May 15th, 2019

Spark for Social Science

Kyle Ueyama, Senior Programmer & Data Scientist

URBAN INSTITUTE · ELEVATE · THE · DEBATE

Overview

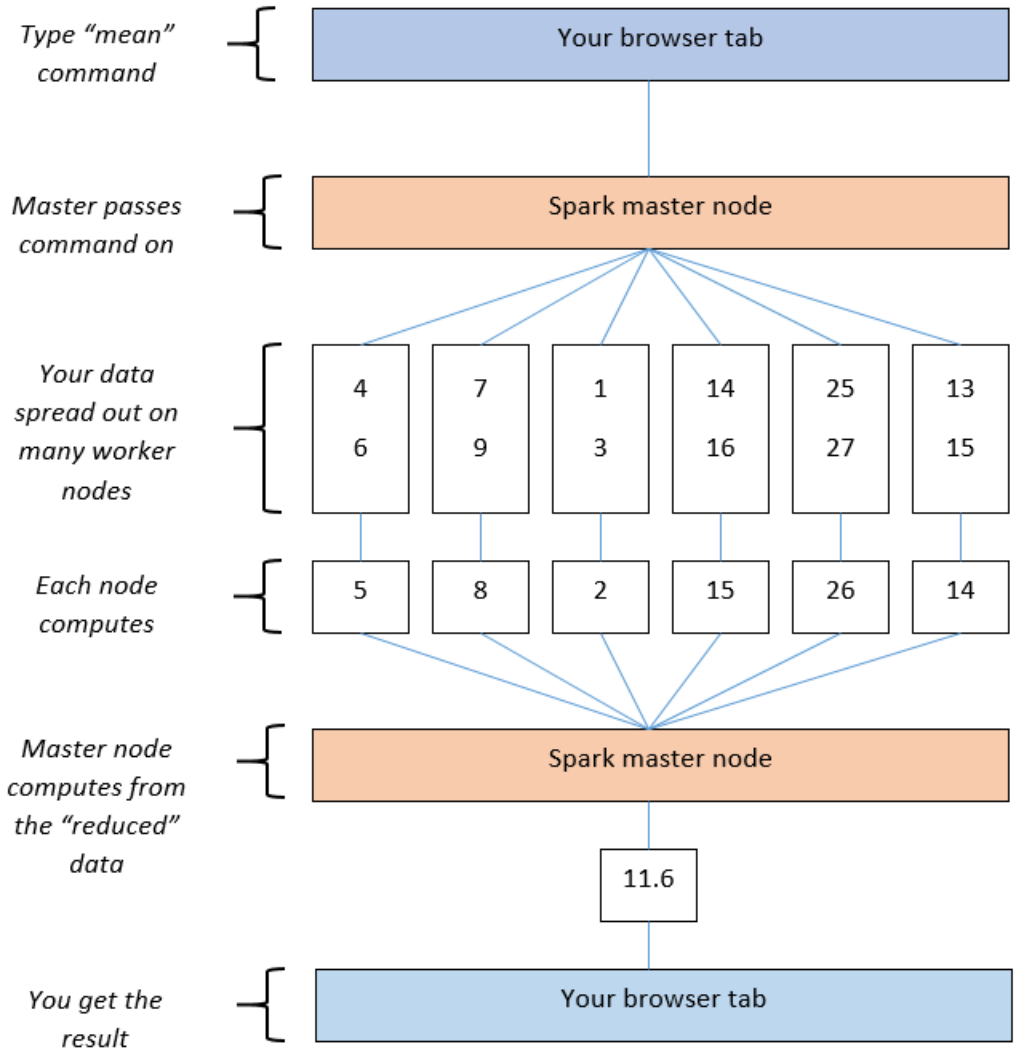
- About Urban and the Data Science Team
- Administrative Data can be Big Data
- Spark Makes Processing Big Data Really Fast
- Spark for Social Science
- Pluses and Minuses
- When to Use Spark
- Stay in Touch!

About Urban and the Data Science Team

Our Administrative Data can be “Big Data”

Spark Makes Processing Big Data Really Fast

- Outgrowing on-premise servers: sometimes a single computer just isn't powerful enough
- Distribute data storage and processing tasks across a cluster of machines
- Can be scaled to a massive degree
- 500 hours processing time to 10 minutes in some use cases



Spark for Social Science

- Manual
  - https://urbaninstitute.github.io/spark-social-science-manual/
- Technical Details
  - https://github.com/UrbanInstitute/spark-social-science
- SparkR Tutorials
  - https://github.com/UrbanInstitute/sparkr-tutorials
- PySpark Tutorials
  - https://github.com/UrbanInstitute/pyspark-tutorials

Pluses

- Speed
- No sharing
- No learning curve (for some)

Minuses

- Supports only R & Python
- 10-15 minute spin-up time
- Learning curve (for some)

When to Use Spark

- Data > 5-10GB
- Data takes too long to process
- Have or willing to acquire R/Python expertise
- You/IT Staff has cloud experience

Stay in Touch!

- Data@Urban on Medium
  - https://medium.com/@urban_institute
- Twitter: @khueyama
- Email: kueyama@urban.org