Module Title : Advanced Methods in Data Science and Big Data Analytics
Duration : 5 days

Overview

This course builds on skills developed in the Data Science and Big Data Analytics course. The main focus areas cover Hadoop (including Pig, Hive, and HBase), Natural Language Processing, Social Network Analysis, Simulation, Random Forests, Multinomial Logistic Regression, and Data Visualization. Taking an "Open" or technology-neutral approach, this course utilizes several open-source tools to address big data challenges.

Audience

This course is intended for aspiring Data Scientists, data analysts that have completed the associate level Data Science and Big Data Analytics course, and computer scientists wanting to learn MapReduce and methods for analyzing unstructured data such as text.

Prerequisite Knowledge/Skills

- Completion of the Data Science and Big Data Analytics course
- Proficiency in at least one programming language such as Java or Python

Course Objectives

Upon successful completion of this course, participants should be able to:

- Develop and execute MapReduce functionality
- Gain familiarity with NoSQL databases and Hadoop Ecosystem tools for analyzing large-scale, unstructured data sets
- Develop a working knowledge of Natural Language Processing, Social Network Analysis, and Data Visualization concepts
- Use advanced quantitative methods, and apply one of them in a Hadoop environment
- Apply advanced techniques to real-world datasets in a final lab

Course Outline

The content of this course is designed to support the course objectives.

Module 1: MapReduce and Hadoop

Lesson 1: The MapReduce Framework

Lesson 2: Apache Hadoop

Lesson 3: Hadoop Distributed File System

Lesson 4: YARN

Module 2: Hadoop Ecosystem and NoSQL

Lesson 1: Hadoop Ecosystem

Lesson 2: Pig

Lesson 3: Hive

Lesson 4: NoSQL - Not Only SQL

Lesson 5: HBase

Lesson 6: Spark

Module 3: Natural Language Processing

Lesson 1: Introduction to NLP

Lesson 2: Text Preprocessing

Lesson 3: TFIDF

Lesson 4: Beyond Bag of Words

Lesson 5: Language Modeling

Lesson 6: POS Tagging and HMM

Lesson 7: Sentiment Analysis and Topic Modeling

Module 4: Social Network Analysis

Lesson 1: Introduction to SNA and Graph Theory

Lesson 2: Most Important Nodes

Lesson 3: Communities and Small World

Lesson 4: Network Problems and SNA Tools

Module 5: Data Science Theory and Methods

Lesson 1: Simulation

Lesson 2: Random Forests

Lesson 3: Multinomial Logistic Regression

Module 6: Data Visualization

Lesson 1: Perception and Visualization

Lesson 2: Visualization of Multivariate Data