International Journal of Population Data Science

Journal Website: www.ijpds.org

Spark for Social Science

MacDonald, G[1*], Engler, A[2], Levy, J[1], and Armstrong, S[2]

[1]Urban Institute

[2]University of Chicago

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.

*Corresponding Author:
Email Address: GMacDonald@urban.org (G MacDonald)

http://dx.doi.org/10.23889/ijpds.v3i5.1044