KDD2016

KDD 2016 Tutorials

Scalable R on Spark

John-Mark Agosta / Microsoft

Debraj GuhaThakurta / Microsoft

Robert Horton / Microsoft

Mario Inchiosa / Microsoft

Srini Kumar / Microsoft

Vanja Paunic / Microsoft

Hang Zhang / Microsoft

Mengyue Zhao / Microsoft

Tutorials

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 data-sets 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 hands-on 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.

Prerequisites: A laptop with a web browser and an ssh client that supports port forwarding. Access to cloud-based clusters will be provided. For R scripts, download details, and suggested reading, see the Readme.md file at https://github.com/Azure/Azure-MachineLearning-DataScience/tree/master/Misc/KDDCup2016.

PROGRAM SKETCH

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

ML models

3. Practical examples and case studies
    - a. Exploration and visualization using SparkSQL and R
    - b. Parallel models: training many parallel models for hierarchical time series optimization
    - c. Distributed model training and parameter optimization: Learning Curves on Big Data
    - d. Overview of open libraries of R templates and examples
4. Q&A