max
  - Sum
  - Mean
  - Standard Deviation
  - Variance

- **Text Statistics**
  - Min (alphabetical ordering)
  - Max (alphabetical ordering)
  - Any
Aggregation
Point counts and attribute means
Path generation

Vertex count aggregation
Spatio-temporal big data store

Management

• Distributed data store for high velocity, high volume data

• Available to GeoAnalytics and GeoEvent
  - Supports high velocity continuous analytics with GeoEvent services
  - Supports high volume batch analytics with GeoAnalytics services

• Accessible through feature services

• Based upon Elasticsearch for storage and indexing
  - Open-source real-time distributed search engine and data store built on top of Apache Lucene
Integration with GeoEvent

ArcGIS 10.4 Release

- Enhanced GeoEvent service integration

- Partnership to better support persisting high velocity, high volume streaming data into the Big Data cluster
  - Spatio-temporal Big Data Store

- Shared platform service for distributed computation
GeoAnalytics capabilities for server

Summary

- Allows you to run GeoAnalytics on dedicated server nodes
- Uses services and data stores to expose the results of analyses
- Supports management and analytics against massive spatio-temporal datasets
Why would I want to use it?

Summary

• Functionality available out of the box in Portal; no need to publish

• Runs on big data collections (observational data)
  - Data collections whose size was previously problematic

• Runs fast, and is scalable

• I don’t need to learn anything new; I use it just like existing GP tools
Understanding our world.