Instructor Name:
Weijia Xu

Course Name:
Large Scale Data Analysis with Hadoop and Spark

Course Description: This course will introduce attendee how to use two most popular big data processing frameworks, Hadoop and Spark, for big data analysis tasks. The course will introduce basic system architecture and core components of each system in order to give beginner a clear picture on basics of the two systems. The course will feature clear instructions and a test system access for attendee to get started on using those systems from day 1. The course will give a grand tour of the data analysis capability to show how common data analysis needs for large data can be met with those platforms. Useful libraries and existing tools will also be introduced including Mahout, MLlib, GrpahX and SparkSQL. Those tools and libraries include a set of implementations of a wide range of analysis algorithms. Finally, the course will also introduce components and applications that enable utilization of the Hadoop and Spark through other programming language and interface including Hadoop Streaming, Spark-Shell and Hive. The course materials will include exemplar problems, hands-on exercises and demonstrations.

Day 1: Getting ready with big data analysis
Introduction to MapReduce Programming model
  Basic concepts
  Why MapReduce is good for big data problem.
Overview of Hadoop Cluster
  Basic components,
  System overview
  Hadoop distributed file system.
Getting started with hadoop cluster
  Key requirements of hadoop
How to run it on your own laptop
How to access Hadoop cluster support at TACC.

**Day 2: Common big data analysis tasks in Hadoop**
Association analysis
  - Goal of the analysis task
  - Common algorithms
  - How it is supported in big data
  - Things to consider when conducting large scale association analysis
Classification analysis
  - Goal of the analysis task
  - Common algorithms
  - How it is supported in big data
  - Things to consider when conducting large scale classification analysis

**Day 3: Introduction on Spark programming model**
(Brief daily outline or expectations.)
Introduction to Spark Programming model
  - Why we need Spark,
  - Spark vs. Hadoop,
  - How to get started with Spark
Data analysis support with Spark,
  - MLlib, a library for machine learning needs.
  - GraphX, network graph analysis support.

**Day 4: Using Hadoop and Spark without Java programming**
Interfacing Hadoop and Spark with other programming language.
  - Hadoop Streaming
  - Spark-shell and PySpark
  - Spark and R
Big data analysis with SQL using Hive
  - Hive introduction,
  - Example workflow of using Hive,
  - Exemplar problems and demonstrations.