Analyzing Big Data with Microsoft R

Modality: On Demand

Duration: 16 Hours

About this course:

The open-source programming language R has for a long time been popular (particularly in academia) for data processing and statistical analysis. Among R's strengths are that it's a succinct programming language and has an extensive repository of third party libraries for performing all kinds of analyses. Together, these two features make it possible for a data scientist to very quickly go from raw data to summaries, charts, and even full-blown reports. However, one deficiency with R is that traditionally it uses a lot of memory, both because it needs to load a copy of the data in its entirety as a data.frame object, and also because processing the data often involves making further copies (sometimes referred to as copy-on-modify). This is one of the reasons R has been more reluctantly received by industry compared to academia.

The main component of Microsoft R Server (MRS) is the RevoScaleR package, which is an R library that offers a set of functionalities for processing large datasets without having to load them all at once in the memory. RevoScaleR offers a rich set of distributed statistical and machine learning algorithms, which get added to over time. Finally, RevoScaleR also offers a mechanism by which we can take code that we developed on our laptop and deploy it on a remote server such as SQL Server or Spark (where the infrastructure is very different under the hood), with minimal effort.

In this course, we will show you how to use MRS to run an analysis on a large dataset and provide some examples of how to deploy it on a Spark cluster or a SQL Server database. Upon completion, you will know how to use R for big-data problems.

Since RevoScaleR is an R package, we assume that the course participants are familiar with R. A solid understanding of R data structures (vectors, matrices, lists, data frames, environments) is required. Familiarity with 3rd party packages such as dplyr is also helpful.

Course Objective:

You will learn how to use MRS to read, process, and analyze large datasets including:

- Read data from flat files into R’s data frame object, investigate the structure of the dataset and make corrections, and store prepared datasets for later use
- Prepare and transform the data
- Calculate essential summary statistics, do crosstabulation, write your own summary functions, and visualize data with the ggplot2 package
- Build predictive models, evaluate and compare models, and generate predictions on new data

Audience:

- Data scientist

- Data Analyst

Prerequisite:

- Familiarity with R

Course Outline:

1. Introduction

- Introduction

2.Reading and Preparing Data

- Reading the Data
- Preparing the Data
- LAB

3. Examining and Visualizing Data

- Examining the Data
- Visualizing the Data
- LAB

4. Clustering and Modeling

- Clustering
- Predictive Modelling
- LAB

5. Deploying and Scaling

- Deploying and Scaling

Final Exam and Wrap-up

- Final Exam
- Exam Wrap-up!