Big Data Science

Course description:

This course is designed to provide students with fundamental understanding of Big Data and help develop skills necessary to handle and implement various aspects of big data projects. The course has an additional focus on Machine Leaning & Data Science developed to advance data science skills that are often required to implement big data projects. Students, while taking the course, will have a unique exposure to both data science and big data technologies that will provide them with skills necessary to implement data science projects in Hadoop as well as support teams executing big data projects.

Three areas of focus for the proposed course:

- Hadoop Architecture
- Big data programming
- Machine Learning in Big Data environment:
    - SQL to manipulate data
    - Feature preparation & Feature engineering
    - Machine Learning methods

Prerequisites and evaluation:

Students are not expected to be proficient in any of the big data languages, experience with SQL and functional programming in form of Python, R and Matlab should be helpful.

Given the wide range of programming languages deployed in big data, students will be provided with an option to choose two languages they wish to use to implement assignments and final project. One language will be used for scripting and preparing the data, another language will used to implement a ML algorithm in a cluster environment. It should be noted that the focus of the course is not on coding skills but rather on fundamental understanding of what languages are most suitable for particular data science tasks. Students will be marked based on their ability to implement a data science solution rather than eloquence of their code.

Course outline with deliverables and dates

Lecture #1 (May 6th):
Introduction to Hadoop

1. Definition of big data
2. Big data industry review
3. History of Hadoop and distributed computing
4. Brief overview of HDFS, Yarn, MapReduce, Spark

5. Review of the big data resources that will be used for this course
6. Final project: objective and requirements
7. Review of additional resources to help with big data skills

Lecture 2 (May 8th):
Enterprise Hadoop and Hadoop basics

1. HDFS, YARN in detail
2. Launch of the cluster environment and practice of HDFS and YARN commands

3. Invited lecture by an industry practitioner responsible for establishing an enterprise Hadoop architecture: Melissa Singh
    a. Implementation of a big data project: review of a real business case with focus on understanding of the complexity of interactions and convergence of DevOps, legal, data management and data science teams

Lecture 3 (May 13th):
SQL: Hive and Impala

1. Programming languages: overview, object-oriented vs functional, scripting languages
2. Big data languages: overview of Spark, Hive & Impala
3. SQL: most important commands
4. Hive vs Impala: comparison between two SQL languages and environments
5. Assignment 1: data mining using Hive

Lecture 4 (May 15th):
Spark, introduction

1. Introduction into Spark
    a. History of Spark, its evolution, libraries
    b. Review of Spark APIs: PySpark, Scala, R and Java
    c. Spark backend transformations
    d. Practice sessions with DataBricks, Scala

Lecture 5 (May 22nd):
Spark, syntax

1. Spark session
2. Import/export of data
3. Spark SQL: functional and scripting
    a. 1 student paper review: Spark SQL original paper
4. Working with SQL images
5. Most important Scala functions
6. Assignment #2 (Scala)

Lecture 6 (May 27th):
Spark: Data Preparation and Feature Transformation for ML

1. Feature preparation: overview of main steps (in Spark)
    a. Handling nulls/NaNs, missing values, outliers, normalization

1. Spark libraries for feature transformation
2. Feature transformation: dimensionality reduction, clustering, conversion from categorical to numerical
3. Principle component analysis
4. Regularization: L1, L2

Lecture 7 (May 29th):
Spark, model development, pipelines, evaluation metrics

1. Model development process and necessary steps
2. Spark pipelines
3. TrainTestSplit and CrossValidation
4. Regularization: L1, L2
5. Model evaluation metrics for different classes of models

Lecture 8 (June 3rd):
Spark ML: Multivariate analysis and time-series analysis

- Multivariate ML models in Spark: regression trees
- Time-series analysis
- Invited speaker: Rogelio Cuevas from TD Securities (H2O?)

Lecture 9 (June 10th):
Advanced data science topics

- Graph theory
- Deep learning libraries

Lecture 10 (June 12th):
Cloud for Big Data, two invited speakers

1. Invited lecture by Microsoft's Azure Deep Learning team lead Dr. Faraz Rashid (confirmed)

1. Cloud for big data:
    a. Invited lecture by Anuj (Cloud architect at TD)

Week (June 22nd):

1. Final exam

Week (June 24th):

1. Final project presentations by student teams

Marking scheme:
Assignments, Exams and Grading:

- Assignment 1 – Hive/Impala: 15%
- Assignment 2 - Spark: 15%
- Final Exam – theory and understanding of big data: 35%

- Final Project: 55%
    - Final Project Presentation: 15%
    - Final project report including analysis and code: 20%
