Analyzing Big Data in R using Apache Spark

About This Course

Master **Apache Spark**, a popular cluster computing framework used for performing large scale data analysis. **SparkR** provides a distributed data frame API that enables structured data processing with a syntax familiar to **R** users.

- Learn why R is a popular statistical programming language with a number of extensions that support data processing and machine learning tasks.
- Learn how SparkR, an R package that provides a light-weight frontend, uses Apache Spark from R.

Course Syllabus

- **Module 1 - Introduction to SparkR**
    1. Learn what SparkR is
    2. Understand why you would use SparkR
    3. List the features of SparkR
    4. Understand the interfaces into SparkR
- **Module 2 - Data manipulation in SparkR**
    1. Understand how to use dataframes
    2. Learn to select data
    3. Learn to filter data
    4. Learn to aggregate data
    5. Learn to operate on columns
    6. Understand how to write SQL queries
- **Lab 1 - Getting started with SparkR**
- **Lab 2 - Data manipulation in SparkR**
- **Module 3 - Machine learning in SparkR**
    1. Understand machine learning
    2. Learn how to use GLM model
- **Lab 3 - Linear models in SparkR**