Highly Scalable Tile-Based Visualization for Exploratory Data Analysis

Strata NY: Hadoop and Beyond, 10/17/2014 David Jonker, Rob Harper





Making Sense of Big Data

Big Data for us is in complexity.
 We have complex data.
 We have big complexity.
 Analyst

No one wants more data, everyone wants better data. - Analyst



More Information Available To make an effective decision

anertureT

Challenges of Effective Visual Analytics

ESSENTIAL

- Richly **informative** and true to reality.
- Answers easily arrived at, easily understood.



More Information Available To make an effective decision

ALSO, every business problem is unique, but cannot afford an entirely unique solution for every problem.

• Need relatively universal, **repeatable** technical approaches.



Tile-Based Visualization for Big Data

WEB MAPPING APPS have an established model for intuitively navigating and understanding big data, based on zoomable multiresolution **tiles** and **layers**.

BUT geospatial data is relatively static. How can a tiled, layered approach be applied with **dynamic** data, and non-geospatial problems, at scale?





Aperture Tiles

TILE-BASED VISUAL ANALYTICS

- Hierarchical data tiling using cluster computing.
- Interactive on-demand image tile generation.
- Layers of raw data and derivative analytics.

OPEN SOURCE

- Oculus research product.
- Built on Apache Spark and Hadoop.
- Preparing for version one product release.





Example Applications



Geospatial

Social Media

Biomedical



Financial



OCULUS @ 2014 OCULUS INFO INC.

187,000,000 NYC Taxi Trips

Photo by <u>Sakeeb Sabakka</u>, CC

NYC Taxi Data

FOIL request by Chris Whong, March 2014



Taxi ID	Hack, Medallion, Vendor
Origin	Location, Time
Dest	Location, Time
Trip	Duration, Distance, # Passengers
Fare \$	Base, Tip, Tolls, Taxes, Payment method

a©ertureTiles 🖽

OCULUS © 2014 OCULUS INFO INC.



Visualizing NYC Taxi Data

Eric Fischer, MapBox

www.mapbox.com/blog/nyc-taxi/

+

agertureTiles Demo





GPS blurring likely caused by fall buildings

Visualizing all of the data including "bad data" helps identify source of errors, understand how errors manifest, and how they may affect analysis





Comparing Pick-up and Drop-off Locations

Red – Pick-up Locations Blue – Drop-off Locations

Multi-scale visualization allows exploration of macro and micro trends.

Average Tip % by Pick-up Location*

Red	~15%
Green	~20%
Yellow	~25%

*CC payments only

Using a spectral color ramp helps us compare differences across the data (luminance better for multi-series).

Tiles uses aggregate data to generate rasters, not precomputed images

Downtown bankers tip less than mid-town shoppers?

Queens

LaGuardia passengers tip more than those

from JFK?

Easily switching between fully zoomable multi-scale layers makes the task of understanding the data easier.

Very short ride anomaly in upper east side exposed by range filtering – hospital sending patients on a short ride to clinic?





De Hot zones at ends of tunnels due to GPS fix

Average \$/hr by **Pick-up Location**

Red \$30/hr Yellow \$75/hr

Calculated using: (fare + tip) / (duration of trip + time until next fare)

Harrow

С

lag observed earlier?

Profitability correlates with pickup likelihood on one-way streets

Brooklyn



Up until now have focused on pick-up location, what about drop-off?

Average \$/hr by Drop-off Location

Red \$30/hr Yellow \$60/hr

Calculated using: (fare + tip) / (duration of trip + time until next fare)



Average annual taxi income by pick-up location

Red \$80k Yellow \$100k



Tiles approach isn't limited to geographic plots

X-Y crossplots can show 3 arbitrary dimensions and expose macro and micro patterns.

e.g. Time + Fare Total + Frequency

World Cup Tweets

player @mentions

Photo by paulisson miura, CC





Analytic overlays showing @mention summaries



Tiles from 10,000'

Fille A. Control of O

0

Brew Link Din

Three Tiers



a@ertureTiles 🛱

Tile Generation



oculus © 2014 OCULUS INFO INC.

25

apertureTiles

Tiled Data Views

Aggregate views of source data designed for answering analytic questions

Optimized for query speed

Indexed by tile key - one view/row per tile

AVRO / Thrift



26



Tiled Data Views

Taxi Tip %

16	%	179	7		14%]()°	%							
18	%	20%	7		17%	1;	59	%		12%	1	4%			2
21	%	229	7	í	21%	10	99	%		14%		7%	15%		ļ
	<i>/</i>		Ľ			<u> </u>	_	, o				1%	19%		
20	%	21%	6		17%	1	5	%				7%	15%	12%	
21	%	199	7		18%							8%	14%	13%	
21	/0	1//	<u> </u>		1070			1 1 70		1 707	_	407	1007	1.407	1
		17%	14)	%	10%		ļ	167	ò	17%	Ľ	Ň.	10%	14%	ļ
		20%	175		15%	20%		18%	7 0	20%	1	7%	15%		
\setminus	21%	22%	215	%	19%	16%	ĺ	219	70	22%	2	1%	$\overline{\}$		ĺ
\setminus	20%	21%	175	%	15%	18%		20%	8	21%					
N	21%	19%	18	%	18%	19%		219	6	19%					
															K

apertureTiles 🗄

OCULUS © 2014 OCULUS INFO INC.

Tiled Data Views

World Cup Sentiment

Top Mentions: [...]

Sentiment Values: [...]

Sentiment Values: [...]

Tweets per minute: [...]

Sentiment Values: [...]

Tweets per minute: [...]

agertureTiles

Tile Generation

	16%	17%	14%	10%	
	18%	20%	17%	15%	12%
Γ	21%	22%	21%	19%	14%
	20%	21%	17%	15%	
	21%	19%	18%		







OCULUS © 2014 OCULUS INFO INC.

Tile Client

Tile Map Service (TMS)

http://{rootURL}/{layer}/{zoom}/{x}/{y}.png



Leaflet 🏑







Tile Client

Twitter Big Data Heatmap Demo

This Aperture Tiles demo illustrates the integration of advanced analytics into an interactive tile-based visual analytics demo. The demo features multiple tiles layers:

1. A base google map

Tweet heatmap

Base Layer

4

2

Descity

2. A heatmap of geo-coded twitter messages ("tweets") across the World

The dataset used for this demonstration consists of 292,722,821 Twitter messages.

Controls

Google Black

settings

Google Gray



(Debelointig

anorth Scena

a@ertureTiles 🖽

http://tiles.oculusinfo.com



Key Points of Value

PLOTTING ALL of the data for richer, truer insights.

- Reveals **truths** in the data that summaries cannot.
- Provides important **context** and **evidence** for analytics.

TILED, LAYERED user interface for natural, scalable exploration.

- Intuitive navigation and understanding.
- Analytic layers at **levels of detail**, down to raw provenance.

INTERACTIVE analysis with data tiling and analytics.

• Rapidly repeatable **tailored** solutions, with filtering + selection.

50,000,000 link interactive graph Upmc Presbytenan Shadyside 1 trillion+ pixels of resolution

Quest Diagnostics Incorp.

Prompt Medical Transcontation Metwest Inc.

Kan di-k

Sonora Quest Laboratorie

Chicago Clinical Laboratorie

Quest Diagnostics Incorrection of America Hor Paterson (America Horpital Symphony Diagnostic S... Laboratory Corporation Of Americ

Quest Diagnostics Clinic...

Aculabs Inc

Laboratory Corporationatory Corporation O... Biological Technology Laboratory

Mercy Health S

hs Hospital

State University Of Iowa

Southern Baptist Hospital Of Florida (200 Boyce - Bynum Pathology Lab Laboratory Gorporation

Quest Diagnostics Incorporat...

Eaboratory Atlanc

Quest Diagnostics Lic

Dia

Christiana-Care Health Service

What's Next?

pration O... Diagnostic Labo U.s. Lab& Radiolog sosvenezuel: ostics Clinical E colectivos otesta. Jucha Paz chavitz 10 st Diagnostics Incorporat INCOME. maduro and the second se qobierno Jro ente venezolanos Servichcock Memorial Hospit Stimute 01010100110 >

Interactive Tiled Graphs Streaming Data and live query

More advanced filtering and analytics

Questions?

Oculus gratefully acknowledges the support of DARPA in research and development of this work.



35