APPLIED DATA SCIENCE AND BIG DATA ANALYTICS – GTBD7

Course Description

This intensive training course provides theoretical and technical aspects of Data Science and Business Analytics. The course covers the fundamental and advanced concepts and methods of deriving business insights from big" and/or "small" data. This training course is supplemented by hands-on labs that help attendees reinforce their theoretical knowledge of the learned material.

Business success in the information age is predicated on the ability of organizations to convert raw data coming from various sources into high-grade business information.

To stay competitive, organizations have started adopting new approaches to data processing and analysis. For example, data scientists are turning to Apache Spark for processing massive amounts of data using Spark's distributed compute capability along with its built-in machine learning library, or switching from proprietary and costly solutions to the free R programming language.

TOPICS

- Applied Data Science and Business Analytics
- Algorithms, Techniques and Common Analytical Methods
- Machine Learning Introduction
- Visualizing and Reporting Processed Results
- The R Programming Language
- Data Analysis with R
- Elements of Functional Programming
- Apache Spark Introduction
- Spark SQL
- ETL with Spark
- MLlib Machine Learning Library
- Graph Processing with GraphX

Target Audience

Data Scientists, Software Developers, IT Architects, and Technical Managers

Prerequisites

Participants should have the general knowledge of statistics and programming

Duration

4 Days

Course Content Summary

Chapter 1. Applied Data Science

- What is Data Science?
- Data Science Ecosystem
- Data Mining vs. Data Science
- Business Analytics vs. Data Science
- Who is a Data Scientist?
- Data Science Skill Sets Venn Diagram
- Data Scientists at Work
- Examples of Data Science Projects
- An Example of a Data Product
- Applied Data Science at Google
- Data Science Gotchas

Chapter 2. Getting Started with R

- Introduction
- Positioning of R in the Data Science Arena
- R Integrated Development Environments
- Running R
- Running RStudio
- Ending the Current R Session
- Getting Help
- Getting System Information
- General Notes on R Commands and Statements
- R Data Structures
- R Objects and Workspace
- Assignment Operators
- Assignment Example
- Arithmetic Operators
- Logical Operators
- System Date and Time
- Operations
- User-defined Functions
- User-defined Function Example
- R Code Example
- Type Conversion (Coercion)
- Control Statements
- Conditional Execution
- Repetitive Execution
- Repetitive execution

- Built-in Functions
- Reading Data from Files into Vectors
- Example of Reading Data from a File
- Writing Data to a File
- Example of Writing Data to a File
- Logical Vectors
- Character Vectors
- Matrix Data Structure
- Creating Matrices
- Working with Data Frames
- Matrices vs Data Frames
- A Data Frame Sample
- Accessing Data Cells
- Getting Info About a Data Frame
- Selecting Columns in Data Frames
- Selecting Rows in Data Frames
- Getting a Subset of a Data Frame
- Sorting (ordering) Data in Data Frames by Attribute(s)
- Applying Functions to Matrices and Data Frames
- Using the apply() Function
- Example of Using apply()
- Executing External R commands
- Loading External Scripts in RStudio
- Listing Objects in Workspace
- Removing Objects in Workspace
- Saving Your Workspace in R
- Saving Your Workspace in RStudio
- Saving Your Workspace in R GUI
- Loading Your Workspace
- Hands-on Exercises
- Getting and Setting the Working Directory
- Getting the List of Files in a Directory
- Diverting Output to a File
- Batch (Unattended) Processing
- Importing Data into R
- Exporting Data from R
- Hands-on Exercise
- Standard R Packages
- Extending R
- Extending R in R GUI
- Extending R in RStudio
- CRAN Page

Chapter 3. R Statistical Computing Features

- Statistical Computing Features
- Descriptive Statistics
- Basic Statistical Functions
- Examples of Using Basic Statistical Functions
- Non-uniformity of a Probability Distribution

- Writing Your Own skew and kurtosis Functions
- Hands-on Exercise
- Generating Normally Distributed Random Numbers
- Generating Uniformly Distributed Random Numbers
- Using the () Function
- Math Functions Used in Data Analysis
- Examples of Using Math Functions
- Correlations
- Correlation Example
- Testing Correlation Coefficient for Significance
- The cor.test() Function
- The cor.test() Example
- Regression Analysis
- Types of Regression
- Simple Linear Regression Model
- Least-Squares Method (LSM)
- LSM Assumptions
- Fitting Linear Regression Models in R
- Example of Using lm()
- Confidence Intervals for Model Parameters
- Example of Using lm() with a Data Frame
- Regression Models in Excel
- Hands-on Exercise
- Multiple Regression Analysis
- Finding the Best-Fitting Regression Model
- Comparing Regression Models
- Hands-on Exercise

Chapter 4. Data Analytics Life-cycle Phases

- Big Data Analytics Pipeline
- Data Discovery Phase
- Data Harvesting Phase
- Data Priming Phase
- Exploratory Data Analysis
- Model Planning Phase
- Model Building Phase
- Communicating the Results
- Production Roll-out

Chapter 5. Data Science Algorithms and Analytical Methods

- Supervised vs Unsupervised Machine Learning
- Supervised Machine Learning Algorithms
- Unsupervised Machine Learning Algorithms
- Choose the Right Algorithm
- Life-cycles of Machine Learning Development
- Classifying with k-Nearest Neighbors (SL)
- k-Nearest Neighbors Algorithm
- k-Nearest Neighbors Algorithm

- The Error Rate
- Hands-on Exercise
- Decision Trees (SL)
- Decision Tree Terminology
- Decision Trees in Pictures
- Decision Tree Classification in Context of Information Theory
- Information Entropy Defined
- The Shannon Entropy Formula
- The Simplified Decision Tree Algorithm
- Using Decision Trees
- Random Forests
- Naive Bayes Classifier (SL)
- Naive Bayesian Probabilistic Model in a Nutshell
- Bayes Formula
- Classification of Documents with Naive Bayes
- Unsupervised Learning Type: Clustering
- K-Means Clustering (UL)
- K-Means Clustering in a Nutshell
- Regression Analysis
- Simple Linear Regression Model
- Linear vs Non-Linear Regression
- Linear Regression Illustration
- Major Underlying Assumptions for Regression Analysis
- Least-Squares Method (LSM)
- Locally Weighted Linear Regression
- Regression Models in Excel
- Multiple Regression Analysis
- Logistic Regression
- Regression vs Classification
- Time-Series Analysis
- Decomposing Time-Series
- Monte-Carlo Simulation (Method)
- Who Uses Monte-Carlo Simulation?
- Monte-Carlo Simulation in a Nutshell
- Monte-Carlo Simulation Example
- Monte-Carlo Simulation Example
- Hands-on Exercise

Chapter 6. Visualizing and Reporting Processed Results

- Data Visualization
- Data Visualization in R
- The ggplot2 Data Visualization Package
- Creating Bar Plots in R
- Creating Horizontal Bar Plots
- Using barplot() with Matrices
- Using barplot() with Matrices Example
- Customizing Plots
- Histograms in R
- Building Histograms with hist()

- Example of using hist()
- Pie Charts in R
- Examples of using pie()
- Generic X-Y Plotting
- Examples of the plot() function
- Dot Plots in R
- Saving Your Work
- Supported Export Options
- Plots in RStudio
- Saving a Plot as an Image
- The BIRT Project
- JavaFX
- Data Visualization with JavaFX
- Visualization with D3 JavaScript Library
- Examples of D3 Visualization
- Google Charts

Chapter 7. Text Mining

- What is Text Mining?
- The Common Text Mining Tasks
- What is Natural Language Processing (NLP)?
- Some of the NLP Use Cases
- Machine Learning in Text Mining and NLP
- Machine Learning in NLP
- TF-IDF
- The Feature Hashing Trick
- Stemming
- Example of Stemming
- Stop Words
- Popular Text Mining and NLP Libraries and Packages

Chapter 8. Introduction to Functional Programming

- What is Functional Programming (FP)?
- Terminology: First-Class and Higher-Order Functions
- Terminology: Lambda vs Closure
- A Short List of Languages that Support FP
- FP with Java
- FP With JavaScript
- Imperative Programming in JavaScript
- The JavaScript map (FP) Example
- The JavaScript reduce (FP) Example
- Using reduce to Flatten an Array of Arrays (FP) Example
- The JavaScript filter (FP) Example
- Common High-Order Functions in Python
- Common High-Order Functions in Scala
- Elements of FP in R

Chapter 9. Big Data Business Intelligence and Analytics

- Traditional Business Intelligence and Analytics
- OLAP Tasks
- Data Mining Tasks
- Big Data / NoSQL Solutions
- NoSQL Data Querying and Processing
- The UnQL Specification
- MapReduce Defined
- MapReduce Explained
- Hadoop
- Hadoop-based Systems for Data Analysis
- Hadoop's Streaming MapReduce
- Streaming Use Cases
- Making things simpler with Hadoop Pig Latin
- Pig Latin Script Example
- SQL Equivalent
- What is Hive?
- Interfacing with Hive
- Hive Data Definition Language
- Business Analytics with Hive
- What is Spark?
- Amazon Elastic MapReduce
- Big Data with Google App Engine (GAE)
- GAE Dashboard
- Example of Google AppEngine Java Datastore API
- Google Cloud Prediction API

Chapter 10. Introduction to Apache Spark

- What is Spark
- A Short History of Spark
- Where to Get Spark?
- The Spark Platform
- Spark Logo
- Common Spark Use Cases
- Languages Supported by Spark
- Running Spark on a Cluster
- The Driver Process
- Spark Applications
- Spark Shell
- The spark-submit Tool
- The spark-submit Tool Configuration
- The Executor and Worker Processes
- The Spark Application Architecture
- Interfaces with Data Storage Systems
- Limitations of Hadoop's MapReduce
- Spark vs MapReduce
- Spark as an Alternative to Apache Tez
- The Resilient Distributed Dataset (RDD)

- Spark Streaming (Micro-batching)
- Spark SQL
- Example of Spark SQL
- Spark Machine Learning Library
- GraphX
- Spark vs R

Chapter 11. The Spark Shell

- The Spark Shell
- The Spark Shell UI
- Spark Shell Options
- Getting Help
- The Spark Context (sc) and SQL Context (sqlContext)
- The Shell Spark Context
- Loading Files
- Saving Files
- Basic Spark ETL Operations

Chapter 12. Spark RDDs

- The Resilient Distributed Dataset (RDD)
- Ways to Create an RDD
- Custom RDDs
- Supported Data Types
- RDD Operations
- RDDs are Immutable
- Spark Actions
- RDD Transformations
- Other RDD Operations
- Chaining RDD Operations
- RDD Lineage
- The Big Picture
- What May Go Wrong
- Checkpointing RDDs
- Local Checkpointing
- Parallelized Collections
- More on parallelize() Method
- The Pair RDD
- Where do I use Pair RDDs?
- Example of Creating a Pair RDD with Map
- Example of Creating a Pair RDD with keyBy
- Miscellaneous Pair RDD Operations
- RDD Caching
- RDD Persistence
- The Tachyon Storage

Chapter 13. Parallel Data Processing with Spark

- Running Spark on a Cluster
- Spark Stand-alone Option
- The High-Level Execution Flow in Stand-alone Spark Cluster
- Data Partitioning
- Data Partitioning Diagram
- Single Local File System RDD Partitioning
- Multiple File RDD Partitioning
- Special Cases for Small-sized Files
- Parallel Data Processing of Partitions
- Spark Application, Jobs, and Tasks
- Stages and Shuffles
- The "Big Picture"

Chapter 14. Introduction to Spark SQL

- What is Spark SQL?
- Uniform Data Access with Spark SQL
- Hive Integration
- Hive Interface
- Integration with BI Tools
- Spark SQL is No Longer Experimental Developer API!
- What is a DataFrame?
- The SQLContext Object
- The SQLContext API
- Changes Between Spark SQL 1.3 to 1.4
- Example of Spark SQL (Scala Example)
- Example of Working with a JSON File
- Example of Working with a Parquet File
- Using JDBC Sources
- JDBC Connection Example
- Performance & Scalability of Spark SQL

Chapter 15. Graph Processing with GraphX

- What is GraphX?
- Supported Languages
- Vertices and Edges
- Graph Terminology
- Example of Property Graph
- The GraphX API
- The GraphX Views
- The Triplet View
- Graph Algorithms
- Graphs and RDDs
- Constructing Graphs
- Graph Operators
- Example of Using GraphX Operators
- GraphX Performance Optimization

- The PageRank Algorithm
- GraphX Support for PageRank

Chapter 16. The Spark Machine Learning Library

- What is MLlib?
- Supported Languages
- MLlib Packages
- Dense and Sparse Vectors
- Labeled Point
- Python Example of Using the LabeledPoint Class
- LIBSVM format
- An Example of a LIBSVM File
- Loading LIBSVM Files
- Local Matrices
- Example of Creating Matrices in MLlib
- Distributed Matrices
- Example of Using a Distributed Matrix
- Classification and Regression Algorithm
- Clustering

Chapter 17. Machine Learning with BigML

- What is BigML?
- How BigML Service Works
- Data Files
- Data Sets
- Data Sets Example
- Models
- Predictions
- The Prediction UI Form
- Text Analysis in BigML
- REST API

Lab Exercises

Lab 1. Learning the Lab Environment
Lab 2. Getting Started with R
Lab 3. Working with R
Lab 4. Data Import and Export in R
Lab 5. Creating Your Own Statistical Functions
Lab 6. Simple Linear Regression
Lab 7. Multiple Linear Regression
Lab 8. k-Nearest Neighbors Algorithm
Lab 9. Monte-Carlo Simulation (Method)

Lab 10. Using R Graphics Package
Lab 11. Using the D3 JavaScript Visualization Library
Lab 12. Common Text Mining Tasks with the tm Library
Lab 13. Elements of Functional Programming with Python
Lab 14. The Spark Shell
Lab 15. RDD Performance Improvement Techniques
Lab 16. Spark ETL and HDFS Interface
Lab 17. Common Map / Reduce Programs in Spark
Lab 18. Spark SQL
Lab 19. Getting Started with GraphX
Lab 20. PageRank with GraphX
Lab 21. Using k-means Algorithm from MLlib
Lab 22. Using Random Forests for Classification with Spark MLlib
Lab 23. Text Classification with Spark ML Pipeline