Visualizing big data in the browser using Spark

Hossein Falaki  @mhfalaki
Spark Summit East – March 18, 2015
Exploratory Visualization

“Critical part of data analysis”
—William S. Cleveland

Put visualization back in the normal workflow of data analysis regardless of data size.

• Interactive
• Collaborative
• Reproducible
Expository Visualization

Communication is often the bottleneck in data science, and a graph is worth a thousand words.

- Control over details
- Shareable
Requirements

- Interactive
- Collaborative
- Shareable
- Reproducible
- Control over details

} \hspace{1cm} \text{Use the browser}

} \hspace{1cm} \text{Use visualization libraries}
Visualization as programming

- For complex tasks point and click may not be enough
- Best expressed with a grammar (API)
- Scripts are reproducible
- Control over all details
- Data scientists are already familiar with these tools

D3.js, Three.js, matplotlib, ggplot, Bokeh, Vincent, …
Do it in the browser

- Output of these tools can be readily used on the web (PNG, SVG, Canvas, WebGL)
- No need to transfer data and results
- Browser is conducive to collaboration (e.g., Notebooks)
- Separating data manipulation from rendering enables users to freely choose the best tool for each job
Challenges with big data visualization

1. Manipulating large data can take a long time
2. We have more data points than pixels

Apache Spark can help solve both problems
Challenges

1. Manipulating large data can take a long time
   > Memory
   > CPU
Reducing latency: caching

Take advantage of memory and storage hierarchy

- Serialized storage levels (for memory)
- Memory & GC tuning
Reducing latency: parallelism

Increase number of CPUs
  > Get more executors with Mesos or Yarn
  > Click a button to increase cluster size in DBC

- Control level of parallelism for map and reduce tasks
- Configure spark locality if needed
Challenges

1. Manipulating large data can take a long time

2. We have more data points than possible pixels

  > Summarize
  > Model
  > Sample
More data than pixels? Summarize

- Extensively used by BI tools
  > Aggregation
  > Pivoting
- Most data scientists’ nightly jobs summarize data
More data than pixels? Model

MLLib supports a large (and growing) set of distributed algorithms

- Clustering: k-means, GMM, LDA
- Classification and regression: LM, DT, NB
- Dimensionality reduction: SVD, PCA
- Collaborative filtering: ALS
- Correlation, hypothesis testing
More data than pixels? Sample

Extensively used in statistics

Spark offers native support for:
- Approximate and exact sampling
- Approximate and exact stratified sampling

Approximate sampling is faster and is good enough in most cases
Demo
Summary

Using Spark we can extend interactive visualization of large data

Reduce interaction latency to seconds
  - Cache data in memory
  - Increase parallelism

To visualize millions of points in the browser
  - Summarize
  - Model
  - Sample
Visualizing big data in the browser using Spark