Scalable Data Analytics Using R: Single Machines to Hadoop Spark Clusters

John-Mark Agosta, Debraj GuhaThakurta, Robert Horton, Mario Inchiosa, Srini Kumar, and Mengyue Zhao
Microsoft
One Microsoft Way
Redmond, Washington, 98052, USA
{joagosta,debraj.guhathakurta,rhorton,marinch,srini.kumar,mez}@microsoft.com

ABSTRACT

R is one of the most popular languages in the data science, statistical and machine learning (ML) community. However, when it comes to scalable data analysis and ML using R, many data scientists are blocked or hindered by (a) its limitations of available functions to handle large datasets efficiently, and (b) knowledge about the appropriate computing environments to scale R scripts from desktop exploratory analysis to elastic and distributed cloud services. In this tutorial we will discuss solutions that demonstrate the use of distributed compute environments and end to end solutions for R. We will present the topics through presentations and worked-out examples with sample code. In addition, we will provide a public code repository that attendees will be able to access and adapt to their own practice. We believe this tutorial will be of strong interest to a large and growing community of data scientists and developers using R for data analysis and modeling.

Keywords

Distributed Systems; Machine Learning; Advanced Analytics; Predictive Analytics; Statistical modeling; Statistical Computing; Parallel computing; Scalability; Spark; Hadoop; YARN; R; SQL; Visualization; Learning Curves; Hierarchical Time Series.

1. TUTORIAL OUTLINE

1. Introduction: scaling your R scripts - issues and solutions
    a. What limits the scalability of R scripts?
    b. What functions and techniques can be used to overcome those limits?
    c. How do the base and scalable approaches compare?
2. End to end scalable data analysis in R: Data exploration, visualization, modeling and deployment using distributed R functions and Hadoop/Spark
    a. Scalable analysis on single nodes
    b. Integrated analytics in Spark: Data exploration, distributed training and deployment of ML models
3. Practical examples and case studies
    a. Exploration and visualization using SparkSQL and R
    b. Distributed model training and parameter optimization: A parallelized investigation into Learning Curves
    c. Parallelizing models: training and deploying many parallel models for hierarchical time series optimization
    d. Overview of an open library of R templates and examples
4. Q&A

KDD '16, August 13-17, 2016, San Francisco, CA, USA.
ACM 978-1-4503-4232-2/16/08.
http://dx.doi.org/10.1145/2939672.2945398