BLM5127- Big Data Analytics Syllabus

Instructor	Name:	Asst. Prof. Serkan Ayvaz
	Tel:	(212) 383 5780
	Email:	sayvaz@yildiz.edu.tr
	Website:	https://avesis.yildiz.edu.tr/17212
	Office:	D-220

Course Hours	Tues	12:00--14:50PM
Office Hours	Tues	10:00--12:00PM or by appointment

Text Books

- **B1:** Hadoop: The Definitive Guide, *Tom White.2012.ISBN-13: 978-1449311520*
- **B2:** Big Data SMACK: A Guide to Apache Spark, Mesos, Akka, Cassandra, and Kafka. *Isaac Ruiz and Raul Estrada. 2016.*
- **B3:** Big Data Analytics with Spark A Practitioner's Guide to Using Spark for Large Scale Data Analysis. Mohammed Guller
- **B4:** Data-Intensive Text Processing with MapReduce, Jimmy Lin and Chris Dyer

Reference Books

MapReduce Design Patterns, by Donald Miner and Adam Shook

O'Reilly Media. ISBN: 978-1-4493-2717-0

Learning Spark, by Holden Karau, Andy Konwinsky, Patrick Wendell, Matei Zaharia. O'Reilly Media. ISBN: 978-1-4493-5862-4

Big Data Science & Analytics: A Hands-On Approach. Bahga, A. and Madisetti, V., 2016.

Course Description

Big data analytics is the process of examining massive amounts of ever-growing data to discover hidden patterns and useful insights. The goal of this course is to provide the students with the fundamental concepts and methods of big data analytics and to help them learn big data analytics approaches and technologies applied to big data solutions.

Upon completion of the course, students will:

1. Understand the architectural components and programming models used for scalable big data analytics.
2. Learn the fundamentals of the distributed file systems and the MapReduce programming model.
3. Be able to apply data analytics methods in big data environments.
4. Learn of concepts of distributed data storage and NoSQL databases.

Course Format

The lectures will include readings tutorials. The PowerPoint presentations/class notes will also be available on the website following each class. During class meetings, there will be some lecturing. You will tackle different Big Data analysis problems individually.

Prerequisites: The course assumes that the students have the understanding of basic knowledge of programming. However, previous experience with Hadoop, Spark or distributed computing is NOT required.

Participation: Participation will be a part of your course grade. You will be responsible for recovering any information you have missed if you do not attend a particular day's lecture.

Exams: The exams will be 90-120 minutes, closed book, closed notes exam. The use of any reference material is strictly forbidden.

Project: There will be one project for the course. It can be a group project or an individual project. For group project, the students are expected to provide more comprehensive work. For group project, all study group members will receive the same score. In general, students may freely communicate within their group, but you may not discuss group work with members of other groups. Details of this assignment will be discussed later in the semester.

Students are responsible for forming and managing their study groups. We expect that students will manage their study groups so that everyone performs a fair share of the work, and that all perspectives are heard and considered. The assignments should be turned in electronically by end of the due date.

Correspondence: If you want to get response to your e-mails, always include your name, your course name. Observe grammatical rules while composing your e-mails.

Grading

Final course grades will be based on:

Homework Assignment	10%
Project	30%
Midterm Examination	20%
Final Examination	40%

No late assignments will be accepted. No make-up will be administered for the midterm.

TENTATIVE SCHEDULE

WEEK 1	**October /5**
Due:	
Topics to be Covered:	Introduction General Introduction

WEEK 2	**October /12**
Due:	B1 Chapter 1
Topics to be Covered:	Hadoop Ecosystem Fundamentals Preliminary discussions of Hadoop /Hadoop Installation

WEEK 3	**October /19**
Due:	B1 Chapter 2
Topics to be Covered:	Hadoop architecture and HDFS Definining Hadoop Distributed File System

WEEK 4	**October /26**
Due:	B1 Chapter 6-7-8
Topics to be Covered:	Introduction to MapReduce MapReduce Programming

WEEK 5	**November /2**
Due:	B1 Chapter 10
Topics to be Covered:	Hadoop administration Configuring, Deploying, and Maintaining a Hadoop Cluster

WEEK 6	**November /9**
Due:	B1 Chapter 20, B2 Chapter 5
Topics to be Covered:	NoSQL Databases and distributed data storage Cassandra system and architecture

WEEK 7	**November /16**
Due:	B1 Chapter 19, B2 Chapter 6
Topics to be Covered:	Introduction to Apache Spark In-Memory Computation with Spark MIDTERM EXAM

WEEK 8	**November /23**
Due:	
Topics to be Covered:	MIDTERM EXAM

WEEK 9	**November /30**
Due:	B2 Chapter 3
Topics to be	In-Memory Computation and RDDs

Covered:	Hands-on experience

WEEK 10	**December /7**
Due:	B2 Chapter 5
Topics to be Covered:	PySpark Parallel Programming

WEEK 11	**December /14**
Due:	B2 Chapter 3, B3 Chapter 2
Topics to be Covered:	Parallel Programming with Scala

WEEK 12	**December/21**
Due:	B3 Chapter 8
Topics to be Covered:	Machine Learning with Spark

WEEK 13	**December /28**
Due:	B2 Chapter 9, B3 Chapter 6
Topics to be Covered:	Fast Data and Stream Processing with Spark

WEEK 14	**January /4**
Due:	
Topics to be Covered:	PRESENTATIONS

WEEK 15	**January 11-18**
Due:	
Topics to be Covered:	FINAL