MOC 20773A: Analyzing Big Data with Microsoft R

Course Summary

Description
The main purpose of the course is to give students the ability to use Microsoft R Server to create and run an analysis on a large dataset, and show how to utilize it in Big Data environments, such as a Hadoop or Spark cluster, or a SQL Server database.

Objectives
By the end of this course, students will be able to:

- Explain how Microsoft R Server and Microsoft R Client work
- Use R Client with R Server to explore big data held in different data stores
- Visualize data by using graphs and plots
- Transform and clean big data sets
- Implement options for splitting analysis jobs into parallel tasks
- Build and evaluate regression models generated from big data
- Create, score, and deploy partitioning models generated from big data
- Use R in the SQL Server and Hadoop environments

Topics
- Microsoft R Server and R Client
- Exploring Big Data
- Visualizing Big Data
- Processing Big Data
- Parallelizing Analysis Operations
- Creating and Evaluating Regression Models
- Creating and Evaluating Partitioning Models
- Processing Big Data in SQL Server and Hadoop

Audience
The primary audience for this course is people who wish to analyze large datasets within a big data environment.
The secondary audience are developers who need to integrate R analyses into their solutions.

Prerequisite
In addition to their professional experience, students who attend this course should have:

- Programming experience using R, and familiarity with common R packages
- Knowledge of common statistical methods and data analysis best practices.
- Basic knowledge of the Microsoft Windows operating system and its core functionality.

Duration
Three Days
MOC 20773A: Analyzing Big Data with Microsoft R

Course Outline

I. Microsoft R Server and R Client
   A. Explain how Microsoft R Server and Microsoft R Client work.
   B. What is Microsoft R server
   C. Using Microsoft R client
   D. The ScaleR functions

Lab: Exploring Microsoft R Server and Microsoft R Client
   - Using R client in VSTR and RStudio
   - Exploring ScaleR functions
   - Connecting to a remote server

II. Exploring Big Data
    At the end of this module the student will be able to use R Client with R Server to explore big data held in different data stores.
    A. Understanding ScaleR data sources
    B. Reading data into an XDF object
    C. Summarizing data in an XDF object

Lab: Exploring Big Data
   - Reading a local CSV file into an XDF file
   - Transforming data on input
   - Reading data from SQL Server into an XDF file
   - Generating summaries over the XDF data

III. Visualizing Big Data
     Explain how to visualize data by using graphs and plots.
     A. Visualizing In-memory data
     B. Visualizing big data

Lab: Visualizing data
   - Using ggplot to create a faceted plot with overlays
   - Using rxlinePlot and rxHistogram

IV. Processing Big Data
    Explain how to transform and clean big data sets.
    A. Transforming Big Data
    B. Managing datasets

Lab: Processing big data
   - Transforming big data
   - Sorting and merging big data
   - Connecting to a remote server

V. Parallelizing Analysis Operations
   Explain how to implement options for splitting analysis jobs into parallel tasks.
   Using the RxLocalParallel compute context with rxExec
   A. Using the revoPemaR package

Lab: Using rxExec and RevoPemaR to parallelize operations
   - Using rxExec to maximize resource use
   - Creating and using a PEMA class

VI. Creating and Evaluating Regression Models
    Explain how to build and evaluate regression models generated from big data
    A. Clustering Big Data
    B. Generating regression models and making predictions

Lab: Creating a linear regression model
   - Creating a cluster
   - Creating a regression model
   - Generate data for making predictions
   - Use the models to make predictions and compare the results

VII. Creating and Evaluating Partitioning Models
     Explain how to create and score partitioning models generated from big data.
     A. Creating partitioning models based on decision trees.
     B. Test partitioning models by making and comparing predictions

Lab: Creating and evaluating partitioning models
   - Splitting the dataset
   - Building models
   - Running predictions and testing the results

VIII. Processing Big Data in SQL Server and Hadoop
      Explain how to transform and clean big data sets.
      A. Using R in SQL Server
      B. Using Hadoop Map/Reduce
      C. Using Hadoop Spark