MET CS777 A1 (Spring 2022) - Big Data Analytics

Instructor

Suresh Kalathur, Ph.D.
Assistant Professor, Computer Science Dept.
Boston Univeristy Metropolitan College
1010 Commonwealth Ave, Room 304
Boston, MA 02215

Email: kalathur@bu.edu
URL: http://kalathur.com/bu
Phone: 617-358-0006
Fax: 617-353-2367

Course Description

This course is an introduction to large-scale data analytics. Big Data analytics is the study of how to extract actionable, non-trivial knowledge from massive amount of data sets. This class will focus both on the cluster computing software tools and programming techniques used by data scientists, as well as the important mathematical and statistical models that are used in learning from large-scale data processing. On the tools side, we will cover the basics systems and techniques to store large-volumes of data, as well as modern systems for cluster computing based on Map-Reduce pattern such as Hadoop MapReduce, Apache Spark and Flink. Students will implement data mining algorithms and execute them on real cloud systems like Amazon AWS, Google Cloud or Microsoft Azure by using educational accounts. On the data mining models side, this course will cover the main standard supervised and unsupervised models and will introduce improvement techniques on the model side.

Course Prerequisites

MET CS 521, MET CS 544 and MET CS 555. Or, MET CS 677. Or, Instructor's consent

Course Grading Policy

The course grade will be based on assignments (40%), mid term exam (30%), and a term project (30%). Assignments are expected to be submitted by their respective due dates.

Course Web Site

- **https://learn.bu.edu**

References

Coming soon...

Student Conduct Code

Please review the academic conduct code

Tentative Course Schedule

- Module1 -- Overview
    - Python & Pandas Review
    - Overview of Big Data Analytics
    - MapReduce paradigm
    - Introduction to Apache Spark & pySpark
    - RDDs (Resilient Distributed Datasets)
- Module2 -- Large-Scale Data Processing
    - Structured data processing of big datasets
    - Datasets and DataFrames
    - Spark SQL Overview
    - Aggregation, Grouping and Join operations
    - Data formats (JSON, Parquet, etc)
    - Stream Processing
- Module 3 -- Machine Learning for Big Data
    - Spark MLlib framework
    - Pipelines, ETL
    - Classification & Regression
    - Clustering
    - Frequent Pattern mining
    - Model selection and hyperparameter tuning
- Module 4 -- Streaming Analytics
    - Event-driven applications
    - Stream & Batch analytics
- Module 5 -- Machine Learning for Streaming Data
    - Scikit-Multiflow framework
    - Concept drift detection in data streams
    - Supervised Learning for streaming data
    - Unsupervised Learning for streaming data
- Module 6 -- Text Analytics and NLP
    - Latent Semantic Indexing
    - Topic modeling
    - Latent Dirichlet Allocation
- Mid Term Exam - March 24th
- Term Project Presentations - April 28th