Interactive Visual Data Exploration with Spark

Hossein Falaki
@mhfalaki
Databricks Inc.
Exploratory vs. Expository
Large data
“Visualization is critical to data analysis.”

William S. Cleveland

But we often skip it for large data
Challenges

1. Interactivity

   interactivity with large data is challenging
   Use Apache Spark

2. Visual medium

   More data points than we have pixels
   Summarize, Sample, Model
Apache Spark

• General computing engine for batch, streaming, and iterative jobs over distributed data

• Capable of handling peta bytes of data

• Enables caching distributed data

• Versatile programming interfaces
Versatile programming interface

• SQL, Scala, Python, Java and (experimental) R API

• Libraries for distributed statistics and learning

• Exploratory data visualization is very much like programming
  • Point and click doesn’t really cut it
  • Requires an API (grammar): ggplot, matplotlib, bokeh, etc.
Interactively & Iterative jobs

- Cache data in memory to reduce latency
- Control data partitioning and parallelization to reduce latency
- Powerful API for data manipulation
  - Mix SQL with other languages
  - Create Hive tables from data in any language
Can we visualize 200GB of multidimensional data?

**Short answer:** no

**Long answer:**
- Summarize & visualize
- Sample & visualize
- Model & visualize
Summarize & visualize
Sample & visualize
Model & visualize
Iterative analysis