11-2018

Spark for Social Science

Graham MacDonald
Urban Institute

DOI https://doi.org/10.23889/ijpds.v3i5.1044

Spark for Social Science

Abstract
Urban has developed an elastic and powerful approach to the analysis of massive datasets using Amazon Web Services' Elastic MapReduce (EMR) and the Spark framework for distributed memory and processing. The goal of the project is to deliver powerful and elastic Spark clusters to researchers and data analysts with as little setup time and effort possible, and at low cost. To do that, at the Urban Institute, we use two critical components: (1) an Amazon Web Services (AWS) CloudFormation script to launch AWS Elastic MapReduce (EMR) clusters (2) a bootstrap script that runs on the Master node of the new cluster to install statistical programs and development environments (RStudio and Jupyter Notebooks). The Urban Institute's Spark for Social Science Github page holds code used to setup the cluster and tutorials for learning how to program in R and Python.

Comments
DOI https://doi.org/10.23889/ijpds.v3i5.1044

November 13th, 2018

Spark for Social Science

Graham MacDonald, Chief 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

Top 4 reasons Spark is so fast for us

Spark for Social Science

Pluses and Minuses

When to use Spark

Stay in touch!

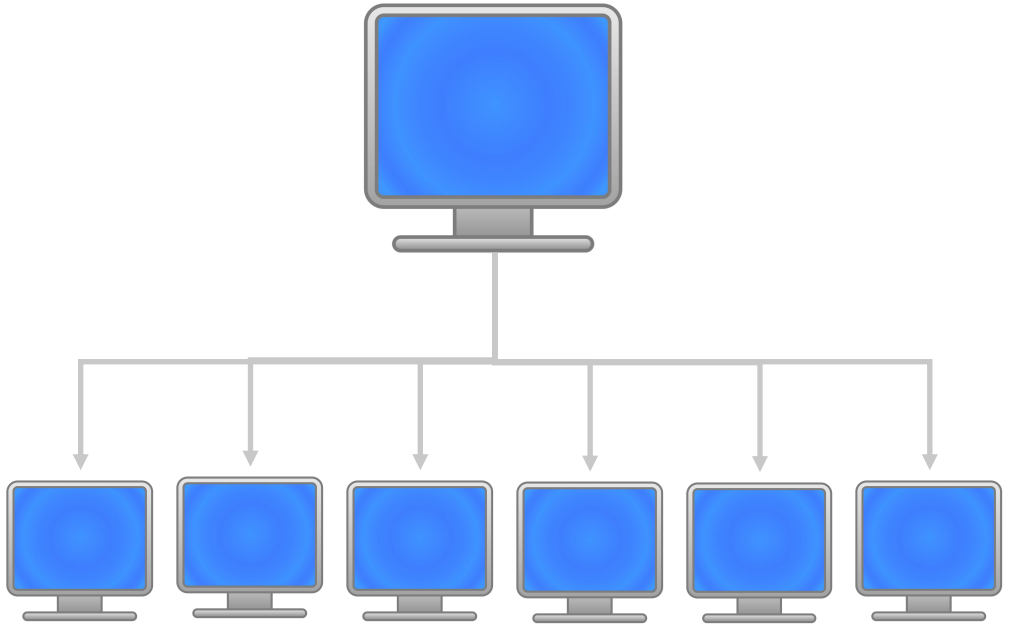
About Urban and the Data Science Team

Our Administrative Data can be “Big Data”

Spark makes processing Big Data really fast

- 500 hours

- 10 minutes



Top 4 reasons Spark is so fast for us

- Many powerful machines running at once
- Cloud machines with the fastest processors
- Compression for much faster read times
- Spark Optimization of code

Spark for Social Science

https://urbaninstitute.github.io/spark-social-science-manual/

Pluses

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

Minuses

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

When to Use Spark

- Data > 5-10GB
- Process takes too long
- 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
- @grahamimac on Twitter