DA512 - Big Data Processing using Hadoop

Course Description

This course will provide the essential background to start developing programs that will run on Hadoop Distributed File System (HDFS). The course will also show the students the limitations of traditional programming techniques and how Hadoop addresses these problems. After learning the basics of a Hadoop Cluster and Hadoop Ecosystem, students will learn to write programs using Apache Spark framework and run these programs on a Hadoop Cluster.

Topics Covered

- Introduction to Hadoop File system
- Introduction to MapReduce Framework
- MapReduce vs Spark
- Introduction to Spark
- Spark Basics
- Working with RDDs
- Running Spark Applicatons on Hadoop
- Paralel Programming with Spark
- Writing Spark Applications
- Caching and Persistence
- Spark Streaming

Instructor

Ahmet Demirelli
Office : FENS L025
E-mail : ahmetdemirelli@sabanciuniv.edu
Phone : x9516
Web : http://myweb.sabanciuniv.edu/ahmetdemirelli

Reading List

http://hadoop.apache.org
http://hadoop.apache.org/docs/current/hadoop-yarn/hadoop-yarn-site/YARN.html
http://www.bthaber.com/nedir-bu-hadoop
http://en.wikipedia.org/wiki/Apache_Hadoop
http://radar.oreilly.com/2012/02/what-is-apache-hadoop.html
http://readwrite.com/2013/05/23/hadoop-what-it-is-and-how-it-works
http://www.slideshare.net/hortonworks/apache-hadoop-yarn-understanding-the-data-operating-system-of-hadoop

Grading

Homework 1	%15
Homework 2	%15
Midterm	%30
Final	%40